Molecular Oncology

FEBS PRESS science publishing by scientists

REVIEW

Bioinformatics roadmap for therapy selection in cancer genomics

María José Jiménez-Santos, Santiago García-Martín, Coral Fustero-Torre, Tomás Di Domenico, Gonzalo Gómez-López and Fátima Al-Shahrour

Bioinformatics Unit, Spanish National Cancer Research Centre (CNIO), Madrid, Spain

Keywords
bioinformatics; drug prioritisation; next-generation sequencing; precision oncology; treatment selection; tumour heterogeneity

Correspondence
F. Al-Shahrour, Bioinformatics Unit, Spanish National Cancer Research Centre (CNIO), Calle Melchor Fernandez Almagro, 3, 28029 Madrid, Spain
Tel: + 34 917 328 000
E-mail: falshahrour@cnio.es

(Received 26 April 2022, revised 22 June 2022, accepted 8 July 2022, available online 20 August 2022)

doi:10.1002/1878-0261.13286

Tumour heterogeneity is one of the main characteristics of cancer and can be categorised into inter- or intratumour heterogeneity. This heterogeneity has been revealed as one of the key causes of treatment failure and relapse. Precision oncology is an emerging field that seeks to design tailored treatments for each cancer patient according to epidemiological, clinical and omics data. This discipline relies on bioinformatics tools designed to compute scores to prioritise available drugs, with the aim of helping clinicians in treatment selection. In this review, we describe the current approaches for therapy selection depending on which type of tumour heterogeneity is being targeted and the available next-generation sequencing data. We cover intertumour heterogeneity studies and individual treatment selection using genomics variants, expression data or multi-omics strategies. We also describe intratumour dissection through clonal inference and single-cell transcriptomics, in each case providing bioinformatics tools for tailored treatment selection. Finally, we discuss how these therapy selection workflows could be integrated into the clinical practice.

1. Introduction

Over the past few years, our understanding of cancer disease has enabled advances in diagnosis and treatment, contributing to improving survival rates in many tumour types. Current therapeutic management of primary and disseminated tumours includes surgical resection, radiotherapy, hormonal therapy, chemotherapy, targeted therapies and immunotherapy. Targeted therapies are considered a cornerstone of precision oncology, that is the use of cancer genomic information as a means to stratify individual patients for the administration of optimal therapeutic modalities [1,2]. Targeted therapies have been conceived on the basis of the druggable genome paradigm (Box 1), that is the genes and gene products known (or predicted) to interact with available compounds [3]. In the recent years, efforts have been focused on defining new predictive biomarkers of anticancer drug efficacy, and as a consequence, the number of predictive biomarkers approved by the Food and Drug Administration (FDA) has increased from 39 in 2013 to 214 in 2022 (i.e. greater than fivefold in the last 10 years) [4]. Common examples of targeted therapies are the use of *BRAF* V600E

Abbreviations
ADR, adverse drug reaction; CNV, copy-number variation; COSMIC, Catalogue Of Somatic Mutations in Cancer; DGE, differential gene expression; DNA-seq, DNA sequencing; FCS, functional class scoring; FDA, Food and Drug Administration; GATK, Genome Analysis Toolkit; ICGC, International Cancer Genome Consortium; Indel, small insertions and deletions; ITH, intratumour heterogeneity; MTB, Molecular Tumour Board; NGS, next-generation sequencing; ORA, over-representation analysis; PCAWG, Pan-Cancer Analysis of Whole Genomes; QC, quality control; RNA-seq, RNA sequencing; scRNA-seq, single-cell RNA sequencing; SNV, single nucleotide variant; ST, spatial transcriptomics; SV, structural variant; sWGS, shallow whole-genome sequencing; TCGA, The Cancer Genome Atlas; TMB, tumour mutational burden; TME, tumour microenvironment; UMAP, Uniform Manifold Approximation and Projection; VCF, variant calling file; WES, whole-exome sequencing; WGS, whole-genome sequencing.

Box 1. Druggable genome.

The druggable genome is formed by the set of genes encoding proteins that are or potentially can be targeted by drugs. Of the ~ 20 000 coding genes present in the human genome, ~ 3000 have been estimated to be druggable and less than 700 are currently targeted by FDA-approved drugs [223].

inhibitors in melanoma patients, imatinib to target *BCR-ABL* translocations in chronic myeloid leukaemia and PD1/PD-L1 inhibitors for the immunotherapeutic treatment of melanoma, lung, renal and other cancer types. In addition, next-generation sequencing (NGS) technologies have driven the discovery and development of new pharmacogenetic biomarkers, which play crucial roles in identifying drug responders and nonresponders, avoiding adverse effects and optimising drug dosage. Nevertheless, targeted therapy development is challenging since most of the druggable genome remains unstudied and the clinical setting of targeted therapies is still underdeveloped. Moreover, even with the consideration of genomic and transcriptomic patients' profiles, some patients may not respond to a genomically guided treatment. Furthermore, a prominent caveat of current targeted therapies is the onset of acquired resistance and thus clinical relapse, despite favourable initial responses in advanced disease [5,6].

Tumour heterogeneity has been revealed as a novel key factor in the failure of anticancer therapies. The findings provided by large-scale cancer genomics projects such as The Cancer Genome Atlas (TCGA), the International Cancer Genome Consortium (ICGC) and the Pan-Cancer Analysis of Whole Genomes (PCAWG) consortia [7–9] have clearly revealed a high multidimensional genomic heterogeneity among different tumour types but also within the same patient, thus underlining the idea that cancers are not single diseases but rather an array of disorders with distinct molecular mechanisms [10]. The concept of tumour heterogeneity encompasses both inter- and intratumour heterogeneity (ITH). The former refers to the existence of different genomic alterations among cancer patients or within the same individual (i.e. primary vs metastatic tumour), while the latter describes the intrinsic clonal diversity found within tumours occurring as a consequence of cancer somatic evolution and natural selection. Tumour heterogeneity has been related to different treatment responses [11,12], the appearance of drug resistance [13,14] and therefore the patients' clinical outcome [15,16]. In order to reveal the relationships between ITH and clinical outcome, the TRAcking Cancer Evolution through therapy (Rx) (TRACERx) initiative is performing an extensive multi-omics profiling of ITH in NSCLC, melanoma, prostate and renal cancer [17]. Deceased patients are corecruited to the Posthumous Evaluation of Advanced Cancer Environment (PEACE) (NCT03004755) study, which allows for metastatic sampling from multiple tumour sites. The Glioma Longitudinal Analysis Consortium (GLASS) is another international effort whose goal is the molecular characterisation of gliomas over several time points in order to understand tumour evolution and identify therapeutic vulnerabilities [18]. The characterisation of ITH has also benefited from single-cell techniques that have allowed high-resolution dissection of both tumour and tumour microenvironment (TME) cell composition. In this sense, the Human Tumour Atlas Network (HTAN) and other initiatives [19,20] are generating single-cell three-dimensional atlases of tumour transitions, ITH and TME landscapes for a diverse set of tumour types. Undoubtedly, these valuable efforts have the potential to improve cancer detection, prevention and therapeutic discovery. However, the ITH, the tumour evolution and the potential for competitive release of resistant tumour subclones are not yet addressed in cancer therapeutics and clinical practice [21].

Bioinformatics provides a vast catalogue of methodologies and databases required to analyse, integrate and interpret cancer multi-omics data. Remarkably, *in silico* drug prioritisation approaches (Box 2) have recently emerged to evaluate tumours' specific genomic alterations and transcriptomic profiles, matching them with tailored candidate treatments [22–26]. This review aims to provide a bioinformatics roadmap and general guidelines to propose anticancer data-driven treatment strategies for bulk and single-cell omics data covering both cancer research and precision oncology scenarios.

Box 2. *In silico* prioritisation.

Precision medicine aims to make tailored prescriptions based on individual omics data. In order to do so, epidemiological, clinical and response data from previous patients are required. Drug prioritisation methods can integrate these sources of data and compute scores to rank the available treatments based on the predicted efficacy. These bioinformatics tools provide clinicians with evidence-based guidance to prescribe the drug that better matches the characteristics of each patient.

Computational approaches required to generate tumour genomic and transcriptomic profiles and explore a tumour's functional activity are also discussed. Current algorithms for characterising tumour heterogeneity and dissecting ITH from multi-omics sources will be addressed together with cutting-edge methods that exploit the drug sensitivity of tumour cell subpopulations. Finally, the current limitations and perspectives in the development and improvement of novel computational approaches for precision medicine-based therapies will also be discussed.

2. Genomics-based drug selection

NGS has been widely adopted for the analysis of tumour DNA extracted from clinical and biological samples with the aim of detecting clinically relevant genomic alterations for cancer diagnosis and treatment guidance. This section describes the computational workflow to analyse, detect and interpret DNA alterations (Box 3) (i.e. short variants and structural variants) that can guide cancer therapy selection using data generated by targeted, whole-exome (WES) and whole-genome sequencing (WGS) experiments. Cancer somatic mutations are the main focus of bioinformatics analyses aimed at identifying important druggable alterations, since targeted therapies directed against these variants would less likely affect healthy cells. However, germinal variants that affect drug substrates or metabolising enzymes can play an important role in drug effectiveness and toxicity and should not be overlooked when designing tailored treatment strategies.

Box 3. Genomic variants.

According to their extent, genomic alterations can be classified into short or structural variants (SVs). Short variants can be subdivided into single nucleotide variants (SNVs) or small insertions and deletions of < 50 bp (indels). On the contrary, SVs affect genomic regions of ≥ 50 bp and can be further classified depending on whether the genetic material is conserved or lost. Balanced SVs arise as a consequence of inversions and translocations, whereas unbalanced SVs, also known as copy-number variations (CNVs), are due to big insertions, deletions or duplications.

Genomic variants can also be classified as germline or somatic, depending on their origin and extent of the affected tissue. While germline mutations are inherited by the progeny and affect the whole organism, somatic mutations arise spontaneously and are localised in a specific tissue [224].

2.1. Short variants to guide therapies

The first step for genomics-based drug selection is to identify clinically relevant alterations in cancer patients via a variant calling analysis. According to GATK Best Practices [27], the general workflow of variant calling consists of nine steps: quality control (QC) and trimming, alignment, marking duplicates, local realignment of indels, base quality score recalibration (BQSR), variant calling, filtering and annotation of variants [28] (Fig. 1). Briefly, after performing sample QC and trimming, the raw reads are aligned to the reference genome with tools such as BWA-MEM [29]. Then, duplicated reads must be removed with Picard [30]. In order to reduce alignment artefacts and obtain more accurate sequencing quality estimations, further processing can be done using GATK tools [31]. There is a broad selection of variant calling tools such as Mutect2 and HaplotypeCaller [31], VarScan 2 [32], VarDict [33] or SomaticSniper [34] that can be used to identify short variants, which are comprised of single nucleotide variants (SNVs) and insertions or deletions (indels) of less than 50 base pairs (bp). The reported variants must be filtered in order to remove low-quality calls and subsequently annotated with information about their biological impact, their frequency in the population and their clinical relevance. This type of analysis mainly focuses on somatic variants occurring in coding regions. Nonsynonymous SNVs are considered as more damaging, since they alter the final sequence of the encoded protein and might affect its correct folding and function [35]. Furthermore, somatic genomic alterations can be classified according to their frequency in the population as rare variants or polymorphisms, which are considered clinically benign due to their high frequency ($> 1\%$). In most patients, at least one detected somatic alteration is potentially clinically relevant [36,37] since it either changes the gene function, suggests the use of surveillance measures for prevention or early detection, helps to establish a diagnosis, influences the prognosis or guides the selection of therapies.

Some tools for automatic variant annotation are SnpEff [38], which determines the biological impact of candidate variants; or ANNOVAR [39] and the Variant Effect Predictor (VEP) [40], which additionally provide information about the frequency of each variant in the population. On top of that, there are many public data repositories of acquired knowledge about

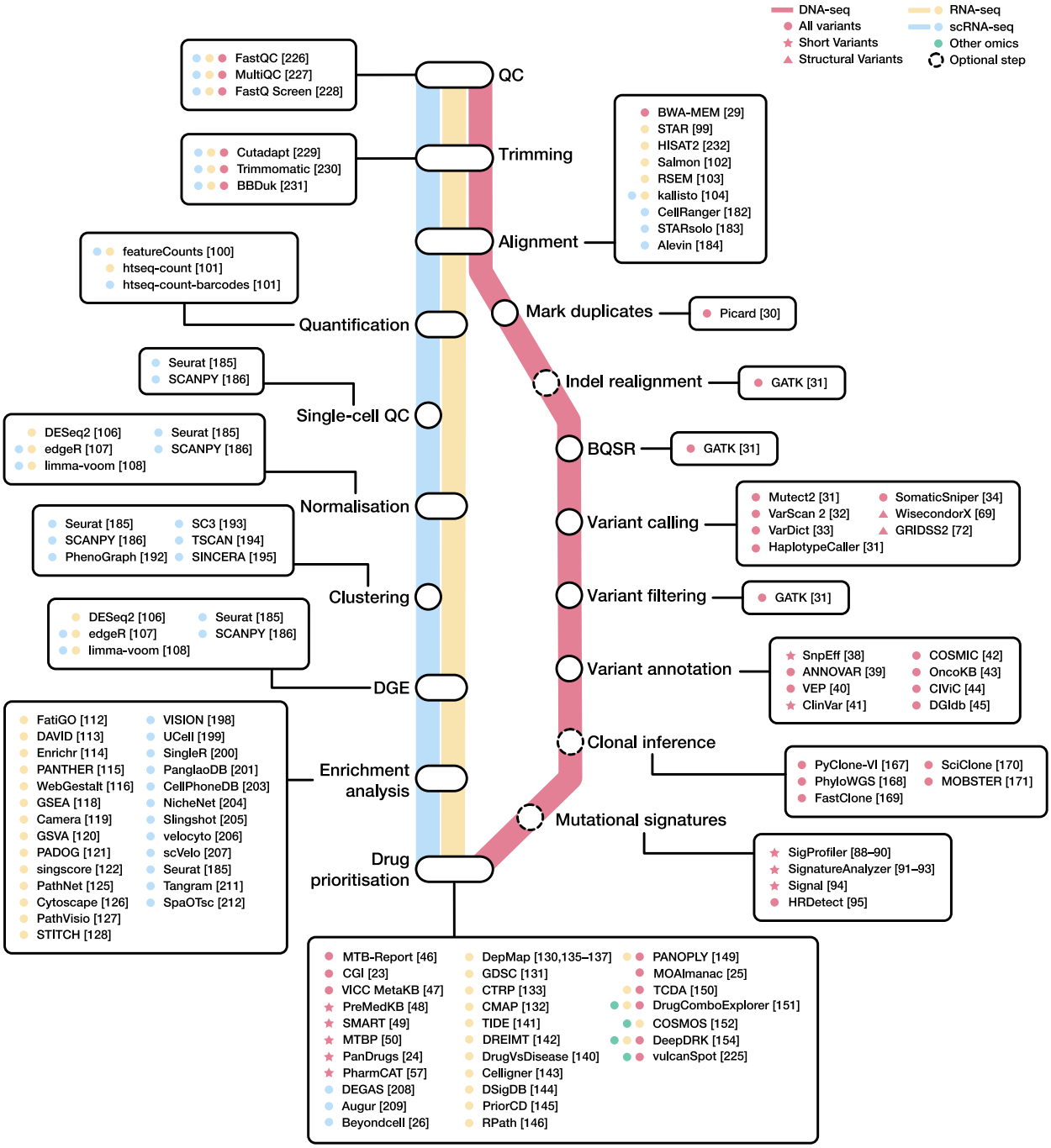

Fig. 1. Roadmap for drug prioritisation from different omics profiles. The roadmap is represented as an underground map in which each next-generation sequencing (NGS) technique is a different line and each step in the workflow is a station. Common steps between workflows are displayed as interchange stations. The names of the available tools are preceded by coloured symbols that indicate in which technique they can be applied. BQSR, base quality score recalibration; DGE, differential gene expression; QC, quality control.

variants, drugs and their interconnections that are useful to annotate candidate somatic variants with. Some examples are ClinVar [41], a database of genetic variants and their clinical repercussions; the Catalogue Of Somatic Mutations In Cancer (COSMIC) [42], a knowledgebase with information about the impact of somatic variations in cancer; OncoKB [43] and CIViC [44], two resources that link somatic cancer variants

with their clinical and therapeutic implications; or DGIdb [45], a database of gene–drug associations.

