biomedicines

MDPI

Review

Bioinformatics: From NGS Data to Biological Complexity in Variant Detection and Oncological Clinical Practice

Serena Dotolo [1], Riziero Esposito Abate [1], Cristin Roma [1], Davide Guido [2], Alessia Preziosi [2], Beatrice Tropea [2], Fernando Palluzzi [2], Luciano Giacò [2] and Nicola Normanno [1,*]

[1] Cell Biology and Biotherapy Unit, Istituto Nazionale Tumori—IRCCS—Fondazione G. Pascale, 80131 Naples, Italy
[2] Bioinformatics Research Core Facility, Gemelli Science and Technology Park (GSTeP), Fondazione Policlinico Universitario Agostino Gemelli IRCCS, Largo A. Gemelli, 8, 00168 Rome, Italy
* Correspondence: n.normanno@istitutotumori.na.it

Abstract: The use of next-generation sequencing (NGS) techniques for variant detection has become increasingly important in clinical research and in clinical practice in oncology. Many cancer patients are currently being treated in clinical practice or in clinical trials with drugs directed against specific genomic alterations. In this scenario, the development of reliable and reproducible bioinformatics tools is essential to derive information on the molecular characteristics of each patient's tumor from the NGS data. The development of bioinformatics pipelines based on the use of machine learning and statistical methods is even more relevant for the determination of complex biomarkers. In this review, we describe some important technologies, computational algorithms and models that can be applied to NGS data from Whole Genome to Targeted Sequencing, to address the problem of finding complex cancer-associated biomarkers. In addition, we explore the future perspectives and challenges faced by bioinformatics for precision medicine both at a molecular and clinical level, with a focus on an emerging complex biomarker such as homologous recombination deficiency (HRD).

Keywords: NGS; variant calling; cancer; biological complexity; ML/AI algorithms; network analysis; homologous recombination deficiency; targeted therapy; big data

Citation: Dotolo, S.; Esposito Abate, R.; Roma, C.; Guido, D.; Preziosi, A.; Tropea, B.; Palluzzi, F.; Giacò, L.; Normanno, N. Bioinformatics: From NGS Data to Biological Complexity in Variant Detection and Oncological Clinical Practice. *Biomedicines* **2022**, *10*, 2074. https://doi.org/10.3390/biomedicines10092074

Academic Editor: Thomas Mohr

Received: 13 July 2022
Accepted: 22 August 2022
Published: 24 August 2022

1. Introduction

Genomic profiling has assumed an increasing role in the clinical management of cancer patients, thanks to the approval of numerous drugs showing demonstrated activity in patients with specific genomic alterations [1–4]. The need to identify an increasing number of complex genomic biomarkers led to the introduction of next-generation sequencing (NGS) technologies in clinical practice. The massive amount of data generated by NGS experiments required the development of algorithms based on computationally efficient statistical methods and artificial intelligence, to improve the processes of genomic variant detection, visualization, and interpretation in terms of pathogenicity [5–8]. The implementation of bioinformatics approaches has a clear intention to elaborate and process data in an efficient and fast way to avoid turnaround times and increase detection accuracy [9–13]. Although new computational solutions have gradually been proposed, showing incredibly high levels of accuracy, many issues in clinical practice remain unsolved. The most important of these challenges is represented by our ability to interpret the potential impact of genomic alterations on a patient's health, and how this information can be used to tune personalized therapies [14,15]. In this review, we will describe some important technologies, computational algorithms and models that can be applied to NGS data ranging from Whole Genome to Targeted Sequencing to address the biological complexity of cancer. Finally, we will explore the future perspectives and challenges faced by bioinformatics for precision medicine both at a molecular and clinical level, with a special focus on Homologous Recombination Deficiency (HRD) as an emerging biomarker in clinical practice.

2. NGS Technologies and Bioinformatics Tools for DNA Sequencing Data

NGS Approaches in Cancer Patients' Management

Different sequencing approaches have been recently used to analyze the genomic landscape of tumors [16]. Traditionally, Sanger sequencing based on electrophoresis and involving the random incorporation of chain-terminated dideoxynucleotides by DNA polymerase during in vitro DNA replication requires time-consuming analyses on a single DNA fragment [17]. On the other hand, NGS can sequence millions of fragments simultaneously per run during automated cycles of synthesis, scanning and washing, playing a critical role in reading the mutational landscape of a patient in clinical routine. This process, applied to hundreds to thousands of loci per subject, generates an enormous amount of sequence data, with many possible applications in research and diagnostic settings including sequence variation detection (referred to as "variant calling"), epigenetic and transcriptional regulation, chromatin conformation, its 3D architecture, and how these phenomena influence each other [18,19]. This review will focus on variant calling and its preprocessing stages.

The primary output of NGS data analysis is the alignment of small DNA or RNA fragments (known as "reads") using either a pre-assembled reference genome sequence (or "assembly") or concatenating input reads using "de novo" strategies, in which no reference sequence is used. Diverse approaches are optimized for genomic analysis at different scales: from a few genes to the whole genome [20–22].

Whole Genome Sequencing (WGS) covers the whole genome, and it is used to investigate previously undescribed genomic alterations, requiring more time and higher costs [23,24]. Whole Exome Sequencing (WES) may cover protein-coding genes only, representing 3% of the whole genome, but with reduced costs and under the assumption that protein-associated alterations have often a deleterious impact on genome regulation [25–28]. Still, the complexity of data interpretation limits significantly the use of WES in clinical research and practice. Consequently, Targeted Sequencing (TS) was introduced for analyzing specific mutational hotspots, for a given genome [29–31]. This approach is used to detect disease-causing genomic alterations with described or suspected pathogenicity [32–36].

The typical NGS workflow is divided into several steps, including sample preprocessing, library preparation, sequencing and bioinformatics analysis. Each step plays a critical role and might hide sources of error that could propagate to the final output. Currently, variant detection and annotation are well-standardized procedures, performed through the following steps [37–39]

A typical variant calling workflow involves the following data preprocessing steps: (i) quality filtering of the raw FASTQ files; (ii) read alignment through either reference-based or de novo alignment (BAM or CRAM files); (iii) duplicate reads removal from the alignment and mapping quality filtering; and (iv) local realignment and/or haplotype determination (phasing) [40–43].

Modern variant callers usually take BAM files from read aligners as input and perform only the last preprocessing step. As explained in Section 4, recent deep learning-based callers also include image-like alignment preprocessing (image pileup) and chromatin conformation analysis. Although the specific variant detection methods may vary for different algorithms, variant calling methods can be grouped on the basis of the input sample origin (germline or somatic), and by variant type: small nucleotide variants (SNVs); small insertions/deletions (indels); and structural variants, including copy number variants (CNVs) and large genomic rearrangements (such as insertions, deletions, and translocations) [44–49]. (Figure 1).

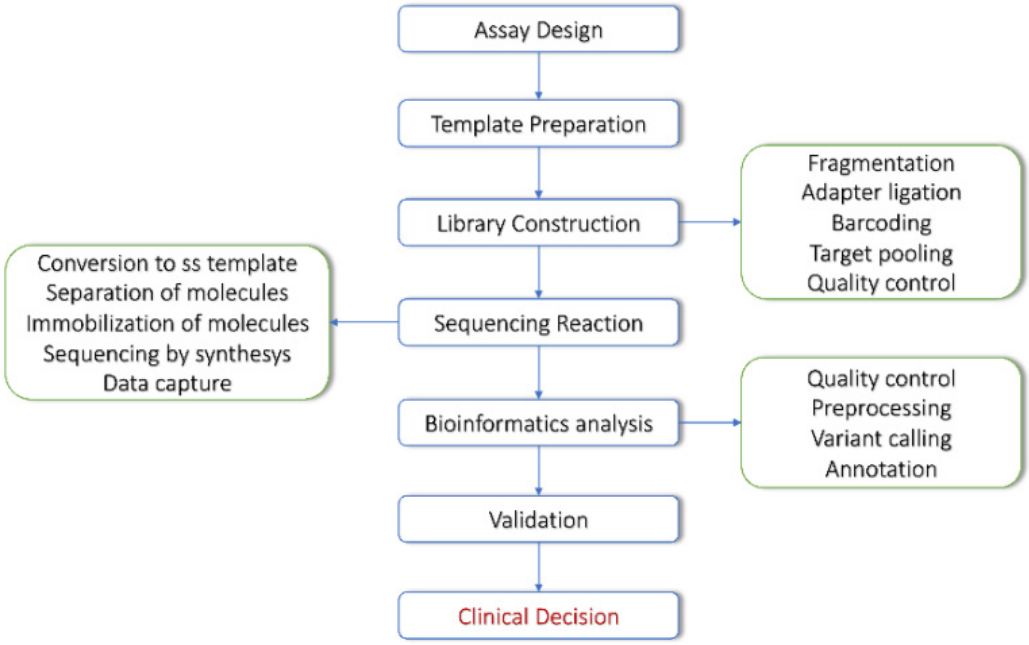

Figure 1. NGS workflow.

3. From NGS Data to Variant Discovery

3.1. Variant Types Relevant for Precision Oncology

The type of alterations that must be reported in clinical diagnostics and that might have a relevant clinical value include Single Nucleotide Variants (SNV), small insertion/deletion (InDel), Copy Number Variants (CNV) and genomic rearrangements that can lead to gene fusions. In addition, cancer-related complex biomarkers, such as Tumor Mutational Burden (TMB) and Microsatellite Instability (MSI) and HRD, are becoming frequently included in clinical reports for their clinical value.

3.2. Variant Discovery Workflow

The variant detection workflow is a sequence of steps, which include the sequencing quality control, the preparation of data (pre-processing) and the use of algorithms able to detect the genomic alterations. To date, the tools used for the variant discovery on tumor samples include three main steps: preprocessing, calling of variants, and annotation.

3.3. Sequencing Quality Control

A flexible, robust and most used tool in quality control is FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ accessed on 12 July 2022), developed at the Babraham Institute to examine sequencing quality, starting from fastq files. It works on all Operating Systems (OS) and can be used with both GUI interface and command line.

This option is commonly used by bioinformaticians, to add the quality control step in a custom pipeline. The latest versions of FastQC include Picard (https://github.com/broadinstitute/picard/ accessed on 12 July 2022), a tool developed by the Broad Institute to manage SAM, BAM, and VCF files, and to perform the quality control at different steps of the bioinformatics pipeline. A fast and simple tool to calculate the coverage starting from BAM files is Mosdepth [50]. It may calculate the coverage depth for both whole genome and exome sequencing data. It is also able to limit the analysis to a specific genomic region providing a bed file. This application could be useful also for targeted sequencing, especially for custom panels.

3.4. Pre Processing

The data preparation is described in the GATK Best Practices [51] and it is a compulsory step in order to provide the correct input to the variant detection algorithms. Many tools

are used to provide the alignment, and to ensure the management of the duplicates and the recalibration phase. The final output is a BAM file ready to be analyzed for the variant's identification. The preprocessing step has been consolidated over time thanks to The GATK Best Practices, developed by Broad Institute, and implemented in the GATK ToolKit [52]. This procedure allows us to produce the alignment files in the best possible way to investigate the presence of alterations in the sequenced genome.

3.5. Variant Calling

The variant calling process is the main step for DNA alteration discovery. It includes different algorithms able to find potentially pathogenic mutations across the human genome. The College of American Pathologists and the American Medical Informatics defined a list of 17 recommendations for clinical NGS bioinformatics pipelines [7]. These statements include, but are not limited to: (i) the involvement of medical personnel, (ii) stage design, (iii) version control, and (iv) reproducibility.

