


Review

Next Generation Sequencing Technology in the Clinic and Its Challenges

Lau K. Vestergaard ¹, Douglas N. P. Oliveira ¹, Claus K. Høgdall ² and Estrid V. Høgdall ^{1,*} 

¹ Molecular Unit, Department of Pathology, Herlev Hospital, University of Copenhagen, DK-2730 Herlev, Denmark; lau.kraesing.vestergaard@regionh.dk (L.K.V.); douglas.nogueira.perez.de.oliveira@regionh.dk (D.N.P.O.)

² Juliane Marie Centre, Department of Gynecology, Rigshospitalet, University of Copenhagen, DK-2100 Copenhagen, Denmark; claus.hogdall@regionh.dk

* Correspondence: estrid.hoegdall@regionh.dk; Tel.: +45-3868-9132

Simple Summary: Precise identification and annotation of mutations are of utmost importance in clinical oncology. Insights of the DNA sequence can provide meaningful knowledge to unravel the underlying genetics of disease. Hence, tailoring of personalized medicine often relies on specific genomic alteration for treatment efficacy. The aim of this review is to highlight that sequencing harbors much more than just four nucleotides. Moreover, the gradual transition from first to second generation sequencing technologies has led to awareness for choosing the most appropriate bioinformatic analytic tools based on the aim, quality and demand for a specific purpose. Thus, the same raw data can lead to various results reflecting the intrinsic features of different datamining pipelines.

Abstract: Data analysis has become a crucial aspect in clinical oncology to interpret output from next-generation sequencing-based testing. NGS being able to resolve billions of sequencing reactions in a few days has consequently increased the demand for tools to handle and analyze such large data sets. Many tools have been developed since the advent of NGS, featuring their own peculiarities. Increased awareness when interpreting alterations in the genome is therefore of utmost importance, as the same data using different tools can provide diverse outcomes. Hence, it is crucial to evaluate and validate bioinformatic pipelines in clinical settings. Moreover, personalized medicine implies treatment targeting efficacy of biological drugs for specific genomic alterations. Here, we focused on different sequencing technologies, features underlying the genome complexity, and bioinformatic tools that can impact the final annotation. Additionally, we discuss the clinical demand and design for implementing NGS.

Keywords: bioinformatic pipeline; cancer; next-generation sequencing; alignment; variant calling; clinical application



Citation: Vestergaard, L.K.; Oliveira, D.N.P.; Høgdall, C.K.; Høgdall, E.V. Next Generation Sequencing Technology in the Clinic and Its Challenges. *Cancers* **2021**, *13*, 1751. <https://doi.org/10.3390/cancers13081751>

Academic Editor: Katia Nones

Received: 9 March 2021

Accepted: 5 April 2021

Published: 7 April 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Insights into the sequence of DNA can provide meaningful knowledge to unravel the genetics of disease. This approach has propelled diagnostic and treatment strategies to a new level, where personalized medicine is gradually becoming adopted in the clinic. The advent of second-generation sequencing technologies, also known as next-generation sequencing (NGS), has contributed remarkably with the demand for more economical and faster sequencing technologies. NGS performs massive parallel sequencing and is steadily replacing its predecessor, the traditional Sanger sequencing (Sanger et al., 1977) [1]. Its technologies have made it possible to resolve billions of sequencing reactions in few days from library preparations to end results. Nonetheless, the handling of such substantial amount of data poses a current challenge regarding their interpretation in a clinically meaningful way. Hence, that demand was shortly followed by the development of a plethora of devised NGS bioinformatic tools, each serving its own purpose. Most computational tools

such as Bowtie2 [2], Burrows–Wheeler Aligner (BWA) [3], Mutect2 [4] and Strelka2 [5] are freely available for processing NGS data in the scopes of (i) sequence mapping, (ii) base calling and (iii) variant calling. The application of different tools have shown to vary in consistency [6,7], highlighting the necessity of caution and experience, as the output could lead to misguidance of diagnosis, prognosis and personalized treatment in a clinical setting. In this review, we focused on the many factors that influence data interpretation and its application in oncology. This covers sequencing technologies, data output from sequencing, pitfalls and bioinformatics concerns. Finally, we discussed the increasing clinical demand for the implementation of NGS.

2. Sequencing Technologies

Sequencing is the process of determining the order of Adenine (A), Guanine (G), Cytosine (C) and Thymine (T) bases, which makes up the primary structure of DNA. The first two DNA sequencing methodologies are known as Maxam–Gilbert sequencing—a chemical approach [8]—and Sanger sequencing—a chain termination approach [9]. The clinical utility of Maxam–Gilbert sequencing is unknown; hence the latter will be further addressed herein. Sanger sequencing provided the basis for The Human Genome Project [10] given its accuracy, robustness and simplicity [11]. Briefly, the method is based on 4 polymerase chain reactions (PCR), where in each reaction one of the nucleotides is incorporated by a specific fluorescent chain-terminating dideoxynucleotide (ddNTP). The ddNTP incorporation during the *in vitro* DNA replication is random, producing fragments with varying length. Subsequently size separation via gel electrophoresis reveal the arrangement of the nucleotides based on where the fragment was terminated [9]. The specific fluorescence embedded in each ddNTP (ddATP, ddGTP, ddCTP, ddTTP) allows to read and annotate the sequence. Despite being a very robust and precise method, the Sanger sequencing can only be performed for a single target at a time. Hence, assessing even a small panel of targets makes this approach cost- and time-inefficient.

The application of faster and lower cost sequencing was introduced by the NGS technologies. Such platforms, as the Illumina and Ion Torrent, are predominantly being used in clinical settings. These two platforms are different in their underlying chemistry of determining the base sequences. In Illumina platforms, a library is first prepared (DNA templates with barcodes and adaptors attached), followed by denaturation to single strands and immobilization on a flow cell. Next, the templates are amplified to form clusters of clonal fragments by bridge amplification [1]. That step is important to yield enough signal for detection during sequencing. The sequencing methodology of Illumina is based upon cyclic reversible dye chain termination in which ddNTP contains different cleavable fluorescence and a reversible blocking group. For each round of sequencing (1) a nucleotide is added to the flow cell; (2) the fluorescent signals are captured and converted into A, T, C or G; (3) the blocking groups are removed; and (4) the process is repeated with a new round of nucleotide incorporation until the strands are synthesized. The Ion Torrent sequencing technology is based on the incorporation of a DNA template from a prepared library together with a single bead into a droplet, referred as bead emulsion. Each reaction unit allows for emulsion PCR to clonally amplify the template until it covers the entire surface of the corresponding bead. This step is analogous to Illumina’s flow cell, it is yielding sufficient signal during sequencing. The covered beads are loaded onto a chip constructed with millions of micro wells to harbor a single bead in each. The sequencing process is based on pH. Each time a nucleotide is added to the synthesizing template a H⁺ (hydrogen ion) is released and detected as a change in pH, the chip is flooded with one nucleotide at a time and the process is repeated hundreds of times.

For the past few years, a new generation of sequencing methods has been under development. In contrast to NGS technologies generating short reads, third generation sequencing (TGS) aims to generate long reads up to 30,000 bases (30 kb) in length in real time [12]. The MinION from Oxford Nanopore Technologies (ONT) and the single molecule real-time (SMRT) technology from Pacific Bioscience (PacBio) are two types of

a TGS technology. TGS bypass the prerequisite for DNA amplification underlying NGS technologies [11]. Hence, avoiding the inherited errors from the amplification step and creating a fast transition from sample acquisition to sequencing. Yet, the sequencing error rate is still high, 10–15% for SMRT [13] and 5–20% for MinION [13] challenging its utility in a clinical setting. A brief comparison of the technologies is presented in Table 1.

Table 1. Sequencing platforms & comparison.

Platform	Immobilization	Amplification	Sequencing Technology	Limitations & Error Rate	Read Length (bp *)	Run Time (h **)	Output (Gb ***)
1st generation technologies							
Sanger	N/A	PCR with dNTPs and ddNTPs	Irreversible chain termination	0.001%	≤900	~4	≤0.002
Maxam-Gilbert	N/A	N/A	Chemical termination of ³² P labeled ssDNA	0.001%	≤900	N/A	≤0.002
2nd generation technologies							
HiSeq2000	Flow cell	Bridge amplification	Cyclic reversible dye chain termination	GC-rich regions 0.2%	≤125	7–144	≤1600
MiSeq	Flow cell	Bridge amplification	Cyclic reversible dye chain termination	GC-rich regions 0.2%	≤300	4–55	≤15
Ion Torrent PGM	Bead emulsion	Emulsion PCR	Synthesis depended H ⁺ detection	Homo-polymers 1% Indel	≤400	2–7.5	≤2
Ion Torrent S5XL	Bead emulsion	Emulsion PCR	Synthesis depended H ⁺ detection	Homo-polymers 1% Indel	≤600	2.5–4	≤25
3rd generation technologies							
ONT MinION	Processive enzyme	N/A	Monitoring the current of a nucleotide in ssDNA	5–20%	10,000–30,000	Real time	≤25
PacBio SMRT	DNA attachment to the bottom of each Zero Mode Waveguide	N/A	Detection of incorporation of fluorescent nucleotides during real time synthesis	10–15%	10,000–30,000	Real time	≤4

* bp = base pair, ** h = hours, *** Gb = gigabyte.