Furthermore, a number of methodologies and bioinformatics tools have been developed with the objective of making cancer variant interpretation easier and suggest possible treatments based on previous evidence (Table 1). Most of these resources are patient-centred, require the somatic variants from the tumour and can be classified depending on the nature of the input data. If a list of variants is available, resources such as MTB-Report [46], the Cancer Genome Interpreter (CGI) [23], the Variant Interpretation for Cancer Consortium meta-knowledgebase (VICC metaKB) [47], PremedKB [48] or the SMART Cancer Navigator [49] can be useful. Some of these tools accept not only a list of short variants, but also disease and/or drug queries as input. In case a variant calling file (VCF) is available, the user can opt for MTBP [50] or PanDrugs [24]. The latter accepts both types of inputs and drug and gene queries. There have also been other approaches to guide therapy prescription on a large scale in order to obtain a general view and observe trends in cohorts of different tumour types [51].

Most of these approaches aim to prioritise drugs based only on somatic variants. However, germinal variants are also crucial in drug metabolism, and therefore in drug effectiveness and toxicity [52]. Thus, patients can have different responses to the same treatment, ranging from responsiveness to ineffectiveness or even adverse drug reactions (ADRs), which are important causes of morbidity and mortality and represent a source of financial burden to healthcare systems [53]. Differences in drug response are mainly due to genetic variation in genes encoding for drug substrates or genes that participate in the metabolism and transport of xenobiotics [54]. By assessing the germinal variants of each patient and mining pharmacogenomic databases such as DrugBank [55], PharmGKB [56], or the Table of Pharmacogenomic Biomarkers in Drug Labeling [4], effective compounds could be prioritised over ineffective or ADR-causing drugs. PharmCAT [57] is a tool developed for suggesting tailored treatments based on germinal variants found in a VCF. Moreover, some resources such as the MTBP have been developed to account for both germinal and somatic variants.

Recent publications [58,59] provide comprehensive lists of variant annotation knowledge bases and bioinformatics tools for variant interpretation, biomarker identification, drug prioritisation and response prediction. All these resources were conceived as supporting tools to inform clinicians of the available treatment options for their patients.

Interestingly, the identification of novel biomarkers of immunotherapy response has become one of the great challenges in oncology. Tumour mutational burden (TMB) has established itself as a promising genomic biomarker that may help identify patients who are most likely to benefit from immunotherapy in a wide range of tumour types [60]. TMB is calculated by counting the total number of somatic alterations divided by the total size in Mbp of the regions that have been sequenced. Nevertheless, there is a lack of standardisation for TMB assessment, which makes it difficult to use as a biomarker. High TMB is associated with improved or clinically relevant patient response to immunotherapy; however, the utility of this biomarker has not been fully demonstrated across all cancer types [61]. Moreover, using bioinformatics techniques, it is now possible to unravel the TMB content and generate *in silico* hypotheses beyond the TMB-based stratification of patients. This way, we can prioritise and select targeted therapies based on the presence of mutations for which treatments already exist. This is the case of PanDrugs [24], a platform that prioritises drug treatments based on actionable mutations present in TMB.

2.2. Structural variants and mutational signatures to guide therapies

Genomic sequence alterations that affect large regions ($\geq$ 50 bp) fall under the umbrella of what is known as structural variation (SV). A SV is composed of several types of events arising from different mutational mechanisms. Some of these events, such as deletions, insertions or duplications, result in changes in the amount of genomic sequence. These changes are known as copy-number variations (CNVs) [62–65]. Throughout history, a series of different techniques have been applied to study CNVs. The decreasing costs of WGS experiments combined with the constant improvement of variant calling methods are positioning WGS-based CNV calling as the preferred technique for the analysis of CNV [66]. CNV can be studied through WGS experiments by detecting areas in the genome that have more or less reads than would be normally expected. This method is commonly known as depth of coverage (DOC) analysis. CNV has seen an increase in its applicability in the clinical diagnostics environment given its robustness to produce results with shallow levels of sequencing depth, usually defined as 0.1× to 1.0× coverage of the genome [67]. Even though most of the currently available CNV characterisation tools are aimed at the research environment [68], tools such as WisecondorX have been created with the

Table 1. Bioinformatics tools for genomics-based drug prioritisation.

Name	Description	Input	Output	URL
MTB-Report [46]	R script that filters and classifies cancer variants into levels of evidence using gene-drug databases	Tables with SNVs, CNVs and gene fusions (somatic)	Molecular Tumour Board (MTB) report with actionable variants in PDF	https://github.com/jperera-bel/MTB-Report
Cancer Genome Interpreter (CGI) [23]	Web tool that annotates cancer variants and identifies potential oncogenic alterations and genomic biomarkers of drug response	List of SNVs, indels, CNVs and/or gene fusions (somatic)	Downloadable tables of (a) annotated variants, including information about the oncogenicity and biological consequence, and (b) drug–variant associations with evidence level and response prediction	https://www.cancergenomeinterpreter.org/home
VICC MetaKB [47]	Web tool for cancer variant interpretation that harmonises six different variant annotation knowledgebases with information about variant, gene, disease and drug associations and their corresponding evidence levels	List of variants (somatic), including gene fusions, genes, diseases and/or drugs	Interactive report with variant–gene–disease–drug associations, each one with its evidence label and supporting links	https://search.cancervariants.org/#*
PreMedKB [48]	Web tool for integrating information on diseases, genes, variants, drugs and the relationships between any two or more of these four components	List of short variants (somatic), genes, drugs and/or diseases	Interactive semantic network displaying components as nodes and their relationships as edges. Results can be downloaded in either JSON or PNG format	http://www.fudan-pgx.org/premedkb/index.html#/home
SMART Cancer Navigator [49]	Web application for variant interpretation that associates the corresponding genes to diseases, known drugs and relevant clinical trials	List of short variants (somatic and germline)	Interactive report with variant, gene, disease and drug information	https://smart-cancer-navigator.github.io/home
PanDrugs [24]	Web tool to prioritise anticancer drug treatments according to individual genomics data. pandrugs computes two scores, the Gene Score (GScore) and the Drug Score (DScore). The GScore ranges from 0 to 1 and is estimated according to gene essentiality and tumoral vulnerability, gene relevance in cancer, the biological impact of mutations, the frequency of gene alterations and their clinical implications. The DScore ranges from −1 to 1, considers drug indication and status, gene–drug associations and number of hits and estimates resistance (negative values) or sensitivity (positive values)	VCF, a list or a ranking of genes or a drug query (somatic)	Report with a prioritised list of anticancer therapies. pandrugs resolves the Best Therapeutic Candidates based on the accumulated and weighted scoring of the GScore and the DScore	https://www.pandrugs.org/#!/

Table 1. (Continued).

Name	Description	Input	Output	URL
MTBP [50]	Web tool that annotates somatic and germline short variants (SNVs and indels) functionally and clinically, categorising the cancer biomarkers (diagnosis, prognosis and drug response) found in the tumour	VCF or a list of short variants (somatic and germline)	HTML report with annotated variants, the evidence supporting the variants' functional classification and their associated actionability	https://mtbp.org/
PHARMCAT [57]	A tool for identifying germinal variants, inferring patient's haplotypes and diplotypes and suggesting treatments following the Clinical Pharmacogenetics Implementation Consortium (CPIC) guidelines	VCF (germline)	HTML/JSON report with drug suggestions based on germinal variants	https://pharmcat.org/

specific goal of exploiting CNV calling using shallow WGS (sWGS) in the clinical setting [69]. An essential requirement for the study of mutational events is the ability to distinguish potentially significant events from those present in the healthy population [65]. Several approaches exist to achieve this, with the two main categories being those that do not require a normal reference, or reference-free, and those that do [69]. Reference-free approaches normalise the samples using known features of the human genome such as GC content and mappability. Reference-based tools require as a reference either a single normal sample associated with the sample of interest or what is sometimes known as a Panel of Normals (PON) [70]. These normal samples have the goal of removing variation from the results, which could in fact be caused by experimental procedures (e.g. sample handling, preparation and sequencing equipment).

Structural variant events, whether CNV-causing or not, can be extremely complex in terms of the changes they produce on the sequence of the genome [63,71]. The characterisation of these events depends on methods such as the analysis of paired reads and split reads, and the *de novo* assembly of the genome of the sample of interest. The short nature of NGS reads imposes some limitations on these types of analyses [63]. Long-read sequencing has given rise to a new generation of tools and approaches that aims at filling this gap in our ability to understand SVs [72]. Furthermore, unlike next-generation machines, nanopore-based sequencers offer great portability and the possibility of analysing data as it is generated (i.e. in a streaming fashion). Tools are already being developed with the aim of enabling the characterisation of SV for clinical diagnostics [73].

Structural variation has a potential impact on both germline and somatic genomic instability that affects disease development and might help to select therapies and report on patients' drug response. For instance, chromosomal translocations are relevant in the diagnosis of haematological malignancies but also lead to therapeutic approaches targeting fusion proteins such as BCR-ABL1 in chronic myeloid leukaemia [74]. Some bioinformatics tools designed to prioritise drugs based on short variants also accept CNVs [23,46] and/or gene fusions [47] as input. More advanced approaches for taking advantage of sWGS CNV calling for diagnostic purposes include efforts towards generating CNV-based signatures, which may allow for more precise diagnostics and treatment selection [75]. Nevertheless, SVs are not yet commonly being used as molecular targets or biomarkers to guide patient-specific treatment [76].

On the contrary, mutational signatures identified in genomic DNA can reveal unique patterns of mutational processes that occurred during the course of cancer development [77,78]. These mutational signatures can include single base substitutions (SBS), doublet base substitutions (DBS), indels, CNV and genome rearrangements [79]. Interestingly, mutational signatures may be informative for guiding the identification of therapeutically targetable biomarkers, suggesting their application in personalised therapeutic approaches. In particular, several studies have found that tumours harbouring mutational signatures of DNA damage repair deficiency may show therapeutic responsiveness for either DNA damaging agents or immunotherapy [80–82]. For example, a mutational signature associated with pathogenic mutations in *BRCA1* and *BRCA2* genes has been identified in several cancer types, including breast and ovarian cancer, suggesting deficient homologous recombination (HR) and sensitivity to PARP inhibitors [83]. On the contrary, previous exposure to DNA-damaging agents such as chemotherapy has been associated with drug resistance [84]. Interestingly, mutational signatures can be used as the mutational footprints of cancer therapies to estimate the contribution of different treatments to the TMB and can reveal their long-term side effects in the genome [85].

There are several computational strategies for performing mutational signature analyses that in general differ on the mathematical properties of mutational signature discovery and can be grouped into two categories: methods that aim to discover novel signatures (*de novo*) or methods to detect already known and validated mutational signatures (refitting) [86,87]. SigProfiler [88–90], a framework used for the previous version of COSMIC, and SignatureAnalyzer [91–93] were the two *de novo* tools used to analyse a large collection of cancer genomes from PCAWG, TCGA and ICGC projects [79]. It is worth noting Signal, a recently published web-based tool for mutational signatures that also calculates the associations between gene drivers and mutational signatures that could provide novel therapeutic dependencies [94]. Moreover, HRDetect is a predictor of HR deficiency that can be useful to stratify patients based on their expected sensitivity to PARP inhibitors [95].

3. Transcriptomics-based drug selection

Transcriptomics profiling has a wide range of applications in cancer research, from tumour classification, diagnosis and prognosis to therapeutic selection of drug candidates. Gene expression measures have already been incorporated in molecular diagnostics techniques such as the MammaPrint expression panel or the Oncotype DX Breast Recurrence Score to guide clinical decision-making [96,97].

This section focuses on bioinformatics approaches to prioritise therapeutic candidates based on gene expression (Fig. 1). First, we briefly summarise the most common steps in RNA sequencing (RNA-seq) workflows. Next, we discuss functional enrichment approaches aimed at revealing biological patterns underlying gene expression. Finally, we review bioinformatics strategies for transcriptome-based drug prioritisation depending on the data format and type.

RNA-seq has become the most used NGS technique to detect and quantify the presence of RNA in biological samples. One of the first steps in a standard RNA-seq data analysis is to generate a matrix of unnormalised gene counts by aligning raw sequence reads to a reference genome or transcriptome [98]. The STAR aligner [99], and aggregation packages such as featureCounts [100] or htseq-count [101] are some of the most widely employed tools to achieve the above step. Alternatively, alignment-free methods such as Salmon [102], RSEM [103] or kallisto [104] output transcript-level estimates, are then summarised to the gene level with R packages such as tximeta [105]. Next, the raw gene expression matrix must be normalised and transformed to stabilise intersample variance. Afterwards, differential gene expression (DGE) analysis extracts significant differences in RNA abundance between experimental conditions. Well-established methodologies such as DESeq2 [106], edgeR [107] or limma-voom [108] perform both normalisation and DGE in tandem.