Quality control of each step of a bioinformatics pipeline is crucial to setting optimal parameters, achieving the best possible variant calling performance, and passing the validation step using a representative set of known variants across the samples. Special attention must be given to software versioning and data integrity, in order to track any changes/updates and prevent loss of information/data.

The validation procedure involves variants selection, sequencing quality control, algorithms, filtering and annotation. While the variants selection is a process carried out by scientists with different skills and training, the analysis performed in the sequencing data quality control are consolidated. Concerning the variant calling, there are as many tools as there are alterations to investigate. The main difference is between germline and somatic variants. Although the available algorithms are quite different, due to the intrinsic difference between germinal and somatic variants, the Input/Output file formats are the same: input BAM file(s) and output VCF file(s). The critical point in the variant calling is the filters applied to prevent false positive and false negative events. This depends also on the sequencing coverage/depth and the length of sequenced genomic DNA (e.g., panels vs. exome).

Although many SNV/indel detection algorithms have now reached high accuracy in several benchmarking tests [53], combining the results of multiple algorithms may increase sensitivity, thus reducing the rate of false negatives [54]. Different tools, such as BCFtools [55], enable the merging of multiple calls into a single VCF file. However, clinical practice generally requires clear and portable workflows, generating reproducible results. Furthermore, standard clinical procedures should encompass multiple variant types (SNVs, indels, and structural variants) and genomic alteration measures, including LOH and MSI, readily usable for targeted therapy. To this end, a huge effort has been provided by the nf-core group [56] through the Sarek pipeline [50] for germline and somatic variant detection. Although these tools often use different methods, they generally achieve comparable accuracy. The Sarek pipeline uses all the tools shown in Table 1, depending on the purpose (germline or somatic variant call) and the variant type (small variants or large genomic rearrangements).

3.6. Variant Discovery Pipelines: Tools and Algorithms

In this scenario, commercial and open-source pipelines are often integrated to provide efficient and customizable solutions. One of the players in the development of these tools, offering licensed software, is the Illumina® company. Illumina® adopted the GATK best practices as a consequence of a partnership with the Broad Institute. This produced a series of Illumina licensed and Broad open-source tools derived, available on the Broad Institute repository (https://broadinstitute.github.io/warp/ accessed on 12 July 2022). Both licensed and open-source solutions are released as "only for research" software and need to be validated by the institutions adopting them. The advantage of licensed software consists of its "ready-to-use" design. In fact, it does not need additional software and

database installation procedures, thanks to container technology. On the other hand, licensed software offers less flexibility during the analysis flow than open-source tools. This limitation arises when, starting from raw data (usually, bcl or fastq files), they run almost uninterruptedly through preprocessing, quality control, variant calling, biomarker analysis (including MSI, TMB), and reporting. The consequence is that the whole analysis cannot be split for custom pipeline development and integration with other (possibly newer and more efficient) resources. In addition, combined output reports from licensed software often include only minimal SNV information, are uneasy to read for clinicians, and require further annotation and processing of the vcf files. Consequently, Illumina® offers commercial solutions to produce clinical reports that make use of collaborations with other companies. Conversely, open-source solutions enable user control over each analysis step, allowing single analyses to run as independent modules. By definition, open-source tools need clinical validation to be adopted in diagnostics.

Table 1. Tools included in the nf-core Sarek framework. The table reports the tool name, the sample type (G: germline, S: somatic), the variant type (SNV: small nucleotide variant, indel: small insertion/deletion, SV: structural variant, CNV: copy number variant, MSI: microsatellite instability), a small description of the core method, and the latest literature reference link.

Tool	Sample Type	Variant Type	Method	Ref
Manta	G, S	SV, indels	Graph-based breakend analysis	[57]
TIDDIT	G, S	SV	Coverage-based genome scan	[58]
Cnvkit	G, S	CNV	Coverage-based genome scan	[59]
Freebayes	G, S	SNV, indels	Haplotype-based Bayes theorem	[60]
Strelka2	G, S	SNV, indels	Haplotype-based mixture modeling	[61]
DeepVariant	G	SNV, indels	Pileup image CNN classification	[62]
HaplotypeCaller	G	SNV, indels	Haplotype re-assembly, likelihood	[63]
Mpileup	G	SNV, indels	Local re-alignment, likelihood	[55]
Mutect2	S	SNV, indels	GATK + read-to-haplotype alignment	[64]
Ascat	S	CNV	Signal intensity and allele frequency	[65]
Control-FREEC	S	CNV	LASSO-based genome segmentation	[66]
MSIsensor-pro	S	MSI	Multinomial distribution	[67]

Note. G: germline, S: Somatic; SV: Structural variant; SNV: small nucleotide variant, indel: small insertion/deletion; CNV: copy number variant, MSI: microsatellite instability.

Recently, the nf-core community aimed at defining standard procedures to be included within bioinformatics analysis workflows. The goal of nf-core is to adopt the best practices in bioinformatics pipeline development, through an open-source and peer-reviewed community, in order to offer a reproducible, portable and robust solution in different fields of application. In this context, the Sarek pipeline [68] was developed for germline and somatic variants calling (Figure 2).

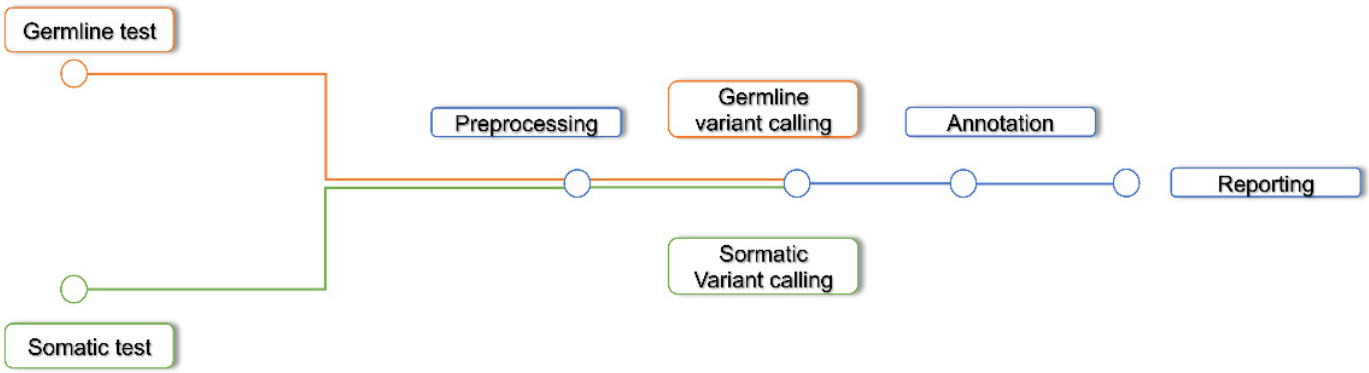

Figure 2. Nf-core Sarek pipeline.

Sarek is designed to analyze data from WGS, WES, and TS. It allows us to perform the analysis starting from several intermediate steps, such as preprocessing or variant calling. In addition, the tools included in the pipeline workflow several variant calling types: germline, somatic, and tumor-only somatic. Consequently, researchers can break the analysis at any step, and, taking advantage of object-oriented programming, every result is an object that could be reused for separate analyses. Thanks to this philosophy, several analyses or quality controls are included and may be added in a continuous process of integration. As a result of this flexibility, the end user can either run the pipeline in a default mode, using a single command line or build a complex job, by using different tools as independent objects in several pipelines. Both nf-core Sarek and GATK are developed as object-oriented software, where the main difference is that the nf-core community allows to include other validated and peer-reviewed algorithms, with the advantage of complete reproducibility.

3.7. Annotation

Variant annotation is the process of assigning information to DNA variants and evaluating their possible pathogenicity. The annotation step is another crucial point and it is the bridge between machine and human-readable format. Furthermore, a punctual annotation of the variants is strongly linked to the query of updated databases. To this end, different open-source and licensed solutions offering reviewed and updated annotations, as VEP [69], VarSome [70], ClinVar [71], OncoKB [72], have been developed.

4. Machine Learning Applied to Next-Generation Sequencing and Variant Discovery

4.1. Sequencing Technology Issues and Machine Learning

The widespread of bio-medical machine learning (ML) methods followed the evolution of high throughput sequencing (such as NGS) technologies. This opened up a new class of diagnostic tools, drug discovery methods, novel patient stratification approaches, and personalized therapies. Besides the large amount of publicly accessible NGS data from consortia [73], the availability of low-cost sequencing platforms caused a worldwide growth of in-house data, and the consequent increasing demand for computationally-efficient, yet accurate and reusable, ML-based software. One of the toughest challenges tackled by ML in clinical genomics is to model what (and possibly how) genomic variants and their interactions influence cell development and fate, leading to cancer transformation [74]. Although classical inference methods can be highly flexible and interpretable in terms of causality, they are often constrained by linearity or model-based assumptions, providing aggregate (e.g., averaged) or partial descriptions of the underlying biological processes (e.g., neglecting nonlinear systems properties) due to a strong dependence on current knowledge or field-expert validation. Traditional methods, such as GATK [75], make extensive use of different statistical models and heuristics based on calling quality, allele frequency, and sequencing coverage, to estimate the likelihood of variation at each genomic position. However, this task is severely hampered by the presence of sequencing artifacts deriving, for instance, from polymerase chain reaction (PCR) errors, DNA synthesis dephasing and inefficiency, and low-complexity or repetitive genomic sequences, that are only partially manageable [76,77]. Moreover, sequencing data is inherently high-dimensional (i.e., the number of observed variants is way bigger than the number of sequenced genomes), the same phenotype can be caused by different combinations of variants (heterogeneity), and often only a few individuals carry the variation associated with the observed disease (sparseness). In addition, a phenotype rarely arises due to the presence of a single deleterious variant, but rather from (hierarchical) interactions among variants with marginal effects [78].

Collectively, these issues motivated the dissemination of ML algorithms, for several reasons. Firstly, their ability to learn patterns of interactions directly from data and, secondly, the capability to model complex hierarchical and nonlinear interactions without specific statistical modeling assumptions. The main costs for these advantages are the need

for very large labeled training sets (i.e., supervised learning), with an obvious impact on the computational demand, and potential vulnerability to the training set compositional biases, often caused by incomplete or imbalanced knowledge of specific domains [62].

4.2. ML and Deep Neural Networks for Variant Discovery

Given their capability of modeling a very large number of features and parameters, deep neural networks have been extensively used for variant discovery. The main idea beneath convolutional neural networks (CNNs) is to convert pileups of aligned reads into patterns of an image, resulting in groups of interconnected variants that might have a pathogenic effect [79].

CNN-based algorithms include the general-purpose software DeepVariant [62], the specialized Clairvoyante [79], NeuSomatic [80], and DeepSV [81], designed for single-molecule technologies, somatic variants, and structural variants, respectively. In many cases, improvements in predictive accuracy have been achieved using ensemble methods, where the learning process takes advantage of different integrated models. For instance, CNNScoreVariants [82] exploits pre-trained models in GATK to discover SNV and indels from short-read sequencing data.

The Clair algorithm [79], the successor of Clairvoyante, uses CNNs in combination with recurrent neural networks (RNNs) and feedforward networks to refine germline SNV and indel discovery. While these methods generally outperform traditional inference methods for SNVs and indels, much less effort has been spent on the more complex structural variants (SVs). To this end, DeepSV [83] uses CNNs to find large (>50 bp) genomic rearrangements, including insertions, deletions, and inversions.