3. Extend of Sequencing

The human genome is constituted by approximately 3 billion nucleotides, of which around 1% encode for protein-coding genes [14]. Mutations introduced into these genes may show consequences by malfunctioning (loss-of-function) or dysregulation (gain-of-function) of proteins crucial for homeostasis, leading to cancer. Mutations are defined as driver mutations when acquiring a cellular phenotype that contributes with an advantage of proliferation and/or survival [15]. Besides driver mutations, the genome is lodging thousands of mutations that are randomly dispersed throughout the genome. These are referred to as passenger mutations and exhibits no immediate phenotype and/or beneficial advantage [16]. In 80% of the cases, cancer is a multifactorial and non-mendelian disease, with somatic mutations found in associated genes at disease [17]. In the remaining 20%, germline mutations are identified [17]. Identification of rare events in genes contributing to tumorigenesis is important for the ongoing understanding of cancer [18]. NGS has led to the discovery of numerous candidates associated with cancer [19]. Noteworthy, detection

of variants via bioinformatic methods can only prioritize novel findings of mutations and genes for functional testing. Hence, can only mutations as drivers of tumorigenesis, which needs validation on experimental settings [20]. Thus, bioinformatic tools should be perceived more as a predictor than a validator.

NGS enables the generation of data from a full genome (Whole Genome Sequencing, WGS) in a few working days. Sequencing of the protein-coding genes (Whole Exome Sequencing, WES) and exome sequencing of selected genes alone or in combination with hot spots regions (Targeted Exome Sequencing, TES and Panel sequencing, PS) has also become available. In clinical oncology, the latter approach is employed extensively for its ability to target cancer related gene-panels with a fast response time. The broad scope of WGS and WES exhibits some challenges to be implemented into clinical setting (discussed later). Thus, to provide information of diagnostic classification, guide therapeutic decisions and/or enlighten prognostic of tumor in shorter time, assessment of gene panels is an informative approach.

4. Targeted Drug Therapies

Several cancer therapies rely on a certain genomic profile to obtain treatment efficacy. Therefore, precise detection of mutations is critical. Sanger sequencing has until recently been used in diagnostic, despite being restricted to few genes. Hence, providing oncologists with limited information about the tumor mutational profile, leading to a more one-fits-all type of therapies [21]. However, given its massive generation of information, NGS promoted the foundation for targeting disease based on individual genomic profile, referred to as Personalized Medicine or Precision Medicine (PM). This concept gives the opportunity for an accurate and effectively treatment strategy [22].

Precise annotation of mutations is required to transform staggering amount of sequencing data into clinically relevant variants with high confidence. Hence, pushing optimal tailoring of a therapeutic course. For instance, poly (ADP-ribose) polymerase (PARP) inhibiting drugs, is used in managing patients with ovarian cancer [23] and breast cancer [24] in cases of pathogenic *BRCA-1/2* -gene mutations.

Additionally, Imatinib, a small molecule that competitively binds to the active site of a tyrosine kinase, is used in treatment of gastrointestinal stromal tumors in cases of *c-KIT* gene mutations [25]. Moreover, competitive kinase inhibitors, are directed towards the increased activity of *BRAF* induced by a specific alteration (V600E/K) in the *BRAF* -gene in patients diagnosed with malignant melanoma [26]. *KRAS* mutations are observed in 15–25% of all cancers, whither 30–40% of colorectal cancer harbor at least a single mutation on that gene [27]. On the other hand, epidermal growth factor receptor (EGFR)-inhibitor, targeting the EGFR on the surface of cell is used in cases of colorectal cancer without mutations in the *RAS*-genes [28]. A selection of specific genetic alterations to guide treatment is presented in Table 2.

Additional Annotation Tool—Drug Databases

Many variants have well-established clinical relevance with targets for molecular therapy. Nevertheless, guiding treatment decisions for novel or rare somatic mutations might be challenging. In that regard, growing databases of variants and putatively beneficial drug molecular targeting are in constant development and can be useful tools to assist guidance, such as the Catalogue Of Somatic Mutations In Cancer (COSMIC) [29,30], ClinVar [31,32] and Precision Oncology Knowledge Base (OncoKB) [33]. Contents of these databases are derived from in vitro and/or in vivo validation studies and clinical investigation expert panels. With the incorporation of NGS into the clinic, such databases are steadily growing and improving as means to assist the treatment of future patients.

Table 2. Examples of genetic aberrations in cancers to guide personalized medicine. The list is devised from information collected from COSMIC, ClinVar and OncoKB.

Gene	Aberration	Targeting Drug	Cancer Type
<i>BRCA-1/2</i>	Loss-of-function	PARP-inhibitor	Breast cancer, Ovarian cancer, Prostate cancer
<i>ERBB2/HER2</i>	Amplification	Dimerization- inhibitor of HER2-HER3	Breast cancer
<i>PIK3CA</i>	Gain-of-function	PIK3 kinase-inhibitor	Breast cancer
<i>BCL2</i>	Gain-of-function (17 bp deletion)	Blocker of Bcl-2	Chronic lymphocytic leukemia
<i>RAS</i>	Wild type	EGFR-inhibitor	Colorectal cancer
<i>c-KIT</i>	Gain-of-function (exon 9, 11, 13, and 17)	Tyrosine kinase-inhibitors	Gastrointestinal Stromal Tumor
<i>EGFR</i>	Gain-of-function (exon 19 deletion and/or L858R)	Tyrosine kinase-inhibitors	Lung cancer, Brain cancer
<i>BRAF</i>	Gain-of-function (V600E/K)	Kinase-inhibitor	Melanoma
<i>CDK12</i>	Loss-of-function	PARP-inhibitor	Prostate cancer

5. Precautions of Data Output from Sequencing

The overall demand for sequencing is to annotate accurate mutations, such as single nucleotide variants (SNV), insertions/deletions (indels), copy number variation (CNV) and structural variation (SV). That should be acquired ideally with high sensitivity (true positives) and specificity (true negatives). A general principle of sequencing is that the broader the scope the lesser the read depth. WGS is on average sequenced to depths of 30–50x [34], making it more explorative oriented, but efficiently enough to identify most germline mutations including, SNV and indels. It also allows for a comprehensive large scale genomic detection of relevant variants, such as large SV or CNV across the whole genome [35]. However, WGS may be insufficient in detecting rare somatic mutations harboring a cancer genome.

The therapeutics available today are exclusively directed against pathogenic alterations in the coding genome. Thus, knowledge of mutations in intronic regions are less informative in clinical oncology. In that regard, WES or large gene panels are more suited for this purpose, where regions can reach an average coverage of 200x [34]. However, targeted sequencing focus on sequencing regions of choice, often gene panels associated with cancer in specific organs with clinical impact. By narrowing the scope to a selected panel, genes and hot-spot regions can be sequenced to depths with more than 1000x coverage [34]. That entails its capability to reach depth able to detect unique low-frequency allele somatic mutations.

NGS is a multifactorial technology, and wariness is important when interpreting results. Factors that may influence results include; type of biological specimen; preanalytical treatment; pseudogenes and repetitive regions; bioinformatic challenges dealing with alignment and variant calling.

5.1. Type of Biological Specimen

Biological specimens vary according to whether the material is collected in a research facility or in a clinical setting. Research facilities most often deal with cultured cells and/or xenograft models leaving high quality DNA to be subjected for sequencing. When human tissue is collected from biopsy or radical surgery in a clinical setting, it is commonly

subjected to formalin fixation and embedded into paraffin blocks (FFPE) or may optimally be collected as fresh frozen (FF) tissue.

5.1.1. FF and FFPE Tissue

At pathology departments tissues are routinely FFPE-prepared, which are stored at room temperature and confers more flexibility in applications. FFPE samples are used in histopathological examination, immunohistochemistry and/or in situ hybridization. Thus, allowing for further molecular characterization of the tumor. However, the chemical procedure underlying FFPE samples facilitates significant fragmentation and chemically artificial modifications to the nucleic acids, altering the DNA sequence [36]. Formalin-fixation is the primary cause of fragmentation, degradation and incorporation of alterations caused by deamination resulting in C:G > T:A transitions [37]. Hence, these modifications can mislead interpretation of NGS results and potentially guide an inaccurate therapeutic course in a clinical setting. A recent study from Gao and colleagues reports a high mutational concordance comparing FF and FFPE in an NGS multi-gene panel [38]. However, some mutations are introduced due to the higher level of false-positive variants. Kerick and collaborators further reported that one strategy to deal with FFPE artefacts is to increase the sequencing depth [39]. Hence, the increment of depth decreases the number of false positives and false negatives. More stringent alignment filtering is another option, accommodating removal of putative false positives calls with lower quality scores, but this approach will on the other hand compromise read depth.

