current issues in molecular biology

MDPI

Review

A Bioinformatics Toolkit for Next-Generation Sequencing in Clinical Oncology

Simon Cabello-Aguilar [1,2,*], Julie A. Vendrell [2] and Jérôme Solassol [2]

[1] Montpellier BioInformatics for Clinical Diagnosis (MOBIDIC), Molecular Medicine and Genomics Platform (PMMG), CHU Montpellier, 34295 Montpellier, France
[2] Laboratoire de Biologie des Tumeurs Solides, Département de Pathologie et Oncobiologie, CHU Montpellier, Université de Montpellier, 34295 Montpellier, France; j-vendrell@chu-montpellier.fr (J.A.V.); j-solassol@chu-montpellier.fr (J.S.)
* Correspondence: s-cabelloaguilar@chu-montpellier.fr

Abstract: Next-generation sequencing (NGS) has taken on major importance in clinical oncology practice. With the advent of targeted therapies capable of effectively targeting specific genomic alterations in cancer patients, the development of bioinformatics processes has become crucial. Thus, bioinformatics pipelines play an essential role not only in the detection and in identification of molecular alterations obtained from NGS data but also in the analysis and interpretation of variants, making it possible to transform raw sequencing data into meaningful and clinically useful information. In this review, we aim to examine the multiple steps of a bioinformatics pipeline as used in current clinical practice, and we also provide an updated list of the necessary bioinformatics tools. This resource is intended to assist researchers and clinicians in their genetic data analyses, improving the precision and efficiency of these processes in clinical research and patient care.

Keywords: bioinformatics; clinical oncology; targeted therapy; pipeline; NGS; SNV; CNV; MSI

check for updates

Citation: Cabello-Aguilar, S.; Vendrell, J.A.; Solassol, J. A Bioinformatics Toolkit for Next-Generation Sequencing in Clinical Oncology. *Curr. Issues Mol. Biol.* **2023**, *45*, 9737–9752. https://doi.org/10.3390/cimb45120608

Academic Editors: Quan Zou and Giulia Fiscon

Received: 6 November 2023
Revised: 28 November 2023
Accepted: 2 December 2023
Published: 4 December 2023

1. Introduction

Progress in next-generation sequencing (NGS), including an increase in its accessibility and cost effectiveness, has enabled comprehensive genetic testing in many cancer centers and transformed cancer treatment. In particular, NGS has permitted the advancement of precision oncology focused on identifying genetic changes in tumors that include single-nucleotide variants (SNVs), copy number variations (CNVs), small insertions and deletions (indels), structural variants (SVs), and microsatellite instability (MSI) [1,2]. Such valuable insights into the molecular characteristics of tumors provided by NGS have made it an essential tool for the diagnosis and treatment of cancer [3].

Robust and reliable bioinformatics pipelines able to organize, interpret, and accurately identify these molecular alterations from within sequencing datasets are crucial in the treatment decision-making process. The robustness ensures that the pipeline can handle variations in the data and produce consistent results, while the reproducibility ensures that the same results can be obtained when the pipeline is run multiple times. In addition, the comprehensive traceability and understanding of how the pipeline works ensure that others are able to reproduce the results. To this end, a well-designed and well-documented bioinformatics pipeline can provide reliable and accurate guidance for oncologists.

In this review, we focus on the role of bioinformatics in NGS-based precision oncology. Specifically, we explore the bioinformatics steps involved in this process, including the calling of genetic alterations, their annotation, and interpretation. To provide a practical example of how each step is implemented, we describe a typical bioinformatics pipeline and reporting workflow for targeted sequencing analysis of solid tumors.

Of note, we have specifically focused this review on the analysis of data from Illumina sequencing, given its widespread adoption in the scientific community. It is noteworthy that various sequencing platforms with unique strengths and applications are available. For instance, Oxford Nanopore Technologies offers long-read sequencing, providing valuable insights into structural variations. Pacific Biosciences (PacBio) is recognized for its ability to generate long reads, facilitating the resolution of complex genomic regions. A thorough understanding of the strengths and limitations of different platforms is essential for making informed choices when implementing a NGS bioinformatic pipeline in clinical oncology. While Illumina is extensively utilized, readers are encouraged to assess their specific needs and explore alternative platforms that may better align with their objectives.

2. Workflow Management

In clinical oncology, the rapid evolution of high-throughput sequencing technologies has increased data generation, necessitating robust and efficient bioinformatic pipelines for analysis. Command-line tools [4,5] offer a flexible and efficient means to handle these data. These tools enable bioinformaticians to construct intricate pipelines that encompass various stages of analysis. The command-line interface, with its text-based interaction, allows for precise control over parameters, facilitating the customization and optimization of workflows to suit the specific requirements of clinical oncology research. However, command-line tools rely solely on text-based interfaces, requiring users to input commands in a terminal or console, while workflow management tools commonly provide users with a graphical or text-based interface to design workflows, offering a more visually intuitive experience. Workflow management tools [6] also ensure the automation and standardization of the bioinformatics process and allow the user to define the order, parameters, and input data for a sequence of tools. They directly take care of the correct execution and documentation of the intermediate steps. Several workflow managers are available, including Snakemake and Nextflow, among others [7–11]. Such systems help bioinformaticians save time, reduce errors, and ensure the accuracy and reliability of their analyses. In cancer genomics, a bioinformatics pipeline is executed by the workflow manager such as that described in Figure 1 and comprises different steps: (i) quality control, (ii) adapter trimming, (iii) alignment, (iv) variant calling, (v) variant annotation, (vi) variant filtering, (vii) CNV calling, (viii) MSI status calling, and (ix) interface generation.

An up-to-date compilation of available tools for each step of the pipeline is provided in Table 1. It is important to mention that the Broad Institute provides a Genome Analysis Toolkit (GATK) [12], which contains a wide variety of tools designed for variant discovery and genotyping that covers the steps described in Figure 1. Moreover, the nf-core community project [13] has assembled a curated collection of analysis pipelines constructed with Nextflow including a somatic variant calling workflow, SAREK [14,15], available at "https://nf-co.re/sarek/3.4.0 (accessed on 1 December 2023)". Nf-core offers portable and reproducible analysis pipelines and the support of an active community.

Galaxy [16] and Taverna [17] are both noteworthy platforms in the field of bioinformatics analysis. Galaxy, as an open-source platform, offers a web-based interface for analyzing high-throughput genomics data, especially NGS data. It accommodates users with varying levels of bioinformatics expertise, allowing them to create, execute, and share workflows for diverse bioinformatics analyses. Featuring a user-friendly graphical interface, Galaxy is accessible to a broad audience, providing tools and workflows for tasks such as sequence alignment, variant calling, and various genomic analyses. The platform emphasizes reproducibility, enabling users to systematically save and share their analyses. Taverna serves as a distinct workflow management system designed for various scientific applications, including bioinformatics. It facilitates the design and execution of workflows, providing a flexible environment for scientific analysis and automation. Additionally, Tavaxy [18] shortens the workflow development cycle by incorporating workflow patterns to streamline the creation process. It facilitates the reuse and integration of existing (sub-) workflows from Taverna and Galaxy, while also supporting the creation of hybrid workflows.

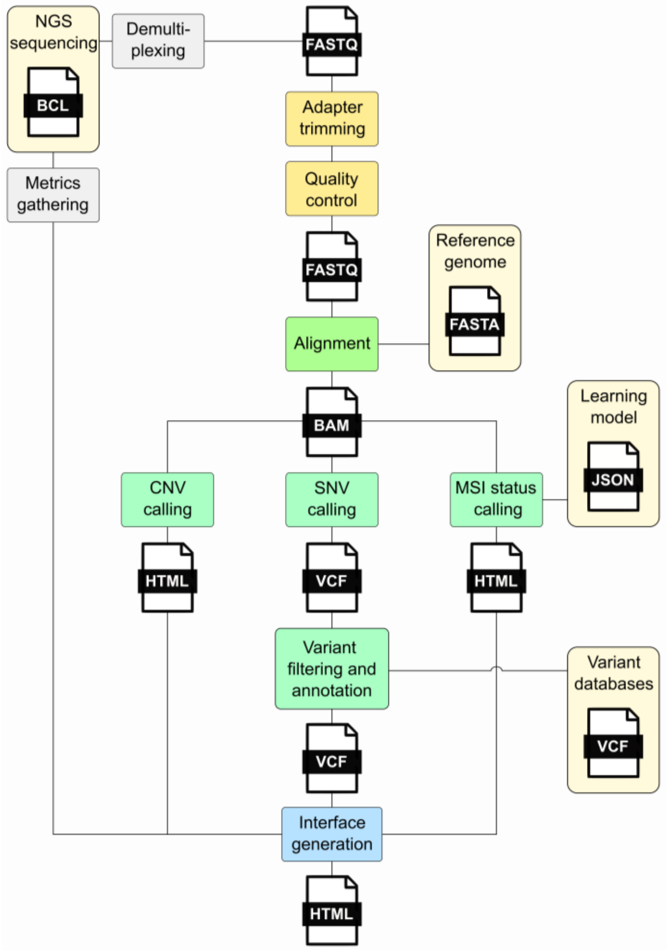


Figure 1. Major steps of an NGS bioinformatics pipeline. This diagram illustrates the processes forming the pipeline and the files generated during its execution. The gray segments denote processes that exist independently of the pipeline. Light yellow signifies external prerequisites, while yellow represents the initial pipeline stages involving FastQ processing. The alignment stage is highlighted in green, while light green indicates the analyses conducted, encompassing SNV, CNV, and MSI status calling. The final step, interface generation, is illustrated in blue. Acronyms: FASTQ—a text-based file storing nucleotide sequences and corresponding quality scores; BAM—Binary Alignment Map; VCF—Variant Call Format; CNV—Copy Number Variation; SNV—Single-Nucleotide Variant; MSI—Micro Satellite Instability.

Noteworthy, private solutions also exist. For example, the DRAGEN secondary analysis pipeline ensures all the steps from sequencing files to annotated and filtered genetic alterations. It was recently benchmarked, and the authors claim its value in a preprint that came out this year [19].

Table 1. List of commonly used bioinformatic tools.

Process	Tools	References
Workflow managers	Nextflow, Snakemake	[7,8]
Quality control	fastp, FastQC *, Picard, MultiQC	[20–23]
Adapter trimming	fastp, trimmomatic, cutadapt *, BBDuk	[20,24–26]
Reads alignment	BWA *, Bowtie, HISAT2, STAR	[27–30]
Variant calling	HaplotypeCaller, freebayes, mutect2, verdict *	[31–34]
Variant filtering	dbSNP, 1000G, GnomAD *	[35–37]
Variant annotation	VEP *, MobiDetails, ANNOVAR, SnpEff	[38–41]

Table 1. *Cont.*

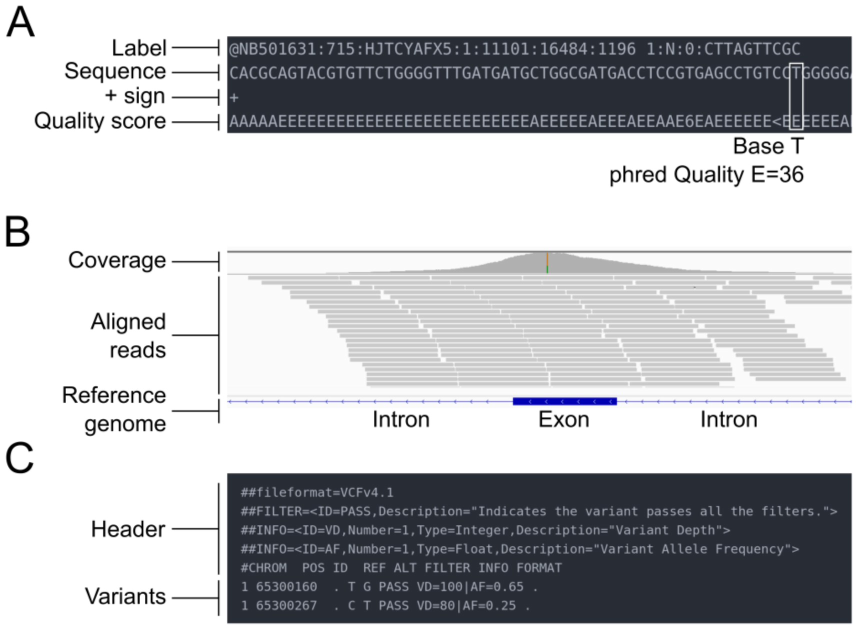
Process	Tools	References
CNV calling	CNV-LOF, CoverageMaster, CNV-RF, DeepCNV, CNV_IFTV, HBOS-CNV, CNV-MEANN, ControlFREEC, ifCNV *, mcna	[42–51]
MSI status calling	MIAmS *, MSIsensor, MSIdetect, deltaMSI	[52–55]

* Used in our in-house bioinformatics pipeline.

3. FastQ Processing

3.1. Quality Control

NGS sequencing produces binary base call sequence files (BCL) that are demultiplexed into FASTQ format sequencing files for each sample. The FASTQ format is a text-based format designed to store nucleotide sequences, along with their corresponding quality scores (Figure 2A). The initial stage of all bioinformatics pipelines is to assess the quality of the data. Indeed, sequence quality control is an essential step in the analysis of NGS data, which are generated in large volumes and can be prone to various types of errors, such as sequencing errors, adapter contamination, and sample cross-contamination. Sequence quality control aims to ensure that the sequencing data are accurate, reliable, and free from technical artifacts that could affect downstream analysis. It aims to identify low-quality bases, sequence bias, and over-representation of certain sequences. Quality assessment can be performed using tools such as fastp [20] or FastQC [21], a flexible and widely used tool for quality control, developed at the Babraham Institute to assess the quality of sequencing data in fastq files. This tool is robust, can be used on all operating systems, and offers both a graphical user interface and a command line interface. It is commonly incorporated by bioinformaticians as a quality control step in customized pipelines. The latest versions of FastQC include Picard [22], a tool developed by the Broad Institute that can manage SAM, BAM, and VCF files and perform quality control at different stages of the bioinformatics pipeline. An example of good and bad sequence quality profiles (i.e., the mean quality value across each base position in the read) obtained using FastQC is provided in Figure 3A. Moreover, MultiQC [23] consolidates data from various QC tools to create a cohesive report, complete with interactive plots, spanning multiple samples.



```
@NB501631:715:HJTCYAFX5:1:11101:16484:1196 1:N:0:CTTAGTTCGC
CACGCAGTACGTGTTCTGGGGTTTGATGATGCTGGCGATGACCTCCGTGAGCCTGTCCTGGGGG
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEAEEEEEEAEEEAEEAAE6EAEEEEEEE<EEEEEEA
```

```
##fileformat=VCFv4.1
##FILTER=<ID=PASS,Description="Indicates the variant passes all the filters.">
##INFO=<ID=VD,Number=1,Type=Integer,Description="Variant Depth">
##INFO=<ID=AF,Number=1,Type=Float,Description="Variant Allele Frequency">
#CHROM  POS ID  REF ALT FILTER INFO FORMAT
1 65300160  . T G PASS VD=100|AF=0.65 .
1 65300267  . C T PASS VD=80|AF=0.25 .
```

Figure 2. Overview of the different file types mentioned in the pipeline. (**A**) FASTQ file. (**B**) SAM/BAM file. (**C**) VCF file.