Functional enrichment is often performed following a DGE analysis with the aim of revealing biological relationships in the differentially expressed genes list and of identifying underlying coordinated patterns (e.g. functional pathways and regulatory modules) in the expression matrix. Functional enrichment methods can be categorised into three main types: (a) over-representation analysis (ORA); (b) functional class scoring (FCS); and (c) pathway topology [109]. They commonly exploit annotations from public databases, ontologies or related gene terms (gene sets) based on their involvement in a pathway, biological function or specific cellular compartment [110,111]. ORA methods statistically evaluate the proportion of genes that share a particular annotation in a gene list of interest with respect to what is expected by chance. Web tools such as FatiGO [112], DAVID [113], Enrichr [114], PANTHER [115], WebGestalt [116] and others [117]

follow this approach. As an alternative to ORA methods, FCS methods consider that coordinated changes in functionally related genes are as important as large expression changes in individual genes. To this end, FCS tools rely on ranked lists (i.e. gene expression rankings) to generate a single pathway-level enrichment score, which is tested for statistical significance. Widely employed methods include GSEA [118], CAMERA [119], GSVA [120], PADOG [121], SINGSCORE [122] and others [123]. Finally, pathway topology analysis adds another information layer by taking into account gene–gene interactions along with gene-level statistics to identify regulatory changes in pathways. Pathway topology methods have been extensively reviewed by Ihnatova et al. [124]. Remarkable contributions include PATHNET [125], which leverages the connectivity between genes of the same pathway along with the differences in gene expression between conditions. Cytoscape [126] and PathVisio [127] offer powerful visualisation and analysis tools tailored to biological interaction networks. Moreover, STITCH [128] integrates information from metabolic pathways, compound structures and drug–target relationships to generate a network of compound–protein interactions.

Drug prioritisation methods can employ transcriptomic profiles as input to suggest which treatments will be most effective for a given tumour sample. Multiple bioinformatics strategies are available depending on the nature and complexity of the data source, ranging from individual genes (e.g. a single overexpressed gene from a DGE analysis) to whole expression data matrices.

The integration of genomic and transcriptional profiles together with drug response profiles has allowed the advancement of drug repositioning and drug combination predictions [129]. By finding its druggable weak spots, pharmacogenomics studies (Box 4) have been demonstrated to be useful in aiding treatment selection in cancer cell lines [130–137]. Resources and tools such as the DEPMAP [130,135–137], the GDSC [131] and the CTRP [133] are remarkable efforts for drug prioritisation using transcriptomics. In all cases, the input is a single gene that can be queried against large databases of pharmacogenomics assays, allowing researchers to correlate the expression level of a gene of interest with the susceptibility to drugs in thousands of cancer cell lines.

Gene expression signatures (a list of genes whose expression is associated with a given condition) can be interrogated for drug prioritisation applying the signature reversal approach, which relies on the fact that the expression pattern of drugs indicated for a disease is often negatively correlated with the changes in gene expression induced by that disease [138]. For instance, Cheng and colleagues mined TCGA to generate an expression signature of *EGFR* activity, which they associated with tumour sensitivity to EGFR inhibitors and other tyrosine kinase inhibitors [139]. Similarly, the CONNECTIVITY MAP (CMAP) [132] project is generating a comprehensive catalogue of cellular signatures representing systematic perturbation to pharmacologic and genetic perturbagens. Researchers can freely access the CMAP database or interrogate signatures of interest through its Web portal (Table 2). DRUGVSDISEASE mines microarray databases to generate ranked expression profiles for the comparison of drug and disease gene expression profiles [140]. It features a precalculated ranked list of differentially expressed genes for 1309 drug compounds applied to cancer cell lines readily available for signature reversal. Expression signatures have also been used to predict response and prioritise compounds for immunotherapy. TIDE evaluates biomarkers to predict immune check blockade clinical response for patient stratification [141]. DREIMT performs drug prioritisation analysis for immunomodulation suggesting candidate immunomodulatory drugs targeting user-supplied gene expression signatures [142].

Box 4. High-throughput drug screenings.

High-throughput screenings are assays in which large libraries of compounds are tested in order to discover candidate drugs with activity against a target.

The whole normalised expression data matrix can also be used to prioritise drugs. For instance, following the CELLIGNER methodology [143], the transcriptomic profile of individual samples can be aligned to the most similar cancer cell line, allowing researchers to harness the extensive pharmacogenomics profiling of said models to draw hypotheses about drug susceptibility. Moreover, GSEA can be used in conjunction with DSigDB [144], a database of drug gene sets, to find whether an experimental condition is enriched in genes participating in a given drug response. On the contrary, single sample enrichment methods such as GSVA perform the aforementioned enrichment sample-wise instead of per condition, transforming a gene matrix to a drug signature enrichment matrix. Then, this matrix can be used for clustering, applying linear models or other approaches.

Finally, network-based algorithms leveraging pathway topology have also been used for drug prioritisation. PRIORCD [145] makes use of a network propagation algorithm and a drug–drug similarity

Table 2. Bioinformatics tools for transcriptomics-based drug prioritisation.

Name	Description	Input	Output	URL
The Connectivity Map (CMAP) [132]	Catalogue of gene expression signatures representing systematic perturbations with genetic and pharmacologic perturbagens. Features a Python library for programmatic access and cloud powered tools for quick interrogation of signatures, genes and compounds	Multiple inputs, depending on the Web tool	Multiple outputs, depending on the Web tool	https://clue.io/
The Cancer Dependency Map (DepMap) [130,135–137]	Systematic study with the aim of uncovering genetic dependencies, small molecule sensitivities and discovering the biomarkers that predict them. Web portal powered by freely available multi-omics data sets	Single gene, compound, cell line or lineage as plain text	Scatterplots, linear models and correlations between features	https://depmap.org/portal/
DREIMT [142]	Web tool/RESTful API for hypothesis generation and prioritisation of drugs capable of modulating immune cell activity from transcriptomics data	Gene list as plain text or as a comma-separated file	Prioritised list of drug candidates for immunomodulation	http://dreimt.org/
DSigDB [144]	Gene set collection of annotated compounds and drugs	None	None	http://dsigdb.tanlab.org/DSigDBv1.0/
RPath [146]	Network-based approach that proposes drugs for a disease by means of a knowledge graph and perturbation signatures	Knowledge graph with source, target and polarity as a tab-separated file. Gene expression data as a dictionary of keys (genes) and values depending on the gene expression	Prioritised list of drug candidates for a given disease	https://github.com/enveda/RPath
The Cancer Druggable Atlas [150]	Comprehensive catalogue of potential druggable genes across cancers. Publicly available through the Functional Cancer Genome data portal	HNGC symbol as plain text	Multi-omics profile of the target gene and its predicted druggability	http://fcgportal.org/TCDA/
vulcanSpot [225]	Web tool/RESTful API which mines massive screenings data to identify genetic dependencies and prioritise therapeutic candidates using a combination of known drug–gene relationships and drug repositioning strategies	Single gene as plain text	Ranked list of genetic dependencies alongside therapeutic candidates	http://vulcanspot.org/
PANOPLY [149]	R package that uses machine learning and knowledge-driven network analysis to identify and analyse patient-specific alterations (CNVs, germline and somatic short variants, fusion transcripts, gene expression and expressed mutations) driving oncogenesis and prioritise drugs that target the networks and pathways associated with these alterations	CNVs, SNVs (somatic and germline), expressed SNVs and expressed genes	Integrated multi-omics case report of the patient with prioritisation of anticancer drugs	http://kalarikrlab.org/Software/Panoply.html

Table 2. (Continued).

Name	Description	Input	Output	URL
COSMOS [152]	R package that leverages extensive prior knowledge of signalling pathways, metabolic networks and gene regulation with computational methods to estimate activities of transcription factors and kinases, as well as network-level causal reasoning	At least two sources of information from human transcriptomics, phosphoproteomics, metabolomics or fluxomics assays	Integrated trans-omics network of estimated activities of kinases and transcription factors	https://saezlab.github.io/cosmosR/
CellPhoneDB [203]	Publicly available repository of curated receptors, ligands and their interactions	Counts data (either a TXT, H5AD or H5 file) or a path to the folder containing a 10× output with mtx/barcode/features files	Multiple comma-separated files detailing ligand–receptor interactions along with statistical significance metrics for each pair	https://www.cellphonedb.org/
NicheNet [204]	R package that predicts ligand–receptor interactions between sender and target cell subpopulations that might drive gene expression changes	NicheNet's tables with prior information about ligand–receptor pairs, a preprocessed scRNA-seq matrix (or a Seurat object) and a list of genes in the target subpopulation whose expression might be influenced by cell-to-cell communication	Table with the probability of each ligand in the sender subpopulation of driving the expression changes of the target genes	https://github.com/saeyslab/nichenetr
Beyondcell [26]	R package for single-cell-based drug prioritisation	Preprocessed scRNA-seq matrix (or a Seurat object)	Prioritised ranking of the differential sensitivity drugs between chosen conditions	https://bioinformatics.cnio.es/tools/
Augur [209]	R package for prioritisation of cell types based on the response to an experimental perturbation. Augur trains a machine learning model for each cell type and quantifies the separability of perturbed and nonperturbed cells. The most separable cell type is assumed to be the most responsive to the perturbation	Preprocessed scRNA-seq object (either Seurat, Monocle 3 or SingleCellExperiment) containing metadata associated with each cell, including the cell type annotations and sample labels to be predicted	Rank of cell types according to the amount of responsiveness to the perturbation	https://github.com/neurorestore/Augur
DEGAS [208]	R package that implements a deep transfer learning framework for prioritising cells in relation to disease attributes (such as diagnosis, prognosis and response to therapy) retrieved from patients	scRNA-seq and gene expression matrices, as well as patient metadata with clinical information	Single-cell metadata with clinical annotations for each cell and/or patient metadata with cell compositions	https://github.com/tsteelejohnson91/DEGAS

network, along with pathway activity profiles to prioritise candidate drugs in cancer. Similarly, RPath [146] relies on a knowledge graph built from disease, protein and drug causal relations along with disease and perturbed expression signatures to prioritise compounds for a given disease.

4. Integrative multi-omics strategies for drug selection

High-throughput technologies have opened up the possibility of integrating orthogonal omics layers for a more comprehensive understanding of biological systems [147]. Drug prioritisation could also benefit from such integration [148]. Methods such as panoply [149] or MOAlmanac [25] (Table 2) integrate genomic and transcriptomic data to identify and prioritise drug targets. The Cancer Druggable Gene Atlas (TCDA) [150] is a recently published database with information about genomic alterations including short variants, CNVs and gene fusions, expression, gene dependency and druggability. DrugComboExplorer [151] takes into account DNA sequencing, gene copy number, methylation and expression data from cancer patients to (a) identify driver signalling pathways and (b) propose anticancer drug combinations.

Transcriptomic networks can also be enriched with subsequent omics layers to provide functional insight transcending individual layers. COSMOS [152] integrates phosphoproteomics, transcriptomics and metabolomics to estimate the activity of kinases and transcription factors. Finally, deep learning algorithms are becoming promising approaches for multi-omics integration thanks to their capability of capturing nonlinear and hierarchical features [153]. For instance, DeepDRK [154] leverages genomics, transcriptomics, epigenomics and chemical properties of compounds to predict drug susceptibility in both cancer cell lines and patients.

The application of bioinformatics methodologies to immunotherapy as part of precision oncology is still in its early stages. However, tools already exist that allow the design of personalised vaccines [155]. From the large lists of potential neoantigens generated from NGS, it is possible to select those with the highest probability of success, that is to find an optimal design to generate efficient vaccines based on patient-specific neoantigen profiles. Neoantigen prediction pipelines such as pVACtools [156] include different computational tools to detect neoantigens from tumour DNA-seq and RNA-seq data. They also estimate the individual's HLA class and prioritise neoantigens based on the molecular match with the patient's MHC and other parameters [157]. Moreover, there are programs such as CIBERSORTX [158] or MCP-counter [159] capable of inferring the presence of immune infiltrates in tissue from expression data. Knowledge of the type of immune infiltration present in a tumour might serve as a guide, together with TMB values, for treatment selection. Finally, it should be noted that most of the proposed methodologies are still far from being applied in the clinic, although some such as the prioritisation of drug treatments or neoantigens based on TMB content are beginning to be used in clinical trials [155].

5. Targeting tumour heterogeneity: ith and drug selection

ITH functional diversity within individual tumours has been related to somatic SNVs, SVs, transcriptomic and epigenetic changes influencing gene expression levels, the TME status and the antitumour immune response [160,161]. ITH can be spatial if it occurs at different regions of the tumour and temporal when it is related to clonal evolution [14]. We can currently determine the degree of ITH and characterise each subset of clonal subpopulations [Box 5] based on their specific mutational or transcriptomic profiles. The knowledge of ITH can be of great help in prioritising drug treatments or understanding tumour response to treatment. This section provides an overview of relevant methodologies for the dissection of ITH for guiding drug selection (Fig. 1).

5.1. Genome profiling for targeting tumour clonality

Tumours can harbour clonal mutations, which are present in all cells, and subclonal mutations, which only affect a subset of them. The prevalence of subclonal mutations can be used to infer the tumour's phy-

Box 5. Clonal and subclonal subpopulations.

ITH is characterised by the presence of different tumour subclones, each one of them exhibiting a fitness that daughter cells inherit. Subclones can harbour clonal or trunk mutations, which are present in all cells, and subclonal mutations, which only affect a subset of cancer cells. The prevalence of subclonal alterations can be used to infer a tumour's phylogeny. Treatment can be a source of selective pressure that performs purifying selection on sensitive subclones and increments the fitness of resistant ones.

logeny, which allows to decipher the order of these mutations and to identify the current subclones and the relationships between them. However, one must take into account that the ability to distinguish between truly clonal and subclonal mutations appearing to be clonal (pseudo-clonal mutations) depends largely on the number of regions sequenced, the sequencing depth and sample purity [21]. Cancer subclones are subject to Darwinian evolution, and each one of them exhibits a fitness that daughter cells can inherit. Some studies have suggested that increased levels of CNV might be advantageous for the subclone that bears these mutations, which ultimately outcompetes its neighbours [162]. Anticancer drug administration creates a selective pressure that alters subclonal fitness. Drug-sensitive cells will die, but some subclones, which are usually a minority, may acquire resistance to the treatment and increase their fitness. This resistance might be due to pre-existent resistant subclones or can arise by *de novo* drug-induced mutations in drug-tolerant cells. Eventually, resistant subclones may expand and cause relapse [14]. An interesting example of this behaviour is represented by the stem cell division dynamics described by Xie et al. In this work, they characterised a subgroup of quiescent glioblastoma cancer stem cells (CSC) that evaded antiproliferative chemotherapy and re-entered the cell cycle, promoting tumour growth and ultimately leading to conventional treatment failure and relapse [163]. Some authors have proposed a combination of multiregion sampling to dissect spatial ITH coupled with monitoring of circulating tumour DNA (ctDNA) via liquid biopsies to measure clonal evolution in real time and adapt the therapy accordingly [164,165]. Other approaches rely on a Bayesian evolutionary framework to study the spatio-temporal dynamics of cancer subclones within a single patient [166]. Subclones can be identified using several approaches, including genome profiling and single-cell sequencing.