One main limitation of many deep learning algorithms resides in possible information biases within their training sets [84]. The goal of variant discovery is to find genomic loci that are causally associated with the disruption of one or more molecular functions and pathways. For coding DNA, pathogenic alterations are likely to alter the structure and function of the encoded protein and therefore are much easier to be associated with a diseased phenotype. Accordingly, most of the diagnostic procedures in clinical and cancer genomics are either based on panels of a limited set of exons or WES [85–87]. However, many of the deleterious traits of disease are caused by noncoding variants that are likely to be located at regulatory elements [87].

To cope with this possible bias, the DeepSEA [88] and Basset [89] algorithms use CNNs to predict the chromatin state and chromatin accessibility that may reveal the presence of regulatory elements. Both DeepSEA and Basset learn the regulatory sequence code from genomic sequence by training a deep CNN over large chromatin-profiling data from ENCODE and Roadmap Epigenomics consortia. These data include transcription factor binding, DNase I sensitivity and histone-mark profiles. Learning from data-driven features rather than annotations (e.g., exons) allows these algorithms to detect noncoding variants with a possible regulatory role. In addition, the deep neural network structure allows us to scale on sequence length, enabling the use of large contextual genomic regions and further improving noncoding variant function interpretation. An alternate application of these methods consists in validating and converting ML results in current knowledge, improving and speeding up current clinical trial protocols, providing tools for efficient (low attrition) patient stratification strategies, and feature reduction (denoising) [82,85]. More recently, Hi-C, a sequencing-based technique to detect the three-dimensional architecture of the nuclear genome, has been shown to effectively detect structural variants in B-cell acute lymphoblastic leukemia, a form of cancer that is frequently characterized by translocations [90]. Although promising, this approach is still non-standard in clinical practice and is currently used for research-only purposes.

4.3. Machine Learning Development Frameworks

The use of NGS technologies in clinical practice introduced the need for variant calling methods reproducibility and reusability. This favored the development of several dedicated

open-source ML development libraries, and the diffusion of object-oriented programming languages in bioinformatics, including Python and R.

Convolutional kernels are the most exploited for variant calling [62,78], often combined with other architectures in ensemble methods [86]. Less frequently, other paradigms are used, including support vector machines (SVM) [91] and non-supervised learning [92], although they are generally restricted to a limited range of applications.

The landscape of ML-based variant discovery methods is dominated by deep neural networks, mainly due to the large amount of publicly available NGS data that can be used for training, validation/testing, and benchmarking against several gold standards (i.e., manually validated datasets with well-known outcomes), allowing these methodologies to outperform other competing methods (e.g., support vector machines, naïve Bayes, or random forests) [62]. This favored the creation of environments to easily develop custom NGS objects, methods, and tools for ML-based analyses.

TensorFlow [93] is the most used environment for the development of variant discovery AI-based software, followed by PyTorch [94].

Often, these environments make use of tools, such as Keras (url: https://github.com/fchollet/keras/ accessed on 12 July 2022) and Nucleus (https://blog.tensorflow.org/using-nucleus-and-tensorflow-for-dna.html accessed on 12 July 2022), offering a user-friendly experience and dedicated objects for sequencing data analysis. The bioinformatics community also uses the R environment, with dedicated packages and development tools, including the R port Torch [95]. However, R libraries are generally used for statistical computing, and currently could be less performant with respect to Python for ML development, which lists a much larger number of dedicated ML solutions.

5. Network-Based Approaches Applied to Cancer Research: Graph Theory and Causality for Analyzing the Biological Complexity

5.1. Graph Theory

Networks can be explored by the graph theory, useful to shed a light on their structure–function relationships [83]. In fact, graph-based approaches have been applied in extensive ways in different frameworks such as biology, chemistry, medicine, etc., [96] by providing a number of characterizations. Specifically, as described by Lecca et al., systems biology conceptualizes the networks of interacting molecules, and graph theory gives the mathematical tools to analyze them [96].

In particular, network analysis can use quantitative approaches to also model interactions between genes, proteins and other biological elements [97–99]. A general expression to refer to the investigation and modeling of these interactions is "molecular network", which is becoming very important in cancer research as demonstrated in the applications against different types of neoplasm [100–102]: pancreas, gastric, lung, ovarian cancers and others are applications of graph theory in this framework.

Moreover, the graph theory allows us to decompose molecular networks in different subnetworks by directed subgraphs and multigraphs as demonstrated by Huang et al., modeling cancer networks, signal transduction networks, and cellular processes [100,101]. Over the years, different software has emerged in order to analyze the biological networks by the graph theory. A number of these are included in the R environment (https://www.R-project.org/ accessed on 12 July 2022), such as igraph (https://igraph.org/ accessed on 12 July 2022) [102], graph (https://github.com/ accessed on 12 July 2022), QuACN [103], network [104], Statnet [104] (https://statnet.org/ accessed on 12 July 2022) and NetBioV [105].

They are free packages that provide many functions to manage network systems, also by the Bioconductor platform (https://www.bioconductor.org/ accessed on 12 July 2022). These have many graphical functions, often inherited by the R environment, and the interesting advantage to follow object-oriented programming, that is suitable to use the elementary elements of a network in an independent and customized way. On the commer-

cial side, the Dragon [106] software includes thousands of molecular descriptors to be used to analyze the biological network.

5.2. Causality

In the last decade, several methods have tried to model and quantify the causality in the biological and molecular networks, especially by considering the relationships among genes in a framework of perturbation of experiments and in presence of unfavorable factors. As a matter of fact, many phenomena in biology, medicine and other disciplines consider relationships among variables in a multivariate causal context. Hence, investigation and analysis of cause–effect relationships through statistical methods are incrementing, in order to explain how to test causal hypotheses, especially with a lack of randomized experiments [107]. Specifically, the methods try to translate the causal network into mathematical equations by generating assumptions on the nature -random or deterministic- of the variables (nodes of the network), and on the type -unidirectional or bidirectional- of the relationships (edges).

However, as suggested by Palluzzi et al. [108], a number of algorithms have been recommended to model and quantify causality in (biological/molecular) networks but they have low reproducibility and robustness, dependence on user-defined setup, and poor interpretability. In this framework, the structural equation models (SEM) provide a favorable methodology able to model and quantify the causality by an inferential approach, with an immediate and easy interpretation of the results [109]. In the SEM the relationships are assumed to be linear and the (response) variables supposed random are assumed to be multivariate normal. In the past few years, the SEM is catching on in cancer research as demonstrated by articles related to the modeling of the molecular networks in breast cancer [110], colorectal cancer [109], neuroblastoma [111] and leukemia, and more in general, in precision medicine [112]. At the same time, different packages emerged to analyze the causality of biological and molecular networks by the SEM. The majority is developed in the R environment. Firsts among everything, the lavaan [112] and SEMgraph [113] packages allow us to convert the causal diagram of the network into linear equations containing free (to be estimated) and fixed parameters. Of note, SEMgraph is a lavaan-based package that specifically manages complex biological systems as multivariate networks ensuring robustness and reproducibility through data-driven evaluation of model architecture and perturbation; that is readily interpretable in terms of causal effects among system components [108]. Finally, the other two R packages apply SEM specifically in a biological/molecular framework, GenomicSEM and GW-SEM [114], useful for modeling (i) the multivariate genetic structure of correlated traits by using a multivariate GWAS framework, and (ii) the associations of SNPs with phenotypes or hidden constructs on a genome-wide scale.

6. Homologous Recombination Deficiency: A New Bioinformatics Challenge

The complexity of biomarkers to be analyzed in clinical research and clinical practice is progressively increasing. This evolution also creates a new challenge for the bioinformatics analysis of sequencing data. As an example, in the following paragraphs, we describe the different strategies for the analysis of HRD, an emerging biomarker in oncology.

6.1. Testing Strategies for HRD Detection Based on Causes and Effects in the Genome

The homologous recombination repair (HRR) pathway is an error-free mechanism able to repair the DNA double-strand breaks (DSBs) during the S/G2 phases of the cell cycle [108,115]. In cells with a deficit of the HRR mechanism, the repair of DSBs occurs through alternative, error-prone methods, resulting in a high degree of genomic instability and in the accumulation of different alterations, including Single Nucleotide Variations (SNVs), small insertions and deletions (InDels), Copy number variation (CNV) or large-scale chromosomal rearrangements. Tumors with HRD are sensitive to poly (ADP-ribose) polymerase (PARP) inhibitors (PARPi), which suppress a second key DNA repair pathway

of DNA single-strand breaks (SSBs) and create synthetic lethality in cancer cells with defects in HRR [116]. In particular, the PARPi Olaparib has been approved by the US Food and Drug Administration (FDA), in combination with bevacizumab, and by the European Medicines Agency (EMA), as a single agent, for the treatment of patients with advanced ovarian cancer associated with HRD-positive status and who are in complete or partial response to first-line platinum-based chemotherapy. FDA recently approved the PARPi Niraparib for the treatment of HRD-positive ovarian cancers, who have been treated with three or more prior chemotherapy regimens and who have progressed more than six months after responding to the last platinum-based chemotherapy [117–119]. In agreement with the FDA and EMA indications, the HRD status is defined by either a deleterious or suspected deleterious mutation in BRCA1 or BRCA2 genes, and/or genomic instability (GIS). To date, there are two principal strategies to identify tumors with HRD. The first strategy is focused on the identification of HRD causes using targeted sequencing with multi-gene panels able to evaluate alterations in the different genes of the HRR [120–122]. However, this approach has several limitations. Multi-gene panels are able to identify only the fraction of HRD cases related to genetic alteration of HRR genes. Indeed, the HRD might be caused by both genetic and epigenetic events. In addition, only some genomic alterations of HRR genes are associated with HRD [123]. Furthermore, the pathogenic role of many mutations of HRR genes is not known, due also to their very low frequency [124]. The second strategy is based on the study of the effects that HRD causes in the genome and looks for the genomic damage induced, independently from the originating mechanism [125–127]. These approaches vary from the analysis of genomic scars to the assessment of mutational signatures [128,129]. Genomic scars are the complex genomic alterations caused by HRD and represent a biomarker to identify patients who may benefit from treatment with PARPi [130]. Two commercial genomic scar assays have been developed to identify tumors with HRD, the "Myriad myChoice HRD" (Myriad Genetics; Salt Lake City, UT, USA) and the "FoundationOne CDx" (Foundation Medicine; Cambridge, MA, USA) tests.

The "myChoice HRD" assay is an NGS-based test able to detect variants in BRCA1 and BRCA2 genes and to determine a GIS score by measurement of three biomarkers: telomeric allelic imbalance (TAI), loss of heterozygosity (LOH) and large-scale transitions (LST) [131]. The HRD score is calculated by combining the LOH, TAI, and LST scores: tumors with a score $\geq$42 are classified as HRD-positive. This assay is the only one FDA approved for use in clinical practice [130]. The FoundationOne CDx is able to identify patients with HRD-positive status combining tumor BRCA1/2 mutational status with the rate of LOH [132,133]. The HRD score is measured as the percent of LOH in the tumor genome: genomic LOH $\geq$16% is classified as HRD-positive [132].