5.1.2. Liquid Biopsies

Liquid biopsy is another preparation used for clinical NGS. Whole blood collection is a non-invasive procedure and has been used for supporting diagnostics and/or monitoring circulating biomarkers. For instance; cancer antigen 125 (CA-125) for ovarian cancer [40]; CA-11-19 for colorectal cancer [41], CA-19-9 for lung cancer [42], and CA-15-3 for breast cancer [43], are well established markers in clinical practice. Circulating tumor DNA (ctDNA), tumor-derived cell-free DNA, may be promising in diagnosis of cancer and/or monitoring of relapse or progression [44]. Hence, also applicable for guiding therapeutic and monitoring in patients with known cancer. Elevated levels of ctDNA is present in plasma of cancer patients [45]. Nonetheless, the amount is still only a fraction in the pool of circulating cell-free DNA, challenging the current utility of ctDNA as a biomarker. Noteworthy, detection of somatic mutations in ctDNA for application of diagnosis of cancer, supportive guidance for optimal treatment strategy, and for surveillance of progression or recurrence is extensively under investigation [45]. ctDNA enters the plasma due to apoptosis and/or necrosis. A hallmark of apoptosis is the cleavage of DNA orchestrated by activated caspase activity. A study from Mouliere and collaborators, observed that the fragment size across 18 cancer types showed enrichment of ctDNA in lengths shorter than 167 bp and a notably enrichment ranging from fragment from 250 bp to 320 bp in size [46]. Analysis of ctDNA requires highly sensitive techniques for its detection an enrichment owing to the relative low fraction of tumor DNA dispersed within background levels of normal circulating free DNA [47]. The examination of ctDNA from liquid biopsies may be an alternative in the management of metastatic cancers, where no tumor tissue can be obtained.

5.2. Homopolymers, Repetitive Regions and Pseudogenes

The genome harbors areas that are difficult to interpret, due to the presence of homopolymers, G/C rich regions, repetitive regions and pseudogenes [48]. This results in substantial differences concerning sequencing depth and the uniformity in sequencing coverage, making these regions difficult for alignment and variant calling [49]. For instance, homopolymer regions are a challenge for the Ion Torrent sequencing platform and it introduces systematic errors due to loss of resolution above 6 nucleotides, for which may cause mis-alignments [21]. Regions with increased G/C content can be lost and subsequently

often observed as higher background due to their ability to form secondary structures [48], thus affecting the uniformity of sequencing coverage in these regions. Repetitive regions are widespread in the genome and encode tandemly repeated or close-identical sequences of variable length, often located in regions of introns [50]. Thus, being a hotspot-entry for genomic rearrangement. The concern regarding repetitive regions is to deal with alignment uncertainties due to reads that subsequently align to multiple regions, instead to an unique location. The multi-alignment reads are an obstacle that may affect variant calling as it can originate from multiple sites. Another type of structural feature embedded in the genome is pseudogenes derived from gene duplications. Pseudogenes are sequences that resembles their protein-coding counterparts with high similarity. Those are however, non-functional due to impairing mutations [51]. Albeit non-functional, some pseudogenes are transcriptionally active and act to regulate their parent protein-coding gene through the microRNA pathway [52,53]. Reads from a desired region of interest might have decreased mapping-quality, due to the presence of a pseudogene homolog causing reads to be misaligned to the pseudogenes or vice versa. Some clinically relevant genes may encounter pseudogenes, such as *KRAS* [53] for colorectal cancer and *BRCA1* [54] for ovarian cancer and breast cancer. A targeted strategy is therefore required to avoid interference from their pseudogenes that might challenge the interpretation during NGS analyses.

5.3. Bioinformatics

A large number of bioinformatic tools have become available in recent years with the aim to navigate and handle the large quantity of raw data generated by the NGS technology [55].

Raw reads from NGS platforms undergo several bioinformatic processes including base calling; quality check; adaptor trimming; read alignment and post processing; variant calling; and finally, variant annotation for functional interpretation of results. An overall of the bioinformatic pipeline available for analyzing NGS is shown in Figure 1. The two most prominent bioinformatic processes that might influence the final interpretation are the tools used for alignment and variant calling [6,7]. Both are numerous and diverse in their underlying algorithms as their original design can often be reserved for a specific purpose, such as WGS, WES, or TES/hot-spot [56]. The challenge associated with artefacts from the material used, library preparation, sequencing technologies and regions selected for sequencing, all underscore the importance of selecting appropriate benchmark tools for specific aims. The different structural peculiarities of the four groups of genomic alterations (SNV, indels, CNV, SV), excludes the possibility of one versatile tool for identifying all variants within the four groups [48]. Improper alignment to the reference genome can significantly constitute discovery of false positive and/or exclusion of disease relevant variants in downstream analyses. During the years, identification of discordance between aligned reads and the reference genome has greatly improved due to progression of variant callers and their ability to handle large amounts of data [57]. However, calling SNVs still remains a challenge after all, as various tools can result in a divergent outcome [6].

5.3.1. Alignment

Bioinformatic approaches are centered around alignment. Variants refer to identifying deviations from a non-cancerous normal reference genome. Alignment to a reference genome is a prerequisite for optimal analysis of NGS data. In that regard, for the human genome there are essentially two main reference builds currently employed: (a) The Genome Reference Consortium Human build 37 (GRCh37 or hg19), published in 2009; and (b) GRCh38 (hg38), released in 2013. The latter is built on data from many donors, subsequently altering 8000 SNV, correction of misassembled hard accessible region, filled-in gaps and added sequences for centromeres [58]. The GRCh38 improvements over GRCh37 have been reported by Pan et al., to give a more accurate genomic analysis results [59]. Additional studies from Guo et al. and Kumaran et al., examined 30 WES data sets examined WES data from NA12878 [6,58], respectively. In both studies they concluded

better accuracy and performance using hg38 as reference genome. It has been reported that aligners can affect the variant calling, when dealing with low quality base scores [57,60].

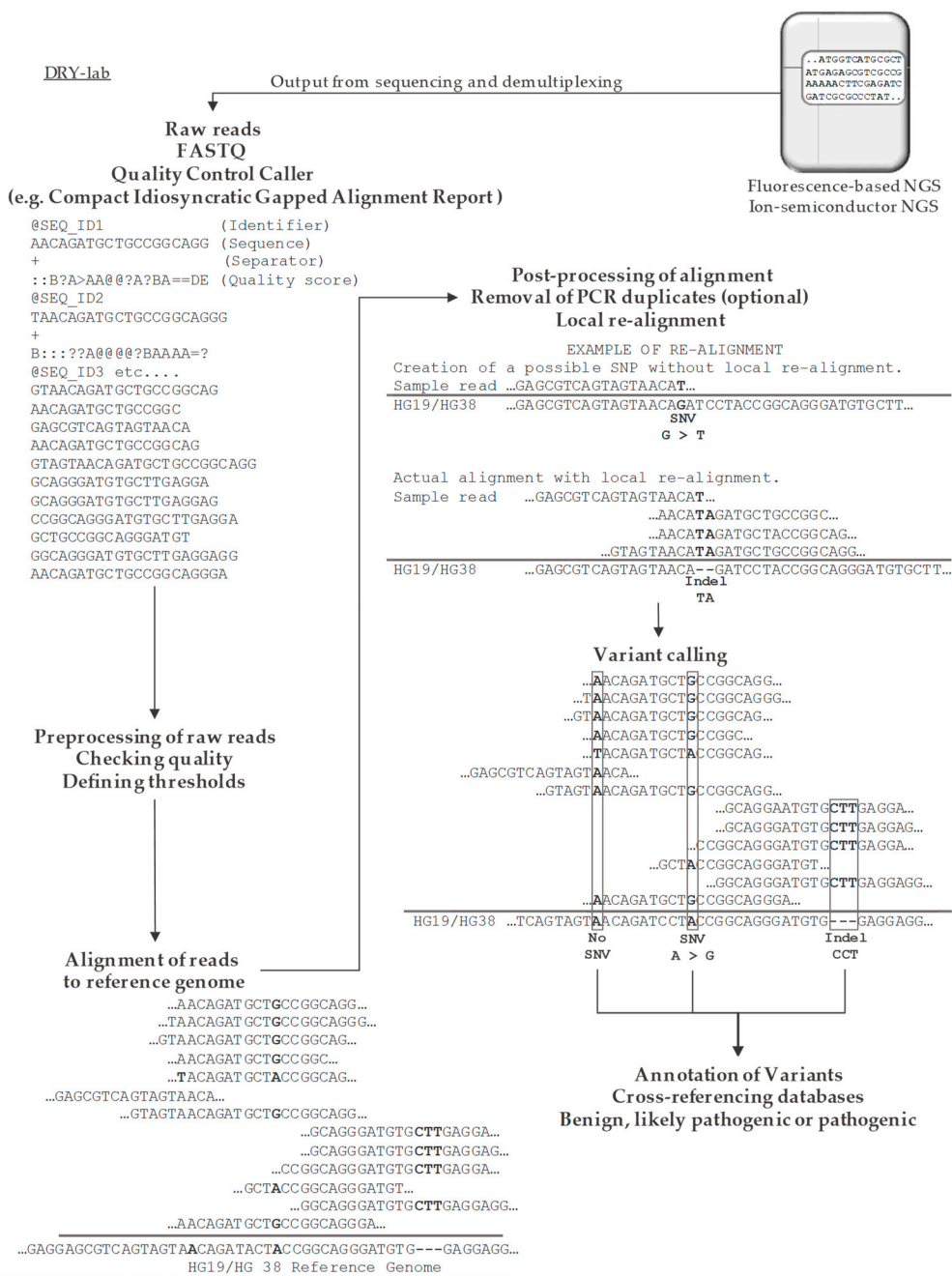


Figure 1. Schematic representation of the overall steps in the workflow of NGS analysis.