Genome profiling is the preferred strategy to study clonal evolution. Several bioinformatics tools have been developed to infer cancer subclones using SNV allele frequencies, CNV profiles and tumour purity measures as input. The most remarkable examples are PyClone-VI [167], PhyloWGS [168], FastClone [169], SciClone [170] or MOBSTER [171]. However, this approach has several limitations. First, only the mutations that are present in all or the majority of cells will be detected. Moreover, stromal contamination may alter mutation frequencies. Finally, these bioinformatics tools perform many prior inference steps that may introduce errors, which can be propagated in subsequent steps [172].

Some drug prioritisation tools initially designed for intertumour heterogeneity have been used for targeting ITH as well [173]. In this work, PanDrugs was run independently for each inferred subclone and the results aggregated in order to prioritise drugs that hit both clonal and subclonal alterations. The term 'clonetherapy' was introduced to define the optimal treatment regime that would cover patient ITH by targeting all subclones, including the minority ones with the ability to relapse (Fig. 2).

5.2. Single-cell transcriptomics-based drug selection

Bulk RNA-seq allows for the use of the transcriptome as a proxy for elucidating cellular phenotypic traits. This has demonstrated to be useful in uncovering genes important for cancer progression and possible drug targets. However, it involves the averaging of the expression levels in a heterogeneous subpopulation of cells, hiding what might result in important patterns defining tissue dynamics, cell fate and transitions. The idea that the study of predominant cancer subpopulations is insufficient for informing precision oncology has been already suggested in several publications [174,175]. To prevent or to overcome resistance, we need to implement more accurate molecular profiling techniques. Single-cell technologies are able to dissect ITH at the scale of individual tumour cells, revealing rare subpopulations and enhancing our understanding of drug resistance and relapse [176,177]. In this context, the development of single-cell RNA-seq (scRNA-seq) technologies has been seen as a new stepping stone towards an increased understanding of cancer biology.

In recent years, there have been significant advances in the generation of computational tools capable of addressing ITH from a single-cell point of view. However, the current lack of gold standard analysis guidelines is one of the biggest challenges in the field [178]. Importantly, the generation of community-maintained and versatile analysis pipelines could help in solving this issue. The bollito pipeline [179], the Web-Accessible Single Cell RNA-Seq Processing Platform (WASP) [180] and the Single Cell Interactive Application (SCiAp) [181] are some of the latest efforts in this direction. In general terms, current single-cell analysis workflows can be subdivided into three main steps: the raw data processing steps or primary analysis; the normalisation and clustering steps, also known as secondary analysis, and the tertiary analysis that

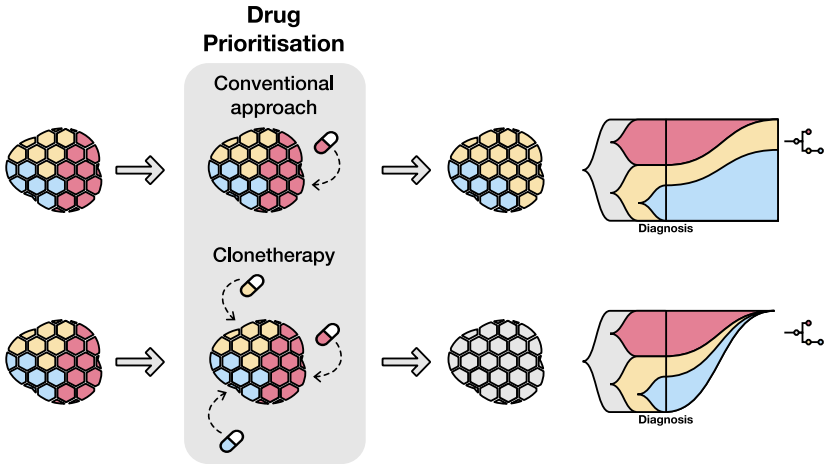

Fig. 2. Concept of clonetherapy. The conventional approach in cancer treatment is to target the major subclone, since it is the most represented in a bulk sample. However, if the drug does not hit clonal alterations, other subclones might survive the treatment and expand. Clonetherapy aims to hit both clonal and subclonal alterations identified in a deconvoluted data set in order to target all subclones, thus avoiding potential relapse.

involves the functional interpretation of the results. Depending on the sequencing platform, it will be necessary to implement some particularities, although the steps will maintain the same goal (Fig. 1).

The primary analysis is the foundation step in the single-cell analysis pipeline and refers to the processing of the raw data. It can be subdivided into sample demultiplexing, alignment, QC and quantification. This is a computationally demanding step with the goal of obtaining a matrix of the gene expression profiles for each cell in the experiment. Demultiplexing is usually performed by the sequencer's built-in software while the most commonly used aligners include Cell Ranger [182] and STARsolo [183]. Pseudo-aligners such as kallisto or Alevin [184], capable of performing an accurate quantification by mapping the reads directly to the transcriptome, are also frequently used because of their increased speed.

The secondary analysis steps include cell-based QC, normalisation, dimensionality reduction and clustering of the samples. They can be performed using popular single-cell analysis toolkits such as Seurat [185] or SCANPY [186]. Nevertheless, it is important to remark that these are not the only methodologies available and that the sparsity (meaning the high fraction of zeroes present in single-cell matrices) of the analysed data set or the amount of sequenced cells should guide the algorithm selection. Interesting reviews from Duò et al. [187] and Yu et al. [188] could be helpful for performing this algorithm selection. The manifold learning algorithms are recommended for further exploratory single-cell data visualisation [189]. For instance, the Uniform Manifold Approximation and Projection (UMAP) [190] is considered best practice thanks to its capacity for preserving both global and local structure [178]. The final goal of all these preprocessing steps is to cluster cells based on the identification of distinct biological patterns, cell types and cell states. Here, it is important to note that the selection of a clustering algorithm will strongly impact further downstream analyses [191]. Common clustering methodologies include PhenoGraph [192], SC3 [193] TSCAN [194] or SINCERA [195].

Finally, the tertiary analysis helps to describe and interpret the functional processes that define the biology of each cell subpopulation and thus enable the study of ITH and facilitate finding suitable therapeutic candidates. These downstream steps involve classic DGE methods such as the Wilcoxon rank-sum test, which has been shown to have an overall robust performance in single-cell data sets [196,197] together with bulk-based methods such as edgeR or limma-voom. Also, single-cell specific functional enrichment methodologies such as VISION [198] and UCell [199] apply bulk-design approaches to individual cells or groups of cells, generating gene signature scores in a similar fashion to bulk methods. Further methodological developments in this context involve prior knowledge-driven cell-type annotation using built-in reference marker collections with SingleR [200] or PanglaoDB [201]. In addition, the study of ligand–receptor interactions between cancer cells and the TME can be crucial for studying the extrinsic factors that contribute to ITH [202] and improving our treatment selections. Tools such as CellPhoneDB [203] or NicheNet [204] are useful for modelling intercellular

communication and linking ligands to target genes. Moreover, trajectory inference and expression dynamic methodologies can help us understand both 'present' and 'future' of the selected subpopulations. While the 'present' of a cell is represented by the captured spliced mRNA transcripts, the analysis of the unspliced mRNA can also be used to predict the cell's future transcriptome, the direction and the speed of that change. SLINGSHOT [205] and VELOCYTO [206] or SCVELO [207] have been developed to recapitulate the transcriptional dynamics within a data set and furthering our understanding of cell transitions. Such methods are complementary, as SLINGSHOT allows the ordering of cells based on their current snapshot, while VELOCYTO or SCVELO facilitates the study of gene regulation by predicting its future steps. Together, all these methodologies facilitate our understanding of the analysed cells, resulting in a better selection of subpopulations of interest and helping in the study of possible clinical targets.

Expression-based drug prioritisation algorithms are usually applied after functional characterisation of cell subpopulations. Recent methods such as DEGAS associate individual cells with disease attributes such as diagnosis, prognosis and response to therapy [208], whereas AUGUR prioritises cell types involved in response to perturbations [209]. BEYONDCELL is a computational method for identifying tumour cell subpopulations with distinct drug responses and proposing cancer-specific treatments. In order to do this, Beyondcell calculates a drug susceptibility score for each cell, delineates therapeutic clusters defined as groups of cells with a similar drug response and generates a prioritised sensitivity-based ranking in order to guide drug selection [26].

Still, the lack of information about the spatial context is one the main drawbacks of scRNA-seq methodologies. This information is of special importance when characterising new subpopulations, since it allows to determine whether the observed differences in expression are a consequence of functional differences or they rely on different interactions with the TME. Additionally, establishing how the TME is going to affect drug tolerance in these subpopulations will be crucial for selecting suitable drug candidates. Spatial transcriptomics (ST) profiling techniques have been recently developed to tackle this question and hold promise of generating much more informed tumoral maps. However, major caveats of this new approach are a lower resolution (still not at the level of single cells) and a lower number of captured genes than scRNA-seq [210]. In this context, integrative analysis methods for ST are part of a trend aimed at generating a common framework of spatial annotation that will help further enriching scRNA-seq data sets. Tools such as TANGRAM [211] or SPAOTSC [212], which map scRNA-seq data to spatial data collected from the same region, could be used to achieve this goal.

6. Incorporating drug prioritisation tools into the clinical practice

Therapy selection guided by bioinformatics approaches is still in its infancy. To date, drug prioritisation methods face technical and biological challenges (Box 6) that constitute clear bottlenecks for their application in routine clinical practice. However, there are currently remarkable efforts to translate these methodologies into medical practices for the benefit of patients.

The patient journey defines the evolution of cancer patients, describing the different stages from disease prevention to detection, diagnosis, treatment and follow-up. To diagnose and decide on the best available treatment options, physicians will need integrated patient's information in a clear and interpretable way through clinical decision support systems. Such systems will be able to efficiently access electronic medical records containing multiple data types, including individual genomic data at different points in the patient journey.

Computational methodologies to analyse and interpret NGS data, including drug prioritisation algorithms, will be incorporated in the clinical decision support systems relying on broad interoperability of data, metadata, research software and computational infrastructure. This will require harmonised nomenclatures, large and well-annotated genomic data sets linked to patients' clinicopathological information and efficient data exchange (Fig. 3). To address this challenge, multimodal cancer data must be meaningfully connected; thus, data harmonisation and standardisation are crucial. There are several ongoing efforts towards this direction. For instance, the Findable, Accessible, Interoperable, Reusable (FAIR) principles have been proposed to facilitate an efficient clinical data exchange [213]. The NIH Data Commons (https://commonfund.nih.gov/commons) and the Cancer Research Data Commons (CRDC, https://datacommons.cancer.gov/) are further examples of data harmonisation initiatives. On the contrary, initiatives promoting the availability of genomic data linked to enriched clinical annotation have been recently launched such as the ICGC-ARGO (https://platform.icgc-argo.org/) [214], which aims to collect a richer data set of cancer genomes with clinical information, health and response to therapy, and the Beyond 1 Million Genomes initiative (B1MG, https://b1mg-project.

Box 6. Main biological and technical challenges associated with the problem of drug prioritisation.

Biological

1 Incomplete dissection of inter- and intratumour heterogeneity and lack of knowledge of somatic evolutionary processes.
2 Poor understanding of the interface between clonal expansion and cancer initiation.
3 Poor understanding of tumour and TME topological relationships and cell–cell cross-talking.
4 Exhaustion of antitumor immunity during disease progression.
5 Poor understanding of the specific events leading to the onset and expansion of drug resistant subclones.
6 Incomplete categorisation of short variants, SVs, epigenomic and transcriptional driver alterations and their relationship with drug response.
7 Poor understanding of interrelations between ageing, senescence and drug response.
8 Missing information about the association of germinal variants and ADRs for most anticancer drugs.

Technical

1 FFPE preparation of samples, which favours DNA fragmentation, degradation and alterations that are difficult to identify as artefacts during variant calling.
2 Trade-off between scope and read depth. In genomics, the broader the region sequenced (all regions using WGS, coding-only using WES or specific genes using targeted sequencing), the lower the coverage. Similarly, the higher the number of sequenced cells, the lower the read depth in single-cell technologies.
3 Dealing with multi-alignment reads due to repetitive regions in the genome.
4 Short reads are not sufficient to resolve large SVs, and long-read sequencing strategies have higher error rates.
5 Lack of gold standard guidelines and information about the spatial context in single-cell technologies.
6 Predicting toxic interactions or synergistic effects of combination therapies.

eu/), which provides a framework for access and interoperability of genomic and medical data [215]. In addition to computational implementations, the incorporation of multi-omics approaches and *in silico* drug prioritisation tools into routine clinical practice will require further efforts in the healthcare scenario (Box 7).

In silico drug prioritisation tool performance would also greatly benefit by extensive and standardised clinical, pathological and genomic annotations integrated in a federated data-sharing model, storing retrospective treatment response information while preserving patients' data privacy. Such a federated data-sharing framework would also provide benchmarking, training and validation data sets for the evaluation of reliability of novel drug response prediction methods and to identify new predictive biomarkers based on retrospective data [173,216]. Thus, clinical decision-making regarding a particular patient would be supported by a genomic report integrating comparative studies of treatment and clinical response obtained from multiple patients with similar genomic profiles. In this sense, the Global Alliance for Genomics and Health (GA4GH) outlines a framework of international policies and standards for the responsible access to genomic and health-related data [217]. Projects such as the GA4GH Genome Beacons provide a pioneer bioinformatics framework for hospitals to interrogate clinicogenomics data without compromising the privacy and the ownership of the data set [218]. Importantly, such a scenario with controlled accession to clinicogenomics information and secure data sharing would also allow for more robust training, testing and validation of novel drug prioritisation methods, ultimately resulting in direct benefit to patients.