Despite the excellent results within several clinical trials, the HRD test based on genomic scar still has some technical limitations [134]. In addition, the presence of GIS based on a genomic scar can only indicate that at the time of testing the tumor had HRD. Indeed, the HRD scoring method is unable to account for reversion mutations that are predictive of platinum and PARPi resistance [135]. Conversely, newer approaches to HRD detection, including the identification of mutational signatures in sequencing data, potentially provide a dynamic readout of the current HRR status [136–138]. Multiple mutational processes generate a characteristic pattern of somatic mutations, termed "signatures" [139]. Multiple studies showed that one of these signatures, namely Signature 3 (Sig3), is associated with a deficiency in the HRR mechanism [140,141]. This mutational signature is based on SNVs and might only in part represent the complex genomic alterations associated with HRD. To further decode this complexity, a method based on the identification of signatures from copy-number (CN) features was developed. In the study by Macintyre et al., the CN signature 3, characterized by a distribution of breaks across all chromosomes and LOH, was significantly enriched in cases that displayed HRD caused by mutations in BRCA1/2 [142]. Moreover, signature 7 was associated with HRD and mutations in other HR genes, including BARD1, PALB2 and ATR, and loss of function mutations in PTEN [141]. Although these approaches, based on mutation or CNA signatures, have shown a good

correlation with HR status, it is likely that a combination of different parameters can more accurately identify all cases with HRD [142].

6.2. Computational Tools for HRD Assessment

In recent years, machine learning algorithms and new computational tools have also been developed to perform a more complete and detailed analysis of the genomic alterations related to HRD, in order to better identify HRD-positive tumors [143–145]. Each algorithm has its own characteristics and specifications, applied to sequencing data from WGS, WES and TS. For the complete and detailed study of mutational signatures, it is possible to adopt different algorithms, such as HRDetect [146], Mutalisk [147], SigMA [148,149]. The application of each of these tools depends on the origin of the sequencing data, and therefore the available information.

HRDetect is primarily a mutational signature-based classifier designed to predict BRCA1 and BRCA2 deficiency based on six mutational signatures. It uses a lasso logistic regression model starting from sequencing data of WGS for HRD detection [146]. In particular, it allows us to calculate the HRD score recognizing the patterns of substitution base signatures and structural rearrangements. The HRDetect pipeline works on mutational data, such as: segments.tsv, somatic_indels.vcf, somatic_snvs.vcf, somatic_sv.tsv. The segments.tsv is used to identify and analyze the CNVs and the LOH score; while somatic_indels.vcf and somatic_snvs.vcf are used to study indels and SNVs in detail. Finally, the somatic_sv.tsv file is used to analyze the structural data of the variants. HRDetect has already been used on cohorts of patients with ovarian cancer, breast cancer and pancreatic cancer. The parameters used for HRD assessment include the evaluation of the main mutational signatures (such as Sig3), large deletions (>3 bp) with microhomology at the junction of the deletion, Rearrangement Signatures 3 and 5, and copy number profiles associated with widespread LOH [146]. The final output is a probability of BRCA1/2 mutation. The sensitivity and reliability of the results obtained with HRDetect changes according to the source of the data. By analyzing WGS data, the HRDetect reaches a sensitivity of 86%, setting the cut-off at 0.7 and the level of agreement at r = 0.96 as optimal parameters. By contrast, when HRDetect is applied to data obtained by WES, the sensitivity of detection is 46.8% [146,150].

To improve the results of HRDetect it is possible to use two different tools named Mutalisk and SigMA. Mutalisk (Mutation AnaLyIs ToolKit: www.mutalisk.org/ accessed on 12 July 2022) is an online computational framework used to investigate the signatures at a somatic level. This tool can be applied to genomic data generated by WGS, WES and TS sequencing using as input data the standard vcf file. There are two versions of Mutalisk, one is the web server and the other one is the R vs. 4.1.1. This algorithm identifies a maximum of seven mutational signatures at most from a specific somatic tissue [147]. For each signature set, a decomposition model can be generated using the maximum likelihood estimation method, or the multinomial test. It can be applied to HRD analysis for the identification of mutational signatures and for the classification of molecular processes mainly involved in the generation of pathogenic or benign mutations [147].

By contrast, SigMA (Signature Multivariate Analysis) algorithm performs a mapping of the most important mutational signatures from the SNV calls of WGS, WES or TS data associated with the HRD pathway [151]. This algorithm has a high sensitivity of 74% in identifying Sig3-derived rearrangements in HRD-positive tumors. The novelty of this algorithm is the application of the likelihood method, which allows us to associate a mutational spectrum to each patient [152].

Recently, a new tool named Classifier of Homologous Recombination Deficiency (**CHORD**) was developed (https://github.com/UMCUGenetics/CHORD/ accessed on 12 July 2022) for the detection of HRD status by BRCA1 and BRCA2 deficiency. CHORD is a random forest model used as a benchmark developed to detect pan-cancer HRD based on genome-wide mutational profiles using specific SNV, indels, and structural variants (SV) [153]. The CHORD algorithm uses deletions with flanking microhomology

and 1–100 kb structural duplications to distinguish BRCA1-type HRD from BRCA2-type HRD [154]. The analysis is divided into two steps: the first step is based on the extraction of mutation contexts required by CHORD to create a matrix with all data. The second step is based on the prediction of HRD probabilities based on the calculation of an HRD score. Initially, this approach was used to calculate the HRD score in ovarian and breast cancer samples. Later, it was also extended to the analysis of other tumors in which BRCA1 and BRCA2 alterations are involved, such as pancreatic and prostate cancer [155] (Table 2).

Table 2. HRD computational tools.

Tools	Applications	Variants Type	References
HRDetect	WGS	indels, snv, sv and CNV	[146]
Mutalisk	WGS, WES and TS	Mutational signatures	[147]
SigMA	WGS, WES and TS	SNV	[148]
CHORD	WGS	SNV, indels	[153,154]
PathAI	WGS, WES and TS	Indels, snv	https://www.pathai.com (accessed on 12 July 2022)
GSA	WGS, WES and TS	CNV	[156]
AcornHRD	WGS, WES and TS	Indels, CNV and snv	[157]

Other more sophisticated and robust methods based on artificial intelligence have been introduced to identify and investigate the mutational signatures associated with HRD starting from WGS, WES and TS sequencing data. **PathAI** (https://www.pathai.com/ accessed on 12 July 2022) employs machine learning models that predict the HRD status by studying how the disease evolves and making dynamic models of it using the mutational signatures. The **GSA** (genomic scar analysis) algorithm was developed and vali. dated to calculate the HRD score and the LOH score [156]. This approach is characterized by the presence of two submodules: tree recursion (TR) segmentation and filtering, and the estimation and correction of the tumor purity and ploidy. These elements are important for a better analysis of the HRD/LOH score. The input data formats of GSA are (i) BAF data and (ii) LRR data. BAF (B allele frequency) represents the median SNP genotype frequency of each capture region while LRR (Log R ratio) is the normalized depth ratio of the tumor and the normal sample (or blood cell control set) in each capture region after GC-bias correction. Currently, none of the above-described gene signatures have been widely adopted in clinical practice because they were identified based solely on a single dataset and did not take into consideration the heterogeneity of patient cohorts [158]. A recent approach under investigation, called **AcornHRD** (https://ascopubs.org/ accessed on 12 July 2022), enables the calculation of an HRD score associated with the efficacy of PARP inhibition and platinum-based chemotherapy in a variety of cancer types. The aim of this approach is to extend the range of patients that might benefit from targeted therapy. A current limitation of the above-described methods is the impossibility to capture tumor evolution processes, such as a restoration of HRR function in response to therapy-selective pressure. Therefore, it could be useful to incorporate functional biomarkers based on dynamic changes in DNA repair that occur throughout tumor evolution for the identification of HRD-positive tumors [157,159].

7. Discussion

The use of NGS technologies in clinical practice and clinical research is progressively increasing. The guidelines of the main scientific societies recommend the use of NGS in the diagnosis of numerous human cancers [160,161]. Furthermore, the availability of clinical studies for often rare and complex genomic alterations requires the use of large TS panels to facilitate the enrollment of patients in studies with new drugs. In this complex scenario, in which the therapeutic decision depends mainly on the genomic landscape of the tumor of each individual patient, the quality and accuracy of the NGS analysis are essential to guarantee the appropriateness of the treatments. Therefore, having a robust, reliable and

validated bioinformatics pipeline available is a necessary requirement to be able to analyze genomic data and provide useful results for the clinical decision.

The introduction first in research and then in clinical practice of complex genomic markers, such as MSI, TMB and HRD, has made the analysis of the sequencing data even more complex. As we have discussed for HRD, it is essential to identify bioinformatics tools capable of deriving these complex biomarkers also from TS data, in order to favor the implementation of these new biomarkers in the field of diagnostics and clinical research, with sustainable costs and times and methods of analysis compatible with clinical needs.

All studies so far have evaluated the correlation between genomic instability and response to platinum and/or PARPi. However, genomic instability could also represent an important marker of response to immunotherapy, as suggested by preliminary data [162]. Studies in this direction are certainly needed. It is clear that HRD plays a crucial role in cancer pathogenesis and progression. Hence, accurate estimation of HRD status is essential, not only to guide treatment decisions but also for the development of novel therapeutic strategies, with the ultimate objective of expanding the pool of patients who may derive clinical benefit from such approaches. Therefore, there is an urgent need to further develop reliable HRD detection methodologies that are comprehensive, cost-effective, and minimally invasive with a high predictive value for treatment response and disease progression [163].

8. Conclusions

In conclusion, the field of genomic biomarkers in oncology is constantly evolving and we expect that they will become increasingly important for precision oncology. The use of AI techniques not only for the interpretation of these data but also for their integration with the clinical and pathological characteristics of the patient represents a future challenge for cancer research.

Author Contributions: Conceptualization: N.N., L.G., F.P. and D.G.; methodology: S.D., R.E.A., C.R., L.G., F.P. and D.G.; investigation: S.D., C.R., L.G., F.P. and D.G.; writing—original draft preparation: S.D., R.E.A., C.R., L.G., F.P. and D.G.; writing—review and editing: S.D., R.E.A., C.R., L.G., F.P., D.G., B.T. and A.P.; supervision: N.N. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the ACC Research Fellowship of GerSom project. Moreover, this work is supported also by 5 × 1000 project "Implementazione di un percorso di oncologia di precisione per pazienti con neoplasie avanzate" and by Ricerca Corrente Pascale "Analisi mediante pannello multigenico per la predisposizione a sindromi ereditarie di pazienti con tumore gastrico".