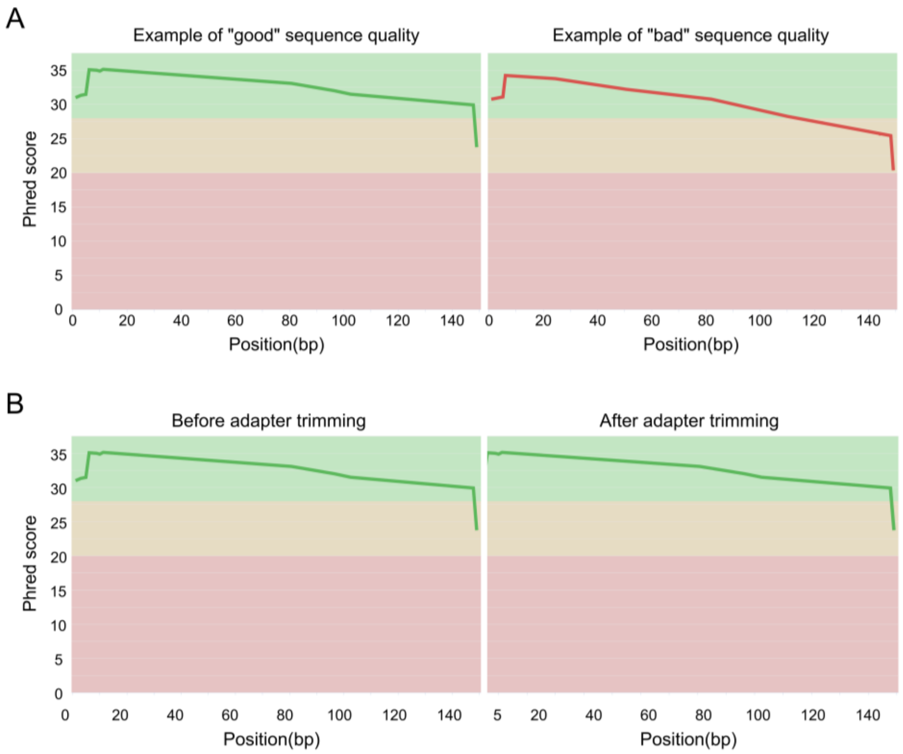


Figure 3. FastQC mean quality scores. (**A**) Examples of "good" and "bad" sequence quality. (**B**) Overview of the adapter trimming impact.

3.2. Adapter Trimming

Another preprocessing step is the adapter trimming, which involves removing adapter sequences, low-quality reads, and contaminating sequences from the raw sequencing data. The most widely used tools for data preprocessing are fastp [20], Trimmomatic [24], Cutadapt [25], and BBDuk [26]. In Figure 3B, we present quality profiles obtained using FastQC, illustrating the impact of adapter trimming with Cutadapt.

4. Alignment of the Nucleotide Sequence on a Reference Genome

After adapter trimming, the next step is to align the reads to a reference genome. The Genome Reference Consortium introduced the latest human reference genome, GRCh38 [56], in 2017, followed by subsequent improvements, the latest being GRCh38.p14 in March 2022, which remarkably reduced the number of gaps in the assembly to 349 compared to the initial version's approximately 150,000 gaps. Notably, these gaps were predominantly found in regions like telomeres, centromeres, and long repetitive sequences. Last year, the Telomere-to-Telomere (T2T) Consortium presented the first fully assembled reference genome [57], T2T-CHM13, eliminating all gaps.

The alignment step is performed by read mapper software, which assigns a location on the reference genome to each read based on its sequence. Since the reads do not contain information about their location in the genome, the mapper infers this information by comparing the read sequence to the reference genome. Essentially, it checks which parts of the reference genome match the sequences in the reads, determining where these reads originated in the genome. However, this seemingly straightforward task is computationally intensive and time-consuming because the software must meticulously compare each read to the entire reference genome and assign a precise position for it. The computational demand arises from the need for high accuracy and reliability in determining the origin of each read, a fundamental step in understanding the genetic information contained within the sequenced sample. There are many different read mappers available, each with its own strengths and weaknesses. Common examples include BWA [27] for genome and

Bowtie2 [28] for transcriptome. These tools employ a Burrows–Wheeler transform, a computational method invented by Michael Burrows and David Wheeler in 1994. This method involves rearranging character strings into sequences of similar characters, which offers significant computational benefits. Indeed, strings with repeated characters are easily compressible using techniques like move-to-front transform and run-length encoding. Various aligners employ distinct strategies; for instance, HISAT2 [29] is a graph-based genome alignment tool. The utilization of a graph-based approach allows leveraging theoretical advancements in computer science, resulting in a rapid and memory-efficient search algorithm. In transcriptome alignment, STAR [30] is also widely employed, using the Maximal Exact (Unique) Match concept for seed searching, it proves particularly advantageous for aligning long reads (>200 bp), such as those generated by third-generation sequencing.

The results of the read mapping step are usually provided in SAM format files, which can be converted to BAM format for more efficient storage and processing. SAM/BAM files can be accessed through the Integrative Genomics Viewer (IGV), allowing visualization of the reads (Figure 2B). The BAM files undergo different modifications during the alignment post-processing step, which includes tasks such as sorting, marking duplicate reads, and recalibrating base quality scores. The goal of these post-processing steps is to improve the accuracy and reliability of the final variant calls.

After the read mapping step, the resulting SAM/BAM files are sorted according to their genomic coordinates. This sorting is important because downstream analysis often relies on the order of the aligned reads. PCR duplicates are then commonly removed using tools such as Picard [22,58] or SAMtools [5]. PCR duplicates are identical copies of the same genomic fragment and can be introduced during sample preparation and PCR amplification steps. They can bias the analysis and lead to overrepresentation of certain regions of the genome. However, it is important to note that duplicated reads can also be biological copies originating from the same genomic location of chromosomes of different cells. For deep-coverage targeted sequencing approaches the probability of a duplicate read to be a biological copy increases with coverage, and therefore, the removal of duplicates is typically not performed in these cases.

5. Genetic Alterations Detection

5.1. SNV Calling

Variant calling is the critical step in identifying DNA alterations such as SNV or indels. This process involves comparing the DNA sequence of a sample (e.g., tumor tissue) to a reference genome or another sample from the same individual (e.g., normal tissue or blood). By detecting differences between the two sequences, variants can be identified. This is also a computationally intensive and time-consuming step, as the algorithms must compare each base to the reference. To perform this analysis, specialized software tools called variant callers are utilized. Called variants are usually stored in Variant Call Format (VCF) files. They consist of a header with various metadata, along with eight mandatory data columns, each row corresponding to a unique variant (Figure 2C).

Numerous variant callers are available, consolidating various statistical methods for variant detection. Noteworthy among them are GATK's variant callers, HaplotypeCaller [31] and UnifiedGenotyper [59]. It is worth mentioning that with the transition from GATK3 to GATK4, UnifiedGenotyper was discontinued as HaplotypeCaller demonstrated superior performance, outperforming it across various metrics [60]. Also, among widely used variant callers for somatic variant calling are FreeBayes [32], mutect2 [33], and VarDict [34]. Those variant callers were benchmarked using synthetic datasets [61] and differences in true positives were minor, but the number of false positives could vary significantly. FreeBayes and VarDict exhibited notably higher false positives, despite VarDict also having the highest number of true positives. A joint approach, combining several variant callers outperforms individual tools, showing increased specificity, balanced accuracy, and fewer false positives [62,63]. However, it is worth noting that each variant caller generates a distinct VCF file with its unique nomenclature. To combine outcomes from multiple variant callers on the same sample, custom-made scripts are necessary. However, the appropriate choice of variant caller depends on the data type and the biological problems addressed. For further information regarding somatic variant calling algorithms, interested readers may consult the latest reviews [64,65].

5.2. Variant Filtering

In the context of somatic variant calling, germline variants and polymorphisms, must be filtered. To that end, the variants found in the tumor sample are compared to a database of known germline variants, such as dbSNP [35], 1000 Genomes Project [36] or GnomAD [37]. Any variants present in this database are likely to be germline variants and are filtered out. The remaining variants are considered potential somatic variants and undergo further analysis. This approach is not as reliable for rare variants or variants in poorly annotated regions of the genome. Furthermore, these algorithmic solutions for identifying somatic mutations have limitations, especially given the Eurocentric bias of many population-based allele frequency databases. Accuracy may be diminished for underrepresented minorities, where allele frequency data are more limited.

Another approach consists in using a normal control sample, involving the sequencing of DNA from both the tumor sample and a sample of normal tissue from the same patient, such as blood or normal tissue adjacent to the tumor. The variants identified in the normal sample are then compared to the variants identified in the tumor sample. Variants that are present in the tumor sample but not in the normal sample are considered potential somatic variants. This approach has higher specificity, but it requires sequencing of both tumor and normal samples, which increases the cost and complexity of the analysis.

5.3. Variant Annotation

Variant annotation is the process of compiling pertinent information to make informed decisions about a given variant, while minimizing the amount of manual parsing required. This includes basic annotations such as the affected gene, whether it is in a coding or noncoding region, and whether it is synonymous or nonsynonymous. This step can be conducted by various software including VEP [38], AnnoVar [40], or SnpEff [41], for example. Additionally, more complex annotations such as clinical significance can also be included. The clinical significance of a variant holds particular importance for clinicians as it can aid in the decision making regarding patient care, including treatment options and risk assessment. The classification of variants is generally based on their association with specific diseases or phenotypes and includes categories such as pathogenic, likely pathogenic, of unknown significance, likely benign, or benign. However, the classification of variants may differ among various databases and tools, which can result in difficulties when interpreting and comparing results obtained from different sources of information. For instance, ClinVar [66], a freely accessible and public archive of reports links particular variants to known functional or clinical features, or the *TP53* Database that compiles *TP53* variant data reported in the published literature since 1989 [67]. Similarly, the database offered by the ENIGMA consortium provides annotations for *BRCA1/2* and *CHEK2* [68]. In contrast, other tools, such as SIFT [69] or Polyphen [70], categorize variants based on their in silico predicted impact on protein function. Recently, Chen et al. introduced AlphaMissense [71], an adaptation of AlphaFold [72], a neural network-based model, specifically designed for predicting missense variant pathogenicity. AlphaMissense demonstrated superior performance with an area under the receiver operator curve (auROC) of 0.940, evaluated on 18,924 ClinVar test variants. It outperformed models that were not trained directly on ClinVar and even surpassed models trained directly on ClinVar data. The emergence of these tools highlights the evolving landscape of the field. Consequently, it is crucial to meticulously evaluate the sources of annotation data employed in variant interpretation. Recently, an aggregator called MobiDetails [39] was developed to provide comprehensive and up-to-date variant annotation. It displays the most pertinent annotation databases and in silico effect predictors in a single web page.

Online databases such as DGIdb [73], OncoKB [74], and CIViC [75] are commonly utilized for querying drug–gene interactions. These databases also function as robust resources for extracting insights into the potential diagnostic implications and prognostic value of identified variants. Such information can be particularly beneficial for physicians, enabling them to adapt therapeutics and optimize patient care. In addition to direct interactions, it would also be advantageous to annotate genes with indirectly interacting drugs, i.e., drugs that target proteins upstream or downstream of the gene within the relevant pathway. Of note, customized in-house databases can be utilized for variant annotation. For instance, annotating a variant if it has been previously observed in another patient or sequencing experiment can provide valuable insights.

5.4. CNV Calling

In clinical oncology, CNV as biomarkers can help predict how a patient will respond to specific therapies. For instance, several targeted therapies are FDA-approved for the treatment of breast cancer patients with ERBB2 amplification [76,77], while MET amplification in non-small-cell lung carcinomas is a known resistance mechanism to tyrosine kinase inhibitors [78,79]. As a result, incorporating CNVs into a laboratory pipeline is critical for improving patient outcomes. There exist three primary methods for identifying CNV from NGS data: read-pair (RP), split-read (SR), and read-depth (RD).

RP methods such as BreakDancer [80], compare the average insert size of sequenced read-pairs to an expected size based on a reference genome. Variations from the predetermined average insert size are used to detect gain or loss of genomic materials. Shorter or longer insert sizes compared to the expected size correspond to the loss or gain of materials, respectively. SR methods evaluate CNV using paired reads where only one read of the pair has a reliable mapping quality while the other one partially fails to map to the reference sequence. These discrepancies within a read pair can potentially provide the precise position of insertion/deletion events. Tools implementing SR strategies (e.g., SVseq2, Gustaf, PRISM [81–83]) enable the detection of these breakpoints but are limited to short insertions or deletions. The RD approach consists in counting the aligned reads overlapping a genomic region and comparing the read counts between the sample of interest and a reference to determine CNV. A local decrease or increase in sequencing depth will correlate to loss or gain/amplification of loci, respectively.

Numerous tools for CNV detection are available, employing diverse algorithms including artificial intelligence, intricate statistical modeling, and more [42–49]. Recent RD-based methods such as ifCNV [50] and mCNA [51] have demonstrated remarkable performance in clinical practice, offering both fast computational times and high sensitivity and specificity.

5.5. MSI Status Calling

MSI is a biomarker of DNA mismatch repair deficiency commonly observed in cancer [84]. Accurate determination of MSI status is important for prognostic and therapeutic purposes. For instance, MSI status can predict the response to immunotherapy in colorectal cancer [85]. Traditional methods for analyzing microsatellite status involve length distribution analysis of multiplex-PCR generated DNA fragments from tumor samples, which can be labor-intensive and time-consuming [86]. NGS technology offers an alternative method for MSI determination. NGS-based applications such as MIAmS [52], MSISensor [53], deltaMSI [54] or more recently the solution published by Sophia genetics, MSIdetect [55], can determine MSI status. It requires specific spiking of microsatellite loci in the targeted panel. This approach offers several advantages over traditional methods, including high accuracy and higher efficiency. MIAmS is a scalable application that does not require paired normal tissue for comparison and generates a user-friendly report for interpretation. The use of NGS-based applications for MSI determination is increasingly being adopted in clinical practice due to their improved performance and convenience.

5.6. Implementation of a Pipeline

Typically, developing a robust NGS analysis pipeline in clinical oncology demands a rigorous scientific approach. It is imperative for medical oncologists to clearly convey their requirements to both biologists and bioinformaticians who can propose effective solutions. It is important to note that any pipeline needs to be adjusted based on specific experimental conditions. Moreover, adapting the pipeline to the computing architecture is crucial for optimal performance. Additionally, specific variant filtering and annotation criteria can be established by the bioinformatician in collaboration with the medical oncologists, tailored to the biological problem being addressed.

For illustration purposes, we provided a list of tools used in our bioinformatics pipeline, and we expect it may aid those faced with numerous options (Table 1). The selection of tools was guided by subjective considerations including the ease of implementation, the utilization in other pipelines for computing harmonization and inter-pipeline compatibility, and a proven track record in efficiently handling large volumes of clinical samples. All the tools mentioned in this review are regularly maintained and kept up to date. It is essential for individuals considering the implementation of a pipeline in their laboratory to consult the documentation of each tool, as each tool has its unique strengths and weaknesses. In recent years, best practices for the implementation of a bioinformatic pipeline have been published [87]. Physicians and bioinformaticians seeking to implement a new pipeline should familiarize themselves with this literature.

6. Future Developments

Moving forward, further developments in bioinformatics are crucial for the advancement of clinical oncology. These ongoing efforts aim to address emerging challenges, refine existing methodologies, and improve the effectiveness of precision medicine in cancer care. The tools discussed herein offer a snapshot of the current state of the field but are designed to evolve. Bioinformaticians, staying abreast of the constantly changing technologies and tools, play a central role in the realm of precision oncology.

The application of deep learning methods in the field has only just begun, with AlphaMisense serving as an illustrative example of how this technological gap is starting to revolutionize various aspects of data analysis, including bioinformatics. The next phase of developments will likely involve the application of advanced AI algorithms to aligners and variant callers. While reference genomes are evolving, aligners have remained unchanged for several years and are due for an update. Additionally, DeepVariant [88], a deep learning-based variant caller currently not applied to somatic variant calling, is expected to be adapted to this specific case in the coming years.

Moreover, while tumor mutational burden (TMB), representing the total count of DNA mutations detected in cancer cells and an important biomarker for immunotherapy [89–91], traditionally relied on whole genome sequencing or whole exome sequencing, it can now be estimated through targeted sequencing of a focused gene panel [92]. However, a recent study by Fang et al. [93] revealed that panels focusing on cancer genes tend to overestimate TMB in comparison to whole exome sequencing. This overestimation is mainly due to the positive selection for mutations in cancer genes. While the complete resolution of this issue remains elusive through the removal of mutational hotspots alone, a meticulous calibration process can enable a truthful TMB calculation within a clinical context. Its seamless integration into somatic pipelines is anticipated in the near future.

Of particular significance is also the development of a user-friendly interface essential to ensure accessibility and effective analysis by physicians of the outcomes yielded by the delineated pipeline, including the genotyping results, the sequencing quality metrics, and the run quality metrics. To our knowledge, no reports of a tool offering this type of interface have been published, and additional work seems necessary to create one. The genotyping results from the various analyses are aggregated to provide a comprehensive overview of the patients' genotype. This aggregation facilitates precision medicine approaches by offering a holistic understanding of individual genetic profiles. Additionally, the sequencing run metrics furnish insights into diverse aspects of the sequencing process, encompassing the quantity of generated reads, read length, read quality, and the coverage level. They thus offer the opportunity to evaluate the performance of the sequencing apparatus and the caliber of the generated sequencing data. Through careful examination of these metrics, both bioinformaticians and physicians can detect potential issues that might affect data quality. Subsequently, this information can be exploited to optimize sequencing conditions, potentially conduct a re-run if warranted, or adapt downstream analysis methodologies to account for identified issues. While such reports thus play essential roles in important patient management decisions, they are often overlooked. Such an interface would need to meet the specific needs of laboratory-based physicians analyzing several thousand samples annually. Its development would thus require the close collaboration between bioinformaticians and physicians.