7. Conclusions

Cancer is a complex disease that results from the interaction of multiple layers of information. The relationship between tumour origin, the appearance of genomic and transcriptomic variations or microenvironment interactions, all play a role in making tumour treatment challenging. Moreover, cancer is characterised by inter- and intratumour heterogeneity, meaning that molecular alterations at multiple levels vary among tumours from different patients, within the same patient or even among cells within the same tumour. For all these reasons, patients may exhibit different responses to the same treatment. As a consequence, there is an urgent need to develop computational methodologies addressing the design of personalised anticancer treatment regimens [173]. Precision oncology aims to address this scenario by proposing patient-specific treatments tailored to the multi-omics profiles of individual tumours and the

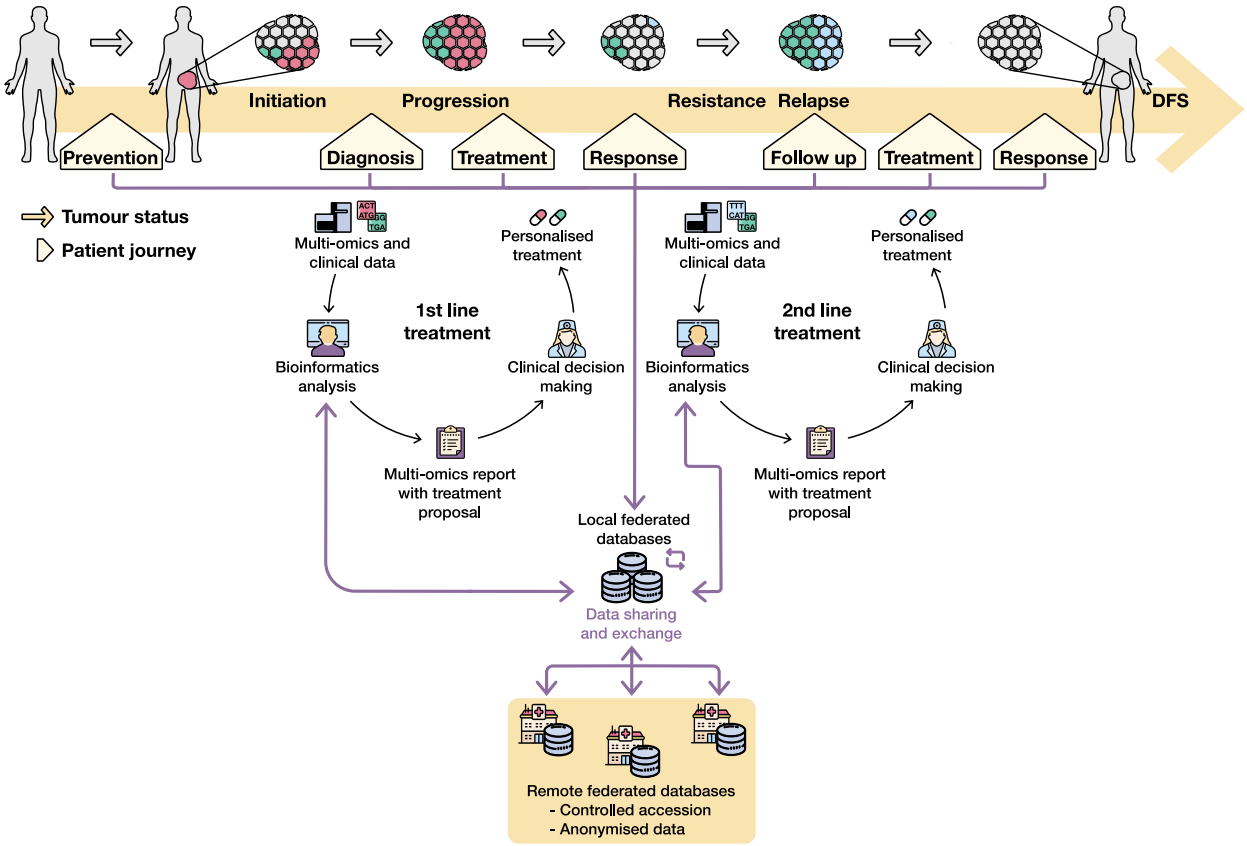

Fig. 3. Integration of drug prioritisation methods for clinical decision-making during the cancer patient journey. Integrated bioinformatics analysis of clinical and multi-omics data from individual cancer patients would generate a report that includes tumour genomics profiling during the patient journey. Based on such profiles, drug prioritisation methods would provide predictions to propose tailored treatments during the different stages of disease progression. The genomics report would be completed with retrospective treatment response information obtained by comparison with other patients with similar clinical and genomic profiles. Data retrieved at each step of the patient journey would be stored in federated databases for aiding future clinical decisions. DFS, disease-free survival.

clinical characteristics of each patient. This challenge cannot be met without bioinformatics, since it requires the development, testing and application of algorithms to interpret multi-source patient data and guide clinical decision-making. There is an extensive catalogue of drug prioritisation methods pursuing to respond to this demand by proposing tailored treatments based on lists of tumour genetic alterations, gene lists or expression profiles.

Box 7. Main challenges for the integration of multi-omics data into routine clinical care.

1 Lack of trained and specialised professionals.
2 Clinical sample accessibility, availability and lack of unified sample processing protocols.
3 Clinical scalability.
4 Lack of standardised gold standard data sets for training and validation of multi-omics data analysis methods.
5 Deficient computational infrastructures.
6 Implementation of data privacy policies.
7 Implementation of legal and ethical frameworks.

This article has reviewed the state of the art in computational drug selection methodologies. It also reviewed the bioinformatics methods currently available for the processing, analysis and interpretation of genomics and transcriptomics data. In particular, the computational approaches used for the dissection, characterisation and drug prioritisation for the therapeutic management of ITH, a major cause of variability in responses to cancer treatment, were also described.

Overall, these computational drug prioritisation methods still rely on the one target–one drug–one disease notion, in contrast to current therapeutic

approaches, which often combine a rational and drug-based synergistic therapeutic regime [219]. Moreover, cancer treatment research has shifted from a cancer-centred model to an TME-centred model [220] and there are still a few methodologies oriented in this direction. Some bioinformatics efforts predict drug combination therapies [221] or suggest TME drug immunomodulators [142] based on omics profiles, but to date, very few methods exist as these areas are underexplored and the challenge remains unsolved. Bioinformatics is crucial to meet the goal of designing precision medicine-based therapies [222] being capable of selecting tailored treatments targeting tumour heterogeneity efficiently and playing a key role in its incorporation into the clinical practice.

Acknowledgements

We thank the whole Bioinformatics Unit staff for useful discussions. CNIO Bioinformatics Unit was supported by the Instituto de Salud Carlos III (ISCIII); Project IMPaCT-Data (Exp. IMP/00019) funded by the Instituto de Salud Carlos III (ISCIII) and cofunded by EU–FEDER 'Una manera de hacer Europa'; Project RETOS RTI2018-097596-B-I00 (AEI/10.13039/501100011033 MCI/FEDER, UE); and Paradifference Foundation.

Conflict of interest

The authors declare no conflict of interest.

Author contributions

MJJ-S and TDD wrote the short variants and structural variants sections. MJJ-S and SG-M described the multi-omics section. MJJ-S also wrote the tumour clonality subsection. SG-M wrote about the transcriptomics-based drug selection. CF-T wrote about the single-cell RNA-seq subsection. MJJ-S, GG-L and FA-S wrote the outline and contributed to all sections. MJJ-S, GG-L and FA-S conceptualised the figures. MJJ-S created the figures. MJJ-S, SG-M and CF-T wrote the tables. All authors read and approved the final manuscript.

References

1 Garraway LA. Genomics-driven oncology: framework for an emerging paradigm. *J Clin Oncol*. 2013;**31**:1806–14. https://doi.org/10.1200/JCO.2012.46.8934

2 Mateo J, Steuten L, Aftimos P, André F, Davies M, Garralda E, et al. Delivering precision oncology to patients with cancer. *Nat Med*. 2022;**28**:658–65. https://doi.org/10.1038/s41591-022-01717-2

3 Hopkins AL, Groom CR. The druggable genome. *Nat Rev Drug Discov*. 2002;**1**:727–30. https://doi.org/10.1038/nrd892

4 US Food and Drug Administration. Table of pharmacogenomic biomarkers. 2022. [cited 2022 April 20]. Available from: http://www.fda.gov/drugs/science-and-research-drugs/table-pharmacogenomic-biomarkers-drug-labeling

5 Vander Velde R, Yoon N, Marusyk V, Durmaz A, Dhawan A, Miroshnychenko D, et al. Resistance to targeted therapies as a multifactorial, gradual adaptation to inhibitor specific selective pressures. *Nat Commun*. 2020;**11**:2393. https://doi.org/10.1038/s41467-020-16212-w

6 Zhong L, Li Y, Xiong L, Wang W, Wu M, Yuan T, et al. Small molecules in targeted cancer therapy: advances, challenges, and future perspectives. *Signal Transduct Target Ther*. 2021;**6**:201. https://doi.org/10.1038/s41392-021-00572-w

7 Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet*. 2013;**45**:1113–20. https://doi.org/10.1038/ng.2764

8 Zhang J, Bajari R, Andric D, Gerthoffert F, Lepsa A, Nahal-Bose H, et al. The International Cancer Genome Consortium Data Portal. *Nat Biotechnol*. 2019;**37**:367–9. https://doi.org/10.1038/s41587-019-0055-9

9 ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature*. 2020;**578**:82–93. https://doi.org/10.1038/s41586-020-1969-6

10 Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*. 2013;**499**:214–8. https://doi.org/10.1038/nature12213

11 Gundem G, Van Loo P, Kremeyer B, Alexandrov LB, Tubio JMC, Papaemmanuil E, et al. The evolutionary history of lethal metastatic prostate cancer. *Nature*. 2015;**520**:353–7. https://doi.org/10.1038/nature14347

12 Ben-David U, Siranosian B, Ha G, Tang H, Oren Y, Hinohara K, et al. Genetic and transcriptional evolution alters cancer cell line drug response. *Nature*. 2018;**560**:325–30. https://doi.org/10.1038/s41586-018-0409-3

13 Saeed K, Ojamies P, Pellinen T, Eldfors S, Turkki R, Lundin J, et al. Clonal heterogeneity influences drug responsiveness in renal cancer assessed by ex vivo drug testing of multiple patient-derived cancer cells. *Int J Cancer*. 2019;**144**:1356–66. https://doi.org/10.1002/ijc.31815

14 Dagogo-Jack I, Shaw AT. Tumour heterogeneity and resistance to cancer therapies. *Nat Rev Clin Oncol*. 2018;**15**:81–94. https://doi.org/10.1038/nrclinonc.2017.166
15 Fisher R, Pusztai L, Swanton C. Cancer heterogeneity: implications for targeted therapeutics. *Br J Cancer*. 2013;**108**:479–85. https://doi.org/10.1038/bjc.2012.581
16 Mroz EA, Tward AD, Pickering CR, Myers JN, Ferris RL, Rocco JW. High intratumor genetic heterogeneity is related to worse outcome in patients with head and neck squamous cell carcinoma. *Cancer*. 2013;**119**:3034–42. https://doi.org/10.1002/cncr.28150
17 Bailey C, Black JRM, Reading JL, Litchfield K, Turajlic S, McGranahan N, et al. Tracking cancer evolution through the disease course. *Cancer Discov*. 2021;**11**:916–32. https://doi.org/10.1158/2159-8290.CD-20-1559
18 GLASS Consortium. Glioma through the looking GLASS: molecular evolution of diffuse gliomas and the Glioma Longitudinal Analysis Consortium. *Neuro Oncol*. 2018;**20**:873–84. https://doi.org/10.1093/neuonc/noy020
19 Rozenblatt-Rosen O, Regev A, Oberdoerffer P, Nawy T, Hupalowska A, Rood JE, et al. The Human Tumor Atlas Network: charting tumor transitions across space and time at single-cell resolution. *Cell*. 2020;**181**:236–49. https://doi.org/10.1016/j.cell.2020.03.053
20 Gavish A, Tyler M, Simkin D, Kovarsky D, Nicolas Gonzalez Castro L, Halder D, et al. The transcriptional hallmarks of intra-tumor heterogeneity across a thousand tumors. *bioRxiv*. 2021;2021.12.19.473368. https://doi.org/10.1101/2021.12.19.473368
21 McGranahan N, Swanton C. Clonal heterogeneity and tumor evolution: past, present, and the future. *Cell*. 2017;**168**:613–28. https://doi.org/10.1016/j.cell.2017.01.018
22 Gohlke B-O, Nickel J, Otto R, Dunkel M, Preissner R. CancerResource – updated database of cancer-relevant proteins, mutations and interacting drugs. *Nucleic Acids Res*. 2016;**44**:D932–7. https://doi.org/10.1093/nar/gkv1283
23 Tamborero D, Rubio-Perez C, Deu-Pons J, Schroeder MP, Vivancos A, Rovira A, et al. Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med*. 2018;**10**:25. https://doi.org/10.1186/s13073-018-0531-8
24 Piñeiro-Yáñez E, Reboiro-Jato M, Gómez-López G, Perales-Patón J, Troulé K, Rodríguez JM, et al. PanDrugs: a novel method to prioritize anticancer drug treatments according to individual genomic data. *Genome Med*. 2018;**10**:41. https://doi.org/10.1186/s13073-018-0546-1
25 Reardon B, Moore ND, Moore NS, Kofman E, AlDubayan SH, Cheung ATM, et al. Integrating molecular profiles into clinical frameworks through the Molecular Oncology Almanac to prospectively guide precision oncology. *Nat Cancer*. 2021;**2**:1102–12. https://doi.org/10.1038/s43018-021-00243-3
26 Fustero-Torre C, Jiménez-Santos MJ, García-Martín S, Carretero-Puche C, García-Jimeno L, Ivanchuk V, et al. Beyondcell: targeting cancer therapeutic heterogeneity in single-cell RNA-seq data. *Genome Med*. 2021;**13**:187. https://doi.org/10.1186/s13073-021-01001-x
27 DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. 2011;**43**:491–8. https://doi.org/10.1038/ng.806
28 Koboldt DC. Best practices for variant calling in clinical sequencing. *Genome Med*. 2020;**12**:91. https://doi.org/10.1186/s13073-020-00791-w
29 Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:13033997v2 [q-bioGN]. 2013. https://doi.org/10.48550/arXiv.1303.3997
30 Broad Institute. Picard toolkit. 2018. [cited 2022 March 28]. Available from: https://github.com/broadinstitute/picard
31 McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;**20**:1297–303. https://doi.org/10.1101/gr.107524.110
32 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. https://doi.org/10.1101/gr.129684.111
33 Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, McEwen R, et al. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res*. 2016;**44**:e108. https://doi.org/10.1093/nar/gkw227
34 Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, Dooling DJ, et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics*. 2012;**28**:311–7. https://doi.org/10.1093/bioinformatics/btr665
35 Sun H, Yu G. New insights into the pathogenicity of non-synonymous variants through multi-level analysis. *Sci Rep*. 2019;**9**:1667. https://doi.org/10.1038/s41598-018-38189-9
36 Bieg-Bourne CC, Millis SZ, Piccioni DE, Fanta PT, Goldberg ME, Chmielecki J, et al. Next-generation sequencing in the clinical setting clarifies patient characteristics and potential actionability. *Cancer Res*. 2017;**77**:6313–20. https://doi.org/10.1158/0008-5472.CAN-17-1569