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hussen, B.M.; Abdullah, S.T.; Salihi, A.; Sabir, D.K.; Sidiq, K.R.; Rasul, M.F.; Hidayat, H.J.; Ghafouri-Fard, S.; Taheri, M.; Jamali, E. The emerging roles of NGS in clinical oncology and personalized medicine. *Pathol. Res. Pract.* **2022**, *230*, 153760. [CrossRef] [PubMed]
2. Malone, E.R.; Oliva, M.; Sabatini, P.J.B.; Stockley, T.; Siu, L.L. Molecular profiling for precision cancer therapies. *Genome Med.* **2020**, *12*, 8. [CrossRef] [PubMed]
3. Mateo, J.; Steuten, L.; Aftimos, P.; André, F.; Davies, M.; Garralda, E.; Geissler, J.; Husereau, D.; Martinez-Lopez, I.; Normanno, N.; et al. Delivering precision oncology to patients with cancer. *Nat. Med.* **2022**, *28*, 658–665. [CrossRef]
4. Normanno, N.; Apostolidis, K.; de Lorenzo, F.; Beer, P.A.; Henderson, R.; Sullivan, R.; Biankin, A.V.; Horgan, D.; Lawler, M. Cancer Biomarkers in the era of precision oncology: Addressing the needs of patients and health systems. *Semin. Cancer Biol.* **2021**, *84*, 293–301. [CrossRef]
5. Hu, T.; Chitnis, N.; Monos, D.; Dinh, A. Next-generation sequencing technologies: An overview. *Hum. Immunol.* **2021**, *82*, 801–811. [CrossRef] [PubMed]

6. Gómez-López, G.; Dopazo, J.; Cigudosa, J.C.; Valencia, A.; Al-Shahrour, F. Precision medicine needs pioneering clinical bioinformaticians. *Brief. Bioinform.* **2017**, *20*, 752–766. [CrossRef] [PubMed]
7. Roy, S.; Coldren, C.; Karunamurthy, A.; Kip, N.S.; Klee, E.W.; Lincoln, S.E.; Leon, A.; Pullambhatla, M.; Temple-Smolkin, R.L.; Voelkerding, K.V.; et al. Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines: A Joint Recommendation of the Association for Molecular Pathology and the College of American Pathologists. *J. Mol. Diagn.* **2018**, *20*, 4–27. [CrossRef]
8. Singer, J.; Irmisch, A.; Ruscheweyh, H.-J.; Singer, F.; Toussaint, N.C.; Levesque, M.P.; Stekhoven, D.J.; Beerenwinkel, N. Bioinformatics for precision oncology. *Brief. Bioinform.* **2017**, *20*, 778–788. [CrossRef]
9. Gauthier, J.; Vincent, A.T.; Charette, S.J.; Derome, N. A brief history of bioinformatics. *Brief. Bioinform.* **2018**, *20*, 1981–1996. [CrossRef]
10. Wang, G.; Liu, Y.; Zhu, N.; Klau, G.W.; Feng, W. Bioinformatics Methods and Biological Interpretation for Next-Generation Sequencing Data. *BioMed Res. Int.* **2015**, *2015*, 690873. [CrossRef]
11. Nik-Zainal, S. Insights into cancer biology through next-generation sequencing. *Clin. Med.* **2014**, *14*, s71–s77. [CrossRef] [PubMed]
12. Nones, K.; Patch, A.-M. The Impact of Next Generation Sequencing in Cancer Research. *Cancers* **2020**, *12*, 2928. [CrossRef] [PubMed]
13. Maljkovic Berry, I.; Melendrez, M.C.; Bishop-Lilly, K.A.; Rutvisuttinunt, W.; Pollett, S.; Talundzic, E.; Morton, L.; Jarman, R.G. Next Generation Sequencing and Bioinformatics Methodologies for Infectious Disease Research and Public Health: Approaches, Applications, and Considerations for Development of Laboratory Capacity. *J. Infect. Dis.* **2019**, *221*, S292–S307. [CrossRef]
14. Slatko, B.E.; Gardner, A.F.; Ausubel, F.M. Overview of Next-Generation Sequencing Technologies. *Curr. Protoc. Mol. Biol.* **2018**, *122*, e59. [CrossRef] [PubMed]
15. Pereira, R.; Oliveira, J.; Sousa, M. Bioinformatics and Computational Tools for Next-Generation Sequencing Analysis in Clinical Genetics. *J. Clin. Med.* **2020**, *9*, 132. [CrossRef]
16. Pareek, C.S.; Smoczynski, R.; Tretyn, A. Sequencing technologies and genome sequencing. *J. Appl. Genet.* **2011**, *52*, 413–435. [CrossRef]
17. Beck, T.F.; Mullikin, J.C.; NISC Comparative Sequencing Program. Systematic Evaluation of Sanger Validation of Next-Generation Sequencing Variants. *Clin. Chem.* **2016**, *62*, 647–654. [CrossRef]
18. Bewicke-Copley, F.; Kumar, A.E.; Palladino, G.; Korfi, K.; Wang, J. Applications and analysis of targeted genomic sequencing in cancer studies. *Comput. Struct. Biotechnol. J.* **2019**, *17*, 1348–1359. [CrossRef]
19. Keshavan, A.; Poline, J.-B. From the Wet Lab to the Web Lab: A Paradigm Shift in Brain Imaging Research. *Front. Neuroinform.* **2019**, *13*, 3. [CrossRef]
20. Lightbody, G.; Haberland, V.; Browne, F.; Taggart, L.; Zheng, H.; Parkes, E.; Blayney, J.K. Review of applications of high-throughput sequencing in personalized medicine: Barriers and facilitators of future progress in research and clinical application. *Brief. Bioinform.* **2019**, *20*, 1795–1811. [CrossRef]
21. Mardis, E.R. Next-Generation Sequencing Platforms. *Annu. Rev. Anal. Chem.* **2013**, *6*, 287–303. [CrossRef] [PubMed]
22. Tucker, T.; Marra, M.; Friedman, J. Massively Parallel Sequencing: The Next Big Thing in Genetic Medicine. *Am. J. Hum. Genet.* **2009**, *85*, 142–154. [CrossRef] [PubMed]
23. Zhao, E.Y.; Jones, M.; Jones, S. Whole-Genome Sequencing in Cancer. *Cold Spring Harb. Perspect. Med.* **2018**, *9*, a034579. [CrossRef] [PubMed]
24. Nakagawa, H.; Fujita, M. Whole genome sequencing analysis for cancer genomics and precision medicine. *Cancer Sci.* **2018**, *109*, 513–522. [CrossRef] [PubMed]
25. Balloux, F.; Brynildsrud, O.B.; van Dorp, L.; Shaw, L.P.; Chen, H.; Harris, K.A.; Wang, H.; Eldholm, V. From Theory to Practice: Translating Whole-Genome Sequencing (WGS) into the Clinic. *Trends Microbiol.* **2018**, *26*, 1035–1048. [CrossRef]
26. Ghazani, A.A.; Oliver, N.M.; Pierre, J.P.S.; Garofalo, A.; Rainville, I.R.; Hiller, E.; Treacy, D.J.; Rojas-Rudilla, V.; Wood, S.; Bair, E.; et al. Assigning clinical meaning to somatic and germ-line whole-exome sequencing data in a prospective cancer precision medicine study. *Genet. Med.* **2017**, *19*, 787–795. [CrossRef]
27. Van Allen, E.M.; Wagle, N.; Stojanov, P.; Perrin, D.L.; Cibulskis, K.; Marlow, S.; Jane-Valbuena, J.; Friedrich, D.C.; Kryukov, G.; Carter, S.L.; et al. Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. *Nat. Med.* **2014**, *20*, 682–688. [CrossRef]
28. Ulintz, P.J.; Wu, W.; Gates, C.M. Bioinformatics Analysis of Whole Exome Sequencing Data. In *Methods in Molecular Biology*; Humana Press Inc.: Totowa, NJ, USA, 2019; Volume 1881, pp. 277–318. [CrossRef]
29. Schultzhaus, Z.; Wang, Z.; Stenger, D. CRISPR-based enrichment strategies for targeted sequencing. *Biotechnol. Adv.* **2020**, *46*, 107672. [CrossRef]
30. Roca, I.; González-Castro, L.; Fernandez-Lopez, H.; Couce, M.L.; Fernández-Marmiesse, A. Free-access copy-number variant detection tools for targeted next-generation sequencing data. *Mutat. Res. Mutat. Res.* **2019**, *779*, 114–125. [CrossRef]
31. Van Dijk, E.L.; Jaszczyszyn, Y.; Naquin, D.; Thermes, C. The Third Revolution in Sequencing Technology. *Trends Genet.* **2018**, *34*, 666–681. [CrossRef]
32. Liu, L.; Li, Y.; Li, S.; Hu, N.; He, Y.; Pong, R.; Lin, D.; Lu, L.; Law, M. Comparison of Next-Generation Sequencing Systems. *J. Biomed. Biotechnol.* **2012**, *2012*, 251364. [CrossRef] [PubMed]

33. Bruijns, B.; Tiggelaar, R.M.; Gardeniers, J. Massively parallel sequencing techniques for forensics: A review. *Electrophoresis* **2018**, *39*, 2642–2654. [CrossRef] [PubMed]
34. Pirooznia, M.; Kramer, M.; Parla, J.; Goes, F.S.; Potash, J.B.; McCombie, W.R.; Zandi, P.P. Validation and assessment of variant calling pipelines for next-generation sequencing. *Hum. Genom.* **2014**, *8*, 14. [CrossRef]
35. Santani, A.; Murrell, J.; Funke, B.; Yu, Z.; Hegde, M.; Mao, R.; Ferreira-Gonzalez, A.; Voelkerding, K.V.; Weck, K.E. Development and Validation of Targeted Next-Generation Sequencing Panels for Detection of Germline Variants in Inherited Diseases. *Arch. Pathol. Lab. Med.* **2017**, *141*, 787–797. [CrossRef] [PubMed]
36. Zheng, J.; Zhang, H.; Banerjee, S.; Li, Y.; Zhou, J.; Yang, Q.; Tan, X.; Han, P.; Fu, Q.; Cui, X.; et al. A comprehensive assessment of Next-Generation Sequencing variants validation using a secondary technology. *Mol. Genet. Genom. Med.* **2019**, *7*, e00748. [CrossRef]
37. Ilyas, M. Next-Generation Sequencing in Diagnostic Pathology. *Pathobiology* **2017**, *84*, 292–305. [CrossRef]
38. Rossing, M.; Sørensen, C.S.; Ejlertsen, B.; Nielsen, F.C. Whole genome sequencing of breast cancer. *APMIS* **2019**, *127*, 303–315. [CrossRef]
39. Garagnani, P.; Marquis, J.; Delledonne, M.; Pirazzini, C.; Marasco, E.; Kwiatkowska, K.M.; Iannuzzi, V.; Bacalini, M.G.; Valsesia, A.; Carayol, J.; et al. Whole-genome sequencing analysis of semi-supercentenarians. *eLife* **2021**, *10*, e57849. [CrossRef]
40. Yoshinaga, Y.; Daum, C.; He, G.; O'Malley, R. Genome Sequencing. In *Methods in Molecular Biology*; Humana Press Inc.: Totowa, NJ, USA, 2018; Volume 1775, pp. 37–52. [CrossRef]
41. Matthijs, G.; Souche, E.; Alders, M.; Corveleyn, A.; Eck, S.; Feenstra, I.; Race, V.; Sistermans, E.; Sturm, M.; Weiss, M.; et al. Guidelines for diagnostic next-generation sequencing. *Eur. J. Hum. Genet.* **2016**, *24*, 2–5. [CrossRef]
42. Aly, S.M.; Sabri, D.M. Next generation sequencing (NGS): A golden tool in forensic toolkit. *Arch. Forensic Med. Criminol.* **2015**, *4*, 260–271. [CrossRef]
43. Verma, M.; Kulshrestha, S.; Puri, A. Genome Sequencing. In *Methods in Molecular Biology*; Humana Press Inc.: Totowa, NJ, USA, 2017; Volume 1525, pp. 3–33. [CrossRef]
44. Horner, D.S.; Pavesi, G.; Castrignano, T.; De Meo, P.D.; Liuni, S.; Sammeth, M.; Picardi, E.; Pesole, G. Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. *Brief. Bioinform.* **2009**, *11*, 181–197. [CrossRef]
45. Wang, M.D. In the Spotlight: Bioinformatics. *IEEE Rev. Biomed. Eng.* **2012**, *6*, 3–8. [CrossRef]
46. Hwang, B.; Lee, J.H.; Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.* **2018**, *50*, 96. [CrossRef]
47. SoRelle, J.A.; Wachsmann, M.; Cantarel, B.L. Assembling and Validating Bioinformatic Pipelines for Next-Generation Sequencing Clinical Assays. *Arch. Pathol. Lab. Med.* **2020**, *144*, 1118–1130. [CrossRef]
48. Gullapalli, R.R. Evaluation of Commercial Next-Generation Sequencing Bioinformatics Software Solutions. *J. Mol. Diagn.* **2019**, *22*, 147–158. [CrossRef]
49. Schwarz, U.I.; Gulilat, M.; Kim, R.B. The Role of Next-Generation Sequencing in Pharmacogenetics and Pharmacogenomics. *Cold Spring Harb. Perspect. Med.* **2019**, *9*, a033027. [CrossRef]
50. Pedersen, B.S.; Quinlan, A.R. Mosdepth: Quick coverage calculation for genomes and exomes. *Bioinformatics* **2018**, *34*, 867–868. [CrossRef]
51. Van der Auwera, G.A. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*; O'Reilly Media: Sebastopol, CA, USA, 2020.
52. Van Der Auwera, G.A.; Carneiro, M.O.; Hartl, C.; Poplin, R.; Del Angel, G.; Levy-Moonshine, A.; Jordan, T.; Shakir, K.; Roazen, D.; Thibault, J.; et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Curr. Protoc. Bioinform.* **2013**, *43*, 11.10.1–11.10.33. [CrossRef]
53. Krusche, P.; Trigg, L.; Boutros, P.C.; Mason, C.E.; De La Vega, F.M.; Moore, B.L.; Gonzalez-Porta, M.; Eberle, M.A.; Tezak, Z.; Lababidi, S.; et al. Best practices for benchmarking germline small-variant calls in human genomes. *Nat. Biotechnol.* **2019**, *37*, 555–560. [CrossRef]
54. Koboldt, D.C. Best practices for variant calling in clinical sequencing. *Genome Med.* **2020**, *12*, 91. [CrossRef]
55. Danecek, P.; Bonfield, J.K.; Liddle, J.; Marshall, J.; Ohan, V.; Pollard, M.O.; Whitwham, A.; Keane, T.; McCarthy, S.A.; Davies, R.M.; et al. Twelve years of SAMtools and BCFtools. *Gigascience* **2021**, *10*, giab008. [CrossRef]
56. Lieven, C.; Lieven, C.; Beber, M.E.; Beber, M.E.; Olivier, B.G.; Olivier, B.G.; Bergmann, F.T.; Bergmann, F.T.; Ataman, M.; Ataman, M.; et al. MEMOTE for standardized genome-scale metabolic model testing. *Nat. Biotechnol.* **2020**, *38*, 272–276. [CrossRef]
57. Garrison, E.; Marth, G. Haplotype-Based Variant Detection from Short-Read Sequencing. Available online: https://arxiv.org/abs/1207.3907 (accessed on 1 January 2016).
58. Chen, X.; Schulz-Trieglaff, O.; Shaw, R.; Barnes, B.; Schlesinger, F.; Källberg, M.; Cox, A.J.; Kruglyak, S.; Saunders, C.T. Manta: Rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **2016**, *32*, 1220–1222. [CrossRef]
59. Kim, S.; Scheffler, K.; Halpern, A.L.; Bekritsky, M.A.; Noh, E.; Källberg, M.; Chen, X.; Kim, Y.; Beyter, D.; Krusche, P.; et al. Strelka2: Fast and accurate calling of germline and somatic variants. *Nat. Methods* **2018**, *15*, 591–594. [CrossRef]
60. Eisfeldt, J.; Vezzi, F.; Olason, P.; Nilsson, D.; Lindstrand, A. TIDDIT, an efficient and comprehensive structural variant caller for massive parallel sequencing data. *F1000Res* **2017**, *6*, 664. [CrossRef]