A list of alignment tools is presented in Table 3. The alignment tools Bowtie2 [2] and BWA [3] are known as the two most efficient aligners to date, with most studies using BWA-MEM (maximal exact matches) as the preferred alignment tool [6,57,59,61–67]. A study from Yu and collaborators investigated the three full-text index in minute space (FM)-indexing aligners (Bowtie2, BWA-MEM, SOAPv2) and one hash-table algorithm aligner (Novoalign) [60]. Preconditions with relatively good base-quality showed similar performance on alignment [60]. However, removal of low quality base improved the alignment performance for Novoalign [60]. The quality of called bases might significantly

impact the performance of aligners, where assessment of Phred scores can support choosing the best fitting alignment tool (FM-index- or Hash-table-based algorithm).

Briefly explained, FM-index alignment tools are derived from the Burrows-Wheeler Transform [68]—a method to sufficiently compress large amount of data and finding approximate matches of sequences in the reference genome [69]. Hash table-based aligners uses the seed-and-extend method in combination with additional alignment algorithms [68,70,71].

Table 3. Short read alignment tools.

Alignment Tools	Model	Latest Version	Ref.
Bowtie2	FM-index	v2.4.2	[2]
BWA-MEM	FM-index	v0.7.17	[3]
CUSHAW3	FM-index	v3.0.3	[72]
GSNAP	Hash-table	N/A	[73]
ISAAC	FM-index	v4	[74]
MOSAIK	Hash-table	v2.6.0	[75]
Novoalign	Hash-table	v4.03.01	http://www.novocraft.com/ *
SOAPv2	FM-index	v2.20	[76]

* accessed date: 6 April 2021.

Thus far, short-read alignment tools are commonly challenged by encountering reads that maps to multiple locations in the reference genome [50]. Hence, 3 strategies are proposed to deal with multi-reads [50]. First being to discard all multi-reads leaving only unique mapped reads to be processed. However, this strategy would consequently leave out reads with repetitive regions and gene-families, putatively harboring significance. Second is the best matching strategy, reporting reads to the location(s) with the smallest number of mismatches. The third being to report all reads and their location(s) up to a desired threshold.

5.3.2. Variant Calling for SNV and Small Indels

Many tools covering SNV detection have likewise been developed. A list of supported variant-calling tools is presented in Table 4. Consequently, their underlying algorithm of error models and assumptions for identifying mutations result in diverse variant calling across tools [67]. Hence, the methods used for variant calling are an important factor that influence mutational calling when aiming at high sensitivity and specificity. Studies subjecting sample NA12878 [6,57,59,61–67], shows that GATK-Haplotype Caller (GATK-HC), Mutect2, SAMtools and Strelka2 are among the best performing variant callers for identifying SNV and small indels [6,57,59,61–67]. The study of Chen et al., additionally revealed that Strelka2 showed better variant calling and sensitivity with a mutation frequency of $\geq 20\%$ whereas Mutect2 performed better at $\leq 10\%$ [62]. Strelka2 was developed to be fast and accurate in calling somatic variations [5]. A fast resolution time encounter an important aspect when employed in clinical oncology. Strelka2 showed to be 18–22 times faster than Mutect2 when processing 100–800x WES samples [62]. Nevertheless, a comprehensive comparison evaluating GATK-HC, Mutect2, SAMtools and Strelka2 against each other remains to be elucidated. An overview of tools and their combination in these studies are shown in Table 5.

Table 4. Variant calling tools.

Variant Calling Tools	Variant Detection	Latest Version	Ref.
Atlas2 suite	SNV, indels	v1.4.1	[77]
CONTRA	CNV, SV	v2.0.8	[78]
CNVnator	CNV, SV	v0.4	[79]
CoNVEX	CNV, SV	N/A	[80]
DeepVariant	SNV, indels	v1.0	[81]
DELLY	CNV, SV	v0.8.7	[82]
ExomeCNV	CNV, SV	v1.4	[83]
FreeBayes	SNV, indels	v1.3.4	[84]
GATK Haplotype Caller (GATK-HC)	SNV, indels	v4.1.9.0	[85,86]
Glfsingle	SNV, indels	N/A	N/A
ISAAC Variant Caller (IVC)	SNV, indels	V2.0.13	[74]
LUMPY	CNV, SV	v0.3.1	[87]
Magnolya	CNV, SV	v0.15	[88]
Mutect	SNV, indels	v1.1.5	[89]
Mutect2	SNV, indels	v4.1.9.0	[4]
Pindel	CNV, SV	N/A	[90]
Platypus	SNV, indels, SV	N/A	[91]
SAMtools	SNV, indels	v1.11	[92]
SNPSVM	SNV	N/A	[93]
SomaticSniper	SNV	v1.0.5.0	[94]
SpeedSeq	SNV, indels	v0.1.2	[95]
Strelka	SNV, indels	N/A	[96]
Strelka2	SNV, indels	v2.9.10	[5]
SVMerge	CNV, SV	v1.2	[97]
Torrent Variant Caller (TVC)	SNV, indels, SV	v5.12.0	N/A
Ulysses	CNV, SV	v1.0	[98]
Varscan2	SNV, Indel	v2.4.4	[99]

Table 5. Overview of alignment tools and variant calling tools in research papers subjecting sample NA12878.

Research Paper	Subjected Sample	Reference Genome	Alignment Tool	Variant Calling Tool
Chen et al. 2019 [61]	NA12878	WES WGS	BWA-MEM	GATK-HC SAMtools Strelka2 VarScan2
Chen et al. 2020 [62]		WES	BWA-MEM	Mutect2 Strelka2
Cornish et al. 2015 [57]		WES	Bowtie2 BWA_MEM CUSHAW3 MOSAİK Novoalign	SAMtools SNPSVM

Table 5. Cont.

Research Paper	Subjected Sample	Reference Genome	Alignment Tool	Variant Calling Tool
Hwang et al. 2015 [64]		WES WGS	Bowtie2 BWA-MEM Novoalign	FreeBayes GATK-HC SAMtools TVC
Hwang et al. 2019 [63]		WES WGS	Bowtie2 BWA-MEM GSNAP ISAAC Novoalign SOAP2	Atlas2 FreeBayes GATK-HC glfSingle IVC Platypus SAMtools suite VarScan2
Kumaran et al. 2019 [6]		WES	Bowtie2 BWA-MEM MOSAIC Novoalign SOAP2	GATK-HC DeepVariant FreeBayes SAMtools
Meng et al. 2018 [65]		TES WES WGS	BWA-MEM	DeepVariant Lancet Strelka2 VarScan2
Pan et al. 2019 [59]		WGS	Bowtie2 BWA-MEM ISAAC Novoalign	FreeBayes GATK-HC IVC SAMtools
Supernat et al. 2018 [66]		WES WGS	BWA-MEM	DeepVariant GATK-HC SpeedSeq
Xu et al. 2014 [67]		WES	BWA-MEM	Mutect SomaticSniper Strelka VarScan2

5.3.3. Variant Calling for CNV and SV

Both CNVs and SVs are commonly found associated with cancer incidences [100], [101]. CNVs covers somatic structural changes of amplification and/or deletion of DNA regions in a chromosomal region [48]. *ERBB2* (*HER2*) is an example of a gene often associated with increased copy-number in breast cancer and clinical relevance for detection [102], whilst *TP53* variants are often observed as loss of the wildtype allele [103]. SVs covers structural changes in terms of large translocations and chromosomal rearrangements [48]. Creation of known and novel tumorigenic fusion proteins as well as de-positioning the proximity of regulatory elements for mRNA transcription might indirectly affect cell function contributing to cancer [104].

Tools for identifying SNVs and small indels are not suited for calling variants of CNVs and SVs. A number of tools exists for calling CNV and SV, as shown in Table 4. However, certain challenges are peculiar to those chromosomal changes. Moreover, the technological limitations of short reads generated by NGS (~150 bp) is not sufficient to resolve long insertions and duplicated regions [105]. Thus, the ongoing development of TGS technologies will contribute to unravel CNVs and SVs more accurately. Implementation of the variant allele frequency may benefit to provide hints of CNV and SVs as the variant allele frequency will increase or decrease with the number of copies [106].

The variant calling for CNV and SV harbors different strategies to identify modifications including read-pairing, read-depth, split-read and read-assembly. Read-pairing is the detection of which read pairs are aligned with increased or decreased distance and/or orientations [105]. This method is largely dependent on the insertion size, as small insertion can be ignored or missed by the algorithm [107].