37 Sanchez-Vega F, Mina M, Armenia J, Chatila WK, Luna A, La KC, et al. Oncogenic signaling pathways in The Cancer Genome Atlas. *Cell*. 2018;**173**:321–37.e10. https://doi.org/10.1016/j.cell.2018.03.035
38 Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, 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. https://doi.org/10.4161/fly.19695
39 Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;**38**:e164. https://doi.org/10.1093/nar/gkq603
40 McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl variant effect predictor. *Genome Biol*. 2016;**17**:122. https://doi.org/10.1186/s13059-016-0974-4
41 Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res*. 2018;**46**:D1062–7. https://doi.org/10.1093/nar/gkx1153
42 Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res*. 2019;**47**:D941–7. https://doi.org/10.1093/nar/gky1015
43 Chakravarty D, Gao J, Phillips SM, Kundra R, Zhang H, Wang J, et al. OncoKB: a precision oncology knowledge base. *JCO Precis Oncol*. 2017;**2017**:PO.17.00011. https://doi.org/10.1200/PO.17.00011
44 Griffith M, Spies NC, Krysiak K, McMichael JF, Coffman AC, Danos AM, et al. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nat Genet*. 2017;**49**:170–4. https://doi.org/10.1038/ng.3774
45 Cotto KC, Wagner AH, Feng Y-Y, Kiwala S, Coffman AC, Spies G, et al. DGIdb 3.0: a redesign and expansion of the drug-gene interaction database. *Nucleic Acids Res*. 2018;**46**:D1068–73. https://doi.org/10.1093/nar/gkx1143
46 Perera-Bel J, Hutter B, Heining C, Bleckmann A, Fröhlich M, Fröhling S, et al. From somatic variants towards precision oncology: evidence-driven reporting of treatment options in molecular tumor boards. *Genome Med*. 2018;**10**:18. https://doi.org/10.1186/s13073-018-0529-2
47 Wagner AH, Walsh B, Mayfield G, Tamborero D, Sonkin D, Krysiak K, et al. A harmonized meta-knowledgebase of clinical interpretations of somatic genomic variants in cancer. *Nat Genet*. 2020;**52**:448–57. https://doi.org/10.1038/s41588-020-0603-8
48 Yu Y, Wang Y, Xia Z, Zhang X, Jin K, Yang J, et al. PreMedKB: an integrated precision medicine knowledgebase for interpreting relationships between diseases, genes, variants and drugs. *Nucleic Acids Res*. 2019;**47**:D1090–101. https://doi.org/10.1093/nar/gky1042
49 Warner JL, Prasad I, Bennett M, Arniella M, Beeghly-Fadiel A, Mandl KD, et al. SMART Cancer Navigator: a framework for implementing ASCO workshop recommendations to enable precision cancer medicine. *JCO Precis Oncol*. 2018;**2018**:PO.17.00292. https://doi.org/10.1200/PO.17.00292
50 Tamborero D, Dienstmann R, Rachid MH, Boekel J, Lopez-Fernandez A, Jonsson M, et al. The Molecular Tumor Board Portal supports clinical decisions and automated reporting for precision oncology. *Nat Cancer*. 2022;**3**:251–61. https://doi.org/10.1038/s43018-022-00332-x
51 Rubio-Perez C, Tamborero D, Schroeder MP, Antolín AA, Deu-Pons J, Perez-Llamas C, et al. In silico prescription of anticancer drugs to cohorts of 28 tumor types reveals targeting opportunities. *Cancer Cell*. 2015;**27**:382–96. https://doi.org/10.1016/j.ccell.2015.02.007
52 Menden MP, Casale FP, Stephan J, Bignell GR, Iorio F, McDermott U, et al. The germline genetic component of drug sensitivity in cancer cell lines. *Nat Commun*. 2018;**9**:3385. https://doi.org/10.1038/s41467-018-05811-3
53 Khalil H, Huang C. Adverse drug reactions in primary care: a scoping review. *BMC Health Serv Res*. 2020;**20**:5. https://doi.org/10.1186/s12913-019-4651-7
54 Roden DM, McLeod HL, Relling MV, Williams MS, Mensah GA, Peterson JF, et al. Pharmacogenomics. *Lancet*. 2019;**394**:521–32. https://doi.org/10.1016/S0140-6736(19)31276-0
55 Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res*. 2018;**46**:D1074–82. https://doi.org/10.1093/nar/gkx1037
56 Whirl-Carrillo M, Huddart R, Gong L, Sangkuhl K, Thorn CF, Whaley R, et al. An evidence-based framework for evaluating pharmacogenomics knowledge for personalized medicine. *Clin Pharmacol Ther*. 2021;**110**:563–72. https://doi.org/10.1002/cpt.2350
57 Sangkuhl K, Whirl-Carrillo M, Whaley RM, Woon M, Lavertu A, Altman RB, et al. Pharmacogenomics Clinical Annotation Tool (PharmCAT). *Clin Pharmacol Ther*. 2020;**107**:203–10. https://doi.org/10.1002/cpt.1568
58 Borchert F, Mock A, Tomczak A, Hügel J, Alkarkoukly S, Knurr A, et al. Knowledge bases and software support for variant interpretation in precision oncology. *Brief Bioinform*. 2021;**22**:bbab134. https://doi.org/10.1093/bib/bbab134
59 Yao H, Liang Q, Qian X, Wang J, Sham PC, Li MJ. Methods and resources to access mutation-dependent

effects on cancer drug treatment. *Brief Bioinform*. 2020;**21**:1886–903. https://doi.org/10.1093/bib/bbz109
60 Hellmann MD, Paz-Ares L. Lung cancer with a high tumor mutational burden. *N Engl J Med*. 2018;**379**:1093–4. https://doi.org/10.1056/NEJMc1808566
61 McGrail DJ, Pilié PG, Rashid NU, Voorwerk L, Slagter M, Kok M, et al. High tumor mutation burden fails to predict immune checkpoint blockade response across all cancer types. *Ann Oncol*. 2021;**32**:661–72. https://doi.org/10.1016/j.annonc.2021.02.006
62 Escaramís G, Docampo E, Rabionet R. A decade of structural variants: description, history and methods to detect structural variation. *Brief Funct Genomics*. 2015;**14**:305–14. https://doi.org/10.1093/bfgp/elv014
63 Li Y, Roberts ND, Wala JA, Shapira O, Schumacher SE, Kumar K, et al. Patterns of somatic structural variation in human cancer genomes. *Nature*. 2020;**578**:112–21. https://doi.org/10.1038/s41586-019-1913-9
64 Ho SS, Urban AE, Mills RE. Structural variation in the sequencing era. *Nat Rev Genet*. 2020;**21**:171–89. https://doi.org/10.1038/s41576-019-0180-9
65 Zarrei M, MacDonald JR, Merico D, Scherer SW. A copy number variation map of the human genome. *Nat Rev Genet*. 2015;**16**:172–83. https://doi.org/10.1038/nrg3871
66 Pös O, Radvanszky J, Styk J, Pös Z, Buglyó G, Kajsik M, et al. Copy number variation: methods and clinical applications. *NATO Adv Sci Inst Ser E Appl Sci*. 2021;**11**:819. https://doi.org/10.3390/app11020819
67 Dong Z, Xie W, Chen H, Xu J, Wang H, Li Y, et al. Copy-number variants detection by low-pass whole-genome sequencing. *Curr Protoc Hum Genet*. 2017;**94**:8.17.1–16. https://doi.org/10.1002/cphg.43
68 Roca I, González-Castro L, Fernández H, Couce ML, Fernández-Marmiesse A. Free-access copy-number variant detection tools for targeted next-generation sequencing data. *Mutat Res Rev Mut Res*. 2019;**779**:114–25. https://doi.org/10.1016/j.mrrev.2019.02.005
69 Raman L, Dheedene A, De Smet M, Van Dorpe J, Menten B. WisecondorX: improved copy number detection for routine shallow whole-genome sequencing. *Nucleic Acids Res*. 2019;**47**:1605–14. https://doi.org/10.1093/nar/gky1263
70 Panel of Normals (PON). 2021. [cited 2022 April 11]. Available from: https://gatk.broadinstitute.org/hc/en-us/articles/360035890631-Panel-of-Normals-PON
71 Baca SC, Prandi D, Lawrence MS, Mosquera JM, Romanel A, Drier Y, et al. Punctuated evolution of prostate cancer genomes. *Cell*. 2013;**153**:666–77. https://doi.org/10.1016/j.cell.2013.03.021
72 Cameron DL, Baber J, Shale C, Valle-Inclan JE, Besselink N, van Hoeck A, et al. GRIDSS2: comprehensive characterisation of somatic structural variation using single breakend variants and structural variant phasing. *Genome Biol*. 2021;**22**:202. https://doi.org/10.1186/s13059-021-02423-x
73 Valle-Inclan JE, Stangl C, de Jong AC, van Dessel LF, van Roosmalen MJ, Helmijr JCA, et al. Optimizing nanopore sequencing-based detection of structural variants enables individualized circulating tumor DNA-based disease monitoring in cancer patients. *Genome Med*. 2021;**13**:86. https://doi.org/10.1186/s13073-021-00899-7
74 Schütte J, Reusch J, Khandanpour C, Eisfeld C. Structural variants as a basis for targeted therapies in hematological malignancies. *Front Oncol*. 2019;**9**:839. https://doi.org/10.3389/fonc.2019.00839
75 Macintyre G, Goranova TE, De Silva D, Ennis D, Piskorz AM, Eldridge M, et al. Copy number signatures and mutational processes in ovarian carcinoma. *Nat Genet*. 2018;**50**:1262–70. https://doi.org/10.1038/s41588-018-0179-8
76 van Belzen IAEM, Schönhuth A, Kemmeren P, Hehir-Kwa JY. Structural variant detection in cancer genomes: computational challenges and perspectives for precision oncology. *NPJ Precis Oncol*. 2021;**5**:15. https://doi.org/10.1038/s41698-021-00155-6
77 Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, et al. Patterns of somatic mutation in human cancer genomes. *Nature*. 2007;**446**:153–8. https://doi.org/10.1038/nature05610
78 Degasperi A, Zou X, Dias Amarante T, Martinez-Martinez A, Koh GCC, Dias JML, et al. Substitution mutational signatures in whole-genome–sequenced cancers in the UK population. *Science*. 2022;**376**. https://doi.org/10.1126/science.abl9283
79 Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. *Nature*. 2020;**578**:94–101. https://doi.org/10.1038/s41586-020-1943-3
80 Waddell N, Pajic M, Patch A-M, Chang DK, Kassahn KS, Bailey P, et al. Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature*. 2015;**518**:495–501. https://doi.org/10.1038/nature14169
81 Ma J, Setton J, Lee NY, Riaz N, Powell SN. The therapeutic significance of mutational signatures from DNA repair deficiency in cancer. *Nat Commun*. 2018;**9**:3292. https://doi.org/10.1038/s41467-018-05228-y
82 Connor AA, Denroche RE, Jang GH, Timms L, Kalimuthu SN, Selander I, et al. Association of distinct mutational signatures with correlates of increased immune activity in pancreatic ductal adenocarcinoma. *JAMA Oncol*. 2017;**3**:774–83. https://doi.org/10.1001/jamaoncol.2016.3916
83 Lord CJ, Ashworth A. BRCAness revisited. *Nat Rev Cancer*. 2016;**16**:110–20. https://doi.org/10.1038/nrc.2015.21