61. Talevich, E.; Shain, A.H.; Botton, T.; Bastian, B.C. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput. Biol.* **2015**, *12*, e1004873. [CrossRef]
62. Poplin, R.; Chang, P.-C.; Alexander, D.; Schwartz, S.; Colthurst, T.; Ku, A.; Newburger, D.; Dijamco, J.; Nguyen, N.; Afshar, P.T.; et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **2018**, *36*, 983–987. [CrossRef]
63. Poplin, R.; Ruano-Rubio, V.; DePristo, M.A.; Fennell, T.J.; Carneiro, M.O.; Van der Auwera, G.A.; Kling, D.E.; Gauthier, L.D.; Levy-Moonshine, A.; Roazen, D.; et al. Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv* **2017**. [CrossRef]
64. Van Loo, P.; Nordgard, S.H.; Lingjærde, O.C.; Russnes, H.G.; Rye, I.H.; Sun, W.; Weigman, V.J.; Marynen, P.; Zetterberg, A.; Naume, B.; et al. Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 16910–16915. [CrossRef]
65. Boeva, V.; Popova, T.; Bleakley, K.; Chiche, P.; Cappo, J.; Schleiermacher, G.; Janoueix-Lerosey, I.; Delattre, O.; Barillot, E. Control-FREEC: A tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics* **2012**, *28*, 423–425. [CrossRef]
66. Benjamin, D.; Sato, T.; Cibulskis, K.; Getz, G.; Stewart, C.; Lichtenstein, L. Calling Somatic SNVs and Indels with Mutect2. *BioRxiv* **2019**. [CrossRef]
67. Jia, P.; Yang, X.; Guo, L.; Liu, B.; Lin, J.; Liang, H.; Sun, J.; Zhang, C.; Ye, K. MSIsensor-pro: Fast, Accurate, and Matched-normal-sample-free Detection of Microsatellite Instability. *Genom. Proteom. Bioinform.* **2020**, *18*, 65–71. [CrossRef]
68. Garcia, M.; Juhos, S.; Larsson, M.; Olason, P.I.; Martin, M.; Eisfeldt, J.; DiLorenzo, S.; Sandgren, J.; De Ståhl, T.D.; Ewels, P.; et al. Sarek: A portable workflow for whole-genome sequencing analysis of germline and somatic variants. *F1000Research* **2020**, *9*, 63. [CrossRef]
69. McLaren, W.; Gil, L.; Hunt, S.E.; Riat, H.S.; Ritchie, G.R.S.; Thormann, A.; Flicek, P.; Cunningham, F. The Ensembl Variant Effect Predictor. *Genome Biol.* **2016**, *17*, 122. [CrossRef]
70. Kopanos, C.; Tsiolkas, V.; Kouris, A.; Chapple, C.E.; Aguilera, M.A.; Meyer, R.; Massouras, A. VarSome: The human genomic variant search engine. *Bioinformatics* **2019**, *35*, 1978–1980. [CrossRef]
71. Landrum, M.J.; Lee, J.M.; Benson, M.; Brown, G.R.; Chao, C.; Chitipiralla, S.; Gu, B.; Hart, J.; Hoffman, D.; Jang, W.; et al. ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **2017**, *46*, D1062–D1067. [CrossRef]
72. OncoKB: A Precision Oncology Knowledge Base. 2017. Available online: http://oncokb.org (accessed on 12 July 2022).
73. Gao, J.; Aksoy, B.A.; Dogrusoz, U.; Dresdner, G.; Gross, B.E.; Sumer, S.O.; Sun, Y.; Jacobsen, A.; Sinha, R.; Larsson, E.; et al. Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. *Sci. Signal.* **2013**, *6*, pl1. [CrossRef]
74. Zitnik, M.; Nguyen, F.; Wang, B.; Leskovec, J.; Goldenberg, A.; Hoffman, M.M. Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. *Inf. Fusion* **2018**, *50*, 71–91. [CrossRef]
75. McKenna, A.; Hanna, M.; Banks, E.; Sivachenko, A.; Cibulskis, K.; Kernytsky, A.; Garimella, K.; Altshuler, D.; Gabriel, S.; Daly, M.; et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **2010**, *20*, 1297–1303. [CrossRef]
76. Stoler, N.; Nekrutenko, A. Sequencing error profiles of Illumina sequencing instruments. *NAR Genom. Bioinform.* **2021**, *3*, lqab019. [CrossRef]
77. Zhou, W.; Kang, L.; Duan, H.; Qiao, S.; Tao, L.; Chen, Z.; Huang, Y. A virtual sequencer reveals the dephasing patterns in error-correction code DNA sequencing. *Natl. Sci. Rev.* **2020**, *8*, nwaa227. [CrossRef]
78. Mazlan, A.U.; Sahabudin, N.A.; Remli, M.A.; Ismail, N.S.N.; Mohamad, M.S.; Nies, H.W.; Warif, N.B.A. A Review on Recent Progress in Machine Learning and Deep Learning Methods for Cancer Classification on Gene Expression Data. *Processes* **2021**, *9*, 1466. [CrossRef]
79. Luo, R.; Sedlazeck, F.J.; Lam, T.-W.; Schatz, M.C. A multi-task convolutional deep neural network for variant calling in single molecule sequencing. *Nat. Commun.* **2019**, *10*, 998. [CrossRef]
80. Sahraeian, S.M.E.; Liu, R.; Lau, B.; Podesta, K.; Mohiyuddin, M.; Lam, H.Y.K. Deep convolutional neural networks for accurate somatic mutation detection. *Nat. Commun.* **2019**, *10*, 1041. [CrossRef]
81. Cai, L.; Wu, Y.; Gao, J. DeepSV: Accurate calling of genomic deletions from high-throughput sequencing data using deep convolutional neural network. *BMC Bioinform.* **2019**, *20*, 665. [CrossRef]
82. Friedman, S.; Gauthier, L.; Farjoun, Y.; Banks, E. Lean and deep models for more accurate filtering of SNP and INDEL variant calls. *Bioinformatics* **2019**, *36*, 2060–2067. [CrossRef]
83. Kohestani, H.; Giuliani, A. Organization principles of biological networks: An explorative study. *Biosystems* **2016**, *141*, 31–39. [CrossRef]
84. Norori, N.; Hu, Q.; Aellen, F.M.; Faraci, F.D.; Tzovara, A. Addressing bias in big data and AI for health care: A call for open science. *Gene Expr. Patterns* **2021**, *2*, 100347. [CrossRef]
85. Weissler, E.H.; Naumann, T.; Andersson, T.; Ranganath, R.; Elemento, O.; Luo, Y.; Freitag, D.F.; Benoit, J.; Hughes, M.C.; Khan, F.; et al. The role of machine learning in clinical research: Transforming the future of evidence generation. *Trials* **2021**, *22*, 537. [CrossRef]
86. Beaulieu-Jones, B.K.; Yuan, W.; Brat, G.A.; Beam, A.L.; Weber, G.; Ruffin, M.; Kohane, I.S. Machine learning for patient risk stratification: Standing on, or looking over, the shoulders of clinicians? *NPJ Digit. Med.* **2021**, *4*, 62. [CrossRef]