The read-depth method assumes a Poisson distribution in the depth of aligned reads and examines the distribution of reads to reveal duplications and/or deletions. Thus, duplicated regions show increased read depth whereas deleted regions show decreased read depth compared to normal diploid regions [108]. Mapped reads with low confidence in regions of repetitive DNA challenges the accuracy of copy-number status and may introduce a biased output [107].

The split-read approach utilizes read pairs to define breakpoints of structural variants. The concept of the method relies on reads that align with high confidence to the reference. Hence, the unaligned read(s) may potentially define the breaking point of the insertion [109]. However, the split-reads approach shows limitation, as reads below 1000 bases affect both sensitivity and specificity [109]. Finally, the read-assembly method is in theory the most versatile approach for identifying variants. As the method suggests, it is based on assembling a read base scaffold genome that is subsequently compared to the reference genome to identify variants [109]. Nevertheless, the method requires a significant demand on computational resources and longer-length reads, hence it is not advantageous used for the detection of CNV and SV, yet.

6. Clinical Demand

NGS assistance to guide diagnostics, prognostics and improve precision medicine are being progressively adopted into the clinic. Furthermore, genomic research has become an area of impact to prioritize mutations for functional testing. Thus, contributing to reveal mechanisms and better understanding of cancer, allowing for the development of new targeted drugs. In clinical oncology, attention to specific genes/hot spots are used for treatment decision. A PCR amplicon-based enrichment strategy underlying gene panels has several benefits for clinical utility. Here, it can be mentioned its low requirement of input DNA, fast resolution time, application to FFPE and the ability to reach greater sequencing depth [62,67,110]. In addition, this strategy can handle multiple samples (patients) in one workflow. Furthermore, assumptions of primary tumor being homogenous holds little promise, as it has been shown that a primary tumor often harbors subclones of heterogenous and/or evolutionary origin [111]. Therefore, close collaboration with pathologist is important for obtaining the right tissue for NGS analysis. Reaching greater sequencing depth allows to potentially explore low frequent mutations in low tumor cellularity and/or in subclones. This identification might contribute to the refinement of diagnosis, clinical management and/or prognosis, owing to knowledge of drug-resistance before initial therapy [112]. A critical element of variant detection is the accuracy and reproducibility of the identified variants called. Hence, evaluation and validation of tools/pipelines must be compared to clear cut variants from previously well-defined samples [106].

The rate of false positives can be handled both by the specimen examined and the filters applied during bioinformatic analyses, minimizing its effects. FFPE samples harbor thousands of artifacts [113], which may remove low-frequency true variants during filtering. It is therefore crucial to consider the type of specimen, in order to deploy more accurate filters and the importance of validating the pipeline. Many studies have subjected aligners and variant callers concerning their performance, concluding limited concordance [6,57]. Filters embedded in aligners focusing on mismatches can be adjusted to allow or exclude fewer or more mismatches during alignment. With a greater number of mismatches, it also increases the likelihood of DNA fragment to align to regions with similarity. Thus, the increment of false-positive findings throughout variant calling. On the other hand, narrowing alignment filters to only accept few mismatches will potentially leave out true-

positives variants with greater amounts of mismatches. Hard filters can be adjusted in most alignment and variant calling -tools to deal with difficult regions (homopolymeric and repetitive-regions). This is solved by either completely rejecting variants in these areas or apply manual empirical filters such as thresholds of coverage, Phred-score and p -values [114]. It is important to note that increasing or decreasing filters identifying less or additional variants in a clinical setting is not necessarily beneficial for the patient. If the identified additional variants are artifacts in relevant genes, then these could potentially lead to misguidance of therapeutic course. Moreover, an extra challenge in the clinical setting is the handling of germline mutations. Tumor samples are usually examined without a germline counterpart (i.e., normal tissue). Hence, by applying normal-like sample within the same patient, may help to reduce the number of germline variants.

Special caution must be taken setting up and interpreting NGS analysis from clinical data, as many factors can interplay with the outcome. Many answers can be intriguing, but the right answer is the one beneficial for the patient outcome. Hence, critical validation of pipelines for clinical utility is of utmost importance.

7. Conclusions

The advent of NGS greatly improved the study of genetics, as well as the diagnosis and treatment of genetic diseases. However, with the ability to sequence billions of reactions in a short period of time creates a demand for analytic tools to overcome these large data sets generated. Robust pipelines for NGS analysis are in constant demand, thus alignment tools and variant calling tools are still improving and are an active area of research. From the literature it has been reported that the combination of alignment tools and variant calling tools tends to vary in consistency. Interestingly, we find that from 10 studies subjecting sample NA12878, BWA-MEM or Bowtie2 in combination with Strelka2 or Mutect2 are among the best performing pipelines for SNV detection. CNV and SV detection are challenging due to the duplications/deletions of regions within a gene and/or translocation. The read-assembly approach is promising for detecting CNV/SV. Nevertheless, that requires longer spans of reads than currently provided by the present NGS technology and extensive computational resources to function ideally. The ongoing improvements to decrease the high error-rate underlying the TGS-technologies so far, will preferably solve the problem with longer reads.

The NGS technologies are increasingly being implemented into diagnostic routine settings, along with a diversity of bioinformatic. Different specimens are subjected for sequencing, according to their purpose and origin. To provide an optimal answer in patient's course of disease, precise annotation of mutations is mandatory. Hence, the prerequisite of evaluating and validating bioinformatic pipelines used for the analysis. Laboratories conducting NGS should as a minimum participate in quality trials as documentation for applied competence conducting NGS analysis. Furthermore, consideration concerning the usage of NGS in relation to the timepoint of treatment should be taken into account.

Evidence-based biological treatment can optimally be supported by using panel sequencing (TES and/or hot-spot) ensuring fast throughput, focused datamining and high sensitivity and specificity. In contrast to WES, that can be used for exploring and prioritizing new relevant drug targets across disease used for experimental treatment of patients. However, Opposite, WGS contributes to large amounts of data, whereas 99% is information about non-coding regions. Due to the large amount of data in WGS results may be presented with low sequencing depth. Hence, the risk of not identifying relevant actionable clinical targets. Although, WGS is still beneficial of gaining new knowledge from research studies, that might benefit future patient.

Author Contributions: L.K.V. contributed to the conceptualization and writing (original draft and figure preparation, review and editing), D.N.P.O. contributed to the conceptualization and writing (review and editing), C.K.H. contributed to the conceptualization and writing (review and editing), E.V.H. contributed to the conceptualization and writing (review and editing). 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.