84 Levatić J, Salvadores M, Fuster-Tormo F, Supek F. Mutational signatures are markers of drug sensitivity of cancer cells. *Nat Commun*. 2022;**13**:2926. https://doi.org/10.1038/s41467-022-30582-3
85 Pich O, Muiños F, Lolkema MP, Steeghs N, Gonzalez-Perez A, Lopez-Bigas N. The mutational footprints of cancer therapies. *Nat Genet*. 2019;**51**:1732–40. https://doi.org/10.1038/s41588-019-0525-5
86 Baez-Ortega A, Gori K. Computational approaches for discovery of mutational signatures in cancer. *Brief Bioinform*. 2019;**20**:77–88. https://doi.org/10.1093/bib/bbx082
87 Omichessan H, Severi G, Perduca V. Computational tools to detect signatures of mutational processes in DNA from tumours: a review and empirical comparison of performance. *PLoS One*. 2019;**14**: e0221235. https://doi.org/10.1371/journal.pone.0221235
88 Bergstrom EN, Huang MN, Mahto U, Barnes M, Stratton MR, Rozen SG, et al. SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events. *BMC Genomics*. 2019;**20**:685. https://doi.org/10.1186/s12864-019-6041-2
89 Islam SMA, Díaz-Gay M, Wu Y, Barnes M, Vangara R, Bergstrom EN, et al. Uncovering novel mutational signatures by *de novo* extraction with SigProfilerExtractor. *bioRxiv*. 2020. https://doi.org/10.1101/2020.12.13.422570
90 Bergstrom EN, Barnes M, Martincorena I, Alexandrov LB. Generating realistic null hypothesis of cancer mutational landscapes using SigProfilerSimulator. *BMC Bioinformatics*. 2020;**21**:438. https://doi.org/10.1186/s12859-020-03772-3
91 Kim J, Mouw KW, Polak P, Braunstein LZ, Kamburov A, Kwiatkowski DJ, et al. Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat Genet*. 2016;**48**:600–6. https://doi.org/10.1038/ng.3557
92 Kasar S, Kim J, Improgo R, Tiao G, Polak P, Haradhvala N, et al. Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nat Commun*. 2015;**6**:8866. https://doi.org/10.1038/ncomms9866
93 Haradhvala NJ, Kim J, Maruvka YE, Polak P, Rosebrock D, Livitz D, et al. Distinct mutational signatures characterize concurrent loss of polymerase proofreading and mismatch repair. *Nat Commun*. 2018;**9**:1746. https://doi.org/10.1038/s41467-018-04002-4
94 Degasperi A, Amarante TD, Czarnecki J, Shooter S, Zou X, Glodzik D, et al. A practical framework and online tool for mutational signature analyses show intertissue variation and driver dependencies. *Nat Cancer*. 2020;**1**:249–63. https://doi.org/10.1038/s43018-020-0027-5
95 Davies H, Glodzik D, Morganella S, Yates LR, Staaf J, Zou X, et al. HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat Med*. 2017;**23**:517–25. https://doi.org/10.1038/nm.4292
96 Byron SA, Van Keuren-Jensen KR, Engelthaler DM, Carpten JD, Craig DW. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nat Rev Genet*. 2016;**17**:257–71. https://doi.org/10.1038/nrg.2016.10
97 Cieślik M, Chinnaiyan AM. Cancer transcriptome profiling at the juncture of clinical translation. *Nat Rev Genet*. 2018;**19**:93–109. https://doi.org/10.1038/nrg.2017.96
98 Corchete LA, Rojas EA, Alonso-López D, De Las Rivas J, Gutiérrez NC, Burguillo FJ. Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis. *Sci Rep*. 2020;**10**:1–15. https://doi.org/10.1038/s41598-020-76881-x
99 Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;**29**:15–21. https://doi.org/10.1093/bioinformatics/bts635
100 Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*. 2014;**30**:923–30. https://doi.org/10.1093/bioinformatics/btt656
101 Putri GH, Anders S, Pyl PT, Pimanda JE, Zanini F. Analysing high-throughput sequencing data in Python with HTSeq 2.0. *Bioinformatics*. 2022;**38**:2943–5. https://doi.org/10.1093/bioinformatics/btac166
102 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. https://doi.org/10.1038/nmeth.4197
103 Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics*. 2011;**12**:323. https://doi.org/10.1186/1471-2105-12-323
104 Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol*. 2016;**34**:525–7. https://doi.org/10.1038/nbt.3519
105 Love MI, Soneson C, Hickey PF, Johnson LK, Pierce NT, Shepherd L, et al. Tximeta: reference sequence checksums for provenance identification in RNA-seq. *PLoS Comput Biol*. 2020;**16**:e1007664. https://doi.org/10.1371/journal.pcbi.1007664
106 Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;**15**:550. https://doi.org/10.1186/s13059-014-0550-8

107 Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;**26**:139–40. https://doi.org/10.1093/bioinformatics/btp616

108 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. https://doi.org/10.1186/gb-2014-15-2-r29

109 Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol*. 2012;**8**:e1002375. https://doi.org/10.1371/journal.pcbi.1002375

110 The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res*. 2019;**47**:D330–8. https://doi.org/10.1093/nar/gky1055

111 Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst*. 2015;**1**:417–25. https://doi.org/10.1016/j.cels.2015.12.004

112 Al-Shahrour F, Díaz-Uriarte R, Dopazo J. FatiGO: a web tool for finding significant associations of gene ontology terms with groups of genes. *Bioinformatics*. 2004;**20**:578–80. https://doi.org/10.1093/bioinformatics/btg455

113 Huang DW, Sherman BT, Zheng X, Yang J, Imamichi T, Stephens R, et al. Extracting biological meaning from large gene lists with DAVID. *Curr Protoc Bioinformatics*. 2009;**Chapter 13**:Unit 13.11. https://doi.org/10.1002/0471250953.bi1311s27

114 Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*. 2013;**14**:128. https://doi.org/10.1186/1471-2105-14-128

115 Mi H, Muruganujan A, Thomas PD. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res*. 2013;**41**:D377–86. https://doi.org/10.1093/nar/gks1118

116 Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res*. 2019;**47**:W199–205. https://doi.org/10.1093/nar/gkz401

117 Nguyen T-M, Shafi A, Nguyen T, Draghici S. Identifying significantly impacted pathways: a comprehensive review and assessment. *Genome Biol*. 2019;**20**:203. https://doi.org/10.1186/s13059-019-1790-4

118 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 USA*. 2005;**102**:15545–50. https://doi.org/10.1073/pnas.0506580102

119 Wu D, Smyth GK. Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Res*. 2012;**40**:e133. https://doi.org/10.1093/nar/gks461

120 Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics*. 2013;**14**:7. https://doi.org/10.1186/1471-2105-14-7

121 Tarca AL, Draghici S, Bhatti G, Romero R. Down-weighting overlapping genes improves gene set analysis. *BMC Bioinformatics*. 2012;**13**:136. https://doi.org/10.1186/1471-2105-13-136

122 Foroutan M, Bhuva DD, Lyu R, Horan K, Cursons J, Davis MJ. Single sample scoring of molecular phenotypes. *BMC Bioinformatics*. 2018;**19**:404. https://doi.org/10.1186/s12859-018-2435-4

123 Geistlinger L, Csaba G, Santarelli M, Ramos M, Schiffer L, Turaga N, et al. Toward a gold standard for benchmarking gene set enrichment analysis. *Brief Bioinform*. 2021;**22**:545–56. https://doi.org/10.1093/bib/bbz158

124 Ihnatova I, Popovici V, Budinska E. A critical comparison of topology-based pathway analysis methods. *PLoS One*. 2018;**13**:e0191154. https://doi.org/10.1371/journal.pone.0191154

125 Dutta B, Wallqvist A, Reifman J. PathNet: a tool for pathway analysis using topological information. *Source Code Biol Med*. 2012;**7**:10. https://doi.org/10.1186/1751-0473-7-10

126 Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;**13**:2498–504. https://doi.org/10.1101/gr.1239303

127 Kutmon M, van Iersel MP, Bohler A, Kelder T, Nunes N, Pico AR, et al. PathVisio 3: an extendable pathway analysis toolbox. *PLoS Comput Biol*. 2015;**11**:e1004085. https://doi.org/10.1371/journal.pcbi.1004085

128 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. https://doi.org/10.1093/nar/gkv1277

129 Menden MP, Wang D, Mason MJ, Szalai B, Bulusu KC, Guan Y, et al. Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. *Nat Commun*. 2019;**10**:2674. https://doi.org/10.1038/s41467-019-09799-2

130 Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of

anticancer drug sensitivity. *Nature*. 2012;**483**:603–7. https://doi.org/10.1038/nature11003

131 Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, 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. https://doi.org/10.1093/nar/gks1111

132 Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*. 2017;**171**:1437–52.e17. https://doi.org/10.1016/j.cell.2017.10.049

133 Rees MG, Seashore-Ludlow B, Cheah JH, Adams DJ, Price EV, Gill S, et al. 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

134 Yu C, Mannan AM, Yvone GM, Ross KN, Zhang Y-L, Marton MA, et al. 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

135 Tsherniak A, Vazquez F, Montgomery PG, Weir BA, Kryukov G, Cowley GS, et al. Defining a cancer dependency map. *Cell*. 2017;**170**:564–76.e16. https://doi.org/10.1016/j.cell.2017.06.010

136 Ghandi M, Huang FW, Jané-Valbuena J, Kryukov GV, Lo CC, McDonald ER 3rd, et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature*. 2019;**569**:503–8. https://doi.org/10.1038/s41586-019-1186-3

137 Corsello SM, Nagari RT, Spangler RD, Rossen J, Kocak M, Bryan JG, et al. Discovering the anti-cancer potential of non-oncology drugs by systematic viability profiling. *Nat Cancer*. 2020;**1**:235–48. https://doi.org/10.1038/s43018-019-0018-6

138 Chen B, Ma L, Paik H, Sirota M, Wei W, Chua M-S, et al. Reversal of cancer gene expression correlates with drug efficacy and reveals therapeutic targets. *Nat Commun*. 2017;**8**:16022. https://doi.org/10.1038/ncomms16022

139 Cheng C, Zhao Y, Schaafsma E, Weng Y-L, Amos C. An EGFR signature predicts cell line and patient sensitivity to multiple tyrosine kinase inhibitors. *Int J Cancer*. 2020;**147**:2621–33. https://doi.org/10.1002/ijc.33053

140 Pacini C, Iorio F, Gonçalves E, Iskar M, Klabunde T, Bork P, et al. DvD: an R/Cytoscape pipeline for drug repurposing using public repositories of gene expression data. *Bioinformatics*. 2013;**29**:132–4. https://doi.org/10.1093/bioinformatics/bts656

141 Fu J, Li K, Zhang W, Wan C, Zhang J, Jiang P, et al. Large-scale public data reuse to model immunotherapy response and resistance. *Genome Med*. 2020;**12**:21. https://doi.org/10.1186/s13073-020-0721-z

142 Troulé K, López-Fernández H, García-Martín S, Reboiro-Jato M, Carretero-Puche C, Martorell-Marugán J, et al. DREIMT: a drug repositioning database and prioritization tool for immunomodulation. *Bioinformatics*. 2021;**37**:578–9. https://doi.org/10.1093/bioinformatics/btaa727

143 Warren A, Chen Y, Jones A, Shibue T, Hahn WC, Boehm JS, et al. Global computational alignment of tumor and cell line transcriptional profiles. *Nat Commun*. 2021;**12**:22. https://doi.org/10.1038/s41467-020-20294-x

144 Yoo M, Shin J, Kim J, Ryall KA, Lee K, Lee S, et al. DSigDB: drug signatures database for gene set analysis. *Bioinformatics*. 2015;**31**:3069–71. https://doi.org/10.1093/bioinformatics/btv313

145 Di J, Zheng B, Kong Q, Jiang Y, Liu S, Yang Y, et al. Prioritization of candidate cancer drugs based on a drug functional similarity network constructed by integrating pathway activities and drug activities. *Mol Oncol*. 2019;**13**:2259–77. https://doi.org/10.1002/1878-0261.12564

146 Domingo-Fernández D, Gadiya Y, Patel A, Mubeen S, Rivas-Barragan D, Diana CW, et al. Causal reasoning over knowledge graphs leveraging drug-perturbed and disease-specific transcriptomic signatures for drug discovery. *PLoS Comput Biol*. 2022;**18**:e1009909. https://doi.org/10.1371/journal.pcbi.1009909

147 Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. *Genome Biol*. 2017;**18**:83. https://doi.org/10.1186/s13059-017-1215-1

148 do Valle ÍF, Menichetti G, Simonetti G, Bruno S, Zironi I, Durso DF, et al. Network integration of multi-tumour omics data suggests novel targeting strategies. *Nat Commun*. 2018;**9**:4514. https://doi.org/10.1038/s41467-018-06992-7

149 Kalari KR, Sinnwell JP, Thompson KJ, Tang X, Carlson EE, Yu J, et al. PANOPLY: omics-guided drug prioritization method tailored to an individual patient. *JCO Clin Cancer Inform*. 2018;**2**:1–11. https://doi.org/10.1200/CCI.18.00012

150 Jiang J, Yuan J, Hu Z, Zhang Y, Zhang T, Xu M, et al. Systematic illumination of druggable genes in cancer genomes. *Cell Rep*. 2022;**38**:110400. https://doi.org/10.1016/j.celrep.2022.110400

151 Huang L, Brunell D, Stephan C, Mancuso J, Yu X, He B, et al. Driver network as a biomarker: systematic integration and network modeling of multi-omics data to derive driver signaling pathways for drug combination prediction. *Bioinformatics*. 2019;**35**:3709–17. https://doi.org/10.1093/bioinformatics/btz109

152 Dugourd A, Kuppe C, Sciacovelli M, Gjerga E, Gabor A, Emdal KB, et al. Causal integration of

multi-omics data with prior knowledge to generate mechanistic hypotheses. *Mol Syst Biol*. 2021;**17**:e9730. https://doi.org/10.15252/msb.20209730

153 Kang M, Ko E, Mersha TB. A roadmap for multi-omics data integration using deep learning. *Brief Bioinform*. 2022;**23**:bbab454. https://doi.org/10.1093/bib/bbab454

154 Wang Y, Yang Y, Chen S, Wang J. DeepDRK: a deep learning framework for drug repurposing through kernel-based multi-omics integration. *Brief Bioinform*. 2021;**22**:bbab048. https://doi.org/10.1093/bib/bbab048

155 Keskin DB, Anandappa AJ, Sun J, Tirosh I, Mathewson ND, Li S, et al. Neoantigen vaccine generates intratumoral T cell responses in phase Ib glioblastoma trial. *Nature*. 2019;**565**:234–9. https://doi.org/10.1038/s41586-018-0792-9

156 Hundal J, Kiwala S, McMichael J, Miller CA, Xia H, Wollam AT, et al. pVACtools: a computational toolkit to identify and visualize cancer neoantigens. *Cancer Immunol Res*. 2020;**8**:409–20. https://doi.org/10.1158/2326-6066.CIR-19-0401

157 Hackl H, Charoentong P, Finotello F, Trajanoski Z. Computational genomics tools for dissecting tumour-immune cell interactions. *Nat Rev Genet*. 2016;**17**:441–58. https://doi.org/10.1038/nrg.2016.67

158 Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat Biotechnol*. 2019;**37**:773–82. https://doi.org/10.1038/s41587-019-0114-2

159 Becht E, Giraldo NA, Lacroix L, Buttard B, Elarouci N, Petitprez F, et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol*. 2016;**17**:218. https://doi.org/10.1186/s13059-016-1070-5

160 Black JRM, McGranahan N. Genetic and non-genetic clonal diversity in cancer evolution. *Nat Rev Cancer*. 2021;**21**:379–92. https://doi.org/10.1038/s41568-021-00336-2

161 Nguyen PHD, Ma S, Phua CZJ, Kaya NA, Lai HLH, Lim CJ, et al. Intratumoural immune heterogeneity as a hallmark of tumour evolution and progression in hepatocellular carcinoma. *Nat Commun*. 2021;**12**:227. https://doi.org/10.1038/s41467-020-20171-7

162 Salehi S, Kabeer F, Ceglia N, Andronescu M, Williams MJ, Campbell KR, et al. Clonal fitness inferred from time-series modelling of single-cell cancer genomes. *Nature*. 2021;**595**:585–90. https://doi.org/10.1038/s41586-021-03648-3

163 Xie XP, Laks DR, Sun D, Ganbold M, Wang Z, Pedraza AM, et al. Quiescent human glioblastoma cancer stem cells drive tumor initiation, expansion, and recurrence following chemotherapy. *Dev Cell*. 2022;**57**:32–46.e8. https://doi.org/10.1016/j.devcel.2021.12.007

164 Amirouchene-Angelozzi N, Swanton C, Bardelli A. Tumor evolution as a therapeutic target. *Cancer Discov*. 2017;**7**:805–17. https://doi.org/10.1158/2159-8290.CD-17-0343

165 Siravegna G, Marsoni S, Siena S, Bardelli A. Integrating liquid biopsies into the management of cancer. *Nat Rev Clin Oncol*. 2017;**14**:531–48. https://doi.org/10.1038/nrclinonc.2017.14

166 Alves JM, Prado-López S, Cameselle-Teijeiro JM, Posada D. Rapid evolution and biogeographic spread in a colorectal cancer. *Nat Commun*. 2019;**10**:5139. https://doi.org/10.1038/s41467-019-12926-8

167 Gillis S, Roth A. PyClone-VI: scalable inference of clonal population structures using whole genome data. *BMC Bioinformatics*. 2020;**21**:571. https://doi.org/10.1186/s12859-020-03919-2

168 Deshwar AG, Vembu S, Yung CK, Jang GH, Stein L, Morris Q. PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol*. 2015;**16**:35. https://doi.org/10.1186/s13059-015-0602-8

169 Xiao Y, Wang X, Zhang H, Ulintz PJ, Li H, Guan Y. FastClone is a probabilistic tool for deconvoluting tumor heterogeneity in bulk-sequencing samples. *Nat Commun*. 2020;**11**:4469. https://doi.org/10.1038/s41467-020-18169-2

170 Miller CA, White BS, Dees ND, Griffith M, Welch JS, Griffith OL, et al. SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS Comput Biol*. 2014;**10**:e1003665. https://doi.org/10.1371/journal.pcbi.1003665

171 Caravagna G, Sanguinetti G, Graham TA, Sottoriva A. The MOBSTER R package for tumour subclonal deconvolution from bulk DNA whole-genome sequencing data. *BMC Bioinformatics*. 2020;**21**:531. https://doi.org/10.1186/s12859-020-03863-1

172 Turajlic S, Sottoriva A, Graham T, Swanton C. Resolving genetic heterogeneity in cancer. *Nat Rev Genet*. 2019;**20**:404–16. https://doi.org/10.1038/s41576-019-0114-6

173 Piñeiro-Yáñez E, Jiménez-Santos MJ, Gómez-López G, Al-Shahrour F. In silico drug prescription for targeting cancer patient heterogeneity and prediction of clinical outcome. *Cancer*. 2019;**11**:1361. https://doi.org/10.3390/cancers11091361

174 Hata AN, Niederst MJ, Archibald HL, Gomez-Caraballo M, Siddiqui FM, Mulvey HE, et al. Tumor cells can follow distinct evolutionary paths to become resistant to epidermal growth factor receptor inhibition. *Nat Med*. 2016;**22**:262–9. https://doi.org/10.1038/nm.4040

175 Marchetti A, Milella M, Felicioni L, Cappuzzo F, Irtelli L, Del Grammastro M, et al. Clinical implications of KRAS mutations in lung cancer patients treated with tyrosine kinase inhibitors: an important role for mutations in minor clones. *Neoplasia*. 2009;**11**:1084–92. https://doi.org/10.1593/neo.09814

176 Kan T, Zhang S, Zhou S, Zhang Y, Zhao Y, Gao Y, et al. Single-cell RNA-seq recognized the initiator of epithelial ovarian cancer recurrence. *Oncogene*. 2022;**41**:895–906. https://doi.org/10.1038/s41388-021-02139-z

177 Candelli T, Schneider P, Garrido Castro P, Jones LA, Bodewes E, Rockx-Brouwer D, et al. Identification and characterization of relapse-initiating cells in MLL-rearranged infant ALL by single-cell transcriptomics. *Leukemia*. 2022;**36**:58–67. https://doi.org/10.1038/s41375-021-01341-y

178 Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol Syst Biol*. 2019;**15**:e8746. https://doi.org/10.15252/msb.20188746

179 García-Jimeno L, Fustero-Torre C, Jiménez-Santos MJ, Gómez-López G, Di Domenico T, Al-Shahrour F. bollito: a flexible pipeline for comprehensive single-cell RNA-seq analyses. *Bioinformatics*. 2021;**38**:1155–6. https://doi.org/10.1093/bioinformatics/btab758

180 Hoek A, Maibach K, Özmen E, Vazquez-Armendariz AI, Mengel JP, Hain T, et al. WASP: a versatile, web-accessible single cell RNA-seq processing platform. *BMC Genomics*. 2021;**22**:195. https://doi.org/10.1186/s12864-021-07469-6

181 Moreno P, Huang N, Manning JR, Mohammed S, Solovyev A, Polanski K, et al. User-friendly, scalable tools and workflows for single-cell RNA-seq analysis. *Nat Methods*. 2021;**18**:327–8. https://doi.org/10.1038/s41592-021-01102-w

182 Zheng GXY, 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

183 Kaminow B, Yunusov D, Dobin A. STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. *bioRxiv*. 2021;2021.05.05.442755. https://doi.org/10.1101/2021.05.05.442755

184 Srivastava A, Malik L, Smith T, Sudbery I, Patro R. Alevin efficiently estimates accurate gene abundances from dscRNA-seq data. *Genome Biol*. 2019;**20**:65. https://doi.org/10.1186/s13059-019-1670-y

185 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;**184**:3573–87.e29. https://doi.org/10.1016/j.cell.2021.04.048

186 Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol*. 2018;**19**:15. https://doi.org/10.1186/s13059-017-1382-0

187 Duò A, Robinson MD, Soneson C. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Res*. 2018;**7**:1141. https://doi.org/10.12688/f1000research.15666.3

188 Yu L, Cao Y, Yang JYH, Yang P. Benchmarking clustering algorithms on estimating the number of cell types from single-cell RNA-sequencing data. *Genome Biol*. 2022;**23**:49. https://doi.org/10.1186/s13059-022-02622-0

189 Moon KR, Stanley JS, Burkhardt D, van Dijk D, Wolf G, Krishnaswamy S. Manifold learning-based methods for analyzing single-cell RNA-sequencing data. *Curr Opin Syst Biol*. 2018;**7**:36–46. https://doi.org/10.1016/j.coisb.2017.12.008

190 McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction. arXiv:180203426 [statML]. 2018. https://doi.org/10.48550/arXiv.1802.03426

191 Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat Rev Genet*. 2019;**20**:273–82. https://doi.org/10.1038/s41576-018-0088-9

192 Levine JH, Simonds EF, Bendall SC, Davis KL, Amir E-AD, Tadmor MD, et al. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell*. 2015;**162**:184–97. https://doi.org/10.1016/j.cell.2015.05.047

193 Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, et al. SC3: consensus clustering of single-cell RNA-seq data. *Nat Methods*. 2017;**14**:483–6. https://doi.org/10.1038/nmeth.4236

194 Ji Z, Ji H. TSCAN: pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res*. 2016;**44**:e117. https://doi.org/10.1093/nar/gkw430

195 Guo M, Wang H, Potter SS, Whitsett JA, Xu Y. SINCERA: a pipeline for single-cell RNA-seq profiling analysis. *PLoS Comput Biol*. 2015;**11**:e1004575. https://doi.org/10.1371/journal.pcbi.1004575

196 Squair JW, Gautier M, Kathe C, Anderson MA, James ND, Hutson TH, et al. Confronting false discoveries in single-cell differential expression. *Nat Commun*. 2021;**12**:5692. https://doi.org/10.1038/s41467-021-25960-2

197 Soneson C, Robinson MD. Bias, robustness and scalability in single-cell differential expression analysis. *Nat Methods*. 2018;**15**:255–61. https://doi.org/10.1038/nmeth.4612

198 DeTomaso D, Jones MG, Subramaniam M, Ashuach T, Ye CJ, Yosef N. Functional interpretation of single cell similarity maps. *Nat Commun*. 2019;**10**:4376. https://doi.org/10.1038/s41467-019-12235-0

199 Andreatta M, Carmona SJ. UCell: robust and scalable single-cell gene signature scoring. *Comput Struct Biotechnol J*. 2021;**19**:3796–8. https://doi.org/10.1016/j.csbj.2021.06.043
200 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**:163–72. https://doi.org/10.1038/s41590-018-0276-y
201 Franzén O, Gan L-M, Björkegren JLM. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database*. 2019;**2019**: baz046. https://doi.org/10.1093/database/baz046
202 Lüönd F, Tiede S, Christofori G. Breast cancer as an example of tumour heterogeneity and tumour cell plasticity during malignant progression. *Br J Cancer*. 2021;**125**:164–75. https://doi.org/10.1038/s41416-021-01328-7
203 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. https://doi.org/10.1038/s41596-020-0292-x
204 Browaeys R, Saelens W, Saeys Y. NicheNet: modeling intercellular communication by linking ligands to target genes. *Nat Methods*. 2020;**17**:159–62. https://doi.org/10.1038/s41592-019-0667-5
205 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**:477. https://doi.org/10.1186/s12864-018-4772-0
206 La Manno G, Soldatov R, Zeisel A, Braun E, Hochgerner H, Petukhov V, et al. RNA velocity of single cells. *Nature*. 2018;**560**:494–8. https://doi.org/10.1038/s41586-018-0414-6
207 Bergen V, Lange M, Peidli S, Wolf FA, Theis FJ. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat Biotechnol*. 2020;**38**:1408–14. https://doi.org/10.1038/s41587-020-0591-3
208 Johnson TS, Yu CY, Huang Z, Xu S, Wang T, Dong C, et al. Diagnostic Evidence GAuge of Single cells (DEGAS): a flexible deep transfer learning framework for prioritizing cells in relation to disease. *Genome Med*. 2022;**14**:11. https://doi.org/10.1186/s13073-022-01012-2
209 Squair JW, Skinnider MA, Gautier M, Foster LJ, Courtine G. Prioritization of cell types responsive to biological perturbations in single-cell data with Augur. *Nat Protoc*. 2021;**16**:3836–73. https://doi.org/10.1038/s41596-021-00561-x
210 Longo SK, Guo MG, Ji AL, Khavari PA. Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. *Nat Rev Genet*. 2021;**22**:627–44. https://doi.org/10.1038/s41576-021-00370-8
211 Biancalani T, Scalia G, Buffoni L, Avasthi R, Lu Z, Sanger A, et al. Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nat Methods*. 2021;**18**:1352–62. https://doi.org/10.1038/s41592-021-01264-7
212 Cang Z, Nie Q. Inferring spatial and signaling relationships between cells from single cell transcriptomic data. *Nat Commun*. 2020;**11**:2084. https://doi.org/10.1038/s41467-020-15968-5
213 Kush RD, Warzel D, Kush MA, Sherman A, Navarro EA, Fitzmartin R, et al. FAIR data sharing: the roles of common data elements and harmonization. *J Biomed Inform*. 2020;**107**:103421. https://doi.org/10.1016/j.jbi.2020.103421
214 ICGC ARGO. ICGC ARGO data platform. 2022. [cited 2022 April 19]. Available from: https://platform.icgc-argo.org/
215 ELIXIR Europe. Beyond 1 million genomes. 2020. [cited 2022 April 19]. Available from: https://b1mg-project.eu/
216 Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med*. 2022;**28**:31–8. https://doi.org/10.1038/s41591-021-01614-0
217 Rehm HL, Page AJH, Smith L, Adams JB, Alterovitz G, Babb LJ, et al. GA4GH: international policies and standards for data sharing across genomic research and healthcare. *Cell Genom*. 2021;**1**:100029. https://doi.org/10.1016/j.xgen.2021.100029
218 Fiume M, Cupak M, Keenan S, Rambla J, de la Torre S, Dyke SOM, et al. Federated discovery and sharing of genomic data using Beacons. *Nat Biotechnol*. 2019;**37**:220–4. https://doi.org/10.1038/s41587-019-0046-x
219 Gilad Y, Gellerman G, Lonard DM, O'Malley BW. Drug combination in cancer treatment-from cocktails to conjugated combinations. *Cancer*. 2021;**13**:669. https://doi.org/10.3390/cancers13040669
220 Jin M-Z, Jin W-L. The updated landscape of tumor microenvironment and drug repurposing. *Signal Transduct Target Ther*. 2020;**5**:166. https://doi.org/10.1038/s41392-020-00280-x
221 Preuer K, Lewis RPI, Hochreiter S, Bender A, Bulusu KC, Klambauer G. DeepSynergy: predicting anti-cancer drug synergy with deep learning. *Bioinformatics*. 2018;**34**:1538–46. https://doi.org/10.1093/bioinformatics/btx806
222 Gómez-López G, Dopazo J, Cigudosa JC, Valencia A, Al-Shahrour F. Precision medicine needs pioneering clinical bioinformaticians. *Brief Bioinform*. 2017;**20**:752–66. https://doi.org/10.1093/bib/bbx144
223 Rodgers G, Austin C, Anderson J, Pawlyk A, Colvis C, Margolis R, et al. Glimmers in illuminating the

druggable genome. *Nat Rev Drug Discov*. 2018;**17**:301–2. https://doi.org/10.1038/nrd.2017.252

224 Moore L, Cagan A, Coorens THH, Neville MDC, Sanghvi R, Sanders MA, et al. The mutational landscape of human somatic and germline cells. *Nature*. 2021;**597**:381–6. https://doi.org/10.1038/s41586-021-03822-7

225 Perales-Patón J, Di Domenico T, Fustero-Torre C, Piñeiro-Yáñez E, Carretero-Puche C, Tejero H, et al. vulcanSpot: a tool to prioritize therapeutic vulnerabilities in cancer. *Bioinformatics*. 2019;**35**:4846–8. https://doi.org/10.1093/bioinformatics/btz465

226 Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010. [cited 2022 March 25]. Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

227 Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*. 2016;**32**:3047–8. https://doi.org/10.1093/bioinformatics/btw354

228 Wingett SW, Andrews S. FastQ screen: a tool for multi-genome mapping and quality control. *F1000Res*. 2018;**7**:1338. https://doi.org/10.12688/f1000research.15931.2

229 Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J*. 2011;**17**. https://doi.org/10.14806/ej.17.1.200

230 Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;**30**:2114–20. https://doi.org/10.1093/bioinformatics/btu170

231 Bushnell B. BBMap: a fast, accurate, splice-aware aligner. 2014. [cited 2022 March 25]. Available from: https://www.osti.gov/servlets/purl/1241166

232 Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol*. 2019;**37**:907–15. https://doi.org/10.1038/s41587-019-0201-4