87. Bartha, Á.; Győrffy, B. Comprehensive Outline of Whole Exome Sequencing Data Analysis Tools Available in Clinical Oncology. *Cancers* **2019**, *11*, 1725. [CrossRef] [PubMed]
88. Zhou, J.; Troyanskaya, O.G. Predicting effects of noncoding variants with deep learning–based sequence model. *Nat. Methods* **2015**, *12*, 931–934. [CrossRef] [PubMed]
89. Kelley, D.R.; Snoek, J.; Rinn, J.L. Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* **2016**, *26*, 990–999. [CrossRef]
90. Mallard, C.; Johnston, M.J.; Bobyn, A.; Nikolic, A.; Argiropoulos, B.; Chan, J.A.; Guilcher, G.M.; Gallo, M. Hi-C detects genomic structural variants in peripheral blood of pediatric leukemia patients. *Mol. Case Stud.* **2022**, *8*, a006157. [CrossRef]
91. Shigaki, D.; Adato, O.; Adhikari, A.N.; Dong, S.; Hawkins-Hooker, A.; Inoue, F.; Juven-Gershon, T.; Kenlay, H.; Martin, B.; Patra, A.; et al. Integration of multiple epigenomic marks improves prediction of variant impact in saturation mutagenesis reporter assay. *Hum. Mutat.* **2019**, *40*, 1280–1291. [CrossRef]
92. Tan, J.; Doing, G.; Lewis, K.A.; Price, C.E.; Chen, K.M.; Cady, K.C.; Perchuk, B.; Laub, M.T.; Hogan, D.A.; Greene, C.S. Unsupervised Extraction of Stable Expression Signatures from Public Compendia with an Ensemble of Neural Networks. *Cell Syst.* **2017**, *5*, 63–71.e6. [CrossRef]
93. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M.; et al. Ten-sorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv* **2004**, arXiv:1603.04467.
94. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. 2019. Available online: http://arxiv.org/abs/1912.01703 (accessed on 12 July 2022).
95. Collobert, R.; Kavukcuoglu, K.; Farabet, C. Torch7: A Matlab-Like Environment for Machine Learning. Available online: http://numpy.scipy.org (accessed on 12 July 2022).
96. Chartrand, G.; Zhang, P. Introduction to Graphs. In *Chromatic Graph Theory*; Chapman and Hall/CRC: New York, NY, USA, 2019; pp. 27–52. [CrossRef]
97. Ghosh, S.; Mukherjee, S.; Sengupta, N.; Roy, A.; Dey, D.; Chakraborty, S.; Chattopadhyay, D.; Banerjee, A.; Basu, A. Network analysis reveals common host protein/s modulating pathogenesis of neurotropic viruses. *Sci. Rep.* **2016**, *6*, 32593. [CrossRef]
98. Huang, C.-H.; Zaenudin, E.; Tsai, J.J.; Kurubanjerdjit, N.; Dessie, E.Y.; Ng, K.-L. Dissecting molecular network structures using a network subgraph approach. *PeerJ* **2020**, *8*, e9556. [CrossRef]
99. Huang, C.-H.; Zaenudin, E.; Tsai, J.J.; Kurubanjerdjit, N.; Ng, K.-L. Network subgraph-based approach for analyzing and comparing molecular networks. *PeerJ* **2022**, *10*, e13137. [CrossRef]
100. Torshizi, A.D.; Petzold, L.R. Graph-based semi-supervised learning with genomic data integration using condition-responsive genes applied to phenotype classification. *J. Am. Med. Inform. Assoc.* **2017**, *25*, 99–108. [CrossRef] [PubMed]
101. Mentzelopoulos, A.; Karanasiou, I.; Papathanasiou, M.; Kelekis, N.; Kouloulias, V.; Matsopoulos, G.K. A Comparative Analysis of White Matter Structural Networks on SCLC Patients After Chemotherapy. *Brain Topogr.* **2022**, *35*, 352–362. [CrossRef] [PubMed]
102. Csardi, G. The Igraph Software Package for Complex Network Research. Available online: https://www.researchgate.net/publication/221995787 (accessed on 12 July 2022).
103. Mueller, L.A.J.; Kugler, K.G.; Dander, A.; Graber, A.; Dehmer, M. QuACN: An R package for analyzing complex biological networks quantitatively. *Bioinformatics* **2010**, *27*, 140–141. [CrossRef] [PubMed]
104. Handcock, M.S.; Hunter, D.R.; Butts, C.T.; Goodreau, S.M.; Morris, M. Analysis and Simulation of Network Data. Available online: http://CRAN.R-project.org/ (accessed on 12 July 2022).
105. Tripathi, S.; Dehmer, M.; Emmert-Streib, F. NetBioV: An R package for visualizing large network data in biology and medicine. *Bioinformatics* **2014**, *30*, 2834–2836. [CrossRef]
106. Mauri, A.; Consonni, V.; Pavan, M.; Todeschini, R. Dragon software: An easy approach to molecular descriptor calculations. *Match* **2006**, *56*, 237–248.
107. Bollen, K.A. *Structural Equations with Latent Variables*; Wiley: Hoboken, NJ, USA, 1989.
108. Dellino, G.I.; Palluzzi, F.; Chiariello, A.M.; Piccioni, R.; Bianco, S.; Furia, L.; De Conti, G.; Bouwman, B.A.M.; Melloni, G.; Guido, D.; et al. Release of paused RNA polymerase II at specific loci favors DNA double-strand-break formation and promotes cancer translocations. *Nat. Genet.* **2019**, *51*, 1011–1023. [CrossRef]
109. Saranya, A.; Venkatesan, S. A Model Based Approach on Gene Expression Profiling of Colorectal Cancer and Normal Mucosa Using Logistic Regression, Artificial Neural Network and Structural Equation Modelling. *Turk. J. Comput. Math. Educ.* **2021**, *12*, 2585–2593.
110. Pepe, D.; Do, J.H. Estimation of dysregulated pathway regions in MPP+ treated human neuroblastoma SH-EP cells with structural equation model. *BioChip J.* **2015**, *9*, 131–138. [CrossRef]
111. Mogaka, J.J.O.; Chimbari, M.J. The mediating effects of public genomic knowledge in precision medicine implementation: A structural equation model approach. *PLoS ONE* **2020**, *15*, e0240585. [CrossRef]
112. Rosseel, Y. Journal of Statistical Software lavaan: An R Package for Structural Equation Modeling. 2012. Available online: http://www.jstatsoft.org/ (accessed on 12 July 2022).
113. Palluzzi, F.; Grassi, M. SEMgraph: An R Package for Causal Network Analysis of High-Throughput Data with Structural Equation Models. 2021. Available online: http://arxiv.org/abs/2103.08332 (accessed on 12 July 2022).

114. Verhulst, B.; Maes, H.H.; Neale, M.C. GW-SEM: A Statistical Package to Conduct Genome-Wide Structural Equation Modeling. *Behav. Genet.* **2017**, *47*, 345–359. [CrossRef]
115. Roy, R.; Chun, J.; Powell, S.N. BRCA1 and BRCA2: Different roles in a common pathway of genome protection. *Nat. Rev. Cancer* **2011**, *12*, 68–78. [CrossRef] [PubMed]
116. Zheng, F.; Zhang, Y.; Chen, S.; Weng, X.; Rao, Y.; Fang, H. Mechanism and current progress of Poly ADP-ribose polymerase (PARP) inhibitors in the treatment of ovarian cancer. *Biomed. Pharmacother.* **2020**, *123*, 109661. [CrossRef] [PubMed]
117. Scott, L.J. Niraparib: First Global Approval. *Drugs* **2017**, *77*, 1029–1034. [CrossRef] [PubMed]
118. Ethier, J.-L.; Lheureux, S.; Oza, A.M. The role of niraparib for the treatment of ovarian cancer. *Future Oncol.* **2018**, *14*, 2565–2577. [CrossRef]
119. Ison, G.; Howie, L.J.; Amiri-Kordestani, L.; Zhang, L.; Tang, S.; Sridhara, R.; Pierre, V.; Charlab, R.; Ramamoorthy, A.; Song, P.; et al. FDA Approval Summary: Niraparib for the Maintenance Treatment of Patients with Recurrent Ovarian Cancer in Response to Platinum-Based Chemotherapy. *Clin. Cancer Res.* **2018**, *24*, 4066–4071. [CrossRef]
120. Kumagai, A.; Lee, J.; Yoo, H.Y.; Dunphy, W.G. TopBP1 Activates the ATR-ATRIP Complex. *Cell* **2006**, *124*, 943–955. [CrossRef]
121. Hoppe, M.M.; Sundar, R.; Tan, D.S.P.; Jeyasekharan, A.D. Biomarkers for Homologous Recombination Deficiency in Cancer. *JNCI J. Natl. Cancer Inst.* **2018**, *110*, 704–713. [CrossRef]
122. Kang, H.G.; Hwangbo, H.; Kim, M.J.; Kim, S.; Lee, E.J.; Park, M.J.; Kim, J.-W.; Kim, B.-G.; Cho, E.-H.; Chang, S.; et al. Aberrant Transcript Usage Is Associated with Homologous Recombination Deficiency and Predicts Therapeutic Response. *Cancer Res.* **2022**, *82*, 142–154. [CrossRef]
123. Takaya, H.; Nakai, H.; Takamatsu, S.; Mandai, M.; Matsumura, N. Homologous recombination deficiency status-based classification of high-grade serous ovarian carcinoma. *Sci. Rep.* **2020**, *10*, 2757. [CrossRef]
124. Foote, J.R.; Lopez-Acevedo, M.; Buchanan, A.H.; Secord, A.A.; Lee, P.S.; Fountain, C.; Myers, E.R.; Cohn, D.E.; Reed, S.D.; Havrilesky, L.J. Cost Comparison of Genetic Testing Strategies in Women with Epithelial Ovarian Cancer. *J. Oncol. Pract.* **2017**, *13*, e120–e129. [CrossRef]
125. McLaughlin, L.J.; Stojanovic, L.; Kogan, A.A.; Rutherford, J.L.; Choi, E.Y.; Yen, R.-W.C.; Xia, L.; Zou, Y.; Lapidus, R.G.; Baylin, S.B.; et al. Pharmacologic induction of innate immune signaling directly drives homologous recombination deficiency. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 17785–17795. [CrossRef] [PubMed]
126. Wagener-Ryczek, S.; Merkelbach-Bruse, S.; Siemanowski, J. Biomarkers for Homologous Recombination Deficiency in Cancer. *J. Pers. Med.* **2021**, *11*, 612. [CrossRef] [PubMed]
127. Shirts, B.H.; Casadei, S.; Jacobson, A.L.; Lee, M.K.; Gulsuner, S.; Bennett, R.L.; Miller, M.; Hall, S.A.; Hampel, H.; Hisama, F.M.; et al. Improving performance of multigene panels for genomic analysis of cancer predisposition. *Genet. Med.* **2016**, *18*, 974–981. [CrossRef] [PubMed]
128. Walsh, C.S. Two decades beyond BRCA1/2: Homologous recombination, hereditary cancer risk and a target for ovarian cancer therapy. *Gynecol. Oncol.* **2015**, *137*, 343–350. [CrossRef]
129. Kurian, A.W.; Kingham, K.E.; Ford, J.M. Next-generation sequencing for hereditary breast and gynecologic cancer risk assessment. *Curr. Opin. Obstet. Gynecol.* **2015**, *27*, 23–33. [CrossRef]
130. Telli, M.L.; Stover, D.G.; Loi, S.; Aparicio, S.; Carey, L.A.; Domchek, S.M.; Newman, L.; Sledge, G.W.; Winer, E.P. Homologous recombination deficiency and host anti-tumor immunity in triple-negative breast cancer. *Breast Cancer Res. Treat.* **2018**, *171*, 21–31. [CrossRef]
131. Abkevich, V.; Timms, K.M.; Hennessy, B.T.; Potter, J.; Carey, M.S.; Meyer, L.A.; Smith-McCune, K.; Broaddus, R.; Lu, K.H.; Chen, J.; et al. Patterns of genomic loss of heterozygosity predict homologous recombination repair defects in epithelial ovarian cancer. *Br. J. Cancer* **2012**, *107*, 1776–1782. [CrossRef]
132. Swisher, E.M.; Lin, K.K.; Oza, A.M.; Scott, C.L.; Giordano, H.; Sun, J.; Konecny, G.E.; Coleman, R.L.; Tinker, A.V.; O'Malley, D.M.; et al. Rucaparib in relapsed, platinum-sensitive high-grade ovarian carcinoma (ARIEL2 Part 1): An international, multicentre, open-label, phase 2 trial. *Lancet Oncol.* **2017**, *18*, 75–87. [CrossRef]
133. Marquard, A.M.; Eklund, A.C.; Joshi, T.; Krzystanek, M.; Favero, F.; Wang, Z.C.; Richardson, A.L.; Silver, D.P.; Szallasi, Z.; Birkbak, N.J. Pan-cancer analysis of genomic scar signatures associated with homologous recombination deficiency suggests novel indications for existing cancer drugs. *Biomark. Res.* **2015**, *3*, 9. [CrossRef]
134. De Luca, X.M.; Newell, F.; Kazakoff, S.H.; Hartel, G.; Reed, A.E.M.; Holmes, O.; Xu, Q.; Wood, S.; Leonard, C.; Pearson, J.V.; et al. Using whole-genome sequencing data to derive the homologous recombination deficiency scores. *NPJ Breast Cancer* **2020**, *6*, 33. [CrossRef]
135. Weigelt, B.; Comino-Méndez, I.; de Bruijn, I.; Tian, L.; Meisel, J.L.; García-Murillas, I.; Fribbens, C.; Cutts, R.; Martelotto, L.G.; Ng, C.K.; et al. Diverse *BRCA1* and *BRCA2* Reversion Mutations in Circulating Cell-Free DNA of Therapy-Resistant Breast or Ovarian Cancer. *Clin. Cancer Res.* **2017**, *23*, 6708–6720. [CrossRef] [PubMed]
136. Cruz, C.; Castroviejo-Bermejo, M.; Gutiérrez-Enríquez, S.; Llop-Guevara, A.; Ibrahim, Y.; Gris-Oliver, A.; Bonache, S.; Morancho, B.; Bruna, A.; Rueda, O.; et al. RAD51 foci as a functional biomarker of homologous recombination repair and PARP inhibitor resistance in germline BRCA-mutated breast cancer. *Ann. Oncol.* **2018**, *29*, 1203–1210. [CrossRef] [PubMed]
137. Tumiati, M.; Hietanen, S.; Hynninen, J.; Pietilä, E.; Färkkilä, A.; Kaipio, K.; Roering, P.; Huhtinen, K.; Alkodsi, A.; Li, Y.; et al. A Functional Homologous Recombination Assay Predicts Primary Chemotherapy Response and Long-Term Survival in Ovarian Cancer Patients. *Clin. Cancer Res.* **2018**, *24*, 4482–4493. [CrossRef] [PubMed]

138. Balmus, G.; Pilger, D.; Coates, J.; Demir, M.; Sczaniecka-Clift, M.; Barros, A.C.; Woods, M.; Fu, B.; Yang, F.; Chen, E.; et al. ATM orchestrates the DNA-damage response to counter toxic non-homologous end-joining at broken replication forks. *Nat. Commun.* **2019**, *10*, 87. [CrossRef]
139. Alexandrov, L.B.; Stratton, M.R. Mutational signatures: The patterns of somatic mutations hidden in cancer genomes. *Curr. Opin. Genet. Dev.* **2014**, *24*, 52–60. [CrossRef]
140. Nik-Zainal, S.; Davies, H.; Staaf, J.; Ramakrishna, M.; Glodzik, D.; Zou, X.; Martincorena, I.; Alexandrov, L.B.; Martin, S.; Wedge, D.C.; et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **2016**, *534*, 47–54. [CrossRef]
141. Polak, P.; Kim, J.; Braunstein, L.Z.; Karlic, R.; Haradhavala, N.J.; Tiao, G.; Rosebrock, D.; Livitz, D.; Kübler, K.; Mouw, K.W.; et al. A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. *Nat. Genet.* **2017**, *49*, 1476–1486. [CrossRef]
142. MacIntyre, G.; Goranova, T.E.; De Silva, D.; Ennis, D.; Piskorz, A.M.; Eldridge, M.; Sie, D.; Lewsley, L.-A.; Hanif, A.; Wilson, C.; et al. Copy-number signatures and mutational processes in ovarian carcinoma. *Nat. Genet.* **2018**, *50*, 1262–1270. [CrossRef]
143. Staaf, J.; Glodzik, D.; Bosch, A.; Vallon-Christersson, J.; Reuterswärd, C.; Häkkinen, J.; Degasperi, A.; Amarante, T.D.; Saal, L.H.; Hegardt, C.; et al. Whole-genome sequencing of triple-negative breast cancers in a population-based clinical study. *Nat. Med.* **2019**, *25*, 1526–1533. [CrossRef]
144. Sztupinszki, Z.; Diossy, M.; Borcsok, J.; Prosz, A.; Cornelius, N.; Kjeldsen, M.K.; Mirza, M.R.; Szallasi, Z. Comparative Assessment of Diagnostic Homologous Recombination Deficiency–Associated Mutational Signatures in Ovarian Cancer. *Clin. Cancer Res.* **2021**, *27*, 5681–5687. [CrossRef]
145. Golan, T.; O'Kane, G.M.; Denroche, R.E.; Raitses-Gurevich, M.; Grant, R.C.; Holter, S.; Wang, Y.; Zhang, A.; Jang, G.H.; Stossel, C.; et al. Genomic Features and Classification of Homologous Recombination Deficient Pancreatic Ductal Adenocarcinoma. *Gastroenterology* **2021**, *160*, 2119–2132.e9. [CrossRef] [PubMed]
146. Davies, H.; Glodzik, D.; Morganella, S.; Yates, L.R.; Staaf, J.; Zou, X.; Ramakrishna, M.; Martin, S.; Boyault, S.; Sieuwerts, A.M.; et al. HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat. Med.* **2017**, *23*, 517–525. [CrossRef] [PubMed]
147. Lee, J.; Lee, A.J.; Lee, J.-K.; Park, J.; Kwon, Y.; Park, S.; Chun, H.; Ju, Y.S.; Hong, D. Mutalisk: A web-based somatic MUTation AnaLyIS toolKit for genomic, transcriptional and epigenomic signatures. *Nucleic Acids Res.* **2018**, *46*, W102–W108. [CrossRef] [PubMed]
148. Ledermann, J.A.; Drew, Y.; Kristeleit, R.S. Homologous recombination deficiency and ovarian cancer. *Eur. J. Cancer* **2016**, *60*, 49–58. [CrossRef] [PubMed]
149. Valerie, K.; Povirk, L.F. Regulation and mechanisms of mammalian double-strand break repair. *Oncogene* **2003**, *22*, 5792–5812. [CrossRef]
150. Chopra, N.; Tovey, H.; Pearson, A.; Cutts, R.; Toms, C.; Proszek, P.; Hubank, M.; Dowsett, M.; Dodson, A.; Daley, F.; et al. Homologous recombination DNA repair deficiency and PARP inhibition activity in primary triple negative breast cancer. *Nat. Commun.* **2020**, *11*, 2662. [CrossRef]
151. Gulhan, D.C.; Lee, J.J.-K.; Melloni, G.E.M.; Cortés-Ciriano, I.; Park, P.J. Detecting the mutational signature of homologous recombination deficiency in clinical samples. *Nat. Genet.* **2019**, *51*, 912–919. [CrossRef]
152. Matondo, A.; Jo, Y.H.; Shahid, M.; Choi, T.G.; Nguyen, M.N.; Nguyen, N.N.Y.; Akter, S.; Kang, I.; Ha, J.; Maeng, C.H.; et al. The Prognostic 97 Chemoresponse Gene Signature in Ovarian Cancer. *Sci. Rep.* **2017**, *7*, 9689. [CrossRef]
153. Leibowitz, B.D.; Dougherty, B.V.; Bell, J.S.K.; Kapilivsky, J.; Michuda, J.; Sedgewick, A.J.; Munson, W.A.; Chandra, T.A.; Dry, J.R.; Beaubier, N.; et al. Validation of genomic and transcriptomic models of homologous recombination deficiency in a real-world pan-cancer cohort. *BMC Cancer* **2022**, *22*, 587. [CrossRef]
154. Nguyen, L.; Martens, J.W.M.; Van Hoeck, A.; Cuppen, E. Pan-cancer landscape of homologous recombination deficiency. *Nat. Commun.* **2020**, *11*, 5584. [CrossRef]
155. Li, Y.; Zhao, Z.; Ai, L.; Wang, Y.; Liu, K.; Chen, B.; Chen, T.; Zhuang, S.; Xu, H.; Zou, M.; et al. Discovering a qualitative transcriptional signature of homologous recombination defectiveness for prostate cancer. *iScience* **2021**, *24*, 103135. [CrossRef] [PubMed]
156. Chen, D.; Shao, M.; Meng, P.; Wang, C.; Li, Q.; Cai, Y.; Song, C.; Wang, X.; Shi, T. GSA: An independent development algorithm for calling copy number and detecting homologous recombination deficiency (HRD) from target capture sequencing. *BMC Bioinform.* **2021**, *22*, 562. [CrossRef] [PubMed]
157. Gonzalez Bosquet, J.; Newtson, A.M.; Chung, R.K.; Thiel, K.W.; Ginader, T.; Goodheart, M.J.; Leslie, K.K.; Smith, B.J. Prediction of chemo-response in serous ovarian cancer. *Mol. Cancer* **2016**, *15*, 66. [CrossRef] [PubMed]
158. Chao, A.; Lai, C.-H.; Wang, T.-H.; Jung, S.-M.; Lee, Y.-S.; Chang, W.-Y.; Yang, L.-Y.; Ku, F.-C.; Huang, H.-J.; Chao, A.-S.; et al. Genomic scar signatures associated with homologous recombination deficiency predict adverse clinical outcomes in patients with ovarian clear cell carcinoma. *Klin. Wochenschr.* **2018**, *96*, 527–536. [CrossRef]
159. Peng, G.; Lin, C.C.-J.; Mo, W.; Dai, H.; Park, Y.-Y.; Kim, S.M.; Peng, Y.; Mo, Q.; Siwko, S.; Hu, R.; et al. Genome-wide transcriptome profiling of homologous recombination DNA repair. *Nat. Commun.* **2014**, *5*, 3361. [CrossRef]

160. Mosele, F.; Remon, J.; Mateo, J.; Westphalen, C.; Barlesi, F.; Lolkema, M.; Normanno, N.; Scarpa, A.; Robson, M.; Meric-Bernstam, F.; et al. Recommendations for the use of next-generation sequencing (NGS) for patients with metastatic cancers: A report from the ESMO Precision Medicine Working Group. *Ann. Oncol.* **2020**, *31*, 1491–1505. [CrossRef]
161. Chakravarty, D.; Johnson, A.; Sklar, J.; Lindeman, N.I.; Moore, K.; Ganesan, S.; Lovly, C.M.; Perlmutter, J.; Gray, S.W.; Hwang, J.; et al. Somatic Genomic Testing in Patients with Metastatic or Advanced Cancer: ASCO Provisional Clinical Opinion. *J. Clin. Oncol.* **2022**, *40*, 1231–1258. [CrossRef]
162. Miller, R.; Leary, A.; Scott, C.; Serra, V.; Lord, C.; Bowtell, D.; Chang, D.; Garsed, D.; Jonkers, J.; Ledermann, J.; et al. ESMO recommendations on predictive biomarker testing for homologous recombination deficiency and PARP inhibitor benefit in ovarian cancer. *Ann. Oncol.* **2020**, *31*, 1606–1622. [CrossRef]
163. Liu, Y.L.; Selenica, P.; Zhou, Q.; Iasonos, A.; Callahan, M.; Feit, N.Z.; Boland, J.; Vazquez-Garcia, I.; Mandelker, D.; Zehir, A.; et al. BRCA Mutations, Homologous DNA Repair Deficiency, Tumor Mutational Burden, and Response to Immune Checkpoint Inhibition in Recurrent Ovarian Cancer. *JCO Precis. Oncol.* **2020**, *4*, 665–679. [CrossRef]