7. Conclusions

Access to dependable bioinformatics pipelines is imperative in precision oncology. They facilitate the accurate identification and interpretation of genomic alterations on which treatment decisions are based (Table 2). However, bioinformatics pipelines often entail computationally intensive steps, often requiring high-performance computing clusters or robust cloud computing resources. Such computational demands must be meticulously considered by bioinformaticians and medical staff when planning to implement such an approach, as poorly designed architecture can result in delays in obtaining results or, in some cases, a failure to obtain any results. It is noteworthy that private solutions such as Sentieon [94] or NVIDIA Parabricks [95] propose to accelerate large-scale data analyses, resulting in overall pipeline execution time savings ranging from three- to eightfold [96].

Table 2. Latest NGS DNA analyses recommended by international guidelines. ESMO: European Society for Medical Oncology; NCCN: National Comprehensive Cancer Network; EANO: European Association of Neuro-Oncology; ESGO: European Society of Gynaecological Oncology; ESTRO: European SocieTy for Radiotherapy and Oncology; ESP: European Society of Pathology.

Tumor Type	Alterations	Guidelines	References
Metastatic non-small-cell lung cancer	*EGFR* exons 18–21 mutations	ESMO 2023	[97]
	BRAF V600 mutation		
	NTRK rearrangement		
	KRAS G12C mutation		
	HER2 exon 20 mutations		
	MET exon 14 skipping mutations		
	MET amplifications		
Cutaneous melanoma	*BRAF* V600 mutation	NCCN 2023	NCCN Guidelines Version 2.2023 Melanoma: cutaneous
	NRAS G12, G13, Q61 mutations		
	KIT Exons 8, 9, 11, 13 and 17 mutations		
High-grade glioma	*IDH1* R132 mutation	EANO 2021	[98]
	IDH2 R172 mutation		
	TERT promotor mutation		
	TP53 mutations		
	Histone H3 K27M mutations		
	CDKN2A/B deletions		
	EGFR amplification		

Table 2. *Cont.*

Tumor Type	Alterations	Guidelines	References
High-grade serous ovarian cancer	*BRCA1* and *BRCA2* mutations	ESMO 2019	[99]
	Homologous recombination deficiency		
Endometrial carcinoma	*BRAF* V600 mutation	ESGO/ESTRO/ESP 2021	[100]
	POLE mutations		
	TP53 mutations		
	Microsatellite instability		
Metastatic colorectal carcinoma	KRAS Exons 2–4 mutations	NCCN 2023	NCCN Guidelines Version 3.2023 Colon Cancer
	NRAS Exons 2–4 mutations		
	BRAF V600 mutation		
	Microsatellite instability		
	HER2 amplification		
Thyroid carcinoma	*BRAF* V600 mutation	NCCN 2023	NCCN Guidelines Version 4.2023 Thyroid Carcinoma
	RET mutations		
GIST	*KIT* Exons 8–11 mutations	ESMO 2022	[101]
	PDGFRA D842V mutation		
Pancreatic adenocarcinoma	Microsatellite instability	NCCN 2023	NCCN Guidelines Version 2.2023 Pancreatic adenocarcinoma
	KRAS G12C mutation		
	BRCA1 and *BRCA2* mutations		
	BRAF V600 mutation *		

* Not included in European guidelines to date.

A well-designed and well-documented bioinformatics pipeline provides reliable and accurate guidance for oncologists, ultimately leading to better outcomes for patients. Variant calling, interpretation, and annotation represent critical steps in precision oncology, and rely on bioinformatics expertise and technology. They are altogether aimed at providing personalized cancer treatment dependent on the tumor-specimen-specific genetic alteration revealed.

Variant calling is a complex and challenging task due to the high levels of background noise and variation present in NGS data, as well as the need to distinguish true cancer-related alterations from germline or benign variants. To address these challenges, advanced bioinformatics tools and algorithms have been developed that exploit various strategies, such as statistical modeling or machine and deep learning, to improve the sensitivity, specificity, and reproducibility of variant calling.

Once the genomic variants have been called, the next step is to annotate and interpret them in the context of known biological and clinical knowledge. This includes identifying the functional impact of the variants on proteins and related biological pathways, as well as assessing their potential relevance to cancer development and treatment. In this context, bioinformatics resources such as public databases, biological pathway analysis tools, and drug–gene interaction databases are indispensable to prioritize and contextualize the genomic findings.

By integrating multiple sources of genomic and clinical data, bioinformatics can help identify the most relevant molecular targets and therapeutic options for cancer patients, ultimately improving their outcomes and quality of life. A crucial step in precision oncology is the clinical reporting of molecular findings, which involves the translation of complex genomic data into meaningful clinical implications that can guide patient care. The clinical report should provide clear and concise information on the identified molecular alterations, their relevance to the patient's disease, potential therapeutic options, and any relevant clinical trials. It should also highlight any issues in the data, as well as provide recommendations for further testing or monitoring. The entire process must ensure that the clinical report accurately reflects the molecular landscape of the patient's disease and provides actionable information to guide personalized treatment decisions.

Author Contributions: Conceptualization, S.C.-A. and J.S.; writing—original draft preparation, S.C.-A.; writing—review and editing, S.C.-A., J.A.V. and J.S.; supervision, J.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors acknowledge David Baux for proof reading and insightful suggestions.

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

References

1. Prasad, V.; Fojo, T.; Brada, M. Precision oncology: Origins, optimism, and potential. *Lancet Oncol.* **2016**, *17*, e81–e86. [CrossRef] [PubMed]
2. Buermans, H.P.J.; Den Dunnen, J.T. Next generation sequencing technology: Advances and applications. *Biochim. Biophys. Acta (BBA)—Mol. Basis Dis.* **2014**, *1842*, 1932–1941. [CrossRef] [PubMed]
3. Arora, N.; Chaudhary, A.; Prasad, A. Editorial: Methods and applications in molecular diagnostics. *Front. Mol. Biosci.* **2023**, *10*, 1239005. [CrossRef] [PubMed]
4. Brandies, P.A.; Hogg, C.J. Ten simple rules for getting started with command-line bioinformatics. *PLoS Comput. Biol.* **2021**, *17*, e1008645. [CrossRef] [PubMed]
5. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079. [CrossRef] [PubMed]
6. Leipzig, J. A review of bioinformatic pipeline frameworks. *Brief. Bioinform.* **2016**, *18*, bbw020. [CrossRef]
7. Di Tommaso, P.; Chatzou, M.; Floden, E.W.; Barja, P.P.; Palumbo, E.; Notredame, C. Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **2017**, *35*, 316–319. [CrossRef]
8. Mölder, F.; Jablonski, K.P.; Letcher, B.; Hall, M.B.; Tomkins-Tinch, C.H.; Sochat, V.; Forster, J.; Lee, S.; Twardziok, S.O.; Kanitz, A.; et al. Sustainable data analysis with Snakemake. *F1000Research* **2021**, *10*, 33. [CrossRef]
9. Sadedin, S.P.; Pope, B.; Oshlack, A. Bpipe: A tool for running and managing bioinformatics pipelines. *Bioinformatics* **2012**, *28*, 1525–1526. [CrossRef]
10. Crusoe, M.R.; Abeln, S.; Iosup, A.; Amstutz, P.; Chilton, J.; Tijanić, N.; Ménager, H.; Soiland-Reyes, S.; Gavrilović, B.; Goble, C.; et al. Methods included: Standardizing computational reuse and portability with the Common Workflow Language. *Commun. ACM* **2022**, *65*, 54–63. [CrossRef]
11. Voss, K.; der Auwera, G.V.; Gentry, J. Full-stack genomics pipelining with GATK4 + WDL + Cromwell. *F1000Research* **2017**, *6*. [CrossRef]
12. 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] [PubMed]
13. Ewels, P.A.; Peltzer, A.; Fillinger, S.; Patel, H.; Alneberg, J.; Wilm, A.; Garcia, M.U.; Di Tommaso, P.; Nahnsen, S. The nf-core framework for community-curated bioinformatics pipelines. *Nat. Biotechnol.* **2020**, *38*, 276–278. [CrossRef] [PubMed]
14. Hanssen, F.; Garcia, M.U.; Folkersen, L.; Pedersen, A.S.; Lescai, F.; Jodoin, S.; Miller, E.; Wacker, O.; Smith, N.; Community, N.-C.; et al. Scalable and efficient DNA sequencing analysis on different compute infrastructures aiding variant discovery. *bioRxiv* **2023**, 549462. [CrossRef]
15. Garcia, M.; Juhos, S.; Larsson, M.; Olason, P.I.; Martin, M.; Eisfeldt, J.; DiLorenzo, S.; Sandgren, J.; Ståhl, T.D.D.; Ewels, P.; et al. Sarek: A portable workflow for whole-genome sequencing analysis of germline and somatic variants. *F1000Research* **2020**, *9*, 63. [CrossRef] [PubMed]
16. The Galaxy Community. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Res.* **2022**, *50*, W345–W351. [CrossRef] [PubMed]
17. Oinn, T.; Addis, M.; Ferris, J.; Marvin, D.; Senger, M.; Greenwood, M.; Carver, T.; Glover, K.; Pocock, M.R.; Wipat, A.; et al. Taverna: A tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* **2004**, *20*, 3045–3054. [CrossRef]

18. Abouelhoda, M.; Issa, S.A.; Ghanem, M. Tavaxy: Integrating Taverna and Galaxy workflows with cloud computing support. *BMC Bioinform.* **2012**, *13*, 77. [CrossRef]
19. Scheffler, K.; Catreux, S.; O'Connell, T.; Jo, H.; Jain, V.; Heyns, T.; Yuan, J.; Murray, L.; Han, J.; Mehio, R. Somatic small-variant calling methods in Illumina DRAGEN™ Secondary Analysis. *bioRxiv* **2023**, 534011. [CrossRef]
20. Chen, S.; Zhou, Y.; Chen, Y.; Gu, J. fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **2018**, *34*, i884–i890. [CrossRef]
21. Andrews, S. FastQC: A Quality Control Tool for High Throughput Sequence Data. 2010. Available online: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed on 1 December 2023).
22. Broad Institute Picard Toolkit. 2019. Available online: http://broadinstitute.github.io/picard/ (accessed on 1 December 2023).
23. Ewels, P.; Magnusson, M.; Lundin, S.; Käller, M. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **2016**, *32*, 3047–3048. [CrossRef] [PubMed]
24. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [CrossRef] [PubMed]
25. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **2011**, *17*, 10–12. [CrossRef]
26. Bushnell, B. BBDuk. 2018. Available online: https://sourceforge.net/projects/bbmap/ (accessed on 1 December 2023).
27. Jung, Y.; Han, D. BWA-MEME: BWA-MEM emulated with a machine learning approach. *Bioinformatics* **2022**, *38*, 2404–2413. [CrossRef] [PubMed]
28. Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **2012**, *9*, 357–359. [CrossRef] [PubMed]
29. Kim, D.; Paggi, J.M.; Park, C.; Bennett, C.; Salzberg, S.L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **2019**, *37*, 907–915. [CrossRef] [PubMed]
30. Dobin, A.; Davis, C.A.; Schlesinger, F.; Drenkow, J.; Zaleski, C.; Jha, S.; Batut, P.; Chaisson, M.; Gingeras, T.R. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **2013**, *29*, 15–21. [CrossRef]
31. Poplin, R.; Ruano-Rubio, V.; DePristo, M.A.; Fennell, T.J.; Carneiro, M.O.; der Auwera, G.A.V.; 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* **2018**, 201178. [CrossRef]
32. Garrison, E.; Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv* **2012**, arXiv:1207.3907.
33. Benjamin, D.; Sato, T.; Cibulskis, K.; Getz, G.; Stewart, C.; Lichtenstein, L. Calling Somatic SNVs and Indels with Mutect2. *bioRxiv* **2019**, 861054. [CrossRef]
34. Lai, Z.; Markovets, A.; Ahdesmaki, M.; Chapman, B.; Hofmann, O.; McEwen, R.; Johnson, J.; Dougherty, B.; Barrett, J.C.; Dry, J.R. VarDict: A novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res.* **2016**, *44*, e108. [CrossRef] [PubMed]
35. Sherry, S.T.; Ward, M.; Sirotkin, K. dbSNP—Database for Single Nucleotide Polymorphisms and Other Classes of Minor Genetic Variation. *Genome Res.* **1999**, *9*, 677–679. [CrossRef] [PubMed]
36. Auton, A.; Abecasis, G.R.; Altshuler, D.M.; Durbin, R.M.; Abecasis, G.R.; Bentley, D.R.; Chakravarti, A.; Clark, A.G.; Donnelly, P.; Eichler, E.E.; et al. A global reference for human genetic variation. *Nature* **2015**, *526*, 68–74. [CrossRef] [PubMed]
37. Karczewski, K.J.; Francioli, L.C.; Tiao, G.; Cummings, B.B.; Alföldi, J.; Wang, Q.; Collins, R.L.; Laricchia, K.M.; Ganna, A.; Birnbaum, D.P.; et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **2020**, *581*, 434–443. [CrossRef] [PubMed]
38. 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] [PubMed]
39. Baux, D.; Van Goethem, C.; Ardouin, O.; Guignard, T.; Bergougnoux, A.; Koenig, M.; Roux, A.-F. MobiDetails: Online DNA variants interpretation. *Eur. J. Hum. Genet.* **2021**, *29*, 356–360. [CrossRef] [PubMed]
40. Wang, K.; Li, M.; Hakonarson, H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **2010**, *38*, e164. [CrossRef]
41. Cingolani, P.; Platts, A.; Wang, L.L.; Coon, M.; Nguyen, T.; Wang, L.; Land, S.J.; Lu, X.; Ruden, D.M. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w ^{1118}; iso-2; iso-3. *Fly* **2012**, *6*, 80–92. [CrossRef]
42. Breunig, M.M.; Kriegel, H.-P.; Ng, R.T.; Sander, J. LOF: Identifying Density-Based Local Outliers. *ACM SIGMOD Rec.* **2000**, *29*, 93–104. [CrossRef]
43. Rapti, M.; Zouaghi, Y.; Meylan, J.; Ranza, E.; Antonarakis, S.E.; Santoni, F.A. CoverageMaster: Comprehensive CNV detection and visualization from NGS short reads for genetic medicine applications. *Brief. Bioinform.* **2022**, *23*, bbac049. [CrossRef]
44. Onsongo, G.; Baughn, L.B.; Bower, M.; Henzler, C.; Schomaker, M.; Silverstein, K.A.T.; Thyagarajan, B. CNV-RF Is a Random Forest-Based Copy Number Variation Detection Method Using Next-Generation Sequencing. *J. Mol. Diagn.* **2016**, *18*, 872–881. [CrossRef] [PubMed]
45. Glessner, J.T.; Hou, X.; Zhong, C.; Zhang, J.; Khan, M.; Brand, F.; Krawitz, P.; Sleiman, P.M.A.; Hakonarson, H.; Wei, Z. DeepCNV: A deep learning approach for authenticating copy number variations. *Brief. Bioinform.* **2021**, *22*, bbaa381. [CrossRef] [PubMed]
46. Yuan, X.; Yu, J.; Xi, J.; Yang, L.; Shang, J.; Li, Z.; Duan, J. CNV_IFTV: An Isolation Forest and Total Variation-Based Detection of CNVs from Short-Read Sequencing Data. *IEEE ACM Trans. Comput. Biol. Bioinf.* **2021**, *18*, 539–549. [CrossRef] [PubMed]
47. Guo, Y.; Wang, S.; Yuan, X. HBOS-CNV: A New Approach to Detect Copy Number Variations From Next-Generation Sequencing Data. *Front. Genet.* **2021**, *12*, 642473. [CrossRef] [PubMed]

48. Huang, T.; Li, J.; Jia, B.; Sang, H. CNV-MEANN: A Neural Network and Mind Evolutionary Algorithm-Based Detection of Copy Number Variations From Next-Generation Sequencing Data. *Front. Genet.* **2021**, *12*, 700874. [CrossRef] [PubMed]
49. 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] [PubMed]
50. Cabello-Aguilar, S.; Vendrell, J.A.; Van Goethem, C.; Brousse, M.; Gozé, C.; Frantz, L.; Solassol, J. ifCNV: A novel isolation-forest-based package to detect copy-number variations from various targeted NGS datasets. *Mol. Ther.—Nucleic Acids* **2022**, *30*, 174–183. [CrossRef] [PubMed]
51. Viailly, P.-J.; Sater, V.; Viennot, M.; Bohers, E.; Vergne, N.; Berard, C.; Dauchel, H.; Lecroq, T.; Celebi, A.; Ruminy, P.; et al. Improving high-resolution copy number variation analysis from next generation sequencing using unique molecular identifiers. *BMC Bioinform.* **2021**, *22*, 120. [CrossRef]
52. Escudié, F.; Van Goethem, C.; Grand, D.; Vendrell, J.; Vigier, A.; Brousset, P.; Evrard, S.M.; Solassol, J.; Selves, J. MIAmS: Microsatellite instability detection on NGS amplicons data. *Bioinformatics* **2019**, *36*, btz797. [CrossRef]
53. Niu, B.; Ye, K.; Zhang, Q.; Lu, C.; Xie, M.; McLellan, M.D.; Wendl, M.C.; Ding, L. MSIsensor: Microsatellite instability detection using paired tumor-normal sequence data. *Bioinformatics* **2014**, *30*, 1015–1016. [CrossRef]
54. Swaerts, K.; Dedeurwaerdere, F.; De Smet, D.; De Jaeger, P.; Martens, G.A. DeltaMSI: Artificial intelligence-based modeling of microsatellite instability scoring on next-generation sequencing data. *BMC Bioinform.* **2023**, *24*, 73. [CrossRef] [PubMed]
55. Marques, A.C.; Ferraro-Peyret, C.; Michaud, F.; Song, L.; Smith, E.; Fabre, G.; Willig, A.; Wong, M.M.L.; Xing, X.; Chong, C.; et al. Improved NGS-based detection of microsatellite instability using tumor-only data. *Front. Oncol.* **2022**, *12*, 969238. [CrossRef] [PubMed]
56. Schneider, V.A.; Graves-Lindsay, T.; Howe, K.; Bouk, N.; Chen, H.-C.; Kitts, P.A.; Murphy, T.D.; Pruitt, K.D.; Thibaud-Nissen, F.; Albracht, D.; et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **2017**, *27*, 849–864. [CrossRef] [PubMed]
57. Aganezov, S.; Yan, S.M.; Soto, D.C.; Kirsche, M.; Zarate, S.; Avdeyev, P.; Taylor, D.J.; Shafin, K.; Shumate, A.; Xiao, C.; et al. A complete reference genome improves analysis of human genetic variation. *Science* **2022**, *376*, eabl3533. [CrossRef] [PubMed]
58. Robinson, J.T.; Thorvaldsdóttir, H.; Winckler, W.; Guttman, M.; Lander, E.S.; Getz, G.; Mesirov, J.P. Integrative Genomics Viewer. *Nat. Biotechnol.* **2011**, *29*, 24–26. [CrossRef] [PubMed]
59. DePristo, M.A.; Banks, E.; Poplin, R.; Garimella, K.V.; Maguire, J.R.; Hartl, C.; Philippakis, A.A.; del Angel, G.; Rivas, M.A.; Hanna, M.; et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **2011**, *43*, 491–498. [CrossRef] [PubMed]
60. Warden, C.D.; Adamson, A.W.; Neuhausen, S.L.; Wu, X. Detailed comparison of two popular variant calling packages for exome and targeted exon studies. *PeerJ* **2014**, *2*, e600. [CrossRef]
61. Bian, X.; Zhu, B.; Wang, M.; Hu, Y.; Chen, Q.; Nguyen, C.; Hicks, B.; Meerzaman, D. Comparing the performance of selected variant callers using synthetic data and genome segmentation. *BMC Bioinform.* **2018**, *19*, 429. [CrossRef]
62. Zook, J.M.; McDaniel, J.; Olson, N.D.; Wagner, J.; Parikh, H.; Heaton, H.; Irvine, S.A.; Trigg, L.; Truty, R.; McLean, C.Y.; et al. An open resource for accurately benchmarking small variant and reference calls. *Nat. Biotechnol.* **2019**, *37*, 561–566. [CrossRef]
63. Ellrott, K.; Bailey, M.H.; Saksena, G.; Covington, K.R.; Kandoth, C.; Stewart, C.; Hess, J.; Ma, S.; Chiotti, K.E.; McLellan, M.; et al. Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Syst.* **2018**, *6*, 271–281.e7. [CrossRef]
64. Xu, C. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Comput. Struct. Biotechnol. J.* **2018**, *16*, 15–24. [CrossRef] [PubMed]
65. Chen, Z.; Yuan, Y.; Chen, X.; Chen, J.; Lin, S.; Li, X.; Du, H. Systematic comparison of somatic variant calling performance among different sequencing depth and mutation frequency. *Sci. Rep.* **2020**, *10*, 3501. [CrossRef] [PubMed]
66. 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.* **2018**, *46*, D1062–D1067. [CrossRef] [PubMed]
67. Olivier, M.; Eeles, R.; Hollstein, M.; Khan, M.A.; Harris, C.C.; Hainaut, P. The IARC TP53 database: New online mutation analysis and recommendations to users. *Hum. Mutat.* **2002**, *19*, 607–614. [CrossRef] [PubMed]
68. Spurdle, A.B.; Healey, S.; Devereau, A.; Hogervorst, F.B.L.; Monteiro, A.N.A.; Nathanson, K.L.; Radice, P.; Stoppa-Lyonnet, D.; Tavtigian, S.; Wappenschmidt, B.; et al. ENIGMA-Evidence-based network for the interpretation of germline mutant alleles: An international initiative to evaluate risk and clinical significance associated with sequence variation in BRCA1 and BRCA2 genes. *Hum. Mutat.* **2012**, *33*, 2–7. [CrossRef] [PubMed]
69. Vaser, R.; Adusumalli, S.; Leng, S.N.; Sikic, M.; Ng, P.C. SIFT missense predictions for genomes. *Nat. Protoc.* **2016**, *11*, 1–9. [CrossRef] [PubMed]
70. Adzhubei, I.A.; Schmidt, S.; Peshkin, L.; Ramensky, V.E.; Gerasimova, A.; Bork, P.; Kondrashov, A.S.; Sunyaev, S.R. A method and server for predicting damaging missense mutations. *Nat. Methods* **2010**, *7*, 248–249. [CrossRef]
71. Cheng, J.; Novati, G.; Pan, J.; Bycroft, C.; Žemgulytė, A.; Applebaum, T.; Pritzel, A.; Wong, L.H.; Zielinski, M.; Sargeant, T.; et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **2023**, *381*, eadg7492. [CrossRef]

72. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. [CrossRef]
73. Freshour, S.L.; Kiwala, S.; Cotto, K.C.; Coffman, A.C.; McMichael, J.F.; Song, J.J.; Griffith, M.; Griffith, O.L.; Wagner, A.H. Integration of the Drug–Gene Interaction Database (DGIdb 4.0) with open crowdsource efforts. *Nucleic Acids Res.* **2021**, *49*, D1144–D1151. [CrossRef]
74. Chakravarty, D.; Gao, J.; Phillips, S.; Kundra, R.; Zhang, H.; Wang, J.; Rudolph, J.E.; Yaeger, R.; Soumerai, T.; Nissan, M.H.; et al. OncoKB: A Precision Oncology Knowledge Base. *JCO Precis. Oncol.* **2017**, 1–16. [CrossRef] [PubMed]
75. Griffith, M.; Spies, N.C.; Krysiak, K.; McMichael, J.F.; Coffman, A.C.; Danos, A.M.; Ainscough, B.J.; Ramirez, C.A.; Rieke, D.T.; Kujan, L.; et al. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nat. Genet.* **2017**, *49*, 170–174. [CrossRef] [PubMed]
76. Verma, S.; Miles, D.; Gianni, L.; Krop, I.E.; Welslau, M.; Baselga, J.; Pegram, M.; Oh, D.-Y.; Diéras, V.; Guardino, E.; et al. Trastuzumab Emtansine for HER2-Positive Advanced Breast Cancer. *N. Engl. J. Med.* **2012**, *367*, 1783–1791. [CrossRef]
77. von Minckwitz, G.; Huang, C.-S.; Mano, M.S.; Loibl, S.; Mamounas, E.P.; Untch, M.; Wolmark, N.; Rastogi, P.; Schneeweiss, A.; Redondo, A.; et al. Trastuzumab Emtansine for Residual Invasive HER2-Positive Breast Cancer. *N. Engl. J. Med.* **2019**, *380*, 617–628. [CrossRef] [PubMed]
78. Lee, K.; Kim, D.; Yoon, S.; Lee, D.H.; Kim, S.-W. Exploring the resistance mechanisms of second-line osimertinib and their prognostic implications using next-generation sequencing in patients with non-small-cell lung cancer. *Eur. J. Cancer* **2021**, *148*, 202–210. [CrossRef] [PubMed]
79. Camidge, D.R.; Otterson, G.A.; Clark, J.W.; Ignatius Ou, S.-H.; Weiss, J.; Ades, S.; Shapiro, G.I.; Socinski, M.A.; Murphy, D.A.; Conte, U.; et al. Crizotinib in Patients With MET-Amplified NSCLC. *J. Thorac. Oncol.* **2021**, *16*, 1017–1029. [CrossRef] [PubMed]
80. Chen, K.; Wallis, J.W.; McLellan, M.D.; Larson, D.E.; Kalicki, J.M.; Pohl, C.S.; McGrath, S.D.; Wendl, M.C.; Zhang, Q.; Locke, D.P.; et al. BreakDancer: An algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods* **2009**, *6*, 677–681. [CrossRef] [PubMed]
81. Zhang, J.; Wang, J.; Wu, Y. An improved approach for accurate and efficient calling of structural variations with low-coverage sequence data. *BMC Bioinform.* **2012**, *13*, S6. [CrossRef]
82. Jiang, Y.; Wang, Y.; Brudno, M. PRISM: Pair-read informed split-read mapping for base-pair level detection of insertion, deletion and structural variants. *Bioinformatics* **2012**, *28*, 2576–2583. [CrossRef]
83. Mahmoud, M.; Gobet, N.; Cruz-Dávalos, D.I.; Mounier, N.; Dessimoz, C.; Sedlazeck, F.J. Structural variant calling: The long and the short of it. *Genome Biol.* **2019**, *20*, 246. [CrossRef]
84. Gologan, A.; Sepulveda, A.R. Microsatellite instability and DNA mismatch repair deficiency testing in hereditary and sporadic gastrointestinal cancers. *Clin. Lab. Med.* **2005**, *25*, 179–196. [CrossRef] [PubMed]
85. Motta, R.; Cabezas-Camarero, S.; Torres-Mattos, C.; Riquelme, A.; Calle, A.; Figueroa, A.; Sotelo, M.J. Immunotherapy in microsatellite instability metastatic colorectal cancer: Current status and future perspectives. *J. Clin. Transl. Res.* **2021**, *7*, 511–522. [PubMed]
86. Thibodeau, S.N.; Bren, G.; Schaid, D. Microsatellite instability in cancer of the proximal colon. *Science* **1993**, *260*, 816–819. [CrossRef] [PubMed]
87. Koboldt, D.C. Best practices for variant calling in clinical sequencing. *Genome Med.* **2020**, *12*, 91. [CrossRef] [PubMed]
88. 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] [PubMed]
89. Goodman, A.M.; Kato, S.; Bazhenova, L.; Patel, S.P.; Frampton, G.M.; Miller, V.; Stephens, P.J.; Daniels, G.A.; Kurzrock, R. Tumor Mutational Burden as an Independent Predictor of Response to Immunotherapy in Diverse Cancers. *Mol. Cancer Ther.* **2017**, *16*, 2598–2608. [CrossRef] [PubMed]
90. Chalmers, Z.R.; Connelly, C.F.; Fabrizio, D.; Gay, L.; Ali, S.M.; Ennis, R.; Schrock, A.; Campbell, B.; Shlien, A.; Chmielecki, J.; et al. Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Med.* **2017**, *9*, 34. [CrossRef]
91. Wang, Z.; Duan, J.; Cai, S.; Han, M.; Dong, H.; Zhao, J.; Zhu, B.; Wang, S.; Zhuo, M.; Sun, J.; et al. Assessment of Blood Tumor Mutational Burden as a Potential Biomarker for Immunotherapy in Patients With Non–Small Cell Lung Cancer With Use of a Next-Generation Sequencing Cancer Gene Panel. *JAMA Oncol.* **2019**, *5*, 696–702. [CrossRef]
92. Budczies, J.; Allgäuer, M.; Litchfield, K.; Rempel, E.; Christopoulos, P.; Kazdal, D.; Endris, V.; Thomas, M.; Fröhling, S.; Peters, S.; et al. Optimizing panel-based tumor mutational burden (TMB) measurement. *Ann. Oncol.* **2019**, *30*, 1496–1506. [CrossRef]
93. Fang, H.; Bertl, J.; Zhu, X.; Lam, T.C.; Wu, S.; Shih, D.J.H.; Wong, J.W.H. Tumour mutational burden is overestimated by target cancer gene panels. *J. Natl. Cancer Cent.* **2023**, *3*, 56–64. [CrossRef]
94. Freed, D.; Aldana, R.; Weber, J.; Edwards, J. The Sentieon Genomics Tools—A fast and accurate solution to variant calling from next-generation sequence data. *bioRxiv* **2017**. [CrossRef]
95. O'Connell, K.A.; Yosufzai, Z.B.; Campbell, R.A.; Lobb, C.J.; Engelken, H.T.; Gorrell, L.M.; Carlson, T.B.; Catana, J.J.; Mikdadi, D.; Bonazzi, V.R.; et al. Accelerating genomic workflows using NVIDIA Parabricks. *BMC Bioinform.* **2023**, *24*, 221. [CrossRef] [PubMed]

96. Franke, K.R.; Crowgey, E.L. Accelerating next generation sequencing data analysis: An evaluation of optimized best practices for Genome Analysis Toolkit algorithms. *Genom. Inform.* **2020**, *18*, e10. [CrossRef] [PubMed]
97. Hendriks, L.E.; Kerr, K.M.; Menis, J.; Mok, T.S.; Nestle, U.; Passaro, A.; Peters, S.; Planchard, D.; Smit, E.F.; Solomon, B.J.; et al. Oncogene-addicted metastatic non-small-cell lung cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up. *Ann. Oncol.* **2023**, *34*, 339–357. [CrossRef] [PubMed]
98. Weller, M.; van den Bent, M.; Preusser, M.; Le Rhun, E.; Tonn, J.C.; Minniti, G.; Bendszus, M.; Balana, C.; Chinot, O.; Dirven, L.; et al. EANO guidelines on the diagnosis and treatment of diffuse gliomas of adulthood. *Nat. Rev. Clin. Oncol.* **2021**, *18*, 170–186. [CrossRef]
99. Colombo, N.; Sessa, C.; du Bois, A.; Ledermann, J.; McCluggage, W.G.; McNeish, I.; Morice, P.; Pignata, S.; Ray-Coquard, I.; Vergote, I.; et al. ESMO-ESGO consensus conference recommendations on ovarian cancer: Pathology and molecular biology, early and advanced stages, borderline tumours and recurrent disease. *Ann. Oncol.* **2019**, *30*, 672–705. [CrossRef]
100. Concin, N.; Matias-Guiu, X.; Vergote, I.; Cibula, D.; Mirza, M.R.; Marnitz, S.; Ledermann, J.; Bosse, T.; Chargari, C.; Fagotti, A.; et al. ESGO/ESTRO/ESP guidelines for the management of patients with endometrial carcinoma. *Int. J. Gynecol. Cancer* **2021**, *31*, 12–39. [CrossRef]
101. Casali, P.G.; Blay, J.Y.; Abecassis, N.; Bajpai, J.; Bauer, S.; Biagini, R.; Bielack, S.; Bonvalot, S.; Boukovinas, I.; Bovee, J.V.M.G.; et al. Gastrointestinal stromal tumours: ESMO–EURACAN–GENTURIS Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* **2022**, *33*, 20–33. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.