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

References

1. Metzker, M.L. Sequencing technologies—The next generation. *Nat. Rev. Genet.* **2010**, *11*, 31–46. [[CrossRef](#)] [[PubMed](#)]
2. Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **2012**, *9*, 357–359. [[CrossRef](#)]
3. Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **2009**, *25*, 1754–1760. [[CrossRef](#)] [[PubMed](#)]
4. Benjamin, D.; Sato, T.; Cibulskis, K.; Getz, G.; Stewart, C.; Lichtenstein, L. Calling Somatic SNVs and Indels with Mutect2. *bioRxiv* **2019**, 861054. [[CrossRef](#)]
5. 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](#)] [[PubMed](#)]
6. Kumaran, M.; Subramanian, U.; Devarajan, B. Performance assessment of variant calling pipelines using human whole exome sequencing and simulated data. *BMC Bioinform.* **2019**, *20*, 342. [[CrossRef](#)] [[PubMed](#)]
7. Liu, X.; Han, S.; Wang, Z.; Gelernter, J.; Yang, B.Z. Variant Callers for Next-Generation Sequencing Data: A Comparison Study. *PLoS ONE* **2013**, *8*, e75619. [[CrossRef](#)]
8. Maxam, A.M.; Gilbert, W. A new method for sequencing DNA. *Proc. Natl. Acad. Sci. USA* **1977**, *74*, 560–564. [[CrossRef](#)] [[PubMed](#)]
9. Sanger, F.; Nicklen, S.; Coulson, A.R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* **1977**, *74*, 5463–5467. [[CrossRef](#)]
10. Olson, M.V. The human genome project. *Proc. Natl. Acad. Sci. USA* **1993**, *90*, 4338–4344. [[CrossRef](#)]
11. Heather, J.M.; Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* **2016**, *107*, 1–8. [[CrossRef](#)] [[PubMed](#)]
12. Slatko, B.E.; Gardner, A.F.; Ausubel, F.M. Overview of Next-Generation Sequencing Technologies. *Curr. Protoc. Mol. Biol.* **2018**, *122*, e59. [[CrossRef](#)]
13. Xiao, T.; Zhou, W. The third generation sequencing: The advanced approach to genetic diseases. *Transl. Pediatr.* **2020**, *9*, 163–173. [[CrossRef](#)]
14. Pertea, M.; Shumate, A.; Pertea, G.; Varabyou, A.; Breitwieser, F.P.; Chang, Y.; Madugundu, A.K.; Pandey, A.; Salzberg, L.S. CHES: A new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biol.* **2018**, *19*, 208. [[CrossRef](#)]
15. Pon, J.R.; Marra, M.A. Driver and passenger mutations in cancer. *Annu. Rev. Pathol. Mech. Dis.* **2015**, *10*, 25–50. [[CrossRef](#)]
16. McFarland, C.D.; Mirny, L.A.; Korolev, K.S. Tug-of-war between driver and passenger mutations in cancer and other adaptive processes. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 15138–15143. [[CrossRef](#)]
17. Futreal, P.A.; Coin, L.; Marshall, M.; Down, T.; Hubbard, T.; Wooster, R.; Rahman, N.; Stratton, M.R. A census of human cancer genes. *Nat. Rev. Cancer* **2004**, *4*, 177–183. [[CrossRef](#)] [[PubMed](#)]
18. Dietlein, F.; Weghorn, D.; Taylor-Weiner, A.; Richters, A.; Reardon, B.; Liu, D.; Lander, E.S.; Van Allen, E.M.; Sunyaev, S.R. Identification of cancer driver genes based on nucleotide context. *Nat. Genet.* **2020**, *52*, 208–218. [[CrossRef](#)]
19. Bailey, M.H.; Tokheim, C.; Porta-Pardo, E.; Sengupta, S.; Bertrand, D.; Weerasinghe, A.; Colaprico, A.; Wendl, M.C.; Kim, J.; Reardon, B.; et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **2018**, *173*, 371–385. [[CrossRef](#)] [[PubMed](#)]
20. Gonzalez-Perez, A.; Mustonen, V.; Reva, B.; Ritchie, G.R.S.; Creixell, P.; Karchin, R.; Vazquez, M.; Fink, J.L.; Kassahn, K.S.; Pearson, J.V.; et al. Computational approaches to identify functional genetic variants in cancer genomes. *Nat. Methods* **2013**, *10*, 723–729. [[CrossRef](#)]
21. Kumar, K.R.; Cowley, M.J.; Davis, R.L. Next-Generation Sequencing and Emerging Technologies. *Semin. Thromb. Hemost.* **2019**, *45*, 661–673. [[CrossRef](#)]
22. Shin, S.H.; Bode, A.M.; Dong, Z. Precision medicine: The foundation of future cancer therapeutics. *Npj Precis. Oncol.* **2017**, *1*, 12. [[CrossRef](#)] [[PubMed](#)]
23. Cortez, A.J.; Tudrej, P.; Kujawa, K.A.; Lisowska, K.M. Advances in ovarian cancer therapy. *Cancer Chemother. Pharmacol.* **2018**, *81*, 17–38. [[CrossRef](#)]
24. Caulfield, S.E.; Davis, C.C.; Byers, K.F. Olaparib: A Novel Therapy for Metastatic Breast Cancer in Patients With a BRCA1/2 Mutation. *J. Adv. Pract. Oncol.* **2019**, *10*, 167–174. [[CrossRef](#)] [[PubMed](#)]
25. DeMatteo, R.P.; Ballman, K.V.; Antonescu, C.R.; Make, R.G.; Pisters, P.W.T.; Demetri, G.D.; Blackstein, M.E.; Blanke, C.D.; Von Mehren, M.; Brennan, M.F.; et al. Adjuvant imatinib mesylate after resection of localised, primary gastrointestinal stromal tumour: A randomised, double-blind, placebo-controlled trial. *Lancet* **2009**, *373*, 1097–1104. [[CrossRef](#)]

26. Chapman, P.B.; Hauschild, A.; Robert, C.; Haanen, J.B.; Ascierto, P.; Larkin, J.; Dummer, R.; Garbe, C.; Testori, A.; Maio, M.; et al. Improved Survival with Vemurafenib in Melanoma with BRAF V600E Mutation. *N. Engl. J. Med.* **2011**, *364*, 2507–2516. [[CrossRef](#)] [[PubMed](#)]
27. Dinu, D.; Dobre, M.; Panaitescu, E.; Bîrlă, R.; Iosif, C.; Hoara, P.; Caragui, A.; Boeriu, M.; Constantinoiu, S.; Ardeleanu, C. Prognostic significance of KRAS gene mutations in colorectal cancer—Preliminary study. *J. Med. Life* **2014**, *7*, 581–587.
28. Maus, M.K.H.; Grimminger, P.P.; Mack, P.C.; Astrow, S.H.; Stephens, C.; Zeger, G.; Hsiang, J.; Brabender, J.; Friedrich, M.; Alakus, H.; et al. KRAS mutations in non-small-cell lung cancer and colorectal cancer: Implications for EGFR-targeted therapies. *Lung Cancer* **2014**, *83*, 163–167. [[CrossRef](#)]
29. Forbes, S.A.; Beare, D.; Boutselakis, H.; Bamford, S.; Bindal, N.; Tate, J.; Cole, C.G.; Ward, S.; Dawson, E.; Ponting, L.; et al. COSMIC: Somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **2017**, *45*, 777–783. [[CrossRef](#)]
30. Tate, J.G.; Bamford, S.; Jubb, H.C.; Sondka, Z.; Beare, D.M.; Bindal, N.; Boutselakis, H.; Cole, C.G.; Creatore, C.; Dawson, E.; et al. COSMIC: The Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **2019**, *47*, 941–947. [[CrossRef](#)]
31. Landrum, M.J.; Lee, J.M.; Riley, G.R.; Jang, W.; Rubinstein, W.S.; Church, D.M.; Maglott, D.R. ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **2014**, *42*, 980–985. [[CrossRef](#)]
32. 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*, 1062–1067. [[CrossRef](#)]
33. 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*, 1–16. [[CrossRef](#)]
34. Bewicke-Copley, F.; Arjun Kumar, 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](#)] [[PubMed](#)]
35. Griffith, M.; Miller, C.A.; Griffith, O.L.; Krysiak, K.; Skidmore, Z.L.; Ramu, A.; Walker, J.R.; Dang, H.X.; Trani, L.; Larson, D.E.; et al. Optimizing Cancer Genome Sequencing and Analysis. *Cell Syst.* **2015**, *1*, 210–223. [[CrossRef](#)] [[PubMed](#)]
36. Williams, C.; Pontén, F.; Moberg, C.; Söderkvist, P.; Uhlén, M.; Pontén, J.; Sitbon, G.; Lundeberg, J. A high frequency of sequence alterations is due to formalin fixation of archival specimens. *Am. J. Pathol.* **1999**, *155*, 1467–1471. [[CrossRef](#)]
37. Kim, S.; Park, C.; Ji, Y.; Kim, D.G.; Bae, H.; Vrancken, M.; Kim, D.; Kim, K. Deamination Effects in Formalin-Fixed, Paraffin-Embedded Tissue Samples in the Era of Precision Medicine. *J. Mol. Diagn.* **2017**, *19*, 137–146. [[CrossRef](#)] [[PubMed](#)]
38. Gao, X.H.; Li, J.; Gong, H.F.; Yu, G.Y.; Liu, P.; Hao, L.Q.; Liu, L.J.; Bai, C.G.; Zhang, W. Comparison of Fresh Frozen Tissue With Formalin-Fixed Paraffin-Embedded Tissue for Mutation Analysis Using a Multi-Gene Panel in Patients with Colorectal Cancer. *Front. Oncol.* **2020**, *10*, 1–8. [[CrossRef](#)] [[PubMed](#)]
39. Kerick, M.; Isau, M.; Timmermann, B.; Sültmann, H.; Herwig, R.; Krobtsch, S.; Schaefer, G.; Verdorfer, I.; Bartsch, G.; Klocker, H.; et al. Targeted high throughput sequencing in clinical cancer Settings: Formaldehyde fixed-paraffin embedded (FFPE) tumor tissues, input amount and tumor heterogeneity. *BMC Med. Genom.* **2011**, *4*, 1–13. [[CrossRef](#)] [[PubMed](#)]
40. Bast, R.C.; Feeney, M.; Lazarus, H.; Nadler, L.M.; Colvin, R.B.; Knapp, R.C. Reactivity of a Monoclonal Antibody with Human Ovarian Carcinoma. *J. Clin. Investig.* **1981**, *68*, 1331–1337. [[CrossRef](#)]
41. Loktionov, A. Biomarkers for detecting colorectal cancer non-invasively: DNA, RNA or proteins. *World J. Gastrointest. Oncol.* **2020**, *12*, 124–128. [[CrossRef](#)]
42. Mamdani, H.; Ahmed, S.; Armstrong, S.; Mok, T.; Jalal, S.I. Blood-based tumor biomarkers in lung cancer for detection and treatment. *Transl. Lung Cancer Res.* **2017**, *6*, 648–660. [[CrossRef](#)]
43. Oloomi, M.; Moazzezy, N.; Bouzari, S. Comparing blood versus tissue-based biomarkers expression in breast cancer patients. *Heliyon* **2020**, *6*, 1–7. [[CrossRef](#)]
44. Dagogo-Jack, I.; Shaw, A.T. Tumour heterogeneity and resistance to cancer therapies. *Nat. Rev. Clin. Oncol.* **2018**, *15*, 81–94. [[CrossRef](#)] [[PubMed](#)]
45. Keller, L.; Belloum, Y.; Wikman, H.; Pantel, K. Clinical relevance of blood-based ctDNA analysis: Mutation detection and beyond. *Br. J. Cancer* **2020**, *6*, 1–14. [[CrossRef](#)] [[PubMed](#)]
46. Moulriere, F.; Chandrananda, D.; Piskorz, A.M.; Moore, E.K.; Morris, J.; Ahlborn, L.B.; Mair, R.; Goranova, T.; Marass, F.; Heider, K.; et al. Enhanced detection of circulating tumor DNA by fragment size analysis. *Sci. Transl. Med.* **2018**, *10*, 1–14. [[CrossRef](#)]
47. Ignatiadis, M.; Dawson, S.J. Circulating tumor cells and circulating tumor DNA for precision medicine: Dream or reality? *Ann. Oncol.* **2014**, *25*, 2304–2313. [[CrossRef](#)]
48. Jennings, L.J.; Arcila, M.E.; Corless, C.; Kamel-Reid, S.; Lubin, I.M.; Pfeifer, J.; Temple-Smolkin, R.L.; Voelkerding, K.V.; Nikiforova, M.N. Guidelines for Validation of Next-Generation Sequencing–Based Oncology Panels: A Joint Consensus Recommendation of the Association for Molecular Pathology and College of American Pathologists. *J. Mol. Diagn.* **2017**, *19*, 341–365. [[CrossRef](#)] [[PubMed](#)]
49. 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](#)]
50. Treangen, T.J.; Salzberg, S.L. Repetitive DNA and next-generation sequencing: Computational challenges and solutions. *Nat. Rev. Genet.* **2012**, *13*, 36–46. [[CrossRef](#)]
51. Lee, C.Y.; Yen, H.Y.; Zhong, A.W.; Gao, H. Resolving misalignment interference for NGS-based clinical diagnostics. *Hum. Genet.* **2020**, *9*, 1–16. [[CrossRef](#)] [[PubMed](#)]

52. Pink, R.C.; Wicks, K.; Caley, D.P.; Punch, E.K.; Jacobs, L.; Carter, D.R.F. Pseudogenes: Pseudo-functional or key regulators in health and disease? *RNA* **2011**, *11*, 792–798. [[CrossRef](#)]
53. Han, L.; Yuan, Y.; Zheng, S.; Yang, Y.; Li, J.; Edgerton, M.E.; Diao, L.; Xu, Y.; Verhaak, R.G.W.; Liang, H. The Pan-Cancer analysis of pseudogene expression reveals biologically and clinically relevant tumour subtypes. *Nat. Commun.* **2014**, *5*, 1–9. [[CrossRef](#)] [[PubMed](#)]
54. Puget, N.; Gad, S.; Perrin-Vidoz, L.; Sinilnikova, O.M.; Stoppa-Lyonnet, D.; Lenoir, G.M.; Mazoyer, S. Distinct BRCA1 rearrangements involving the BRCA1 pseudogene suggest the existence of a recombination hot spot. *Am. J. Hum. Genet.* **2002**, *70*, 858–865. [[CrossRef](#)] [[PubMed](#)]
55. 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](#)] [[PubMed](#)]
56. Meynert, A.M.; Ansari, M.; FitzPatrick, D.R.; Taylor, M.S. Variant detection sensitivity and biases in whole genome and exome sequencing. *BMC Bioinform.* **2014**, *15*, 1–11. [[CrossRef](#)] [[PubMed](#)]
57. Cornish, A.; Guda, C. A Comparison of Variant Calling Pipelines Using Genome in a Bottle as a Reference. *BioMed Res. Int.* **2015**, *456479*, 1–11. [[CrossRef](#)]
58. Guo, Y.; Dai, Y.; Yu, H.; Zhao, S.; Samuels, D.C.; Shyr, Y. Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. *Genomics* **2017**, *109*, 83–90. [[CrossRef](#)]
59. Pan, B.; Kusko, R.; Xiao, W.; Zheng, Y.; Liu, Z.; Xiao, C.; Sakkiah, S.; Guo, W.; Gong, P.; Zhang, C.; et al. Similarities and differences between variants called with human reference genome HG19 or HG38. *BMC Bioinform.* **2019**, *20*, 17–29. [[CrossRef](#)] [[PubMed](#)]
60. Yu, X.; Guda, K.; Willis, J.; Veigl, M.; Wang, Z.; Markowitz, S.; Adams, M.D.; Sun, S. How do alignment programs perform on sequencing data with varying qualities and from repetitive regions? *BioData Min.* **2012**, *5*, 1–12. [[CrossRef](#)]
61. Chen, J.; Li, X.; Zhong, H.; Meng, Y.; Du, H. Systematic comparison of germline variant calling pipelines cross multiple next-generation sequencers. *Sci. Rep.* **2019**, *9*, 1–13. [[CrossRef](#)] [[PubMed](#)]
62. 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*, 1–9. [[CrossRef](#)] [[PubMed](#)]
63. Hwang, K.B.; Lee, I.; Li, H.; Won, D.; Hernandez-Ferrer, C.; Negron, J.A.; Kong, S.W. Comparative analysis of whole-genome sequencing pipelines to minimize false negative findings. *Sci. Rep.* **2019**, *9*, 1–10. [[CrossRef](#)] [[PubMed](#)]
64. Hwang, S.; Kim, E.; Lee, I.; Marcotte, E.M. Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Sci. Rep.* **2015**, *5*, 1–8. [[CrossRef](#)]
65. Meng, J.; Chen, Y.P.P. A database of simulated tumor genomes towards accurate detection of somatic small variants in cancer. *PLoS ONE* **2018**, *13*, e202982. [[CrossRef](#)]
66. Supernat, A.; Vidarsson, O.V.; Steen, V.M.; Stokowy, T. Comparison of three variant callers for human whole genome sequencing. *Sci. Rep.* **2018**, *8*, 1–6. [[CrossRef](#)]
67. Xu, H.; DiCarlo, J.; Satya, R.V.; Peng, Q.; Wang, Y. Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. *BMC Genom.* **2014**, *15*, 1–10. [[CrossRef](#)]
68. Mielczarek, M.; Szyda, J. Review of alignment and SNP calling algorithms for next-generation sequencing data. *J. Appl. Genet.* **2016**, *57*, 71–79. [[CrossRef](#)]
69. Kuhnle, A.; Mun, T.; Boucher, C.; Gagie, T.; Langmead, B.; Manzini, G. Efficient Construction of a Complete Index for Pan-Genomics Read Alignment. *J. Comput. Biol.* **2020**, *27*, 500–513. [[CrossRef](#)]
70. Lindner, R.; Friedel, C.C. A Comprehensive Evaluation of Alignment Algorithms in the Context of RNA-Seq. *PLoS ONE* **2012**, *7*, e52403. [[CrossRef](#)]
71. Zhang, H.; Chan, Y.; Fan, K.; Schmidt, B.; Liu, W. Fast and efficient short read mapping based on a succinct hash index. *BMC Bioinform.* **2018**, *19*, 1–14. [[CrossRef](#)] [[PubMed](#)]
72. Liu, Y.; Popp, B.; Schmidt, B. CUSHAW3: Sensitive and accurate base-space and color-space short-read alignment with hybrid seeding. *PLoS ONE* **2014**, *9*, e86869. [[CrossRef](#)]
73. Wu, T.D.; Nuca, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **2010**, *26*, 873–881. [[CrossRef](#)]
74. Raczy, C.; Petrovski, R.; Saunders, C.T.; Chorny, I.; Kruglyak, S.; Margulies, E.H.; Chuang, H.; Källberg, M.; Kumar, S.A.; Liao, A.; et al. Isaac: Ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. *Bioinformatics* **2013**, *29*, 2041–2043. [[CrossRef](#)]
75. Lee, W.P.; Stromberg, M.P.; Ward, A.; Stewart, C.; Garrison, E.P.; Marth, G.T. MOSAIK: A hash-based algorithm for accurate next-generation sequencing short-read mapping. *PLoS ONE* **2014**, *9*, e90581. [[CrossRef](#)]
76. Li, R.; Yu, C.; Li, Y.; Lam, T.; Yiu, S.; Kristiansen, K.; Wang, J. SOAP2: An improved ultrafast tool for short read alignment. *Bioinformatics* **2009**, *25*, 1966–1967. [[CrossRef](#)] [[PubMed](#)]
77. Challis, D.; Yu, J.; Evani, U.S.; Jackson, A.R.; Paithankar, S.; Coarfa, C.; Milosavljevic, A.; Gibbs, R.A.; Yu, F. An integrative variant analysis suite for whole exome next-generation sequencing data. *BMC Bioinform.* **2012**, *8*, 1–12. [[CrossRef](#)] [[PubMed](#)]
78. Li, J.; Lupat, R.; Amarasinghe, K.C.; Thompson, E.R.; Doyle, M.A.; Ryland, G.L.; Tothill, R.W.; Halgamuge, S.K.; Campbell, I.G.; Gorringe, K.L. CONTRA: Copy number analysis for targeted resequencing. *Bioinformatics* **2012**, *28*, 1307–1313. [[CrossRef](#)] [[PubMed](#)]

79. Abyzov, A.; Urban, A.E.; Snyder, M.; Gerstein, M. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* **2011**, *21*, 974–984. [[CrossRef](#)]
80. Amarasinghe, K.C.; Li, J.; Halgamuge, S.K. Correction to CoNVEX: Copy number variation estimation in exome sequencing data using HMM. *BMC Bioinform.* **2013**, *14*, 1–9. [[CrossRef](#)]
81. Poplin, R.; Chang, P.; 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](#)]
82. Rausch, T.; Zichner, T.; Schlattl, A.; Stütz, A.M.; Benes, V.; Korbel, J.O. DELLY: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **2012**, *28*, 333–339. [[CrossRef](#)] [[PubMed](#)]
83. Sathirapongsasuti, J.F.; Lee, H.; Horst, B.A.J.; Brunner, G.; Cochran, A.J.; Binder, S.; Quackenbush, J.; Nelson, S.F. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics* **2011**, *27*, 2648–2654. [[CrossRef](#)]
84. Garrison, E.; Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv* **2012**, arXiv:1207.3907.
85. McKenna, A.; Hanna, M.; Banks, E.; Sivachenko, A.; Cibulskis, K.; Kernysky, 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.* **2009**, *20*, 1297–1303. [[CrossRef](#)] [[PubMed](#)]
86. Ren, S.; Bertels, K.; Al Ars, Z. Efficient Acceleration of the Pair-HMMs Forward Algorithm for GATK HaplotypeCaller on Graphics Processing Units. *Evol. Bioinform.* **2018**, *14*, 1–12. [[CrossRef](#)]
87. Layer, R.M.; Chiang, C.; Quinlan, A.R.; Hall, I.M. LUMPY: A probabilistic framework for structural variant discovery. *Genome Biol.* **2014**, *15*, 1–19. [[CrossRef](#)] [[PubMed](#)]
88. Nijkamp, J.F.; Van Den Broek, M.A.; Geertman, J.M.A.; Reinders, M.J.T.; Daran, J.M.G.; De Ridder, D. De novo detection of copy number variation by co-assembly. *Bioinformatics* **2012**, *28*, 3195–3202. [[CrossRef](#)]
89. Cibulskis, K.; Lawrence, M.S.; Carter, S.L.; Sivachenko, A.; Jaffe, D.; Sougnez, C.; Gabriel, S.; Meyerson, M.; Lander, E.S.; Getz, G. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **2013**, *31*, 213–219. [[CrossRef](#)]
90. Ye, K.; Schulz, M.H.; Long, Q.; Apweiler, R.; Ning, Z. Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **2009**, *25*, 2865–2871. [[CrossRef](#)]
91. Rimmer, A.; Phan, H.; Mathieson, I.; Iqbal, Z.; Twigg, S.R.F.; Wilkie, A.O.M.; McVean, G.; Lunter, G. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* **2014**, *46*, 912–918. [[CrossRef](#)]
92. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079. [[CrossRef](#)]
93. O’Fallon, B.D.; Wooderchak-Donahue, W.; Crockett, D.K. A support vector machine for identification of single-nucleotide polymorphisms from next-generation sequencing data. *Bioinformatics* **2013**, *29*, 1361–1366. [[CrossRef](#)] [[PubMed](#)]
94. Larson, D.E.; Harris, C.C.; Chen, K.; Koboldt, D.C.; Abbott, T.E.; Dooling, D.J.; Ley, T.J.; Mardis, E.R.; Wilson, R.K.; Ding, L. Somaticsniper: Identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* **2012**, *28*, 311–317. [[CrossRef](#)] [[PubMed](#)]
95. Chiang, C.; Layer, R.M.; Faust, G.G.; Lindberg, M.R.; Rose, D.B.; Garrison, E.P.; Marth, G.T.; Quinlan, A.R.; Hall, I.M. SpeedSeq: Ultra-fast personal genome analysis and interpretation. *Nat. Methods* **2015**, *12*, 966–968. [[CrossRef](#)] [[PubMed](#)]
96. Saunders, C.T.; Wong, W.S.W.; Swamy, S.; Becq, J.; Murray, L.J.; Cheetham, R.K. Strelka: Accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* **2012**, *28*, 1811–1817. [[CrossRef](#)] [[PubMed](#)]
97. Wong, K.; Keane, T.M.; Stalker, J.; Adams, D.J. Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. *Genome Biol.* **2010**, *11*, 1–9. [[CrossRef](#)]
98. Gillet-Markowska, A.; Richard, H.; Fischer, G.; Lafontaine, I. Ulysses: Accurate detection of low-frequency structural variations in large insert-size sequencing libraries. *Bioinformatics* **2015**, *31*, 801–808. [[CrossRef](#)]
99. Koboldt, D.C.; Zhang, Q.; Larson, D.E.; Shen, D.; McLellan, M.D.; Lin, L.; Miller, C.A.; Mardis, E.R.; Ding, L.; Wilson, R.K. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **2012**, *22*, 568–576. [[CrossRef](#)]
100. Beroukhi, R.; Mermel, C.H.; Porter, D.; Wei, G.; Raychaudhuri, S.; Donovan, J.; Barretina, J.; Boehm, J.S.; Dobson, J.; Urashima, M.; et al. The landscape of somatic copy-number alteration across human cancers. *Nature* **2010**, *463*, 899–905. [[CrossRef](#)]
101. Li, Y.; Roberts, N.D.; Wala, J.A.; Shapira, O.; Schumacher, S.E.; Kumar, K.; Khurana, E.; Waszak, S.; Korbel, J.O.; Haber, J.E.; et al. Patterns of somatic structural variation in human cancer genomes. *Nature* **2020**, *578*, 112–121. [[CrossRef](#)] [[PubMed](#)]
102. Marotta, M.; Chen, X.; Inoshita, A.; Stephens, R.; Budd, G.T.; Crowe, J.P.; Lyons, J.; Kondratova, A.; Tubbs, R.; Tanaka, H. A common copy-number breakpoint of ERBB2 amplification in breast cancer colocalizes with a complex block of segmental duplications. *Breast Cancer Res.* **2012**, *14*, 1–19. [[CrossRef](#)] [[PubMed](#)]
103. Demidenko, Z.N.; Fojo, T.; Blagosklonny, M.V. Complementation of two mutant p53: Implications for loss of heterozygosity in cancer. *FEBS Lett.* **2005**, *579*, 2231–2235. [[CrossRef](#)] [[PubMed](#)]
104. Alaei-Mahabadi, B.; Bhadury, J.; Karlsson, J.W.; Nilsson, J.A.; Larsson, E. Global analysis of somatic structural genomic alterations and their impact on gene expression in diverse human cancers. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 13768–13773. [[CrossRef](#)]

105. Alkan, C.; Coe, B.P.; Eichler, E.E. Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* **2011**, *12*, 363–376. [[CrossRef](#)] [[PubMed](#)]
106. Koboldt, D.C. Best practices for variant calling in clinical sequencing. *Genome Med.* **2020**, *12*, 1–13. [[CrossRef](#)]
107. Le Scouarnec, S.; Gribble, S.M. Characterising chromosome rearrangements: Recent technical advances in molecular cytogenetics. *Heredity* **2012**, *108*, 75–85. [[CrossRef](#)]
108. Escaramís, G.; Docampo, E.; Rabionet, R. A decade of structural variants: Description, history and methods to detect structural variation. *Brief. Funct. Genom.* **2015**, *14*, 305–314. [[CrossRef](#)] [[PubMed](#)]
109. Pirooznia, M.; Goes, F.; Zandi, P.P. Whole-genome CNV analysis: Advances in computational approaches. *Front. Genet.* **2015**, *6*, 1–9. [[CrossRef](#)]
110. Quail, M.A.; Smith, M.; Coupland, P.; Otto, T.D.; Harris, S.R.; Connor, T.R.; Bertoni, A.; Swerdlow, H.P.; Gu, Y. A tale of three NGS sequencing platforms. *BMC Genom.* **2012**, *13*, 1–13. [[CrossRef](#)]
111. Chowell, D.; Napier, J.; Gupta, R.; Anderson, K.S.; Maley, C.C.; Wilson Sayres, M.A. Modeling the subclonal evolution of cancer cell populations. *Cancer Res.* **2018**, *78*, 830–839. [[CrossRef](#)]
112. Stratton, M.R.; Campbell, P.J.; Futreal, P.A. The cancer genome. *Nature* **2009**, *458*, 719–724. [[CrossRef](#)] [[PubMed](#)]
113. Gaffney, E.F.; Riegman, P.H.; Grizzle, W.E.; Watson, P.H. Factors that drive the increasing use of FFPE tissue in basic and translational cancer research. *Biotech. Histochem.* **2018**, *93*, 373–386. [[CrossRef](#)] [[PubMed](#)]
114. Reumers, J.; Rijk, P.D.; Zhao, H.; Liekens, A.; Smeets, D.; Cleary, J.; