Advances in Experimental Medicine and Biology 1361

Alessandro Laganà *Editor*

Computational Methods for Precision Oncology

Springer

Advances in Experimental Medicine and Biology

Volume 1361

Series Editors

Wim E. Crusio, Institut de Neurosciences Cognitives et Intégratives d'Aquitaine, CNRS and University of Bordeaux, Pessac Cedex, France

Haidong Dong, Departments of Urology and Immunology, Mayo Clinic, Rochester, MN, USA

Heinfried H. Radeke, Institute of Pharmacology & Toxicology, Clinic of the Goethe University Frankfurt Main, Frankfurt am Main, Hessen, Germany

Nima Rezaei, Research Center for Immunodeficiencies, Children's Medical Center, Tehran University of Medical Sciences, Tehran, Iran

Ortrud Steinlein, Institute of Human Genetics, LMU University Hospital, Munich, Germany

Junjie Xiao, Cardiac Regeneration and Ageing Lab, Institute of Cardiovascular Science, School of Life Science, Shanghai University, Shanghai, China

Advances in Experimental Medicine and Biology provides a platform for scientific contributions in the main disciplines of the biomedicine and the life sciences. This series publishes thematic volumes on contemporary research in the areas of microbiology, immunology, neurosciences, biochemistry, biomedical engineering, genetics, physiology, and cancer research. Covering emerging topics and techniques in basic and clinical science, it brings together clinicians and researchers from various fields.

Advances in Experimental Medicine and Biology has been publishing exceptional works in the field for over 40 years, and is indexed in SCOPUS, Medline (PubMed), Journal Citation Reports/Science Edition, Science Citation Index Expanded (SciSearch, Web of Science), EMBASE, BIOSIS, Reaxys, EMBiology, the Chemical Abstracts Service (CAS), and Pathway Studio.

2020 Impact Factor: 2.622

More information about this series at https://link.springer.com/bookseries/5584

Alessandro Laganà
Editor

Computational Methods for Precision Oncology

Springer

Editor
Alessandro Laganà
Department of Genetics and Genomic Sciences
Department of Oncological Sciences
Mount Sinai Icahn School of Medicine
New York, NY, USA

ISSN 0065-2598 ISSN 2214-8019 (electronic)
Advances in Experimental Medicine and Biology
ISBN 978-3-030-91835-4 ISBN 978-3-030-91836-1 (eBook)
https://doi.org/10.1007/978-3-030-91836-1

© Springer Nature Switzerland AG 2022
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland


Medical
Records

Preface

Cancer encompasses a large number of complex malignancies characterized by extensive clonal heterogeneity and variations in the microenvironment that influence response to treatment. Recent advances in sequencing technologies have enabled the generation of large amount of data, which, in turn, has prompted the development of multi-omics integrative computational tools aimed at dissecting the biology of individual tumors. This has given rise to the new field of precision oncology, a fast-evolving multidisciplinary research field whose goal is to leverage the genomic and molecular features of the individual patient's cancer cells and microenvironment to efficiently inform therapy selection. The key idea underpinning precision oncology's paradigm is that treatment may be given based on the specific alterations observed in the patient, rather than on the tumor histology or tissue type. Recent studies and clinical trials have shown that genomic profiling benefits patients whose cancer is driven by specific targetable alterations. However, the widespread heterogeneity of cancer genomes and drug responses still poses a significant challenge in the design of effective personalized treatments. Responses to therapies are often short-lived, and many patients lack well-defined genetic aberrations targetable by currently available drugs. The goal of precision oncology is to integrate different low- and high-throughput technologies and data modalities to capture the genetic, molecular, and cellular complexity of the patient's tumor and microenvironment, identify dysregulated mechanisms and potential vulnerabilities, and design and prioritize treatments accordingly.

This volume provides a comprehensive state-of-the-art overview of the computational approaches, methodologies, and tools that enable precision oncology, as well as the most promising and exciting innovations that are advancing the field and are likely to be incorporated into clinical practice soon.

The first two chapters cover the basic concepts in computational precision oncology, providing a comprehensive overview of the architecture and components of a precision oncology platform, the data analyzed, and the insights that can be gained from data analysis. In particular, Chap. 1 introduces the fundamental concepts in multi-omics data analysis for precision oncology and illustrates the typical architecture of a precision oncology platform, including both basic and advanced components, while Chap. 2 focuses on the more technical aspects of software infrastructure and workflow implementation.

Chapters 3, 4, 5, 6, 7, 8, 9, and 10 dive into each component of a precision oncology workflow and describe them in terms of their objectives, the related biological background, and the most efficient solutions, along with their strengths and weaknesses. Chapters 3 and 4 present current workflows and algorithms for calling single-nucleotide variations (SNV) and copy number alterations (CNA), respectively, from next-generation sequencing (NGS) data, and discuss their merits and limitations. Chapter 5 introduces the problem of assessing microsatellite instability (MSI), its relevance to precision oncology applications, and the methods and tools available for determining MSI status from NGS data in cancer. Chapter 6 introduces the challenge of dissecting intra-tumor heterogeneity in cancer samples and provides a state-of-the-art review of the statistical and computational approaches and tools for the analysis of NGS data to infer the clonal landscape of a tumor and its temporal and spatial evolution, in the context of precision oncology applications. Chapter 7 covers data resources and computational tools for drug repurposing, which consists in finding novel uses for existing drugs. The chapter presents and discusses several categories of methods, including target-based, knowledge-based, signature-based, and pathway-based methods, as well as the data resources specialized in gene expression changes induced by drugs in cell lines and that are used in drug repurposing applications. Chapter 8 focuses on the analysis of pathways for precision oncology applications. It provides a comprehensive description of cancer omics projects and data sources, a survey of the main biological pathway databases, and a detailed review of the most used pathway analysis tools, discussed in the context of their potential applications in precision oncology. Chapter 9 introduces gene fusions, which are a hallmark of several cancer types, and discusses the strategies for their identification from patient's RNA-Seq data, along with the tools for their annotation and visualization. Chapter 10 provides a detailed overview of knowledge bases and tools for the prioritization and interpretation of variants in cancer, which are essential tasks in a precision oncology platform.

Chapters 11, 12, 13, 14, 15, and 16 present advanced cutting-edge research topics that represent promising future directions in the field of precision oncology. Chapter 11 provides an introduction to network-based approaches used to integrate multi-modal data sources for patient stratification and classification, and discusses challenges and opportunities for the application of these approaches in precision oncology. Chapter 12 describes patient-derived models of cancer, such as organoids, chorioallontoic membranes, tumor slice cultures, microfluidic platforms, and xenograft models, and discusses their potential implementation in clinical settings to inform novel therapeutic options. Chapter 13 covers liquid biopsies, an advanced testing technology allowing non-invasive detection of biomarkers for cancer screening and real-time monitoring of disease progression. The chapter provides a comprehensive overview of the methods to perform molecular profiling of circulating tumor cells, cell-free DNA, and extra-cellular vesicles, reviews the ongoing clinical trials for liquid biopsies, and discusses the future directions for their full clinical implementation. Chapter 14 surveys data resources and several applications of artificial intelligence (AI) in cancer research and precision

oncology, from cancer subtype identification to drug prioritization and image analysis, and introduces the first diagnostic FDA-approved AI-powered tools. Chapter 15 reviews the state-of-the-art of single-cell DNA and RNA sequencing technologies and analysis tools, and discusses their potential uses in precision oncology, including the dissection of intra-tumor heterogeneity, multi-omics data integration for the full characterization of a tumor, and the implementation of targeted drug repurposing approaches. Finally, Chap. 16 discusses the importance of profiling the tumor immune microenvironment. While current precision oncology applications focus on the dissection of tumor genomics and transcriptomics to identify actionable alterations, research has shown that investigating the immune microenvironment is fundamental to understand cancer initiation, progression, and response to therapy. Chapter 16 summarizes the technologies and computational methodologies available to study the microenvironment and discusses the implications for the identification of effective treatments for patients.

Assembling and curating this volume has been an inspiring and stimulating experience that has allowed me to connect and collaborate with valuable colleagues, and to widen my knowledge and perspectives in this exciting field. I am thankful for having been given this unique opportunity and hopeful that this volume will be of help to many researchers who are approaching precision oncology and its related areas.

I would like to express my gratitude to all the authors for the effort they put in this project during this challenging and uncertain time of global pandemic, and for the valuable content they contributed. My sincere gratitude also goes to the Springer team, particularly to Associate Editor Larissa Albright and Project Coordinator Shabib Shaikh for their guidance and support throughout the development of this book.

New York, NY, USA Alessandro Laganà

Acknowledgments

The front matter illustration was drawn by Dr. Francesco Russo, from the Section for Clinical Mass Spectrometry, Danish Center for Neonatal Screening, Department of Congenital Disorders at Statens Serum Institut in Copenhagen, Denmark. The figure represents the data integration process to obtain a better understanding of diseases and find personalized treatments.

Contents

1 The Architecture of a Precision Oncology Platform 1
Alessandro Laganà

2 Software Workflows and Infrastructures for Precision Oncology 23
Waleed Osman and Alessandro Laganà

3 Somatic and Germline Variant Calling from Next-Generation Sequencing Data 37
Ti-Cheng Chang, Ke Xu, Zhongshan Cheng, and Gang Wu

4 Identification of Copy Number Alterations from Next-Generation Sequencing Data 55
Sheida Nabavi and Fatima Zare

5 Assessment of Microsatellite Instability from Next-Generation Sequencing Data 75
Victor Renault, Emmanuel Tubacher, and Alexandre How-Kit

6 Computational Approaches for the Investigation of Intra-tumor Heterogeneity and Clonal Evolution from Bulk Sequencing Data in Precision Oncology Applications 101
Alessandro Laganà

7 Computational Methods for Drug Repurposing 119
Rosaria Valentina Rapicavoli, Salvatore Alaimo, Alfredo Ferro, and Alfredo Pulvirenti

8 Pathway Analysis for Cancer Research and Precision Oncology Applications 143
Alessandro La Ferlita, Salvatore Alaimo, Alfredo Ferro, and Alfredo Pulvirenti

9 RNA-seq Fusion Detection in Clinical Oncology 163
Dale J. Hedges

10 Computational Resources for the Interpretation of Variations in Cancer 177
Grete Francesca Privitera, Salvatore Alaimo, Alfredo Ferro, and Alfredo Pulvirenti

11 Network Approaches for Precision Oncology 199
Shraddha Pai

12 Patient-Derived In Vitro and In Vivo Models of Cancer 215
Sally E. Claridge, Julie-Ann Cavallo, and Benjamin D. Hopkins

13 Molecular Profiling of Liquid Biopsies for Precision Oncology 235
Edgar E. Gonzalez-Kozlova

14 Artificial Intelligence for Precision Oncology 249
Sherry Bhalla and Alessandro Laganà

15 Single-Cell Sequencing Technologies in Precision Oncology 269
David T. Melnekoff and Alessandro Laganà

16 Multi-Omics Profiling of the Tumor Microenvironment 283
Oliver Van Oekelen and Alessandro Laganà

Index 327

The original version of the book frontmatter was revised: An acknowledgment has been added.

The Architecture of a Precision Oncology Platform

1

Alessandro Laganà

Abstract

Precision oncology is a novel research field and approach to cancer care which leverages high-throughput sequencing technologies and bioinformatics pipelines to determine diagnosis, prognosis, and treatment of patients in a personalized manner. This chapter provides an overview of a typical precision oncology software platform, from raw data to patient reports. Standard and advanced analytical components are described and discussed, along with their strengths and limitations, in general and in the context of a precision oncology application for advanced cancer patients.

Introduction

Precision oncology is a novel area of biomedicine, a specialization of precision medicine aimed at determining diagnosis, prognosis, and therapy in cancer patients in a highly personalized manner, based on the specific characteristics of the individual tumor rather than on the cancer type. A key feature of precision oncology is the actionable alteration, that is, a genetic or molecular alteration that can be directly targeted by a drug or a biomarker that indicates sensitivity to a specific drug.

The development of highly accurate and affordable high-throughput sequencing technologies in the past decade has fueled significant progress in precision oncology, which is now gaining momentum and increasingly being incorporated into mainstream clinical practices.

The two main elements enabling precision oncology are the technologies to produce the data and the computational pipelines to analyze the data. DNA and RNA sequencing allow to detect pathological variations in the genome and the transcriptome of cancer cells, which are then used to determine the cancer subtype, assess risk, and design a specific therapy. Naturally, our knowledge of the mechanisms driving cancer progression is still incomplete, and only a fraction of patients benefits from this approach today. Nevertheless, the translation of research findings into actionable strategies is now faster than ever, thus promoting unprecedented progress in the development of novel therapies and their personalized applications.

This chapter focuses on the software architecture of a precision oncology platform, from patient data to reports. Figure 1.1 illustrates the general schema of a precision oncology pipeline, including both standard and advanced components. The implementation of such pipelines is

A. Laganà (✉)
Department of Genetics and Genomic Sciences, Department of Oncological Sciences, Mount Sinai Icahn School of Medicine, New York, NY, USA
e-mail: alessandro.lagana@mssm.edu

© Springer Nature Switzerland AG 2022

A. Laganà (ed.), *Computational Methods for Precision Oncology*, Advances in Experimental Medicine and Biology 1361, https://doi.org/10.1007/978-3-030-91836-1_1

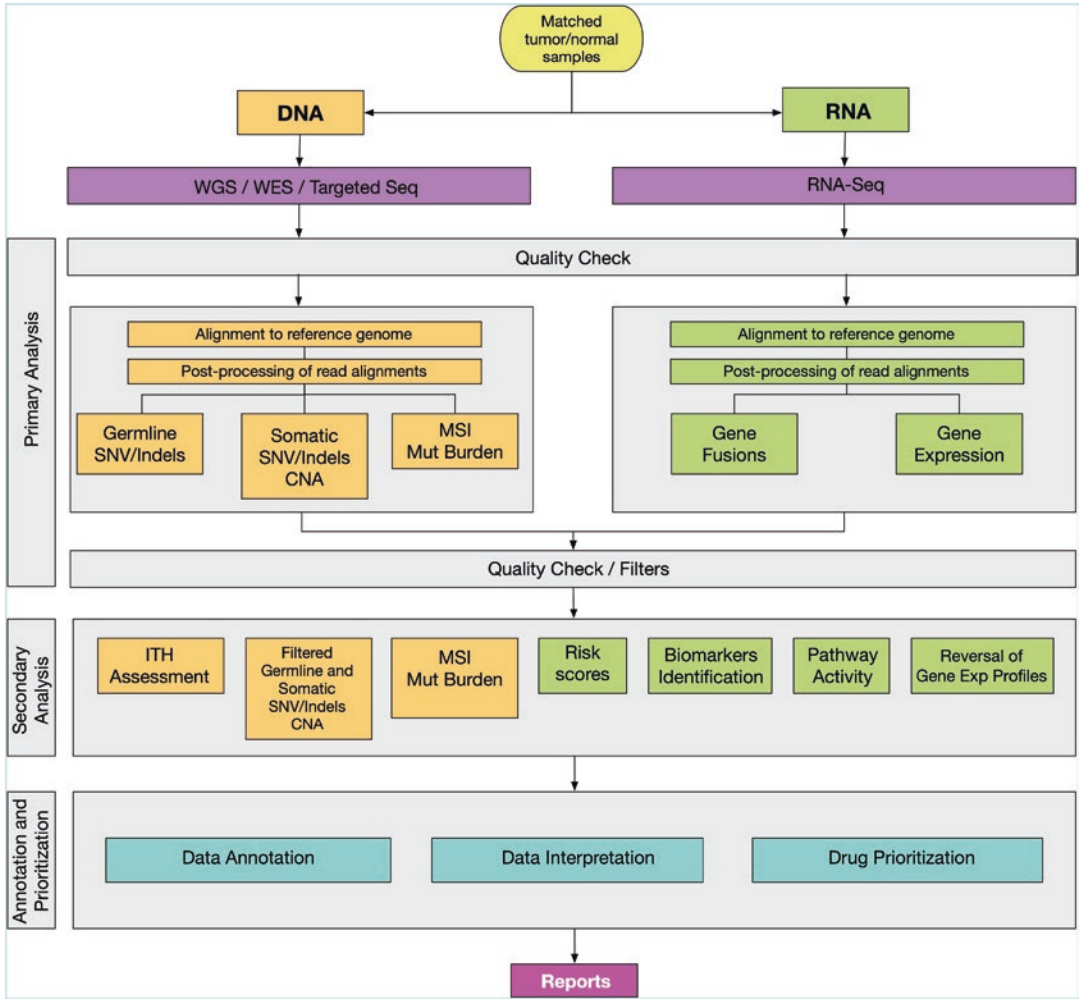

Fig. 1.1 Schema of a precision oncology platform. A typical precision oncology platform is composed of three macro-components: primary analysis, secondary analysis, and annotation/prioritization. After matched normal and tumor samples from a patient have been subjected to DNA and RNA sequencing, DNA primary analysis extracts primary data, that is, germline and somatic mutations (SNVs), copy number alterations (CNAs), and genomic metrics such as microsatellite instability (MSI) and mutational burden, through alignment and variant calling. Similarly, RNA primary analysis performs alignment and extracts primary data such as gene expression profile and gene fusions. Primary data is then assessed for quality and filtered, for example, variants are filtered based on their allele frequency (VAF), or gene expression is filtered to remove lowly expressed genes. Secondary alignment consists of several independent tasks to extract a second layer of more abstract/complex information from primary data, such as subclonal landscape/intratumor heterogeneity (ITH), risk scores and biomarkers, pathway activity measures, and drug associations based on reversal of gene expression profiles (drug repurposing). Finally, all the produced data are annotated with external databases to assess their pathogenicity and actionability, then the potential associated therapeutic options are prioritized based on the level of clinical evidence, cancer type, and other metrics. A summary report of the main alterations and drug recommendation is then produced for the physician and the patient

facilitated by special data management platforms, like Arvados and Terra.bio, and workflows management systems, like CWL and Nextflow [1, 2], which provide high-level systems and tools for defining and running standard and custom bioinformatics analyses. These pipelines start with raw sequencing data, usually from matched tumor and normal samples, which are then pro-

cessed to identify genetic alterations like somatic mutations and copy number variations, quantify gene transcripts and pathway activity, and assess genomic instability and intratumor heterogeneity. All the findings are then annotated and evaluated for pathogenicity and actionability, and reports with a summary of the findings along with possible drug recommendation are generated for the physicians and the patients.

The next sections of this chapter will describe these steps and components and provide general guidance for their implementation and interpretation. An example of a precision oncology platform is then given, along with the results obtained in a pilot clinical trial with relapsed/refractory cancer patients.

Data Sources and Assays

Precision oncology platforms are typically based on high-throughput targeted sequencing, whole-genome sequencing (WGS) or whole-exome sequencing (WES) data, which is often complemented with transcriptomic data from RNA-Seq as well as clinical data [3–5]. DNA sequencing is usually performed on matched tumor and normal tissue (e.g., saliva or peripheral blood) and allows to identify somatic single nucleotide variations (SNVs) and short indels as well as larger structural changes including broad and focal amplifications, deletions, and chromosomal rearrangements. When normal tissue is not present, the data from the tumor is compared to the reference genome. This, however, can lead to inaccuracies as it is not possible to discriminate between truly somatic alterations and germline single nucleotide polymorphisms (SNPs), although databases with known SNPs may be used to mitigate this problem [6, 7].

Several variables should be considered when choosing the most appropriate assay for a precision oncology platform, such as turn-around time, cost, sequencing depth, coverage, and clinical questions. Sequencing depth is the average number of reads covering each DNA base, while coverage refers to the width of the genome area sequenced. Targeted panels are the most popular and widely used assays in oncological applications, because of fast turn-around time, which is typically less than a week, and high depth, usually greater than 500×. The latter, particularly, increases confidence in the discovered variants, especially when they are present in a small fraction of tumor cells. This variable is called variant allele frequency (VAF) and is defined as the number of tumor reads carrying a specific variant (alternate allele) divided by the total number of reads covering the locus. Targeted panels usually cover a few hundred loci including mutational hotspots, drivers, and genes that are frequently altered in cancer and carry prognostic impact and/or therapeutic actionability. Popular options are the Foundation Medicine panels and MSKCC-Impact, among others [8, 9]. It is also possible to develop custom panels targeting specific cancer types and genes of interests.

In recent years, the use of WES for clinical applications has also gained popularity [5]. WES covers the coding regions of the genome, the exome, which is where the majority of known deleterious and actionable variations occur, and constitutes ~2% of the human genome. Such regions are captured through hybridization of genomic DNA to biotinylated oligonucleotide probes (baits) complementary to targeted exons [10]. Sequencing depth of WES typically ranges between 100× and 150× for the tumor, which allows discovery of most variants with VAF of 10% and above, while normal samples are typically sequenced at lower depth [11]. One important reason to prefer WES over targeted panels is the extended coverage to most coding regions, which allows the discovery of variants of likely pathogenic significance beyond the genes included in the panels. This may be especially important in applications that couple clinical care with research [12].

WGS provides the widest coverage of the genome, albeit at higher cost and lower depth, which is typically in the range of 30× to 50×, depending on the application [13–15]. WGS is mostly used in research, since it allows to explore variations in the non-coding portion of the genome, which may help elucidate mechanisms of cancer pathogenesis and progression, for

example, in the promoters. However, there are clinical applications that may benefit from WGS as it allows the identification of large structural variations, such as chromosomal translocations, and a better characterization of copy number alterations than WES [16, 17]. Previous studies have suggested that WGS-based translocation calls may be more accurate than FISH/Cytogenetics assays [18, 19]; thus, it is reasonable to expect WGS to gain more popularity in the near future, especially considering the potential therapeutic actionability of structural alterations which cannot be confidently discovered with WES.

While DNA sequencing is essential for the discovery of actionable variations in the genome, RNA sequencing (RNA-Seq) is fundamental to study the variations in the transcriptome and quantify gene expression. The analysis of gene expression has been the focus of thousands of studies in the past 20 years and has been enabled by microarrays before RNA-Seq. Most gene expression studies in oncology have allowed the discovery of biomarker of sensitivity and resistance to specific drugs, to elucidate the mechanisms of action of drugs and their effect on signaling and metabolic pathways, to investigate the cellular response to different conditions (e.g., hypoxia), and to dissect the role of oncogenes and tumor suppressors and their genetic variations [20–22]. RNA-Seq is not yet widely used in precision oncology, but some platforms use it to identify chimeric transcript that results from gene fusions, which is currently the most common application of this technology in precision oncology [23]. However, recent studies suggest that there are other advantages to incorporating RNA-Seq in clinical care applications [24]. A pilot study that we have conducted at Mount Sinai, NY, in 2017, has demonstrated that RNA-Seq can help identify viable therapeutic options beyond what can be detected through DNA sequencing [25]. More specifically, RNA-Seq can identify actionable altered pathways and measure biomarkers of drug sensitivity.

The next sections of this chapter will provide an overview of the different standard and advanced components of a precision oncology platform, from DNA and RNA sequencing analysis workflows and methods to prioritization of findings and generation of clinical reports.

Analysis of DNA Sequencing Data

This section provides an overview of the steps and methods involved in processing DNA sequencing data, from raw reads to variants. These steps generally work for both targeted and WGS/WES data, but a distinction will be made where necessary. Most of the pipelines discussed here refer to the Broad Institute Best Practices Workflows (BroadBPW), which provide useful step-by-step recommendations for performing variant discovery analysis in high-throughput sequencing (HTS) data [26, 27] (https://gatk.broadinstitute.org/hc/en-us/sections/360007226651-Best-Practices-Workflows). However, when designing a precision oncology platform, it is always important to consider the specific technologies used for data generation along with the goals of downstream analysis.

Pre-processing of Sequencing Data

The output of sequencers consists of raw reads organized in text files in the FASTQ format, where each read is annotated with its quality score. It is always advisable to perform quality check (QC) on these files, to ensure that the downstream analysis produces reliable data and high-confidence calls [28]. Library preparation and sequencing can, indeed, introduce biases, errors, and contamination, which, in turn, can affect variant identification and lead to inaccurate results. Several tools for quality check and data pre-processing have been developed in the past years, aimed at providing comprehensive quality profiles including basic statistics such as total number of reads and their length, GC content, per-base and per-sequence quality scores, as well as more sophisticated metrics such as sequence duplication levels, overrepresented sequences,

and k-mer content. FastQC is a popular and widely used QC option [29]. In some cases, reads need to be trimmed to remove adapter sequences and low-quality bases. These tasks can be performed using tools such as Cutadapt and Trimmomatic [30, 31]. The more recent Fastp conveniently incorporates both QC and reads trimming as well as advanced filtering features, including base correction in the overlapping fragments of paired reads, and provides before- and after-processing reports [32]. Other tools can additionally be used to estimate tumor purity, which is the proportion of cancer cells in the sample and, therefore, provides an estimate of possible contamination with normal cells [33]. Several tools for the analysis of copy number alterations, which are introduced in section "Copy Number Alteration Calling", provide estimates of tumor purity. Finally, another crucial aspect in data pre-processing is to make sure that the matched tumor and normal samples analyzed are indeed from the same individual. Sample mismatches, in fact, can happen at different steps of the experimental and data analysis pipeline, and tools such as NGS Checkmate can help validating sample identity and detect such mismatches [34].

Data QC and pre-processing are crucial steps and can help identify and mitigate problems in downstream analyses. Problematic samples can be flagged as such, so that the results of the analysis are interpreted with caution, or discarded altogether, when not of sufficient quality for reliable conclusions to be drawn.

Reads Mapping

The steps in this and next sections can be applied to any type of DNA sequencing data, whether it is WGS, WES, or targeted. Once QC has been performed and the reads have been pre-processed and filtered, the next step is to map each individual read or read pair to the reference genome (e.g., human GRCh38/hg38), which is a representative common genome sequence of the analyzed species in string format, to correctly identify their origins [35, 36]. This process, which is referred to as *alignment*, is carried out by specific tools such as Bowtie, BWA, and GMAP/GSNAP, which implement efficient methods to index the reference genome and deal with mapping ambiguities and handling mismatches and indels [37–39]. BWA consists of three methods based on the Burrows-Wheeler Transform (BWT) algorithm, which enables counting the number of exact hits of a string in the genome independently of the size of the genome. While BWA-backtrack is specifically designed for Illumina sequence reads up to 100 bp, BWA-MEM and BWA-SW support long reads and split alignment. BWA-MEM is the most efficient and accurate option and is recommended in several workflows, including BroadBPW. GSNAP is another suitable option, albeit slower than BWA, which organizes the output in different files that include uniquely mapped reads, multiply mapped reads, and unmapped reads. All read aligners use multi-threading to independently process multiple reads simultaneously and, therefore, significantly reduce computing time. Alignments are usually outputted in the text-based SAM format (Sequence Alignment Map) or its compressed version called BAM (Binary Alignment Map).

Post-processing of Read Alignments

Read alignments are further processed to mitigate biases introduced by library preparation steps, such as PCR amplification, and recalibrate the base quality score. This step involves tools such as Picard, the Genome Analysis ToolKit (GATK), and Samtools [26, 40]. Here is an example of a typical post-processing protocol based on BroadBPW: (1) ensure that all mate-pair information is consistent between each read and its mate pair (tool: Picard-FixMateInformation), (2) match the contig ordering in the reference genome file (tool: Picard-ReorderSam), (3) soft-clip beyond-end-of-reference alignments (tool: Picard-CleanSam), (4) identify and tag duplicate reads, that is, reads originating from a single

fragment of DNA (tool: Picard-MarkDuplicates), (5) build the index of the alignment files (tool: Picard-BuildBamIndex, (6) perform local realignment in order to minimize the mismatches introduce by indels (tools: GATK-RealignerTargetCreator and GATK:IndelRealigner), (7) adjust base quality scores (tools: GATK-BaseRecalibrator and GATK-PrintReads). The output of this workflow produces alignment files in the BAM format that are ready to be analyzed for variant discovery.

Somatic Single Nucleotide Variants (SNVs) and Short Indels Calling

This task can be performed on either targeted sequencing, WES, or WGS data. However, the typical read depth provided by WGS may not be sufficient to enable the discovery of low frequency variants. Once the BAM files have been filtered and cleaned up, they are ready to be processed for variant discovery [41]. This is the fundamental step where reads from a tumor and a matched normal sample from the same individual are compared to the reference genome to identify somatic single nucleotide variations (SNVs) and short indels, that is, short insertion or deletion of bases, which are then typically outputted in VCF (Variant Call Format) text files. Somatic variants are found in the tumor but neither in the normal control nor in the reference genomes. Thus, they are more likely to have an impact in the oncogenic process. Several tools have been developed in the past decade to address this task. MuTect2, Strelka2, VarDict, and Lancet are among the most popular and widely used options [42–45]. Providing sequencing data from both the tumor and normal cells is crucial to ensure the discovery of true somatic variants. While some tools (e.g., MuTect2) are designed to identify variants in tumors even when a matched normal sample is not available, they may produce false positives at higher rates; thus, the results should be interpreted with caution [46]. MuTect2 has a tumor-only mode which compares a tumor sample with an unmatched panel of normals (PoN) [42, 47]. This modality relies on multiple samples from healthy tissues, for example, obtained from public databases such as the 1000 Genomes Project, which should be processed as similarly as possible to the tumor, for example, in terms of library preparation method and sequencing technology, in order to minimize artifacts and technical noise. Although there is no definitive rule for how many samples should be used to create a PoN, Broad Institute guidelines suggest using at least 40 samples (https://gatk.broadinstitute.org/hc/en-us/articles/360035890631-Panel-of-Normals-PON-). A study evaluating several tools for variant discovery in non-matched sequencing samples using real and simulated data has shown that no tool was able to call all the mutations, indicating that further improvements are necessary to ensure reliable calls [48].

Because of the different algorithms, statistical models, and filtering strategies that each tool for variant calling implements, the identified variants may be significantly different. Studies comparing different callers using both real and synthetic data have shown various degrees of agreement between them, where many calls were reported only by one or two tools [49, 50]. Since there is no definitive answer to which caller is the most accurate and reliable in all cases, a good rule of thumb is to employ a consensus calling strategy, where several tools are interrogated and the union or (weighted) overlap of their calls considered for downstream analysis [51]. For example, our approach to somatic variant calling in multiple myeloma uses MuTect2, Strelka2, and Lancet and considers as true somatic calls those produced by at least two of the three tools. This mitigates both the false negative problem, where a true variant may be missed by a tool, and the false positive problem, where a false variant may be called by a single tool. It is worth pointing out that false negatives may be the greatest challenge, as false positives may still be discovered and filtered out in downstream analyses. Somatic variant calling is covered in greater details in Chap. 3 of this volume.

Copy Number Alteration Calling

Somatic copy number alterations (CNAs) are changes in the structure of somatic cells that result in the gain or loss in copies of chromosome segments and are a common feature of cancer genomes [6, 7]. Some events may involve a whole chromosome or one of its arms and are termed broad CNAs, others are restricted to smaller fragments and are termed focal CNAs. It is common to refer to single-copy alterations as *gain* and *loss*, while *amplification* and *deletion* are used for multi-copy gain and loss of both copies (homozygous) of a chromosome fragment. CNAs can play a critical role in cancer pathogenesis and progression as they may activate oncogenes and inactivate tumor suppressors [52]. The most obvious outcome of a change in the number of copies of a gene is a corresponding change in gene expression. Several studies have demonstrated that variations in copy numbers significantly correlate with differential gene expression [53], which may help to distinguish between driver and passenger CNAs. While SNV/indel calling is most accurate when performed on WES and targeted data, CNA calling is typically performed on WGS data, as such alterations are often broad and span both coding and non-coding areas. Tools such as Battenberg have been designed for this specific task [54]. However, it is still possible to perform CNA calling on WES data and special targeted panels designed to cover common chromosome breakpoints. Facets is one such tool designed to identify CNAs in WES data [55]. Chapter 4 provides an in-depth review of the methods and tools available for the identification of CNA in sequencing data.

Microsatellite and Genomic Instability

Microsatellite instability (MSI) is a genetic condition where microsatellites, which are short regions of repeated DNA interspersed throughout the genome, accumulate mutations as a result of a deficiency in the DNA mismatch repair (MMR) system. MMR corrects errors that may spontaneously occur during DNA replication, and its impairment may lead to widespread mutagenesis and neoplastic development [56]. In recent years, MSI has been shown to be a major predictor of response to immune blockade therapy, and the Food and Drug Administration (FDA) has approved immune checkpoint inhibitors for the treatment of solid tumors with high MSI, such as pembrolizumab, regardless of their type [57]. Thus, MSI is now included in precision oncology platforms as it may help the oncologist to identify patients that may benefit from immune checkpoint inhibitors. Several computational methods, such as MSIsensor and MSI-seq, have been developed to detect MSI from WGS and WES data [58, 59]. Some tools are based on mutation burden as a measure of MSI status, while others compare the percentage of unstable microsatellites in a tumor compared to a matched normal sample. As for variant calling, the input of these tools are post-processed BAM files. The output is usually a score that indicates if the tumor is MSI-high or MSI-low (or stable). For example, a pan-cancer study on >15,000 solid cancers reliably inferred MSI status using the tool MSIsensor and found that MSI-high was predictive of Lynch syndrome-associated cancer predisposition [60]. More details on MSI and computational approaches for its assessment in sequencing data are given in Chap. 5.

Germline Variant Calling in Precision Oncology

Germline variants are changes that occur in reproductive cells (i.e., egg or sperm) and, therefore, can be passed from parent to offspring, where it is incorporated into the DNA of every cell of the body. Some germline mutations may cause hereditary cancer syndromes, which predispose the carriers to certain types of cancers. For example, mutations in the BRCA1 or BRCA2 genes are often associated with hereditary breast and ovarian cancer, while mutations affecting the DNA mismatch repair mechanism may cause Lynch syndrome, which increases the risk of developing colorectal cancer. It is estimated that

between 5% and 10% of all cancers are hereditary. While precision oncology platforms are mostly focused on somatic variants, which are the driver alterations in most cancers, they can also incorporate the analysis of control samples to identify variants associated with cancer predisposition. Like for the detection of somatic mutations, there is a BroadBPW for the discovery of germline short variants [27]. The processing pipeline includes the same steps described in sections "Pre-processing of Sequencing Data" and "Post-processing of Read Alignments" for somatic variant calling, such as reads mapping, duplicate marking, and base quality score recalibration. Then, GATK HaplotypeCaller is used to identify variants in the sample, which are outputted in a GVCF (Genomic VCF) file. Multiple GVCFs from different patients can then be consolidated and processed by GenotypeGVCFs, which performs cohort joint variant calling. This step is highly recommended because it improves the accuracy and sensitivity of variant detection, particularly at low-coverage or low-quality sites, by leveraging population-level information. The resulting raw VCF file is then filtered using different tools in GATK, such as VariantRecalibrator, to remove variants that are likely to be false positive. These tools make use of machine learning algorithms that are trained on large high-quality datasets of known variants and then applied to identify variants that are likely to be real in the target sample. The filtered VCF is then ready for annotation, evaluation, and downstream analyses. Other popular tools for germline variant discovery are Freebayes, VarScan, and DeepVariant [61–63].

Intratumor Heterogeneity

Inter-tumor heterogeneity, that is the genetic and phenotypic variations observed in different patients affected by the same cancer, is a well-established factor which constitutes the basis for tumor subtyping and patient classification [64–66]. For example, there are four main subtypes of breast cancer, each characterized by specific molecular alterations and level of aggressiveness: luminal A, luminal B, triple-negative/basal-like, and HER2-enriched. This explains how different tumors respond to therapy and poses a significant challenge for the discovery of reliable prognostic and therapeutic biomarkers as well as the design of clinical trials. However, another level of heterogeneity, observed within the individual tumor and, therefore, called intratumor heterogeneity (ITH), is a consequence of tumor evolution and introduces further challenges related to resistance to therapy [67]. Each tumor, indeed, consists of heterogeneous cell populations resulting from selection and expansion of clones and subclones carrying specific mutations, which confer them selective growth advantage, often following exposure to drugs [68]. Subclonal cell populations may develop from a primary tumor clone either in a linear or branching fashion, that is, by acquiring additional alterations sequentially or in parallel. Thus, they can be represented in a phylogenetic tree which describes their relationships in the context of tumor evolution. Several computational tools to define the tumor clonal landscape and evolution using WES data have been developed, based on the observation that the SNVs affecting a tumor clone and its descendants, have correlated variant allele frequency (VAF), which can then be used to define the clones through clustering techniques and probabilistic modeling [69–71]. It has been reported that the clonal landscape of a tumor may change dramatically following therapy, where different subclonal populations resistant to the specific therapy expand and cause relapse and disease progression. Therefore, assessing clonality before and after a treatment may help to identify better drug options based on the alterations driving each subclonal population. This is a challenging task, since it is not always easy to identify drivers of clonal and subclonal expansion and there may not be drugs which are known to be active against specific driver lesions. Nevertheless, characterizing the clonal landscape of a tumor can provide meaningful information that can inform prognosis and therapy design. Tools such as PyClone, SciClone, and PhyloWGS employ probabilistic modeling to infer the clonal landscape that best explains the observed mutations [72–74]. While

some tools additionally overlay CNAs on copy number neutral SNVs (e.g., SciClone), others integrate them in the inference process, as their presence can alter the VAF of mutations and, therefore, needs to be considered for accurate subclonal reconstruction (e.g., PhyloWGS and QuantumClone) [75]. Moreover, several tools are designed to utilize multiple samples from the same patient, separated either temporally (e.g., different time points during treatment) or spatially (e.g., primary tumor and metastasis), which can improve the accuracy of subclonal reconstruction and provide important information about tumor evolution [76–79]. The typical output of an ITH analysis tool consists of a list of SNVs and/or CNAs in the samples of interest annotated with the subclones they belong to. In addition, they can provide useful plots to visualize the different clonal and subclonal cell populations and the tumor phylogenetic tree. The problem of dissecting ITH from bulk sequencing samples and the methods and tools available to solve it are described in greater detail in Chap. 6.

Analysis of RNA Sequencing Data

Precision oncology has mostly focused so far on the identification of actionable genetic alterations, thus relying on DNA sequencing analysis. Recently, RNA-Seq has emerged as a promising complementary tool for clinical decision-making, as proven in several studies. One typical application of RNA-Seq analysis in precision oncology is the identification of chimeric transcripts arising from gene fusions, hybrid genes created from the fusion of two separate genes formed as a product of chromosomal rearrangements. Gene fusions can be functional and encode chimeric proteins with oncogenic potential [80]. For example, *BCR-ABL*, a gene fusion found in most patients with chronic myelogenous leukemia (CML), is created from the translocation between the long arms of chromosomes 9 and 22 t(9; 22) (known as the Philadelphia chromosome) which merges the 5' part of the BCR gene, normally located on chromosome 22, with the 3' part of the ABL1 gene, located on chromosome 9, and is translated into a hybrid protein with constitutive kinase and oncogenic activity. The BCR-ABL chimeric protein is also a prime example of actionable alteration, since it is the target of the tyrosine kinase inhibitors imatinib and nilotinib [81]. Other emerging applications of RNA-Seq in precision oncology include the identification of specific prognostic and therapeutic gene expression biomarkers and the calculation of pathway activity. This section will provide an overview of the steps and methods involved in processing RNA-Seq data. Since many steps are virtually the same as for DNA analysis, only the RNA-specific steps will be discussed, while the previous sections will be referenced for the common steps.

Pre-processing, Mapping, and Filtering of RNA-Seq Data

RNA-Seq raw data is organized in text files in the FASTQ format, exactly like DNA sequencing data. The steps for QC, trimming, and filtering described in section "Pre-processing of Sequencing Data" are applicable to RNA-Seq as well. While the goal of RNA reads alignment is to identify the genomic location where these reads originated from, like in the case of DNA data, this task is performed using tools specifically designed to handle RNA-Seq data, as it is complicated by the non-contiguous nature of RNA transcripts resulting from splicing. A fast an accurate tool widely used for the alignment of RNA-Seq data is STAR [82]. STAR uses gene structure annotations provided in the gene transfer format (GTF) to extract known splice junctions and build spliced sequences by deleting intron sequences. This significantly improves the mapping of canonical spliced reads. However, STAR is also capable of discovering non-canonical splices and chimeric transcripts. STAR has several basic and advanced parameters that can be fine-tuned to optimize its speed and accuracy [83]. Like in the case of DNA mapping, RNA-Seq alignments are outputted as BAM files, which are usually sorted and indexed using tools such as Samtools [40].

Quantification and Normalization of Gene Expression

Once reads have been aligned, the next step in the analysis of RNA-Seq data is to quantify gene expression. The raw counts are obtained by assigning the reads originating from a specific genomic location to the overlapping gene, providing an estimate of the transcript abundance for each gene. Tools such as featureCounts and HT-Seq perform this task [84, 85]. Since genes may be expressed as different isoforms due to alternative splicing events, it is also possible to use this information to quantify the abundance of specific transcripts, rather than assigning all the overlapping reads to a unique gene entity. While featureCounts and HT-Seq return transcript abundance as the raw number of reads, that is, counts, which is a format accepted by many tools for RNA-Seq analysis such as DESeq2 for differential expression [86], it is also common to perform a within-sample normalization and use units such as Reads Per Kilobase per Million (RPKM)/Fragments per Kilobase per Million (FPKM) and Transcripts Per Million (TPM). These three units are similar to one another and account for both gene/transcript length and sequencing depth, allowing to compare features of different length.

It has been shown that different alignment and quantification strategies may result in inaccurate gene expression estimates, where the expression of a gene may be over- or underestimated, for example, in the case of reads overlapping multiple genes or when features share high sequence similarity [87]. If a read is mapped to more than one gene, tools such as HT-Seq and featureCounts will discard it. However, alternative approaches that virtually use all the sequenced reads have been proposed, such as mmquant, which assigns multi-mapping reads to groups of genes instead [88].

Finally, alignment-free approaches for accurate and fast transcript quantification, such as pseudoalignment and quasi-mapping, could be used when the objective is to quantify transcript abundance. Pseudoalignment, for example, which is implemented in the tool Kallisto, consists in matching the reads to the transcriptome, rather than the entire genome, and using specific data structures (e.g., a De Bruijn graph) to optimize the search for the set of transcripts compatible with a given read [89]. Other tools implementing rapid quantification approaches include Sailfish and Salmon [90, 91].

Gene Fusion Identification

The identification of gene fusions resulting from chromosomal translocations such as the Philadelphia chromosome can be effectively performed by the analysis of the transcriptome, which reflects these genetic abnormalities. The chimeric transcripts originating from gene fusions can be detected by either mapping the reads to the genome capturing discordant read pairs and chimeric alignments or by performing de novo RNA-seq assembly of the transcripts and then identifying the chimeric alignments. Numerous tools have been developed implementing these approaches, many of which are described by Haas and colleagues in a comprehensive review and benchmarking study [92]. This study showed that read mapping-based approaches, and particularly the tools STAR-Fusion, Arriba, and STAR-SEQR, had the best performances overall in terms of both accuracy and speed, while de novo assembly-based tools were unable to achieve the same sensitivity in discovering fusions in cancer transcriptomes and were more successful in other applications instead, such as reconstruction of tumor viruses [93–95]. A comprehensive review of tools for gene fusion detection in RNA-Seq data is also given in Chap. 9 of this volume.

Gene Expression Analysis and Biomarker Identification

One of the most common applications of gene expression analysis, whether based on microarray or RNA-Seq, is the identification of prognostic and therapeutic biomarkers. Thousands of scientific articles have been published in the past

two decades reporting gene expression signatures and machine learning models discriminating between responders and non-responders to specific therapies and stratifying patients into prognostic classes [96–101]. Such signatures and models can be easily incorporated into a precision oncology platform for risk assessment and predict response to drugs. Naturally, this process depends on the specific signatures and whether they are transferable to the case in analysis. Gene expression signatures are typically determined through differential expression and Cox regression analyses. Such signatures may discriminate between cancer and normal tissues, identify patients in specific disease stages, or identify responders to a specific therapy [98, 102]. Specific diagnostic, prognostic, and response scores can then be derived from these signatures and implemented in precision oncology applications.

Single-Sample Approaches to Gene Expression Analysis and N-of-1 Studies

While the traditional gene expression analysis performed on groups of tumor and control samples can inform precision oncology applications by enabling the discovery of prognostic and therapeutic biomarkers and general mechanisms of dysregulation, the individual nature of a precision oncology analysis require single-sample and N-of-1 based approaches [103]. In fact, although cohort-level analyses benefit from statistical assessment of the observed changes, thus reducing the chances of false positives and artifacts, their findings are not always generalizable and genes that are perturbed on average in specific conditions may not be perturbed at all in the individual patients. To overcome such limitations and enable gene expression studies at the individual level, novel approaches have been recently developed, which mitigate lack of statistical power and exploit the variation observed within a sample or in multiple samples from the same patient. One obvious approach would be to obtain at least three replicates from the same individual for each of the conditions tested, that is, tumor and normal tissue. This would allow to employ statistical tests to assess significance of the differentially expressed genes. However, such approach may not be feasible, due to limited tissue availability, or cost-effective. One alternative approach for individual-level analysis of gene expression relies on the relative ranking of gene expression within a sample, which has been shown to be robust to batch effect and normalization. The tool RankComp, and the more recent method PenDA (Personalized Differential Analysis), are based on the observation that the relative ordering of gene expression across normal tissue samples is much more stable than in diseased tissue [104, 105]. RankComp uses accumulated gene expression data from normal samples, which may be obtained from different sources, to identify statistically stable gene pairs, that is, genes which have the same order relationship ($g1<g2$ or $g1>g2$), in terms of expression, in most of the samples. Next, reversal gene pairs are identified in the individual tumor sample, that is, genes whose ranking within the sample is reversed compared to the cohort of normal samples. A statistical test is then employed to determine if a given gene g is significantly up- or downregulated in the tumor sample based on the number of genes whose expression is lower or higher than g in the normal and tumor samples. A more recent method called iDEG (individualized Differentially Expressed Genes) bypasses the individual-sample limitation by modeling read counts for each gene in the normal and tumor samples based on other genes with similar baseline expression and applying a localized version of the variance-stabilizing transformation used in cohort-based gene expression analysis in different windows of genes with similar expression at the baseline [106]. A different approach was implemented in the tool PePPeR (PErsonliazed Perturbation ProfilER), which constructs personalized perturbation profiles that reflect expression changes within a single subject [107]. Genes whose expression level in an individual sample is far from the range of values observed in a panel of control samples are defined as perturbed in the individual.

Reversal of Gene Expression Profiles

In the past decade, a novel drug repurposing paradigm based on the analysis of the transcriptome has been proposed and demonstrated effective in several studies. This approach leverages libraries of drug-induced gene expression profiles in cell lines to identify those drugs whose profile is the inverse of that of an individual patient. More specifically, given a set of up- and downregulated genes identified in a specific patient's tumor, this methodology search for drugs inducing gene expression changes that are opposite to those observed in the patient, that is, genes that are upregulated by the drug are downregulated in the patient and vice versa [108]. The assumption behind this approach is that such drugs may reverse the gene expression changes induced by the disease state. Several studies have applied this approach to determine novel drug candidates for various cancers, including lung, renal, and colorectal cancer. The Connectivity Map (CMap) project and its next generation called L1000 established "a comprehensive catalog of cellular signatures representing systematic perturbation with genetic and pharmacologic perturbagens," which can then be leveraged to perform gene expression reversal drug repurposing using the Reverse Gene Expression Score (RGES), which is a measure of potency to reverse disease gene expression [109, 110]. A recent study found that the RGES positively correlates with the half-maximal inhibitory concentration (IC50), which is a measure used to estimate drug efficacy in vitro [108]. The study determined four compounds with the potential of reversing gene expression in liver cancer and validated them in cell lines. They further validated one of the drugs, *pyrvinium pamoate*, in vivo in a xenograft model. We recently incorporated this approach to drug repurposing in our pipeline for precision medicine of multiple myeloma, where potential drug candidates were identified using the tool L1000CDS [2], a fast L1000 search engine based on characteristic direction method [25, 111].

Pathway Analysis

Turning lists of mutated or differentially expressed genes into meaningful biological insights can be a complex task, which computational functional analysis seeks to address using a variety of algorithms and statistical approaches. The aim of such analysis is to leverage knowledge bases of curated gene-set collections to assess the impact of the observed genetic alterations and/or gene expression changes on the activity of molecular and cellular pathways and understand the functional, clinical, and therapeutic implications of such events [112]. For example, the dysregulation of cell cycle is a hallmark of cancer cells and can be effectively assessed by evaluating the changes in the expression of cell cycle regulator genes, which are described and summarized in curated gene sets. Such gene sets may also include the prognostic and drug response-related signatures discussed in sections "Gene Expression Analysis and Biomarker Identification" and "Reversal of Gene Expression Profiles". Numerous computational tools have been developed to perform functional analysis of gene sets, the majority of which have been designed specifically for gene expression studies. However, several tools have also been recently proposed which focus on genomic rather than gene expression changes or which combine both types of measurements. Over-representation analysis (ORA) and functional class scoring (FCS) are the most popular classes of tools for functional analysis [113, 114]. The ORA approach aims at assessing the over-representation of a list of genes, for example, genes differentially expressed in a group of patients, in a list of gene sets, for example, genes involved in specific pathways. ORA tools typically implement the hypergeometric and Fisher's exact tests. FCS methods represent an improvement on ORA, as they consider the full gene expression profiles rather than a list of DE genes previously identified. This allows to identify sets of functionally related genes whose expression may not change significantly individually but whose coordinated variation has significant impact on a specific

pathway. Both ORA and FCS approaches have been developed for patient cohort analysis, although in principle pure ORA tools can be applied to DE genes from an individual patient. A further refinement of functional enrichment tools has been specifically developed to perform single-sample analysis. Given a gene expression matrix from multiple samples, these tools calculate an enrichment score for each patient and gene-set pair, thus allowing to perform downstream analysis such as unsupervised clustering, comparing a tumor with a normal sample or multiple tumor samples with one another at the gene-set or pathway level, rather than at the individual gene level. In this sense, this approach may also be thought of as a dimensional reduction technique, where a large gene expression matrix is replaced by a smaller one focused on gene sets. Several tools have been developed implementing different solutions for this task. Single-sample GSEA (ssGSEA) and Gene Set Variation Analysis (GSVA) perform a relative enrichment of pathways across the sample space by evaluating whether each gene is highly or lowly expressed in each sample in the context of the sample population distribution [115, 116]. Other more recent tools, such as Singscore and MixEnrich, implement true single-sample methods, where the enrichment scores calculated for each tumor sample do not depend on the other samples analyzed. Singscore is a rank-based method which calculates a score corresponding to the relative mean percentile rank of the analyzed gene sets within each sample [117, 118]. The ideal application of Singscore is to assess the enrichment in a single sample for curated gene signatures of up- and downregulated genes. Thus, if the goal of the analysis is to determine the activation of a specific pathway, it is important to first determine the list of dysregulated genes, along with their directions, which are expected to be observed when the pathway is activated. MixEnrich implements a mixture model clustering of transcripts, that is, a mixture of the distributions of dysregulated vs. unaltered mRNAs, followed by an enrichment analysis [119]. While MixEnrich is robust against bidirectional dysregulation, since genes within a pathway that are dysregulated in both directions contribute additively to the over-representation of such genes, it can only identify significantly dysregulated pathways, where the direction of the dysregulation, that is, activation or inhibition, is not given. To truly quantify and characterize the impact of dysregulated genes on relevant pathways, topology-based (TB) methods have been developed, which take into account not only the list of genes in a pathway but the structure of the pathway as well, including the type of interactions and dependencies between the genes in the pathway [120, 121]. Most TB tools available so far are cohort-based and calculate activation and inhibition of pathways in each provided sample based on the gene expression distribution in the patient population, for example, from the lists of DE genes between tumors and normal controls. They can be a suitable choice when both tumor and normal samples for multiple patient samples are available. Recently, a single-sample TB tool for pathway enrichment has been developed specifically for precision medicine applications. PerPAS quantifies pathway activity at the individual sample level by quantifying gene contribution to a pathway and calculating a personalized pathway activity score [122]. However, to calculate this score, PerPAS standardizes gene expression of a tumor sample to the mean and standard deviation of a group of control samples, if available, or to a cohort of tumor samples, basically measuring gene expression difference between the tumor and the mean of all the other tumor samples in the cohort. For each pathway, gene contribution is quantified based on topology measures, such as betweenness centrality and hubness. For example, genes with high betweenness centrality, that is, bottleneck genes, are considered to have high contribution to a pathway. The personalized pathway activity score for a given sample s and pathway p is then calculated by summing the contribution of each gene in the sample to the pathway by considering its standardized expression in the sample.

Finally, PROGENy is a recent tool which leverages a large database of publicly available perturbation experiments to define a core of pathway responsive genes [123]. The authors of PROGENy have collected gene expression data

from perturbation experiments related to 10 pathways that are relevant to cancer, such as MAPK, Hypoxia, and JAK-STAT, and extracted signatures corresponding to the activated state of these pathways. Thus, rather than measuring the genes that define each pathway, PROGENy calculates a sample-specific enrichment for downstream signatures of pathway activation. The authors have shown that PROGENy can accurately infer pathway activity based on gene expression and effectively recover the effect of known driver mutations on pathway activation. The current limitation of PROGENy is the small number of pathways available, which may not be sufficient depending on the cancer analyzed.

Annotation, Interpretation, and Prioritization of Actionable Findings

Variant Annotation

Once DNA variants, for example, SNVs and CNAs, have been identified, the next step is to annotate them with information on their potential oncogenic impact. This is a complex task that leverages data from multiple sources such as databases of variations observed in cancer and other diseases (e.g., COSMIC, ClinVar), in the general population (e.g., gnomAD) and in GWAS studies [124–127]. This helps to determine which mutations are likely pathogenic, for example, recurring in the same or other cancers, or likely benign, for example, observed frequently in healthy individuals. Additionally, several tools have been developed to predict the impact of a mutation based on the specific amino acid substitution and whether it is likely to affect the function of the protein. SIFT, PolyPhen, MutationAssessor, CADD, and FATHMM are popular tools for this task [128–133]. In general, variants that change the coding sequence of a gene and the corresponding amino acid sequence (e.g., missense mutations) are the most obvious oncogenic candidates. Other relevant variants include nonsense mutations, which result in a premature stop codon and in a truncated, likely non-functional, protein, and splice site mutations, which may alter the coding sequence by extending or reducing exons. Other genetic variants may also have deleterious effects on the corresponding proteins. For example, mutations in the 3' untranslated region (UTR) may affect post-transcriptional regulation by microRNAs (miRNAs) and RNA-binding proteins (RBPs), but the functional implications of such events are not easy to determine; therefore, they are not typically flagged as deleterious. Naturally, such decisions depend on the task at hand. While investigating the impact of non-coding or passenger mutations and their role in the oncogenic process can yield significant novel insights into the pathogenesis of cancer and advance the field, the goal of a precision oncology platform is to support clinical decision-making. Therefore, it is essential to remove as much noise as possible and focus on what is most likely deleterious and actionable, that is, supported by convincing evidence. The annotation task is facilitated by tools like Annovar, ClassifyCNV, Oncotator, and Funcotator, which conveniently include several data sources and impact-assessment tools [134–136]. Tools for assessing and annotating variants are described in greater detail in Chap. 10 of this volume.

Variant Interpretation

The crucial task at the end of a precision oncology workflow is to connect the annotated variants with actionable clinical data retrieved from the literature and clinical trials. Actionable variants are those associated with sensitivity to one or more drugs in one or more cancer types. For example, a phase 3 randomized clinical trial in patients with metastatic melanoma showed that the V600E variant in the gene BRAF predicts sensitivity to the BRAF inhibitor vemurafenib [137]. This type of information is essential to provide a clinically meaningful interpretation of all the alterations detected in a patient and represents the ultimate output of a precision oncology pipeline, which guides the physician's decision on what potential therapeutic options might ben-

efit the patient. Actionable clinical data is available from several public sources as manually curated and scored variant-drug associations based on their level of evidence. Recently, the Variant Interpretation for Cancer Consortium (VICC), a project of the Global Alliance for Genomics and Health, has developed a harmonized meta-knowledgebase (VICC KB) of clinical interpretation of somatic genomic variants in cancer [138]. The VICC KB is curated from six different independent data sources of clinically relevant evidence associated with genomic variation in cancers: Clinical Interpretation of Variants in Cancer (CIViC), the Jackson Laboratory Clinical Knowledgebase (JAX CKB), MolecularMatch, the Memorial Sloan-Kettering OncoKB, the Weill-Cornell Precision Medicine Knowledgebase (PMKB), and the Cancer Genome Interpreter (CGI) [139–143]. The initial release of VICC KB v.0.10 contained 12,856 aggregate interpretations supported by 4354 unique publications, for a total of 3439 variants. Surprisingly, over 70% of variants were described by only one of the six KBs, with less than 10% described in at least three. The authors reported that this lack of overlap was in part due to the different forms used to identify the same alterations and that their harmonization method improved consensus across the different sources and increased findings of clinical significance. Each entry in the VICC KB corresponds to a specific interpretation. For example, at the time of this writing a search for the BRAF V600E mutation returned 682 different entries. Each one of these entries correspond to a record from one of the six data original sources, providing details on the associated drug, the type of association, that is, sensitivity or resistance, the level of evidence, the cancer type and a list with the relevant publications supporting the association, along with a summary extracted from these publications. The type of alterations documented includes SNVs, CNAs, and differentially expressed genes. In the case of BRAF V600E, for example, the first hit is an entry from the JAX CKB describing a level A association supporting sensitivity to the MEK inhibitor Trametinib in patients with melanoma, based on a publication by Flaherty et al. from 2012 [144]. Another entry from JAX CKB, instead, describes a level A association supporting resistance to the EGFR inhibitor cetuximab in patients with colon carcinoma, based on guidelines from the National Comprehensive Cancer Network (NCCN). Entries are annotated using four different levels of evidence, A to D. Level A is the strongest and indicates "evidence from professional guidelines or FDA-approved therapies relating to a biomarker and disease." Level B indicates "evidence from clinical trials or other well-powered studies in clinical populations, with expert consensus." Level C, instead, corresponds to "Evidence for therapeutic predictive markers from case studies, or other biomarkers from several small studies" as well as "evidence for biomarker therapeutic predictions for established drugs for different indications." Finally, level D indicates "preclinical findings or case studies of prognostic or diagnostic biomarkers," including indirect and inferential findings as well. The VICC KB is accessible via an API and can thus be easily incorporated programmatically into any precision oncology pipeline. Chap. 10 of this volume provides a detailed overview of methods and databases for clinical interpretation of variants.

Precision Oncology Reports

Reports are a key component and the main output and of a precision oncology platform which summarize clinically relevant findings and present them to the referring physicians and the patients. When compiling a report, one should aim for completeness, clarity, and brevity: all the meaningful data that can inform patient stratification, prognosis, and therapy should be included, organized coherently in tables, lists, and plots, and annotated with essential information from the literature and other relevant sources.

Prioritizing the findings is crucial to generate effective reports. In this chapter, the typical standard components of a precision oncology platform were introduced, as well as other potential sources of clinically relevant information that may not yet be mature for a clinical decision-

making system but could be in the (near) future. Clinically actionable data, associating the variants detected in a patient with specific treatments and clinical trials, such as level A and B evidence from the VICC KB, represent the essential information to include in a report and should be further prioritized by cancer type. An actionable mutation is particularly relevant when identified in the same type of cancer affecting the patient, while may not be as relevant when discovered in a distant type, for example, solid vs. liquid tumors. Nevertheless, one of the key concepts of precision medicine is that patients may be matched with treatments on the basis of specific genomic alterations rather than on the tumor histology or tissue type; therefore, it is always recommendable to include all the data supported by strong evidence. If available, information on clinical trials for which the patient is eligible should also be provided. While data on SNV, CNA, gene fusions, and specific gene expression biomarkers included in the VICC KB and similar sources is annotated with evidence level and, therefore, relatively easy to prioritize, the results of the optional components discussed in this chapter, such as pathway analysis and intratumor heterogeneity, may not be interpreted and incorporated as easily and should be provided, if available, as additional information. For example, while a specific level A SNV, like BRAF V600E, can be immediately actionable and targeted with a specific drug, the knowledge that the patient's gene expression profile indicates activation of the MAPK pathway can increase confidence in the recommendation when associated with the mutation, but can be less specific on its own. Similarly, the additional information provided by the assessment of the clonal landscape as estimated indirectly based on SNV and CNA, may not represent a strong actionable finding. However, knowing that the patient has a tumor subclone harboring a mutation that confers resistance to a specific drug which is otherwise recommended, may help the physician to consider alternative treatments or a combination therapy additionally targeting the problematic subclone, if available.

An Example: A Precision Oncology Platform for Multiple Myeloma

In 2018, we published the results obtained in a pilot precision medicine clinical trial with relapsed and refractory multiple myeloma (MM) patients at the Mount Sinai Hospital, NY [25]. This trial leveraged our novel precision oncology platform which integrated WES and targeted DNA panels and RNA-Seq to generate drug recommendation for patients with advanced disease. MM is a cancer of antibody-secreting plasma cells in the bone marrow, and, although generally manageable in the early stages, it becomes increasingly difficult to treat as patients progress and become refractory to therapy. The patients that were treated following the recommendations of our platform had already received 7 lines of therapy, on average, and were mostly refractory to standard of care and other approved options for myeloma.

The platform consisted of several components for the analysis of DNA sequencing data, whether produced by WES or targeted panels, and RNA-Seq data. Tumor plasma cells from the bone marrow (CD138+ cells) were subjected to DNA sequencing (WES or targeted sequencing) and RNA-Seq. Granulocytes from peripheral blood were also sequenced (WES) as a control for DNA analysis, which consisted of mapping of raw sequencing data using BWA, data post-processing using Samtools and GATK/Picard, SNV calling using MuTect, and CNA calling using Battenberg [26, 38, 40, 47, 54].

The RNA pipeline consisted of raw data mapping with STAR and gene quantification with featureCounts [82, 84]. Raw counts were normalized using the TMM function from the R package edgeR and voom from the R package limma [145, 146]. Since there was no normal control available for RNA-Seq data, DE genes were identified in each sample as the genes that were over- or underexpressed compared to the other samples in the cohort, as determined by z-scores. Pathway activity was calculated by applying single sample variation analysis (GSVA) on a set of pathways of interest [116]. Finally, the tool

L1000CDS [2] was used to perform reverse-matching of patient samples with gene expression profiles induced by drugs on cell lines retrieved from the L1000 project [110, 111]. The methods are described in greater detail in the article [25].

DNA and RNA actionable findings (SNV, CNA, and DE genes) were identified and prioritized using the knowledge base CIViC, which provided association with drugs, along with the L1000 analysis on gene expression profiles [142]. Additional RNA findings from pathway analysis were annotated based on a curated set of drugs targeting dysregulated pathways.

The platform generated recommendations for 63 of the 64 patients whose sequencing data was analyzed, and 21 of these received at least one of the recommendations and were evaluable for response. The clinical benefit rate, that is, minimal response or greater, was 76%. Successful drug options included the MEK inhibitor Trametinib, recommended because of SNVs in NRAS or KRAS, Panobinostat, recommended based on activation of the HDAC pathway, and the BCL2 inhibitor Venetoclax, recommended based on the upregulation of BCL2.

The results of this study demonstrated that a comprehensive precision medicine approach based on DNA and RNA sequencing in advanced myeloma is feasible and can identify valuable therapeutic options beyond standard of care. Furthermore, the study provided proof of principle that the analysis of RNA can successfully complement the standard DNA-based approach, providing additional actionable evidence not captured by genomic assays.

Conclusion

Precision oncology is a fast-evolving field which enables rapid translation of biomedical research discoveries into clinical cancer care. This chapter has provided an overview of the architecture of a precision oncology software platform, describing both standard components currently implemented in commercial and research settings, and more advanced components that are not yet mature for full clinical application. While the use of actionable DNA mutations to determine prognosis and therapy has now become a component of many clinical trials and of routine clinical decision-making, the use of other data modalities, such as RNA biomarkers, and advanced secondary information, like intratumor heterogeneity and pathway activity, is still being investigated for its accuracy and clinical potential. The TRACERx clinical trials in lung and renal cell cancers, for example, are evaluating the impact of intratumor heterogeneity in cancer progression and treatment resistance to determine novel actionable markers of drug response [147–150]. Therefore, it is reasonable to expect that these and other studies will enable the incorporation of additional data layers into precision oncology platforms for improved dissection of the disease and design of therapy.

It is also reasonable to expect future expansions of precision oncology systems to include other sequencing technologies, such as single-cell DNA and RNA sequencing, which could help to better dissect the complexity of intratumor heterogeneity and of the tumor microenvironment, and other types of omics, such as radiomics, which allows to extract quantitative features from medical images and is already showing great performance in diagnosis and disease staging [151–155]. Finally, the integration of more sophisticated artificial intelligence and machine learning tools into clinical decision-making platforms will further accelerate progress in cancer treatment and assist physicians to provide better and focused care [156–160].

References

1. Amstutz P. Portable, reproducible analysis with arvados. F1000Res. 2015;4
2. Di Tommaso P, et al. Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017;35:316–9.
3. Rusch M, et al. Clinical cancer genomic profiling by three-platform sequencing of whole genome, whole exome and transcriptome. Nat Commun. 2018;9:3962.
4. Sboner A, Elemento O. A primer on precision medicine informatics. Brief Bioinform. 2016;17:145–53.

5. Réda M, et al. Implementation and use of whole exome sequencing for metastatic solid cancer. EBioMedicine. 2020;51:102624.
6. Speleman F, et al. Copy number alterations and copy number variation in cancer: close encounters of the bad kind. Cytogenet Genome Res. 2008;123:176–82.
7. Shlien A, Malkin D. Copy number variations and cancer. Genome Med. 2009;1:62.
8. Frampton GM, et al. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. Nat Biotechnol. 2013;31:1023–31.
9. Brannon AR, et al. Comparative sequencing analysis reveals high genomic concordance between matched primary and metastatic colorectal cancer lesions. Genome Biol. 2014;15:454.
10. Chilamakuri CSR, et al. Performance comparison of four exome capture systems for deep sequencing. BMC Genomics. 2014;15:449.
11. Malone ER, Oliva M, Sabatini PJB, Stockley TL, Siu LL. Molecular profiling for precision cancer therapies. Genome Med. 2020;12:8.
12. Strom SP. Current practices and guidelines for clinical next-generation sequencing oncology testing. Cancer Biol Med. 2016;13:3–11.
13. Mandelker D, et al. Germline-focussed analysis of tumour-only sequencing: recommendations from the ESMO Precision Medicine Working Group. Ann Oncol. 2019;30:1221–31.
14. Nakagawa H, Fujita M. Whole genome sequencing analysis for cancer genomics and precision medicine. Cancer Sci. 2018;109:513–22.
15. Priest JR. A primer to clinical genome sequencing. Curr Opin Pediatr. 2017;29:513–9.
16. Rosenquist R, et al. Clinical utility of whole-genome sequencing in precision oncology. Semin Cancer Biol. 2021; https://doi.org/10.1016/j.semcancer.2021.06.018.
17. Zhao EY, Jones M, Jones SJM. Whole-genome sequencing in cancer. Cold Spring Harb Perspect Med. 2019;9:a034579.
18. Laganà A, et al. Integrative network analysis identifies novel drivers of pathogenesis and progression in newly diagnosed multiple myeloma. Leukemia. 2018;32:120–30.
19. Höllein A, et al. The combination of WGS and RNA-Seq is superior to conventional diagnostic tests in multiple myeloma: ready for prime time? Cancer Genet. 2020;242:15–24.
20. Monforte J, McPhail S. Strategy for gene expression-based biomarker discovery. Biotechniques. 2005;38(S4):25–9.
21. Yang X, et al. High-throughput transcriptome profiling in drug and biomarker discovery. Front Genet. 2020;11:19.
22. Goossens N, Nakagawa S, Sun X, Hoshida Y. Cancer biomarker discovery and validation. Transl Cancer Res. 2015;4:256–69.
23. Heyer EE, et al. Diagnosis of fusion genes using targeted RNA sequencing. Nat Commun. 2019;10:1388.
24. Rodon J, et al. Genomic and transcriptomic profiling expands precision cancer medicine: the WINTHER trial. Nat Med. 2019;25:751–8.
25. Laganà A, et al. Precision medicine for relapsed multiple myeloma on the basis of an integrative multiomics approach. JCO Precis Oncol. 2018;2018:1–17.
26. DePristo MA, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43:491–8.
27. Van der Auwera GA, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013;43:11.10.1–11.10.33.
28. Li X, Nair A, Wang S, Wang L. Quality control of RNA-seq experiments. Methods Mol Biol. 2015;1269:137–46.
29. Andrews, S. et al. FastQC: a quality control tool for high throughput sequence data. (2010).
30. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17:10.
31. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.
32. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34:i884–90.
33. Aran D, Sirota M, Butte AJ. Systematic pan-cancer analysis of tumour purity. Nat Commun. 2015;6:8971.
34. Lee S, et al. NGSCheckMate: software for validating sample identity in next-generation sequencing studies within and across data types. Nucleic Acids Res. 2017;45:e103.
35. Li H, Homer N. A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform. 2010;11:473–83.
36. Reinert K, Langmead B, Weese D, Evers DJ. Alignment of next-generation sequencing reads. Annu Rev Genomics Hum Genet. 2015;16:133–51.
37. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
38. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25:1754–60.
39. Wu TD, Reeder J, Lawrence M, Becker G, Brauer MJ. GMAP and GSNAP for genomic sequence alignment: enhancements to speed, accuracy, and functionality. Methods Mol Biol. 2016;1418:283–334.
40. Li H, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.
41. Koboldt DC. Best practices for variant calling in clinical sequencing. Genome Med. 2020;12:91.
42. Benjamin D, et al. Calling Somatic SNVs and Indels with Mutect2. bioRxiv. 2019:861054. https://doi.org/10.1101/861054.
43. Kim S, et al. Strelka2: fast and accurate calling of germline and somatic variants. Nat Methods. 2018;15:591–4.

44. Narzisi G, et al. Genome-wide somatic variant calling using localized colored de Bruijn graphs. Commun Biol. 2018;1:20.
45. Lai Z, et al. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Res. 2016;44:e108.
46. Sun JX, et al. A computational approach to distinguish somatic vs. germline origin of genomic alterations from deep sequencing of cancer specimens without a matched normal. PLoS Comput Biol. 2018;14:e1005965.
47. Cibulskis K, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013;31:213–9.
48. Sandmann S, et al. Evaluating variant calling tools for non-matched next-generation sequencing data. Sci Rep. 2017;7:43169.
49. Krøigård AB, Thomassen M, Lænkholm A-V, Kruse TA, Larsen MJ. Evaluation of nine somatic variant callers for detection of somatic mutations in exome and targeted deep sequencing data. PLoS One. 2016;11:e0151664.
50. Bian X, et al. Comparing the performance of selected variant callers using synthetic data and genome segmentation. BMC Bioinformatics. 2018;19:429.
51. Wang M, et al. SomaticCombiner: improving the performance of somatic variant calling based on evaluation tests and a consensus approach. Sci Rep. 2020;10:12898.
52. Zack TI, et al. Pan-cancer patterns of somatic copy number alteration. Nat Genet. 2013;45:1134–40.
53. Shao X, et al. Copy number variation is highly correlated with differential gene expression: a pan-cancer study. BMC Med Genet. 2019;20:175.
54. Nik-Zainal S, et al. The life history of 21 breast cancers. Cell. 2015;162:924.
55. Shen R, Seshan VE. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. Nucleic Acids Res. 2016;44:e131.
56. Gologan A, Sepulveda AR. Microsatellite instability and DNA mismatch repair deficiency testing in hereditary and sporadic gastrointestinal cancers. Clin Lab Med. 2005;25:179–96.
57. Chang L, Chang M, Chang HM, Chang F. Microsatellite instability: a predictive biomarker for cancer immunotherapy. Appl Immunohistochem Mol Morphol. 2018;26:e15–21.
58. Niu B, et al. MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. Bioinformatics. 2014;30:1015–6.
59. Huang MN, et al. MSIseq: software for assessing microsatellite instability from catalogs of somatic mutations. Sci Rep. 2015;5:13321.
60. Latham A, et al. Microsatellite instability is associated with the presence of Lynch syndrome pancancer. J Clin Oncol. 2019;37:286–95.
61. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv [q-bio.GN]. 2012;
62. Koboldt DC, Larson DE, Wilson RK. Using VarScan 2 for germline variant calling and somatic mutation detection. Curr Protoc Bioinformatics. 2013;44:15.4.1-17.
63. Yun T, et al. Accurate, scalable cohort variant calls using DeepVariant and GLnexus. Bioinformatics. 2021;36:5582–9.
64. Sutherland KD, Visvader JE. Cellular mechanisms underlying intertumoral heterogeneity. Trends Cancer. 2015;1:15–23.
65. Cavalli FMG, et al. Intertumoral heterogeneity within medulloblastoma subgroups. Cancer Cell. 2017;31:737–754.e6.
66. Bhalla S, et al. Patient similarity network of multiple myeloma identifies patient sub-groups with distinct genetic and clinical features. bioRxiv. 2020; https://doi.org/10.1101/2020.06.02.129767.
67. Marusyk A, Janiszewska M, Polyak K. Intratumor heterogeneity: the Rosetta stone of therapy resistance. Cancer Cell. 2020;37:471–84.
68. McGranahan N, Swanton C. Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. Cancer Cell. 2015;28:141.
69. Rosenthal R, McGranahan N, Herrero J, Swanton C. Deciphering genetic intratumor heterogeneity and its impact on cancer evolution. Annu Rev Cancer Biol. 2017;1:223–40.
70. Dentro SC, Wedge DC, Van Loo P. Principles of reconstructing the subclonal architecture of cancers. Cold Spring Harb Perspect Med. 2017;7:a026625.
71. Vandin F. Computational methods for characterizing cancer mutational heterogeneity. Front Genet. 2017;8:83.
72. Roth A, et al. PyClone: statistical inference of clonal population structure in cancer. Nat Methods. 2014;11:396–8.
73. Miller CA, et al. SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. PLoS Comput Biol. 2014;10:e1003665.
74. Deshwar AG, et al. PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. Genome Biol. 2015;16:35.
75. Deveau P, et al. QuantumClone: clonal assessment of functional mutations in cancer based on a genotype-aware method for clonal reconstruction. Bioinformatics. 2018;34:1808–16.
76. Jiang Y, Qiu Y, Minn AJ, Zhang NR. Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. Proc Natl Acad Sci U S A. 2016;113:E5528–37.
77. Ricketts C, et al. Meltos: multi-sample tumor phylogeny reconstruction for structural variants. Bioinformatics. 2020;36:1082–90.
78. Myers MA, Satas G, Raphael BJ. CALDER: inferring phylogenetic trees from longitudinal tumor samples. Cell Syst. 2019;8:514–522.e5.

79. El-Kebir M, Satas G, Oesper L, Raphael BJ. Inferring the mutational history of a tumor using multi-state perfect phylogeny mixtures. Cell Syst. 2016;3:43–53.
80. Wu H, Li X, Li H. Gene fusions and chimeric RNAs, and their implications in cancer. Genes Dis. 2019;6:385–90.
81. An X, et al. BCR-ABL tyrosine kinase inhibitors in the treatment of Philadelphia chromosome positive chronic myeloid leukemia: a review. Leuk Res. 2010;34:1255–68.
82. Dobin A, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.
83. Dobin A, Gingeras TR. Optimizing RNA-seq mapping with STAR. Methods Mol Biol. 2016;1415:245–62.
84. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30:923–30.
85. Anders S, Pyl PT, Huber W. HTSeq–a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31:166–9.
86. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.
87. Robert C, Watson M. Errors in RNA-Seq quantification affect genes of relevance to human disease. Genome Biol. 2015;16:177.
88. Zytnicki M. Mmquant: how to count multi-mapping reads? BMC Bioinformatics. 2017;18(**1**):**1–6**.
89. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34:525–7.
90. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14:417–9.
91. Patro R, Mount SM, Kingsford C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nat Biotechnol. 2014;32:462–4.
92. Haas BJ, et al. Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. Genome Biol. 2019;20:213.
93. Haas BJ, et al. STAR-fusion: fast and accurate fusion transcript detection from RNA-Seq. bioRxiv. 2017; https://doi.org/10.1101/120295.
94. Jasper J, Powers JG, Weigman VJ. Abstract 2296: STAR-SEQR: accurate fusion detection and support for fusion neoantigen applications. In: Bioinformatics and systems biology. American Association for Cancer Research; 2018. https://doi.org/10.1158/1538-7445.am2018-2296.
95. Uhrig S, et al. Accurate and efficient detection of gene fusions from RNA sequencing data. Genome Res. 2021;31:448–60.
96. Han L-O, Li X-Y, Cao M-M, Cao Y, Zhou L-H. Development and validation of an individualized diagnostic signature in thyroid cancer. Cancer Med. 2018;7:1135–40.
97. Paquet ER, Lesurf R, Tofigh A, Dumeaux V, Hallett MT. Detecting gene signature activation in breast cancer in an absolute, single-patient manner. Breast Cancer Res. 2017;19:1–**5**.
98. Liu X, et al. A prognostic gene expression signature for oropharyngeal squamous cell carcinoma. EBioMedicine. 2020;61:102805.
99. Chen H-Y, et al. A five-gene signature and clinical outcome in non-small-cell lung cancer. N Engl J Med. 2007;356:11–20.
100. Zhou J-G, et al. Development and validation of an RNA-Seq-based prognostic signature in neuroblastoma. Front Oncol. 2019;9:1361.
101. van Laar R, et al. Translating a gene expression signature for multiple myeloma prognosis into a robust high-throughput assay for clinical use. BMC Med Genet. 2014;7:25.
102. Kwa M, Makris A, Esteva FJ. Clinical utility of gene-expression signatures in early stage breast cancer. Nat Rev Clin Oncol. 2017;14:595–610.
103. Rachid Zaim S, et al. Evaluating single-subject study methods for personal transcriptomic interpretations to advance precision medicine. BMC Med Genet. 2019;12:96.
104. Geman D, d'Avignon C, Naiman DQ, Winslow RL. Classifying gene expression profiles from pairwise mRNA comparisons. Stat Appl Genet Mol Biol. 2004;3:Article19.
105. Tan AC, Naiman DQ, Xu L, Winslow RL, Geman D. Simple decision rules for classifying human cancers from gene expression profiles. Bioinformatics. 2005;21:3896–904.
106. Li Q, et al. Interpretation of 'Omics dynamics in a single subject using local estimates of dispersion between two transcriptomes. AMIA Annu Symp Proc. 2019;2019:582–91.
107. Menche J, et al. Integrating personalized gene expression profiles into predictive disease-associated gene pools. NPJ Syst Biol Appl. 2017;3:10.
108. Chen B, et al. Reversal of cancer gene expression correlates with drug efficacy and reveals therapeutic targets. Nat Commun. 2017;8:16022.
109. Lamb J. The connectivity map: a new tool for biomedical research. Nat Rev Cancer. 2007;7:54–60.
110. Subramanian A, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. Cell. 2017;171:1437–1452.e17.
111. Duan Q, et al. L1000CDS2: LINCS L1000 characteristic direction signatures search engine. NPJ Syst Biol Appl. 2016;2(**1**):1–2.
112. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol. 2012;8:e1002375.
113. Tarca AL, Bhatti G, Romero R. A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. PLoS One. 2013;8:e79217.

114. Nguyen T-M, Shafi A, Nguyen T, Draghici S. Identifying significantly impacted pathways: a comprehensive review and assessment. Genome Biol. 2019;20:203.
115. Barbie DA, et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature. 2009;462:108–12.
116. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013;14:7.
117. Bhuva DD, et al. Using singscore to predict mutation status in acute myeloid leukemia from transcriptomic signatures. F1000Res. 2019;8:776.
118. Foroutan M, et al. Single sample scoring of molecular phenotypes. BMC Bioinformatics. 2018;19:404.
119. Li Q, et al. N-of-1-pathways MixEnrich: advancing precision medicine via single-subject analysis in discovering dynamic changes of transcriptomes. BMC Med Genet. 2017;10(1):5–16.
120. Ma J, Shojaie A, Michailidis G. A comparative study of topology-based pathway enrichment analysis methods. BMC Bioinformatics. 2019;20:546.
121. Ihnatova I, Popovici V, Budinska E. A critical comparison of topology-based pathway analysis methods. PLoS One. 2018;13:e0191154.
122. Liu C, Lehtonen R, Hautaniemi S. PerPAS: topology-based single sample pathway analysis method. IEEE/ACM Trans Comput Biol Bioinform. 2018;15:1022–7.
123. Schubert M, et al. Perturbation-response genes reveal signaling footprints in cancer gene expression. Nat Commun. 2018;9:20.
124. Tate JG, et al. COSMIC: the catalogue of somatic mutations in cancer. Nucleic Acids Res. 2019;47:D941–7.
125. Rehm HL, Harrison SM, Martin CL. ClinVar is a critical resource to advance variant interpretation. Oncologist. 2017;22:1562.
126. Landrum MJ, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 2016;44:D862–8.
127. Koch L. Exploring human genomic diversity with gnomAD. Nat Rev Genet. 2020;21:448.
128. Flanagan SE, Patch A-M, Ellard S. Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations. Genet Test Mol Biomarkers. 2010;14:533–7.
129. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome Res. 2001;11:863–74.
130. Adzhubei IA, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7:248–9.
131. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. 2011;39:e118.
132. Kircher M, et al. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46:310–5.
133. Shihab HA, et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum Mutat. 2013;34:57–65.
134. Yang H, Wang K. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. Nat Protoc. 2015;10:1556–66.
135. Ramos AH, et al. Oncotator: cancer variant annotation tool. Hum Mutat. 2015;36:E2423–9.
136. Gurbich TA, Ilinsky VV. ClassifyCNV: a tool for clinical annotation of copy-number variants. Sci Rep. 2020;10:20375.
137. Chapman PB, et al. Improved survival with vemurafenib in melanoma with BRAF V600E mutation. N Engl J Med. 2011;364:2507–16.
138. Wagner AH, et al. A harmonized meta-knowledgebase of clinical interpretations of somatic genomic variants in cancer. Nat Genet. 2020;52:448–57.
139. Huang L, et al. The cancer precision medicine knowledge base for structured clinical-grade mutations and interpretations. J Am Med Inform Assoc. 2017;24:513–9.
140. Chakravarty D, et al. OncoKB: a precision oncology knowledge base. JCO Precis Oncol. 2017;1:1–16.
141. Patterson SE, et al. The clinical trial landscape in oncology and connectivity of somatic mutational profiles to targeted therapies. Hum Genomics. 2016;10:4.
142. Griffith M, et al. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. Nat Genet. 2017;49:170–4.
143. Tamborero D, et al. Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. Genome Med. 2018;10(1):1–8.
144. Flaherty KT, et al. Combined BRAF and MEK inhibition in melanoma with BRAF V600 mutations. N Engl J Med. 2012;367:1694–703.
145. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40.
146. Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014;15:R29.
147. TRACERx Renal consortium. TRACERx Renal: tracking renal cancer evolution through therapy. Nat Rev Urol. 2017;14:575–6.
148. Bailey C, et al. Tracking cancer evolution through the disease course. Cancer Discov. 2021;11:916–32.
149. Jamal-Hanjani M, et al. Tracking genomic cancer evolution for precision medicine: the lung TRACERx study. PLoS Biol. 2014;12:e1001906.
150. Jamal-Hanjani M, et al. Tracking the evolution of non-small-cell lung cancer. N Engl J Med. 2017;376:2109–21.
151. Wiedmeier JE, Noel P, Lin W, Von Hoff DD, Han H. Single-cell sequencing in precision medicine. Cancer Treat Res. 2019;178:237–52.
152. Winterhoff B, Talukdar S, Chang Z, Wang J, Starr TK. Single-cell sequencing in ovarian cancer: a new

frontier in precision medicine. Curr Opin Obstet Gynecol. 2019;31:49–55.
153. Valdes-Mora F, et al. Single-cell transcriptomics in cancer immunobiology: the future of precision oncology. Front Immunol. 2018;9:2582.
154. Nieto P, et al. A single-cell tumor immune atlas for precision oncology. bioRxiv. 2020; https://doi.org/10.1101/2020.10.26.354829.
155. Lee G, Lee HY, Ko ES, Jeong WK. Radiomics and imaging genomics in precision medicine. Precis Futur Med. 2017;1:10–31.
156. Parekh VS, Jacobs MA. Deep learning and radiomics in precision medicine. Expert Rev Precis Med Drug Dev. 2019;4:59–72.
157. Azuaje F. Artificial intelligence for precision oncology: beyond patient stratification. NPJ Precis Oncol. 2019;3:6.
158. Del Giudice M, et al. Artificial intelligence in bulk and single-cell RNA-sequencing data to foster precision oncology. Int J Mol Sci. 2021;22(9):4563.
159. Dlamini Z, Francies FZ, Hull R, Marima R. Artificial intelligence (AI) and big data in cancer and precision oncology. Comput Struct Biotechnol J. 2020;18:2300–11.
160. Ding MQ, Chen L, Cooper GF, Young JD, Lu X. Precision oncology beyond targeted therapy: combining omics data with machine learning matches the majority of cancer cells to effective therapeutics. Mol Cancer Res. 2018;16:269–78.

2 Software Workflows and Infrastructures for Precision Oncology

Waleed Osman and Alessandro Laganà

Abstract

Precision oncology mainly relies on genetic and molecular patient profiling from high-throughput sequencing data. The necessity to process and analyze large volumes of data has led to the development of robust computational tools and methods. The most challenging aspect in the implementation of a precision oncology workflow involves proper handling of large volume of data, while ensuring the results are reproducible and replicable. In this chapter, we provide a detailed description of the various tools available for the design and implementation of a precision oncology pipeline along with the technical considerations to make to utilize these tools effectively. We then provide a guide to the development of a precision oncology pipeline, with a specific emphasis on the software workflows and infrastructure needed.

W. Osman
Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA

A. Laganà (✉)
Department of Genetics and Genomic Sciences, Department of Oncological Sciences, Mount Sinai Icahn School of Medicine, New York, NY, USA
e-mail: alessandro.lagana@mssm.edu

Introduction

Precision oncology is an innovative research area that has introduced a novel approach to cancer care, where diagnosis, prognosis, and therapy are informed by genetic and molecular profiling of the individual patient, rather than being based on a *one-size-fits-all* approach [1–4]. This landmark paradigm shift has been enabled in recent years by the reduced cost of next-generation sequencing (NGS) technologies and a myriad of ad hoc tools and software applications developed in order to analyze the data generated [5, 6]. The explosion of tools and methods as a response to the more widely available multi-omic data sets has created a challenge in terms of reproducibility, interoperability, and standardization. Tools created for the analysis of genomic, proteomic, transcriptomic, and other omic data are typically written in one or a combination of three different styles: Command Line Interface (CLI), Application Programming Interface (API), or Graphical User Interface (GUI) [6]. Combining and ensuring reproducibility of these disparate application types has proven to be a major challenge for biologists as they often will require a deeper knowledge of software application development norms and techniques as well as greater computational capabilities. The absence of widely accepted best practices regarding software and database

© Springer Nature Switzerland AG 2022

A. Laganà (ed.), *Computational Methods for Precision Oncology*, Advances in Experimental Medicine and Biology 1361, https://doi.org/10.1007/978-3-030-91836-1_2

utilization has contributed greatly to irreproducibility, resulting in many man hours and compute cycles wasted in attempting to recreate past efforts [7].

As a remedy to this, a number of workflow management systems (WMS) and executors for running these workflow systems have been developed, such as Snakemake, Nextflow, the Workflow Description Language (WDL) (https://openwdl.org/), and The Common Workflow Language (CWL) [8–10]. Infrastructure enabling the execution of these workflows have also been developed such as Arvados (stand-alone, deployable, open-source), and Broad Institute's Terra Bio Cloud Platform (web based) [11, 12].

These infrastructure and software solutions are able to organize and process large volumes of genomics data enabling scientists to discover ever deeper insight into biological data. Today, with the use of CWL, Arvados, and Cromwell (https://github.com/broadinstitute/cromwell), and facilitated by virtual servers on cloud infrastructure, bioinformaticians and savvy data engineers can write and implement a precision medicine pipeline while maintaining reproducibility and interoperability. In this chapter, we will introduce several bioinformatics workflow management systems and the infrastructures to execute them.

Workflow Management Systems and Languages

Workflow management systems (WMS) are essential in the processing of large sets of patient's genomic data. WMS are tools developed to facilitate the orchestration and execution of computational processes in an optimal and efficient manner. In bioinformatics, these systems integrate various discrete command-line tools into one workflow for the rapid development of pipelines, which can be deployed across a variety of infrastructures and environments. Utilizing a WMS ensures ease of set-up and the ability to monitor performance of individual predefined tasks. These workflows are often linear but can also be dynamic or run in parallel. Table 2.1 provides a list with the most used WMS along with their URLs.

CWL: Common Workflow Language

The first of several bioinformatic workflow management languages and systems discussed here is the Common Workflow Language (CWL; https://github.com/common-workflow-language) [8]. CWL is an open standard for describing analysis workflows and tools in a way that makes them portable and scalable across a variety of software and hardware environments, from workstations to cluster, cloud, and high-performance computing (HPC) environments. It can be applied to a number of different scientific domains including Bioinformatics, Medical Imaging, Astronomy, High Energy Physics, and Machine Learning. CWL sets itself apart from most other workflow languages by attempting to adopt open-source principles and standards such as open-stand.org, which advocates for cooperation, adherence to principles, collective empowerment, availability, and voluntary adoption. CWL is not a software, but a specification which describes command line tools and allows them to be connected together to form a workflow. CWL's commitment to creating a community which focuses on standardization and other open-source principles has led to its adoption by a number of workflow execution programs such as Toil, Arvados, Rabix, Cromwell, and Bcbio (See Tables 2.1 and 2.2). Rabix, for example, is a powerful open-source suite of tools for CWL, which include Rabix Composer, a graphical editor enabling visual programming in CWL, Rabix Benten, a language server for CWL documents, and Rabix Executor, a workflow runner that can execute CWL pipelines (https://rabix.io/). Figure 2.1 shows an example of graph generated with Rabix Composer.

The use of CWL to create tools and workflows facilitates the ease of future repeatability and reproducibility of results. This leads to greater cooperation between standard organizations,

Table 2.1 Workflow management systems

Name	Description	Website
Nextflow	Domain-specific language	http://nextflow.io
Toil	Pipeline management system	https://toil.ucsc-cgl.org
Snakemake	Domain-specific language	https://snakemake.github.io
Bpipe	Domain-specific language	http://docs.bpipe.org
WDL	Workflow specification language	https://openwdl.org/
CWL	Workflow specification language	https://www.commonwl.org/

building a foundation for collaboration. The development of CWL into a standard was made possible by adhering to five fundamental principles of standard development [13]. First, decisions regarding the direction and development of the standard must be made with equity and fairness, implementing a well-defined *due process* by which participating parties have the ability to appeal decisions made. Next, a *broad consensus* must be made in order to facilitate agreement across a range of interests. A general agreement, incorporating all views, is paramount to the establishment and persistence of an open standard. Third, activities and work being undertaken must be recorded for posterity with those records open and easily accessible to all. A consistent *transparency* must be maintained by giving advance notice of new proposals and activities. Fourth, a certain *balance* must be struck among all parties involved. No one entity involved in the development of the standard may have disproportionate influence on its direction or activities. Finally, the processes by which the standards are developed must be *open* to all. CWL stands out by encompassing all these principles and enabling cross-collaboration.

WDL: Workflow Description Language

WDL (Workflow Description Language) is a community-driven open-development workflow language developed by the Broad Institute [14]. WDL specifies data processing workflows with a human-readable and writable syntax very similarly to CWL. WDL was ostensibly developed to support Terra, a platform developed by the Broad Institute of MIT and Harvard in collaboration with Verily Life Sciences. Terra is not open-source platform and requires users to purchase credits for compute cycles. Similar to CWL, the WDL scripts are not executable and require an execution engine, such as Cromwell, MiniWDL or dxWDL, and an environment to be runnable.

NextFlow

NextFlow is a popular workflow system developed by Seqera Labs in Barcelona, Spain, designed to address numerical instability, efficient parallel execution, error tolerance, execution provenance, and traceability [9]. Similar to CWL, this domain-specific language (DSL) utilizes software containers to create scalable and reproducible workflows, enabling rapid pipeline development through the adaptation of existing pipelines written in any scripting language. NextFlow also supports GitHub and BitBucket integration, which allows for the consistent tracking of software changes and versions. Containerization, enabled by utilizing container platforms such as Docker (https://www.docker.com/) or Singularity (https://singularity.hpcng.org/), ensures numerical stability [15, 16]. It can be executed on Sun Grid Engine (SGE) (http://star.mit.edu/cluster/docs/0.93.3/guides/sge.html), Load Sharing Facility (LSF) (https://www.ibm.com/docs/en/spectrum-lsf/10.1.0), SLURM workload manager (https://slurm.schedmd.com/overview.html), Portable Batch System (PBS) (https://www.nas.nasa.gov/hecc/support/kb/portable-batch-system-(pbs)-overview_126.html) and for Kubernetes (https://kubernetes.io/), Amazon Web Services (AWS) (https://aws.amazon.com/), and Google Cloud platforms (https://cloud.google.com/) for rapid computation and

Table 2.2 Data processing platforms

Platform	Description	Website
Arvados	Open-source platform for managing, processing, and sharing genomic and other large scientific and biomedical data	https://arvados.org/
Terra	A scalable platform for biomedical research which allows data access, running analysis tools, and collaboration	https://terra.bio/
Galaxy	Open, web-based platform for accessible, reproducible, and transparent computational biomedical research.	https://github.com/galaxyproject/galaxy
bcbio-nextgen	Validated, scalable, community developed variant calling, RNA-seq, and small RNA analysis platform	https://github.com/bcbio/bcbio-nextgen
DolphinNext	A graphical user interface for distributed data processing of high-throughput genomics	https://github.com/UMMS-Biocore/dolphinnext
Sequanix	GUI for the Snakemake pipeline	https://github.com/sequana/sequana/
DNAnexus	A cloud-based data analysis and management platform for DNA sequence data	https://www.dnanexus.com/

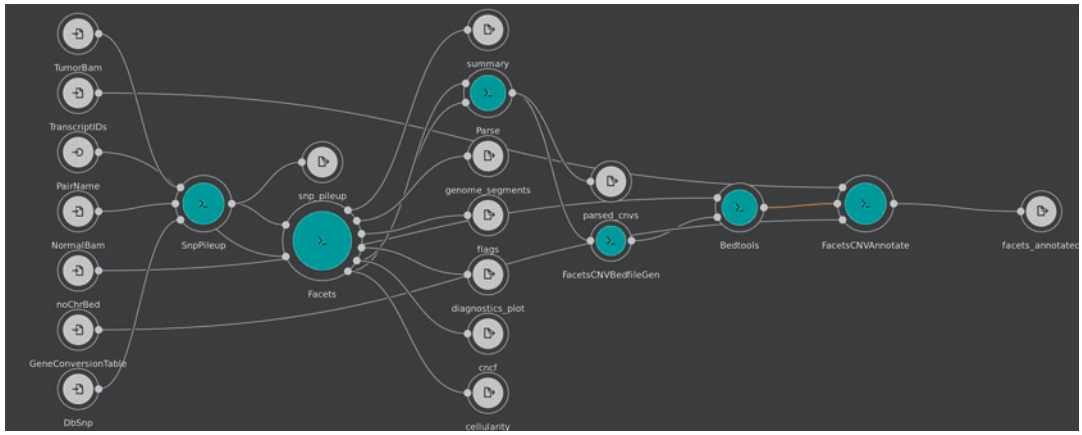

Fig. 2.1 Example of graph generated with CWL Rabix Composer

the ability to scale up projects manyfold. NextFlow also takes advantage of the "dataflow programming paradigm," where execution tasks are started automatically as soon as data is received through input channels. The Make-like approach adopted by tools such as CWL require pre-estimation of all computational dependencies as well as a directed acyclic graph (DAG). NextFlow, however, utilizes a top to bottom approach which mimics the natural flow of data.

Data Processing Platforms

The main data processing platform we will be discussing in this section is Arvados, which has been deployed in our lab and has shown great utility for our genomic processing needs. Table 2.2 summarizes the main data processing platforms.

Arvados

Arvados is a free and open-source platform for processing large volumes of genomic data [11]. This distributed computing platform for data analysis on massive data sets also enables users to share and manage their data with ease. It is licensed under the GNU Affero General Public License version 3. Two key features of Arvados are *provenance* and *reproducibility*. Arvados maintains integrity of data by recording its history and place of origin, which also reduces the incidence of replication of intermediate files. Arvados retains the history of jobs run in its infrastructure and recognizes when to re-use existing files, a cost-saving measure valuable to system administrators and informaticists alike. This is all enabled in-part by Arvados's *keep store*, a content-addressable storage system designed to run on low-cost commodity hardware or cloud services.

Other Platforms

While Arvados is free and open source, other platforms require a payment or subscription, where billing is incorporated directly into the application software.

DNAnexus (http://www.dnanexus.com/) and Terra.bio (http://terra.bio) both require the user to pay for storage and processing costs; the Galaxy project stands out with a strong, knowledgeable, and supportive online community [17] (https://usegalaxy.org/); Bcbio-nextgen is focused mainly on RNA genomic data analysis and lacks the flexibility of the other platforms mentioned in this paper [18] (https://github.com/bcbio/bcbio-nextgen); DolphinNext (https://dolphinnext.umassmed.edu/) and Sequanix (https://github.com/sequana/sequana/) are two GUIs developed specifically for Snakemake and Nextflow DSLs,

respectively [19, 20]. These platforms attempt to ease the process of generating workflows by providing users with a web interface, expanding access to users with limited bioinformatics experience.

Implementation of a Precision Oncology Workflow

Designing and implementing a precision oncology pipeline requires several of the abovementioned components and entails the coordination of many tools which are then combined to create explicit workflows, relaying and processing data until it is collected and presented in a final report form. Compute and data intensive processing steps often require infrastructure consisting of large compute clusters, multiple processors, and large amounts of disc space in order to ensure reliability, efficiency, availability, and scalability. A comprehensive description of a precision oncology pipeline is provided in Chap. 1. Here, we introduce the basic syntax of CWL scripts, describe the basic steps in the design of a precision medicine workflow for DNA variant calling, and provide an overview of the software infrastructures necessary for the implementation of such workflows.

Introduction to CWL Scripting

The first step in writing a precision medicine workflow is to select the command-line tools intended for integration. This usually comprises several steps including, but not limited to, a raw read QC step, alignment, variant calling, annotation, and secondary analysis. We will use CWL as the specification for the workflow in a few examples. Figure 2.2 illustrates how inputs and outputs are isolated for reproducibility. This simple "hello world" program accepts one input parameter, writes a message to the terminal or job log, and subsequently will produce no permanent output. Several of these tools can then be written together in conjunction to form a "workflow." Figure 2.3 shows a sample workflow which extracts a java source file from a tar file and then compiles it.

There are several key considerations to make when writing and executing a workflow. First, every step in a workflow will require its own CWL description. The final inputs and outputs of the workflow are listed in the inputs and outputs section. The steps are specified under steps. The order of execution is determined by the specified connections between steps.

After writing the workflows, one has to choose an appropriate method for running them. In the example shown in Fig. 2.4, we use the cwl-runner. Since CWL is highly portable, the compute environment chosen to run the workflows will be up to user discretion.

Finally, Fig. 2.5 displays a more complex example of a script implementing the workflow shown in Fig. 2.1, with steps from a precision oncology pipeline which include the analysis of Copy Number Alterations (CNA) (tool: Facets [21]) and the reconstruction of tumor sub-clonal composition (tool: PhyloWGS [22]).

The Typical Steps of a Precision Oncology Pipeline

Figure 2.6 shows a typical schema for a precision oncology pipeline. After sample collection, processing and sequencing has occurred, the raw sequencing data in the form of Fastq files are used as inputs into the pipeline. Next, a series of quality control metrics are generated from the data to help determine in which areas there may be problems or poor-quality data. Metrics included in the evaluation of quality include raw sequencing data quality and depth, alignment quality, GC content, adapter contamination, and reads duplication rates [23, 24]. Evaluating these metrics allows for the identification and flagging of poor-quality data and to avoid potentially expensive and computationally intensive steps. Checking alignment quality can prevent potential false-positive single nucleotide polymorphism calls. Furthermore, it is important to verify that paired files generated from samples from the same individual, for example, normal and tumor WES samples, are indeed from the same individual, by using a tool like NGSCheckMate [25].

Next, reads are aligned to a common reference genome. Alignment algorithms such as the

Code

```
#!/usr/bin/env cwl-runner

cwlVersion: v1.0
class: CommandLineTool
baseCommand: echo
inputs:
  message:
    type: string
    inputBinding:
      position: 1
outputs: []
```

Fig. 2.2 Example of simple CWL demonstrating input/output

Burrows–Wheeler transform can be utilized to rearrange raw sequencing data and prepare it for downstream analysis and mutational calling [26]. The resulting file produced is typically a Sequence Alignment Map (SAM) or its binary version (Binary Alignment Map, BAM) file.

Following sequence alignment and the generation of a BAM/SAM file, a typical precision medicine pipeline would then perform variant calling by identifying where the aligned reads differ from the reference genome, producing a variant call file to be used in further downstream analysis [27] (see also Chaps. 1 and 3). After the variants have been annotated using various online databases, additional pertinent information is assigned to each variant call [28]. This information may include the definition of a variant and its genotype, basic information regarding whether it lies in a coding region, its impact on the corresponding protein (e.g., missense or synonymous mutation), or whether the variant is an insertion or a deletion. Those variants are then classified based on ACMG guidelines as pathogenic, likely pathogenic, uncertain significance, likely benign, or benign [29]. Additionally, structural variation analysis may be conducted to identify genomic alterations such as duplications, inversions, translocations, and copy number variants (CNVs) (See also Chap. 4).

The variants are then collected and classified based on whether they are actionable or not, using different databases for clinical interpretation, then summarized into reports, often after being reviewed and further annotated by pathologists [28].

In more advanced settings, the variants data can be inputted into a rule-based engine which will select and prioritize drugs matching the alterations. These "drug recommendation engines" are still in early-phase development and are typically ad hoc applications which draw on experts with domain-specific knowledge in order to auto-generate drugs with the expectation of affecting the deleterious variants in a positive manner [30–32]. Many iterations and versions of this ad hoc pipeline are being developed across academia and medical institutions for the treatment of various cancers. Each pipeline with its own unique set of rules and considerations based on the model-disease specifications.

Software Infrastructures for Precision Oncology Platforms

Here we provide some background on the software infrastructure for a precision oncology pipeline. The diagram in Fig. 2.7 illustrates the

Code

```
#!/usr/bin/env cwl-runner

cwlVersion: v1.0
class: Workflow
inputs:
  tarball: File
  name_of_file_to_extract: string

outputs:
  compiled_class:
    type: File
    outputSource: compile/classfile

steps:
  untar:
    run: tar-param.cwl
    in:
      tarfile: tarball
      extractfile: name_of_file_to_extract
    out: [extracted_file]

  compile:
    run: arguments.cwl
    in:
      src: untar/extracted_file
    out: [classfile]
```

Fig. 2.3 Example of CWL workflow which extracts a java source file from a tar file and compiles it

components which comprise the Arvados technical architecture. It can be deployed locally, or on a number of different cloud providers such as Amazon Web Services (AWS) (https://aws.amazon.com/), the Google Cloud Platform (GCP) (https://cloud.google.com/), or on Microsoft Azure (https://azure.microsoft.com/). Several key components work together in harmony to create an elastic computing environment where the overall resource footprint available or consumed by a specific job can grow or shrink on demand. The ability of Arvados to quickly expand or decrease computer processing, memory, and storage resources as well as manage data through a content-addressable distributed storage system sets it apart from its competitors. These components are the container orchestration system called "Crunch," the distributed storage system "Keep," the REST API Server, the CLI, the GUI "Workbench," native language SDKs, Data Manager, Node Manager, and Keep proxy.

The main two innovations of the Arvados platform are "Crunch" and "Keep." The Crunch container orchestration management engine executes

Output

```
$ echo "public class Hello {}" > Hello.java && tar -cvf hello.tar He
llo.java
$ cwl-runner 1st-workflow.cwl 1st-workflow-job.yml
[job untar] /tmp/tmp94qFiM$ tar --create --file /home/example/hello.
tar Hello.java
[step untar] completion status is success
[job compile] /tmp/tmpu1iaKL$ docker run -i --volume=/tmp/tmp94qFiM/
Hello.java:/var/lib/cwl/job301600808_tmp94qFiM/Hello.java:ro --volum
e=/tmp/tmpu1iaKL:/var/spool/cwl:rw --volume=/tmp/tmpfZnNdR:/tmp:rw -
-workdir=/var/spool/cwl --read-only=true --net=none --user=1001 --rm
--env=TMPDIR=/tmp java:7 javac -d /var/spool/cwl /var/lib/cwl/job301
600808_tmp94qFiM/Hello.java
[step compile] completion status is success
[workflow 1st-workflow.cwl] outdir is /home/example
Final process status is success
{
  "compiled_class": {
    "location": "/home/example/Hello.class",
    "checksum": "sha1$e68df795c0686e9aa1a1195536bd900f5f417b18",
    "class": "File",
    "size": 416
  }
}
```

Fig. 2.4 Example of cwl-runner execution

CWLs while maintaining provenance and reproducibility. It accomplishes this by automatically tracking the origin of result data; therefore, it is able to compare workflows to one another, avoiding the need to repeat previously performed data analysis. This saves on cost and time, two significant considerations when executing a workflow or data analysis. Crunch also provides the ability to scale horizontally by provisioning compute nodes upon demand, delivering cost-effective performance. Finally, the Crunch engine isolates workloads by running jobs inside of Docker containers, a standard unit of software that packages up code and all its dependencies [15].

The Keep system efficiently handles data storage and management using a content-addressable distributed storage system. It is able to handle petabyte-sized data sets, scaling accordingly by utilizing location-addressed storage. A permanent universally unique identifier (UUID) is then given to each content address. This creates a highly scalable flat address space, virtualizing storage access. The benefits of the keep store system include, elimination of duplication, canonical records, provenance, easy management of temporary data, flexible organization, high reliability, security and access control, POSIX interface, data sharing, and versioning.

```
#!/usr/bin/env cwl-runner
cwlVersion: v1.0

class: Workflow
label: facets
requirements:
  InlineJavascriptRequirement: {}

inputs:
  PairName: string
  NormalBam: File
  TumorBam: File
  noChrBed: File
  GeneConversionTable: File
  TranscriptIDs: File
  DbSnp: File

steps:
  SnpPileup:
    run: ../Tools/SnpPileup.cwl
    in:
      TumorBam: TumorBam
      NormalBam: NormalBam
      DbSnp: DbSnp
      isGzip:
        default: true
    out:
      [pileup]

  Facets:
    run: ../Tools/Facets/Facets.cwl
    in:
      PairName: PairName
      PileUp: SnpPileup/pileup
      CritValue:
        default: 150
    out:
      [genome_segments, diagnostics_plot, cncf, summary, flags, cellularity]

  Parse:
    run: ../Tools/PhyloWGS/PhyloWGS-Parse.cwl
    in:
      InputFile: Facets/cncf
      CNVFormat:
        default: "facets"
      Cellularity: Facets/cellularity
    out:
      [parsed_cnvs]

  FacetsCNVBedfileGen:
    run: ../Tools/Facets/FacetsCNVBedfileGen.cwl
    in:
      ParsedCNVs: Parse/parsed_cnvs
    out:
      [CNV_bed_file]

outputs:
  snp_pileup:
    type: File
    outputSource: SnpPileup/pileup

  genome_segments:
    type: File
    outputSource: Facets/genome_segments

  diagnostics_plot:
    type: File
    outputSource: Facets/diagnostics_plot

  cncf:
    type: File
    outputSource: Facets/cncf

  summary:
    type: File
    outputSource: Facets/summary

  flags:
    type: File
    outputSource: Facets/flags

  cellularity:
    type: File
    outputSource: Facets/cellularity

  facets_annotated:
    type: File
    outputSource: FacetsCNVAnnotate/facets_annotated

  parsed_cnvs:
    type: File
    outputSource: Parse/parsed_cnvs
```

Fig. 2.5 Example of a CWL script from a precision oncology pipeline. The script defines the step to run a CNV analysis using the tool Facets. The class field indicates that this document describes a command line tool. The three main sections describe the inputs, steps, and outputs of the pipeline

The installation and deployment of such infrastructures can be accomplished on GNU/Linux systems either bare metal, or on AWS, GCP, and Azure cloud services. The multi-host installation provides the highest throughput and can be accomplished using Salt, an automated infrastructure management software [26]. The Arvados salt formula can be found at https://github.com/saltstack-formulas/arvados-formula.git, and the steps for deployment are as follows:

1. Fork/copy the formula to your Salt master host.
2. Edit the Arvados, nginx, postgres, locale, and docker pillars to match your desired configuration.
3. Run a state.apply to get it deployed.

After this step, the cloud/software engineer will then need to set up the DNS in order to access the cluster's nodes. Typical operations include running a workflow, uploading, and downloading data from keep. Periodically, Arvados releases new versions of the platform which will require a short maintenance window where data processing will need to be suspended.

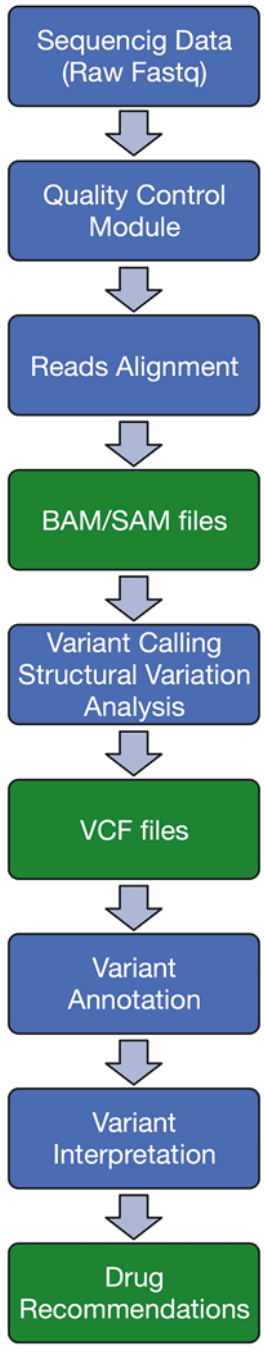

Fig. 2.6 A typical schema of a precision oncology pipeline

Conclusion

Utilizing an appropriate domain-specific language for workflow development and execution is a necessity. The adept bioinformatics engineer/analyst will require the combination of many tools and that combination will need to be seamless. CWL, WDL, Snakemake, and NextFlow all provide the portability and flexibility needed for precision oncology workflows. When the requisite components for a robust pipeline are in place, the effort to scale up your workload will be minimal.

Although many workflow systems are available, we have found that the combination of CWL and Arvados serve for the most comprehensive platform for genomics data processing at large scale. CWL's requirements for explicitness and isolation lead to more flexibility, portability, and scalability for your workloads. With a large user base, CWL is and will continue to be supported and updated on a regular basis. This will ensure the resilience and longevity of pipelines and precision medicine platforms.

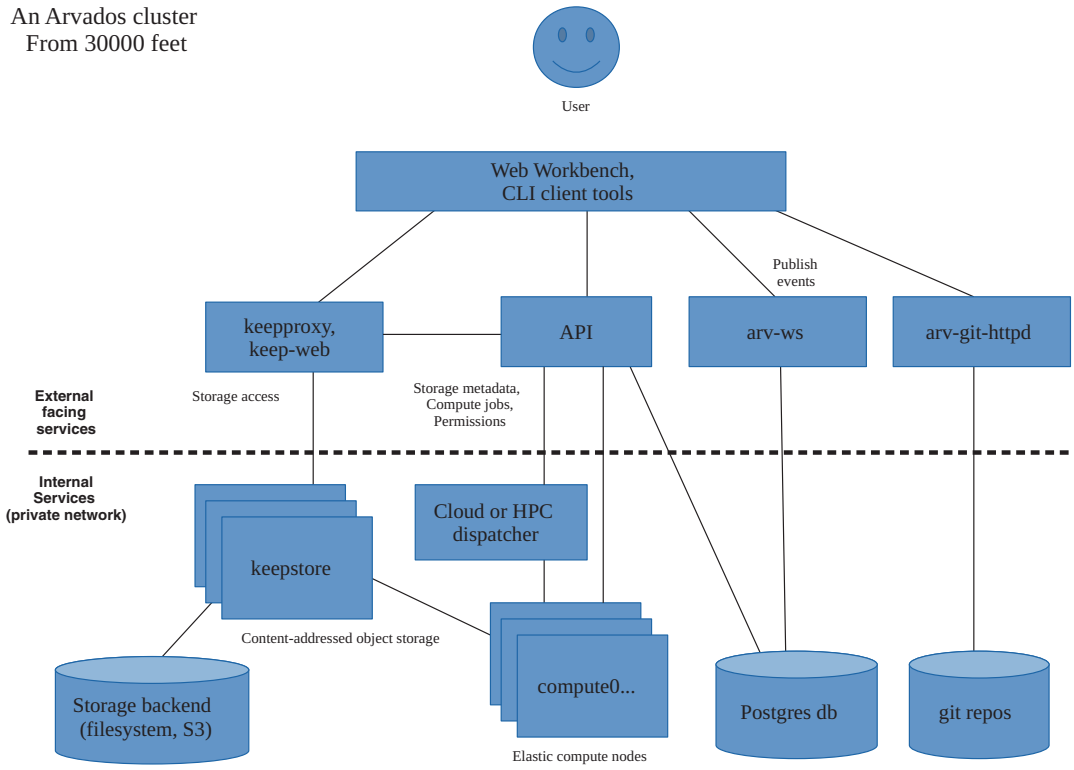

Fig. 2.7 The Arvados technical architecture

References

1. Laganà A, et al. Precision medicine for relapsed multiple myeloma on the basis of an integrative multiomics approach. JCO Precis Oncol. 2018;2018:1–17.
2. Berger MF, Mardis ER. The emerging clinical relevance of genomics in cancer medicine. Nat Rev Clin Oncol. 2018;15:353–65.
3. Johnson TM. Perspective on precision medicine in oncology. Pharmacotherapy. 2017;37:988–9.
4. Odle TG. Precision medicine in breast cancer. Radiol Technol. 2017;88:401M–21M.
5. Bødker JS, et al. Development of a precision medicine workflow in hematological cancers, Aalborg University Hospital, Denmark. Cancers (Basel). 2020;12:312.
6. Jäger N. Bioinformatics workflows for clinical applications in precision oncology. In: Seminars in cancer biology. Academic Press; 2021. https://doi.org/10.1016/j.semcancer.2020.12.020.
7. Altintas I, et al. Understanding collaborative studies through interoperable workflow provenance. In: Lecture notes in computer science. Berlin Heidelberg: Springer; 2010. p. 42–58.
8. Amstutz P, et al. Common workflow language, v1. 0. 2016.
9. Di Tommaso P, et al. Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017;35:316–9.
10. Mölder F, et al. Sustainable data analysis with Snakemake. F1000Res. 2021;10:33.
11. Amstutz P. Portable, reproducible analysis with arvados. F1000Res. 2015;4
12. Terra. https://terra.bio/
13. The Modern Standards Paradigm – 5 Key Principles. https://open-stand.org/about-us/principles/
14. Workflow Description Language (WDL). *OpenWDL* https://openwdl.org/
15. Boettiger C. An introduction to Docker for reproducible research. Oper Syst Rev. 2015;49:71–9.
16. Kurtzer GM, Sochat V, Bauer MW. Singularity: scientific containers for mobility of compute. PLoS One. 2017;12:e0177459.
17. Blankenberg D, Hillman-Jackson J. Analysis of next-generation sequencing data using Galaxy. Methods Mol Biol. 2014;1150:21–43.
18. Guimera RV. bcbio-nextgen: automated, distributed next-gen sequencing pipeline. EMBnet J. 2012;17:30.
19. Yukselen O, Turkyilmaz O, Ozturk AR, Garber M, Kucukural A. DolphinNext: a distributed data processing platform for high throughput genomics. BMC Genomics. 2020;21:310.

20. Desvillechabrol D, et al. Sequanix: a dynamic graphical interface for Snakemake workflows. Bioinformatics. 2018;34:1934–6.
21. Shen R, Seshan VE. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. Nucleic Acids Res. 2016;44:e131.
22. Deshwar AG, et al. PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. Genome Biol. 2015;16:35.
23. Andrews S, et al. FastQC: a quality control tool for high throughput sequence data. 2010.
24. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34:i884–90.
25. Lee S, et al. NGSCheckMate: software for validating sample identity in next-generation sequencing studies within and across data types. Nucleic Acids Res. 2017;45:e103.
26. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25:1754–60.
27. Koboldt DC. Best practices for variant calling in clinical sequencing. Genome Med. 2020;12:91.
28. Li X, Warner JL. A review of precision oncology knowledgebases for determining the clinical actionability of genetic variants. Front Cell Dev Biol. 2020;8:48.
29. Li MM, et al. Standards and guidelines for the interpretation and reporting of sequence variants in cancer: a joint consensus recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. J Mol Diagn. 2017;19:4–23.
30. Piñeiro-Yáñez E, et al. PanDrugs: a novel method to prioritize anticancer drug treatments according to individual genomic data. Genome Med. 2018;10(1):1–11.
31. Yu Y, et al. PreMedKB: an integrated precision medicine knowledgebase for interpreting relationships between diseases, genes, variants and drugs. Nucleic Acids Res. 2019;47:D1090–101.
32. Xu Q, et al. OncoPDSS: an evidence-based clinical decision support system for oncology pharmacotherapy at the individual level. BMC Cancer. 2020;20:740.

Check for updates

3 Somatic and Germline Variant Calling from Next-Generation Sequencing Data

Ti-Cheng Chang, Ke Xu, Zhongshan Cheng, and Gang Wu

Abstract

Re-sequencing of the human genome by next-generation sequencing (NGS) has been widely applied to discover pathogenic genetic variants and/or causative genes accounting for various types of diseases including cancers. The advances in NGS have allowed the sequencing of the entire genome of patients and identification of disease-associated variants in a reasonable timeframe and cost. The core of the variant identification relies on accurate variant calling and annotation. Numerous algorithms have been developed to elucidate the repertoire of somatic and germline variants. Each algorithm has its own distinct strengths, weaknesses, and limitations due to the difference in the statistical modeling approach adopted and read information utilized. Accurate variant calling remains challenging due to the presence of sequencing artifacts and read misalignments. All of these can lead to the discordance of the variant calling results and even misinterpretation of the discovery. For somatic variant detection, multiple factors including chromosomal abnormalities, tumor heterogeneity, tumor-normal cross contaminations, unbalanced tumor/normal sample coverage, and variants with low allele frequencies add even more layers of complexity to accurate variant identification. Given the discordances and difficulties, ensemble approaches have emerged by harmonizing information from different algorithms to improve variant calling performance. In this chapter, we first introduce the general scheme of variant calling algorithms and potential challenges at distinct stages. We next review the existing workflows of variant calling and annotation, and finally explore the strategies deployed by different callers as well as their strengths and caveats. Overall, NGS-based variant identification with careful consideration allows reliable detection of pathogenic variant and candidate variant selection for precision medicine.

Introduction

Germline variants are nucleotide changes in a germ or egg cells and can be passed to a child from parents during conception. Since the variants are in reproductive cells, they are hereditary mutations and can be passed to future genera-

Ti-Cheng Chang and Ke Xu contributed equally

T.-C. Chang (✉) · K. Xu · Z. Cheng · G. Wu
Center for Applied Bioinformatics, St Jude Children's Research Hospital, Memphis, TN, USA
e-mail: ti-cheng.chang@stjude.org

© Springer Nature Switzerland AG 2022

A. Laganà (ed.), *Computational Methods for Precision Oncology*, Advances in Experimental Medicine and Biology 1361, https://doi.org/10.1007/978-3-030-91836-1_3

tions. Germline mutations account for ~5–10% of cancers [1]. Somatic variants are variants that arose in any cells except germline cells, i.e., sperm and egg, and cannot be transmitted to progeny. Somatic variants include mosaicisms in different subsets of somatic cells including clonal hematopoiesis of indeterminant potential (CHIP). Somatic variants are of particular interests because they are associated with various human diseases, including cancers.

Traditional germline/somatic genetic testing relied on a "panel" of gene testing with a focus on hotspot variants in a number of well-characterized driver genes, such as BRCA1 and BRCA2 [2]. With the advances and reduced cost of the next-generation sequencing (NGS) technology, whole exome/genome sequencing (WES/WGS) and targeted sequencing have become an option for detecting variants on a much larger scale and higher definition. A major challenge of WGS/WES analysis is the accuracy of mutation calling analyses on single nucleotide variants (SNVs) and small *in*sertions and *del*etions (indels).

Development of SNV/Indel Variant Calling in the Past Years

NGS workflow usually starts with the fragmentation of the genome or targeted regions of genomes into small fragments, followed by alignments to reference genomes or genome re-assembly. The aligned/piled-up segments are used subsequently for variant detection. In early studies, the variant calling was performed by counting alleles at each site with simple cutoff rules to determine a variant call, which oftentimes lacks sensitivity to detect heterozygous alleles and does not provide confidence level of the genotype calls [3].

Uncertainties of variant calls arise when a sample's coverage is shallow, sequencing read quality is poor, or a variant site has low allele count support [4]. After variant calling, layers of filters are therefore suggested to be applied to filter the variant calls to reduce the likelihood of sequencing artifacts in the call sets and increase the confidence of variant calls. An in-depth overview of filters that can be considered is described in section "Contributing Factors for Bogus Somatic Variant Calling" of this chapter.

Germline and somatic variant calling algorithms differ in the assumption of expected allele frequency. Germline variants are expected to have 50% or 100% allele frequencies to differentiate three basic genotypes harbor at each variant site, e.g., homozygous allele A (AA), heterozygous (AB), or homozygous allele B (BB). On the contrary, for somatic variant calling, the allele frequency displays a larger spectrum of variations symbolizing distinct stages of cell development. An increasing number of algorithms have been developed in the past decades to enhance the calling accuracy by incorporating error rate estimation and probability frameworks to model the genotyping and phasing likelihoods. Given the complexity of genomes, local re-assembly was also placed into the calling scheme to increase the confidence of variant calling. Table 3.1 provides a summary of available tools for somatic and/or germline variant calling to date. In the following section, we will introduce the algorithms implemented in a few popular variant callers.

Algorithm Basis of Germline SNV/Indel Variant Calling

Samtools mpileup [5] deployed the approach of read coverage depth counting to identify coverage characteristics of potential SNVs/indel sites. The coverage information was then fed into BCFtools [6] for variant calling based on general Bayesian likelihood. This approach is usually used for germline variant calling.

GATK HaplotypeCaller [7] is a widely used germline variant caller. An advantage of GATK is that the algorithm can be applied for the joint calling of a group of samples at the same time to control the false discovery rate and increase the sensitivity of low-frequency variant detection. In addition, GATK allows the re-assembly of reads to re-construct the real allelic segment or haplotype, which will be realigned to the reference genome to identify the variant sites. GATK HaplotypeCaller begins with defining active

Table 3.1 List of publicly available tools for variant calling in chronological order

Software	Algorithm detail	Type of variant	Single-sample mode	Year published	References
GATK	Haplotype analysis and Joint genotype analysis	SNV/indel	Yes	2010	[7]
SAMtools	Joint genotype analysis	SNV/indel	Yes	2011	[6]
SomaticSniper	Joint genotype analysis	SNV	No	2011	[14]
MutationSeq	Machine learning	SNV	No	2012	[33]
JointSNVMix2	Joint genotype analysis	SNV	No	2012	[79]
VarScan2	Heuristic threshold	SNV/indel	Yes	2012	[16]
deepSNV	Allele frequency analysis	SNV	No	2012	[80]
LoFreq	Allele frequency analysis	SNV/indel	Yes	2012	[81]
FreeBayes	Haplotype analysis	SNV/indel	Yes	2012	[9]
EBCall	Allele frequency analysis	SNV/indel	No	2013	[82]
Shimmer	Heuristic threshold improved for highly contaminated or heterogeneous samples	SNV/indel	No	2013	[83]
Seurat	Bayesian-based analysis of sequenced genome pairs	SNV/indel, SV	No	2013	[84]
Virmid	Joint genotype analysis improved by inferring sample impurity	SNV	No	2013	[85]
qSNP	Heuristic threshold with low tumor content	SNV	No	2013	[86]
MuTect2	Allele frequency analysis	SNV/indel	Yes	2013	[13]
BAYSIC	Machine learning (ensemble caller)	SNV	No	2014	[87]
FaSD-somatic	Joint genotype analysis	SNV	Yes	2014	[88]
Platypus	Haplotype analysis	SNV/indel, SV	Yes	2014	[89]
HapMuC	Haplotype analysis	SNV/indel	Yes	2014	[90]
RADIA	Heuristic threshold with RNA and DNA integrated analysis	SNV	No	2014	[91]
SOAPsnv	An integrated tool for somatic single-nucleotide variants detection with or without normal tissues in cancer genome	SNV	No	2014	[92]
SomaticSeq	Machine learning (an ensemble approach to detect somatic mutations)	SNV	No	2015	[34]
LocHap	Haplotype analysis	SNV/indel	No	2015	[93]
VarDict	Heuristic threshold	SNV/indel, SV	Yes	2016	[94]
SNVSniffer	An integrated caller for germline and somatic single-nucleotide and indel mutations	SNV/indel	Yes	2016	[95]
MuSE	Markov chain model	SNV	No	2016	[17]
SNooPer	Machine learning for low-pass next-generation sequencing	SNV/indel	Yes	2016	[96]

(continued)

Table 3.1 (continued)

Software	Algorithm detail	Type of variant	Single-sample mode	Year published	References
CaVEMan	Joint genotype analysis	SNV	No	2016	[97]
LoLoPicker	Allele frequency analysis	SNV	No	2017	[98]
Strelka2	Mixture-model-based estimation for calling of germline and somatic variants	SNV/indel	No	2018	[18]
Cerebro	Machine learning (random forest)	SNV/indel	No	2018	[35]
DeepVariant	Deep convolutional neural network (CNN) to call germline SNV/indel	SNV/indel	No	2018	[12]
NeuSomatic	Convolutional neural network	SNV/indel	No	2019	[99]
NeoMutate	An ensemble machine learning framework	SNV/indel	No	2019	[29]
SMuRF	Machine learning	SNV/indel	No	2019	[31]
DeepSSV	Convolutional neural network	SNV/indel	No	2020	[100]

regions where abundant evidence has shown the presence of variants. Only the active region is used for variant calling to reduce the time on the assembly. With the assembly step, the variant calling is not only dependent on the read alignment against the reference genome but also the reconstructed haplotype. The overall GATK algorithm takes a divide-and-conquer concept by shredding the sequencing data into small chunks for parallel processing; however, its efficiency is still a concern when processing a large collection of samples for joint calling. Approaches have been proposed to address the performance issue when dealing with a large number of samples [8].

FreeBayes [9] applied a Bayesian framework to relate the likelihood of sequencing errors of the reads and the prior likelihood of a particular genotype. Also, the phase of haplotypes was inferred from the reads, and the non-uniform copy number of samples was taken into consideration. FreeBayes is usually used for germline variant calling, while it has been expanded for somatic calling [10]. FreeBayes shows good performance across sequencing platforms for SNV calling, but it tends to have a higher false-positive rate for indel sites [11].

DeepVariant [12] performs variant detection using a convolutional neural network (CNN) learning model implemented via the python TensorFlow library. DeepVariant identifies variants through learning the features in images of pileup reads surrounding putative variants and true genotypes. A version of DeepVariant for somatic calling is still under development.

Algorithm Basis of Somatic SNV/Indel Variant Calling

Mutect2 [13] as a part of the GATK toolkit shares a similar process of variant calling with GATK and is mainly used for somatic calling with matched, paired tumor-normal samples. Mutect2 also allows tumor only calling (see section "SNV/Indel Variant Calling"). Mutect2 calls SNVs and indels simultaneously via the local de novo assembly of haplotypes in an active region as described previously. Mutect2 reassembles the reads present in the active regions to candidate variant haplotypes. Each read is then aligned to each haplotype via the Pair-HMM algorithm to obtain a matrix of likelihoods. Finally, log odds were derived to distinguish somatic variants from sequencing errors by a Bayesian somatic likelihood model.

SomaticSniper [14] is another somatic variant caller. SomaticSniper determines the somatic status of a variant site by comparing the site's genotyping likelihood between normal and tumor derived from the MAQ tool [15] using a Bayesian approach. SomaticSniper implemented internal filters to exclude the sites with poor read/base quality or with low read support to reduce calling artifacts.

VarScan2 [16] relies on the results from SAMtools pileup or mpileup for somatic variant calling. At each variant site, VarScan2 compares the genotypes and supporting read counts between tumor and normal to determine the somatic status, and the call-set is refined with post-calling filters including the variant position in a read, strand bias, read coverage depth, variant frequency, homopolymer, mapping quality, and so on [16]. Of note, VarScan2 also allows the germline variant calling and detection of somatic copy number abnormality (SCNA).

MuSE [17] somatic calling starts with matched tumor-normal alignment BAM files. The alignment is first filtered for sequencing artifacts. The evolutionary F81 Markov substitution model of DNA is applied to describe the changes from reference to tumor allele compositions with estimates of equilibrium frequencies for all alleles and evolutionary distance. With the frequencies, MuSE derived a sample-specific error model and five-tier-based cutoffs to address the variations present in the frequency distribution in tumor and normal samples. The tier-based approach allows the MuSE to retain variants with low variant allele frequency to achieve a higher sensitivity.

Strelka2 [18] is an open-source somatic/germline variant caller developed by Illumina®. The somatic calling algorithm of Strelka2 is enhanced based on the original Strelka [19] method to account for tumor-in-normal contamination that is essential for liquid tumor variant analyses.

Strelka first identifies indel regions and performs realignment. After realignments, Strelka derives a somatic variant probability using the tumor and normal samples and deduces the somatic status of a site after accounting for the status of loss of heterozygosity (LOH) or copy number change regions. Strelka applied a two-tier-based filtering strategy with distinct filters and sensitivity. Similar to other tools, post-filtering is applied by Strelka2 to handle different types of potential calling errors.

The variant calling is usually computationally intensive, particularly when the sample number is large. To improve efficiency, Illumina® has released a Dynamic Read Analysis for GENomics (DRAGEN) platform using a highly configurable field-programmable gate arrays (FPGAs) hardware to accelerate the analysis processes [20]. DRAGEN first identifies callable regions and assembles the haplotypes using *De Bruijn* graph method. The reassembly is aligned to the reference genome to identify the variants. The probability of all read alignments to the haplotype is calculated via the pair hidden Markov model that is speeded up using the FPGA and summed up for each read. In the end, the diploid genotype is calculated to determine the variant calls.

In the past few years, GPU-based read alignment and variant calling solutions have also been developed to reduce the WGS data processing time to a couple of hours. For example, NVIDIA Clara Parabricks pipelines include a somatic variant calling workflow that integrates GPU-based alignments by BWA-MEM and downstream somatic variant calling by Mutect2 [13] or DeepVariant [12]. Parabrick also allows germline calling using GATK HaplotypeCaller [7]. The pipeline reduces the time taken for a typical 30× WGS data by over an order of magnitude.

SNV/Indel Variant Calling Workflows

Variant calling workflow can be compartmentalized into four steps: data preprocessing, variant calling, variant filtering, and variant annotation. Each step has its challenges and strategies. We detail these steps as follows.

Data Preprocessing

The raw read quality can be examined using FastQC [21]. FastQC identifies the potential read issues before mapping. A good WGS/WES read library usually has an average read base quality >20 and a low level of duplicated or overrepresented sequences.

Selection of the reference genome is the first step for correct variant calling. The latest version of the human reference genome GRCh38 (Hg38) with improved resolution [22] is suggested for human variant analyses. Also, the reference is recommended to include decoy genome sequences for the alignment purpose to reduce misalignments, as well as virus sequences that are known in human to attract the viral reads. In addition, the alternative contigs from highly complex loci, such as the human HLA allele region, should be included to reduce SNV/indel calling artifacts. For read alignments, frequently used aligners are BWA [5], Bowtie2 [23], and Novoalign (http://www.novocraft.com/products/novoalign/). Benchmarks of short-read aligners indicated that the MEM algorithm implemented in BWA achieved a better balance between specificity and sensitivity [24, 25]. BWA-MEM is suggested to use when read length is greater than 70, while BWA-ALN for shorter reads [26].

Following alignments, duplicate reads generated from PCR artifacts are flagged using tools such as GATK MarkDuplicates to prevent downstream variant calling errors. Incorrect read alignment surrounding the indel regions frequently causes inaccurate substitution calls. These alignment artifacts can be reduced through indel realignments by GATK IndelRealigner or similar tools. Furthermore, the base quality produced by different library preparation protocols and sequencing instruments would have different levels of technical or chemistry errors. GATK toolkits comprised two tools, BaseRecalibrator and ApplyBQSR, to facilitate the correction of these systematic errors. These tools implemented

machine learning approaches to model errors and adjust base qualities to obtain a more accurate overall base quality profile. Figure 3.1a shows a general workflow for the data preprocessing.

SNV/Indel Variant Calling

The next step is to choose appropriate variant callers. The GATK tool suite is well performed for the germline SNV/indel calling. A number of best practices for variant callings have been provided by GATK (https://gatk.broadinstitute.org/hc/en-us/sections/360007226651-Best-Practices-Workflows). For somatic variant calling, accurate identification of a somatic variant is still not trivial due to varied caller performance and tumor heterogeneity. Below we describe three common scenarios in somatic and germline variant calling as well as variant prioritization in cancer genomics.

Somatic Mutation Calling on Matched Tumor-Normal Pairs

Variant calling with matched tumor-normal sample pairs is the most common scenario for the identification of somatic variants (Fig. 3.1b). Most of the callers use the aligned BAM files of paired tumor and normal samples as the standard inputs. To identify low-frequency variants, a caller that can model the allele frequency is suggested, such as Mutect2, MuSE, and Strelka2 as detailed in the Introduction. Due to the differences of underlying algorithms and statistic modeling, the somatic variant callers differ in sensitivity and specificity when detecting variants at different levels of variant allele frequencies (VAF) [27]. Compared with Strelka and Mutect, SomaticSniper has a lower sensitivity and specificity when calling the variants with VAF <8%. However, the performance of SomaticSniper is comparable with Strelka and Mutect for variants with VAF >18%. The sensitivity of VarScan2 was increased with lower minimum allele fraction thresholds, which was however compromised with reduced specificity [28]. Therefore, a careful setting of thresholds to

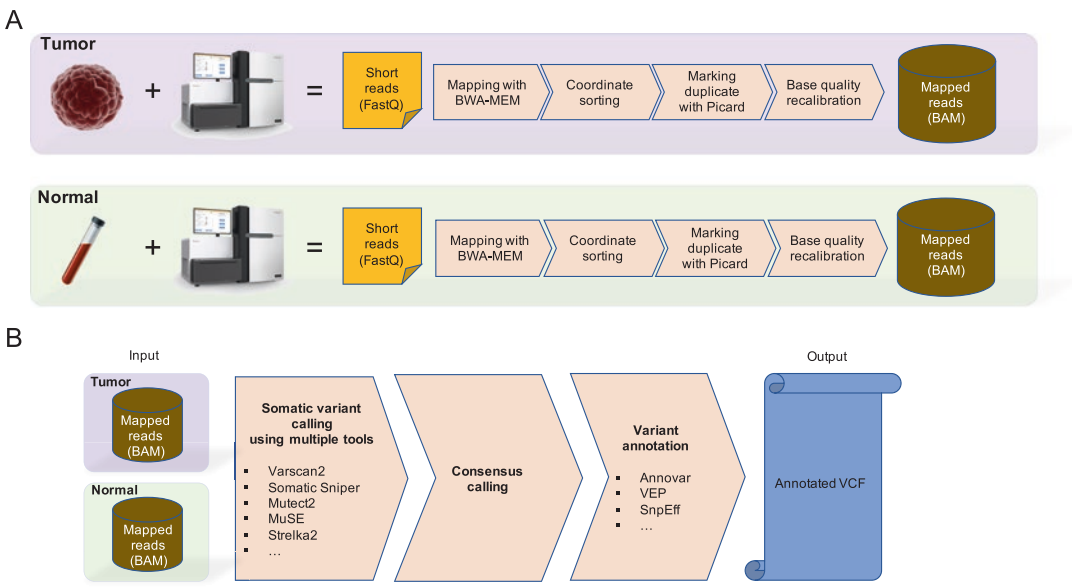

Fig. 3.1 The workflow of the somatic variant calling of paired tumor-normal samples. (**a**) Data preprocessing steps from sample preparation to short reads mapping and calibration into binary version of Sequence Alignment/Map (BAM) files for paired tumor and normal samples. (**b**) Variant calling and annotation steps from paired tumor-normal BAM files to annotated somatic variants in VCF format

achieve a balance between sensitivity and specificity for each caller and a well-considered post-calling filtering strategy play important roles to assure the validity of final call sets.

Given the complex heterogeneity and structural rearrangements of tumor tissue, finding an appropriate somatic variant caller along with parameter fine-tuning and development of a solid calling strategy remain a major challenge for cancer genomics. To tackle this complexity and exploit each caller's strength, a consensus voting to determine a valid variant call by multiple callers has gradually become a prevalent strategy in studies [29–33]. In addition to a simple voting strategy, machine learning has been incorporated into the consensus calling steps to improve calling performance. MutationSeq incorporated multiple sequence quality features derived from normal data based on Samtools and GATK, along with several sequence artifacts and low-frequency variant features to build classifiers to determine the somatic variants [33]. SomaticSeq [34] integrated five somatic callers from which feature sets were identified for each candidate variant position to build a classifier using a stochastic boosting machine-learning algorithm. Cerebro [35] applied a random forest classification model to generate a confidence score for each candidate variant derived from whole-exome sequencing data, which is limited to the coding region with >150× coverage. These approaches generally lack portability, i.e., users are required to obtain appropriate training data and have knowledge about the machine learning to re-train the models. In light of these issues, SMuRF [31] was developed and generalized for either WGS or WES data. SMuRF implemented a supervised machine learning using features derived from four variant callers along with mapping auxiliary features. NeoMutate [29], as another machine learning based caller, profiled a collection of seven distinct classifiers based on a training dataset of >3000 cancer variants from the Catalogue of Somatic Mutations in Cancer (COSMIC) database [36].

Machine learning–based callers determine the somatic status of a variant through different features of a variant harbors and therefore offer a higher level of flexibility than rule-based filtering strategy, especially for the tumor samples with intra-heterogeneity and normal tissue admixtures. However, a detailed curation of a set of ground-truth training data including both true-positive and true-negative variants is the key to optimize and refine the training models.

Mutation Calling and Prioritization on Tumor Sample Without Matched Normal Sample

In large-scale cancer genomic projects, it is common to have tumor samples without matched normal samples or with tumor-contaminated adjacent normal samples, due to the difficulties to collect patients' blood samples. In these cases, the somatic variant calling oftentimes has a high rate of false positives, because it is almost impossible to confidently determine whether a called variant is of germline origin or somatically acquired. Mutect2 can call somatic mutations in tumor-only mode; however, the calling results require careful filtering for false positives due to the deficiency of corresponding germline information. Common germline SNPs can be eliminated by filtering against appropriate human genome variation databases such as Genome Aggregation Database (gnomAD). To date, limited number of studies have compared the performance of Mutect2 tumor-only and tumor/normal calling modes when both tumor/normal WGS/WES data are available. A tool designed specifically for somatic mutation calling on tumor-only WES samples is ISOWN [37], which utilizes a family of supervised learning classifications to distinguish somatic SNVs in NGS data from SNPs in the absence of normal samples. In terms of performance, the F1-measure of ISOWN is between 75.9% and 98.6% across different cancer types, cell lines, fresh frozen tissues, and formalin-fixed paraffin-embedded tissues. Calling somatic variants in tumor only WGS/WES data still warrants further improvement.

Due to these challenges, one can consider focusing on identifying putatively pathogenic variants in a set of genes of interest to specific tumors, irrespective of their germline or somatic origin (Fig. 3.2). Specifically, after basic variant

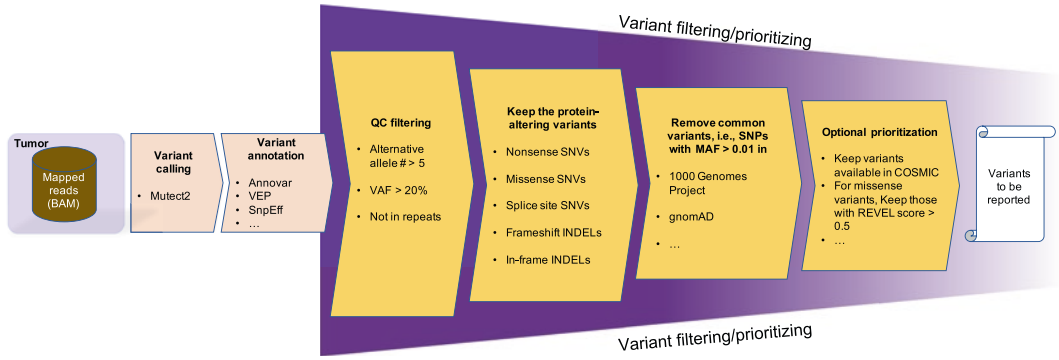

Fig. 3.2 The workflow of the variant calling of tumor sample without a matched normal sample. The workflow focuses on reporting potentially pathogenic variants regardless of their tumor or germline origin

quality filtering such as keeping variants with higher alternative allele count (>5) and VAF (>20%), and excluding those located in regions of low complexity or regions with extreme GC content, additional filters can be applied for the variant class and population frequency filter, i.e., only keeping protein-altering variants with minor allele frequency <0.01 in population frequency databases such as 1000 Genomes [38] and gnomAD [39]. In addition, optional filters can be added to increase the calling confidence such as keeping any variants that are available in the COSMIC catalog of somatic mutations or missense variants with a REVEL score >0.5 [36, 40].

Germline Mutation Calling and Prioritization

Identifying germline mutations in cancer predisposition genes has important implications in understanding tumorigenesis and guiding clinical practice. A common germline mutation calling workflow is illustrated in Fig. 3.3a. The recommended germline variant calling follows the GATK best practices including read mapping, alignment sorting, duplicated reads marking, and variant calling by GATK HaplotypeCaller [7]. Also, joint variant calling in multiple germline samples is recommended whenever possible because the genotype information at the population level can be leveraged to rescue the variant at a site with low coverage or with lower quality in a sample. The efficiency of GATK calling can be enhanced by a divide-and-conquer strategy, i.e., splitting the genomes into multiple small chunks for parallel variant calling followed by merging the output variant files (VCFs). After variant calling, the GATK Variant Quality Score Recalibration (VQSR) method is the suggested approach to filter the germline variants. VQSR relies on a deep learning method and therefore requires a sufficient amount of the variant sites to establish a reliable training model. The variant number for a single-sample WGS is usually sufficient for VQSR; however, for WES data, at least 30 samples are required to perform VQSR. When the sample size is limited, the variant call set can be filtered by the GATK VariantFiltration tool.

To narrow down from the vast amount of germline variants reported by germline variant caller, usually only rare, non-silent coding variants in cancer-related genes, such as autosomal dominant or autosomal recessive cancer-predisposition genes, or genes that are recurrently mutated in tumors, are considered. For example, Zhang et al. evaluated germline mutations in a cohort of pediatric cancers in a curated list of 565 cancer-related genes based on expert reviews of the genes from American College of Medical Genetics and Genomics (ACMG) and genes from related literatures [41]. Specifically, after germline variant calling, QC-passed variants are shortlisted based on their frequencies in human populations such that only novel variants or the variants with minor allele frequency <0.001 in NHLBI Exome Sequencing Project (ESP) are kept [42]. These shortlisted variants

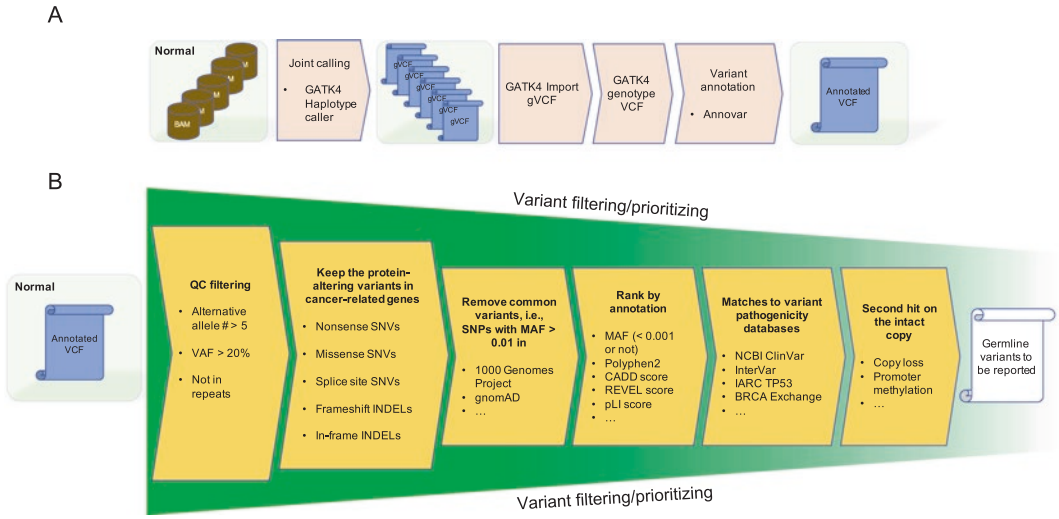

Fig. 3.3 The workflow of germline variant calling and prioritization. (**a**) Steps of the joint calling of germline variants from pooled germline BAM files. (**b**) Steps of filtering and prioritizing potentially pathogenic germline variants or variants of unknown significance

can be then ranked based on (1) mutational class such as nonsense SNVs, missense SNVs, splice site SNVs, frameshift indels, or in-frame indels; (2) functional annotation databases such as PolyPhen2 and MutationAssessor [43, 44], (3) matches to curated variant pathogenicity databases such as NCBI ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/), locus-specific databases such as IARC TP53 (https://p53.iarc.fr/) and BRCA Exchange (https://brcaexchange.org/); and (4) second hit on the intact copy in the tumor genome due to one copy loss or promoter methylation of the intact copy. Other popular databases for germline variant classification and prioritization include pLI and LOFTEE scores for loss-of-function variant prioritization [39, 45]; REVEL and CADD scores for missense variant prioritization [40, 46]; and dbscSNV scores for splice variant prioritization [47]. In addition, InterVar, an automatic interpretation of variants based on dozens of criteria laid out by ACMG and Association for Molecular Pathology (AMP), can be included to aid manual review of clinical significance [48]. Figure 3.3b summarizes the filtering steps to prioritize germline variants to be reported. The final ranked list of putatively pathogenic germline variants will then need to be manually reviewed and validated based on phenotype data, RNA-seq, and literature review. The whole prioritization process before manual reviews can be automated. For example, St. Jude Pediatric Cancer Variant Pathogenicity Information Exchange (PeCan PIE, https://pecan.stjude.cloud/pie), a free cloud service for non-commercial use, offer variant annotation and ranking service based on MedalCeremony pipeline to triage the germline variants into three categories, including Gold, Silver, and Bronze [41, 49].

Variant Annotation

To understand the context of the germline variants and somatic mutations, several tools are available to perform variant annotation on the called variants. Typically, the genomic locations of the variants are compared against a gene-based annotation database such as a GENCODE release (https://www.gencodegenes.org/pages/data_access.html) to determine if a variant is exonic, intronic, or intergenic [50]. Variants in exonic regions are further classified as missense

variants, nonsense variants, silent variants, splice acceptor variants, splice donor variants, splice region variants, in-frame indels, and frameshift indels. Some annotation tools such as ANNOVAR [51], VEP [52], and SnpEff [53] also add population allele frequency from 1000 Genomes Project [38], NHLBI ESP [42], Exome Aggregation Consortium (ExAC) [45], and gnomAD [39]; and provide comparative genomics-based scores such as GERP++ [54], SIFT [55], PolyPhen2 [43]; and include machine learning–based pathogenicity scores such as CADD [46, 56] and REVEL [40].

ANNOVAR [51] is an annotation pipeline to functionally annotate variants. The workflow can be performed for either gene-based coding change annotations or region-based non-genic genomic element annotations. Moreover, ANNOVAR has extended functionality to identify and filter variants documented in specific databases, which can be used for enriching causal variants in diseases. ANNOVAR allows the annotation of SNVs and structural variants from a standard VCF. A web interface is available via wANNOVAR (http://wannovar.wglab.org/).

VEP [52] is another popular toolkit for variant annotation. Compared to ANNOVAR, VEP provides cell-line-based annotation. VEP generates transcript-level annotations, while ANNOVAR gives gene-level annotations. LOFTEE (Loss-Of-Function Transcript Effect Estimator, https://github.com/konradjk/loftee) is a very useful VEP plugin to evaluate the loss of function of splice variant [39]. VEP also allows the variant annotation of species other than human and mouse. In addition to local installation, users can perform annotations through the VWP web interface (https://uswest.ensembl.org/info/docs/tools/vep/online/index.html) or cloud virtual machine.

SnpEff [53] implements an interval forest algorithm to efficiently query, annotate, and predict the effect of the variants. SnpEff can run locally or via a Galaxy instance. Similar to VEP, SnpEff also provides a cloud VM for users. SnpEff allows the assessment of nonsense mediated decay (NMD), a functionality absent from ANNOVAR and VEP.

Contributing Factors for Bogus Somatic Variant Calling

Somatic variants generated from the variant callers oftentimes include false positives due to various types of contributing factors. Below we describe four common scenarios that cause bogus somatic variants calling and need to be considered in postprocessing.

Strand Bias

Strand bias is observed when reads are favorably sequenced for one strand over the other; only one strand of the DNA has reads covered in extreme cases. The sources of this type of artifact remain elusive but may be relevant to library preparation of analytic procedures [57]. This bias raises the concerns of variant call accuracy. GATK and Samtools both implement functionality to calculate strand bias scores.

Repetitive DNA Sequences

Repetitive DNA sequences are sequences that are identical or similar across the genome. They vary in sizes and frequencies and cause mapping ambiguities. RepeatMasker [58] can be used to mark or mask the repetitive sequences in the genome to reduce such ambiguities. The error rate of short reads sequencing has been shown to increase in genomic regions with high- and low-GC content or with long homopolymer runs [59]. Also, the GC-rich regions frequently suffered from low coverage issues. Segmental duplication can also cause some reads mapped to multiple places in the genome and give rise to unusual coverage. A BLAT (BLAST-like alignment tool, available at http://genome.ucsc.edu/cgi-bin/hgBlat) search can be used to determine if the flanking sequence of a variant with high coverage is uniquely mapped to a locus or multiple different loci. Those that can be mapped to multiple loci in the genome are recommended to be reviewed manually.

Variants in simple repeats or homopolymer regions, such as CCCCCCCC or ACGACGACGACG ($[ACG]_n$), often lead to false-positive variant calls due to sequencing errors and following read misalignments. Indels

in repetitive regions coupled with low alternative reads count support are usually filtered out. However, frameshift indels in disease-causing genes (e.g., ATRX, PMS2) require careful visual inspection and perhaps validation with an orthogonal sequencing approach to avoid missing important findings.

Low-Frequency Variants

VAF is the number of reads supporting the alternative allele divided by the total number of reads covering the genomic location. For germline samples, a heterozygous germline variant would have an approximately 50% VAF. Germline variants with significantly low VAF and a low number of alternative reads count could be due to sequencing errors. Germline variants with sufficient alternative read count and total read count but with low VAF may indicate mutation mosaicism [60]. If a large number of germline variants have low VAFs, it may suggest that the normal sample is contaminated by the tumor sample, which sometimes happens when the normal sample is collected as tissue adjacent to the tumor or blood after treatment. Paralogous mapping can also lead to VAF ranging from 10% to 25%.

Somatic mutations, on the other hand, exhibit a broader range of VAFs. A heterozygous somatic mutation in a copy-intact region would have an approximately 50% VAF. However, since tumor genomes are frequently subject to copy number alteration, the VAF of a somatic mutation could be around 33% or 67% due to one copy gain and could be close to 100% because of LOH. In addition, since patient tumor samples are rarely 100% pure, low tumor purity may further contribute to the global dilution of VAFs of somatic mutations in a tumor genome. Mutations with significantly lower VAFs than the truncal mutations in a tumor genome but with sufficient mutant read counts may suggest that they are subclonal. Somatic mutations with significantly low VAF and few alternative allele read counts could be due to sequencing error/artifacts and are recommended to be filtered out.

Germline Variant Contamination

A few somatic SNV callers, e.g., Mutect, have implemented specific filters to eliminate the potential germline variant contamination in somatic variants calling. Mutect allows the inclusion of a panel of normal samples (PON) and dbSNP database to exclude germline variants. The germline variant contamination can also be reduced by checking minor allele frequencies of mutations across different population frequency databases such as gnomAD and the 1000 Genome Project database. A recent study [61] reported that there would be one germline SNP among a median somatic SNVs prediction set containing 4325 somatic SNVs; the study also reported a negative correlation between germline SNP contamination and tumor purity.

Concluding Notes

Somatic variant calling from WGS/WES is critical for cancer genomics as it not only depicts the mutational landscape for a tumor sample but also serves as input data for downstream analyses such as mutational signature and clonal evolution. Consequently, there has been great interest in developing fast, accurate, and scalable methodologies and tools for variant calling across academia and industry. In addition to the tools mentioned above, there are also other variant calling tools acting on different data types and different platforms as described below.

Mitochondria Mutation Calling

Variants present in the mitochondria genome (mtDNA) is implicated in a wide spectrum of human disorders and diseases with highly divergent phenotypes and penetrance. The challenges of mtDNA variant calling arise from the circular topology of mtDNA as well as the homology between mtDNA and a part of the nuclear genome with mitochondrial origin (nuMTs). The mtDNA mutation load also varies greatly among tissues

and organs from heteroplasmy (<100%) to homoplasmy (100%). The Human Mitochondrial Genome Database, Mitomap [62], provides a repertoire of reported mtDNA variants. Nuclear genome variant callers such as VarScan and LoFreq have been used for identifying the somatic mtDNA variants [63, 64]. MitoCaller [65] of the MitoAnalyzer toolkit was designed specifically to infer the mutation status of each position of the mitochondria genome using likelihood-based models and adapted an iterative alignment strategy to account for the circularity of the mtDNA genome. Importantly, discrepancies of mtDNA variant calling have been reported when using different reference genome and enrichment strategies [64], which should be taken into consideration when performing mtDNA variant calling and interpretation.

Long-Read Variant Calling

While short reads from paired-end sequencing were used by most state-of-the-art SNV callers to accurately detect variations in diploid genomes, they provide limited haplotype information that is required by some SNV callers, such as GATK HaplotyperCaller and FreeBayes. In addition, the accurate calling of SNVs in repetitive regions of the human genome is another challenge. Third-generation sequencing (TGS) technologies, including Pacific Biosciences and Oxford Nanopore (ONT), have the potential to overcome the limitations of short-read sequencing. Nevertheless, compared to short-read sequencing, long-read sequencing usually costs more and generates less-accurate long reads (e.g., sporadic indels in ONT data), posing challenges for accurate variant detection [66]. Current SNV callers using TGS data are mostly designed for germline variants calling and usually optimized based on the publicly available data from the Genome in a Bottle (GIAB) Consortium. Somatic SNV calling based on long reads technology is still underdeveloped.

NGS-based mapping tool such as BWA-mem is not suitable for long reads mapping. Instead, new mapping tools such as Minimap2 [67] and NGMLR [68] have been developed specifically for long reads mapping. Similarly, NGS-based SNV calling tools such as GATK HaplotyperCaller and FreeBayes are not recommended for variant calling on long-reads sequencing data. Instead, several variant callers have been developed specifically for long-reads data to leverage haplotype information available in long reads to improve the accuracy to call and phase SNVs in diploid genomes, as well as mapping variants in duplicated regions of the genome that are not possibly mapped using short reads. For example, Longshot [66] takes advantage of the haplotype information present in PacBio long reads to improve the SNV calling accuracy [69]. WhatsApp [69] introduces a novel statistical framework for the joint inference of haplotypes and genotypes from noisy long reads, which takes full advantage of linkage information provided by PacBio long reads. Clairvoyante [70] uses a multi-task five-layer convolutional neural network model to predict variants. Other tools include DeepVariant for variant calling on PacBio data [12] and MarginPhase (https://github.com/benedictpaten/marginPhase) for simultaneous haplotyping and genotyping on Oxford Nanopore data.

Different tools differ in their precision and recall rate. In a benchmark study using PacBio data from GIAB, three callers, including Longshot, WhatsApp, and Clairvoyante, demonstrating very similar performance [66]. Compared to the previous three tools, MarginPhase performed moderately when focused on GIAB high confidence regions [69]. Another software, HELLO [71], has been created to integrate the short read and long read data to improve the robustness of SNV calling by leveraging the Mixture of Experts paradigm that uses an ensemble of deep neural networks (DNNs).

Variant Calling in Single-Cell Data

Single-cell sequencing has been the hotspot of functional genomics to elucidate the heterogeneity of cell compositions. Variant calling of single-cell data can aid the inference of the lineage relationship of cells. Although challenges remain

for large-scale single-cell WGS/WES in terms of experimental design complexity and sequencing cost currently, single-cell RNA sequencing (scRNA) has been applied broadly to examine cell population dynamics and track the development of cell lineages. The preprocessing steps for scRNA data are relatively similar to the usual practice of WGS/WES calling. However, splicing-aware aligners, e.g., STAR [72] or GSNAP [73], are suggested for the read alignment. There are still not many callers designed specifically for single-cell data [74]. Trinity Cancer Transcriptome Analysis Toolkit (CTAT) is one caller with extended functionality for scRNA-seq SNV detection. SCIΦ is another tool that can perform jointly calling of mutations in individual cells followed by an estimation of the tumor phylogeny [75]. SSrGE [76] is an integrative workflow to connect genotype and phenotype in single-cell data which implemented GATK best practice and FreeBayes for variant inference. A few other studies used SAMtools mpileup approach for variant identification [77, 78]. Solid variant calling strategies in single-cell data will be of great needs in the following years.

References

1. Hampel H, Bennett RL, Buchanan A, Pearlman R, Wiesner GL, Guideline Development Group, American College of Medical Genetics and Genomics Professional Practice and Guidelines Committee and National Society of Genetic Counselors Practice Guidelines Committee. A practice guideline from the American College of Medical Genetics and Genomics and the National Society of Genetic Counselors: referral indications for cancer predisposition assessment. Genet Med. 2015;17(1):70–87.
2. Velazquez C, Lastra E, Avila Cobos F, Abella L, de la Cruz V, Hernando BA, Hernandez L, Martinez N, Infante M, Duran M. A comprehensive custom panel evaluation for routine hereditary cancer testing: improving the yield of germline mutation detection. J Transl Med. 2020;18(1):232.
3. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet. 2011;12(6):443–51.
4. Li H. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics. 2014;30(20):2843–51.
5. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
6. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27(21):2987–93.
7. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297–303.
8. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, Kling DE, Gauthier LD, Levy-Moonshine A, Roazen D, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv. 2018;2018:201178.
9. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:12073907. 2012.
10. Chiang C, Layer RM, Faust GG, Lindberg MR, Rose DB, Garrison EP, Marth GT, Quinlan AR, Hall IM. SpeedSeq: ultra-fast personal genome analysis and interpretation. Nat Methods. 2015;12(10):966–8.
11. Hwang S, Kim E, Lee I, Marcotte EM. Systematic comparison of variant calling pipelines using gold standard personal exome variants. Sci Rep. 2015;5:17875.
12. Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, Newburger D, Dijamco J, Nguyen N, Afshar PT, et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol. 2018;36(10):983–7.
13. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, Getz G. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013;31(3):213–9.
14. Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, Dooling DJ, Ley TJ, Mardis ER, Wilson RK, Ding L. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics. 2012;28(3):311–7.
15. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 2008;18(11):1851–8.
16. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22(3):568–76.
17. Fan Y, Xi L, Hughes DS, Zhang J, Zhang J, Futreal PA, Wheeler DA, Wang W. MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. Genome Biol. 2016;17(1):178.
18. Kim S, Scheffler K, Halpern AL, Bekritsky MA, Noh E, Kallberg M, Chen X, Kim Y, Beyter D,

Krusche P, et al. Strelka2: fast and accurate calling of germline and somatic variants. Nat Methods. 2018;15(8):591–4.
19. Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics. 2012;28(14):1811–7.
20. Miller NA, Farrow EG, Gibson M, Willig LK, Twist G, Yoo B, Marrs T, Corder S, Krivohlavek L, Walter A, et al. A 26-hour system of highly sensitive whole genome sequencing for emergency management of genetic diseases. Genome Med. 2015;7:100.
21. Andrews S. FastQC: a quality control tool for high throughput sequence data [Online]. Available online at: http://www.bioinformaticsbabrahamacuk/projects/fastqc/ 2010.
22. Pan B, Kusko R, Xiao W, Zheng Y, Liu Z, Xiao C, Sakkiah S, Guo W, Gong P, Zhang C, et al. Similarities and differences between variants called with human reference genome HG19 or HG38. BMC Bioinformatics. 2019;20(Suppl 2):101.
23. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–9.
24. Highnam G, Wang JJ, Kusler D, Zook J, Vijayan V, Leibovich N, Mittelman D. An analytical framework for optimizing variant discovery from personal genomes. Nat Commun. 2015;6:6275.
25. Thankaswamy-Kosalai S, Sen P, Nookaew I. Evaluation and assessment of read-mapping by multiple next-generation sequencing aligners based on genome-wide characteristics. Genomics. 2017;109(3–4):186–91.
26. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:13033997. 2013.
27. Xu C. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. Comput Struct Biotechnol J. 2018;16:15–24.
28. Xu H, DiCarlo J, Satya RV, Peng Q, Wang Y. Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. BMC Genomics. 2014;15:244.
29. Anzar I, Sverchkova A, Stratford R, Clancy T. NeoMutate: an ensemble machine learning framework for the prediction of somatic mutations in cancer. BMC Med Genet. 2019;12(1):63.
30. Shin HT, Choi YL, Yun JW, Kim NKD, Kim SY, Jeon HJ, Nam JY, Lee C, Ryu D, Kim SC, et al. Prevalence and detection of low-allele-fraction variants in clinical cancer samples. Nat Commun. 2017;8(1):1377.
31. Huang W, Guo YA, Muthukumar K, Baruah P, Chang MM, Jacobsen Skanderup A. SMuRF: portable and accurate ensemble prediction of somatic mutations. Bioinformatics. 2019;35(17):3157–9.
32. Wang M, Luo W, Jones K, Bian X, Williams R, Higson H, Wu D, Hicks B, Yeager M, Zhu B. SomaticCombiner: improving the performance of somatic variant calling based on evaluation tests and a consensus approach. Sci Rep. 2020;10(1):12898.
33. Ding J, Bashashati A, Roth A, Oloumi A, Tse K, Zeng T, Haffari G, Hirst M, Marra MA, Condon A, et al. Feature-based classifiers for somatic mutation detection in tumour-normal paired sequencing data. Bioinformatics. 2012;28(2):167–75.
34. Fang LT, Afshar PT, Chhibber A, Mohiyuddin M, Fan Y, Mu JC, Gibeling G, Barr S, Asadi NB, Gerstein MB, et al. An ensemble approach to accurately detect somatic mutations using SomaticSeq. Genome Biol. 2015;16:197.
35. Wood DE, White JR, Georgiadis A, Van Emburgh B, Parpart-Li S, Mitchell J, Anagnostou V, Niknafs N, Karchin R, Papp E, et al. A machine learning approach for somatic mutation discovery. Sci Transl Med. 2018;10(457):eaar7939.
36. Bamford S, Dawson E, Forbes S, Clements J, Pettett R, Dogan A, Flanagan A, Teague J, Futreal PA, Stratton MR, et al. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. Br J Cancer. 2004;91(2):355–8.
37. Kalatskaya I, Trinh QM, Spears M, McPherson JD, Bartlett JMS, Stein L. ISOWN: accurate somatic mutation identification in the absence of normal tissue controls. Genome Med. 2017;9(1):59.
38. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68–74.
39. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434–43.
40. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, Musolf A, Li Q, Holzinger E, Karyadi D, et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. Am J Hum Genet. 2016;99(4):877–85.
41. Zhang J, Walsh MF, Wu G, Edmonson MN, Gruber TA, Easton J, Hedges D, Ma X, Zhou X, Yergeau DA, et al. Germline mutations in predisposition genes in pediatric cancer. N Engl J Med. 2015;373(24):2336–46.
42. Auer PL, Reiner AP, Wang G, Kang HM, Abecasis GR, Altshuler D, Bamshad MJ, Nickerson DA, Tracy RP, Rich SS, et al. Guidelines for large-scale sequence-based complex trait association studies: lessons learned from the NHLBI exome sequencing project. Am J Hum Genet. 2016;99(4):791–801.
43. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248–9.
44. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. 2011;39(17):e118.
45. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, et al. Analysis of protein-

coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285–91.
46. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46(3):310–5.
47. Jian X, Boerwinkle E, Liu X. In silico prediction of splice-altering single nucleotide variants in the human genome. Nucleic Acids Res. 2014;42(22):13534–44.
48. Li Q, Wang K. InterVar: clinical interpretation of genetic variants by the 2015 ACMG-AMP guidelines. Am J Hum Genet. 2017;100(2):267–80.
49. Edmonson MN, Patel AN, Hedges DJ, Wang Z, Rampersaud E, Kesserwan CA, Zhou X, Liu Y, Newman S, Rusch MC, et al. Pediatric Cancer Variant Pathogenicity Information Exchange (PeCanPIE): a cloud-based platform for curating and classifying germline variants. Genome Res. 2019;29(9):1555–65.
50. Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 2019;47(D1):D766–73.
51. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164.
52. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. The Ensembl variant effect predictor. Genome Biol. 2016;17(1):122.
53. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012;6(2):80–92.
54. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol. 2010;6(12):e1001025.
55. Sim NL, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res. 2012;40(Web Server issue):W452–7.
56. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019;47(D1):D886–94.
57. Guo Y, Li J, Li CI, Long J, Samuels DC, Shyr Y. The effect of strand bias in Illumina short-read sequencing data. BMC Genomics. 2012;13:666.
58. Smit A, Hubley R, Green P. RepeatMasker Open-4.0. http://www.repeatmaskerorg 2013–2015.
59. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, Nusbaum C, Jaffe DB. Characterizing and measuring bias in sequence data. Genome Biol. 2013;14(5):R51.
60. Dou Y, Kwon M, Rodin RE, Cortes-Ciriano I, Doan R, Luquette LJ, Galor A, Bohrson C, Walsh CA, Park PJ. Accurate detection of mosaic variants in sequencing data without matched controls. Nat Biotechnol. 2020;38(3):314–9.
61. Sendorek DH, Caloian C, Ellrott K, Bare JC, Yamaguchi TN, Ewing AD, Houlahan KE, Norman TC, Margolin AA, Stuart JM, et al. Germline contamination and leakage in whole genome somatic single nucleotide variant detection. BMC Bioinformatics. 2018;19(1):28.
62. Ruiz-Pesini E, Lott MT, Procaccio V, Poole JC, Brandon MC, Mishmar D, Yi C, Kreuziger J, Baldi P, Wallace DC. An enhanced MITOMAP with a global mtDNA mutational phylogeny. Nucleic Acids Res. 2007;35(Database issue):D823–8.
63. Payne BA, Wilson IJ, Hateley CA, Horvath R, Santibanez-Koref M, Samuels DC, Price DA, Chinnery PF. Mitochondrial aging is accelerated by anti-retroviral therapy through the clonal expansion of mtDNA mutations. Nat Genet. 2011;43(8):806–10.
64. Santibanez-Koref M, Griffin H, Turnbull DM, Chinnery PF, Herbert M, Hudson G. Assessing mitochondrial heteroplasmy using next generation sequencing: a note of caution. Mitochondrion. 2019;46:302–6.
65. Ding J, Sidore C, Butler TJ, Wing MK, Qian Y, Meirelles O, Busonero F, Tsoi LC, Maschio A, Angius A, et al. Assessing mitochondrial DNA variation and copy number in lymphocytes of ~2,000 Sardinians using tailored sequencing analysis tools. PLoS Genet. 2015;11(7):e1005306.
66. Edge P, Bansal V. Longshot enables accurate variant calling in diploid genomes from single-molecule long read sequencing. Nat Commun. 2019;10(1):4660.
67. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–100.
68. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz MC. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods. 2018;15(6):461–8.
69. Ebler J, Haukness M, Pesout T, Marschall T, Paten B. Haplotype-aware diplotyping from noisy long reads. Genome Biol. 2019;20(1):116.
70. Luo R, Sedlazeck FJ, Lam TW, Schatz MC. A multi-task convolutional deep neural network for variant calling in single molecule sequencing. Nat Commun. 2019;10(1):998.
71. Ramachandran A, Lumetta SS, Klee E, Chen D. HELLO: a hybrid variant calling approach. bioRxiv. 2003;2020(2020):2023.004473.
72. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21.

73. Wu TD, Reeder J, Lawrence M, Becker G, Brauer MJ. GMAP and GSNAP for genomic sequence alignment: enhancements to speed, accuracy, and functionality. Methods Mol Biol. 2016;1418:283–334.
74. Liu F, Zhang Y, Zhang L, Li Z, Fang Q, Gao R, Zhang Z. Systematic comparative analysis of single-nucleotide variant detection methods from single-cell RNA sequencing data. Genome Biol. 2019;20(1):242.
75. Singer J, Kuipers J, Jahn K, Beerenwinkel N. Single-cell mutation identification via phylogenetic inference. Nat Commun. 2018;9(1):5144.
76. Poirion O, Zhu X, Ching T, Garmire LX. Using single nucleotide variations in single-cell RNA-seq to identify subpopulations and genotype-phenotype linkage. Nat Commun. 2018;9(1):4892.
77. Rodriguez-Meira A, Buck G, Clark SA, Povinelli BJ, Alcolea V, Louka E, McGowan S, Hamblin A, Sousos N, Barkas N, et al. Unravelling intratumoral heterogeneity through high-sensitivity single-cell mutational analysis and parallel RNA sequencing. Mol Cell. 2019;73(6):1292–305, e1298.
78. Tirosh I, Venteicher AS, Hebert C, Escalante LE, Patel AP, Yizhak K, Fisher JM, Rodman C, Mount C, Filbin MG, et al. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. Nature. 2016;539(7628):309–13.
79. Roth A, Ding J, Morin R, Crisan A, Ha G, Giuliany R, Bashashati A, Hirst M, Turashvili G, Oloumi A, et al. JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. Bioinformatics. 2012;28(7):907–13.
80. Gerstung M, Beisel C, Rechsteiner M, Wild P, Schraml P, Moch H, Beerenwinkel N. Reliable detection of subclonal single-nucleotide variants in tumour cell populations. Nat Commun. 2012;3:811.
81. Wilm A, Aw PP, Bertrand D, Yeo GH, Ong SH, Wong CH, Khor CC, Petric R, Hibberd ML, Nagarajan N. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res. 2012;40(22):11189–201.
82. Shiraishi Y, Sato Y, Chiba K, Okuno Y, Nagata Y, Yoshida K, Shiba N, Hayashi Y, Kume H, Homma Y, et al. An empirical Bayesian framework for somatic mutation detection from cancer genome sequencing data. Nucleic Acids Res. 2013;41(7):e89.
83. Hansen NF, Gartner JJ, Mei L, Samuels Y, Mullikin JC. Shimmer: detection of genetic alterations in tumors using next-generation sequence data. Bioinformatics. 2013;29(12):1498–503.
84. Christoforides A, Carpten JD, Weiss GJ, Demeure MJ, Von Hoff DD, Craig DW. Identification of somatic mutations in cancer through Bayesian-based analysis of sequenced genome pairs. BMC Genomics. 2013;14:302.
85. Kim S, Jeong K, Bhutani K, Lee J, Patel A, Scott E, Nam H, Lee H, Gleeson JG, Bafna V. Virmid: accurate detection of somatic mutations with sample impurity inference. Genome Biol. 2013;14(8):R90.
86. Kassahn KS, Holmes O, Nones K, Patch AM, Miller DK, Christ AN, Harliwong I, Bruxner TJ, Xu Q, Anderson M, et al. Somatic point mutation calling in low cellularity tumors. PLoS One. 2013;8(11):e74380.
87. Cantarel BL, Weaver D, McNeill N, Zhang J, Mackey AJ, Reese J. BAYSIC: a Bayesian method for combining sets of genome variants with improved specificity and sensitivity. BMC Bioinformatics. 2014;15:104.
88. Wang W, Wang P, Xu F, Luo R, Wong MP, Lam TW, Wang J. FaSD-somatic: a fast and accurate somatic SNV detection algorithm for cancer genome sequencing data. Bioinformatics. 2014;30(17):2498–500.
89. Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SRF, WGS500 Consortium, Wilkie AOM, McVean G, Lunter G. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nat Genet. 2014;46(8):912–8.
90. Usuyama N, Shiraishi Y, Sato Y, Kume H, Homma Y, Ogawa S, Miyano S, Imoto S. HapMuC: somatic mutation calling using heterozygous germ line variants near candidate mutations. Bioinformatics. 2014;30(23):3302–9.
91. Radenbaugh AJ, Ma S, Ewing A, Stuart JM, Collisson EA, Zhu J, Haussler D. RADIA: RNA and DNA integrated analysis for somatic mutation detection. PLoS One. 2014;9(11):e111516.
92. Shi Y. SOAPsnv: an integrated tool for somatic single-nucleotide variants detection with or without normal tissues in cancer genome. J Clin Oncol. 2014;32(15_suppl):e22086.
93. Sengupta S, Gulukota K, Zhu Y, Ober C, Naughton K, Wentworth-Sheilds W, Ji Y. Ultra-fast local-haplotype variant calling using paired-end DNA-sequencing data reveals somatic mosaicism in tumor and normal blood samples. Nucleic Acids Res. 2016;44(3):e25.
94. Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, McEwen R, Johnson J, Dougherty B, Barrett JC, Dry JR. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Res. 2016;44(11):e108.
95. Liu Y, Loewer M, Aluru S, Schmidt B. SNVSniffer: an integrated caller for germline and somatic single-nucleotide and indel mutations. BMC Syst Biol. 2016;10(Suppl 2):47.
96. Spinella JF, Mehanna P, Vidal R, Saillour V, Cassart P, Richer C, Ouimet M, Healy J, Sinnett D. SNooPer: a machine learning-based method for somatic variant identification from low-pass next-generation sequencing. BMC Genomics. 2016;17(1):912.
97. Jones D, Raine KM, Davies H, Tarpey PS, Butler AP, Teague JW, Nik-Zainal S, Campbell PJ. cgpCaVEManWrapper: simple execution of CaVEMan in order to detect somatic single nucleotide variants in NGS data. Curr Protoc Bioinformatics. 2016;56:15.

98. Carrot-Zhang J, Majewski J. LoLoPicker: detecting low allelic-fraction variants from low-quality cancer samples. Oncotarget. 2017;8(23):37032–40.
99. Sahraeian SME, Liu R, Lau B, Podesta K, Mohiyuddin M, Lam HYK. Deep convolutional neural networks for accurate somatic mutation detection. Nat Commun. 2019;10(1):1041.
100. Meng J, Victor B, He Z, Liu H, Jiang T. DeepSSV: detecting somatic small variants in paired tumor and normal sequencing data with convolutional neural network. Brief Bioinform. 2020;22(4):bbaa272.

4 Identification of Copy Number Alterations from Next-Generation Sequencing Data

Sheida Nabavi and Fatima Zare

Abstract

Copy number variation (CNV), which is deletion and multiplication of segments of a genome, is an important genomic alteration that has been associated with many diseases including cancer. In cancer, CNVs are mostly somatic aberrations that occur during cancer evolution. Advances in sequencing technologies and arrival of next-generation sequencing data (whole-genome sequencing and whole-exome sequencing or targeted sequencing) have opened up an opportunity to detect CNVs with higher accuracy and resolution. Many computational methods have been developed for somatic CNV detection, which is a challenging task due to complexity of cancer sequencing data, high level of noise and biases in the sequencing process, and big data nature of sequencing data. Nevertheless, computational detection of CNV in sequencing data has resulted in the discovery of actionable cancer-specific CNVs to be used to guide cancer therapeutics, contributing to significant progress in precision oncology. In this chapter, we start by introducing CNVs. Then, we discuss the main approaches and methods developed for detecting somatic CNV for next-generation sequencing data, along with its challenges. Finally, we describe the overall workflow for CNV detection and introduce the most common publicly available software tools developed for somatic CNV detection and analysis.

Introduction

Recently, copy number variation (CNV) has drawn much attention in biomedical fields. Copy number variation is a form of structural variation of a DNA sequence that includes amplification and deletion of a particular segment of DNA (shown in Fig. 4.1). It features a higher mutation rate than single-nucleotide polymorphisms (SNPs) and affects a larger fragment of genomes. There is no precise definition for the minimum length of CNVs in research, although a minimum length of 1 kb is commonly used for clinical applications.

Researchers have considered the impact of genomic variations on human diseases as it provides valuable insight into functional elements and disease-causing regulatory variants [1–11]. Previously, SNPs were considered as the predominant form of genomic variation associated with phenotypes [12, 13]. However, more recent studies show the widespread existence of CNV in

S. Nabavi (✉) · F. Zare
Computer Science and Engineering, Institute for Systems Genomics, University of Connecticut, Storrs, CT, USA
e-mail: Sheida.nabavi@uconn.edu; fatima.zare@uconn.edu

© Springer Nature Switzerland AG 2022
A. Laganà (ed.), *Computational Methods for Precision Oncology*, Advances in Experimental Medicine and Biology 1361, https://doi.org/10.1007/978-3-030-91836-1_4

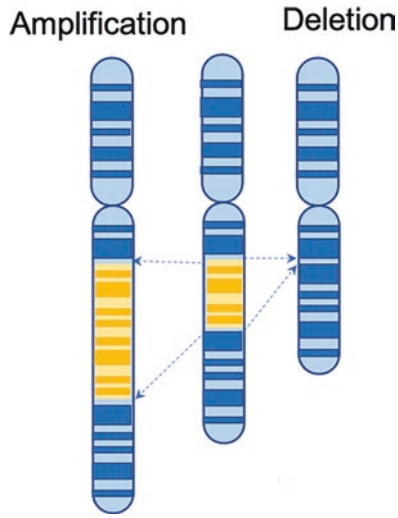

Fig. 4.1 Amplification and deletion of a segment of a genome

individuals and their association with complex diseases. These studies have shown that there is a close correlation between CNV and gene expression and the copy number has influence on gene expression for the majority of genes [10].

The interest in and importance of CNVs has risen in a wide collection of diseases including Parkinson [14], Hirschsprung [15], diabetes mellitus [16], autism [17–19], Alzheimer [20], and schizophrenia [21]. Specifically, significant effort has found associations between CNVs and cancers [22–27]. The importance of CNV can be seen by the trend in the number of scientific articles about CNV stored in PubMed per year, which has significantly increased from a few tens in 1990 to more than 800 in 2018 [28].

Copy number variations usually are mentioned in two contexts: germline CNVs, refer to inherited variants, many of which are polymorphic at the population level, and somatic CNVs, refer to changes resulting from somatic mutations, such as those commonly observed in cancer. Somatic CNVs are also called copy number aberrations or CNAs. In this chapter, for simplicity, we call somatic CNVs or CNAs just CNVs. Cancer is well known as a disease of genome, and genomic variations in cancer are mostly somatic variations, since tumors arise from normal cells with acquired aberrations in their genomic materials [25, 29]. Copy number variation is one of the most important genomic variations in cancer [22, 27, 29–31], since oncogene activation is often attributed to chromosomal copy number amplification, and tumor suppressor gene inactivation is often caused by either heterozygous deletion associated with mutation or by homozygous deletion. Studies show that CNVs significantly contribute to cancer cell growth, drug sensitivity, and resistance. Thus, identification of somatic CNVs can have an important role in cancer prognosis and treatment improvement [32].

Over the years, several methods have been used to detect CNVs including cytogenetics, fluorescence in situ hybridization (FISH), polymerase chain reaction (PCR), comparative genomic hybridization (CGH), and microarrays or SNP arrays. Some of these methods such as PCR and FISH are fast, but they are not readily scalable to many genomic targets. Array-based technologies have been used widely since late 1990s for more than a decade as an affordable and relatively high-resolution assay for CNV detection targeting a large scale of genomic regions [33]. In standard practice, chromosomal microarray has been the main clinical test for CNV detection. However, array-based technologies have limitations associated with hybridization, which results in poor sensitivity and precision, and with resolution, related to the coverage and density of the arrays' probes.

With the arrival of next-generation sequencing (NGS) technologies [34], sequence-based CNV detection has rapidly emerged as a viable option to identify CNVs with higher resolution and accuracy [23, 35, 36]. As a result, recently whole-genome sequencing (WGS), whole-exome sequencing (WES), and targeted sequencing have emerged as primary strategies for NGS technologies in CNV detection and for studying human diseases. Using sequencing data, it is currently feasible to not only detect a rapidly growing set of known clinically relevant mutations but also identify novel or unexpected important variations.

Whole-genome sequencing offers an unbiased genome-wide approach to detect CNVs, while

WES and targeted sequencing allow the identification of genomic variants in protein coding regions (less than 2% of the genome) that can provide direct functional interpretation. Whole-genome sequencing is the most comprehensive platform for cancer genome profiling, and in many studies, CNVs are identified from WGS data. However, WGS is considered too expensive for research involving large cohort and WES or targeted sequencing is becoming an alternative, cost-effective strategy [37]. In clinical practice, WGS has been employed in pilot studies involving a few cases [38] and WES has been used in large cohort in clinical studies of genetic disorders [39]. Large research studies such as The Cancer Genome Atlas (TCGA) and Stand Up To Cancer (SU2C) have employed WES. Even though WES is widely used in clinical genetics [40], it has recently emerged as one of the most popular techniques for identifying clinically relevant genomic variations in cancer [41]. Studies have shown the feasibility of using WES into clinical practice for precision cancer care [42, 43]. Exome represents a highly function-enriched subset of the human genome, and CNVs in exome are more likely to be disease-causing variations than those in nongenic regions [44, 45]. WES can offer lower cost, higher coverage, and less complex data analysis, which are appealing for clinical applications when there are several samples available. However, WES impose new challenges to CNV detection analysis compared to WGS and has several technical issues [46]. Methods developed for CNV detection for WGS data might not be proper for WES data, since their main assumptions on read distributions and continuity of data do not hold for WES data. In addition, WES data introduce biases due to hybridization. As a result, different methodologies need to be employed for CNV detection using WES/targeted and WGS data.

In general, many tools (>150) have been developed for CNV detection using WGS and WES/targeted data. However, not all of them are appropriate for detecting somatic mutations in cancer. Germline and somatic CNVs are very different in their overall coverage of the genome and their frequency across population. The characteristics of somatic CNVs need special consideration in algorithms and strategies in which germline CNV detection methods are usually not suited for. In general, germline CNVs cover small portion of a genome (about 4%) [47], they are more often deletion, and they are common among different people. However somatic CNVs can cover a majority part of a genome, can be focal, and are unique for each tumor. As a result, CNV detection methods that are developed for identifying population CNVs or germline CNVs are not suitable for identifying somatic variations. Also, identifying somatic CNVs in cancer is very challenging because of the tumor heterogeneity and complexity: tumor samples are contaminated by normal tissue, the ploidy of tumors is unknown, and there are multiple clones in tumor samples. On top of the tumor samples' complexity there are experimental, technical, and sequencing noise and biases which make somatic CNV detection very challenging.

In this chapter, we first briefly discuss the important role of CNVs in precision oncology. We then introduce CNV detection methods and challenges and describe the most commonly used CNV detection tools for WES/targeted and WGS data.

CNV and Precision Oncology

The introduction of NGS technologies and the increasing number of large-scale tumor molecular profiling studies have revolutionized the field of precision oncology. According to the European Society for Medical Oncology (ESMO), cancer precision medicine or precision oncology is defined as "the use of therapeutics that are expected to confer benefit to a subset of patients whose cancer displays specific molecular or cellular features (most commonly genomic changes and changes in gene or protein expression patterns)" [48]. Precision oncology involves the detection of tumor-specific somatic genomic variations, followed by treatment with therapeutics that specifically target identified actionable variations. CNVs as major genomic variations

and proven actionable biomarkers play an important role in precision oncology.

With emerging NGS technologies and the decreasing cost of generating NGS data, comprehensive genomic analyses have become increasingly available due to advances in development of accessible and applicable bioinformatics tools for detecting genomic and molecular variants from NGS data. This availability is not just limited to research settings, as we observe increasing availability of comprehensive genomic analyses in clinical settings as well. As a result, the number of identified actionable and druggable tumor-specific genomic variations has grown substantially in the past decade and we expect rapid emergence of additional biomarkers as well. Studies show that many tumor-specific molecular variations in cancer driver genes (including SNPs, CNVs, translocations, and gene fusions) are well-proven predictive biomarkers of response to selective targeted therapies.

Precision oncology has transformed cancer care, which is moving from standard treatments based on cancer types and increasing focus on personalizing treatments based on genomic variants. Genomic variants can now be targeted by specific therapies to improve clinical outcomes in patients. Several studies show that a significant survival benefit has been obtained from biomarker matching therapies compared with standard therapies in several cancer types [49–52]. It is also shown that employing NGS data and bioinformatics tools for medical diagnostics has no significant impact on the costs of cancer care by [49]. These studies found that patients who received therapy on the basis of specific molecular variations, independent of tumor type, experienced improved survival. It is also shown that molecularly guided therapies improve survival in patients with advanced refractory cancer [49]. CNVs as major genomic variations have been used as important actionable biomarkers to advance precision oncology. Many studies reported actionable CNVs that have been used to select effective molecularly guided therapies. For example, in [49], EGFR3 amplification in bladder cancer, FGFR1 amplification in colon cancer and lung cancer, FGF4 amplification in gastric cancer, and PDGFRA amplification in lung cancer are treated with Pazopanib; PIK3CA amplification in breast cancer and head/neck cancers, MTOR amplification in lung cancer, and PIK3R2 deletion in ovary cancer are treated with Everolimus; HRAS amplification in colon cancer is treated with Trametinib; and ERBB2 amplification in colon cancer is treated with Ado-trastuzumab, in advanced refractory cancer patients. In this study, all the patients received the genomic variant-guided therapies show improved survival compared to whom treated with standard medications.

To facilitate annotating genomic variations to actionable variants and translating results to clinical actions, many publicly accessible data resources have been developed. Most of these databases include CNVs. These databases compile evidence and associations with a specific histology or disease, as well as their prognostic and/or predictive value of response to specific therapies to aid clinical decision-making [53]. Example of such databases are Precision Oncology Knowledge Base (OncoKB) (http://oncokb.org/#/) and Personalized Cancer Therapy Knowledge Base for Precision Oncology (PCT) (from MD Anderson Cancer Centre (https://pct.mdanderson.org/#/). These databases integrate variant-specific recommendations from clinical practice guidelines in their annotation and rank relevant matched therapies by evidence of clinical benefit. There are also other databases for cancer somatic variants such as the Catalogue of Somatic Mutations in Cancer (COSMIC) [54] database. COSMIC is the largest and most comprehensive resource for exploring the impact of somatic variations (including CNVs) in human cancer. However, it does not rank relevant matched therapies.

To enable precision oncology, few NGS diagnostics assays (or gene panels) for targeted deep sequencing of key cancer genes have been developed. These assays are designed for calling actionable genomic variations in cancer, including CNVs [55]. Two of such panels that are also approved by the Food and Drug Administration (FDA) in 2017 are Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT)

[56] developed at the Memorial Sloan Kettering Cancer Center (MSKCC) that leverages 486 cancer genes, and FoundationOne CDx (F1CDx) developed at the Foundation Medicine Inc. that includes 324 cancer genes. These assays are increasingly used in clinical settings for personalizing cancer treatment in a precision oncology fashion.

CNV Detection

CNV Detection Methods

In general, there are four main approaches to identify CNV from next-generation sequencing data: (1) read depth (RD), (2) paired-end (PE), (3) split read (SR), and (4) assembly [57] as shown in Fig. 4.2.

In read depth (RD) approaches, mostly a non-overlapping sliding window is used to count the number of short reads that are mapped to a genomic region overlapped with the window. Then, these read count values are used to identify CNV regions. RD methods are based on the hypothesis that there is a correlation between depth of coverage of a genomic region and the copy number of the region [58]. In most RD-based approaches, a statistical method is used to merge neighboring genomic regions with similar read counts and to identify genomic regions where read counts are significantly different from their adjacent regions. The overall pipeline for detecting CNVs using the RD-based approach is shown in Fig. 4.3.

Paired-end (PE) approaches, which are applied to paired-end NGS data, identify genomics aberration based on the distances between the pairs of reads. In PE sequencing data, reads from the two ends of the genomics segments are available. The distance between a pair of paired-end reads is used as an indicator of a genomic aberration including CNV. A genomic aberration is detected when the distance is significantly different from the predetermined average insert size, where a larger distance indicates amplification, and a shorter distance indicates deletion. This approach is mostly used for identifying other types of structural variation (SV), beyond CNVs, such as inversion and translocation.

Split read (SR) methods also applied to paired-end NGS data and use pairs of reads where only

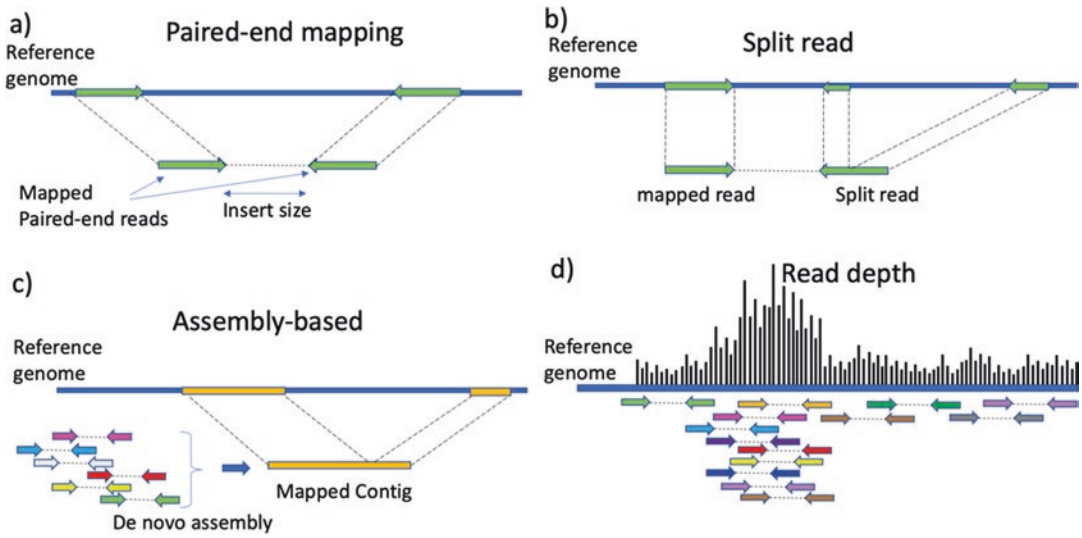

Fig. 4.2 CNV detection methods for NGS data. (**a**) In paired-end mapping methods the distance between a pair of paired-end reads indicates CNVs. (**b**) In split read methods the locations of mapped split reads indicate breakpoints of genomic structural variations. (**c**) In Assembly-based methods the comparison between generated contigs with the reference genome indicates structural variations (**d**) in read depth methods the coverages of genomic regions indicate CNVs

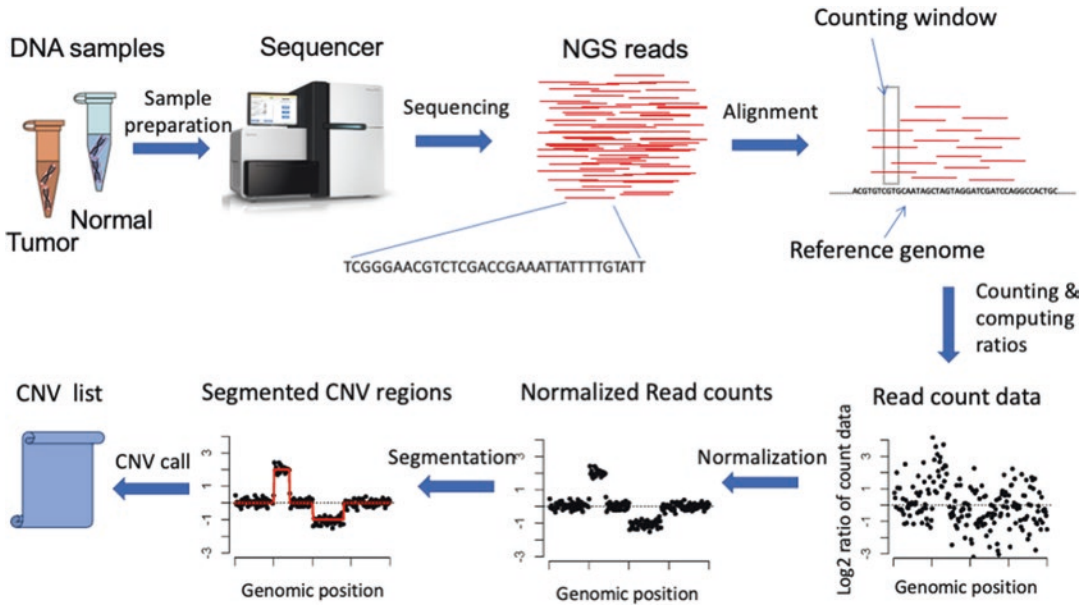

Fig. 4.3 Overall pipeline for detecting CNVs using read count data

one read of a pair is mapped and the other one either completely or partially fails to map to the genome [59]. The unmapped reads are split, and fractions are mapped to the genome. The locations of mapped split reads indicate breakpoints of SVs. Split-read-based methods have limited ability to identify large-scale SVs. They can detect, at least in theory, deletions without size limitation, while they cannot detect insertions larger than the read length, because the insertion cannot be contained in a single read. Split-read-based methods are more effective for detecting indels (small insertion and deletion) and breakpoints.

Compared to PE and SR methods, RD-based methods can detect the exact number of CNVs, as PE, and SR can only report the position of potential CNVs and not the counts. In addition, RD-based methods can work better on large size CNVs, which are hard to detect with PE and SR methods [60].

In assembly approaches, short reads are used to assemble the genomics regions by connecting overlapping short reads (contigs). Assembly-based methods first generate a contig/scaffold, then compare the contig with the reference genome to detect SVs [61]. Assembly-based methods are computationally very expensive. Moreover, eukaryotic genomes are very complex and contain a significant fraction of repeats and segmental duplications and assembly-based methods perform poorly in these complex regions. Another issue with the assembly-based methods is that they are unable to handle haplotype sequences and therefore only homozygous structural variations can be detected [62]. Due to the above limitations, assembly-based methods are less used in CNV detection for eukaryotic genomes.

Because each of these approaches has limitations, several methods have been developed that combine multiple aforementioned approaches to detect CNVs for WGS data [63].

Due to the availability of high coverage sequencing data and the limitation of PE, SR, and assembly-based CNV detection methods, RD-based methods have recently become a major approach to identify CNVs, especially in cancer studies, where the number of copies is important. However, PE, SR, and assembly methods cannot distinguish somatic and germline structural variations, and do not provide copy numbers. In RD-based methods, the absolute number of DNA copy of any genomic region can be inferred by

counting the number of reads/bases aligned to that particular region. As a result, RD-based methods are mostly used in cancer research employing both WGS and WES/targeted data, and especially for WES/targeted data. In WES, targeted regions are exonic regions that are very short and discontinuous across the genome. As a result, the PE, SP, and assembly approaches for identifying CNVs are not proper for WES data. Also, high coverage of WES data makes the RD approach more practical. Therefore, all CNV detection tools for WES are based on the RD approach.

Read depth–based CNV detection methods can be categorized into three classes: single sample, paired case/control samples, and population samples. In the single sample category, as there is no other subject available, the absolute copy number will be reported. In the paired case/control samples category, the relative copies compared to the control will be reported. And, in population samples category, the RD data across all the samples will be considered to report CNVs [57]. In cancer studies, the paired case/control samples category of CNV detection methods that specially use match tumor-normal paired samples are more proper for detecting somatic CNVs. This is because germline CNVs are excluded from consideration and biases such as GC content and mappability are reduced due to the comparison of read counts from the same genomic regions. Generally, these methods use the ratio of normalized read counts between tumor and normal samples in a given genomic window. In theory, assuming a diploid genome, the ratio of 1 (or log2 ratio of 0) represents no copy-number change; the ratio of 2 (log2 ratio of 1) represents a two-copy gain; the ratio of 1.5 (log2 ratio of 0.58) represents a one-copy gain, and the ratio of 0.5 (log2 ratio of −1) represents a one-copy loss. However, because most tumor samples are mixtures of normal and cancer cells and there are tumor subclones, the read ratios tend to deviate from the expected values. This problem is more challenging for loss or gain of one copy, which can be deviated from normal easily (from 0.5 or 1.5 toward 1). Thus, a threshold is used to call amplification and deletion. Also, tumor sample complexity affects the performance of CNV detection. Therefore, taking into account sample purity and ploidy is important for accurate detection of somatic CNVs.

Challenges in Developing Computational Methods for Detecting Somatic CNVs in Cancer

Despite improvements in sequencing technologies and CNV detection methods, identifying CNV is still a challenging task, and even more in cancer samples because of the complexity of tumors and heterogeneity in CNV characteristics [57, 64]. In this section, we briefly explain the challenges that somatic CNV identification is faced with in cancer when using sequencing data. We divide these challenges into five classes: (1) noisy sequencing data, (2) sequencing technical problems, (3) tumor complexity, (4) lack of ground truth, and (5) CNV heterogeneity.

Noisy Sequencing Data

The main assumption of the RD-based CNV detection algorithms is that the read counts and CNV for a particular region are correlated. However, there are biases and noise that distort the relationship between the read count and copy number. These biases and noise include GC bias, mappability bias, experimental noise, and technical (sequencing) noise. GC content varies significantly along the genome and has been found to influence read coverage on most sequencing platforms [58, 65]. In the alignment step, a huge number of reads are mapped to multiple positions due to the short read length and the presence of repetitive regions in the reference genome [58, 66]. These ambiguities in alignment can produce unavoidable biases and errors in RD-based CNV detection methods [58].

Sequencing Technical Problems

In the process of generating sequencing reads from samples, sample preparation, library preparation, and sequencing process introduce experimental and systematic noise that can hinder CNV detection [58, 67].

One of the major causes of sequencing noise is PCR hybridization, which is commonly used in WES. Hybridization causes generation of many reads for a specific region (not because of amplification) and, as a result there are less reads available to ensure evenly distribution of reads to other regions. It is very common that in some genomic regions the read count is very low, and this problem is more severe in WES data. The exome capture procedure in the library preparation process for WES introduces more biases and noise [68], causing not even distribution of reads in the exonic regions. These biases affect the statistical analysis for calling CNVs and distort the relation between read counts and CNVs; as a result, they present noise to CNV detection algorithms. Moreover, in many cases, samples are extracted from formalin-fixed (FFPE) tissues, from which only a small amount of poor quality RNA is usually extracted.

Tumor Complexity

The complexity of cancer tumors also distorts the relationship between read count and CNV and, as a result, it affects the performance of CNV detection methods. The tumor complexity includes tumor purity, tumor ploidy, and tumor subclonal heterogeneity.

Therefore, reads mapped to a particular region do not all belong to tumor cells. As a result, read count values do not completely reflect copy number of tumor cells, and the tumor normal copy number ratio is less than the real value. This introduces difficulties in calling copy number segments. A threshold for calling CNV will depend on tumor purity, which is usually unknown. Few tools are available to estimate tumor purity such as Absolute [69], which are designed for array-based data, THetA2 [70], Accurity [71], BubbleTree [72], AbsCN-seq [73], and MixClone [74], which are designed for sequencing data.

Aneuploidy of the tumor genome is observed in almost all cancer tumors [75], which creates difficulties in determining the copy number values. The tumor-normal read count ratio corresponds to the average ploidy, which is usually unknown in the tumor sample. So far, few tools, such as Patchwork [76], AbsCN-seq [73], Absolute [69], and Sequenza [77], have been developed to identify tumor ploidy. They mostly incorporate the fraction of nonreference allele (B allele frequency or BAF) to identify the tumor ploidy, since different tumor ploidy exhibits distinct BAF signatures.

It is also observed that multiple clonal subpopulations of cells are present in tumors [78]. Due to their low percentage in a sample, it is hard to determine the subclones. This intra-tumor heterogeneity or multiple clonality distorts the estimate of copy number values and makes calling CNV segments complicated.

Lack of Gold Standard

Benchmark analyses are necessary to evaluate the performance of CNV detection methods. However, there is no gold standard WES or WGS data with a validated CNV list that can be used for benchmarking the developed CNV detection methods. Comparative studies of CNV detection tools have shown that while they can detect CNVs, the concordance of the results is quite low [57, 64, 79, 80].

To evaluate the performance of CNV detection methods, simulated WGS and WES data, and/or data from other CNV profiling platforms, such as array-based CNV from the same samples, are used. Large data repositories such as TCGA or the1000 Genome Project contain genomic data from array-based and sequencing technologies that can be used for benchmarking.

A few simulators have been developed to generate synthesized genomes harboring CNVs. For example, RSVSim [81] and SVsim [82] simulate genome with SV including CNVs (deletions, insertions, inversions, tandem duplications and translocations), and SCNVSim [83] simulates genomes with CNVs. Since these tools only generate genomes, short read simulators need to be used to generate simulated sequencing data. A few tools have been developed for synthesizing WGS data from a genome with SV (Pysim-sv [84]) and with CNVs (SinC [85]); and for synthesizing WES data from a genome with CNV (CNV-Sim [86] and SECNVs [87]).

Using different assays or simulated sequencing data can be very useful for benchmarking CNV detection methods. However, it is unclear how well the CNV results from different assays, such as array-based technology, with lower resolution and precision, and simulated CNVs can capture the complexity of real tumor samples.

Heterogeneity of CNV Profiles

CNVs are very prevalent in cancer, can cover most of a genome, and have very heterogeneous profiles. While CNVs affecting large genomic segments are often frequent in cancer genomes, unlike in normal genomes, focal or small CNVs are also observed frequently. CNV detection methods are designed based on some assumption about characteristics of CNVs and the read count distributions. It has been shown that consistency among the CNV detection methods in calling CNVs is not high. This is mainly because CNV profile for a cancer sample is very heterogenous and complex, and each method is strong in capturing some characteristics of CNVs but not all. For example, a method that can detect large CNV segments cannot capture focal CNV segments precisely. The heterogeneity in the CNV profile of a genome adds challenges in developing an effective CNV detection method. To address this challenge, some studies suggest using multiple CNV detection methods that are designed based on different characteristics of CNVs and consolidate their results. A few ensemble approaches have been proposed (such as CN_Learn [88] and Anaconda [89]) that combine the results of several tools instead of using a single tool.

CNV Detection Algorithms and Tools

So far, more than 150 software tools have been developed to analyze CNVs [28] using next-generation sequencing data or microarray data. Among them about 60 tools have been developed for analyzing and detecting CNVs in cancer data, or have been used in cancer studies, using WGS, WES, and targeted sequencing data, as shown in Tables 4.1, 4.2, and 4.3. Several review papers have been published discussing and comparing CNV detection tools and methods including [57, 63, 79, 80, 90], which we point interested readers to for more information.

As mentioned in the previous section, the most appropriate approach to detect somatic CNVs and to provide absolute copy numbers is the read depth–based approach, which is used by almost all of the available CNV detection methods for cancer. In general, an RD approach consists of three major steps: (1) preprocessing, (2) segmentation, and (3) CNV calling, as shown in Fig. 4.4.

The input data are usually aligned short reads in the BAM, SAM, or Pileup formats from tumor and matched normal. A few tools such as VCF2CNA [91], sCNAphase [92], and SAAS-CNV [93] also use VCF format for input data. As mentioned in the previous section, using match normal samples is very important to reduce sequencing biases and to eliminate germline CNVs. However matched normal samples may not always be available and, as a result, some tools have been developed which rely on a group of normal samples instead of a matched normal, such as AluScanCNV2 [94] and RefCNV [95], or on a "synthetic-normal" such as SynthEx [96]. Moreover, since generating sequencing data from normal samples increases the sequencing cost, some tools offer to detect somatic CNVs without using normal samples, as indicated in Tables 4.1, 4.2, and 4.3. In addition to aligned sequencing data as input, some tools take BAF information as well to estimate tumor ploidy and adjusting copy numbers for more accurate CNV detection.

In the preprocessing step, outlier reads (such as low-quality reads, repeated reads, unmapped reads) are filtered out, read depth data are generated by using a sliding window, and sequencing data's biases and noise are reduced. Normalization and de-noising algorithms are important in this step. A typical strategy to reduce biases in cancer read count data for CNV studies is to use sequencing data from the matched normal tissue or germline of the same patient generated under the identical experimental conditions. Using matched normal samples helps to identify heterozygous SNPs for calculating BAF and to filter out germ-

Table 4.1 CNV detection software tools for WES data

Tool Name (WES)	Preprocessing	Programing Language	Year	Segmentation method	Detecting LOH	Need match norma/control	Using BAF	Purity/ploidy estimation
VarScan2 [108]	Normalized read depth	Java	2012	Needs segmentation external such as CBS	Y	Y	N	None
Ioncopy [115]	Sample normalization, Amplicon normalization	R	2018	Statistical analysis to call CNV for each gene	N	N	N	None
CNV-RF [109]	Read count normalization and GC correction	R/Perl	2016	Random forest for CNV call	Y	Y	N	None
ADTEX [37]	DWT	R/Python	2014	HMM	Y	Y	Y	Both
CloneCNA [116]	GC correction & read normalization	Matlab	2016	HMM	N	Y	Y	Purity
DEFOR [117]	GC correction	Perl	2019	Combining allelic frequency and read depth data	Y	Y	Y	Purity
EXCAVATOR2 [118]	GC correction & read normalization	R/Fortran/ bash	2016	Shifting level model segmentation algorithm	N	Y	N	None
ExomeCNV [119]	Read normalization	R	2011	CBS	Y	Y	Y	None
exomeCopy [120]	GC correction & read normalization	R	2011	HMM	N	Y	Y	None
ExomeDepth [121]	GC correction & read normalization	R	2012	Beta-binomial model for read count	N	N	N	None
hsegHMM [122]	_	R	2018	HMM	Y	N	Y	Both
CODEX2 [123]	Singular value decomposition (SVD)-based SVD GC correction & read normalization	R	2018	Recursive Poisson-likelihood segmentation algorithm	N	N	N	None
RefCNV [95]	_	R	2016	Linear regression	N	Y	N	None
CNVkit [124]	GC correction & read normalization	Python	2016	CBS	N	N	N	Ploidy
ONCOCNV [125]	GC correction & read normalization	R	2014	CBS	N	Y	N	None
PureCN [126]	GC correction	R	2016	CBS	Y	Y	Y	Both
Contra [114]	GC correction & read normalization	Python, R	2012	CBS	N	Y	N	None

Table 4.2 CNV detection software tools for WGS data

Tool Name (WGS)	Preprocessing	Programing Language	Year	Segmentation method	Detecting LOH	Need match norma/control	Using BAF	Purity/ploidy estimation
ACE [110]	GC correction and mappability correction	R	2018	CBS	N	Y	N	Both
CNV-seq [105]	Global normalization	R/Perl	2009	Heuristic statistical method assuming Gaussian distribution	N	Y	N	None
CLImAT-HET [127]	Correct GC content and mappability bias, quantile normalization of read depths	C/Matlab	2017	HMM	Y	N	Y	Purity
CNV_IFTV [128]	GC correction	Python	2019	Isolation forest algorithm and TV	N	Y	N	Both
CNVnator [129]	RD signal calculation and correction of GC-bias	Python	2011	Change-point analysis, mean shift	Y	N	Y	None
CONDEL [130]	GC correction	C++, Perl	2020	Bayesian Inference, mixture model	N	Y	N	None
COPS [131]	Read normalization	R/Perl	2012	Boundary segmentation module	N	Y	N	None
FALCON [92]	NO	R	2015	Change-point model on a bivariate mixed binomial process	Y	Y	Y	Both
Patchwork [76]	GC correction & read normalization	R	2013	CBS	Y	Y	Y	Both
ReadDepth [132]	GC correction	R	2011	CBS	N	Y	N	None
sCNAphase [92]	NO	R	2017	HMM, regional haplotype depth	Y	Y	Y	Purity
SEG [133]	GC correction & read normalization	C	2018	Change-point analysis	N	Y	N	None
seqCNA [134]	GC correction	R	2014	Change-point analysis	N	N	N	None
TITAN [135]	GC correction	R	2014	HMM	Y	Y	Y	Ploidy
WaveCNV [136]	Read count normalization	Matlab	2014	Translation-invariant discrete wavelet transforms	Y	Y	Y	Ploidy
Xcavator [137]	Read count normalization and GC correction	R; Perl; Fortran	2017	Shifting level model	N	Y	N	None
BiC-seq2 [138]	Read count normalization and GC correction	C	2016	Bayesian information criterion	N	Y	N	None
BagGMM [139]	Read count normalization and GC correction	Matlab	2019	Gaussian Mixture model	N	Y	N	None

Table 4.3 CNV detection software tools for both WGS and WES data

Tool Name (WES and WGS)	Preprocessing	Programing Language	Year	Segmentation method	Detecting LOH	Need match norma/control	Using BAF	Purity/ploidy estimation
AluScanCNV2 [94]	Normalization, GC correction	R	2019	Geary-Hinkley transformation	N	Y	N	None
cn.mops [140]	Sample normalization, GC correction	R	2012	HMM	N	Y	N	None
Control-FREEC [101]	GC correction	C++	2011	LASSO-based algorithm	Y	Y	Y	None
FACETS [141]	GC correction	R	2016	CBS	Y	Y	Y	Both
SAAS-CNV [93]	GC correction	R	2015	CBS, allelic frequency	Y	Y	Y	Both
SynthEx [96]	GC correction	R	2017	CBS	N	N	Y	Both
VCF2CNA [91]	Read count normalization	R	2019	Recursive partitioning–based segmentation using SNP allele counts	Y	Y	Y	Purity

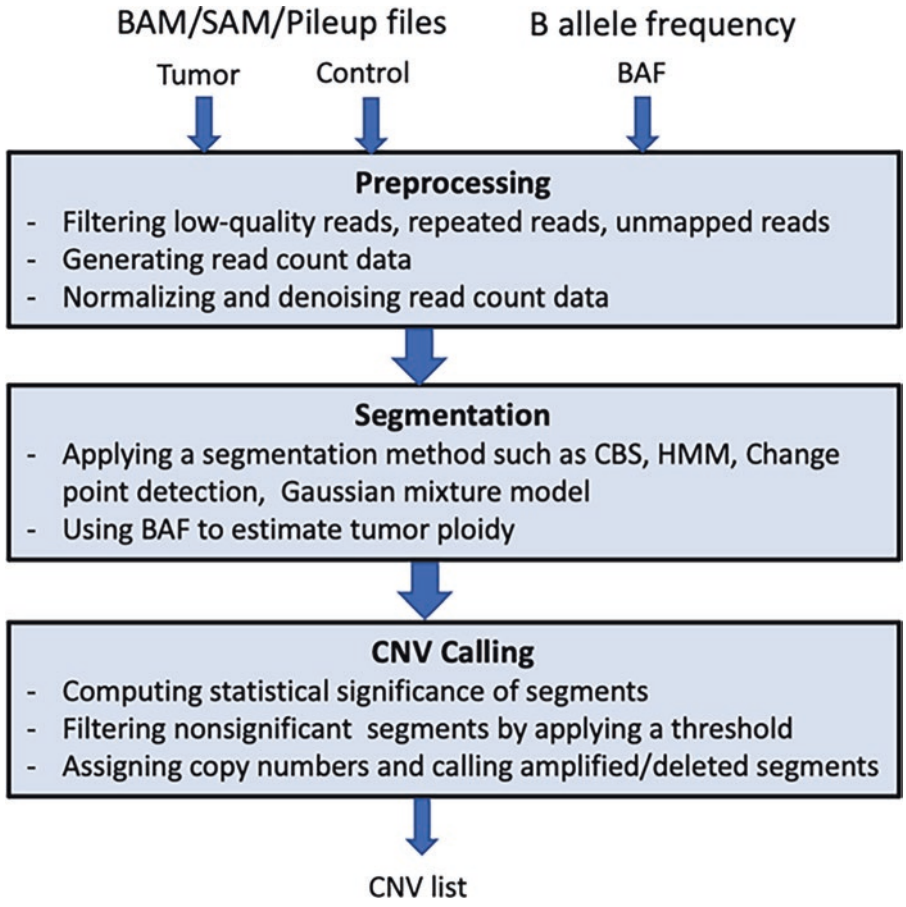

Fig. 4.4 Overall workflow for detecting CNVs that consists of three main parts: preprocessing, segmentation, and CNV Calling

line CNVs in patients. While the inclusion of a matched normal sample is a powerful strategy for somatic CNV detection and to reduce some biases, it cannot remove all the biases. The main sources of bias are mappability and GC content. GC content varies significantly along the genome and has been found to influence read coverage on most sequencing platforms [65, 97, 98]. Sequencing technologies behave differently on sequences with different GC content due to biochemical differences in the sequenced DNA [99]. It has been observed that regions with low or high GC content have low read counts compared to other regions. In fact, there is a unimodal relationship between read counts and G and C bases in a genome [100–102]. The GC content bias is neither linear nor consistent among different samples. Although the global structure of the distribution of read counts with respect to the GC content (GC bias curve) is consistent, the exact shape varies considerably across samples. Furthermore, a huge number of sequencing reads cannot be clearly mapped to the reference genome due to short length of reads and the presence of repetitive regions within the reference genome. Especially in WES data, some regions of the genome have low or no coverage. Mutations and sequencing errors can lead to incorrectly mapped reads as well. These errors introduce a challenge to the alignment process resulting in a mappability bias [58].

To compensate for GC and mappability biases, several RD-based methods [103–106], use the ratio of tumor to normal read counts. Many stud-

ies have been conducted on addressing GC and mappability biases by employing smoothing techniques. Several methods have been proposed, of which the most popular is the Loess (locally estimated scatterplot smoothing) regression method [102, 106–108]. Loess regression is a nonparametric technique that uses local weighted regression to fit a smooth curve through data points. Other methods such as moving average, discrete Wavelet transform, and Total Variation methods have also been used for smoothing read count data. Preprocessing and normalization have a big impact on the total performance of CNV detection tools and need to be considered when choosing a tool for CNV analysis or developing a new CNV detection method.

In the segmentation step, a common strategy is to make an assumption about the distribution of read count data and to apply a statistical approach to merge the neighboring regions (adjacent exonic regions in WES data) with similar read counts to estimate a CNV segment. Most of the tools assume that the read count distribution is Gaussian, whereas a few assume a negative binomial distribution or a Poisson distribution. Different tools are designed for different characteristics of read count data, for example some tools are designed for detecting small CNVs such as CNV-RF [109], and some tools are designed for low coverage data such as ACE [110]. As shown in Tables 4.1, 4.2, and 4.3, variants of few statistical approaches, such as circular binary segmentation (CBS), hidden Markov model (HMM), change-point detection, Gaussian mixture model, regression, Bayesian analysis, and Random Forest, have been used for calling CNVs. Some heuristic statistical analyses have been also used in a few tools.

The most commonly used statistical methods for segmentation are CBS, HMM, change-point analysis, and the Gaussian mixture model. In CBS, the algorithm recursively localizes the breakpoints by changing genomic positions until the chromosomes are divided into segments with equal copy numbers that are significantly different from the copy numbers of their adjacent genomic regions. CBS can also be seen as a change-point detection method. In HMM, the read count data points are sequentially binned along the chromosome according to whether they are likely to measure an amplification, a deletion, or a region in which no copy number change occurred. In change-point analysis, genomic locations at which the probability distribution of the sequence of read counts changes are identified. These change points are marked as CNV breakpoints and the means or medians of read count values between the breakpoints are computed as CNV values. This approach can also offer significance values (p-values) for the detected breakpoints. In the Gaussian mixture approach, it is assumed that all the read counts are generated from a mixture of a finite number of Gaussian distributions and a Bayesian approach is used to assign read counts to CNV levels.

Finally, in the CNV calling step, statistical significance of detected segments and their copy numbers are evaluated for calling segments as amplified, deleted (with their copy number), or normal. In this step, mostly a user-defined threshold is used to filter out nonsignificant CNV segments.

While these steps are used in CNV detection methods for both WES and WGS data, CNV detection methods developed for WGS data are not suitable to WES data. This is because the main assumptions on read distributions and continuity of read count data points do not hold in WES. In addition, WES data introduce biases due to hybridization, which do not exist in WGS data and are not considered in the CNV detection methods for WGS data. Some CNV detection methods for WES data provide copy numbers only for exonic regions, while some merge adjacent exonic regions to provide genomic regions with CNVs.

In addition to CNV detection, visualization and analyzing CNV profiles of cohorts of patients are important. Few tools have been developed for analyzing CNV results such as GISTIC2 [111] and CNspector [112]. GISTIC is a commonly used method for identifying and visualizing regions of the genome that are significantly amplified or deleted across a set of samples. CNspector is a more recent tool which is a web-

based browser to visualize CNV calls. A few tools such as RUbioSeq+ [113] have also been developed that use other CNV detection methods, for example, Contra [114], and provide a user friendly pipeline for visualization and analyzing sequencing data. Some tools have also gone further providing cancer susceptibility risk using CNV profiles (such as AluScanCNV2 [94]).

Conclusion

Next-generation sequencing technologies have a great potential to change cancer research and clinical practice toward precision oncology. NGS offers cheap, fast, and accurate results, and generates huge amounts of data. This has now shifted the challenge from generating data to analyzing them. Consequently, many software tools have been developed for analyzing sequencing data, providing a robust means to make important genomic discoveries and to identify novel genomic biomarkers, which could have not been possible without them.

The emerging cancer care approach is moving from standard treatments based on cancer types, to focus on molecularly guided therapies based on each individual patient's genomic variants. CNVs, which are deletion or multiplication of segments of genomes, are major genomic variants and have been associated with cancer prognosis. Many studies have identified actionable and druggable CNVs and have reported survival improvements in patients who received CNV guided (variant matching) therapies. CNV detection algorithms have had a significant impact on advancing precision oncology by facilitating identification of actionable CNVs accurately and precisely.

As explained in this chapter, CNV detection is complex and faces computational and analytical challenges. Many tools and pipelines have been developed for CNV detection that have contributed significantly to advancing cancer studies and cancer care. Even though CNV detection has made significant progress and many tools are available, computational pipelines for detecting CNVs are not well standardized and vary significantly from study to study and even from sample to sample. Also, consistency among methods in calling CNVs is not high. As a result, precise and accurate CNV detection needs extra care and consideration. However, even though there are some limitations in CNV detection, the crucial role of these tools in discovering actionable genomic biomarkers and in advancing the field of cancer genomics and precision oncology cannot be overlooked. With the emergence of new statistical and computational methods, such as machine learning approaches especially now that increasing amount of cancer sequencing data are available, and with the development of new sequencing technologies, such as single-cell sequencing, long-read sequencing, and linked-read sequencing, we expect to see more improvement in CNV detection. Therefore, we expect to achieve improved identification of more actionable CNVs and significant progress in cancer treatment and precision oncology.

References

1. Barnes MR. Genetic variation analysis for biomedical researchers: a primer. Methods Mol Biol. 2010;628:1–20.
2. Stankiewicz P, Lupski JR. Structural variation in the human genome and its role in disease. Annu Rev Med. 2010;61:437–55.
3. Wain LV, Armour JA, Tobin MD. Genomic copy number variation, human health, and disease. Lancet. 2009;374:340–50.
4. Menghi F, Barthel FP, Yadav V, Tang M, Ji B, Tang Z, et al. The tandem duplicator phenotype is a prevalent genome-wide cancer configuration driven by distinct gene mutations. Cancer Cell. 2018;34:197–210.e5.
5. Henrichsen CN, Chaignat E, Reymond A. Copy number variants, diseases and gene expression. Hum Mol Genet. 2009;18:R1–8.
6. Beckmann JS, Estivill X, Antonarakis SE. Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. Nat Rev Genet. 2007;8:639–46.
7. Weischenfeldt J, Symmons O, Spitz F, Korbel JO. Phenotypic impact of genomic structural variation: insights from and for human disease. Nat Rev Genet. 2013;14:125–38.
8. Almal SH, Padh H. Implications of gene copy-number variation in health and diseases. J Hum Genet. 2012;57:6–13.

9. Fanciulli M, Petretto E, Aitman T. Gene copy number variation and common human disease. Clin Genet. 2010;77:201–13.
10. Shao X, Lv N, Liao J, Long J, Xue R, Ai N, et al. Copy number variation is highly correlated with differential gene expression: a pan-cancer study. BMC Med Genet. 2019;20:175.
11. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature. 2009;458:719–24.
12. The International SNP Map Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature. 2001;409:928–33.
13. †The International HapMap Consortium. The International HapMap Project. Nature. 2003;426:789–96.
14. Pankratz N, Dumitriu A, Hetrick KN, Sun M, Latourelle JC, Wilk JB, et al. Copy number variation in familial Parkinson disease. PLoS One. 2011;6:e20988.
15. Jiang Q, Ho Y-Y, Hao L, Nichols Berrios C, Chakravarti A. Copy number variants in candidate genes are genetic modifiers of Hirschsprung disease. PLoS One. 2011;6:e21219.
16. Grayson BL, Smith ME, Thomas JW, Wang L, Dexheimer P, Jeffrey J, et al. Genome-wide analysis of copy number variation in type 1 diabetes. PLoS One. 2010;5:e15393.
17. Levy D, Ronemus M, Yamrom B, Lee Y, Leotta A, Kendall J, et al. Rare De Novo and transmitted copy-number variation in autistic spectrum disorders. Neuron. 2011;70:886–97.
18. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, et al. Strong Association of De Novo Copy Number Mutations with Autism. Science. 2007;316:445–9.
19. Pinto D, Pagnamenta AT, Klei L, Anney R, Merico D, Regan R, et al. Functional impact of global rare copy number variation in autism spectrum disorders. Nature. 2010;466:368–72.
20. Brouwers N, Van Cauwenberghe C, Engelborghs S, Lambert J-C, Bettens K, Le Bastard N, et al. Alzheimer risk associated with a copy number variation in the complement receptor 1 increasing C3b/C4b binding sites. Mol Psychiatry. 2012;17:223–33.
21. Kirov G. The role of copy number variation in schizophrenia. Expert Rev Neurother. 2010;10:25–32.
22. Shlien A, Malkin D. Copy number variations and cancer. Genome Med. 2009;1:62.
23. Meyerson M, Gabriel S, Getz G. Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet. 2010;11:685–96.
24. Speleman F, Kumps C, Buysse K, Poppe B, Menten B, De Preter K. Copy number alterations and copy number variation in cancer: close encounters of the bad kind. Cytogenet Genome Res. 2008;123:176–82.
25. Weir B, Zhao X, Meyerson M. Somatic alterations in the human cancer genome. Cancer Cell. 2004;6:433–8.
26. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, et al. Pan-cancer patterns of somatic copy number alteration. Nat Genet. 2013;45:1134–40.
27. Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, et al. The landscape of somatic copy-number alteration across human cancers. Nature. 2010;463:899–905.
28. CNV Tools and Software. https://bioinformaticshome.com/tools/cnv/cnv.html.
29. Albertson DG, Collins C, McCormick F, Gray JW. Chromosome aberrations in solid tumors. Nat Genet. 2003;34:369–76.
30. Kircher M, Kelso J. High-throughput DNA sequencing – concepts and limitations. BioEssays. 2010;32:524–36.
31. PCAWG Structural Variation Working Group, PCAWG Consortium, Li Y, Roberts ND, Wala JA, Shapira O, et al. Patterns of somatic structural variation in human cancer genomes. Nature. 2020, 578:112–21.
32. Dancey JE, Bedard PL, Onetto N, Hudson TJ. The genetic basis for cancer treatment decisions. Cell. 2012;148:409–20.
33. Carter NP. Methods and strategies for analyzing copy number variation using DNA microarrays. Nat Genet. 2007;39 7 Suppl:S16–21.
34. Metzker ML. Sequencing technologies — the next generation. Nat Rev Genet. 2010;11:31–46.
35. Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K, et al. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature. 2008;456:66–72.
36. Ku CS, Loy EY, Salim A, Pawitan Y, Chia KS. The discovery of human genetic variations and their use as disease markers: past, present and future. J Hum Genet. 2010;55:403–15.
37. Amarasinghe KC, Li J, Hunter SM, Ryland GL, Cowin PA, Campbell IG, et al. Inferring copy number and genotype in tumour exome data. BMC Genomics. 2014;15:732.
38. Ruan J, Liu Z, Sun M, Wang Y, Yue J, Yu G. DBS: a fast and informative segmentation algorithm for DNA copy number analysis. BMC Bioinformatics. 2019;20:1.
39. Pfundt R, del Rosario M, Vissers LELM, Kwint MP, Janssen IM, de Leeuw N, et al. Detection of clinically relevant copy-number variants by exome sequencing in a large cohort of genetic disorders. Genet Med. 2017;19:667–75.
40. de Ligt J, Boone PM, Pfundt R, Vissers LELM, Richmond T, Geoghegan J, et al. Detection of clinically relevant copy number variants with whole-exome sequencing. Hum Mutat. 2013;34:1439–48.
41. Rabbani B, Tekin M, Mahdieh N. The promise of whole-exome sequencing in medical genetics. J Hum Genet. 2014;59:5–15.
42. Rennert H, Eng K, Zhang T, Tan A, Xiang J, Romanel A, et al. Development and validation of a whole-exome sequencing test for simultaneous detection

of point mutations, indels and copy-number alterations for precision cancer care. NPJ Genomic Med. 2016;1:16019.
43. Van Allen EM, Wagle N, Stojanov P, Perrin DL, Cibulskis K, Marlow S, et al. Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. Nat Med. 2014;20:682–8.
44. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, et al. Origins and functional impact of copy number variation in the human genome. Nature. 2010;464:704–12.
45. Gonzaga-Jauregui C, Lupski JR, Gibbs RA. Human genome sequencing in health and disease. Annu Rev Med. 2012;63:35–61.
46. Belkadi A, Bolze A, Itan Y, Cobat A, Vincent QB, Antipenko A, et al. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. Proc Natl Acad Sci U S A. 2015;112:5473–8.
47. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, et al. Mapping and sequencing of structural variation from eight human genomes. Nature. 2008;453:56–64.
48. Yates LR, Seoane J, Le Tourneau C, Siu LL, Marais R, Michiels S, et al. The European Society for Medical Oncology (ESMO) precision medicine glossary. Ann Oncol. 2018;29:30–5.
49. Haslem DS, Van Norman SB, Fulde G, Knighton AJ, Belnap T, Butler AM, et al. A retrospective analysis of precision medicine outcomes in patients with advanced cancer reveals improved progression-free survival without increased health care costs. JOP. 2017;13:e108–19.
50. Kris MG, Johnson BE, Berry LD, Kwiatkowski DJ, Iafrate AJ, Wistuba II, et al. Using multiplexed assays of oncogenic drivers in lung cancers to select targeted drugs. JAMA. 2014;311:1998.
51. Tsimberidou A-M, Iskander NG, Hong DS, Wheler JJ, Falchook GS, Fu S, et al. Personalized medicine in a phase I clinical trials program: the MD Anderson Cancer Center initiative. Clin Cancer Res. 2012;18:6373–83.
52. Wilson MA, Zhao F, Khare S, Roszik J, Woodman SE, D'Andrea K, et al. Copy number changes are associated with response to treatment with carboplatin, paclitaxel, and sorafenib in melanoma. Clin Cancer Res. 2016;22:374–82.
53. Prawira A, Pugh TJ, Stockley TL, Siu LL. Data resources for the identification and interpretation of actionable mutations by clinicians. Ann Oncol. 2017;28:946–57.
54. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. 2015;43:D805–11.
55. Allegretti M, Fabi A, Buglioni S, Martayan A, Conti L, Pescarmona E, et al. Tearing down the walls: FDA approves next generation sequencing (NGS) assays for actionable cancer genomic aberrations. J Exp Clin Cancer Res. 2018;37:47, s13046-018-0702–x.
56. Cheng DT, Mitchell TN, Zehir A, Shah RH, Benayed R, Syed A, et al. Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT). J Mol Diagn. 2015;17:251–64.
57. Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. BMC Bioinformatics. 2013;14 Suppl:11.
58. Teo SM, Pawitan Y, Ku CS, Chia KS, Salim A. Statistical challenges associated with detecting copy number variations with next-generation sequencing. Bioinformatics. 2012;28:2711–8.
59. Zhang ZD, Du J, Lam H, Abyzov A, Urban AE, Snyder M, et al. Identification of genomic indels and structural variations using split reads. BMC Genomics. 2011;12:375.
60. Yoon S, Xuan Z, Makarov V, Ye K, Sebat J. Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res. 2009;19:1586–92.
61. Nijkamp JF, van den Broek MA, Geertman J-MA, Reinders MJT, Daran J-MG, de Ridder D. De novo detection of copy number variation by co-assembly. Bioinformatics. 2012;28:3195–202.
62. Xi R, Lee S, Park PJ. A survey of copy-number variation detection tools based on high-throughput sequencing data. Curr Protoc Hum Genet. 2012;75 https://doi.org/10.1002/0471142905.hg0719s75.
63. Pirooznia M, Goes FS, Zandi PP. Whole-genome CNV analysis: advances in computational approaches. Front Genet. 2015;6:138.
64. Liu B, Morrison C, Johnson C, Trump D, Qin M, Conroy J, et al. Computational methods for detecting copy number variations in cancer genome using next generation sequencing: principles and challenges. Oncotarget. 2013;4. http://www.impactjournals.com/oncotarget/index.php?journal=oncotarget&page=article&op=view&path%5B%5D=1537.
65. Minoche AE, Dohm JC, Himmelbauer H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. Genome Biol. 2011;12:R112.
66. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet. 2011; https://doi.org/10.1038/nrg3117.
67. Ulahannan D, Kovac MB, Mulholland PJ, Cazier J-B, Tomlinson I. Technical and implementation issues in using next-generation sequencing of cancers in clinical practice. Br J Cancer. 2013;109:827–35.
68. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet. 2014;15:121–32.
69. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T. Absolute quantification of somatic

DNA alterations in human cancer. Nat Biotechnol. 2012;30 https://doi.org/10.1038/nbt.2203.
70. Oesper L, Satas G, Raphael BJ. Quantifying tumor heterogeneity in whole-genome and whole-exome sequencing data. Bioinformatics. 2014;30 https://doi.org/10.1093/bioinformatics/btu651.
71. Luo Z, Fan X, Su Y, Huang YS. Accurity: accurate tumor purity and ploidy inference from tumor-normal WGS data by jointly modelling somatic copy number alterations and heterozygous germline single-nucleotide-variants. Bioinformatics. 2018;34:2004–11.
72. Zhu W, Kuziora M, Creasy T, Lai Z, Morehouse C, Guo X, et al. BubbleTree: an intuitive visualization to elucidate tumoral aneuploidy and clonality using next generation sequencing data. Nucleic Acids Res. 2016;44:e38.
73. Bao L, Pu M, Messer K. AbsCN-seq: a statistical method to estimate tumor purity, ploidy and absolute copy numbers from next-generation sequencing data. Bioinformatics. 2014;30:1056–63.
74. Li Y, Xie X. MixClone: a mixture model for inferring tumor subclonal populations. BMC Genomics. 2015;16:S1.
75. Rajagopalan H, Lengauer C. Aneuploidy and cancer. Nature. 2004;432:338–41.
76. Mayrhofer M, DiLorenzo S, Isaksson A. Patchwork: allele-specific copy number analysis of whole-genome sequenced tumor tissue. Genome Biol. 2013;14:R24.
77. Favero F, Joshi T, Marquard AM, Birkbak NJ, Krzystanek M, Li Q, et al. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. Ann Oncol. 2015;26:64–70.
78. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, et al. Tumour evolution inferred by single-cell sequencing. Nature. 2011;472:90–4.
79. Zare F, Dow M, Monteleone N, Hosny A, Nabavi S. An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. BMC Bioinformatics. 2017;18:286.
80. Tan R, Wang Y, Kleinstein SE, Liu Y, Zhu X, Guo H, et al. An evaluation of copy number variation detection tools from whole-exome sequencing data. Hum Mutat. 2014;35:899–907.
81. Bartenhagen C, Dugas M. RSVSim: an R/Bioconductor package for the simulation of structural variations. Bioinformatics. 2013;29:1679–81.
82. Greg Faust. SVsim: a tool that generates synthetic Structural Variant calls as benchmarks to test/evaluate SV calling pipelines. https://github.com/GregoryFaust/SVsim.
83. Qin M, Liu B, Conroy JM, Morrison CD, Hu Q, Cheng Y, et al. SCNVSim: somatic copy number variation and structure variation simulator. BMC Bioinformatics. 2015;16:66.
84. Xia Y, Liu Y, Deng M, Xi R. Pysim-sv: a package for simulating structural variation data with GC-biases. BMC Bioinformatics. 2017;18:53.
85. Pattnaik S, Gupta S, Rao AA, Panda B. SInC: an accurate and fast error-model based simulator for SNPs, Indels and CNVs coupled with a read generator for short-read sequence data. BMC Bioinformatics. 2014;15:40.
86. Abdelrahman Hosny. Copy Number Variation Simulator (CNV-Sim). https://nabavilab.github.io/CNV-Sim/.
87. Xing Y, Dabney AR, Li X, Wang G, Gill CA, Casola C. SECNVs: a simulator of copy number variants and whole-exome sequences from reference genomes. Front Genet. 2020;11:82.
88. Pounraja VK, Jayakar G, Jensen M, Kelkar N, Girirajan S. A machine-learning approach for accurate detection of copy number variants from exome sequencing. Genome Res. 2019;29:1134–43.
89. Gao J, Wan C, Zhang H, Li A, Zang Q, Ban R, et al. Anaconda: AN automated pipeline for somatic COpy number variation detection and annotation from tumor exome sequencing data. BMC Bioinformatics. 2017;18:436.
90. Zhang L, Bai W, Yuan N, Du Z. Comprehensively benchmarking applications for detecting copy number variation. PLoS Comput Biol. 2019;15:e1007069.
91. Putnam DK, Ma X, Rice SV, Liu Y, Newman S, Zhang J, et al. VCF2CNA: a tool for efficiently detecting copy-number alterations in VCF genotype data and tumor purity. Sci Rep. 2019;9:10357.
92. Chen H, Bell JM, Zavala NA, Ji HP, Zhang NR. Allele-specific copy number profiling by next-generation DNA sequencing. Nucleic Acids Res. 2015;43:e23.
93. Zhang Z, Hao K. SAAS-CNV: a joint segmentation approach on aggregated and allele specific signals for the identification of somatic copy number alterations with next-generation sequencing data. PLoS Comput Biol. 2015;11:e1004618.
94. Hu T, Chen S, Ullah A, Xue H. AluScanCNV2: an R package for copy number variation calling and cancer risk prediction with next-generation sequencing data. Genes Dis. 2019;6:43–6.
95. Chang L-C, Das B, Lih C-J, Si H, Camalier CE, McGregor PM, et al. RefCNV: identification of gene-based copy number variants using whole exome sequencing. Cancer Inform. 2016:15, CIN. S36612.
96. Silva GO, Siegel MB, Mose LE, Parker JS, Sun W, Perou CM, et al. SynthEx: a synthetic-normal-based DNA sequencing tool for copy number alteration detection and tumor heterogeneity profiling. Genome Biol. 2017;18:66.
97. Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. Nat Methods. 2009;6:291–5.
98. Dohm JC, Lottaz C, Borodina T, Himmelbauer H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res. 2008;36:e105.

99. Rieber N, Zapatka M, Lasitschka B, Jones D, Northcott P, Hutter B, et al. Coverage bias and sensitivity of variant calling for four whole-genome sequencing technologies. PLoS One. 2013;8:e66621.
100. Iakovishina D, Janoueix-Lerosey I, Barillot E, Regnier M, Boeva V. SV-Bay: structural variant detection in cancer genomes using a Bayesian approach with correction for GC-content and read mappability. Bioinformatics. 2016;32:984–92.
101. Boeva V, Zinovyev A, Bleakley K, Vert J-P, Janoueix-Lerosey I, Delattre O, et al. Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. Bioinformatics (Oxford, England). 2011;27:268–9.
102. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. 2012;40:e72.
103. Chiang DY, Getz G, Jaffe DB, O'Kelly MJT, Zhao X, Carter SL, et al. High-resolution mapping of copy-number alterations with massively parallel sequencing. Nat Methods. 2009;6:99–103.
104. Xi R, Hadjipanayis AG, Luquette LJ, Kim T-M, Lee E, Zhang J, et al. Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. Proc Natl Acad Sci U S A. 2011;108:E1128–36.
105. Xie C, Tammi MT. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. BMC Bioinformatics. 2009;10:80.
106. Gusnanto A, Wood HM, Pawitan Y, Rabbitts P, Berri S. Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data. Bioinformatics. 2012;28:40–7.
107. Liao C, Yin A-H, Peng C-F, Fu F, Yang J-X, Li R, et al. Noninvasive prenatal diagnosis of common aneuploidies by semiconductor sequencing. Proc Natl Acad Sci. 2014;111:7415–20.
108. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22:568–76.
109. Onsongo G, Baughn LB, Bower M, Henzler C, Schomaker M, Silverstein KAT, et al. CNV-RF is a random forest–based copy number variation detection method using next-generation sequencing. J Mol Diagn. 2016;18:872–81.
110. Poell JB, Mendeville M, Sie D, Brink A, Brakenhoff RH, Ylstra B. ACE: absolute copy number estimation from low-coverage whole-genome sequencing data. Bioinformatics. 2019;35:2847–9.
111. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011;12:R41.
112. Markham JF, Yerneni S, Ryland GL, Leong HS, Fellowes A, Thompson ER, et al. CNspector: a web-based tool for visualisation and clinical diagnosis of copy number variation from next generation sequencing. Sci Rep. 2019;9:6426.
113. Rubio-Camarillo M, López-Fernández H, Gómez-López G, Carro Á, Fernández JM, Torre CF, et al. RUbioSeq+: a multiplatform application that executes parallelized pipelines to analyse next-generation sequencing data. Comput Methods Prog Biomed. 2017;138:73–81.
114. Li J, Lupat R, Amarasinghe KC, Thompson ER, Doyle MA, Ryland GL, et al. CONTRA: copy number analysis for targeted resequencing. Bioinformatics. 2012;28:1307–13.
115. Budczies J, Pfarr N, Romanovsky E, Endris V, Stenzinger A, Denkert C. Ioncopy: an R Shiny app to call copy number alterations in targeted NGS data. BMC Bioinformatics. 2018;19:157.
116. Yu Z, Li A, Wang M. CloneCNA: detecting subclonal somatic copy number alterations in heterogeneous tumor samples from whole-exome sequencing data. BMC Bioinformatics. 2016;17:310.
117. Zhang H, Zhan X, Brugarolas J, Xie Y. DEFOR: depth- and frequency-based somatic copy number alteration detector. Bioinformatics. 2019;35:3824–5.
118. D'Aurizio R, Pippucci T, Tattini L, Giusti B, Pellegrini M, Magi A. Enhanced copy number variants detection from whole-exome sequencing data using EXCAVATOR2. Nucleic Acids Res. 2016:gkw695.
119. Sathirapongsasuti JF, Lee H, Horst BAJ, Brunner G, Cochran AJ, Binder S, et al. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. Bioinformatics. 2011;27:2648–54.
120. Love MI, Myšičková A, Sun R, Kalscheuer V, Vingron M, Haas SA. Modeling read counts for CNV detection in exome sequencing data. Stat Appl Genet Mol Biol. 2011;10 https://doi.org/10.2202/1544-6115.1732.
121. Plagnol V, Curtis J, Epstein M, Mok KY, Stebbings E, Grigoriadou S, et al. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. Bioinformatics. 2012;28:2747–54.
122. Choo-Wosoba H, Albert PS, Zhu B. hsegHMM: hidden Markov model-based allele-specific copy number alteration analysis accounting for hypersegmentation. BMC Bioinformatics. 2018;19:424.
123. Jiang Y, Wang R, Urrutia E, Anastopoulos IN, Nathanson KL, Zhang NR. CODEX2: full-spectrum copy number variation detection by high-throughput DNA sequencing. Genome Biol. 2018;19:202.
124. Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing. PLoS Comput Biol. 2016;12:e1004873.
125. Boeva V, Popova T, Lienard M, Toffoli S, Kamal M, Le Tourneau C, et al. Multi-factor data normalization enables the detection of copy number aberrations in amplicon sequencing data. Bioinformatics. 2014;30:3443–50.

126. Riester M, Singh AP, Brannon AR, Yu K, Campbell CD, Chiang DY, et al. PureCN: copy number calling and SNV classification using targeted short read sequencing. Source Code Biol Med. 2016;11:13.
127. Yu Z, Li A, Wang M. CLImAT-HET: detecting subclonal copy number alterations and loss of heterozygosity in heterogeneous tumor samples from whole-genome sequencing data. BMC Med Genet. 2017;10:15.
128. Yuan X, Yu J, Xi J, Yang L, Shang J, Li Z, et al. CNV_IFTV: an isolation forest and total variation-based detection of CNVs from short-read sequencing data. IEEE/ACM Trans Comput Biol Bioinf. 2019:1.
129. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 2011;21:974–84.
130. Yuan X, Bai J, Zhang J, Yang L, Duan J, Li Y, et al. CONDEL: detecting copy number variation and genotyping deletion zygosity from single tumor samples using sequence data. IEEE/ACM Trans Comput Biol Bioinf. 2018:1.
131. Krishnan NM, Gaur P, Chaudhary R, Rao AA, Panda B. COPS: a sensitive and accurate tool for detecting somatic copy number alterations using short-read sequence data from paired samples. PLoS One. 2012;7:e47812.
132. Miller CA, Hampton O, Coarfa C, Milosavljevic A. ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. PLoS One. 2011;6:e16327.
133. Zhang M, Liu D, Tang J, Feng Y, Wang T, Dobbin KK, et al. SEG – a software program for finding somatic copy number alterations in whole genome sequencing data of cancer. Comput Struct Biotechnol J. 2018;16:335–41.
134. Mosen-Ansorena D, Telleria N, Veganzones S, la Orden V, Maestro M, Aransay AM. seqCNA: an R package for DNA copy number analysis in cancer using high-throughput sequencing. BMC Genomics. 2014;15:178.
135. Ha G, Roth A, Khattra J, Ho J, Yap D, Prentice LM, et al. TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. Genome Res. 2014;24:1881–93.
136. Holt C, Losic B, Pai D, Zhao Z, Trinh Q, Syam S, et al. WaveCNV: allele-specific copy number alterations in primary tumors and xenograft models from next-generation sequencing. Bioinformatics. 2014;30:768–74.
137. Magi A, Pippucci T, Sidore C. XCAVATOR: accurate detection and genotyping of copy number variants from second and third generation whole-genome sequencing experiments. BMC Genomics. 2017;18:747.
138. Xi R, Lee S, Xia Y, Kim T-M, Park PJ. Copy number analysis of whole-genome data using BIC-seq2 and its application to detection of cancer susceptibility variants. Nucleic Acids Res. 2016;44:6274–86.
139. Li Y, Zhang J, Yuan X. BagGMM: calling copy number variation by bagging multiple Gaussian mixture models from tumor and matched normal next-generation sequencing data. Digit Signal Process. 2019;88:90–100.
140. Klambauer G, Schwarzbauer K, Mayr A, Clevert D-A, Mitterecker A, Bodenhofer U, et al. cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. Nucleic Acids Res. 2012;40:e69.
141. Shen R, Seshan VE. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. Nucleic Acids Res. 2016;44:e131.

Check for updates

5 Assessment of Microsatellite Instability from Next-Generation Sequencing Data

Victor Renault, Emmanuel Tubacher, and Alexandre How-Kit

Abstract

Microsatellite instability (MSI) is a genetic alteration due to a deficiency of the DNA mismatch repair system, where microsatellites accumulate insertions/deletions. This phenotype has been extensively characterized in colorectal cancer and is also sought in the context of Lynch syndrome diagnosis. It has recently been described in dozens of cancer types from whole genome/exome sequencing data, bearing some prognostic information. Moreover, MSI has also proven to be a major predicator of the response to immune checkpoint blockade therapy in solid cancer patients. Among the different methods developed for MSI detection in cancer, next-generation sequencing (NGS) is a promising and versatile technology offering many possibilities and advantages in diverse clinical applications compared to the gold standard PCR and capillary electrophoresis approach. NGS could notably increase the number of analyzed microsatellites and potentially be used to analyze other genetic alterations required for precision oncology. However, it requires the development of robust new computational algorithms for the analysis of NGS microsatellite data. In this chapter, we describe the different approaches developed for the assessment of MSI from NGS data in cancer, including the different microsatellite panels and computational algorithms proposed, highlighting their advantages and drawbacks, and their evaluation in different clinical applications.

Introduction

Microsatellites are short tandem repeats of 1–6 nucleotides ubiquitously present throughout the genome, and whose polymorphism is based on the number of repetitions of the repeat motif [22]. These highly polymorphic sequences have been used since the 1990s for cancer diagnosis to detect microsatellite instability (MSI/MSI-H) [11]. MSI is a widespread genomic alteration caused by a deficiency of the DNA mismatch repair system (dMMR) that normally allows the repair and correction of DNA strands in the presence of DNA mismatches introduced during DNA replication due to polymerase slippage [11, 22]. Thus, the MSI phenotype is characterized by the accumulation of microsatellite mutations corresponding to insertions or deletions of several nucleotides that are multiples of the repeated unit [11, 18, 38]. MSI has been extensively studied in

V. Renault · E. Tubacher
Laboratory for Bioinformatics, Foundation Jean Dausset – CEPH, Paris, France

A. How-Kit (✉)
Laboratory for Genomics, Foundation Jean Dausset – CEPH, Paris, France
e-mail: alexandre.how-kit@fjd-ceph.org

© Springer Nature Switzerland AG 2022

A. Laganà (ed.), *Computational Methods for Precision Oncology*, Advances in Experimental Medicine and Biology 1361, https://doi.org/10.1007/978-3-030-91836-1_5

colorectal cancer (CRC), where 15–20% of tumors present this phenotype associated with better patient survival in stage II and III CRCs [11]. MSI CRC can have a sporadic origin or arise in the context of Lynch syndrome and more rarely CMMRD syndrome that are caused by mono- and bi-allelic germline mutations of one of the four MMR genes (*MLH1*, *MSH2*, *MSH6*, or *PMS2*), respectively [11, 93]. The presence of MSI was also well known for a long time in other cancer types including gastric [81] and endometrial [98] cancers, but it was the recent comprehensive pan-cancer analyses of whole genomes/exomes data of The Genome Cancer Atlas program (TGCA) that showed the genome-wide occurrence of MSI in several other cancer types [13, 18, 38]. Moreover, it has been shown that MSI also bears some prognostic information and is also a predicator of the efficacy of immune checkpoint blockade therapy in solid tumors, which presents a great clinical interest for cancer patients and might help clinicians for therapeutic decision making [38, 59, 60].

The gold standard approach for MSI detection in cancer relies on PCR followed by capillary electrophoresis fragment analysis (MSI-PCR) using different panels of microsatellites aiming to provide the highest degree of sensitivity and specificity [6, 55, 86]. Other methods relying on the modification of standard procedures have also been developed to improve the limit of detection (LOD) of MSI, as required for some applications (e.g., to detect MSI in blood and in tumors with a high level of normal cell contamination or in precancerous lesions) [7, 19, 43, 45, 54]. Recently, next-generation sequencing (NGS) has also been used for the detection of MSI in cancer, which required the development of new computational algorithms allowing the analysis of sequencing data from a far greater number of microsatellites (from less than ten to tens of thousands) [6]. These bioinformatic methods allowed the detection of MSI using different approaches either based on (i) the average mutational burdens, (ii) the percentage of unstable loci determined by comparison with non-tumoral samples, or (iii) more complex machine learning classifiers. The three approaches have been applied to different types of clinical samples [6]. These NGS-based MSI detection methods showed comparable or sometimes better sensitivity for MSI detection compared to the gold standard approach and have been evaluated in different types of cancers and for diverse clinical applications. In this chapter, we describe the different panels of microsatellites and computational approaches developed to date for the detection of MSI in cancer from NGS experiments. We highlight and compare the advantages and drawbacks of these methods and demonstrate their potential applications in routine clinical testing.

The Need of Sensitive and Specific Microsatellite Panels for MSI Assessment from NGS Data

Sensitive Microsatellite Markers Used for MSI Detection in Cancer with the Gold Standard Method

MSI was initially described in 1993 in which deletions in several microsatellites were found in CRC samples [1, 46, 84]. Some studies also showed that the percentage of tumors with MSI could greatly vary depending on the microsatellites used for MSI detection, suggesting that they presented different stabilities for MMR deficiency [1, 9, 17, 20, 85]. Thus, the first parameter influencing the sensitivity and specificity of MSI detection in cancer is the microsatellite markers used whatever the analytical method chosen, including NGS. A combination of the most sensitive and specific microsatellites whose instability reflects the MMR deficiency should therefore be used for MSI detection in cancer.

A first standardized panel of microsatellites, known as the Bethesda/NCI panel, to be used with PCR and capillary electrophoresis was proposed at the National Cancer Institute (NCI) workshop held in Bethesda (Maryland, USA). The panel was composed of two mono-nucleotide (BAT-25 and BAT-26) and three di-nucleotide (D2S123, D5S346, and D17S250) repeat microsatellites [12]. CRC samples presenting two or more unstable markers should be defined as MSI/

MSI-H, while the others should be classified as MSS (microsatellite stable) or MSI-L (MSI-low) if no markers or only one marker was unstable, respectively [12]. Though still under debate, MSS and MSI-L tumors are generally considered and treated clinically as a single group, while some studies showed that most MSS tumors could be classified as MSI-L tumors if a sufficient number of microsatellite markers are tested [34, 52, 55, 57, 76].

Although the Bethesda/NCI panel is still the current gold standard used in several laboratories for MSI detection, several concerns have arisen regarding this panel including notably the presence of di-nucleotide microsatellites that are less sensitive and less specific compared to mono-nucleotide microsatellites [37, 56, 74]. The use of mono-nucleotide repeat microsatellites have been recommended as they are more sensitive and less polymorphic [16, 37, 42, 56, 74]. The use of individual mono-nucleotide microsatellite markers such as CAT25 and HT17 or inclusion into new panels such as the gold standard pentaplex panel (NR-21, NR-24, BAT-25, BAT-26, and NR-27/MONO-27) have been proposed due to showing higher sensitivities compared to the Bethesda/NCI panel [4, 8, 15, 16, 21, 25, 33, 67].

Pan-Cancer Microsatellite Instability Analysis at a Genome-Wide Scale Using NGS: Some Basic Insights

Since the advent of NGS, almost all microsatellites in the genome can now be analyzed and used to assess MSI in cancer compared to the gold standard approach, which should further improve the sensitivity of MSI detection. The number of microsatellites that can be simultaneously detected and analyzed is dependent on the protocol used for library preparation, ranging from less than ten to several hundred thousand [18, 30, 79]. Recently, two genome-wide studies based on whole genome sequencing (WGS) and whole exome sequencing (WES) data from TCGA assessing MSI in several types of cancers developed their own MSI classifier and presented a comprehensive landscape of microsatellite instability across cancer [18, 38]. These studies showed that most microsatellites in the genome are stable and therefore would not be informative in the case of MMR deficiency [18, 38]. Moreover, most tumors presented some unstable microsatellites, whose overall burden could distinguish MSI/MSI-H from MSS and MSI-L tumors but not MSI-L from MSS tumors, thereby confirming that MSI-L and MSS tumors could be considered as a same phenotype [18, 38]. The studies also showed that some microsatellites could be unstable in both MSS and MSI tumors indicating that they could not be used for MSI/MSI-H assessment. Moreover, MSI preferentially affected di-nucleotide repeat microsatellites in frequency and mono-nucleotide repeat microsatellites in quantity as well as microsatellites located in intergenic, intronic, and 5' and 3'UTR regions [18, 38]. Finally, both studies showed that some unstable microsatellites were inter- and/or intra-tumor-type specific and identified lists of microsatellites that were the most frequently and specifically unstable in MSI-H cancers, which could potentially be used to define new pan-cancer panels with improved sensitivity and specificity for MSI detection [18, 38]. As the microsatellites of the Bethesda/NCI and pentaplex panels were principally developed and recommended for CRC and have shown poor performances in other cancer types [24, 35, 73], there was a need to develop new microsatellite panels for NGS with improved sensitivity and specificity that could be used in several types of malignancies.

Refined Microsatellite Panels for MSI Detection in Cancer Using NGS: Toward Precision Oncology

To date, most studies based on NGS experiments evaluating MSI in cancer used microsatellite sequencing data available from WGS, WES, or targeted gene sequencing (TGS) of panels of genes implicated in cancer that were not initially designed for MSI detection, either in several types of microsatellites (mono- to penta-nucleotide repeat) or refining their analysis to

mono-nucleotide repeat microsatellites [13, 18, 38, 50, 52, 79]. Although genome-wide analyses of MSI in all microsatellites contained in WGS and WES data present a great interest for basic research and the general comprehension of MSI mechanism in tumorigenesis, they are less adapted for translational and clinical applications in oncology due to their costs and low coverage (30 X for WGS and 100 X for WES in average). Thus, these applications require small panels of microsatellites with high clinical sensitivity and specificity for MSI detection that could be applicable to high throughput and to degraded samples such as FFPE or cell-free circulating DNA samples for a reduced cost. Another parameter to consider is the coverage used for the NGS experiments, which was shown to negatively influence the detection of MSI events from 30 X or lower in high purity tumor samples [18]. The coverage should therefore be greatly increased to ensure a sufficient LOD of MSI in clinical samples [75], notably those with a high level of normal DNA contamination in heterogeneous tumors with minor clones bearing the mutations, tumors with a high level of stromal cell contamination, precancerous lesions, and blood or plasma samples.

In a proof-of-concept study, ultra-deep amplicon sequencing (5000–8000 X) of 2 or 5 microsatellites, which included some of the Bethesda/NCI and pentaplex panels microsatellites, has been performed on FFPE tumor samples for MSI detection [30]. Results from NGS experiments presented 100% concordance with the gold standard MSI-PCR method using the NCI and pentaplex panel (Table 5.1), thereby demonstrating the applicability of NGS for MSI testing [30]. Other studies have used sets of microsatellites identified from cancer gene panels initially developed for the detection of single nucleotide variations, fusions, and copy number alterations in the context of precision oncology to guide the therapeutic decision-making. Among them, we can site ColoSeq, UW-OncoPlex, BROCA, MSK-IMPACT, and Caris MI TumorSeek 592-Gene NGS panels capturing 50, 194, 53, 341–468, and 592 genes implicated in cancer where 146, 15, 146, and more than 7000 microsatellites have been used to detect MSI on tumor samples, respectively (Table 5.1), with almost or 100% agreement with the gold standard MSI-PCR using the pentaplex panel [59, 66, 77, 79, 87].

Other microsatellite panels have also been specifically designed and recently published for MSI detection in cancer (Table 5.1). Thus, the ColonCore panel was principally proposed for CRC and allowed the simultaneous detection of MSI and mutations in 36 CRC-related genes and presented high concordance with MSI-PCR and immunohistochemical (IHC) analysis (Table 5.1) [97]. A similar group proposed two panels for MSI detection: the MSIplus panel [40] and a pan-cancer panel of 111 microsatellites [88]. The first panel evaluated MSI in 17 microsatellites and mutations in 3 oncogenes (*KRAS*, *BRAF*, and *NRAS*) and presented slightly improved performances compared to MSI-PCR [40], while the latter panel showed performances comparable to and better than the pentaplex MSI-PCR in colorectal and non-colorectal tumor samples, respectively [88]. Another group also proposed MSI detection by two panels of 17 and 6 microsatellites optimized from two larger panels of 120 and 24 microsatellites, respectively (Table 5.1). They presented 97–100% concordance with the pentaplex MSI-PCR in CRC samples [29, 78]. Recently, three panels of a small number of selected microsatellites (5, 24, and 90) have also been used to detect MSI in samples with very little tumor DNA content. These samples included blood and plasma samples (Table 5.1) and also required the use of unique molecular identifiers (UMI) in the library preparation to sufficiently improve the LOD of MSI [28, 31, 91].

In summary, the choice of microsatellites greatly impacts the sensitivity and specificity of the detection of MSI and should be performed carefully. Although easily feasible with NGS technologies, there is sometimes no need to drastically increase the number of microsatellite loci used for precision oncology purposes. To demonstrate this point, panels with as few as 6 microsatellites allowed highly sensitive MSI detection in clinical samples [29, 39]. Moreover, due to the tumor-type specificity of some unstable microsatellites, the microsatellites used for MSI detection should be selected according to the studied

Table 5.1 Microsatellite panels proposed for MSI detection by NGS

Study	Types of microsatellites[a]	Number of microsatellites in the panel	Panel name	Library type[b]	Computational method used	Sample types[c]	Sensitivity	Specificity	Reference methods
Gan et al. [30]	Mono- and di-NRMS	5 and 2	–	PCR amplicon	In-house scoring method	FFPE	100	100	MSI-PCR using the pentaplex or 9 microsatellites including the NCI/Bethesda panel
Salipante et al. [79]	Mono-NRMS	146	ColoSeq	Gene capture by RNA bait	mSINGs	FFPE	96.4	97.2	MSI-PCR using the pentaplex panel
	Mono-NRMS	15	UW-OncoPlex	Gene capture by RNA bait	mSINGs	FFPE	100	100	
Prichard et al. [77]	Mono-NRMS	146	BROCA	Gene capture by RNA bait	mSINGs	FF and FFPE	–	–	MSI-PCR using the pentaplex panel
Middha et al. [66]	–	≈1000–1500	MSK-IMPACT	Gene capture by RNA bait	MSIsensor	–	96.1	98.5	MSI-PCR using the pentaplex panel, dMMR IHC
Le et al. [59]	–	–	Caris MI TumorSeek	Gene capture by RNA bait	In-house scoring method	FF and FFPE	95.8	99.4	MSI-PCR using the pentaplex panel
Vanderwalde et al. [87]	–	7317	Caris MI TumorSeek	Gene capture by RNA bait	In-house scoring method	FF and FFPE	95.8	99.4	MSI-PCR using the pentaplex panel
Hempelman et al. [40]	Mono-NRMS	17	MSIplus	PCR amplicon	mSINGs	FF and FFPE	97	100	MSI-PCR using the pentaplex panel, dMMR IHC
Zhu et al. [97]	Mono-NRMS	22	MSI-ColonCore	–	MSI-ColonCore	FFPE	97.9	100	MSI-PCR using the pentaplex panel, dMMR IHC
Wallkes et al. [88]	Mono-NRMS	111	–	Gene capture by smMIP	mSINGs	FF and FFPE	95.8–100	100	MSI-PCR using the pentaplex panel
Redford et al. [78]	Mono-NRMS	17	–	PCR amplicon	In-house scoring method	FFPE	100	100	MSI-PCR using the pentaplex panel

(continued)

Table 5.1 (continued)

Study	Types of microsatellites[a]	Number of microsatellites in the panel	Panel name	Library type[b]	Computational method used	Sample types[c]	Sensitivity	Specificity	Reference methods
Gallon et al. [29]	Mono-NRMS	6	–	Gene capture by smMIP	In-house scoring method	FFPE	100	100	MSI-PCR using the pentaplex panel
Willis et al. [91]	Mono- and tri-NRMS	90	Guardant360	Gene capture by RNA bait	In-house scoring method	cfDNA	86.6	99.5	MSI-PCR using the pentaplex panel, dMMR IHC & MSI-NGS on tumors
Georgiadis et al. [31]	Mono-NRMS	5	–	Gene capture by RNA bait	In-house scoring method	cfDNA	–	–	MSI-PCR using the pentaplex panel, dMMR IHC & MSI-NGS on tumors
Gallon et al. [28]	Mono-NRMS	24	–	Gene capture by smMIP	In-house scoring method	wbDNA	97–100	98–100	Germline gMSI and MMR sequencing & clinical diagnosis

[a]*NRMS* nucleotide repeat microsatellites
[b]*smMIP* single-molecule molecular inversion probes
[c]*FF* fresh frozen, *FFPE* formalin-fixed paraffin-embedded, *cfDNA* cell-free DNA, *wbDNA* whole-blood DNA

tumor type or be simply part of a pan-cancer panel [88]. The detection of MSI by NGS has required the development of specific algorithms and computational methods whose choice also affects the sensitivity of MSI detection despite the use of the same microsatellite panel [48, 50].

Computational Approaches for MSI Assessment in Cancer from NGS Data

Challenges and Difficulties of Microsatellite Next-Generation Sequencing Data Analysis

Compared to the gold standard capillary electrophoresis procedures, NGS experiments are much more expensive to perform and the NGS-based computational approaches developed for the detection of MSI require more time to generate results due to the more complex analyses required. Thus, the different methods developed to date needed to take into account several intrinsic parameters and characteristics for the analysis of microsatellite data generated from NGS (WGS, WES, and TGS) experiments. These methods had to manage the formation of stutter artifacts introduced during the PCR amplification steps included in NGS experiments, the homopolymer-induced sequencing errors and the subsequent difficulties for accurate indel calling, the shortcomings of short read sequencing limiting the length of the microsatellites analyzed as well as the high level of microsatellite polymorphism across individuals [23, 47, 83, 95]. The assessment of MSI status using NGS data (mostly WES and WGS) thereby requires calibration and false-positive filtering steps, where local or global realignment may improve MSI calling.

To deal with the difficulties previously mentioned, each computational approach developed their own specificities for microsatellite data analysis and MSI detection. In general, the use of a high number of microsatellite loci for MSI detection by NGS in most approaches could be particularly useful as it bases MSI detection on a high number of MSI events rather than a small number, thus limiting the effect of some false-positive calls that might arise at certain microsatellite loci. This is particularly true for the methods based on mutation burden for the detection of MSI, where specific loci are not considered as the MSI status depends on an average burden of mutation in the samples. Moreover, in most approaches, the exclusion of microsatellite loci with insufficient coverage might also improve the MSI calling accuracy.

The polymerase slippage may create artificial instability (stutter artifact) in microsatellite sequences introducing noise in length distributions mostly on long microsatellites (15 and more repetitions) where numerous alleles are present at different frequencies for a same microsatellite. Several methods, including mSINGs [79] and MSIsensor [69], thereby compared the differences between the microsatellite allele distributions based on the read counts of tumoral and normal samples that include the stutter artifact and the genuine microsatellite allele using different statistical tests to detect unstable microsatellites. The presence of these stutter artifacts could also hide some small microsatellite instabilities (-1 or -2 deletions) that are present at low frequency in some samples with a high level of normal DNA such as in plasma or blood samples of cancer patients. To address this difficulty and improve the limit of detection of MSI, the use of molecular barcodes (UMI) in NGS experiments has been developed for microsatellites and allowed after a bioinformatics correction to identify true MSI events present at low frequency [28, 31, 88].

In addition, some microsatellites present a naturally high level of polymorphism, notably in black African populations, and this naturally high polymorphism could sometimes be confused with MSI and result in false positives [53, 55]. This bias can notably be reduced by the comparison of normal and tumoral samples of a same individual as performed by MSIsensor [69] and MANTIS [50], which will thereby filter out the natural polymorphism and possibly reduce reads errors due to the polymerase slippage. When matched normal samples were not available,

some approaches such as mSINGS [79] and MSI-ColonCore [97] allowed the calibration of instability baseline across a set of samples sequenced using the same NGS protocol.

The following parts are a catalog of the different NGS-based computational approaches developed to date for the detection of MSI in cancer that are also summarized in Table 5.2. Although they all presented good performances, a systematic evaluation will be required to delineate their platform or software dependencies. The availability and requirements of each method are also indicated in Table 5.3.

Methods Using the Percentage of Unstable Microsatellite Loci Compared to Normal Samples

These methods rely on the comparison of the percentage of microsatellite instability of a defined set of microsatellites (mainly mono-nucleotide repeat microsatellites, Table 5.1) between a tumoral sample and a paired normal sample or a set of normal samples, using the same NGS protocol. They all require BAM alignment files as the main input and the corresponding genome reference sequence. To determine the MSI status of each microsatellite locus of a tumor sample, its length distribution giving the frequency (read count) of each possible allele is compared to that of its paired normal or normal baseline/reference samples using different statistical tests that distinguish the different methods (Table 5.2). An empiric threshold of microsatellite instability is then used to classify the samples into MSI or MSS tumor samples. A first study on endometrial and colorectal cancer proposed a bioinformatics approach for MSI detection using WGS and WES data from TCGA, where unstable mono- to tetra-nucleotide repeat microsatellites were detected in tumor samples by comparison of their allele length distribution to those of matched normal samples using the Kolmogorov–Smirnov statistical test [52]. However, the authors did not propose a classifier for the MSI/MSS status in this study.

MSIsensor

MSIsensor is the first publicly available tool performing this kind of analysis on paired tumoral and non-tumoral samples. The first step is to scan the reference genome to find all the microsatellites, by default mono- to penta-nucleotide repeat microsatellite of at least five repetitions. MSIsensor gives a per microsatellite MSI status based on the chi square statistic, resulting in a MSI score where a threshold over 3.5% indicates the presence of MSI phenotype [69]. MSIsensor can be run on WES and WGS data and is thus a very powerful tool for discovery studies as it gives a complete landscape of microsatellite instability across genome or exome [3, 49, 64].

mSINGS

mSINGS first creates a baseline reference panel of mono-nucleotide repeat microsatellites describing basic instability from a set of normal samples [79]. Then, tumoral samples can be individually compared to this reference in order to determine their MSI status using a Z-score approach and a threshold of 20% of unstable markers to define MSI in tumors. This software is intended to be run on targeted sequencing assays, but it is also compatible with WES data [79]. It has been used for MSI detection in many studies using different panels of microsatellites including pan-cancer panels. mSING has also been integrated into another MSI/MSS tumor classifier called MOSAIC [38] and in MSIplus [40].

MANTIS

In MANTIS, Kautto et al. use a set of mono- to penta-nucleotide repeat microsatellites from WES data to detect MSI [50]. Their approach individually computes and aggregates the differences between the allele length distribution of every locus of matched normal and tumor samples to obtain an average distance score varying from 0 (fully stable) to 2 (fully unstable). A score threshold of 0.4 is recommended by the authors to consider that the sample presents the MSI phenotype. In MANTIS, smaller custom panels can also be designed that could be particularly useful for rapid diagnosis purposes. When compared to

Table 5.2 List of computational methods for MSI detection from NGS data

Study	Method name	Method type	Detection of indels	Target microsatellites		MSI/MSS status per locus		MSI/MSS status per sample
				Type	Tool used for microsatellite identification	Comparison of MS allele length distribution	Statistic used	
Niu et al. [69]	MSIsensor	% of unstable loci classifier	MSIsensor	Mono- to penta-nucleotide repeat	MSIsensor	Tumor vs. paired normal sample	$\chi 2$	Binary (MSI/MSS) classifier based on a threshold of 3.5% of instable microsatellites for MSI
Salipante et al. [79]	mSINGs	% of unstable loci classifier	VarScan2	Mono-nucleotide repeat	mSINGS (Perl script)	Tumor vs. baseline normal samples	Z-score [mean number of allele + (3 × SD)]	Binary (MSI/MSS) classifier based on a threshold of 20% of instable microsatellites for MSI
Kautto et al. [50]	MANTIS	% of unstable loci classifier	MANTIS	Mono- to penta-nucleotide repeat	RepeatFinder	Tumor vs. paired normal sample	Average distance	Binary (MSI/MSS) classifier based on a threshold of 0.4 based on the average aggregate MSI score for MSI
Zu et al. [97]	MSI-ColonCore	% of unstable loci classifier	As MSIsensor	Mono-nucleotide repeat	As MSIsensor	Tumor vs. baseline normal samples	Z-score [mean − (3 × SD)]	Ternary (MSI-H/MSI-L/MSS) classifier based on a threshold of 40% of instable microsatellites for MSI
Hempelmann et al. [40]	MSIplus using mSINGs	% of unstable loci classifier	VarScan2	–	–	mSINGs	mSINGs	MSI status was determined using the mSINGS package
Hirostu et al. [41]	MSIcall	% of unstable loci classifier	–	Mono-nucleotide repeat	Targeted seq	Tumor vs. baseline normal samples	Allele length distance	MSI if score ≥40 with score = sum(marker score). Marker score relies on the distance between the alleles in the tumor vs the control.
Jia et al. [48]	MSIsensor-pro	% of unstable loci classifier	MSIsensor	Mono- to penta-nucleotide repeat	MSIsensor	Tumor vs. paired normal sample	Kolmogorov–Smirnov	Score = %MSI markers in sample with instability for each marker derived from probability p of deletion using a multinomial distribution

(continued)

Table 5.2 (continued)

Study	Method name	Method type	Detection of indels	Target microsatellites		MSI/MSS status per locus		MSI/MSS status per sample
				Type	Tool used for microsatellite identification	Comparison of MS allele length distribution	Statistic used	
Lu et al. [62]	MSI-seq Index	MSI burden classifier	Dindel	Mono- to hexa-nucleotide repeat	Tandem Repeats Finder	–	–	Binary MSI/MSS classifier based a threshold of PI (proportion of insertions in microsatellite over all insertions)/PD (proportion of deletions in microsatellite over all insertions) ratio lower than 0.9 for MSI
Nowak et al. [70]	Unnamed	MSI burden classifier	Indelocator	Mono-nucleotide repeat	Indelocator	–	–	Binary MSI/MSS classifier based a threshold of 40 total mutations per Mb or 5 indels per Mb in microsatellites for MSI
Fujimoto et al. [27]	Unnamed	MSI burden classifier	MIMcall	Mono- to hepta-nucleotide repeat	MsDetector, Tandem Repeats Finder, and MISA	Tumor vs. paired normal sample	Binomial distribution	MSI if %MSI markers in sample ≥3%. A MS is instable for a sample if MIMcall, a somatic indel caller relying on a binomial distribution using microsatellite error rates, detects a somatic indel for that sample on that marker
Cortes-Ciriano et al. [18]	Unnamed	Complex classifier	Sputnik	Mono- to tetra-nucleotide repeat	Sputnik	Tumor vs. paired normal sample	Kolmogorov–Smirnov	Binary (MSI/MSS) classifier based on machine learning random forest model
Huang et al. [44]	MSIseq/ NGS classifier	Complex classifier	GATK	Mono- to tetra-nucleotide repeat	MSIseq	–	–	Binary MSI/MSS classifier based on machine learning decision tree and using the number of indels in microsatellites per Mb
Hause et al. [38]	mSINGs + MOSAIC	Complex classifier	MISA	Mono- to penta-nucleotide repeat	MISA	mSINGs	mSINGs	Binary MSI/MSS classifier using parsimonious, weighted-tree microsatellite instability classifier (MOSAIC)
[89]	MSIpred	Complex classifier	–	–	–	–	–	Binary MSI/MSS SVM classifier with a radial basis function kernel using 22 variant count features

Study	Method name	Method type	Detection of indels	Target microsatellites		MSI/MSS status per locus		MSI/MSS status per sample
				Type	Tool used for microsatellite identification	Comparison of MS allele length distribution	Statistic used	
Foltz et al. [26]	MIRMMR	Complex classifier	–	–	–	–	–	MSI status given as floating value between 0 and 1 using a penalized logistic regression model. Cutoff score of 0.1922.
Redford et al. [78]	Unnamed	Complex classifier	Tandem Repeat Annotator	Mono-nucleotide repeat	Tandem Repeat Annotator	Tumor vs. baseline normal samples	Bayesian model	Score = log10(r) with $r = P(MSI\|O) / P(MSS\|O)$ and O is the observed deletion frequency and allelic bias. r is calculated using a Bayesian model
Gallon et al. [28]	Unnamed/ CMMRD PBL	Complex classifier	–	Mono-nucleotide repeat	–	Tumor vs. baseline normal samples	Beta distributions + Fisher's method	Score = −log10(probability p not belonging to control group). p calculated using Fisher's method by combining probabilities of each of the 24 markers that sample if not from control group. Single marker probability is derived using beta distribution
Gallon et al. [29]	Unnamed/ Tumor	Complex classifier	–	Mono-nucleotide repeat	–	Tumor vs. baseline normal samples	Bayesian model	Same as Redford et al. [78]

Table 5.3 Software availability and accessibility

Study	Method name	Programming language	Input files	Read alignment	Availability	Accessibility	Last version	Last update
Niu et al. [69]	MSIsensor	C++/R	Raw NGS file (BAM)	NI	Online	https://github.com/ding-lab/msisensor	v0.5	19-Sep-18
Salipante et al. [79]	mSINGs	Python	Raw NGS file (BAM)	BWA	Online	https://bitbucket.org/uwlabmed/msings	v3.4	10-Aug-18
Kautto et al. [50]	MANTIS	Python	Raw NGS file (BAM)	Any sequence aligner	Online	https://github.com/OSU-SRLab/MANTIS	v1.0.4	19-Jun-18
Zu et al. [97]	MSI-ColonCore	NA	Raw NGS file (BAM)	BWA	–	No download link	–	–
Hempelmann et al. [40]	MSIplus using mSINGs	Python	–	BWA	See mSINGs	See mSINGs	–	–
Hirostu et al. [41]	MSIcall	–	Raw NGS file (BAM)	Torrent Suite	Available from the authors upon request	Available from the authors upon request	–	–
Jia et al. [48]	MSIsensor-pro	C++	Raw NGS file (BAM)	NI	Online	https://github.com/xjtu-omics/msisensor-pro	v1.0.a	23-Apr-20
Lu et al. [62]	MSI-seq Index	R	Raw NGS file (BAM)	BWA	–	No download link	–	–
Nowak et al. [70]	Unnamed		Raw NGS file (BAM)	BWA	–	No download link	–	–
Fujimoto et al. [27]	Unnamed	Perl	Raw NGS file (BAM) wGS data	BWA	Online	https://github.com/afujimoto/MIMcall	–	30-Aug-18
Cortes-Ciriano et al. [18]	Unnamed	NA	Raw NGS file (BAM)	BWA	Available from the authors upon request	Available from the authors upon request	–	–
Huang et al. [44]	MSIseq/NGS classifier	R	List of somatic mutations	–	Online	https://CRAN.R-project.org/package=MSIseq	v1.0.0	15-Jun-15
Hause et al. [38]	mSINGs + MOSAIC	Python	Raw NGS file (BAM)	BWA	Online	https://github.com/ronaldhause/mosaic	–	12-Dec-17
[89]	MSIpred	Python	MAF file	–	Online	https://github.com/wangc29/MSIpred	–	17-Aug-18

Study	Method name	Programming language	Input files	Read alignment	Availability	Accessibility	Last version	Last update
Foltz et al. [26]	MIRMMR	R	Point mutation rate, methylation & MMR genes CADD scores files	–	Online	https://github.com/ding-lab/MIRMMR	–	31-May-17
Redford et al. [78]	Unnamed	R	Raw NGS file (BAM)	BWA	Supplementary file	https://doi.org/10.1371/journal.pone.0203052.s004	–	–
Gallon et al. [28]	Unnamed	–	Same as Redford et al. [78]	BWA	R scripts available upon request	R scripts available upon request	–	–
Gallon et al. [29]	Unnamed	–	Same as Redford et al. [78]	BWA	Supplementary file	https://doi.org/10.1371/journal.pone.0203052.s004	–	–

MSIsensor and mSINGs, MANTIS generally showed a higher overall sensitivity and specificity [50]. However, Kautto et al. also showed that the number of microsatellite loci used for MSI detection (from 10 to more than 2000) differentially influenced the accuracy of MSI detection depending on the method (mSINGs, MSIsenor, or MANTIS) used [50], thereby demonstrating the extreme importance of the panel of microsatellites selected. MANTIS has been used for the detection of MSI in 39 cancer types from TCGA [13].

MSI-ColonCore

MSI-ColonCore was developed as a rapid MSI diagnosis tool for routine clinical testing using a read-count–distribution-based method based on 22 mono-nucleotide repeat microsatellites of the MSI-ColonCore panel. The allele length distribution of the microsatellites from tumor samples are compared to a baseline composed of normal MSS reference samples predicting the MSI status of each microsatellite using a Z-score approach where a microsatellite locus was considered as unstable with a coverage ratio (ratio of the read count of the reference length over all other possible lengths for a microsatellite) lower than threshold (mean minus 3 standard deviation) of the baseline reference ratio [97]. MSI-ColonCore could thus classify each tumor sample into three possible phenotypes: MSI/MSI-H, MSI-L, and MSS if more than 40%, between 15% and 40% and less than 15% of the microsatellites were unstable, respectively [97]. The authors found that MSI-ColonCore presented an overall better accuracy than MSIsensor and mSINGs for MSI detection [97].

MSIsensor-pro

MSIsensor-pro is an updated version of MSIsensor, using a multinomial distribution model to quantify polymerase slippages for each tumor sample and a discriminative site selection method to enable MSI detection without matched normal samples [48]. MSIsensor-pro first scans the genome to identify all the microsatellites, then models an event at each base of a repeat sequence using a multinoulli distribution with 3 states: a deletion with a probability p, an insertion with a probability q, and the normal state with probability 1-p-q [48]. Using 1532 TCGA samples, the parameters p and q are calculated for each microsatellite, and all the microsatellites can subsequently be sorted according to their respective AUC. MSIsensor-pro then selects a panel composed of the most discriminative microsatellites thus removing the need for a matched normal sample. MSIsensor-pro finally defines the percentage of unstable microsatellites among the covered microsatellites within the panel as the score which can then be used to discriminate MSI-H from MSS samples. The authors compared their method on the same TCGA samples with Mantis, MSIsensor, and mSINGS and found MSIsensor-pro outperformed these methods in terms of AUC and used computing resources (peak RAM and run time) while requiring only the tumor sample [48].

MSIcall

MSIcall is based on a targeted sequencing panel of 76 mono- to tri-nucleotide repeat microsatellites used to calculate a MSI score corresponding to the weighted normalized sum of the marker scores [41]. The marker score is based on the comparison of the distances of the mean homopolymer signal of a tumor sample with a control sample (a CEPH DNA in the original study). MSIcall does not require paired normal and tumoral samples and predicted the MSI status in 25 cancer types with an overall accuracy higher than 98% when using a MSI score threshold of 40 [41].

Methods Directly Based on Mutation Burden

These methods incorporated classifiers that can determine the MSI status of a tumor sample directly from the mutation burden observed in all sequences and/or the burden of indels in microsatellites from TGS, WES, or RNAseq data.

MSI-seq Index

MSI-seq Index is an approach for MSI detection that is based on RNA sequencing data. It uses the ratio of two measures called PI and PD, corresponding to the proportion of insertions and deletions in mono- to hexa-nucleotide repeat

microsatellites among all insertions and deletions found in RefSeq RNA transcripts, respectively [62]. The authors proposed a PI/PD ratio threshold of 0.9 to distinguish between MSS and MSI tumor samples without the use of matched normal samples [62].

Nowak Method

Nowak et al. developed an approach for MSI tumor classification based on targeted sequencing data from 275 genes implicated in cancer. They proposed a total mutation burden threshold of 40 per Mb and a threshold of indels in mononucleotide microsatellites of 5 per Mb from which a tumor sample could be considered as MSI-H, with 100% concordance with MSI-PCR [70]. They also showed that their approach presented some false-positive MSI-H tumor calls for MSI-H tumors, which were attributed to *POLE* mutated tumors [70].

Fujimoto Method

Fujimoto et al. estimated microsatellite error rates for the various microsatellite patterns on chromosome X only as it is hemizygotic in males using WGS data [27]. They considered, at each nucleotide base, the most frequently occurring base to be correct and other calls to be errors. Using these estimated error rates in a binomial distribution, Fujimoto et al. implemented MIMcall a somatic indel caller and applied their algorithm on ~3000 paired bams from ICGC and TCGA to detect microsatellites showing instability among a list of microsatellites generated using MsDetector, Tandem Repeats Finder, and MISA [27]. They selected about 200,000 microsatellites showing instability in at least 2–3 samples, and they considered a sample to be MSI if the proportion of microsatellites showing instability among the selected microsatellites was ≥3%. Their approach allowed the identification of MSI across 21 different types of cancers [27].

Complex Methods Based on Machine Learning Approaches

The methods presented in this section are based on different machine learning approaches for MSI detection. The various MSI classifiers first vary according to the underlying model chosen in each paper. MSIpred relies on a SVM model, while MIRMMR used a penalized logistic regression. In Cortes-Ciriano et al., they used random forests [18], whereas decision trees were used in MSIseq/NGS classifier and in Hause et al. (C4.5 algorithm for MSIseq/NGS classifier and recursive partitioning trees in Hause et al. [38]). Finally a score relying on a Bayesian model was used in Redford et al. and Gallon et al. [29, 78], while a score was built from beta distributions in Gallon et al. [28]. The Hause et al. and Cortes-Ciriano et al. approaches are quite similar as they both start by identifying microsatellites using a reference sequence, they then proceed by extracting features from these markers and finally these features are fed, respectively, to random forests and recursive partitioning trees in Cortes-Ciriano et al. and Hause et al., respectively. The second main difference between all these models is the input features they used. Some of them use only very specific features. Foltz et al. trained and validated MIRMMR on CADD scores and methylation values from genes involved in the MMR pathway only [26]. In Redford et al., the read distribution of only 17 microsatellites (and their flanking SNPs) were used in their final Bayesian model [78], while Gallon et al. used the same model and showed comparable performances with a mere 6 microsatellites. In Gallon et al. [29], they built a score using the inferred beta distributions of 24 microsatellites. In all the remaining approaches, pan-genomic features were incorporated into their model. MSIseq and MSIpred both use global variant type counts, but MSIpred also uses the same features by genomic location and effect type [89]. Cortes-Ciriano et al. used the instability status of all the covered microsatellites [18]. Hause et al. used only two features in their final model: the instability status of one locus within *DEFB105A/B* and the average gain in unique alleles in tumors relative to matched normal tissue across all interrogated microsatellites [38].

MSIpred

MSIpred uses 22 features extracted from the variation information contained in a MAF file

produced using paired whole exome data [89]. These features include the global SNP and indel count per Mb, the same counts within simple repeat regions (UCSC "Simple Repeats" track) and various counts based on the variant location (splicing sites, UTR, flanking sites) and effect (silent, missense and nonsense). Wang et al. then trained an SVM classifier with a radial basis function kernel on ~1000 TCGA samples with a known MSI status [89]. The gamma and C parameters from the kernel were chosen using a grid search approach in order to maximize the average accuracy during a tenfold cross validation. MSIpred showed an overall accuracy of 98.3% compared to MSI-PCR on 358 TCGA tumors [89]. Wang et al. also demonstrated MSIpred had a higher sensitivity and accuracy compared to MSIseq [44, 89].

The MSIseq/NGS Classifier

The MSIseq/NGS classifier allows for direct MSI assessment using WES somatic mutation data (small nucleotide substitutions and indels). The method was developed from four machine learning frameworks, including random forest, logistic regression, naïve Bayes, and decision tree [44]. This approach uses the rate and ratio of the small nucleotide substitutions in all sequence types as well as the indels found in mono- to tetra-nucleotide repeat microsatellites to classify the tumor samples into "MSI" and "non-MSI" phenotypes. The authors showed that the MSIseq/NGS classifier gave the best results with the decision tree classifier [44].

MIRMMR (Microsatellite Instability Regression Using Methylation and Mutations in R)

In the MIRMMR paper, Foltz et al. built a model relying on the gene CADD scores calculated from variants in 35 genes involved in the MMR pathway, the methylation values in these genes, and the point mutation rate as features using TCGA data [26]. These features were fed into a penalized logistic regression with a tenfold cross validation. This function was run 1000 times, and the best performing lambda (the lambda which minimizes the mean cross-validation error) was chosen for the final model. With a cutoff score of 0.1922, MIRMMR presented a maximized sum of sensitivity and specificity for MSI detection, with performances similar to those obtained with mSINGs and MSIsensor [26].

Cortes-Ciriano Method

Cortes-Ciriano et al. proposed a model based on random forests to distinguish MSI-H and MSS samples using paired whole exome data from TCGA [18]. First, they scanned a transcriptome reference sequence to identify microsatellites using Sputnik. For each tumor and for each of these microsatellites with a minimum coverage of 5 reads for the tumor and normal samples, a Kolmogorow–Smirnov test was applied to the distribution of read lengths extracted from the normal and tumor bams in order to detect individual specific sites of instability (FDR <0.05) [18]. Second, for each tumor, a vector was constructed with the total number of unstable sites and binary values indicating whether the microsatellite was unstable for the patient for all the microsatellites showing instability in at least one sample. Random forests along with conformal prediction were used to build a binary MSI status predictor. This approach has been used to detect MSI from WES and WGS data of TCGA in a pan-cancer study including 23 types of cancer [18].

Hause Method (mSING + 1 Locus) MOSAIC

Hause et al. implemented a pipeline similar to the one from Cortes-Ciriano et al. Indeed, they first established a list of microsatellites to investigate using MISA instead of Sputnik in Cortes-Ciriano method. From this list, they also extracted a list of unstable microsatellites using mSINGS [79], a simpler approach which considered a microsatellite to be unstable if there was at least one additional length in the read length distribution of the tumor sample compared to the normal [38]. The main difference between the approach described by Cortes-Ciriano et al. and Hause et al. lies in the model used for the classifier itself and the features that were fed to their respective model. Indeed, Cortes-Ciriano et al. used all the microsatellites showing instability in at least one sam-

ple as features [18]. Hause et al., on the other hand, extracted summary features from their MSI calls. In their final model called MOSAIC, Hause et al. used the average gain in unique alleles in tumor relative to matched normal tissue across all interrogated microsatellites (peak_avg) and the instability status of the most discriminating microsatellite, a locus within DEFB105A/B, chr. 8:7679723–7679741 on hg19, between MSI-H and MSS tumors using Fisher's exact tests [38]. These features were fed into recursive partitioning trees using a leave-one-sample-out cross-validation strategy to optimize the parameters. MOSAIC has notably allowed the detection of MSI from TCGA data in 14 out of 18 cancer types [38].

Redford and Gallon Methods

These methods were developed from a small number of microsatellite loci, with the same scoring approach [29, 78]. Redford et al. investigated mononucleotide repeat microsatellites with sizes 7–12 bp as they are less subject to experimental biases [78]. They used TGCA CRC WGS alignments merged into a control and MSI-H group to identify 120 microsatellites showing instability in the MSI-H group compared to the control group and with a flanking SNP in dbsnp within 30 bp of the repeat. These markers were typed in a discovery cohort on the Illumina Miseq. Using the percentage of reads presenting a deletion as a threshold to classify MSI-H and MSS samples, Redford et al. were able to draw a ROC curve and calculate the associated AUC for each marker [78]. These AUCs along with the amplicon length were used to select a final list of 17 markers. A Bayesian model relying on the deletion frequency and allelic bias for these markers was trained using the discovery cohort to produce a MSI score which could in turn be used to classify a sample as MSI-H or MSS. Redford et al. showed their model had the same classification performance compared to fragment analysis on ~200 CRC tumor samples [78]. In a later publication, Gallon et al. [29] reused the same panel of 24 markers from Gallon et al. [28], a subset of the 120 markers mentioned earlier, and extracted 6 markers with a backward–forward stepwise selection showing the same accuracy than the 24-marker panel using ~100 CRC samples for training and ~200 CRC samples for testing from 3 independent cohorts.

Gallon Method 2

Gallon et al. selected 24 markers from the 120 microsatellites described by Redford et al. and built a model to predict whether a sample has CMMRD using non-neoplastic tissues, which are characterized by low-level microsatellite instability [28]. For each of these markers, they fit the distribution of percentage of wild type reads (WTP) on that microsatellite over all the control samples to a beta distribution. Given a new sample with an observed percentage of wild type reads p_i for marker i, the probability to observe WTP $\geq p_i$ can be calculated. These probabilities were then combined using Fisher's method to produce a probability that the sample is from the control group [28]. Their model was initially built on 40 control samples and 5 CMMRD patients. On a second cohort of 27 CMMRD patients and 54 controls, their model achieved 100% sensitivity and 98% specificity across all samples [28]. By fine-tuning parameters to make the model more conservative, they reached 97% sensitivity and 100% specificity misclassifying one CMMRD patient [28].

Examples of Clinical Applications Using MSI Detection by NGS

Implications of the MSI Phenotype for Cancer Diagnosis, Prognosis, Prediction of Treatment Response and Therapeutic Decision-Making

The determination of MSI status has many clinical implications for cancer patients. MSI testing by MSI-PCR is recommended for the diagnosis of two inherited cancer-predisposing syndromes known as Lynch and CMMRD syndromes, alongside dMMR testing with IHC. Approximatively 3% of colorectal tumors (and 2% of endometrial tumors) arise in the context of Lynch syndrome where a constitutional mutation

of a MMR gene leads to an increased risk of cancer incidence (10% at 50 years and 40% at 70 years) and requires a specific management of the affected patients as well as members of their family [2, 5, 32, 63]. In 2015, the European Society for Medical Oncology therefore recommended that every CRC patient should be tested for MSI at the time of diagnosis as a first screen for Lynch syndrome [82]. Moreover, a more severe syndrome known as CMMRD caused by bi-allelic germline mutations of one of the four MMR genes is characterized by the appearance of colorectal cancer in childhood and also requires a specific management similar to Lynch syndrome [92, 93].

Concerning the predictive and prognosis value of MSI phenotype, it was shown in stage II/III CRC that MSI was associated with a better prognosis and was also predictive of the response to different chemotherapy combinations [51, 96]. Thus, adjuvant 5'fluorouracil chemotherapy presented no benefits for stage II CRC patients, while it showed improved response in combination with oxaliplatin in stage III CRC patients [51, 96]. MSI status might thus be used to guide the choice of a tailored treatment for stage II and III CRC patients. Moreover, in metastatic CRC as well as in other metastatic solid cancers patients, MSI was shown to be a major predictor of response to immune blockade therapy [59, 60, 71, 72]. Thus, the Food and Drug Administration (FDA) approved in 2017 the administration of immune checkpoint inhibitors (nivolumab and pembrolizumab) for the treatment of every solid cancer with this genetic feature, regardless of their tumor type [14, 65].

Assessment of MSI Status by NGS in Tumors

The gold standard method for MSI testing in the clinic is based on MSI-PCR using either NCI/Betesda or pentaplex panel in FFPE tumor samples [86]. As an emerging new technology for MSI detection in cancer, NGS has been evaluated in tumor samples for routine clinical testing and compared to MSI-PCR in several studies [94]. However, MSI-PCR was used in most of these studies as the reference method, and it was therefore not possible to know which method actually reflected a genuine MMR deficiency when discrepancies were observed between MSI-PCR and NGS (Table 5.1), except when a third method such as dMMR IHC or full sequencing of MMR genes was included. We thereby focused on these latter studies allowing the comparison of performances of MSI detection by NGS and MSI-PCR.

A first study using the MSIplus panel and the mSING algorithm showed that NGS presented slightly better performances (97% sensitivity and 100% specificity) compared to MSI-PCR (97% sensitivity and 95% specificity) on a cohort of 78 FFPE CRC samples without the need of matched normal samples [40]. The MSI-ColonCore panel was used to assess MSI in 91 FFPE CRC samples and presented slightly lower concordance rates (92.3%) with IHC than MSI-PCR (93.4%) with IHC [97]. In prostate cancer, MSI plus, BROCA, and UW-OncoPlex panels combined with the mSING algorithm have been used to detect MSI in a set of 71 FFPE and 20 fresh frozen prostate tumors [39]. MSIplus and the larger panels presented 96.6% and 93.1% sensitivity and 100% and 98.4% specificity, respectively, while MSI-PCR had 72.4% sensitivity and 100% specificity [39]. Using the pan-cancer panel of 111 microsatellites and mSING, Waalkes et al. evaluated the detection of MSI in three types of cancer and compared the performances of their method to MSI-PCR. Their approach showed equal performances for colorectal cancer (100% sensitivity and specificity) and improved performances for prostate (100% sensitivity and specificity vs. 81.8% sensitivity and 100% specificity) and endometrial (95.8% sensitivity and 100% specificity vs. 75.0% sensitivity and 100% specificity) cancer compared to MSI-PCR using the pentaplex panel [88].

The recommended algorithms for the diagnosis of Lynch syndrome are complex, time-consuming, and involve multiple sequential steps and different techniques including MSI-PCR, dMMR IHC, and/or Sanger sequencing of *BRAF* and/or MMR genes [32]. With the advent of NGS, the direct next-generation sequencing of the MMR genes in CRC tumors has been proposed as a replacement of the standard multi-test

approaches for Lynch screening and presented improved sensitivity and simplified steps [36]. Moreover, a study based on the detection of MSI by NGS using a pan-cancer microsatellite panel from MSK-IMPACT assay and MSI-sensor algorithm aimed to determine the prevalence of Lynch syndrome across multiple types of cancers (more than 15,000 tumors and 50 cancer types) [58]. Their results showed that the detection of MSI using NGS was predictive of Lynch syndrome in all types of tumors, half of the cases concerned tumor types not previously or rarely being associated with this syndrome and a little less than half not meeting the clinical criteria (familial and personal cancer history) for Lynch syndrome genetic testing [58]. These results suggested that the assessment of MSI by NGS could thereby be used for Lynch syndrome screening prior to germline screening [58].

These studies illustrated the clinical potential of MSI detection using NGS for MSI testing in tumors and for Lynch syndrome diagnosis. MSI detection using NGS approaches generally presented better performances (sensitivity and specificity) compared to MSI-PCR, notably in non-CRC tumors and could potentially replace the latter method. The main advantage of NGS is that it can combine different types of genetic analyses in a same gene panel-based experiment including MSI testing, tumor mutation burden (TMB) assessment, single nucleotide variations, fusions, and/or copy number alterations in cancer-related genes. The analysis of these genetic alterations could thereby guide the therapeutic decision-making and management of the patients. For example, Vanderwalde et al. combined, in a recent study based on NGS of a panel of 592 genes in thousands of FFPE tumor samples from 26 cancer types, the detection of MSI and the evaluation of TMB, which is another predictive biomarker of the response to immune checkpoint inhibitors whose status can only be assessed to date by NGS [87]. As MSI and high TMB were only partially overlapping in tumor samples, they could potentially be used together to identify more patients that may benefit from these treatments than when each marker is used alone [87]. Pan-cancer panels of genes allowing the analysis of multiple types of genetic alterations in cancer by NGS are therefore of great interest for precision oncology and need to be developed.

Assessment of MSI Status by NGS in Blood and Plasma Samples

Besides the development of novel NGS-based approaches for the detection of MSI in tumor samples, some studies focused on the capacity to detect MSI in non-tumoral samples, notably blood and plasma samples. The benefits for the patients could be the early and non-invasive diagnosis of MSI in patients with suspected MSI/dMMR cancer or Lynch and/or CMMRD syndromes and the prediction and/or monitoring of treatment response.

In the context of CMMRD syndrome diagnosis, several approaches have been proposed for the detection of MSI from whole blood DNA. gMSI is based on the analysis of stutter peaks following MSI-PCR on di-nucleotide repeat microsatellites using a freely available software but is unable to detect MSI due to MSH6 deficiency [45]. A second approach named evMSI proposed to detect MSI in lymphoblastoid cell lines derived from the blood of CMMRD patients using MSI-PCR and presented 100% sensitivity and specificity but required 120 days of in vitro culture after immortalization [10]. Recently, Gallon et al. proposed a simple method for CMMRD diagnosis based on a panel of 24 mono-nucleotide repeat microsatellites combined with NGS library preparation that included unique molecular identifiers (UMI) for slippage error correction and using a bioinformatic algorithm for MSI scoring [28]. The use of UMI allowed PCR and sequencing slippage error correction and to identify genuine MSI events in the panel, while the MSI score allowed the correct detection of MSI in all CMMRD and suspected-CMMRD samples including those with MSH6 deficiency and thus outperforming the gMSI method [28]. This simple method may lead to use in clinics for rapid CMMRD diagnosis and screening of at-risk populations.

Serum and plasma have been extensively studied for decades and used for cell-free nucleid

acid-based biormarkers analysis in cancer patients which presents great interest for numerous non-invasive clinical applications such as early diagnosis and the prediction and/or monitoring of treatment response [61, 80]. Very few studies have evaluated MSI in plasma samples of MSI cancer patients, although the presence of MSI was demonstrated by MSI-PCR in cell-free DNA of serum from head and neck cancer patients, since 1996 [68]. In two recent and simulteneous studies, MSI has been assessed using NGS in cell-free DNA of plasma from cancer patients including some receiving immune checkpoint inhibitors [90]. Willis et al. used a plasma-based cancer genotyping assay (Guardant360 assay) and NGS digital sequencing for slippage error correction to detect MSI in more than 28,000 cfDNA samples. Their approach showed a LOD of MSI of 0.1% and a high concondance between MSI statuts of paired plasma and tumor, where the overall accuracy for MSI detection in plasma was evaluated at 98.4% when compared to the status of matched tumor tissues [91]. Alternatively Georgiadis et al. also developed an approach for the detection of MSI and TMB-H in plasma using a 58 gene panel, where UMI barcoding and a digital peak finding algorithm allowed slippage error correction and the accurate detection of MSI events with a LOD of 1% [31]. Their approach presented 78% and 67% sensitivities and >99% specificity for MSI detection and TMB-H, respectively. It also showed that the presence of MSI in pretreatment plasma can predict progression-free survival, whereas the disapearance of MSI in post-treatment plasma is associated with progression-free and overall surival [31].

Conclusion

Since its discovery almost 30 years ago, MSI was mainly studied and used as a genetic biomarker in CRC and Lynch syndrome. Recently, whole genome and whole exome studies showed the presence of MSI in dozens of cancer types correlating to patient survival outcome in a positive dose-effect manner. Moreover, MSI was also identified as a major predicator of the response of immune checkpoint blockade therapy in solid cancers, which marked a renewed interest in the study of this genetic alteration in cancer. Consequently, in 2017, the FDA approved the use of MSI status for the administration of immune checkpoint inhibitors in advanced solid cancers, regardless of the type of cancer. Therefore, there was a need for the development of new tools for pan-cancer MSI detection as the gold standard methods based on PCR and capillary electrophoresis generally presented poor performances for non-colorectal cancers. Although more complex, NGS could be a powerful method for the detection and analysis of MSI in cancer compared to MSI-PCR, as it could drastically increase the number of interrogated microsatellite loci for sensitive pan-cancer MSI detection and combine in a same experiment the analyses of several other types of genetic alteration (single nucleotide variations, fusions, copy number alterations, TMB), which are required for the stratification of patient tumors for precision oncology. Several panels of microsatellites including pan-cancer panels as well as new bioinformatics algorithms have been proposed to assess MSI in cancer samples. The computational methods generally took into account the particularities of microsatellites, including their highly polymorphic nature and polymerase slippage errors, as well as the difficulties for the analysis of microsatellite data due to the errors induced during the sequencing-by-synthesis of homopolymers by NGS, the alignement errors induced by the short read length, and the low accuracy of indel calling. As a result, these methods proposed different approaches for the determination of MSI status that sometimes showed better performances than MSI-PCR. Future perspectives should include the development of pan-cancer panels of genes and of highly sensitive and specific microsatellite markers allowing the simultaneous detection and identification of multiple genetic alterations including MSI in all types of cancer samples for precision oncology. Moreover, a global evaluation, comparison, and validation of the different microsatellite panels and computational algorithms proposed to date for MSI detection should

also be performed in the different types of cancer, prior to their implementation in routine clinical testing.

References

1. Aaltonen LA, Peltomaki P, Leach FS, Sistonen P, Pylkkanen L, Mecklin JP, Jarvinen H, Powell SM, Jen J, Hamilton SR, et al. Clues to the pathogenesis of familial colorectal cancer. Science. 1993;260(5109):812–6.
2. Aaltonen LA, Salovaara R, Kristo P, Canzian F, Hemminki A, Peltomaki P, Chadwick RB, Kaariainen H, Eskelinen M, Jarvinen H, Mecklin JP, de la Chapelle A. Incidence of hereditary nonpolyposis colorectal cancer and the feasibility of molecular screening for the disease. N Engl J Med. 1998;338(21):1481–7. https://doi.org/10.1056/NEJM199805213382101.
3. Abida W, Cheng ML, Armenia J, Middha S, Autio KA, Vargas HA, Rathkopf D, Morris MJ, Danila DC, Slovin SF, Carbone E, Barnett ES, Hullings M, Hechtman JF, Zehir A, Shia J, Jonsson P, Stadler ZK, Srinivasan P, Laudone VP, Reuter V, Wolchok JD, Socci ND, Taylor BS, Berger MF, Kantoff PW, Sawyers CL, Schultz N, Solit DB, Gopalan A, Scher HI. Analysis of the prevalence of microsatellite instability in prostate cancer and response to immune checkpoint blockade. JAMA Oncol. 2019;5(4):471–8. https://doi.org/10.1001/jamaoncol.2018.5801.
4. Bacher JW, Flanagan LA, Smalley RL, Nassif NA, Burgart LJ, Halberg RB, Megid WM, Thibodeau SN. Development of a fluorescent multiplex assay for detection of MSI-high tumors. Dis Markers. 2004;20(4–5):237–50.
5. Barnetson RA, Tenesa A, Farrington SM, Nicholl ID, Cetnarskyj R, Porteous ME, Campbell H, Dunlop MG. Identification and survival of carriers of mutations in DNA mismatch-repair genes in colon cancer. N Engl J Med. 2006;354(26):2751–63. https://doi.org/10.1056/NEJMoa053493.
6. Baudrin LG, Deleuze JF, How-Kit A. Molecular and computational methods for the detection of microsatellite instability in cancer. Front Oncol. 2018a;8:621. https://doi.org/10.3389/fonc.2018.00621.
7. Baudrin LG, Duval A, Daunay A, Buhard O, Bui H, Deleuze JF, How-Kit A. Improved microsatellite instability detection and identification by nuclease-assisted microsatellite instability enrichment using HSP110 T17. Clin Chem. 2018b; https://doi.org/10.1373/clinchem.2018.287490.
8. Bianchi F, Galizia E, Catalani R, Belvederesi L, Ferretti C, Corradini F, Cellerino R. CAT25 is a mononucleotide marker to identify HNPCC patients. J Mol Diagn. 2009;11(3):248–52. https://doi.org/10.2353/jmoldx.2009.080155.
9. Bocker T, Diermann J, Friedl W, Gebert J, Holinski-Feder E, Karner-Hanusch J, von Knebel-Doeberitz M, Koelble K, Moeslein G, Schackert HK, Wirtz HC, Fishel R, Ruschoff J. Microsatellite instability analysis: a multicenter study for reliability and quality control. Cancer Res. 1997;57(21):4739–43.
10. Bodo S, Colas C, Buhard O, Collura A, Tinat J, Lavoine N, Guilloux A, Chalastanis A, Lafitte P, Coulet F, Buisine MP, Ilencikova D, Ruiz-Ponte C, Kinzel M, Grandjouan S, Brems H, Lejeune S, Blanche H, Wang Q, Caron O, Cabaret O, Svrcek M, Vidaud D, Parfait B, Verloes A, Knappe UJ, Soubrier F, Mortemousque I, Leis A, Auclair-Perrossier J, Frebourg T, Flejou JF, Entz-Werle N, Leclerc J, Malka D, Cohen-Haguenauer O, Goldberg Y, Gerdes AM, Fedhila F, Mathieu-Dramard M, Hamelin R, Wafaa B, Gauthier-Villars M, Bourdeaut F, Sheridan E, Vasen H, Brugieres L, Wimmer K, Muleris M, Duval A. Diagnosis of constitutional mismatch repair-deficiency syndrome based on microsatellite instability and lymphocyte tolerance to methylating agents. Gastroenterology. 2015;149(4):1017–1029 e1013. https://doi.org/10.1053/j.gastro.2015.06.013.
11. Boland CR, Goel A. Microsatellite instability in colorectal cancer. Gastroenterology. 2010;138(6):2073–2087 e2073. https://doi.org/10.1053/j.gastro.2009.12.064.
12. Boland CR, Thibodeau SN, Hamilton SR, Sidransky D, Eshleman JR, Burt RW, Meltzer SJ, Rodriguez-Bigas MA, Fodde R, Ranzani GN, Srivastava S. A National Cancer Institute Workshop on Microsatellite Instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer. Cancer Res. 1998;58(22):5248–57.
13. Bonneville R, Krook MA, Kautto EA, Miya J, Wing MR, Chen HZ, Reeser JW, Yu L, Roychowdhury S. Landscape of microsatellite instability across 39 cancer types. JCO Precis Oncol. 2017, 2017; https://doi.org/10.1200/PO.17.00073.
14. Boyiadzis MM, Kirkwood JM, Marshall JL, Pritchard CC, Azad NS, Gulley JL. Significance and implications of FDA approval of pembrolizumab for biomarker-defined disease. J Immunother Cancer. 2018;6(1):35. https://doi.org/10.1186/s40425-018-0342-x.
15. Buhard O, Lagrange A, Guilloux A, Colas C, Chouchene M, Wanherdrick K, Coulet F, Guillerm E, Dorard C, Marisa L, Bokhari A, Greene M, El-Murr N, Bodo S, Muleris M, Sourouille I, Svrcek M, Cervera P, Blanche H, Lefevre JH, Parc Y, Lepage C, Chapusot C, Bouvier AM, Gaub MP, Selves J, Garrett K, Iacopetta B, Soong R, Hamelin R, Garrido C, Lascols O, Andre T, Flejou JF, Collura A, Duval A. HSP110 T17 simplifies and improves the microsatellite instability testing in patients with colorectal cancer. J Med Genet. 2016;53(6):377–84. https://doi.org/10.1136/jmedgenet-2015-103518.
16. Buhard O, Suraweera N, Lectard A, Duval A, Hamelin R. Quasimonomorphic mononucleotide repeats for high-level microsatellite instability analysis. Dis Markers. 2004;20(4–5):251–7.

17. Cawkwell L, Li D, Lewis FA, Martin I, Dixon MF, Quirke P. Microsatellite instability in colorectal cancer: improved assessment using fluorescent polymerase chain reaction. Gastroenterology. 1995;109(2):465–71.
18. Cortes-Ciriano I, Lee S, Park WY, Kim TM, Park PJ. A molecular portrait of microsatellite instability across multiple cancers. Nat Commun. 2017;8:15180. https://doi.org/10.1038/ncomms15180.
19. Daunay A, Duval A, Baudrin LG, Buhard O, Renault V, Deleuze JF, How-Kit A. Low temperature isothermal amplification of microsatellites drastically reduces stutter artifact formation and improves microsatellite instability detection in cancer. Nucleic Acids Res. 2019; https://doi.org/10.1093/nar/gkz811.
20. Dietmaier W, Wallinger S, Bocker T, Kullmann F, Fishel R, Ruschoff J. Diagnostic microsatellite instability: definition and correlation with mismatch repair protein expression. Cancer Res. 1997;57(21):4749–56.
21. Dorard C, de Thonel A, Collura A, Marisa L, Svrcek M, Lagrange A, Jego G, Wanherdrick K, Joly AL, Buhard O, Gobbo J, Penard-Lacronique V, Zouali H, Tubacher E, Kirzin S, Selves J, Milano G, Etienne-Grimaldi MC, Bengrine-Lefevre L, Louvet C, Tournigand C, Lefevre JH, Parc Y, Tiret E, Flejou JF, Gaub MP, Garrido C, Duval A. Expression of a mutant HSP110 sensitizes colorectal cancer cells to chemotherapy and improves disease prognosis. Nat Med. 2011;17(10):1283–9. https://doi.org/10.1038/nm.2457.
22. Ellegren H. Microsatellites: simple sequences with complex evolution. Nat Rev Genet. 2004;5(6):435–45. https://doi.org/10.1038/nrg1348.
23. Fang H, Wu Y, Narzisi G, O'Rawe JA, Barron LT, Rosenbaum J, Ronemus M, Iossifov I, Schatz MC, Lyon GJ. Reducing INDEL calling errors in whole genome and exome sequencing data. Genome Med. 2014;6(10):89. https://doi.org/10.1186/s13073-014-0089-z.
24. Faulkner RD, Seedhouse CH, Das-Gupta EP, Russell NH. BAT-25 and BAT-26, two mononucleotide microsatellites, are not sensitive markers of microsatellite instability in acute myeloid leukaemia. Br J Haematol. 2004;124(2):160–5.
25. Findeisen P, Kloor M, Merx S, Sutter C, Woerner SM, Dostmann N, Benner A, Dondog B, Pawlita M, Dippold W, Wagner R, Gebert J, von Knebel DM. T25 repeat in the 3′ untranslated region of the CASP2 gene: a sensitive and specific marker for microsatellite instability in colorectal cancer. Cancer Res. 2005;65(18):8072–8. https://doi.org/10.1158/0008-5472.CAN-04-4146.
26. Foltz SM, Liang WW, Xie M, Ding L. MIRMMR: binary classification of microsatellite instability using methylation and mutations. Bioinformatics. 2017;33(23):3799–801. https://doi.org/10.1093/bioinformatics/btx507.
27. Fujimoto A, Fujita M, Hasegawa T, Wong JH, Maejima K, Oku-Sasaki A, Nakano K, Shiraishi Y, Miyano S, Yamamoto G, Akagi K, Imoto S, Nakagawa H. Comprehensive analysis of indels in whole-genome microsatellite regions and microsatellite instability across 21 cancer types. Genome Res. 2020; https://doi.org/10.1101/gr.255026.119.
28. Gallon R, Muhlegger B, Wenzel SS, Sheth H, Hayes C, Aretz S, Dahan K, Foulkes W, Kratz CP, Ripperger T, Azizi AA, Baris Feldman H, Chong AL, Demirsoy U, Florkin B, Imschweiler T, Januszkiewicz-Lewandowska D, Lobitz S, Nathrath M, Pander HJ, Perez-Alonso V, Perne C, Ragab I, Rosenbaum T, Rueda D, Seidel MG, Suerink M, Taeubner J, Zimmermann SY, Zschocke J, Borthwick GM, Burn J, Jackson MS, Santibanez-Koref M, Wimmer K. A sensitive and scalable microsatellite instability assay to diagnose constitutional mismatch repair deficiency by sequencing of peripheral blood leukocytes. Hum Mutat. 2019;40(5):649–55. https://doi.org/10.1002/humu.23721.
29. Gallon R, Sheth H, Hayes C, Redford L, Alhilal G, O'Brien O, Spiewak H, Waltham A, McAnulty C, Izuogu OG, Arends MJ, Oniscu A, Alonso AM, Laguna SM, Borthwick GM, Santibanez-Koref M, Jackson MS, Burn J. Sequencing-based microsatellite instability testing using as few as six markers for high-throughput clinical diagnostics. Hum Mutat. 2020;41(1):332–41. https://doi.org/10.1002/humu.23906.
30. Gan C, Love C, Beshay V, Macrae F, Fox S, Waring P, Taylor G. Applicability of next generation sequencing technology in microsatellite instability testing. Genes. 2015;6(1):46–59. https://doi.org/10.3390/genes6010046.
31. Georgiadis A, Durham JN, Keefer LA, Bartlett BR, Zielonka M, Murphy D, White JR, Lu S, Verner EL, Ruan F, Riley D, Anders RA, Gedvilaite E, Angiuoli S, Jones S, Velculescu VE, Le DT, Diaz LA Jr, Sausen M. Noninvasive detection of microsatellite instability and high tumor mutation burden in cancer patients treated with PD-1 blockade. Clin Cancer Res. 2019;25(23):7024–34. https://doi.org/10.1158/1078-0432.CCR-19-1372.
32. Giardiello FM, Allen JI, Axilbund JE, Boland CR, Burke CA, Burt RW, Church JM, Dominitz JA, Johnson DA, Kaltenbach T, Levin TR, Lieberman DA, Robertson DJ, Syngal S, Rex DK. Guidelines on genetic evaluation and management of Lynch syndrome: a consensus statement by the US Multi-society Task Force on colorectal cancer. Am J Gastroenterol. 2014;109(8):1159–79. https://doi.org/10.1038/ajg.2014.186.
33. Goel A, Nagasaka T, Hamelin R, Boland CR. An optimized pentaplex PCR for detecting DNA mismatch repair-deficient colorectal cancers. PLoS One. 2010;5(2):e9393. https://doi.org/10.1371/journal.pone.0009393.
34. Gonzalez-Garcia I, Moreno V, Navarro M, Marti-Rague J, Marcuello E, Benasco C, Campos O, Capella G, Peinado MA. Standardized approach for microsat-

ellite instability detection in colorectal carcinomas. J Natl Cancer Inst. 2000;92(7):544–9.
35. Hampel H, Frankel W, Panescu J, Lockman J, Sotamaa K, Fix D, Comeras I, La Jeunesse J, Nakagawa H, Westman JA, Prior TW, Clendenning M, Penzone P, Lombardi J, Dunn P, Cohn DE, Copeland L, Eaton L, Fowler J, Lewandowski G, Vaccarello L, Bell J, Reid G, de la Chapelle A. Screening for Lynch syndrome (hereditary nonpolyposis colorectal cancer) among endometrial cancer patients. Cancer Res. 2006;66(15):7810–7. https://doi.org/10.1158/0008-5472.CAN-06-1114.
36. Hampel H, Pearlman R, Beightol M, Zhao W, Jones D, Frankel WL, Goodfellow PJ, Yilmaz A, Miller K, Bacher J, Jacobson A, Paskett E, Shields PG, Goldberg RM, de la Chapelle A, Shirts BH, Pritchard CC. Assessment of tumor sequencing as a replacement for lynch syndrome screening and current molecular tests for patients with colorectal cancer. JAMA Oncol. 2018;4(6):806–13. https://doi.org/10.1001/jamaoncol.2018.0104.
37. Hatch SB, Lightfoot HM Jr, Garwacki CP, Moore DT, Calvo BF, Woosley JT, Sciarrotta J, Funkhouser WK, Farber RA. Microsatellite instability testing in colorectal carcinoma: choice of markers affects sensitivity of detection of mismatch repair-deficient tumors. Clin Can Res. 2005;11(6):2180–7. https://doi.org/10.1158/1078-0432.CCR-04-0234.
38. Hause RJ, Pritchard CC, Shendure J, Salipante SJ. Classification and characterization of microsatellite instability across 18 cancer types. Nat Med. 2016;22(11):1342–50. https://doi.org/10.1038/nm.4191.
39. Hempelmann JA, Lockwood CM, Konnick EQ, Schweizer MT, Antonarakis ES, Lotan TL, Montgomery B, Nelson PS, Klemfuss N, Salipante SJ, Pritchard CC. Microsatellite instability in prostate cancer by PCR or next-generation sequencing. J Immunother Cancer. 2018;6(1):29. https://doi.org/10.1186/s40425-018-0341-y.
40. Hempelmann JA, Scroggins SM, Pritchard CC, Salipante SJ. MSIplus for integrated colorectal cancer molecular testing by next-generation sequencing. J Mol Diagn. 2015;17(6):705–14. https://doi.org/10.1016/j.jmoldx.2015.05.008.
41. Hirotsu Y, Nagakubo Y, Amemiya K, Oyama T, Mochizuki H, Omata M. Microsatellite instability status is determined by targeted sequencing with MSIcall in 25 cancer types. Clin Chim Acta. 2020;502:207–13. https://doi.org/10.1016/j.cca.2019.11.002.
42. Hoang JM, Cottu PH, Thuille B, Salmon RJ, Thomas G, Hamelin R. BAT-26, an indicator of the replication error phenotype in colorectal cancers and cell lines. Cancer Res. 1997;57(2):300–3.
43. How-Kit A, Daunay A, Buhard O, Meiller C, Sahbatou M, Collura A, Duval A, Deleuze JF. Major improvement in the detection of microsatellite instability in colorectal cancer using HSP110 T17 E-ice-COLD-PCR. Hum Mutat. 2018;39(3):441–53. https://doi.org/10.1002/humu.23379.
44. Huang MN, McPherson JR, Cutcutache I, Teh BT, Tan P, Rozen SG. MSIseq: software for assessing microsatellite instability from catalogs of somatic mutations. Sci Rep. 2015;5:13321. https://doi.org/10.1038/srep13321.
45. Ingham D, Diggle CP, Berry I, Bristow CA, Hayward BE, Rahman N, Markham AF, Sheridan EG, Bonthron DT, Carr IM. Simple detection of germline microsatellite instability for diagnosis of constitutional mismatch repair cancer syndrome. Hum Mutat. 2013;34(6):847–52. https://doi.org/10.1002/humu.22311.
46. Ionov Y, Peinado MA, Malkhosyan S, Shibata D, Perucho M. Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis. Nature. 1993;363(6429):558–61. https://doi.org/10.1038/363558a0.
47. Ivady G, Madar L, Dzsudzsak E, Koczok K, Kappelmayer J, Krulisova V, Macek M Jr, Horvath A, Balogh I. Analytical parameters and validation of homopolymer detection in a pyrosequencing-based next generation sequencing system. BMC Genomics. 2018;19(1):158. https://doi.org/10.1186/s12864-018-4544-x.
48. Jia P, Yang X, Guo L, Liu B, Lin J, Liang H, Sun J, Zhang C, Ye K. MSIsensor-pro: fast, accurate, and matched-normal-sample-free detection of microsatellite instability. Genomics Proteomics Bioinformatics. 2020; https://doi.org/10.1016/j.gpb.2020.02.001.
49. Johansen AFB, Kassentoft CG, Knudsen M, Laursen MB, Madsen AH, Iversen LH, Sunesen KG, Rasmussen MH, Andersen CL. Validation of computational determination of microsatellite status using whole exome sequencing data from colorectal cancer patients. BMC Cancer. 2019;19(1):971. https://doi.org/10.1186/s12885-019-6227-7.
50. Kautto EA, Bonneville R, Miya J, Yu L, Krook MA, Reeser JW, Roychowdhury S. Performance evaluation for rapid detection of pan-cancer microsatellite instability with MANTIS. Oncotarget. 2017;8(5):7452–63. https://doi.org/10.18632/oncotarget.13918.
51. Kawakami H, Zaanan A, Sinicrope FA. Microsatellite instability testing and its role in the management of colorectal cancer. Curr Treat Options in Oncol. 2015;16(7):30. https://doi.org/10.1007/s11864-015-0348-2.
52. Kim TM, Laird PW, Park PJ. The landscape of microsatellite instability in colorectal and endometrial cancer genomes. Cell. 2013;155(4):858–68. https://doi.org/10.1016/j.cell.2013.10.015.
53. Kinney N, Titus-Glover K, Wren JD, Varghese RT, Michalak P, Liao H, Anandakrishnan R, Pulenthiran A, Kang L, Garner HR. CAGm: a repository of germline microsatellite variations in the 1000 genomes project. Nucleic Acids Res. 2019;47(D1):D39–45. https://doi.org/10.1093/nar/gky969.
54. Ladas I, Yu F, Leong KW, Fitarelli-Kiehl M, Song C, Ashtaputre R, Kulke M, Mamon H, Makrigiorgos GM. Enhanced detection of microsatellite instability using pre-PCR elimination of wild-type DNA

homo-polymers in tissue and liquid biopsies. Nucleic Acids Res. 2018; https://doi.org/10.1093/nar/gky251.
55. Laghi L, Bianchi P, Malesci A. Differences and evolution of the methods for the assessment of microsatellite instability. Oncogene. 2008;27(49):6313–21. https://doi.org/10.1038/onc.2008.217.
56. Laghi L, Bianchi P, Roncalli M, Malesci A. Re: revised Bethesda guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability. J Natl Cancer Inst. 2004;96(18):1402–3; author reply 1403–1404. https://doi.org/10.1093/jnci/djh280.
57. Laiho P, Launonen V, Lahermo P, Esteller M, Guo M, Herman JG, Mecklin JP, Jarvinen H, Sistonen P, Kim KM, Shibata D, Houlston RS, Aaltonen LA. Low-level microsatellite instability in most colorectal carcinomas. Cancer Res. 2002;62(4):1166–70.
58. Latham A, Srinivasan P, Kemel Y, Shia J, Bandlamudi C, Mandelker D, Middha S, Hechtman J, Zehir A, Dubard-Gault M, Tran C, Stewart C, Sheehan M, Penson A, DeLair D, Yaeger R, Vijai J, Mukherjee S, Galle J, Dickson MA, Janjigian Y, O'Reilly EM, Segal N, Saltz LB, Reidy-Lagunes D, Varghese AM, Bajorin D, Carlo MI, Cadoo K, Walsh MF, Weiser M, Aguilar JG, Klimstra DS, Diaz LA Jr, Baselga J, Zhang L, Ladanyi M, Hyman DM, Solit DB, Robson ME, Taylor BS, Offit K, Berger MF, Stadler ZK. Microsatellite instability is associated with the presence of Lynch syndrome pan-cancer. J Clin Oncol. 2019;37(4):286–95. https://doi.org/10.1200/JCO.18.00283.
59. Le DT, Durham JN, Smith KN, Wang H, Bartlett BR, Aulakh LK, Lu S, Kemberling H, Wilt C, Luber BS, Wong F, Azad NS, Rucki AA, Laheru D, Donehower R, Zaheer A, Fisher GA, Crocenzi TS, Lee JJ, Greten TF, Duffy AG, Ciombor KK, Eyring AD, Lam BH, Joe A, Kang SP, Holdhoff M, Danilova L, Cope L, Meyer C, Zhou S, Goldberg RM, Armstrong DK, Bever KM, Fader AN, Taube J, Housseau F, Spetzler D, Xiao N, Pardoll DM, Papadopoulos N, Kinzler KW, Eshleman JR, Vogelstein B, Anders RA, Diaz LA Jr. Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. Science. 2017;357(6349):409–13. https://doi.org/10.1126/science.aan6733.
60. Le DT, Uram JN, Wang H, Bartlett BR, Kemberling H, Eyring AD, Skora AD, Luber BS, Azad NS, Laheru D, Biedrzycki B, Donehower RC, Zaheer A, Fisher GA, Crocenzi TS, Lee JJ, Duffy SM, Goldberg RM, de la Chapelle A, Koshiji M, Bhaijee F, Huebner T, Hruban RH, Wood LD, Cuka N, Pardoll DM, Papadopoulos N, Kinzler KW, Zhou S, Cornish TC, Taube JM, Anders RA, Eshleman JR, Vogelstein B, Diaz LA Jr. PD-1 blockade in tumors with mismatch-repair deficiency. N Engl J Med. 2015;372(26):2509–20. https://doi.org/10.1056/NEJMoa1500596.
61. Louveau B, Tost J, Mauger F, Sadoux A, Podgorniak MP, How-Kit A, Pages C, Roux J, Da Meda L, Lebbe C, Mourah S. Clinical value of early detection of circulating tumour DNA-BRAF(V600mut) in patients with metastatic melanoma treated with a BRAF inhibitor. ESMO Open. 2017;2(2):e000173. https://doi.org/10.1136/esmoopen-2017-000173.
62. Lu Y, Soong TD, Elemento O. A novel approach for characterizing microsatellite instability in cancer cells. PLoS One. 2013;8(5):e63056. https://doi.org/10.1371/journal.pone.0063056.
63. Lynch HT, de la Chapelle A. Hereditary colorectal cancer. N Engl J Med. 2003;348(10):919–32. https://doi.org/10.1056/NEJMra012242.
64. Mandal R, Samstein RM, Lee KW, Havel JJ, Wang H, Krishna C, Sabio EY, Makarov V, Kuo F, Blecua P, Ramaswamy AT, Durham JN, Bartlett B, Ma X, Srivastava R, Middha S, Zehir A, Hechtman JF, Morris LG, Weinhold N, Riaz N, Le DT, Diaz LA Jr, Chan TA. Genetic diversity of tumors with mismatch repair deficiency influences anti-PD-1 immunotherapy response. Science. 2019;364(6439):485–91. https://doi.org/10.1126/science.aau0447.
65. Marcus L, Lemery SJ, Keegan P, Pazdur R. FDA approval summary: Pembrolizumab for the treatment of microsatellite instability-high solid tumors. Clin Cancer Res. 2019;25(13):3753–8. https://doi.org/10.1158/1078-0432.CCR-18-4070.
66. Middha S, Zhang L, Nafa K, Jayakumaran G, Wong D, Kim HR, Sadowska J, Berger MF, Delair DF, Shia J, Stadler Z, Klimstra DS, Ladanyi M, Zehir A, Hechtman JF. Reliable pan-cancer microsatellite instability assessment by using targeted next-generation sequencing data. JCO Precis Oncol. 2017, 2017; https://doi.org/10.1200/PO.17.00084.
67. Murphy KM, Zhang S, Geiger T, Hafez MJ, Bacher J, Berg KD, Eshleman JR. Comparison of the microsatellite instability analysis system and the Bethesda panel for the determination of microsatellite instability in colorectal cancers. J Mol Diagn. 2006;8(3):305–11. https://doi.org/10.2353/jmoldx.2006.050092.
68. Nawroz H, Koch W, Anker P, Stroun M, Sidransky D. Microsatellite alterations in serum DNA of head and neck cancer patients. Nat Med. 1996;2(9):1035–7. https://doi.org/10.1038/nm0996-1035.
69. Niu B, Ye K, Zhang Q, Lu C, Xie M, McLellan MD, Wendl MC, Ding L. MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. Bioinformatics. 2014;30(7):1015–6. https://doi.org/10.1093/bioinformatics/btt755.
70. Nowak JA, Yurgelun MB, Bruce JL, Rojas-Rudilla V, Hall DL, Shivdasani P, Garcia EP, Agoston AT, Srivastava A, Ogino S, Kuo FC, Lindeman NI, Dong F. Detection of mismatch repair deficiency and microsatellite instability in colorectal adenocarcinoma by targeted next-generation sequencing. J Mol Diagn. 2017;19(1):84–91. https://doi.org/10.1016/j.jmoldx.2016.07.010.
71. Overman MJ, Lonardi S, Wong KYM, Lenz HJ, Gelsomino F, Aglietta M, Morse MA, Van Cutsem E, McDermott R, Hill A, Sawyer MB, Hendlisz A, Neyns B, Svrcek M, Moss RA, Ledeine JM, Cao ZA, Kamble S, Kopetz S, Andre T. Durable clinical benefit with nivolumab plus Ipilimumab in DNA mismatch repair-deficient/microsatellite instability-high metastatic

colorectal cancer. J Clin Oncol. 2018;36(8):773–9. https://doi.org/10.1200/JCO.2017.76.9901.
72. Overman MJ, McDermott R, Leach JL, Lonardi S, Lenz HJ, Morse MA, Desai J, Hill A, Axelson M, Moss RA, Goldberg MV, Cao ZA, Ledeine JM, Maglinte GA, Kopetz S, Andre T. Nivolumab in patients with metastatic DNA mismatch repair-deficient or microsatellite instability-high colorectal cancer (CheckMate 142): an open-label, multicentre, phase 2 study. Lancet Oncol. 2017;18(9):1182–91. https://doi.org/10.1016/S1470-2045(17)30422-9.
73. Pagin A, Zerimech F, Leclerc J, Wacrenier A, Lejeune S, Descarpentries C, Escande F, Porchet N, Buisine MP. Evaluation of a new panel of six mononucleotide repeat markers for the detection of DNA mismatch repair-deficient tumours. Br J Cancer. 2013;108(10):2079–87. https://doi.org/10.1038/bjc.2013.213.
74. Perucho M. Correspondence re: C.R. Boland et al., A National Cancer Institute workshop on microsatellite instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer. Cancer Res. 58: 5248–5257, 1998. Cancer Res. 1999;59(1):249–56.
75. Petrackova A, Vasinek M, Sedlarikova L, Dyskova T, Schneiderova P, Novosad T, Papajik T, Kriegova E. Standardization of sequencing coverage depth in NGS: recommendation for detection of clonal and subclonal mutations in cancer diagnostics. Front Oncol. 2019;9:851. https://doi.org/10.3389/fonc.2019.00851.
76. Phipps AI, Limburg PJ, Baron JA, Burnett-Hartman AN, Weisenberger DJ, Laird PW, Sinicrope FA, Rosty C, Buchanan DD, Potter JD, Newcomb PA. Association between molecular subtypes of colorectal cancer and patient survival. Gastroenterology. 2015;148(1):77–87 e72. https://doi.org/10.1053/j.gastro.2014.09.038.
77. Pritchard CC, Morrissey C, Kumar A, Zhang X, Smith C, Coleman I, Salipante SJ, Milbank J, Yu M, Grady WM, Tait JF, Corey E, Vessella RL, Walsh T, Shendure J, Nelson PS. Complex MSH2 and MSH6 mutations in hypermutated microsatellite unstable advanced prostate cancer. Nat Commun. 2014;5:4988. https://doi.org/10.1038/ncomms5988.
78. Redford L, Alhilal G, Needham S, O'Brien O, Coaker J, Tyson J, Amorim LM, Middleton I, Izuogu O, Arends M, Oniscu A, Alonso AM, Laguna SM, Gallon R, Sheth H, Santibanez-Koref M, Jackson MS, Burn J. A novel panel of short mononucleotide repeats linked to informative polymorphisms enabling effective high volume low cost discrimination between mismatch repair deficient and proficient tumours. PLoS One. 2018;13(8):e0203052. https://doi.org/10.1371/journal.pone.0203052.
79. Salipante SJ, Scroggins SM, Hampel HL, Turner EH, Pritchard CC. Microsatellite instability detection by next generation sequencing. Clin Chem. 2014;60(9):1192–9. https://doi.org/10.1373/clinchem.2014.223677.
80. Schwarzenbach H, Hoon DS, Pantel K. Cell-free nucleic acids as biomarkers in cancer patients. Nat Rev Cancer. 2011;11(6):426–37. https://doi.org/10.1038/nrc3066.
81. Seo HM, Chang YS, Joo SH, Kim YW, Park YK, Hong SW, Lee SH. Clinicopathologic characteristics and outcomes of gastric cancers with the MSI-H phenotype. J Surg Oncol. 2009;99(3):143–7. https://doi.org/10.1002/jso.21220.
82. Stoffel EM, Mangu PB, Limburg PJ. Hereditary colorectal cancer syndromes: American Society of Clinical Oncology clinical practice guideline endorsement of the familial risk-colorectal cancer: European Society for Medical Oncology clinical practice guidelines. J Oncol Pract. 2015;11(3):e437–41. https://doi.org/10.1200/JOP.2015.003665.
83. Tae H, Kim DY, McCormick J, Settlage RE, Garner HR. Discretized Gaussian mixture for genotyping of microsatellite loci containing homopolymer runs. Bioinformatics. 2014;30(5):652–9. https://doi.org/10.1093/bioinformatics/btt595.
84. Thibodeau SN, Bren G, Schaid D. Microsatellite instability in cancer of the proximal colon. Science. 1993;260(5109):816–9.
85. Thibodeau SN, French AJ, Cunningham JM, Tester D, Burgart LJ, Roche PC, McDonnell SK, Schaid DJ, Vockley CW, Michels VV, Farr GH Jr, O'Connell MJ. Microsatellite instability in colorectal cancer: different mutator phenotypes and the principal involvement of hMLH1. Cancer Res. 1998;58(8):1713–8.
86. Umar A, Boland CR, Terdiman JP, Syngal S, de la Chapelle A, Ruschoff J, Fishel R, Lindor NM, Burgart LJ, Hamelin R, Hamilton SR, Hiatt RA, Jass J, Lindblom A, Lynch HT, Peltomaki P, Ramsey SD, Rodriguez-Bigas MA, Vasen HF, Hawk ET, Barrett JC, Freedman AN, Srivastava S. Revised Bethesda guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability. J Natl Cancer Inst. 2004;96(4):261–8.
87. Vanderwalde A, Spetzler D, Xiao N, Gatalica Z, Marshall J. Microsatellite instability status determined by next-generation sequencing and compared with PD-L1 and tumor mutational burden in 11,348 patients. Cancer Med. 2018;7(3):746–56. https://doi.org/10.1002/cam4.1372.
88. Waalkes A, Smith N, Penewit K, Hempelmann J, Konnick EQ, Hause RJ, Pritchard CC, Salipante SJ. Accurate pan-cancer molecular diagnosis of microsatellite instability by single-molecule molecular inversion probe capture and high-throughput sequencing. Clin Chem. 2018;64(6):950–8. https://doi.org/10.1373/clinchem.2017.285981.
89. Wang C, Liang C. MSIpred: a python package for tumor microsatellite instability classification from tumor mutation annotation data using a support vector machine. Sci Rep. 2018;8(1):17546. https://doi.org/10.1038/s41598-018-35682-z.
90. Wang L, Ajani JA. Ushering in liquid biopsy for the microsatellite status: advantages and caveats.

Clin Cancer Res. 2019;25(23):6887–9. https://doi.org/10.1158/1078-0432.CCR-19-2585.

91. Willis J, Lefterova MI, Artyomenko A, Kasi PM, Nakamura Y, Mody K, Catenacci DVT, Fakih M, Barbacioru C, Zhao J, Sikora M, Fairclough SR, Lee H, Kim KM, Kim ST, Kim J, Gavino D, Benavides M, Peled N, Nguyen T, Cusnir M, Eskander RN, Azzi G, Yoshino T, Banks KC, Raymond VM, Lanman RB, Chudova DI, Talasaz A, Kopetz S, Lee J, Odegaard JI. Validation of microsatellite instability detection using a comprehensive plasma-based genotyping panel. Clin Can Res. 2019;25(23):7035–45. https://doi.org/10.1158/1078-0432.CCR-19-1324.
92. Wimmer K, Etzler J. Constitutional mismatch repair-deficiency syndrome: have we so far seen only the tip of an iceberg? Hum Genet. 2008;124(2):105–22. https://doi.org/10.1007/s00439-008-0542-4.
93. Wimmer K, Kratz CP, Vasen HF, Caron O, Colas C, Entz-Werle N, Gerdes AM, Goldberg Y, Ilencikova D, Muleris M, Duval A, Lavoine N, Ruiz-Ponte C, Slavc I, Burkhardt B, Brugieres L. Diagnostic criteria for constitutional mismatch repair deficiency syndrome: suggestions of the European consortium 'care for CMMRD' (C4CMMRD). J Med Genet. 2014;51(6):355–65. https://doi.org/10.1136/jmedgenet-2014-102284.
94. Yamamoto H, Imai K. An updated review of microsatellite instability in the era of next-generation sequencing and precision medicine. Semin Oncol. 2019;46(3):261–70. https://doi.org/10.1053/j.seminoncol.2019.08.003.
95. Zavodna M, Bagshaw A, Brauning R, Gemmell NJ. The accuracy, feasibility and challenges of sequencing short tandem repeats using next-generation sequencing platforms. PLoS One. 2014;9(12):e113862. https://doi.org/10.1371/journal.pone.0113862.
96. Zeinalian M, Hashemzadeh-Chaleshtori M, Salehi R, Emami MH. Clinical aspects of microsatellite instability testing in colorectal cancer. Adv Biomed Res. 2018;7:28. https://doi.org/10.4103/abr.abr_185_16.
97. Zhu L, Huang Y, Fang X, Liu C, Deng W, Zhong C, Xu J, Xu D, Yuan Y. A novel and reliable method to detect microsatellite instability in colorectal cancer by next-generation sequencing. J Mol Diagn. 2018;20(2):225–31. https://doi.org/10.1016/j.jmoldx.2017.11.007.
98. Zighelboim I, Goodfellow PJ, Gao F, Gibb RK, Powell MA, Rader JS, Mutch DG. Microsatellite instability and epigenetic inactivation of MLH1 and outcome of patients with endometrial carcinomas of the endometrioid type. J Clin Oncol. 2007;25(15):2042–8. https://doi.org/10.1200/JCO.2006.08.2107.

6

Computational Approaches for the Investigation of Intra-tumor Heterogeneity and Clonal Evolution from Bulk Sequencing Data in Precision Oncology Applications

Alessandro Laganà

Abstract

While the clonal model of cancer evolution was first proposed over 40 years ago, only recently next-generation sequencing has allowed a more precise and quantitative assessment of tumor clonal and subclonal landscape. Consequently, a plethora of computational approaches and tools have been developed to analyze this data with the goal of inferring the clonal landscape of a tumor and characterize its temporal or spatial evolution. This chapter introduces intra-tumor heterogeneity (ITH) in the context of precision oncology applications and provides an overview of the basic concepts, algorithms, and tools for the dissection, analysis, and visualization of ITH from bulk DNA sequencing.

A. Laganà (✉)
Department of Genetics and Genomic Sciences, Department of Oncological Sciences, Mount Sinai Icahn School of Medicine, New York, NY, USA
e-mail: alessandro.lagana@mssm.edu

Intra-tumor Heterogeneity and Cancer Evolution

Intra-tumor heterogeneity (ITH) refers to the genetic diversity observed in the population of cancer cells that make up an individual tumor. Such heterogeneity is the result of an evolutionary process that begins with one or more genetic alterations acquired by a single cell and passed on to its offspring [1]. While most alterations are likely passenger events that do not confer a selective growth advantage to the tumor, a small subset of driver alterations lead to the activation of oncogenes and/or the inactivation of tumor suppressors, promoting uncontrolled cell proliferation and immortality. As the disease progresses, tumor cells acquire additional alterations following a linear or a branched evolution pattern [2] (Fig. 6.1). While linear evolution involves the sequential acquisition of alterations over time, branched evolution is characterized by the emergence of subclonal cell populations which acquire different alterations and evolve independently [1]. Subclones may either cooperate or compete. Clonal cooperation can support tumor growth and progression. For example, a study on colorectal cancer demonstrated that RAS-mutant cells

© Springer Nature Switzerland AG 2022

A. Laganà (ed.), *Computational Methods for Precision Oncology*, Advances in Experimental Medicine and Biology 1361, https://doi.org/10.1007/978-3-030-91836-1_6

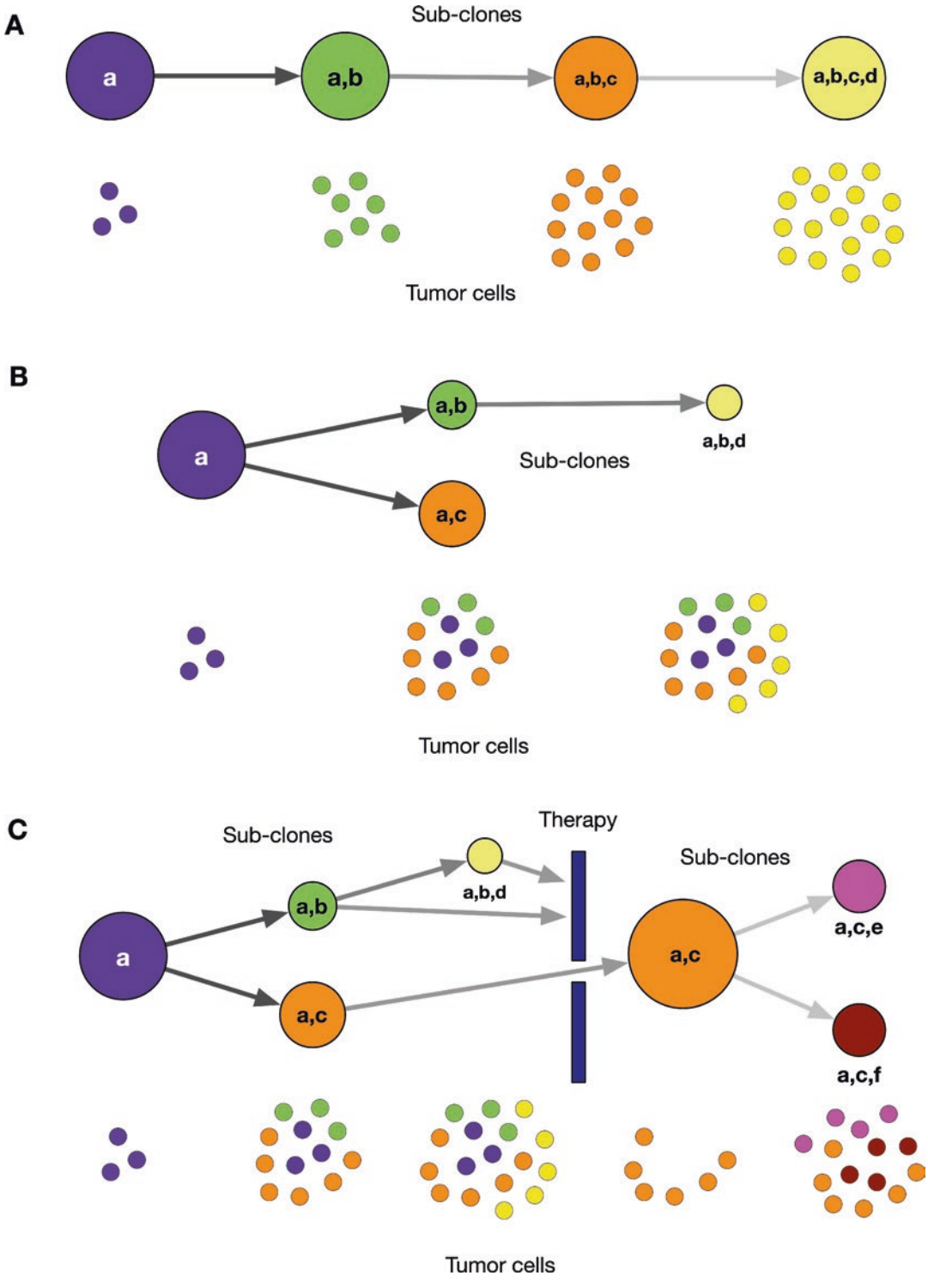

Fig. 6.1 Models of tumor clonal evolution. (**a**) Linear evolution: cells making up the tumor mass acquire and accumulate alterations (*a*, *b*, *c*, *d*) sequentially. (**b**) Branching evolution: subclones emerge by independent acquisition of additional alterations. The founding clone shown in purple is characterized by alteration *a*

(continued)

resistant to EGFR blockade by cetuximab created a protective microenvironment for sensitive cells by secretion of TGFα and amphiregulin, whose production was further increased by cetuximab treatment [3]. On the other hand, subclones may compete for space and vital resources and even alternate in a back-and-forth fashion for dominance with therapy [1, 4]. Furthermore, if a subclone promoting tumor growth through its interactions with the microenvironment is outcompeted by a more proliferative subclone which also depends on that favorable microenvironment, it can cause tumor collapse [5].

As ITH affects tumor behavior and growth, it may carry prognostic significance and drive drug resistance. A study by Landau et al. demonstrated the impact of subclonal heterogeneity on clinical outcome in chronic lymphocytic leukemia (CLL) patients, where the presence of a subclonal driver detected before treatment was associated with shorter progression-free survival, independently of the treatment received [6]. In our recent study on the characterization of newly diagnosed multiple myeloma (MM), we found that tumors characterized by complex subclonal landscape had significantly higher mutation burden and were less responsive to standard of care treatment [7]. A study on 54 childhood cancers, including neuroblastoma and Wilms tumors, revealed that the most dynamic landscapes characterized by clonal sweep, that is, the process through which a subclone outcompetes all the others and become dominant, or by the local emergence of numerous small clones, were associated with the poorest prognosis [8].

ITH is also one of the hallmarks of metastasis. For example, multi-region sequencing of primary renal cell carcinomas and associated metastatic sites revealed branched evolutionary tumor growth and significant mutational and gene expression heterogeneity between the different areas of the same tumors, characterized by different prognostic features [9]. This clearly indicates that a single biopsy may not provide sufficient information about an individual cancer and represents a significant challenge for the design of an effective treatment.

Intra-tumor Heterogeneity and Precision Oncology

Precision oncology is an emerging and fast evolving research field introducing a novel approach to cancer patient's care where diagnosis, prognosis, and therapy are informed by the specific genetic and molecular features of the individual patient's cancer, rather than by the cancer type [10–14]. A key concept in precision oncology is the *actionable alteration*, which is an alteration (e.g., a mutation, a copy number change or another structural variation) that can be targeted with a specific drug or drug combination [15]. A typical example is the BRAF V600E mutation in melanoma, which can be targeted by BRAF and/or MEK inhibitors [16]. While several successful cases of targeted therapies have been reported, many patients only experience partial and/or short-lived benefits, and ITH is one of the likely causes for treatment failure. In fact, while the presence of a specific alteration in a tumor may indicate a potential benefit from a specific targeted treatment, it does not account for other subclones, often barely detectable before treatment, which may already have or develop resistance to the same therapy, and which would then emerge and re-shape the tumor landscape once the sensitive subclones have been eliminated [17]. Furthermore, initially sensitive clones may adapt to selective pressure imposed by therapy and develop de novo

Fig. 6.1 (continued) (e.g., a mutation). The green and orange subclones acquire alterations *b* and *c*, respectively, and inherit alteration *a* from the parent clone. A further subclone, represented in yellow, emerges from the green subclone, carrying an additional alteration *d*. The tumor mass is then composed of a mixture of cells harboring different alterations. Subclones in different branches share all the alterations of their ancestors and carry unique alterations specific to the branch. (**c**) The impact of therapy on the tumor's clonal landscape. In this example, therapy eradicates the green and yellow subclones. However, the orange subclone is resistant and expands following treatment. Two new subclones, magenta and brown, emerge acquiring additional alterations. The tumor mass after therapy is quite different than the founding one

resistance. Increasing evidence suggests that therapy in the presence of resistant subclones may actually accelerate tumor progression [1]. For example, HER2 targeted therapies in breast cancer are only successful when all the tumor cells express and are dependent on HER2, and a recent study reported that breast tumors with high degree of ITH and HER2 status heterogeneity are associated with shorter disease-free survival [18]. Therefore, ITH poses a significant challenge in the successful implementation of targeted therapies. Ideally, one way to prevent the scenario illustrated above is to target multiple subclones simultaneously, assuming that the subclones harbor targetable alterations and that the corresponding drug combination is implementable. Another possible approach is to prevent the emergence of resistance subclones by targeting multiple pathways simultaneously, like in the case of BRAF mutant melanoma, where both the MAPK and PI3K pathways upfront may block or delay the emergence of resistance clones [19].

Novel clinical trials have been designed to investigate tumor's ITH and evolution and their clinical implications. For example, the TRACERx trial for non-small cell lung cancer aims at defining the evolutionary trajectories of individual cancers both spatially and temporally through multi-region and longitudinal tumor sampling [20–22]. Interim findings from the trial revealed widespread ITH in terms of both SNVs and CNAs and that higher ITH was associated with increased risk of relapse and death. Moreover, the analysis suggested an important role of ongoing chromosomal instability in subclonal selection and tumor evolution, which was then further investigated and demonstrated in a follow-up pan-cancer study [23]. The TRACERx trial for renal cell carcinoma showed similar findings and provided evidence of specific alterations, such as loss of chromosome 3p, preceding the growth of a clinically detectable tumor by 30–50 years [24–26].

These and similar studies not only allow to dissect the genetic and molecular mechanisms driving cancer evolution, but also demonstrate the clinical implications of ITH, from prognosis assessment to therapy design and management. The DARWIN and DARWIN II trials (Deciphering Anti-tumor Response With INtratumor Heterogeneity) (http://clinicaltrials.gov/show/NCT02183883) were designed for patients enrolled in TRACERx and aim to assess the impact of ITH in lung cancer patients with specific actionable mutations (e.g., activating EGFR mutations, BRAF V600 mutations, or ALK/RET rearrangements) assigned to study arms according to their alterations (e.g., afatinib for EGFR, vemurafenib for BRAF V600, or alectinib for ALK/RET rearrangements). The data from the study will be used to determine the impact of a targeted clonal versus subclonal mutation on treatment outcome and to dissect the dynamics of subclonal changes through therapy. Moreover, in DARWIN II, patients without actionable mutations receive the anti-PD-L1 monoclonal antibody atezolizumab and their disease is tracked to investigate genomic and immune markers that may predict response to immune checkpoint inhibitors (https://clinicaltrials.gov/ct2/show/NCT02314481).

To account for ITH and its clinical consequences, precision oncology platforms should be equipped with specific tools to reconstruct and characterize the individual tumor's clonal landscape and model tumor's evolution by inferring its clonal temporal and spatial trajectory. Specialized tools addressing these tasks are discussed in the next section, along with practical considerations for their implementation in precision oncology pipelines.

Inferring Tumor's Clonal Landscape from Bulk DNA Sequencing

While the clonal model of cancer evolution was first proposed over 40 years ago, only recently next-generation sequencing has allowed a more precise and quantitative assessment of tumor clonal and subclonal landscape. Consequently, a plethora of computational approaches and tools have been developed to analyze this data with the goal of inferring the clonal landscape of a tumor and characterize its temporal and/or spatial evolution. This section covers the basic concepts behind the computational approaches for the dissection of ITH from bulk whole-genome (WGS)

and whole-exome (WES) sequencing. A more comprehensive and in-depth review of the principles and methods for cancer subclonal reconstruction can be found elsewhere [27–29].

The main goal of ITH analysis in bulk sequencing data is to identify the set of clonal and subclonal cell populations in a tumor sample and annotate each clone/subclone with the specific alterations that they carry. Additionally, many tools also address the problem of reconstructing the tumor phylogeny, which organizes the identified clones/subclones in tree structures, where downstream nodes inherit all the alterations present in their ancestor nodes and carry additional ones (Fig. 6.1). For practical purpose and increased readability, moving forward we will refer to both clones and subclones simply as *clones*, except where a distinction is necessary.

The main idea that most algorithms for the identification of tumor clones implement is that single nucleotide variants (SNVs) that co-occur in the same clone have similar allele frequency (VAF); therefore, such values can be used to infer the proportion of cells carrying a specific SNV, or its cellular prevalence (CP) (See also Glossary in Box 6.1). Moreover, since tumor samples are often contaminated by non-cancer cells, tumor purity (TP), which is the proportion of cancer cells in the sample, allows to calculate the proportion of tumor cells which carry the SNV, or cancer cell fraction (CCF). Methods for tumor clonal landscape reconstruction rely on the infinite-sites assumption (ISA), which states that a site does note mutate twice during the evolutionary history of a tumor. Although this assumption may not be true at all times, it is necessary in order to solve the problem of inferring tumor clones, which would otherwise be computationally intractable. Furthermore, the ISA supports the pigeonhole principle used for the reconstruction of tumor phylogeny, which states that the sum of CCFs of subclones cannot exceed the CCF of their ancestors [27, 28, 30].

Cancer genomes are often characterized by widespread copy number alterations (CNAs), and knowing the number of copies of a chromosome segment bearing an SNV, which is known as multiplicity, is necessary to accurately infer tumor

Box 6.1. Key Terms

Branched evolution: the independent acquisition of different alterations by subclones, which causes the emergence of subclonal cell populations described by a branching structure in the clonal tree.

Cancer cell fraction (CCF): the fraction of tumor cells harboring a set of SNVs.

Cellular prevalence (CP): the fraction of cells harboring a set of SNVs. It includes both tumor and normal cells in the sample.

Clonal tree: a phylogenetic tree structure describing the evolutionary relationships between clones and subclones.

Clone: a population of mutant cells that expands to form a neoplasm. Multiple clones can cooperate or compete to dominate the tumor ecosystem.

Copy number alteration (CNA): somatic changes in the number of copies of a chromosome area, which can be gained (>2 copies) or lost (<2 copies). CNA can affect small sections of a chromosome (focal) or wide areas, including whole chromosome arms (broad).

Infinite sites assumption (ISA): a site in the genome does note mutate twice during the evolutionary history of a tumor. This assumption is essential to reduce the complexity of computational subclonal reconstruction and holds true in most cases, although it can be occasionally violated for SNVs.

Linear evolution: the sequential acquisition of alterations over time, where each node (clone/subclone) in the clonal tree can only have one parent and one child.

Multiplicity: the number of copies of a chromosome segment carrying an SNV.

Pigeonhole principle: the sum of CCFs of tumor subclones must be less than the CCF of their ancestor clones.

Single nucleotide variation (SNV): a variation in a single nucleotide, also known as a point mutation. SNVs can be germline, that is, present in all cells and hereditary, or

(continued)

Box 6.1 (continued)

somatic, that is, only present in specific cells, such as a tumor clone.

Structural variation (SV): a large genomic alteration, including deletions, inversions, insertions, and translocations.

Subclone: a population of mutant cells that arise from the main tumor clone or another subclone by acquiring additional alterations, including drivers of tumor expansion and/or drug resistance.

Tumor purity (TP): the fraction of tumor cells in a sample.

Variant allele frequency (VAF): the percentage of sequence reads harboring a variation, for example, an SNV, divided by number of total reads covering the specific genomic locus.

clones. For example, if a site mutates and is next duplicated, its VAF will be different than it would be if the duplication had occurred before the mutation. For this reason, most tools for ITH reconstruction leverage both SNVs and CNAs, often pre-calculated using other specific tools. Besides identifying chromosome sections with copy number changes, CNA callers can additionally infer allele-specific copy number and estimate tumor purity and ploidy (i.e., the number of sets of chromosomes in the tumor). Therefore, the information they provide is employed to estimate the CCF of clones.

Modeling ITH is a very complex problem, since the SNV and CNA profile of a tumor can often be explained by several equally likely models. For example, two different clonal populations with the same CCF are likely to be identified as a unique clone. While mutation phasing, that is, determining whether SNVs are co-occurring or mutually exclusive, can often be performed on a single tumor sample, thus enabling the separation of distinct clones, adding more information is often necessary to accurately dissect the clonal landscape of a tumor. This can be done by providing multiple samples from the same cancer, separated either longitudinally (i.e., sampled at different times through disease evolution and treatment course) or spatially (i.e., sampled from different areas of the tumor mass or from metastatic sites). The additional information provided by multi-sample sequencing may indeed allow better mutation phasing and to distinguish between similar cell populations that may have different CCFs in different regions and/or at different times.

Numerous tools have been developed in the past decade to reconstruct the tumor landscape and evolutionary phylogeny from bulk sequencing data. The remainder of this section present the most popular tools for ITH deconvolution, introduced chronologically. For each tool, the main features, algorithm, and limitations, along with the type of input accepted and output generated, are provided. Tools are also summarized in Table 6.1, along with their distinctive features and URLs.

PyClone

PyClone is a tool that implements a hierarchical Bayes statistical model for the identification and quantification of subclonal tumor cell populations based on WES data [31]. It requires a coverage of at least 100× and works on both single and multiple samples from the same patient. The algorithm works under the assumption that mutations with similar VAF belong to the same subclone but uses a Bayesian beta-binomial model to estimate the proportion of tumor cells harboring a mutation using its VAF. PyClone incorporates allele-specific copy number at each mutation locus in each sample, obtained from either genotyping array (e.g., Array Comparative Genomic Hybridization or aCGH) or WGS, which enables it to cluster mutations occurring in regions with copy number variations. It outputs posterior densities of cellular prevalences for each mutation and the clustering structure over the mutations. A new version of PyClone, called PyClone-VI, is orders of magnitude faster than the original PyClone, while maintaining the same accuracy [32]. PyClone-VI enables the analysis of WGS data from large cohorts of tumor samples harboring hundreds of thousands of mutations.

Table 6.1 Computational tools for the analysis of ITH and subclonal reconstruction

Tool	Single/multi samples[a]	Seq type[b]	Input[c]	Phylogeny[d]	Notes	URL	Refs.
CALDER	Multiple (temporal)	WES/WGS	SNV	Yes	Does not correct for CNA	https://github.com/raphael-group/calder	[44]
Canopy	Multiple (spatial or temporal)	WES/WGS	SNV, CNA	Yes	–	https://github.com/yuchaojiang/Canopy	[37]
CloneHD	Single or multiple (spatial or temporal)	WES/WGS	SNV, read depth, B-allele counts	No	–	https://github.com/andrej-fischer/cloneHD	[34]
FastClone	Single	WES/WGS	SNV, CNA	Yes	–	https://github.com/GuanLab/FastClone_GuanLab	[48]
Meltos	Multiple (spatial)	WGS	SNV, SV, BAM, phylogeny tree	Yes	Extracts read counts directly from BAM	https://github.com/ih-lab/Meltos	[46]
Palimpsest	Single or multiple	WES/WGS	SNV, CNA, purity	No	Generates a comprehensive oncogenic timeline annotated with the clonal and subclonal mutations and the timing of the predicted driver mutations	https://github.com/FunGeST/Palimpsest	[38]
PhyloWGS	Single or multiple (temporal)	WES/WGS	SNV, CNA	Yes	–	https://github.com/morrislab/phylowgs	[35]
PyClone	Single or multiple	WES/WGS	SNV, CNA	No	–	https://github.com/Roth-Lab/pyclone	[31]
PyClone-VI	Single or multiple	WES/WGS	SNV, CNA	No	Faster version of PyClone	https://github.com/Roth-Lab/pyclone-vi	[32]
QuantumClone	Single or multiple (temporal)	WES/WGS	SNV, CNA	No	–	https://github.com/DeveauP/QuantumClone	[39]
SciClone	Single or multiple	WES/WGS	SNV, CNA	No	Does not correct for CNA	https://github.com/genome/sciclone	[33]
SPRUCE	Multiple (temporal)	WES/WGS	SNV, CNA	Yes	–	https://github.com/raphael-group/spruce	[36]
SubMARine	Single or multiple	WES/WGS	SNV, CNA	Yes	–	https://github.com/morrislab/submarine	[53]

(continued)

Table 6.1 (continued)

Tool	Single/multi samples[a]	Seq type[b]	Input[c]	Phylogeny[d]	Notes	URL	Refs.
SuperFreq	Single or multiple	WES/ WGS	SNV, CNA, BAM	Yes	Does not need a matched normal sample; uses quality scores from BAM file	https://github.com/ChristofferFlensburg/superFreq	[50]
SVclone	Single	WGS	SV, BAM	No	SNV, CNA and TP can be optionally provided	https://github.com/mcmero/SVclone	[42]
Tusv	Single or Multiple	WGS	SV, CNA	Yes	Does not use SNV data	https://github.com/jaebird123/tusv	[40]

[a]Whether the tool runs on a single or multiple tumor samples, either spatially or temporally separated, or both
[b]Type of sequencing data necessary. WGS = whole genome sequencing; WES = whole exome sequencing
[c]Type of input accepted. SNV = single nucleotide variants; CAN = copy number alterations; SV = structural variants; BAM = binary alignment map (file format of reads alignment)
[d]Whether the tool generates a tumor phylogenetic tree

SciClone

SciClone employs a variational Bayesian mixture model to identify the number and composition of tumor subclones in single or multiple samples from the same patient, based on the VAF of SNVs [33]. Like PyClone, SciClone assumes that mutations with similar VAF belong to the same subclone; thus, neither tool can distinguish between different subclones with similar CP. While SciClone does take both SNVs and CNAs as input, it only overlays CNAs on copy number neutral SNVs. Therefore, it cannot cluster mutations in regions with CNAs, which makes it not applicable to many types of tumors which recurrently harbor such genetic events.

CloneHD

CloneHD reconstructs the subclonal structure of a tumor cell population employing a set of coupled Hidden Markov Models (HMMs) jointly across multiple sources of information, such as read depth, B-allele counts, and SNV VAFs, in a way that facilitates ruling out different competing models [34]. Read depth, which refers to the number of reads mapping to different loci in the genome, is used to infer the copy number profiles of the different cell population. B-allele counts, that is, the number of reads reporting a minor allele at an originally heterozygous locus, help differentiating between balanced and unbalanced copy number changes. SNV data provides information about the size and genotype of the different subclonal fractions. Using multiple samples from the same patient, either spatially or longitudinally correlated, can improve inference of the tumor structure, under the assumption that the same subclones are present in all samples, possibly at different CCF.

PhyloWGS

PhyloWGS was specifically designed to infer the subclonal landscape of tumor cells using SNVs and CNAs obtained from WGS samples, where the decreased read depth, which can complicate subclonal reconstruction, is compensated by the larger number of mutations [35]. However, PhyloWGS can be also applied to WES data or a mixture of WES and WGS data, for example, SNVs from WES and CNAs from WGS.

Similar to PyClone, PhyloWGS accounts for CNAs but it integrates it with SNVs in a phylogenic reconstruction. In fact, the observed VAF of SNVs in a sample is affected by the phylogenetic relationships between SNVs and CNAs. For example, a heterozygous SNV which occurs before a copy number gain of its locus, will be now present in two out of the three copies of the area. However, if the copy number gain occurs before the SNV, this will be present in one copy only (Fig. 6.1). Another scenario can present when the SNV and CNA occur in the same locus but on different branches of the subclonal populations. Therefore, the VAF of the SNV will reflect such differences. PhyloWGS models all these scenarios to provide a more accurate landscape of the tumor subclonal composition. It employs a generative probabilistic model of VAFs that incorporates a non-parametric Bayesian prior over tree. The output includes the list of subclonal cell populations with their estimated CCF and their SNVs and CNAs. It also provides samples of the phylogenetic trees explaining the data, ranked by their likelihood. When multiple longitudinal samples from the same patient are provided, the phylogenetic trees visualize the changes in CCF across time points.

SPRUCE

SPRUCE defines the problem of inferring tumor evolution as a perfect phylogeny mixture deconvolution problem, where the goal is to reconstruct a tumor phylogenetic tree given mixtures of its leaves, that is, the observed mixtures of cancer cells with different SNVs and CNAs, under the infinite alleles model [36]. This means that while a genomic locus may mutate multiple times in the tree, each specific mutation may only appear once. The algorithm solves the problem by jointly modeling SNV and CNA data from multiple tumor samples using a combinatorial enumeration approach. It first computes a set of compatible

trees for each altered genomic locus, where a locus is described by a multi-state character comprising the number of wild type and mutated copies. Then, it derives a compatibility graph whose edges are pairs of compatible trees for pairs of altered genomic loci. All the evolutionary trees that are compatible with the observed data are then obtained by enumerating all the multi-state perfect phylogeny trees on the largest subset of characters. While potentially high, the number of phylogenetic trees is dramatically reduced when considering multiple samples for a tumor, thus reducing the ambiguity associated with many potential solutions.

Canopy

Canopy implements a probabilistic model and performs joint phylogenetics and deconvolution using a Markov chain Monte Carlo (MCMC) sampling procedure [37]. It is specifically designed for the analysis of ITH in multiple spatially or temporally separated samples from the same patient; thus, it is unsuited for cases where only one tumor sample is available. Canopy takes SNV VAFs and allele-specific CNA estimates and, like PhyloWGS, jointly models them so that SNVs that fall within CNA regions can be phased and temporally ordered. However, while PhyloWGS requires pre-processed CNA data and uses the absolute copy number of each allele to first estimate the subclonal structure, which is then integrated with SNVs, Canopy takes raw CNA data and performs a truly joint inference of subclones and their evolutionary history, which allows it to achieve greater accuracy in complex scenarios. The algorithm includes a pre-clustering initialization step aimed at improving robustness to noise and reducing computation time. The output consists of one or multiple evolutionary models explaining the data along with their posterior confidence assessment.

Palimpsest

Palimpsest implements an automated comprehensive workflow for the integrative analysis of mutational signatures and clonality analysis aimed at reconstructing tumor phylogeny [38]. Palimpsest takes as input SNVs, CNAs, and tumor purity from one or multiple tumor sample from the same patient. The algorithm first estimates CCF based on TP and CNA, then classifies each variant as clonal or subclonal and uses SNVs to extract patterns of mutational signatures in early and late subclonal mutations. Additionally, users can provide data on structural variations (SVs), which are then classified in 38 different categories according to their type (e.g., deletion, inversion, chromosomal translocation) and size. Then the algorithm uses a Bayesian statistic to estimate the probability of each SNV and SV being due to each identified signature and to predict the mechanism at the origin of each driver event. The integrated analysis of SNVs and CNAs then allows to estimate the molecular timing of chromosomal gains using the proportion of duplicated/non-duplicated SNVs. Finally, Palimpsest generates a comprehensive oncogenic timeline annotated with the clonal and subclonal mutations and the timing of the predicted driver mutations.

QuantumClone

Like PhyloWGS, SPRUCE, and Canopy, QuantumClone uses both SNV and CNA data to improve the accuracy of tumor clonal reconstruction [39]. It can be applied to single or multiple samples from the same patient, either spatially or temporally separated. The algorithm performs clustering of cellular prevalence of SNVs, which are calculated as a function of VAF, number of copies of the corresponding locus in tumor and normal cells, and TP. As the number of copies carrying each variant is unknown, the Expectation Maximization (EM) algorithm is used to identify the most probable cellular prevalence value based on the probability to observe a specific number of reads supporting a mutation given the number of reads overlapping the locus, the purity, and the cellularity of a clone. The authors of QuantumClone showed that their method outperforms PhyloWGS, as well as PyClone and SciClone, both in terms of accuracy of clonal

reconstruction and computation time. However, QuantumClone does not model tumor phylogeny.

Tusv

Tusv aims at resolving tumor deconvolution and phylogeny based on SVs, which are defined as pairs of breakpoints found adjacent to one another in the tumor sample but at non-adjacent positions in the reference genome [40]. The algorithm uses CNA data in the specific form produced by the tool Weaver, which returns allele-specific copy numbers of regions and phased breakpoints supporting the SVs [41]. Tusv employs *coordinate descent*, an optimization algorithm, to solve copy number profiles at subclonal level, the distribution of clones, and the phylogeny describing the ancestral relationships between clones by leveraging the mixed copy number profile and breakpoints from the tumor sample. Because SVs can be accurately inferred only from WGS data, Tusv cannot be applied to WES data, therefore limiting its range of applicability. Another limitation of Tusv, with regard to precision medicine applications, is that it does not take into account SNV data, thus reducing the clinical actionability of the inferred subclonal cell populations to specific cases supported by SV-related data.

SVclone

Similar to Tusv, SVclone aims at inferring the CCF of SV breakpoints, including CNAs, from WGS data of a single tumor sample [42]. The algorithm requires pre-calculated SV calls, for example by the tool Socrates, and the alignment file in BAM format, which it uses to determine the directionality of SVs and to classify them according to their type (inversions, deletions, tandem duplications, interspersed duplications, and intra- and inter-chromosomal translocations) [43]. Then, it estimates the VAF of SVs based on the number of supporting reads and removes low-quality SVs. If CNVs, SNVs, and TP are provided, SVclone infers the background copy number for each break-end and match SNVs with the SV loci. Next, the algorithm uses purity, ploidy, and copy number status of the normal and tumor cell populations to estimate the SV CCFs and then cluster them based on these values. To compensate for possible small number of SVs and improve the inference of clonal composition, SVclone additionally derives clusters from SNVs and then reassigns SV cluster memberships to either an SNV model, or to a joint SV + SNV model. The output consists of the estimated subclonal composition of the tumor, that is, number of clusters (subclones), the subclonal multiplicity, the variants assigned to each subclone, and their CCF. SVclone does not infer tumor phylogeny.

CALDER

CALDER is a tool that reconstructs tumor phylogenetic trees from longitudinal samples based on a vertex-colored tree model, where colors encoding temporal order are assigned to each node, that is, cell population [44]. The problem of reconstructing the tree is formulated as the Longitudinal Variant Allele Frequency Factorization Problem (LVAFFP), where SNVs are first clustered based on their VAFs, under the assumptions that they are in copy-neutral regions. SNVs affected by CNA should be excluded from the analysis or their read counts be corrected accordingly prior to running the analysis. The clonal tree is reconstructed using a matrix factorization approach, where the goal is to determine a perfect phylogeny matrix B, which describes the evolutionary relationships between clones, and a clone proportion matrix U, which describes the composition of a tumor at each time point, such that the matrix of the observed mutation frequencies at different time points F is a product of U and B, that is, $F = U*B$. The output of CALDER consists of a file containing the inferred phylogenetic tree T, which can be visualized using tools such as graphviz, and a file containing the clone proportion matrix U [45].

Meltos

Meltos is a computational probabilistic framework that uses somatic structural variants (SVs) from multiple spatially separated WGS samples from the same patient to reconstruct tumor phylogeny trees [46]. Meltos leverages phylogeny trees inferred more accurately from somatic SNVs to identify high confidence SVs and learn about the evolution of SVs in a multiple-sample scenario. The authors show that, although the evolutionary trajectory of SVs is not necessarily the same as for SNVs, the phylogeny tree inferred from SNVs can guide the assignment of somatic SVs to the tree. Meltos extracts read counts directly from BAM files and takes as input SNV and SV calls along with a phylogeny tree already inferred from SNV by the tool LICHeE [47]. It then applies quality filters and calculates VAFs for SNVs and SVs, which are then matched and clustered together. SVs with VAFs that do not match any cluster form new potential nodes and place in the tree using evolutionary constraints, under the assumption that a tree inferred based on SNVs only is a subtree of the true clonal tree.

FastClone

FastClone implements a probabilistic model for inferring tumor heterogeneity similar to previous tools such as PyClone, SciClone, and PhyloWGS [48]. It ranked first in the DREAM SMC-Het challenge, which is discussed in section "Visualizing Clonal Landscape and Tumor Clonal Evolution", and its main innovation is that it extends previous approaches with the scenario in which different subclones have independent CNV events within the same chromosome sections [49]. Furthermore, the algorithm is extremely fast and can analyze a tumor sample with tens of thousands of mutations in a few seconds. FastClone's input consists of SNV calls in VCF format and CNV calls from the tool Battenberg [30]. The algorithm first identifies subclones via Kernel density estimation based on the VAF and copy number of the SNVs located on non-ambiguous chromosome regions. Then, it assigns all the SNVs, regardless of their CNA status, to the subclones by maximizing an SNV/subclone association score. Its output consists of a list of subclones annotated with their CCF and the assigned mutations, as well as the structure of the phylogenetic tree with the highest likelihood.

SuperFreq

SuperFreq is a WES analysis pipeline that integrates the identification of SNVs and CNAs with the inference of intra-tumoral heterogeneity and clonal evolution over multiple samples from the same individual [50]. Notably, SuperFreq does not require a matched normal and instead relies on unrelated controls to separate somatic and germline SNVs. This makes it a suitable choice for those applications where it is not easy to obtain normal samples. SuperFreq takes BAM files for tumor and reference normal samples and preliminary SNV calls as input, for example obtained using samtools or varscan, then filter the SNVs using quality scores in the BAM file and through comparison with the reference normal samples [51, 52].

SubMARine

SubMARine is a tool that reconstruct the evolutionary history of a tumor by calculating a partial clone tree, which is a polynomial-space representation of a potentially exponentially sized set of clone trees [53]. More specifically, it defines pairwise ancestral relationships between subclones (e.g., subclone A is an ancestor of subclone B) and select all the trees that are consistent with such relationship as well as their parents. The Maximally Constrained Ancestral Reconstruction (MAR) is the unique partial clone tree which defines the maximal set of all the ancestral relationships constrained by the input data. The SubMARine algorithm identifies the subMAR, which is a unique partial clone tree that approximates the MAR and whose relationships are guaranteed to be a subset of those present in the

MAR, in polynomial time. SubMARine models both SNVs and CNAs and, like FastClone, can infer the clonal landscape of a tumor in a very short time.

Measuring ITH from Bulk RNA-Seq

As explained in the previous sections, ITH is defined based on genomic alterations, for example, SNVs, CNAs, and SVs; therefore, it is straightforward to solve it by modeling these types of alterations obtained from WES or WGS data. Nevertheless, a few approaches have been proposed to evaluate ITH levels using RNA-Seq data. In fact, genomic alterations often lead to changes in gene expression profiles. Moreover, being able to dissect ITH from RNA-Seq data would have the advantage of simultaneously describing the transcriptomic alterations associated with clonal heterogeneity and, ultimately, provide a more comprehensive landscape of a tumor. However, reconstructing ITH using RNA-Seq is a challenging task. While gene expression deconvolution has been successfully employed to identify different cell populations in a mixture sample, that is, a sample comprising both tumor and stromal cells, by relying on certain signatures specific of the different cell populations that may be present, it is not a viable strategy in the context of ITH, where the different cell populations are for the most part arbitrary and specific to the individual patient [54]. Currently, there are no methods or tools available to dissect ITH using bulk RNA-Seq data, but a few approaches have been proposed to measure and quantify levels of ITH in tumor RNA-Seq samples.

Park et al. have proposed a method called tITH (transcriptome-ITH), which models ITH in RNA-Seq samples as an entropy-based distance between two protein-protein interaction networks (PPIs) [55]. More specifically, tITH uses PPI networks to model gene-gene relationships and pathway information. The assumption is that pathway ambiguity increases along with tumor's clonal complexity as subclones arise, which can be expressed as network perturbation and measured by nJSD (network-based Jensen-Shannon divergence), which is the sum of entropy values measured at each of the genes in a PPI network. In their article, the authors show that their method effectively measures levels of ITH, as tITH correlates with genomic ITH during tumor progression, and higher tITH is significantly associated with worse prognosis. A more recent work proposed DEPTH (Deviating gene Expression Profiling Tumor Heterogeneity), a novel algorithm to evaluate ITH levels in RNA-Seq samples [56]. DEPTH is based on the observation that consistently high or low deviation of gene expression from their mean values correspond to low ITH, while mixed deviation values correspond to high ITH. The authors showed that the DEPTH score positively correlates with measures of genomic instability such as tumor mutation burden (TMB), microsatellite instability (MSI), and homologous recombination deficiency (HRD), and that higher DEPTH scores correspond with worse survival in multiple cancer types.

While RNA-Seq-based approaches for the analysis of ITH cannot dissect the clonal landscape of a tumor and identify clonal cell populations, they can still provide a measure of ITH in the tumor, particularly where DNA data is not available. However, it is not yet clear whether RNA-based measure provides additional information compared to other established metrics of genomic instability and tumor complexity, especially in the context of a precision oncology application.

Visualizing Clonal Landscape and Tumor Clonal Evolution

Intuitive and informative diagrams have been developed to represent ITH and visualize the clonal landscape of a tumor and its evolution in time and space. Such diagrams can be very useful features in precision oncology, allowing to convey the complexity of ITH clearly and concisely in patient genomic reports.

A *perspective* article, from a few years ago, discusses how to effectively visualize temporal and spatial clonal evolution in 2D plots [57]. Qualitative tumor clone evolution diagrams are

introduced and dissected, presenting different scenarios describing tumor progression, genetic and therapeutic bottlenecks, and clonal changes following clinical response. Graphical strategies for drawing clone evolution diagrams are presented and discussed, as well as how to combine quantitative displays of evolutionary relationships and clone populations.

Many of the tools for ITH reconstruction and modeling of tumor evolution introduced in this chapter provide both textual and graphical output, where the latter may consist of a plot displaying clusters of cell populations or a tree describing the hierarchical relationship between the different clonal populations and the linear and branching patterns of tumor evolution.

Figure 6.2 displays three different representations of clonal plots. In Fig. 6.2a, an example of output from SciClone is shown, where the top plot displays the kernel density plot across regions of one to four copies, identifying five different clones (each clone corresponding to a peak in the model fit) and the plots below display mutation clusters as VAFs versus read depth for each of the four copy number regions [33]. SciClone does not infer clonal evolution; therefore, the graphical output is limited to cell clustering.

Figure 6.2b shows an example of tumor clonal tree and trajectory from diagnosis (sample 1) to relapse (sample 2) inferred by PhyloWGS [35]. In the upper panel, clone 0 represents a normal cell, clone 1 represents the parent tumor clone and the children nodes represent the subclones, which are connected with clone 1 and with one another via edges describing their evolutionary relationships. The size of each node is proportional to the CCF of the clone. In the specific example, the parent clone has a mutation in BRAF, which is then inherited by all children subclones. Additionally, subclone 3 has acquired a TP53 mutation, which is passed to subclone 4. In the other branch, instead, subclone 6 has acquired a mutation in EGFR. The bottom panel shows the trajectories of subclonal expansion from diagnosis to relapse, where subclones 2 and 3 have a slightly lower CCF at relapse, subclone 4 is almost wiped out, and subclone 5 has expanded.

Finally, Fig. 6.2c shows an example of tumor evolutionary trajectory generated by the R package *fishplot* [58]. While *fishplot* automatically imports tumor phylogenies inferred by the tool ClonEvol, it can also be easily applied to data generated by any ITH tool, which must be provided as a matrix with the fraction of each cell population at each time point [59]. In the figure, the founding clone with a BRAF mutation is represented in gray and two children subclones are shown in green and red. A very small subclone is represented in yellow. A further subclone has emerged from the red population, represented in orange, harboring an additional mutation in MET. At relapse, the tumor is made up by the yellow subclone, barely present at diagnosis and carrying a mutation in KRAS, and the red subclone, while the green and orange subclones have been eradicated by therapy.

Assessment of ITH Estimates

Assessing the accuracy of the different methods and tools for clonal reconstruction is a difficult task, given the lack of a gold-standard or ground truth reference for benchmarking as well as of objective metrics and scores to use for the assessment. A study published a few years ago performed a systematic evaluation of six computational methods for ITH reconstruction from bulk WES data on a large dataset with >1600 samples of breast invasive carcinoma, bladder urothelial carcinoma, and head and neck squamous cell carcinoma from The Cancer Genome Atlas (TCGA) [60]. The study included the tools SciClone, PyClone, and PhyloWGS, which were presented and discussed earlier in this chapter. Not all tools were able to produce outputs for all the samples, for different reasons such as insufficient number of SNVs in regions without CNA or LOH events (SciClone) or extremely long runtime (PyClone and PhyloWGS); therefore, the analysis of the results was limited to a subset of 686 samples for which all tools completed successfully. The study

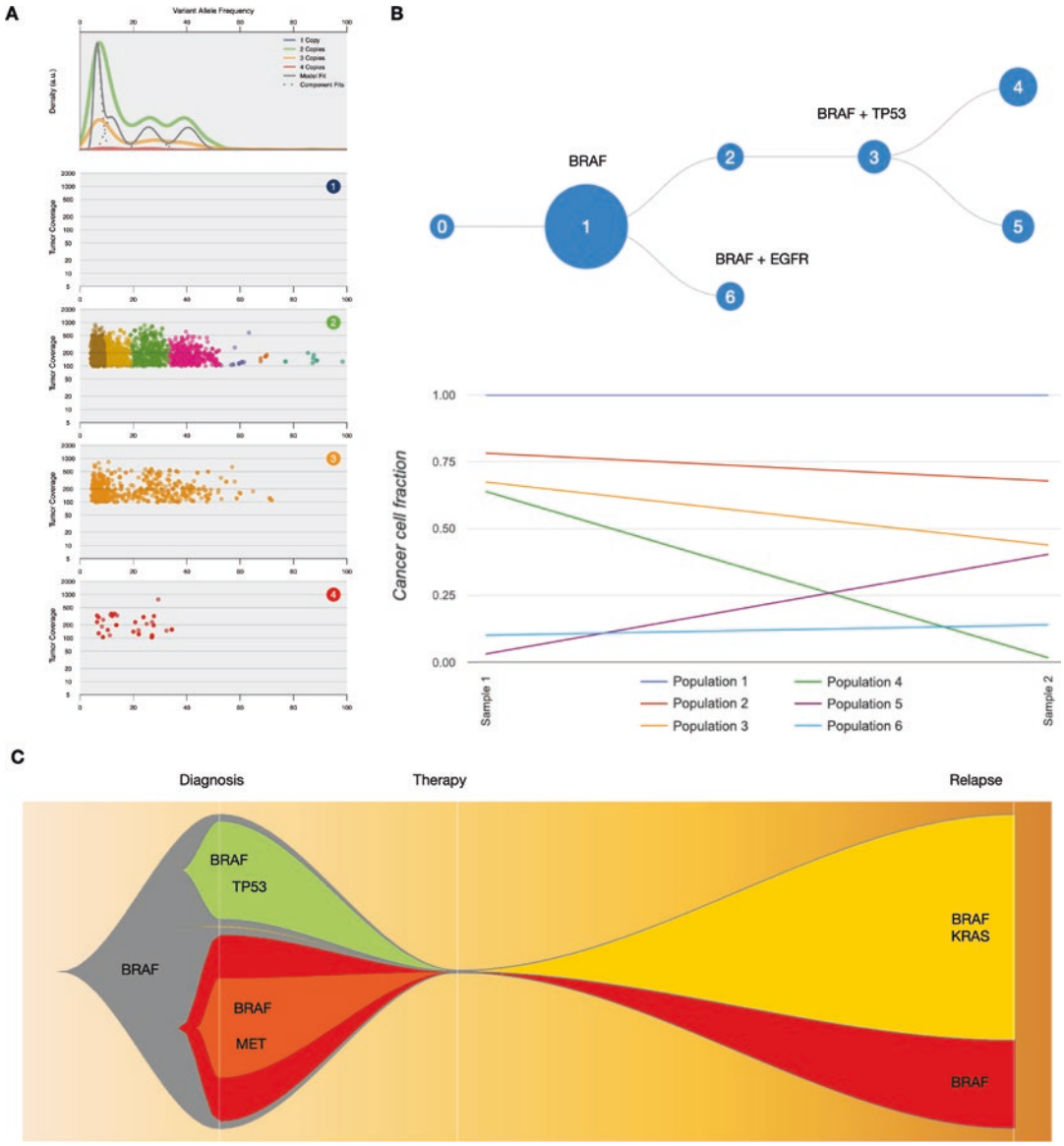

Fig. 6.2 Visualization of ITH and tumor evolution. (**a**) Plots generated by the tool SciClone. The top plot displays the kernel density plot across regions of one to four copies and identifies several different clones, each corresponding to a peak in the model fit. The plots below display mutation clusters as VAFs versus read depth for each of the four copy number regions. (**b**) The upper panel shows a tumor phylogenetic tree generated by PhyloWGS with two samples from the same patient collected at diagnosis and relapse, where each subclones inherit the alterations present in their ancestors and acquire additional ones. The lower panel shows the trajectory of the CCF of subclones from diagnosis to relapse

showed that runtime of PyClone and PhyloWGS increased dramatically with the number of mutations, while SciClone was able to finish in very short time even for heavily mutated tumors. Quantification of ITH was inconsistent across the tools assessed, where the number of clonal popu-

lations identified in the same samples were rather different between tools, in some cases even negatively correlated. Furthermore, there was limited prognostic value of ITH estimates in terms of metrics such as the number of subclonal populations and the evolutionary patterns (early vs. late clonal diversification) and demonstrated no improvement over established clinical factors.

More recently, the ICGC-TCGA DREAM Somatic Mutation Calling Tumor Heterogeneity Challenge (SMC-Het) was developed to address the problem of creating standards for evaluating tumor subclonal reconstruction and to compare the different methods available through a crowdsourced benchmarking effort. The organizers of the challenge generated simulated realistic tumors and developed robust scoring metrics which were then employed to evaluate the methods developed by the challenge participants and provided as re-distributable software containers [49]. SMC-Het consisted of seven subchallenges. Subchallenges 1A, 1B, and 1C evaluated the performances of the algorithms in terms of global characteristics of tumor composition, such as TP, the number of subclonal lineages, and CP. Subchallenges 2A and 2B evaluated hard and soft assignment (i.e., absolute and through probability) of SNVs to subclones. Subchallenges 3A and 3B evaluated the ability of the algorithms to recover the ancestral relationships between subclones. To assess the performance of the algorithms, four metrics were identified: Matthew's correlation coefficient (MCC), area under the precision-recall curve (AUPR), Jensen-Shannon divergence (AJSD), and Pearson's correlation coefficient (PCC). A fifth metric, clonal fraction (CF), was added to subchallenges 3A and 3B, to evaluate the accuracy of the predicted fraction of mutations assigned to the subclones. These metrics were tested by evaluating different mistake scenarios in several tree topologies and compared with the ranking of the scenarios by a panel of nine experts. The challenge was run on simulated tumor data, which was generated using a variant of the tool BAMSurgeon developed specifically for the challenge to create realistic tumors with accurate SNVs, indels, large-scale allele-specific copy number changes, translocations, and other cancer genome features [61]. The top-performing algorithm of SMC-Het was FastClone, described in section "FastClone" of this chapter [48]. The metrics and evaluation model developed for SMC-Het provide a basis for the establishment of gold-standard methods for the analysis of ITH and represent a useful resource for the evaluation of novel methodologies for subclonal reconstruction.

Conclusions and Perspectives

Despite over 10 years of research and a plethora of computational tools developed to measure, quantify, and dissect ITH from bulk DNA sequencing, the remarkable differences observed between the results generated by different methods indicate that there is still a clear need for improvement. While many of the approaches discussed in this chapter have been successfully used to dissect and describe important features in the spatial and temporal subclonal evolution of several cancers, as well as the impact of clonality on therapy resistance, their application in the context of a precision oncology pipeline remains a challenge. Nevertheless, having information regarding the clonal or subclonal nature of variants and being able to attribute variants to different subclones has the potential to significantly impact patient care by enabling the design of clone-aware therapies that may delay or even avoid the development of drug resistance. Therefore, the incorporation of one or more tools for subclonal reconstruction in a clinical decision-making pipeline and the critical interpretation of their outputs may add an important layer of information to a patient's profile and help improve outcomes. Novel approaches based on single-cell sequencing, where clonality can be directly measured and dissected rather than imputed, are likely to take over and become a standard analytical tool in a not so far future, as technology improves and costs decrease. In the meantime, as clinical sequencing is increasingly adopted as a routine standard of care, the computational approaches described in this chapter and their future iterations will likely be an important addi-

tion to the arsenal of tools for genomic profiling and personalized treatment of cancer patients.

References

1. McGranahan N, Swanton C. Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. Cancer Cell. 2015;28:141.
2. Rosenthal R, McGranahan N, Herrero J, Swanton C. Deciphering genetic intratumor heterogeneity and its impact on cancer evolution. Annu Rev Cancer Biol. 2017;1:223–40.
3. Hobor S, et al. TGFα and amphiregulin paracrine network promotes resistance to EGFR blockade in colorectal cancer cells. Clin Cancer Res. 2014;20:6429–38.
4. Keats JJ, et al. Clonal competition with alternating dominance in multiple myeloma. Blood. 2012;120:1067–76.
5. Marusyk A, et al. Non-cell-autonomous driving of tumour growth supports sub-clonal heterogeneity. Nature. 2014;514:54–8.
6. Landau DA, et al. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell. 2013;152:714–26.
7. Laganà A, et al. Integrative network analysis identifies novel drivers of pathogenesis and progression in newly diagnosed multiple myeloma. Leukemia. 2018;32:120–30.
8. Karlsson J, et al. Four evolutionary trajectories underlie genetic intratumoral variation in childhood cancer. Nat Genet. 2018;50:944–50.
9. Gerlinger M, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med. 2012;366:883–92.
10. Laganà A, et al. Precision medicine for relapsed multiple myeloma on the basis of an integrative multiomics approach. JCO Precis Oncol. 2018;2018:1–17.
11. Malone ER, Oliva M, Sabatini PJB, Stockley TL, Siu LL. Molecular profiling for precision cancer therapies. Genome Med. 2020;12:8.
12. Uzilov AV, et al. Development and clinical application of an integrative genomic approach to personalized cancer therapy. Genome Med. 2016;8:1–20.
13. Bode AM, Dong Z. Precision oncology- the future of personalized cancer medicine? NPJ Precis Oncol. 2017;1:2.
14. Remon J, Dienstmann R. Precision oncology: separating the wheat from the chaff. ESMO Open. 2018;3:e000446.
15. Li X, Warner JL. A review of precision oncology knowledgebases for determining the clinical actionability of genetic variants. Front Cell Dev Biol. 2020;8:48.
16. Warburton L, et al. Stopping targeted therapy for complete responders in advanced BRAF mutant melanoma. Sci Rep. 2020;10:18878.
17. Marusyk A, Janiszewska M, Polyak K. Intratumor heterogeneity: the Rosetta stone of therapy resistance. Cancer Cell. 2020;37:471–84.
18. Rye IH, et al. Intratumor heterogeneity defines treatment-resistant HER2+ breast tumors. Mol Oncol. 2018;12:1838–55.
19. Shi H, et al. Acquired resistance and clonal evolution in melanoma during BRAF inhibitor therapy. Cancer Discov. 2014;4:80–93.
20. Jamal-Hanjani M, et al. Tracking the evolution of non-small-cell lung cancer. N Engl J Med. 2017;376:2109–21.
21. Bailey C, et al. Tracking cancer evolution through the disease course. Cancer Discov. 2021;11:916–32.
22. Jamal-Hanjani M, et al. Tracking genomic cancer evolution for precision medicine: the lung TRACERx study. PLoS Biol. 2014;12:e1001906.
23. Watkins TBK, et al. Pervasive chromosomal instability and karyotype order in tumour evolution. Nature. 2020;587:126–32.
24. TRACERx Renal consortium. TRACERx renal: tracking renal cancer evolution through therapy. Nat Rev Urol. 2017;14:575–6.
25. Turajlic S, et al. Tracking cancer evolution reveals constrained routes to metastases: TRACERx renal. Cell. 2018;173:581–94, e12.
26. Turajlic S, et al. Deterministic evolutionary trajectories influence primary tumor growth: TRACERx renal. Cell. 2018;173:595–610, e11.
27. Dentro SC, Wedge DC, Van Loo P. Principles of reconstructing the subclonal architecture of cancers. Cold Spring Harb Perspect Med. 2017;7:a026625.
28. Tarabichi M, et al. A practical guide to cancer subclonal reconstruction from DNA sequencing. Nat Methods. 2021;18:144–55.
29. Vandin F. Computational methods for characterizing cancer mutational heterogeneity. Front Genet. 2017;8:83.
30. Nik-Zainal S, et al. The life history of 21 breast cancers. Cell. 2015;162:924.
31. Roth A, et al. PyClone: statistical inference of clonal population structure in cancer. Nat Methods. 2014;11:396–8.
32. Gillis S, Roth A. PyClone-VI: scalable inference of clonal population structures using whole genome data. BMC Bioinformatics. 2020;21:571.
33. Miller CA, et al. SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. PLoS Comput Biol. 2014;10:e1003665.
34. Fischer A, Vázquez-García I, Illingworth CJR, Mustonen V. High-definition reconstruction of clonal composition in cancer. Cell Rep. 2014;7:1740–52.
35. Deshwar AG, et al. PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. Genome Biol. 2015;16:35.
36. El-Kebir M, Satas G, Oesper L, Raphael BJ. Inferring the mutational history of a tumor using multi-state perfect phylogeny mixtures. Cell Syst. 2016;3:43–53.

37. Jiang Y, Qiu Y, Minn AJ, Zhang NR. Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. Proc Natl Acad Sci U S A. 2016;113:E5528–37.
38. Shinde J, et al. Palimpsest: an R package for studying mutational and structural variant signatures along clonal evolution in cancer. Bioinformatics. 2018;34:3380–1.
39. Deveau P, et al. QuantumClone: clonal assessment of functional mutations in cancer based on a genotype-aware method for clonal reconstruction. Bioinformatics. 2018;34:1808–16.
40. Eaton J, Wang J, Schwartz R. Deconvolution and phylogeny inference of structural variations in tumor genomic samples. Bioinformatics. 2018;34:i357–65.
41. Li Y, Zhou S, Schwartz DC, Ma J. Allele-specific quantification of structural variations in cancer genomes. Cell Syst. 2016;3:21–34.
42. Cmero M, et al. Inferring structural variant cancer cell fraction. Nat Commun. 2020;11:730.
43. Schröder J, et al. Socrates: identification of genomic rearrangements in tumour genomes by re-aligning soft clipped reads. Bioinformatics. 2014;30:1064–72.
44. Myers MA, Satas G, Raphael BJ. CALDER: inferring phylogenetic trees from longitudinal tumor samples. Cell Syst. 2019;8:514–22, e5.
45. Ellson J, Gansner E, Koutsofios L, North SC, Woodhull G. Graphviz—open source graph drawing tools. In: Graph drawing. Berlin Heidelberg: Springer; 2002. p. 483–4.
46. Ricketts C, et al. Meltos: multi-sample tumor phylogeny reconstruction for structural variants. Bioinformatics. 2020;36:1082–90.
47. Ricketts C, Popic V, Toosi H, Hajirasouliha I. Using LICHeE and BAMSE for reconstructing cancer phylogenetic trees. Curr Protoc Bioinformatics. 2018;62:e49.
48. Xiao Y, et al. FastClone is a probabilistic tool for deconvoluting tumor heterogeneity in bulk-sequencing samples. Nat Commun. 2020;11:4469.
49. Salcedo A, et al. A community effort to create standards for evaluating tumor subclonal reconstruction. Nat Biotechnol. 2020;38:97–107.
50. Flensburg C, Sargeant T, Oshlack A, Majewski IJ. SuperFreq: integrated mutation detection and clonal tracking in cancer. PLoS Comput Biol. 2020;16:e1007603.
51. Li H, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.
52. Koboldt DC, Larson DE, Wilson RK. Using VarScan 2 for germline variant calling and somatic mutation detection. Curr Protoc Bioinformatics. 2013;44(1):15-4.
53. Sundermann LK, Wintersinger J, Rätsch G, Stoye J, Morris Q. Reconstructing tumor evolutionary histories and clone trees in polynomial-time with SubMARine. PLoS Comput Biol. 2021;17:e1008400.
54. Avila Cobos F, Vandesompele J, Mestdagh P, De Preter K. Computational deconvolution of transcriptomics data from mixed cell populations. Bioinformatics. 2018;34:1969–79.
55. Park Y, Lim S, Nam J-W, Kim S. Measuring intratumor heterogeneity by network entropy using RNA-seq data. Sci Rep. 2016;6:37767.
56. Li M, Zhang Z, Li L, Wang X. An algorithm to quantify intratumor heterogeneity based on alterations of gene expression profiles. Commun Biol. 2020;3:505.
57. Krzywinski M. Visualizing clonal evolution in cancer. Mol Cell. 2016;62:652–6.
58. Miller CA, et al. Visualizing tumor evolution with the fishplot package for R. BMC Genomics. 2016;17:880.
59. Dang HX, et al. ClonEvol: clonal ordering and visualization in cancer sequencing. Ann Oncol. 2017;28:3076–82.
60. Abécassis J, et al. Assessing reliability of intratumor heterogeneity estimates from single sample whole exome sequencing data. PLoS One. 2019;14:e0224143.
61. Ewing AD, et al. Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. Nat Methods. 2015;12:623–30.

Computational Methods for Drug Repurposing

7

Rosaria Valentina Rapicavoli, Salvatore Alaimo, Alfredo Ferro, and Alfredo Pulvirenti

Abstract

The wealth of knowledge and multi-omics data available in drug research has allowed the rise of several computational methods in the drug discovery field, resulting in a novel and exciting strategy called drug repurposing. Drug repurposing consists in finding new applications for existing drugs. Numerous computational methods perform a high-level integration of different knowledge sources to facilitate the discovery of unknown mechanisms. In this chapter, we present a survey of data resources and computational tools available for drug repositioning.

R. V. Rapicavoli
Department of Physics and Astronomy, University of Catania, Catania, Italy

Department of Clinical and Experimental Medicine, Bioinformatics Unit, University of Catania, Catania, Italy

S. Alaimo · A. Ferro · A. Pulvirenti (✉)
Department of Clinical and Experimental Medicine, Bioinformatics Unit, University of Catania, Catania, Italy
e-mail: alfredo.pulvirenti@unict.it

Introduction

Systematic drug repurposing, also known as drug repositioning, is the re-evaluation of known, pharmaceutically relevant compounds to identify new therapeutic applications.

Finding alternative uses for old drugs has the advantage of optimizing the discovery and development research process, yielding high cost and time savings in drug development. Since in vitro and in vivo screening, chemical optimization, toxicology, mass production, and clinical trials have already been completed and can be bypassed, substantial risks and "overheads" are removed from the path to market (these are known as Bioavailability and Absorption, Distribution, Metabolism, Excretion and Toxicity—ADMET profiles) [1].

Ideal drug candidates for repurposing are those that have passed Phase III, in terms of the American Food and Drug Administration (FDA) system, as this implies that they have proven to be effective in larger populations and verified to be safe. In this way, clinical trials can proceed at a much faster rate [1].

A repurposed drug does not need the initial 6–9 (or more) years, neither 2–3 billion dollars typically required for new drug development [2, 3], but it will proceed directly to preclinical testing and clinical trials, resulting in reduced risks and costs.

© Springer Nature Switzerland AG 2022

A. Laganà (ed.), *Computational Methods for Precision Oncology*, Advances in Experimental Medicine and Biology 1361, https://doi.org/10.1007/978-3-030-91836-1_7

Among the best-known examples, sildenafil citrate (brand name: Viagra) [3] has been repurposed from a common hypertension drug to therapy for erectile dysfunction.

There are many successes in repositioning old drugs, and what was initially driven by serendipity is now operated by focused and systematic computational explorations that precede shorter experimental design cycles [1].

In a world where thousands of therapeutic molecules are known, drug repositioning is becoming an attractive form of drug discovery with a significant impact on personalized medicine.

Customizing or optimizing repositioning methods into efficient drug repositioning pipelines requires a comprehensive understanding of the available methods obtained by evaluating both biological and pharmaceutical knowledge and the mechanism of action of drugs [3].

In addition, the advent of high-throughput technologies to explore biological systems (drug-related data, high-throughput genomic screens, protein structures) resulted in the generation of an impressive amount of data that requires computational analysis and mining tools to be explored and used. Methods and tools available in chemoinformatics, bioinformatics, network biology, and systems biology play a key role in making full use of known targets, drugs, and biomarkers or disease pathways, thus leading to the development of proof-of-concept methods and accelerated timeframe clinical trial design.

Repositioning involves a deep synergy of investigators and computational scientists to develop relevant and realistic exploration tools. However, advanced computational tools are often difficult to understand or use, limiting their accessibility to scientists without a solid computational background [1]. For example, life scientists will find it challenging to use many of the computational tools that require data preparation, installation, and execution of packaged software; computer scientists, on the other hand, will not be able to make experimental validations of predictions.

In this chapter, we describe how to choose a proper drug repositioning approach based on information and knowledge, focusing on prioritizing the methods. Then, we discuss some of the tools built to facilitate the approach to this research field for both life scientists and computer scientists, bridging the gap due to different cultural backgrounds.

Drug Repositioning Methods and Approaches

Over the past few years, the number of drug repositioning methods has increased dramatically. Applying an efficient drug repositioning pipeline to a specific study requires identifying suitable methods based on available information about the drugs or diseases of interest [1, 3]. Therefore, it becomes essential to understand these existing methods better and prioritize them based on specific studies.

Computational drug repositioning methods can be classified as target-based, knowledge-based, signature-based, pathway- or network-based, and mechanism-targeted methods.

According to the information available and the elicited mechanisms, methods can be defined as drug-oriented, disease-oriented, and treatment-oriented. Therefore, these computational drug repositioning methods allow researchers to screen almost any drug candidate and test it on a large number of diseases in a relatively short time [1].

Because repositioning studies are tied to prior knowledge and available information, this will guide the choice of a drug repositioning methodology, and therefore the prioritization: (i) when there is limited information available for the studied disease, phenotypic screening or off-label FDA use would be the best option; (ii) if a protein biomarker exists for the studied disease, target-based or knowledge-based methods should be prioritized; (iii) when disease information is available, knowledge-based or signature-based methods can be used to integrate available disease pathways or disease-related omics data into the drug repositioning process; (iv) when omics data related to drug treatment are available, signature-based or mechanism-targeted methods can be used to elucidate unknown targeted mechanisms, such as off-targets and targeted signaling pathways [1, 3].

Screening Methods or Blinded Research

Blind drug repositioning methods mainly depend on serendipitous identification from targeted disease- and drug-specific assays and do not involve pharmaceutical or biological information. These methods can be applied to a large number of drugs or diseases [3].

Target-Based Methods

Target-based drug repurposing methods involve in vitro and in vivo high-throughput or high-content screening (HTS/HCS) of drugs for a protein or biomarker of interest and an in silico screening of drugs or compounds from drug libraries. The use of target information in drug repurposing ensures a greater chance of finding valuable drugs than blind methods. These methods allow researchers to screen almost any drug or compound with known chemical structure information within days [3].

Knowledge-Based Methods

Knowledge-based drug repositioning methods apply bioinformatics or cheminformatics approaches to integrate available drug information, drug-target networks, chemical structures of targets and drugs, clinical trial information (including adverse effects), FDA approval labels, and signaling or metabolic pathways. This knowledge is then used to predict unknown mechanisms, unknown drug similarities, and new biomarkers for diseases [3].

Signature-Based Methods

These methods rely on the use of gene signatures derived from disease omics data (i.e., microarray, RNA-seq), with or without treatments, to uncover unknown off-targets or unknown disease mechanisms. This type of data is now available on various databases, including NCBI Gene Expression Omnibus (GEO), Connectivity Map (CMap), and Cancer Cell Line Encyclopedia (CCLE). Signature-based methods can support discovering unknown mechanisms of action of molecules and drugs because they are supported by molecular information from which valuable information can be extracted (i.e., differential expression of genes concerning disease or drug administration). This method is advantageous when, for example, drugs need to be repurposed for a large number of diseases. Since the required knowledge (biomarkers, targets) may not be available or may be difficult to derive from available literature or databases, deriving gene signatures for those diseases from publicly available genomic data becomes the best option [3].

Pathway- or Network-Based Methods

Pathway or network-based drug repositioning methods use available disease omics data, signaling or metabolic pathways, and protein interaction networks to reconstruct disease-specific pathways that provide key targets for repositioned drugs. These methods are beneficial in identifying, within extensive pathways, subnetworks, or a small number of crucial, targetable proteins [3].

Targeted Mechanism-Based Methods

Targeted mechanism-based methods use treatment omics data, known signaling pathway information, or protein interaction networks to delineate unknown drug action mechanisms. The application of these approaches involves the use of sophisticated computational models that are characteristic of Systems Biology. Such models find vast space in the era of precision medicine and can also be a valuable support in clinical practice [3]. One potential application is studying the molecular mechanisms that lead cancer patients to drug resistance after a few months of treatment [3]. The methods described above

demonstrate that the success of drug repositioning is closely related to the complexity and richness of the available information [3].

Drug Repurposing Tools: Web-Based Solutions

The field of drug repositioning requires the close collaboration of scientists belonging to different fields. Life scientists, experimental and clinical scientists, evaluate and interpret data and results. Computer scientists and bioinformaticians provide powerful computational software and systems to model the intrinsic complexity of biological models and make predictions to acquire novel knowledge.

The correct use of this type of software may be complex when the appropriate bioinformatics skills are lacking. For this reason, in the last few years, several tools available on the web have emerged. These provide easy-to-use computational solutions to bridge the gap between wet-lab scientists and the software tools available for drug discovery.

It is possible to consider three main categories of web-based platforms that help in drug repurposing based on the type of interaction used to perform repositioning studies: predicting drug-target interactions and using drug-induced gene expression to predict new connections and link drugs to disease.

Web-Based Tools: Predicting Drug-Target Interactions

Within this category, the various tools are classified into five subcategories based on the data used to do repositioning and how they are parameterized [1]:

1. Ligand similarity using fingerprint encoding
2. 3D structures of drugs and targets
3. Network-based approaches
4. Binding site parameterization
5. Other

Ligand Similarity Using Fingerprint Encoding

The paradigm underlying ligand-centered predictions is that the structural similarity implies comparable biological functions or properties. Similar compounds will therefore be likely to bind the same target, which is why a priori knowledge of query-binding targets is used to uncover previously unknown leads. In this sense, it becomes essential to know the fingerprint of molecules, be it 1D, 2D, or 3D.

Some tools belonging to this subcategory are indicated below.

ChemMapper

To find similar molecules and target annotations to identify candidate targets for a given query, ChemMapper uses a 3D similarity algorithm called SHAFTS (SHApe-FeaTure Similarity). The usage of 3D similarity metrics has been shown to improve off-target prediction accuracy [4].

SHAFTS relies on a triplet hashing technique for rapid alignment of molecular conformations and uses shape and chemotype to assess similarity [4].

ChemMapper uses drug information and target annotations from various sources such as ChEMBL, DrugBank, BindingDB, KEGG, and the Protein Database (PDB) [4].

ChemMapper offers the possibility to choose the most appropriate application depending on the final goal of the user (list of plausible proteins, related compounds) and the type of input available (protein ID, protein sequence, list of compounds) [4].

ChemProt 3.0

ChemProt 3.0 is a publicly available collection of chemical-protein disease annotation resources enabling the study of systems pharmacology for a small molecule at different levels of complexity (from molecular to clinical level) [5].

The platform allows users to navigate various data and make assessments from the global scale to specific analyses.

ChemProt 3.0 includes several computational approaches: Similarity Ensemble Approach—SEA, Quantitative Structure-Activity Relationship—QSAR, and network biology-based enrichment analysis [5].

These approaches support generating new hypotheses for bioactivity of novel and already annotated compounds and identifying other genes that may play significant roles in modulating chemical perturbations in humans [5].

The user can search for information about a compound, protein, or clinical outcome or can choose to perform a QSAR prediction for a specific compound. Each molecule can be imported as a SMILES (Simplified Molecular-Input Line-Entry System) code, or it can be drawn or uploaded from a compound structure file via the SD file format [5].

Through the "Heatmap" feature, ChemProt 3.0 allows the user to have a general overview of chemical-protein interactions, providing a global map linking bioactivities of compounds and proteins based on more than 7 million stored interactions collected from multiple databases annotating compounds, proteins, and diseases. ChemProt has one of the most extensive databases for each category (drugs, proteins, interactions, diseases) [5].

QSAR prediction can have two types of application cases: (i) Comparison of the query molecule with the drug set and thus the map will provide a method to navigate through known interactions. (ii) Prediction of new interactions. In this case, the similarity of the molecule fingerprints is used to generate similar drugs and predict the activity of the new compound [5].

HitPick

HitPick is a web server for identifying hits in high-throughput chemical screenings and predicting their molecular targets. It is currently the only resource that can process hits from chemical biology screening experiments and provide target prediction. Indeed, the user can upload the results of the biological assay [6].

HitPick applies the B-score method for identifying high-quality hits based on a statistical evaluation of many screening parameters and an integrative approach that combines 1-nearest-neighbor (1NN) similarity metrics and Laplacian-modified naïve Bayesian target models to predict the targets of identified hits [6].

Targets are predicted based on 2D molecular fingerprints.

The most similar compound from the compound-target interactions is identified using the pairwise Tanimoto coefficient. A ranking of target predictions will then be performed based on the Laplacian-modified Naive Bayesian method-based score.

iDrug-Target

iDrug-Target comprises four subpredictors: iDrug-GPCR, iDrug-Chl, iDrug-Ezy, and iDrug-NR, focusing, respectively, on the identification of drug interactions with G protein-coupled receptors (iDrug-GPCR), ion channels (iDrug-Chl), enzymes (iDrug-Ezy), and nuclear receptors (iDrug-NR) based on KEGG data. The predictions attempt to avoid oversampling due to non-interacting drug-target pairs. The Neighborhood Cleaning Rule and the Synthetic Minority Over-Sampling Technique are used to eliminate redundant negative samples, and some hypothetical positive samples are also added [7]. The Neighborhood Cleaning Rule (NCL) method is among the most popular under-sampling methods. All the samples of the class of interest are maintained, whereas those from the rest of the data are reduced [8]. On the other hand, Synthetic Minority Over-sampling TEchnique (SMOTE) is an over-sampling method that addresses this problem by creating synthetic minority samples to balance the data set. The minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors [9].

iDrug-Target combines protein sequence encoding, using pseudo amino acid composition, with a 256-component 2D fingerprint representation of the ligand. This molecular signature must also be generated to construct a query. iDrug-Target uses a Support Vector Machine (SVM) to classify inputs as interactive or non-interactive [7].

Polypharmacology Browser—PPB

Merged footprints combining features between different footprints can also be generated. PPB searches through 4613 groups of at least ten annotated targets of bioactive molecules from ChEMBL and returns a list of predicted targets ranked by consensus voting scheme and *p*-value [10].

Targets can be ranked by their *p*-values. Indeed, it was found that the pairwise overlap between high confidence (low *p*-value) targets of different fingerprints was significantly higher than low confidence (high *p*-value) targets. In PPB, the similarity is calculated using city-block distances. This tool reports better performance from fusion and pairwise combination fingerprints than single fingerprints [10].

Similarity Ensemble Approach—SEA

SEA was the first tool that used ligand similarity to cluster proteins. The protein clusters thus formed represented functional themes that were potentially useful in predicting the polypharmacology of ligands [11].

The ligands were grouped according to the minimum coverage tree, while Tanimoto coefficients (TC) were used to determine similarity and Daylight 2D fingerprints. The encoding of the ligands is done through 2D fingerprints. The pipeline suggested by SEA, which leads to the identification of new suggestions of repositioned drugs, ultimately provides for validation with experimental techniques [11].

SuperPred

SuperPred is a prediction web server able to connect the chemical similarity of compounds to drugs with molecular targets and a therapeutic approach based on the principle of similar property [11, 12].

The ligand-target interactions are first aggregated by SuperTarget, ChEMBL, and BindingDB, then the set of ligands is normalized/cleaned using JChem to obtain a single set of ligands [11, 12].

Among them, only molecular targets will be extracted through the use of stringent binding affinity thresholds.

Drug-target prediction is achieved by considering the 2D Tanimoto similarity between a query compound and the ligands associated with their respective targets (target sets) [11, 12].

The specificity of each prediction is done through the calculation of two parameters called Z scores and *E*-values. The *E*-value is used as a threshold value. An *E*-value >1 means that the prediction is random. In order to evaluate the similarity between ligands, a weighting factor is calculated. The use of weight improves the accuracy of the predictions.

Thanks to the presence of these thresholds, SuperPred has a prediction success rate of 94.1% [11, 12].

SwissTargetPrediction

SwissTargetPrediction is a web server that has been online since 2014 [13, 14] and whose rationale is based on the observation that similar bioactive molecules are more likely to share similar targets. Thus, identifying proteins with known ligands similar to the query molecule can predict the targets of a given molecule.

This tool combines 2D and 3D similarity metrics to predict targets of bioactive molecules to improve target prediction accuracy. Query molecules can be inputted either as SMILES or drawn in 2D using the javascript-based molecular editor of ChemAxon. This system uses ChEMBL version 23 (the old version was based on ChEMBL version 16) as a data source. The dataset includes 376,342 unique compounds (580,496 binding activities on 3068 protein targets) [13, 14].

SwissTargetPrediction offers the possibility to perform predictions in different organisms, and mapping predictions by homology within and between different species is enabled for close paralogs and orthologs. The updated version can choose among humans, rats, and mice [13, 14].

The similarity quantification consists of calculating a pairwise comparison of 1D vectors describing molecular structures. The 2D measure uses the Tanimoto coefficient between path-based binary footprints (FP2), while the 3D measure is based on a Manhattan similarity distance between Electroshape 5D float vectors (ES5D) [13, 14].

Targets are prioritized based on a logistic regression of the 2D–3D similarity values [13, 14].

TarPred

TarPred is an online computational model based on a reference library containing 533 individual targets with 179,807 active ligands [13, 15].

Given a query compound, TarPred provides the first 30 ranked interacting targets. For each of them, the structure of the three most similar ligands is displayed, along with the disease indications associated with each target. This information helps understand the mechanisms of action and toxicity of active compounds and may offer new inputs for drug repositioning [13, 15].

To calculate the similarity of the query with the set of drug-related targets, TarPred also uses a combination of ECFP4 (Extended-Connectivity Fingerprints), designed for molecular characterization, similarity searching, and structure-activity modeling, together with the Tanimoto coefficient. The prioritized list of targets produced is closely associated with FDA-approved drugs [13, 15].

Protein sequences that interact with FDA-approved drugs (FDA-approved drug targets) are retrieved from DrugBank and subjected to BLAST against BindingDB proteins [13, 15].

TarPred calculates ECFP4 similarity scores between the query compound and ligand sets, producing a ranked list of targets [13, 15].

TargetHunter

TargetHunter is a web-based tool that uses an algorithm based on the Tanimoto similarity index, called TAMOSIC (Targets Associated with its MOst SImilar Counterparts) [16].

The similarity to the query compound is calculated using three different 2D fingerprints. The targets associated with the first "N" most similar compounds are shown as possible targets. The data for the compounds are retrieved from ChEMBL [16].

3D Structures of Drugs and Targets

Structure-based design is founded on the knowledge of the three-dimensional structure of the molecular target for the drug. The methods to derive the 3D structure are X-ray crystallography and NMR solution. Alternatively, homology models based on related proteins are commonly used.

This type of approach focuses on exploring the similarity of binding sites from PDB crystal structures.

Structure-based design predominantly uses molecular modeling techniques such as docking and pharmacophore models to calculate binding affinities of leads.

From the computational point of view, these techniques are more expensive. In fact, most of the computational research in this area is used to create predictive software rather than building real-time web-based applications.

Below are some web services that use this type of approach.

idTarget

idTarget can predict possible binding targets of a small chemical molecule via a *divide et impera* docking approach combined with scoring functions based on regression analysis and quantum chemical charge models. The affinity profiles of the protein targets are used to provide the confidence levels of the prediction. The *divide et impera* docking approach uses small overlapping grids adaptively constructed to limit the search space, thus achieving better efficiency in terms of time. idTarget performs screening on almost all protein structures deposited in the Protein Data Bank (PDB) [17].

The search engine of the idTarget web server is MEDock, which generates initial docking poses of the small ligand [17].

Protein-Drug Interaction Database (PDID)

PDID can be used to systematically catalog protein-drug interactions and facilitate various studies related to drug polypharmacology and drug repurposing.

PDID queries the binding sites within the PDB's drug-protein complexes based on stringent filters against all other proteins on the PDB to find likely off-targets of the original drugs [18].

PDID uses experimentally curated interactions present in DrugBank, BindingDB, and Protein Data Bank.

PDID is based on nearly 1.1 million all-atom predictions on the entire human structural proteome (10,000 structures for over 3700 proteins) and provides access to all putative targets (between 4444 and 7184, depending on the prediction method used) of several popular drugs. Therefore, it represents a valid starting point for drug repositioning [18].

TARget FIShing DOCKing (TarFisDock)

TarFisDock is a tool created to automatize the procedure of searching for small molecule-protein interactions on an extensive repertoire of protein structures. It provided a database of potential drug targets (PDTDs) containing 698 protein structures covering 15 therapeutic areas and was one of the first online tools to offer a reverse ligand-protein docking program. Reverse ligand-protein docking aims to search for potential protein targets by examining an appropriate protein database [19].

TarFisDock requests as input the small molecule to be tested in standard mol2 format and performs the docking through the DOCK 4.0 algorithm using protein structures present in PDTD. Targets can be provided by the user or retrieved from PDTD. The ligand-protein interaction energy terms of the DOCK program are adopted to classify proteins [19].

Network-Based Approaches

Many databases store annotations on system-wide biological networks, including information on various entities that interact with drugs (e.g., targets). Integrating these types of biological networks can help understand the pharmacological properties of specific molecules and thus in drug repositioning. However, working with this kind of data poses new challenges related to managing multidimensional interaction networks.

BalestraWeb

BalestraWeb is an online service that allows users to make predictions about potential interactions between a chosen drug and target or predict the most likely interaction partners of any drug or target listed in DrugBank. It also enables to perform similarity search between drugs or determine the most similar targets based on their interaction patterns [20]. The system uses active learning (AL) techniques relying on probabilistic matrix factorization (PMF) to calculate the statistical weight of each approved drug for all targets associated with the entire set of approved drugs. The server allows three types of queries to be submitted: drug-target interaction, drug-drug similarity, and target-target similarity [20].

Predictions made by BalestraWeb are not dependent on structural or chemical similarities [20].

CSNAP

CSNAP (Chemical Similarity Network Analysis Pull-down) is a computational tool for target identification based on network similarity.

The method combines chemical similarity networks (CSNs) and chemical consensus that results in chemotype-based subnetworks, which predict targets for a set of drug classes [21].

The compounds and their information (e.g., bioactivity) are stored in databases such as ChEMBL and PubChem. Such compounds are grouped by CSN, and target prediction will be based on a consensus statistic determined by the target frequency shared by the first neighbors centered on the compound in the query. The resulting subnetwork will consist of nodes representing compounds and edges representing similarity [21].

The S score is used to rank the targets of the first neighbor compounds, and the significance of each composite protein pair is calculated using an H score. CSNAP appears to have greater predictive ability than the SEA approach [21].

DASPfind

DASPfind is a web service for identifying novel drug-target interactions using “simple paths” of particular lengths inferred from a heterogeneous graph composed of three types of subgraphs: drug-target interactions, drug-drug similarities, and similarities between drug-protein targets [22].

The various known interactions were extracted from the KEGG BRITE, BRENDA, SuperTarget, and DrugBank databases. The chemical structures of the drugs were extracted from the KEGG LIGAND database, and the similarities between the drugs were calculated using SIMCOMP. Target similarity scores are calculated using a normalized version of the Smith-Waterman algorithm [22].

DASPfind performs best when a subjective test using only the "top 1 candidate" is used [22].

nAnnoLyze

nAnnoLyze is a web-based tool for target identification centered on the hypothesis that structurally similar binding sites associate with similar ligands and is based on network-based comparative docking called Annolyze. nAnnoLyze integrates structural information into a bipartite network of interactions and similarities to predict compound-protein structural interactions on a proteomic scale [23].

This network consists of compounds found in PDB, protein-binding sites from LigBase, human proteome structure from ModBase, and DrugBank compounds [23].

Then, the protein subnetwork is constructed using targets that bind ligands above a threshold of drug similarity. The network is connected using the structural similarity of the binding sites calculated by ProBis. The two subnetworks are joined if a resolved PDB structure validates a known ligand-target interaction. Only proteins that have a resolved 3D structure are used for nAnnolyze predictions [23].

PROMISCUOUS

PROMISCUOUS is one of the first public network-based Web servers for drug repurposing [24, 25]. The network employed consists of nodes representing drugs, proteins, side effects, and edges representing drug-target, drug-drug, target-target, and drug-side effect interactions. The information to support the network comes from publicly available databases such as SuperDrug, DrugBank, ChEMBL, SIDER, TTD, SuperTarget, and SuperPred [24, 25].

In the updated version of the Promiscuous 2.0 Database, the number of drugs and drug-like compounds has been significantly increased from 25,000 to nearly 1 million (side effects ~110,000, drug-target interactions ~3 million), compared to the first version. Promiscuous 2.0 also includes a section devoted to potential treatments for COVID-19 [24, 25].

Promiscuous is an easy-to-use resource that allows users to interactively create complex interaction networks and infer new indications for existing compounds. Users can also submit new molecular structures and be presented with suggested application areas or, vice versa, get potential drug candidates for disease indications of interest [24, 25].

SLAP

Semantic Link Association Prediction (SLAP) is a web-based tool that predicts associations between drugs and targets through semantic database integration and statistical modeling. SLAP predicts associations using "path models," predefined association paradigms that include nodes and edges [26]. These are part of a semantic network constructed using drug-drug and protein-protein similarity and drug-target interactions from Chem2Bio2RDF during semantic annotations from the Chem2Bio2OWL ontology. The drug-target pairs used to construct the association network are taken from DrugBank [26].

SLAP uses the Heap-based Dijkstra algorithm to find the shortest path length between two nodes (shortest path length < 3). The predicted values are associated with a *p*-value, calculated as the sum of the Z-scores of all valid paths between two nodes, that allows their ranking based on significance [26].

Three types of input can be given to SLAP: drug-pair and predict association; targets predicted by drugs and drugs with similar biological function; proteins alone and get associated ligands/for and obtain associated ligands [26].

The performance of SLAP is comparable to SEA for drug-target predictions and CMap for drug-association predictions [26].

STITCH

STITCH, a search tool for interacting chemicals, is a web service focused on providing the user a comprehensive map of drug-target interactions with sophisticated filters and visualization [27–30].

We can consider this platform as an interface that integrates drug-target interaction data resources derived from high-throughput, manually curated database experiments and many predictive algorithms.

STITCH has been updated many times and, over the many years of development, has been connected to many databases such as DrugBank, GLIDA, MATADOR, TTD, CTD, KEGG, PID, Reactome, BioCyc, ChEMBL, PDSP Ki Database, and PDB. Over time, STITCH has also been implemented with automated text mining algorithms that predict interactions based on co-occurrence in PubMed, MEDLINE, and NIH Re-PORTER [27–30].

A confidence score is given for each interaction indicating its level of significance and certainty.

As input, it is possible to give a chemical name, a gene name, a chemical structure, or a protein sequence, from which a network of interactions with related chemicals and proteins will be generated [27–30]. STITCH is a well-established and widely used resource by many research groups that directly use its results [27–30].

DT-Web

DT-Web [31] is a web-based application to the Domain Tuned-Hybrid (DT-Hybrid) [32], which extends a well-established recommendation technique from domain-based knowledge that includes drug and target similarity.

This method, together with domain-specific knowledge expressing drug-target similarity, is used to calculate recommendations for each drug.

DT-Web can consider different matrices as input: known drug-target matrix, drug-drug similarity matrix, and target-target similarity matrix.

The drug-target interactions are taken from DrugBank, and from this data, an adjacency matrix is constructed. The drug-drug similarity is assessed using SIMCOMP, and then a similarity matrix is constructed. The target similarity matrix can be obtained by performing BLAST or using the Smith-Waterman local alignment technique.

Then, using these three matrices, a drug-target interaction network is constructed. Each target is mapped to its Entrez Identifier and annotated with Gene Ontology (GO) terms in this interaction network. For each pair of GO terms, the similarity score is calculated. Therefore, a p-value is calculated to evaluate the association between the predicted and validated targets.

Another potential of DT-Web is that, given a set of candidate disease genes as input, it can predict drug combinations whose targets are at an optimal distance from those genes. DT-Web shows better results than NBI and Hybrid, network-based interaction prediction algorithms.

Searching off-lAbel dRUg aNd NEtwoRk—SAveRUNNER

SAveRUNNER is a freely available network-based algorithm for drug repurposing to detect potential new indications for existing drugs that could be used for other purposes [2, 33].

Starting from a list of drug-target interactions and disease-gene associations, this tool predicts drug-disease associations by computing a new network-based similarity measure that prioritizes associations between drugs and diseases located in the same neighborhoods [2, 33].

The SAveRUNNER pipeline consists of two macro steps [2, 33]:

1. The construction of the proximity-based drug-disease network
2. The construction of a similarity-based bipartite drug-disease network

The construction of the proximity-based drug-disease network comprises three phases:

Computation of network proximity (p) to measure how close the disease and drug modules are in the human interactome. Given two modules T and S that, respectively, represent the drug module, containing all t targets of the drug, and the disease module, comprising all s genes of the disease, we can describe this measure as the average length of the shortest path between the elements of T and S [2, 33].

Computation of z-score proximity and p values. SAveRUNNER calculates z-scores and their *p*-values by building a reference distance distribution corresponding to the expected distance between two randomly selected sets of proteins with the same size and degree distribution as the original sets of disease proteins and drug targets in the human interactome. The procedure is repeated 1000 times, and the z-score and its *p*-value are calculated through the mean and standard deviation of the reference distance distribution [2, 33].

Selection of statistically significant drug-disease associations by filtering *p*-values (generally, *p*-value ≤0.05) [2, 33].

Next, the pipeline involves the construction of a similarity-based bipartite drug-disease network that comprises the following steps:

Computation of Network Similarity

The similarity measure is calculated from the network proximity measure p through the equation

$$\text{Similarity} = \frac{\max(p) - p}{\max(p)}\, p$$
$$= \text{network proximity}.$$

This measure assumes a value between 0 and 1 [2, 33].

Cluster Detection

SAveRUNNER uses a clustering algorithm based on greedy optimization of the modularity network to define drug and disease groups. Each identified cluster is evaluated by the cluster quality score (QC) [2, 33].

Adjustment of Network Similarity

If a drug and a disease are part of the same cluster, the drug can probably be repurposed for the disease. Thus, the drug-disease pair should have a higher similarity [2, 33].

Therefore, the similarity of a drug-disease pair belonging to the same cluster is increased proportionally to the cluster's QC score. On the other hand, if two nodes do not fall into the same cluster, QC is set to zero and the similarity value does not change [2, 33].

Normalization of Network Similarity by Applying a Sigmoid Function

SAveRUNNER outputs a list of predicted and prioritized drug-disease associations in a weighted bipartite network format, in which nodes represent drugs and diseases. A link between a drug and a disease occurs if the corresponding drug targets and disease genes are close in the interactome with a significant *p*-value ($p \leq 0.05$). Their interactions are represented by weighted edges in which the weight corresponds to the adjusted and normalized similarity value [2, 33].

Binding Site Parametrization

Binding sites are structural regions of macromolecules that bind ligands through interactions that are almost always reversible and can often be accompanied by conformational changes in the molecules. These are often conserved regions that can be used to search for other ligand-binding proteins that generally bind to other molecules by exploiting the structural similarity of these binding regions. Below, we explore some of the methods designed to predict targets based on the binding sites of query molecules.

ProBis

The ProBiS-ligands Web server predicts the binding of ligands to a protein structure. Starting with a protein structure or binding site, ProBiS-ligands identify model proteins in the Protein Data Bank (PDB) that share similar binding sites to the query [34].

The algorithm uses the structure and physicochemical properties of the constituent amino acids and their backbones to compare two protein-binding sites [34].

Then, it detects structures sharing similar 3D amino acid motifs to the searched protein within the PDB [34].

ProBiS-Database is a repository of non-redundant-binding sites and associated PDB structures, which is updated weekly. ProBiS can be used through pre-calculated data to get results faster or by starting from scratch by looking for a specific protein [34].

PoSSuM

Pocket Similarity Search using Multiple-sketches, PoSSuM, searches the entire PDB database for binding similarity of all coupling molecules. PoSSuM accepts three types of input: a protein structure; a ligand-binding site; and a ligand [35, 36].

Given a protein query, PoSSuM will search for all known ligand-binding sites with a structure similar to the input. PoSSuM can search for any known ligand-binding site or putative-binding site [35, 36]. It uses a neighbor-searching algorithm called SketchSort. The similarity measure is determined based on cosine similarity and a *p*-value indicating significance [35, 36]. On the other hand, dissimilarity values are given by the mean square deviation [35, 36].

Other Web-Based Tools

This section is dedicated to tools that use disease association-dependent annotations. Disease-based approaches are used when drug pharmacology is not present or not considered.

MeSHDD

MeSHDD is a literature-based repositioning methodology that leverages drug-drug similarity based on the MeSH term co-occurrence [37]. MeSHDD clusters drugs based on disease-centered Medical Subject Heading (MeSH) terms found in the MEDLINE Baseline Repository, which contains manually annotated MeSH terms for over 20 million biomedical articles, to predict shared indications [37].

MeSHDD uses drugs from DrugBank, including manually curated information on approved, investigational, and illicit drugs and their targets, mechanisms of action, and indications. Co-occurrence of drug-MeSH terms is calculated using a hypergeometric *P*-value, followed by a Bonferroni correction [37]. The drug-drug similarity is measured by calculating the bitwise distance from converting *p*-values to a binary representation. Drugs are clustered based on pairwise distances and bootstrap-means clustering techniques (implemented in *R*), and the Jaccard index was used to compare the clustering of various *k*-values [37].

RE:fine Drugs

RE:fine drugs is a freely available interactive dashboard for integrated search and discovery of drug repurposing candidates from GWAS and PheWAS repurposing datasets constructed using previously reported methods in Nature Biotechnology [38].

Given a disease as input to the web server, users receive a list of drugs that can potentially treat that disease [38].

The output predictions are classified as known/discovered if present in DrugBank, strongly supported if present in the NIH clinical trial registry and biomedical literature, probable if the evidence is in the NIH clinical trial registry or biomedical literature, and novel if not present in either [38].

Bayesian ANalysis to Determine Drug Interaction Targets—BANDIT

BANDIT is a machine learning algorithm that uses a Bayesian approach to integrate multiple data types to predict possible interactions with therapeutic effects. The rationale for this approach is integrating multiple data types to significantly improve the accuracy of target prediction [39].

BANDIT integrates over 20,000,000 data points from six distinct data types (drug efficacy, post-treatment transcriptional responses, drug structures, reported adverse effects, bioassay results, and known targets) [39]. The tool is based on a database containing approximately 2000 different drugs with 1670 different known targets and over 100,000 compounds without known targets (orphans) [39].

For each data type, a similarity score is calculated for all drug pairs with known targets. For

each pair, BANDIT converts the similarity score into a likelihood ratio. These ratios are then combined to obtain a total likelihood ratio (TLR) proportional to the probability that two drugs share a target, given all available evidence [39].

The integrative approach of BANDIT can accurately identify drugs that share targets, discern the mechanisms of approved drugs, explain existing but not fully known clinical phenotypes, and repurpose drugs for new therapeutic indications [39]. Finally, BANDIT is a dynamic system that can be continuously updated [39]. BANDIT showed high accuracy in identifying shared target interactions and discovering novel targets for cancer treatment [39]. The use of this tool led to the identification of 14 novel microtubule inhibitors, including 3 with activity on resistant cancer cells [39].

Using Drug-Induced Gene Expression to Predict New Connections and Link Drugs to Disease

Drug-induced gene expression refers to the differential mRNA expression profiles in a cell line before and after drug treatment. This repurposing approach is accomplished by comparing disease-associated expression signatures with these drug-induced expression signatures, looking for drugs that have opposite effects on the disease and may be effective.

CMap

Connectivity Map (CMap) relies on a database of pre- and post-gene expression profiles from cellular samples in response to various types of perturbation, e.g., genetic perturbations in response to drug administration. CMap provides mRNA expression data from DNA microarrays for researchers who want to monitor differential expression to identify drugs that produce reverse signatures to query expression signatures. Connectivities are measured using the Kolmogorov-Smirnov statistical test. To date, CMap has generated a library containing over 1.5M gene expression profiles from ~5000 small molecule compounds and ~3000 gene reagents, tested in multiple cell types [40, 41]. CMap has profoundly impacted therapeutic research and has opened new challenges in scientific investigations in drug repurposing, MoA elucidation, biological understanding, and systems biology [40, 41]. It provides one of the most valuable and direct methods to investigate the alternative therapeutic potential of drugs [40, 41].

DeSigN

DeSigN (differentially expressed gene signatures—inhibitors) associates disease signatures with drug response signatures based on IC50 (quantitative measure of drug efficacy often used to prioritize compounds in vitro) data. Unlike CMap, which uses pre- and post-gene expression profiles, DeSigN uses only baseline gene expression profiles. DeSigN is constructed using GDSC [42].

GoPredict

GoPredict uses gene expression data integrated with heterogeneous public information, such as signaling pathways and drug-target information. It takes gene expression data as input and returns drug predictions as output. The reference databases used in GoPredict are TCGA, KEGGDrug, DrugBank, and Gene Ontology [43].

MANTRA 2.0

MANTRA 2.0 predicts molecular drug targets from gene expression profiles before and after drug perturbation in a collaborative and additive learning environment [44].

An automated pipeline of MANTRA 2.0 transforms the gene expression profiles into a single drug “node” in the network and allows users to explore their neighbors to find new indications and interactions. They calculate a proto-

type ranked list (PRL) for each drug, followed by a method to compare two PRLs using a Gene Set Ensemble Approach (GSEA) based method [44].

NFFinder

NFFinder uses the MARQ method to compare molecular signatures. Performing this analysis requires two sets of expression data, up- and downregulated genes compared to GEO, CMap, and DrugMatrix data [45].

PDOD

The online server Prediction of Drugs with Opposing Effects on Disease Genes—PDOD uses gene expression data and associates to them information regarding "effect-type" and "effect-direction" using pathway information (KEGG) and drug-target information from DrugBank [46]. It uses case/control expression datasets published in GEO to determine which gene expression changes happen due to a specific disease and looks for a drug that can counteract them [46].

To extract the gene signature, PDOD draws differentially expressed genes from the expression data by applying Limma and a function that evaluates the drug-disease score based on the parameterization of relationships [46].

RGES

The Reverse Gene Expression Score—RGES is a system providing a predictive measure on how a given drug could reverse the gene expression profile for a given disease. The principle consists of contrasting overexpressed while increasing weakly expressed ones, thus restoring gene expression to levels closer to normal tissue [33].

First, the computational pipeline needs to compute disease gene expression signatures and drug-induced gene expression signatures [33]. From these two molecular signatures, it can calculate the Reverse gene expression score (RGES) between disease and drug. This score ranges from −1 to 1, and it represents a measure of how much the drug under consideration can counteract the changes in expression due to disease. A low RGES value indicates higher potency to reverse disease gene expression and vice versa [33].

RGES is hence dependent on biological conditions. It is also reported that it is positively correlated with drug efficiency and, therefore, the IC50. RGES could also be used to provide insights into drug candidates' mechanisms [33].

The required data to perform the analysis can be taken from various publicly available databases such as TCGA, which includes gene expression profiles of tissue samples, LINCS, which includes perturbagen-mediated gene expression profiles, ChEMBL, which includes drug activity in cancer cells, and CCLE, which includes gene expression profiles of cancer cells [33].

Thanks to the progressively decreasing cost of many profiling technologies, large volumes of gene expression profiles of drugs in different biological conditions can be produced and made available to apply various drug repositioning and compound screening techniques such as RGES [33].

Data Sources for Drug Repurposing

During the past decade, the rapid collection of genomic data has brought an explosion of new insights into the genetic basis of diseases. It is enough to mention the numerous studies through which the association of gene loci with the risk of developing certain diseases has been discovered or the sequencing of human tumors, thanks to which somatic mutations underlying many types of cancer have been identified.

The acquisition of new knowledge about some disease phenotypes and drug-induced perturbations has increased the interest in new computational methods that can analyze and integrate large amounts of data to uncover new disease targets.

In general, applying these approaches on drug perturbation datasets has helped improve the understanding of the connection between genes, drugs, and diseases, as these methodologies can lead to the generation of novel hypotheses.

Drug repurposing tools: web-based solutions			
Categories and core concepts	Tools	Features	Web link
Ligand similarity using fingerprint encoding Core concept: structural similarity implies comparable biological function or properties	*ChemMapper*	3D similarity algorithm SHAFTS (SHApe-FeaTure Similarity). It uses shape and chemotype to assess similarity	http://lilab.ecust.edu.cn/chemmapper/
	ChemProt 3.0	2D similarity-based algorithm. It includes several computational approaches: Similarity Ensemble Approach—SEA, Quantitative Structure-Activity Relationship—QSAR, and network biology-based enrichment analysis	http://potentia.cbs.dtu.dk/ChemProt/
	HitPick	Prediction based on 2D molecular fingerprints. B-score method based on a statistical evaluation of screening parameters for hits identification and integrative approach combining 1-nearest-neighbor (1NN) similarity metrics and Laplacian-modified naïve Bayesian target models to predict the targets of identified hits	http://mips.helmholtz-muenchen.de/hitpick
	iDrug-Target	2D molecular fingerprint-based approach. It uses Support Vector Machine (SVM) to classify inputs as interactive or non-interactive	http://www.jci-bioinfo.cn/iDrug-Target/
	Polypharmacology browser—PPB	Multi fingerprint-based approach. Similarity is calculated using city-block distances	http://gdbtools.unibe.ch:8080/PPB/
	Similarity ensemble approach—SEA	2D molecular fingerprint-based approach	http://sea.bkslab.org/
	SuperPred	2D molecular fingerprint-based approach (2D Tanimoto strategy)	http://prediction.charite.de
	SwissTargetPrediction	Combination of 2D (2D Tanimoto strategy) and 3D (Manhattan similarity distance) similarity approach. It is possible to perform predictions in different organisms (human, rat, and mouse)	http://www.swisstargetprediction.ch
	TarPred	Combination of ECFP4 (Extended-Connectivity Fingerprints), designed for molecular characterization, similarity searching, and structure-activity modeling, and Tanimoto coefficient (2D fingerprint similarity)	http://www.dddc.ac.cn/tarpred
	TargetHunter	Tanimoto similarity index—TAMOSIC (Targets Associated with its MOst SImilar Counterparts). 2D molecular fingerprint-based approach	http://www.cbligand.org/TargetHunter/
3D structures of drugs and targets knowledge Core concept: knowledge of the three-dimensional structure of the molecular target for the drug	*idTarget*	Divide et impera docking approach in combination with scoring functions based on regression analysis and quantum chemical charge models	http://idtarget.rcas.sinica.edu.tw/
	Protein-Drug Interaction Database—PDID	It is based on all-atom predictions on the entire human structural proteome and provides access to all putative targets (depending on the prediction method used: ILbind, SMAP, and eFindSite) of several popular drugs	http://biomine.ece.ualberta.ca/PDID/
	TARget FIShing DOCKing—TarFisDock	Reverse ligand-protein docking approach. The docking is performed through the DOCK 4.0 algorithm using protein structures present in PDTD	http://www.dddc.ac.cn/tarfisdock/

Drug repurposing tools: web-based solutions			
Categories and core concepts	Tools	Features	Web link
Biological networks *Core concept: integrating these types of biological networks can be of great help in understanding the pharmacological properties of certain molecules and thus in drug repositioning*	*Balestra Web*	It uses active learning (AL) techniques based on probabilistic matrix factorization (PMF) to calculate the statistical weight of approved drugs for all targets associated with the entire set of approved drugs. Predictions made by BalestraWeb do not depend on structural or chemical similarities	http://balestra.csb.pitt.edu/
	CSNAP	Combination of chemical similarity networks (CSNs) and chemical consensus from chemotype-based subnetworks to predict targets for a set of drug classes	https://services.mbi.ucla.edu/CSNAP/index.html
	DASPfind	Network-based approach. Drug similarities are calculated using SIMCOMP. Target similarity is calculated using a normalized version of the Smith-Waterman algorithm	http://www.cbrc.kaust.edu.sa/daspfind/
	nAnnolyze	Network-based comparative docking approach. Structural information, interactions, and similarities are integrated to predict compound-protein structural interactions on a proteomic scale	http://www.marciuslab.org/services/nAnnoLyze
	PROMISCUOUS	Network-based approach wherein nodes represent drugs, proteins, and side effects, and edges represent drug-target, drug-drug, target-target, and drug-side effect interactions	https://bioinformatics.charite.de/promiscuous2/
	SLAP	Semantic Link Association Prediction. Drug-target associations are predicted through semantic database integration and statistical modeling. Semantic network is constructed using Chem2Bio2OWL ontology drug-drug and drug-target interactions using Chem2Bio2RDF and drug-target pairs used to construct the association network from DrugBank	http://chem2bio2rdf.org/slap
	STITCH	It provides a comprehensive map of drug-target interactions. It integrates drug-target interaction data resources derived from high-throughput, manually curated database experiments and predictive algorithms	http://stitch.embl.de/
	DT-Web	Recommendation-based algorithm. Domain-specific knowledge expressing drug-target similarity is used to calculate recommendations for each drug	http://alpha.dmi.unict.it/dtweb/
	SAveRUNNER	Network-based algorithm for drug repurposing. It provides an R code. From a list of drug-target interactions and disease-disease associations requested as input, it predicts drug-disease associations by computing a network-based similarity measure	https://github.com/giuliafiscon/SAveRUNNER.git
Binding site parametrization Core concept: methods designed to predict targets based on the binding sites of query molecules	*ProBis*	It identifies model proteins in the Protein Data Bank (PDB) that share similar binding sites to the query (3D structure). It uses the ProBiS algorithm	http://probis.cmm.ki.si/
	PoSSuM	All-pairs similarity	http://possum.cbrc.jp/PoSSuM/

Drug repurposing tools: web-based solutions			
Categories and core concepts	Tools	Features	Web link
Using drug-induced gene expression to predict new connections Core concept: mRNA expression profiles in a cell line, before and after drug treatment	*Connectivity Map—CMap*	Predictions are based on a database of pre- and post-gene expression profiles from cellular samples in response to various types of perturbation (drug effects)	http://www.broad.mit.edu/cmap
	DeSigN	It associates disease signatures with drug response signatures based on IC50 data. It does not use pre- and post-gene expression profiles but only baseline gene expression profiles	http://design.cancerresearch.my/
	GoPredict	It integrates gene expression data with heterogeneous public information (signaling pathways, drug-target information, etc.).	http://csblcanges.fimm.fi/GOPredict/
	MANTRA 2.0	It uses gene expression profiles before and after drug perturbation to define the drug behavior into the network as a new "node." MANTRA 2.0 allows to explore the network to find new indications and interactions	http://mantra.tigem.it/
	NFFinder	It uses MARQ method to compare molecular signatures	http://nffinder.cnb.csic.es/
	PDOD	Prediction of Drugs with Opposing Effects on Disease Genes. It uses gene expression data (GEO) and associates to them pathway information from KEGG and drug-target information from DrugBank	http://gto.kaist.ac.kr/pdod/index.php/main
	RGES	Reverse Gene Expression Score involves using gene expression data to predict how a given drug might reverse the gene expression profile for a given disease by antagonizing genes that are overexpressed (underexpressed) due to the disease and thereby reverts gene expression closer to normal tissue levels	https://github.com/Bin-Chen-Lab/RGES
Others Core concept: disease association–dependent annotations	*MeSHDD*	Literature-based repositioning methodology	http://apps.chiragjpgroup.org/MeSHDD/
	RE:fine drugs	It integrates GWAS and PheWAS reposition datasets using drug-gene-disease triads	
	BANDIT	Machine learning algorithm using a Bayesian approach to integrate multiple data types to predict to which target (enzyme, receptor, or other) a drug may interact to have its therapeutic effect. The rationale for this approach is that the integration of multiple data types improves the accuracy of target prediction since each data type captures different aspects of a molecule's activity	Not publicly available. Select pieces of custom code can be made available upon request

Machine learning techniques and biomedical text mining approaches have been crucial in discovering hidden relationships between drugs and potential new therapeutic indications.

Systematic collection and analysis of gene expression data from human cell lines before and after drug treatment can be used to identify new opportunities for drug repurposing, discover new mechanisms of action for compounds, make small-molecule mimics of endogenous ligands, and predict side effects of such compounds [47].

This approach was initially enabled by the *Connectivity Map* that contains data on transcriptional responses of human cancer cell lines to various drugs/compounds and other small molecules.

The first version of this database had limitations due to its small scale, leading to the extension of the *Connectivity Map* project through the NIH *Library of Integrated Network–based Cellular Signatures* (LINCS) program. A new approach was introduced to increase the available experimental data. A cheaper technology than the classic RNA-seq, called L1000, was employed. The LINCS-L1000 provides the signatures of ~50 human cell lines in response to ~20,000 drugs (at various concentrations) for a total of over a million experiments [47].

In this section, we will provide an overview of CMap and its evolution LINCS L1000. These "big data" resources provide essential but straightforward platforms for characterizing small molecule–induced changes in gene expression and determining connections, similarities, or dissimilarities among diseases, drugs, genes, and pathways.

CMap

The *Connectivity Map* (CMap), introduced in 2006 by Lamb et al., is a database collecting gene-expression profiles of drug-treated human cell cultures, which has been used for investigation of polypharmacology and drug repurposing.

Gene expression profiles are a series of experiments conducted using a microarray platform (Affymetrix HT_HG_U133 and HG_U133A) and standardized preprocessing (MAS 5.0). Experiments were done on different cell lines at different vehicle concentrations and time points compared to controls [48].

In the original CMap study, the initial reference database (Build 1) included 455 treatment-control pairs, where treatment constitutes a selection of 165 drugs, 42 different concentrations, 2-time points, and four human cell lines (MCF7, PC3, SKMEL5, and HL60). Subsequently, the database was significantly extended (Build 2), adding 1309 drugs with 156 different concentrations for a total of about 7000 gene expression profiles [48].

An "instance identifier" uniquely identifies each instance within the database. Thus, there is an instance representation in the reference database for each drug corresponding to treatment and control conditions [48].

The Connectivity Mapping Methods

CMap's rationale is to use a reference database containing disease-specific gene expression profiles and compare it to the gene signature of a given drug. This approach is aimed to predict potential therapeutic candidate drugs. It also allows the identification of connections between drugs, genes, and diseases.

The CMap workflow comprises an initial query consisting of a set of gene signatures highly representative of a given biological state (e.g., disease). Although there is no definite way to generate the optimal gene signatures, the conventional approach identifies and uses a statistically significant list of differentially expressed genes (DEGs) calculated from disease and control samples. This list of genes will delineate the characteristic phenotype for a particular disease [48].

This kind of approach is platform-independent, allowing users to create query signatures from any gene expression platform [40]. Then, the query is used to interrogate the CMap catalog.

Within the database, each of the signatures consists of a weighted average of the three biological replicate perturbations to mitigate the effects of unrelated replicates or outliers [40].

At this point, a connectivity score with a *p*-value is estimated using a non-parametric

rank-ordered Kolmogorov-Smirnov (KS) test. The "*connectivity score*" is normalized through the random permutation described by Lamb et al., assuming values from 1 to −1 to reflect the closeness between expression profiles [40, 48].

A positive correlation indicates the degree of similarity between a query signature and a perturbation-derived profile after specific treatment, whereas a negative correlation denotes an inverse similarity. These correlations are used to determine how exposure to a particular chemical may mimic or reverse the signature of the biological sample of interest.

A false discovery rate (FDR), which adjusts the *p*-value considering multiple hypothesis testing, and a t-parameter, which compares an observed enrichment score to all others in the database, are also calculated [40]. These metrics allow a comprehensive assessment of the relationship between a query and a perturbation, rather than just sorting by similarity.

Since the methodology behind CMap involves using expression profiles to define molecular signatures, it does not require prior knowledge of the detailed mechanism of action (MoA) or drug targets [40, 48]. This advantage makes it a widely used method in drug discovery and repositioning.

The original CMap database had limited chemical and genetic perturbation data due to the high cost of commercial gene expression microarrays and RNA sequencing (RNA-seq). In addition, the expression profiles looked only at a few cell lines leaving the uncertainty of applicability to other cell lines, animal models, or human systems.

To improve the system and overcome these significant limitations, the same team of researchers developed a new simplified platform called L1000 to facilitate rapid and high-throughput gene expression profiles at a lower cost.

L1000

The L1000 platform, developed at the Broad Institute by the CMap team, is a method to facilitate high-throughput, low-cost gene expression profiling and is suitable for extending CMap at a large scale [40, 48]. The development of this method was part of the NIH LINCS (*Library of Integrated Cellular Signatures*) consortium, which funds the generation of expression profiles across multiple cell types and perturbations. To date, through L1000 technology, over 1 million gene expressions have been profiled and collected.

Its name, L1000, is because it contains several reference transcripts equal to 1000, used to estimate the signature of the whole genome gene expression generated by microarrays. Effectively, the basic idea is that it is possible to capture any cellular state by starting from a certain number of representative transcripts at a low cost.

The authors used a set (12,031) of Affymetrix HGU133A expression profiles available in the Gene Expression Omnibus (GEO) to define the threshold for the number of transcripts. From this analysis, it was estimated that 1000 landmarks were sufficient to recover 82% of the information in the entire transcriptome [40].

The L1000 platform combines ligation-mediated amplification, optically addressed and barcoded microspheres (beads), and a flow cytometric detection system for gene expression signature analysis [40]. The L1000 platform is based on hybridization, making the detection of non-abundant transcripts feasible and with a substantial degree of similarity to the profiles obtained with RNA-seq platforms while bypassing the problem of prohibitive costs inherent in this conventional technique.

CMap and its updated versions provide a hypothesis-generating tool to identify new therapeutic targets (drug repositioning), signaling pathways affected by a compound, and search for new mechanisms of action (MoA), including potential side effects. It allows identifying new or known disease-gene-drug connections, depending on the observed level of changes.

Among the most exciting uses is the functional annotation of previously uncharacterized small molecules. For example, using the new-generation CMAp, a new inhibitor of casein kinase CSNK1A1 (compound BRD-1868) was discovered. CSNK1A1 is a protein essential for

the survival of some myeloid malignancies. It is also implicated in resistance to EGFR inhibitors [40].

To facilitate the fruition and use of this system, a platform called CLUE—CMap Linked User Environment has been developed. It can provide several analyses and allow access to all data at multiple levels of pre-processing via Gene Expression Omnibus (GEO: GSE92742) [40].

The L1000 LINCS currently includes over 1 million gene expression profiles of chemically disrupted human cell lines. Several resources and databases derived from L1000 LINCS data are available, for example, the L1000 Characteristic Direction Signature (L1000CDS2) search engine described below.

L1000CDS2

L1000CDS2 is a web-based search engine software designed to query gene expression signatures versus LINCS data to discover and prioritize small molecules that reverse or mimic the entered gene expression profile [47].

To compute the signatures, the L1000CDS2 uses a multivariate method called the Characteristic Direction (CD). Processing L1000 data with the Characteristic Direction (CD) method significantly improves the signature noise compared to the MODZ method used to calculate L1000 signatures [47]. The L1000CDS2 tool can be applied in many biological and biomedical contexts, improving knowledge extraction from the LINCS L1000 resource.

The L1000CDS2 search engine prioritizes thousands of small molecule signatures and their pairwise combinations predicted to mimic or reverse an input gene expression signature. The L1000CDS2 search engine also predicts drug targets for all small molecules profiled by the L1000 assay [47].

Rather than giving relevance to fold-change and assigning greater weight to single genes that show a big fold-change, the CD method assigns a higher weight to genes that move together in the same direction. Thus, a gene that changes less but "moves" along with a large group of other genes may have more weight than a single gene that has changed more in magnitude [47].

The method first identifies the linear hyperplane that best separates control samples from treatment samples using linear discriminant analysis and then uses the normal to this hyperplane to define the direction of change in expression space for each gene [47]. The CD method is more sensitive in identifying "correct" differentially expressed genes than the other alternative methods [47]. CD L1000 signatures can be accessed through an advanced web-based application called L1000CDS2 [47].

When accessing L1000CDS2, there are five sections on the application's home page [47]: the first section on the left consists of two text boxes to enter up- or downregulated genes. The application also gives the possibility to insert an input signature [47]. In this case, the signature should be pasted in the upregulated gene textbox and expression values. The search can be started by clicking on the "search" button once the text boxes are filled [47].

In the central part of the home page, there is a section dedicated to some examples, a configuration section, a section dedicated to metadata, and a section dedicated to recent searches [47].

Optional parameters provided in the configuration section offer several possibilities to customize a search process. For example, through the *mimic/reverse* cursor, it is possible to look for small molecules that mimic or reverse the input signature. The default search mode is *reverse*. The system also supports searching for paired combinations of small molecules [47].

In the *metadata* section, any metadata associated with the input signature can be entered. In the recent searches section, the last 20 queries are stored and are easily accessible by clicking on each entry [47].

Interestingly, there is a function that allows users to share their input signatures and metadata so that others can query those signatures [47].

After starting the search by clicking the Search button, the first 50 signatures are shown in a table on the results page (14 entries for each page) [47].

Each entry provides seven columns of signature information: rank, score, perturbation, cell

line, dose, time point, and overlap with input [47].

Clicking the overlap button, the overlapped genes (and their values) will be shown in the two text boxes. If the user had given up/down genes as input, then the first box will show the overlapping genes between the up input and the up signature, while the second box will show the overlap between the down input and the down signature. If the input is a signature, then the first box will show the genes with a positive value from the input signature, and the second box will have negative values [47].

It is possible to download all the information about a signature as a JavaScript Object Notation file (JSON) by clicking on the download button. Through the tag button, it is possible to view the inserted metadata [47].

Clicking the diamond icon button, it is possible to execute the enrichment analysis on the substructures of the best classified small molecules. The enrichment analysis results are displayed as a table where each row provides three pieces of information: the substructure, the *p*-value (calculated using Fisher's exact test), and the perturbation count. The substructure is represented as a string in the SMARTS format [47].

The cloud icon is used to download the results in table format to a .csv file. Clicking on the share icon provides a permanent URL that can be used to share the enrichment analysis results through an email, publication, or other documentation [47].

If the user chooses to search for combinations of small molecules, then a table of signature combinations will appear below the table of single perturbation results. Each entry provides information about the identified combinations: rank, synergy score, and combinations [47].

When looking for combinations, L1000CDS2 compares each possible pair among the top 50 matching signatures and calculates the potential synergy between each pair by examining the level of orthogonality. The synergy score is calculated as the combined overlap of the differentially expressed genes of the two drug signatures with the input gene lists [47].

Clicking on a perturbation will highlight that perturbation in the single signature results table so that the user can learn more details about that particular perturbation. Clicking the cloud download button in the upper right will download the combination table in a .csv file [47].

In summary, L1000CDS2 is a computational method that potentially elevates the usefulness of a subset of the newly generated publicly available LINCS-L1000 data set to rapidly prioritize small molecules that could reverse or mimic expression in disease and other biological settings [47].

Thanks to L1000CDS2, kenpaullone has been identified as a small molecule that can potentially interfere with the infectious process caused by Ebola by inhibiting GSK3B. Kenpaullone induces the expression of immune response genes and, as such, is a potential antiviral candidate [47].

Conclusion

In this chapter we have reviewed data resources and computational tools available for drug repositioning with the aim of providing a comprehensive guide for researchers and practitioners interested in such a topic. The survey highlights the content and the limitations of each tool or database and compares their content.

References

1. Sam E, Athri P. Web-based drug repurposing tools: a survey. Brief Bioinform. 2019;20:299–316.
2. Fiscon G, Paci P. SAveRUNNER: an R-based tool for drug repurposing. BMC Bioinformatics. 2021;22:150.
3. Jin G, Wong STC. Toward better drug repositioning: prioritizing and integrating existing methods into efficient pipelines. Drug Discov Today. 2014;19:637–44.
4. Gong J, Cai C, Liu X, Ku X, Jiang H, Gao D, Li H. ChemMapper: a versatile web server for exploring pharmacology and chemical structure association based on molecular 3D similarity method. Bioinformatics. 2013;29:1827–9.
5. Kringelum J, Kjaerulff SK, Brunak S, Lund O, Oprea TI, Taboureau O. ChemProt-3.0: a global chemical biology diseases mapping. Database. 2016;2016:bav123. https://doi.org/10.1093/database/bav123.

6. Liu X, Vogt I, Haque T, Campillos M. HitPick: a web server for hit identification and target prediction of chemical screenings. Bioinformatics. 2013;29:1910–2.
7. Xiao X, Min J-L, Lin W-Z, Liu Z, Cheng X, Chou K-C. iDrug-Target: predicting the interactions between drug compounds and target proteins in cellular networking via benchmark dataset optimization approach. J Biomol Struct Dyn. 2015;33:2221–33.
8. Abdouli NOA, Al Abdouli NO, Aung Z, Woon WL, Svetinovic D. Tackling class imbalance problem in binary classification using augmented neighborhood cleaning algorithm. In: Kim K, editor. Information science and applications. Lecture notes in electrical engineering. Berlin, Heidelberg: Springer; 2015. p. 827–34.
9. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling TEchnique. J Artif Intell Res. 2002;16:321–57.
10. Awale M, Reymond J-L. The polypharmacology browser: a web-based multi-fingerprint target prediction tool using ChEMBL bioactivity data. J Cheminform. 2017;9:11.
11. Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK. Relating protein pharmacology by ligand chemistry. Nat Biotechnol. 2007;25:197–206.
12. Nickel J, Gohlke B-O, Erehman J, Banerjee P, Rong WW, Goede A, Dunkel M, Preissner R. SuperPred: update on drug classification and target prediction. Nucleic Acids Res. 2014;42:W26–31.
13. Gfeller D, Grosdidier A, Wirth M, Daina A, Michielin O, Zoete V. SwissTargetPrediction: a web server for target prediction of bioactive small molecules. Nucleic Acids Res. 2014;42:W32–8.
14. Daina A, Michielin O, Zoete V. SwissTargetPrediction: updated data and new features for efficient prediction of protein targets of small molecules. Nucleic Acids Res. 2019;47:W357–64.
15. Liu X, Gao Y, Peng J, Xu Y, Wang Y, Zhou N, Xing J, Luo X, Jiang H, Zheng M. TarPred: a web application for predicting therapeutic and side effect targets of chemical compounds. Bioinformatics. 2015;31:2049–51.
16. Wang L, Ma C, Wipf P, Liu H, Su W, Xie X-Q. TargetHunter: an in silico target identification tool for predicting therapeutic potential of small organic molecules based on chemogenomic database. AAPS J. 2013;15:395–406.
17. Wang J-C, Chu P-Y, Chen C-M, Lin J-H. idTarget: a web server for identifying protein targets of small chemical molecules with robust scoring functions and a divide-and-conquer docking approach. Nucleic Acids Res. 2012;40:W393–9.
18. Wang C, Hu G, Wang K, Brylinski M, Xie L, Kurgan L. PDID: database of molecular-level putative protein–drug interactions in the structural human proteome. Bioinformatics. 2016;32:579–86.
19. Li H, Gao Z, Kang L, et al. TarFisDock: a web server for identifying drug targets with docking approach. Nucleic Acids Res. 2006;34:W219–24.
20. Cobanoglu MC, Oltvai ZN, Taylor DL, Bahar I. BalestraWeb: efficient online evaluation of drug-target interactions. Bioinformatics. 2015;31:131–3.
21. Lo Y-C, Senese S, Li C-M, Hu Q, Huang Y, Damoiseaux R, Torres JZ. Large-scale chemical similarity networks for target profiling of compounds identified in cell-based chemical screens. PLoS Comput Biol. 2015;11:e1004153.
22. Ba-Alawi W, Soufan O, Essack M, Kalnis P, Bajic VB. DASPfind: new efficient method to predict drug-target interactions. J Cheminform. 2016;8:15.
23. Martínez-Jiménez F, Marti-Renom MA. Ligand-target prediction by structural network biology using nAnnoLyze. PLoS Comput Biol. 2015;11:e1004157.
24. von Eichborn J, Murgueitio MS, Dunkel M, Koerner S, Bourne PE, Preissner R. PROMISCUOUS: a database for network-based drug-repositioning. Nucleic Acids Res. 2011;39:D1060–6.
25. Gallo K, Goede A, Eckert A, Moahamed B, Preissner R, Gohlke B-O. PROMISCUOUS 2.0: a resource for drug-repositioning. Nucleic Acids Res. 2021;49:D1373–80.
26. Chen B, Ding Y, Wild DJ. Assessing drug target association using semantic linked data. PLoS Comput Biol. 2012;8:e1002574.
27. Kuhn M, Szklarczyk D, Pletscher-Frankild S, Blicher TH, von Mering C, Jensen LJ, Bork P. STITCH 4: integration of protein-chemical interactions with user data. Nucleic Acids Res. 2014;42:D401–7.
28. Kuhn M, Szklarczyk D, Franceschini A, von Mering C, Jensen LJ, Bork P. STITCH 3: zooming in on protein-chemical interactions. Nucleic Acids Res. 2012;40:D876–80.
29. Szklarczyk D, Santos A, von Mering C, Jensen LJ, Bork P, Kuhn M. STITCH 5: augmenting protein–chemical interaction networks with tissue and affinity data. Nucleic Acids Res. 2016;44:D380–4.
30. Kuhn M, Szklarczyk D, Franceschini A, Campillos M, von Mering C, Jensen LJ, Beyer A, Bork P. STITCH 2: an interaction network database for small molecules and proteins. Nucleic Acids Res. 2010;38:D552–6.
31. Alaimo S, Bonnici V, Cancemi D, Ferro A, Giugno R, Pulvirenti A. DT-Web: a web-based application for drug-target interaction and drug combination prediction through domain-tuned network-based inference. BMC Syst Biol. 2015;9(Suppl 3):S4.
32. Alaimo S, Pulvirenti A, Giugno R, Ferro A. Drug-target interaction prediction through domain-tuned network-based inference. Bioinformatics. 2013;29:2004–8.
33. Chen B, Ma L, Paik H, Sirota M, Wei W, Chua M-S, So S, Butte AJ. Reversal of cancer gene expression correlates with drug efficacy and reveals therapeutic targets. Nat Commun. 2017;8:16022.
34. Konc J, Janezic D. ProBiS-2012: web server and web services for detection of structurally similar binding sites in proteins. Nucleic Acids Res. 2012;40:W214–21.
35. Ito J-I, Tabei Y, Shimizu K, Tsuda K, Tomii K. PoSSuM: a database of similar protein-ligand binding and putative pockets. Nucleic Acids Res. 2012;40:D541–8.

36. Ito J-I, Ikeda K, Yamada K, Mizuguchi K, Tomii K. PoSSuM v.2.0: data update and a new function for investigating ligand analogs and target proteins of small-molecule drugs. Nucleic Acids Res. 2015;43:D392–8.
37. Brown AS, Patel CJ. MeSHDD: literature-based drug-drug similarity for drug repositioning. J Am Med Inform Assoc. 2017;24:614–8.
38. Moosavinasab S, Patterson J, Strouse R, Rastegar-Mojarad M, Regan K, Payne PRO, Huang Y, Lin SM. "RE:fine drugs": an interactive dashboard to access drug repurposing opportunities. Database. 2016;2016:baw083. https://doi.org/10.1093/database/baw083.
39. Madhukar NS, Khade PK, Huang L, Gayvert K, Galletti G, Stogniew M, Allen JE, Giannakakou P, Elemento O. A Bayesian machine learning approach for drug target identification using diverse data types. Nat Commun. 2019;10:5221.
40. Subramanian A, Narayan R, Corsello SM, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. Cell. 2017;171:1437–1452.e17.
41. Lamb J, Crawford ED, Peck D, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science. 2006;313:1929–35.
42. Lee BKB, Tiong KH, Chang JK, Liew CS, Abdul Rahman ZA, Tan AC, Khang TF, Cheong SC. DeSigN: connecting gene expression with therapeutics for drug repurposing and development. BMC Genomics. 2017;18:934.
43. Louhimo R, Laakso M, Belitskin D, Klefström J, Lehtonen R, Hautaniemi S. Data integration to prioritize drugs using genomics and curated data. BioData Min. 2016;9:21.
44. Carrella D, Napolitano F, Rispoli R, Miglietta M, Carissimo A, Cutillo L, Sirci F, Gregoretti F, Di Bernardo D. Mantra 2.0: an online collaborative resource for drug mode of action and repurposing by network analysis. Bioinformatics. 2014;30:1787–8.
45. Setoain J, Franch M, Martínez M, Tabas-Madrid D, Sorzano COS, Bakker A, Gonzalez-Couto E, Elvira J, Pascual-Montano A. NFFinder: an online bioinformatics tool for searching similar transcriptomics experiments in the context of drug repositioning. Nucleic Acids Res. 2015;43:W193–9.
46. Yu H, Choo S, Park J, Jung J, Kang Y, Lee D. Prediction of drugs having opposite effects on disease genes in a directed network. BMC Syst Biol. 2016;10:S2. https://doi.org/10.1186/s12918-015-0243-2.
47. Duan Q, Reid SP, Clark NR, et al. L1000CDS2: LINCS L1000 characteristic direction signatures search engine. npj Syst Biol Appl. 2016;2:16015. https://doi.org/10.1038/npjsba.2016.15.
48. Musa A, Ghoraie LS, Zhang S-D, Glazko G, Yli-Harja O, Dehmer M, Haibe-Kains B, Emmert-Streib F. A review of connectivity map and computational approaches in pharmacogenomics. Brief Bioinform. 2018;19:506–23.

8

Pathway Analysis for Cancer Research and Precision Oncology Applications

Alessandro La Ferlita, Salvatore Alaimo, Alfredo Ferro, and Alfredo Pulvirenti

Abstract

With the advent of OMICs technologies, several bioinformatics methods have been developed to infer biological knowledge from such data. Pathway analysis methodologies help integrate multi-OMICs data and find altered function in known metabolic and signaling pathways. As widely known, such alterations promote the cancer cells' progression and the maintenance of the malignant state. In this chapter, we provide (i) a comprehensive description of the primary data sources for omics data, cancer "omics" projects, and precision oncology knowledge bases; (ii) a survey of the main biological pathway databases; (iii) and a global view of the principal pathway analysis tools and methodologies, describing their main characteristics and shortcomings highlighting their potential applications in cancer research and precision oncology.

Introduction: From Patients to Pathways

In medicine, due to human physiology complexity, clinicians rarely have enough data to make fully informed decisions. Significant contributors to this complexity are genetic mutations, epigenetic modifications, alteration in gene expression, and metabolite levels. For this reason, multi-omics approaches (e.g., genomics, transcriptomics, proteomics, metabolomics) are playing a pivotal role in modern medicine [1, 2]. Moreover, in diseases like cancer, this knowledge is quite relevant since many diverse alterations can establish abnormal cell growth, leading to radically different treatments.

With the term "multi-omics," we mean using more than one of the current high-throughput biomolecular experimental techniques to characterize biological systems at the phenomenological level [2]. It is well known that every omic contributes in a specific way to describe the biological mechanism underlying the phenotype under study. For this reason, it has become evident that there is a need for novel integrative systems that gather and organize such information into mechanistic or semi-mechanistic descriptions of the biological phenomenon [2]. This issue has been particularly relevant for studying complex phenotypes, such as cancer [2]. Indeed, many factors are involved in developing and maintaining the malignant state of cancer cells,

A. La Ferlita · S. Alaimo · A. Ferro · A. Pulvirenti (✉)
Department of Clinical and Experimental Medicine, Bioinformatics Unit, University of Catania, Catania, Italy
e-mail: alfredo.pulvirenti@unict.it

© Springer Nature Switzerland AG 2022

A. Laganà (ed.), *Computational Methods for Precision Oncology*, Advances in Experimental Medicine and Biology 1361, https://doi.org/10.1007/978-3-030-91836-1_8

such as genetic aberrations, epigenetic alterations, changes in response to cellular signaling, metabolic alterations, and beyond [2]. Hence, by analyzing cancer as a complex pathology, we try to gain insight into the cancer cells' molecular mechanisms by looking at their different components. A strategy to integrate such multi-omics data in an in silico model is presented by pathway analysis approaches.

Pathway analysis is an extensive class of methods that can determine biological processes' status, identifying altered functionalities in complex diseases [3]. Specifically, they use knowledge from pathway databases such as the Kyoto Encyclopedia of Gene and Genomes (KEGG) [4–10], Reactome [11–15], WikiPathways [16–19], and Pathway Commons [20, 21] to identify perturbed pathways associated with a specific phenotype or condition starting from a combination of several different types of omics data such as genomics, transcriptomics, proteomics, and metabolomics data [22]. Indeed, growing pieces of evidence suggest that cancer can be better understood through dysregulated pathways rather than individual mutations [23]. However, biological pathways are still partial and incomplete. Therefore, their annotation with other experimentally validated interactions may increase the reliability of pathway-based analysis methods [24].

In this chapter, we will provide (i) a description of the primary data sources for omics data, cancer "omics" projects, and precision oncology knowledge bases; (ii) a summary of the main biological pathway databases; (iii) and a global view of the principal pathway analysis methodologies, describing their main characteristics and shortcomings highlighting their applications in cancer research and precision oncology.

Omics Data Source for Pathway Analysis

Different types of omics data such as genomics, transcriptomics, proteomics, and metabolomics are recommended to perform pathway analysis. Indeed, each omic contributes to highlighting essential aspects of the development and maintenance of cancer cells' malignant state. Therefore, the availability of such heterogeneous data is crucial for cancer research. Fortunately, the amount of such data in public repositories is rising. However, since NGS technologies are cheap and widely used, genomic and transcriptomic data sources are more widespread than proteomics and metabolomics since they rely on mass spectrometry and Nuclear Magnetic Resonance (NMR) spectroscopy.

In what follows, we survey (i) the central sequencing data repositories for genomics and transcriptomics data; (ii) some of the most widely known repositories for proteomics and metabolomics data; (iii) some cancer "omics" projects which are landmarks for cancer research; (iv) and finally some knowledge bases which are useful for precision oncology applications.

Sequencing Data Repositories

Historically, databases have been pivotal in biology and biomedicine research advancements. Since its creation in 1982, GenBank (previously known as Los Alamos Sequence Database), which is now available through the platform of the National Center for Biotechnology Information (NCBI), has been a seminal resource in its field. After that, a joint effort between NCBI, the European Molecular Biology Laboratory (EMBL), and the DNA Databank of Japan (DDBJ) created the International Nucleotide Sequence Database Collaboration (INSDC) to collect the nucleotide and amino acid sequence data that was becoming available. Since then, the INSDC databases have grown every day, reflecting an exponential growth rate in which the amount of stored data has doubled every 18 months.

A terrific contribution to the growth of such databases has been the advent of NGS technologies, which are exponentially increasing the volume and complexity of these sequence data collections [25]. Indeed, these technologies allow the sequencing of entire genomes in a few days, yielding the possibility to detect gene mutations

or polymorphisms (e.g., CNV, SNPs, INDEL, STR) potentially associated with different diseases [26]. Moreover, NGS technologies are also extensively used for transcriptome profiling (RNA-Seq), allowing the identification of differentially expressed coding-protein genes and non-coding RNAs (ncRNAs), splicing variants, or complex gene rearrangements which could represent driver events in specific diseases [27].

Due to the broad spectrum of biological, biotechnological, and biomedical research applications, NGS data are continuously generated and then discussed and analyzed through scientific papers. Consequently, before the publication of study results, a mandatory step is to share the data produced during the research by submitting raw DNA-Seq and RNA-Seq data into an INSDC database such as NCBI SRA, EBI ENA, and DDBJ DRA (Table 8.1).

The availability of such data is essential to the reproducibility of the results. Furthermore, it enables other researchers to reuse such data, perhaps focusing on different angles. Due to sequencing data's plasticity, they can then be used for projects and purposes different from those designed by the original authors of the data.

Proteomics and Metabolomics Data Repositories

Like sequencing data, proteomics and metabolomics data also have specific repositories. However, compared to other data-intensive disciplines such as genomics and transcriptomics, public resources of mass spectrometry (MS)–based proteomics and metabolomics data are fewer due to the complexity of the data [28]. Several public repositories for MS proteomics and metabolomics experiments have been developed to address this need, each with different purposes. Concerning proteomics, the most established resources are the Global Proteome Machine Database (GPMDB), PeptideAtlas, and PRIDE, while for metabolomics data, we cite MetaboLights and The Human Metabolome Database (HMDB) (Table 8.2).

GPMDB [29] is one of the most well-known protein expression databases. The GPMDB pipeline reprocesses the MS data provided by users or raw data stored in other repositories using the popular open-source search engine X!Tandem [30]. Peptide and protein identifications are generated and stored in XML files indexed in a MySQL database.

PeptideAtlas [31–33] was created to serve as the endpoint for the trans-proteomic pipeline (TPP) processing software [34]. More recently, PeptideAtlas has grown as a data reprocessing resource, and it has served as a research database for the development of spectral libraries [35] and SRM-related tools [36, 37]. Currently, PeptideAtlas is one of the most extensive and well-curated protein expression data resources.

PRIDE [38] was initially developed at the European Bioinformatics Institute (EBI, Cambridge, UK) to store the experimental data included in publications, supporting the manuscript review process. The primary data types stored in PRIDE are peptide/protein identifications, peptide/protein expression values, the analyzed mass spectra (both as raw data and peak lists), and the related technical/biological metadata.

MetaboLights [39] is a database for metabolic experiments recommended by several leading

Table 8.1 Main international sequencing data repositories

Repository name	Link	Raw read file downloadable	Download
Sequencing Read Archive (SRA)	https://www.ncbi.nlm.nih.gov/sra	SRA	Command-line SRA toolkit
European Nucleotide Archive (ENA)	https://www.ebi.ac.uk/ena/browser/home	FASTQ, SRA	ENA browser WEB GUI
DDBJ Sequence Read Archive (DRA)	https://www.ddbj.nig.ac.jp/dra/index-e.html	FASTQ, SRA	DRASearch WEB GUI

Table 8.2 Proteomics and metabolomics data repositories

Repository name	Link	OMICS data	Type of data	Technology
Global Proteome Machine Database (GPMDB)	http://gpmdb.thegpm.org/	Proteomics	Peptide/protein expressions 2D page blots Technical/biological metadata GO analysis results	Mass spectrometer
PeptideAtlas	http://www.peptideatlas.org/	Proteomics	Peptide identification, peptide/protein expressions	Mass spectrometer
PRIDE	https://www.ebi.ac.uk/pride/	Proteomics	Experimental proteomics data included in the publication. The downloadable data include technical/biological metadata, raw and analyzed mass spectra, and peptide/protein expressions	Any high-throughput proteomics technology
MetaboLights	https://www.ebi.ac.uk/metabolights/index	Metabolomics	Experimental metabolomics data included in the publication. The downloadable data are technical/biological metadata, raw and analyzed mass spectra, and metabolite concentrations	Mass spectrometer, NMR spectroscopy
The Human Metabolome Database (HMDB)	https://hmdb.ca/	Metabolomics	MetaboCards containing several information for each metabolite found in the human body	Mass spectrometer, NMR spectroscopy

journals to store experimental data included in publications. The database is cross-species, cross-technique, and covers metabolite structures and their reference spectra and their biological roles, locations, concentrations, and experimental data from metabolic experiments.

HMDB [40] is a freely available database containing detailed information (e.g., chemical data, clinical data, and biochemistry data) about metabolites found in the human body. Each entry has a MetaboCard containing more than 100 data fields, with 2/3 of the information being devoted to chemical/clinical data and the other 1/3 devoted to enzymatic or biochemical data. Many data fields are linked to other databases (KEGG, PubChem, MetaCyc, ChEBI, PDB, Swiss-Prot, and GenBank) and various structure and pathway viewing apps.

Cancer "Omics" Projects Data

Thanks to the advent of these novel omics technologies, several genomics, and transcriptomics cancer-related projects were started. Precisely, some of them were involved in the molecular characterization of the widely used cancer cell lines. In contrast, others were involved in the characterization of primary tumor samples and new cancer models. As a result of these initiatives, several petabytes of OMICs data are now publicly available to help scientists study different cancer biology aspects. In what follows, we survey some of the most known cancer "omics" projects highlighting the type of data produced by them (Table 8.3).

The Cancer Cell Line Encyclopedia (CCLE) [41, 42] is a database consisting of a detailed genetic and pharmacologic characterization of a large panel of human cancer models (over 1100 cell lines). The goal was to develop integrated computational analyses linking distinct pharmacologic vulnerabilities to genomic patterns and translating cell line integrative genomics into cancer patient stratification [41, 42]. The CCLE portal (https://portals.broadinstitute.org/ccle) provides public access to these datasets: (1) copy number variation, (2) mRNA expression (Affy and RNAseq), (3) reverse phase protein array (RPPA), and (4) reduced-representation bisulfite sequencing (RRBS). Raw sequencing data such as whole genome sequencing (WGS), whole exome sequencing (WXS), and RNA-Seq can also be freely downloaded from GDC Legacy

Table 8.3 Cancer omics data repositories

Project name	Homepage of the project	Sample details	Type of OMICs data available	Availability
The Cancer Cell Line Encyclopedia (CCLE)	https://portals.broadinstitute.org/ccle	1457 different cancer cell lines which span several cancer types	WGS WXS RNA-Seq	BAM files are publicly available without authorization in GDC Legacy Archive https://portal.gdc.cancer.gov/legacy-archive/search/f. Raw sequencing file can also be found in NCBI SRA and EBI ENA with this accession number PRJNA523380
The Cancer Genome Atlas (TCGA)	https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga	20,000 primary cancers and matched normal samples spanning 33 different cancer types	WGS WXS RNA-Seq miRNA-Seq ATAC-Seq Genotyping array Methylation array	Publicly available after authorization in GDC Data Portal https://portal.gdc.cancer.gov/. Processed data do not need authorization
MMRF CoMMpass Study	https://research.themmrf.org/	Plasma cells of more than 1100 patients with newly diagnosed active myeloma tracked over a period of 8 years	WGS WXS RNA-Seq	Publicly available in GDC Data Portal https://portal.gdc.cancer.gov/. Processed data do not need authorization
The Therapeutically Applicable Research to Generate Effective Treatments (TARGET)	https://ocg.cancer.gov/programs/target#:~:text=The%20Therapeutically%20Applicable%20Research%20to,of%20effective%2C%20less%20toxic%20therapies	More than 6000 cancer samples which cover acute lymphoblastic leukemia, acute myeloid leukemia, kidney tumors, neuroblastoma, and osteosarcoma (OS)	WGS WXS RNA-Seq miRNA-Seq Genotyping array	Publicly available in GDC Data Portal https://portal.gdc.cancer.gov/. Processed data do not need authorization
The Human Cancer Models Initiative (HCMI)	https://ocg.cancer.gov/programs/HCMI#:~:text=The%20Human%20Cancer%20Models%20Initiative,available%20as%20a%20community%20resource	More than 1000 next-generation cancer models generated from several cancer types such as breast, colorectal, glioblastoma, gastroesophageal, lung, melanoma, pancreas, neuroblastoma, osteosarcoma, Wilms tumor, rhabdomyosarcoma, and Ewing sarcoma	WGS WXS RNA-Seq	Publicly available in GDC Data Portal https://portal.gdc.cancer.gov/. Processed data do not need authorization

Archive (https://portal.gdc.cancer.gov/legacy-archive/search/f) or NCBI SRA/EBI ENA databases (accession number PRJNA523380).

The Cancer Genome Atlas (TCGA) is a landmark cancer genomics program designed to molecularly characterize over 20,000 primary cancers and matched normal samples spanning 33 different cancer types [43, 44]. The goal was to apply high-throughput analysis techniques to improve the ability to diagnose, treat, and prevent cancer by better understanding this disease's genetic basis [43, 44]. TCGA generated over 2.5 petabytes of OMICs data publicly available from the NCI GDC data portal (https://portal.gdc.cancer.gov/). TCGA is supervised by the National Cancer Institute's Center for Cancer Genomics and the National Human Genome Research Institute, funded by the US government.

MMRF CoMMpass Study is a significant research project in which more than 1100 patients with newly diagnosed active myeloma are being tracked for 8 years. This study collects information about clinical data, treatments and responses, quality of life data, and cytogenetics immunophenotyping. Blood and bone marrow samples were collected from patients and subjected to WGS, WXS, and RNA-Seq. Notably, the samples were taken from patients at three different times (1) when they entered the study, (2) when they responded to treatment, and (3) when they had a relapse (https://themmrf.org/finding-a-cure/our-work/the-mmrf-commpass-study/). Sequencing data can be downloaded from the NCI GDC data portal (https://portal.gdc.cancer.gov/) and NCBI dbGaP (https://www.ncbi.nlm.nih.gov/gap/. ID of the project: phs000748).

The Therapeutically Applicable Research to Generate Effective Treatments (TARGET) program applies a comprehensive genomic approach to determine molecular changes that drive childhood cancers [45, 46]. The program's goal was to guide the development of effective and less toxic therapies by generating data available to the research community, enabling the identification of therapeutic targets and prognostic markers for novel and more effective treatment strategies. The TARGET initiative originated with two pilot projects characterizing the genomes and transcriptomes of "high-risk" subtypes of acute lymphoblastic leukemia (ALL) [47–52] and neuroblastoma (NBL) [53]. The two pilot project teams' success allowed TARGET to expand its efforts by incorporating additional childhood cancers [54–58]. To date, TARGET researchers have molecularly characterized subtypes of acute myeloid leukemia, osteosarcoma, and select kidney tumors, and additional subtypes of ALL and NBL. Sequencing data are available from such tumors, such as WGS, WXS, and RNA-Seq (mRNA-Seq and miRNA-Seq). All these data are stored in the NCI GDC data portal (https://portal.gdc.cancer.gov/).

The Human Cancer Models Initiative (HCMI) is a project aimed at creating up to 1000 next-generation cancer models (e.g., organoids, conditionally reprogrammed cells, and optimal growth condition models) from patient tumors that are clinically and molecularly characterized [59], with collected data harmonized and accessible through the NCI's GDC data portal (https://portal.gdc.cancer.gov/). Precisely, HCMI next-generation cancer models were generated from parent tumors, which span a range of different cancer subtypes (e.g., breast, colorectal, glioblastoma, gastroesophageal, lung, melanoma, pancreas, neuroblastoma, osteosarcoma, Wilms tumor, rhabdomyosarcoma, and Ewing sarcoma), and annotated with clinical, genomic, and molecular data that include (1) clinical information, (2) biospecimen data, (3) tumor- and model-associated somatic mutations, (4) gene expression data, (5) raw sequencing data for WGS, WXS, or RNA-Seq, and (6) harmonized datasets which contain germline variants. These models represent valuable resources for translational cancer research and may contribute to developing innovative therapeutic strategies, identifying novel diagnostic markers, and individualized patient treatment plans.

Knowledge Bases for Precision Oncology

As we described above, omics technologies are extensively used today for cancer research. In fact, they have allowed us to discover new prognostic and diagnostic biomarkers for several can-

cer types, molecularly characterize cancer models, and identify the molecular mechanisms of cancer resistance to several anti-tumor drugs. In this direction, the previously discussed data sources, together with the considerable amount of publicly available omics data produced by the several multi-omics cancer projects, had an invaluable contribution to this cause. Currently, however, growing attention is given on translating such knowledge and omics technologies into clinical practice. For this reason, a new oncological discipline named "precision oncology" is rising. The premise of precision oncology is to develop treatments that target the tumor's molecular characteristics. The emergence of this kind of targeted therapy is an exciting moment in the battle against cancer. However, for the precision-oncology dream to be fully realized, the treatments must help more people with cancer than the 5–10% who currently benefit [60]. One way to do this is to identify more molecular targets. With this purpose, several tumor molecular profiling such as DNA-Seq, RNA-Seq, and mass spectrometry for proteomics and metabolomics analyses are currently used to identify new options for cancer treatment. However, today, the most widely used molecular profiling method is DNA-Seq. The introduction of DNA-Seq into clinical oncology has provided oncologists with a large amount of genomic information, which can be used in clinical decision-making [61]. Not all genomic variants identified by DNA-Seq analyses are clinically relevant. In fact, after detecting such variants, several additional steps are necessary. For example, germline polymorphisms, false-positive artifacts, and clinically insignificant synonymous variants must be filtered out from the final report. Moreover, the clinical significance of the remaining variants must be assessed by oncologists to identify potential treatments [61]. With the introduction of large numbers of targeted therapy drugs and genotype-selected clinical trials, it is challenging for oncologists to fully explore all appropriate treatment options [61]. For this purpose, several precision oncology knowledge bases have been developed to provide clinical decision support for oncologists in interpreting genomic data and identifying therapy targets. We describe some examples of the most common and used knowledge bases for precision oncology in what follows (Table 8.4).

Cancer Genome Interpreter (CGI) (https://www.cancergenomeinterpreter.org/home) incorporates several different databases for the annotation of alterations, the identification of driver mutations, the determination of variant actionability, and exploration of biomarker interactions [62]. Also, CGI tries to assess the tumorigenic potential of Variants of Unknown Significance (VUS) by using a rule-based approach. It combines several VUS features, including the gene's action, the consequence of the mutation, its position within the transcript, its prevalence within the human population, and whether the mutation occurs in a domain of the protein that is depleted of germline variants. This aspect is essential since the clinical relevance of VUS represents one of the most challenging aspects during the interpretation of genomic information influencing clinical decisions. CGI is hosted at the Barcelona Biomedical Genomics Lab.

Clinical Interpretations of Variants in Cancer (CIViC) (https://civicdb.org/home) is a knowledge base for the clinical implications of cancer genome variants [63]. It contains 7575 curated clinical evidence records for 2602 variants affecting 431 genes at the time of writing. Each evidence record is associated with a specific gene

Table 8.4 Knowledge bases for precision oncology

Precision oncology knowledge base	References	Link
CGI	[62]	https://www.cancergenomeinterpreter.org/home
CIViC	[63]	https://civicdb.org/home
MCG	[66]	https://www.mycancergenome.org/
PMKB	[67]	https://pmkb.weill.cornell.edu/
OncoKB	[68]	https://www.oncokb.org/

variant and contains information related to therapy, prognosis, diagnosis, and cancer predisposition. Genetic variants are ranked accordingly with the evidence of their clinical utility from level A (established clinical utility) to E (inferential) [63]. CIViC is hosted at the Washington University in St. Louis School of Medicine.

My Cancer Genome (MCG) (https://www.mycancergenome.org/) is a knowledge base that offers information on targeted therapies and clinical trials for several genetic variants that are involved in tumorigenesis [64]. Unlike many other precision oncology knowledge bases, MCG organizes clinical evidence for genomic variants in a disease-centric approach instead of a gene-centric approach [65]. MCG distinguishes genetic variants within the same gene [66] that could potentially discourage the use of specific targeted agents or not. MCG is hosted at Vanderbilt University Medical Center and has a commercial relationship with GenomOncology LLC (Cleveland, OH, United States).

Precision Medicine Knowledge Base (PMKB) (https://pmkb.weill.cornell.edu/) is a knowledge base for cancer mutation interpretations [67]. Like CIViC, PMKB rates variant interpretations by a numeric tier, indicating the clinical actionability: Tier 1, strong evidence of clinical utility, Tier 2, potential clinical relevance, and Tier 3, undetermined clinical significance [67]. PMKB is hosted at Weill Cornell Medicine Englander Institute for Precision Medicine.

Precision Oncology Knowledge Base (OncoKB) (https://www.oncokb.org/) is a precision oncology knowledge base of tumor variants and their related FDA-approved therapies and/or other drugs that are under study in clinical trials [68]. OncoKB offers information for 5425 genetic variants in 682 cancer-associated genes from 56 tumor types. Interestingly, OncoKB also highlights adverse outcomes of off-label drugs in specific mutational contexts [65]. OncoKB is hosted at the Memorial Sloan Kettering Cancer Center.

Biological Pathways Databases

Information derived from the analysis of genomics, transcriptomics, proteomics, and metabolomics experiments could be integrated into pathways that describe the samples under study to understand the interrelations between the different components and their effects. Pathway-centric approaches are widely used to interpret and contextualize omics data. Several pathway databases that describe the already known signaling and metabolic pathways in humans and other organisms have been developed for this purpose. Examples include KEGG [4–10], Reactome [11–15], WikiPathways [16–19], and Pathway Commons [20, 21] (Table 8.5). However, these databases contain different representations of the same biological pathway, leading to varying results during pathway analysis [69]. Besides, pathways are often also described at different levels of detail [69]. Nonetheless, most pathway analyses are conducted extensively by using pathway information retrieved from such databases. Some well-known examples of pathway databases are briefly described below.

Kyoto Encyclopedia of Genes and Genomes (KEGG) (https://www.genome.jp/kegg/) is an online resource consisting of 18 databases used to study biological systems from large-scale molecular datasets generated by high-throughput experimental technologies [4–10]. Among these databases, there is KEGG pathway, a collection of manually drawn pathways representing our knowledge of the molecular interaction, reaction,

Table 8.5 Biological pathway databases

Pathway database	References	Link
KEGG	[4–10]	https://www.genome.jp/kegg/
Reactome	[11–15]	https://reactome.org/
Pathway Commons	[20, 21]	https://www.pathwaycommons.org
WikiPathways	[16–19]	https://www.wikipathways.org/index.php/WikiPathways

and relation networks for many metabolic, signaling, and disease pathways.

Reactome (https://reactome.org/) is an open-source, open access, and manually curated pathway database that helps scientists find, organize, and utilize biological information to support data visualization, integration, and analysis [11–15]. Indeed, Reactome could be used (i) to interpret the results of high-throughput experimental studies; (ii) to develop novel algorithms for pathway analysis; (iii) and to implement predictive models of normal and disease pathways [11–15]. The core unit of the Reactome data model is the reaction. Entities (proteins, enzymes, metabolites, anti-cancer drugs, etc.) participating in reactions form a network of biological interactions and are grouped into pathways [11–15]. Examples of biological pathways in Reactome include metabolic pathways, signaling pathways, transcriptional regulation, apoptosis, and disease.

Pathway Commons (https://www.pathwaycommons.org) is another publicly available database of biological pathways. It collects data from different pathway databases [20, 21]. Pathway Commons does not compete with other pathway databases, but it adds value to these existing ones by providing a unique online resource for sharing and querying pathway information [20, 21].

WikiPathways was established to facilitate the contribution and maintenance of pathway information by biologists [16–19]. Indeed, WikiPathways is an open, collaborative platform dedicated to curating biological pathways that enhance and complement ongoing efforts, such as KEGG, Reactome, and Pathway Commons [16–19]. The easy-to-use web-based interface of WikiPathways was specifically developed to reduce the obstacles to participate in pathway curation. Any pathway can be edited from within its wiki page by using the pathway editor. More importantly, the open approach of WikiPathways encourages broader participation of the scientific community [16–19]. Finally, pathways and their content can be downloaded in several data and image formats, including GPML, which can be used by pathway visualization and analysis tools such as Cytoscape, GenMAPP, and PathVisio.

Strategies for Pathway Analysis and Their Applications in Cancer Research and Precision Oncology

Pathway Analysis Methods

Initially, pathway analysis identified a class of techniques for (i) the study of ontological terms and protein-protein interaction (PPI) networks; and (ii) the inference of gene regulatory networks from expression data. The aim was to use ontologies and/or pathways as knowledge bases for grouping genes or proteins into smaller subsets according to some relationships, reducing the dimensionality of expression data. However, more recently, research effort has been devoted to deploying a novel class of knowledge base–driven pathway analysis methods. Such methods leverage existing databases such as the previously discussed KEGG [4], Reactome [11], WikiPathways [16–19], and Pathway Commons [20, 21], to identify perturbed pathways associated with a specific phenotype or condition. A typical knowledge base–driven pathway analysis method starts from two types of input data: (i) a set of pathways representing the molecular interaction knowledge base, and (ii) experimental OMICs or multi-omics data containing measurements of gene expressions, protein abundance, or metabolite concentration in two or more conditions [24]. A graph model is then built to represent pathways. Models depend on pathway type: (i) signaling pathway where nodes are gene (or gene products), and edges represent signals, such as activation or repression, (ii) metabolic pathway in which nodes are biochemical compounds, and enzymes and edges represent reactions that transform one or more compounds into another one. Pathways are then ranked according to the perturbation level, which is computed through a scoring scheme [24]. Following temporal criteria, knowledge base–driven pathway analysis methods can be classified into three generations of approaches: (i) Over-Representation Analysis (ORA), (ii) Functional Class Scoring (FCS), and (iii) Pathway Topology (PT)-based analysis [24]. More recently, new

approaches have been proposed to analyze pathways augmented with missing regulatory elements, such as microRNAs and their post-transcriptional regulatory interactions with genes [3].

Over-Representation Analysis (ORA)

ORA methods are the first generation of pathway analysis models. ORA techniques measure pathway perturbation considering only the number of DIfferentially Expressed Genes (DEGs) present in the pathway. Their primary hypothesis is that a statistically significant pathway contains more DEGs than those that would appear by chance. Therefore, ORA strategies typically divide the list of genes according to the pathway each gene belongs to. Then, they compute the probability of observing a certain number of altered genes in a pathway by chance applying a hypothesis test [24]. However, these methods have several drawbacks. First, they ignore the expression of genes and the magnitude of their change. Second, they analyze only DEGs, missing genes with coordinated alterations, which may lead to remarkable effects. Finally, pathways are analyzed independently from the surrounding biological context, ignoring the dependence from other pathways encoding different molecular processes. Unfortunately, the hypothesis behind ORA methods is a very simplified representation of what happens in reality, where several biological processes are accomplished by chains of reactions involving two or more pathways [24]. Some examples of ORA methods follow below.

DIANA-miRPath [70] is an example of a first-generation pathway analysis that assesses miRNAs' impact on biological processes by identifying the pathways they are significantly involved in. The tool functionally annotates one or more miRNAs through a hypergeometric distribution, an unbiased empirical distribution, or a statistical meta-analysis. Moreover, it allows identifying subsets of miRNAs, which significantly regulate a collection of pathways, starting from experimental data.

Onto-Express [71, 72] is a Java-based tool to automatically translate a list of differentially regulated genes into functional profiles characterizing their impact. More specifically, Onto-Express uses public data and GO categories to create functional profiles that correlate expression profiles with the following categories: cytogenetic locations, biochemical and molecular functions, biological processes, cellular components, and cellular roles of the translated proteins. For each pathway and category, statistical significance values are calculated by using a user-chosen binomial or χ^2 test. In the case of χ^2 test unreliability, Onto-Express automatically determines expected values and uses Fisher's exact test.

FuncAssociate [73] is a web-based tool to characterize large gene sets starting from GO attributes. The input is a list of genes. For each attribute, the algorithm detects the number of genes annotated with such an attribute. Finally, multiple hypotheses testing is performed to establish the statistical significance of this number. FuncAssociate can also handle ranked input lists. This option is useful when the user wants to rank genes according to some criterion of interest, e.g., their significance or their fold-change in mRNA abundance between two different conditions.

GeneMerge [74] is a web-based and standalone program that returns a range of functional and genomic data for a given set of study genes and provides statistical rank scores for over-representing particular functions or categories in the data set.

GOToolBox [75] is a set of methods and tools to process GO annotations. The user can find all GO terms associated with each gene in the input dataset, rank all annotation terms, evaluate the significance of their occurrences within the dataset, group together functionally related genes based on their GO terms, and find genes sharing GO terms with a user-given gene, based on a functional similarity calculation.

MAPPFinder [76] dynamically links gene expression data to the GO hierarchy. The algorithm calculates the percentage of genes that meet a user-defined criterion for a meaningful gene expression change.

Functional Class Scoring (FCS)

Second-generation tools, named Functional Class Scoring, consider both changes in gene expressions and their correlations with a phenotype of interest. FCS methods' central hypothesis is that pathway status is affected by significant gene expression changes as well as smaller changes with a combined significant contribution [24]. FCS methods typically start by computing a gene-level statistic from gene expression values, such as the correlation of molecular measurements with the phenotype (i.e., ANOVA, Q-statistic, signal-to-noise ratio, t-test, and Z-score). Next, gene-level statistics for all genes in a pathway are aggregated into a single pathway-level statistic, such as Kolmogorov-Smirnov statistic; sum, mean, or median of gene-level statistics; Wilcoxon rank-sum; or max-mean statistic. Finally, the pathway-level statistic's significance is assessed through an appropriate null hypothesis, which can be self-contained or competitive [24]. In the first case, class labels (i.e., phenotypes) are permuted for each sample, and the set of genes in a given pathway is compared to itself, ignoring genes that are not present in the pathway. In competitive null hypotheses, gene labels are permuted for each pathway, and the set of genes in the pathway is compared to the set of genes that are not in the pathway. FCS approaches can rank genes through their expression levels and consider the dependencies within a pathway. This idea allows going beyond some of the limitations of ORA methods. However, they do not consider either genes' deregulation magnitude for pathway activity estimation, or the interactions between genes, or their direction, type, and strength. Therefore, FCS methods, as well as ORA methods, treat pathways as simple sets of genes [24]. Examples of FCS methods are listed below.

Gene Set Enrichment Analysis (GSEA) [77] ranks genes according to the correlation between gene expression and phenotype. It computes a score that expresses how much the pathway is related to the phenotypic class distinction.

GSA [78] is an extension of GSEA, improved by using a max-mean statistic to summarize gene-sets and adding a restandardization procedure. Restandardization consists of centering and scaling the max-mean statistic by its mean and standard deviation under row randomizations. This standardized max-mean statistic is then computed both on the original data and on the permuted datasets.

GlobalANCOVA [79] is a general methodology that uses gene-wise linear models and aggregates their information in a multivariate test procedure. It can be used to study how expression structure within a group of genes is influenced by design aspects of the study, such as group membership, time course, group by time course interaction, dosage, group by dose interaction, etc. Gene-wise linear models are used to formalize the relationship of gene expression with phenotypic or genomic covariates. An ANOVA-based sum of squares summarizes individual gene-wise linear models to a group statement. A permutation test and an asymptotic distribution of the test statistics under the null hypothesis are available to calculate p-values.

Pathway Topology (PT)-Based Analysis

The third generation of pathway analysis systems is the so-called Topology-based methods. They fully exploit the topological information encoded by pathways when computing perturbation scores. Pathways are modeled as complex graphs where each node is a gene or a protein, and each edge is an interaction between them. Even though thousands of genes are not annotated in pathways, and existing annotations may be inaccurate, graphs in these databases provide a more detailed view of biological processes within the cell, helping the interpretation of high-throughput experiments [24]. Some examples of PT methods are listed below.

ScorePAGE [80] computes similarities between each pair of genes in a pathway (e.g., correlation, covariance). The similarity is averaged to calculate a pathway-level score. A weight is given to pairwise similarities, dividing similarities by the number of reactions needed to connect two genes in a given pathway.

In [81], *Draghici et al.* introduce a technique called impact factor (IF). The impact factor is a pathway-level score that considers the magnitude

of changes in gene expression, the type of interaction, and the location of genes in the pathway graph. Authors also define a gene-level statistic called perturbation factor (PF), which is a linear function of the change in gene expression and the perturbation of its neighborhood. This statistic is then combined for each element in the pathway, and a *p*-value is computed using an exponential distribution. The IF technique has been implemented as a web-based tool, called Pathway-Express, and freely available as part of Onto-Tools (http://vortex.cs.wayne.edu).

SPIA [82] improves Draghici's method by attenuating the effect of expression changes in the PF computation and lowering the high rate of false positives when the input list of genes is small.

NetGSA [83] considers both the change in correlation and the change in the network structure as experimental conditions change. Like the IF technique, NetGSA models gene expression as a linear function of other network genes. It considers a gene baseline expression by representing it as a latent variable in the model. However, pathways must be defined as directed acyclic graphs (DAGs).

PARADIGM [84] can predict the degree of alteration in the patient-specific genetic activity of a pathway by employing a probabilistic inference algorithm. Each pathway is converted into a factor graph that includes both hidden and observed states. The factor graph integrates observations on gene- and biological process-related state information with a structure describing known interactions among the entities. Variables of the model describe the states of entities in a cell, such as mRNAs or complexes, and factors represent the interactions and information flow between these entities. These variables represent the differential state of each entity compared to a "control" or normal level rather than the direct concentrations of the molecular entities. Parameters of the observation factors are estimated using an EM algorithm. Authors show that their model achieves more reliable results than SPIA, but Mitrea et al. in [85] state they could not reproduce the results reported in [84].

pDis [86] is another PT method that can identify significantly impacted pathways using the entire set of genes, rather than focusing only on DE genes. To reduce false positives and false negatives, they propose a scoring scheme that can distinguish between genes that are sources of primary deregulation due to mutations, copy number variations, epigenetic changes, etc., and genes that merely respond to perturbation signals from upstream genes. The method yields significant improvements to SPIA, GSEA, and GSA in terms of both ranks and *p*-values of perturbed pathways.

miRNA-Sensitive Topological Pathway Analysis

Most pathway analysis methods do not consider the effects of post-transcriptional regulatory interactions involving microRNAs. Recently, new methodologies have been proposed in this direction. In [3], the authors present MITHrIL, a tool that extends the method in [81] and SPIA [82]. The method returns a list of pathways sorted according to their deregulation degree and the corresponding statistical significance starting from expression values of genes and microRNAs. A predicted degree of alteration for each endpoint (i.e., a pathway node whose alteration, based on current knowledge, affects the phenotype in a specific way) is computed. Validated inhibition interactions between miRNA and targets are taken from miRTarBase [87] and miRecords [88]. Endpoints in each pathway are found through a Depth-First Search (DFS) algorithm to automatically mark genes located at the end of the chains of reactions in the pathway. Putative endpoints are then manually screened to determine if they are associated with phenotypic changes as stated in the KEGG database. For each gene in a pathway, MITHrIL computes a perturbation factor (PF), which estimates how much its activity is altered considering its expression and its immediate neighbors. By appropriately combining each PF of a pathway, MITHrIL can compute an impact factor (IF) and an accumulator (Acc). IF indicates how important the pathway changes are, while Acc measures the

total level of perturbation in the pathway and the tendency to have a majority of activated or inhibited genes. Next, a *p*-value is associated with the Acc measure to estimate the probability of getting such an accumulator by chance. Finally, the false discovery rate is calculated, and *p*-values are adjusted on multiple hypotheses. The output of MITHrIL consists of a list of pathways along with their impact factor, accumulator, and adjusted *p*-values. Such a list is sorted by *p*-value and Acc.

A summary of all surveyed pathway analysis methods is shown in Table 8.6.

Applications of Pathway Analysis Methods in Cancer Research

As we discussed above, pathway analysis is an essential step for interpreting omics data to understand the phenotype under study. Hence, the increasing amount of data available is fostering rapid advances in accurate and reliable pathway analysis tools. This aspect is quite relevant in complex diseases like cancer, where alterations can establish abnormal cell growth. Indeed, many factors are involved in developing and maintaining the malignant state of cancer cells, such as genetic aberrations, epigenetic alterations, changes in gene expression, metabolic alterations, and beyond [2]. Therefore, by analyzing cancer as a complex pathology, we try to gain insight into the cancer cells' molecular mechanisms by looking at their different components. Such knowledge can be extrapolated from the different OMICs data types and integrated into pathway analysis approaches through in silico models representing the biological system under study. Few examples of pathway analyses in cancer research include the following: (i) the identification of driver genes and pathways [89, 90]; (ii) the discovery of novel tumor subtypes [91]; (iii) understanding cancer mechanisms and biomarkers [90, 91]; and (iv) the identification of key regulators in cancer gene networks [92, 93]. Many scientific studies show successful applications of pathway analysis approaches to get insights into several cancer biology aspects. This aspect is straightforward if we consider that many critical pathways' functions are altered during cancer initiation and progression.

In this context, PT methods play a more pivotal role than previous generation systems such as ORA and FCS methods. In fact, PT methods fully exploit the topological information encoded by pathways when computing perturbation scores and not only the presence of particular sets of DEGs for specific pathways and the magnitude of their dysregulation to identify the altered pathways. This aspect should not be underestimated since PT methods allow integrating high-throughput data into a more realistic model where each element (genes, proteins, metabolites, etc.) is connected accordingly with functional interaction available on the several pathway databases previously discussed.

Although biological pathways are partial and incomplete, pathway analyses are still conducted extensively to interpret the results of high-throughput experiments. In any case, the annotation of pathways with other experimentally validated interactions may increase the reliability of PT-based analysis methods [24]. For example, many regulatory ncRNAs' functional interactions are still missing in pathway databases. Among the plethora of different ncRNA classes already discovered, miRNAs have been revealed to be important in modulating several pathways via the exertion of their regulatory function when targeting essential genes [94, 95]. Indeed, the deregulation of even a single miRNA is capable of causing cancer, as in the case of miR-155, which is responsible for the onset of acute lymphoblastic leukemia/high-grade lymphoma in mice [96]. Additionally, the predominant roles played by miRs 21, 221, and 222 in several cancer types prove the importance these small RNA molecules have in tumor pathogenesis and progression while also being a determining factor in drug resistance [95]. In light of this and many other pieces of evidence discovered in more recent years, the integration of miRNA expression when evaluating cancer pathway perturbation has become of utmost crucial importance. Indeed, considering the effects of miRNAs on overall gene expression contributes to a more

Table 8.6 Methods for pathway analysis

Pathway analysis tools	Category	References	Link	Deployment
DIANA-miRPath	ORA	[70]	http://snf-515788.vm.okeanos.grnet.gr/	Web-based application
Onto-Express	ORA	[71, 72]	N\A	N\A
FuncAssociate	ORA	[73]	http://llama.mshri.on.ca/funcassociate/	Web-based application
GeneMerge	ORA	[74]	http://www.genemerge.net/	Stand-alone command-line tool
GOToolBox	ORA	[75]	N\A	N\A
MAPPFinder	ORA	[76]	http://www.genmapp.org/	N\A
GSEA	FCS	[77]	https://www.gsea-msigdb.org/gsea/index.jsp	Java desktop application
GSA	FCS	[78]	N\A	N\A
GlobalANCOVA	FCS	[79]	https://www.bioconductor.org/packages/release/bioc/html/GlobalAncova.html	R package
ScorePAGE	PT	[80]	https://rdrr.io/github/zhilongjia/scorePage/	R package
Draghici et al.	PT	[81]	https://rdrr.io/bioc/ROntoTools/man/pe.html	R package
SPIA	PT	[82]	https://bioconductor.org/packages/release/bioc/html/SPIA.html	R package
NetGSA	PT	[83]	https://github.com/drjingma/netgsa	R package
PARADIGM	PT	[84]	https://sbenz.github.io/Paradigm/	Stand-alone command-line tool
pDis	PT	[86]	https://rdrr.io/bioc/ROntoTools/man/pDis.html	R package
MITHrIL	miRNA-sensitive PT	[3]	https://alpha.dmi.unict.it/mithril/	Command-line Java application

comprehensive depiction of the biological reality, providing a more accurate means for pathway assessment and phenotype categorization [3]. Leveraging on the potential offered by miRNA enrichment in pathway analysis, MITHrIL [3] represents a bioinformatic resource capable of a far more accurate evaluation of pathway deregulation in cancer. This feature could provide a decisive contribution to cancer research in terms of directing researchers more effectively, reducing costs and time requirements [3]. However, additional classes of ncRNAs such as tRNA-derived small ncRNAs (tsRNAs) are still missing in biological pathways despite their roles in cancer development and progression have been extensively assessed in recent years [97–100]. Indeed, among their functions, it seems that tsRNAs are also complexed with AGO proteins, and they might act as negative regulators of gene expression in a miRNA-like manner [99, 101, 102]. Therefore, in the upcoming years, it is reasonable to think that they will soon be annotated in biological pathways, increasing our knowledge of the molecular interaction acting in cancer cells and the accuracy of pathway analysis methods.

Phenotype and Therapy Predictions: Applications for Precision Oncology

In addition to pathway analysis, the knowledge retrieved from pathway databases can also be used in system biology to predict the effect of gene dysregulation, gene mutations, chromosomal deletion, and cellular response to specific signals and drugs on phenotypes. As widely known, all these alterations can be efficiently assessed by the new OMICs technologies. However, despite the improvements in our understanding of cell biology, it is challenging to link OMICs data to the physiopathological status. In any case, systems biology computational approaches have emerged as efficient means capable of bridging the gap between experimental biology at the system-level and quantitative sciences [103]. Indeed, such methods can be used as time- and cost-saving solutions for efficient in silico predictions [103, 104]. At present, several simulation models have been developed, and they can be mainly grouped into two broad categories: (i) discrete/logic or (ii) continuous models [105]. Briefly, discrete models represent each element's state in a biological network as discrete levels, and the temporal dynamic is also discretized. At each time step, the state is updated according to a function, determining how an entity's state depends on the state of other (usually connected) entities. Boolean networks [106, 107] and Petri nets [108] represent two types of discrete models. On the other hand, continuous models usually produce real continuous measurements, instead of discretized values, simulating network dynamics over a continuous timescale. Although they could provide a greater degree of accuracy, these methods are limited by our current description of the biological systems and our measurement techniques' capabilities. Continuous linear models [109, 110] and flux balance analysis [111] are the most representative continuous models. A new web-based and user-friendly system named PHENSIM [112] has been recently released for phenotype prediction among these systems. Specifically, PHENSIM allows phenotype predictions on selected cell lines or tissues in three different organisms (*Homo sapiens*, *Mus musculus*, and *Rattus norvegicus*) using a probabilistic algorithm to compute the effect of dysregulated genes, proteins, miRNAs, and metabolites on KEGG pathways. Results are then summarized through a Perturbation value, which represents the expected magnitude of the alteration, and an Activity Score, which is an index of both the predicted effect of a gene dysregulation on a node (up- or downregulation) and its likelihood. All values are also summarized at the pathway level. To achieve better performance, PHENSIM performs all calculations in a more realistic model that is the KEGG meta-pathway, obtained by merging all pathways [113] and integrating information on miRNA-target and transcription factor (TF)-miRNA extracted from online public knowledge bases [3]. Intuitively, such a method has many potential applications in cancer research and precision oncology. First, these techniques can simulate the effect of mutations

on normal cells or the influence of drugs (or drug combinations) on tumor cell lines. It is also possible to predict the probable result of laboratory experiments, prioritize the most promising ones, and optimize resource use. Finally, thanks to the possibility of simulating the effect of drugs, these techniques can be applied to data from individual patients to predict which approved drugs may be potentially the most appropriate for the treatment of the disease or repositioning if no suitable therapy is available.

Conclusion

It is widely known that oncological research and new disciplines such as precision oncology rely on OMICs data, computational tools, and knowledge bases for data analysis and results interpretation. However, in contexts like biology and biomedicine, the literature rises fast, and, therefore, the ecosystem of up-to-date models increases rapidly. In this chapter, we gave a comprehensive survey of the main data sources for publicly available OMICs data (e.g., genomics, transcriptomics, proteomics, and metabolomics), together with the main cancer-related OMICs projects and knowledge bases for precision oncology applications. Then, we provided a user-view of relevant examples of pathway databases and pathway analysis methods that can be used for multi-OMICs integration and analysis. Ultimately, this chapter aimed to provide a guide for researchers interested in using omics data and pathway analysis methods for cancer research and precision oncology application.

References

1. Aronson SJ, Rehm HL. Building the foundation for genomics in precision medicine. Nature. 2015;526:336–42.
2. de Anda-Jáuregui G, Hernández-Lemus E. Computational oncology in the multi-omics era: state of the art. Front Oncol. 2020;10:423.
3. Alaimo S, Giugno R, Acunzo M, et al. Post-transcriptional knowledge in pathway analysis increases the accuracy of phenotypes classification. Oncotarget. 2016;7:54572–82.
4. Kanehisa M. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28:27–30.
5. Kanehisa M, Goto S, Kawashima S, Nakaya A. The KEGG databases at GenomeNet. Nucleic Acids Res. 2002;30:42–6.
6. Kanehisa M. The KEGG resource for deciphering the genome. Nucleic Acids Res. 2004;32:277D–280.
7. Kanehisa M, Goto S, Hattori M, et al. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res. 2006;34:D354–7.
8. Kanehisa M, Goto S, Furumichi M, et al. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. 2010;38:D355–60.
9. Kanehisa M, Goto S, Sato Y, et al. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 2012;40:D109–14.
10. Kanehisa M, Furumichi M, Tanabe M, et al. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017;45:D353–61.
11. Joshi-Tope G, Gillespie M, Vastrik I, et al. Reactome: a knowledgebase of biological pathways. Nucleic Acids Res. 2005;33:D428–32.
12. Jassal B, Matthews L, Viteri G, et al. The reactome pathway knowledgebase. Nucleic Acids Res. 2020;48:D498–503.
13. Fabregat A, Korninger F, Viteri G, et al. Reactome graph database: efficient access to complex pathway data. PLoS Comput Biol. 2018;14:e1005968.
14. Fabregat A, Sidiropoulos K, Viteri G, et al. Reactome diagram viewer: data structures and strategies to boost performance. Bioinformatics. 2018;34:1208–14.
15. Sidiropoulos K, Viteri G, Sevilla C, et al. Reactome enhanced pathway visualization. Bioinformatics. 2017;33:3461–7.
16. Kelder T, van Iersel MP, Hanspers K, et al. WikiPathways: building research communities on biological pathways. Nucleic Acids Res. 2012;40:D1301–7.
17. Pico AR, Kelder T, van Iersel MP, et al. WikiPathways: pathway editing for the people. PLoS Biol. 2008;6:e184.
18. Kutmon M, Riutta A, Nunes N, et al. WikiPathways: capturing the full diversity of pathway knowledge. Nucleic Acids Res. 2016;44:D488–94.
19. Slenter DN, Kutmon M, Hanspers K, et al. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. Nucleic Acids Res. 2018;46:D661–7.
20. Cerami EG, Gross BE, Demir E, et al. Pathway Commons, a web resource for biological pathway data. Nucleic Acids Res. 2011;39:D685–90.
21. Rodchenkov I, Babur O, Luna A, et al. Pathway Commons 2019 Update: integration, analysis and exploration of pathway data. Nucleic Acids Res. 2020;48:D489–97.
22. Jin L, Zuo X-Y, Su W-Y, et al. Pathway-based analysis tools for complex diseases: a review. Genomics Proteomics Bioinformatics. 2014;12:210–20.

23. Zhang W, Chien J, Yong J, Kuang R. Network-based machine learning and graph theory algorithms for precision oncology. npj Precis Oncol. 2017;1:25.
24. Alaimo S, Micale G, La Ferlita A, et al. Computational methods to investigate the impact of miRNAs on pathways. Methods Mol Biol. 2019;1970:183–209.
25. Siva N. 1000 Genomes project. Nat Biotechnol. 2008;26:256.
26. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. Trends Genet. 2014;30:418–26.
27. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10:57–63.
28. Perez-Riverol Y, Alpi E, Wang R, et al. Making proteomics data accessible and reusable: current state of proteomics databases and repositories. Proteomics. 2015;15:930–49.
29. Craig R, Cortens JP, Beavis RC. Open source system for analyzing, validating, and storing protein identification data. J Proteome Res. 2004;3:1234–42.
30. Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. Bioinformatics. 2004;20:1466–7.
31. Farrah T, Deutsch EW, Omenn GS, et al. State of the human proteome in 2013 as viewed through PeptideAtlas: comparing the kidney, urine, and plasma proteomes for the biology- and disease-driven human proteome project. J Proteome Res. 2014;13:60–75.
32. Farrah T, Deutsch EW, Hoopmann MR, et al. The state of the human proteome in 2012 as viewed through PeptideAtlas. J Proteome Res. 2013;12:162–71.
33. Desiere F, Deutsch EW, Nesvizhskii AI, et al. Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. Genome Biol. 2005;6:R9.
34. Deutsch EW, Mendoza L, Shteynberg D, et al. A guided tour of the Trans-Proteomic Pipeline. Proteomics. 2010;10:1150–9.
35. Lam H, Aebersold R. Building and searching tandem mass (MS/MS) spectral libraries for peptide identification in proteomics. Methods. 2011;54:424–31.
36. Picotti P, Rinner O, Stallmach R, et al. High-throughput generation of selected reaction-monitoring assays for proteins and proteomes. Nat Methods. 2010;7:43–6.
37. Deutsch EW, Lam H, Aebersold R. PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. EMBO Rep. 2008;9:429–34.
38. Vizcaíno JA, Côté RG, Csordas A, et al. The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. Nucleic Acids Res. 2013;41:D1063–9.
39. Haug K, Cochrane K, Nainala VC, et al. MetaboLights: a resource evolving in response to the needs of its scientific community. Nucleic Acids Res. 2020;48:D440–4.
40. Wishart DS, Feunang YD, Marcu A, et al. HMDB 4.0: the human metabolome database for 2018. Nucleic Acids Res. 2018;46:D608–17.
41. Barretina J, Caponigro G, Stransky N, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012;483:603–7.
42. Ghandi M, Huang FW, Jané-Valbuena J, et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature. 2019;569:503–8.
43. Weinstein JN, Collisson EA, Mills GB, et al. The cancer genome atlas pan-cancer analysis project. Nat Genet. 2013;45:1113.
44. Hutter C, Zenklusen JC. The cancer genome atlas: creating lasting value beyond its data. Cell. 2018;173:283–5.
45. Ma X, Liu Y, Liu Y, et al. Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. Nature. 2018;555:371–6.
46. Smith MA, Seibel NL, Altekruse SF, et al. Outcomes for children and adolescents with cancer: challenges for the twenty-first century. J Clin Oncol. 2010;28:2625–34.
47. Mullighan CG, Su X, Zhang J, et al. Deletion of IKZF1 and prognosis in acute lymphoblastic leukemia. N Engl J Med. 2009;360:470–80.
48. Mullighan CG, Zhang J, Harvey RC, et al. JAK mutations in high-risk childhood acute lymphoblastic leukemia. Proc Natl Acad Sci U S A. 2009;106:9414–8.
49. Kang H, Chen I-M, Wilson CS, et al. Gene expression classifiers for relapse-free survival and minimal residual disease improve risk classification and outcome prediction in pediatric B-precursor acute lymphoblastic leukemia. Blood. 2010;115:1394–405.
50. Yang JJ, Cheng C, Devidas M, et al. Ancestry and pharmacogenomics of relapse in acute lymphoblastic leukemia. Nat Genet. 2011;43:237–41.
51. Zhang J, Mullighan CG, Harvey RC, et al. Key pathways are frequently mutated in high-risk childhood acute lymphoblastic leukemia: a report from the Children's Oncology Group. Blood. 2011;118:3080–7.
52. Loh ML, Zhang J, Harvey RC, et al. Tyrosine kinome sequencing of pediatric acute lymphoblastic leukemia: a report from the Children's Oncology Group TARGET Project. Blood. 2013;121:485–8.
53. Pugh TJ, Morozova O, Attiyeh EF, et al. The genetic landscape of high-risk neuroblastoma. Nat Genet. 2013;45:279–84.
54. Walz AL, Ooms A, Gadd S, et al. Recurrent DGCR8, DROSHA, and SIX homeodomain mutations in favorable histology Wilms tumors. Cancer Cell. 2015;27:286–97.
55. Gooskens SL, Gadd S, Guidry Auvil JM, et al. TCF21 hypermethylation in genetically quiescent clear cell sarcoma of the kidney. Oncotarget. 2015;6:15828–41.
56. Chun H-JE, Lim EL, Heravi-Moussavi A, et al. Genome-wide profiles of extra-cranial malignant rhabdoid tumors reveal heterogeneity and dys-

regulated developmental pathways. Cancer Cell. 2016;29:394–406.
57. Nakka M, Allen-Rhoades W, Li Y, et al. Biomarker significance of plasma and tumor miR-21, miR-221, and miR-106a in osteosarcoma. Oncotarget. 2017;8:96738–52.
58. Brunner AM, Graubert TA. Genomics in childhood acute myeloid leukemia comes of age. Nat Med. 2018;24:7–9.
59. Ledford H. Global initiative seeks 1,000 new cancer models. Nature. 2016. http://www.nature.com/news/global-initiative-seeks-1-000-new-cancer-models-1.20242. Accessed 3 Nov 2020.
60. Hodson R. Precision oncology. Nature. 2020;585:S1.
61. Li X, Warner JL. A review of precision oncology knowledgebases for determining the clinical actionability of genetic variants. Front Cell Dev Biol. 2020;8:48.
62. Tamborero D, Rubio-Perez C, Deu-Pons J, et al. Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. Genome Med. 2018;10:25.
63. Griffith M, Spies NC, Krysiak K, et al. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. Nat Genet. 2017;49:170–4.
64. Taylor AD, Micheel CM, Anderson IA, et al. The path(way) less traveled: a pathway-oriented approach to providing information about precision cancer medicine on my cancer genome. Transl Oncol. 2016;9:163–5.
65. Gao P, Zhang R, Li J. Comprehensive elaboration of database resources utilized in next-generation sequencing-based tumor somatic mutation detection. Biochim Biophys Acta Rev Cancer. 2019;1872:122–37.
66. Swanton C. My Cancer Genome: a unified genomics and clinical trial portal. Lancet Oncol. 2012;7:668–9.
67. Huang L, Fernandes H, Zia H, et al. The cancer precision medicine knowledge base for structured clinical-grade mutations and interpretations. J Am Med Inform Assoc. 2017;24:513–9.
68. Chakravarty D, Gao J, Phillips SM, et al. OncoKB: a precision oncology knowledge base. JCO Precis Oncol. 2017;2017:PO.17.00011. https://doi.org/10.1200/PO.17.00011.
69. Mubeen S, Hoyt CT, Gemünd A, et al. The impact of pathway database choice on statistical enrichment analysis and predictive modeling. Front Genet. 2019;10:1203.
70. Vlachos IS, Zagganas K, Paraskevopoulou MD, et al. DIANA-miRPath v3.0: deciphering microRNA function with experimental support. Nucleic Acids Res. 2015;43:W460–6.
71. Khatri P, Draghici S, Ostermeier GC, Krawetz SA. Profiling gene expression using onto-express. Genomics. 2002;79:266–70.
72. Drăghici S, Khatri P, Martins RP, et al. Global functional profiling of gene expression. Genomics. 2003;81:98–104.
73. Berriz GF, King OD, Bryant B, et al. Characterizing gene sets with FuncAssociate. Bioinformatics. 2003;19:2502–4.
74. Castillo-Davis CI, Hartl DL. GeneMerge--post-genomic analysis, data mining, and hypothesis testing. Bioinformatics. 2003;19:891–2.
75. Martin D, Brun C, Remy E, et al. GOToolBox: functional analysis of gene datasets based on Gene Ontology. Genome Biol. 2004;5:R101.
76. Doniger SW, Salomonis N, Dahlquist KD, et al. MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. Genome Biol. 2003;4:R7.
77. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102:15545–50.
78. Efron B, Tibshirani R. On testing the significance of sets of genes. Ann Appl Stat. 2007;1:107–29.
79. Hummel M, Meister R, Mansmann U. GlobalANCOVA: exploration and assessment of gene group effects. Bioinformatics. 2008;24:78–85.
80. Rahnenführer J, Domingues FS, Maydt J, Lengauer T. Calculating the statistical significance of changes in pathway activity from gene expression data. Stat Appl Genet Mol Biol. 2004;3:16.
81. Draghici S, Khatri P, Tarca AL, et al. A systems biology approach for pathway level analysis. Genome Res. 2007;17:1537–45.
82. Tarca AL, Draghici S, Khatri P, et al. A novel signaling pathway impact analysis. Bioinformatics. 2009;25:75–82.
83. Shojaie A, Michailidis G. Analysis of gene sets based on the underlying regulatory network. J Comput Biol. 2009;16:407–26.
84. Vaske CJ, Benz SC, Sanborn JZ, et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics. 2010;26:i237–45.
85. Mitrea C, Taghavi Z, Bokanizad B, et al. Methods and approaches in the topology-based analysis of biological pathways. Front Physiol. 2013;4:278.
86. Ansari S, Voichita C, Donato M, et al. A novel pathway analysis approach based on the unexplained disregulation of genes. Proc IEEE Inst Electr Electron Eng. 2017;105(3):482–95.
87. Hsu S-D, Lin F-M, Wu W-Y, et al. miRTarBase: a database curates experimentally validated microRNA-target interactions. Nucleic Acids Res. 2011;39:D163–9.
88. Xiao F, Zuo Z, Cai G, et al. miRecords: an integrated resource for microRNA-target interactions. Nucleic Acids Res. 2009;37:D105–10.
89. Akavia UD, Litvin O, Kim J, et al. An integrated approach to uncover drivers of cancer. Cell. 2010;143:1005–17.
90. Danussi C, Akavia UD, Niola F, et al. RHPN2 drives mesenchymal transformation in malignant glioma by triggering RhoA activation. Cancer Res. 2013;73:5140–50.

91. Hoadley KA, Yau C, Wolf DM, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell. 2014;158:929–44.
92. Sonabend AM, Bansal M, Guarnieri P, et al. The transcriptional regulatory network of proneural glioma determines the genetic alterations selected during tumor progression. Cancer Res. 2014;74:1440–51.
93. Carro MS, Lim WK, Alvarez MJ, et al. The transcriptional network for mesenchymal transformation of brain tumours. Nature. 2010;463:318–25.
94. Calin GA, Dumitru CD, Shimizu M, et al. Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. Proc Natl Acad Sci U S A. 2002;99:15524–9.
95. Acunzo M, Romano G, Wernicke D, Croce CM. MicroRNA and cancer--a brief overview. Adv Biol Regul. 2015;57:1–9.
96. Costinean S, Sandhu SK, Pedersen IM. Src homology 2 domain–containing inositol-5-phosphatase and CCAAT enhancer-binding protein β are targeted by miR-155 in B cells of Eμ-MiR-155 transgenic mice. Blood. 2009;114(7):1374–82.
97. Balatti V, Nigita G, Veneziano D, et al. tsRNA signatures in cancer. Proc Natl Acad Sci U S A. 2017;114:8071–6.
98. Balatti V, Pekarsky Y, Croce CM. Role of the tRNA-derived small RNAs in cancer: new potential biomarkers and target for therapy. Adv Cancer Res. 2017;135:173–87.
99. Pekarsky Y, Balatti V, Palamarchuk A, et al. Dysregulation of a family of short noncoding RNAs, tsRNAs, in human cancer. Proc Natl Acad Sci U S A. 2016;113:5071–6.
100. La Ferlita A, Alaimo S, Veneziano D, et al. Identification of tRNA-derived ncRNAs in TCGA and NCI-60 panel cell lines and development of the public database tRFexplorer. Database. 2019;2019:baz115. https://doi.org/10.1093/database/baz115.
101. Kumar P, Anaya J, Mudunuri SB, Dutta A. Meta-analysis of tRNA derived RNA fragments reveals that they are evolutionarily conserved and associate with AGO proteins to recognize specific RNA targets. BMC Biol. 2014;12:78.
102. Kuscu C, Kumar P, Kiran M, et al. tRNA fragments (tRFs) guide Ago to regulate gene expression post-transcriptionally in a Dicer-independent manner. RNA. 2018;24:1093–105.
103. Wang R-S, Maron BA, Loscalzo J. Systems medicine: evolution of systems biology from bench to bedside. Wiley Interdiscip Rev Syst Biol Med. 2015;7:141–61.
104. Kirchmair J, Göller AH, Lang D, et al. Predicting drug metabolism: experiment and/or computation? Nat Rev Drug Discov. 2015;14:387–404.
105. Karlebach G, Shamir R. Modelling and analysis of gene regulatory networks. Nat Rev Mol Cell Biol. 2008;9:770–80.
106. Cohen DPA, Martignetti L, Robine S, et al. Mathematical modelling of molecular pathways enabling tumour cell invasion and migration. PLoS Comput Biol. 2015;11:e1004571.
107. Sizek H, Hamel A, Deritei D, et al. Boolean model of growth signaling, cell cycle and apoptosis predicts the molecular mechanism of aberrant cell cycle progression driven by hyperactive PI3K. PLoS Comput Biol. 2019;15:e1006402.
108. Barbuti R, Gori R, Milazzo P, Nasti L. A survey of gene regulatory networks modelling methods: from differential equations, to Boolean and qualitative bio-inspired models. J Membr Comput. 2020;2:207–26.
109. Sauer U, Hatzimanikatis V, Hohmann HP, et al. Physiology and metabolic fluxes of wild-type and riboflavin-producing Bacillus subtilis. Appl Environ Microbiol. 1996;62:3687–96.
110. Hellerstein MK. In vivo measurement of fluxes through metabolic pathways: the missing link in functional genomics and pharmaceutical research. Annu Rev Nutr. 2003;23:379–402.
111. Raman K, Chandra N. Flux balance analysis of biological systems: applications and challenges. Brief Bioinform. 2009;10:435–49.
112. Alaimo S, Rapicavoli RV, Marceca GP, et al. PHENSIM: phenotype simulator. PLoS Comput Biol. 2021;17(6):e1009069.
113. Alaimo S, Marceca GP, Ferro A, Pulvirenti A. Detecting disease specific pathway substructures through an integrated systems biology approach. Noncoding RNA. 2017;3:20. https://doi.org/10.3390/ncrna3020020.

RNA-seq Fusion Detection in Clinical Oncology

9

Dale J. Hedges

Abstract

Gene fusions play a prominent role in the oncogenesis of many cancers and have been extensively targeted as biomarkers for diagnostic, prognostic, and therapeutic purposes. Detection methods span a number of platforms, including cytogenetics (e.g., FISH), targeted qPCR, and sequencing-based assays. Before the advent of next-generation sequencing (NGS), fusion testing was primarily targeted to specific genome loci, with assays tailored for previously characterized fusion events. The availability of whole genome sequencing (WGS) and whole transcriptome sequencing (RNA-seq) allows for genome-wide screening for the simultaneous detection of both known and novel fusions. RNA-seq, in particular, offers the possibility of rapid turnaround testing with less dedicated sequencing than WGS. This makes it an attractive target for clinical oncology testing, particularly when transcriptome data can be multipurposed for tumor classification and additional analyses. Despite considerable efforts and substantial progress, however, genome-wide screening for fusions solely based on RNA-seq data remains an ongoing challenge. A host of technical artifacts adversely impact the sensitivity and specificity of existing software tools. In this chapter, the general strategies employed by current fusion software are discussed, and a selection of available fusion detection tools are surveyed. Despite its current limitations, RNA-seq-based fusion detection offers a more comprehensive and efficient strategy as compared to multiple targeted fusion assays. When thoughtfully employed within a wider ecosystem of diagnostic assays and clinical information, RNA-seq fusion detection represents a powerful tool for precision oncology.

Introduction

Importance of Fusions for Clinical Oncology

The significance of oncogenic fusions in clinical oncology has been well established, both in terms of their role in oncogenesis and their use as biomarkers of diagnostic, prognostic, and therapeutic significance [1]. Classic or "canonical" gene fusions in cancer are typified by the in-frame conjunction of two (or more) amino acid-coding frames from distinct genes, yielding chimeric proteins with oncogenic properties (Fig. 9.1). The term "fusion," is also commonly extended to

D. J. Hedges (✉)
Department of Pathology, St. Jude Children's Research Hospital, Memphis, TN, USA
e-mail: Dale.Hedges@STJUDE.ORG

© Springer Nature Switzerland AG 2022

A. Laganà (ed.), *Computational Methods for Precision Oncology*, Advances in Experimental Medicine and Biology 1361, https://doi.org/10.1007/978-3-030-91836-1_9

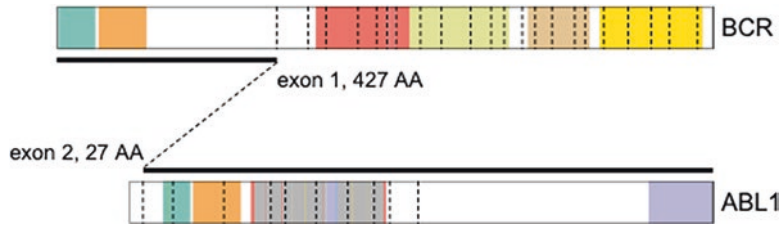

Fig. 9.1 BRC-ABL1 fusion. An example of one observed BCR-ABL1 fusion product, joining exon1 of BCR with exon 2 of ABL1. (Image generated by ProteinPaint (https://proteinpaint.stjude.org/) [3])

molecular phenomenon such as "promoter-swaps" and related regulatory region alterations. In the course of this chapter, both canonical fusions, as well as the larger class of "fusion-like" events, will be discussed. The nomenclature and classification of fusions are briefly addressed in section "Nomenclature of Fusions and Related Phenomena".

While oncogenic fusions contribute to a number of adult cancers, they exhibit a higher prevalence among pediatric cancers, where they comprise a significant fraction of driver alterations in common childhood leukemias. Commonly observed driver fusions include *ETV6-RUNX1*, which is present in approximately 25% of pediatric acute lymphoblastic leukemia (ALL) cases diagnosed between the ages of 2 and 10 [2]. Fusion drivers are also commonly found in pediatric brain tumors and solid tumors, such as *BRAF-KIAA1549* in astrocytoma and *PAX3-FOX01A* in rhabdomyosarcoma.

For the most prevalent and well-characterized fusions, robust targeted molecular assays are readily available based on traditional cytogenetic (e.g., FISH) and PCR-based technologies. Since resource availability typically constrains the number/scope of tests that can be performed on a given sample, targeted assays are often preferred, selected based upon initial clinical, histopathological, and/or laboratory findings. Although this approach has enjoyed considerable success, especially for common fusion lesions, there remains increased risk for false negatives when a patient exhibits atypical clinical and/or molecular presentations. Moreover, even in the case of the most prevalent fusion types, rare junctions between the gene partners can remain undetected, simply due to their falling outside the boundaries of a given targeted assay's design parameters. Targeted PCR-based assays, for example, are generally based on the observed distribution of known fusion junction sites and may not comprehensively cover all possible oncogenic combinations between two partner genes. The difficulty of exhaustively screening a pair of fusion genes can increase with gene length, exon number, and other factors, such as interspersed homology, that can further constrain assay designs. The result is that fusions involving larger genes, or those with significant runs of homology, can be more susceptible to false negatives for atypical junctions. It is also the case that some genes are promiscuous fusion partners, forming chimeras with several partners, and each partner may be observed at different frequencies. One such example is BCR, which has been observed to partner with *ABL1*, *FGFR1, JAK2*, and *PDGFRA*, among others [4]. In these cases, it can be difficult to exhaustively screen all possible combinations without a more comprehensive transcriptome sequencing solution. Although not covered here, the Archer FusionPlex (Illumina) targeted fusion system, based on next-generation sequencing technology, represents one option that lies in between multiple independent tests and full transcriptome surveys.

Although each class of rare or otherwise atypical fusion may on its own represent only a small fraction of oncogenic fusions, in aggregate, these rare events comprise an appreciable portion of cases. It is for these less common, less well-characterized lesions that RNA-seq based approaches hold the most promise. In this chapter, the strategies behind RNA-seq fusion

detection are surveyed, along with the benefits and challenges of a technology that has demonstrable advantages over traditional fusion detection methods, but also one that is not without its limitations.

Nomenclature of Fusions and Related Phenomena

Apart from canonical, coding fusions producing chimeric proteins, related molecular phenomena are also categorized as fusions and targeted by detection software with some degree of regularity. The purpose of this section is not to indicate which molecular lesions should or should not be formally considered fusions. It is, however, important to be aware that such variation in usage exists. Consequently, for any given piece of software, it is imperative to understand the range of molecular phenomena targeted.

The categories of molecular lesions most commonly assessed by fusion software are summarized in Fig. 9.2, along with an indication of their usage. Proto-typical fusions, forming chimeric proteins of canonical coding genes, are reliably targeted by all fusion-focused software applications. Out-of-frame fusions formed from protein-coding transcripts are also commonly detected, but their relative prioritization vs. the prototype fusion may vary from one software to another. The topic of prioritization and ranking is discussed in section "Annotation, Prioritization, and Visualization". Fusions involving the untranslated regions (UTR) of transcripts or nearby regulatory regions, where "in-frame" vs. "out-of-frame" status is no longer applicable, are also commonly detected and reported. In the case of these events, however, both sensitivity and prioritization can vary considerably between detection tools, warranting caution if these types of events are expected to contribute to the phenotypes under examination. Chimeric transcripts in which one or both genes are noncoding RNAs are less consistently targeted, captured, and/or prioritized. Additional related phenomenon, such as internal tandem duplications (ITDs) exist at the periphery of fusion software targeting and are captured with considerably less regularity. ITDs are, in some respects, fusion-like events wherein the "chimera" is formed from the coding region of a single gene. While their status as bona fide proper fusions may be debatable, these clinically important lesions can sometimes fall outside the radar of short indel callers, standard structural variation analysis, and fusion detection software alike. A reliable detection strategy from transcriptome can alleviate the need for independent targeted testing for important

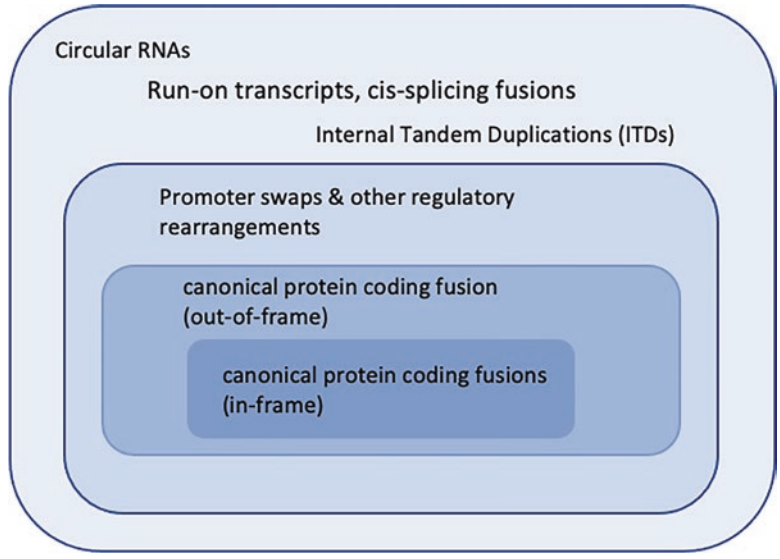

Fig. 9.2 The variety of molecular lesions targeted by fusion tools. In the center, in-frame fusions of protein-coding genes are universally targeted, as expected, while other molecular phenomena are captured and prioritized with differing levels of consistency across applications, indicated by lighter shading radiating out from the center

lesions such as *FLT3* ITDs in pediatric leukemias or *FGFR1* ITDs in brain tumors. *CICERO* [5] and Squid [6] are among the few fusion-oriented tools that explicitly target ITDs from RNA-seq data. It should be noted that, in the case of ITDs, DNA-based detection is currently more robust than RNA-based testing, where seasoned tools, such as Pindel [7], can reliably detect ITD events at known hot-spot loci when sufficient coverage is present. Also on the periphery of the fusion-related events are splicing abnormalities resulting in chimeric transcripts. These are typically observed in the context of adjacent or otherwise neighboring genes. In some instances, aberrant *cis*-splicing between nearby genes can form products that appear to be canonical fusion events without any corresponding indication of structural variation at the DNA level [8]. However, many adjacent genes yield run-on products, comprised of various chimeras of adjacent genes that are also observed within normal tissue expression, generating a multitude of red herring events among fusion detection output. In section "Selections from Available Software" strategies for filtering out innocuous run-on chimeras without, inadvertently, removing legitimate oncogenic fusions at such neighboring regions (e.g., P2RYR8-CRLF2), will be presented. Also included on the periphery of phenomena targeted by fusion-oriented software applications are circular RNA molecules, which are formed through back-splicing of exons (reviewed in [9]). These products, typically associated with no underlying DNA variation, can be difficult to distinguish from other structural variations and sequencing artifacts that can yield similar observable products.

Heterogeneity of Sequence Data Sources

Fusion workflows ultimately benefit from the fact that they can be initiated from multiple input sources and by a variety of platforms. Fusion testing can accept both DNA and RNA, and combined results from the parallel performance of RNA-seq and WGS workflows, while resource-intensive, can be superior to either platform on its own [10]. Other aspects of input data heterogeneity, however, contribute unwelcome complexity to the evaluation of fusion detection software. Even restricting our discussion to RNA-seq for the purposes of this chapter, a fusion detection algorithm can receive data from diverse laboratory workflows and sequencing instruments, contributing to differences in molecular fragment sizes, sequencing lengths, and input sequence quality. Sequence data quality can be further influenced by upstream sample handling and processing, which can vary by locale and are not always in the control of the laboratory testing facility. All this potential for input heterogeneity adds complexity to algorithm design and the assessment of software performance. Ideally, detection software would be evaluated and validated using the laboratory parameters (sample processing, laboratory protocol, sequencing instrument) that will be used in the actual testing environment, performed using a set of positive and negative controls that reflect the targeted population. The degree to which this ideal is achieved will vary with the nature of the testing cohort, including phenotype prevalence and the availability resources such positive control samples harboring an appropriately diverse group of fusion products. Because of these limitations, it is recommended that fusion workflow testing be supplemented with additional synthetically derived controls, such as Seraseq RNA fusion mix v4 (Seracare), which include titrations of 18 clinically relevant fusions. While not capturing the full range of diversity of sample quality, allele frequency, or lesions likely to be encountered during testing, this approach allows the entire workflow to be evaluated, beginning with initial laboratory processing. Several available tools also exist for simulating fusion read data (e.g., FusionSimulator Toolkit (https://github.com/FusionSimulatorToolkit/FusionSimulatorToolkit/wiki)). While such simulated data sets are not a robust substitute for true end-to-end testing beginning at the lab bench, they nevertheless increase the variety of lesions available to challenge software algorithms and evaluate parameterization.

Fusion Detection

Primary Signals

While the tissue sources and granular details of sequencing input may vary, the principal elements involved in fusion detection from RNA-seq can be distilled to a few essential elements. These are "junction" or "clipped" sequencing reads, anomalous "paired-end" reads, and, to a lesser extent, expression levels, as assessed by sequencing coverage depth at putative fusion loci. Each of these fundamental elements of fusion analysis are addressed in section "Fusion Detection Strategies". All of the above can be ultimately derived from primary sequence read inputs, such as the *.fastq* files generated by modern sequencers. These basic units of information determine the strategies available for RNA-seq-based fusion detection. Note that to detect abnormal read pairing, paired-end library preparations and sequencing runs are required. While some fusion tools accept unpaired, single-end sequencing input, current mainstream sequencing platforms such as the Illumina instrument series widely adopted in molecular pathology laboratories, should be run with paired-end libraries for optimal fusion detection. As explained below, paired-end reads increase the variety of information available to algorithms for fusion detection (see section "Anomalous Paired Read Signals"). Single molecule technologies, such as Oxford Nanopore and Pacific Biosystem's SMRT sequencing technology, which provide extended read lengths, are not covered here, as throughput and other considerations have so far limited adoption in oncology testing contexts. Nevertheless, the extended read lengths offered by these systems can be expected to enjoy increased adoption as these technologies are further advanced and refined. Long read analyses ultimately target the same basic signals of chimeric transcript junctions described in section "Fusion Detection Strategies", with the added advantage that the increased read lengths can significantly reduce the level of alignment ambiguity present with shorter read approaches.

Fusion Detection Strategies

De Novo Transcriptome Assembly

The first approach is only briefly addressed here, on account of its limited adoption among current fusion detection pipelines. In principle, given sufficient transcript read data, the entire repertoire of transcriptome products can be reconstructed, including fusion chimeras, from established de novo graph-based assembly algorithms, such as the one reviewed in [11]. Following de novo assembly, annotation and alignment with known transcripts can be used to reveal the presence of anomalous junctions supporting fusion events. More recent efforts to apply assembly strategies to fusion detection include TrinityFusion [12]. Despite considerable progress, however, tools that are primarily reliant on de novo assembly have thus far enjoyed limited adoption. For one, de novo assembly of mammalian genomes is resource-intensive from a computational standpoint and requires substantial amounts of memory and processing time. Many investigators may not have ready access to the requisite computing resources. Another contributing factor is the high degree of interspersed homology present within mammalian genomes, with over 50% of the human genome comprised of repetitive elements, not including homologous gene families, pseudogenes, segmental duplications, and related sources of homology. When this homology is combined alongside with structural genomic variation, inconsistent transcript coverage, and rampant sequencing artifacts, the difficulty of the assembly task and subsequent task of selecting candidate fusions is tremendously increased. The end result, in addition to increased resource demands, is generally poorer sensitivity and specificity, in comparison to alternative strategies discussed below (reviewed in [13]).

Mining Junction or "Clipped Reads"

Perhaps the most definitive RNA-seq evidence for a fusion event is the presence of a breakpoint within a sequencing read, where the sequence of one gene transcript abruptly transitions to that of another gene. These events (among others) are

represented by soft-clipping in modern aligner output, where the portion of the sequencing read that is not aligned to the reference sequence is indicated with an "S" within the CIGAR field of the SAM or BAM alignment file. The location of such a breakpoint can occur anywhere throughout the sequencing read, and such junctions are most informative when there is an adequate number of bases present on both sides of the breakpoint to allow for unique alignment to their respective source genes. One such ideal read is represented in Fig. 9.3. The more RNA-seq coverage of the fusion event, the more likely one or more junctions will be fully informative in this respect, drawing a clear link implicating the two genes in a fusion event. Often, however, the junction observed may fall near the end of a given sequence read. In these cases, the short "stub" of remaining sequence may not be sufficient for unique placement on reference or proposed fusion partner gene transcript. Even fragments of appreciable size (e.g., >15–20 bp) cannot always be uniquely located on the reference due to the presence of interspersed homology throughout the genome. These reads can nevertheless prove informative, as will be described further below, when combined with additional supporting information. It should be noted that the soft-clip signatures described above can be generated for a number of reasons, most of which have nothing to do with fusions. They can be formed, for example, when the sequence read extends into remnant anchor sequences from the library preparation process, particularly if the sequences have not been previously trimmed for such appendages. Soft-clip junctions can also occur when the sequence read quality diminishes to such an extent that the alignment can no longer be extended. A further challenge for the fusion detection workflow is to separate out these additional clipping signatures from those representing legitimate fusion support.

While the basic concept of a sequencing read spanning a breakpoint junction may be straightforward, their representation both within alignment files and visualization can be less so. For example, any given read spanning a fusion breakpoint junction will be anchored by its primary alignment to the location of one gene's transcript in the genome or its fusion partner gene's transcript. The remaining unaligned, soft-clipped sequence portion of the read may not be readily visualized in a browser, such as IGV, depending upon view settings and, even if it is visualized, there is no clear indication of where it might belong (Fig. 9.3). A separate supporting read for the same fusion event may anchored with its primary alignment the partner gene, and thus only be observed when browsing at that gene's location. Further discussion of fusion visualization is provided in section "Visualization".

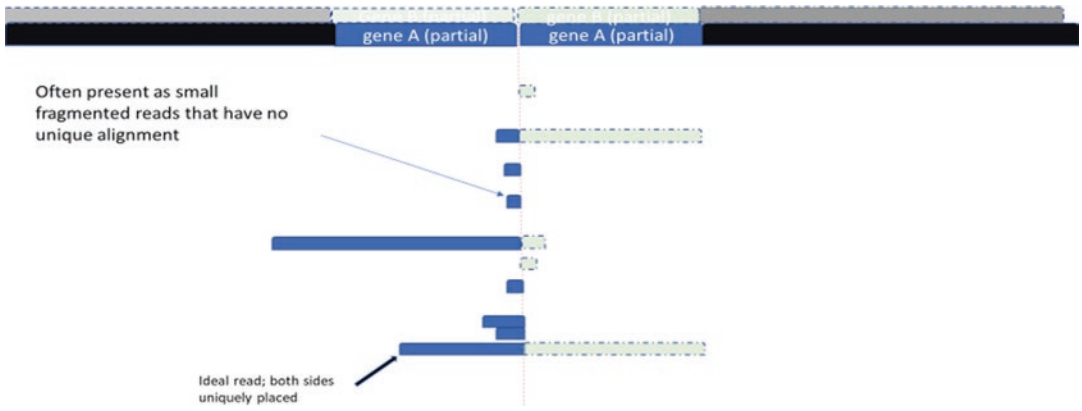

Fig. 9.3 Depiction of fusion evidence from inside a standard alignment viewer. Here, the gene currently in focus in the browser (gene A, blue), is super-imposed (gene B, light green with dashed lines). At the predicted breakpoint junction in gene A, evidence from fusion partner gene B may be visible only as soft-clipped extensions of gene B alignments. The mirror situation exists when gene B is in focus

These considerations extend beyond visualization, however. In both scenarios above, the position of the two fusion supporting reads in a coordinate-sorted BAM or SAM file will be anchored to their respective primary alignment and will not appear in proximity within the alignment file. Fusion detection algorithms must therefore work to aggregate evidence from distant genomic locations to support individual fusion breakpoint candidates. To make the matter more challenging, one side of a clipped read may be of insufficient length for unambiguous alignment to a reference sequence. One such orphaned "stub" is depicted in Fig. 9.3. Finally, the ubiquitous presence of interspersed homology, sequencing artifact, and normal structural variation in the population further frustrate the process of junction detection and support. In the case of interspersed homology, the seeding and extension of an alignment at a highly homologous region will result in an abrupt junction being formed once the alignment reaches the end of the homologous sequence and can no longer extend into the adjacent region. Chimeric artifact molecules generated during library preparation and sequencing can also introduce soft-clipped junction simply by juxtaposing sequence from two unrelated transcripts. Structural variation in the population, such as polymorphic mobile element insertions and segmental duplications, provides yet additional sources of junction signals. The challenges posed by phenomena that mimic fusion evidence, as well as some strategies for addressing them, are covered further in section "Artifacts, Technical Scoring, and Filtration".

Anomalous Paired Read Signals

As indicated above, the ability to obtain abnormal spans is dependent upon a paired-end sequencing run, as available on modern second-generation sequencers. In the context of WGS sequencing, anomalous insert sizes can be flagged for further examination based on the distance between where the reads align in the reference genome. If the distance falls significantly outside the expected fragment length distribution based on the laboratory sequencing protocol and/or the empirical distribution obtained from the sequencing data, the read pair may be flagged as a potential anomaly. A similar process is applied in RNA-seq for fusion detection, based instead on the expected locations of the ends within the existing annotated transcriptome units of expression. In principle, any RNA-seq mapping placing one end of the pair within an annotated gene A, and the other paired end in annotated gene B, can provide evidence for a putative a fusion event. As is so often the case in short read mapping, however, things are rarely that straightforward. Large, repetitive genomes, such as that of *Homo sapiens* and *Mus musculus*, present numerous opportunities for alignment error, yielding phantom chimeras having no underlying molecular basis. Exacerbating the situation, the library preparation and sequencing process itself can produce chimeric sequencing artifacts that combine fragments of two independent molecules. These aberrant molecules are faithfully registered by the sequencer but were never present within the original sample being tested. Such fragments are largely indistinguishable from legitimate chimeras. Both of the above processes can yield spurious fusion evidence. It falls upon the fusion detection software to make an attempt to distinguish legitimate fusion transcripts from artifactual ones, and the difficulty of this process represents one of the primary constraints in current RNA-seq based fusion testing (see section "Artifacts, Technical Scoring, and Filtration").

Expression Abnormalities

Apart from the detecting evidence supporting the primary sequence of chimeric transcripts, the magnitude of expression across participating genes can prove informative in some circumstances. This source of information is not as commonly drawn upon as those listed above, since the nature of RNA-seq coverage support can vary significantly across fusion classes, participating genes, and even among different combinations of the same fusion partners. In the case of a fusion at a high percentage within the tumor, the expression levels before and after breakpoints in the parent genes may indicate a marked transition. More reliably, the target genes of fusions involving regulatory regions, including promoters, may

exhibit dramatic increases in expression compared to a similar sample cohort. Such a shift in expression can, in the context of additional sequencing evidence suggestive of a regulatory fusion, add support for the event. Expression-level-based evidence warrants an amount of caution, however, since a great many factors can lead to aberrant expression within cancerous tissue. Examination of expression, while occasionally used within software algorithms, is most commonly done during visualization and assessment of the technical quality of specific fusion candidates, as described below.

Artifacts, Technical Scoring, and Filtration

Technical Scoring of Fusion Candidates

Following the filtering process, there remains significant work yet to be done by a complete fusion workflow. Depending on parameterization, a typical fusion detection application can yield anywhere from five predicted candidates to several thousand. The first order of business is typically to determine which putative fusions are best supported by the available sequence evidence. Most software tools provide some mechanism of scoring or rating fusions with respect to quality. At a basic level, virtually universal among fusion detection software outputs are counts of the number of supporting junction reads and the number of supporting anomalous paired ends, along with the proposed locations of the junction points within the parent genes. Beyond these staple software output fields, there is little consistency among programs. Most applications have some method for providing the precise sequence read records supporting the candidate fusion, either via supplemental output files or by providing their associated SAM id of the read for later retrieval. Additional output fields that may be provided for each candidate include the longest anchor sequence (primary alignment) supporting the event. Since longer primary alignment lengths reduce the chance that the read evidence is the result of spurious homology, longer anchor sequences are weighted as better supporting technical evidence. The consistency of the breakpoint position itself can also provide critical information. While normal sequencing error can lead to ambiguous placement of breakpoints within a target region, too much heterogeneity at the breakpoint location is often indicative of artifacts caused by interspersed homology. For each software package, these data fields are combined through a number of strategies to produce a technical score. In some software, such as Arriba, this score may be translated to a qualitive rating indicate “high confidence” and “low confidence” predictions [14].

CICERO, as one example, combines several data points to calculate its technical score, including both the number and length of supporting read alignments [5]. The technical scoring above will frequently be augmented with additional annotation and cross-referenced with known artifacts in a filtering procedure.

The result of this “filter” process may either remove candidates from the list or demarcate them in some fashion to indicate their risk of being either an artifact or otherwise undesirable candidate, such as assigning a “neighboring” status to gene partners at risk of forming run-on chimeras. The latter, label-based, approach is generally advised, as it offers more flexibility for attempting different filtering parameters in different contexts or during different phases of review.

While the technical merit of any fusion call is critical, determining which fusion candidates warrant additional scrutiny by a human may sometimes involve additional input from the annotation and prioritization schemes discussed in the next section. Although not every putative candidate from list of several thousand can receive detailed human curation, it is often tractable to prioritize fusions involving genes known to participate in oncogenic fusions. Methods to elevate these candidates, by the means discussed in the next section, can assist in focusing human curator efforts.

Annotation, Prioritization, and Visualization

Annotation and Prioritization

Before fusion candidates can be considered for their possible significance, the participating genes and/or regulatory sequences must be identified and appropriately annotated. In the case of canonical protein-coding fusions, an effort must also be made to determine whether the fusion remains in-frame. In the case of well-supported junctions, this can usually be done without error, making use of the predicted chimeric transcription. However, fusion breakpoints can often coincide with SNV-indel events as well as complex rearrangements in the vicinity of breakpoints. These events can occasionally confound automated frame prediction and require additional manual curation for correction. In general, any promising candidate fusion that is annotated as out-of-frame should receive additional scrutiny to determine if the detection software has made an annotation error. Apart from coding frame status prediction (where relevant), there is a host of additional information that can be layered on to a given fusion candidate to help assess its relevance. Among the more important pieces of information to be collected is which protein domains are retained in the final fusion product. In many cases, established fusions are characterized by the abnormal or unregulated activity of a domain from a source gene (e.g., tyrosine kinase domain). If a fusion's pathogenic mechanism is known to occur via the constitutive activity of a domain, and that domain is included in the predicted fusion product, the candidate warrants further scrutiny. The prediction could be the result of an artifact, but it is also possible that complex structural alterations in the vicinity of the fusion breakpoint have rendered the protein prediction incorrect.

In addition to the standard transcript information provided by genome releases such as GRCh38, a number of public resources are available to assist with this process (Table 9.1), and most fusion software packages draw upon one or more available resources in the course of their execution. In cases where the number of putative fusions is excessively high, the ability to screen events for those with the greatest likelihood to be of clinical significance is paramount, particularly when testing is conducted on restricted set of phenotypes with known candidate fusion genes. In a research setting, a wider selection of candidates would typically be considered; otherwise, there is the risk of missing an important, yet uncharacterized, fusion lesion.

The list of resources provided in Table 9.1 is not exhaustive, and new resources continue to become available. As a general rule in prioritization, previously described fusions with known clinical implication that are present in curated databases are prioritized higher than those that are unknown and/or have no documented pathogenic effect. Following known fusions, any fusion product containing gene that is known to participate in a clinically significant fusion would typically be ranked higher in priority than those predictions containing genes not known to participate in oncogenic fusions. Other factors that may impact prioritization include the coding status of genes, the in-frame or out-of-frame status of the predicted fusion product. As indicated previously, however, automated frame predictions are not universally accurate, and fusions involving noncoding sequence do not fall neatly under in-frame or out-of-frame categories. Prioritization, apart from technical scoring rank, is a process that can be highly individualized to the particular testing environment or research question at hand. This is typically an area where a customized approach suitable to the phenotypes under examination will, in most cases, be superior to a generic attempt to determine possible relevance.

Visualization

Visualization is employed for two main purposes during the fusion detection workflow. Initially, it can be used as part of the technical assessment process. In this instance, a genome viewer such as Integrative Genomics Viewer [15] can be employed to examine the read alignments sup-

Table 9.1 Public fusion annotation resources

Resource	Website
Tumor fusion gene data portal	https://www.tumorfusions.org/
ChimerDB	http://www.kobic.re.kr/chimerdb/
Mitelman database	https://mitelmandatabase.isb-cgc.org/
FusionGDB	https://ccsm.uth.edu/FusionGDB/
Fusion hub	https://fusionhub.persistent.co.in/
Archer quiver	http://quiver.archerdx.com/
COSMIC fusion resource	https://cancer.sanger.ac.uk/cosmic/fusion

porting a given fusion event to assess their quality and survey the local region for possible sources of artifact (e.g., low complexity sequence). As indicated previously, the process is not entirely straightforward and can involve navigating between both candidate genes to assess available evidence. When using IGV, it is recommended to turn on the option to view soft clips, which is not enabled by default. Third party tools are available that can take fusion read data as input and output a IGV input file. One such example is Clinker [16].

Visualization can also be used for prioritization and interpretations purposes, such as when assessing which protein domains remain in the final fusion product. While there is no stand-out leader in fusion visualization, there are several applications available. FusionEditor (https://proteinpaint.stjude.org/examples/fusioneditor.html), published as part of the *CICERO* package, offers one of the more visually appealing and intuitive means of interactively viewing fusion event, including representations on alternate isoforms of the participant genes [5]. It is currently limited to output obtained from *CICERO*, however, the output of other software tools could potentially be transformed to *CICERO* format. The Arriba software package also comes with its own visualization tool, which can be either automatically run or separately invoked with an included R script, *draw_fusions.R*. Although not interactive, Arriba visualization software produces publication quality pdf images, including Circos plot representations of fusion events (Fig. 9.4). The straightforward images produced provide most of the key information necessary for evaluating the impact of a given fusion event, including read support and indication of which protein domains are retained in the fused product. One drawback is that only a single annotated transcript is selected for visualization, although it is possible to compel the use of an alternate transcript by providing annotation that elevates the priority of the desired transcript. Additional visualization options are provided by shinyFuse, published as part of the recently published *annoFuse* application [17].

Available Fusion Detection Software

What follows is by no means an exhaustive listing of the many RNA-seq-based fusion detection tools. Rather, a curated list of tools is discussed that the author has personally observed to provide a reasonable balance of ease of deployment, run times, and, critically, combination of sensitivity and specificity. For a comprehensive examination of the performance attributes of numerous available packages, relative performance of 23 available software packages was extensively reviewed recently in [13]. Although both laboratory technologies and analytical approaches will rapidly date any specific suggestions, the tools listed in Table 9.2 are recommended for consideration when developing a fusion workflow.

Selections from Available Software

The tools listed in Table 9.2 all employ some combination of the strategies outlined in section "Fusion Detection Strategies", although with dif-

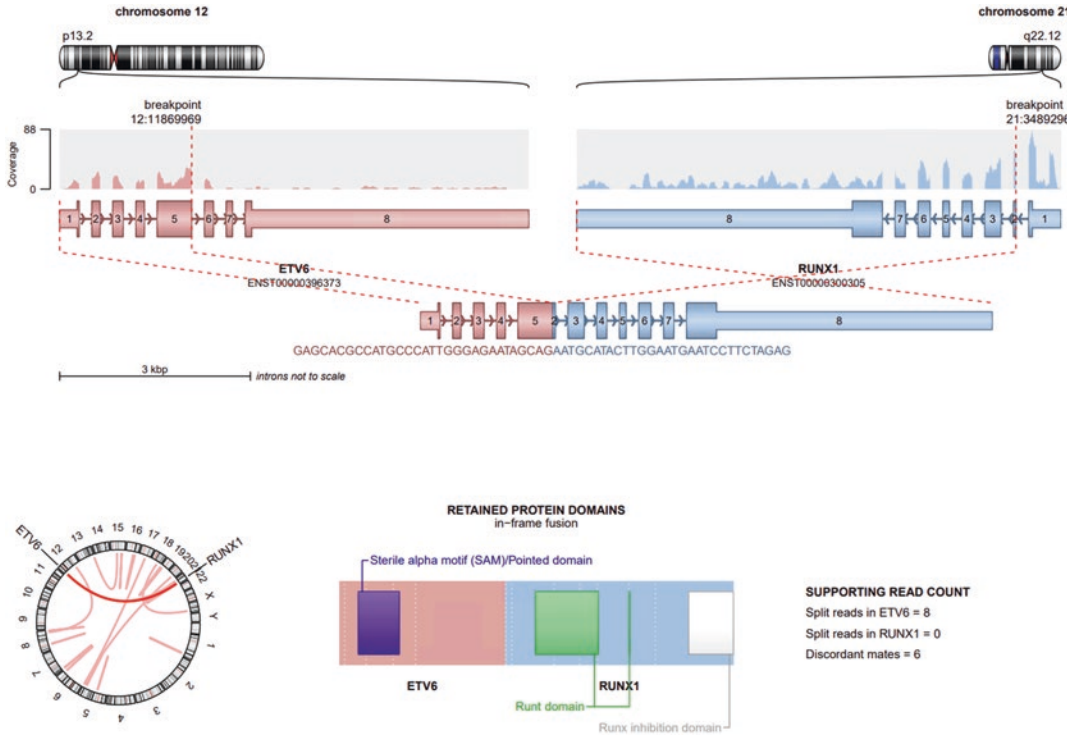

Fig. 9.4 Arriba viewer visualization of ETV6-RUNX1 fusion. An example ETV6-RUNX1 fusion event is depicted to demonstrate Arriba view output. In the upper portion of the image, the location of the fusion breakpoint is depicted with respect to each parent gene. Coverage information is simultaneously overlayed to aid in evidence interpretation. In the lower portion of the figure, the number of supporting reads is provided, along with a depiction of the domains retained in the final fusion product

Table 9.2 Selected fusion detection software

Resource	Reference
Arriba	[14]
CICERO	[5]
DRAGEN-RNA	Illumina
FusionCatcher	https://github.com/ndaniel/fusioncatcher
Pizzly	https://github.com/pmelsted/pizzly
STAR-fusion	[15]

ferences in implementations, points of emphasis, and methods for evaluating and ranking final candidates. One notable exception here is Pizzly, which begins with a k-mer-based pseudo-alignment approach, resulting in rapid run times and relatively low resource requirements (https://github.com/pmelsted/pizzly). Despite incurring a penalty of reduced sensitivity [13], its relatively low resource requirements make it an excellent candidate for addition to ensemble fusion workflows described in the next section, where it can provide additional support for fusions detected by one or more independent applications. One additional caveat, however, is that development on Pizzly has not been active for >3 years, and it may be increasingly difficult to port to newer genome releases and/or annotation sets. STAR-Fusion, Arriba, and DRAGEN-RNA stand out as well-rounded applications, offering a good balance of speed and sensitivity. FusionCatcher and

CICERO excel at detecting non-canonical fusion events often missed by alternative tools, with *CICERO* additionally targeting ITD events that are important in a number of cancers. The potential downside of the latter two applications is that they tend to be more resource-intensive and incur longer run times than some of the other candidates. The *CICERO* authors have made a cloud-based version of their tool available in order to partly mitigate this issue and provide a dynamic fusion viewer, *FusionEditor,* adding value to availability of *CICERO* output [5]. With the exception of Pizzly, the above software packages can perform fairly well on their own. Nevertheless, it is highly recommended that one or more be run in combination as part of an ensemble approach, such as that described in the next section.

Ensemble Strategies

In some respects, most modern fusion detection software employs a type of ensemble strategy, combining evidence from multiple, independent algorithmic approaches to identify and assess candidate fusions. This is particularly true with software such as FusionCatcher, which uses multiple independent alignment tools and a diverse array of heuristics. The ensemble strategies discussed in this section, however, take the process still further by independently running different fusion detection tools and then combining the resulting information to determine fusion support. Such ensemble efforts are not new and have been successfully employed for snv-indel calling [16]. The two principal disadvantages of ensemble strategies are that they expand the number of computational resources required, typically resulting in longer run times. The other is that, depending on the method of evidence combination, an ensemble workflow can easily magnify RNA-seq fusion detection's already substantial with excessive numbers of predicted fusions. A simple union of results, for example, would vastly increase the number of potential candidates and, without further refinement, prove untenable for downstream interpretation. In contrast, requiring multiple callers to support a given event could err in being overly conservative, missing legitimate fusions by failing to capitalize on the possible strength of one caller in a particular edge case. For any in-house workflow development, a customized mechanism for weighing output from different aligners and prioritizing results may be required. Efforts continue to be made in the arena of fusion. One notable example of an existing ensemble fusion caller is the *nf-core/rna-fusion* workflow (https://doi.org/10.5281/zenodo.1400710), based on the Nextflow domain-specific language (https://www.nextflow.io/). As the deployment of multiple software tools in a local high performance computing environment can is not always a smooth or simple process, nf-core/rna-fusion offers a container-based approach that is amenable to several popular commercial cloud systems (e.g., AWS, Google Cloud, and Microsoft Azure). The application also includes additional tools for combining and visualizing fusion evidence. While the specific tools used in the standard distribution may not be appropriate for all testing scenarios, the open-source framework provided can serve as a springboard for established custom workflows using alternative detection tools. Due to the complexity of the RNA-seq fusion detection, with no one algorithm or application able to accommodate the full range of calling challenges, the pursuit and refinement of new ensemble workflows will continue to develop.

Conclusion

Comprehensive fusion screening remains a team effort, with no existing software able to regularly yield results that are directly amenable to clinical interpretation without significant prior human curation. Local bioinformatics personnel are typically required to prioritize and assess the technical merit of fusion evidence for particular events before passing candidates on to molecular pathologists or clinicians for further evaluation and clinical interpretation. The number of candidates generated by the typical whole transcriptome workflow can be daunting. The process of assessing technical support for fusions is, however,

greatly facilitated by the availability of supporting DNA evidence from WGS for cross-reference. The financial, technical, and personnel resources required for combined RNA-seq and WGS analysis remain a substantial barrier to widespread adoption of the parallel approach, but the benefits are nevertheless evident [10]. Even with known limitations, whole transcriptome sequencing nevertheless represents a game-changing advance in oncogenic fusion screening. Testing facilities can achieve greater efficiencies, combining multiple tests into a single procedure, and patients with rarer fusion lesions will be better served. RNA-seq is nevertheless not without its limitations, and it remains only one component of an overall testing ecosystem. Information from parallel laboratory testing, such as cytogenetics and immunohistochemistry, can be used to both corroborate RNA-seq results and raise flags when a potential fusion may be missed. As both laboratory and computational technologies improve, RNA-seq-based fusion screening, with or without accompanying whole genome sequencing, is expected to increasingly become the method of choice for oncogenic fusion screening.

References

1. Mitelman F, Johansson B, Mertens F. The impact of translocations and gene fusions on cancer causation. Nat Rev Cancer. 2007;7:233–45.
2. Sun C, Chang L, Zhu X. Pathogenesis of ETV6/RUNX1-positive childhood acute lymphoblastic leukemia and mechanisms underlying its relapse. Oncotarget. 2017;8(21):35445–59. https://doi.org/10.18632/oncotarget.16367.
3. Zhou X, Edmonson MN, Wilkinson MR, Patel A, Wu G, Liu Y, Li Y, Zhang Z, Rusch MC, Parker M, Becksfort J, Downing JR, Zhang J. Exploring genomic alteration in pediatric cancer using ProteinPaint. Nat Genet. 2016;48(1):4–6. https://doi.org/10.1038/ng.3466. PMID: 26711108; PMCID: PMC4892362.
4. Peiris MN, Li F, Donoghue DJ. BCR: a promiscuous fusion partner in hematopoietic disorders. Oncotarget. 2019;10(28):2738–54.
5. Tian L, Li Y, Edmonson MN, et al. CICERO: a versatile method for detecting complex and diverse driver fusions using cancer RNA sequencing data. Genome Biol. 2020;21:126.
6. Ma C, Shao M, Kingsford C. SQUID: transcriptomic structural variation detection from RNA-seq. Genome Biol. 2018;19:52. https://doi.org/10.1186/s13059-018-1421-5.
7. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics. 2009;25(21):2865–71.
8. Akiva P, Toporik A, Edelheit S, Peretz Y, Diber A, Shemesh R, Novik A, Sorek R. Transcription-mediated gene fusion in the human genome. Genome Res. 2006;16(1):30–6. https://doi.org/10.1101/gr.4137606. Epub 2005 Dec 12. PMID: 16344562; PMCID: PMC1356126.
9. Santer L, Bär C, Thum T. Circular RNAs: a novel class of functional RNA molecules with a therapeutic perspective. Mol Ther. 2019;27(8):1350–63.
10. Rusch M, Nakitandwe J, Shurtleff S, Newman S, Zhang Z, Edmonson MN, Parker M, Jiao Y, Ma X, Liu Y, Gu J, Walsh MF, Becksfort J, Thrasher A, Li Y, McMurry J, Hedlund E, Patel A, Easton J, Yergeau D, Vadodaria B, Tatevossian RG, Raimondi S, Hedges D, Chen X, Hagiwara K, McGee R, Robinson GW, Klco JM, Gruber TA, Ellison DW, Downing JR, Zhang J. Clinical cancer genomic profiling by three-platform sequencing of whole genome, whole exome and transcriptome. Nat Commun. 2018;9(1):3962.
11. Martin JA, Wang Z. Next-generation transcriptome assembly. Nat Rev Genet. 2011;12(10):671–82. https://doi.org/10.1038/nrg3068.
12. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29(7):644–52.
13. Haas BJ, Dobin A, Li B, et al. Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. Genome Biol. 2019;20:213. https://doi.org/10.1186/s13059-019-1842-9.
14. Uhrig S, Ellermann J, Walther T, Burkhardt P, Fröhlich M, Hutter B, Toprak UH, Neumann O, Stenzinger A, Scholl C, Fröhling S, Brors B. Accurate and efficient detection of gene fusions from RNA sequencing data. Genome Res. 2021;31(3):448–60. Jan 13:gr.257246.119
15. Haas BJ. STAR-Fusion code and documentation on GitHub 2019. Available from: https://github.com/STAR-Fusion/STAR-Fusion/wiki.
16. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56–65.
17. Schmidt BM, Davidson NM, Hawkins ADK, Bartolo R, Majewski IJ, Ekert PG, Oshlack A. Clinker: visualizing fusion genes detected in RNA-seq data. Gigascience. 2018;7(7):giy079.

10 Computational Resources for the Interpretation of Variations in Cancer

Grete Francesca Privitera, Salvatore Alaimo, Alfredo Ferro, and Alfredo Pulvirenti

Abstract

A broad ecosystem of resources, databases, and systems to analyze cancer variations is present in the literature. These are a strategic element in the interpretation of NGS experiments. However, the intrinsic wealth of data from RNA-seq, ChipSeq, and DNA-seq can be fully exploited only with the proper skill and knowledge. In this chapter, we survey relevant literature concerning databases, annotators, and variant prioritization tools.

Introduction

Experiments from Next-Generation Sequencing (NGS) technologies, such as RNA-Seq, ChipSeq, and DNA-Seq, represent a common ground for most of the research conducted in biomedical laboratories, and the literature describing molecular variants and their associated treatments is growing rapidly. A downside of such experiments is the vast amount of data they produce. DNA-Seq experiments, for example, produce an enormous amount of variants that need to be interpreted to identify potential disease causal genes and driver or passenger mutations. Furthermore, many somatic alterations identified by whole-exome and gene panel sequencing are likely passenger events with no influence on the patient's prognosis or response to therapy.

Molecular pathologists need to summarize their findings on molecular reports without extensive literature curation. To help them, the Association for Molecular Pathology [1], the American Society of Clinical Oncology [2], and the College of American Pathologists [3] (AMP/ASCO/CAP) have published structured somatic variant clinical interpretation guidelines that specifically address diagnostic, prognostic, and therapeutic implications.

Moving in this direction, many tools and databases have been released. Variant explanatory databases usually provide information ranging from gene and phenotype descriptions to pathogenicity assessment, treatment, and drug resistance.

This chapter reviews the resources to interpret variations in cancer, focusing on databases, annotators, and variants prioritization tools. In particular, the Databases section provides a comprehensive survey of resources used to anno-

G. F. Privitera
Department of Physics and Astronomy, University of Catania, Catania, Italy

Department of Clinical and Experimental Medicine, Bioinformatics Unit, University of Catania, Catania, Italy

S. Alaimo · A. Ferro · A. Pulvirenti (✉)
Department of Clinical and Experimental Medicine, Bioinformatics Unit, University of Catania, Catania, Italy
e-mail: alfredo.pulvirenti@unict.it

© Springer Nature Switzerland AG 2022

A. Laganà (ed.), *Computational Methods for Precision Oncology*, Advances in Experimental Medicine and Biology 1361, https://doi.org/10.1007/978-3-030-91836-1_10

tate the variants with essential information, such as their druggability, their interpretation, and their importance in a specific disease. The Annotator section gives a close look at the annotation tools, which are strategic resources to automatically and reliably annotate variants. The chapter ends with the Variant Prioritization section, which introduces several tools used for variant prioritization and their applications to elucidate the pathogenicity of variants in cancer.

Databases

In this section, we describe variant annotation resources and databases available in the literature. We provide a brief description for each resource with the details on how to access the data (Table 10.1).

VICC Project

The Global Alliance for Genomic and Health (GA4GH) [4] is an international, nonprofit alliance formed in 2013 to accelerate the potential of research and medicine to advance human health. It brings together 500+ leading organizations working in healthcare.

Since variant databases were often redundant and with a limited-access, this alliance aimed at allowing secure sharing of genomic and health data to boost advances in knowledge. In 2016, it supported the creation of the VICC (Variant Interpretation for Cancer Consortium) project [5]. This project brings together the institutions that are developing cancer variant interpretation databases.

The most crucial objective of the VICC project is to increase the confidence of the variant to avoid redundancy and fill the gaps. The institutes need to have permissive licenses to share the variant interpretations. Notwithstanding, at the beginning of the project, the variants used were derived from published findings. Now, they provide comprehensive reports of clinically relevant variants along with their diagnostic, prognostic, and treatment data in patients. Data can be bulk downloaded or searched through proper APIs or web interfaces equipped with query systems.

CIViC

Clinical Interpretation of Variants in Cancer [6] (CIViC) is an open-source database. It was first released in 2015 to centralize variant information. The database consists of data concerning the therapeutic, prognostic, diagnostic, and predisposing relevance of inherited and somatic variants of all types.

The clinical interpretations are displayed for a specific gene, variant, disease, and clinical action. Each clinical interpretation is annotated with an evidence type and an evidence level. The former denotes whether a variant is predictive of response to therapy, prognostic, diagnostic, or predisposing for cancer. The latter indicates whether a variant has an established clinical utility from a case study, has preclinical evidence, or represents only inferential evidence.

The database provides five different evidence levels: (i) A—*Validated association*: those which have proven/consensus associations in human medicine. (ii) B—*Clinical evidence*: associations supported by clinical trials or other primary patients data. (iii) C—*Case study*: variants found in case reports from clinical journals. (iv) D—*Preclinical evidence*: associations supported by in vivo or in vitro models. (v) E—*Inferential association*: indirect evidence.

Curators determine the variants, and editors review them. To be accepted, every variant submitted requires the agreement between at least two independent contributors, where at least one has to be an expert editor. CIViC also accepts public contributions, but these need to be controlled and revised wisely by experts.

Furthermore, a panel of clinical domain experts provides independent guidance on the resource's development and accuracy.

The latest release of the database at the time of writing, contains 2587 variants, 431 gene mutations, and 455 drugs for 7570 evidence items. Additional variant information is imported from

Table 10.1 Databases information

	ClViC	*PharmGKB*	*CGI*	*OncoKB*	*DoCM*	*PMKB*
Evidence Level	YES	YES	YES	YES	NO	NO
Genome version	GRCh37	GRCh37	GRCh37	GRCh37/ GRCh38	GRCh37	GRCh37
License	The Creative Commons Public Domain Dedication or 'CC0' license	Creative Commons Attribution-ShareAlike 4.0 International License	License derived from the databases that it aggregates	GNU Affero General Public License v3.0	Creative Commons Attribution 4.0 International License	Creative Commons Attribution 4.0
VICC	YES	YES	NO	YES	NO	YES
Resistance Level	NO, but it is described	No, but It states the efficacy or toxicity of the drug	No, but it is described	YES	NO	NO
Open Source	YES	YES	YES	YES	YES	YES
Manually Curated	YES	YES	YES	YES	YES	YES
API	YES	YES, Beta Version	YES	YES	YES	YES
Website	https://civicdb.org/home	https://www.pharmgkb.org/	https://www.cancergenomeinterpreter.org/home	https://www.oncokb.org/	http://www.docm.info/	https://pmkb.weill.cornell.edu/
Genes	431	600 (in 2014)	765	682	132	610
Variants	2586	24,021 (variants annotations)	5601	5616	1364	2247
Drugs	455	709	310 (in 2018)	90	/	/

other resources such as ClinVar [7], COSMIC [7, 8], and ExAC [9].

CIViC offers APIs allowing its integration into clinical reports for gene panel, exome, whole-genome, and RNA sequencing of a tumor. Moreover, through https://civicdb.org/releases, it is possible to download directly nightly and monthly releases both as TSV and VCF files.

The database is now part of the VICC project to ensure long-term sustainability and cooperation with other databases.

PharmGKB

Pharmacogenomics Knowledgebase (PharmGKB) [10, 11] is a resource that collects curated information about the following: (i) potentially actionable gene-drug association; (ii) pathways, which are evidence-based diagrams depicting the pharmacokinetics (PK) or pharmacodynamics (PD) of a drug with relevant pharmacogenetic (PGx) associations; (iii) Clinical Guideline Annotations for drugs dosing guidelines, annotated drug labels, and genotype-phenotype relationship.

PharmGKB was created to help precision medicine efforts and understand how genetic variation contributes to different responses to drugs. Gene-drug-disease relationships are extracted from the literature using manual curation and natural-language-processing techniques. Every clinical annotation is linked to PubMed identifiers.

The PharmGKB databases were built to provide a freely available collection of high-quality genotypic and phenotypic data retrieved from pharmacogenetics and pharmacogenomics studies.

PharmaGKB pathways are drug centered to highlight how interacting genes can affect both drug metabolism and drug response.

Like in CIViC, each clinical annotation is associated with a level of evidence score that measures the association's confidence as determined by the PharmGKB curators. This score is based on existing replication of the association in connection to a p-value, odds ratio, and other relevant indexes. There are different levels of annotations.

- Level 1A is associated with variants for which the PharmGKB staff is aware of clinical implementation tests or deployments.
- Level 2 annotations are for variant-drug combinations with moderate evidence of an association. In particular, level 2A is for VIP (Very Important Pharmacogene) genes that are well documented.
- Level 3 annotations are based on a single significant study for a variant-drug combination or variant-drug annotation evaluated in multiple studies but lacking clear evidence of an association.
- Level 4 annotations are based on a case report, on a biologically plausible study even if it does not achieve significance, or is based on in vitro, molecular, or functional assay evidence.

For the variants with no associations in the literature, no clinical annotations are reported.

PharmGKB supports several clinically relevant projects, like the Clinical Pharmacogenetics Implementation Consortium (CPIC) that provides drug-dosing guidelines based on the individual genotype.

The PharmGKB web site, http://www.pharmgkb.org, provides genotype, molecular, and clinical knowledge integrated into pathway representations and Very Important Pharmacogene (VIP) summaries with links to additional external resources.

Some information contained in the database is retrieved from other repositories, like drug names and structures from Drugbank [12], and gene symbols and names from the Human Genome Nomenclature Committee (HGNC) [13].

PharmGKB provides an interactive interface that allows inspecting pathway information, genes, and related drugs that can be downloaded and used in pathway analysis. Currently, it contains 149 curated pathways. Every pathway presents a summary to describe the pathway graphics' content, limitations, and controversial features that are not shown in the representation. The rep-

resentation is a consensus of the opinions of the authors. Currently, these pathways are constructed by hand as graphic images.

OncoKB

OncoKB is a comprehensive precision oncology database and is part of the VICC project. It is publicly available through an interactive website and integrated into the cBioPortal [13–15] for cancer genomics. Genomic alterations are annotated with their biological effects and clinical implications.

OncoKB supports treatment decisions collected by oncologists, evidence-based information about individual somatic mutations and structural alterations. The content of OncoKB is supervised by a physician and cancer biologists. Moreover, thanks to a continuous dialogue with the scientific and medical community, it integrates clinical best practices.

OncoKB includes biological, clinical, and therapeutic information curated from multiple unstructured information resources, including guidelines and recommendations derived from FDA (Food and Drug Administration) labeling, NCCN (National Comprehensive Cancer Network) guidelines [16], other disease-specific expert and advocacy group recommendations, and the medical literature.

It also stores information about FDA-approved therapies, drugs under evaluation in clinical trials, and data about negative clinical results for specific drug-biomarker pairs.

The information is organized by gene, alteration, tumor type, and clinical implications. OncoKB contains 5293 alterations and 682 genes.

It comprises a system of evidence classification levels that communicates the mutation's clinical utility to the user depending on tumor origin.

The evidence is organized into the following levels:

- Level 1 includes genes for which the FDA has recognized specific alterations as predictive of response to an FDA-approved drug for a particular disease context.
- Level 2A includes alterations that are not FDA-recognized biomarkers but are considered standard-of-care predictive biomarkers of response to an FDA-approved therapy in specific cancer types. These alterations are highlighted in the expert panel guidelines, such as the NCCN Compendium or ASCO Clinical Practice Guidelines. Level 2A associations involve rare cancer types or small subpopulations of common cancers and therefore unsuitable for randomized phase III clinical studies.
- Level 2B includes alterations that are standard predictive biomarkers of drug sensitivity in other tumor types.
- Level 3A includes mutations that are candidate predictive biomarkers of drug response based on off-label use of FDA-approved drugs or investigational agents not yet FDA-approved for any indication.
- Level 3B applies to all tumor types for which clinical activity of an off-label drug has not been yet reported.
- Level 4 alterations are candidate predictive biomarkers of response to either FDA-approved or investigational agents based on promising laboratory research data but no relevant and robust clinical data.

A second type of classification, related to therapy-resistant mutations with three levels of evidence, is also available:

- Level R1 includes mutational events associated with drug resistance.
- Levels R2 and R3 include alterations that have hypothetical therapeutic implications and alterations predictive of drug resistance based on clinical or biological data, respectively. However, their usage in guiding treatment decisions is still considered investigational.

The majority of the alterations in OncoKB have curated biological effects and are classified as oncogenic, but their actionability is unknown.

It is possible to leave suggestions for new alterations analyzed by the scientific team of OncoKB and potentially integrated into the database.

Through the VICC project, OncoKB participates in both ClinGen [17] and the Global Alliance for Genomic Health (GA4GH) to promote harmonization of variant annotation.

All the alterations presented in the database are identified by their recurrence, from public variant databases, and by prior knowledge available in the literature. Biological and clinical therapeutic implications of alterations are curated from several public resources, including disease-specific treatment guidelines, abstracts from major conference proceedings, such as ASCO, European Society For Medical Oncology (ESMO) [18], and American Association for Cancer Research (AACR) [19], and the scientific literature through PubMed.

DoCM: Database of Curated Mutations in Cancer

DoCM is a cancer mutation open-source database. It helps the research community to aggregate, store, and track biologically important cancer variants. DoCM facilitates the aggregation of gene/variant information for variants with prognostic, diagnostic, predictive, or functional roles.

DoCM is licensed under the Creative Commons Attribution license (CC BY 4.0), allowing academic and industry researchers to freely access the content.

New variants can be added to DoCM, but they must be formatted and standardized. After submission, they are reviewed and evaluated by DoCM editors for inclusion. DoCM provides easy access to a current and accurate list of functionally important cancer variants with clear provenance, based on peer-reviewed journal citations. The content of DoCM may be accessed through a web interface or a documented application programming interface (API).

The data model and batch submission process used by DoCM places it at a critical intersection between the two major trade-offs of curated resources: comprehensiveness of variants and curation burden.

To be included in DoCM, variants need to be supported by peer-reviewed literature or expert opinions indicating their relevance to cancer or a cancer subtype. Furthermore, variants, single nucleotide substitutions (SNSs), and insertions and deletions (indels) should have published evidence of clinical relevance, such as prognostic, diagnostic information, or response data for targeted therapies.

Additionally, variants whose etiology in cancer has been established by functional experimentation, in either cell lines or model organisms, are included. Finally, variations that have been observed in large-scale sequencing efforts as being significantly associated with a particular cancer type are included in the resource.

DoCM currently holds 1364 variants of 132 genes across 122 cancer subtypes, based on 876 publications.

PMKB

The Precision Medicine Knowledge Base (PMKB) is a database of variant interpretation for oncology developed at Weill Cornell in collaboration with pathologists. All accepted interpretations need to be approved by a board of molecular pathologists. The PMKB interpretations can be accessed either directly using a web interface or through an API. PMKB is designed to provide data granularity to automatize the retrieval of interpretations and provide a convenient experience for pathologists.

Every interpretation in PMKB requires three associations: gene-variant descriptions, cancer and tumor-type descriptions, and tissue-type descriptions. Each interpretation includes the textual interpretation supported by literature and a numeric level indicating how clinically actionable the interpretation is. The variants are described using the Human Genome Variation Society (HGVS) standards, using the gene and its associated variant categories like SNVs, copy number alterations, and gene fusions. SNV and

indels can be described through protein-change and DNA-change notation or the gene region-based description. Gene regions can be further divided into specific codons, specific exons, and the entire gene.

PMKB automatically retrieves specific gene region information from Ensembl based on Ensembl's canonical transcript for a gene and its GRCh37-based API.

Separating variant descriptions into discrete fields facilitates the process of matching them against existing annotations. PMKB's API takes a variant's HGVS protein notation as input and matches that variant against multiple levels of variant descriptions, returning all relevant interpretations. These matches are classified in order of their specificity.

The user can choose a specific tumor type and tissue type for which interpretations are returned. The PMKB possesses a multi-user interface for entering, editing, browsing, and querying variants. The user can search any gene symbol or enter a COSMIC Gene ID. Once a gene is chosen, the user can select the variant type and other mutation details. Once the user submits the variant, PMKB adds region information using the Ensembl's API.

The interface for entering interpretations allows users to first select from any gene in PMKB that has at least one variant description. This feature dramatically facilitates applying a single interpretation to many variants. It also allows the user to easily modify an existing interpretive comment and apply newly edited comments to one or more variants for a gene. Every variant needs at least one PubMed citation. PMKB contains 2247 alterations for 610 genes.

PMKB has three different user levels:

- A high-level "approver," who can review and approve others' entries; this role is reserved to the PMKB's molecular pathologists.
- Standard users, who can submit edits that must be approved and eventually modified by the approver pathologist.
- Guests who cannot make changes.

All interpretations are available free of charge to the community under the Creative Commons Attribution 4.0.

CGI

Cancer Genome Interpreter (CGI) [20] is a free platform that annotates all tumor variants that constitute state-of-the-art biomarkers of drug response organized using other clinical evidence. It comprises 5601 validated oncogenic alterations, 1631 biomarkers of drug response, and 765 cancer genes. It is continuously updated by a board of medical oncologists and cancer genomics experts. It is available through an API or a web interface. The catalog was obtained using various resources such as manually curated resources, literature, and bioinformatic analysis of large tumor cohorts. Each gene is annotated with its mode of action in tumorigenesis based on experimentally verified sources or in silico prediction.

Each entry includes the name of the driver gene, the disease(s) it drives, the alteration type, the source of this information, the context in which these alterations are tumorigenic, and the gene mode of action in cancer. The CGI platform with its catalog identifies all known and likely tumorigenic genomic alterations and annotates the ones that constitute state-of-the-art biomarkers of drug response, organizing them based on clinical evidence.

CGI provides user-friendly reports. With a pan-cancer cohort of 6792 tumors sequenced, the CGI authors noted that only 5360 (916 unique variants) of the 44,601 protein-affecting mutations (PAMs) found in cancer genes appear in this catalog. In other words, 88% of all PAMs that affect cancer genes in this cohort are currently of uncertain significance for tumorigenesis.

Furthermore, CGI assesses the unknown variants' tumorigenic potential, especially for the genes that are therapeutic targets, using OncodriveMut, a bioinformatics method to identify the most likely driver mutations of a tumor.

CGI annotates the mutations that affect cancer genes, identifying the most likely drivers among the unknown significance variants. It uses four databases, with the former two exploring the associations between gene alterations and drug response, and with the latter two exploring the genes tumorigenesis:

1. The Cancer Biomarkers database is a database that is currently being integrated with knowledge databases of other institutions in a collaborative effort of the Global Alliance for Genomics and Health.
2. The Cancer Bioactivities database contains information about 20,243 chemical compound-protein product interactions that may support novel research applications.
3. The Catalog of Validated Oncogenic Mutations is a compiled inventory of mutations in cancer genes demonstrated to drive tumor growth or predispose to cancer. It was built combining the DoCM, ClinVar, and OncoKB databases plus published experimental assays and manually curated results. It also includes germline mutation from ClinVar and IARC that predispose to cancer.
4. The Catalog of Cancer Genes collects the genes driving tumorigenesis in several cancer types through a specific alteration (e.g., gene translocation). These genes are supported by validated data or computational predictions. This information is aggregated from the Cancer Gene Census [20, 21] and a manual curation effort. The computational predictions are made using tools included in the IntOGen resource [20–22].

CGI matches the alterations observed in newly sequenced tumors to the biomarkers or target genes in these two databases.

The CGI also reports co-occurring alterations that affect the response to a given treatment as appropriate, including the co-existence of resistance or sensitivity biomarkers and biomarkers of drug sensitivity that depend upon simultaneous genomic events.

CGI can identify between 5.2% and 3.5% of the tumors with genomic alterations that are biomarkers of drug response with high evidence levels. If we consider biomarkers with low levels of evidence, CGI can identify 62% of tumors with at least one of these biomarkers. So CGI can help the process of decision-making for the therapies of a patient.

Databases Consensus

Figures 10.1, 10.2, 10.3, and 10.4 show the consensus among the different databases. We observe that the databases have few genes in common. This heterogeneity is due to several factors such as the manual curation, the lack of name standardization, and the existence in PharmGKB of different disease associations other than cancer. For this reason, projects like VICC have been established, to represent and share harmonized interpretations.

COSMIC

The Catalogue Of Somatic Mutations In Cancer (COSMIC) [8] is a resource for exploring the effect of somatic mutations in human cancer. The current version (v92) comprises 9,215,470 gene variants curated over 27,724 papers. It covers non-coding mutations, gene fusions, copy number variants, and drug-resistance mutations. It draws together information about somatic mutations across human cancers, deriving it from expert manual curation of scientific literature. It catalogs all genes that are causally implicated in cancer through somatic and germline mutations. Since 2016, COSMIC includes drug-resistance genetics, annotating the novel somatic gene mutations that enable a tumor to evade therapeutic cancer drugs.

COSMIC started in 2004 as a survey of only four genes. It now usually has four releases per year. It is complemented by additional datasets and tools like The COSMIC Cell Lines Project that comprises data from whole-exome sequencing and molecular profiling of 1015 cell lines at the Wellcome Sanger Institute and aims to char-

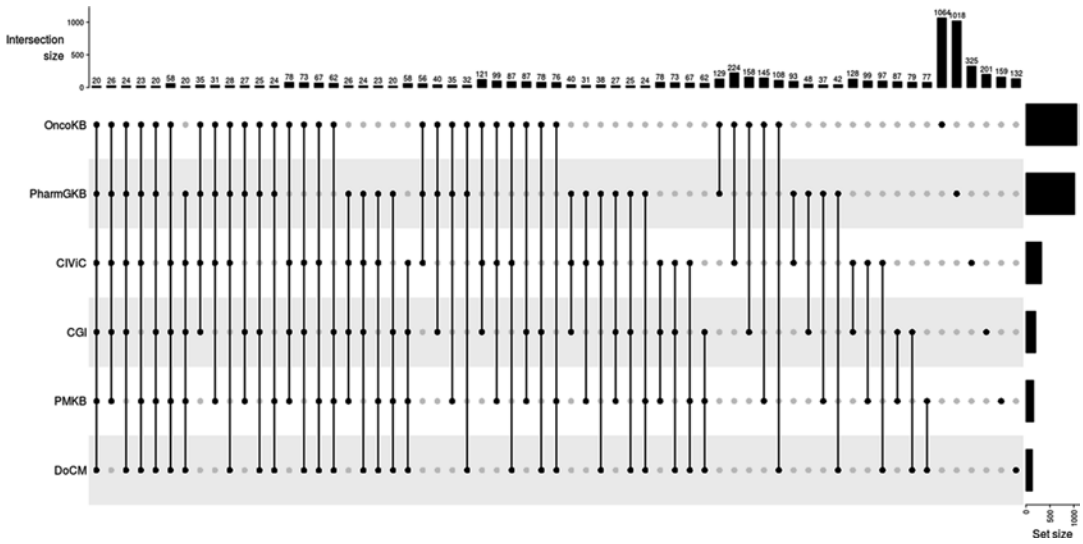

Fig. 10.1 Database Gene Intersection

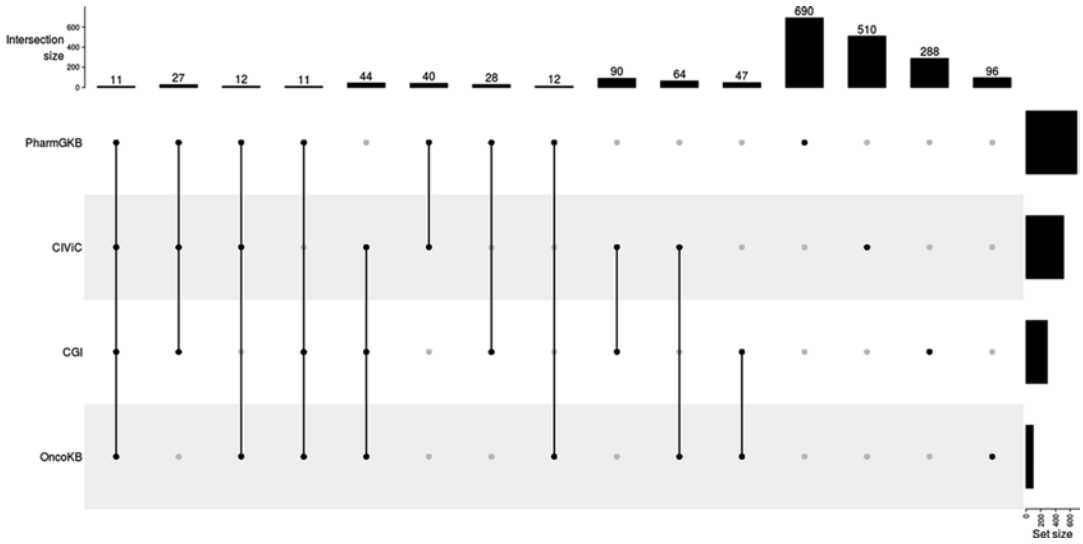

Fig. 10.2 Database Drug and Drug combination Intersection

acterize the genetics and genomics of cancer cell lines systematically.

COSMIC-3D is a tool that links the detailed sequence-level mutation data in COSMIC with the rich protein-structural data in the Protein Data Bank [23], facilitating structure, function, and druggability analysis.

The Cancer Gene Census (CGC) identifies every gene with a demonstrable role across all forms of human cancer and explains how dysfunction of these genes drives cancer. In CGC, a gene is classified as an oncogene, a tumor suppressor gene (TSG), or both, or a fusion gene after an evaluation process. Moreover, the genes are classified into tiers depending on the strength of the evidence supporting their cancer-promoting role. The genes are classified as follows:

- Tier 1, if they present a mutational pattern that strongly supports their involvement in cancer

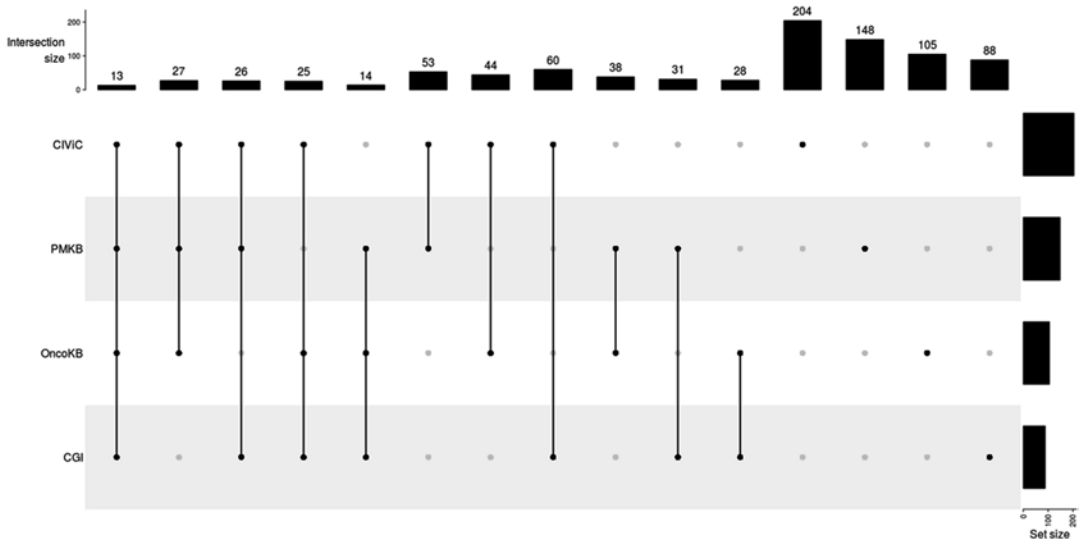

Fig. 10.3 Database Disease Intersection

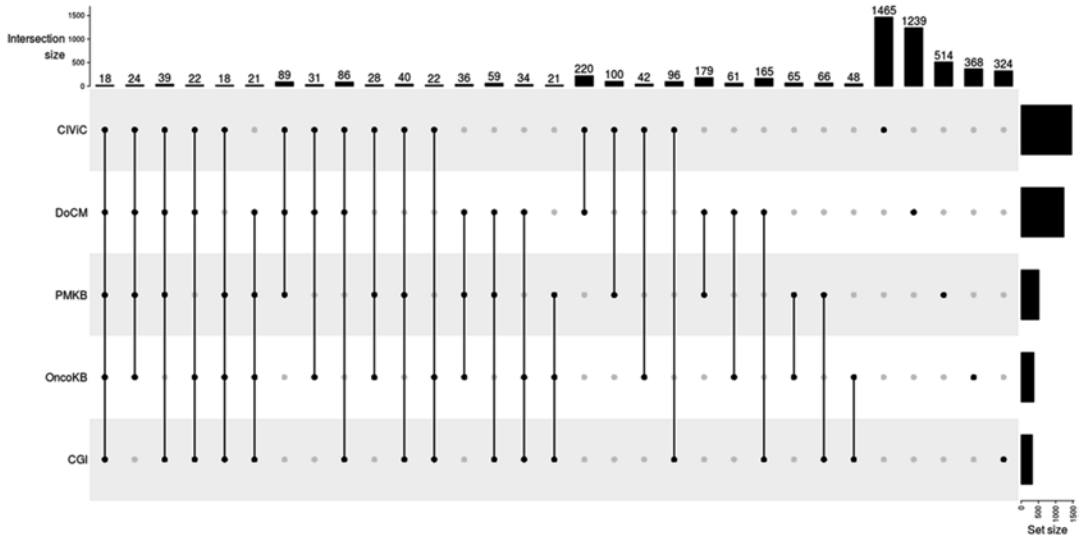

Fig. 10.4 Database Gene-Variant Intersection

etiology. It requires the existence of at least two publications from two independent groups that describe somatic mutations in the gene in at least one type of cancer and at least two independent publications that provide evidence of functional involvement of the gene in biological processes driving cancer.

- Tier 2, if they have extensive literature evidence for their tumor development participation but have less robust evidence supporting mutational patterns or functional consequence. The evidence must be assessed independently by at least two postdoctoral scientists.

A section of CGC is also focused on functional descriptions of cancer genes to characterize each gene's impact on the ten hallmarks of cancer.

COSMIC also includes a wide variety of valuable annotations related to patients' clinical details, diseases, and treatment.

ClinVar

ClinVar is an archive of human genetic variants and their relationships to human health and disease. It was created at the National Center for Biotechnology Information (NCBI) at the National Institutes of Health (NIH) to provide a centralized, public open-access database for data needed to help users interpret variants. It provides a freely available archive of reports of relationships among medically important variants and phenotypes. It gives interpretations of the variation in relation to human health and the evidence supporting each interpretation. ClinVar was first released in November 2012. Its database is part of the NCBI's Entrez system.

ClinVar aims to facilitate evaluating variant-phenotype relationships by archiving submitted interpretations of these relationships with supporting evidence, taking data from multiple groups such as laboratories to search for a consensus about the interpretation. ClinVar is mainly focused on variations that may be medically relevant. ClinVar depends on MedGen to represent phenotype, Gene to represent genes, and Human RefSeq to represent the location of sequence variation.

ClinVar's data model is based on five major categories of content: submitter data for attribution, the definition of the variation, characterization of the phenotype, evidence about the effect of variation on health, and interpretation of that evidence. Whenever possible, the content is highly structured rather than free text and is harmonized to controlled vocabularies or other data standards.

Variations submitted to ClinVar are compared to variations accessed by dbSNP or dbVar. If known, ClinVar adds the rs- (dbSNP) or variant call identifier (dbVar) to the RCV record. If novel, the information is submitted to the appropriate variation database to be accessed so that the identifiers can be added to ClinVar.

Annotators

Identified variants need to be annotated to identify which genes, transcripts, and genomic regions they belong to, in order to understand their impact and role in the tumor. It is then possible to establish whether a gene is an onco-driver or is involved in therapy response or resistance, and how deleterious the variant is for the patient.

Several tools, such as ANNOVAR [24], SnpEff [25], and Variant Effect Predictor (VEP) [26], can annotate sequencing variants (Tables 10.2 and 10.3).

Table 10.2 Annotator comparison

	Annovar	Oncotator	SnpEff/SnpSift	VEP	Funcotator
Input	VCF	VCF, MAF	VCF, BED	VCF	VCF
Output	VCF, TXT	VCF, TSV, MAF	VCF	VCF	VCF, MAF
Human supported genome versions	NCBI36, GRCh37, GRCh38	GRCh37	GRCh37, GRCh38	GRCh37, GRCh38	GRCh37, GRCh38
Web interface	YES, wAnnovar	Dismissed	YES, in Galaxy	YES, Ensembl Tools	Available through GATK
Language	Perl	Python	Java	Perl	Java
Research Availability	Registration required	Free	Free	Free	Free
Commercial Availability	License required	NO	Free	Free	Free

Table 10.3 Annotator databases

	Annovar	Oncotator	Funcotator	VEP	SnpEff/SnpSift
RefSeq	✓	✓	✓	✓	
UCSC Known Gene	✓	✓	✓		
Ensembl gene	✓	✓		✓	✓
GENCODE	✓	✓	✓	✓	✓
Epigenome Roadmap					✓
ESP		✓			
CCLE		✓			
TFBS Transcription factor binding site predictions					✓
NextProt					✓
UniProt		✓	✓		
HGNC			✓		
COSMIC	✓	✓	✓	✓	
Cancer Gene Census		✓	✓		
Familiar Cancer Database		✓			
HGMD-Public				✓	
Oreganno			✓		
Clinvar	✓	✓	✓	✓	✓
Intervar	✓				
Exac	✓			✓	✓
1000Genome	✓	✓		✓	✓
Kaviar	✓				
NCI-60	✓				
dbsnp	✓	✓		✓	✓
GnomAD	✓		✓		
NHLBI-ESP				✓	
SIFT	✓	✓		✓	✓
PolyPhen2	✓	✓		✓	✓
LRT	✓	✓			✓
MutationTaster	✓	✓		✓	✓
PhyloP	✓	✓			✓
MutationAssessor	✓	✓			
GERP++	✓	✓		✓	✓
MetaSVM	✓				
MetaLR	✓				
CADD	✓			✓	
DANN	✓				
fitCons	✓				

(continued)

Table 10.3 (continued)

	Annovar	Oncotator	Funcotator	VEP	SnpEff/SnpSift
GWAS					✓
Condel				✓	
FATHMM	✓	✓		✓	
SiPhy	✓	✓			✓
Interpro					✓
Haploinsufficiency					✓
PhastCons	✓				✓
GWAVA				✓	

ANNOVAR

ANNOVAR (Annotate Variation) [24–27] is a well-known open-source command-line software developed as a configurable, flexible, updated, and cross-platform annotator to overcome all common problems arising after an NGS analysis.

It takes a VCF file and annotates it using several built-in or user-generated databases. ANNOVAR modifies the VCF file to include columns representing the chromosome, start position, end position, the reference nucleotide(s), and the observed nucleotide(s). The user can also supply additional columns that will be printed out in the output files.

When the user downloads ANNOVAR, he needs first to download gene datasets and specify the genome version (hg18, hg19, or hg38).

ANNOVAR can perform three types of annotations: (i) gene-based annotation, (ii) region-based annotation, and (iii) filter-based annotation.

In the gene-based annotation, ANNOVAR splits the variant in intronic, exonic, intergenic, 5'-3'-UTR, splicing site, and upstream/downstream. For intergenic variants, the closest two genes and the distances between them are reported. Another file is created for the exonic variants, which are classified as frameshift insertion/deletion/block substitution, stop-gain, stop-loss, non-frameshift insertion/deletion/block substitution, non-synonymous SNV, synonymous SNV, and unknown. It also reports the amino acid change.

The databases for gene-based annotation are RefSeq gene annotation, UCSC Known Gene, and Ensembl Gene annotation.

The region-based annotation refers to specific elements like conserved genomic regions, predicted transcription factor binding sites, predicted microRNA target sites, and predicted stable RNA secondary structures. The filter-based annotation consists of the comparison between the database and the user VCF. It produces two output files: one with the variants common to both the database and the VCF, and one with the variant existing only in the VCF file. The user can also decide to filter these variants by frequency (MAF—Minor allele frequency) or by SIFT [28] (Sorting Intolerant From Tolerant) score. The former is the frequency of the second most common allele in the population. It helps to differentiate between common and rare variants. The latter is the score predicting the consequence of an amino acid substitution on the protein function.

The databases available for variant filtering are 1000 Genomes Project dataset [29], with allele frequency in six populations; Kaviar database [30] which contains a collection of variants; the Haplotype Reference Consortium (HRC) [31, 32]; Allele frequency in 69 human subjects sequenced by Complete Genomics [33]; Allele frequency in Genome Aggregation Database (gnomAD) [34]; Latest Exome Aggregation Consortium (ExAC) dataset [35]; Latest NHLBI-ESP [36] project with 6500 exomes; GME

(Greater Middle East Variome) allele frequency [37, 38]; Abraom, 2.3 million Brazilian genomic variants [39, 40]; dbnsfp41 [41, 42]: this dataset includes deleteriousness prediction scores SIFT, PolyPhen2 HDIV, PolyPhen2 HVAR [43], LRT (Likelihood Ratio Test) [44], MutationTaster [45], MutationAssessor [46], FATHMM (Functional Analysis through Hidden Markov Models) [47], MetaSVM [48], MetaLR [48], VEST [49], CADD (Combined Annotation–Dependent Depletion) [50], BayesDel_addAF and BayesDel_noAF [51], CADD_hg19 [50], ClinPred [52], DEOGEN2 [53], Eigen and Eigen PC [54], FATHMM-XF [55], GenoCanyon [56], LINSIGHT [57], LIST-S2 [58], M-CAP [59], MPC [60], MutPred [61], MVP [62], PrimateAI [63, 64], REVEL [65], SIFT4G [66], DANN (deleterious annotation of genetic variants using neural networks) [67], fitCons (fitness consequence) [68], conservation scores GERP++ [69], PhyloP [70], SiPhy [71], phyloP17-way_primate [72], phastCons17way_primate [73] and bStatistic [74], and one score for loss of function prediction ALoFT [75]; dbscSNV [76] version 1.1 for splice site prediction by AdaBoost and Random Forest; ClinVar database; Intervar [77] helps in the variants clinical significance interpretation; COSMIC v70; International Cancer Genome Consortium [78, 79] version 21, only for hg19; and NCI-60 human tumor cell line panel exome sequencing allele frequency data [80] and snp142 [81] (Single Nucleotide Polymorphisms), which is a public repository of variants with information about the type of mutation.

Oncotator and Funcotator

Oncotator [82] is a command-line tool for cancer variant annotation, written in Python and recommended for advanced users. It also has a web interface, providing both an interactive user interface and a programmatic web service. Although it is still downloadable, it has been officially dismissed by the Broad Institute and superseded by Funcotator [83].

Oncotator can be included in automated pipelines since the annotation options, selection of data sources, and file formats are flexible. It owns a bundle of cancer-relevant information that users can use in a single step.

Oncotator needs a file with the genomic position, reference allele, and variant allele in VCF or TSV format.

To map variants to specific genes and classify them, Oncotator uses GENCODE [84]. The nomenclature follows the Human Genome Variation Society (http://www.hgvs.org/mutnomen) standards. Moreover, to identify common Single Nucleotide Polymorphism (SNP) variants (which are less likely to contribute to tumorigenesis), Oncotator utilizes data from dbSNP, 1000 Genomes Project [29], and National Heart, Lung, and Blood Institute's Exome Sequence Project.

Oncotator has a unique feature. It can annotate variants with their local GC content and their surrounding nucleotides to discover if these mutations are biological processes or artifactual mutation biases, such as oxidation of guanine bases during sequencing library construction (OxoG). It can also annotate genomic variants with protein-specific annotations derived from UniProt human protein sequence records to predict cancer mutation's functional impact. Protein annotations added include "region" (e.g., protein kinase domain), "site" (e.g., ATP binding site), "natural variation" (e.g., Y → F in Pfeiffer syndrome), and "experimental" (e.g., Y → F: 50% decrease in interaction with PIK3C2B) data.

Gene Ontology annotations are derived from UniProt records. These annotations are categorized as biological process, cellular component, and molecular function. Furthermore, Drugbank annotation for small molecules that target the protein of interest is integrated. The database dbNSFP is used to make functional predictions.

Oncotator also annotates variants with several cancer-specific databases such as COSMIC, the Cancer Cell Line Encyclopedia, Cancer Gene Census, ClinVar, The Familiar Cancer Database, and a curated set of DNA repair genes. COSMIC is used to identify variants reported in published studies and reports their observed frequency across all cancers and within each tissue type, overlapping breakpoint, and fusion genes. The

Cancer Cell Line Encyclopedia is employed to observe if a variant has already been observed in a cell line.

Depending on the type of input, the output is in either MAF or VCF format.

Oncotator uses a three-stage workflow:

1. Convert the input data into an internal model of mutations.
2. Annotate the mutation objects with a collection of pre-processed data sources (which can be locus-, variant- or gene-specific).
3. Render the mutations to the specified output format (VCF or MAF).

Oncotator has been superseded by Funcotator (FUNCtional annOTATOR). It analyzes variants for their function. The user can use their data sources to create a custom annotation. All the databases integrated in the tool can be downloaded in one step with the FuncotatorDataSourceDownloader.

Different from Oncotator, Funcotator supports both GRCh37 and GRCh38 genomes.

Funcotator requires as inputs a reference genome sequence and the VCF sample that needs to be annotated. It performs some processing on the input data to create the GENCODE annotations. The output can be either a MAF or a VCF. The output file contains all the variants from the input with added annotations.

SnpEff and SnpSift

SnpEff [25, 85] (SNP effect) is a multi-platform open-source Variant Effect predictor program written in java able to analyze and annotate thousands of variants per second. It supports different genome versions. VCF and BED are its primary input formats. Furthermore, the user can add custom genomes and annotations from multiple species. SnpEff can also annotate non-coding genes.

The output is a VCF or a TXT file that includes the following:

1. Variant information consisting of the genomic position, the reference and variant sequences, change type, heterozygosity, quality, and coverage.
2. Genetic information consisting of the gene Id, gene name, gene biotype, transcript ID, exon ID, and exon rank.
3. Effect information consisting of the effect type, amino acid changes, codon changes, codon number in CDS, codon degeneracy, etc.

A variant can have more than one line in the output if more than one transcript exists.

Human genome versions GRCh37 and GRCh38 are supported. The variant annotation and filtering are supported through SnpSift, which can also calculate a conservation score annotation through phastCons. Like Funcotator, SnpEff adds annotation information to the INFO field of a VCF file.

The annotation information is divided into the following:

- Allele or ALT for multiple ALT fields.
- Annotation or effect.
- Putative impact or deleteriousness.
- Gene name, the HGNC name. If the variant is intergenic, the closest gene name is used.
- Gene ID.
- Feature type.
- Feature ID.
- Transcript biotype.
- HGVS.c and HGVS.p Variant using HGVS notation, respectively, in DNA level and protein level.
- cDNA position and length.
- CDS position and length.
- Protein position and length.
- Distance to feature.
- Errors, warnings, or information messages

The impact categories (high, moderate, or low) have to be carefully handled since they have been created only to simplify the filtering process. Indeed, there is no way to predict whether a high or low impact variant produces the phenotype of interest.

SnpEff possesses the command-line option—cancer that helps users to compare somatic and germline samples. Furthermore, when multiple

ALTs, somatic, and germline samples are in a VCF file of cancer data, they can be separated using a TXT file with the cancerSample command-line option or using the PEDIGREE meta information of the VCF header.

Furthermore, the tool performs some statistical analysis available as HTML or CSV. Another output is a TXT file that counts the number of variants affecting each transcript and gene.

There is also a commercial version of SnpEff & SnpSift, and it is called ClinEff [86]. It is considered more stable and suitable for clinical and production operations, while SnpEff/SnpSift are designed for research use.

VEP

The Ensembl Variant Effect Predictor (VEP) [26] is a software suite that performs annotation and analysis of most genomic variation types in the genome's coding and non-coding regions. It can annotate and prioritize variants with well-defined changes, such as SNVs, insertions, deletions, and larger structural variants. VEP returns detailed annotation for effects on transcripts, proteins, and regulatory regions for all input variants. For known or overlapping variants, allele frequencies and disease or phenotype information are included. VEP can be used for any species for which an assembled genome and an annotated gene set exist. For human, it supports both GRCh38 and GRCh37 assemblies. VEP results include a wide variety of gene and transcript-related information.

The leading database annotation for Ensembl is the GENCODE gene set. Ensembl transcripts are matched exactly to the reference genome assembly eliminating the potential errors in the annotation. If VEP is configured to use RefSeq, any mismatch is reported to remove possible interpretation confusion.

A variant may have more than one alternative non-reference allele and may overlap with more than one transcript or regulatory region. Therefore, to present the most comprehensive annotation, VEP reports one line (or unit) of annotation per variant alternative allele and genomic feature. When there is no robust annotation of dominant transcript per tissue type available, VEP includes various data to filter transcript isoforms. Cross-references to known proteins in UniProt and the option to filter for variants in protein-coding transcripts are also included.

VEP also indicates the effect of the amino acid change using protein biophysical properties. It contains several pathogenicity predictor scores and conservation scores. The former are the database SIFT for the Top-10 species in Ensembl, PolyPhen2 [43] for human proteins, Condel [87], FATHMM [47], and MutationTaster [45] for human data. GERP [88] and GWAVA [89], CADD [50], and FATHMM-MKL [90] are available as plugins.

The Ensembl Regulatory Build is used for regulatory region annotation. Additional databases are dbSNP, COSMIC, the Human Gene Mutation Database, and the Genomic Variants archive for structural variants and copy number variants. The allele frequencies can be filtered using 1000 Genomes, NHLBI exome sequencing, and ExAC project.

Every variant annotated with VEP has a PubMed identifier, the disease or trait using OMIM, Orphanet, and the Genome-Wide Association Study (GWAS Catalog) [91]. ClinVar reports the clinical significance.

VEP uses an input VCF file. The output consists of an HTML or TXT summary file and a primary results file in VCF, GVF, or JSON format. VEP is available through a web interface, a Perl script, or via the Ensembl API.

Variant Prioritization

A typical NGS assay can detect thousands of genetic variants. However, many variants do not have a specific classification. Such variants are therefore called "Variant of Uncertain (or Unknown) Significance" (VUS). Many tools can prioritize these variants as Pathogenic or Benign. This process is known as "Variant prioritization"

or "Variant filtration." Examples include VINYL [92], KGGSeq [93], and MutationDistiller [94]. These tools filter and evaluate variants using existing databases, creating a pathogenicity score. Some tools like GeneDistiller [95] and Endeavour [96] prioritize the full gene instead of the single variants.

Prioritization helps the clinicians to separate authentic disease-causing variants from others. Unfortunately, prioritization tools are not commonly used. Furthermore, many institutions apply different criteria and filters, limiting the reproducibility of the analysis.

GeneDistiller

GeneDistiller can be used as a prioritization tool or together with other prioritization tools to display rich information on human candidate genes. It offers users different approaches such as Projection, Selection, Sorting, and Prioritization. In the first approach, the user chooses the genes of interest. In the second, the user applies filters to the genes reducing them to a smaller group. In the third approach, the genes are sorted according to certain parameters. The fourth approach, the prioritization one, offers a function that ranks genes according to the researcher's specifications. These methods can be combined. GeneDistiller can be used through a web user interface on http://www.genedistiller.org/.

Endeavour

Endeavour forecasts the most promising candidate genes implicated in a disease. Differently from GeneDistiller, it executes gene prioritization for six species (*Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, and Danio rerio). The prioritization requires four stages: choosing the species, preparing a list of genes already associated with the disease of interest, picking data sources to use for the prioritization, and finally defining the candidate genes. The output is a list of the genes ordered by prioritization with a p-value derived from the combination of rankings. Endeavour can be used through a web user interface on https://endeavour.esat.kuleuven.be/.

MutationDistiller

MutationDistiller prioritizes monogenic disease variants with the help of GeneDistiller. It filters the polymorphisms using ExAC and 1000Genome and uses ClinVar to identify known disease-causing mutations. After the analysis, MutationDistiller presents a prioritized list of the most likely candidate variants with information about them and their genes downloaded as a summary table. The table shows the variant in class; therefore, the user can focus on certain alterations.

VINYL

VINYL derives a pathogenicity score aggregating various public databases. The idea behind the tool's construction is that affected individuals have an excess of the deleterious variant with respect to a matched population of unaffected individuals. The tool is highly flexible, permitting the incorporation of different types of annotation and resources.

KGGSeq

KGGSeq performs an analysis procedure for the discovery of human Mendelian disease genes combining filtration and prioritization functions. It filters and prioritizes the variants at three levels, genetic, variant-gene, and knowledge, according to the resource used. Like MutationDistiller, it filters out common variants using public databases, such as 1000Genome, and the allele frequency threshold.

Pathogenicity Predictors

The pathogenicity of a variant within protein-coding transcripts can be assessed and assigned a score by several tools.

Tools for Functional Prediction Scores

SIFT uses sequence homology to predict if an amino acid change will alter a protein's function. According to SIFT, the amino acid change in a well-conserved position is presumed deleterious. To predict whether a substitution will affect the protein function, SIFT considers the type of the amino acid change and its position, calculating the probability of tolerance. The resulting score is the normalized probability that the amino acid change is tolerated. SIFT can be obtained from https://sift.bii.a-star.edu.sg/.

PolyPhen-2 (Polymorphism Phenotyping v2) is a tool that forecasts the possible impact of an amino acid substitution on a protein. To make the prediction, it uses comparative consideration, such as comparing the property of normal and mutant alleles. It estimates the mutations as benign, possibly damaging, or probably damaging, calculating Naïve Bayes posterior probability that the mutation is damaging. PolyPhen-2 can be obtained from http://genetics.bwh.harvard.edu/pph2/.

LRT can identify deleterious mutations that affect highly conserved amino acids giving protein alterations. It was built to model phylogenetic relationships using a probabilistic framework and closely related species.

MutationTaster2 is a software that estimates the pathogenic potential of DNA sequence alterations. It predicts the functional consequences of amino acid substitutions, intronic, intergenic, synonymous, and indel mutations. It uses a Bayes classifier and three classification models, one for alterations of single amino acid, one for alterations that involve more than one amino acid, and one for non-coding and synonymous alterations. MutationTaster2 can be obtained from http://www.mutationtaster.org/.

MutationAssessor predicts the functional impact of amino-acid substitutions in proteins. The prediction is made using evolutionary conservation information of the protein family through multiple alignment. MutationAssessor can be obtained from http://mutationassessor.org/r3/.

FATHMM is a software for predicting the functional effects of protein missense variants. It can be applied to human and non-human genomes. It possesses several sub-algorithms, one for cancer-specific driver mutations [97] and one, called FATHMM-MKL, for coding and non-coding sequence variants. FATHMM can be obtained from http://fathmm.biocompute.org.uk/.

MetaSVM and MetaLR provide an ensemble score integrating nine scores (SIFT, PolyPhen-2, GERP++, MutationTaster, MutationAssessor, FATHMM, LRT, SiPhy, and PhyloP) with the MMAF (maximum minor allele frequency) of populations. MetaLR uses a logistic regression (LR) algorithm, while MetaSVM uses Support Vector Machine (SVM)

VEST (Variant Effect Scoring Tool) is a supervised machine learning classifier that prioritizes missense mutations that alter protein function. VEST can be obtained from https://karchinlab.org/apps/appVest.html.

CADD integrates multiple annotations for scoring the deleteriousness of single nucleotide variants and indel variants in the human genome. It was built from 60 genomic features. It uses a machine learning model trained with de novo variants and variants fixed in human populations. CADD can be obtained from https://cadd.gs.washington.edu/.

DANN [67] is built using a deep neural network trained using the same feature set and training data as the first CADD version. DANN data can be obtained from https://cbcl.ics.uci.edu/public_data/DANN/.

PROVEAN [98] (Protein Variation Effect Analyzer) is a tool that predicts the functional impact of an amino acid change or an indel on a protein. It was tested on UniProtKB/Swiss-Prot database and experimental datasets. PROVEAN

can be obtained from http://provean.jcvi.org/index.php.

FitCons is a method that estimates the probability that a point mutation in a genome will influence fitness. It computes a score that indicates the evolution-based potential genomic function integrating evolutionary and functional data. fitCons can be obtained from http://compgen.cshl.edu/fitCons/.

Tools for Conservation Score

GERP++ [69] is a program that identifies conservation scores following a comparative genomic approach. It recognizes sites under evolutionary constraint through multiple alignment of the human genome with 33 other mammalian species. GERP++ can be obtained from http://mendel.stanford.edu/SidowLab/downloads/gerp/.

SiPhy [71] identifies conservation scores as a decrease in the rate of mutation and searching for biased substitution patterns. It works with multiple alignment data. SiPhy can be obtained from http://portals.broadinstitute.org/genome_bio/siphy/index.html.

PhastCons [73] is a tool for the identification and scoring of conserved elements in multiple alignment sequences. It is based on a two-state phylogenetic hidden Markov model (phylo-HMM). PhastCons can be obtained from http://compgen.cshl.edu/phast/.

PhyloP computes p-values for conservation or acceleration in functional elements. It was built by implementing four statistical phylogenetic tests, a likelihood ratio test, a score test, a test based on exact distributions of several substitutions, and the genomic evolutionary rate profiling (GERP) test. PhyloP can be obtained from http://compgen.cshl.edu/phast/.

dbNSFP

In the last few years, all these tools have been gathered in a database called dbNSFP [99, 100], where the deleteriousness score has been standardized to facilitate the evaluation of the importance of a mutation in sequencing studies. dbNSFP was built because each algorithm uses different information, and it is based on various training data, outcoming in different results. So, a more reliable prediction can come from an analysis that uses multiple algorithms. dbNSFP, in its latest version (4.1), comprises all the tools of cited above and other 22 tools.

This latest version is based on human reference sequence version hg38 and GENCODE version 29 and includes 81,782,923 nsSNVs and 2,230,170 ssSNVs. It also includes the ExAC database, allele frequencies from the UK10K cohorts, the NHLBI Exome Sequencing Project data, and allele frequencies from the 1000 Genomes Project phase 1 data, ClinVar, and dbSNP.

If one of the scores is missing in a database, it is imputed using BPCAfill [101] which is generally applied for the imputation of missing expression data from microarray analyses.

Conclusion

In this chapter we have surveyed relevant resources for interpreting variations in cancer. We focused on databases, annotators, and variants prioritization tools to provide an up-to-date reference point to help researchers in their studies. Furthermore, we have highlighted the content and limitations of each tool and database, and compared their content.

References

1. Home. https://www.amp.org/.
2. American Society of Clinical Oncology. https://www.asco.org/front.
3. Homepage. https://www.cap.org/.
4. GA4GH. https://www.ga4gh.org/.
5. Standardizing cancer variant knowledge to enable precision oncology. https://cancervariants.org/index.html.
6. Griffith M, et al. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. Nat Genet. 2017;49:170–4.
7. Landrum MJ, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014;42:D980–5.

8. Tate JG, et al. COSMIC: the catalogue of somatic mutations in cancer. Nucleic Acids Res. 2019;47:D941–7.
9. Karczewski KJ, et al. The ExAC browser: displaying reference data information from over 60 000 exomes. Nucleic Acids Res. 2017;45:D840–5.
10. Whirl-Carrillo M, et al. Pharmacogenomics knowledge for personalized medicine. Clin Pharmacol Ther. 2012;92:414–7.
11. Thorn CF, Klein TE, Altman RB. PharmGKB: the pharmacogenomics Knowledge Base. Methods Mol Biol. 2013;1015:311–20.
12. Wishart DS, et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 2006;34:D668–72.
13. Braschi B, et al. Genenames.org: the HGNC and VGNC resources in 2019. Nucleic Acids Res. 2019;47:D786–92.
14. Gao J, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal. 2013;6:l1.
15. Cerami E, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. 2012;2:401–4.
16. NCCN - Evidence-Based Cancer Guidelines, Oncology Drug Compendium, Oncology Continuing Medical Education. https://www.nccn.org/.
17. Rehm HL, et al. ClinGen--the clinical genome resource. N Engl J Med. 2015;372:2235–42.
18. ESMO. ESMO. https://www.esmo.org/.
19. American Association for Cancer Research (AACR). https://www.aacr.org/.
20. Tamborero D, et al. Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. Genome Med. 2018;10:25.
21. Sondka Z, et al. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat Rev Cancer. 2018;18:696–705.
22. Gonzalez-Perez A, et al. IntOGen-mutations identifies cancer drivers across tumor types. Nat Methods. 2013;10:1081–2.
23. Berman HM, et al. The archiving and dissemination of biological structure data. Curr Opin Struct Biol. 2016;40:17–22.
24. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38:e164.
25. Cingolani P, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly. 2012;6:80–92.
26. McLaren W, et al. The Ensembl variant effect predictor. Genome Biol. 2016;17:122.
27. Yang H, Wang K. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. Nat Protoc. 2015;10:1556–66.
28. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31:3812–4.
29. 1000 Genomes Project Consortium et al. A global reference for human genetic variation. Nature. 2015;526:68–74.
30. Glusman G, Caballero J, Mauldin DE, Hood L, Roach JC. Kaviar: an accessible system for testing SNV novelty. Bioinformatics. 2011;27:3216–7.
31. McCarthy S, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet. 2016;48:1279–83.
32. The Haplotype Reference Consortium. http://www.haplotype-reference-consortium.org/.
33. brandonvd. 69 Genomes Data - Complete Genomics. https://www.completegenomics.com/public-data/69-genomes/.
34. Karczewski KJ, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581:434–43.
35. Lek M, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536:285–91.
36. NHLBI Grand Opportunity Exome Sequencing Project (ESP). https://esp.gs.washington.edu/drupal/.
37. GME Variome. http://igm.ucsd.edu/gme/.
38. Scott EM, et al. Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery. Nat Genet. 2016;48:1071–6.
39. ABraOM: Brazilian genomic variants. http://abraom.ib.usp.br/.
40. Naslavsky MS, et al. Whole-genome sequencing of 1,171 elderly admixed individuals from the largest Latin American metropolis (São Paulo, Brazil). Cold Spring Harbor Laboratory 2020.09.15.298026, 2020. https://doi.org/10.1101/2020.09.15.298026.
41. Liu X, Wu C, Li C, Boerwinkle E. dbNSFP v3.0: a one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. Hum Mutat. 2016;37:235–41.
42. Liu X, Li C, Mou C, Dong Y, Tu Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. Genome Med. 2020;12:103.
43. Adzhubei IA, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7:248–9.
44. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res. 2009;19:1553–61.
45. Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: mutation prediction for the deep-sequencing age. Nat Methods. 2014;11:361–2.
46. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. 2011;39:e118.
47. Shihab HA, et al. Predicting the functional, molecular, and phenotypic consequences of amino acid sub-

stitutions using hidden Markov models. Hum Mutat. 2013;34:57–65.
48. Dong C, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum Mol Genet. 2015;24:2125–37.
49. Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics. 2013;14 Suppl 3:S3.
50. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019;47:D886–94.
51. Feng B-J. PERCH: a unified framework for disease gene prioritization. Hum Mutat. 2017;38:243–51.
52. Alirezaie N, Kernohan KD, Hartley T, Majewski J, Hocking TD. ClinPred: prediction tool to identify disease-relevant nonsynonymous single-nucleotide variants. Am J Hum Genet. 2018;103:474–83.
53. Raimondi D, et al. DEOGEN2: prediction and interactive visualization of single amino acid variant deleteriousness in human proteins. Nucleic Acids Res. 2017;45:W201–6.
54. Ionita-Laza I, McCallum K, Xu B, Buxbaum JD. A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat Genet. 2016;48:214–20.
55. Rogers MF, et al. FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. Bioinformatics. 2018;34:511–3.
56. Lu Q, et al. A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. Sci Rep. 2015;5:10576.
57. Huang Y-F, Gulko B, Siepel A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. Nat Genet. 2017;49:618–24.
58. Malhis N, Jacobson M, Jones SJM, Gsponer J. LIST-S2: taxonomy based sorting of deleterious missense mutations across species. Nucleic Acids Res. 2020;48:W154–61.
59. Jagadeesh KA, et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. Nat Genet. 2016;48:1581–6.
60. Qi H, et al. MVP: predicting pathogenicity of missense variants by deep learning, vol. 259390. Cold Spring Harbor Laboratory; 2018. https://doi.org/10.1101/259390.
61. Li B, et al. Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics. 2009;25:2744–50.
62. Qi H, et al. MVP predicts the pathogenicity of missense variants by deep learning. Nat Commun. 2021;12:510.
63. Sundaram L, et al. Predicting the clinical impact of human mutation with deep neural networks. Nat Genet. 2018;50:1161–70.
64. Sundaram L, et al. Author correction: predicting the clinical impact of human mutation with deep neural networks. Nat Genet. 2019;51:364.
65. Ioannidis NM, et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. Am J Hum Genet. 2016;99:877–85.
66. Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC. SIFT missense predictions for genomes. Nat Protoc. 2016;11:1–9.
67. Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics. 2015;31:761–3.
68. Gulko B, Hubisz MJ, Gronau I, Siepel A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. Nat Genet. 2015;47:276–83.
69. Davydov EV, et al. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol. 2010;6:e1001025.
70. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010;20:110–21.
71. Garber M, et al. Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics. 2009;25:i54–62.
72. Siepel A, Pollard KS, Haussler D. New methods for detecting lineage-specific selection. In: M. Research in computational molecular biology: 10th annual international conference, RECOMB 2006, Venice, Italy, April 2–5, 2006, proceedings. Springer; 2006.
73. Siepel A, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005;15:1034–50.
74. McVicker G, Gordon D, Davis C, Green P. Widespread genomic signatures of natural selection in hominid evolution. PLoS Genet. 2009;5:e1000471.
75. Balasubramanian S, et al. Using ALoFT to determine the impact of putative loss-of-function variants in protein-coding genes. Nat Commun. 2017;8:382.
76. Jian X, Boerwinkle E, Liu X. In silico prediction of splice-altering single nucleotide variants in the human genome. Nucleic Acids Res. 2014;42:13534–44.
77. Li Q, Wang K. InterVar: clinical interpretation of genetic variants by the 2015 ACMG-AMP guidelines. Am J Hum Genet. 2017;100:267–80.
78. Data Access Compliance Office (DACO). https://daco.icgc.org/.
79. International Cancer Genome Consortium et al. International network of cancer genome projects. Nature. 2010;464:993–8.
80. Shoemaker RH. The NCI60 human tumour cell line anticancer drug screen. Nat Rev Cancer. 2006;6:813–23.
81. Sherry ST, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29:308–11.
82. Ramos AH, et al. Oncotator: cancer variant annotation tool. Hum Mutat. 2015;36:E2423–9.

83. Website. https://gatk.broadinstitute.org/hc/en-us/articles/360035889931-Funcotator-Information-and-Tutorial.
84. Harrow J, et al. GENCODE: the reference human genome annotation for the ENCODE project. Genome Res. 2012;22:1760–74.
85. Cingolani P, et al. Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. Front Genet. 2012;3:35.
86. DnaMiner - ClinEff. http://www.dnaminer.com/clineff.html.
87. Gonzalez-Perez A, Deu-Pons J, Lopez-Bigas N. Improving the prediction of the functional impact of cancer mutations by baseline tolerance transformation. Genome Med. 2012;4:89.
88. Cooper GM, et al. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 2005;15:901–13.
89. Ritchie GRS, Dunham I, Zeggini E, Flicek P. Functional annotation of noncoding sequence variants. Nat Methods. 2014;11:294–6.
90. Shihab HA, et al. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics. 2015;31:1536–43.
91. Buniello A, et al. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019;47:D1005–12.
92. Chiara M, et al. VINYL: variant prIoritizatioN by survivaL analysis. Bioinformatics. 2020; https://doi.org/10.1093/bioinformatics/btaa1067.
93. Li M-X, Gui H-S, Kwan JSH, Bao S-Y, Sham PC. A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. Nucleic Acids Res. 2012;40:e53.
94. Hombach D, et al. MutationDistiller: user-driven identification of pathogenic DNA variants. Nucleic Acids Res. 2019;47:W114–20.
95. Seelow D, Schwarz JM, Schuelke M. GeneDistiller--distilling candidate genes from linkage intervals. PLoS One. 2008;3:e3874.
96. Tranchevent L-C, et al. Candidate gene prioritization with Endeavour. Nucleic Acids Res. 2016;44:W117–21.
97. Shihab HA, Gough J, Cooper DN, Day INM, Gaunt TR. Predicting the functional consequences of cancer-associated amino acid substitutions. Bioinformatics. 2013;29:1504–10.
98. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. PLoS One. 2012;7:e46688.
99. Liu X, Jian X, Boerwinkle E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Hum Mutat. 2011;32:894–9.
100. Liu X, Jian X, Boerwinkle E. dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. Hum Mutat. 2013;34:E2393–402.
101. Oba S, et al. A Bayesian missing value estimation method for gene expression profile data. Bioinformatics. 2003;19:2088–96.

Network Approaches for Precision Oncology

11

Shraddha Pai

Abstract

The growth of multi-omic tumour profile datasets along with knowledge of genome regulatory networks has created an unprecedented opportunity to advance precision oncology. Achieving this goal requires computational methods that can make sense of and combine heterogeneous data sources. Interpretability and integration of prior knowledge is of particular relevance for genomic models to minimize ungeneralizable models, promote rational treatment design, and make use of sparse genetic mutation data. While networks have long been used to capture genomic interactions at the levels of genes, proteins, and pathways, the use of networks in precision oncology is relatively new. In this chapter, I provide an introduction to network-based approaches used to integrate multi-modal data sources for patient stratification and patient classification. There is a particular emphasis on methods using patient similarity networks (PSNs) as part of the design. I separately discuss strategies for inferring driver mutations from individual patient mutation data. Finally, I discuss challenges and opportunities the field will need to overcome to achieve its full potential, with an outlook towards a clinic of the future.

Introduction

Precision oncology is the goal of using a patient's clinical, genomic, and physiological profile to predict disease outcomes such as prognosis or treatment resistance and decide the course of clinical care. This goal can be achieved by dividing tumour profiles into groups reflecting different types of molecular dysregulation in cancer, which in turn contribute to a unique signature of pathophysiology, outcome, and treatment response. Tumour profiling, particularly in the area of genomics, has dramatically grown in the past decade thanks to international collaborations to pool patient samples, ever-cheaper genomic profiling assays, and the ubiquity of cloud compute. The most recent integrative analysis of The Cancer Genome Atlas and International Cancer Genomics Consortium included nearly 10,000 tumours spanning 33 most common cancers [1], providing a sizable resource for precision oncology. However, an individual genomic assay can contain anywhere from thousands to millions of measures, increasing the challenge of discovering predictive signals in the noise. Computational approaches of unsupervised and supervised learning – or clustering and classification – help solve these problems. These approaches have

S. Pai (✉)
Ontario Institute for Cancer Research, University of Toronto, Toronto, ON, Canada
e-mail: shraddha.pai@utoronto.ca

© Springer Nature Switzerland AG 2022

A. Laganà (ed.), *Computational Methods for Precision Oncology*, Advances in Experimental Medicine and Biology 1361, https://doi.org/10.1007/978-3-030-91836-1_11

been successfully used to identify clinically relevant molecular subtypes in breast cancer, medulloblastoma, ependymoma, and others [2–5], and multiple commercial diagnostic tests exist for breast cancer, which use gene expression (e.g. Oncotype DX, ProSigna, Mammaprint) or immunohistochemistry (MammoStrat, IHC4) [2, 6–11]. However, models based on genomic data require interpretability for several reasons (see discussion in the next section), which involves modelling prior knowledge of molecular interactions within a cell. Networks provide a conceptually intuitive and elegant way to achieve this goal.

Normal genome and cell function is mediated by molecular interactions or networks, resulting in metabolic and physiological outcomes of cell growth and division, maintenance of cellular identity, and energy production. Conversely, these interactions and outcomes are disrupted in cancer, affecting a core set of cellular signalling pathways as well as interactions unique to specific types of cancer [12]. As a data structure, therefore, graphs or networks are well suited to present the correlations of cellular measures that in turn reflect molecular interactions, but which also capture similarities at the level of individual patient tumour profiles. As we shall see, networks also provide the means to encode prior knowledge of gene regulatory models, which can be used to improve inferences and interpretability from patient data.

This chapter will cover state-of-the-art network-based computational algorithms for patient stratification (or tumour subtype discovery) and patient classification. It will begin with a brief background on the value of networks in precision oncology and introduce the reader to the concept of patient similarity networks. Box 1 defines foundational concepts in the field. The chapter will then cover the tasks of patient stratification and classification in turn, discussing major advances in terms of concepts, algorithms, and software in this problem space. It will also cover network-based approaches for inferring driver mutations from individual (N-of-1) patient tumour mutation profiles. Each section first describes the algorithm and associated concepts, and then covers applications to date in precision oncology. Table 11.1 provides a list of all methods discussed in this chapter, accompanied by links to current software implementations. Finally, this chapter discusses existing challenges and opportunities in this relatively new field, and suggests a vision for network-enabled precision oncology in the span of the next decade.

Background

Networks in Precision Oncology

Networks explicitly represent entities and their relationships, the latter of which may be quantitative or qualitative; these are correspondingly depicted as nodes connected by weighted or unweighted edges (see example in Fig. 11.1). For oncology applications, networks can be used to reflect relations at various levels of system organization, including gene-gene or protein-protein interactions [13, 14], sets of pathways with overlapping member genes [15], or similarities in patient profiles [5, 16]. Box 1 outlines key concepts related to the application of networks and network-based methods for precision oncology. While readers may be familiar with gene and protein association networks owing to their long use in understanding signalling pathways and predicting gene function (e.g [17, 18].), the paradigm of patient similarity networks is relatively new. We therefore discuss it in some detail here.

Patient Similarity Networks

The use of patient similarity networks has only recently been used in biomedical applications for stratification and classification [5, 16, 19]. As patient samples in cancer are often tumour samples, one may also think of these as tumour similarity networks. In essence, a patient similarity network (PSN) is a network where the nodes are patients, and the edges are measures of pairwise similarity for the data from which the PSN was derived; an example is shown in Fig. 11.1. For example, in a PSN derived from

Table 11.1 Software for networks in precision oncology

Method	Use case	Approach	URL	References
NBS	Patient stratification / clustering	Smooth patient mutations over gene interaction network before clustering	http://chianti.ucsd.edu/~mhofree/NBS/	[21, 52]
SNF	Patient stratification / clustering	Integrate multi-omic patient similarity networks created from each data layer; edges weighted by cross-layer concordance prior to classification	http://compbio.cs.toronto.edu/SNF/SNF/Software.html	[5]
ndmaSNF	Patient stratification / clustering	SNF adapted for somatic mutations	–	[27]
NetNorM	Patient stratification/ clustering	Normalize patient mutations over gene interaction network before clustering	https://github.com/marineLM/NetNorM	[31]
NCIS	Patient stratification/ clustering	Co-clustering tumour expression data based on gene interaction networks	–	[33]
NetBC	Patient stratification/ clustering	Bi-clustering tumour expression profiles using gene interaction networks	http://mlda.swu.edu.cn/codes.php?name=NetBC	[34]
netDx	Patient classification	Integrate multi-omic patient similarity networks, perform feature selection, classify patients using "guilt by association". Use pathway-level features	http://www.bioconductor.org/packages/release/bioc/html/netDx.html	[16]
DawnRank	N-of-1 driver mutation prediction	Adapt Google's PageRank algorithm to rank potential driver mutations based on position in gene-interaction network	http://genome.compbio.cs.cmu.edu/~jianma/software/DawnRank/	[24]
OncoIMPACT	N-of-1 driver mutation prediction	Ranks driver mutations by minimum path length to misexpressed genes, followed by clustering using minimum set cover	https://github.com/CSB5/OncoIMPACT	[25]
Cytoscape	Network visualization	Visualize and operate on networks. Popular, mature software with app store containing 200+ third-party tools and API integration with R (RCy3). Use cases include visualization of patient similarity networks, gene-interaction networks, networks of feature-selected pathways	https://cytoscape.org	[15, 53, 54]
NDEx	Network repository	Share networks with the community	https://home.ndexbio.org/index/	[50]

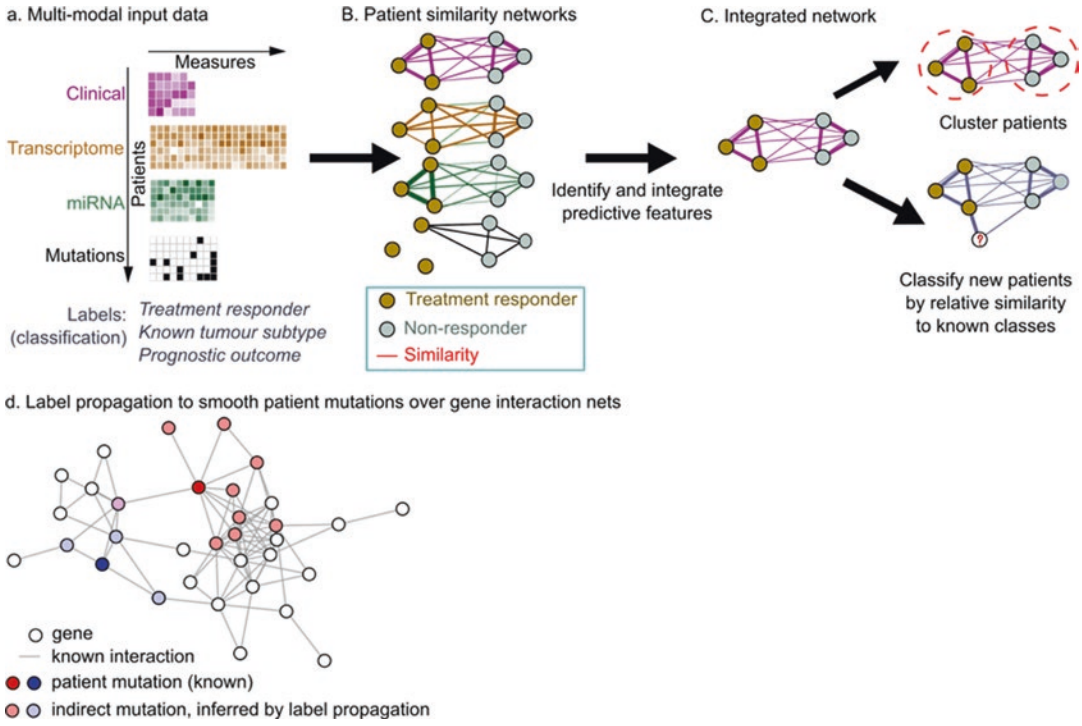

Fig. 11.1 Networks in precision oncology. (**a**) The methods described in this chapter take as input patient data, often from multiple modalities, including clinical data, multiple 'omic layers, sparse genetic data such as somatic mutations or copy number aberrations, and imaging data. Classification methods additionally require labels for each patient, which could reflect clinical outcome. (**b**) Patient similarity networks (PSNs) generated from sample data in (**a**). Nodes are patients, and edges quantify pairwise similarity for a given profile type. Edge weight signifies similarity strength. Node fill indicates known patient label. (**c**) Conversion of multimodal data to a PSN view enables data integration for the purposes of classification or clustering. Classification methods like netDx additionally score input features (represented as PSN) based on ability to predict a given patient label. (**d**) Sparse genetic mutation data can result in overfitting in classification models. To reduce sparsity, prior knowledge about gene-gene interaction networks is used to infer indirect mutations in genes neighbouring those with known mutations. Network-based stratification has demonstrated that inferring indirect mutations using this approach improves tumour subtype discovery in uterine, ovarian, and lung cancers [21]

transcriptomic data, edges quantify how similar the transcriptomic profiles of the corresponding patients are. An unsupervised method, such as one for stratification or clustering, aims to discover patient labels, while a supervised approach aims to maximize discriminability between patients with known labels. In the latter instance, patients may be labelled by outcome of interest, such as prognostic status or treatment response, or by previously identified tumour subtype (Fig. 11.2).

Advantages of PSNs

- The framework provides an advantage for multi-modal data integration, as heterogeneous data types can be converted into a common space of patient similarity; that is, clinical, genomic, and imaging cohort data can be converted to three PSNs, correspondingly representing clinical, genomic, and imaging profile similarity. Conversion to this common patient space allows data integration, which preserves correlation structure of each data layer; this is in contrast to concatenation, which ignores this correlation structure, possibly explaining its worse performance in multi-modal data-based tumour classification [5].
- Machine learning methods that take patient similarity networks as input do not require access to the raw data. In a compute envi-

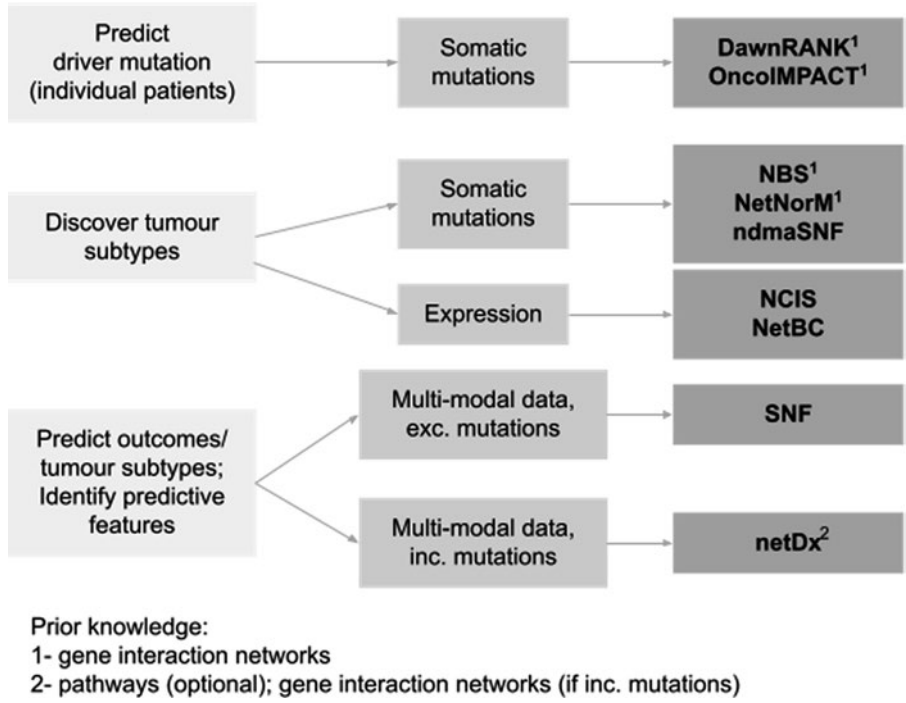

Fig. 11.2 Flowchart to help identify a network-based tool for a precision oncology application

ronment requiring controlled access to patient data – for example, as in the case with genotype data – PSNs can be computed offline and uploaded to the system running the clustering or classification algorithms.

- Similarity networks are an arguably intuitive framework for clinical diagnosis, being conceptually similar to diagnosing a patient based on their profile similarity to patients with known diseases. Interpretability can be improved by feature engineering, such as grouping molecular measures based on prior knowledge of shared regulation; such a design would be more interpretable following feature selection, allowing mechanistic insight.

This chapter will later cover methods that encode multi-modal tumour profiles as PSNs for precision oncology: Similarity Network Fusion for stratification [5], and netDx and MORONET for patient classification [16, 20].

The Value of Interpretability and Prior Knowledge in Genomic Models

There are several reasons for trying to incorporate prior knowledge in genomic models, in order to have interpretability. One main reason is to safeguard against spuriously well-performing models, which may lead to lost research resources pursuing a false lead. A second is to use an interpretable model's insight to drive treatment design. A third still is the need to use prior knowledge to interpret highly sparse data such as patient somatic mutations with low inter-individual recurrence.

Small sample size and overfitting Tumour classification algorithms use a computational strategy called machine learning, in which the algorithm assigns weights to predictive features by partitioning data into a set of training samples and a held-out set of test samples. An iteration of the algorithm involves setting the importance (or weights) for input features based on the training

samples, and evaluating predictive error on the test samples. This process is repeated for different random partitions of training and test data to adjust or "learn" weights, until these have stabilized. Relative to traditional machine learning applications such as computer vision, genomic datasets for a given cancer type tend to be small, on the order of a few hundred samples in total [19]. This small size is particularly problematic when considering that tumour samples can be heterogeneous due to biological sources of variation that may not be directly relevant to the predictive task, including genetics, age, and environmental effects, in addition to disease heterogeneity and technical variation. Models therefore can be prone to overfitting, which occurs when a model demonstrates high performance on the data with which it was trained, but does not generalize to new datasets. Lack of generalization can occur because of bias in the training samples, which can be overcome by better sampling design and increasing sample size. When provided with two models, one which performs well but is a "black box", i.e. lacks transparency about which predictive features contributes to performance, and another which is transparent, the latter is more open to critical evaluation. A transparent model that identifies features consistent with prior knowledge about disruptions in signalling pathways for a given cancer type inspires more confidence in its accuracy, than one which does not.

Hypothesis generation and rational treatment design Learning algorithms, such as clustering or classification algorithms, take features as input. Features may be provided at the level of individual molecules, such as when a regression model identifies weights for each gene in a transcriptomic assay to predict tumour subtype, at the level of biologically meaningful groupings of measures such as pathways, or at the level of an entire data layer [5, 16]. Feature design that is based on prior knowledge of signalling pathways or gene regulatory networks may improve the ability to identify mechanisms contributing to clinical outcome, generating hypotheses for validation and principled treatment design.

Improving signal-to-noise with sparse data Another problem arises with the use of patient somatic mutations. These tend to be sparse and may include no statistically discernible recurrence in individual genes [21, 22]. However, several algorithms have been developed taking advantage of the prior knowledge that oncogenic mutations tend to cluster in a small set of established signalling pathways [21, 23–25]. These algorithms have overlaid patient mutation data with known gene and protein interaction networks, to infer indirect impact of mutations using a guilt-by-association principle. This strategy has been used to improve tumour subtyping in ovarian carcinoma, lung adenocarcinoma, and endometrial carcinoma [21], and predict driver mutations in breast and ovarian cancer [24]. A corollary of this outcome is that prior knowledge of variant significance or non-coding regulation may be useful in selecting which mutations or non-coding genomic measures be included in the model.

These arguments collectively advocate for the use of biologically aware models in precision oncology, an area that network-based representations excel in.

Application Areas

Patient Stratification

When provided with genomic or multi-omic profiles of a tumour type, the first step after data processing is usually to explore data structure and use clustering or class discovery to identify tumour subtypes. Class discovery helps identify subgroups of tumours with shared molecular signatures, which could reflect distinct oncogenic mechanisms and, more relevant to clinical decision-making, outcomes such as survival time, disease progression, and treatment response. It is therefore advisable to cluster even in instances where a project has defined an outcome of interest (e.g. prognosis), as this method provides a means to identify potential alternate sources of biological or technical variation that drive struc-

ture in the data. Once classes are discovered and validated, a classifier may be built for each subtype or alternate methods may be used, such as multi-task learning [26].

Similarity Network Fusion and Similar for Data Integration and Clustering

Similarity Network Fusion (SNF) is an unsupervised or clustering algorithm that integrates multi-modal patient data and identifies patient clusters [5]. It does this by first converting each input data type to a patient similarity network, choosing scaled Euclidean distance as the similarity metric for continuous-valued measures. Input networks are then combined into a single fused network through an iterative step which increases the weight of those edges shared across networks and conversely decreases the weight of unshared edges. The final network is then "cut" into connected communities using spectral clustering, conceptually a dimensionality reduction of network edges. SNF outperformed other approaches in a benchmark to identify clusters in 5 different tumour types by integrating mRNA expression, DNA methylation, and miRNA expression. In particular, SNF-based clustering outperformed the strategy of simply concatenating data measures across all layers, consistent with the hypothesis that creating "views'' of a given data type before integration improves model performance by capturing the correlation structure of each data layer. Yang et al. adapted SNF for use with somatic mutations [27].

Application SNF has been used to discover tumours in pancreatic ductal adenocarcinoma by integrating DNA methylation, mRNA, and miRNA profiles in 150 tumours [28]. This method identified a two-cluster solution consistent with previous characterizations of a basal-like and classical subtype [29], as well as the results obtained by clustering each 'omic layer separately. Separately, the Medulloblastoma Advanced Genomics International Consortium applied SNF to 763 primary frozen medulloblastoma samples with high-quality DNA methylation and transcriptomic profiles, evaluating clustering performance for different choices for the number of clusters [30]. SNF was able to capture previously characterized four subtypes of medulloblastoma, namely, WNT, SSH, Group 3, and Group 4. Importantly, tumours demonstrated robustness of cluster membership across different settings for numbers of clusters.

Network-Based Stratification: Clustering Tumours by Somatic Mutations by Integrating Prior Knowledge

Method While somatic mutations provide a rich source with which to cluster tumours, mutations are sparse – one ovarian cancer cohort typically demonstrated fewer than 100 mutated bases in an entire patient exome [21] – and lack of mutational recurrence is common [22]. This sparsity provides a challenge to clustering tumours. Hofree et al. [21] developed the approach of network-based stratification or NBS, to stratify tumours from somatic mutations, based on the idea that driver mutations impact a regulatory, signalling-related, or metabolic pathway, and therefore, a subnetwork of genes [12]. This idea was originally used in HotNet [23], a computational driver detection algorithm based on a heat diffusion model. Given a gene interaction network constructed from prior knowledge, HotNet first assigns maximum heat to genes with known patient mutations. This heat then diffuses along neighbouring nodes with some decay, and statistical tests are used to identify significantly clustered subnetworks. Similarly, NBS diffuses inferred mutation status from mutated genes in the cohort outward to connected genes in the interaction network, with an applied threshold for the minimum eligible value. The resulting patient-gene mutation matrix, now consisting of direct as well as inferred, indirect mutations is then decomposed using non-negative matrix factorization to identify subgroups with shared commonly mutated genes. NetNorM uses a similar strategy to NBS, except that patient mutations are normalized based on the location of the mutation relative to network topology, instead of smoothing [31]. While NBS used a fixed gene network for all cancer types, He et al. found that creating cancer-specific networks – for example, from co-

expression networks – helped identify subtypes in endometrial carcinomas that were not discovered using original NBS [32]. Similar strategies are used by other algorithms to integrate gene regulatory network information for transcriptomic data, to improve tumour clustering [33, 34].

Application NBS was applied to TCGA exome-sequencing data for ovarian carcinoma, lung adenocarcinoma, and endometrial carcinoma, with consensus clustering used for comparison [21]. In uterine cancer, NBS identified tumour subtypes with significantly higher association with recorded histological subtypes, as compared to those found by consensus clustering. In ovarian cancer, NBS subtypes correlated with overall survival, so that subtypes robustly predicted survival independent of clinical covariates. Importantly, permuting mutated genes in the network, which disrupted the relationship of the mutations to the gene network, abolished the association. Subsequent analysis of mutated subnetworks in ovarian cancer subtypes identified correlates of treatment response and other prior knowledge of cancer networks. For instance, subtype 1 was characterized by shortest platinum-free survival and also contained over 20 mutated genes from the fibroblast growth factor (FGF) pathway, a driver implicated in platinum resistance. NBS and similar algorithms have successfully been used to identify tumour subtypes related to clinical outcomes such as patient survival, treatment response, or tumour histology in ovarian, lung, kidney, prostate, and endometrial cancers [21, 32, 35].

Topological Data Analysis

Li et al. used 73 clinical measures to identify type 2 diabetes subtypes from electronic medical records [36]. In a common approach, patient networks were generated by applying singular value decomposition on variables, with pairwise patient similarity defined as the cosine function. The cosine function is a popular choice in natural language processing applications. Once clusters had been identified, the authors used clinical and genetic data from the same patients to demonstrate enrichment of specific comorbidities and biological pathways in specific subgroups. Although this method was not used in an application of relevance to oncology, it is included for completeness.

Patient Classification

This application area concerns building a predictive model capable of discriminating between tumours with pre-assigned labels, such as molecular subtype, treatment response, or prognosis. These problems are solved by a supervised learning approach called machine learning. As described above, machine learning involves the splitting of samples into a training and a test partition. The training set is used to assign model weights, and the predictive error on the test sample is used to adjust or learn weights, until the predictor error falls below some user-defined threshold. The model's accuracy is then evaluated on an independent validation set.

netDx

netDx is a recently developed supervised learning algorithm for patient classification, which uses the patient similarity network (PSN) paradigm [16, 19]. It additionally provides interpretability through feature engineering, allowing user-defined groupings that incorporate prior biological knowledge such as pathways and sparse somatic mutation data. As input, netDx requires patient labels, multi-modal data for each patient, and user-provided rules for feature engineering. Patient measures can be provided in tabular form (e.g. clinical or transcriptomic measures), or as genomic intervals used to encode sparse somatic mutations or copy number variant (CNV) calls, where each patient has an event at a different genomic locus. Custom similarity metrics can be used, although commonly used metrics such as Pearson correlation and normalized difference also exist.

Model training proceeds with the division of samples into a training and a held-out test partition, with feature selection using only training

samples. Data on training samples are converted to PSN, which serve as input features to the model. A form of constrained regression is used in the PSN edges to score networks, such that networks which tend to connect patients of the same label are upweighted. This step is performed for each patient label in turn, and cross-validation is used to score networks between zero and a user-defined maximum. Networks passing a user-defined threshold are used to classify test patients. Classification of the held-out test set uses a PSN created by integrating selected features, and which contains training as well as test samples. Label propagation is applied to this integrated PSN, starting with training patients, that is, those with known labels. A patient of unknown class is assigned a similarity score for each patient label and patients are classified as the class to which they are most similar. Finally, feature selection and patient scoring are repeated for numerous random train/test splits to identify consistently high-scoring features.

MORONET

MORONET is a recent work which adapts patient similarity networks to a deep learning-based framework, for data integration and patient classification [20]. Deep learning is a relatively recent development in machine learning. In this framework, the classifier is a neural network, consisting of a set of simple non-linear functions (neurons) which are layers in different architectural configurations fine-tuned for specific applications. Deep learning has been particularly successful in computer vision and natural language processing [37]. MORONET uses a graph convolutional network (GCN) to provide input in PSN format to the deep learning model and identifies cross-omics correlations by means of a View Correlation Discovery Network (VCDN), which it uses for label prediction. At the time of this writing, MORONET is limited to being able to handle at most three layers of input data.

Application netDx has been used to predict binarized survival by integrating six data types – clinical, mRNA, miRNA, DNA methylation, proteomics, and somatic copy number aberrations [16]. This design was applied to ovarian cystadenocarcinoma (N = 252 tumours), lung serous carcinoma (N = 77 tumours), glioblastoma (N = 155), and renal clear cell carcinoma (N = 150) data from the Pan-Cancer survival project [16, 38]. In a benchmark, netDx outperformed other machine learning approaches for most applications, with the exception of the small lung cancer dataset where support vector machines were able to find highly non-linear separability. netDx was also used for binary classification of a breast tumour as being either Luminal A or not, using pathway-based features created from mRNA (N = 348 tumours). In addition to excellent performance, netDx identified LumA-predictive pathways consistent with prior knowledge of disrupted regulatory and signalling pathways in this group of breast tumours, including DNA damage repair and cell cycle regulation. Separately, MORONET was able to use mRNA, miRNA, and DNA methylation data to classify tumour subtypes in breast and low-grade glioma, as well as distinguish between three forms of kidney cancer (renal clear cell carcinoma, chromophobe renal cell carcinoma, and papillary renal cell carcinoma) [20]. It outperformed a battery of standard machine learning methods such as K-nearest neighbour, SVM, random forests, and latent models. It was able to identify gene- and miRNA-level biomarkers in breast cancer, including well-known *FOXA1*, *ERBB4*, and *AR* from gene expression and *LRCC25* and *SOSTDC1* from DNA methylation.

Predicting Drivers from N-of-1 Patient Tumour Profiles: DawnRank and OncoIMPACT

While the methods mentioned above require a cohort of samples, another category of methods makes inferences based on single patient profiles using the context of prior knowledge. Current approaches make use of patient-level whole-genome or -transcriptome level data, and sometimes additionally require matched comparisons of tumour and normal samples from the same patient. When provided with somatic genetic

mutations from a single patient, DawnRank uses gene interaction networks to rank genes to infer driver mutations [24]. It borrows the intuition from Google's PageRank algorithm, which ranks websites in a search result based on the number of websites to which a page is linked. In DawnRank, a gene bearing a somatic mutation is ranked more highly if it is connected to genes known to be differentially expressed in cancer, than otherwise. The method uses a damping factor to ensure that the effect of a node drops with graph distance and is applied to the gene interaction network which is modelled as a directed graph.

OncoIMPACT also ranks and calls driver mutations from a single patient's somatic mutation profile by integrating knowledge of gene interactions with tumour-specific gene expression, although its approach is different. When provided with patient somatic mutations, OncoIMPACT first identifies all deregulated genes connected to mutated genes by a minimum path length and where target genes exceed some threshold for tumour-specific differential expression [25]. It then clusters these deregulated genes by the putative causal mutation to which deregulation can be attributed, in a way that parsimoniously explains the deregulation. Generally, these methods represent an advance in providing patient-specific treatment guidance but are limited by requiring whole-genome sequencing of the patient in the clinic.

Applications DawnRank was applied to glioblastoma multiforme (N = 512 samples), breast cancer (N = 504 samples), and ovarian cancer (N = 572) samples from the TCGA, including non-synonymous mutations and insertions/deletions in protein-coding regions and gene expression measures. The underlying gene interaction network was compiled from curated pathway databases such as Reactome [39, 40], NCI-Nature Curated PID [41], and KEGG [42], as well as from MEMo, which generates a single interaction network by integrating multiple functional genomics sources [43]. DawnRank demonstrated greater specificity and sensitivity in prioritizing putative drivers from the Cancer Gene Census (CGC [44]), particularly in breast and ovarian cancers, which may indicate a robustness to a greater number of passenger mutations. In addition to identifying well-known drivers also identified by other methods, including *TP53* and *ATM*, DawnRank additionally discovered *BRCA1*, *CDH1*, and *PIK3R1*, and a novel centromere-associated protein driver in basal breast tumours, *CENPE*. Similarly, OncoIMPACT was able to prioritize candidate driver events in glioblastoma, ovarian, prostate, and bladder cancers, including point mutations and indels, with higher precision than competing methods. OncoIMPACT also identified a previously uncharacterized patient-specific driver mutation in *TRIM24* in melanoma, experimentally validated with siRNA-mediated downregulation in a patient-derived cell line. Subsequent literature search identified a role for TRIM24 in ubiquitin-mediated TP53 degradation in breast cancer.

Challenges, Opportunities, and Perspectives

The use of the PSN framework in biomedical applications is fairly recent, and the strategy needs to be applied widely to develop an appreciation of its relative merits, as compared to more established approaches such as regression, random forests, support vector machines, and even deep learning approaches that do not rely on network-based encoding. Methodological advances as described here will accelerate the ability of these methods to achieve their potential in routine use for precision oncology. Here are some of the challenges and opportunities to better use this paradigm for precision oncology:

Speed, scalability and tunability Analytical methods are needed to improve the ability of PSN-based methods to handle thousands of genomes and thousands of input networks; a pathway-based design currently uses ~2000 input networks, and research is required to identify strategies to maximize signal-to-noise ratio as sample size and feature size increases. Separately, methods are needed to identify predictive fea-

tures when interactions are non-linear. PSN implementations such as netDx need the ability to improve performance by automatically tuning hyperparameters, rather than requiring users to manually try different configurations. This is automatically done for deep learning applications by libraries such as FastAI [45].

Feature engineering to model the non-coding genome Just as gene expression measures can be naturally grouped into pathway-level features as input to classifiers or clustering algorithms, so do filtering and grouping rules need to be developed for other types of data, including somatic mutations, germline variants, miRNA expression, and DNA methylation. A unique challenge is presented by measures in the non-coding genome; 43% disease-associated genetic variants are located in intergenic regions, and this fraction goes up to 88% if one includes introns [46]. Possible solutions for feature engineering of non-coding variants include use of prior knowledge of genome regulation and long-range chromatin structure for grouping by pathways or other meaningful units of genome regulation. Pan-tissue atlases of tissue-specific regulatory regions such as those generated by the NIH Roadmap Consortium, GTEx, and ENCODE3 could be used to tailor feature engineering for cancers affecting specific tissues [47–49]. As an example, predictors for a renal cell carcinoma and pancreatic adenocarcinomas would model non-coding measures correspondingly using kidney- and pancreas-specific genome regulatory maps, rather than using a tissue-agnostic map of genome regulation, such as that general pathways represent. Such prior knowledge is expected to further improve the interpretability of the corresponding features.

Perspectives

Networks are a versatile paradigm to study interactions at different levels of systems biology, from molecules to patients. They have helped classify patients, predict prognosis, identify driver mutations and activated pathways, and integrate multi-modal data. However, the use of networks for precision oncology applications is still in the realm of basic research, with many milestones necessary to develop a mature model used in the clinic, similar to that of OncotypeDx in breast cancer [7]. In particular, models will need to be validated in independent datasets as well as varied ethnicities, be actionable, and have oversight following clinical deployment (see [16] for discussion). At present, networks from published research can be submitted to research repositories such as NDEx [50], creating a reference corpus to find patterns that generalize across different applications in precision oncology. One long-term vision for application of PSNs is in a doctor's clinic of the near future [51]. In this scenario, a physician would be able to generate a PSN profile of a patient, putting the patient in the context of a knowledge bank of PSNs reflecting prognosis and treatment outcome for the cancer type of interest. An interactive web-based interface would automatically classify the patient based on these criteria, quantifying classification uncertainty and listing the features that influenced classification, thereby influencing their treatment plan.

Network-based strategies, and particularly patient similarity networks, provide a means for predictive modelling of genomic data that is conceptually intuitive and biologically grounded in the use of prior biological knowledge for effective model development. The coming years should see an increasing use of this strategy in characterizing tumour profiles for precision oncology.

Box 1: Key Concepts

Networks Data structure representing entities represented as nodes, and quantitative or qualitative relationships between entities represented by weighted or unweighted edges. Commonly used networks in precision oncology include those capturing patient profile similarity, gene-gene or protein-protein interaction networks, or networks of pathways enriched in a given tumour type.

Similarity Metric used to quantify pairwise concordance of profiles in patient similarity networks. Commonly used similarity metrics include correlation for multivariate continuous-valued data such as gene expression, normalized difference for univariate similarity, or cosine similarity for tokenized clinical records. However, the choice of similarity metric can be defined based on the particular application.

Machine learning Algorithms that iteratively fit a mathematical model to data by evaluating goodness-of-fit on a held-out random subsample of data, and do so repeatedly till a user-defined criterion is achieved, either a certain number of cycles or a threshold for error. Common applications of machine learning in precision oncology include patient stratification using unsupervised or clustering algorithms, and tumour classification and outcome prediction using supervised approaches.

Spectral clustering A graph clustering approach used in Similarity Network Fusion [5] for patient stratification. When provided with a similarity network, spectral clustering methods use top eigenvalues to "cut" the network into maximally connected communities.

Deep learning Popular supervised learning approach based on artificial neural network theory that excels at finding non-linear decision boundaries between entities, with successes in image classification and more recently, genomics. The model consists of layers of simple non-linear functions stacked into an overall configuration that is user-definable and tuneable for specific applications. Used by MORONET for patient classification [20].

Label propagation Graph-theoretic algorithm commonly used to "diffuse" values from nodes with data to those without. Applications in precision oncology include classifying patients based on relative similarity to labelled patients in a patient similarity network [16], and inferring impact of a driver mutation on neighbouring nodes or pathway [21, 24, 25].

Training and test set Partitions of patient data used respectively to train model parameters and to evaluate goodness-of-fit on an independent dataset. Used to prevent overfitting.

Overfitting Situation where a machine learning model has artificially high performance owing to having fit biases in the training data, with limited generalizability to other datasets.

Feature A unit of data provided to a machine learning algorithm, which is scored for predictive value and used to interpret a trained model. Features may include individual measures such as expression of a gene or clinical variation. Some network-based methods support grouping of measures into features that reflect prior knowledge. For instance, netDx supports the creation of pathway-level features by grouping gene-level measures [16].

Feature engineering The use of domain-specific prior knowledge to filter or transform data to be provided as input to the model. In precision oncology applications, examples include deciding to group gene-level measures into pathways relevant to cancer, limiting genetic variants to driver mutations or to those correlating with gene expression in a tissue of interest.

Feature selection Machine learning algorithm step where features are scored by predictive value, using training samples.

References

1. Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, Shen R, Taylor AM, Cherniack AD, Thorsson V, Akbani R, Bowlby R, Wong CK, Wiznerowicz M, Sanchez-Vega F, Robertson AG, Schneider BG, Lawrence MS, Noushmehr H, Malta TM, Cancer Genome Atlas Network, Stuart JM, Benz CC, Laird PW. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. Cell. 2018;173:291.
2. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490:61.
3. Sharma T, Schwalbe EC, Williamson D, Sill M, Hovestadt V, Mynarek M, Rutkowski S, Robinson GW, Gajjar A, Cavalli F, Ramaswamy V, Taylor MD, Lindsey JC, Hill RM, Jäger N, Korshunov A, Hicks D, Bailey S, Kool M, Chavez L, Northcott PA, Pfister SM, Clifford SC. Second-generation molecular subgrouping of Medulloblastoma: an International Meta-Analysis of Group 3 and Group 4 subtypes. Acta Neuropathol. 2019;138:309.
4. Mack SC, Witt H, Piro RM, Gu L, Zuyderduyn S, Stütz AM, Wang X, Gallo M, Garzia L, Zayne K, Zhang X, Ramaswamy V, Jäger N, Jones DT, Sill M, Pugh TJ, Ryzhova M, Wani KM, Shih DJ, Head R, Remke M, Bailey SD, Zichner T, Faria CC, Barszczyk M, Stark S, Seker-Cin H, Hutter S, Johann P, Bender S, Hovestadt V, Tzaridis T, Dubuc AM, Northcott PA, Peacock J, Bertrand KC, Agnihotri S, Cavalli FM, Clarke I, Nethery-Brokx K, Creasy CL, Verma SK, Koster J, Wu X, Yao Y, Milde T, Sin-Chan P, Zuccaro J, Lau L, Pereira S, Castelo-Branco P, Hirst M, Marra MA, Roberts SS, Fults D, Massimi L, Cho YJ, Van Meter T, Grajkowska W, Lach B, Kulozik AE, von Deimling A, Witt O, Scherer SW, Fan X, Muraszko KM, Kool M, Pomeroy SL, Gupta N, Phillips J, Huang A, Tabori U, Hawkins C, Malkin D, Kongkham PN, Weiss WA, Jabado N, Rutka JT, Bouffet E, Korbel JO, Lupien M, Aldape KD, Bader GD, Eils R, Lichter P, Dirks PB, Pfister SM, Korshunov A, Taylor MD. Epigenomic alterations define lethal CIMP-positive ependymomas of infancy. Nature. 2014;506:445.
5. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, Haibe-Kains B, Goldenberg A. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods. 2014;11:333.
6. Tang G, Shak S, Paik S, Anderson SJ, Costantino JP, Geyer CE, Mamounas EP, Lawrence Wickerham D, Wolmark N. Comparison of the prognostic and predictive utilities of the 21-gene recurrence score assay and adjuvant! For women with node-negative, ER-positive breast cancer: results from NSABP B-14 and NSABP B-20. Breast Cancer Res Treat. 2011;127:133.
7. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, Baehner FL, Walker MG, Watson D, Park T, Hiller W, Fisher ER, Lawrence Wickerham D, Bryant J, Wolmark N. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med. 2004;351:2817.
8. Wallden B, Storhoff J, Nielsen T, Dowidar N, Schaper C, Ferree S, Liu S, Leung S, Geiss G, Snider J, Vickery T, Davies SR, Mardis ER, Gnant M, Sestak I, Ellis MJ, Perou CM, Bernard PS, Parker JS. Development and verification of the PAM50-based Prosigna breast cancer gene signature assay. BMC Med Genet. 2015;8:54.
9. Cardoso F, van't Veer LJ, Bogaerts J, Slaets L, Viale G, Delaloge S, Pierga J-Y, Brain E, Causeret S, DeLorenzi M, Glas AM, Golfinopoulos V, Goulioti T, Knox S, Matos E, Meulemans B, Neijenhuis PA, Nitz U, Passalacqua R, Ravdin P, Rubio IT, Saghatchian M, Smilde TJ, Sotiriou C, Stork L, Straehle C, Thomas G, Thompson AM, van der Hoeven JM, Vuylsteke P, Bernards R, Tryfonidis K, Rutgers E, Piccart M, MINDACT Investigators. 70-gene signature as an aid to treatment decisions in early-stage breast cancer. N Engl J Med. 2016;375:717.
10. Prat A, Parker JS, Karginova O, Fan C, Livasy C, Herschkowitz JI, He X, Perou CM. Phenotypic and molecular characterization of the Claudin-low intrinsic subtype of breast cancer. Breast Cancer Res. 2010;12:R68.
11. Stephen J, Murray G, Cameron DA, Thomas J, Kunkler IH, Jack W, Kerr GR, Piper T, Brookes CL, Rea DW, van de Velde CJH, Hasenburg A, Markopoulos C, Dirix L, Seynaeve C, Bartlett JMS. Time dependence of biomarkers: non-proportional effects of immunohistochemical panels predicting relapse risk in early breast cancer. Br J Cancer. 2014;111:2242.
12. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW. Cancer genome landscapes. Science. 2013;339:1546.
13. Snel B, Lehmann G, Bork P, Huynen MA. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucleic Acids Res. 2000;28:3442.
14. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, Jensen LJ, von Mering C. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019;47:D607.
15. Merico D, Isserlin R, Stueker O, Emili A, Bader GD. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. PLoS One. 2010;5:e13984.
16. Pai S, Hui S, Isserlin R, Shah MA, Kaka H, Bader GD. netDx: interpretable patient classification using integrated patient similarity networks. Mol Syst Biol. 2019;15:e8497.
17. Tong AHY, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Chang M, Chen Y, Cheng X, Chua G, Friesen H, Goldberg DS, Haynes J, Humphries C, He G, Hussein S, Ke L, Krogan N, Li Z, Levinson JN, Lu H, Ménard P, Munyana C,

Parsons AB, Ryan O, Tonikian R, Roberts T, Sdicu A-M, Shapiro J, Sheikh B, Suter B, Wong SL, Zhang LV, Zhu H, Burd CG, Munro S, Sander C, Rine J, Greenblatt J, Peter M, Bretscher A, Bell G, Roth FP, Brown GW, Andrews B, Bussey H, Boone C. Global mapping of the yeast genetic interaction network. Science. 2004;303:808.

18. Mostafavi S, Morris Q. Fast integration of heterogeneous data sources for predicting gene function with limited annotation. Bioinformatics. 2010;26:1759.
19. Pai S, Bader GD. Patient similarity networks for precision medicine. J Mol Biol. 2018;430:2924.
20. Wang T, Shao W, Huang Z, Tang H, Zhang J, Ding Z, Huang K. MORONET: multi-omics integration via graph convolutional networks for biomedical data classification. bioRXiv Preprint. 2020;184705
21. Hofree M, Shen JP, Carter H, Gross A, Ideker T. Network-based stratification of tumor mutations. Nat Methods. 2013;10:1108.
22. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, Kiezun A, Hammerman PS, McKenna A, Drier Y, Zou L, Ramos AH, Pugh TJ, Stransky N, Helman E, Kim J, Sougnez C, Ambrogio L, Nickerson E, Shefler E, Cortés ML, Auclair D, Saksena G, Voet D, Noble M, DiCara D, Lin P, Lichtenstein L, Heiman DI, Fennell T, Imielinski M, Hernandez B, Hodis E, Baca S, Dulak AM, Lohr J, Landau D-A, Wu CJ, Melendez-Zajgla J, Hidalgo-Miranda A, Koren A, McCarroll SA, Mora J, Crompton B, Onofrio R, Parkin M, Winckler W, Ardlie K, Gabriel SB, Roberts CWM, Biegel JA, Stegmaier K, Bass AJ, Garraway LA, Meyerson M, Golub TR, Gordenin DA, Sunyaev S, Lander ES, Getz G. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013;499:214.
23. Vandin F, Upfal E, Raphael BJ. Algorithms for detecting significantly mutated pathways in cancer. J Comput Biol. 2011;18:507.
24. Hou JP, Ma J. DawnRank: discovering personalized driver genes in cancer. Genome Med. 2014;6:56.
25. Bertrand D, Chng KR, Sherbaf FG, Kiesel A, Chia BKH, Sia YY, Huang SK, Hoon DSB, Liu ET, Hillmer A, Nagarajan N. Patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer omics profiles. Nucleic Acids Res. 2015;43:e44.
26. Ruffalo M, Stojanov P, Pillutla VK, Varma R, Bar-Joseph Z. Reconstructing cancer drug response networks using multitask learning. BMC Syst Biol. 2017;11:96.
27. Yang C, Ge S-G, Zheng C-H. ndmaSNF: cancer subtype discovery based on integrative framework assisted by network diffusion model. Oncotarget. 2017;8:89021.
28. Cancer Genome Atlas Research Network. Electronic address: andrew_aguirre@dfci.harvard.edu and Cancer Genome Atlas Research Network. Integrated genomic characterization of pancreatic ductal adenocarcinoma. Cancer Cell. 2017;32:185.
29. Moffitt RA, Marayati R, Flate EL, Volmar KE, Loeza SGH, Hoadley KA, Rashid NU, Williams LA, Eaton SC, Chung AH, Smyla JK, Anderson JM, Kim HJ, Bentrem DJ, Talamonti MS, Iacobuzio-Donahue CA, Hollingsworth MA, Yeh JJ. Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. Nat Genet. 2015;47:1168.
30. Cavalli FMG, Remke M, Rampasek L, Peacock J, Shih DJH, Luu B, Garzia L, Torchia J, Nor C, Morrissy AS, Agnihotri S, Thompson YY, Kuzan-Fischer CM, Farooq H, Isaev K, Daniels C, Cho B-K, Kim S-K, Wang K-C, Lee JY, Grajkowska WA, Perek-Polnik M, Vasiljevic A, Faure-Conter C, Jouvet A, Giannini C, Nageswara Rao AA, Li KKW, Ng H-K, Eberhart CG, Pollack IF, Hamilton RL, Gillespie GY, Olson JM, Leary S, Weiss WA, Lach B, Chambless LB, Thompson RC, Cooper MK, Vibhakar R, Hauser P, van Veelen M-LC, Kros JM, French PJ, Ra YS, Kumabe T, López-Aguilar E, Zitterbart K, Sterba J, Finocchiaro G, Massimino M, Van Meir EG, Osuka S, Shofuda T, Klekner A, Zollo M, Leonard JR, Rubin JB, Jabado N, Albrecht S, Mora J, Van Meter TE, Jung S, Moore AS, Hallahan AR, Chan JA, Tirapelli DPC, Carlotti CG, Fouladi M, Pimentel J, Faria CC, Saad AG, Massimi L, Liau LM, Wheeler H, Nakamura H, Elbabaa SK, Perezpeña-Diazconti M, de León FCP, Robinson S, Zapotocky M, Lassaletta A, Huang A, Hawkins CE, Tabori U, Bouffet E, Bartels U, Dirks PB, Rutka JT, Bader GD, Reimand J, Goldenberg A, Ramaswamy V, Taylor MD. Intertumoral heterogeneity within Medulloblastoma subgroups. Cancer Cell. 2017;31:737.
31. Le Morvan M, Zinovyev A, Vert J-P. NetNorM: capturing cancer-relevant information in somatic exome mutation data with gene networks for cancer stratification and prognosis. PLoS Comput Biol. 2017;13:e1005573.
32. He Z, Zhang J, Yuan X, Liu Z, Liu B, Tuo S, Liu Y. Network based stratification of major cancers by integrating somatic mutation and gene expression data. PLoS One. 2017;12:e0177662.
33. Liu Y, Gu Q, Hou JP, Han J, Ma J. A network-assisted Co-Clustering algorithm to discover cancer subtypes based on gene expression. BMC Bioinform. 2014;15:37.
34. Yu G, Yu X, Wang J. Network-aided bi-clustering for discovering cancer subtypes. Sci Rep. 2017;7:1046.
35. Zhong X, Yang H, Zhao S, Shyr Y, Li B. Network-based stratification analysis of 13 major cancer types using mutations in panels of cancer genes. BMC Genomics. 2015;16 Suppl 7:S7.
36. Li L, Cheng W-Y, Glicksberg BS, Gottesman O, Tamler R, Chen R, Bottinger EP, Dudley JT. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. Sci Transl Med. 2015;7:311ra174.

37. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436.
38. Yuan Y, Van Allen EM, Omberg L, Wagle N, Amin-Mansour A, Sokolov A, Byers LA, Xu Y, Hess KR, Diao L, Han L, Huang X, Lawrence MS, Weinstein JN, Stuart JM, Mills GB, Garraway LA, Margolin AA, Getz G, Liang H. Assessing the clinical utility of cancer genomic and proteomic data across tumor types. Nat Biotechnol. 2014;32:644.
39. Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, Caudy M, Garapati P, Gillespie M, Kamdar MR, Jassal B, Jupe S, Matthews L, May B, Palatnik S, Rothfels K, Shamovsky V, Song H, Williams M, Birney E, Hermjakob H, Stein L, D'Eustachio P. The Reactome pathway knowledgebase. Nucleic Acids Res. 2014;42:D472.
40. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, Haw R, Jassal B, Korninger F, May B, Milacic M, Roca CD, Rothfels K, Sevilla C, Shamovsky V, Shorser S, Varusai T, Viteri G, Weiser J, Wu G, Stein L, Hermjakob H, D'Eustachio P. The Reactome pathway knowledgebase. Nucleic Acids Res. 2018;46:D649.
41. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH. PID: the pathway interaction database. Nucleic Acids Res. 2009;37:D674.
42. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28:27.
43. Ciriello G, Cerami E, Sander C, Schultz N. Mutual exclusivity analysis identifies oncogenic network modules. Genome Res. 2012;22:398.
44. Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA. The COSMIC cancer gene census: describing genetic dysfunction across all human cancers. Nat Rev Cancer. 2018;18:696.
45. Howard J, Gugger S. Fastai: a layered API for deep learning. Information. 2020;11:108.
46. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, Shafer A, Neri F, Lee K, Kutyavin T, Stehling-Sun S, Johnson AK, Canfield TK, Giste E, Diegel M, Bates D, Hansen RS, Neph S, Sabo PJ, Heimfeld S, Raubitschek A, Ziegler S, Cotsapas C, Sotoodehnia N, Glass I, Sunyaev SR, Kaul R, Stamatoyannopoulos JA. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012;337:1190.
47. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, Amin V, Whitaker JW, Schultz MD, Ward LD, Sarkar A, Quon G, Sandstrom RS, Eaton ML, Wu Y-C, Pfenning AR, Wang X, Claussnitzer M, Liu Y, Coarfa C, Harris RA, Shoresh N, Epstein CB, Gjoneska E, Leung D, Xie W, Hawkins RD, Lister R, Hong C, Gascard P, Mungall AJ, Moore R, Chuah E, Tam A, Canfield TK, Hansen RS, Kaul R, Sabo PJ, Bansal MS, Carles A, Dixon JR, Farh K-H, Feizi S, Karlic R, Kim A-R, Kulkarni A, Li D, Lowdon R, Elliott G, Mercer TR, Neph SJ, Onuchic V, Polak P, Rajagopal N, Ray P, Sallari RC, Siebenthall KT, Sinnott-Armstrong NA, Stevens M, Thurman RE, Wu J, Zhang B, Zhou X, Beaudet AE, Boyer LA, De Jager PL, Farnham PJ, Fisher SJ, Haussler D, Jones SJM, Li W, Marra MA, McManus MT, Sunyaev S, Thomson JA, Tlsty TD, Tsai L-H, Wang W, Waterland RA, Zhang MQ, Chadwick LH, Bernstein BE, Costello JF, Ecker JR, Hirst M, Meissner A, Milosavljevic A, Ren B, Stamatoyannopoulos JA, Wang T, Kellis M. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518:317.
48. ENCODE Project Consortium, Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, Adrian J, Kawli T, Davis CA, Dobin A, Kaul R, Halow J, Van Nostrand EL, Freese P, Gorkin DU, Shen Y, He Y, Mackiewicz M, Pauli-Behn F, Williams BA, Mortazavi A, Keller CA, Zhang X-O, Elhajjajy SI, Huey J, Dickel DE, Snetkova V, Wei X, Wang X, Rivera-Mulia JC, Rozowsky J, Zhang J, Chhetri SB, Zhang J, Victorsen A, White KP, Visel A, Yeo GW, Burge CB, Lécuyer E, Gilbert DM, Dekker J, Rinn J, Mendenhall EM, Ecker JR, Kellis M, Klein RJ, Noble WS, Kundaje A, Guigó R, Farnham PJ, Cherry JM, Myers RM, Ren B, Graveley BR, Gerstein MB, Pennacchio LA, Snyder MP, Bernstein BE, Wold B, Hardison RC, Gingeras TR, Stamatoyannopoulos JA, Weng Z. Expanded Encyclopaedias of DNA elements in the human and mouse genomes. Nature. 2020;583:699.
49. G. Consortium and GTEx Consortium. Genetic effects on gene expression across human tissues. Nature. 2017;550:204.
50. Pratt D, Chen J, Welker D, Rivas R, Pillich R, Rynkov V, Ono K, Miello C, Hicks L, Szalma S, Stojmirovic A, Dobrin R, Braxenthaler M, Kuentzer J, Demchak B, Ideker T. NDEx, the network data exchange. Cell Syst. 2015;1:302.
51. Friend SH, Ideker T. Point: are we prepared for the future doctor visit? Nat Biotechnol. 2011;29:215.
52. Huang JK, Jia T, Carlin DE, Ideker T. pyNBS: a python implementation for network-based stratification of tumor mutations. Bioinformatics. 2018;34:2859.
53. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13:2498.
54. Kucera M, Isserlin R, Arkhangorodsky A, Bader GD. AutoAnnotate: a Cytoscape App for summarizing networks with semantic annotations. F1000Res. 2016;5:1717.

12 Patient-Derived In Vitro and In Vivo Models of Cancer

Sally E. Claridge, Julie-Ann Cavallo, and Benjamin D. Hopkins

Abstract

Over the last two decades, cancer researchers have taken the promise offered by the Human Genome Project and have expanded its capacity to use sequencing to identify the genomic alterations that give rise to and sustain individual tumors. This expansion has allowed researchers to identify and target highly recurrent alterations in specific cancer contexts, such as EGFR mutations in non-small cell lung cancer (Lynch et al, N Engl J Med 350:2129–2139, 2004; Sharifnia et al., Proc Natl Acad Sci U S A 111:18661–18666, 2014), BCR-ABL translocations in chronic myeloid leukemia (Deininger, Pharmacol Rev 55:401–423. https://doi.org/10.1124/pr.55.3.4, 2003; Druker et al, N Engl J Med 344.1038–1042, 2001; Druker et al, N Engl J Med 344:1031–1037. https://doi.org/10.1056/NEJM200104053441401, 2001), or HER2 amplifications in breast cancer (Slamon et al, N Engl J Med 344:783–792. https://doi.org/10.1056/NEJM200103153441101, 2001; Solca et al, Beyond trastuzumab: second-generation targeted therapies for HER-2-positive breast cancer. In: Sibilia M, Zielinski CC, Bartsch R, Grunt TW (eds) Drugs for HER-2-positive breast cancer. Springer, Basel, pp 91–107, 2011). Despite these advances in our capacity to identify the genetic alterations that drive tumor initiation, survival, and proliferation, our ability to target these alterations to provide effective treatment options for patients in need, particularly those with rare or advanced cancers, remains limited (Gould et al, Nat Med 21:431–439. https://doi.org/10.1038/nm.3853, 2015). Patient-derived models of cancer offer one potential mechanism to overcome this barrier between the bench and bedside. Through the development and testing of patient-derived models of cancer, functional genomics efforts can identify tumor-specific drug sensitivities and thereby provide a connection between tumor genetics and effective therapeutics for patients in need of treatment options.

Recognizing that cancer is a multifaceted set of disease states, the development of personalized models of cancer that can be used to compare treatment options, identify tumor-specific vulnerabilities, and guide clinical decision-making has tremendous potential for improving patient outcomes. This chapter will describe a representative set of patient-derived models of cancer, reviewing each of their strengths and weaknesses and highlighting how selecting a model to suit a specific question or context is critical. Each model comes

S. E. Claridge · J.-A. Cavallo · B. D. Hopkins (✉)
Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA
e-mail: benjamin.hopkins@mssm.edu

© Springer Nature Switzerland AG 2022

A. Laganà (ed.), *Computational Methods for Precision Oncology*, Advances in Experimental Medicine and Biology 1361, https://doi.org/10.1007/978-3-030-91836-1_12

with a unique set of pros and cons, making them more or less appropriate for each specific research or clinical question. As each model can be leveraged to gain new insights into cancer biology, the key to their deployment is to identify the most appropriate model for a specific context, while carefully considering the strengths and limitations of the selected model. When used appropriately, patient-derived models may prove to be the missing link needed to bring the promise of personalized oncology to fruition in the clinic.

Introduction: Why Are Patient-Derived Models Important and Useful?

Most cancer subtypes are complex and heterogeneous in histological presentation, genetic variation, and prognostic outcomes. For the most part, engineered models fail to recapitulate this diversity, and natural processes that underpin this diversity, ultimately creating models that fail to recreate the complexity of the disease states. Patient-derived cancer models were developed in order to more closely recapitulate patient tumors and allow us to capture the specifics of individual tumors. They can be loosely defined as any model of cancer that is developed from patient samples. Cancerous patient tissues and/or cells, as compared to genetically engineered cells or animal models, provide the benefit of having evolved in a patient and thereby having the full complement of genomic alterations acquired over time and driven by unique, environmental pressures present in that patient. Every tumor consists of a unique ratio of tumor cells, immune cells, fibroblasts, extracellular scaffolding, and endothelial cells, all of which interact both physically and molecularly. Unlike traditional two-dimensional cancer cell lines, modern patient-derived models of cancer seek to preserve elements of the genetic profile, cell-cell interactions, and physical components of a given tumor.

Patient-derived models of cancer provide tractable platforms with which researchers can test specific hypotheses. No model fully recapitulates the unique context of a patient's tumor and, as a general rule, the greater the complexity of a model, the more limited is the number of ways one can perturb and/or evaluate it (Fig. 12.1). For example, two-dimensional cell line models can be plated in hundreds to thousands of replicates for high-throughput assays, but generating the equivalent number of mouse models is not practically feasible. Even the most complex cancer models do not fully recapitulate the native state of tumors, so when using these models, it is critical to account for the ways in which they do and do not faithfully represent the tumor from which they were derived. Additionally, evaluating agents that target tumor-extrinsic factors, such as the vasculature or immune system, is largely futile in simple systems that lack the complexity to evaluate multicellular (not to mention multisystem) therapeutic responses. Interpretation of the data generated in a specific model and any clinically relevant conclusions thereof are limited to the capacity of said model to recapitulate the tumor from which they were derived. Furthermore, limitations can be present in a variety of different elements within a model system as well as in how it is being assessed. In short, it is critical to recognize the strengths and weaknesses of each model, to evaluate the data generated in each model within the context of its specific capacity and limitations, and to design experiments and workflows accordingly [1].

For decades and at present, cancer modeling has been dominated by the use of two-dimensional cell lines as representations of different tumor types. For much of this time, the focus was war-

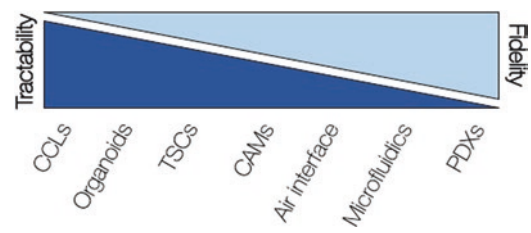

Fig. 12.1 Relative relationship between ease of use in a laboratory setting and fidelity to the original patient across select patient-derived models of cancer. CCLs Cancer cell lines, TSCs tumor slice cultures, CAMs chorioallantoic membrane models, PDXs patient-derived xenografts

ranted since site of origin and pathology were the best available methods for determining diagnosis and treatment. There are many benefits to working with two-dimensional cell lines, since they provide a relatively stable and largely reproducible platform for experimentation and analysis. Unfortunately, they fail to recapitulate many tumor elements that are critical to therapeutic response, e.g., tumor-stromal interactions and microenvironment [2], restricting researchers' capacity to directly translate findings from these models into the clinic. Moreover, over time, cell lines evolve to their culture conditions, losing the heterogeneity and features of the cancer of which they were derived.

While patient-derived models can be powerful tools to study individual tumors, our capacity to use them to study the specific impact of a given gene alteration may be limited since models do not have naturally occurring isogenic controls and, rather, represent the accumulation of all of the alterations in a cell rather than an isolated few. Conversely, however, the genetic complexity of these models may be critical for gaining insight into the tumor's signaling or metabolism or other interactions that play critical role in its sensitivity to therapeutics. One could supplement a patient-derived model with engineered cell lines, e.g., those with targeted oncogenic mutations in genes such as a *KRAS G12D* [3] or deletions of tumor suppressors such as *PTEN* [4], which is beneficial when attempting to demonstrate the relationship between a specific alteration and a given phenotype. Recognizing that each tumor is the result of its own unique environment and the selective pressures to which it was exposed, patient-derived models provide a means to assess each tumor individually. When paired with genomics, this information may prove vital to elucidating the complex interplay between genomics and therapeutic response.

With the advent of rapid next-generation sequencing technologies, there has been a shift in the clinic from a singular focus on tissue of origin toward using molecular diagnostics to inform therapeutic strategies. This shift has allowed for the development and use of agents not focused on tumor type but on targeting the genetic events that drive and sustain individual tumors [5–12] and even agents that target recurrence [13]. While actionable mutations can be predictive of therapeutic response, their predictive power is often context-dependent. To better capture how inter-patient physiological and genetic variation contributes to therapeutic responses, it is imperative to both generate new cancer models and expand upon existing models to more accurately recapitulate what is observed in vivo on a molecular, genetic, histologic, and patient level. This will not only allow for discovery of novel interactions but will help in elucidating drug efficacy and possibly even stratifying patients with similar cancer profiles into new therapeutic groups. In order to continue developing agents that target these tumor-specific alterations, there is a need for the development and use of increasingly personalized and higher-fidelity models. In the next decade, diagnostic approaches that incorporate functional genomics have the potential to become a routine part of patient care, whereby direct assessments of drug sensitivity in patient-derived models could provide an avenue for rapid comparisons and personalization of therapeutics prescribed in the clinic [14–16]. Taken together with our increased capacity to sequence tumors to understand the genetic alterations driving and sustaining tumor growth, development, and treatment response, there has been an expeditious development of myriad patient-derived modeling platforms that allow investigators to assess the effects of different treatment modalities on various aspects of patient tumors (Fig. 12.2). *Combining physiologically relevant information from complimentary model systems could be the key to designing functional pipelines that allow for personalized therapeutic decision-making in the clinic*.

Types of Models

This section will review several patient-derived models of cancer. It is key to note that each model has its own set of strengths and weaknesses and that their use is predicated on understanding and controlling for these system-specific consider-

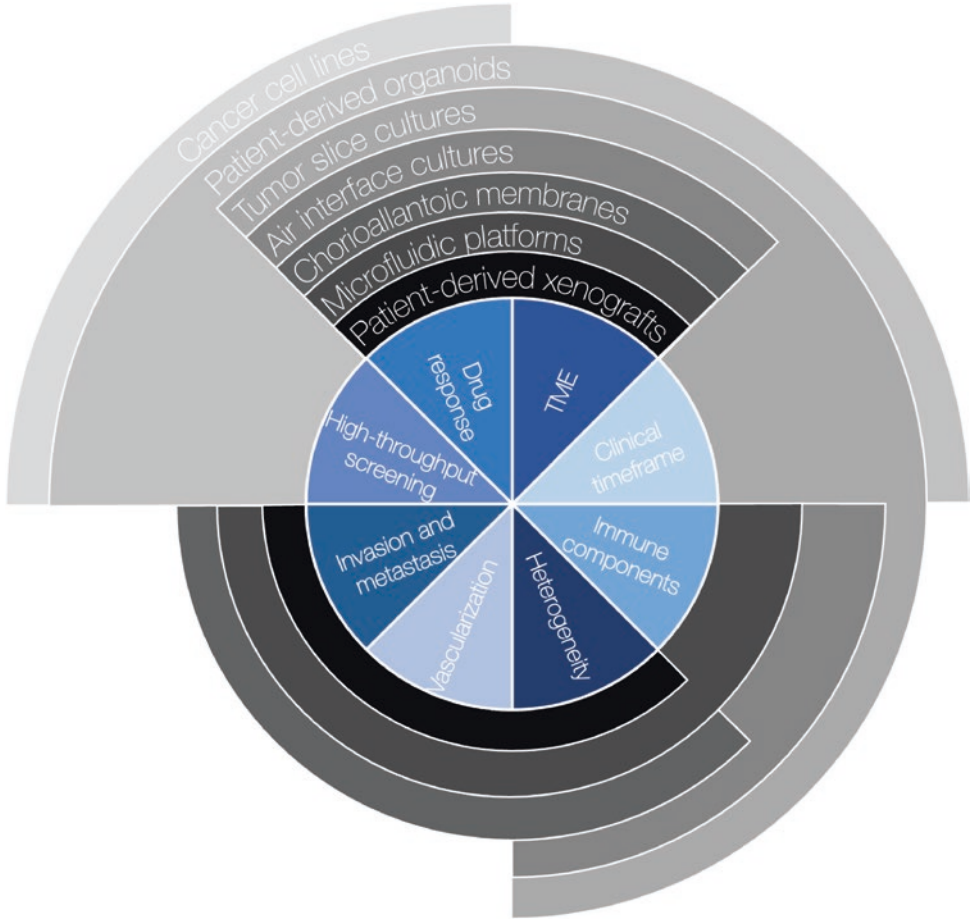

Fig. 12.2 Relative capacity of select patient-derived models of cancer to recapitulate two elements directly related to patient care, high-throughput screening and assessment on a clinical timeframe, and a select six core elements of tumor architecture and proliferation: Ability to quantify response to drugs or other perturbations, tumor microenvironment (TME) factors like stromal cells or extrinsic growth factors, various immune components, intra-tumoral cellular and/or molecular heterogeneity, vascular organization and angiogenesis paradigms, and the ability to invade surrounding tissue or matrix and to metastasize

ations. For example, patient-derived xenografts in humanized mice may recapitulate more elements of the patient condition, but the feasibility, high cost, long timeline, and variable stability of these models limit the applications of their use. Conversely, patient-derived, two-dimensional tumor cell lines are relatively cheap and tractable but fail to reconstitute many of the elements of the tumor and its microenvironment that are targeted by therapeutics, thus limiting the scope and therefore the questions that can be asked with them.

Two-Dimensional Cancer Cell Lines

Two-dimensional cancer cell lines (CCLs) are a well-established model system for testing small molecules, such as the National Cancer Institute 60 (NCI60) project [17, 18]. Over the decades, thousands of commercially available cancer cell lines have been generated [19] and assessed using high-throughput drug and genetic screens. CCLs can also be established directly from dissociated patient tissue. While not always easy to establish, the strength of these models is in the ease of their propagation and culture. Hundreds to thousands

of copies of these lines can be generated in very short timeframes, making them ideal for high-throughput screening. They have been used in pseudo-unselected trials for multiple compounds targeting similar pathways [20] and for pseudo-enrichment trials for drug efficacy in breast cancer [21, 22]. There have also been large-scale efforts by the Dependency Map Consortium (DepMap) to conduct RNAi and CRISPR screens in genomically characterized CCLs to identify cancer type-specific and pan-cancer dependencies [23–26] in conjunction with small molecule screening via the PRISM method that utilizes DNA barcoding to pool cell lines for high-efficiency drug screening [27]. Multiple other groups have also worked to generate publicly available pharmacogenomics datasets, such as the Cancer Cell Line Encyclopedia (CCLE) [28, 29], Genomics of Drug Sensitivity in Cancer (GDSC) [30, 31], and the Cancer Therapeutics Response Portal (CTRP) [32–34].

Established CCLs have been shown to be inconsistent across different laboratories, e.g., in a study of 27 cell lines all labeled as MCF7 (breast cancer), wide genetic variation and response to anticancer therapeutics were recorded [35], but this is an issue inherent in divergent evolution during propagation and variable maintenance practices, which is seen in other in vitro culture systems. As the most broadly used models of cancer, CCLs have been at the center of debates concerning consistency across different datasets, namely, the CCLE and the GDSC. Studies have shown that drug-gene interactions matched between CCLE and GDSC exhibited poor correlation and inconsistencies [36, 37], prompting other groups to join the debate on how best correct for experimental and methodological variation between the original drug screens and subsequent computational analysis [29, 38–43]. In response to these issues, multiple patient-derived tumor modeling platforms, e.g., a customizable Functional Genomics Pipeline [44] and the National Cancer Institute's Patient-Derived Models Repository [45–47], have both implemented fidelity checks that use a combination of genomics and pathology to ensure model fidelity, and also produced datasets that explain the observed differences in drug sensitivity.

As patient-derived culture systems have been developed for the full gambit of tumor types, it is interesting to note that some tumor types, e.g., liver, are highly amenable to growth in two-dimensional culture while being resistant to growth in three dimensions. As such, two-dimensional culture of patient-derived models still has the potential to play a critical role in both precision oncology and cancer research at large. Besides known issues with inconsistent cell line nomenclature and contamination [48], a key outstanding question for the implementation of patient-derived models is how well they model tumor dynamics or recapitulate clinical cancer vulnerabilities. Despite many large-scale grants and clinical trials being fundamentally anchored by results from screens conducted in CCLs, these models tend to require further validation, as their simplicity, which is key to their broad use, also limits their fidelity. Moreover, cancer cell cultures grown in a monolayer lose their three-dimensional architecture and the resultant intercellular interactions. These changes induce changes in gene and protein expression. Thus, patient-derived models from heterogeneous cancers undergo in vitro selection, potentially altering their fidelity to the tumors from which they were derived.

Patient-Derived Organoids

Patient-derived organoids (PDOs) are generated via the dissociation and subsequent expansion of patient tumor samples. Unlike traditional two-dimensional cultures, PDOs are grown in the context of a matrix, e.g., laminin, to generate three-dimensional models of the tumors from which they were derived. Because these cells are grown in a three-dimensional matrix, the models retain some of the structural features and cellular diversity of the tumors from which they were derived (Table 12.1). By culturing PDOs in conditions that mimic the native environment, investigators are able to recapitulate elements of the primary tumor that are lost in two-dimensional

culture, including preserving cell-cell and cell-extracellular matrix interactions.

The PDO system is a middle-ground approach between more complex models and two-dimensional cultures in that it has the capacity for layered complexity through the addition of other cell types (such as T cells) while still allowing investigators to generate thousands of replicates, which can be evaluated in shorter time frames and across more conditions than complex models. PDOs are also amenable to application in other model systems in which the tumor organoids serve as the base patient-derived model that is then made increasingly complex by the addition of alternative platforms, e.g., air interface models, tumor-immune co-cultures, and chorioallantoic membrane models. Furthermore, one can run concurrent-specific and high-throughput drug screens in mice and in three-dimensional culture (with or without co-cultures), respectively. When PDOs are used in co-cultures and other more complex models, one can assess the intricate interplay between tumors and their micro- and macroenvironments, making them a useful tool for the development of both clinical and preclinical pipelines.

Air Interface Cultures

Cultures with an air-liquid interface (ALI) expose three-dimensional organoids to the air rather than encapsulating them in media and allow for long-term propagation of organoids. ALI organoids have been used to study differentiation programs [72], as well as paracrine signaling, and architecture of oncogenically transformed gastrointestinal tissues [73]. This model allows for in vitro profiling of primary tumor epithelium and immune and stromal components from patient biopsies, and it has been shown to accurately model the effects of immunotherapy on endogenous tumor-infiltrating lymphocytes [74]. Non-small cell lung cancer ALI cultures have been used to successfully screen aerosolized drugs, suggesting ALI cultures as a viable alternative to animal studies in regard to studying anticancer drug effects in the respiratory tract and inhalation delivery [75, 76]. These features make ALI models a good fit for studies assessing tumor-immune interactions, but the added complexity of the model increases the barrier to using them at large scale.

Table 12.1 Select publications demonstrating patient-derived organoid development for specific cancer types

Cancer type	References
Bladder	[49, 50]
Brain	[51, 52]
Breast	[53–55]
Colorectal	[56–61]
Gastric	[62]
Gastroesophageal	[61]
Kidney	[63]
Lung	[64, 65]
Ovarian	[66, 67]
Pancreatic	[68, 69]
Prostate	[70, 71]

Chorioallontoic Membrane Models

Chorioallontoic membrane (CAM) models use the developing chicken embryo as the host into which patient-derived tumor models can be implanted in order to evaluate growth interactions with the vasculature and microenvironment. In contrast to the classic in vivo rodent PDX model, CAMs provide a more tractable and cheaper option that is naturally immunodeficient [78], provide an easily manipulated vascular environment, has relatively reduced maintenance requirements, and requires shorter experimental timelines [79]. CAMs have been used to evaluate nanoparticle drug delivery in ovarian cancer [80], in vivo perineural invasion of head and neck squamous cell carcinoma [81], cancer-associated autophagy programs [82], metastatic capacity of non-small cell lung and prostate cancer cells and screening for putative anti-metastatic drugs [83], and the effects of tumor cell invasion on vascular network structure and stability [83]. While CAMs are relatively tractable compared to PDXs, they are not as easily genetically manipulated, and many reagents, e.g., antibodies or cytokines, are incompatible with the avian model. Despite the aforementioned pitfalls of this model, CAMs can be used to provide insight to invasion studies,

elucidate molecular and drug mechanisms, and drug pharmacokinetic and pharmacodynamic studies.

Tumor Slice Cultures

Tumor slice cultures (TSCs) are generated by isolating tumors from patients and creating ex vivo cultures that maintain the composition and orientation of the native tumor microenvironment and extracellular matrix. They are particularly well suited for assessment of the extent to which inter- and intra-tumoral heterogeneity affects the tumor response to therapies. In a TSC, the myriad cell types, vascular networks, and tissue organization from the original tumor are maintained and can be easily imaged and visualized. TSCs are also easily monitored over time, allowing researchers to examine the temporal dynamics of perturbation at the cellular level, a feat that is increasingly complicated in murine models, which cannot be as readily imaged. For example, Minami et al. used real-time and serial imaging and immunohistochemical analysis of drug-treated TSCs from a mouse model of malignant glioma to inform the ways in which organotypic brain slice cultures could be used in testing anti-glioma drugs [84]. Since TSCs maintain the complexity and heterogeneity of the tumor from which they are derived and are rapidly culturable following biopsy, they have potential as personalized preclinical models to stratify individual patients for treatment on a diagnostic timeline. Patient tissue is the limiting reagent in this model, rendering TSCs a low-throughput model. Furthermore, using certain media conditions for each tumor may select for tumor growth over sustaining the growth of other cells (i.e., immune cell populations). Unlike PDXs or cell-based models, these cultures cannot be propagated, only remain viable for limited time windows, and develop abnormal growth kinetics and signaling after approximately 6 days in culture, depending on the tumor type and level of optimization of the culture conditions.

Over time, protocols for slice cultures have evolved to standardize TSC volume and surface area, permitting these models to be used for measuring metabolic activity across TSCs within and across patients, ultimately allowing for quantitative measurements of different biologic activities [85]. Vaira et al. showed that slices taken from human colon, lung, and prostate tumors maintained proliferative capacity and native morphology in culture, and demonstrated reduced proliferation upon treatment with targeted inhibition of Mdm2 and PI3K [86]. Merz et al. demonstrated that patient-derived glioblastoma TSCs recapitulated clinical responses to X-ray, spread-out Bragg-peak carbon irradiation, and temozolomide, suggesting their use in understanding therapeutic effects in glioblastoma and dissecting resistance mechanisms [87]. Similarly, Martin et al. showed that patient-derived TSCs of liver metastases of colorectal cancer could be screened with cetuximab, oxaliplatin, and pembrolizumab in order to identify patient-specific response to these standard of care regimens, suggesting a potential utility for TSCs in personalized oncology [88]. TSCs (ex vivo and from PDX) have also been shown to be a viable system for testing drug efficacy as shown in Table 12.2.

In addition to their potential use in evaluating tumor-specific drug sensitivities, TSCs have the capacity to stably model the tumor micro- and immune environment. Naipal et al. developed culture methods for breast cancer TSCs that maintained tumor and stromal cell morphological and viability characteristics for up to 7 days [93]. TSC treatment with FAC in decreasing dilutions revealed variation in sensitivity to the chemotherapeutic regimen due to morphological and proliferative capacities as well as prior exposure to neo-adjuvant therapy in vivo [93]. Jiang et al. showed that TSCs of pancreatic ductal adenocarcinoma (PDAC) maintained staining for the stromal component α-smooth muscle actin and infiltrating T cells and macrophages between day 1 and 6 of culture [89], while Misra et al. showed that these PDAC models stably preserve the tumor micro- and immune environment, and cancerous cells maintained their proliferative capacity and recapitulated the differentiation grade of the primary tumor, allowing for pharmacologic screening of heterogeneous patient-derived tissue

Table 12.2 Select drug screening experiments in tumor slice culture models

Cancer type	Tumor source	Drugs tested	Reference
Pancreatic ductal adenocarcinoma	Patient	Staurosporine, cycloheximide	[89]
Pancreatic ductal adenocarcinoma	Patient	Rapamycin	[90]
Colon, triple-negative breast cancer	PDX PDX	Staurosporine Panels of FDA-approved drugs	[91]
Colorectal cancer	Patient	Cetuximab, oxaliplatin, and pembrolizumab	[88]
Colorectal cancer	Patient	5-fluorouracil (5-FU) and FOLFOX (5-FU and oxaliplatin)	[92]
Breast	Patient	FAC (5-FU, doxorubicin, 4-HC or preactivated cyclophosphamide)	[93]
Glioma	Mouse	Cisplatin, temozolomide, paclitaxel, tranilast	[84]
Colon, lung, prostate	Human	LY294002 (PI3K inhibitor), Nutlin-3	[86]

[90]. Varying dose responses for 5-FU and FOLFOX were evaluated in patient-derived colorectal cancer TSCs, further underscoring the ability of TSCs to capture inter-patient heterogeneity in drug sensitivity [92]. Sivakumar et al. determined that the immune cell composition in TSCs from syngeneic mouse models of pancreatic, breast, and colon cancer, melanoma, and a primary liver tumor sample remained stable over 7 days in culture and that TSCs from PDX models were valuable models for pharmacologic screening [91]. Taken together, TSC is a high-fidelity model that can be used to address drug sensitivity and mechanisms-of-action, metabolic studies, as well as monitor cell-cell molecular and physical interactions within a patient's tumor, but their use is limited to a relatively small scale and the short time interval for which these cultures remain viable.

Microfluidic Platforms

In microfluidic platforms, synthetic scaffolds made of glass or polymers provide substrate onto which three-dimensional models can be seeded. Three-dimensional microfluidic systems have been used to mimic vascular biology and physiological conditions by flowing fluid through chambers seeded with the cells of interest and monitoring phenotypes of interests [94]. In the cancer context, perfusable microfluidic systems have been shown to recapitulate expected drug toxicities in hepatoblastoma [95], triple-negative breast cancer [96], and head and neck cancer [97], to name a few. These platforms can also be customized to mimic in vivo tumor microenvironment by repopulating decellularized matrix [98], which is an ideal system for pharmacologic screening. Another use case for microfluidic platforms is to test candidate therapeutics against patient-derived single-cell suspensions, which can increase throughput from limited patient sample volume [99]. Lung adenocarcinoma PDX biopsies have been successfully screened with staurosporine in a microfluidic platform, indicating the potential of these systems to maintain physical interactions between cancers and their native tumor microenvironment [100].

Immune checkpoint blockade (ICB) treatment has been tested in murine- and patient-derived organotypic spheroids suspended in perfusable collagen hydrogels, which allows for monitoring of immune cell compositions and profiling of the secretome in response to ICB [101], though this model is restricted to tumor-infiltrating cells and does not address using appropriate immune cell ratios found in the parent patient tissue. Deng et al. showed that patient-derived tumor spheroids in a 3D microfluidic device treated with CDK4/6 inhibitors palbociclib and trilaciclib released increased T-helper 1 cytokines, support-

ing this model system as a future direction for studying the tumor immune microenvironment ex vivo [102]. Microfluidics have been used to identify factors that influence TCR-engineered T cell efficacy against cancerous hepatocytes [103, 104].

More complex microfluidic "organ-on-a-chip" (OOAC) systems can be used to assess how vascularization affects therapeutic efficacy and delivery in co-cultures of endothelial cells, fibroblasts, and tumor cells [105] and have been used to assess dynamics of nanoparticle intravasation from vessels into tumors [106]. Colorectal and breast cancer cells have been shown to grow around and utilize synthetic vasculature and display clinically relevant responses to anticancer therapies, suggesting that these vascularized micro-organs and micro-tumors could be a promising model for therapeutic testing of angiogenesis inhibitors [107]. However, tumor vasculature has been shown to have varying effects on drug delivery (and even anti-angiogenic medications) due to poorly executed neo-angiogenesis and remodeling [108], which may alter drug efficacy modeled in synthetic vasculature. A breast cancer OOAC system successfully modeled ductal carcinoma in situ and mammary tissue layers that recapitulated clinical response to paclitaxel treatment [109]. In another OOAC study, cancer growth and invasion of non-small cell lung cancer was faithfully modeled, as were the effect of mechanical breathing on vascularization and cancer cell response to tyrosine kinase inhibitor rociletinib [110].

While these microfluidic platforms can be difficult and expensive to develop and maintain, the capacity to customize three-dimensional microfluidic platforms makes them ideal for modeling the complex interactions between cell populations and highlights them as a powerful tool for dissecting the molecular mechanisms underpinning treatment efficacy.

Patient-Derived Xenografts

Patient-derived xenografts (PDXs) are generated through the implantation of patient tumor tissue into mice, thus creating a mammalian model of the patient's tumor [111]. Once implanted, these tumors have the capacity to grow, establish a blood supply, and interact with the murine host, thereby providing a living mammalian system in which to run analyses. Once established, PDX models can be propagated; however, much like other patient derived models, they can lose their heterogeneity and be subject to further evolution in their new hosts over time as dominant clones take over and the tumor endures subsequent passaging. While early PDXs can retain elements of the tumor microenvironment, e.g., cancer-associated fibroblasts [112, 113], these elements can be lost over time as the tumors continue to evolve in their new environment. In a study of over a thousand PDXs across more than 18 tumor types, it was found that the selective environment of passaging xenografts in the mouse leads to the accumulation of copy number alterations, which puts evolutionary distance between the primary patient and the model [114]. This supports the notion that passage number may be a critical factor in retaining the complexity of PDX models and underscores the need to evaluate model fidelity not only at the time of development but also at the time of use to ensure that key components of the tumor are being recapitulated as expected.

Increased fidelity to the patient of origin may be obtained for some tumor types by altering the site of xenograft implantation. Traditionally, all PDXs regardless of tumor type have been implanted in the flank of their murine hosts, but through orthotopic implantation into the tissue of origin, researchers may produce a more accurate tumor model [115, 116]. By placing the PDX in a context that more closely resembles its native site of initiation, the model can provide contextual feedbacks specific to the site of origin, e.g., air interface in the lung or microbes in the gut. The type of mouse, e.g., athymic nude or NOD-SCID [111], used can also significantly alter the fidelity of these models and the types of questions that can be addressed. While PDXs have been traditionally generated in immune-compromised mice, in the last decade multiple groups have generated increasingly complex "humanized mice" that are engineered to express human

genes (or mice transplanted with human monocytes) to allow for modeling of interactions between the tumor and the immune system [117]. Adding yet more complexity, other groups are advancing toward being able to generate personalized murine avatars that have patient-matched tumor and immune components [118], though these models are not stable for extended periods, which limits their capacity to model long-term therapeutic responses.

There are several strengths associated with PDX models for drug testing. They represent the gold standard for preclinical models by providing a system where researchers can evaluate the impact of therapies or other perturbations on patient tumors while also evaluating their effects on other organs in a mammalian system. PDXs can be generated in genetically modified animals to evaluate the role of a specific host gene or protein upon tumor development, progression, or drug response [116]. PDX models can be used concurrently with clinical trials to evaluate the impact of drugs on anti-tumor efficacy and to generate a readout of potential toxicities in critical organs [119–121]. Though PDXs can be incredibly powerful in both model capacity and level of fidelity, it is impractical to generate sizeable cohorts of these models to exhaustively test larger drug libraries. Furthermore, their clinical use is limited by factors such as differences in surgical techniques, take rates (i.e., the fraction of tumors that successfully implant to generate a PDX, which is dependent upon multiple factors such as tumor type and implantation site), and timing (the time to generate a cohort of PDXs from a single patient tumor adequate for testing can be months or years depending on growth rates). One could use PDXs to address the efficacy and toxicity of drugs on a patient tumor and organ systems, and supplement this experiment with both genetically similar murine allografts and human xenografts to incorporate the immune system effects and increase the speed and power of the drug study. All together, these features limit the speed at which data from PDX models can be generated and used to identify treatment options for the patient from which they were derived but make them a useful tool for the development of preclinical data sets that can be used to support clinical trials (Box 12.1).

Box 12.1 While not directly derived from patient tissue, it is worth noting that the power of fly genetics allows researchers to design complex *Drosophila* (fly) avatars that represent the individual genetics of a given patient with >15 putative driver mutations. While fly avatars do not provide a mammalian environment, the genomic drivers present in a patient's cancer are modeled in *Drosophila* hindguts, which can then be screened with candidate therapeutics to assess drug toxicity and efficacy in vivo and to quantify animal survival rates in a clinically relevant timeline of 3 weeks [*122*, *123*]. While the fidelity of these models to the initial patient tumor is much lower than that of PDXs, this avatar system is similar in that it can be read out in terms of survival, allowing researchers to gauge tumor-specific treatment efficacy as compared to generic toxicity. Unlike PDXs, these models can be generated and evaluated in clinically relevant time frames and on a sufficiently large scale to allow for broader testing of drug libraries, and ultimately, generating data that can be used to guide clinical decision-making [*124*].

Conclusion

Patient-derived models of cancer have the potential to inform novel therapeutic options and elucidate the complex genetics and multifactorial intercellular interactions underlying cancer phenotypes. Depending on the specific research question, clinical context, or use case, and desired time frame, all patient-derived models have the potential to be the "correct" model, especially when more than one model is used in an orthogonal way or to approach different aspects of the

Table 12.3 Comparison of features of different patient-derived models of cancer

Model	Heterogeneity	Model complexity	Tumor microenvironment	Immune system components	Renewability	Screening capacity	Screening throughput
2-D cell lines	Little to none unless co-cultured	Large preexisting banks can be established from other models or directly form patients	None	None	Can be passaged and banked for future use	Single and combination agent	High
Organoids	Is lost over time, cellular diversity can be enforced through co-culture approaches	Establishment requires customization of culture conditions	TME can be engineered	None	Can be passaged and banked for future use	Single and combination agent	High
Air interface	Low to high	Complex experimental system that requires specialized equipment	TME from patient is maintained	Patient tumor-infiltrating immune cells are maintained	Can be passaged and banked for future use	ICB, single and combination agent	Low
CAM	Low to high	Easy to establish	TME from patient is maintained	Patient tumor-infiltrating immune cells are maintained	Can be regenerated with new eggs and cells from culture	Single and combination agent	Medium
TSC	High	Easy to establish	TME from patient is maintained	Patient tumor-infiltrating immune cells are maintained	Non-renewable	Single and combination agent	Low
Microfluidic co-cultures	Low to high	Complex experimental system that requires specialized equipment	TME is engineered	Immune complexity is engineered	Non-renewable	ICB, single and combination agent	Medium
PDX	High	Labor intensive to establish and grow out	TME from patient is partially maintained	None	Renewable as long as you have mice	ICB, CAR-T, single and combination agent	Low

same question. Key features to consider that were touched on in this chapter are time, cost, scale, and fidelity (Table 12.3). While genomics has rapidly expanded our understanding of tumorigenesis and tumor maintenance, the careful application of appropriate patient-derived models could provide a path to truly personalized oncology by providing platforms to understand the complex interplay between tumors, their environment, and therapeutic sensitivities.

References

1. Gould SE, Junttila MR, de Sauvage FJ. Translational value of mouse models in oncology drug development. Nat Med. 2015;21:431–9. https://doi.org/10.1038/nm.3853.
2. Klemm F, Joyce JA. Microenvironmental regulation of therapeutic response in cancer. Trends Cell Biol. 2015;25:198–213. https://doi.org/10.1016/j.tcb.2014.11.006.
3. Olive KP, Jacobetz MA, Davidson CJ, Gopinathan A, McIntyre D, Honess D, Madhu B, Goldgraben MA, Caldwell ME, Allard D, Frese KK, DeNicola G, Feig C, Combs C, Winter SP, Ireland-Zecchini H, Reichelt S, Howat WJ, Chang A, Dhara M, Wang L, Rückert F, Grützmann R, Pilarsky C, Izeradjene K, Hingorani SR, Huang P, Davies SE, Plunkett W, Egorin M, Hruban RH, Whitebread N, McGovern K, Adams J, Iacobuzio-Donahue C, Griffiths J, Tuveson DA. Inhibition of hedgehog signaling enhances delivery of chemotherapy in a mouse model of pancreatic cancer. Science. 2009;324:1457–61. https://doi.org/10.1126/science.1171362.
4. Podsypanina K, Ellenson LH, Nemes A, Gu J, Tamura M, Yamada KM, Cordon-Cardo C, Catoretti G, Fisher PE, Parsons R. Mutation of Pten/Mmac1 in mice causes neoplasia in multiple organ systems. Proc Natl Acad Sci. 1999;96:1563–8. https://doi.org/10.1073/pnas.96.4.1563.
5. Sos ML, Michel K, Zander T, Weiss J, Frommolt P, Peifer M, Li D, Ullrich R, Koker M, Fischer F, Shimamura T, Rauh D, Mermel C, Fischer S, Stückrath I, Heynck S, Beroukhim R, Lin W, Winckler W, Shah K, LaFramboise T, Moriarty WF, Hanna M, Tolosi L, Rahnenführer J, Verhaak R, Chiang D, Getz G, Hellmich M, Wolf J, Girard L, Peyton M, Weir BA, Chen T-H, Greulich H, Barretina J, Shapiro GI, Garraway LA, Gazdar AF, Minna JD, Meyerson M, Wong K-K, Thomas RK. Predicting drug susceptibility of non–small cell lung cancers based on genetic lesions. J Clin Invest. 2009;119:1727–40. https://doi.org/10.1172/JCI37127.
6. Lynch TJ, Well DW, Sordella R, Gurubhagavatula S, Okimoto RA, Brannigan BW, Harris PL, Haserlat SM, Supko JG, Haluska FG, Louis DN, Christiani DC, Settleman J, Haber DA. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non–small-cell lung cancer to gefitinib. N Engl J Med. 2004;350:2129–39. https://doi.org/10.1056/NEJMoa040938.
7. Sharifnia T, Rusu V, Piccioni F, Bagul M, Imielinski M, Cherniack AD, Pedamallu CS, Wong B, Wilson FH, Garraway LA, Altshuler D, Golub TR, Root DE, Subramanian A, Meyerson M. Genetic modifiers of EGFR dependence in non-small cell lung cancer. Proc Natl Acad Sci U S A. 2014;111:18661–6. https://doi.org/10.1073/pnas.1412228112.
8. Deininger MWN. Specific targeted therapy of chronic myelogenous leukemia with imatinib. Pharmacol Rev. 2003;55:401–23. https://doi.org/10.1124/pr.55.3.4.
9. Druker BJ, Sawyers CL, Kantarjian H, Resta DJ, Reese SF, Ford JM, Capdeville R, Talpaz M. Activity of a specific inhibitor of the BCR-ABL tyrosine kinase in the blast crisis of chronic myeloid leukemia and acute lymphoblastic leukemia with the Philadelphia chromosome. N Engl J Med. 2001;344:1038–42. https://doi.org/10.1056/NEJM200104053441402.
10. Druker BJ, Talpaz M, Resta DJ, Peng B, Buchdunger E, Ford JM, Lydon NB, Kantarjian H, Capdeville R, Ohno-Jones S, Sawyers CL. Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. N Engl J Med. 2001;344:1031–7. https://doi.org/10.1056/NEJM200104053441401.
11. Slamon DJ, Leyland-Jones B, Shak S, Fuchs H, Paton V, Bajamonde A, Fleming T, Eiermann W, Wolter J, Pegram M, Baselga J, Norton L. Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. N Engl J Med. 2001;344:783–92. https://doi.org/10.1056/NEJM200103153441101.
12. Solca FF, Adolf GR, Jones H, Uttenreuther-Fischer MM. Beyond trastuzumab: second-generation targeted therapies for HER-2-positive breast cancer. In: Sibilia M, Zielinski CC, Bartsch R, Grunt TW, editors. Drugs for HER-2-positive breast cancer. Basel: Springer; 2011. p. 91–107. https://doi.org/10.1007/978-3-0346-0094-1_6.
13. Shaw AT, Friboulet L, Leshchiner I, Gainor JF, Bergqvist S, Brooun A, Burke BJ, Deng Y-L, Liu W, Dardaei L, Frias RL, Schultz KR, Logan J, James LP, Smeal T, Timofeevski S, Katayama R, Iafrate AJ, Le L, McTigue M, Getz G, Johnson TW, Engelman JA. Resensitization to crizotinib by the lorlatinib ALK resistance mutation L1198F. N Engl J Med. 2016;374:54–61. https://doi.org/10.1056/NEJMoa1508887.
14. Johannessen CM, Boehm JS. Progress towards precision functional genomics in cancer. Curr Opin

Syst Biol. 2017;2:74–83. https://doi.org/10.1016/j.coisb.2017.02.002.
15. Letai A. Functional precision cancer medicine—moving beyond pure genomics. Nat Med. 2017;23:1028–35. https://doi.org/10.1038/nm.4389.
16. Tyner JW. Integrating functional genomics to accelerate mechanistic personalized medicine. Cold Spring Harb Mol Case Stud. 2017;3:a001370. https://doi.org/10.1101/mcs.a001370.
17. Alley MC, Scudiere DA, Monks A, Hursey ML, Czerwinski MJ, Fine DL, Abbott BJ, Mayo JG, Shoemaker RH, Boyd MR. Feasibility of drug screening with panels of human tumor cell lines using a microculture tetrazolium assay. Cancer Res. 1988;48:589–601. PMID: 3335022.
18. Stinson S, Alley M, Kopp W, Fiebig H, Mullendore L, Pittman A, Kenney S, Keller J, Boyd M. Morphological and immunocytochemical characteristics of human tumor cell lines for use in a disease-oriented anticancer drug screen. Anticancer Res. 1992;12:1035–53. PMID: 1503399.
19. Bairoch A. The cellosaurus, a cell-line knowledge resource. J Biomol Tech. 2018;29:25–38. https://doi.org/10.7171/jbt.18-2902-002.
20. Greshock J, Bachman KE, Degenhardt YY, Jing J, Wen YH, Eastman S, McNeil E, Moy C, Wegrzyn R, Auger K, Hardwicke MA, Wooster R. Molecular target class is predictive of in vitro response profile. Cancer Res. 2010;70:3677–86. https://doi.org/10.1158/0008-5472.CAN-09-3788.
21. Heiser LM, Sadanandam A, Kuo W-L, Benz SC, Goldstein TC, Ng S, Gibb WJ, Wang NJ, Ziyad S, Tong F, Bayani N, Hu Z, Billig JI, Dueregger A, Lewis S, Jakkula L, Korkola JE, Durinck S, Pepin F, Guan Y, Purdom E, Neuvial P, Bengtsson H, Wood KW, Smith PG, Vassilev LT, Hennessy BT, Greshock J, Bachman KE, Hardwicke MA, Park JW, Marton LJ, Wolf DM, Collisson EA, Neve RM, Mills GB, Speed TP, Feiler HS, Wooster RF, Haussler D, Stuart JM, Gray JW, Spellman PT. Subtype and pathway specific responses to anticancer compounds in breast cancer. Proc Natl Acad Sci U S A. 2012;109:2724–9. https://doi.org/10.1073/pnas.1018854108.
22. Marcotte R, Sayad A, Brown KR, Sanchez-Garcia F, Reimand J, Haider M, Virtanen C, Bradner JE, Bader GD, Mills GB, Pe'er D, Moffat J, Neel BG. Functional genomic landscape of human breast cancer drivers, vulnerabilities, and resistance. Cell. 2016;164:293–309. https://doi.org/10.1016/j.cell.2015.11.062.
23. McDonald ER, de Weck A, Schlabach MR, Billy E, Mavrakis KJ, Hoffman GR, Belur D, Castelletti D, Frias E, Gampa K, Golji J, Kao I, Li L, Megel P, Perkins TA, Ramadan N, Ruddy DA, Silver SJ, Sovath S, Stump M, Weber O, Widmer R, Yu J, Yu K, Yue Y, Abramowski D, Ackley E, Barrett R, Berger J, Bernard JL, Billig R, Brachmann SM, Buxton F, Caothien R, Caushi JX, Chung FS, Cortés-Cros M, deBeaumont RS, Delaunay C, Desplat A, Duong W, Dwoske DA, Eldridge RS, Farsidjani A, Feng F, Feng J, Flemming D, Forrester W, Galli GG, Gao Z, Gauter F, Gibaja V, Haas K, Hattenberger M, Hood T, Hurov KE, Jagani Z, Jenal M, Johnson JA, Jones MD, Kapoor A, Korn J, Liu J, Liu Q, Liu S, Liu Y, Loo AT, Macchi KJ, Martin T, McAllister G, Meyer A, Mollé S, Pagliarini RA, Phadke T, Repko B, Schouwey T, Shanahan F, Shen Q, Stamm C, Stephan C, Stucke VM, Tiedt R, Varadarajan M, Venkatesan K, Vitari AC, Wallroth M, Weiler J, Zhang J, Mickanin C, Myer VE, Porter JA, Lai A, Bitter H, Lees E, Keen N, Kauffmann A, Stegmeier F, Hofmann F, Schmelzle T, Sellers WR. Project DRIVE: a compendium of cancer dependencies and synthetic lethal relationships uncovered by large-scale, deep RNAi screening. Cell. 2017;170:577–592.e10. https://doi.org/10.1016/j.cell.2017.07.005.
24. McFarland JM, Ho ZV, Kugener G, Dempster JM, Montgomery PG, Bryan JG, Krill-Burger JM, Green TM, Vazquez F, Boehm JS, Golub TR, Hahn WC, Root DE, Tsherniak A. Improved estimation of cancer dependencies from large-scale RNAi screens using model-based normalization and data integration. Nat Commun. 2018;9:1–13. https://doi.org/10.1038/s41467-018-06916-5.
25. Meyers RM, Bryan JG, McFarland JM, Weir BA, Sizemore AE, Xu H, Dharia NV, Montgomery PG, Cowley GS, Pantel S, Goodale A, Lee Y, Ali LD, Jiang G, Lubonja R, Harrington WF, Strickland M, Wu T, Hawes DC, Zhivich VA, Wyatt MR, Kalani Z, Chang JJ, Okamoto M, Stegmaier K, Golub TR, Boehm JS, Vazquez F, Root DE, Hahn WC, Tsherniak A. Computational correction of copy-number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. Nat Genet. 2017;49:1779–84. https://doi.org/10.1038/ng.3984.
26. Tsherniak A, Vazquez F, Montgomery PG, Weir BA, Kryukov G, Cowley GS, Gill S, Harrington WF, Pantel S, Krill-Burger JM, Meyers RM, Ali L, Goodale A, Lee Y, Jiang G, Hsiao J, Gerath WFJ, Howell S, Merkel E, Ghandi M, Garraway LA, Root DE, Golub TR, Boehm JS, Hahn WC. Defining a cancer dependency map. Cell. 2017;170:564–576.e16. https://doi.org/10.1016/j.cell.2017.06.010.
27. Yu C, Mannan AM, Yvone GM, Ross KN, Zhang Y-L, Marton MA, Taylor BR, Crenshaw A, Gould JZ, Tamayo P, Weir BA, Tsherniak A, Wong B, Garraway LA, Shamji AF, Palmer MA, Foley MA, Winckler W, Schreiber SL, Kung AL, Golub TR. High-throughput identification of genotype-specific cancer vulnerabilities in mixtures of barcoded tumor cell lines. Nat Biotechnol. 2016;34:419–23. https://doi.org/10.1038/nbt.3460.
28. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehár J, Kryukov GV, Sonkin D, Reddy A, Liu M, Murray L, Berger MF, Monahan JE, Morais P, Meltzer J, Korejwa A, Jané-Valbuena J, Mapa FA, Thibault J, Bric-Furlong E, Raman P, Shipway A, Engels IH, Cheng J, Yu GK, Yu J, Aspesi P, de Silva M, Jagtap K, Jones MD, Wang L, Hatton C, Palescandolo E, Gupta

S, Mahan S, Sougnez C, Onofrio RC, Liefeld T, MacConaill L, Winckler W, Reich M, Li N, Mesirov JP, Gabriel SB, Getz G, Ardlie K, Chan V, Myer VE, Weber BL, Porter J, Warmuth M, Finan P, Harris JL, Meyerson M, Golub TR, Morrissey MP, Sellers WR, Schlegel R, Garraway LA. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012;483:603–7. https://doi.org/10.1038/nature11003.

29. Cancer Cell Line Encyclopedia Consortium, Genomics of Drug Sensitivity in Cancer Consortium. Pharmacogenomic agreement between two cancer cell line data sets. Nature. 2015;528:84–7. https://doi.org/10.1038/nature15736.
30. Garnett MJ, Edelman EJ, Heidorn SJ, Greenman CD, Dastur A, Lau KW, Greninger P, Thompson IR, Luo X, Soares J, Liu Q, Iorio F, Surdez D, Chen L, Milano RJ, Bignell GR, Tam AT, Davies H, Stevenson JA, Barthorpe S, Lutz SR, Kogera F, Lawrence K, McLaren-Douglas A, Mitropoulos X, Mironenko T, Thi H, Richardson L, Zhou W, Jewitt F, Zhang T, O'Brien P, Boisvert JL, Price S, Hur W, Yang W, Deng X, Butler A, Choi HG, Chang JW, Baselga J, Stamenkovic I, Engelman JA, Sharma SV, Delattre O, Saez-Rodriguez J, Gray NS, Settleman J, Futreal PA, Haber DA, Stratton MR, Ramaswamy S, McDermott U, Benes CH. Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature. 2012;483:570–5. https://doi.org/10.1038/nature11005.
31. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, Bindal N, Beare D, Smith JA, Thompson IR, Ramaswamy S, Futreal PA, Haber DA, Stratton MR, Benes C, McDermott U, Garnett MJ. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res. 2012;41:D955–61. https://doi.org/10.1093/nar/gks1111.
32. Basu A, Bodycombe NE, Cheah JH, Price EV, Liu K, Schaefer GI, Ebright RY, Stewart ML, Ito D, Wang S, Bracha AL, Liefeld T, Wawer M, Gilbert JC, Wilson AJ, Stransky N, Kryukov GV, Dancik V, Barretina J, Garraway LA, Hon CS, Munoz B, Bittker JA, Stockwell BR, Khabele D, Stern AM, Clemons PA, Shamji AF, Schreiber SL. An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. Cell. 2013;154:1151–61. https://doi.org/10.1016/j.cell.2013.08.003.
33. Rees MG, Seashore-Ludlow B, Cheah JH, Adams DJ, Price EV, Gill S, Javaid S, Coletti ME, Jones VL, Bodycombe NE, Soule CK, Alexander B, Li A, Montgomery P, Kotz JD, Hon CS-Y, Munoz B, Liefeld T, Dančík V, Haber DA, Clish CB, Bittker JA, Palmer M, Wagner BK, Clemons PA, Shamji AF, Schreiber SL. Correlating chemical sensitivity and basal gene expression reveals mechanism of action. Nat Chem Biol. 2016;12:109–16. https://doi.org/10.1038/nchembio.1986.
34. Seashore-Ludlow B, Rees MG, Cheah JH, Cokol M, Price EV, Coletti ME, Jones V, Bodycombe NE, Soule CK, Gould J, Alexander B, Li A, Montgomery P, Wawer MJ, Kuru N, Kotz JD, Hon CS, Munoz B, Liefeld T, Dančík V, Bittker JA, Palmer M, Bradner JE, Shamji AF, Clemons PA, Schreiber SL. Harnessing connectivity in a large-scale small-molecule sensitivity dataset. Cancer Discov. 2015;5:1210–23. https://doi.org/10.1158/2159-8290.CD-15-0235.
35. Ben-David U, Siranosian B, Ha G, Tang H, Oren Y, Hinohara K, Strathdee CA, Dempster J, Lyons NJ, Burns R, Nag A, Kugener G, Cimini B, Tsvetkov P, Maruvka YE, O'Rourke R, Garrity A, Tubelli AA, Bandopadhayay P, Tsherniak A, Vazquez F, Wong B, Birger C, Ghandi M, Thorner AR, Bittker JA, Meyerson M, Getz G, Beroukhim R, Golub TR. Genetic and transcriptional evolution alters cancer cell line drug response. Nature. 2018;560:325–30. https://doi.org/10.1038/s41586-018-0409-3.
36. Haibe-Kains B, El-Hachem N, Birkbak NJ, Jin AC, Beck AH, Aerts HJWL, Quackenbush J. Inconsistency in large pharmacogenomic studies. Nature. 2013;504:389–93. https://doi.org/10.1038/nature12831.
37. Jang IS, Neto EC, Guinney J, Friend SH, Margolin AA. Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. Pac Symp Biocomput. 2014:63–74. PMID: 24297534; PMCID: PMC3995541.
38. Geeleher P, Cox NJ, Huang RS. Cancer biomarker discovery is improved by accounting for variability in general levels of drug sensitivity in pre-clinical models. Genome Biol. 2016;17:190. https://doi.org/10.1186/s13059-016-1050-9.
39. Geeleher P, Gamazon ER, Seoighe C, Cox NJ, Huang RS. Consistency in large pharmacogenomic studies. Nature. 2016;540:E1–2. https://doi.org/10.1038/nature19838.
40. Hatzis C, Bedard PL, Juul Birkbak N, Beck AH, Aerts HJWL, Stern DF, Shi L, Clarke R, Quackenbush J, Haibe-Kains B. Enhancing reproducibility in cancer drug screening: how do we move forward? Cancer Res. 2014;74:4016–23. https://doi.org/10.1158/0008-5472.CAN-14-0725.
41. Haverty PM, Lin E, Tan J, Yu Y, Lam B, Lianoglou S, Neve RM, Martin S, Settleman J, Yauch RL, Bourgon R. Reproducible pharmacogenomic profiling of cancer cell line panels. Nature. 2016;533:333–7. https://doi.org/10.1038/nature17987.
42. Safikhani Z, El-Hachem N, Smirnov P, Freeman M, Goldenberg A, Birkbak NJ, Beck AH, Aerts HJWL, Quackenbush J, Haibe-Kains B. Safikhani et al. reply. Nature. 2016;540:E2–4. https://doi.org/10.1038/nature19839.
43. Safikhani Z, Smirnov P, Freeman M, El-Hachem N, She A, Rene Q, Goldenberg A, Birkbak NJ, Hatzis C, Shi L, Beck AH, Aerts HJWL, Quackenbush J, Haibe-Kains B. Revisiting inconsistency in large pharma-

cogenomic studies. F1000Research. 2017;5:2333. https://doi.org/10.12688/f1000research.9611.3.
44. Pauli C, Hopkins BD, Prandi D, Shaw R, Fedrizzi T, Sboner A, Sailer V, Augello M, Puca L, Rosati R, McNary TJ, Churakova Y, Cheung C, Triscott J, Pisapia D, Rao R, Mosquera JM, Robinson B, Faltas BM, Emerling BE, Gadi VK, Bernard B, Elemento O, Beltran H, Demichelis F, Kemp CJ, Grandori C, Cantley LC, Rubin MA. Personalized in vitro and in vivo cancer models to guide precision medicine. Cancer Discov. 2017;7:462–77. https://doi.org/10.1158/2159-8290.CD-16-1154.
45. Evrard YA, Srivastava A, Randjelovic J, Doroshow JH, Dean DA, Morris JS, Chuang JH. Systematic establishment of robustness and standards in patient-derived xenograft experiments and analysis. Cancer Res. 2020;80:2286–97. https://doi.org/10.1158/0008-5472.CAN-19-3101.
46. Meehan TF, Conte N, Goldstein T, Inghirami G, Murakami MA, Brabetz S, Gu Z, Wiser JA, Dunn P, Begley DA, Krupke DM, Bertotti A, Bruna A, Brush MH, Byrne AT, Caldas C, Christie AL, Clark DA, Dowst H, Dry JR, Doroshow JH, Duchamp O, Evrard YA, Ferretti S, Frese KK, Goodwin NC, Greenawalt D, Haendel MA, Hermans E, Houghton PJ, Jonkers J, Kemper K, Khor TO, Lewis MT, Lloyd KCK, Mason J, Medico E, Neuhauser SB, Olson JM, Peeper DS, Rueda OM, Seong JK, Trusolino L, Vinolo E, Wechsler-Reya RJ, Weinstock DM, Welm A, Weroha SJ, Amant F, Pfister SM, Kool M, Parkinson H, Butte AJ, Bult CJ. PDX-MI: minimal information for patient-derived tumor xenograft models. Cancer Res. 2017;77:e62–6. https://doi.org/10.1158/0008-5472.CAN-17-0582.
47. Tatum JL, Kalen JD, Jacobs PM, Ileva LV, Riffle LA, Hollingshead MG, Doroshow JH. A spontaneously metastatic model of bladder cancer: imaging characterization. J Transl Med. 2019;17:425. https://doi.org/10.1186/s12967-019-02177-y.
48. Yu M, Selvaraj SK, Liang-Chu MMY, Aghajani S, Busse M, Yuan J, Lee G, Peale F, Klijn C, Bourgon R, Kaminker JS, Neve RM. A resource for cell line authentication, annotation and quality control. Nature. 2015;520:307–11. https://doi.org/10.1038/nature14397.
49. Lee SH, Hu W, Matulay JT, Silva MV, Owczarek TB, Kim K, Chua CW, Barlow LJ, Kandoth C, Williams AB, Bergren SK, Pietzak EJ, Anderson CB, Benson MC, Coleman JA, Taylor BS, Abate-Shen C, McKiernan JM, Al-Ahmadie H, Solit DB, Shen MM. Tumor evolution and drug response in patient-derived organoid models of bladder cancer. Cell. 2018;173:515–528.e17. https://doi.org/10.1016/j.cell.2018.03.017.
50. Mullenders J, de Jongh E, Brousali A, Roosen M, Blom JPA, Begthel H, Korving J, Jonges T, Kranenburg O, Meijer R, Clevers HC. Mouse and human urothelial cancer organoids: a tool for bladder cancer research. Proc Natl Acad Sci. 2019;116:4567–74. https://doi.org/10.1073/pnas.1803595116.
51. Ballabio C, Anderle M, Gianesello M, Lago C, Miele E, Cardano M, Aiello G, Piazza S, Caron D, Gianno F, Ciolfi A, Pedace L, Mastronuzzi A, Tartaglia M, Locatelli F, Ferretti E, Giangaspero F, Tiberi L. Modeling medulloblastoma in vivo and with human cerebellar organoids. Nat Commun. 2020;11:583. https://doi.org/10.1038/s41467-019-13989-3.
52. Linkous A, Balamatsias D, Snuderl M, Edwards L, Miyaguchi K, Milner T, Reich B, Cohen-Gould L, Storaska A, Nakayama Y, Schenkein E, Singhania R, Cirigliano S, Magdeldin T, Lin Y, Nanjangud G, Chadalavada K, Pisapia D, Liston C, Fine HA. Modeling patient-derived glioblastoma with cerebral organoids. Cell Rep. 2019;26:3203–3211.e5. https://doi.org/10.1016/j.celrep.2019.02.063.
53. Sachs N, de Ligt J, Kopper O, Gogola E, Bounova G, Weeber F, Balgobind AV, Wind K, Gracanin A, Begthel H, Korving J, van Boxtel R, Duarte AA, Lelieveld D, van Hoeck A, Ernst RF, Blokzijl F, Nijman IJ, Hoogstraat M, van de Ven M, Egan DA, Zinzalla V, Moll J, Boj SF, Voest EE, Wessels L, van Diest PJ, Rottenberg S, Vries RGJ, Cuppen E, Clevers H. A living biobank of breast cancer organoids captures disease heterogeneity. Cell. 2018;172:373–386.e10. https://doi.org/10.1016/j.cell.2017.11.010.
54. Mazzucchelli S, Piccotti F, Allevi R, Truffi M, Sorrentino L, Russo L, Agozzino M, Signati L, Bonizzi A, Villani L, Corsi F. Establishment and morphological characterization of patient-derived organoids from breast cancer. Biol Proced Online. 2019;21:1–3. https://doi.org/10.1186/s12575-019-0099-8.
55. Goldhammer N, Kim J, Timmermans-Wielenga V, Petersen OW. Characterization of organoid cultured human breast cancer. Breast Cancer Res. 2019;21:141. https://doi.org/10.1186/s13058-019-1233-x.
56. Fujii M, Shimokawa M, Date S, Takano A, Matano M, Nanki K, Ohta Y, Toshimitsu K, Nakazato Y, Kawasaki K, Uraoka T, Watanabe T, Kanai T, Sato T. A colorectal tumor organoid library demonstrates progressive loss of niche factor requirements during tumorigenesis. Cell Stem Cell. 2016;18:827–38. https://doi.org/10.1016/j.stem.2016.04.003.
57. Narasimhan V, Wright JA, Churchill M, Wang T, Rosati R, Lannagan TRM, Vrbanac L, Richardson AB, Kobayashi H, Price T, Tye GXY, Marker J, Hewett PJ, Flood MP, Pereira S, Whitney GA, Michael M, Tie J, Mukherjee S, Grandori C, Heriot AG, Worthley DL, Ramsay RG, Woods SL. Medium-throughput drug screening of patient-derived organoids from colorectal peritoneal metastases to direct personalized therapy. Clin Cancer Res. 2020;26:3662–70. https://doi.org/10.1158/1078-0432.CCR-20-0073.
58. Otte J, Dizdar L, Behrens B, Goering W, Knoefel WT, Wruck W, Stoecklein NH, Adjaye J. FGF signalling in the self-renewal of colon cancer organoids.

Sci Rep. 2019;9:17365. https://doi.org/10.1038/s41598-019-53907-7.
59. Roerink SF, Sasaki N, Lee-Six H, Young MD, Alexandrov LB, Behjati S, Mitchell TJ, Grossmann S, Lightfoot H, Egan DA, Pronk A, Smakman N, van Gorp J, Anderson E, Gamble SJ, Alder C, van de Wetering M, Campbell PJ, Stratton MR, Clevers H. Intra-tumour diversification in colorectal cancer at the single-cell level. Nature. 2018;556:457–62. https://doi.org/10.1038/s41586-018-0024-3.
60. Verissimo CS, Overmeer RM, Ponsioen B, Drost J, Mertens S, Verlaan-Klink I, van Gerwen B, van der Ven M, van de Wetering M, Egan DA, Bernards R, Clevers H, Bos JL, Snippert HJ. Targeting mutant RAS in patient-derived colorectal cancer organoids by combinatorial drug screening. eLife. 2016;5:e18489. https://doi.org/10.7554/eLife.18489.
61. Vlachogiannis G, Hedayat S, Vatsiou A, Jamin Y, Fernández-Mateos J, Khan K, Lampis A, Eason K, Huntingford I, Burke R, Rata M, Koh D-M, Tunariu N, Collins D, Hulkki-Wilson S, Ragulan C, Spiteri I, Moorcraft SY, Chau I, Rao S, Watkins D, Fotiadis N, Bali M, Darvish-Damavandi M, Lote H, Eltahir Z, Smyth EC, Begum R, Clarke PA, Hahne JC, Dowsett M, de Bono J, Workman P, Sadanandam A, Fassan M, Sansom OJ, Eccles S, Starling N, Braconi C, Sottoriva A, Robinson SP, Cunningham D, Valeri N. Patient-derived organoids model treatment response of metastatic gastrointestinal cancers. Science. 2018;359:920–6. https://doi.org/10.1126/science.aao2774.
62. Yan HHN, Siu HC, Law S, Ho SL, Yue SSK, Tsui WY, Chan D, Chan AS, Ma S, Lam KO, Bartfeld S, Man AHY, Lee BCH, Chan ASY, Wong JWH, Cheng PSW, Chan AKW, Zhang J, Shi J, Fan X, Kwong DLW, Mak TW, Yuen ST, Clevers H, Leung SY. A comprehensive human gastric cancer organoid biobank captures tumor subtype heterogeneity and enables therapeutic screening. Cell Stem Cell. 2018;23:882–897.e11. https://doi.org/10.1016/j.stem.2018.09.016.
63. Schutgens F, Rookmaaker MB, Margaritis T, Rios A, Ammerlaan C, Jansen J, Gijzen L, Vormann M, Vonk A, Viveen M, Yengej FY, Derakhshan S, de Winter-de Groot KM, Artegiani B, van Boxtel R, Cuppen E, Hendrickx APA, van den Heuvel-Eibrink MM, Heitzer E, Lanz H, Beekman J, Murk J-L, Masereeuw R, Holstege F, Drost J, Verhaar MC, Clevers H. Tubuloids derived from human adult kidney and urine for personalized disease modeling. Nat Biotechnol. 2019;37:303–13. https://doi.org/10.1038/s41587-019-0048-8.
64. Kim M, Mun H, Sung CO, Cho EJ, Jeon H-J, Chun S-M, Jung DJ, Shin TH, Jeong GS, Kim DK, Choi EK, Jeong S-Y, Taylor AM, Jain S, Meyerson M, Jang SJ. Patient-derived lung cancer organoids as in vitro cancer models for therapeutic screening. Nat Commun. 2019;10:3991. https://doi.org/10.1038/s41467-019-11867-6.
65. Li Z, Qian Y, Li W, Liu L, Yu L, Liu X, Wu G, Wang Y, Luo W, Fang F, Liu Y, Song F, Cai Z, Chen W, Huang W. Human lung adenocarcinoma-derived organoid models for drug screening. iScience. 2020;23:101411. https://doi.org/10.1016/j.isci.2020.101411.
66. Hill SJ, Decker B, Roberts EA, Horowitz NS, Muto MG, Worley MJ, Feltmate CM, Nucci MR, Swisher EM, Nguyen H, Yang C, Morizane R, Kochupurakkal BS, Do KT, Konstantinopoulos PA, Liu JF, Bonventre JV, Matulonis UA, Shapiro GI, Berkowitz RS, Crum CP, D'Andrea AD. Prediction of DNA repair inhibitor response in short-term patient-derived ovarian cancer organoids. Cancer Discov. 2018;8:1404–21. https://doi.org/10.1158/2159-8290.CD-18-0474.
67. Kopper O, de Witte CJ, Lõhmussaar K, Valle-Inclan JE, Hami N, Kester L, Balgobind AV, Korving J, Proost N, Begthel H, van Wijk LM, Revilla SA, Theeuwsen R, van de Ven M, van Roosmalen MJ, Ponsioen B, Ho VWH, Neel BG, Bosse T, Gaarenstroom KN, Vrieling H, Vreeswijk MPG, van Diest PJ, Witteveen PO, Jonges T, Bos JL, van Oudenaarden A, Zweemer RP, Snippert HJG, Kloosterman WP, Clevers H. An organoid platform for ovarian cancer captures intra- and interpatient heterogeneity. Nat Med. 2019;25:838–49. https://doi.org/10.1038/s41591-019-0422-6.
68. Nelson SR, Zhang C, Roche S, O'Neill F, Swan N, Luo Y, Larkin A, Crown J, Walsh N. Modelling of pancreatic cancer biology: transcriptomic signature for 3D PDX-derived organoids and primary cell line organoid development. Sci Rep. 2020;10:2778. https://doi.org/10.1038/s41598-020-59368-7.
69. Tiriac H, Belleau P, Engle DD, Plenker D, Deschênes A, Somerville TDD, Froeling FEM, Burkhart RA, Denroche RE, Jang G-H, Miyabayashi K, Young CM, Patel H, Ma M, LaComb JF, Palmaira RLD, Javed AA, Huynh JC, Johnson M, Arora K, Robine N, Shah M, Sanghvi R, Goetz AB, Lowder CY, Martello L, Driehuis E, LeComte N, Askan G, Iacobuzio-Donahue CA, Clevers H, Wood LD, Hruban RH, Thompson E, Aguirre AJ, Wolpin BM, Sasson A, Kim J, Wu M, Bucobo JC, Allen P, Sejpal DV, Nealon W, Sullivan JD, Winter JM, Gimotty PA, Grem JL, DiMaio DJ, Buscaglia JM, Grandgenett PM, Brody JR, Hollingsworth MA, O'Kane GM, Notta F, Kim E, Crawford JM, Devoe C, Ocean A, Wolfgang CL, Yu KH, Li E, Vakoc CR, Hubert B, Fischer SE, Wilson JM, Moffitt R, Knox J, Krasnitz A, Gallinger S, Tuveson DA. Organoid profiling identifies common responders to chemotherapy in pancreatic cancer. Cancer Res. 2018;19:1112–29. https://doi.org/10.1158/2159-8290.CD-18-0349.
70. Drost J, Karthaus WR, Gao D, Driehuis E, Sawyers CL, Chen Y, Clevers H. Organoid culture systems for prostate epithelial tissue and prostate cancer tissue. Nat Protoc. 2016;11:347–58. https://doi.org/10.1038/nprot.2016.006.
71. Puca L, Bareja R, Prandi D, Shaw R, Benelli M, Karthaus WR, Hess J, Sigouros M, Donoghue A,

Kossai M, Gao D, Cyrta J, Sailer V, Vosoughi A, Pauli C, Churakova Y, Cheung C, Deonarine LD, McNary TJ, Rosati R, Tagawa ST, Nanus DM, Mosquera JM, Sawyers CL, Chen Y, Inghirami G, Rao RA, Grandori C, Elemento O, Sboner A, Demichelis F, Rubin MA, Beltran H. Patient derived organoids to model rare prostate cancer phenotypes. Nat Commun. 2018;9:2404. https://doi.org/10.1038/s41467-018-04495-z.

72. Ootani A, Li X, Sangiorgi E, Ho QT, Ueno H, Toda S, Sugihara H, Fujimoto K, Weissman IL, Capecchi MR, Kuo CJ. Sustained in vitro intestinal epithelial culture within a Wnt-dependent stem cell niche. Nat Med. 2009;15:701–6. https://doi.org/10.1038/nm.1951.
73. Li X, Nadauld L, Ootani A, Corney DC, Pai RK, Gevaert O, Cantrell MA, Rack PG, Neal JT, Chan CW-M, Yeung T, Gong X, Yuan J, Wilhelmy J, Robine S, Attardi LD, Plevritis SK, Hung KE, Chen C-Z, Ji HP, Kuo CJ. Oncogenic transformation of diverse gastrointestinal tissues in primary organoid culture. Nat Med. 2014;20:769–77. https://doi.org/10.1038/nm.3585.
74. Neal JT, Li X, Zhu J, Giangarra V, Grzeskowiak CL, Ju J, Liu IH, Chiou S-H, Salahudeen AA, Smith AR, Deutsch BC, Liao L, Zemek AJ, Zhao F, Karlsson K, Schultz LM, Metzner TJ, Nadauld LD, Tseng Y-Y, Alkhairy S, Oh C, Keskula P, Mendoza-Villanueva D, De La Vega FM, Kunz PL, Liao JC, Leppert JT, Sunwoo JB, Sabatti C, Boehm JS, Hahn WC, Zheng GXY, Davis MM, Kuo CJ. Organoid modeling of the tumor immune microenvironment. Cell. 2018;175:1972–1988.e16. https://doi.org/10.1016/j.cell.2018.11.021.
75. Movia D, Bazou D, Volkov Y, Prina-Mello A. Multilayered cultures of NSCLC cells grown at the air-liquid interface allow the efficacy testing of inhaled anti-cancer drugs. Sci Rep. 2018;8:12920. https://doi.org/10.1038/s41598-018-31332-6.
76. Movia D, Bazou D, Prina-Mello A. ALI multilayered co-cultures mimic biochemical mechanisms of the cancer cell-fibroblast cross-talk involved in NSCLC MultiDrug Resistance. BMC Cancer. 2019;19:1–21. https://doi.org/10.1186/s12885-019-6038-x.
77. Leene W, Duyzings MJ, van Steeg C. Lymphoid stem cell identification in the developing thymus and bursa of Fabricius of the chick. Z Zellforsch Mikrosk Anat Vienna Austria 1948. 1973;136:521–33. https://doi.org/10.1007/BF00307368.
78. Nowak-Sliwinska P, Segura T, Iruela-Arispe ML. The chicken chorioallantoic membrane model in biology, medicine and bioengineering. Angiogenesis. 2014;17:779–804. https://doi.org/10.1007/s10456-014-9440-7.
79. Vu BT, Shahin SA, Croissant J, Fatieiev Y, Matsumoto K, Le-Hoang Doan T, Yik T, Simargi S, Conteras A, Ratliff L, Jimenez CM, Raehm L, Khashab N, Durand J-O, Glackin C, Tamanoi F. Chick chorioallantoic membrane assay as an in vivo model to study the effect of nanoparticle-based anticancer drugs in ovarian cancer. Sci Rep. 2018;8:8524. https://doi.org/10.1038/s41598-018-25573-8.
80. Schmitd LB, Liu M, Scanlon CS, Banerjee R, D'Silva NJ. The chick chorioallantoic membrane in vivo model to assess perineural invasion in head and neck cancer. JoVE J Vis Exp. 2019:e59296. https://doi.org/10.3791/59296.
81. Janser FA, Ney P, Pinto MT, Langer R, Tschan MP. The Chick Chorioallantoic Membrane (CAM) assay as a three-dimensional model to study autophagy in cancer cells. Bio-Protoc. 2019;9:e3290.
82. Pawlikowska P, Tayoun T, Oulhen M, Faugeroux V, Rouffiac V, Aberlenc A, Pommier AL, Honore A, Marty V, Bawa O, Lacroix L, Scoazec JY, Chauchereau A, Laplace-Builhe C, Farace F. Exploitation of the chick embryo chorioallantoic membrane (CAM) as a platform for anti-metastatic drug testing. Sci Rep. 2020;10:16876. https://doi.org/10.1038/s41598-020-73632-w.
83. Mangir N, Raza A, Haycock JW, Chapple C, Macneil S. An improved in vivo methodology to visualise tumour induced changes in vasculature using the chick chorionic allantoic membrane assay. In Vivo. 2018;32:461–72. https://doi.org/10.21873/invivo.11262.
84. Minami N, Maeda Y, Shibao S, Arima Y, Ohka F, Kondo Y, Maruyama K, Kusuhara M, Sasayama T, Kohmura E, Saya H, Sampetrean O. Organotypic brain explant culture as a drug evaluation system for malignant brain tumors. Cancer Med. 2017;6:2635–45. https://doi.org/10.1002/cam4.1174.
85. Kenerson HL, Sullivan KM, Seo YD, Stadeli KM, Ussakli C, Yan X, Lausted C, Pillarisetty VG, Park JO, Riehle KJ, Yeh M, Tian Q, Yeung RS. Tumor slice culture as a biologic surrogate of human cancer. Ann Transl Med. 2020;8:114. https://doi.org/10.21037/atm.2019.12.88.
86. Vaira V, Fedele G, Pyne S, Fasoli E, Zadra G, Bailey D, Snyder E, Faversani A, Coggi G, Flavin R, Bosari S, Loda M. Preclinical model of organotypic culture for pharmacodynamic profiling of human tumors. Proc Natl Acad Sci. 2010;107:8352–6. https://doi.org/10.1073/pnas.0907676107.
87. Merz F, Gaunitz F, Dehghani F, Renner C, Meixensberger J, Gutenberg A, Giese A, Schopow K, Hellwig C, Schäfer M, Bauer M, Stöcker H, Taucher-Scholz G, Durante M, Bechmann I. Organotypic slice cultures of human glioblastoma reveal different susceptibilities to treatments. Neuro-Oncol. 2013;15:670–81. https://doi.org/10.1093/neuonc/not003.
88. Martin SZ, Wagner DC, Hörner N, Horst D, Lang H, Tagscherer KE, Roth W. Ex vivo tissue slice culture system to measure drug-response rates of hepatic metastatic colorectal cancer. BMC Cancer. 2019;19:1–14. https://doi.org/10.1186/s12885-019-6270-4.
89. Jiang X, Seo YD, Chang JH, Coveler A, Nigjeh EN, Pan S, Jalikis F, Yeung RS, Crispe IN, Pillarisetty VG. Long-lived pancreatic ductal adenocarci-

noma slice cultures enable precise study of the immune microenvironment. Onco Targets Ther. 2017;6:e1333210. https://doi.org/10.1080/2162402X.2017.1333210.
90. Misra S, Moro CF, Del Chiaro M, Pouso S, Sebestyén A, Löhr M, Björnstedt M, Verbeke CS. Ex vivo organotypic culture system of precision-cut slices of human pancreatic ductal adenocarcinoma. Sci Rep. 2019;9:2133. https://doi.org/10.1038/s41598-019-38603-w.
91. Sivakumar R, Chan M, Shin JS, Nishida-Aoki N, Kenerson HL, Elemento O, Beltran H, Yeung R, Gujral TS. Organotypic tumor slice cultures provide a versatile platform for immuno-oncology and drug discovery. OncoImmunology. 2019;8:e1670019. https://doi.org/10.1080/2162402X.2019.1670019.
92. Sönnichsen R, Hennig L, Blaschke V, Winter K, Körfer J, Hähnel S, Monecke A, Wittekind C, Jansen-Winkeln B, Thieme R, Gockel I, Grosser K, Weimann A, Kubick C, Wiechmann V, Aigner A, Bechmann I, Lordick F, Kallendrusch S. Individual susceptibility analysis using patient-derived slice cultures of colorectal carcinoma. Clin Colorectal Cancer. 2018;17:e189–99. https://doi.org/10.1016/j.clcc.2017.11.002.
93. Naipal KAT, Verkaik NS, Sánchez H, van Deurzen CHM, den Bakker MA, Hoeijmakers JHJ, Kanaar R, Vreeswijk MPG, Jager A, van Gent DC. Tumor slice culture system to assess drug response of primary breast cancer. BMC Cancer. 2016;16:78. https://doi.org/10.1186/s12885-016-2119-2.
94. Bhatia SN, Ingber DE. Microfluidic organs-on-chips. Nat Biotechnol. 2014;32:760–72. https://doi.org/10.1038/nbt.2989.
95. Bhise NS, Manoharan V, Massa S, Tamayol A, Ghaderi M, Miscuglio M, Lang Q, Shrike Zhang Y, Shin SR, Calzone G, Annabi N, Shupe TD, Bishop CE, Atala A, Dokmeci MR, Khademhosseini A. A liver-on-a-chip platform with bioprinted hepatic spheroids. Biofabrication. 2016;8:014101. https://doi.org/10.1088/1758-5090/8/1/014101.
96. Lanz HL, Saleh A, Kramer B, Cairns J, Ng CP, Yu J, Trietsch SJ, Hankemeier T, Joore J, Vulto P, Weinshilboum R, Wang L. Therapy response testing of breast cancer in a 3D high-throughput perfused microfluidic platform. BMC Cancer. 2017;17:1–11. https://doi.org/10.1186/s12885-017-3709-3.
97. Jin D, Ma X, Luo Y, Fang S, Xie Z, Li X, Qi D, Zhang F, Kong J, Li J, Lin B, Liu T. Application of a microfluidic-based perivascular tumor model for testing drug sensitivity in head and neck cancers and toxicity in endothelium. RSC Adv. 2016;6:29598–607. https://doi.org/10.1039/C6RA01456A.
98. Lu S, Cuzzucoli F, Jiang J, Liang L-G, Wang Y, Kong M, Zhao X, Cui W, Li J, Wang S. Development of a biomimetic liver tumor-on-a-chip model based on decellularized liver matrix for toxicity testing. Lab Chip. 2018;18:3379–92. https://doi.org/10.1039/C8LC00852C.
99. Eduati F, Utharala R, Madhavan D, Neumann UP, Longerich T, Cramer T, Saez-Rodriguez J, Merten CA. A microfluidics platform for combinatorial drug screening on cancer biopsies. Nat Commun. 2018;9:2434. https://doi.org/10.1038/s41467-018-04919-w.
100. Holton AB, Sinatra FL, Kreahling J, Conway AJ, Landis DA, Altiok S. Microfluidic biopsy trapping device for the real-time monitoring of tumor microenvironment. PLoS One. 2017;12:e0169797. https://doi.org/10.1371/journal.pone.0169797.
101. Jenkins RW, Aref AR, Lizotte PH, Ivanova E, Stinson S, Zhou CW, Bowden M, Deng J, Liu H, Miao D, He MX, Walker W, Zhang G, Tian T, Cheng C, Wei Z, Palakurthi S, Bittinger M, Vitzthum H, Kim JW, Merlino A, Quinn M, Venkataramani C, Kaplan JA, Portell A, Gokhale PC, Phillips B, Smart A, Rotem A, Jones RE, Keogh L, Anguiano M, Stapleton L, Jia Z, Barzily-Rokni M, Cañadas I, Thai TC, Hammond MR, Vlahos R, Wang ES, Zhang H, Li S, Hanna GJ, Huang W, Hoang MP, Piris A, Eliane J-P, Stemmer-Rachamimov AO, Cameron L, Su M-J, Shah P, Izar B, Thakuria M, LeBoeuf NR, Rabinowits G, Gunda V, Parangi S, Cleary JM, Miller BC, Kitajima S, Thummalapalli R, Miao B, Barbie TU, Sivathanu V, Wong J, Richards WG, Bueno R, Yoon CH, Miret J, Herlyn M, Garraway LA, Allen EMV, Freeman GJ, Kirschmeier PT, Lorch JH, Ott PA, Hodi FS, Flaherty KT, Kamm RD, Boland GM, Wong K-K, Dornan D, Paweletz CP, Barbie DA. Ex vivo profiling of PD-1 blockade using organotypic tumor spheroids. Cancer Discov. 2018;8:196–215. https://doi.org/10.1158/2159-8290.CD-17-0833.
102. Deng J, Wang ES, Jenkins RW, Li S, Dries R, Yates K, Chhabra S, Huang W, Liu H, Aref AR, Ivanova E, Paweletz CP, Bowden M, Zhou CW, Herter-Sprie GS, Sorrentino JA, Bisi JE, Lizotte PH, Merlino AA, Quinn MM, Bufe LE, Yang A, Zhang Y, Zhang H, Gao P, Chen T, Cavanaugh ME, Rode AJ, Haines E, Roberts PJ, Strum JC, Richards WG, Lorch JH, Parangi S, Gunda V, Boland GM, Bueno R, Palakurthi S, Freeman GJ, Ritz J, Nicholas Haining W, Sharpless NE, Arthanari H, Shapiro GI, Barbie DA, Gray NS, Wong K-K. CDK4/6 inhibition augments antitumor immunity by enhancing T-cell activation. Cancer Discov. 2018;8:216–33. https://doi.org/10.1158/2159-8290.CD-17-0915.
103. Lee SWL, Adriani G, Ceccarello E, Pavesi A, Tan AT, Bertoletti A, Kamm RD, Wong SC. Characterizing the role of monocytes in T cell cancer immunotherapy using a 3D microfluidic model. Front Immunol. 2018;9:416. https://doi.org/10.3389/fimmu.2018.00416.
104. Pavesi A, Tan AT, Koh S, Chia A, Colombo M, Antonecchia E, Miccolis C, Ceccarello E, Adriani G, Raimondi MT, Kamm RD, Bertoletti A. A 3D microfluidic model for preclinical evaluation of TCR-engineered T cells against solid tumors. JCI

Insight. 2017;2:e89762. https://doi.org/10.1172/jci.insight.89762.
105. Paek J, Park SE, Lu Q, Park K-T, Cho M, Oh JM, Kwon KW, Yi Y, Song JW, Edelstein HI, Ishibashi J, Yang W, Myerson JW, Kiseleva RY, Aprelev P, Hood ED, Stambolian D, Seale P, Muzykantov VR, Huh D. Microphysiological engineering of self-assembled and perfusable microvascular beds for the production of vascularized three-dimensional human microtissues. ACS Nano. 2019;13:7627–43. https://doi.org/10.1021/acsnano.9b00686.
106. Jarvis M, Arnold M, Ott J, Pant K, Prabhakarpandian B, Mitragotri S. Microfluidic co-culture devices to assess penetration of nanoparticles into cancer cell mass. Bioeng Transl Med. 2017;2:268–77. https://doi.org/10.1002/btm2.10079.
107. Sobrino A, Phan DTT, Datta R, Wang X, Hachey SJ, Romero-López M, Gratton E, Lee AP, George SC, Hughes CCW. 3D microtumors in vitro supported by perfused vascular networks. Sci Rep. 2016;6:31589. https://doi.org/10.1038/srep31589.
108. Seo BR, DelNero P, Fischbach C. In vitro models of tumor vessels and matrix: engineering approaches to investigate transport limitations and drug delivery in cancer. Adv Drug Deliv Rev. 2014;69–70:205–16. https://doi.org/10.1016/j.addr.2013.11.011.
109. Choi Y, Hyun E, Seo J, Blundell C, Kim HC, Lee E, Lee SH, Moon A, Moon WK, Huh D. A microengineered pathophysiological model of early-stage breast cancer. Lab Chip. 2015;15:3350–7. https://doi.org/10.1039/C5LC00514K.
110. Hassell BA, Goyal G, Lee E, Sontheimer-Phelps A, Levy O, Chen CS, Ingber DE. Human organ chip models recapitulate orthotopic lung cancer growth, therapeutic responses, and tumor dormancy in vitro. Cell Rep. 2017;21:508–16. https://doi.org/10.1016/j.celrep.2017.09.043.
111. Hidalgo M, Amant F, Biankin AV, Budinská E, Byrne AT, Caldas C, Clarke RB, de Jong S, Jonkers J, Mælandsmo GM, Roman-Roman S, Seoane J, Trusolino L, Villanueva A. Patient-derived xenograft models: an emerging platform for translational cancer research. Cancer Discov. 2014;4:998–1013. https://doi.org/10.1158/2159-8290.CD-14-0001.
112. Invrea F, Rovito R, Torchiaro E, Petti C, Isella C, Medico E. Patient-derived xenografts (PDXs) as model systems for human cancer. Curr Opin Biotechnol. 2020;63:151–6. https://doi.org/10.1016/j.copbio.2020.01.003.
113. Linxweiler J, Hajili T, Körbel C, Berchem C, Zeuschner P, Müller A, Stöckle M, Menger MD, Junker K, Saar M. Cancer-associated fibroblasts stimulate primary tumor growth and metastatic spread in an orthotopic prostate cancer xenograft model. Sci Rep. 2020;10:12575. https://doi.org/10.1038/s41598-020-69424-x.
114. Ben-David U, Ha G, Tseng Y-Y, Greenwald NF, Oh C, Shih J, McFarland JM, Wong B, Boehm JS, Beroukhim R, Golub TR. Patient-derived xenografts undergo mouse-specific tumor evolution. Nat Genet. 2017;49:1567–75. https://doi.org/10.1038/ng.3967.
115. Erstad DJ, Sojoodi M, Taylor MS, Ghoshal S, Razavi AA, Graham-O'Regan KA, Bardeesy N, Ferrone CR, Lanuti M, Caravan P, Tanabe KK, Fuchs BC. Orthotopic and heterotopic murine models of pancreatic cancer and their different responses to FOLFIRINOX chemotherapy. Dis Model Mech. 2018;11:dmm034793. https://doi.org/10.1242/dmm.034793.
116. Lai Y, Wei X, Lin S, Qin L, Cheng L, Li P. Current status and perspectives of patient-derived xenograft models in cancer research. J Hematol Oncol. 2017;10:106. https://doi.org/10.1186/s13045-017-0470-7.
117. Tian H, Lyu Y, Yang Y-G, Hu Z. Humanized rodent models for cancer research. Front Oncol. 2020;10:1696. https://doi.org/10.3389/fonc.2020.01696.
118. Jespersen H, Lindberg MF, Donia M, Söderberg EMV, Andersen R, Keller U, Ny L, Svane IM, Nilsson LM, Nilsson JA. Clinical responses to adoptive T-cell transfer can be modeled in an autologous immune-humanized mouse model. Nat Commun. 2017;8:707. https://doi.org/10.1038/s41467-017-00786-z.
119. Clohessy JG, Pandolfi PP. Mouse hospital and co-clinical trial project—from bench to bedside. Nat Rev Clin Oncol. 2015;12:491–8. https://doi.org/10.1038/nrclinonc.2015.62.
120. Clohessy JG, Pandolfi PP. The mouse hospital and its integration in ultra-precision approaches to cancer care. Front Oncol. 2018;8:340. https://doi.org/10.3389/fonc.2018.00340.
121. Koga Y, Ochiai A. Systematic review of patient-derived xenograft models for preclinical studies of anti-cancer drugs in solid tumors. Cells. 2019;8:418. https://doi.org/10.3390/cells8050418.
122. Bangi E, Murgia C, Teague AGS, Sansom OJ, Cagan RL. Functional exploration of colorectal cancer genomes using Drosophila. Nat Commun. 2016;7:13615. https://doi.org/10.1038/ncomms13615.
123. Das TK, Esernio J, Cagan RL. Restraining network response to targeted cancer therapies improves efficacy and reduces cellular resistance. Cancer Res. 2018;78:4344–59. https://doi.org/10.1158/0008-5472.CAN-17-2001.
124. Bangi E, Ang C, Smibert P, Uzilov AV, Teague AG, Antipin Y, Chen R, Hecht C, Gruszczynski N, Yon WJ, Malyshev D, Laspina D, Selkridge I, Rainey H, Moe AS, Lau CY, Taik P, Wilck E, Bhardwaj A, Sung M, Kim S, Yum K, Sebra R, Donovan M, Misiukiewicz K, Schadt EE, Posner MR, Cagan RL. A personalized platform identifies trametinib plus zoledronate for a patient with KRAS-mutant metastatic colorectal cancer. Sci Adv. 2019;5:eaav6528. https://doi.org/10.1126/sciadv.aav6528.

Molecular Profiling of Liquid Biopsies for Precision Oncology

13

Edgar E. Gonzalez-Kozlova

Abstract

In recent years, the rapid development of next-generation sequencing (NGS) has led to a significant increase in accuracy toward molecular profiling, allowing noninvasive and real-time detection of novel biomarkers for cancer screening and dynamic monitoring of disease development. Currently, the biggest challenge liquid biopsies face is the selection of the highest signal-bearing tissues (blood/urine or others) and components for diagnosis, being either circulating tumor cells (CTCs), circulating tumor DNA (ctDNA), or extracellular vesicles (EVs). This chapter describes the process of identifying cancer-associated molecular signals from liquid biopsies. First, we address strategies in selecting and processing samples for sequencing, and technical considerations involved in liquid biopsies under three settings: early detection, cancer diagnosis, and metastatic monitoring. Next, we discuss the methods and challenges to identify and validate prognostic signals, such as tumor burden or stage from CTC, targeted and non-targeted mutations from ctDNA, or noncoding RNAs from EVs. Finally, we review the current landscape of novel biomarkers and ongoing clinical trials for liquid biopsies to discuss the potential avenues for future precision medicine and clinical implementation.

E. E. Gonzalez-Kozlova (✉)
Department of Oncological Sciences, Icahn School of Medicine at Mount Sinai,
New York, NY, USA
e-mail: edgar.gonzalez-kozlova@mssm.edu

Circulating Tumor Cells (CTCs)

In 1869, Ashworth et al. published for the first time the existence and isolation of CTCs, describing the study of CTCs as a challenging endeavor due their rarity [5, 82]. Since then, a variety of methods and technologies have been adapted and developed for CTC isolation, such as filtration, chip, ficoll gradient, electric field, and microfluidics [78, 82]. However, the most notorious developments land in the field of microfluidics. The first microfluidic devices processed samples through channels and relied on physical capture and immobilization of CTCs into surfaces coated with specific antibodies, such as μpCTC-Chip and the HBCTC-Chip [5]. Most recent advances allow separating known cellular populations by depletion of leukocytes, erythrocytes, platelets, and noncellular objects, resulting in the enrichment of CTCs [5, 78].

The hypothesis of CTCs preceding metastasis has shown to be true for breast and pancreatic cancer [8, 79, 80]. Additionally, CTCs can be found within the bloodstream with a half-life between 1 and 2.4 hours, which is consistent with the observation that apoptotic CTCs are frequently found in patients with cancer [9, 81].

© Springer Nature Switzerland AG 2022

A. Laganà (ed.), *Computational Methods for Precision Oncology*, Advances in Experimental Medicine and Biology 1361, https://doi.org/10.1007/978-3-030-91836-1_13

Thus, CTCs can be released from either primary or secondary tumors after undergoing strong regulatory conditions, potentially acting as biomarkers for solid tumors and metastatic precursors [1]. Interestingly, CTCs have shown evidence of representing subclones of the primary tumor with the potential of being more invasive [79]. For example, in breast cancer, the median survival of a metastatic patient with epithelial CTCs exceeding a cutoff of 15 CTCs is 6 months. However, for nonmetastatic patients, the survival increased to almost 1.5 years [80]. Their presence is also associated with a higher risk of recurrence and mortality in carcinomas and across all stages of disease [2]. Finally, the presence of notorious molecular markers at the RNA or protein levels such as EPCAM, CK8, CK18, CK19, MET, CD47, or CD44 can be used to determine CTCs; however, their expression can be sparse depending on the type of cancer, which results in false-positive results [82].

Methods for CTCs Identification and Analysis

The sampling of CTCs in the peripheral blood proves to be the first and most challenging step in the study of tumor cells [7, 82] (see Note 1). Most recent advances in microfluidics, microdevices, immunobeads, and functional assays allow the collection of CTCs based on physical properties or/and cellular markers [6, 82]. Physical property-based assays such as dielectrophoretic field flow fractionation use membrane capacitance after polarization to gently filter cells at a flow rate of 1 million cells per minute. However, it requires highly specific parameters such as electric field frequency. Another example is a metacell filtraction device that can isolate CTCs based on cell size, for slightly larger CTCs but fails to isolate smaller CTCs [6, 7].

The cellular heterogeneity is the second obstacle in identifying tumor-specific cells due the sparse expression of the cancer-specific markers, which can lead to falsely identified or misclassified CTCs [10, 81]. Aggressive cancer CTCs have been described to express epithelial cell adhesion molecules (EPCAMs), allowing colonization of multiple tissues such as breast, liver, pancreas, lung, and others. However, they require integrin-based cell adhesion and extracellular matrix degradation mechanisms (RAC and RHOA activity). These molecular mechanisms are variable between CTCs and can be classified into 31 clusters with uniquely associated gene profiles [83]. Finally, sample size is also a complicating factor, which is highly dependent on the downstream analysis, such as whole-genome amplification (WGA) [11, 17], WES [12], and NGS [15, 16]. However, increasing blood sample volumes is a possible solution that provides more accurate measurements, but it comes with its own time constraints and patient care challenges [4]. It is critical to acknowledge that currently there is no single standard regarding the isolation and sequencing of CTCs; however, plenty of technologies are available from standard companies such as Qubit, ThermoFisher, Agilent, and LifeTechnologies [19]. For example, to identify the CNA profiles of CTCs, laser microdissection slides (ThermoFisher) were used to remove CTCs from a slide into a microfluidics syringe for further sequencing with REPLI-g single cell kit (Qiagen) [11]. Ultimately, CTCs can be directly filtered from peripheral blood with a CellCollector (Gilupi) with antibodies against EPCAM, filtering 1.5 liters of blood during 30 minutes. This new device was studied in clinical trials by the European ERA-NET on Translational Cancer Research with promising capture rates (>5 CTCs) per 7.5 ml of blood per patient without adverse effects [11, 84].

The computational burden of analyzing next-generation sequencing data depends entirely on the data quality progressive reduction in costs and accessibility to methods. Thus, NGS is the single most direct and efficient approach to uncover single nucleotide variations (SNVs), copy number variations (CNVs), structure variations (SVs), gene expression, fusions, novel transcripts, alternative splicing, methylation and chromatin patterns on a single cell level with the help of standardized algorithms such as trim_galore, bowtie2, as well as SLIM and dbSUPER, used to identify hypomethylated genes like SOX2

and OCT4, which are associated with poor prognosis in breast cancer [18]. Thus, these results encourage further multi-omic studies in order to identify key signatures and mechanisms associated with cell proliferation and protein synthesis [14]. Moreover, careful combination of multi-omic approaches has great potential to reveal novel biological concepts not previously investigated, such as uncover new cell types in the nervous system [20], immune system, and hematopoietic system [21], as well as new insights into the clonal evolution of cancer [22]. For example, tumor-associated macrophages (TAMs) are major components of the tumor microenvironment, and the CD163+ TAMs correlate with mesenchymal CTC ratio in colorectal cancer metastasis [18, 21, 22]. An in-depth study of these TAMs revealed how they enhance colorectal cancer migration and CTC-mediated metastasis by regulating the JAK/STAT pathway with miR-506-3p/FoxQ1, increasing production of CCL2 and IL6, describing a new cross-talk between immune and tumor cells in the colorectal cancer microenvironment [21].

Successful Studies of CTCs

Single cell RNAseq pipelines for CTCs are very similar to standard guidelines, discussed in greater detail in Chap. 15. These pipelines involve the process of labeling each cell with barcodes using microfluidic devices that isolate a droplet with a single cell next to the barcoded RNAs. This allows to barcode each individual cell RNAs and then sequence all cells in the same machine (Multiplexing). Moreover, it is critical to ensure good RNA quality before sequencing. This can be achieved using a bioanalyzer or a nanodrop to estimate the quality and amount of contaminants in the sample. Later, after sequencing each barcode can be tracked with bioinformatics tools and the transcriptomic profile of each cell can be reconstructed and further analyzed. Briefly, these steps can be summarized as:

- First, rigorous QC of the samples for read quality assessment.

Single cell quality controls are usually embedded with the type of technology used. For example, 10X Genomics performs QC controls and provides detailed statistics of the mapping and barcode ratios in their pipeline Cellranger. Additionally, fastq files can be generated from the raw sequencing outputs and the quality of each read can be assessed through other tools such as trimmomatic [79] or cutadapt [80].

- Second, alignment to the reference genome.

After QC, aligning the resulting reads into a genome of reference is the next step. Although it is default to use the latest reference genome (e.g., Hg38), using alternative, shorter versions of the genome are also possible. This step is achieved with algorithms such as TopHat [81], STAR [82], or Cellranger [83]. Depending on the genome coverage per cell, which is usually linked to the number of cells sequenced, it is possible to use the reads to identify other aspects besides transcription levels, such as clonality in the case of T or B cells for RNAseq. For whole exome sequencing and to approach genomic questions, other tools have proven to be more efficient, such as GATK (any) or DRAGEN (Illumina).

- Third, quantification as Reads-Per-Million (RPM), FPKM (Fragments Per Kilobase Million), or CPM (Counts-Per-Million).

The next step is to quantify the results reads or counts into a measurement that is representative of the sample profile. For single cell transcriptomics, the data can be normalized into RPMs (best to compare genes of similar lengths) or FPKMs (more accurate comparisons between genes of different lengths). However, for visualization, it is common to transform the data into a normal distribution per gene (Z-Score) and clip the maximum values to better illustrate the expression profiles per cell.

- Fourth, application of a preferred unsupervised clustering [24, 25] or trajectory analysis [26] and reduction of dimensionality (UMAP or tSNE).

Finally, the analysis of transcriptomic and genomic alignments can be separated into unsupervised and supervised methods. Unsupervised methods such as hierarchical, spectral, or agglomerative clustering, DBScan or others, allow to define patterns in groups of samples. However, their performance is totally dependent on the amount and type of noise present in each study. By the contrary, supervised methods, such as nearest neighbors, regression, or Bayesian inference approaches can handle noisy datasets better, but have a disadvantage in the discovery of rare profiles. Most methods are available in R, python, or matlab packages and commonly available in repositories like github.

Clustering has been used routinely for individual conditions or punctual comparisons trajectory to infer the topology associations to biological features or time-dependent events [28].

This process is well illustrated in elegant studies done by Miyamoto et al. [16], where CTCs were sampled from 13 prostate cancer patients (under treatment with an AR inhibitor) and untreated cases were collected, sequenced, and analyzed. This retrospective analysis of CTCs indicates activation of noncanonical Wnt signaling and that ectopic expression of Wnt5a in prostate cancer cells attenuates the antiproliferative effect of AR inhibition. Whereas its suppression in drug-resistant cells restores partial sensitivity to treatment. A key highlight from this study is that the authors define CTCs by setting a very clear and stringent thresholding based on the log10(RPM) values for CD45 and CD16 markers, in addition to KRT7, KRT8, KRT18, KRT19, EPCAM, AR, KLK3 (PSA), FOLH1 (PSMA), and AMACR (prostate-specific and epithelial markers), to avoid specimens containing contaminants, leukocytes, and other cells (see Note 2). Finally, the findings are independently validated using a parallel set of samples through qRT-PCR or dPCR.

Another outstanding example is a work by Yu-Heng Cheng et al. [27], demonstrating a creative use of microfluidics through a method labeled "Hydro-Seq," which enables separation of CTCs and normal cells based on morphology or size selection. This approach does not rely on specific markers, which is an advantage in the context of heterogeneous expression of a specific marker. The downstream sequencing and computational analysis followed the same profile as summarized above.

Whole genomes approaches [29, 30] provide a platform to investigate the mutational profiles and genomic modifications CTCs can present as elucidated by Szczerba et al. [12]. Their research isolated CTCs associated to white blood cells, specifically neutrophils. These cells showed a number of differentially expressed genes associated with cycle progression, cell-cell junction, and cytokine receptors. Moreover, whole exome sequencing showed a unique mutational signature from metastatic inducing CTCs, in pair base alterations (C > T). These results suggest that the association between neutrophils and CTCs drives cell cycle progression within the bloodstream and expands the metastatic potential of CTCs, providing a rationale for targeting this interaction in treatment of breast cancer.

Cell-Free DNA (cfDNA) and Circular Tumor DNA (ctDNA)

Cell-free nucleic acids in human blood were first described in 1948 by Mandel Metais [78]. Subsequently, circulating cell-free DNA has been found to range between 1 and 10 ng/ml in plasma of healthy individuals [4]. In 1989, Stroun et al. reported that at least some cfDNA in the plasma of cancer patients originated from cancer cells. Since then, studies have shown to detect mutations in key genes like TP53 and KRAS mutations matching colorectal, pancreatic, and lung cancers from liquid biopsies such as plasma, stool, and sputum samples. Thus, the cfDNA subset of tumoral origins is labeled as circulating tumor DNA or ctDNA.

The presence of ctDNA in human blood was officially first discovered in the 1950s [4]. Later, it was observed that cancer patients had also a higher concentration of cfDNA in serum and plasma probably due to the release of ctDNA from cancer cells [5]. However, the exact origin and release mechanisms are still not fully understood, but there are clear fragment size characteristics of cfDNA (>150 bp) and ctDNA (<150 bp), which suggests that most DNA is released possibly through apoptosis, active release, or similar mechanisms [39].

Methods for cfDNA and ctDNA Identification and Analysis

The main challenge of ctDNA analysis is the identification of a signal that is cancer specific, which can be done by using a series of experimental methods based on the approach described in [23] (see Note 1). The first and most direct method is to test known markers through quantitative PCR (qPCR) or digital PCR (dPCR), where both techniques excel in sensitivity, are rapid and cost-effective [32]. The second direction in ctDNA analysis involves several sequencing strategies. These consist of NGS techniques that can be developed for targeted (panel genes) and nontargeted analysis (WES & sWGS) reviewed in detail elsewhere [31]. However, because of the unique short fragment nature of ctDNA, it is incompatible with long read sequencing. Outside the blood circulation, cfDNA has been detected in various body fluids, including urine, cerebrospinal fluid, pleural fluid, and saliva. These sample types potentially harbor biomarkers and are flexible sample sources that have not been explored in detail.

Moreover, ctDNA approaches can be easily repeated over time due the easy access to samples to monitor the molecular evolution of the disease in the absence of clinical progression, of great interest for precision and personalized medicine. However, using ctDNA for early cancer diagnosis or therapeutic response monitoring can be challenging due to the low amount of tumor DNA detected in the circulation [33, 36]. Additionally, ctDNA may be diluted due the presence of DNA from nontumor cells. Development of new technologies such as dPCR or optimized targeted next-generation sequencing (NGS) has greatly improved the sensitivity, specificity, and precision for the detection of rare DNA sequences and copy number variations [35].

Successful Studies of cfDNA and ctDNA

Early cancer diagnosis may face many technical challenges for ctDNA approaches [38]. However, recent research has demonstrated the efficacy of patient profiling using ctDNA for various cancer types [33], setting the path for its usage in phase I trials. A breast cancer study named "TARGET" built a target panel of 641 cancer-associated genes and identified actionable mutations in 41 out of 100 patients, 11 of whom received matched therapy [37]. Similarly, another randomized clinical trial I/II with over 600 patients showed that ctDNA quantification of PIK3CA levels predicted progression free survival. Additionally, it allowed us to closely monitor the effect of various anti-tumoral drugs over breast cancer [34]. Despite the limitations of extracting ctDNA from liquid biopsies, the detection of ctDNA in itself is a good predictor of disease severity [46] (see Note 3).

FDA-approved panels, such as the Memorial Sloan Kettering–Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT) panel, are designed to identify clinically actionable mutations in tumor tissue [40]. This panel targets 4976 canonical, 104 noncanonical exons, and 33 introns of 341 cancer relevant genes. The mean average of coverage was 700X with an SD of 182, and 97% of all samples used was higher than the average sample (350X). The pipeline (assay and analysis) is a great example of targeted identification of somatic mutations with high levels of accuracy, sensitivity, and reproducibility, feasible in the clinical setting [40]. However, analyses of low-frequency variants in cell-free DNA are much limited by the detection

thresholds of 0.02% and 0.0002% under optimal conditions is possible [41].

The rationale for ctDNA use is the ability to detect minimum residual disease from patients under surveillance noninvasively (TRACERx trial [85]). These clinical trials show that ctDNA detection is possible in over 90% of patients and allow to potentially detect relapse, metastasis, and potentially chemotherapy resistance markers, which could be integrated into a therapeutic application panel. Thus, metastatic monitoring and early identification are two sides of the same coin. Initial efforts using targeted panels and detecting mutations in known metastatic genes showed to be precise (>70%) but not that sensitive considering that about 75% of all patients had detectable ctDNA [35–40] (see Note 4). The most established methodologies are described in several reviews for WES and the slightly more sensitive sWGS (shallow whole genome sequencing) [31, 32, 42, 48]. The most remarkable studies include TRACERx, CancerSEEK, TEC-seq, and CAPP-seq with the focus of quantifying genomic alterations with mPCR enrichment (TRACERx and CancerSEEK) or hybridization enrichment (CAPP-seq and TEC-seq). Interestingly, the correlation between their outcomes can be expressed as a ratio between tumor burden and ctDNA mutant allele frequency (MAF). Although NGS platforms can detect plasma MAF values under 0.1% to 0.01%, panels similar to MSK-IMPACT solid tumor MAF detection limits are around 2% for hotspot mutations, which is insufficient for analysis of low-frequency variants. Overall, ctDNA using NGS approaches showcases MAFs between 0.01 and 9.3% (median 0.31%). Thus, there is a great enthusiasm of implementing ctDNA diagnostics to identify MRDs in order to improve patient stratification, for tumors in stages T3 to T1b (Classification based on tumor volume, diameter and stage) [31].

A longitudinal study regarding ctDNA tracking upon neoadjuvant therapy evaluated the reduction of ctDNA levels on treatment as a successful recovery measure [47]. Moreover, they demonstrated that a careful design of the targeted panel can increase the detection threshold 100 times fold. This could enable individualized clinical management of patients with cancer treated with curative intent. Another remarkable study done by Scherer et al. [49] illustrates NGS profiling through CAPP-seq analysis of B-cell lymphoma from 92 patients and 24 healthy subjects, showing that ctDNA correlates with clinical indices and allows to track multiple somatic mutations outperforming immunoglobulin sequencing and radiographic imaging for detection of minimal residual disease. Moreover, ctDNA allowed the authors to identify patterns of clonal evolution distinguishing indolent follicular lymphomas from treatment resistan[]t. Finally, they demonstrated that ctDNA analysis reveals biological factors that underlie outcomes and could facilitate individualized therapy by applying non-redundant panels in a clinical setting to identify key signatures associated with resistance or metastasis, thus directing the design of treatment.

Extracellular Vesicles

The discovery of the extracellular vesicles (EVs) was first overlooked during the 1980s, when it was thought that they contained waste products. However, with more studies detailing the molecular contents of EVs, it was clear that there were many kinds of EVs with different purposes, such as microvesicles, ectosomes, apoptotic bodies, and exosomes. In 1987, the word “exosomes” was proposed for EVs of endosomal origin that form during the endosomal sorting complexes required for transport (ESCRT) pathway. The existence of this unusual type of EV was confirmed later in antigen-presenting cells such as epithelial cells and tumor cells. Further studies of exosomal contents revealed enrichment for miRNAs and other rare RNA biotypes, sparking interest.

Exosomes (30–150 nm) were indeed a revolutionary contribution to cellular biology. This class of endocytic origin vesicles are secreted by most types of cells and circulate in bodily fluids such as blood, urine, saliva, and breast milk [69]. Their contents have been shown to be broad,

composed of various growth factors, proteins, lipids, and various nucleic acids, microRNAs, long noncoding RNAs, and circular RNAs (circRNAs) [60]. Exosomes start as intraluminal vesicles generated within the endolysosomal system and secreted by the fusion of multivesicular endosomes (MVEs). Their abundance and constant presence in biofluids make them an ideal target to survey in the search for biomarkers, in early and late stages of disease.

Methods for EV Identification and Analysis

The isolation of exosomes for sequencing, proteomics, or lipidomic analyses requires specialized equipment (ultracentrifugation, chromatography, or microfluidics) and validation through MVE-specific markers, to avoid other structures within a similar size range [51, 52] (see Note 1). However, compared to ctDNA or CTCs, exosomes are much easier to access [65].

Early detection is the hallmark of cancer therapy characterized by a high heterogeneity in patient response. In ovarian cancer, exosomes have shown to carry proteins such as CD9, CD81, and CD63 used in screening and diagnosis [63]. For example, it was reported positive expression of claudin-4 in exosomes in the blood of 32 of 63 patients, but in only 1 of 50 samples from healthy controls, with 51% sensitivity and 98% specificity, indicating its clinical significance for diagnosis [64] (see Note 5).

The importance of the study of EVs in a clinical setting to complement the diagnosis and prognosis of several diseases has been well demonstrated [51–69]. Nonetheless, it is critical to highlight that the election of the most appropriate technique to be used in the clinic depends on the required outcome, which could be to obtain the highest concentration of EVs or to select one particular type of EVs (i.e., exosomes, microvesicles, or apoptotic bodies) [67, 73]. This is illustrated in the identification of specific profiles of EVs from gingival crevicular fluid, showing that oral EVs in early pregnancy can identify patients at risk of developing gestational diabetes mellitus [67].

In colorectal, gastric, pancreatic, and lung cancer, a growing number of studies have focused on exosomal cargo, and their use in diagnosis, prognosis, and prediction as biomarkers have also been investigated [68, 71]. However, these studies have a low number of patients and do not control for variation between demographic variables like, age, sex, or race.

Successful Studies on EV Identification and Analysis

The outstanding study done by Hoshino A. et al. [72] details how integrin content from exosomes dictates the target for the metastatic event to occur, suggesting that exosomes play a critical role in delivering specific molecular contents to specific cellular targets. Here, the authors demonstrate that exosomes from mouse and human lung-, liver-, and brain-tropic tumor cells fuse preferentially with resident cells at their predicted destination, namely, lung fibroblasts and epithelial cells, liver Kupffer cells, and brain endothelial cells. Next, the tumor-derived exosomes uptaken by organ-specific cells suggest preparing the pre-metastatic niche. Thus, treatment with exosomes from lung-tropic models redirected the metastasis of bone-tropic tumor cells. The proteomic analysis showed that integrin expression patterns, in which the exosomal integrins α6β4 and α6β1 were associated with lung metastasis, while exosomal integrin αvβ5 was linked to liver metastasis. Targeting the integrins α6β4 and αvβ5 decreased exosome uptake, as well as lung and liver metastasis, respectively. This showcases a groundbreaking example for possible metastatic monitoring and the key role of proteins commonly dismissed as therapeutic targets and opens the possibility of a similar role for other exosomal contents such as RNA or lipids.

The computational methods employed in exosomal studies are no different from WES or single cell RNA profiling. However, the molecular size distributions are much different between cel-

lular transcriptomic profiles (mRNA average of 1.4kpb) and exosomal RNAs (200pb average). Thus, the methods to characterize these small molecules in exosomes are limited [65, 74–77] (see note 6). An attempt to understand their biological relevance has been done by grouping the exosomal RNA reads into small read clusters [86, 87]. The analysis of these small read clusters showed to significantly (FDR < 0.05) predict the survival outcome of liver cancer patients [87], exemplifying a viable alternative for quantification of small RNA. These studies present different approaches to treat RNA or proteomic expression as clusters or vectors rather than individual piles up of genes/proteins, adding new information that correlates with biophysical properties.

Future Directions

The implementation and personalization of exosomes from liquid biopsies have been poorly explored. Sample sizes are below optimal but the findings so far are very promising. Current computational methods are robust enough to provide a characterization of the molecular contents of exosomes; however, more appropriate approaches that allow to understand the biological relevance of these molecules are lacking [65].

Accurate methods for identification of individual traits from tissues, plasma, blood, urine, or other noninvasive liquid biopsies that can lead to diagnosis or patient classification for treatment targeting specific tumor types are needed. Here, we have presented studies and methods evidencing how structural variants, copy number alterations, tumor cells, and even extracellular vesicles can be recovered noninvasively, allowing us to identify a specific cancer. These approaches could potentially direct the course of treatment for individual patients and improve our understanding of these diseases. However, challenges around the appropriate methods and limited sample sizes prevent these technologies from being currently in use. Although clinical trials in phase 3 are promising, more in-depth and rigorous studies that take into account demographics, comorbidities, and reproducibility are needed. The recent breakthroughs in liquid biopsies technologies and the era of personalized medicine are indeed having a great impact in extending overall health, quality of life, and survival of patients.

Notes

Next-generation sequencing and precise targeted digital polymerase chain reaction are the gold standard to identify and validate molecular targets for precision medicine therapies. Advancements in microfluidics and other purification methods allow the careful isolation of specific components from liquid biopsies such as CTCs, ctDNAs, and EVs, boosting the signal quality of downstream analysis. Although these approaches remain costly and under development, the possibilities, improvements, and advancement in materials are transforming the landscape of methods into more accessible and more precise tools (i.e., lab in a chip). Today, the studies are identifying several putative biomarkers in a series of cancer types and diseases. However, the current challenge is to overcome inter-patient and cellular heterogeneity. In this chapter, we reviewed the common considerations, approaches, and resources one should implement when utilizing liquid biopsies for the purpose of uncovering the molecular drivers and mechanisms under the umbrella of precision medicine. Throughout this chapter, we emphasize key points that should be accounted in order to both increase precision, sensitivity and facilitate the quality of data produced by NGS, including the following:

1. The isolation of cfDNA, ctDNA, CTCs, and EVs requires appropriate physical and molecular validations (i.e., DNA quality, fragment size, molecular markers). The isolation selection protocols should reduce the stress caused on the fluid to avoid cellular debris from contaminants.
2. Cellular heterogeneity is the main driver of variance in CTC studies and a critical component to consider in any clinical setting.

Additionally, somatic patient mutations accumulate over time and age is an additional factor to consider. Therefore, targeted panels are the most appropriate approach in a clinical setting, unless a robust validation is performed on an independent dataset.

3. The levels of ctDNA can be variable and undetectable in some patients, which does correlate with therapy or disease stage. However, in patients where ctDNA is detectable, it can be used for monitoring residual disease and progression status.
4. Structural and somatic variants can be called from CTCs and ctDNA; however, precision will scale with the amount of samples and validation is imperative.
5. Patient heterogeneity and sample sizes are a key element in the study of exosomes (EVs). The researchers must ensure appropriate numbers between controls and cases, to distinguish disease from patient-specific signatures.
6. We urge the researchers to incorporate omic approaches, in addition to as many resources as possible in downstream analysis with the objective of identifying the underlying biology and molecular mechanisms behind the identified signatures.
7. The list of resources mentioned throughout this chapter includes highly established methods and references that expand into each individual topic but does not include the entire arsenal of possible available computational resources.

References

1. Poulet G, Massias J, Taly V. Liquid biopsy: general concepts. Acta Cytol. 2019;63(6):449–55.

Liquid Biopsy: Current Status and Future Perspectives

2. Mader S, Pantel K. Liquid biopsy: current status and future perspectives. Oncol Res Treat. 2017;40(7–8):404–8.
3. Alix-Panabières C. The future of liquid biopsy. Nature. 2020;579(7800):S9.
4. Fernandes Marques J, Pereira Reis J, Fernandes G, Hespanhol V, Machado JC, Costa JL. Circulating tumor DNA: a step into the future of cancer management. Acta Cytol. 2019;63(6):456–65.
5. Wan JCM, Massie C, Garcia-Corbacho J, Mouliere F, Brenton JD, Caldas C, Pacey S, Baird R, Rosenfeld N. Liquid biopsies come of age: towards implementation of circulating tumour DNA. Nat Rev Cancer. 2017;17(4):223–38.
6. Miyamoto DT, Ting DT, Toner M, Maheswaran S, Haber DA. Single-cell analysis of circulating tumor cells as a window into tumor heterogeneity. Cold Spring Harb Symp Quant Biol. 2016;81:269–74.
7. Sharma S, Zhuang R, Long M, Pavlovic M, Kang Y, Ilyas A, Asghar W. Circulating tumor cell isolation, culture, and downstream molecular analysis. Biotechnol Adv. 2018;36(4):1063–78.
8. Ebright RY, Lee S, Wittner BS, Niederhoffer KL, Nicholson BT, Bardia A, Truesdell S, Wiley DF, Wesley B, Li S, Mai A, Aceto N, Vincent-Jordan N, Szabolcs A, Chirn B, Kreuzer J, Comaills V, Kalinich M, Haas W, Ting DT, Toner M, Vasudevan S, Haber DA, Maheswaran S, Micalizzi DS. Deregulation of ribosomal protein expression and translation promotes breast cancer metastasis. Science. 2020;367(6485):1468–73.
9. Ma N, Jeffrey SS. Deciphering cancer clues from blood. Science. 2020;367(6485):1424–5.
10. Keller L, Pantel K. Unravelling tumour heterogeneity by single-cell profiling of circulating tumour cells. Nat Rev Cancer. 2019;19(10):553–67.
11. Court CM, Hou S, Liu L, Winograd P, DiPardo BJ, Liu SX, Chen PJ, Zhu Y, Smalley M, Zhang R, Sadeghi S, Finn RS, Kaldas FM, Busuttil RW, Zhou XJ, Tseng HR, Tomlinson JS, Graeber TG, Agopian VG. Somatic copy number profiling from hepatocellular carcinoma circulating tumor cells. NPJ Precis Oncol. 2020;4:16.
12. Szczerba BM, Castro-Giner F, Vetter M, Krol I, Gkountela S, Landin J, Scheidmann MC, Donato C, Scherrer R, Singer J, Beisel C, Kurzeder C, Heinzelmann-Schwarz V, Rochlitz C, Weber WP, Beerenwinkel N, Aceto N. Neutrophils escort circulating tumour cells to enable cell cycle progression. Nature. 2019;566(7745):553–7.
13. Zhu Z, Qiu S, Shao K, Hou Y. Progress and challenges of sequencing and analyzing circulating tumor cells. Cell Biol Toxicol. 2018;34(5):405–15.
14. Payne K, Brooks J, Spruce R, Batis N, Taylor G, Nankivell P, Mehanna H. Circulating tumour cell biomarkers in head and neck cancer: current progress and future prospects. Cancers (Basel). 2019;11:8.
15. Huang X, Liu S, Wu L, Jiang M, Hou Y. High throughput single cell RNA sequencing, bioinformatics analysis and applications. Adv Exp Med Biol. 2018;1068:33–43.
16. Miyamoto DT, Zheng Y, Wittner BS, Lee RJ, Zhu H, Broderick KT, Desai R, Fox DB, Brannigan

BW, Trautwein J, Arora KS, Desai N, Dahl DM, Sequist LV, Smith MR, Kapur R, Wu CL, Shioda T, Ramaswamy S, Ting DT, Toner M, Maheswaran S, Haber DA. RNA-Seq of single prostate CTCs implicates noncanonical Wnt signaling in antiandrogen resistance. Science. 2015;349(6254):1351–6.
17. Court CM, Ankeny JS, Sho S, Hou S, Li Q, Hsieh C, Song M, Liao X, Rochefort MM, Wainberg ZA, Graeber TG, Tseng HR, Tomlinson JS. Reality of single circulating tumor cell sequencing for molecular diagnostics in pancreatic cancer. J Mol Diagn. 2016;18(5):688–96.
18. de Bono JS, Scher HI, Montgomery RB, Parker C, Miller MC, Tissing H, Doyle GV, Terstappen LW, Pienta KJ, Raghavan D. Circulating tumor cells predict survival benefit from treatment in metastatic castration-resistant prostate cancer. Clin Cancer Res. 2008;14(19):6302–9.
19. Thiele JA, Pitule P, Hicks J, Kuhn P. Single-cell analysis of circulating tumor cells. Methods Mol Biol. 2019;1908:243–64.
20. Krol I, Castro-Giner F, Maurer M, Gkountela S, Szczerba BM, Scherrer R, Coleman N, Carreira S, Bachmann F, Anderson S, Engelhardt M, Lane H, Evans TRJ, Plummer R, Kristeleit R, Lopez J, Aceto N. Detection of circulating tumour cell clusters in human glioblastoma. Br J Cancer. 2018;119(4):487–91.
21. Wei C, Yang C, Wang S, Shi D, Zhang C, Lin X, Liu Q, Dou R, Xiong B. Crosstalk between cancer cells and tumor associated macrophages is required for mesenchymal circulating tumor cell-mediated colorectal cancer metastasis. Mol Cancer. 2019;18(1):64.
22. Tsao SC, Wang J, Wang Y, Behren A, Cebon J, Trau M. Characterising the phenotypic evolution of circulating tumour cells during treatment. Nat Commun. 2018;9(1):1482.
23. Chen M, Zhao H. Next-generation sequencing in liquid biopsy: cancer screening and early detection. Hum Genomics. 2019;13(1):34.
24. Liu SX, Gustafson HH, Jackson DL, Pun SH, Trapnell C. Trajectory analysis quantifies transcriptional plasticity during macrophage polarization. Sci Rep. 2020;10(1):12273.
25. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, Hao Y, Stoeckius M, Smibert P, Satija R. Comprehensive integration of single-cell data. Cell. 2019;177(7):1888–1902.e21.
26. Tyler SR, Rotti PG, Sun X, Yi Y, Xie W, Winter MC, Flamme-Wiese MJ, Tucker BA, Mullins RF, Norris AW, Engelhardt JF. PyMINEr finds gene and Autocrine-paracrine networks from human Islet scRNA-Seq. Cell Rep. 2019;26(7):1951–1964.e8.
27. Cheng YH, Chen YC, Lin E, Brien R, Jung S, Chen YT, Lee W, Hao Z, Sahoo S, Min Kang H, Cong J, Burness M, Nagrath S, Wicha S, M, Yoon E. Hydro-Seq enables contamination-free high-throughput single-cell RNA-sequencing for circulating tumor cells. Nat Commun. 2019;10(1):2163.

Methods in Trajectory Analysis

28. Saelens W, Cannoodt R, Todorov H, Saeys Y. A comparison of single-cell trajectory inference methods. Nat Biotechnol. 2019;37(5):547–54.
29. Huang L, Ma F, Chapman A, Lu S, Xie XS. Single-cell whole-genome amplification and sequencing: methodology and applications. Annu Rev Genomics Hum Genet. 2015;16:79–102.
30. Palmirotta R, Lovero D, Silvestris E, Felici C, Quaresmini D, Cafforio P, Silvestris F. Next-generation Sequencing (NGS) analysis on single circulating tumor cells (CTCs) with no need of Whole-genome Amplification (WGA). Cancer Genomics Proteomics. 2017;14(3):173–9.
31. Abbosh C, Birkbak NJ, Swanton C. Early stage NSCLC - challenges to implementing ctDNA-based screening and MRD detection. Nat Rev Clin Oncol. 2018;15(9):577–86.
32. Valpione S, Campana L. Detection of circulating tumor DNA (ctDNA) by digital droplet polymerase chain reaction (dd-PCR) in liquid biopsies. Methods Enzymol. 2019;629:1–15.

Detection of Circulating Tumor DNA in Early- and Late-Stage Human Malignancies

33. Bettegowda C, Sausen M, Leary RJ, Kinde I, Wang Y, Agrawal N, Bartlett BR, Wang H, Luber B, Alani RM, Antonarakis ES, Azad NS, Bardelli A, Brem H, Cameron JL, Lee CC, Fecher LA, Gallia GL, Gibbs P, Le D, Giuntoli RL, Goggins M, Hogarty MD, Holdhoff M, Hong SM, Jiao Y, Juhl HH, Kim JJ, Siravegna G, Laheru DA, Lauricella C, Lim M, Lipson EJ, Marie SK, Netto GJ, Oliner KS, Olivi A, Olsson L, Riggins GJ, Sartore-Bianchi A, Schmidt K, Shih lM, Oba-Shinjo, SM, Siena S, Theodorescu D, Tie J, Harkins TT, Veronese S, Wang TL, Weingart JD, Wolfgang CL, Wood LD, Xing D, Hruban RH, Wu J, Allen PJ, Schmidt CM, Choti MA, Velculescu VE, Kinzler KW, Vogelstein B, Papadopoulos N, Diaz LA. Detection of circulating tumor DNA in early- and late-stage human malignancies. Sci Transl Med. 2014;6:224:224ra24.
34. Hrebien S, Citi V, Garcia-Murillas I, Cutts R, Fenwick K, Kozarewa I, McEwen R, Ratnayake J, Maudsley R, Carr TH, de Bruin EC, Schiavon G, Oliveira M, Turner N. Early ctDNA dynamics as a surrogate for progression-free survival in advanced breast cancer in the BEECH trial. Ann Oncol. 2019;30(6):945–52.
35. Wan JCM, Heider K, Gale D, Murphy S, Fisher E, Mouliere F, Ruiz-Valdepenas A, Santonja A, Morris J, Chandrananda D, Marshall A, Gill AB, Chan PY, Barker E, Young G, Cooper WN, Hudecova I, Marass F, Mair R, Brindle KM, Stewart GD, Abraham JE, Caldas C, Rassl DM, Rintoul RC, Alifrangis C, Middleton MR, Gallagher FA, Parkinson C, Durrani A, McDermott U, Smith CG, Massie C, Corrie PG,

Rosenfeld N. ctDNA monitoring using patient-specific sequencing and integration of variant reads. Sci Transl Med. 2020;12:548.

Direct Detection of Early-Stage Cancers Using Circulating Tumor DNA

36. Phallen J, Sausen M, Adleff V, Leal A, Hruban C, White J, Anagnostou V, Fiksel J, Cristiano S, Papp E, Speir S, Reinert T, Orntoft MW, Woodward BD, Murphy D, Parpart-Li S, Riley D, Nesselbush M, Sengamalay N, Georgiadis A, Li QK, Madsen MR, Mortensen FV, Huiskens J, Punt C, van Grieken N, Fijneman R, Meijer G, Husain H, Scharpf RB, Diaz LA, Jones S, Angiuoli S, Ørntoft T, Nielsen HJ, Andersen CL, Velculescu VE. Direct detection of early-stage cancers using circulating tumor DNA. Sci Transl Med. 2017;9:403.
37. Rothwell DG, Ayub M, Cook N, Thistlethwaite F, Carter L, Dean E, Smith N, Villa S, Dransfield J, Clipson A, White D, Nessa K, Ferdous S, Howell M, Gupta A, Kilerci B, Mohan S, Frese K, Gulati S, Miller C, Jordan A, Eaton H, Hickson N, O'Brien C, Graham D, Kelly C, Aruketty S, Metcalf R, Chiramel J, Tinsley N, Vickers AJ, Kurup R, Frost H, Stevenson J, Southam S, Landers D, Wallace A, Marais R, Hughes AM, Brady G, Dive C, Krebs MG. Utility of ctDNA to support patient selection for early phase clinical trials: the TARGET study. Nat Med. 2019;25(5):738–43.

Enhanced Detection of Circulating Tumor DNA by Fragment Size Analysis

38. Chin RI, Chen K, Usmani A, Chua C, Harris PK, Binkley MS, Azad TD, Dudley JC, Chaudhuri AA. Detection of solid tumor molecular residual disease (MRD) using circulating tumor DNA (ctDNA). Mol Diagn Ther. 2019;23(3):311–31.
39. Mouliere F, Chandrananda D, Piskorz AM, Moore EK, Morris J, Ahlborn LB, Mair R, Goranova T, Marass F, Heider K, Wan JCM, Supernat A, Hudecova I, Gounaris I, Ros S, Jimenez-Linan M, Garcia-Corbacho J, Patel K, Østrup O, Murphy S, Eldridge MD, Gale D, Stewart GD, Burge J, Cooper WN, van der Heijden MS, Massie CE, Watts C, Corrie P, Pacey S, Brindle KM, Baird RD, Mau-Sørensen M, Parkinson CA, Smith CG, Brenton JD, Rosenfeld N. Enhanced detection of circulating tumor DNA by fragment size analysis. Sci Transl Med. 2018;10:466.
40. Cheng DT, Mitchell TN, Zehir A, Shah RH, Benayed R, Syed A, Chandramohan R, Liu ZY, Won HH, Scott SN, Brannon AR, O'Reilly C, Sadowska J, Casanova J, Yannes A, Hechtman JF, Yao J, Song W, Ross DS, Oultache A, Dogan S, Borsu L, Hameed M, Nafa K, Arcila ME, Ladanyi M, Berger MF. Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): a hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. J Mol Diagn. 2015;17(3):251–64.
41. Chae YK, Oh MS. Detection of minimal residual disease using ctDNA in lung cancer: current evidence and future directions. J Thorac Oncol. 2019;14(1):16–24.
42. Rossi D, Condoluci A, Spina V, Gaidano G. Methods for measuring ctDNA in lymphomas. Methods Mol Biol. 2019;1881:253–65.
43. Oellerich M, Schütz E, Beck J, Walson PD. Circulating cell-free DNA-diagnostic and prognostic applications in personalized cancer therapy. Ther Drug Monit. 2019;41(2):115–20.
44. Oellerich M, Schütz E, Beck J, Kanzow P, Plowman PN, Weiss GJ, Walson PD. Using circulating cell-free DNA to monitor personalized cancer therapy. Crit Rev Clin Lab Sci. 2017;54(3):205–18.
45. Gorgannezhad L, Umer M, Islam MN, Nguyen NT, Shiddiky MJA. Circulating tumor DNA and liquid biopsy: opportunities, challenges, and recent advances in detection technologies. Lab Chip. 2018;18(8):1174–96.
46. Garcia-Murillas I, Schiavon G, Weigelt B, Ng C, Hrebien S, Cutts RJ, Cheang M, Osin P, Nerurkar A, Kozarewa I, Garrido JA, Dowsett M, Reis-Filho JS, Smith IE, Turner NC. Mutation tracking in circulating tumor DNA predicts relapse in early breast cancer. Sci Transl Med. 2015;7:302:302ra133.
47. McDonald BR, Contente-Cuomo T, Sammut SJ, Odenheimer-Bergman A, Ernst B, Perdigones N, Chin SF, Farooq M, Mejia R, Cronin PA, Anderson KS, Kosiorek HE, Northfelt DW, McCullough AE, Patel BK, Weitzel JN, Slavin TP, Caldas C, Pockaj BA, Murtaza M. Personalized circulating tumor DNA analysis to detect residual disease after neoadjuvant therapy in breast cancer. Sci Transl Med. 2019;11:504.
48. Barlebo Ahlborn L, Østrup O. Toward liquid biopsies in cancer treatment: application of circulating tumor DNA. APMIS. 2019;127(5):329–36.
49. Scherer F, Kurtz DM, Newman AM, Stehr H, Craig AF, Esfahani MS, Lovejoy AF, Chabon JJ, Klass DM, Liu CL, Zhou L, Glover C, Visser BC, Poultsides GA, Advani RH, Maeda LS, Gupta NK, Levy R, Ohgami RS, Kunder CA, Diehn M, Alizadeh AA. Distinct biological subtypes and patterns of genome evolution in lymphoma revealed by circulating tumor DNA. Sci Transl Med. 2016;8:364:364ra155.
50. Huang C, Liu S, Tong X, Fan H. Extracellular vesicles and ctDNA in lung cancer: biomarker sources and therapeutic applications. Cancer Chemother Pharmacol. 2018;82(2):171–83.
51. Tang YT, Huang YY, Zheng L, Qin SH, Xu XP, An TX, Xu Y, Wu YS, Hu XM, Ping BH, Wang Q. Comparison of isolation methods of exosomes and exosomal RNA from cell culture medium and serum. Int J Mol Med. 2017;40(3):834–44.
52. De Gregorio C, Díaz P, López-Leal R, Manque P, Court FA. Purification of exosomes from primary Schwann cells, RNA extraction, and next-generation sequencing of Exosomal RNAs. Methods Mol Biol. 2018;1739:299–315.

Extracellular Vesicles from Thyroid Carcinoma: The New Frontier of Liquid Biopsy

53. Rappa G, Puglisi C, Santos MF, Forte S, Memeo L, Lorico A. Extracellular vesicles from thyroid carcinoma: the new frontier of liquid biopsy. Int J Mol Sci. 2019;20:5.
54. Rappa G, Puglisi C, Santos MF, Forte S, Memeo L, Lorico A. Extracellular vesicles from thyroid carcinoma: the new frontier of liquid biopsy. Int J Mol Sci. 2019;20:5.
55. Yoshioka Y, Kosaka N, Konishi Y, Ohta H, Okamoto H, Sonoda H, Nonaka R, Yamamoto H, Ishii H, Mori M, Furuta K, Nakajima T, Hayashi H, Sugisaki H, Higashimoto H, Kato T, Takeshita F, Ochiya T. Ultrasensitive liquid biopsy of circulating extracellular vesicles using ExoScreen. Nat Commun. 2014;5:3591.
56. Nazarenko I. Extracellular vesicles: recent developments in technology and perspectives for cancer liquid biopsy. Recent Results Cancer Res. 2020;215:319–44.
57. Shankar GM, Balaj L, Stott SL, Nahed B, Carter BS. Liquid biopsy for brain tumors. Expert Rev Mol Diagn. 2017;17(10):943–7.
58. Klekner Á, Szivos L, Virga J, Árkosy P, Bognár L, Birkó Z, Nagy B. Significance of liquid biopsy in glioblastoma - a review. J Biotechnol. 2019;298:82–7.
59. Ludwig N, Whiteside TL, Reichert TE. Challenges in exosome isolation and analysis in health and disease. Int J Mol Sci. 2019;20:19.
60. Wang Y, Liu J, Ma J, Sun T, Zhou Q, Wang W, Wang G, Wu P, Wang H, Jiang L, Yuan W, Sun Z, Ming L. Exosomal circRNAs: biogenesis, effect and application in human diseases. Mol Cancer. 2019;18(1):116.
61. Wang HX, Gires O. Tumor-derived extracellular vesicles in breast cancer: from bench to bedside. Cancer Lett. 2019;460:54–64.
62. Lucchetti D, Fattorossi A, Sgambato A. Extracellular vesicles in oncology: progress and pitfalls in the methods of isolation and analysis. Biotechnol J. 2019;14(1):e1700716.
63. Eguchi A, Kostallari E, Feldstein AE, Shah VH. Extracellular vesicles, the liquid biopsy of the future. J Hepatol. 2019;70(6):1292–4.
64. Chang L, Ni J, Zhu Y, Pang B, Graham P, Zhang H, Li Y. Liquid biopsy in ovarian cancer: recent advances in circulating extracellular vesicle detection for early diagnosis and monitoring progression. Theranostics. 2019;9(14):4130–40.

Unannotated Small RNA Clusters in Circulating Extracellular Vesicles Detect Early Stage Liver Cancer

65. von Felden J, Garcia-Lezana T, Dogra N, Kozlova E, Ahsen ME, Craig AJ, Gifford S, Wunsch B, Smith JT, Kim S, Diaz JEL, Chen X, Labgaa I, Haber PK, Olsen R, Han D, Restrepo P, D'Avola D, Hernandez-Meza G, Allette K, Sebra R, Saberi B, Tabrizian P, Asgharpour A, Dieterich D, Llovet JM, Cordon-Cardo C, Tewari A, Schwartz M, Stolovitzky G, Losic B, Villanueva A. Unannotated small RNA clusters in circulating extracellular vesicles detect early stage liver cancer. bioRxiv. 2020.04.29.066183 https://doi.org/10.1101/2020.04.29.066183.
66. Lu J, Pang J, Chen Y, Dong Q, Sheng J, Luo Y, Lu Y, Lin B, Liu T. Application of microfluidic chips in separation and analysis of extracellular vesicles in liquid biopsy for cancer. Micromachines (Basel). 2019;10:6.
67. Monteiro LJ, Varas-Godoy M, Monckeberg M, Realini O, Hernández M, Rice G, Romero R, Saavedra JF, Illanes SE, Chaparro A. Oral extracellular vesicles in early pregnancy can identify patients at risk of developing gestational diabetes mellitus. PLoS One. 2019;14(6):e0218616.
68. Cui S, Cheng Z, Qin W, Jiang L. Exosomes as a liquid biopsy for lung cancer. Lung Cancer. 2018;116:46–54.
69. Braicu C, Tomuleasa C, Monroig P, Cucuianu A, Berindan-Neagoe I, Calin GA. Exosomes as divine messengers: are they the Hermes of modern molecular oncology? Cell Death Differ. 2015;22(1):34–45.
70. Zhao C, Gao F, Weng S, Liu Q. Pancreatic cancer and associated exosomes. Cancer Biomark. 2017;20(4):357–67.
71. Wu X, Zhao H, Natalia A, Lim CZJ, Ho NRY, Ong CJ, Teo MCC, So JBY, Shao H. Exosome-templated nanoplasmonics for multiparametric molecular profiling. Sci Adv. 2020;6:19:eaba2556.
72. Hoshino A, Costa-Silva B, Shen TL, Rodrigues G, Hashimoto A, Tesic Mark M, Molina H, Kohsaka S, Di Giannatale A, Ceder S, Singh S, Williams C, Soplop N, Uryu K, Pharmer L, King T, Bojmar L, Davies AE, Ararso Y, Zhang T, Zhang H, Hernandez J, Weiss JM, Dumont-Cole VD, Kramer K, Wexler LH, Narendran A, Schwartz GK, Healey JH, Sandstrom P, Labori KJ, Kure EH, Grandgenett PM, Hollingsworth MA, de Sousa M, Kaur S, Jain M, Mallya K, Batra SK, Jarnagin WR, Brady MS, Fodstad O, Muller V, Pantel K, Minn AJ, Bissell MJ, Garcia BA, Kang Y, Rajasekhar VK, Ghajar CM, Matei I, Peinado H, Bromberg J, Lyden D. Tumour exosome integrins determine organotropic metastasis. Nature. 2015;527(7578):329–35.
73. Dai J, Su Y, Zhong S, Cong L, Liu B, Yang J, Tao Y, He Z, Chen C, Jiang Y. Exosomes: key players in cancer and potential therapeutic strategy. Signal Transduct Target Ther. 2020;5(1):145.
74. Li M, Xi N, Wang YC, Liu LQ. Atomic force microscopy for revealing micro/nanoscale mechanics in tumor metastasis: from single cells to microenvironmental cues. Acta Pharmacol Sin. 2021;42(3):323–39.
75. Goodarzi H, Nguyen HCB, Zhang S, Dill BD, Molina H, Tavazoie SF. Modulated expression of specific tRNAs drives gene expression and cancer progression. Cell. 2016;165(6):1416–27.

Profiling Surface Proteins on Individual Exosomes Using a Proximity Barcoding Assay

76. Wu D, Yan J, Shen X, Sun Y, Thulin M, Cai Y, Wik L, Shen Q, Oelrich J, Qian X, Dubois KL, Ronquist KG, Nilsson M, Landegren U, Kamali-Moghaddam M. Profiling surface proteins on individual exosomes using a proximity barcoding assay. Nat Commun. 2019;10(1):3854.

Computational Tools

77. Rozowsky J, Kitchen RR, Park JJ, Galeev TR, Diao J, Warrell J, Thistlethwaite W, Subramanian SL, Milosavljevic A, Gerstein M. exceRpt: a comprehensive analytic platform for extracellular RNA profiling. Cell Syst. 2019;8(4):352–357.e3.
78. Mandel P, Metais P. Nucleic acids in human blood plasma. C R Seances Soc Biol Fil. 1948;142(3–4):241–3.
79. Anthony M. Bolger, Marc Lohse, Bjoern Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.
80. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J, [Sl]. 2011;17(1):10–2. ISSN 2226-6089
81. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25(9):1105–11.
82. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21.
83. Zheng GXY, Terry JM, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8:1–12.
84. CTC blood filtering system. https://pubmed.ncbi.nlm.nih.gov/22825490/
85. tracerX. https://pubmed.ncbi.nlm.nih.gov/29656895/
86. Exosomes in prostate cancer. https://doi.org/10.1101/2020.09.28.20190009
87. Exosomes read clusters. https://www.biorxiv.org/content/10.1101/2020.04.29.066183v1

Artificial Intelligence for Precision Oncology

14

Sherry Bhalla and Alessandro Laganà

Abstract

Precision oncology is an innovative approach to cancer care in which diagnosis, prognosis, and treatment are informed by the individual patient's genetic and molecular profile. The rapid development of novel high-throughput omics technologies in recent years has led to the generation of massive amount of complex patient data, which in turn has prompted the development of novel computational infrastructures, platforms, and tools to store, retrieve, and analyze this data efficiently. Artificial intelligence (AI), and in particular its subfield of machine learning, is ideal for deciphering patterns in large datasets and offers unique opportunities for advancing precision oncology. In this chapter, we provide an overview of the various public data resources and applications of AI in precision oncology and cancer research, from subtype identification to drug prioritization, using multi-omics datasets. We also discuss the impact of AI-powered medical image analysis in oncology and present the first diagnostic FDA-approved AI-powered tools.

Introduction

The term "precision oncology" describes the genetic and molecular profiling of tumors to determine actionable alterations, which is increasingly being incorporated into mainstream clinical practices. Precision oncology includes a range of strategies such as the use of biomarkers to discern specific tumor subtypes and ascertain reoccurrence, creation of mouse models for testing drugs, genome sequencing, and omics analysis to identify targetable mutations and guide therapy [1]. Since the emergence of imatinib as the first targeted therapy for chronic myeloid leukemia almost 20 years ago, the field of personalized cancer medicine has taken off to achieve greater heights [2]. Novel high-throughput sequencing approaches and the rise of big data in oncology have led to the development of many predictive models based on artificial intelligence (AI), and particularly machine learning (ML) techniques for the discovery of early diagnosis, prognosis, and therapeutic biomarkers in different cancer types. With precision oncology pro-

S. Bhalla
Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA

A. Laganà (✉)
Department of Genetics and Genomic Sciences, Department of Oncological Sciences, Mount Sinai Icahn School of Medicine, New York, NY, USA
e-mail: alessandro.lagana@mssm.edu

© Springer Nature Switzerland AG 2022

A. Laganà (ed.), *Computational Methods for Precision Oncology*, Advances in Experimental Medicine and Biology 1361, https://doi.org/10.1007/978-3-030-91836-1_14

gressively taking center stage, the need to develop predictive models in oncology has increased many folds. The availability of big public data resources like The Cancer Genome Atlas (TCGA), the International Cancer Genome Consortium (ICGC), and the Cancer Cell Line Encyclopedia (CCLE) has fueled the development of the various prediction models that sustain precision oncology. Among the challenges in big data analytics are the proper maintenance and sharing of patients' data in a responsible manner. The FAIR guiding principles of Findability, Accessibility, Interoperability, and Reusability published in 2016 are guiding the data producers and publishers to overcome the hurdles in data management and sharing (https://www.go-fair.org/fair-principles/).

Clinical oncology practice is now starting to reap the benefits of ML, data science (DS), and big data analytics by leveraging models that predict survival in specific cancer subtypes and response to specific therapies created by using single datatypes like genomics, proteomics, metabolomics, imaging and clinical notes, or combinations of these. There has been tremendous development in the literature regarding the role of AI in the early prediction, diagnosis, and prognosis in different cancer types. But the field of personalized oncology is still in its infancy. Tumor heterogeneity and clonality, patient genetic makeup, and body response to drug treatment regimens, collectively make it a complex problem. The development of new techniques, namely, single-cell genomics, spatial transcriptomics, and high-throughput gene expression perturbation assays have gained traction in recent years. With the development of computer hardware, the neural networks envisaged decades ago are now realized on the ground and have paved the way to handle high-dimensional datasets. By using AI, automated, scalable pipelines have been developed to identify cancer types and subtypes and predict prognosis with performance comparable to physicians. In particular, deep learning models based on pathology images have propelled the progress in the field of precision oncology.

In this chapter, we focus on describing the major public data resources that are gold mines for developing precision medicine models. We describe developments in cancer subtype identification analysis and the models that have been learned and leveraged to predict drug sensitivity on cell lines, which can be a very important guide to develop subtype-specific treatment in cancer. Next, we introduce machine learning models based on pathology images, which predict important outcomes in oncology, and describe the predictive models that have gained FDA approval. Finally, we discuss the current challenges in realizing the full potential of machine learning models in clinical practice.

Public Data Resources Powering Precision Oncology

Precision oncology is a data-driven approach to personalized cancer care. With the development of advanced sequencing techniques, data points are available for every step of the central dogma, ranging from genomics and transcriptomics to proteomics and metabolomics. At the genome level, data are generated from whole-genome sequencing (WGS) and whole-exome sequencing (WES), which include single nucleotide polymorphism (SNP), copy number variation (CNV), and structural variation (SV) data. At the transcriptomic level, sequencing techniques provide expression measurements for coding and non-coding genes [3]. The proteomics data generated using protein array and mass spectrometry measure the expression of actionable protein molecules [4]. With the role of epigenetics becoming more well defined in cancer, data illustrating epigenetic regulation and post-translational modifications (i.e., miRNA methylation and other chromosomal modifications) have become pivotal for prognostic models in oncology [5]. In parallel, there has been development in non-omics datasets, e.g., imaging (pathology and medical imaging). Non-omics data, including electronic health records, radiographic, and histologic data, have been widely used to develop cancer diagnostic and prognostic algorithms. These data

have helped to gain a deeper understanding of cancer development and metastasis and have led to personalized and efficient oncologic care.

As cancer is a disease of the genome, publicly accessible genomic resources can undoubtedly promote precision medicine-based scientific discovery (Table 14.1). The National Cancer Institute (NCI) has provided an array of repositories that house omics and non-omics data related to cancer. One of the well-curated data-sharing platforms for cancer supported by NCI is Genomics Data Common (GDC), which hosts genomics data collected from nearly 65 projects and 80,000 patients. The GDC covers the two most extensive data resources of oncology, incorporating The Cancer Genome Atlas (TCGA, https://cancergenome.nih.gov/) and Therapeutically Applicable Research to Generate Effective Therapies (TARGET, https://ocg.cancer.gov/programs/target). GDC aims to provide uniformly processed cancer data to support the development of precision medicine in oncology; it contains clinical, biospecimen, and molecular data for several cancer types. The molecular data in this repository include WGS, WES, transcriptome sequencing (RNA-seq), microRNA sequencing, DNA methylation analysis, and DNA copy number analysis.

The uniform processing of data types in GDC allows the user to combine different datasets for analysis or test machine learning models of different datasets [6]. NCI also maintains a proteomics data repository called Proteomic Data Commons (PDC), comprising raw and processed mass spectrometry data from cancer proteomics experiments. Most of the datasets have corresponding genomic and imaging data available in other Cancer Research Data Commons [7, 8]. PDC contains datasets from three major repositories: NCI's Clinical Proteomic Tumor Analysis Consortium (CPTAC), Children's Brain Tumor Tissue Consortium (CBTTC), and International Cancer Proteogenomic Consortium (ICPC) ([9, 10], https://icpc.cancer.gov/portal/). Currently, around 27 TB of proteomics data is available for 11 major body sites. The major objective of this repository is to make proteomics data available to the research community and support the integration of proteomics data with genomics data to promote precision medicine. Further, the non-omics data for cancer are also archived and stored in various repositories. The NCI hosts The Cancer

Table 14.1 Major data portals that aid the precision medicine in oncology

Name	Portal	Data type
GDC	https://portal.gdc.cancer.gov	mRNA expression, miRNA expression, mutation, CNV, methylation
PDC	https://proteomic.datacommons.cancer.gov	Proteomics expression, spectral data
TCIA	https://www.cancerimagingarchive.net	Medical images
NCI60	https://dtp.cancer.gov/discovery_development/nci-60/	mRNA expression, miRNA expression, protein expression, metabolomics data, methylation data, enzyme activity
GDSC	https://www.cancerrxgene.org	Drug sensitivity, mRNA expression, mutation, CNV, methylaion
CCLE	https://portals.broadinstitute.org/ccle	Drug sensitivity, genomics transcriptomics, protein array, DNA methylation
L1000	https://clue.io	Perturbed gene expression data
cMAP	http://clue.io/cmap	Perturbed gene expression data
ALMANAC	https://dtp.cancer.gov/ncialmanac	Drug pair sensitivity, mRNA expression, miRNA expression, protein expression, metabolomics data, methylation data, enzyme activity
COSMIC	https://cancer.sanger.ac.uk/cosmic	Mutations, CNV, structural variants
CCLE	https://portals.broadinstitute.org/ccle	Drug sensitivity, genomics transcriptomics, protein array, DNA methylation
Depmap	https://depmap.org	Gene expression from CCLE, gene dependency data in form of CRISPR knockout screens

Imaging Archive (TCIA), which archives a large number of publicly accessible medical images from 25 types of cancer for nearly 30,000 patients. Along with image data, the associated clinical data such as patient outcomes, treatment details, genomics, pathology, and expert analyses are also provided when available [7].

Repositories comprising of pharmacogenomic screens of cancer cell lines have emerged as an appealing pre-clinical system for identifying tumor genetic subtypes with selective sensitivity to targeted therapeutic drugs [11]. One of the oldest projects is NCI60, started in the late 1980s with the aim of creating an in vitro drug discovery tool. With time, NCI 60 has become a rich source of information about the mechanisms of growth inhibition [12]. Another project called Cancer Cell Line Encyclopedia (CCLE) has been developed to store comprehensive genetic characterization of a large panel of cancer cell lines. The CCLE provides public access to DNA copy number, mRNA expression, mutation data, and more, for 1000 cancer cell lines [13]. The Genomics of Drug Sensitivity in Cancer (GDSC) is another public resource that contains drug sensitivity data for nearly 138 anticancer drugs across almost 800 cancer cell lines. GDSC is integrated with the Catalogue of Somatic Mutations in Cancer (COSMIC), an essential and comprehensive database of somatic mutations in cancer and represents an important resource for the identification of molecular markers of drug response [14].

Along with the drug sensitivity prediction resources, some repositories store the response to combination of drugs. One such repository is NCI-ALMANAC (A Large Matrix of Anti-Neoplastic Agent Combinations), containing therapeutic activity of over 5000 pairs of FDA-approved cancer drugs against a panel of cell lines in NCI-60 [15].

While the abovementioned drug sensitivity projects store data before treatment with cell line, the Connectivity Map (CMap) and the Library of Integrated Network-Based Cellular Signatures (LINCS) projects are repositories that store data on the transcriptional responses of cancer cell lines after treatment with small molecules. The CMap is a repository of perturbational datasets containing transcriptomic profiles of dozens of cultivated cell lines treated with thousands of chemical compounds serving as reference databases and has been scaled up to 1000-fold by using a low-cost, high-throughput, and reduced representation expression profiling method called L1000 to contain 1.3 million profiles. This method has been shown to be reproducible, analogous to RNA sequencing, and appropriate for computational inference of the expression levels of 81% of non-measured transcripts [16, 17].

Finally, the Cancer Dependency Map (DepMap) is a rolling project aimed at discovering gene dependencies in many cancer cell lines using CRISPR and shRNA genome-wide screens. DepMap can be used to assess targets for highly selective drugs, predict the efficacy and selectivity of candidate drugs, and identify susceptible cell lines for testing them [18].

AI in Cancer Subtype Identification

Molecular subtyping is a process to identify subgroups of samples that share features within a given cancer type. This workflow involves data preprocessing and an unsupervised clustering approach to identify best clusters, perform subtype characterization with clinical metadata, and calculate supervised classification of new samples. Molecular characterization is an essential component of personalized therapy. Subtype characterization is then fundamental as there is no ground truth and unbiased clustering of samples is assessed via statistical metrics and meaningful correlation with clinical outcomes. Clustering approaches range from simple hierarchical clustering to Non-negative Matrix Factorization (NMF), integrative and consensus clustering. Generally, molecular subtypes are identified based purely on genomic information or integration of omics. These may not always be clinically meaningful as they do not always correlate with patient survival, which is an essential criterion to assess the efficacy of therapy. Recent studies have integrated survival information in the clustering approach in a semi-supervised

fashion to make these workflows more robust and clinically actionable. Further, many studies now perform multi-omics data integration to improve the discovery of clinically and biologically meaningful clusters [19–21]. The clusters obtained with this approach are often enriched for clinical outcomes, e.g., survival, morbidity, and mortality,, and therefore may be helpful for personalized diagnosis, prognosis, and therapy. Upadhyaya et al. [22] have identified subgroup-specific clinical outcomes in the prospective multi-institutional trial of atypical teratoid rhabdoid tumor (ATRT). A Pearson correlation-based distance matrix generated from genome-wide methylation probes was used to perform the clustering. One of the three molecular subgroups identified was associated with metastasis and another one with best overall survival. The challenge of integrating ever increasing multi-omics datasets to identify clinically relevant subgroups remains elusive. Similarity network fusion (SNF) [21] is a recent method to generate multi-omics patient-level networks. These are obtained by generating a network from each data type and then merging those networks into a single one by an iterative non-linear optimization method based on message-passing theory. When tested on five cancer types from TCGA, SNF outperformed single omics-based subtype identification and survival prediction. Multiple flagship papers from TCGA have reported molecular subtypes identified based on multi-omics datasets. For instance, TCGA network performed consensus clustering using five omics datasets from breast cancer samples to identify four subtypes. They performed hierarchical and NMF clustering as well and found associations with clinical parameters [23]. In another study, Curtis et al. [24] performed integrative clustering to identify ten subtypes of breast cancer by integrating CNA and gene expression datasets using discovery and validation cohorts.

Though most of the studies identify subgroups within tissue of origin, Hoadley et al. [25] implemented Cluster-of-Cluster-Assignments (COCA) to cluster samples from 12 cancer types and identify similarities and differences among them. While some clusters had significant overlap to their tissue-of-origin counterparts, others included samples from different cancer types enriched for specific alterations and provided independent clinical associations to predict survival.

Cancer subtyping is not only based on omics data but also on secondary information derived from omics data, e.g., pathways and non-omics datasets such as histopathology images, whole-slide imaging, and medical imaging (radiomics, PET, CT scans, etc.). TCIA-based clinical MRI analysis of gliomas has led to a hybrid technique using radiomics and machine learning to classify molecular subtypes of gliomas. The study indicates that a radiomics-based AI approach can be a reliable alternative to identify glioma subtypes with 80% accuracy [26]. Another example of methods based on secondary omics data is PACL, a pathway-based deep clustering method for molecular subtyping of cancer. PACL captures non-linear associations in high-dimensional data, including transcriptomic data and survival information, and outperforms other clustering methods, providing meaningful biological interpretation of clustering outcomes [27]. Finally, DeePaN is a deep patient graph convolution network that stratifies non-small cell lung cancer (NSCLC) patients into subgroups associated with different outcomes from immuno-oncology therapies [28].

AI-Powered Drug Prioritization in Cancer

Computationally driven drug prioritization takes a patient's genetic makeup into account to decide treatment. Recently in 2017, the FDA approved Pembrolizumab for tumor-site agnostic molecular aberration of mismatch repair deficiency or high microsatellite instability, based on clinical trials in 15 cancer types [29]. Another drug called Larotrectinib was approved to target the tropomyosin receptor kinase gene fusion in multiple cancers [30]. These are the first examples of cancer drugs approved based on specific pan-cancer markers rather than tumor type. Machine learning has been extensively applied in this area to

help clinicians guide treatment, and several prediction methods have been developed to facilitate the drug prioritization process (Table 14.2). Gupta et al. developed a model using genomic features to predict drug response and achieved correlations ranging from 0.43 to 0.78 using the CCLE dataset [31]. Similarly, Dong et al. proposed a classification model based on Support Vector Machines (SVM) to predict drug sensitivity using gene expression profile in the CCLE dataset and attained good performance for several drugs [32]. Additional methods have been developed to predict drug response via multi-omics data integration. Recently, a technique called MERGE (mutation, expression hubs, known regulators, genomic CNV, and methylation) has been developed to model the potential of a gene being a reliable marker for drugs based on the novel MERGE score, a weighted combination of the gene's driver features [34]. MERGE concurrently learns the weights of driver features and the influence of the MERGE score on the observed gene-drug associations. The main innovation of this method is that it combines disease-related, multi-omics prior evidence to rank gene-drug associations. Zhang et al. have developed a heterogeneous network-based method for drug response prediction named HNMDRP to predict cell line-drug associations by leveraging genomic information from cell lines, the chemical structure of the drug as well as drug-target and protein-protein interaction information [33]. Another method called PriorCD (prioritization of candidate drugs) has been developed to prioritize cancer drugs based on a global network and a drug-drug functional similarity network generated by integrating pathway and drug activity profiles using NCI-60 data. This approach applies the unique criteria of interpreting drug functional similarities at the pathway level and has been evaluated on drug datasets of ovarian and breast cancer where it achieved a performance of 0.82 AUROC (Area Under the Receiver Operating Characteristic curve) [35]. The tool BMTMKL uses the state-of-the-art kernelized Bayesian matrix factorization (KBMF) method with component-wise multiple kernel learning for drug response prediction. The method also leverages known pathway information to learn drug response associations. The authors have validated their results in the Fully Blinded Experimental settings using an in-house Acute Myeloid Leukemia (AML) cell line panel. The experimental and predicted drug sensitivity score showed correlation of 0.44 on eight compounds and six cell lines, which increased to 0.70 when the outlier drug Venetoclax was removed [44].

Deep learning methods have become prevalent in many fields and, more recently, deep learning-based neural networks have been developed to perform drug response prediction. Ding et al. explored the use of AutoEncoders to learn important information on the state of tumor cells before treatment. In the study, AutoEncoder models were built to derive compressed features from an input dataset comprising somatic mutations, CNVs, and gene expression data. Further, elastic net classifiers were trained on the compressed features to predict drug sensitivity in cancer cell lines. The encoded features resulted in high sensitivity and specificity but a low AUROC of 0.67 on the external validation dataset of CCLE [43]. Another similar method is Cancer Drug Response profile scan (CDRscan), an ensemble deep learning model of five convolutional networks developed using the CCLP1 and GDSC6 drug response assay datasets. Four convolution networks out of five have dual convergence architecture which means that a series of convolutions are applied on mutation and molecular data separately, the data are merged, then convolution is applied again before predicting the IC50 values for cell line-drug pairs. To make the model more generalized and robust, mean prediction values from five models were reported [37]. Another autoencoder-based method for predicting drug sensitivity is DeepDSC, which was trained and evaluated on CCLE and GDSC. The model attained fairly high R2 scores with a ten-fold cross-validation scheme. Still, performance reduced significantly when tested using the leave the drug out method, showing that the model is not robust to the compounds it has not seen before [42].

Although drug response prediction may help identifying the optimal treatment for some cancer

Table 14.2 Major machine learning and deep learning models implicated to predict drug sensitivity

Method	Model	Validation scheme	Datasets	Input data types	Prediction task	Year	Performance
CancerDP Gupta et al. [31]	SVM	None	CCLE	Gene expression, CNV, mutations	Drug sensitivity	2016	R = 0.43–0.78
Dong et al. [32]	SVM	Tenfold cross-validation	CCLE	Gene expression	Drug sensitivity	2018	Acc: ≥70–≥80%
HNMDRP Zhang et al. [33]	Heterogeneous network	Leave-one-out cross-validation	GDSC	Gene expression, drug chemical structure, drug-target interactions and PPIs	Drug sensitivity	2018	AUROC = 0.91–0.93
MERGE Lee et al. [34]	Probabilistic graphical model	Experimental validation	In-house 30 AML patients, Clinical Trial NCT02551718	Mutations, hubness in a gene expression network, the gene's regulatory role, genomic CNV status, methylation status	Drug sensitivity	2018	–
PriorCD Di et al. [35]	Global network propagation algorithm and a drug–drug functional similarity network	Leave-one-out cross-validation	NCI-60	Gene and microRNA Expression	Drug sensitivity	2019	AUROC = 0.82
H-RACS Yan et al. [36]	Gradient Boosting Regression	Fivefold CV	AstraZeneca, the O'Neil, and DREAM challenge dataset	Chemical descriptors, drug similarity, drug targeting network feature, and gene expression	Drug synergy	2020	AUROC = 0.89
CDRscan Chang et al. [37]	Deep neural network	Fivefold CV	GDSC	Somatic mutations and drug compound fingerprints	Drug sensitivity	2018	AUROC = 0.98
Dr.VAE Rampasek et al. [38]	Deep generative model based on variational autoencoders	100 train-validation-test splits (20 × 5-fold CV)	CTRPv2, NIH LINCS Consortium Cmap	Gene expression, drug-induced transcription change	Drug sensitivity	2019	AUROC = 0.706, AUPRC = 0.718
DeepSynergy Preuer et al. [39]	Deep neural network	Leave-cell-lines-out-fold stratified CV	Merck Compound screen	Chemical descriptors and genomic feature	Drug synergy	2018	AUROC = 0.90, AUPRC = 0.59, TPR = 0.57, TNR = 0.95
Xia et al. [40]	Deep neural network	Fivefold CV	NCI-ALMANAC	Gene expression, proteome, miRNA, drug descriptors	Drug synergy	2018	R2 = 0.94, R = 0.97

(continued)

Table 14.2 (continued)

Method	Model	Validation scheme	Datasets	Input data types	Prediction task	Year	Performance
Chen et al. [41]	Restricted Boltzmann Machine (RBM)	Leave-one-out	astraZeneca-Sanger DREAM challenge		Drug synergy	2018	TPR = 0.602
DeepDSC Li et al. [42]	Stacked AE + DNN	Tenfold CV, leave-one-tissue-out	CCLE & GDSC	Gene expression, molecular fingerprints	Drug sensitivity	2021	UPRC
Ding et al. [43]	Deep autoencoders + elastic nets/SVMs	25-fold CV, external test set	GDSC	Mutation, CNV, gene expression	Drug sensitivity	2018	AUROC = 0.67–0.70, AUPRC = 0.718, TPR = 0.80–0.82
BMTMKL Ammad-Ud-Din et al. [44]	Bayesian matrix factorization (KBMF) with component-wise multiple kernel learning (MKL)	Repetitive fivefold cross-validation, experimental validation	GDSC and CTRP	Gene expression, pathways and gene sets from Molecular Signature database MSigDB	Drug sensitivity	2016	R = 0.44

patients, drug resistance often emerges, often via subclonal tumor cell populations. Combining two or more drugs with different mechanisms of action increases the success rate of drug repositioning [45]. Synergetic combinations of drugs can limit toxicity by reducing drugs' dosage and can help overcome drug resistance by targeting multiple pathways. Trial and error combination design has inadequate application in the clinic due to hazardous exposure to toxic combinations without improving efficacy [45–47]. Therefore, many computational methods have been developed to predict anticancer drug combination synergy based on a variety of genomic, drug structure, and biological data, while limiting toxicity. H-RACS is a machine learning technique that uses signature genes of basal cell lines and drug features, such as chemical descriptors, drug similarities, and drug targeting network features as input to compute a synergy score. Among the seven machine learning models tested, Gradient Boosting Regression gave the maximum AUC and lowest RMSE [36]. Combining two or more drugs with different mechanisms of action is an alternative approach to increase the success rate of drug repositioning. DeepSynergy is a deep neural network developed based on the Merk and Co. pharmacological data and Genomics data from GDSC [39]. The complete dataset consists of 23,062 samples, where every selection entail two compounds and one cell line involving 583 distinct combinations, each tested against 39 human cancer cell lines derived from 7 different tissue types. The network uses the gene expression profile of the cell line and the chemical descriptors of two drugs as input. The information is then propagated through the layers of DeepSynergy until the output unit produces the predicted synergy score. The authors compared their deep learning model with other state-of-the-art machine learning methods like gradient boosting, RFs, SVMs, and elastic net and showed that DeepSynergy performs significantly better. Another study by Xia et al. used the subset of drug pairs in the NCI-ALMANAC database to develop neural networks for encoding features as well as predicting tumor growth, explaining 94% of the response variance. This model takes in gene expression, miRNA expression, protein abundance as well as drug descriptors and fingerprints and returns a scalar prediction score of growth inhibition. However, the model in this study has not been validated in unseen drugs or cell lines [40]. Another innovative prediction method by Chen et al. based on the DREAM 2015 dataset consisting of 4999 drug pairs has been built using a deep belief network to predict drug synergy from gene expression profile, pathway, and the ontological profile of genes derived from the literature [41]. This method was evaluated using the leave-one-out approach and was not tested on a separate validation dataset. Additional recent work by Lee et al. has proposed the SELECT (synthetic lethality and rescue-mediated precision oncology via the transcriptome) workflow for guided patient treatment using transcriptome data. This workflow predicted patients' response to therapy in 80% of cancer clinical trials [48]. Finally, Yuan et al. developed an interpretable machine learning tool, which incorporated a mathematical model of cell dynamics, to identify personalized combination therapy for cancer by using combinatorial perturbations [49].

The Impact of AI-Powered Image Analysis on Precision Oncology

Image analysis has emerged as the preferred and most advanced automated task in personalized oncology thanks to the development of novel neural network-based hardware and algorithms. In the deep neural network domain, convolutional neural network (CNN) has become the best option so far due to its capability to handle image data efficiently compared to other deep neural network (DNN) architectures and conventional machine learning algorithms. Image analysis also leverages transfer learning, where a model is trained on huge image datasets and tested on the problem-specific data to predict a defined cancer by just tuning the model's hyperparameters [50]. Machine learning and deep learning algorithms like CNN and autoencoders have been used in multiple medical image analysis applications

(Table 14.3). Interpretation of cancer histopathology images is one of the most challenging tasks in disease detection. Machine learning and deep learning algorithms can aid dissecting the complex patterns in imaging data and offer accurate and reproducible quantitative radiology assessments. In cancer, there are many tasks where AI can assist the radiologist and help saving time. The first and foremost important task is cancer detection, which involves identifying abnormalities based on the change in intensities in the image sections. The second task consists of the characterization of the tumor and involves tasks like segmentation, diagnosis, and staging, while the third task involves tracking the disease at different time points. To aid the pathologist and increase the accuracy of cancer detection, many algorithms have been designed to segregate normal and cancerous tissue through image analysis. Coudray et al. [75] used deep learning to process hematoxylin and eosin (H&E)-stained histopathology whole-slide images (WSIs) from TCGA to distinguish lung adenocarcinoma (LUAD) vs. lung squamous cell carcinoma (LUSC) vs. normal lung tissue. The results of their method were compared with the assessment of three pathologists, reporting comparable performance with an average AUC of 0.97. The model was based on the Inception-v3 architecture to differentiate normal vs. tumor first, and then to further classify tumors into LUSC or LUAD. The authors further extended the application of their method to predict the genotype of an established panel of genes with AUCs ranging from 0.73 to 0.86. Wang et al. [65] developed a fully automated deep CNN-based pipeline to identify prostate cancer patients from those with prostate benign conditions. Their DCNN model used magnetic resonance (MR) images as input and performed better than non-deep learning methods (mean AUC: 0.84 vs 0.70). Haenssle et al. [78] used the Inception-v4 CNN architecture for diagnostic classification of dermoscopic images alone and in combination with clinical information. They also compared their method with the assessment of 58 dermatologists in different analyses and with the top five algorithms from the ISBI 2016 challenge. Their method not only outperformed the panel of dermatologists based on different performance metrics (AUC: 0.86 vs. 0.79) but was also found to be comparable to the top three ISBI 2016 challenge algorithms. In a quest to correlate H&E-stained histopathology images with proteomics data coming from the CPTAC consortium, Azuaje et al. [79] employed transfer learning using the VGG16-CNN model on clear cell renal cell carcinoma (ccRCC) images. They not only found high correlation between image and proteomics features but also found that image-based models could identify ccRCC samples with 0.95 accuracy on a test dataset. Ribli et al. [80] used a previously published method, namely, Faster R-CNN framework, to classify malignant and benign lesions from mammograms. This method reported AUC of 0.95 on the INbreast dataset and 0.85 in the Digital Mammography DREAM Challenge. Similarly, Lu et al. [81] achieved AUC of 0.91 to classify metastatic lymph nodes in rectal cancer using the Faster R-CNN framework on MR images.

To match the other omics data counterparts, the term "radiomics" was coined in 2012. Radiomics involves converting medical image data obtained from computed tomography, positron emission tomography, or magnetic resonance imaging to high-dimensional features involving steps like image acquisition and restoration, image segmentation, feature extraction, and then subsequent informatics analyses to perform correlation with clinical and biological data and derive diagnostic prognostic and predictive biomarkers. Applying these methods on drug perturbation data sets has proven to be beneficial in enhancing our understanding of the connection between genes, drugs, and diseases. Mobadersany et al. developed a unified framework based on survival convolutional neural networks (SCNN) for integrating histology and genomic biomarkers for predicting time-to-event outcomes from histological images of glioma from TCGA. In this framework, SCNN learns visual patterns associated with survival using convolution and pooling operations, then fully connected layers provide additional non-linear transformations of the extracted image features. Finally, a Cox proportional hazard layer is added which models

Table 14.3 Major machine learning and deep learning models in the field of medical image analysis

PMID/DOI	Method	Input data	Cancer type	Purpose	Performance
Mu et al. [51]	SResCNN	18F-FDG- PET/CT scans	Non-small cell lung cancer	EGFR mutation status	AUC ≥ 0.81
Jiang et al. [52]	CNN (PMetNet)	CT scan	Gastric	Occult peritoneal metastasis	AUC (0.92–0.94)
Wang et al. [53]	DL (GoogLeNet/ VGG-19)	FFPE WSI slides	Gastrointestinal	Determine TMB	AUC, 0.75–0.82
Jain and Massoud [54]	Image2TMB CNN (Inception-v3)	Frozen H&E slides lung adenocarcinoma (LUAD) TCGA	Lung adenocarcinoma	Determine TMB	AUPRC: 0.92
Kather et al. [55]	CNN (ResNet18)	H&E slides	Gastric, colorectal, and endometrial	MSI prediction	AUC: 0.75–0.84
Yamashita et al. [56]	MSINet (MobileNetV2/ ImageNet)	H&E-stained histopathology images	Colorectal	MSI prediction	AUC: 0.78–0.93
Saltz et al. [57]	Autoencoder + CNN	H&E-stained histopathology slides	13 TCGA cancer types	Tumor-infiltrating lymphocytes (TILs) mapping	Spearman Corr: 0.10–0.45
Bychkov et al. [58]	CNN + RNN	H&E-stained tumor tissue microarray (TMA)	Colorectal	Survival	HR:2.3, AUC:0.69
Akbar et al. [59]	CNN (InceptionNet)	H&E-stained WSI	Breast	Quantify tumor cellularity	Correlation: 0.82
Skrede et al. [60]	CNN (MobileNetV2)	H&E-stained WSI	Colorectal	Cancer-specific survival	HR:3.04, AUC:0.71
Ehteshami Bejnordi et al. [61]	CNN (VGG-Net)	Histopathology images	Breast	Tumor-associated stroma prediction	ACC: 92%
Mobadersany et al. [62]	CNN + Cox reg (SCNN)	Histopathology slides + genomic marker	Glioma	Overall survival	Median c index: 0.75
Xu et al. [63]	CNN+RNN	Longitudinal CT scans	NSCLC	Overall survival	AUC: 0.74, HR: 6.16
McKinney et al. [64]	DL (ImageNet)	Mammograms	Breast	Predict risk score	AUC: 0.75–0.88
Wang et al. [65]	CNN	MRI	Prostate	Classification	AUC: 0.84
Zhou et al. [66]	CNN (SENet and DenseNet)	MRI	Hepatocellular carcinoma	Predict grade (low vs. high)	ACC: 0.83
Korfiatis et al. [67]	Residual DNN (ResNet)	MRI	Brain tumor	Methylation status of MGMT gene	ACC: 94.90%

(continued)

Table 14.3 (continued)

PMID/DOI	Method	Input data	Cancer type	Purpose	Performance
Shboul et al. [68]	XGBoost	MRI	Diffuse low-grade gliomas	IDH1 mutation, MGMT methylation status, 1p/19q co-deletion, ATRX mutation, and TERT mutations	AUC: 0.70–0.84
Fassler et al. [69]	ColorAE: U-Net	Multiplex IHC histopathology slides	Pancreatic cancer	Evaluates the expression of six biomarkers in mIHC	F1: 0.40–0.84
Choi et al. [70]	CNN	PET/MRI scans	Advanced breast cancer	Neoadjuvant chemotherapy	AUC: 0.81
Wang et al. [71]	CNN (ImageNet)	Preoperative CT scans	Lung adenocarcinoma	EGFR mutation status	AUC $\geq$ 0.81
Esteva et al. [72]	CNN (Inception-v3)	Skin lesion images	Melanoma	NA	AUC: 0.91–0.94
Johannet et al. [73]	CNN (Inception-v3)	Treatment-naïve histopathology slides+clinical	Predict responses to checkpoint immunotherapy	Advanced melanoma	AUC: 0.80
Chen et al. [74]	CNN (Inception-v3)	WSI	Hepatocellular carcinoma	CTNNB1, FMN2, TP53, and ZFX4 mutation status	AUC: 0.71–0.89
Coudray et al. [75]	DeepPATH (Inception-v3)	WSI	Lung (LUSC, LUAD)	Classification	AUC: 0.97
Nagpal et al. [76]	CNN (Inception-v3 + KNN)	WSI for H&E-stained prostatectomy specimens	Prostate	Gleason scores	ACC: 0.70
Khosravi et al. [77]	CNN (GoogLeNet V1 and V3)	WSI from H&E-stained tumor tissues	Bladder/breast/lung cancer	Discriminate tissues, subtypes, biomarkers, and scores	ACC: 100%, 92%, 95%, and 69%

time-to-event data, like overall survival or progression-free survival [62]. Another study by Bychkov and colleagues applied a combination of convolutional and recurrent architectures to train a deep network to predict colorectal cancer outcome based on images of tumor tissue samples. The model was evaluated on 420 colorectal cancer patients with clinicopathological and outcome data available, showing that deep learning-based outcome prediction with only small tissue areas as input outperforms (hazard ratio HR 2.3) visual histological assessment performed by human experts on both TMA spot (HR 1.67) and whole-slide level (HR 1.65) in the stratification into low- and high-risk patients [58].

Recently, deep learning has also been applied to learn mutation status in cancer patients. EGFR mutation status is important to identify lung cancer patients eligible for EGFR-TKI treatment. Image-based assays are an important option in this case as tissue-based *EGFR* testing is invasive and the percentage of cells expressing an *EGFR* mutation changes dynamically. Image-based assays provide non-invasive and reproducible methods to determine EGFR status. Wang et al. developed a deep learning model on preoperative CT scan images to predict EGFR mutations. The model was trained on 14,926 images and obtained an AUC of 0.81 on an independent dataset [71]. Mu et al. developed a 2D small-residual-convolutional-network (SResCNN) based on deep learning to predict EGFR mutation status from ^{18}F-MPG, PET/CT imaging. The network was trained on 429 patients and validated on 187 patients and on an additional external dataset of 65 patients, obtaining an AUC of 0.81 [51].

Tumor mutation burden (TMB) is defined as the total number of mutations carried by tumor cells. Patients with higher TMB have a higher antigen load, which aids the immune system to recognize the tumors, and may benefit from treatment with immune checkpoint inhibitors. Measurement of DNA sequencing-based TMB is costly and time-consuming. Recently, deep learning methods have been developed to predict TMB using pathology images. Jain et al. developed Image2TMB using Inception-v3, a convolutional neural network (CNN) that has achieved state-of-the-art performance on the ImageNet Large Scale Visual Recognition Challenge. Image2TMB combines the prediction values from three deep learning models that operate at diverse resolution scales (×5, ×10 and ×20 magnification). On a held-out set of patients, Image2TMB achieved an area under the precision recall curve (AUPRC) of 0.92 [54]. Another study by Wang et al. led to the development of a transfer deep learning method to extract the characteristics of digital whole-slide image and predict TMB from pathological images of gastric and colon cancer. The experimental validation by the GoogLeNet model achieved an AUC of 0.75 for the gastric patients cohort and 0.82 for the colon adenocarcinoma patients cohort, which is of great value for clinical applications [53].

It has been observed that patients with microsatellite instability (MSI) can benefit from immunotherapy. Although it is possible to assess MSI status using genetic or immunohistochemical tests, direct usage of the ubiquitously available H&E-stained images can serve as a good alternative. Kather et al. [55] have employed a CNN with deep residual learning (resnet18) and the ImageNet database on gastric adenocarcinoma and colorectal cancer, achieving patient-level AUC values greater than 0.77. Yamashita et al. [56] used MSINet, a deep learning method using H&E-stained whole-slide images (WSIs) to identify MSI in a subset of patients undergone primary colorectal cancer resection. Their deep learning model outperformed the prediction performance of a panel of five pathologists. MSINet achieved AUROC of 0.93 and 0.78 on internal and external test datasets.

Finally, a study by Xu et al. used pre-treatment and post-treatment follow-up CT images to predict prognostic and other clinical endpoints for non-small cell lung cancer (NSCLC) patients treated with radiation therapy, using a ResNet CNN combined with an RNN. The CNN extracted features from CT images of each time-point, which were fed into a recurrent network for longitudinal analysis. The model predicted 2-year overall survival with 1-month follow-up scan with an AUC of 0.64. The performance of the model increased with the addition of each

follow-up scan up to 6 months to an AUC of 0.74 [63].

AI-Related Models in Oncology Approved by US FDA

The successful implementation of machine learning and deep neural networks on medical and biological data has recently led to the unprecedented approval of AI-based oncology software by the Food and Drug Administration (FDA) (Table 14.4). Radiology is one of the areas where deep learning algorithms have been the most successful. AI-based algorithms for image analysis have not only achieved accuracy and reproducibility but have also reduced the reading time of bulk images [82, 83]. AmCAD-UT is one of the first AI-based devices approved by FDA for the detection of thyroid cancer. It is designed to characterize thyroid nodule sonographic features using pattern recognition and quantification algorithms. The algorithm assesses the risk of malignancy based on thyroid imaging reporting and data systems (TI-RADS). Reverter et al. performed an external validation study on 300 thyroid nodules, which exhibited a comparable sensitivity but lower specificity and area under the receiver operating characteristics (AUROC), compared to the clinical experts using the American Thyroid Association TI-RADS classification system [84]. Arterys Oncology DL, approved in 2018 by FDA, can identify lung nodules (Lung-RADS) and liver lesions (LI-RADS) in an automated fashion using images from CT or MR scans. It does not only help comparing medical images from diverse modalities via 3D visualization, but it also enables clinicians to edit the automated segmentations, thus complementing the standard protocols.

In 2019, FDA cleared cmTriage**,** a notification-only algorithm used to prioritize specific patients to radiologists based on the presence of at least one suspicious lesion obtained from 2D Full-Field Digital Mammography (FFDM) screening mammograms (https://appliedradiology.com/communities/Artificial-Intelligence/the-potential-and-reality-of-ai-in-clinical-application, https://curemetrix.com/cm-triage-2/). This passive notification algorithm helps the radiologist to prioritize patients based on difficult or suspicious cases with added vigilance via Picture Archiving and Communication System (PACS) worklist.

The ProFound™ AI Software V2.1 was approved by FDA in October, 2019. It is a computer-assisted diagnosis device capable of detecting soft tissue density and calcification by

Table 14.4 List of AI software related to oncology approved by the US FDA

Product name	FDA clearance number	Year	Cancer type	Description
AmCAD-UT	K180006	2018	Thyroid cancer	Computer-assisted detection for thyroid cancer based on ultrasound images
Arterys Oncology DL	K173542	2018	Solid tumors (lung and liver cancer)	Measure, track lesions and nodules using CT/MRI scans
ProFound™ AI V2.1	K191994	2018	Breast	Detect malignant soft tissue densities and calcifications from 3D DBT exams
cmTriage	K183285	2019	Breast	Notification triage algorithm based on presence of suspicious region of interest from 2D FFDM screening mammograms
Transpara™ 1.6.0	K193229	2020	Breast	Convolutional neural networks (CNN)-based AI system that aids radiologists in breast lesion detection, diagnosis, and biopsy guidance from FFDM/DBT systems
QuantX	DEN170022	2020	Breast	Artificial intelligence tool that aids the radiologists in breast lesion detection, diagnosis, and biopsy guidance from MRI data
RayCare 2.3	K191384	2019	Pan-cancer	Medical charged-particle radiation therapy system

reading 3D digital breast tomosynthesis (DBT) exams. The algorithm assigns a score to the detected region as well as case which is reflective of the confidence of these being malignant. In other words, scores range from 0 to 100 with a lower to higher likeliness of malignancy. Another decision support AI software for breast cancer approved by the FDA in 2019 is the Transpara™ system. It helps to screen mammograms obtained from FFDM & DBT systems by identifying suspicious soft tissue lesions and calcifications, and reports a region-based as well as an overall score indicating the likelihood of malignancy. The authors have reported that radiologists achieved better performance in lesser time using Transpara (version 1.3.0), which is based on deep learning convolutional neural networks (CNN) and was trained and validated on nearly 9000 mammograms with cancer and a matched number of non-malignant mammograms [85].

QuantX is another FDA-approved computer-aided diagnostic system for breast cancer which aids the radiologists in the assessment and characterization of breast aberrations in MRI data. The QuantX algorithm gives the output in the form of a QI Score, based on the features obtained from the characteristics of the region of interest, which helps the radiologist to compare the lesion with known ground truth available in the form of a reference database. In a clinical study, it showed a 20% improvement in accuracy for breast MRI interpretation over conventional software (https://www.qlarityimaging.com/quantx-research-study). Finally, RayCare is a non-diagnostic information management system including patient data transfer, storage, conversion, and visualization. This system supports scheduling and workflow management, including patient follow-up across different cancer modalities in medical/surgical oncology and radiation therapy.

Current Challenges and Future Perspectives

AI has increasingly become a powerful and indispensable tool for advanced data analysis and inference. In this chapter, we summarized different studies where AI-based techniques have been successfully applied to address different types of problems in cancer research with important applications in precision oncology. The identification of multiple clinically actionable subtypes within a larger cohort is being assessed in prospective clinical trials reflecting the need for different treatment regimens owing to differential response to therapies [22]. The successful applications of AI-based methods for biomedical image analysis and the FDA approvals of several such methods are encouraging and furthering the development of more sophisticated learning and inference tools. Although these efforts are paving the way for real-time AI-based pipelines to be employed in clinic and hospital settings, a lot more needs to be done to establish AI as a precision-point-of-care modality. Effective AI and machine learning rely on big data, and not all precision oncology applications have access to large datasets. For many therapeutics, the only data available from patients are from small clinical trials, where patients' samples are not always subject to high-throughput sequencing. In such cases, it is difficult to build a dataset large enough for a machine learning tool to detect any signal, which is further complicated by the inherent heterogeneity of cancer genomes. For AI applications to be successful, genomic data must be generated properly and abundantly, which is not always possible in many clinical settings. A different way to address this problem, which is also an active area of research, is to devise novel AI strategies that can successfully handle relatively smaller datasets.

Another challenge for the implementation of AI in clinical setting is interpretability. Explainable AI is the branch of AI dedicated to explaining the predictions, but it is still in its infancy. Physicians are not just happy with a *black box* predicting prognosis or treatment for a patient, and understandably so. It is instead crucial to build trust and confidence in recommendation systems by providing explanations supporting diagnostic and prognostic predictions and rationales for the suggested therapies.

In conclusion, AI-based technologies are rapidly reshaping clinical care and becoming a crucial developing part of the precision oncology

field. There is great optimism that AI-powered applications will soon be integrated in clinical practice, providing invaluable insights into patients' disease and assistance to medical personnel.

References

1. Doroshow DB, Doroshow JH. From the broad phase II trial to precision oncology: a perspective on the origins of basket and umbrella clinical trial designs in cancer drug development. Cancer J. 2019;25(4):245–53.
2. Deininger MW, Druker BJ. Specific targeted therapy of chronic myelogenous leukemia with imatinib. Pharmacol Rev. 2003;55(3):401–23.
3. Slatko BE, Gardner AF, Ausubel FM. Overview of next-generation sequencing technologies. Curr Protoc Mol Biol. 2018;122(1):e59.
4. Li X, Wang W, Chen J. Recent progress in mass spectrometry proteomics for biomedical research. Sci China Life Sci. 2017;60(10):1093–113.
5. Werner RJ, Kelly AD, Issa JJ. Epigenetics and precision oncology. Cancer J. 2017;23(5):262–9.
6. Jensen MA, Ferretti V, Grossman RL, Staudt LM. The NCI genomic data commons as an engine for precision medicine. Blood. 2017;130(4):453–9.
7. Grossman RL, Heath A, Murphy M, Patterson M, Wells W. A case for data commons: toward data science as a service. Comput Sci Eng. 2016;18(5):10–20.
8. Prior FW, Clark K, Commean P, Freymann J, Jaffe C, Kirby J, Moore S, Smith K, Tarbox L, Vendt B, Marquez G. TCIA: An information resource to enable open science. Annu Int Conf IEEE Eng Med Biol Soc. 2013;2013:1282–5.
9. Chen F, Zhang Y, Creighton CJ. Systematic identification of non-coding somatic single nucleotide variants associated with altered transcription and DNA methylation in adult and pediatric cancers. NAR Cancer. 2021;3(1):zcab001.
10. Rudnick PA, Markey SP, Roth J, Mirokhin Y, Yan X, Tchekhovskoi DV, Edwards NJ, Thangudu RR, Ketchum KA, Kinsinger CR, Mesri M, Rodriguez H, Stein SE. A description of the clinical proteomic tumor analysis consortium (CPTAC) common data analysis pipeline. J Proteome Res. 2016;15(3):1023–32.
11. Jang IS, Neto EC, Guinney J, Friend SH, Margolin AA. Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. Pac Symp Biocomput. 2014:63–74.
12. Shoemaker RH. The NCI60 human tumour cell line anticancer drug screen. Nat Rev Cancer. 2006;6(10):813–23.
13. Ghandi M, Huang FW, Jane-Valbuena J, Kryukov GV, Lo CC, McDonald ER 3rd, Barretina J, Gelfand ET, Bielski CM, Li H, Hu K, Andreev-Drakhlin AY, Kim J, Hess JM, Haas BJ, Aguet F, Weir BA, Rothberg MV, Paolella BR, Lawrence MS, Akbani R, Lu Y, Tiv HL, Gokhale PC, de Weck A, Mansour AA, Oh C, Shih J, Hadi K, Rosen Y, Bistline J, Venkatesan K, Reddy A, Sonkin D, Liu M, Lehar J, Korn JM, Porter DA, Jones MD, Golji J, Caponigro G, Taylor JE, Dunning CM, Creech AL, Warren AC, McFarland JM, Zamanighomi M, Kauffmann A, Stransky N, Imielinski M, Maruvka YE, Cherniack AD, Tsherniak A, Vazquez F, Jaffe JD, Lane AA, Weinstock DM, Johannessen CM, Morrissey MP, Stegmeier F, Schlegel R, Hahn WC, Getz G, Mills GB, Boehm JS, Golub TR, Garraway LA, Sellers WR. Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature. 2019;569(7757):503–8.
14. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, Bindal N, Beare D, Smith JA, Thompson IR, Ramaswamy S, Futreal PA, Haber DA, Stratton MR, Benes C, McDermott U, Garnett MJ. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res. 2013;41(Database issue):D955–61.
15. Holbeck SL, Camalier R, Crowell JA, Govindharajulu JP, Hollingshead M, Anderson LW, Polley E, Rubinstein L, Srivastava A, Wilsker D, Collins JM, Doroshow JH. The National Cancer Institute ALMANAC: a comprehensive screening resource for the detection of anticancer drug pairs with enhanced therapeutic activity. Cancer Res. 2017;77(13):3564–76.
16. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN, Reich M, Hieronymus H, Wei G, Armstrong SA, Haggarty SJ, Clemons PA, Wei R, Carr SA, Lander ES, Golub TR. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. Science. 2006;313(5795):1929–35.
17. Vidovic D, Koleti A, Schurer SC. Large-scale integration of small molecule-induced genome-wide transcriptional responses, Kinome-wide binding affinities and cell-growth inhibition profiles reveal global trends characterizing systems-level drug action. Front Genet. 2014;5:342.
18. Meyers RM, Bryan JG, McFarland JM, Weir BA, Sizemore AE, Xu H, Dharia NV, Montgomery PG, Cowley GS, Pantel S, Goodale A, Lee Y, Ali LD, Jiang G, Lubonja R, Harrington WF, Strickland M, Wu T, Hawes DC, Zhivich VA, Wyatt MR, Kalani Z, Chang JJ, Okamoto M, Stegmaier K, Golub TR, Boehm JS, Vazquez F, Root DE, Hahn WC, Tsherniak A. Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. Nat Genet. 2017;49(12):1779–84.
19. Bhalla S, Melnekoff DTM, Keats J, Onel K, Madduri D, Richter J, Richard S, Chari A, Cho HJ, Dudley JT, Jagannath S, Laganà A, Parekh S. Patient similarity network of multiple myeloma identifies patient sub-groups with distinct genetic

and clinical features. bioRxiv. 2020; https://doi.org/10.1101/2020.06.02.129767.
20. Cavalli FMG, Remke M, Rampasek L, Peacock J, Shih DJH, Luu B, Garzia L, Torchia J, Nor C, Morrissy AS, Agnihotri S, Thompson YY, Kuzan-Fischer CM, Farooq H, Isaev K, Daniels C, Cho BK, Kim SK, Wang KC, Lee JY, Grajkowska WA, Perek-Polnik M, Vasiljevic A, Faure-Conter C, Jouvet A, Giannini C, Nageswara Rao AA, Li KKW, Ng HK, Eberhart CG, Pollack IF, Hamilton RL, Gillespie GY, Olson JM, Leary S, Weiss WA, Lach B, Chambless LB, Thompson RC, Cooper MK, Vibhakar R, Hauser P, van Veelen MC, Kros JM, French PJ, Ra YS, Kumabe T, Lopez-Aguilar E, Zitterbart K, Sterba J, Finocchiaro G, Massimino M, Van Meir EG, Osuka S, Shofuda T, Klekner A, Zollo M, Leonard JR, Rubin JB, Jabado N, Albrecht S, Mora J, Van Meter TE, Jung S, Moore AS, Hallahan AR, Chan JA, Tirapelli DPC, Carlotti CG, Fouladi M, Pimentel J, Faria CC, Saad AG, Massimi L, Liau LM, Wheeler H, Nakamura H, Elbabaa SK, Perezpena-Diazconti M, Chico Ponce de Leon F, Robinson S, Zapotocky M, Lassaletta A, Huang A, Hawkins CE, Tabori U, Bouffet E, Bartels U, Dirks PB, Rutka JT, Bader GD, Reimand J, Goldenberg A, Ramaswamy V, Taylor MD. Intertumoral heterogeneity within Medulloblastoma subgroups. Cancer Cell. 2017;31(6):737–754 e736.
21. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, Haibe-Kains B, Goldenberg A. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods. 2014;11(3):333–7.
22. Upadhyaya SA, Robinson GW, Onar-Thomas A, Orr BA, Johann P, Wu G, Billups CA, Tatevossian RG, Dhanda SK, Srinivasan A, Broniscer A, Qaddoumi I, Vinitsky A, Armstrong GT, Bendel AE, Hassall T, Partap S, Fisher PG, Crawford JR, Chintagumpala M, Bouffet E, Gururangan S, Mostafavi R, Sanders RP, Klimo P Jr, Patay Z, Indelicato DJ, Nichols KE, Boop FA, Merchant TE, Kool M, Ellison DW, Gajjar A. Relevance of molecular groups in children with newly diagnosed atypical teratoid rhabdoid tumor: results from prospective St. Jude multi-institutional trials. Clin Cancer Res. 2021;27(10):2879–89.
23. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490(7418):61–70.
24. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, Graf S, Ha G, Haffari G, Bashashati A, Russell R, McKinney S, METABRIC Group, Langerod A, Green A, Provenzano E, Wishart G, Pinder S, Watson P, Markowetz F, Murphy L, Ellis I, Purushotham A, Borresen-Dale AL, Brenton JD, Tavare S, Caldas C, Aparicio S. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012;486(7403):346–52.
25. Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, Leiserson MDM, Niu B, McLellan MD, Uzunangelov V, Zhang J, Kandoth C, Akbani R, Shen H, Omberg L, Chu A, Margolin AA, Van't Veer LJ, Lopez-Bigas N, Laird PW, Raphael BJ, Ding L, Robertson AG, Byers LA, Mills GB, Weinstein JN, Van Waes C, Chen Z, Collisson EA, Cancer Genome Atlas Research Network, Benz CC, Perou CM, Stuart JM. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell. 2014;158(4):929–44.
26. Lu CF, Hsu FT, Hsieh KL, Kao YJ, Cheng SJ, Hsu JB, Tsai PH, Chen RJ, Huang CC, Yen Y, Chen CY. Machine learning-based radiomics for molecular subtyping of gliomas. Clin Cancer Res. 2018;24(18):4429–36.
27. Mallavarapu T, Hao J, Kim Y, Oh JH, Kang M. Pathway-based deep clustering for molecular subtyping of cancer. Methods. 2020;173:24–31.
28. Fang C, Xu D, Su J, Dry JR, Linghu B. DeePaN: deep patient graph convolutional network integrating clinico-genomic evidence to stratify lung cancers for immunotherapy. NPJ Digit Med. 2021;4(1):14.
29. Le DT, Durham JN, Smith KN, Wang H, Bartlett BR, Aulakh LK, Lu S, Kemberling H, Wilt C, Luber BS, Wong F, Azad NS, Rucki AA, Laheru D, Donehower R, Zaheer A, Fisher GA, Crocenzi TS, Lee JJ, Greten TF, Duffy AG, Ciombor KK, Eyring AD, Lam BH, Joe A, Kang SP, Holdhoff M, Danilova L, Cope L, Meyer C, Zhou S, Goldberg RM, Armstrong DK, Bever KM, Fader AN, Taube J, Housseau F, Spetzler D, Xiao N, Pardoll DM, Papadopoulos N, Kinzler KW, Eshleman JR, Vogelstein B, Anders RA, Diaz LA Jr. Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. Science. 2017;357(6349):409–13.
30. Drilon A, Laetsch TW, Kummar S, DuBois SG, Lassen UN, Demetri GD, Nathenson M, Doebele RC, Farago AF, Pappo AS, Turpin B, Dowlati A, Brose MS, Mascarenhas L, Federman N, Berlin J, El-Deiry WS, Baik C, Deeken J, Boni V, Nagasubramanian R, Taylor M, Rudzinski ER, Meric-Bernstam F, Sohal DPS, Ma PC, Raez LE, Hechtman JF, Benayed R, Ladanyi M, Tuch BB, Ebata K, Cruickshank S, Ku NC, Cox MC, Hawkins DS, Hong DS, Hyman DM. Efficacy of larotrectinib in TRK fusion-positive cancers in adults and children. N Engl J Med. 2018;378(8):731–9.
31. Gupta S, Chaudhary K, Kumar R, Gautam A, Nanda JS, Dhanda SK, Brahmachari SK, Raghava GP. Prioritization of anticancer drugs against a cancer using genomic features of cancer cells: a step towards personalized medicine. Sci Rep. 2016;6:23857.
32. Dong Z, Zhang N, Li C, Wang H, Fang Y, Wang J, Zheng X. Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection. BMC Cancer. 2015;15:489.
33. Zhang F, Wang M, Xi J, Yang J, Li A. A novel heterogeneous network-based method for drug response prediction in cancer cell lines. Sci Rep. 2018;8(1):3355.
34. Lee SI, Celik S, Logsdon BA, Lundberg SM, Martins TJ, Oehler VG, Estey EH, Miller CP, Chien S, Dai J, Saxena A, Blau CA, Becker PS. A machine learning approach to integrate big data for precision

medicine in acute myeloid leukemia. Nat Commun. 2018;9(1):42.
35. Di J, Zheng B, Kong Q, Jiang Y, Liu S, Yang Y, Han X, Sheng Y, Zhang Y, Cheng L, Han J. Prioritization of candidate cancer drugs based on a drug functional similarity network constructed by integrating pathway activities and drug activities. Mol Oncol. 2019;13(10):2259–77.
36. Yan X, Yang Y, Chen Z, Yin Z, Deng Z, Qiu T, Tang K, Cao Z. H-RACS: a handy tool to rank anti-cancer synergistic drugs. Aging (Albany NY). 2020;12(21):21504–17.
37. Chang Y, Park H, Yang HJ, Lee S, Lee KY, Kim TS, Jung J, Shin JM. Cancer drug response profile scan (CDRscan): a deep learning model that predicts drug effectiveness from cancer genomic signature. Sci Rep. 2018;8(1):8857.
38. Rampasek L, Hidru D, Smirnov P, Haibe-Kains B, Goldenberg A. Dr.VAE: improving drug response prediction via modeling of drug perturbation effects. Bioinformatics. 2019;35(19):3743–51.
39. Preuer K, Lewis RPI, Hochreiter S, Bender A, Bulusu KC, Klambauer G. DeepSynergy: predicting anti-cancer drug synergy with deep learning. Bioinformatics. 2018;34(9):1538–46.
40. Xia F, Shukla M, Brettin T, Garcia-Cardona C, Cohn J, Allen JE, Maslov S, Holbeck SL, Doroshow JH, Evrard YA, Stahlberg EA, Stevens RL. Predicting tumor cell line response to drug pairs with deep learning. BMC Bioinform. 2018;19(Suppl 18):486.
41. Chen G, Tsoi A, Xu H, Zheng WJ. Predict effective drug combination by deep belief network and ontology fingerprints. J Biomed Inform. 2018;85:149–54.
42. Li M, Wang Y, Zheng R, Shi X, Li Y, Wu FX, Wang J. DeepDSC: a deep learning method to predict drug sensitivity of cancer cell lines. IEEE/ACM Trans Comput Biol Bioinform. 2021;18(2):575–82.
43. Ding MQ, Chen L, Cooper GF, Young JD, Lu X. Precision oncology beyond targeted therapy: combining omics data with machine learning matches the majority of cancer cells to effective therapeutics. Mol Cancer Res. 2018;16(2):269–78.
44. Ammad-Ud-Din M, Khan SA, Malani D, Murumagi A, Kallioniemi O, Aittokallio T, Kaski S. Drug response prediction by inferring pathway-response associations with kernelized Bayesian matrix factorization. Bioinformatics. 2016;32(17):i455–63.
45. Sun W, Sanderson PE, Zheng W. Drug combination therapy increases successful drug repositioning. Drug Discov Today. 2016;21(7):1189–95.
46. Hecht JR, Mitchell E, Chidiac T, Scroggin C, Hagenstad C, Spigel D, Marshall J, Cohn A, McCollum D, Stella P, Deeter R, Shahin S, Amado RG. A randomized phase IIIB trial of chemotherapy, bevacizumab, and panitumumab compared with chemotherapy and bevacizumab alone for metastatic colorectal cancer. J Clin Oncol. 2009;27(5):672–80.
47. Tol J, Koopman M, Cats A, Rodenburg CJ, Creemers GJ, Schrama JG, Erdkamp FL, Vos AH, van Groeningen CJ, Sinnige HA, Richel DJ, Voest EE, Dijkstra JR, Vink-Borger ME, Antonini NF, Mol L, van Krieken JH, Dalesio O, Punt CJ. Chemotherapy, bevacizumab, and cetuximab in metastatic colorectal cancer. N Engl J Med. 2009;360(6):563–72.
48. Lee JS, Nair NU, Dinstag G, Chapman L, Chung Y, Wang K, Sinha S, Cha H, Kim D, Schperberg AV, Srinivasan A, Lazar V, Rubin E, Hwang S, Berger R, Beker T, Ronai Z, Hannenhalli S, Gilbert MR, Kurzrock R, Lee SH, Aldape K, Ruppin E. Synthetic lethality-mediated precision oncology via the tumor transcriptome. Cell. 2021;184(9):2487–2502 e2413.
49. Yuan B, Shen C, Luna A, Korkut A, Marks DS, Ingraham J, Sander C. CellBox: interpretable machine learning for perturbation biology with application to the design of cancer combination therapy. Cell Syst. 2021;12(2):128–140 e124.
50. Kim YG, Kim S, Cho CE, Song IH, Lee HJ, Ahn S, Park SY, Gong G, Kim N. Effectiveness of transfer learning for enhancing tumor classification with a convolutional neural network on frozen sections. Sci Rep. 2020;10(1):21899.
51. Mu W, Jiang L, Zhang J, Shi Y, Gray JE, Tunali I, Gao C, Sun Y, Tian J, Zhao X, Sun X, Gillies RJ, Schabath MB. Non-invasive decision support for NSCLC treatment using PET/CT radiomics. Nat Commun. 2020;11(1):5228.
52. Jiang Y, Liang X, Wang W, Chen C, Yuan Q, Zhang X, Li N, Chen H, Yu J, Xie Y, Xu Y, Zhou Z, Li G, Li R. Noninvasive prediction of occult peritoneal metastasis in gastric cancer using deep learning. JAMA Netw Open. 2021;4(1):e2032269.
53. Wang L, Jiao Y, Qiao Y, Zeng N, Yu R. A novel approach combined transfer learning and deep learning to predict TMB from histology image. Pattern Recogn Lett. 2020;135:244–8.
54. Jain MS, Massoud TF. Predicting tumour mutational burden from histopathological images using multiscale deep learning. Nat Mach Intell. 2020;2:356–62.
55. Kather JN, Pearson AT, Halama N, Jager D, Krause J, Loosen SH, Marx A, Boor P, Tacke F, Neumann UP, Grabsch HI, Yoshikawa T, Brenner H, Chang-Claude J, Hoffmeister M, Trautwein C, Luedde T. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nat Med. 2019;25(7):1054–6.
56. Yamashita R, Long J, Longacre T, Peng L, Berry G, Martin B, Higgins J, Rubin DL, Shen J. Deep learning model for the prediction of microsatellite instability in colorectal cancer: a diagnostic study. Lancet Oncol. 2021;22(1):132–41.
57. Saltz J, Gupta R, Hou L, Kurc T, Singh P, Nguyen V, Samaras D, Shroyer KR, Zhao T, Batiste R, Van Arnam J, Cancer Genome Atlas Research Network, Shmulevich I, Rao AUK, Lazar AJ, Sharma A, Thorsson V. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. Cell Rep. 2018;23(1):181–193 e187.
58. Bychkov D, Linder N, Turkki R, Nordling S, Kovanen PE, Verrill C, Walliander M, Lundin M, Haglund

C, Lundin J. Deep learning based tissue analysis predicts outcome in colorectal cancer. Sci Rep. 2018;8(1):3395.
59. Akbar S, Peikari M, Salama S, Panah AY, Nofech-Mozes S, Martel AL. Automated and manual quantification of tumour cellularity in digital slides for tumour burden assessment. Sci Rep. 2019;9(1):14099.
60. Skrede OJ, De Raedt S, Kleppe A, Hveem TS, Liestol K, Maddison J, Askautrud HA, Pradhan M, Nesheim JA, Albregtsen F, Farstad IN, Domingo E, Church DN, Nesbakken A, Shepherd NA, Tomlinson I, Kerr R, Novelli M, Kerr DJ, Danielsen HE. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. Lancet. 2020;395(10221):350–60.
61. Ehteshami Bejnordi B, Mullooly M, Pfeiffer RM, Fan S, Vacek PM, Weaver DL, Herschorn S, Brinton LA, van Ginneken B, Karssemeijer N, Beck AH, Gierach GL, van der Laak J, Sherman ME. Using deep convolutional neural networks to identify and classify tumor-associated stroma in diagnostic breast biopsies. Mod Pathol. 2018;31(10):1502–12.
62. Mobadersany P, Yousefi S, Amgad M, Gutman DA, Barnholtz-Sloan JS, Velazquez Vega JE, Brat DJ, Cooper LAD. Predicting cancer outcomes from histology and genomics using convolutional networks. Proc Natl Acad Sci U S A. 2018;115(13):E2970–9.
63. Xu Y, Hosny A, Zeleznik R, Parmar C, Coroller T, Franco I, Mak RH, Aerts H. Deep learning predicts lung cancer treatment response from serial medical imaging. Clin Cancer Res. 2019;25(11):3266–75.
64. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, Back T, Chesus M, Corrado GS, Darzi A, Etemadi M, Garcia-Vicente F, Gilbert FJ, Halling-Brown M, Hassabis D, Jansen S, Karthikesalingam A, Kelly CJ, King D, Ledsam JR, Melnick D, Mostofi H, Peng L, Reicher JJ, Romera-Paredes B, Sidebottom R, Suleyman M, Tse D, Young KC, De Fauw J, Shetty S. International evaluation of an AI system for breast cancer screening. Nature. 2020;577(7788):89–94.
65. Wang X, Yang W, Weinreb J, Han J, Li Q, Kong X, Yan Y, Ke Z, Luo B, Liu T, Wang L. Searching for prostate cancer by fully automated magnetic resonance imaging classification: deep learning versus non-deep learning. Sci Rep. 2017;7(1):15415.
66. Zhou Q, Zhou Z, Chen C, Fan G, Chen G, Heng H, Ji J, Dai Y. Grading of hepatocellular carcinoma using 3D SE-DenseNet in dynamic enhanced MR images. Comput Biol Med. 2019;107:47–57.
67. Korfiatis P, Kline TL, Lachance DH, Parney IF, Buckner JC, Erickson BJ. Residual deep convolutional neural network predicts MGMT methylation status. J Digit Imaging. 2017;30(5):622–8.
68. Shboul, Z. A., J. Chen and , KM Iftekharuddin (2020). Prediction of molecular mutations in diffuse low-grade gliomas using MR imaging features. Sci Rep 10(1): 3711.
69. Fassler DJ, Abousamra S, Gupta R, Chen C, Zhao M, Paredes D, Batool SA, Knudsen BS, Escobar-Hoyos L, Shroyer KR, Samaras D, Kurc T, Saltz J. Deep learning-based image analysis methods for brightfield-acquired multiplex immunohistochemistry images. Diagn Pathol. 2020;15(1):100.
70. Choi JH, Kim HA, Kim W, Lim I, Lee I, Byun BH, Noh WC, Seong MK, Lee SS, Kim BI, Choi CW, Lim SM, Woo SK. Early prediction of neoadjuvant chemotherapy response for advanced breast cancer using PET/MRI image deep learning. Sci Rep. 2020;10(1):21149.
71. Wang S, Shi J, Ye Z, Dong D, Yu D, Zhou M, Liu Y, Gevaert O, Wang K, Zhu Y, Zhou H, Liu Z, Tian J. Predicting EGFR mutation status in lung adenocarcinoma on computed tomography image using deep learning. Eur Respir J. 2019;53(3):1800986.
72. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115–8.
73. Johannet P, Coudray N, Donnelly DM, Jour G, Illa-Bochaca I, Xia Y, Johnson DB, Wheless L, Patrinely JR, Nomikou S, Rimm DL, Pavlick AC, Weber JS, Zhong J, Tsirigos A, Osman I. Using machine learning algorithms to predict immunotherapy response in patients with advanced melanoma. Clin Cancer Res. 2021;27(1):131–40.
74. Chen M, Zhang B, Topatana W, Cao J, Zhu H, Juengpanich S, Mao Q, Yu H, Cai X. Classification and mutation prediction based on histopathology H&E images in liver cancer using deep learning. NPJ Precis Oncol. 2020;4:14.
75. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyo D, Moreira AL, Razavian N, Tsirigos A. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat Med. 2018;24(10):1559–67.
76. Nagpal K, Foote D, Liu Y, Chen PC, Wulczyn E, Tan F, Olson N, Smith JL, Mohtashamian A, Wren JH, Corrado GS, MacDonald R, Peng LH, Amin MB, Evans AJ, Sangoi AR, Mermel CH, Hipp JD, Stumpe MC. Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. NPJ Digit Med. 2019;2:48.
77. Khosravi P, Kazemi E, Imielinski M, Elemento O, Hajirasouliha I. Deep convolutional neural networks enable discrimination of heterogeneous digital pathology images. EBioMedicine. 2018;27:317–28.
78. Haenssle HA, Fink C, Schneiderbauer R, Toberer F, Buhl T, Blum A, Kalloo A, Hassen ABH, Thomas L, Enk A, Uhlmann L, Reader study level-I and level-II Groups, Alt C, Arenbergerova M, Bakos R, Baltzer A, Bertlich I, Blum A, Bokor-Billmann T, Bowling J, Braghiroli N, Braun R, Buder-Bakhaya K, Buhl T, Cabo H, Cabrijan L, Cevic N, Classen A, Deltgen D, Fink C, Georgieva I, Hakim-Meibodi LE, Hanner S, Hartmann F, Hartmann J, Haus G, Hoxha E, Karls R, Koga H, Kreusch J, Lallas A, Majenka P, Marghoob A, Massone C, Mekokishvili L, Mestel D, Meyer V, Neuberger A, Nielsen K, Oliviero M, Pampena R, Paoli J, Pawlik E, Rao B, Rendon A, Russo T, Sadek

A, Samhaber K, Schneiderbauer R, Schweizer A, Toberer F, Trennheuser L, Vlahova L, Wald A, Winkler J, Wolbing P, Zalaudek I. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann Oncol. 2018;29(8):1836–42.
79. Azuaje F, Kim SY, Perez Hernandez D, Dittmar G. Connecting histopathology imaging and proteomics in kidney cancer through machine learning. J Clin Med. 2019;8(10):1535.
80. Ribli D, Horvath A, Unger Z, Pollner P, Csabai I. Detecting and classifying lesions in mammograms with deep learning. Sci Rep. 2018;8(1):4165.
81. Lu Y, Yu Q, Gao Y, Zhou Y, Liu G, Dong Q, Ma J, Ding L, Yao H, Zhang Z, Xiao G, An Q, Wang G, Xi J, Yuan W, Lian Y, Zhang D, Zhao C, Yao Q, Liu W, Zhou X, Liu S, Wu Q, Xu W, Zhang J, Wang D, Sun Z, Gao Y, Zhang X, Hu J, Zhang M, Wang G, Zheng X, Wang L, Zhao J, Yang S. Identification of metastatic lymph nodes in MR imaging with faster region-based convolutional neural networks. Cancer Res. 2018;78(17):5135–43.
82. Transin S, Souchon R, Gonindard-Melodelima C, de Rozario R, Walker P, Funes de la Vega M, Loffroy R, Cormier L, Rouviere O. Computer-aided diagnosis system for characterizing ISUP grade>/=2 prostate cancers at multiparametric MRI: a cross-vendor evaluation. Diagn Interv Imaging. 2019;100(12):801–11.
83. Wang S, Burtt K, Turkbey B, Choyke P, Summers RM. Computer aided-diagnosis of prostate cancer on multiparametric MRI: a technical review of current research. Biomed Res Int. 2014;2014:789561.
84. Reverter JL, Vazquez F, Puig-Domingo M. Diagnostic performance evaluation of a computer-assisted imaging analysis system for ultrasound risk stratification of thyroid nodules. AJR Am J Roentgenol. 2019;213(1):169–74.
85. Rodriguez-Ruiz A, Krupinski E, Mordang JJ, Schilling K, Heywang-Kobrunner SH, Sechopoulos I, Mann RM. Detection of breast cancer with mammography: effect of an artificial intelligence support system. Radiology. 2019;290(2):305–14.

15 Single-Cell Sequencing Technologies in Precision Oncology

David T. Melnekoff and Alessandro Laganà

Abstract

Single-cell sequencing technologies are revolutionizing cancer research and are poised to become the standard for translational cancer studies. Rapidly decreasing costs and increasing throughput and resolution are paving the way for the adoption of single-cell technologies in clinical settings for personalized medicine applications. In this chapter, we review the state of the art of single-cell DNA and RNA sequencing technologies, the computational tools to analyze the data, and their potential application to precision oncology. We also discuss the advantages of single-cell over bulk sequencing for the dissection of intra-tumor heterogeneity and the characterization of subclonal cell populations, the implementation of targeted drug repurposing approaches, and describe advanced methodologies for multi-omics data integration and to assess cell signaling at single-cell resolution.

D. T. Melnekoff
Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA

A. Laganà (✉)
Department of Genetics and Genomic Sciences, Department of Oncological Sciences, Mount Sinai Icahn School of Medicine, New York, NY, USA
e-mail: alessandro.lagana@mssm.edu

Introduction

Single-cell sequencing technologies have revolutionized the way we investigate disease processes. The ability to measure omics data at single-cell resolution has led to great advances in our understanding of the immune system, developmental biology, bacterial-host interactions, and cancer mechanisms of action. In the context of genomics-based precision cancer medicine, most platforms have relied on bulk genomic sequencing technologies such as whole exome sequencing (WES), RNA sequencing (RNA-seq), or whole genome sequencing (WGS), to identify druggable targets. These technologies rely on the pooling of genomic materials from tumor samples, and then subsequent sequencing, giving you a snapshot of the mutational and transcriptomic landscape of the disease. But these technologies fail to address a major confounding factor in cancer precision medicine: intra-tumor heterogeneity (ITH).

Intra-tumor heterogeneity, or tumor clonality, is not a new finding, and was discovered well before the advent of next-generation sequencing (NGS) [1]. While the concept of variations between patients has been well known for

© Springer Nature Switzerland AG 2022

A. Laganà (ed.), *Computational Methods for Precision Oncology*, Advances in Experimental Medicine and Biology 1361, https://doi.org/10.1007/978-3-030-91836-1_15

decades, the idea that within the same individual there are multiple subsets of cells which have unique genetic aberrations is quickly becoming a popular area of focus in cancer research and treatment [2, 3]. There are multiple models of intra-tumor heterogeneity, but all focus on cellular evolution of different populations of cancer cells with distinct genetic features which impart survival advantage in the tumor environment, whether in a solid tumor mass [4–6], or hematological malignancies [7]. In the context of treatment resistance and relapse, intra-tumor heterogeneity can lead to the "bottleneck effect," which selects for resistant clonal populations and leads to the expansion of resistant downstream disease clones. Thus, it is extremely important to take intra-tumor heterogeneity into account when attempting to perform precision and personalized cancer medicine, in order to obtain superior outcomes and possibly full eradication of disease.

Bulk Versus Single-Cell Sequencing for ITH Estimates in Precision Oncology

There are now many commercially available precision diagnostic and therapeutic tests for cancer including products from Foundation Medicine, Tempus Labs, and Memorial Sloan Kettering [8–10]. All of these tests rely on bulk sequencing of tumor DNA and RNA, to identify DNA aberrations such as mutations, or RNA aberrations such as gene fusions. While these tools have greatly expanded the therapeutic toolbox available to physicians, they lack the sensitivity to profile patient disease at the clonal level and can be improved by taking into account multiple disease clones with multiple actionable targets.

Disease clones are defined by cellular populations which have distinct DNA alterations [11]. In an attempt to leverage existing NGS platforms for investigation of cancer at the clonal level, methodologies have been developed for WES to describe tumor clones. Tools such as PhyloWGS, CANOPY, and SciClone are publicly available for the estimation of clonal composition of tumor samples [12–14]. These tools mainly rely on clustering of the Variant Allele Frequency (VAF) of mutations. The VAF is the proportion of alternate (mutated) alleles to reference (non-mutated) alleles and describes the prevalence of specific mutations within a tumor sample. If mutations are acquired in a temporal manner, and cells containing the mutations expand, mutations with higher VAF are assumed to have occurred earlier in tumor growth, while mutations with lower VAF are assumed to have occurred later on in the tumor lifespan. While these tools rely on different sets of probabilistic assumptions, such as the infinite-sites assumption[1] in the case of PhyoWGS, they all rely on a clustering of mutations based on VAF, the rationale being that mutations with similar VAF were acquired at the same time. While these methods provide varying levels of estimation of tumor heterogeneity, they can also be misleading. This is due to multiple factors, such as the inability to detect very small/rare subclones at low VAF, copy number alterations which can also confer survival advantage and confound VAF of driver mutations, and an inability to separate possible co-occurring clones, also known as different disease clones which evolved at the same time. Alternatively, single-cell sequencing technologies directly measure each cell's genomic content, at a resolution which far surpasses WES and even targeted bulk sequencing approaches. This allows for the clustering of cells, not mutations, and thus removes much of the uncertainty around clonal estimates.

Another aspect of ITH which is critical to measure for effective precision medicine is tumor phylogeny, or the temporal and structural map of disease clones. Some of the tools mentioned previously, such as PhyloWGS, perform both VAF clustering and phylogeny estimations on WES data, while there are other tools which only perform phylogeny prediction such as CALDER and CloneEvol [15, 16]. Phylogeny estimates show the relationship of disease clones to each other, and an understanding of this structure would allow for effective tailoring of treatments. Downstream or children disease clones will har-

[1] The infinite-sites assumption posits that a site does note mutate twice during the evolutionary history of a tumor.

bor all the aberrations of the parent clone, and thus multiple disease clones may be targeted by a single therapeutic. In order to estimate tumor phylogeny from bulk sequencing technologies, probability modeling and statistical functions need to be designed to determine the "most likely" tree. These models may not be completely accurate, especially if there are minor subclones or LoH events, which are difficult to capture from bulk sequencing modalities. Using single-cell sequencing technologies we can easily see the temporal acquisition of mutations by measuring the presence and absence of mutations and copy number alterations within each cell. The ability to limit the amount of treatment a patient receives by intelligently designing combinatorial therapies based on tumor phylogeny would lead to better outcomes and less side effects.

RNA-seq for ITH in Precision Oncology

While most precision medicine tools and workflows focus on matching drugs to specific genomic alterations, very few leverage RNA transcript expression information to recommend therapeutics. It has been shown that using RNA expression to determine pathway activation, or dysregulation of specific genes, can lead to effective treatment recommendations with patients that lack actionable genomic alterations [17]. In the context of ITH though, RNA-seq technology does not allow for the estimation of clonal expression profiles. This is due to the measurement of pooled transcript abundance from an entire tumor sample. Reads from RNA-seq do not correspond directly to number of copies of DNA alleles, since a single region of the genome is transcribed repeatedly to generate multiple mRNAs, and ultimately proteins. Furthermore, mapping of RNA transcripts to the human genome often discards mutated mRNAs, and mutation detection from RNA is notoriously unreliable. Deconvolution methods to segment RNA-seq data into distinct expression signatures of different cell populations exist, such as xCell and CIBERSORT, but require known expression signatures [18, 19]. This is normally reserved for samples of diverse cellular composition, such as whole blood, where there exist known expression signatures of different white blood cell populations. In the context of tumor cells, deconvolution would not be able to resolve disease clones, because a known signature would need to be extracted from the RNA-seq measurements, and this is exactly what is trying to be determined. Using single-cell RNA sequencing (scRNA-seq) allows for the direct measurement of RNA transcript abundance in single cells and allows for the clustering of cells with similar expression profiles. These clustered expression profiles can then be used to perform cluster specific analysis, such as pathway analysis or RNA-based drug repurposing. It is important to reiterate that RNA-based cell clusters do not represent tumor disease clones, which are strictly defined by genomic alterations. Regardless, disease clones may evolve from genomic alterations which have no known targeted therapeutic. In this case, we can utilize the scRNA-seq data to supplement our precision medicine predictions. This will allow for better combinatorial therapy prediction, and hopefully avoid the selection of resistant, and non-targetable, disease clones.

Isolation of Tumor Cells Versus Tumor Microenvironment

Precision medicine is dependent on profiling patient-derived samples for analysis. The sample collection process is different for each different type of cancer. In the case of hematological malignancies, samples can be collected from the blood or the bone marrow. In the case of solid tumors, biopsies are resected from patient tumors surgically or via fine-needle aspirate. In both cases, it is important to have pure tumor sample so that results are not confounded from normal cells. In the context of hematological malignancies, pure tumor can normally be isolated by cell sorting techniques based on cell surface markers, such as CD138+ selection of plasma cells in multiple myeloma. In the context of solid tumors, it may be unclear if only tumor cells were resected.

Furthermore, even though when performing WES-matched normal sample are used to filter out germline mutations, the tumor microenvironment may have been included in the tumor biopsy. The tumor microenvironment harbors its own distinct alterations which would be considered to be somatic alterations within the tumor and confound results. Utilizing single-cell sequencing techniques would allow for the isolation and removal of tumor microenvironment cells from the analysis, leading to more accurate therapeutic recommendations. Moreover, this would allow for the analysis of the tumor microenvironment separately since cells from this niche have been implicated in tumor growth and immune suppression.

Single-Cell RNA-seq

The Technology

scRNA-seq has advanced dramatically over the past decade, and currently there are a plethora of different scRNA-seq technologies available. These technologies can be broadly divided into two classes: full-length transcript capture, and 3′ or 5′ end capture platforms [20]. Many reviews have been published which outline the specific strength and weaknesses of specific platforms, which differ in transcript capture class, transcript abundance quantification, and upstream sample preparation [21–23]. In short, the differences between scRNA-seq platforms can be distilled to the comparison of sensitivity vs. throughput. Full-length transcript capture platforms such as Smart-seq2 and MATQ-seq allow for the more precise analysis of individual transcripts including the measurement of allelic imbalance, RNA editing, and distinct transcript isoforms, with the penalty of lower throughput and a higher sequencing cost per cell [24]. 3′ and 5′ end capture platforms such as the 10× Genomics CHROMIUM platform allows for massive high-throughput sequencing of thousands of cells based on a microfluidic platform [25]. This platform, and other droplet sequencing-based platforms such as MARS-seq [26], can allow for the profiling of much larger cell counts vs. full-transcript capture platforms which are often plate based. Droplet-based assays are normally cheaper to perform as well. The combined increased cellular throughput and reduced cost make droplet-based platforms more suitable for precision medicine (Table 15.1).

Table 15.1 Single-cell sequencing technologies

Method	Name	Transcript measurement	Throughput	Cell isolation	Reference
scRNA-seq	Smart-seq2	Full-length	100–1000 cells	Well-based	Picelli et al. [22]
	MATQ-seq	Full-length	100–1000 cells	Plate-based	Sheng et al. (2017)
	MARS-seq	3′	100–1000 cells	Plate-based	Jaitin et al. [26]
	Cel-seq2	3′	100–1000 cells	Plate-based	Hashimshony et al. (2016)
	Chromium	3′ or 5′	~10,000 cells	Droplet-based	Zheng et al. (2017)
	SPLiT-seq	3′	1000–100,000 cells	Plate-based	Rosenberg et al. (2018)
	Quartz-seq2	3′	1000–100,000 cells	Plate-based	Sasagawa et al. (2018)
scDNA-seq	Sic-seq	WGA	>50,000 cells	Droplet-based	Lan et al. [61]
	Mission bio Tapestri	Amplicon	>50,000 cells	Droplet-based	Pellegrino et al [62]
	10× Genomics scCNV	WGA	~10,000 cells	Droplet-based	10xgenomics.com

Advantages of scRNA-seq in Precision Oncology

The paradigm of precision, or personalized, oncology treatment is dependent on in-depth characterization of patient-specific disease factors. As mentioned previously, bulk sequencing technologies provide an approximation, or average measurement, of RNA or DNA alterations in a single patient sample, whereas single-cell sequencing technologies allow the characterization of disease at cellular resolution. Therefore, bulk sequencing technologies may overstate the alterations present in dominant disease clones and minimize those in minor clones. scRNA-seq measurement would allow for the direct measurement of disease cellular states and allow for alternative methodologies to increase the therapeutic toolbox for oncologists.

Investigating Intra-Tumor Heterogeneity with scRNA-seq

scRNA-seq has been shown in the research setting to be a significant tool for investigating ITH. Guan et al. showed marked heterogeneity in the triple negative breast cancer cell line SUM149, including the expression of other classical breast cancer subtype markers such as HER2 and ER [27]. In patient samples, marked heterogeneity across tumor and immune subsets have been described in lung, bladder, and skin [28–30]. This is because scRNA-seq allows for the clustering of cellular populations based on similar transcriptomic expression. Many software packages have been developed for this distinct purpose, including the Seurat R package and Bioconductor scRNA-seq R workflow. However, cellular populations identified by transcriptomic clustering may not represent "disease clones," which are defined by genomic alterations. There are many confounding effects which may lead to inaccurate measurement of disease clones based on transcriptomic clustering, especially in scRNA-seq data, which is exceptionally sensitive to biological noise. One of the best documented effects is the cell cycle state of the cells at the time of sequencing, and multiple tools have been developed to regress out these effects [31–33]. Due to the rapid evolution and cellular turn-over of cancer cells, tumor biopsies may contain cells at various stages within the cell cycle. Therefore, the downstream transcriptomic measurements may be dominated by cell cycle-specific effects, leading to clustering of cells into cellular states rather than tumor clones or transcriptional programs. Regression of cell cycle signals is therefore crucial for accurate cellular clustering into functionally linked groups.

Even after appropriate QC, scRNA-seq clusters may not be definitive disease clones. In practice scRNA-seq would often overestimate the number of clones within a patient sample, due to the larger variation in transcriptome state vs. genomic state. This could be detrimental in the determination of accurate treatment recommendations for precision oncology, especially when drug toxicity and interaction must be considered. One methodology to further refine tumor scRNA-seq data, is the inference of copy number alterations/variations (CNA/CNV) from the data, to inform and refine cellular clustering.

While the downstream transcriptomic effects of certain oncogenic drivers such as RAS family mutations have been well studied, the extent that other oncogene and tumor suppressor aberrations drive transcriptomic changes are not yet well understood. The ability to investigate distinct transcriptomic signatures of genomic alterations is one of the holy grails of scRNA-seq analysis and would lead to a plethora of new data points for possible therapeutic recommendations at clonal resolution. However, identifying mutations from scRNA-seq data is challenging. Fan et al. showed that the ability to detect known SNPs from paired WES in scRNA-seq data was only 0.34 [34]. However, CNV changes are also a well-documented source of ITH [35, 36]. CNV changes are more likely to have a direct effect in the expression profile of the genes contained within the CNV region, and thus provide an avenue to map genomic alteration onto single-cell transcriptomic data. Multiple methods have been developed to estimate CNV data from scRNA-seq data with and without paired bulk DNA

sequencing, including HoneyBADGER and inferCNV, respectively [34, 37–40]. After mapping CNV states onto scRNA-seq data, cells can be clustered into genomically informed groups which would lead to more accurate clonal estimates. Using these cell clusters, we can generate clone-specific transcriptomic profiles, and utilize these signatures with downstream analysis to identify possibly therapeutic targets (Table 15.2).

Pathway Analysis Using scRNA-seq

Pathway analysis is a common methodology to determine biological and functional insights from sets of differentially expressed genes. An abundance of methods has been developed to determine pathway activation or enrichment from bulk RNA-seq data, such as gene set enrichment analysis (GSEA), gene set variation analysis (GSVA), and signaling pathway impact analysis (SPIA) [41–43]. These tools allow for group and sample level pathway analysis, and have been shown to be effective in classifying patients into disease subgroups, and even informing treatment decision-making [17]. Pathway analysis in scRNA-seq data can be more challenging due to the technical limitations of the platform, such as a high drop-out rate and low coverage of certain regions of the transcriptome. Due to these issues, there are primarily two different methodologies for pathway analysis in scRNA-seq data: those which depend on a pre-determined differential expression (DE) profile, and those which determine pathways from raw scRNA-seq counts matrices and cluster cells based on those pathways. Tools which have been developed specifically for scRNA-seq, and rely on input of the entire cell-count matrix without a priori clustering, include PAGODA2, SCENIC, and iDEA [44–46]. An excellent in-depth review of the accuracy of many of these tools can be found by Zhang et al. [47] In short, tools varied greatly in their ability to cluster cells and define pathway enrichment within the same scRNA-seq dataset. Data preprocessing was also critical for influencing the outcomes of pathway analysis. In the context of clinical precision medicine, these differences could be detrimental to patient outcomes. Thus, we will focus on pathway activation/enrichment tools which use predetermined DE profiles, for example, those determined by pre-clustering based on available expert knowledge (such as grouping of immune/tumor subsets on known marker expression) and inferred genomic data. This will increase the likelihood of developing clinically relevant and actionable pathway estimates for disease clones.

An advantage of using pre-clustered cells with DE expression profiles as our input for pathway analysis is that it allows for the usage of bulk RNA-seq pathway tools. Zhang et al. showed that the liger R package, which is an implementation of GSEA algorithm, was successful in identifying relevant pathways for immune subsets in scRNA-seq data of rheumatoid arthritis joint synovial tissues [48]. GSEA and GSVA also performed well in identifying immune cell subsets from their differential expression profiles derived from cellular clustering [49]. GSEA was also used to characterize newly identified subsets of chondrocytes in osteoarthritis, showing its utility outside of well-defined cell populations [50]. Topology-based pathway methods such as SPIA can also be used with scRNA-seq data, but they have yet to be benchmarked (Table 15.2).

RNA-Based Drug Repurposing

Another avenue to therapeutic decision-making based on transcriptomics is RNA signature-based drug repurposing. This methodology has been used to match RNA signatures of disease to the reverse signature of drugs in cell lines, under the assumption that the drug would “reverse” the signature of the disease. Using DE profiles derived from cell clusters in scRNA-seq, RNA signature data bases can be queried for reverse signatures to form clone specific drug repurposing recommendations. The LINCS L1000 database (L1000), Cancer Cell Line Encyclopedia (CCLE), and Genomics of Drug Sensitivity in Cancer (GDSC) datasets all contain RNA-profiles of cell line sensi-

Table 15.2 Common tools for data analysis in bulk and single-cell sequencing data

Function	Name	Implementation	Description	Reference
Pathway analysis: Bulk tools	GSEA	Multiple including desktop client, web, and R	Enrichment for pathways between defined groups	Subramanian et al. [42]
	GSVA	R	Rank-based pathway activation	Hänzelmann et al. [43]
	SPIA	R	Topology-based method	Tarca et al. [41]
Pathway analysis: scRNA-seq clustering and pathway identification	Pagoda2	R	Clustering, visualization, and gene-set/ overdispersion analysis	Yung et al. (2018)
	SCENIC	R	Cell clustering based on gene-regulatory networks	Aibar et al. [45]
	iDEA	R	Integrated differential expression and gene set enrichment analysis	Ma et al. [46]
	scTPA	Web client/R	Pathway activation signature in a web-based client to identify functionally different cell types	Zhang et al. [47]
Drug repurposing	L1000FWD	Python/R/web client	Gene signature mapping of drug treated cell lines	Wang et al. (2018)
	GDSC tools	Python	Analyze drug interactions with genomic/ transcriptomic features	Cokelaer et al. [55]
	CaDRReS-Sc	Python	Machine learning algorithm to match scRNA-seq data with GDSC cell line data	Suphavilai et al. (2018)
Identification of CNV from scRNA-seq	Infer-CNV	R	Generate CNV profile from scRNA-seq reads	https://github.com/broadinstitute/infercnv
	Honey-badger	R	Generate CNV profile from scRNA-seq and matched bulk-exome	Fan et al. [34]
Cell-cell interaction	SingleCellSignalR	R	LR interactions in scRNA-seq data using curated LR database	Cabello-Aguilar et al. [73]
	CellPhoneDB	R/web client	Cell-cell interactions from scRNA-seq data including cell type clustering	Efremova et al. [74]

tivity data to thousands of compounds [51–53]. Each database has its own web-based portal for browsing and querying the data using a gene or gene list as input. Tools have also been developed to directly query these databases through command line tools, such as L1000FWD and GDSCtools [54, 55]. These tools allow for the direct attribution of therapeutics, even those outside of the realm of cancer, to specific cancer clones based on transcriptomic profiles, and may greatly expand the therapeutic options available for patients. Furthermore, these tools may reveal patterns of therapeutics within cancer subtypes that lead to hypothesis generation for further functional and mechanistic validation (Table 15.2).

Single-Cell DNA-seq

The Technology

While scRNA-seq has become a widespread paradigm in biological research over the past decade, scDNA-seq has remained more elusive. This is due to the significant variation and difficulty in nucleic acid amplification and extraction methodologies. One would ideally seek to characterize all possible genomic alterations in single cells using scDNA-seq technology using whole genome amplification (WGA), including both SNVs and CNVs. The difficulty in using single-cell WGA with respect to variant calling has been described in depth [56–58]. Each of the two major classes of single-cell WGA, PCR and multiple displacement amplification (MDA), are more suitable for CNV and SNV detection, respectively [59]. However, as stated previously, both SNVs and CNVs can be drivers of ITH and tumor evolution. Thus, the ability to measure both type of alterations accurately is crucial for accurate tumor clone identification. Another challenge in scDNA-seq is scalability. Up until recently, most scDNA-seq technologies have been limited to hundreds of cells, as opposed to thousands of cells [60]. The scalability of the technology must be a similar cellular throughput as drop-seq methods in scRNA-seq, to allow for the isolation and identification of rare subclones, which can measure <1% of tumor cells. Finally, for application in personalized medicine, the scDNA-seq technology should be an out-of-the-box solution which does not require significant wet lab intervention or customization. Two recent microfluidic technologies, SiC-seq and the Tapestri Platform from Mission Bio, both achieve cellular throughput of >50,000 cells using WGA or targeted amplicon sequencing, respectively [61, 62]. While both platforms still harbor the pitfalls of their respective amplification technologies, namely, lack of specificity vs. lack of coverage, both have been used to characterize ITH in cancer. The ability to resolve genomically distinct populations within a single tumor sample, at high resolution, makes these platforms ideal for personalized cancer therapy prediction.

Advantages of scDNA-seq in Precision Oncology

Single-cell DNA-seq solves multiple problems with regard to measuring ITH: (1) the determination of co-occurrence of genomic aberrations, (2) the temporal relationship between distinct tumor clones, (3) resolution of uncertainty in VAF in regions with an overlapping CNV and SNV, and (4) the identification of rare populations below the threshold of NGS sequencing. The direct measurement of DNA sequences from single cells allows for direct measurement of ITH, rather than imputation from bulk WES or WGS. This allows for a greater chance of assigning treatment regimens that hit all disease clones with the minimal amount of therapeutics, thus increasing treatment efficacy and reducing side effects.

Investigating Intra-Tumor Heterogeneity with scDNA-seq

The specificity and confidence in ITH measurements is greatly enhanced by scDNA-seq. The methods used to call mutations are similar, if not identical to, the methods used for bulk DNA sequencing. Once mutations and CNV values are assigned per cell, disease clones can be identi-

fied. Instead of taking the totality of mutations and CNV events and merging them together by a common metric such as VAF, we can select mutations and CNVs which are known or suspected to be deleterious. Similar to the analysis performed by Ediriwickrema, A. et al., mutations found in scDNA-seq data were selected by their appearance in clinical cancer databases such as ClinVAR [63]. These mutations were then grouped in the context of specific cellular populations, leading to the definition of distinct disease clones. The temporal relationship between clones is fairly simple to ascertain, by appreciating the appearance, disappearance, and co-occurrence of SNVs and CNVs within each cell. VAFs of SNPs within each cell are given as output, so CNVs can be inferred by VAFs which fall between homozygous (100%) and heterozygous (50%). For example, an amplification of an SNV region after a mutation would lead to a VAF of 66% (2 mutated, 1 wt allele), while an SNV which occurred after a CNV amplification would lead to a VAF of ~33% (1 mutated allele, 1 wt allele). This further allows for the reconciliation of the temporal nature of CNVs and SNVs within a single tumor sample. Finally, the high sensitivity and throughput of modern microfluidic scDNA-seq platforms allows for the characterization of very rare subclonal populations. A single tumor cell was identified out of a total cell output of ~8 thousand cells from a remission biopsy in AML, which would require a read depth greater than 16,000 to confidently identify it through bulk sequencing [63]. The ability to not only capture, but genomically profile such rare cell populations would be invaluable for precision oncology, because these rare populations could be the difference between a long-term cure and relapse.

Assigning Therapeutics in scDNA-seq Data

As mentioned previously, multiple precision medicine pipelines rely on matching genomic alterations found with bulk sequencing to specific therapies. scDNA-seq supplies a similar output, but in a cell population-specific manner. However the same resources can still be used for drug recommendations, such as CivicDB, OncoKB, and PMKB [64–66]. In short, these databases contain information from clinical and research trials which associate genomic alterations with approved and research therapeutics, along with an evidence score which indicates the level of robustness of the study supporting the use of the therapeutics. By querying each clones' alterations separately, we can ascertain combinatorial therapies that would attack multiple, if not all, disease clones. This would hopefully lead to complete remission, or at least, a reduction in the possibility of treatment resistance (Table 15.2).

Single-Cell Multi-Omics and Integration of Single-Cell Platforms

While RNA and DNA sequencing are the most common measurements used to characterize patient disease, it has been shown that bulk multi-omics analysis is beneficial in generating a comprehensive view of mechanistic pathways and the relationships between cellular processes. As described in a review by Lee et al., assays such as chromatin accessibility, proteomics, and DNA methylation have been performed in concert with bulk DNA or RNA measurements to further elucidate disease processes and lead to a better understanding of cancer biology [67]. For example, multi-omics assays of the combined effect of CNVs and DNA methylation on the transcriptional profiles of liver cancer samples revealed three new prognostic groups [68]. Furthermore, as scRNA-seq and scDNA-seq have expanded the resolution and specificity of transcriptome and genomic measurements, single-cell multi-omics technologies have recently been developed to measure multiple cellular properties simultaneously from single cells. These single-cell multi-omics technologies vary in their throughput, sensitivity, and analytes, the majority of which can measure two distinct analytes from single cells. In the context of precision oncology,

having the ability to measure all aspects of a cell's biology would lead to more avenues to predicting successful therapeutics; however, no such technology exists yet. We have shown that both RNA and DNA information can be used to predict effective therapeutics, but single-cell technologies which measure both RNA and DNA at scale have remained elusive [67]. Therefore, the integration of paired and separate single-cell assays, similar to bulk multi-omics experiments, would allow for a more comprehensive view of patient disease, and would be scalable for use in precision oncology.

Integration of scDNA-Seq and scRNA-Seq

As explained previously, matching genomic features from bulk WES to scRNA-seq data can be helpful in defining tumor clones at the single-cell level. scDNA-seq can also directly measure distinct tumor populations. Therefore, the integration of scDNA-seq and scRNA-seq is possible and would be effective in defining the genome and transcriptome of disease clones, including rare clonal populations which are only found with the high resolution of scDNA-seq. Single-cell CNV profiles can be determined from scDNA-seq by measuring the allelic depths between each cell at specific loci. In order to generate accurate CNV profiles for each cell, a normal diploid cell population must be present in order to normalize the read depth, since the read depth will be specific to each run. In the specific case where purified tumor is used, spiking in a normal diploid cell line in a known concentration is advised to generate absolute, and not relative, CNV profiles. After determining cell-specific CNV profiles, cells can be clustered into different clonal populations based on CNVs. These clonal CNV profiles can then be mapped to scRNA-seq data using statistical or machine learning methods which compare regions of known CNVs and scRNA-seq transcript depths. Clone-Align is an example of a machine learning method which uses a Bayesian Inference to map clone-specific CNV profiles, generated from scDNA-seq data, to scRNA-seq data performed on the same sample [69]. This methodology would provide more accurate clonal CNV profiles than inferred from scRNA-seq data and allow for the merging of both genomic and transcriptomic information at single-cell resolution.

SNVs identified from scDNA-seq data can also be computationally matched to scRNA-seq data. Specifically, the mutations can be investigated directly within the individual reads from scRNA-seq data. In the case of 10x Genomics Chromium scRNA-seq data, the tool Vartrix can be used to query the data for known variants [70]. In the case of other scRNA-seq technologies, common tools to investigate BAM files may be used, such as Samtools and Picard [71, 72]. While allelic drop out and variable coverage in the paired scRNA-seq data may be an issue, high confidence variants called from scDNA-seq should overcome the uncertainty in scRNA-seq data. Furthermore, cells which cluster closely to cells with measured mutations may be inferred to also harbor the same mutations. Finally, leveraging CNV data may allow for the assignment of mutations to scRNA-seq data where SNVs are not detectably, mainly in the case of co-occurring CNVs and SNVs as measured by scDNA-seq. A schematic for the workflow proposed can be seen in Fig. 15.1.

The integration of single-cell platforms is also a strategy to circumvent the current limit of measuring two types of omics data from a single cell. Single-cell multi-omics platforms like CITE-seq (RNA-Protein), MissionBio Tapestri (DNA-Protein), or Paired-seq (RNA-Chromatin) can generate data corresponding to the "same" cell, which can then be integrated using the same approaches to merge DNA and RNA features discussed above. This is especially exciting in the context of novel immune therapies, which are currently the focus of many clinical trials in cancer. Immune therapies are normally molecules with specific affinity to tumor cell surface proteins and require an interaction between surface proteins of immune effector cells and target malignant cells. Using a platform such as CITE-seq can show the relationship between transcriptomic profiles and the surface expression of

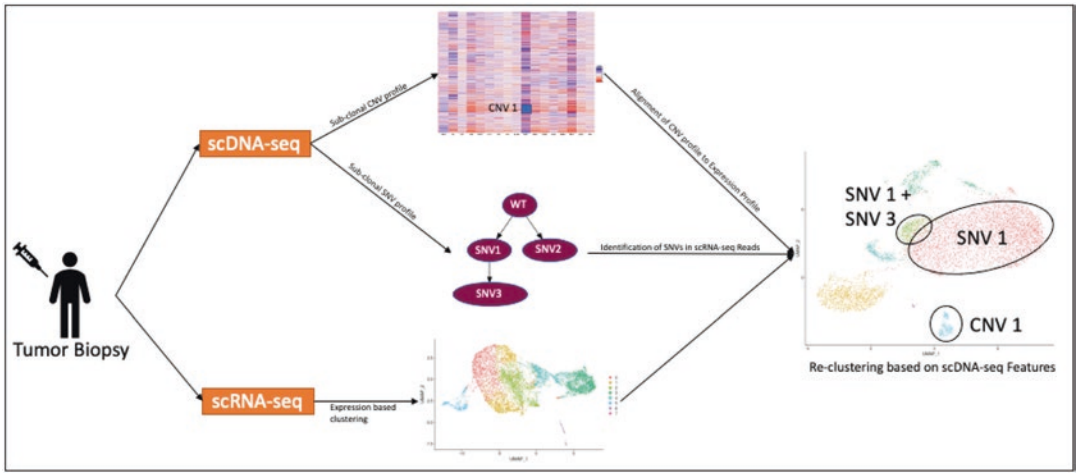

Fig. 15.1 Workflow for the integration of scDNA-seq and scRNA-seq data

transcript proteins. Mission Bio Tapestri platform also allows for simultaneous protein expression and genomic measurements. The ability to relate transcriptomic or genomic features to external protein expression would be extremely useful for predicting immune therapy responsiveness, determine factors for relapse, and possible alternative targets or complimentary therapies. By integrating one multi-omics platform with a single-omics platform, we can generate a conglomerate picture of disease state, spanning three or more different omics technologies.

Cell-Cell Signaling

Single-cell measurements can also be leveraged to determine interactions between cells, which can be useful when interrogating the specific cell types involved in disease processes or therapeutic action. For example, with the increasing popularity and approval of immune-based therapies for cancer, the interaction between tumor and immune-microenvironment cells would be invaluable to determine the reasons for therapeutic sensitivity or resistance. While extra-cellular measurements of circulating proteins can be performed using clinical or research assays such as Ella or O-Link Proteomics, this does not provide directional or cell-specific interactions. New software tools such as SingleCellSignalR [73] or CellPhoneDB [74] have been developed to use single-cell transcriptomics to map Ligand-Receptor(LR) interactions between cells within a single-cell experiment. Leveraging this information, with the potential of multi-omics single-cell experiments, would allow for a more comprehensive understanding of tumor-microenvironment interactions, for example, the ability to determine if specific tumor clones are resistant to immunetherapy due to lacking specific LR interactions with effector T cells, or the secretion of inhibitory cytokines from specific tumor cells. Understanding which tumor populations need to be targeted with secondary agents to supplement immune-oncology therapeutics is currently a major unmet need in the field, and utilizing cell-cell interaction networks may provide the information necessary to lead to precision immune-oncology solutions.

Conclusion

Just as the cost reduction and out-of-the-box next-generation sequencing solutions ushered in an explosion of genomic, transcriptomic, and epigenomic data in a clinical setting, single-cell sequencing is fast becoming a viable option as a clinically focused technology. Single-cell sequencing technologies have already transformed cancer research as we know it and are

poised to become the standard for cancer analysis. The unparalleled resolution and sensitivity of single-cell sequencing can be exploited to better profile patients prior to therapy, track patient's disease progression, and characterize patients at relapse, all while providing vital information to further our understanding of cancer processes and inform clinical decision-making. The inclusion of single-cell sequencing into clinical cancer sequencing pipelines is obviously imminent and could revolutionize patient care in the coming decade.

References

1. Potter VR. Biochemical uniformity and heterogeneity in cancer tissue (further discussion). Cancer Res. 1956;16(7):658–67.
2. Winterhoff BJ, et al. Single cell sequencing reveals heterogeneity within ovarian cancer epithelium and cancer associated stromal cells. Gynecol Oncol. 2017;144:598–606.
3. Dong X, et al. The impact of intratumoral metabolic heterogeneity on postoperative recurrence and survival in resectable esophageal squamous cell carcinoma. Oncotarget. 2017;8:14969–77.
4. González-García I, Solé RV, Costa J. Metapopulation dynamics and spatial heterogeneity in cancer. Proc Natl Acad Sci. 2002;99:13085–9.
5. Giaretti W, et al. Intratumor heterogeneity of K-ras2 mutations in colorectal adenocarcinomas: association with degree of DNA aneuploidy. Am J Pathol. 1996;149:237–45.
6. Shipitsin M, et al. Molecular definition of breast tumor heterogeneity. Cancer Cell. 2007;11:259–73.
7. Campbell PJ, et al. Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. Proc Natl Acad Sci U S A. 2008;105:13081–6.
8. Frampton GM, et al. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. Nat Biotechnol. 2013;31:1023–31.
9. Cheng DT, et al. Memorial Sloan Kettering-integrated mutation profiling of actionable cancer targets (MSK-IMPACT): a hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. J Mol Diagn. 2015;17:251–64.
10. Beaubier N, et al. Integrated genomic profiling expands clinical options for patients with cancer. Nat Biotechnol. 2019;37:1351–60.
11. Marusyk A, Polyak K. Tumor heterogeneity: causes and consequences. Biochim Biophys Acta. 2010;1805:105.
12. PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors | Genome Biology | Full Text. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-015-0602-8.
13. Jiang Y, Qiu Y, Minn AJ, Zhang NR. Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. Proc Natl Acad Sci. 2016;113:E5528–37.
14. SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003665.
15. Hx D, et al. ClonEvol: clonal ordering and visualization in cancer sequencing. Ann Oncol. 2017;28. https://pubmed.ncbi.nlm.nih.gov/28950321/
16. Myers MA, Satas G, Raphael BJ. CALDER: Inferring phylogenetic trees from longitudinal tumor samples. Cell Syst. 2019;8:514–522.e5.
17. Laganà A, et al. Precision medicine for relapsed multiple myeloma on the basis of an integrative multiomics approach. JCO Precis Oncol. 2018;1–17 https://doi.org/10.1200/PO.18.00019.
18. Aran D, Hu, Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biol. 2017;18(1):220. https://doi.org/10.1186/s13059-017-1349-1.
19. Chen B, Khodadoust MS, Liu CL, Newman AM, Alizadeh AA. Profiling tumor infiltrating immune cells with CIBERSORT. Methods Mol Biol Clifton NJ. 2018;1711:243.
20. Chen G, Ning B, Shi T. Single-cell RNA-Seq technologies and related computational data analysis. Front Genet. 2019;10:317.
21. Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA. The technology and biology of single-cell RNA sequencing. Mol Cell. 2015;58:610–20.
22. Picelli S. Single-cell RNA-sequencing: the future of genome biology is now. RNA Biol. 2017;14:637–50.
23. Haque A, Engel J, Teichmann SA, Lönnberg T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Med. 2017;9:75.
24. Ziegenhain C, et al. Comparative analysis of single-cell RNA sequencing methods. Mol Cell. 2017;65:631–643.e4.
25. Massively parallel digital transcriptional profiling of single cells | Nature Communications. https://www.nature.com/articles/ncomms14049.
26. Jaitin DA, et al. Massively parallel single-cell RNA-Seq for marker-free decomposition of tissues into cell types. Science. 2014;343:776–9.
27. Wu S, et al. Cellular, transcriptomic and isoform heterogeneity of breast cancer cell line revealed by full-length single-cell RNA sequencing. Comput Struct Biotechnol J. 2020;18:676–85.
28. Lee HW, et al. Single-cell RNA sequencing reveals the tumor microenvironment and facilitates strategic choices to circumvent treatment failure in a chemorefractory bladder cancer patient. Genome Med. 2020;12:47.

29. He D, et al. Single-cell RNA sequencing reveals heterogeneous tumor and immune cell populations in early-stage lung adenocarcinomas harboring EGFR mutations. Oncogene. 2021;40:355–68.
30. Tirosh I, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science. 2016;352:189–96.
31. Leng N, et al. OEFinder: a user interface to identify and visualize ordering effects in single-cell RNA-seq data. Bioinforma Oxf Engl. 2016;32:1408–10.
32. Tsang JCH, et al. Single-cell transcriptomic reconstruction reveals cell cycle and multi-lineage differentiation defects in Bcl11a-deficient hematopoietic stem cells. Genome Biol. 2015;16:178.
33. Juliá M, Telenti A, Rausell A. Sincell: an R/Bioconductor package for statistical assessment of cell-state hierarchies from single-cell RNA-seq. Bioinforma Oxf Engl. 2015;31:3380–2.
34. Fan J, et al. Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data. Genome Res. 2018;28:1217–27.
35. Vogelstein B, et al. Cancer genome landscapes. Science. 2013;339:1546–58.
36. Melchor L, et al. Single-cell genetic analysis reveals the composition of initiating clones and phylogenetic patterns of branching and parallel evolution in myeloma. Leukemia. 2014;28:1705–15.
37. Serin Harmanci A, Harmanci AO, Zhou X. CaSpER identifies and visualizes CNV events by integrative analysis of single-cell or bulk RNA-sequencing data. Nat Commun. 2020;11:89.
38. Visualizing Large-scale Copy Number Variation in Single-Cell RNA-Seq Expression Data. https://bioconductor.org/packages/devel/bioc/vignettes/infercnv/inst/doc/inferCNV.html.
39. Feng X, et al. SCYN: single cell CNV profiling method using dynamic programming. bioRxiv. 2020;2020.03.27.011353. https://doi.org/10.1101/2020.03.27.011353.
40. Madipour-Shirayeh A, et al. Simultaneous profiling of DNA copy number variations and transcriptional programs in single cells using RNA-seq. bioRxiv. 2020;2020.02.10.942607. https://doi.org/10.1101/2020.02.10.942607.
41. Tarca AL, et al. A novel signaling pathway impact analysis. Bioinformatics. 2009;25:75–82.
42. Subramanian A, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci. 2005;102:15545–50.
43. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinform. 2013;14:7.
44. Lake BB, et al. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. Nat Biotechnol. 2018;36:70–80.
45. Aibar S, et al. SCENIC: single-cell regulatory network inference and clustering. Nat Methods. 2017;14:1083–6.
46. Ma Y, et al. Integrative differential expression and gene set enrichment analysis using summary statistics for scRNA-seq studies. Nat Commun. 2020;11:1585.
47. Zhang Y, et al. Benchmarking algorithms for pathway activity transformation of single-cell RNA-seq data. Comput Struct Biotechnol J. 2020;18:2953–61.
48. Zhang F, et al. Defining inflammatory cell states in rheumatoid arthritis joint synovial tissues by integrating single-cell transcriptomics and mass cytometry. Nat Immunol. 2019;20:928–42.
49. Diaz-Mejia JJ, Meng EC, Pico AR, MacParland SA, Ketela T, Pugh TJ, Bader GD, Morris JH. Evaluation of methods to assign cell type labels to cell clusters from single-cell RNA-sequencing data. F1000Res. 2019;8:ISCB Comm J-296. https://doi.org/10.12688/f1000research.18490.3. eCollection 2019
50. Ji Q, et al. Single-cell RNA-seq analysis reveals the progression of human osteoarthritis. Ann Rheum Dis. 2019;78:100–10.
51. Stathias V, et al. LINCS Data Portal 2.0: next generation access point for perturbation-response signatures. Nucleic Acids Res. 2020;48:D431–9.
52. Barretina J, et al. The cancer cell line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012;483:603–7.
53. Yang W, et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res. 2013;41:D955–61.
54. L1000FWD: fireworks visualization of drug-induced transcriptomic signatures | Bioinformatics | Oxford Academic. https://academic.oup.com/bioinformatics/article/34/12/2150/4840732.
55. Cokelaer T, et al. GDSCTools for mining pharmacogenomic interactions in cancer. Bioinformatics. 2018;34:1226–8.
56. Hou Y, et al. Comparison of variations detection between whole-genome amplification methods used in single-cell resequencing. GigaScience. 2015;4:37.
57. Huang L, Ma F, Chapman A, Lu S, Xie XS. Single-cell whole-genome amplification and sequencing: methodology and applications. Annu Rev Genomics Hum Genet. 2015;16:79–102.
58. Estévez-Gómez N, et al. Comparison of single-cell whole-genome amplification strategies. bioRxiv. 2018;443754 https://doi.org/10.1101/443754.
59. Lähnemann D, et al. Eleven grand challenges in single-cell data science. Genome Biol. 2020;21:31.
60. Andor N, et al. Joint single cell DNA-Seq and RNA-Seq of gastric cancer reveals subclonal signatures of genomic instability and gene expression. bioRxiv. 2018;445932 https://doi.org/10.1101/445932.
61. Lan F, Demaree B, Ahmed N, Abate AR. Single-cell genome sequencing at ultra-high-throughput with microfluidic droplet barcoding. Nat Biotechnol. 2017;35:640–6.
62. Pellegrino M, et al. High-throughput single-cell DNA sequencing of acute myeloid leukemia tumors with droplet microfluidics. Genome Res. 2018;28:1345–52.

63. Ediriwickrema A, et al. Single-cell mutational profiling enhances the clinical evaluation of AML MRD. Blood Adv. 2020;4:943–52.
64. Chakravarty D, et al. OncoKB: a precision oncology knowledge base. JCO Precis Oncol. 2017;1–16 https://doi.org/10.1200/PO.17.00011.
65. Huang L, et al. The cancer precision medicine knowledge base for structured clinical-grade mutations and interpretations. J Am Med Inform Assoc. 2017;24:513–9.
66. Griffith M, et al. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. Nat Genet. 2017;49:170–4.
67. Lee J, Hyeon DY, Hwang D. Single-cell multiomics: technologies and data analysis methods. Exp Mol Med. 2020;52:1428–42.
68. Woo HG, et al. Integrative analysis of genomic and epigenomic regulation of the transcriptome in liver cancer. Nat Commun. 2017;8:839.
69. Campbell KR, et al. Clonealign: statistical integration of independent single-cell RNA and DNA sequencing data from human cancers. Genome Biol. 2019;20:54.
70. 10XGenomics/vartrix. (10x Genomics, 2021).
71. Li H, et al. The sequence alignment/map format and SAMtools. Bioinforma Oxf Engl. 2009;25:2078–9.
72. Picard Tools – By Broad Institute. http://broadinstitute.github.io/picard/.
73. Cabello-Aguilar S, et al. SingleCellSignalR: inference of intercellular networks from single-cell transcriptomics. Nucleic Acids Res. 2020;48:e55.
74. Efremova M, Vento-Tormo M, Teichmann SA, Vento-Tormo R. CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. Nat Protoc. 2020;15:1484–506.

Multi-Omics Profiling of the Tumor Microenvironment

16

Oliver Van Oekelen and Alessandro Laganà

Abstract

All solid tumors and many hematological malignancies grow and proliferate in a tumor microenvironment (TME), a spectrum of continuous and highly dynamic interactions with different immune and stromal cells. This ecosystem contributes to the extensive heterogeneity that exists between and within cancer patients. Understanding the characteristics of this intricate network could significantly improve cancer prognosis, as was demonstrated already for a subset of patients by the advent of immunotherapies (including monoclonal antibodies, bispecific antibodies, and chimeric antigen receptor (CAR) T cells. The development of multimodal *omics* technologies has allowed researchers to document and characterize the TME at single-cell resolution, which provides an unprecedent opportunity to understand the full complexity of the tumor microenvironment. In this chapter, we highlight the paradigm shift that has brought the TME to the forefront of cancer research and discuss its composition. In addition, we summarize the available multimodal single-cell *omics* methods that allow studying the TME from different angles, as well as their advantages and limitations. We discuss computational analysis tools, data integration, and methods to specifically study crosstalk between TME components. Finally, we touch upon the implications of studying the TME for ongoing or future clinical studies and how these can lead to more effective treatments for cancer patients.

Introduction

For many decades, cancer research has been primarily focused on understanding the distinct characteristics and specific vulnerabilities of tumor cells. This has led to the development of a wide variety of targeted and non-targeted therapeutic interventions [1]. It has become increasingly clear, however, that this approach fails to fully capture disease heterogeneity.

O. Van Oekelen
Department of Genomics and Data Science, Icahn School of Medicine at Mount Sinai, New York, NY, USA

A. Laganà (✉)
Department of Genetics and Genomic Sciences, Department of Oncological Sciences, Mount Sinai Icahn School of Medicine, New York, NY, USA
e-mail: alessandro.lagana@mssm.edu

© Springer Nature Switzerland AG 2022

A. Laganà (ed.), *Computational Methods for Precision Oncology*, Advances in Experimental Medicine and Biology 1361, https://doi.org/10.1007/978-3-030-91836-1_16

More recently, the pivotal role of targeting non-malignant components of the tumor microenvironment (TME) has become evident, most explicitly heralded by impressive and durable responses to immune-mediated therapies such as checkpoint inhibitors against PD-1, PD-L1, and CTLA-4 across different tumor types, albeit only in a subset of patients [2–6]. This suggests a largely untapped potential and emphasizes the unmet need to better understand the composition and dynamics of the TME, with the ultimate goal to develop more effective anticancer treatments and rational treatment combinations.

Concurrently, biomedical research has seen a rapid rise in the development and propagation of so-called *omics* technologies that allow the comprehensive study of biological molecules that determine the structure and function of cells (DNA, RNA, protein, metabolites, etc.). Whereas initially these technologies were applied to *bulk* data (i.e., averaged across a large population of input cells), recent advances in molecular biology have led to the availability of technologies that allow profiling of (epi)genetic, proteomic, and spatial data modalities in individual cells. Today, the objective of simultaneously characterizing multiple data types in the same cell to achieve true *single-cell multi-omics* profiling is quickly becoming reality.

It should come as no surprise that the paradigm shift toward increased interest in the TME coincided with the development of effective single-cell *omics* methods. It is precisely those technologies that have allowed us to query the heterogeneity present in the TME ecosystem. In this chapter, we set out to discuss how the microenvironment has come to play an increasingly central role in oncology. We will summarize the approaches to study the composition of and crosstalk within the TME from multi-omics data sources. We believe that further developments and efforts in this field will lead to key insights that can help bring durable advantages to patients with cancer.

The Central Role of Microenvironment in Oncology

The tumor microenvironment (TME) has increasingly come into focus as a major determinant of clinical phenotype and prognosis in multiple cancer types [2, 5]. An appreciation of the significance of the tumor environment is in itself not new. Paget's *seed and soil theory* posited in the nineteenth century that cancer metastases are critically dependent on the properties of the organ where they arise [7, 8]. The TME is best understood as an intricately connected network in which cellular and non-cellular components together create a dynamical niche where cancer cells can thrive [4]. The different components of the TME interact with each other and with the cancer cells directly or indirectly, thereby affecting tumor growth and proliferation.

This contextual view replaces a somewhat outdated tumor cell-intrinsic perspective in which cancer development is largely or exclusively driven by a stepwise process of increasingly complex genomic aberrations that lead to uncontrolled cell growth [9, 10]. A better understanding of signal transduction in general, and the notion that the TME provides signals that impact cancer gene expression in a way that is comparable to the impact of oncogenes and tumor suppressor genes was important for this change in mindset [8]. Admittedly, the appreciation of cancer as a genetic disease has led to significant advances in treatment. However, it ignores the critical differential influence of cell-extrinsic factors on a (cancer) cell's behavior. It should come as no surprise that an overly narrow focus on tumor genetics has not been able to provide effective treatments for a significant fraction of patients. A similar argument could be made about focusing exclusively on non-malignant components such as the immune system, which right now also provides benefit only for a subset of patients and tumor types. It is reasonable to expect that as our understanding of the intricate ecosystem of the TME increases, we will be able to design and recommend rational treatment combinations with clinical benefit in a larger group.

The TME is composed of the tumor itself in a context of cellular and non-cellular elements. The cellular fraction consists of resident and recruited immune cells (e.g., T and B lymphocytes, monocytes, macrophages, NK cells) and stromal cells (e.g., cancer-associated fibroblasts, endothelial cells, pericytes, adipocytes, mesenchymal stem cells). The non-cellular components consist of chemicals secreted by the aforementioned cell types (e.g., cytokines/chemokines), as well as the extracellular matrix (ECM), local metabolites, and environmental conditions (acidosis, hypoxia) present in the surrounding tissue. The bidirectional interactions between these components determine the balance between a tumorigenic/antitumoral niche. Therefore, it comes as no surprise that the microenvironment of different organs and different tumor types can be remarkably different and highly variable, as is the microenvironment of primary tumors versus later metastases [11]. These multiple levels of variability, in part, help explain differences in response and resistance to treatment between patients (even with similar tumor types) and within patients over time [12].

The Immune Landscape of the Tumor Microenvironment

Despite this high degree of heterogeneity, general (albeit overly simplified) trends arise across multiple cancer types that are of interest for any scientist trying to study the TME. Here, we will attempt to summarize these basic rules of thumb. A more comprehensive and nuanced description is not the scope of this chapter and can be found elsewhere [6, 8, 13]. T cells have received most attention as it is clear that the proportion and functional status of T cells within the TME are major factors in determining tumor progression. CD8+ (cytotoxic) T cells are considered major effector cells that have the potential to kill tumor cells after mounting an adaptive immune response. It has been shown, however, that these T cells often develop a dysfunctional state over time that shares many similarities with the exhausted phenotype observed in chronic viral infections, characterized by high expression of inhibitory checkpoints (e.g., PD-1, CTLA-4, TIGIT, TIM-3, and LAG-3) and loss of proliferative capacity and effector functions, including the production of cytokines (e.g., IFN-γ, TNF-α). Modifying or reversing this dysfunctional state of exhaustion is one of the approaches that has shown effect in a subset of patients with checkpoint inhibitors, that is, monoclonal antibodies blocking PD-1, PD-L1, or CTLA-4, approved for different cancer types. The plasticity of the exhausted phenotype is still not fully understood and might be irreversible. It should be noted that while high CD8+ T cell infiltration is associated with improved overall survival in many tumor types (e.g., melanoma, colorectal, non-small cell lung, breast, and bladder cancer), it is actually associated with worse overall survival in renal cell carcinoma and prostate cancer, highlighting the complexity of studying the TME and the importance of studying the functional characteristics of cells present [5, 14, 15]. CD4+ (helper) T cells play a coordinating role in the adaptive immune response by secreting cytokine combinations that impact a wide range of immune cells and can help mount a coordinated antitumor response. A subset of CD4+ T cells, regulatory T cells (Treg), are known to promote a suppressed and tolerogenic immune environment that generally benefits tumor growth and infiltration. Attempts to selectively target and eliminate these cells have been made using metronomic (i.e., low-dose) cyclophosphamide with some success [16–21]. T cells recognize a target by the highly specific interaction between the T cell receptor (TCR) and a peptide antigen bound in the major histocompatibility complex (MHC) on other cells, including cancer cells. Cancer cells, however, have developed strategies, such as downregulation of MHC expression, to bypass this interaction and escape T cell killing.

Natural killer (NK) cells are innate lymphoid cells that can recognize aberrant cells and subsequently kill them without antigen-specific binding. They can specifically target cells that lack the MHC complex and therefore might have an important role in supporting or initiating the anticancer immune response [22]. Strategies to

enhance or leverage T/NK cell function (e.g., with bispecific antibodies) and cellular therapies where (genetically modified) T or NK cells (e.g., chimeric antigen receptor (CAR) cells) are infused have been approved in multiple hematological malignancies and are currently being studied across the full cancer spectrum [23–26].

CD8+ T cells recognize cognate antigens in the context of MHC-I which is present on all nucleated cells of the body and presents cytoplasmic peptides. Some immune cells, including macrophages and dendritic cells (DCs), have acquired a role as professional antigen-presenting cells (APCs) that take up antigens from their environment via endocytosis and present them in the context of MHC-II to CD4+ T cells. Notably, some DC have developed *cross-presentation* which allows them to transfer antigens to the cytoplasm after endocytosis and present them in the context of MHC-I to CD8+ cytotoxic T cells. Given the crucial effector role of these T cells in tumor killing, efforts have been made to use DC to help mount an effective immune response, for example, in cancer vaccine and immunotherapy studies. Differences in composition and state of APC between tissues that are frequently exposed to environmental microbial antigens (e.g., lung, gut, skin) and tissues that are typically sterile (e.g., pancreas, brain) are potentially driving the variations in T cell infiltration across diverse cancer types [11].

Other cells have also been implicated in the cancer-immune cycle, but their role is not as straightforward. Cells of the myeloid lineage, that is, monocytes and macrophages, play a dual role in tumor-immune control. Classical CD14+ monocytes are recruited to the TME where they differentiate into tumor-associated macrophages (TAMs). Macrophages are innate immune cells with a phagocytic potential that could help kill tumors. In general, an important distinction has been made between M1 (classical, proinflammatory) and M2 (alternative, immunosuppressive) macrophages [27]. These two classes have a markedly different expression profile with the M1 macrophages often being characterized tumoricidal versus the tumorigenic M2 macrophages. It has been shown that a polarization of the macrophage phenotype occurs in the TME of different cancer types, contributing to tumor progression, although studying the diversity of macrophage states in vivo is an ongoing effort and phenotypes probably exist on a more continuous spectrum. This is further complicated by the existence and contribution of tissue-resident macrophages (i.e., not derived from monocytes) that have self-renewal characteristics independent of typical hematopoiesis and are important in tissue homeostasis. Other myeloid cells, for example, neutrophils and eosinophils, are also being considered as contributors to the cancer immune cycle, although their role is currently even less clear [28, 29]. A special group of immature myeloid cells, myeloid-derived suppressor cells (MDSCs), expands in the context of chronic inflammation and cancer. These cells are strongly immunosuppressive, and their presence is associated with poor prognosis and resistance to cancer treatment [30]. Subsets of MDSC with more monocytic characteristics (M-MDSC) versus more polymorphonuclear or granulocytic characteristics (PMN-MDSC) have been described.

The Stromal Compartment in the Tumor Microenvironment

Stromal factors are being implicated as well and they interact with both the tumor and with infiltrating immune cells [11, 31]. Endothelial cells (ECs) that make up the lining of blood vessels are highly variable across anatomic sites to allow for differences in permeability or adhesion that determine biological function. In the TME, new blood vessels are formed that are characterized by structural abnormalities and generally increased permeability. The formation of such a tumor-associated vasculature also occurs in a tissue-specific manner, reflecting the local microenvironment's signals and intrinsic differences between EC at different anatomic locations. Notably, the established tumor vasculature then has a profound impact on the subsequent immune infiltration as it controls homing and extravasation of specific leukocytes subsets at the TME. Direct immunosuppressive interaction of

EC with infiltrating immune cells (e.g., via expression of the FAS ligand) also contributes to modulation of an effective immune response [32]. Cancer-associated fibroblasts (CAFs) are mesenchymal cells that contribute to tumor immunity indirectly by modulating vessel permeability and depositing matrix components, both of which can act as a physical barrier for attracted immune cells. They also directly act on the immune system by the secretion of chemokines and cytokines that hold back immune system components with antitumor properties while promoting a more immunosuppressive environment. The deposition and organization of the extracellular matrix by mesenchymal cells also has impact on the permeability of the TME to infiltrating immune cells, mainly in solid tumors. Recently, the presence of nerve fibers within the TME has been associated with adverse prognosis in different tumor types, both due to a direct supportive effect on the tumor via release of neurotransmitters, as well as through indirect effect on angiogenesis and interaction with immune cells [33, 34]. Although the full heterogeneity of stromal components needs further investigation, these examples all support a common concept in which the tissue-specific stroma of the TME has an impact on tumor growth directly, as well as through its contribution to a particular immune microenvironment.

Qualitative and Quantitative Description of the TME

In terms of the immune microenvironment, attempts have been made to come up with a classification system that somewhat reliably reflects the presence of the aforementioned cell types [6, 35]. Tumors characterized by high infiltration of cytotoxic T cells across the entire tumor are often referred to as immunologically "hot" or inflamed, whereas a TME broadly populated with immune cells but with a sparsity of cytotoxic T cells in the tumor core itself has been described as immune-excluded or "cold." In the latter case, cytotoxic T cells are often found restricted to the periphery of the tumor by putative mechanisms that involve fibroblasts and macrophages, but that remain to be clearly elucidated. Notably, tumor specimens that lack appreciable infiltration by cytotoxic T cells are often referred to as immune deserts. Within each of these major categories, more specific cancer-immune phenotypes can be delineated, each having its unique underlying pathophysiology with consequences for potential treatment strategies. For example, a subset of the inflamed tumors histologically develop tertiary lymphoid structures (TLSs). These structured aggregates of lymphoid cells recapitulate the organization of lymph nodes and can act as preferential sites of immune activation and recruitment of adaptive anticancer immunity via antigen-presenting DC. Different classification and scoring systems exist, but their exact prognostic significance and/or therapeutic relevance is currently still limited to specific clinical contexts. The Immunoscore [14, 15] quantifies the CD8+ T cell infiltrate in the TME and was demonstrated to be superior to traditional TNM (tumor-node-metastasis) staging in patients with colorectal cancer. The effort to come up with a widely applicable system across cancer types including both solid tumors and hematologic malignancies will undoubtedly benefit the undergoing characterization efforts.

Therapeutic Implications of the TME

Understanding the full scope of cancer's complexity will be relevant to provide significant therapeutic improvement for those patients with the direst prognosis. Indeed, it is now increasingly clear that the TME characteristics influence response to cancer therapy. Likewise, cancer therapy can shape the TME [5, 16], for example, chemotherapy and radiotherapy are known to recruit immune cells and trigger their activation and maturation in the TME but also to recruit bone marrow-derived mesenchymal stem cells (MSCs) that produce chemoprotective factors and anti-angiogenic agents result in the attraction of circulating endothelial cells and the induction of tumor-supportive fibroblasts. This reciprocal interaction opens opportunities for researches to

use their increased understanding to improve therapeutic outcome by explicitly targeting TME components [36]. The stromal and immune cells in the TME are genetically stable, arguably making them a more attractive target for drugs than cancer cells (i.e., less susceptible to developing resistance mechanisms through evolution). Strengthening interactions between the tumor and the TME that keep tumor growth in check while at the same time blocking interactions that support tumor growth is a promising approach.

Clearly, targeting the TME is a potentially valuable strategy for the design of drugs and drug combinations, although many challenges and bottlenecks remain that have made this effort particularly challenging [36, 37]. Animal studies play an important role in preclinical research but fail to accurately mimic the interactions with the TME, especially when tumors are implanted ectopically. Targeting the TME may be associated with higher and unintended toxicities as drugs distort the normal homeostatic balance and disrupt a complex network of signal transduction. It is also conceivable that tumors could develop strategies to evade the disruption of a single microenvironmental factor which would lead to resistance. Also, the mechanisms of resistance to therapies targeting the TME are not completely understood. Furthermore, it has been suggested that the traditional paradigm of determining the maximal tolerated dose (MTD) as is common in clinical oncological trials might not be suitable for treatments targeting the TME where it might be more relevant to determine the optimal biological dose (OBD, i.e., the lowest dose that achieves maximal efficacy). This has also been suggested by the paradigm of *metronomic chemotherapy* showing an effect of cyclophosphamide on the immune microenvironment and endothelial cells at doses much lower than the MTD [17–20]. An important obstacle is the lack of clear biomarkers that can be assessed at baseline and predict response [5]. As a first step, attempts to classify different tumor immune microenvironments have been proposed to predict and guide immunotherapeutic responsiveness [6]. Finally, as our arsenal of cancer therapeutics expands, we need to think about treatment combination and sequencing in order to achieve cure or the most durable response with optimal quality of life [38].

This effort requires researchers to be able to capture and characterize the many different levels of heterogeneity: within a tumor (tumor cell genetic/epigenetic + local microenvironment differences), between different lesions in patients with metastatic disease and, importantly, between different patients [39]. Studying cancer cell genetics in isolation merely touches upon a superficial layer of complexity. Although it is relevant in determining drug responses to targeted therapies, genetic features alone cannot fully characterize the dynamic behavior of cancer cells, underscoring the importance of integrating other omics data.

From Bulk to Single-Cell Omics Data Analysis

Omics data analysis allows the unbiased assessment of different biological modalities and provides a rich and comprehensive picture in any biological context. Whole genome sequencing (WGS) and whole exome sequencing (WES) comprehensively capture the DNA of a sample (genome/exome), whereas RNA sequencing (RNA-seq) captures the transcriptional state (transcriptome) [40]. Other assays have been developed that allow characterizing the chemical state of DNA/RNA (epigenetics), proteins, and metabolites, respectively. The rapidly decreasing cost of high-throughput sequencing and development of novel massively parallel technologies now allow to study the genome, epigenome, transcriptome, proteome, metabolome, and other proposed *omics* modalities. Traditionally, profiling was performed on a grouped collection of cells from a particular tissue sample, in so-called *bulk* analysis. This essentially results in the registration of the average genome and/or the average expression level across a large population of cells and fails to trace back expression to individual cells, the fundamental biological unit. Despite this limitation, RNA-seq has been particularly useful for comparative transcriptomics, and

robust analytical tools have been developed to conduct differential gene expression [41–43].

The blended nature of RNA-seq data led to the development of tools that use bulk data to estimate relative cell composition, either through deconvolution [44–47] or gene set enrichment analysis (GSEA) [48], which has been used to determine the composition of tumor-infiltrating immune cells semi-quantitatively [49]. GSEA-based methods compute the enrichment score (ES) for a set of genes, for example, a gene set that is characteristically expressed in a particular cell type. The ES is high when the particular set of genes is overrepresented among the top highly expressed transcripts in a sample, which suggests the cell type is enriched in the sample. Deconvolution algorithms use a signature matrix of cell-type-specific expression values to quantitatively reconstruct the contributions of the different cell types to the heterogeneous sample [49]. Many computational tools have been developed to characterize other aspects of (potential) tumor-immune cell interactions [50, 51]. Bulk RNA-seq data of the TME has been used to evaluate diversity and heterogeneity of the antigen receptor repertoire of B and T lymphocytes (MiXCR [52, 53], TRUST [54]) that can be used on bulk RNA-seq data of the TME. Similarly, RNA-seq has been used (in combination with genomics data) to determine the presence of cancer neoantigens in patients as a result of mutations, deletions, insertions, alternative splicing, and gene fusion events. When combined with (genomics-based) typing of the human leukocyte antigen (HLA) locus, this technique can be applied to identify cancer neoantigen peptides that are likely to bind to a patient's antigen receptor, elicit an adaptive immune response, and can therefore be used for rational cancer vaccine development [55].

As mentioned before, a disadvantage of studying bulk populations is the fact that it blends out and potentially masks signals that are present in individual cell types. Despite the aforementioned tools, bulk sequencing is not sufficient to study heterogeneous biological system that contains multiple (uncommon) cell types and does not fully characterize the stochastic nature of gene expression. In practice, the diversity of individual cells in the TME exceeds what can be measured by merely studying a mixture of these cells. Furthermore, the genomes of individual cells are not always the same, especially in cancer. Studying any modality at single-cell resolution, in contrast, allows identification of rare cell types that potentially drive cancer progression, invasion or response, and whose transcriptomic/epigenetic/etc. signal would be averaged out and lost in bulk analyses.

Single-Cell RNA Sequencing: The Paradigm of Single-Cell Technology

Single-cell RNA sequencing (scRNA-seq) has been the frontrunner in the transition to single-cell analysis with the first single-cell transcriptome published around 2009 [56]. It has rapidly become more widespread available with commercial platforms allowing characterization of up to thousands of genes in more than 10,000 cells in a single experiment [57, 58]. Cells need to be dissociated from tissue into a single-cell suspension in a process that can have a considerable impact on the molecular profile and relative cell-type abundance, especially in the case of tissue from solid tumors. A reverse transcription and amplification process, conceptually similar to that of bulk RNA-seq, can be achieved by different experimental approaches, resulting in different strengths and weaknesses [59–62]. There are two main technological strategies for RNA capture: microfluidic (including droplet-based systems) and microwell (plate-based). Microdroplet-based technologies capture individual cells in lipid droplets together with barcoded beads on which RNA capture and barcoded reverse transcription occur. They generally allow for the analysis of a large number of cells (thousands) (i.e., highest throughput) but with a more restricted limited number of reads per cell (depth). Microwell-based protocols, on the other hand, are more limited in terms of the number of cells that can be analyzed (up to hundreds) but do provide much deeper sequencing of full-length

transcripts and methods that can be combined effectively with cell sorting techniques, for example, to select cells of interest based on surface markers. Other microfluidic platforms bring a more integrated system to the table, which automates the reactions necessary for library preparation and generally provides an intermediary throughput. Arguably more important is the distinction between full-length and tag-based technologies. Full-length platforms (e.g., C1/Smart-seq [63], MATQ-seq [64], Smart-seq2 [65, 66]) try to achieve a uniform read coverage over the whole transcript, whereas tag-based protocols (e.g., CEL-seq2 [67], Chromium [68], ddSEQ [69], DroNc-seq [70], Drop-seq [71], inDrop [72], MARS-seq [73], Nx1-seq [74], Quartz-Seq2 [75], Seq-Well [76], STRT-seq [77]) only capture the 3′ or 5′ end of the transcript, limiting its use for the characterization of splice variants, RNA editing, or detection of mutations. By incorporating unique molecular identifiers (UMIs) that act as barcodes added before PCR amplification, these technologies can reduce errors and amplification bias. Although most platforms focus exclusively on capturing and studying mRNA, methods to sequence RNA more broadly (i.e., both polyadenylated and non-polyadenylated) have been proposed (e.g., SUPeR-seq [78], RamDa-seq [79], Small-seq [80], and Smart-seq-total [81]). Depending on the exact platform used, researchers should be especially mindful of potential problems including limited cell capture, biases in cell survival, reverse transcription efficiency, and cDNA amplification that affect distinct protocols differently. Head-to-head comparisons have elucidated many of these weaknesses, stressing the importance of orthogonal validation as these technologies continue to mature [57, 82, 83].

The Analysis of Single-Cell RNA Sequencing Data

Data from scRNA-seq platforms are generally noisier, and its downstream analysis is considered to be more challenging than bulk RNA-seq. The relative absence of computational standards for analysis and interpretation after library generation and sequencing limits reproducibility. In terms of computational data analysis tools, bulk RNA-seq methods (for differential gene expression, regulatory network inference, etc.) have been applied to scRNA-seq datasets with variable success [84]. The presence of unique technical noise and extensive biological variability (including stochastic transcription), however, raises the question whether this approach leads to meaningful and optimal results. Recently, a growing number of tools specifically aimed at scRNA-seq datasets have come forward, although they each come with their (dis)advantages and proper head-to-head comparisons are rare, which makes method/parameter selection and reproducibility problematic [84, 85]. As the number of available tools grows rapidly, it becomes increasingly difficult to navigate the full spectrum of methods and generate an up-to-date and reproducible workflow. Several groups have begun to conduct cross-platform benchmarking studies to help address the critical problem [86].

A full review of the bioinformatics pipelines and applications available for analyzing scRNA-seq data is outside the scope of this chapter, and excellent reviews exist that explain specific challenges and current best practices [84, 87]. We do, however, want to highlight some aspects that are important in the context of studying the TME. Whereas data analysis initially required considerable computational experience, research groups and companies that sell hardware and reagents for scRNA-seq have made efforts to release software and packages that provide tools for integrated quality control, dimensionality reduction, visualization, and data analysis, often with minimal parameter tuning or coding requirements. Researchers should carefully review the methods, including the underlying algorithms and parameters to understand the introduction of potential bias, especially in the absence of a gold-standard analysis platform. The most widely used packages are probably Seurat [88–90] in R, and Scanpy [91] in Python, the two most prominently used programming languages in the field. Recently, much effort has been invested in establishing proper cross-environment support.

Quality Control and Preparation of Single-Cell RNA Sequencing Data

Before downstream biological analysis, scRNA-seq data generally undergoes a series of quality control (QC) checks. This is particularly relevant when studying heterogeneous samples taken from the TME. Poor quality from single cells can be the result of poor cell viability, limited mRNA recovery, ambient RNA, poor cDNA production, etc. There is no clear consensus, and the best filtering strategy often relies on the tissue type and sequencing platform used. Library size, number of (housekeeping) genes detected per cell, and fraction of mitochondrial-encoded mRNA molecules are most commonly used, but thresholds vary widely and using them in isolation is especially problematic when studying the diverse set of cell types and metabolic states present in the TME. More integrated computational tools have been proposed to manage QC [92, 93] and remove contamination with ambient RNA (e.g., SoupX [94] and DecontX [95]). QC also involves detection and removal of doublets, for which several tools are available [96–101]. A recent benchmarking analysis [102] suggested that DoubletFinder [98] excels in detection accuracy, whereas scds [100] provides the best computational efficiency. The efficacy of any QC approach is generally determined by judging the quality and plausibility of downstream analytical performance (e.g., cluster annotation and differential expression). In the context of studying the TME, there are often diverse cell populations with widely varying expression characteristics and viabilities that risk being removed as low-quality outliers. Therefore, in our experience, it is often valuable to start with a more permissive QC threshold and rerun a more stringent approach later, based on the initial results. This iterative strategy, however, introduces bias, and QC should not be exploited to improve downstream statistical tests.

When high-quality data are filtered, other preprocessing steps involve normalization, feature selection, and dimensionality reduction, which ultimately allow downstream cell- and gene-level analyses. Bulk expression methods have been attempted, but the specific sources of variation and noise has led to development of single-cell-specific methods. Whereas normalization attempts to remove some of the effects of sampling individual mRNA molecules, it cannot account for all technical and biological covariates that might still be present (including batch, dropout, and cell cycle effects). This can be done using a simple linear regression as well as more complicated mixture models (e.g., f-scLVM [103, 104]) but should be considered carefully, since correcting for biological covariates can obscure interpretability and might unintentionally remove other relevant signals. For technical covariates, batch correction methods such as ComBat [105] have been shown to be applicable for scRNA-seq data [106]. A related problem is the integration of data from different experiments which poses additional challenges as the cell-type composition might not be identical. Multiple methods including the Seurat implementation using canonical correlation analysis (CCA) [88], LIGER [107], and Harmony [108] have been released, although head-to-head comparison is limited.

Dropouts have been a particular challenge of scRNA-seq analysis. They refer to zero counts that are the result of sampling only a fraction of the mRNA present in any single cell (i.e., technical dropouts) but also can be caused by biological phenomena, such as transcriptional bursting (i.e., pulses of transcriptional activity are followed by inactive periods in which mRNA cannot be detected). These zero counts result in a unique data distribution that is different from bulk RNA-seq. Attempts to recover these (technical) dropouts (referred to as denoising or data imputation) have been proposed with tools such as scVI [109], scImpute [110], and DCA [111]. Although these methods generally result in appealing data (and visualizations), we want to emphasize that there are obvious risks of under- or overcorrection which could result in spurious correlations, as well as lack of reproducibility. In light of the current lack of standards, we argue that denoising or imputation should be clearly disclosed and used with caution.

Clustering and Compositional Analysis

These preparatory steps lead into cell- and gene-level analyses that are the highlight of scRNA-seq data. A first step is splitting the cells into biologically meaningful clusters. The standard approach (implemented in most packages including Seurat and Scanpy) implements the Louvain algorithm [112] to identify communities on a single-cell KNN (K-Nearest Neighbors) graph. Careful cell-type annotation can be achieved by manual inspection of signature *marker genes* of each cluster. Notably, the resolution of clustering (i.e., the number of clusters) can be changed to alter the granularity. While this can be particularly useful for sub-clustering major cell types present in the TME, it is important to realize that the clusters identified using an unsupervised approach do not necessarily coincide with biologically valid cell types. Furthermore, the definition of what constitutes a cell type also depends on the type of experiment and the context, as cell types in different developmental or metabolic stages often appear as separate clusters [113]. For example, when considering T cells in the TME, it might be sufficient to classify them into CD4+ and CD8+ T cells for some biological context, whereas in other situations it might be necessary to divide these "major" populations into subpopulations representing developmental (e.g., naïve versus memory) or functional (e.g., effector versus exhausted) states. It is conceivable that these cell states represent a transcriptional profile change that is continuous to some extent, making clustering (i.e., splitting them into distinct groups) somewhat arbitrary. This realization has been the inspiration for so-called trajectory analysis methods that explicitly try to capture transitions between cell identities. Tools such as Monocle [114] and Wanderlust [115] have first established the feasibility and have sparked off a set of other methods that have reviewed comprehensively [116], suggesting the use of Slingshot [117] when expecting simple trajectories and PAGA [118] when dealing with complex bifurcations, although ultimately the best algorithm varies depending on the underlying data.

The reliance on manual inspection for cell-type annotation has prompted the development of reference databases (e.g., the Human Cell Atlas [119]) that can be used to provide guidance. Alternatively, researchers rely on prior knowledge and literature to come up with marker genes, although this approach does not take into account the distinctive nature of single-cell data. Importantly, there is a weak or no correlation between mRNA expression and protein abundance on the cell surface, restricting the use of traditional cytometry-based phenotyping markers, for example, when identifying immune cells in the TME. To limit bias, methods for automated cluster annotation have been presented, including Garnett [120], scmap [121], scMatch [122], SingleCellNet [123], and singleR [124]. Different tools were recently evaluated [125] showing the advantages and limitations of this approach. Whereas a manual annotation relies on prior knowledge and is intrinsically biased, automated cluster annotation offers speed, simplicity, and flexibility. When studying large and complex datasets such as those derived from TME samples, however, reference atlases often do not properly represent the full composition of cell types and states. It is therefore advisable to use a combination of both strategies (e.g., start with a "rough" automated cell-type annotation and complement this with manual sub-clustering).

When cell types and states are determined and annotated, compositional analysis (i.e., compare proportions of different cell types and states between samples of different patients or in response to disease/treatment/etc.) is a logical next step. Many analysis tools developed for (mass) cytometry data are being adapted for use on scRNA-seq datasets. It is important to understand that analysis of TME samples usually is associated with a large variation of input material so that it is typical not meaningful to look at absolute cell number differences when comparing samples. Instead, statistical tests focus on relative abundances (i.e., percentages of total cells), but of course tests for different cell types in the same data set are not independent. This is an important challenge that complicates interpretation of such changes.

Whereas previously we used gene-level data to describe and understand the cellular heterogeneity, now we can leverage this context for a deeper dissection of the data and comparison of different experimental conditions. This includes analysis like differential expression testing, gene set analysis, and gene regulatory network inference, originally developed and used on bulk gene expression data. The opportunity to discriminate the expression changes in different cell types separately is one of the main arguments to prefer single-cell transcriptomics over a bulk approach in analysis of the TME. Because of the unique nature of scRNA-seq data, novel differential expression (DE) methods were developed that take into account dropouts and other sources of cell-cell variability that are not typically present in bulk RNA-seq data [126, 127]. Somewhat surprisingly, a recent comparison of methods [128] found single-cell-specific DE methods provide no advantage over their bulk-derived counterparts, with the latter even achieving superior results when adapted using a gene-weighting strategy to properly model single-cell data [129]. This suggests that the widely used analysis tools EdgeR [130] and DESeq2 [41] can be adapted and used on scRNA-seq data. However, because bulk DE tools were originally designed for comparing a limited number of samples (instead of thousands of individual cells), this approach can quickly become computationally prohibitive. In this case, it is advisable to use either MAST [127], a single-cell DE method shown to achieve reliable results [86] or the limma-voom pipeline which optimizes computational speed [131]. To interpret differentially expressed genes, it is advisable to summarize the collection of significant genes and search for overrepresented sets of genes based on biological process. Databases used in bulk transcriptomic data (e.g., GO, MSigDB, KEGG, Reactome) can be readily queried using the gene lists from scRNA-seq data. At a higher level, it is possible to construct gene regulatory networks by studying co-expression of transcripts as a proxy for causal relationships. Tools such as SCODE [132] and SCENIC [133] were designed for single-cell transcriptomics data specifically, but they were shown to have a relatively poor performance (as did bulk methods) in a recent benchmark study [134]. Until more data becomes available, these methods should be applied with caution and their results interpreted prudently.

Other Dimensions of Single-Cell Profiling in the Tumor Microenvironment

Single-cell RNA sequencing has cleared the path for other high-throughput omics technologies at the single-cell level, each of which contributes to a comprehensive and multifaceted snapshot of the TME. The data structures of these omics technologies and their associated analytical challenges share many similarities with scRNA-seq and will not be discussed in detail. Instead, we focus on what each of these data types can bring to the table to increase our understanding of the TME.

Single-Cell DNA Sequencing

Single-cell DNA sequencing (scDNA-seq) allows investigation of genomic heterogeneity at single-cell resolution. It allows studying single nucleotide variants (SNVs), copy number alterations (CNAs), as well as more complex chromosomal rearrangements [135]. This in turn provides an essential contribution to reliably characterize tumor clonality and evolution. Understanding the clonal make-up of a tumor (and its metastatic lesions) has gained particular interest in the context of personalized or precision oncology, with the expectation that a combination of (targeted) therapies aimed at the individual clones present might achieve better clinical outcomes. Single-cell whole-genome sequencing is achieved with different methods including MDA [136], DOP-PCR [137, 138], MALBAC [139], and TnBC [140], although other methods are available [135]. Over the last decade, scDNA-seq has increasingly been applied to characterize genetic heterogeneity within tumors, on circu-

lating tumor cells and metastases to better understand the evolution and mutation rate as well as development of resistance to therapy [137, 141–148]. The analysis of scDNA-seq data is complicated by problems of non-uniform coverage, mutation bias, doublet detection, and the risk of allelic dropouts, all leading to false positive/negative results [135, 149, 150]. Whole-genome coverage is often not essential when studying a particular tumor type with a known set of commonly mutated genes. To delineate the clonal architecture and determine CNAs, it might be sufficient and more efficient to target a limited panel of genes as has been proposed with the Tapestri platform [151, 152].

Single-Cell Epigenomics

Single-cell epigenomics provides an extra perspective on the activity of genes in individual cells [153]. When profiling tumor cells, it can be used to detect genes or chromosomal regions that have been silenced through hypermethylation. Various cancer drugs are known to target epigenetic regulators that result in altered histone modification patterns. However, epigenetic profiling at single-cell resolution is particularly important for characterizing functional capacity present among different (immune) cells. Its use in the context of the TME has mostly focused on T cell differentiation. It is well understood that T cell development is characterized by profound epigenetic changes, some of which are deemed irreversible. This has been of particular interest when studying CD8+ T cell dysfunction ("exhaustion") in the TME and its contribution to immunotherapy resistance. Multiple platforms exist that provide different information, including scATAC-seq [154] (assay for transposase-accessible chromatin; i.e., chromatin accessibility), scBS-seq [155] and scRRBS [156] (DNA methylation), scChIC-seq [157] (histone modifications), and scHi-C [158] (chromatin configuration).

Single-Cell Proteomics

Complementing genomic studies with protein expression at single-cell resolution is important as we know that the correlation between mRNA expression (as detected in scRNA-seq data sets) and protein abundance is limited. There is another clear benefit to detecting protein as this readout provides a more direct assessment of cell functionality. In contrast to single-cell genomics and transcriptomics, high-throughput single-cell proteomics (i.e., the detection of all proteins present at single-cell resolution) is not yet commercially viable or widely available, so measurements are limited to a prespecified panel of proteins of interest [159]. Nevertheless, single-cell protein detection is arguably not a novel technology. Flow cytometry has been available for decades to study surface and intracellular protein abundance at the level of individual cells. A limiting factor in terms of dimensionality has been the overlap in the spectrum of different fluorophores used, resulting in complex deconvolution methods for panels with more than a handful markers.

A workaround has been the development of mass cytometry (CyTOF) in which this problem is solved by conjugating antibodies with heavy metal isotopes that can be distinguished by mass spectrometry with minimal overlap [160–162]. This allowed using panels of up to about 50 antigens on millions of cells in a single experiment and has facilitated using the technology to identify novel (immune) cell subsets or detect rare cell populations based on a combination of protein markers. Note that the technology is not restricted to surface protein detection but has been adapted to measure intracellular markers and cell signaling processes [162]. CyTOF has been especially transformational for research characterizing the TME in both solid and hematologic malignancies and for immunophenotyping more generally in the context of immunotherapy. Limitations (in comparison to traditional flow cytometry) are the high fraction of cells lost during staining and acquisition, lower throughput, cost and availability of the technology, as well as the fact that no viable cells can be retained. This makes CyTOF an attractive

methodology for hypothesis generation and biomarker discovery that should be complemented with functional studies that confirm the phenotypic observations. Another more recent workaround of the dimensionality limitation is the use of spectral flow cytometry in which the full spectrum of emitted fluorescence is taken into account for each probe [163, 164]. The data can later be unmixed based on reference spectra and known autofluorescence. This offsets some of the disadvantages of CyTOF, although panel design is generally more complicated and requires careful selection of markers and appropriate fluorophores in order to acquire clean data with minimal artifacts.

The rapid increase of protein markers in cytometry experiments has complicated traditional analysis workflows depending on manual gating. This traditional approach is still useful and powerful when conducting supervised cell-type annotation. However, it is intrinsically biased and cannot be relied upon for discovery of novel and unknown phenotypic populations (as the number of binary phenotypes increases exponentially). This has sparked off the implementation of computational tools aimed at unsupervised clustering and visualization of multidimensional cytometry data, with algorithms used most prominently including SPADE [165], Phenograph [166], and FlowSOM [167]. The large number of cells (often millions) presents a computational challenge for most single-cell methods proposed for the analysis of scRNA-seq data. Packages that provide integrated workflows in R [168, 169] and Python [170] for cytometry data analysis are available.

Whereas protein detection methods for surface and intracellular markers are restricted by the limitations specified earlier, the detection of soluble proteins in plasma can be expanded to assess hundreds of proteins simultaneously, bringing the ultimate goal of characterizing the full plasma proteome closer. Technological methods exist based on either mass spectrometry or affinity proteomics platforms (i.e., the proximity extension assay (PEA) [171] or SOMAscan [172], commercialized by Olink and SomaLogic, respectively) [173]. Although the plasma protein repertoire obviously does not reflect the single-cell paradigm, it provides an extensive and high-level overview of the pathophysiological state of an organism. The array of circulating cytokines and chemokines in the blood can be queried to characterize the functional status of the immune system holistically and to discover biomarkers for response and resistance in all many different tumor types [174–176]. Additionally, for tumors that reside in the bone marrow (i.e., hematological malignancies such as multiple myeloma or myelodysplastic syndrome) or metastasize to the bone marrow niche, these technologies can measure said markers directly in the TME and contribute information on the ongoing crosstalk [177].

The Multi-Omics Paradigm

The full impact of single-cell omics will become more obvious as new technologies emerge that allow concurrent measurement of different modalities from the same cell. The simultaneous characterization of genotype, epigenetic status, transcriptional program, and phenotype of individual cells in the TME provides insights that would allow a more comprehensive perspective on the fundamental biological unit: the cell. Even integration of multiple omics data from different cells (from the same sample), for example, the combination of scRNA-seq with CyTOF can help to confirm new cellular phenotypes and better understand their transcriptional behavior at the same time as has been showcased successfully in different tumor types. Single-cell multimodal omics have attracted acclaim and attention across the entire biological spectrum which promoted Nature to name single-cell multimodal omics as its *method of the year 2019* [178].

An important aspect of measuring multi-omics in the TME is the consideration of the spatial context of cells. Single-cell experiments often rely on digesting (solid) tumors into soluble cell suspensions for subsequent analysis. Tumors, however, are complex 3D structures in which soluble mediators might indeed impact cells at a distance, although individual cells are affected mainly by interactions with their direct neigh-

bors. This underscores the essential role of imaging techniques to capture the spatial organization. Combining aforementioned omics technologies with the localization of tumor cells, non-tumor (immune, vascular, stromal) cells, and matrix components provides a unique window on the structural organization at a (sub)cellular resolution.

Methods focusing on protein detection in situ often build upon established imaging techniques such as immunohistochemistry (IHC) or immunofluorescence (IF). Multiplexed IHC by iterative staining of single slides (MICSSS) [179, 180] is the most widely used chromogen-based example. It offers the ability to stain slides for a handful of different antigens using a protocol that can be readily implemented by most pathology labs and is complemented with automated digital analysis tools to automatically map the TME. To expand the number of proteins that can be measured, novel antibody-based techniques including imaging mass cytometry (IMC) [181, 182] and multiplexed ion beam imaging (MIBI) [183] have been showcased, which combine spatial data collection with the dimensionality of mass cytometry (i.e., 20–40 markers) [184, 185]. These platforms are promising and have obvious potential to characterize the spatial heterogeneity present in the TME. There are other methods that use oligonucleotide-tagged antibodies to achieve multiplexed protein detection (of up to 50 markers), such as co-detection by indexing (CODEX) [186] and Digital Spatial Profiling (DSP) [187]. DSP also allows RNA detection of up to 100 genes. However, most of these methods are considered lower throughput due to the time necessary to acquire and process the images, precluding analysis of large surfaces and multiple slides per specimen. Given the fact that it is currently unknown whether single slides can provide an accurate representation for the whole tumor context, this limitation becomes even more relevant. IMC has been expanded to incorporate detection of RNA using metal-labeled oligonucleotides [188]. Specific computational methods have been developed to enable analysis of spatially resolved cytometry data (e.g., HistoCAT [189] and ImaCytE [190]).

Spatial transcriptomics methods are generally divided into two main categories. Targeted detection of a small panel of DNA/RNA molecules has been in use for years via fluorescence in situ hybridization (FISH). These FISH-based methods have been expanded to allow detection of a more expansive set of genes, but ultimately, these methods are limited by the molecular crowding leading to spatial overlap of the fluorescence signal and intrinsically biased. Recently, in situ hybridization has been modified through multiplexing and super-resolution imaging (e.g., seqFISH+ [191]) to achieve transcriptome-scale detection of mRNA. A second group of methods for spatial transcriptomics relies on single-cell sequencing to achieve unbiased coverage of the full transcriptome. In this case, tissues are still dissociated, so connecting the measured transcriptomes back to their original spatial context leads to unique challenges. Technologies using a barcoded oligonucleotide capture array have been developed and commercially implemented, albeit with limited resolution of around 100 micrometer [192]. Alternatively, a microdroplet-based method has been adapted to accommodate spatial information (i.e., Slide-seq [193]).

Incorporation of spatial information is only one example of a valuable addition to transcriptomics when studying the TME, but many other multimodal single-cell technologies are rapidly emerging. Most of these methods share (single-cell) RNA-seq as the common scaffold and combine this with additional information sources to systemically cover different aspects of biology as shown in Fig. 16.1. A short description of these methods, including their potential use in characterization of the TME and relevant references, is provided in Table 16.1 [191–227]. The crowded and quickly evolving field emphasizes the great excitement present in the scientific community but also make it difficult to navigate. In general, a trade-off exists between in-depth analysis of cells one by one, which often allows acquisition of a more comprehensive molecular profile (with often cleaner data) but suffers from low throughput and high cost per cell versus more scalable platforms that offer the advantage of profiling large numbers

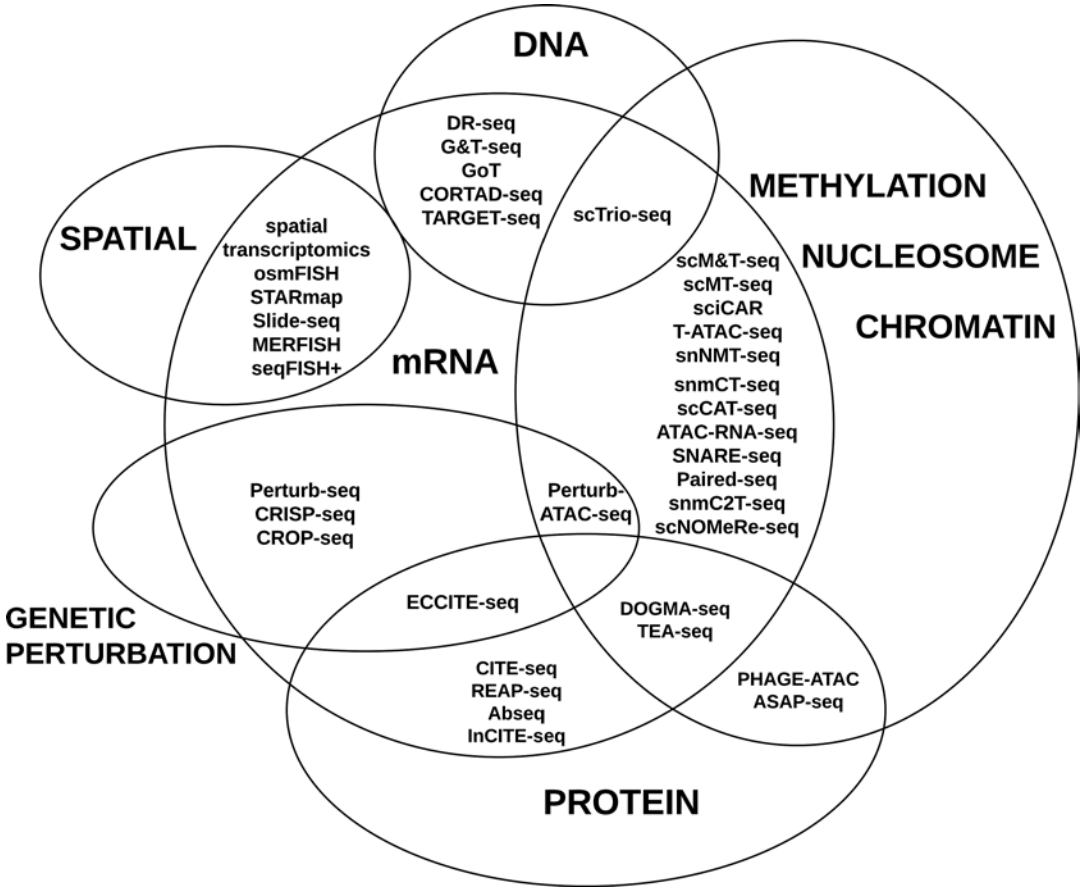

Fig. 16.1 Multimodal omics methods build on single-cell RNA sequencing as a common scaffold

of cells, which could prove advantageous in more heterogeneous contexts like the TME. The latter methods, however, tend to suffer from data sparsity (i.e., incomplete coverage, technical noise, high cell-to-cell variability) which complicates downstream interpretation. As technology evolves, it is now conceivable that, in the near future, we will have access to a platform that allows characterization of all the different levels of the central dogma of biology (DNA-RNA-protein), all at the level of individual cells. Bridging these different omics would offer a rich data source that could help elucidate complex regulatory mechanisms underlying cancer (and other diseases), for example, by providing important insights into the mechanisms related to tumorigenesis, metastasis, and (immune-mediated) response to treatment.

Multi-Omics Data Integration

Despite great promise, many of these technologies present with a great degree of experimental and analytical complexity. Standardization of technologies and computational tools therefore remains a concern. Initially, different data types or data from different sources were typically processed and analyzed separately and correlated at a later stage. This method is highly flexible and allows relying on existing methods developed for that specific data type, but it can also introduce biases or systematic errors and often fails to acknowledge the dependencies between different data types of the same cell. Methods for data integration are particularly critical to gain the most valuable insights from multimodal data [228, 229]. There is a growing list of recent tools and

Table 16.1 Selection of current experimental methods for multimodal single-cell measurements

Technology	Full name or description	Genome	Epig-enome	Tran-scriptome	Pro-teome	Pertur-bation	Spatial	Scalability	Preprint	Reference	Notes	Potential use(s) in characterization of the TME
DR-seq	gDNA-mRNA sequencing	x		x				Limited		Dey et al. Nat Biotech 2015	Amplification of genomic DNA and mRNA without prior separation, subsequent modified CEL-seq and modified MALBAC	Identify causative genetic variations for variable expression in tumors or non-tumor cells Determine accurate tumor clonality linked to expression
G&T-seq	Genome and transcriptome sequencing	x		x				Limited		Macaulay et al. Nat Methods 2015	Full-length transcriptome analysis using a modified Smart-seq2 protocol and separate whole-genome amplification	
GoT	Genotyping of transcriptomes	x*		x				Yes		Nam et al. Nature 2019	Adds genotyping information to single-cell transcriptomics for limited number of genes and/or specific genetic alterations	
CORTAD-seq	Concurrent sequencing of the transcriptome and targeted genomic regions	x*		x				Limited		Kong et al. Clin Chem 2019	Concurrent evaluation of the transcriptome and targeted genomic features within the same single cell on the Fluidigm C1 platform	
TARGET-seq	High-sensitivity genomic DNA and cDNA genotyping with scRNA-seq	x*		x				Limited		Rodriguez-Meira et al. Mol Cell 2019	Method for the high-sensitivity detection of multiple mutations within single cells from both genomic and coding DNA, in parallel with unbiased whole-transcriptome analysis	

scM&T-seq	Single-cell methylome and transcriptome sequencing		a	x				Limited		Angermueller et al. Nat Methods 2016	Modification of the G&T-seq protocol adding methylome sequencing using scBS-seq	Connect methylated elements and chromatin accessibility with variable gene expression in immune cells, e.g., in T-cell exhaustion. Uncover functional regulators of dynamic (immune) cell states, e.g., T-cell differentiation
scM&T-seq	Single-cell methylome and transcriptome sequencing		a	x				Limited		Hu et al. Genome Biol 2016	Transcriptional sequencing via Smart-seq2 and methylome sequencing via modified scRRBS	
scTrio-seq	Single-cell triple omics sequencing	CNV	a	x				Limited		Hou et al. Cell Res 2016	Simultaneous analysis of copy number variations (CNV) and methylome (modified scRRBS), as well as transcriptome of a single mammalian cell	
sci-CAR	Single-cell combinatorial indexing of chromatin accessibility and mRNA		b	x				Yes		Cao et al. Science 2018	Pooled barcode method that jointly analyzes RNA transcripts and chromatin accessibility	
T-ATAC-seq	Transcript-indexed ATAC-seq		b	x*				Limited		Satpathy et al. Nat Med 2018	Sequencing of TCR-encoding gene mRNA in combination with ATAC-seq at single-cell level	

(continued)

Table 16.1 (continued)

Technology	Full name or description	Genome	Epig-enome	Tran-scriptome	Pro-teome	Pertur-bation	Spatial	Scalability	Preprint	Reference	Notes	Potential use(s) in characterization of the TME
snNMT-seq	Single-cell nucleosome, methylation, and transcription sequencing		a,c	x				Limited		Clark et al. Nat Commun 2018	NOME-seq adaptation, transcriptome via Smart-seq2 protocol; methylation and chromatin accessibility are separated bioinformatically	
snmCT-seq	Single-nucleus methylcytosine and transcriptome sequencing		a	x				Yes	x	Luo et al. bioRxiv 2018	Joint capture of cytosine DNA methylome (5mC) and transcriptome profiles (based on Smar-seq2) from single cells/nuclei requiring no physical separation of RNA and DNA	
scCAT-seq	Single-cell chromatin accessibility and transcriptome sequencing		b	x				Limited		Liu et al. Nat Commun 2019	Integration of scATAC-seq and scRNA-seq using modified Smart-seq2 protocol	
ATAC-RNA-seq	Combined ATAC sequencing and RNA sequencing		b	x				Limited		Reyes et al. Adv Biosystems 2019	Integration of chromatin accessibility with bulk tagmentation and scRNA-seq using modified Smart-seq2 protocol	
SNARE-seq	Single-nucleus chromatin accessibility and mRNA expression sequencing		b	x				Yes		Chen et al. Nat Biotech 2019	Highly parallel profiling of chromatin accessibility and mRNA from individual nuclei, implemented on a micro-droplet platform	

Paired-seq	Parallel analysis of individual cells for RNA expression and DNA accessibility by sequencing		b	x				Yes		Zhu et al. Nat Struct Mol Biol 2019	Ligation-based combinatorial indexing strategy to simultaneously tag both open chromatin fragments generated by transposase and cDNA molecules generated from reverse transcription (RT) of RNA in millions of cells	
snmC2T-seq	Single-nucleus methylcytosine, chromatin accessibility, and transcriptome sequencing		a,b	x				Yes	x	Luo et al. bioRxiv 2019	snmCT-seq with chromatin accessibility, method based on scNOMe-seq	
scNOMeRe-seq	Single-cell nucleosome occupancy, methylome and RNA expression sequencing		a,c	x				Limited		Wang et al. Nat Commun 2021	Combination of scNOMe-seq with MATQ-seq (multiple annealing and dC-tailing-based quantitative single-cell RNA sequencing)	
ORCA	Optical reconstruction of chromatin architecture		d	x*			x	Limited		Mateo et al. Nature 2019	In situ detection of DNA folding and selected gene expression (RNA-FISH) in single cells	

(continued)

Table 16.1 (continued)

Technology	Full name or description	Genome	Epig-enome	Tran-scriptome	Pro-teome	Pertur-bation	Spatial	Scalability	Preprint	Reference	Notes	Potential use(s) in characterization of the TME
CITE-seq	Cellular indexing of transcriptomes and epitopes			x	x*			Yes		Stoeckius et al. Nat Methods 2017	Cell surface protein detection with barcoded oligonucleotide linked to antibody	Accurate cell type and state annotation for immune and non-immune cell subsets in TME Direct read-out of surface protein levels that are potential targets of immunotherapy (e.g., checkpoint inhibitors and other monoclonal antibodies) or direct readout of intracellular proteins (e.g., transcription factors) in combination with gene expression
REAP-seq	RNA expression and protein sequencing assay			x	x*			Yes		Peterson et al. Nat Biotech 2017	Cell surface protein detection with barcoded oligonucleotide linked to antibody	
Abseq				x	x*			Yes		Shahi et al. Sci Rep 2017	Antibodies to detect epitopes of interest labeled with sequence tags that can be read out with microfluidic barcoding and DNA sequencing	
inCITE-seq	Intranuclear cellular indexing of transcriptomes and epitopes			x	x*			Yes	x	Chung et al. bioRxiv 2021	Measuring multiplexed intranuclear protein levels and the transcriptome in parallel in thousands of cells	
PHAGE-ATAC			b		x*			Yes	x	Fiskin et al. bioRxiv 2020	Uses engineered nanobody-displaying phages for simultaneous single-cell measurement of surface proteins, chromatin accessibility profiles, and mtDNA-based clonal tracing through a massively parallel droplet-based ATAC-seq assay	Technologies are emerging that capture measurements of gene activity ranging from chromatin accessibility over mRNA expression to protein levels, allowing a more comprehensive reconstruction of regulatory networks at single-cell level

ASAP-seq	ATAC with select antigen profiling by sequencing		b		x*			Yes		Mimitou et al. Nat Biotech 2021	Pairs sparse scATAC-seq data with robust detection of hundreds of cell surface and intracellular protein markers (and optional capture of mitochondrial DNA for clonal tracking)	
DOGMA-seq			b	x	x*			Yes		Mimitou et al. Nat Biotech 2021	A variant of CITE-seq, allowing co-measurement of chromatin accessibility, gene expression, and protein from the same cells	
TEA-seq	Trimodal assay that simultaneously measures transcriptomics, epitopes, and chromatin accessibility		b	x	x*			Yes		Swanson et al. eLife 2021	Adaptation of CITE-seq protocol that includes chromatin accessibility	

(continued)

Table 16.1 (continued)

Technology	Full name or description	Genome	Epig-enome	Tran-scriptome	Pro-teome	Pertur-bation	Spatial	Scalability	Preprint	Reference	Notes	Potential use(s) in characterization of the TME
Perturb-seq	Pooled, combinatorial CRISPR screens with scRNA-seq readout			x		x		Yes		Dixit et al. Cell 2016	Pooled CRISPR screen with (droplet-based) scRNA-seq readout	Single-cell reporting of genetic perturbation Separate perturbation responses from potential confounders. Help elucidate molecular circuits, e.g., immune response in TME or perturbation associated with drug response/resistance in tumor and non-tumor cells
CRISP-seq	An integrated method for single-cell RNA-seq and CRISPR-pooled screens			x		x		Yes		Jaitin et al. Cell 2016	Pooled CRISPR screen with (droplet-based) scRNA-seq readout	
CROP-seq	CRISPR droplet sequencing			x		x		Yes		Datlinger et al. Nat Methods 2017	Pooled CRISPR screen with (droplet-based) scRNA-seq readout	
Perturb-ATAC-seq	Simultaneous CRISPR guide detection and epigenome profiling in single cells		b	x		x		Yes		Rubin et al. Cell 2018	Combines multiplexed CRISPR interference or knockout with genome-wide chromatin accessibility profiling in single cells based on the simultaneous detection of CRISPR guide RNAs and open chromatin sites by ATAC-seq	
ECCITE-seq	Expanded CRISPR-compatible cellular indexing of transcriptomes and epitopes by sequencing			x	x*	x		Yes		Mimitou et al. Nat Methods 2019	Detection of surface proteins similar to CITE-seq together with the scRNA-seq and clonotype features; system adapted to enable direct and robust capture of sgRNAs from existing guide libraries and commonly used vectors compatible with pooled cloning	

Spatial transcriptomics				x			x	Yes		Stahl et al. Science 2016	Histological sections on arrayed reverse transcription primers with unique positional barcodes; no true single-cell resolution	Provides insight into the spatial architecture of the TME which can benefit patient subtyping (e.g., reliable identification of immune-excluded tumors)
osmFISH	Cyclic-ouroboros single-molecule fluorescence in situ hybridization method			x*			x	Limited		Codeluppi et al. Nat Methods 2018	Cyclic single-molecule fluorescence in situ hybridization methodology, number of targets scales linearly with the number of hybridization rounds, limited number of transcripts	
STARmap	Spatially resolved transcript amplicon readout mapping			x*			x	Yes		Wang et al. Science 2018	Combination of in situ sequencing approach with hydrogel-tissue chemistry to develop technology for three-dimensional (3D) intact-tissue RNA sequencing of up to 1000 genes	
Slide-seq				x			x	Yes		Rodriques et al. Science 2019	RNA spatially resolved from tissue sections by transfer onto a surface covered with DNA-barcoded beads, sequencing similar to Drop-seq	

(continued)

Table 16.1 (continued)

Technology	Full name or description	Genome	Epig-enome	Tran-scriptome	Pro-teome	Pertur-bation	Spatial	Scalability	Preprint	Reference	Notes	Potential use(s) in characterization of the TME
MERFISH	Multiplexed error-robust FISH			x*			x	Yes		Xia et al. PNAS 2019	Near-genome-wide (~10,000 genes), spatially resolved RNA profiling of individual cells	
seqFISH+	Sequential fluorescence in situ hybridization			x			x	Yes		Eng et al. Nature 2019	Image mRNAs for 10,000 genes in single cells with high accuracy and sub-diffraction-limit resolution	

a Methylome profiling
b Chromatin accessibility
c Nucleosome occupancy
d Chromatin structure
*Does not provide whole-genome coverage

methods that can be used depending on the data type, as well as the analysis goal (e.g., disease subtyping versus biomarker discovery). A selection of methods which have R or Python implementation is highlighted in Table 16.2 [90, 107, 108, 220, 230–250].

Crosstalk in the Tumor Microenvironment

Detecting and quantifying the presence of cell types in the TME within a particular spatial context leaves out an important and critical aspect of heterogeneity: the interaction that occurs between cells in the TME, carried out by molecules that are either secreted (including metabolites, ions, hormones, extracellular matrix components) or expressed on the surface (mostly (glyco)proteins). These interactions can play a role in structure and communication, triggering downstream signaling pathways and changing transcriptional activity. Furthermore, they have been shown to be a major contributor to tumor phenotype [251, 252], prognosis [253] and, therefore, an important target for (novel) anticancer agents [5, 254, 255]. The majority of such interactions, including various oncogenic pathways, growth receptor signaling, and immune modulation, obviously occur in a cell-type-specific fashion (i.e., different cell types provide different signals), so that the use of single-cell omics technologies provides a valuable source of information that can be used to model the intricate communication network that is typical for the TME.

A proteomic assessment, that is, direct measurement of protein levels, would provide the most straightforward read-out to interpret ligand-receptor pairs present in the TME, but it is difficult to achieve technically [159]. Computational methods have been developed in recent years to use gene expression data from interacting individual cells to profile intercellular communication using multiple (single-cell) omics modalities [256, 257]. Early attempts using bulk RNA sequencing data to systematically characterize human ligand-receptor pairs established a reference and determined expression thresholds that lead to acceptable false-positive rates [258]. This reference has been used in different disease context on scRNA-seq data to determine significant receptor-ligand interactions [259, 260]. CellPhoneDB [261, 262], currently one of the most used tools, uses such a predefined repository of ligands, receptors, and their interactions and is available online or as a Python package. It takes into account the structural stoichiometry of ligands and receptors, which distinguishes it from many earlier approaches [251, 252, 263–266]. This is particularly relevant for multimeric receptors in the TME (e.g., cytokine receptors) for which expression of all subunits is required in order to allow a functional interaction and communication. A statistical framework allows to predict enriched cellular interactions between cell types based on single-cell transcriptomic data. The use of a curated database is a potential downside as it leads to bias but also prevents detection of spurious interactions. It has been used to characterize intercellular interactions in the TME of various cancer types as well as in non-oncological biological contexts [262, 267–273]. The same approach with implementation of multimeric interactions has been expanded in more recent implementations, including CellChat [274] and ICELLNET [275]. NATMI (Network Analysis Toolkit for Multicellular Interactions) uses its own curated ligand-receptor database to create an asymmetric directional network and is slightly faster and results in the selection of more specific interactions, although it lacks the capacity to take multimeric ligands/receptors into consideration [276]. SingleCellSignalR is an R-based implementation that similarly uses a curated database of ligand-receptor interactions and links the ligand-receptor pairs to intracellular networks rooted at the receptors identified in a particular context [277, 278]. Its statistical implementation allows to determine the false discovery rate explicitly. A subset of methods relies on differential gene expression of both ligand and receptors between cell-type clusters in scRNA-seq data to determine potentially relevant pairs, including CellTalker [279], CCCExplorer [280], iTALK [281], and PyMINER [282].

Table 16.2 Selection of computational tools and methods implemented in R/Python for the analysis and integration of single-cell multimodal omics data

Name	Description	Coding language	Clustering, subtyping, classification	Biomarker prediction	Disease biology	Supported data types	Reference
iCluster	Joint latent variable model for integrative clustering	R	x			Numerical	Shen et al. Bioinformatics 2009
PARADIGM	PAthway Recognition Algorithm using Data Integration on Genomic Models	Python		x		Numerical	Vaske et al. Bioinformatics 2010
iClusterPlus	Joint latent variable model for integrative clustering with use of generalized linear regression for the formulation of a joint model	R	x	x	x	Numerical and categorical	Mo et al. Proc Natl Acad Sci USA 2013
SNF	Similarity network fusion; two-step process with nonlinear integration of sample-similarity networks for each data type	R	x			Numerical	Wang et al. Nat Methods 2014
LRAcluster	Low-rank approximation based multi-omics data clustering	R	x			Numerical	Wu et al. BMC Genomics 2015
MIMOSCA	Multi-input Multi-output Single-Cell Analysis; computational framework for the analysis of Perturb-seq data	Python			x	Numerical	Dixit et al. 2016

iNMF	Non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data	Python	x			Numerical	Yang et al. Bioinformatics 2016
MATCHER	Manifold Alignment to Characterize Experimental Relationships: low-dimensional representation (manifold) calculated for each data type and projected into a common space	Python	x			Numerical	Welch et al. Genome Biol 2017
Clusternomics	Integrative context-dependent clustering; probabilistic clustering method with clusters on both local and global level modeled using a hierarchical Dirichlet mixture model to identify structure	R	x			Numerical	Gabasova et al. PLoS Comput Biol 2017
mixOmics	R package with implementation of wide range of methods for statistical integration of multiple data types	R	x	x		Numerical and categorical	Rohart et al. PLoS Comput Biol 2017
AMARETTO	Multi-omics data fusion to construct models of co-expressed genes	R		x	x	Numerical	Champion et al. EBioMedicine 2018

(continued)

Table 16.2 (continued)

Name	Description	Coding language	Clustering, subtyping, classification	Biomarker prediction	Disease biology	Supported data types	Reference
PINSPlus	Perturbation clustering for data INtegration and disease Subtyping; unsupervised similarity-based clustering	R	x			numerical	Nguyen et al. Bioinformatics 2018
MOFA	Multi-Omics Factor Analysis; probabilistic Bayesian framework	R/Python		x	x	Numerical and categorical	Argelaguet et al. Mol Syst Biol 2018
Seurat v3	Dimensionality reduction of multiple data sets/types jointly via canonical correlation analysis (CCA) identifies mutual nearest neighbors that act as anchors	R	x	x	x	Numerical	Stuart et al. Cell 2019
LIGER	Dimensionality reduction of multiple data sets/types via integrative non-negative matrix factorization (iNMF) to identify shared or dataset-specific metagenes; cell types identified in metagene space	R	x	x	x	Numerical	Welch et al. Cell 2019
Harmony	Robust, scalable, and flexible multi-dataset integration using dimensionality reduction (PCA) followed by iterative soft clustering strategy	R	x	x	x	Numerical	Korsunsky et al. Nature Methods 2019

NEMO	Neighborhood-based multi-omics clustering: similarity-based clustering of multi-omic data points (e.g., cancer subtypes), aimed at partial (i.e., incomplete) data sources	R	x			Numerical	Rappoport et al. Bioinformatics 2019
MOGSA	Multi-omics gene-set analysis: integrative single sample gene-set analysis of multiple omics data	R	x	x		Numerical	Meng et al. Mol Cell Proteomics 2019
MCIA	multiple co-inertia analysis	R	x		x	Numerical	Meng et al. Mol Cell Proteomics 2019
MMD-MA	Maximum Mean Discrepancy Manifold Alignment: unsupervised manifold alignment algorithm for integrating multiple measurements carried out on disjoint aliquots of a given population of cells	Python	x			Numerical	Liu et al. bioRxiv 2019
MOFA+	Multi-Omics Factor Analysis v2; Bayesian framework with stochastic variational inference framework amenable to GPU computations, enabling analysis of larger datasets and incorporating flexible priors for different data types	R/Python		x	x	Numerical and categorical	Argelaguet et al. Genome Biol 2020

(continued)

Table 16.2 (continued)

Name	Description	Coding language	Clustering, subtyping, classification	Biomarker prediction	Disease biology	Supported data types	Reference
UnionCom	Embeds intrinsic low-dimensional structure of each single-cell dataset into a distance matrix of cells within the same dataset and aligns cells across single-cell multi-omics datasets by matching the distance matrices via a matrix optimization method	Python	x			Numerical	Cao et al. Bioinformatics 2020
CiteFuse	SNF implementation package for CITE-seq data	R	x	x	x	Numerical	Kim et al. Bioinformatics 2020
Stereoscope	Model-based probabilistic method that uses single-cell data to deconvolve the cell mixtures in spatial data	Python			x	Spatial	Andersson et al. Commun Biol 2020
Seurat v4	"Weighted nearest neighbor" analysis integrates multimodal single-cell data; provides resource to map query datasets onto multimodal reference atlas	R	x	x	x	Numerical	Hao et al. Cell 2021
totalVI	Joint probabilistic modeling of single-cell omics data	Python	x	x	x	Numerical	Gayoso et al. Nat Methods 2021

Incorporation of information beyond ligand-receptor pairs is an obvious next step to improve the output. Spatial relations influence potential interactions between cell types, particularly in the case of cell surface ligand-receptor interactions. Attempts to explicitly include spatial transcriptomic/proteomic information in the analysis (SpaOTsc [283], SVCA [284]) infer three-dimensional organization (RNA-Magnet [285]) or validate existing tools like CellPhoneDB [271] using spatial data have all been proposed recently. Inferring interactions in the TME from transcriptome data ultimately relies on the interpretation of co-expression data. Social network-style gene co-expression graphs, with nodes in the graph representing genes and edges referring to the strength of co-expression, have been used extensively to analyze bulk RNA-seq data [286, 287]. PyMINER [282] tries to integrate the structure of a gene co-expression graph with information about protein-protein interaction into a scRNA-seq analysis pipeline. NicheNet builds upon the sender-receiver framework to infer the effects of sender-cell ligands on a receiver-cell's gene expression more comprehensively by integrating prior knowledge on ligand-target downstream signaling pathways [288]. It goes beyond ligand-receptor interactions to predict which ligands influence the transcriptome of another cell and which target genes and signaling mediators may be involved. This methodology has been successfully used to characterize interactions with the TME in squamous cell carcinoma [288, 289] and colorectal cancer [290]. Other computational tools with similar objectives have been proposed [291] and which tool to use depends on whether there is previous knowledge of interactions of interest (versus a more unbiased characterization) in the TME.

Differences in ligand-receptor pairs that are used as ground truth as well as profound differences in statistical methodology make a direct comparison particularly difficult. This has resulted in efforts to collate publicly available list into a single ligand-receptor repository [257]. Predicting cell-cell interactions from single-cell omics data is no easy task as it requires inference of protein levels from mRNA expression and integration of various inferred properties including cell-type annotation, marker gene identification, and dropout correction, in part explaining the abundance in variability in currently proposed methods. There is an urgent need for benchmarking of computational prediction of crosstalk in the tissue microenvironment, and it is important to consider assumptions and limitations when selecting which tool to use. Furthermore, experimental validation of interactions discovered *in silico* (through protein detection, visualization, and/or functional assays) remains crucial, even if false discovery rates can be controlled. It is clear that single-cell omics data modalities can help gain a better understanding of the crosstalk that occurs between components of the TME, which would inevitably lead to a better mechanistic understanding of the processes driving response and resistance in diverse tumor types and can guide the rational development of synergistic or targeted treatment regimens. Other aspects of crosstalk in the TME like extracellular vesicles and direct cell contact (e.g., via gap junctions) are more difficult to model, but efforts to study this comprehensively are underway.

Conclusion

The increased interest into the role of the TME, driven in part by recent developments of single-cell omics technologies, emphasizes the importance of expanding the concept of tumor heterogeneity to include aspects of the non-tumoral context. Immune responses in cancer have been shown to be extremely variable within tumors from individual patients with same tumor type. The cancer-immune set point exists on a continuous spectrum, rather than as a discrete phenotype. The identification of factors that determine the immune profile and set point of individual patients represents a crucial goal. Biomarkers for response to immunotherapy include immune cell infiltration, cytokine profile, germline and tumor genetics, age, microbiome, (viral) infections, UV exposure, and (previous) exposure to immune-modifying drugs. The balance between anticancer activity and tolerance

that determines the efficacy of immunotherapy in individual patients may be primarily due to small differences in each of these factors, rather than dramatic splits. Another aspect that is currently incompletely understood is the dynamics and trafficking of immune cell types between the TME, the peripheral blood, and lymphoid tissues (lymph nodes, TLS, bone marrow) during cancer development, treatment, and eventual relapse.

This variability complicates biomarker discovery and underlines the importance of incorporating rich clinical and demographic data to mitigate confounding factors. Thus far, the speed by which clinical trials involving cancer immunotherapy are conducted has exceeded the pace of our progress to understand the basic science behind the role of the (immune) microenvironment in response and resistance. This observation suggests a critical opportunity for researchers to use novel single-cell technologies on trial patient samples, in order to reconcile scientific and clinical insights synergistically as trials proceed. The lessons learned will not only greatly increase our understanding of the cancer-immune cycle but can also guide the identification of new targets and help set up a systematic framework for personalization of cancer treatment.

In conclusion, multimodal single-cell technologies offer an unprecedented insight to improve the understanding of physiology and disease. Analysis of genome, epigenome, transcriptome, and proteome together in individual cells has the power to comprehensively characterize cell identity and state, as well as reveal gene regulatory networks. The incorporation of spatial data and methods to specifically study crosstalk between cells offers an important advantage to capture the complexities present in the TME. In this chapter, we have highlighted a range of experimental methods and analytical tools that help acquire and leverage this data. Important challenges remain in terms of standardization, technical performance, integration, and systematic computational analysis. As data from individual studies and cell atlases becomes available, these tools will become more robust. We hope and we believe that this broader scope will translate to clinical improvements and more effective personalized cancer therapy.

References

1. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646–74. https://doi.org/10.1016/j.cell.2011.02.013.
2. Fridman WH, Pages F, Sautes-Fridman C, Galon J. The immune contexture in human tumours: impact on clinical outcome. Nat Rev Cancer. 2012;12(4):298–306. https://doi.org/10.1038/nrc3245.
3. Chen DS, Mellman I. Oncology meets immunology: the cancer-immunity cycle. Immunity. 2013;39(1):1–10. https://doi.org/10.1016/j.immuni.2013.07.012.
4. Schiavoni G, Gabriele L, Mattei F. The tumor microenvironment: a pitch for multiple players. Frontiers. Oncology. 2013;3(90) https://doi.org/10.3389/fonc.2013.00090.
5. Fridman WH, Zitvogel L, Sautès–Fridman C, Kroemer G. The immune contexture in cancer prognosis and treatment. Nat Rev Clin Oncol. 2017;14(12):717–34. https://doi.org/10.1038/nrclinonc.2017.101.
6. Binnewies M, Roberts EW, Kersten K, Chan V, Fearon DF, Merad M, et al. Understanding the tumor immune microenvironment (TIME) for effective therapy. Nat Med. 2018;24(5):541–50. https://doi.org/10.1038/s41591-018-0014-x.
7. Talmadge JE, Fidler IJ. AACR centennial series: the biology of cancer metastasis: historical perspective. Cancer Res. 2010;70(14):5649–69. https://doi.org/10.1158/0008-5472.CAN-10-1040.
8. Maman S, Witz IP. A history of exploring cancer in context. Nat Rev Cancer. 2018;18(6):359–76. https://doi.org/10.1038/s41568-018-0006-7.
9. Vogelstein B, Kinzler KW. The multistep nature of cancer. Trends Genet. 1993;9(4):138–41. https://doi.org/10.1016/0168-9525(93)90209-z.
10. Vogelstein B, Kinzler KW. Cancer genes and the pathways they control. Nat Med. 2004;10(8):789–99. https://doi.org/10.1038/nm1087.
11. Salmon H, Remark R, Gnjatic S, Merad M. Host tissue determinants of tumour immunity. Nat Rev Cancer. 2019;19(4):215–27. https://doi.org/10.1038/s41568-019-0125-9.
12. Gentles AJ, Newman AM, Liu CL, Bratman SV, Feng W, Kim D, et al. The prognostic landscape of genes and infiltrating immune cells across human cancers. Nat Med. 2015;21(8):938–45. https://doi.org/10.1038/nm.3909.
13. Garner H, de Visser KE. Immune crosstalk in cancer progression and metastatic spread: a complex conversation. Nat Rev Immunol. 2020;20(8):483–97. https://doi.org/10.1038/s41577-019-0271-z.

14. Pagès F, Mlecnik B, Marliot F, Bindea G, Ou FS, Bifulco C, et al. International validation of the consensus Immunoscore for the classification of colon cancer: a prognostic and accuracy study. Lancet. 2018;391(10135):2128–39. https://doi.org/10.1016/s0140-6736(18)30789-x.
15. Bruni D, Angell HK, Galon J. The immune contexture and Immunoscore in cancer prognosis and therapeutic efficacy. Nat Rev Cancer. 2020;20(11):662–80. https://doi.org/10.1038/s41568-020-0285-7.
16. Galluzzi L, Buqué A, Kepp O, Zitvogel L, Kroemer G. Immunological effects of conventional chemotherapy and targeted anticancer agents. Cancer Cell. 2015;28(6):690–714. https://doi.org/10.1016/j.ccell.2015.10.012.
17. Hanahan D, Bergers G, Bergsland E. Less is more, regularly: metronomic dosing of cytotoxic drugs can target tumor angiogenesis in mice. J Clin Invest. 2000;105(8):1045–7. https://doi.org/10.1172/jci9872.
18. Hughes E, Scurr M, Campbell E, Jones E, Godkin A, Gallimore A. T-cell modulation by cyclophosphamide for tumour therapy. Immunology. 2018;154(1):62–8. https://doi.org/10.1111/imm.12913.
19. Kerbel RS, Kamen BA. The anti-angiogenic basis of metronomic chemotherapy. Nat Rev Cancer. 2004;4(6):423–36. https://doi.org/10.1038/nrc1369.
20. Pasquier E, Kavallaris M, Andre N. Metronomic chemotherapy: new rationale for new directions. Nat Rev Clin Oncol. 2010;7(8):455–65. https://doi.org/10.1038/nrclinonc.2010.82.
21. Zitvogel L, Apetoh L, Ghiringhelli F, Kroemer G. Immunological aspects of cancer chemotherapy. Nat Rev Immunol. 2008;8(1):59–73. https://doi.org/10.1038/nri2216.
22. Chiossone L, Dumas PY, Vienne M, Vivier E. Natural killer cells and other innate lymphoid cells in cancer. Nat Rev Immunol. 2018;18(11):671–88. https://doi.org/10.1038/s41577-018-0061-z.
23. Feins S, Kong W, Williams EF, Milone MC, Fraietta JA. An introduction to chimeric antigen receptor (CAR) T-cell immunotherapy for human cancer. Am J Hematol. 2019;94(S1):S3–s9. https://doi.org/10.1002/ajh.25418.
24. Rafiq S, Hackett CS, Brentjens RJ. Engineering strategies to overcome the current roadblocks in CAR T cell therapy. Nat Rev Clin Oncol. 2019; https://doi.org/10.1038/s41571-019-0297-y.
25. Labrijn AF, Janmaat ML, Reichert JM, Parren P. Bispecific antibodies: a mechanistic review of the pipeline. Nat Rev Drug Discov. 2019;18(8):585–608. https://doi.org/10.1038/s41573-019-0028-1.
26. Myers JA, Miller JS. Exploring the NK cell platform for cancer immunotherapy. Nat Rev Clin Oncol. 2021;18(2):85–100. https://doi.org/10.1038/s41571-020-0426-7.
27. Wynn TA, Chawla A, Pollard JW. Macrophage biology in development, homeostasis and disease. Nature. 2013;496(7446):445–55. https://doi.org/10.1038/nature12034.
28. Grisaru-Tal S, Itan M, Klion AD, Munitz A. A new dawn for eosinophils in the tumour microenvironment. Nat Rev Cancer. 2020; https://doi.org/10.1038/s41568-020-0283-9.
29. Jaillon S, Ponzetta A, Di Mitri D, Santoni A, Bonecchi R, Mantovani A. Neutrophil diversity and plasticity in tumour progression and therapy. Nat Rev Cancer. 2020; https://doi.org/10.1038/s41568-020-0281-y.
30. Diaz-Montero CM, Finke J, Montero AJ. Myeloid-derived suppressor cells in cancer: therapeutic, predictive, and prognostic implications. Semin Oncol. 2014;41(2):174–84. https://doi.org/10.1053/j.seminoncol.2014.02.003.
31. Turley SJ, Cremasco V, Astarita JL. Immunological hallmarks of stromal cells in the tumour microenvironment. Nat Rev Immunol. 2015;15(11):669–82. https://doi.org/10.1038/nri3902.
32. Motz GT, Santoro SP, Wang LP, Garrabrant T, Lastra RR, Hagemann IS, et al. Tumor endothelium FasL establishes a selective immune barrier promoting tolerance in tumors. Nat Med. 2014;20(6):607–15. https://doi.org/10.1038/nm.3541.
33. Silverman DA, Martinez VK, Dougherty PM, Myers JN, Calin GA, Amit M. Cancer-associated neurogenesis and nerve-cancer cross-talk. Cancer Res. 2021;81(6):1431–40. https://doi.org/10.1158/0008-5472.Can-20-2793.
34. Wang W, Li L, Chen N, Niu C, Li Z, Hu J, et al. Nerves in the tumor microenvironment: origin and effects. Front Cell Dev Biol. 2020;8(1630) https://doi.org/10.3389/fcell.2020.601738.
35. Chen DS, Mellman I. Elements of cancer immunity and the cancer-immune set point. Nature. 2017;541(7637):321–30. https://doi.org/10.1038/nature21349.
36. Chen F, Zhuang X, Lin L, Yu P, Wang Y, Shi Y, et al. New horizons in tumor microenvironment biology: challenges and opportunities. BMC Med. 2015;13:45. https://doi.org/10.1186/s12916-015-0278-7.
37. Fang H, Declerck YA. Targeting the tumor microenvironment: from understanding pathways to effective clinical trials. Cancer Res. 2013;73(16):4965–77. https://doi.org/10.1158/0008-5472.Can-13-0661.
38. Zappasodi R, Merghoub T, Wolchok JD. Emerging concepts for immune checkpoint blockade-based combination therapies. Cancer Cell. 2018;33(4):581–98. https://doi.org/10.1016/j.ccell.2018.03.005.
39. Bedard PL, Hansen AR, Ratain MJ, Siu LL. Tumour heterogeneity in the clinic. Nature. 2013;501(7467):355–64. https://doi.org/10.1038/nature12627.
40. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10(1):57–63. https://doi.org/10.1038/nrg2484.
41. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with

DESeq2. Genome Biol. 2014;15(12):550. https://doi.org/10.1186/s13059-014-0550-8.
42. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47-e. https://doi.org/10.1093/nar/gkv007.
43. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40. https://doi.org/10.1093/bioinformatics/btp616.
44. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015;12:453. https://doi.org/10.1038/nmeth.3337. https://www.nature.com/articles/nmeth.3337#supplementary-information
45. Li B, Severson E, Pignon J-C, Zhao H, Li T, Novak J, et al. Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. Genome Biol. 2016;17(1):174. https://doi.org/10.1186/s13059-016-1028-7.
46. Racle J, de Jonge K, Baumgaertner P, Speiser DE, Gfeller D. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. elife. 2017;6 https://doi.org/10.7554/eLife.26476.
47. Finotello F, Mayer C, Plattner C, Laschober G, Rieder D, Hackl H, et al. Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. Genome Med. 2019;11(1):34. https://doi.org/10.1186/s13073-019-0638-6.
48. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci. 2005;102(43):15545–50. https://doi.org/10.1073/pnas.0506580102.
49. Finotello F, Trajanoski Z. Quantifying tumor-infiltrating immune cells from transcriptomics data. Cancer Immunol Immunother. 2018;67(7):1031–40. https://doi.org/10.1007/s00262-018-2150-z.
50. Hackl H, Charoentong P, Finotello F, Trajanoski Z. Computational genomics tools for dissecting tumour–immune cell interactions. Nat Rev Genet. 2016;17(8):441–58. https://doi.org/10.1038/nrg.2016.67.
51. Liu XS, Mardis ER. Applications of immunogenomics to cancer. Cell. 2017;168(4):600–12. https://doi.org/10.1016/j.cell.2017.01.014.
52. Bolotin DA, Poslavsky S, Davydov AN, Frenkel FE, Fanchi L, Zolotareva OI, et al. Antigen receptor repertoire profiling from RNA-seq data. Nat Biotechnol. 2017;35(10):908–11. https://doi.org/10.1038/nbt.3979.
53. Bolotin DA, Poslavsky S, Mitrophanov I, Shugay M, Mamedov IZ, Putintseva EV, et al. MiXCR: software for comprehensive adaptive immunity profiling. Nat Methods. 2015;12(5):380–1. https://doi.org/10.1038/nmeth.3364.
54. Li B, Li T, Wang B, Dou R, Zhang J, Liu JS, et al. Ultrasensitive detection of TCR hypervariable-region sequences in solid-tissue RNA–seq data. Nat Genet. 2017;49(4):482–3. https://doi.org/10.1038/ng.3820.
55. Richters MM, Xia H, Campbell KM, Gillanders WE, Griffith OL, Griffith M. Best practices for bioinformatic characterization of neoantigens for clinical utility. Genome Med. 2019;11(1):56. https://doi.org/10.1186/s13073-019-0666-2.
56. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods. 2009;6(5):377–82. https://doi.org/10.1038/nmeth.1315.
57. Svensson V, Natarajan KN, Ly LH, Miragaia RJ, Labalette C, Macaulay IC, et al. Power analysis of single-cell RNA-sequencing experiments. Nat Methods. 2017;14(4):381–7. https://doi.org/10.1038/nmeth.4220.
58. Svensson V, Vento-Tormo R, Teichmann SA. Exponential scaling of single-cell RNA-seq in the past decade. Nat Protoc. 2018;13(4):599–604. https://doi.org/10.1038/nprot.2017.149.
59. Chen G, Ning B, Shi T. Single-cell RNA-Seq technologies and related computational data analysis. Front Genet. 2019;10(317) https://doi.org/10.3389/fgene.2019.00317.
60. Haque A, Engel J, Teichmann SA, Lonnberg T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Med. 2017;9(1):75. https://doi.org/10.1186/s13073-017-0467-4.
61. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med. 2018;50(8):96. https://doi.org/10.1038/s12276-018-0071-8.
62. Kashima Y, Sakamoto Y, Kaneko K, Seki M, Suzuki Y, Suzuki A. Single-cell sequencing techniques from individual to multiomics analyses. Exp Mol Med. 2020;52(9):1419–27. https://doi.org/10.1038/s12276-020-00499-2.
63. Ramsköld D, Luo S, Wang Y-C, Li R, Deng Q, Faridani OR, et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nat Biotechnol. 2012;30(8):777–82. https://doi.org/10.1038/nbt.2282.
64. Sheng K, Cao W, Niu Y, Deng Q, Zong C. Effective detection of variation in single-cell transcriptomes using MATQ-seq. Nat Methods. 2017;14(3):267–70. https://doi.org/10.1038/nmeth.4145.
65. Picelli S, Bjorklund AK, Faridani OR, Sagasser S, Winberg G, Sandberg R. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat Methods. 2013;10(11):1096–8. https://doi.org/10.1038/nmeth.2639.
66. Picelli S, Faridani OR, Bjorklund AK, Winberg G, Sagasser S, Sandberg R. Full-length RNA-seq from single cells using smart-seq2. Nat Protoc.

2014;9(1):171–81. https://doi.org/10.1038/nprot.2014.006.
67. Hashimshony T, Senderovich N, Avital G, Klochendler A, de Leeuw Y, Anavy L, et al. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. Genome Biol. 2016;17(1):77. https://doi.org/10.1186/s13059-016-0938-8.
68. Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8:14049. https://doi.org/10.1038/ncomms14049.
69. Bernard V, Semaan A, Huang J, San Lucas FA, Mulu FC, Stephens BM, et al. Single-cell transcriptomics of pancreatic cancer precursors demonstrates epithelial and microenvironmental heterogeneity as an early event in neoplastic progression. Clin Cancer Res. 2019;25(7):2194–205. https://doi.org/10.1158/1078-0432.Ccr-18-1955.
70. Habib N, Avraham-Davidi I, Basu A, Burks T, Shekhar K, Hofree M, et al. Massively parallel single-nucleus RNA-seq with DroNc-seq. Nat Methods. 2017;14(10):955–8. https://doi.org/10.1038/nmeth.4407.
71. Macosko Evan Z, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly parallel genome-wide expression profiling of individual cells using Nanoliter droplets. Cell. 2015;161(5):1202–14. https://doi.org/10.1016/j.cell.2015.05.002.
72. Klein Allon M, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015;161(5):1187–201. https://doi.org/10.1016/j.cell.2015.04.044.
73. Keren-Shaul H, Kenigsberg E, Jaitin DA, David E, Paul F, Tanay A, et al. MARS-seq2.0: an experimental and analytical pipeline for indexed sorting combined with single-cell RNA sequencing. Nat Protoc. 2019;14(6):1841–62. https://doi.org/10.1038/s41596-019-0164-4.
74. Hashimoto S. Nx1-Seq (well based single-cell analysis system). Adv Exp Med Biol. 2019;1129:51–61. https://doi.org/10.1007/978-981-13-6037-4_4.
75. Sasagawa Y, Danno H, Takada H, Ebisawa M, Tanaka K, Hayashi T, et al. Quartz-Seq2: a high-throughput single-cell RNA-sequencing method that effectively uses limited sequence reads. Genome Biol. 2018;19(1):29. https://doi.org/10.1186/s13059-018-1407-3.
76. Gierahn TM, Wadsworth MH, Hughes TK, Bryson BD, Butler A, Satija R, et al. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. Nat Methods. 2017;14(4):395–8. https://doi.org/10.1038/nmeth.4179.
77. Islam S, Kjällquist U, Moliner A, Zajac P, Fan J-B, Lönnerberg P, et al. Highly multiplexed and strand-specific single-cell RNA 5′ end sequencing. Nat Protoc. 2012;7(5):813–28. https://doi.org/10.1038/nprot.2012.022.
78. Fan X, Zhang X, Wu X, Guo H, Hu Y, Tang F, et al. Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos. Genome Biol. 2015;16(1):148. https://doi.org/10.1186/s13059-015-0706-1.
79. Hayashi T, Ozaki H, Sasagawa Y, Umeda M, Danno H, Nikaido I. Single-cell full-length total RNA sequencing uncovers dynamics of recursive splicing and enhancer RNAs. Nat Commun. 2018;9(1):619. https://doi.org/10.1038/s41467-018-02866-0.
80. Hagemann-Jensen M, Abdullayev I, Sandberg R, Faridani OR. Small-seq for single-cell small-RNA sequencing. Nat Protoc. 2018;13(10):2407–24. https://doi.org/10.1038/s41596-018-0049-y.
81. Isakova A, Neff N, Quake SR. Single cell profiling of total RNA using Smart-seq-total. bioRxiv. 2020:2020.06.02.131060. https://doi.org/10.1101/2020.06.02.131060.
82. Mereu E, Lafzi A, Moutinho C, Ziegenhain C, McCarthy DJ, Álvarez-Varela A, et al. Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. Nat Biotechnol. 2020; https://doi.org/10.1038/s41587-020-0469-4.
83. Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, et al. Comparative analysis of single-cell RNA sequencing methods. Mol Cell. 2017;65(4):631–43.e4. https://doi.org/10.1016/j.molcel.2017.01.023.
84. Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. Nat Rev Genet. 2015;16(3):133–45. https://doi.org/10.1038/nrg3833.
85. Bacher R, Kendziorski C. Design and computational analysis of single-cell RNA-sequencing experiments. Genome Biol. 2016;17(1):63. https://doi.org/10.1186/s13059-016-0927-y.
86. Vieth B, Parekh S, Ziegenhain C, Enard W, Hellmann I. A systematic evaluation of single cell RNA-seq analysis pipelines. Nat Commun. 2019;10(1):4667. https://doi.org/10.1038/s41467-019-12266-7.
87. Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. Mol Syst Biol. 2019;15(6):e8746. https://doi.org/10.15252/msb.20188746.
88. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018;36(5):411–20. https://doi.org/10.1038/nbt.4096.
89. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol. 2015;33(5):495–502. https://doi.org/10.1038/nbt.3192.
90. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM III, et al. Comprehensive integration of single-cell data. Cell. 2019;177(7):1888–902.e21. https://doi.org/10.1016/j.cell.2019.05.031.
91. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome

Biol. 2018;19(1):15. https://doi.org/10.1186/s13059-017-1382-0.
92. Jiang P, Thomson JA, Stewart R. Quality control of single-cell RNA-seq by SinQC. Bioinformatics. 2016;32(16):2514–6. https://doi.org/10.1093/bioinformatics/btw176.
93. McCarthy DJ, Campbell KR, Lun ATL, Wills QF. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. Bioinformatics. 2017;33(8):1179–86. https://doi.org/10.1093/bioinformatics/btw777.
94. Young MD, Behjati S. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. GigaScience. 2020;9(12) https://doi.org/10.1093/gigascience/giaa151.
95. Yang S, Corbett SE, Koga Y, Wang Z, Johnson WE, Yajima M, et al. Decontamination of ambient RNA in single-cell RNA-seq with DecontX. Genome Biol. 2020;21(1):57. https://doi.org/10.1186/s13059-020-1950-6.
96. Stoeckius M, Zheng S, Houck-Loomis B, Hao S, Yeung BZ, Mauck WM 3rd, et al. Cell hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. Genome Biol. 2018;19(1):224. https://doi.org/10.1186/s13059-018-1603-1.
97. DePasquale EAK, Schnell DJ, Van Camp PJ, Valiente-Alandí Í, Blaxall BC, Grimes HL, et al. DoubletDecon: deconvoluting doublets from single-cell RNA-sequencing data. Cell Rep. 2019;29(6):1718–27.e8. https://doi.org/10.1016/j.celrep.2019.09.082.
98. McGinnis CS, Murrow LM, Gartner ZJ. DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. Cell Syst. 2019;8(4):329–37.e4. https://doi.org/10.1016/j.cels.2019.03.003.
99. Wolock SL, Lopez R, Klein AM. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. Cell Syst. 2019;8(4):281–91 e9. https://doi.org/10.1016/j.cels.2018.11.005.
100. Bais AS, Kostka D. scds: computational annotation of doublets in single-cell RNA sequencing data. Bioinformatics. 2019;36(4):1150–8. https://doi.org/10.1093/bioinformatics/btz698.
101. Xin H, Lian Q, Jiang Y, Luo J, Wang X, Erb C, et al. GMM-Demux: sample demultiplexing, multiplet detection, experiment planning, and novel cell-type verification in single cell sequencing. Genome Biol. 2020;21(1):188. https://doi.org/10.1186/s13059-020-02084-2.
102. Xi NM, Li JJ. Benchmarking computational doublet-detection methods for single-cell RNA sequencing data. Cell Syst. 2021;12(2):176–94.e6. https://doi.org/10.1016/j.cels.2020.11.008.
103. Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, Theis FJ, et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat Biotechnol. 2015;33(2):155–60. https://doi.org/10.1038/nbt.3102.
104. Buettner F, Pratanwanich N, McCarthy DJ, Marioni JC, Stegle O. f-scLVM: scalable and versatile factor analysis for single-cell RNA-seq. Genome Biol. 2017;18(1):212. https://doi.org/10.1186/s13059-017-1334-8.
105. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8(1):118–27. https://doi.org/10.1093/biostatistics/kxj037.
106. Buttner M, Miao Z, Wolf FA, Teichmann SA, Theis FJ. A test metric for assessing single-cell RNA-seq batch correction. Nat Methods. 2019;16(1):43–9. https://doi.org/10.1038/s41592-018-0254-1.
107. Welch JD, Kozareva V, Ferreira A, Vanderburg C, Martin C, Macosko EZ. Single-cell multi-omic integration compares and contrasts features of brain cell identity. Cell. 2019;177(7):1873–87.e17. https://doi.org/10.1016/j.cell.2019.05.006.
108. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods. 2019;16(12):1289–96. https://doi.org/10.1038/s41592-019-0619-0.
109. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. Nat Methods. 2018;15(12):1053–8. https://doi.org/10.1038/s41592-018-0229-2.
110. Li WV, Li JJ. An accurate and robust imputation method scImpute for single-cell RNA-seq data. Nat Commun. 2018;9(1):997. https://doi.org/10.1038/s41467-018-03405-7.
111. Eraslan G, Simon LM, Mircea M, Mueller NS, Theis FJ. Single-cell RNA-seq denoising using a deep count autoencoder. Nat Commun. 2019;10(1):390. https://doi.org/10.1038/s41467-018-07931-2.
112. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Statist Mech Theory Exp. 2008;2008:10008. https://doi.org/10.1088/1742-5468/2008/10/p10008.
113. Wagner A, Regev A, Yosef N. Revealing the vectors of cellular identity with single-cell genomics. Nat Biotechnol. 2016;34(11):1145–60. https://doi.org/10.1038/nbt.3711.
114. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. 2014;32(4):381–6. https://doi.org/10.1038/nbt.2859.
115. Bendall SC, Davis KL, Amir el AD, Tadmor MD, Simonds EF, Chen TJ, et al. Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. Cell. 2014;157(3):714–25. https://doi.org/10.1016/j.cell.2014.04.005.
116. Saelens W, Cannoodt R, Todorov H, Saeys Y. A comparison of single-cell trajectory inference methods. Nat Biotechnol. 2019;37(5):547–54. https://doi.org/10.1038/s41587-019-0071-9.

117. Street K, Risso D, Fletcher RB, Das D, Ngai J, Yosef N, et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genomics. 2018;19(1):477. https://doi.org/10.1186/s12864-018-4772-0.
118. Wolf FA, Hamey FK, Plass M, Solana J, Dahlin JS, Göttgens B, et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. Genome Biol. 2019;20(1):59. https://doi.org/10.1186/s13059-019-1663-x.
119. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, et al. The human cell atlas. Elife. 2017;6 https://doi.org/10.7554/eLife.27041.
120. Pliner HA, Shendure J, Trapnell C. Supervised classification enables rapid annotation of cell atlases. Nat Methods. 2019;16(10):983–6. https://doi.org/10.1038/s41592-019-0535-3.
121. Kiselev VY, Yiu A, Hemberg M. scmap: projection of single-cell RNA-seq data across data sets. Nat Methods. 2018;15(5):359–62. https://doi.org/10.1038/nmeth.4644.
122. Hou R, Denisenko E, Forrest ARR. scMatch: a single-cell gene expression profile annotation tool using reference datasets. Bioinformatics. 2019;35(22):4688–95. https://doi.org/10.1093/bioinformatics/btz292.
123. Tan Y, Cahan P. SingleCellNet: a computational tool to classify single cell RNA-Seq data across platforms and across species. Cell Syst. 2019;9(2):207–13.e2. https://doi.org/10.1016/j.cels.2019.06.004.
124. Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat Immunol. 2019;20(2):163–72. https://doi.org/10.1038/s41590-018-0276-y.
125. Huang Q, Liu Y, Du Y, Garmire LX. Evaluation of cell type annotation R packages on single-cell RNA-seq data. Genomics Proteomics Bioinformatics. 2020; https://doi.org/10.1016/j.gpb.2020.07.004.
126. Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nat Methods. 2014;11(7):740–2. https://doi.org/10.1038/nmeth.2967.
127. Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol. 2015;16(1):278. https://doi.org/10.1186/s13059-015-0844-5.
128. Soneson C, Robinson MD. Bias, robustness and scalability in single-cell differential expression analysis. Nat Methods. 2018;15(4):255–61. https://doi.org/10.1038/nmeth.4612.
129. Van den Berge K, Perraudeau F, Soneson C, Love MI, Risso D, Vert J-P, et al. Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. Genome Biol. 2018;19(1):24. https://doi.org/10.1186/s13059-018-1406-4.
130. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11(3):R25. https://doi.org/10.1186/gb-2010-11-3-r25.
131. Law CW, Chen Y, Shi W, Smyth GK. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014;15(2):R29. https://doi.org/10.1186/gb-2014-15-2-r29.
132. Matsumoto H, Kiryu H, Furusawa C, Ko MSH, Ko SBH, Gouda N, et al. SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. Bioinformatics. 2017;33(15):2314–21. https://doi.org/10.1093/bioinformatics/btx194.
133. Aibar S, Gonzalez-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, et al. SCENIC: single-cell regulatory network inference and clustering. Nat Methods. 2017;14(11):1083–6. https://doi.org/10.1038/nmeth.4463.
134. Chen S, Mar JC. Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. BMC Bioinform. 2018;19(1):232. https://doi.org/10.1186/s12859-018-2217-z.
135. Mallory XF, Edrisi M, Navin N, Nakhleh L. Methods for copy number aberration detection from single-cell DNA-sequencing data. Genome Biol. 2020;21(1):208. https://doi.org/10.1186/s13059-020-02119-8.
136. Dean FB, Hosono S, Fang L, Wu X, Faruqi AF, Bray-Ward P, et al. Comprehensive human genome amplification using multiple displacement amplification. Proc Natl Acad Sci U S A. 2002;99(8):5261–6. https://doi.org/10.1073/pnas.082089499.
137. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, et al. Tumour evolution inferred by single-cell sequencing. Nature. 2011;472(7341):90–4. https://doi.org/10.1038/nature09807.
138. Telenius H, Carter NP, Bebb CE, Nordenskjöld M, Ponder BA, Tunnacliffe A. Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. Genomics. 1992;13(3):718–25. https://doi.org/10.1016/0888-7543(92)90147-k.
139. Zong C, Lu S, Chapman AR, Xie XS. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. Science. 2012;338(6114):1622–6. https://doi.org/10.1126/science.1229164.
140. Xi L, Belyaev A, Spurgeon S, Wang X, Gong H, Aboukhalil R, et al. New library construction method for single-cell genomes. PLoS One. 2017;12(7):e0181163. https://doi.org/10.1371/journal.pone.0181163.
141. Xu X, Hou Y, Yin X, Bao L, Tang A, Song L, et al. Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. Cell. 2012;148(5):886–95. https://doi.org/10.1016/j.cell.2012.02.025.

142. Francis JM, Zhang C-Z, Maire CL, Jung J, Manzo VE, Adalsteinsson VA, et al. EGFR variant heterogeneity in glioblastoma resolved through single-nucleus sequencing. Cancer Discov. 2014;4(8):956–71. https://doi.org/10.1158/2159-8290.Cd-13-0879.
143. Hughes AEO, Magrini V, Demeter R, Miller CA, Fulton R, Fulton LL, et al. Clonal architecture of secondary acute myeloid leukemia defined by single-cell sequencing. PLoS Genet. 2014;10(7):e1004462. https://doi.org/10.1371/journal.pgen.1004462.
144. Gawad C, Koh W, Quake SR. Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics. Proc Natl Acad Sci. 2014;111(50):17947–52. https://doi.org/10.1073/pnas.1420822111.
145. Casasent AK, Schalck A, Gao R, Sei E, Long A, Pangburn W, et al. Multiclonal invasion in breast tumors identified by topographic single cell sequencing. Cell. 2018;172(1):205–17.e12. https://doi.org/10.1016/j.cell.2017.12.007.
146. Leung ML, Davis A, Gao R, Casasent A, Wang Y, Sei E, et al. Single-cell DNA sequencing reveals a late-dissemination model in metastatic colorectal cancer. Genome Res. 2017;27(8):1287–99. https://doi.org/10.1101/gr.209973.116.
147. Heitzer E, Auer M, Gasch C, Pichler M, Ulz P, Hoffmann EM, et al. Complex tumor genomes inferred from single circulating tumor cells by Array-CGH and next-generation sequencing. Cancer Res. 2013;73(10):2965–75. https://doi.org/10.1158/0008-5472.Can-12-4140.
148. Lohr JG, Adalsteinsson VA, Cibulskis K, Choudhury AD, Rosenberg M, Cruz-Gordillo P, et al. Whole-exome sequencing of circulating tumor cells provides a window into metastatic prostate cancer. Nat Biotechnol. 2014;32(5):479–84. https://doi.org/10.1038/nbt.2892.
149. Wang Y, Navin NE. Advances and applications of single-cell sequencing technologies. Mol Cell. 2015;58(4):598–609. https://doi.org/10.1016/j.molcel.2015.05.005.
150. Luquette LJ, Bohrson CL, Sherman MA, Park PJ. Identification of somatic mutations in single cell DNA-seq using a spatial model of allelic imbalance. Nat Commun. 2019;10(1):3908. https://doi.org/10.1038/s41467-019-11857-8.
151. Miles LA, Bowman RL, Merlinsky TR, Csete IS, Ooi AT, Durruthy-Durruthy R, et al. Single-cell mutation analysis of clonal evolution in myeloid malignancies. Nature. 2020; https://doi.org/10.1038/s41586-020-2864-x.
152. Pellegrino M, Sciambi A, Treusch S, Durruthy-Durruthy R, Gokhale K, Jacob J, et al. High-throughput single-cell DNA sequencing of acute myeloid leukemia tumors with droplet microfluidics. Genome Res. 2018;28(9):1345–52. https://doi.org/10.1101/gr.232272.117.
153. Shema E, Bernstein BE, Buenrostro JD. Single-cell and single-molecule epigenomics to uncover genome regulation at unprecedented resolution. Nat Genet. 2019;51(1):19–25. https://doi.org/10.1038/s41588-018-0290-x.
154. Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, et al. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature. 2015;523(7561):486–90. https://doi.org/10.1038/nature14590.
155. Smallwood SA, Lee HJ, Angermueller C, Krueger F, Saadeh H, Peat J, et al. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. Nat Methods. 2014;11(8):817–20. https://doi.org/10.1038/nmeth.3035.
156. Guo H, Zhu P, Wu X, Li X, Wen L, Tang F. Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. Genome Res. 2013;23(12):2126–35. https://doi.org/10.1101/gr.161679.113.
157. Ku WL, Nakamura K, Gao W, Cui K, Hu G, Tang Q, et al. Single-cell chromatin immunocleavage sequencing (scChIC-seq) to profile histone modification. Nat Methods. 2019;16(4):323–5. https://doi.org/10.1038/s41592-019-0361-7.
158. Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature. 2013;502(7469):59–64. https://doi.org/10.1038/nature12593.
159. Marx V. A dream of single-cell proteomics. Nat Methods. 2019;16(9):809–12. https://doi.org/10.1038/s41592-019-0540-6.
160. Bandura DR, Baranov VI, Ornatsky OI, Antonov A, Kinach R, Lou X, et al. Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. Anal Chem. 2009;81(16):6813–22. https://doi.org/10.1021/ac901049w.
161. Bendall SC, Simonds EF, Qiu P, Amir el AD, Krutzik PO, Finck R, et al. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. Science. 2011;332(6030):687–96. https://doi.org/10.1126/science.1198704.
162. Bodenmiller B, Zunder ER, Finck R, Chen TJ, Savig ES, Bruggner RV, et al. Multiplexed mass cytometry profiling of cellular states perturbed by small-molecule regulators. Nat Biotechnol. 2012;30(9):858–67. https://doi.org/10.1038/nbt.2317.
163. Futamura K, Sekino M, Hata A, Ikebuchi R, Nakanishi Y, Egawa G, et al. Novel full-spectral flow cytometry with multiple spectrally-adjacent fluorescent proteins and fluorochromes and visualization of in vivo cellular movement. Cytometry A. 2015;87(9):830–42. https://doi.org/10.1002/cyto.a.22725.
164. Ferrer-Font L, Pellefigues C, Mayer JU, Small SJ, Jaimes MC, Price KM. Panel design and optimization for high-dimensional Immunophenotyping assays using spectral flow cytometry. Curr Protoc

Cytom. 2020;92(1):e70. https://doi.org/10.1002/cpcy.70.
165. Qiu P, Simonds EF, Bendall SC, Gibbs KD Jr, Bruggner RV, Linderman MD, et al. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. Nat Biotechnol. 2011;29(10):886–91. https://doi.org/10.1038/nbt.1991.
166. Levine JH, Simonds EF, Bendall SC, Davis KL, Amir el AD, Tadmor MD, et al. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. Cell. 2015;162(1):184–97. https://doi.org/10.1016/j.cell.2015.05.047.
167. Gassen SV, Callebaut B, Helden MJV, Lambrecht BN, Demeester P, Dhaene T, et al. FlowSOM: using self-organizing maps for visualization and interpretation of cytometry data. Cytometry A. 2015;87(7):636–45. https://doi.org/10.1002/cyto.a.22625.
168. Krieg C, Nowicka M, Guglietta S, Schindler S, Hartmann FJ, Weber LM, et al. High-dimensional single-cell analysis predicts response to anti-PD-1 immunotherapy. Nat Med. 2018;24(2):144–53. https://doi.org/10.1038/nm.4466.
169. Nowicka M, Krieg C, Weber LM, Hartmann FJ, Guglietta S, Becher B, et al. CyTOF workflow: differential discovery in high-throughput high-dimensional cytometry datasets. F1000Res. 2017;6:748. https://doi.org/10.12688/f1000research.11622.2.
170. Burton RJ, Ahmed R, Cuff S, Baker S, Artemiou A, Eberl M. CytoPy: an autonomous cytometry analysis framework. bioRxiv. 2021:2020.04.08.031898. https://doi.org/10.1101/2020.04.08.031898.
171. Assarsson E, Lundberg M, Holmquist G, Björkesten J, Thorsen SB, Ekman D, et al. Homogenous 96-plex PEA immunoassay exhibiting high sensitivity, specificity, and excellent scalability. PLoS One. 2014;9(4):e95192. https://doi.org/10.1371/journal.pone.0095192.
172. Lollo B, Steele F, Gold L. Beyond antibodies: new affinity reagents to unlock the proteome. Proteomics. 2014;14(6):638–44. https://doi.org/10.1002/pmic.201300187.
173. Petrera A, von Toerne C, Behler J, Huth C, Thorand B, Hilgendorff A, et al. Multiplatform approach for plasma proteomics: complementarity of Olink proximity extension assay technology to mass spectrometry-based protein profiling. J Proteome Res. 2021;20(1):751–62. https://doi.org/10.1021/acs.jproteome.0c00641.
174. Enroth S, Berggrund M, Lycke M, Broberg J, Lundberg M, Assarsson E, et al. High throughput proteomics identifies a high-accuracy 11 plasma protein biomarker signature for ovarian cancer. Commun Biol. 2019;2(1):221. https://doi.org/10.1038/s42003-019-0464-9.
175. Graumann J, Finkernagel F, Reinartz S, Stief T, Brödje D, Renz H, et al. Multi-platform affinity proteomics identify proteins linked to metastasis and immune suppression in ovarian cancer plasma. Front Oncol. 2019;9(1150) https://doi.org/10.3389/fonc.2019.01150.
176. Jurisic V. Multiomic analysis of cytokines in immuno-oncology. Expert Rev Proteomics. 2020;17(9):663–74. https://doi.org/10.1080/14789450.2020.1845654.
177. Bigenwald C, Horowitz A, Navada SC, Odchimar-Reissig R, Rai R, Melana S, et al. Cross talk between the immune compartment and the tumor cells in myelodysplastic syndromes (MDS). Blood. 2019;134(Supplement_1):2986. https://doi.org/10.1182/blood-2019-124887.
178. Method of the year 2019: single-cell multimodal omics. Nat Methods. 2020;17(1):1. https://doi.org/10.1038/s41592-019-0703-5.
179. Remark R, Merghoub T, Grabe N, Litjens G, Damotte D, Wolchok JD, et al. In-depth tissue profiling using multiplexed immunohistochemical consecutive staining on single slide. Sci Immunol. 2016;1(1):aaf6925-aaf. https://doi.org/10.1126/sciimmunol.aaf6925.
180. Tsujikawa T, Kumar S, Borkar RN, Azimi V, Thibault G, Chang YH, et al. Quantitative multiplex immunohistochemistry reveals myeloid-inflamed tumor-immune complexity associated with poor prognosis. Cell Rep. 2017;19(1):203–17. https://doi.org/10.1016/j.celrep.2017.03.037.
181. Giesen C, Wang HAO, Schapiro D, Zivanovic N, Jacobs A, Hattendorf B, et al. Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. Nat Methods. 2014;11(4):417–22. https://doi.org/10.1038/nmeth.2869.
182. Chang Q, Ornatsky OI, Siddiqui I, Loboda A, Baranov VI, Hedley DW. Imaging mass cytometry. Cytometry A. 2017;91(2):160–9. https://doi.org/10.1002/cyto.a.23053.
183. Angelo M, Bendall SC, Finck R, Hale MB, Hitzman C, Borowsky AD, et al. Multiplexed ion beam imaging of human breast tumors. Nat Med. 2014;20(4):436–42. https://doi.org/10.1038/nm.3488.
184. Baharlou H, Canete NP, Cunningham AL, Harman AN, Patrick E. Mass cytometry imaging for the study of human diseases—applications and data analysis strategies. Front Immunol. 2019;10(2657) https://doi.org/10.3389/fimmu.2019.02657.
185. Bodenmiller B. Multiplexed epitope-based tissue imaging for discovery and healthcare applications. Cell Syst. 2016;2(4):225–38. https://doi.org/10.1016/j.cels.2016.03.008.
186. Goltsev Y, Samusik N, Kennedy-Darling J, Bhate S, Hale M, Vazquez G, et al. Deep profiling of mouse splenic architecture with CODEX multiplexed imaging. Cell. 2018;174(4):968–81.e15. https://doi.org/10.1016/j.cell.2018.07.010.
187. Merritt CR, Ong GT, Church SE, Barker K, Danaher P, Geiss G, et al. Multiplex digital spatial profiling of proteins and RNA in fixed tissue. Nat Biotechnol. 2020;38(5):586–99. https://doi.org/10.1038/s41587-020-0472-9.

188. Schulz D, Zanotelli VRT, Fischer JR, Schapiro D, Engler S, Lun XK, et al. Simultaneous multiplexed imaging of mRNA and proteins with subcellular resolution in breast cancer tissue samples by mass cytometry. Cell Syst. 2018;6(1):25–36.e5. https://doi.org/10.1016/j.cels.2017.12.001.
189. Schapiro D, Jackson HW, Raghuraman S, Fischer JR, Zanotelli VRT, Schulz D, et al. histoCAT: analysis of cell phenotypes and interactions in multiplex image cytometry data. Nat Methods. 2017;14(9):873–6. https://doi.org/10.1038/nmeth.4391.
190. Somarakis A, Van Unen V, Koning F, Lelieveldt B, Hollt T. ImaCytE: visual exploration of cellular micro-environments for imaging mass cytometry data. IEEE Trans Vis Comput Graph. 2021;27(1):98–110. https://doi.org/10.1109/tvcg.2019.2931299.
191. Eng C-HL, Lawson M, Zhu Q, Dries R, Koulena N, Takei Y, et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. Nature. 2019;568(7751):235–9. https://doi.org/10.1038/s41586-019-1049-y.
192. Ståhl PL, Salmén F, Vickovic S, Lundmark A, Navarro JF, Magnusson J, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science. 2016;353(6294):78–82. https://doi.org/10.1126/science.aaf2403.
193. Rodriques SG, Stickels RR, Goeva A, Martin CA, Murray E, Vanderburg CR, et al. Slide-seq: a scalable technology for measuring genome-wide expression at high spatial resolution. Science. 2019;363(6434):1463–7. https://doi.org/10.1126/science.aaw1219.
194. Dey SS, Kester L, Spanjaard B, Bienko M, van Oudenaarden A. Integrated genome and transcriptome sequencing of the same cell. Nat Biotechnol. 2015;33(3):285–9. https://doi.org/10.1038/nbt.3129.
195. Macaulay IC, Haerty W, Kumar P, Li YI, Hu TX, Teng MJ, et al. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. Nat Methods. 2015;12(6):519–22. https://doi.org/10.1038/nmeth.3370.
196. Nam AS, Kim K-T, Chaligne R, Izzo F, Ang C, Taylor J, et al. Somatic mutations and cell identity linked by genotyping of transcriptomes. Nature. 2019; https://doi.org/10.1038/s41586-019-1367-0.
197. Kong SL, Li H, Tai JA, Courtois ET, Poh HM, Lau DP, et al. Concurrent single-cell RNA and targeted DNA sequencing on an automated platform for Comeasurement of genomic and transcriptomic signatures. Clin Chem. 2019;65(2):272–81. https://doi.org/10.1373/clinchem.2018.295717.
198. Rodriguez-Meira A, Buck G, Clark S-A, Povinelli BJ, Alcolea V, Louka E, et al. Unravelling Intratumoral heterogeneity through high-sensitivity single-cell mutational analysis and parallel RNA sequencing. Mol Cell. 2019;73(6):1292–305.e8. https://doi.org/10.1016/j.molcel.2019.01.009.
199. Angermueller C, Clark SJ, Lee HJ, Macaulay IC, Teng MJ, Hu TX, et al. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. Nat Methods. 2016;13(3):229–32. https://doi.org/10.1038/nmeth.3728.
200. Hu Y, Huang K, An Q, Du G, Hu G, Xue J, et al. Simultaneous profiling of transcriptome and DNA methylome from a single cell. Genome Biol. 2016;17(1):88. https://doi.org/10.1186/s13059-016-0950-z.
201. Hou Y, Guo H, Cao C, Li X, Hu B, Zhu P, et al. Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. Cell Res. 2016;26(3):304–19. https://doi.org/10.1038/cr.2016.23.
202. Cao J, Cusanovich DA, Ramani V, Aghamirzaie D, Pliner HA, Hill AJ, et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. Science. 2018;361(6409):1380–5. https://doi.org/10.1126/science.aau0730.
203. Satpathy AT, Saligrama N, Buenrostro JD, Wei Y, Wu B, Rubin AJ, et al. Transcript-indexed ATAC-seq for precision immune profiling. Nat Med. 2018;24(5):580–90. https://doi.org/10.1038/s41591-018-0008-8.
204. Clark SJ, Argelaguet R, Kapourani C-A, Stubbs TM, Lee HJ, Alda-Catalinas C, et al. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. Nat Commun. 2018;9(1):781. https://doi.org/10.1038/s41467-018-03149-4.
205. Luo C, Liu H, Wang B-A, Bartlett A, Rivkin A, Nery JR, et al. Multi-omic profiling of transcriptome and DNA methylome in single nuclei with molecular partitioning. bioRxiv. 2018:434845. https://doi.org/10.1101/434845.
206. Liu L, Liu C, Quintero A, Wu L, Yuan Y, Wang M, et al. Deconvolution of single-cell multi-omics layers reveals regulatory heterogeneity. Nat Commun. 2019;10(1):470. https://doi.org/10.1038/s41467-018-08205-7.
207. Reyes M, Billman K, Hacohen N, Blainey PC. Simultaneous profiling of gene expression and chromatin accessibility in single cells. Adv Biosyst. 2019;3(11):1900065. https://doi.org/10.1002/adbi.201900065.
208. Chen S, Lake BB, Zhang K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. Nat Biotechnol. 2019;37(12):1452–7. https://doi.org/10.1038/s41587-019-0290-0.
209. Zhu C, Yu M, Huang H, Juric I, Abnousi A, Hu R, et al. An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. Nat Struct Mol Biol. 2019;26(11):1063–70. https://doi.org/10.1038/s41594-019-0323-x.
210. Luo C, Liu H, Xie F, Armand EJ, Siletti K, Bakken TE, et al. Single nucleus multi-omics links human cortical cell regulatory genome diversity to disease risk variants. bioRxiv. 2019:2019.12.11.873398. https://doi.org/10.1101/2019.12.11.873398.
211. Wang Y, Yuan P, Yan Z, Yang M, Huo Y, Nie Y, et al. Single-cell multiomics sequencing reveals the func-

tional regulatory landscape of early embryos. Nat Commun. 2021;12(1):1247. https://doi.org/10.1038/s41467-021-21409-8.
212. Mateo LJ, Murphy SE, Hafner A, Cinquini IS, Walker CA, Boettiger AN. Visualizing DNA folding and RNA in embryos at single-cell resolution. Nature. 2019;568(7750):49–54. https://doi.org/10.1038/s41586-019-1035-4.
213. Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, et al. Simultaneous epitope and transcriptome measurement in single cells. Nat Methods. 2017;14(9):865–8. https://doi.org/10.1038/nmeth.4380.
214. Peterson VM, Zhang KX, Kumar N, Wong J, Li L, Wilson DC, et al. Multiplexed quantification of proteins and transcripts in single cells. Nat Biotechnol. 2017;35(10):936–9. https://doi.org/10.1038/nbt.3973.
215. Shahi P, Kim SC, Haliburton JR, Gartner ZJ, Abate AR. Abseq: ultrahigh-throughput single cell protein profiling with droplet microfluidic barcoding. Sci Rep. 2017;7(1):44447. https://doi.org/10.1038/srep44447.
216. Chung H, Parkhurst C, Magee EM, Phillips D, Habibi E, Chen F, et al. Simultaneous single cell measurements of intranuclear proteins and gene expression. bioRxiv. 2021:2021.01.18.427139. https://doi.org/10.1101/2021.01.18.427139.
217. Fiskin E, Lareau CA, Eraslan G, Ludwig LS, Regev A. Single-cell multimodal profiling of proteins and chromatin accessibility using PHAGE-ATAC. bioRxiv. 2020:2020.10.01.322420. https://doi.org/10.1101/2020.10.01.322420.
218. Mimitou EP, Lareau CA, Chen KY, Zorzetto-Fernandes AL, Hao Y, Takeshima Y, et al. Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. Nat Biotechnol. 2021; https://doi.org/10.1038/s41587-021-00927-2.
219. Swanson E, Lord C, Reading J, Heubeck AT, Genge PC, Thomson Z, et al. Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq. elife. 2021;10 https://doi.org/10.7554/eLife.63632.
220. Dixit A, Parnas O, Li B, Chen J, Fulco CP, Jerby-Arnon L, et al. Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. Cell. 2016;167(7):1853–66.e17. https://doi.org/10.1016/j.cell.2016.11.038.
221. Jaitin DA, Weiner A, Yofe I, Lara-Astiaso D, Keren-Shaul H, David E, et al. Dissecting immune circuits by linking CRISPR-pooled screens with single-cell RNA-Seq. Cell. 2016;167(7):1883–96.e15. https://doi.org/10.1016/j.cell.2016.11.039.
222. Datlinger P, Rendeiro AF, Schmidl C, Krausgruber T, Traxler P, Klughammer J, et al. Pooled CRISPR screening with single-cell transcriptome readout. Nat Methods. 2017;14(3):297–301. https://doi.org/10.1038/nmeth.4177.
223. Rubin AJ, Parker KR, Satpathy AT, Qi Y, Wu B, Ong AJ, et al. Coupled single-cell CRISPR screening and Epigenomic profiling reveals causal gene regulatory networks. Cell. 2019;176(1):361–76.e17. https://doi.org/10.1016/j.cell.2018.11.022.
224. Mimitou EP, Cheng A, Montalbano A, Hao S, Stoeckius M, Legut M, et al. Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. Nat Methods. 2019;16(5):409–12. https://doi.org/10.1038/s41592-019-0392-0.
225. Codeluppi S, Borm LE, Zeisel A, La Manno G, van Lunteren JA, Svensson CI, et al. Spatial organization of the somatosensory cortex revealed by osmFISH. Nat Methods. 2018;15(11):932–5. https://doi.org/10.1038/s41592-018-0175-z.
226. Wang X, Allen WE, Wright MA, Sylwestrak EL, Samusik N, Vesuna S, et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. Science. 2018;361(6400):eaat5691. https://doi.org/10.1126/science.aat5691.
227. Xia C, Fan J, Emanuel G, Hao J, Zhuang X. Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. Proc Natl Acad Sci. 2019;116(39):19490–9. https://doi.org/10.1073/pnas.1912459116.
228. Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics data integration, interpretation, and its application. Bioinform Biol Insights. 2020;14:1177932219899051. https://doi.org/10.1177/1177932219899051.
229. Peng A, Mao X, Zhong J, Fan S, Hu Y. Single-cell multi-omics and its prospective application in cancer biology. Proteomics. 2020;20(13):1900271. https://doi.org/10.1002/pmic.201900271.
230. Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics. 2009;25(22):2906–12. https://doi.org/10.1093/bioinformatics/btp543.
231. Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics. 2010;26(12):i237–45. https://doi.org/10.1093/bioinformatics/btq182.
232. Mo Q, Wang S, Seshan VE, Olshen AB, Schultz N, Sander C, et al. Pattern discovery and cancer gene identification in integrated cancer genomic data. Proc Natl Acad Sci U S A. 2013;110(11):4245–50. https://doi.org/10.1073/pnas.1208949110.
233. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods. 2014;11(3):333–7. https://doi.org/10.1038/nmeth.2810.
234. Wu D, Wang D, Zhang MQ, Gu J. Fast dimension reduction and integrative clustering of multi-omics

data using low-rank approximation: application to cancer molecular classification. BMC Genomics. 2015;16(1):1022. https://doi.org/10.1186/s12864-015-2223-8.
235. Yang Z, Michailidis G. A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. Bioinformatics. 2016;32(1):1–8. https://doi.org/10.1093/bioinformatics/btv544.
236. Welch JD, Hartemink AJ, Prins JF. MATCHER: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. Genome Biol. 2017;18(1):138. https://doi.org/10.1186/s13059-017-1269-0.
237. Gabasova E, Reid J, Wernisch L. Clusternomics: integrative context-dependent clustering for heterogeneous datasets. PLoS Comput Biol. 2017;13(10):e1005781. https://doi.org/10.1371/journal.pcbi.1005781.
238. Rohart F, Gautier B, Singh A, KA LC. mixOmics: An R package for 'omics feature selection and multiple data integration. PLoS Comput Biol. 2017;13(11):e1005752. https://doi.org/10.1371/journal.pcbi.1005752.
239. Champion M, Brennan K, Croonenborghs T, Gentles AJ, Pochet N, Gevaert O. Module analysis captures Pancancer genetically and epigenetically deregulated cancer driver genes for smoking and antiviral response. EBioMedicine. 2018;27:156–66. https://doi.org/10.1016/j.ebiom.2017.11.028.
240. Nguyen H, Shrestha S, Draghici S, Nguyen T. PINSPlus: a tool for tumor subtype discovery in integrated genomic data. Bioinformatics. 2018;35(16):2843–6. https://doi.org/10.1093/bioinformatics/bty1049.
241. Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, et al. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. Mol Syst Biol. 2018;14(6):e8124. https://doi.org/10.15252/msb.20178124.
242. Rappoport N, Shamir R. NEMO: cancer subtyping by integration of partial multi-omic data. Bioinformatics. 2019;35(18):3348–56. https://doi.org/10.1093/bioinformatics/btz058.
243. Meng C, Basunia A, Peters B, Gholami AM, Kuster B, Culhane AC. MOGSA: integrative single sample gene-set analysis of multiple omics data. Mol Cell Proteomics. 2019;18(8 suppl 1):S153–s68. https://doi.org/10.1074/mcp.TIR118.001251.
244. Liu J, Huang Y, Singh R, Vert J-P, Noble WS. Jointly embedding multiple single-cell omics measurements. bioRxiv. 2019:644310. https://doi.org/10.1101/644310.
245. Argelaguet R, Arnol D, Bredikhin D, Deloro Y, Velten B, Marioni JC, et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biol. 2020;21(1):111. https://doi.org/10.1186/s13059-020-02015-1.
246. Cao K, Bai X, Hong Y, Wan L. Unsupervised topological alignment for single-cell multi-omics integration. Bioinformatics. 2020;36(Supplement_1):i48–i56. https://doi.org/10.1093/bioinformatics/btaa443.
247. Kim HJ, Lin Y, Geddes TA, Yang JYH, Yang P. CiteFuse enables multi-modal analysis of CITE-seq data. Bioinformatics. 2020; https://doi.org/10.1093/bioinformatics/btaa282.
248. Andersson A, Bergenstråhle J, Asp M, Bergenstråhle L, Jurek A, Fernández Navarro J, et al. Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. Commun Biol. 2020;3(1):565. https://doi.org/10.1038/s42003-020-01247-y.
249. Hao Y, Hao S, Andersen-Nissen E, Mauck WM 3rd, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021; https://doi.org/10.1016/j.cell.2021.04.048.
250. Gayoso A, Steier Z, Lopez R, Regier J, Nazor KL, Streets A, et al. Joint probabilistic modeling of single-cell multi-omic data with totalVI. Nat Methods. 2021;18(3):272–82. https://doi.org/10.1038/s41592-020-01050-x.
251. Puram SV, Tirosh I, Parikh AS, Patel AP, Yizhak K, Gillespie S, et al. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. Cell. 2017;171(7):1611–24.e24. https://doi.org/10.1016/j.cell.2017.10.044.
252. Kumar MP, Du J, Lagoudas G, Jiao Y, Sawyer A, Drummond DC, et al. Analysis of single-cell RNA-Seq identifies cell-cell communication associated with tumor characteristics. Cell Rep. 2018;25(6):1458–68.e4. https://doi.org/10.1016/j.celrep.2018.10.047.
253. Yuan D, Tao Y, Chen G, Shi T. Systematic expression analysis of ligand-receptor pairs reveals important cell-to-cell interactions inside glioma. Cell Commun Signal. 2019;17(1):48. https://doi.org/10.1186/s12964-019-0363-1.
254. Finotello F, Rieder D, Hackl H, Trajanoski Z. Next-generation computational tools for interrogating cancer immunity. Nat Rev Genet. 2019;20(12):724–46. https://doi.org/10.1038/s41576-019-0166-7.
255. Yeung T-L, Sheng J, Leung CS, Li F, Kim J, Ho SY, et al. Systematic identification of Druggable epithelial–stromal crosstalk signaling networks in ovarian cancer. J Natl Cancer Inst. 2018;111(3):272–82. https://doi.org/10.1093/jnci/djy097.
256. Shao X, Lu X, Liao J, Chen H, Fan X. New avenues for systematically inferring cell-cell communication: through single-cell transcriptomics data. Protein Cell. 2020;11(12):866–80. https://doi.org/10.1007/s13238-020-00727-5.
257. Armingol E, Officer A, Harismendy O, Lewis NE. Deciphering cell–cell interactions and communication from gene expression. Nat Rev Genet. 2020; https://doi.org/10.1038/s41576-020-00292-x.
258. Ramilowski JA, Goldberg T, Harshbarger J, Kloppmann E, Lizio M, Satagopam VP, et al. A draft network of ligand–receptor-mediated multicellular signalling in human. Nat Commun. 2015;6(1):7866. https://doi.org/10.1038/ncomms8866.

259. Fernandez DM, Rahman AH, Fernandez NF, Chudnovskiy A, Amir ED, Amadori L, et al. Single-cell immune landscape of human atherosclerotic plaques. Nat Med. 2019;25(10):1576–88. https://doi.org/10.1038/s41591-019-0590-4.
260. Martin JC, Chang C, Boschetti G, Ungaro R, Giri M, Grout JA, et al. Single-cell analysis of Crohn's disease lesions identifies a pathogenic cellular module associated with resistance to anti-TNF therapy. Cell. 2019;178(6):1493–508 e20. https://doi.org/10.1016/j.cell.2019.08.008.
261. Efremova M, Vento-Tormo M, Teichmann SA, Vento-Tormo R. CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. Nat Protoc. 2020; https://doi.org/10.1038/s41596-020-0292-x.
262. Vento-Tormo R, Efremova M, Botting RA, Turco MY, Vento-Tormo M, Meyer KB, et al. Single-cell reconstruction of the early maternal-fetal interface in humans. Nature. 2018;563(7731):347–53. https://doi.org/10.1038/s41586-018-0698-6.
263. Camp JG, Sekine K, Gerber T, Loeffler-Wirth H, Binder H, Gac M, et al. Multilineage communication regulates human liver bud development from pluripotency. Nature. 2017;546(7659):533–8. https://doi.org/10.1038/nature22796.
264. Pavličev M, Wagner GP, Chavan AR, Owens K, Maziarz J, Dunn-Fletcher C, et al. Single-cell transcriptomics of the human placenta: inferring the cell communication network of the maternal-fetal interface. Genome Res. 2017;27(3):349–61. https://doi.org/10.1101/gr.207597.116.
265. Zhou JX, Taramelli R, Pedrini E, Knijnenburg T, Huang S. Extracting intercellular signaling network of cancer tissues using ligand-receptor expression patterns from whole-tumor and single-cell transcriptomes. Sci Rep. 2017;7(1):8815. https://doi.org/10.1038/s41598-017-09307-w.
266. Boisset J-C, Vivié J, Grün D, Muraro MJ, Lyubimova A, van Oudenaarden A. Mapping the physical network of cellular interactions. Nat Methods. 2018;15(7):547–53. https://doi.org/10.1038/s41592-018-0009-z.
267. Xiong Z, Yang Q, Li X. Effect of intra- and inter-tumoral heterogeneity on molecular characteristics of primary IDH-wild type glioblastoma revealed by single-cell analysis. CNS Neurosci Ther. 2020;26(9):981–9. https://doi.org/10.1111/cns.13396.
268. Corridoni D, Antanaviciute A, Gupta T, Fawkner-Corbett D, Aulicino A, Jagielowicz M, et al. Single-cell atlas of colonic CD8+ T cells in ulcerative colitis. Nat Med. 2020;26(9):1480–90. https://doi.org/10.1038/s41591-020-1003-4.
269. Ghorani E, Reading JL, Henry JY, Massy MR, Rosenthal R, Turati V, et al. The T cell differentiation landscape is shaped by tumour mutations in lung cancer. Nat Cancer. 2020;1(5):546–61. https://doi.org/10.1038/s43018-020-0066-y.
270. Lee H-O, Hong Y, Etlioglu HE, Cho YB, Pomella V, Van den Bosch B, et al. Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. Nat Genet. 2020;52(6):594–603. https://doi.org/10.1038/s41588-020-0636-z.
271. Park J-E, Botting RA, Domínguez Conde C, Popescu D-M, Lavaert M, Kunz DJ, et al. A cell atlas of human thymic development defines T cell repertoire formation. Science. 2020;367(6480):eaay3224. https://doi.org/10.1126/science.aay3224.
272. Chua RL, Lukassen S, Trump S, Hennig BP, Wendisch D, Pott F, et al. COVID-19 severity correlates with airway epithelium–immune cell interactions identified by single-cell analysis. Nat Biotechnol. 2020;38(8):970–9. https://doi.org/10.1038/s41587-020-0602-4.
273. Zhang M, Yang H, Wan L, Wang Z, Wang H, Ge C, et al. Single-cell transcriptomic architecture and intercellular crosstalk of human intrahepatic cholangiocarcinoma. J Hepatol. 2020;73(5):1118–30. https://doi.org/10.1016/j.jhep.2020.05.039.
274. Jin S, Guerrero-Juarez CF, Zhang L, Chang I, Myung P, Plikus MV, et al. Inference and analysis of cell-cell communication using CellChat. bioRxiv. 2020:2020.07.21.214387. https://doi.org/10.1101/2020.07.21.214387.
275. Noël F, Massenet-Regad L, Carmi-Levy I, Cappuccio A, Grandclaudon M, Trichot C, et al. ICELLNET: a transcriptome-based framework to dissect intercellular communication. bioRxiv. 2020:2020.03.05.976878. https://doi.org/10.1101/2020.03.05.976878.
276. Hou R, Denisenko E, Ong HT, Ramilowski JA, Forrest ARR. Predicting cell-to-cell communication networks using NATMI. Nat Commun. 2020;11(1):5011. https://doi.org/10.1038/s41467-020-18873-z.
277. Cabello-Aguilar S, Alame M, Kon-Sun-Tack F, Fau C, Lacroix M, Colinge J. SingleCellSignalR: inference of intercellular networks from single-cell transcriptomics. Nucleic Acids Res. 2020; https://doi.org/10.1093/nar/gkaa183.
278. Alame M, Cornillot E, Cacheux V, Rigau V, Costes-Martineau V, Lacheretz-Szablewski V, et al. The immune landscape of primary central nervous system diffuse large B cell lymphoma. bioRxiv. 2020:2020.08.17.254284. https://doi.org/10.1101/2020.08.17.254284.
279. Cillo AR, Kürten CHL, Tabib T, Qi Z, Onkar S, Wang T, et al. Immune landscape of viral- and carcinogen-driven head and neck cancer. Immunity. 2020;52(1):183–99.e9. https://doi.org/10.1016/j.immuni.2019.11.014.
280. Choi H, Sheng J, Gao D, Li F, Durrans A, Ryu S, et al. Transcriptome analysis of individual stromal cell populations identifies stroma-tumor crosstalk in mouse lung cancer model. Cell Rep. 2015;10(7):1187–201. https://doi.org/10.1016/j.celrep.2015.01.040.

281. Wang Y, Wang R, Zhang S, Song S, Jiang C, Han G, et al. iTALK: an R package to characterize and illustrate intercellular communication. bioRxiv. 2019:507871. https://doi.org/10.1101/507871.
282. Tyler SR, Rotti PG, Sun X, Yi Y, Xie W, Winter MC, et al. PyMINEr finds gene and Autocrine-paracrine networks from human islet scRNA-Seq. Cell Rep. 2019;26(7):1951–64.e8. https://doi.org/10.1016/j.celrep.2019.01.063.
283. Cang Z, Nie Q. Inferring spatial and signaling relationships between cells from single cell transcriptomic data. Nat Commun. 2020;11(1):2084. https://doi.org/10.1038/s41467-020-15968-5.
284. Arnol D, Schapiro D, Bodenmiller B, Saez-Rodriguez J, Stegle O. Modeling cell-cell interactions from spatial molecular data with spatial variance component analysis. Cell Rep. 2019;29(1):202–11.e6. https://doi.org/10.1016/j.celrep.2019.08.077.
285. Baccin C, Al-Sabah J, Velten L, Helbling PM, Grünschläger F, Hernández-Malmierca P, et al. Combined single-cell and spatial transcriptomics reveal the molecular, cellular and spatial bone marrow niche organization. Nat Cell Biol. 2019; https://doi.org/10.1038/s41556-019-0439-6.
286. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinform. 2008;9(1):559. https://doi.org/10.1186/1471-2105-9-559.
287. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol. 2005;4:Article17. https://doi.org/10.2202/1544-6115.1128.
288. Browaeys R, Saelens W, Saeys Y. NicheNet: modeling intercellular communication by linking ligands to target genes. Nat Methods. 2020;17(2):159–62. https://doi.org/10.1038/s41592-019-0667-5.
289. Ji AL, Rubin AJ, Thrane K, Jiang S, Reynolds DL, Meyers RM, et al. Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. Cell. 2020;182(2):497–514.e22. https://doi.org/10.1016/j.cell.2020.05.039.
290. Sathe A, Grimes SM, Lau BT, Bai X, Chen J, Suarez C, et al. The cellular genomic diversity, regulatory states and networking of the metastatic colorectal cancer microenvironment. bioRxiv. 2020;2020.09.01.273672 https://doi.org/10.1101/2020.09.01.273672.
291. Wang S, Karikomi M, MacLean AL, Nie Q. Cell lineage and communication network inference via optimization for single-cell transcriptomics. Nucleic Acids Res. 2019;47(11):e66–e. https://doi.org/10.1093/nar/gkz204.

Index

A
AbsCN-seq, 62
Absolute, 62
Actionable alteration, 1, 9, 103
Activity score, 157
Acute lymphoblastic leukemia (ALL), 164
Acute myeloid leukemia (AML), 254
Ado-trastuzumab, 58
AI-powered image analysis, 257, 258, 261
Air-liquid interface (ALI), 220
Alignment, 167, 168
AluScanCNV2, 63
Amazon Web Services (AWS), 25, 30, 32
AmCAD-UT, 262
Amphiregulin, 103
Aneuploidy, 62
Annotation
  RNA-seq fusion detection, 171
Annotators
  cancer, 187
    ANNOVAR, 189, 190
    Oncotator and Funcotator, 190, 191
    SnpEff and SnpSift, 191, 192
    VEP, 187–189, 192
ANNOVAR, 47, 189, 190
Anomalous paired read signals, RNA-seq fusion detection, 169
Arriba, 170
Arterys Oncology DL, 262
Artifacts, 170
Artificial intelligence (AI), 249, 250
  AI-powered image analysis, 257, 258, 261
  cancer, powered drug prioritization in, 253, 254, 257
  in cancer subtype identification, 252, 253
  challenges, 263
  drug sensitivity, 255–256
  medical image analysis, field of, 259–260
  public data resources powering precision oncology, 250–252
  US FDA, AI-related models in, 262, 263
Arvados, 24, 27, 34
Atypical teratoid rhabdoid tumor (ATRT), 253
Autoencoder-based method, 254

B
BalestraWeb, 126, 134
BAMSurgeon, 116
Battenberg, 16
Bayesian ANalysis to Determine Drug Interaction Targets (BANDIT), 130, 131, 135
Bayesian mixture model, 109
Bcbio-nextgen, 27
BCR-ABL chimeric protein, 9
BCR-ABL gene, 9
Bethesda/NCI panel, 76
Binary Alignment Map (BAM), 5, 29, 43
Binary version, 29
Binding site parametrization, drug repurposing, 129, 130
Bioinformatics, 76, 81, 82, 94
Biological pathways databases
  pathway analysis, for cancer research and precision oncology applications, 150, 151
Biomarkers, 4, 8–11, 15–17
BitBucket integration, 25
BMTMKL, 254
BRC-ABL1 fusion, 164
Breakpoint, 168–171
Broad consensus, 25
Broad Institute Best Practices Workflows (BroadBPW), 4, 5
Broad Institute's Terra Bio Cloud Platform, 24
Bulk DNA sequencing
  inferring tumor's clonal landscape from, 104–106
    CALDER, 111
    Canopy, 110
    CloneHD, 109
    FastClone, 112
    Meltos, 112
    Palimpsest, 110
    PhyloWGS, 109
    PyClone, 106
    QuantumClone, 110
    SciClone, 109
    SPRUCE, 109, 110
    SubMARine, 112, 113
    SuperFreq, 112
    SVclone, 111
    Tusv, 111

© Springer Nature Switzerland AG 2022

A. Laganà (ed.), *Computational Methods for Precision Oncology*, Advances in Experimental Medicine and Biology 1361, https://doi.org/10.1007/978-3-030-91836-1

Bulk RNA-seq, ITH from, 113
Bulk sequencing for ITH estimates, in precision oncology, 270, 271
Bulk whole-genome (WGS), 104
Burrows–Wheeler Transform (BWT) algorithm, 5

C
CADD, 14, 194
CALDER, 111
Cancer
  annotators, 187
    ANNOVAR, 189, 190
    Oncotator and Funcotator, 190, 191
    SnpEff and SnpSift, 191, 192
    VEP, 187–189, 192
  interpretation of variations in, 177, 178
    CGI, 179, 183, 184
    CIViC, 178, 180
    ClinVar, 187
    COSMIC, 184, 185, 187
    databases, 178
    databases consensus, 184–186
    DoCM, 182
    OncoKB, 181, 182
    pharmacogenomics knowledgebase, 180, 181
    PMKB, 182, 183
    VICC project, 178
  pathogenicity predictors, 194
    conservation score, tools for, 195
    functional prediction scores, tools for, 194, 195
  patient-derived in vitro and in vivo models of, 216–218
  powered drug prioritization in, 253, 254, 257
  variant prioritization, 192, 193
    Endeavour, 193
    GeneDistiller, 193
    KGGSeq, 193
    MutationDistiller, 193
    VINYL, 193
Cancer Biomarkers database, 184
Cancer cell fraction (CCF), 105
Cancer Cell Line Encyclopedia (CCLE), 146, 191, 250
Cancer evolution, 101, 103
Cancer Gene Census (CGC), 185
Cancer Genome Atlas and International Cancer Genomics Consortium, 199
Cancer Genome Interpreter (CGI), 149, 179, 183, 184
Cancer genomes, 105
Cancer research and precision oncology applications, pathway analysis for
  biological pathways databases, 150, 151
  omics data source, 144
    cancer “omics” projects data, 146, 148
    precision oncology, knowledge bases for, 148–150
    proteomics and metabolomics data repositories, 145, 146
    sequencing data repositories, 144, 145
  patients to pathways, 143, 144
  strategies
    applications, 155, 157
    pathway analysis methods, 151–155
    phenotype and therapy predictions, 157, 158
Cancer subtype identification, AI, 252, 253
Cancer Transcriptome Analysis Toolkit (CTAT), 50
Catalog of Cancer Genes, 184
Catalog of Validated Oncogenic Mutations, 184
Catalogue of Somatic Mutations In Cancer (COSMIC), 58, 184, 185, 187
Cell-cell signaling
  single-cell sequencing technologies, 279
Cell-free DNA (cfDNA)
  precision oncology, molecular profiling, 238, 239
    identification and analysis, methods, 239
    studies of, 239, 240
CellPhoneDB, 279
Cellular heterogeneity, 236
Cerebro, 44
Change-point detection, 68
Chemical Similarity Network Analysis Pull-down (CSNAP), 126, 134
ChemMapper, 122, 133
ChemProt 3.0, 122, 123, 133
Chimeric artifact molecules, 169
Chimeric proteins, 165
Chimeric transcripts, 166
Chorioallontoic membrane (CAM), 220, 221
Chronic lymphocytic leukemia (CLL), 103
Chronic myelogenous leukemia (CML), 9
CICERO, 166, 170
Circular binary segmentation (CBS), 68
Circular tumor DNA (ctDNA)
  precision oncology, molecular profiling, 238, 239
    identification and analysis, methods, 239
    studies of, 239, 240
Circulating tumor cells (CTCs)
  precision oncology, molecular profiling, 235, 236
    identification and analysis, methods for, 236, 237
    studies of, 237, 238
Clinical evidence, 178
Clinical Interpretations of Variants in Cancer (CIViC), 15, 149, 178, 180
ClinVar, 187
Clonal hematopoiesis of indeterminant potential (CHIP), 38
Clonal landscape, 113, 114
Clone-Align, 278
CloneHD, 109
CLUE—CMap Linked User Environment, 138
Clustering, 199, 202–206, 238
Cluster-of-Cluster-Assignments (COCA), 253
CMMRD syndrome, 76, 91, 93
cmTriage, 262
Common Workflow Language (CWL), 24, 25, 27, 28, 31, 33
Comparative genomic hybridization (CGH), 56
Computational methods, 81, 83–85, 94
Connectivity Map (CMap) project, 12, 131, 135, 136, 252

L1000, 137, 138
L1000CDS2, 138, 139
method, 136, 137
Containerization, 25
Convolutional neural networks (CNN), 257, 261, 263
Copy number alterations (CNAs), 7, 28, 105
Copy number variation (CNV)
detection algorithms and tools, 63–66, 68
detection methods
paired-end (PE) approaches, 59
read depth (RD) approaches, 59
read depth–based CNV detection methods, 61
split read (SR) methods, 59
heterogeneity of CNV profiles, 63
lack of gold standard, 62
noisy sequencing data, 61
precision oncology, 57–59
sequencing technical problems, 61
tumor complexity, 62
Cortes-Ciriano method, 90
COSMIC-3D, 185
CWL-runner execution, 31
CWL script, 32

D
DASPfind, 126, 127, 134
Database gene-variant intersection, 186
Database of curated mutations in cancer (DoCM), 182
Databases consensus, 184–186
Dataflow programming paradigm, 27
Data preprocessing, 274
Data processing platform
arvados, 27
Bcbio-nextgen, 27
DNAnexus, 27
Terra.bio, 27
DawnRank, 207, 208
dbNSFP, 195
Deleterious annotation of genetic variants using neural networks (DANN), 194
DeepDSC, 254
Deep learning, 210, 254
drug sensitivity, 255–256
medical image analysis, field of, 259–260
Deep neural networks (DNNs), 49, 257
DeepSynergy, 257
DeepVariant, 41
DEFB105A/B, 89
De novo graph-based assembly algorithms, 167
De novo transcriptome assembly
RNA-seq fusion detection, 167
DepMap, 252
DeSigN, 131, 135
DIANA-miRPath, 152
Digital breast tomosynthesis (DBT), 263
Directed acyclic graph (DAG), 27
Disease clones, 270
Divide et impera docking approach, 125
DNA Databank of Japan (DDBJ), 144
DNA mismatch repair system (dMMR), 75, 91–93
DNAnexus, 27
DNA sequencing data
copy number alterations, 7
genomic instability, 7
germline variant calling in precision oncology, 7, 8
inter-tumor heterogeneity, 8, 9
microsatellite instability, 7
post-processing of read alignments, 5
pre-processing of sequencing data, 4, 5
reads mapping, 5
short indels calling, 6
somatic single nucleotide variants, 6
DolphinNext, 27
Domain-specific language (DSL), 25
Droplet-based assays, 272
Drug-induced gene expression
drug repurposing using, 131
CMap, 131
DeSigN, 131
GoPredict, 131
MANTRA 2.0, 131, 132
NFFinder, 132
PDOD, 132
RGES, 132
Drug prioritization in cancer, 253, 254, 257
Drug recommendation, 2, 3, 16
Drug repurposing, 119, 120
data sources for, 132, 136
CMap, 136, 137
L1000, 137, 138
L1000CDS2, 138, 139
using drug-induced gene expression, 131
CMap, 131
DeSigN, 131
GoPredict, 131
MANTRA 2.0, 131, 132
NFFinder, 132
PDOD, 132
RGES, 132
methods and approaches, 120
screening methods/blinded research
knowledge -based drug repurposing methods, 121
pathway-or network-based drug repurposing methods, 121
signature-based drug repurposing methods, 121
target-based drug repurposing methods, 121
targeted mechanism-based drug repurposing methods, 121, 122
web-based solutions, 122
web-based tools, 122
BANDIT, 130, 131
binding site parametrization, 129, 130
ligand similarity using fingerprint encoding, 122–125
MeSHDD, 130
network-based approaches, 126–129
RE:fine drugs, 130
3D structures of drugs and targets, 125, 126
Drug response prediction, 254

Drug sensitivity, AI, 255–256
Drug sensitivity prediction resources, 252
DT-Web, 128, 134
Due process, 25
Dynamic Read Analysis for GENomics (DRAGEN) platform, 42

E
EGFR mutation, 261
Endeavour, 193
Endosomal sorting complexes required for transport (ESCRT), 240
Ensembl, 192
Ensembl Regulatory Build, 192
Ensembl variant effect predictor (VEP), 192
Ensemble machine learning, 49
Ensemble strategies, RNA-seq fusion detection, 174
Epithelial cell adhesion molecules (EPCAMs), 236
ETV6-RUNX1 fusion, 173
European Molecular Biology Laboratory (EMBL), 144
European Society for Medical Oncology (ESMO), 57
Exosomes, 240
Expectation Maximization (EM) algorithm, 110
Expression abnormalities, RNA-seq fusion detection, 169, 170
Extracellular vesicles (EVs)
  precision oncology, molecular profiling, 240, 241
    identification and analysis, methods, 241
    studies on, 241, 242

F
False discovery rate (FDR), 137
FastClone, 112
FATHMM, 14, 194
Feature engineering, 210
Feature selection, 210
Fibroblast growth factor (FGF), 206
Filtration, 170
FISH/Cytogenetics assays, 4
FitCons, 195
Fluorescence in situ hybridization (FISH), 296
5'fluorouracil chemotherapy, 92
FOLFOX, 222
FoundationOne CDx (F1CDx), 59
Fragments per Kilobase per Million (FPKM), 10
Freebayes, 41
Fujimoto method, 89
5-FU, 222
FuncAssociate, 152
Funcotator, 190, 191
Functional class scoring (FCS, 12, 151, 153
Functional genomics, 217

G
Gallon method 2, 91
GATK HaplotypeCaller, 8, 38, 45
GATK:IndelRealigner, 6
GATK-RealignerTargetCreator, 6
GATK VQSR method, 45
Gaussian mixture, 68
GeneDistiller, 193
Gene Expression Omnibus (GEO), 137
Gene interaction network, 201, 205, 208
GeneMerge, 152
Gene Ontology annotations, 190
Gene set enrichment analysis (GSEA), 153, 274
Gene Set Ensemble Approach (GSEA) based method, 132
Gene set variation analysis (GSVA), 13, 274
Gene transfer format (GTF), 9
Genome Analysis ToolKit (GATK), 5
Genomic alterations, 273
Genomic models
  value of interpretability and prior knowledge in, 203
    hypothesis generation and rational treatment design, 204
    signal-to-noise with sparse data, 204
    small sample size and overfitting, 203, 204
Genomic VCF (GVCF) file, 8
Germline and somatic variant
  SNV/indel variant
    algorithm basis of, 38, 41
    Bogus somatic variant calling, factors for, 47, 48
    data preprocessing, 42
    development of, 38
    germline mutation calling and prioritization, 45, 46
    matched tumor-normal pairs, 43, 44
    mutation calling and prioritization, 44, 45
    variant annotation, 46, 47
  somatic SNV/Indel variant
    algorithm basis of, 41, 42
Germline CNVs, 56
Germline variants, 7, 8, 37, 48
GERP++, 195
GISTIC, 68
GISTIC2, 68
GitHub, 25
Global Alliance for Genomic and Health (GA4GH), 15, 178, 182
GlobalANCOVA, 153
Global Proteome Machine Database (GPMDB), 145
Google Cloud Platform (GCP), 25, 30, 32
GoPredict, 131, 135
GOToolBox, 152
GRCh38, 42, 171
GSA, 153

H
Hause method (mSING + 1 Locus) MOSAIC, 90
HER2, 104
Heterogeneity, of sequence data sources, 166
Hidden Markov models (HMMs), 68, 109
High-throughput drug screening, 220
High-throughput sequencing (HTS), 4, 249, 250, 263
HitPick, 123, 133

Homology, 164, 167
H-RACS, 257
Human Cancer Models Initiative (HCMI), 148
Human Genome Variation Society (HGVS), 182
Human Metabolome Database (HMDB), 146

I
iDrug-Target, 123, 125, 133
Illumina®, 42
Immune checkpoint blockade (ICB), 222
Immunohistochemical (IHC) analysis, 78, 91, 92
Inception-v3, 261
Individualized Differentially Expressed Genes (iDEG), 11
Inferential association, 178
Inferring tumor's clonal landscape
  from bulk DNA sequencing, 104–106
    CALDER, 111
    Canopy, 110
    CloneHD, 109
    FastClone, 112
    Meltos, 112
    Palimpsest, 110
    PhyloWGS, 109
    PyClone, 106
    QuantumClone, 110
    SciClone, 109
    SPRUCE, 109, 110
    SubMARine, 112, 113
    SuperFreq, 112
    SVclone, 111
    Tusv, 111
Infinite-sites assumption (ISA), 105
Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT), 58
Internal tandem duplications (ITDs), 165
International Cancer Genome Consortium (ICGC), 250
International Nucleotide Sequence Database Collaboration (INSDC), 144
Inter-tumor heterogeneity, 8, 9
Intra-tumor heterogeneity (ITH), 8, 101, 103, 269, 270
ITH
  assessment of, 114, 116
  bulk DNA sequencing, inferring tumor's clonal landscape from, 104–106
    CALDER, 111
    CloneHD, 109
    FastClone, 112
    Meltos, 112
    Palimpsest, 110
    PhyloWGS, 109
    PyClone, 106
    QuantumClone, 110, 111
    SciClone, 109
    SPRUCE, 109, 110
    SubMARine, 112, 113
    SuperFreq, 112
    SVclone, 111
    Tusv, 111
  from Bulk RNA-Seq, 113
  bulk *vs.* single-cell sequencing for, 270, 271
  and precision oncology, 103, 104
  RNA-seq for, 271
  with scDNA-seq, 276, 277
  with single-cell RNA-seq, 273, 274
  visualization of, 115
  visualizing clonal landscape and tumor clonal evolution, 113, 114

J
Jackson Laboratory Clinical Knowledgebase (JAX CKB), 15
JavaScript Object Notation file (JSON), 139

K
Kenpaullone, 139
Kernelized Bayesian matrix factorization (KBMF) method, 254
KGGSeq, 193
Knowledge-based drug repurposing methods, 121
Kolmogorov–Smirnov statistical test, 82, 90
KRAS G12D, 217
Kyoto Encyclopedia of Genes and Genomes (KEGG), 144, 150

L
L1000, 137, 138
L1000CDS2, 138, 139
Label propagation, 207, 210
Larotrectinib, 253
Learning algorithms, 204
Library of Integrated Network–based Cellular Signatures (LINCS) program, 136, 252
Ligand-Receptor(LR) interactions, 279
Ligand similarity
  using fingerprint encoding
    ChemMapper, 122
    ChemProt 3.0, 122, 123
    HitPick, 123
    iDrug-Target, 123
    Polypharmacology Browser, 124
    SEA, 124
    SuperPred, 124
    SwissTargetPrediction, 124
    TargetHunter, 125
    TarPred, 125
Likelihood Ratio Test (LRT), 194
Limit of detection (LOD) of MSI, 76, 78, 94
Liquid biopsies
  molecular profiling of
    cfDNA and ctDNA, 238–240
    circulating tumor cells, 235–238
    extracellular vesicles, 240–242
Load Sharing Facility (LSF), 25
Locally estimated scatterplot smoothing, 68
Loess regression, 68

Long-read variant calling, 49
Loss-Of-Function Transcript Effect Estimator (LOFTEE), 47
Low-frequency variants, 48
Lung adenocarcinoma (LUAD), 258
Lung adenocarcinoma PDX biopsies, 222
Lung squamous cell carcinoma (LUSC), 258
Lynch syndrome, 91

M
Machine learning (ML), 44, 76, 89, 210
  drug sensitivity, 255–256
  medical image analysis, field of, 259–260
  ML–based callers, 44
MANTIS, 81, 82
MANTRA 2.0, 131, 132, 135
MAPPFinder, 152
Mapping, 169
MarginPhase, 49
Markov chain Monte Carlo (MCMC), 110
MATQ-seq, 272
Medical image analysis, AI, 259–260
Meltos, 112
Memorial Sloan Kettering Cancer Center (MSKCC), 59
Memorial Sloan Kettering–Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT), 239
Memorial Sloan-Kettering OncoKB, 15
MeSHDD, 135
MetaLR, 194
MetaSVM, 194
Microfluidic platforms, 222, 223
MicroRNAs (miRNAs), 14
Microsatellite instability (MSI) from NGS data
  blood and plasma samples assessment, 93, 94
  computational approaches for challenges and difficulties, 81, 82
  Cortes-Ciriano method, 90
  Fujimoto method, 89
  Gallon method 2, 91
  Hause method, 90
  MANTIS, 82
  MIRMMR, 90
  MSIcall, 88
  MSI-ColonCore, 88
  mSINGS, 82
  MSIpred, 89
  MSIsensor, 82
  MSIsensor-pro, 88
  MSI-seq Index, 88
  MSIseq/NGS classifier, 90
  Nowak method, 89
  pan-cancer at WGS, 77
  phenotype for cancer diagnosis, 91
  Redford and Gallon method, 91
  refined microsatellite panels for MSI, 77, 81
  sensitive microsatellite markers, 76, 77
  treatment response and therapeutic decision-making, 92
  tumor assessment, 92, 93
Microsatellite instability regression using methylation and mutations in R (MIRMMR), 90
Microsoft Azure, 30
Mining junction/clipped reads, RNA-seq fusion detection, 167–169
miRNA-sensitive topological pathway analysis, 154, 155
Mitochondria mutation calling, 48, 49
MixEnrich, 13
MMR genes, 76, 87, 92
MMRF CoMMpass Study, 148
Modeling ITH, 106
MolecularMatch, 15
Mono-nucleotide microsatellites, 77, 89
MORONET, 207
MOSAIC, 91
MSH6 deficiency, 93
MSIcall, 88
MSI-ColonCore, 82, 88
MSI detection method, 76
mSINGs, 81, 82, 90
MSIpred, 89
MSIsensor, 81, 82
MSIsensor-pro, 88
MSIseq/NGS classifier, 90
Multimodal omics methods, 297
Multiple myeloma (MM), 16, 17, 103
Multiplexed IHC by iterative staining of single slides (MICSSS), 296
Multivesicular endosomes (MVEs), 241
MuSE, 41
MutationAssessor, 14, 194
MutationDistiller, 193
Mutations, 3, 6–9, 14, 17
MutationTaster2, 194
MuTect2, 6, 41, 44
My cancer genome (MCG), 150

N
nAnnolyze, 127, 134
National Cancer Institute (NCI), 76–78, 92
National Center for Biotechnology Information (NCBI), 144, 187
National Comprehensive Cancer Network (NCCN), 15
NCI-ALMANAC database, 257
netDx, 206, 207
NetGSA, 154
Network-based approaches
  drug repurposing
    BalestraWeb, 126
    CSNAP, 126
    DASPfind, 126, 127
    DT-Web, 128
    nAnnoLyze, 127
    PROMISCUOUS, 127
    SAveRUNNER, 128, 129
    SLAP, 127
    STITCH, 128
Network-based strategies, 209
Network-based stratification, 205, 206

Networks in precision oncology, 200, 202, 209
Nextflow, 24, 25, 27
Next-generation sequencing (NGS), 38, 242
NFFinder, 132, 135
NGS-based mapping tool, 49
Noisy sequencing data, 61
Non-negative matrix factorization (NMF), 252
Normal genome and cell function, 200

O
O-Link Proteomics, 279
Omics technologies, 284
OncoIMPACT, 208
OncoKB, 181, 182
Oncotator, 190, 191
Onto-Express, 152
Organ-on-a-chip (OOAC) systems, 223
Overfitting, 210
Over-representation analysis (ORA), 12, 151, 152

P
Paired-end (PE) approaches, 59
Palimpsest, 110
Pan-cancer microsatellite instability, 77
Pancreatic ductal adenocarcinoma (PDAC), 221
PARADIGM, 154
Patchwork, 62
Pathway analysis, 151–155
  cancer research and precision oncology applications, 151, 152
    biological pathways databases, 150, 151
    cancer "omics" projects data, 146, 148
    FCS, 153
    miRNA-sensitive topological, 154, 155
    omics data source, 144
    ORA, 152
    phenotype and therapy predictions, 157, 158
    precision oncology, knowledge bases for, 148–150
    proteomics and metabolomics data repositories, 145, 146
    PT-based analysis, 153, 154
    sequencing data repositories, 144, 145
    strategies, applications, 155, 157
  patients to pathways, 143, 144
  using scRNA-seq, 274
Pathway Commons, 151
Pathway-or network-based drug repurposing methods, 121
Pathway topology (PT)-based analysis, 151, 153, 154
Patient classification
  precision oncology, 206
    MORONET, 207
    netDx, 206, 207
Patient-derived culture systems, 219
Patient-derived in vitro and in vivo models of cancer, 216–218
  features of, 225
  types of, 217
    chorioallontoic membrane, 220, 221
    microfluidic platforms, 222, 223
    patient-derived organoids, 219, 220
    patient-derived xenografts, 223, 224
    tumor slice cultures, 221, 222
    two-dimensional cancer cell lines, 218, 219
Patient-derived organoids (PDOs), 219, 220
Patient-derived xenografts (PDXs), 223, 224
Patient similarity networks (PSNs), 200, 202, 203
Patient stratification, precision oncology, 204
  data integration and clustering, similarity network fusion and similar for, 205
  network-based stratification, 205, 206
  topological data analysis, 206
PCR-based technologies, 164
pDis, 154
Pembrolizumab, 253
Pentaplex panels microsatellites, 78
PeptideAtlas, 145
PerPAS, 13
Personalized Differential Analysis (PenDA), 11
PErsonliazed Perturbation ProfilER (PePPeR), 11
Pharmacogenomics knowledgebase (PharmGKB), 180, 181
PhastCons, 195
PHENSIM, 157
Philadelphia chromosome, 9
Phylogeny, 270
PhyloP, 195
PhyloWGS, 8, 9, 109, 114
Picard-BuildBamIndex, 6
Picard-CleanSam, 5
Picard-FixMateInformation, 5
Picard-MarkDuplicates, 6
Picard-ReorderSam, 5
Picture Archiving and Communication System (PACS), 262
Polymerase chain reaction (PCR), 56
Polypharmacology browser (PPB), 124, 133
PolyPhen, 14
PolyPhen-2 (Polymorphism Phenotyping v2), 194
Portable Batch System (PBS), 25
PoSSuM, 130, 134
Precision medicine knowledge base (PMKB), 150, 182, 183
Precision oncology, 58, 199, 200
  AI (*see* Artificial intelligence (AI))
  application areas
    DawnRank and OncoIMPACT, 207, 208
    patient classification, 206, 207
    patient stratification, 204–206
  challenges and opportunities, 208
    feature engineering, 209
    speed, scalability and tunability, 208
  genomic models, value of interpretability and prior knowledge in, 203
    hypothesis generation and rational treatment design, 204
    signal-to-noise with sparse data, 204
    small sample size and overfitting, 203, 204

Precision oncology (*cont.*)
intra-tumor heterogeneity and, 103, 104
molecular profiling, liquid biopsies for
cfDNA and ctDNA, 238–240
circulating tumor cells, 235–238
extracellular vesicles, 240–242
network, 200, 202, 209
network-based strategies, 209
network-based tool for, 203
patient similarity networks, 200, 202, 203
software for networks in, 201
Precision Oncology Knowledge Base (OncoKB), 58, 150
Precision oncology pipeline, 33
Precision oncology reports, 15, 16
Precision oncology workflow
CWL scripting, 28
software infrastructures, 29, 30, 32
typical steps, 28, 29
Preclinical evidence, 178
Prediction of Drugs with Opposing Effects on Disease Genes (PDOD), 132, 135
PRIDE, 145
Primary signals, RNA-seq fusion detection, 167
ProBis, 129, 130, 134
ProFound™ AI Software, 262
PROGENy, 13, 14
PROMISCUOUS, 127, 134
Protein-Drug Interaction Database (PDID), 125, 126, 133
Protein-protein interaction (PPI), 113, 151
Protein Variation Effect Analyzer (PROVEAN), 194
Provenance, 27
Pseudoalignment, 10
PTEN, 217
Public data resources powering precision oncology, 250–252
PyClone, 8, 106
Python, 190

Q
QuantumClone, 9, 110, 111
QuantX, 263

R
Radiomics, 258
RankComp, 11
Rapid next-generation sequencing technologies, 217
RayCare, 263
Reactome, 144, 151
Read depth (RD) approaches, 59
Reads Per Kilobase per Million (RPKM), 10
Redford and Gallon methods, 91
RefCNV, 63
RE:fine drugs, 130, 135
Refractory multiple myeloma, 16
Region-based annotation, 189
Repetitive DNA sequences, 47, 48
Reproducibility, 27
Reversal of gene expression profiles, 12
Reverse gene expression score (RGES), 12, 132, 135
RNA-based drug repurposing, 272, 274, 276
RNA-binding proteins (RBPs), 14
RNA-seq-based approaches, 113
RNA-seq fusion detection
annotation and prioritization, 171
anomalous paired read signals, 169
for clinical oncology, importance, 163–165
de novo transcriptome assembly, 167
expression abnormalities, 169, 170
fusion candidates, technical scoring of, 170
mining junction/clipped reads, 167–169
nomenclature, 165, 166
primary signals, 167
sequence data sources, heterogeneity of, 166
software, 172
ensemble strategies, 174
selections, 173, 174
visualization, 171–173
RNA-seq raw data, 9
RNA sequencing (RNA-seq), 1, 4, 9, 88
biomarker identification, 11
gene expression analysis, 10, 11
gene fusion identification, 10
for ITH, 271
processing, mapping, and filtering of, 9
quantification and normalization of gene expression, 10
reversal of gene expression profiles, 12
single-sample approaches, 11
RSVSim, 62
RUbioSeq+, 69

S
SAAS-CNV, 63
SAveRUNNER, 128, 129, 134
scDNA-seq and scRNA-seq, 278, 279
SciClone, 8, 109, 114, 115
sCNAphase, 63
SCNVSim, 62
ScorePAGE, 153
scRNA-seq and scDNA-seq, 278, 279
Semantic Link Association Prediction (SLAP), 127, 134
Sequanix, 27
Sequence Alignment Map (SAM), 5, 29
Sequence data sources, heterogeneity of, 166
Sequencing artifacts, 167, 169
Sequencing technical problems, 61
Sequenza, 62
SIFT, 14, 194
Signal-to-noise, 204
Signature-based drug repurposing methods, 121
Sildenafil citrate, 120
Similarity, 210
Similarity ensemble approach (SEA), 124, 133
Similarity network fusion (SNF), 205, 253
Single-cell DNA-seq
advantages, 276
assigning therapeutics in, 275, 277

intra-tumor heterogeneity, 276, 277
technology, 276
Single-cell omics data analysis, 288, 289, 308–312
Single-cell platforms
single-cell multi-omics and integration of, 277, 278
cell-cell signaling, 279
scDNA-seq and scRNA-seq, 278, 279
Single-cell RNA sequencing (scRNA-seq), 50, 289
advantages of, 273
analysis of, 290
clustering and compositional analysis, 292
intra-tumor heterogeneity with, 273, 274
pathway analysis, 274
QC checks, 291
RNA capture, 289
technology, 272
Single-cell sequencing, 49, 50, 269, 270
DNA-seq
advantages, 276
assigning therapeutics in, 275, 277
intra-tumor heterogeneity, 276, 277
technology, 276
ITH
bulk *vs.* single-cell sequencing for, 270, 271
RNA-seq for, 271
RNA-based drug repurposing, 272, 274, 276
RNA-seq
advantages of, 273
intra-tumor heterogeneity with, 273, 274
pathway analysis, 274
technology, 272
single-cell platforms, single-cell multi-omics and integration of, 277, 278
cell-cell signaling, 279
scDNA-seq and scRNA-seq, 278, 279
tumor cells *vs.* tumor microenvironment, isolation of, 271, 272
SingleCellSignalR, 279
Single-nucleotide polymorphisms (SNPs), 3, 55, 58, 63, 190
Single nucleotide variants (SNVs), 3, 6, 105
Single-sample GSEA (ssGSEA), 13
Singscore, 13
SiPhy, 195
Smart-seq2, 272
SMuRF, 44
Snakemake, 24
SnpEff, 47, 191, 192
SnpSift, 191, 192
SNV/indel variant
algorithm basis of, 38, 41
Bogus somatic variant calling, factors for, 47, 48
data preprocessing, 42
development of, 38
germline mutation calling and prioritization, 45, 46
matched tumor-normal pairs, 43, 44
mutation calling and prioritization, 44, 45
variant annotation, 46, 47
Soft-clip junctions, 168
Solid tumor mass, 270
Somatic Mutation Calling Tumor Heterogeneity Challenge (SMC-Het), 116
Somatic mutations, clustering tumours by, 205, 206
SomaticSeq, 44
SomaticSniper, 41
Somatic SNV/indel variant, algorithm basis of, 41, 42
Specificity, 167, 172
Spectral clustering, 210
SPIA, 154, 274
Split read (SR) methods, 59
SPRUCE, 109, 110
Squid, 166
Stand Up To Cancer (SU2C), 57
STAR, 9, 16
STAR-SEQR, 10
STITCH, 128, 134
Strand bias, 47
Strelka2, 6, 41
SubMARine, 112, 113
Sun Grid Engine (SGE), 25
SuperFreq, 112
SuperPred, 124, 133
Supervised learning, 199, 206, 210
Survival convolutional neural networks (SCNN), 258
SVclone, 111
SVsim, 62
SwissTargetPrediction, 124, 133
Synthetic lethality and rescue-mediated precision oncology via the transcriptome (SELECT), 257
Systematic drug repurposing, 119

T
TarFisDock, 126
Target-based drug repurposing methods, 121
Targeted gene sequencing (TGS), 77, 81, 88
Targeted mechanism-based drug repurposing methods, 121, 122
Targeted PCR-based assays, 164
TARget FIShing DOCKing (TarFisDock), 126, 133
TargetHunter, 125, 133
Targets Associated with its MOst SImilar Counterparts (TAMOSIC), 125
TarPred, 125, 133
Terra.bio, 27
TGFα, 103
The Cancer Genome Atlas (TCGA), 57, 148, 250, 251
The Genome Cancer Atlas program (TGCA), 76
Therapeutically Applicable Research to Generate Effective Treatments (TARGET) program, 148
Third-generation sequencing (TGS) technologies, 49
Thyroid imaging reporting and data systems (TI-RADS), 262
Topological data analysis, 206
Topology-based (TB) methods, 13, 274
Total variation methods, 68
TRACERx trial, 104
Training and test set, 210
Trametinib, 58
Transcripts Per Million (TPM), 10

Transpara, 263
Tumor cells, isolation of, 271, 272
Tumor clonal evolution, 102–103, 113, 114
Tumor clonality, 269
Tumor complexity, 62
Tumor evolution, 104, 109, 114, 115
Tumor heterogeneity, 250
Tumor microenvironment (TME), 222, 284
  cellular and non-cellular elements, 285
  clinical phenotype and prognosis, 284
  contextual view, 284
  crosstalk, 307
    ligand-receptor pairs, 313
    proteomic assessment, 307
  immune landscape, 285
    cancer-immune cycle, 286
    CD8+ T cells, 286
    NK cells, 285
  isolation, 271, 272
  multiple omics data integration, 295
    experimental and analytical complexity, 297
    protein detection, 296
    spatial transcriptomics methods, 296
  qualitative and quantitative description, 287
  scDNA-seq, 293
  single-cell epigenomics, 294
  single-cell measurements, 298–306
  single-cell multi-omics profiling, 284
  single-cell proteomics, 294
  stromal compartment, 286
  therapeutic implications, 287, 288
Tumor mutation burden (TMB), 261
Tumor phylogeny, 105, 112
Tumor slice cultures (TSCs), 221, 222
Tusv, 111
Two-dimensional cancer cell lines (CCLs), 218, 219
2D Full-Field Digital Mammography (FFDM), 262

U
Ultra-deep amplicon sequencing, 78
UniProt, 190
Unique molecular identifiers (UMI), 78, 93, 94
Universally unique identifier (UUID), 31
Untranslated regions (UTR) of transcripts, 165
US FDA, AI-related models in, 262, 263

V
Validated association, 178
Variant allele frequency (VAF), 3, 8, 43, 48, 270
Variant annotation, 14, 46, 47
Variant Call Format (VCF) text files, 6
Variant Effect predictor (VEP), 47, 192
Variant effect scoring tool (VEST), 194
Variant filtration, 193
Variant interpretation, 14, 15
Variant Interpretation for Cancer Consortium (VICC), 15, 178
Variant prioritization
  cancer, 192, 193
    Endeavour, 193
    GeneDistiller, 193
    KGGSeq, 193
    MutationDistiller, 193
    VINYL, 193
VarScan2, 41
VCF2CNA, 63
Very Important Pharmacogene (VIP), 180
VICC KB, 15
VINYL, 193
Visualization, RNA-seq fusion detection, 171–173

W
Web-based solutions, 122
Web-based tools
  drug repurposing, 122
    BANDIT, 130, 131
    binding site parametrization, 129, 130
    ligand similarity using fingerprint encoding, 122–125
    MeSHDD, 130
    network-based approaches, 126–129
    RE:fine drugs, 130
    3D structures of drugs and targets, 125, 126
Weill-Cornell Precision Medicine Knowledgebase (PMKB), 15
Whole-exome sequencing (WES), 3, 56, 57, 61–64, 66, 68, 77, 78, 81, 82, 88, 90, 105, 236, 240, 250
Whole-genome amplification (WGA), 236
Whole-genome sequencing (WGS), 3, 56, 57, 61–63, 65, 68, 77, 78, 81, 82, 89–91, 250
Whole-slide images (WSIs), 258
WikiPathways, 144, 151
Wild type reads (WTP), 91
Workflow Description Language (WDL), 24, 25
Workflow management systems (WMS)
  CWL, 24, 25
  NextFlow, 25, 27
  URLs, 25
  WDL, 25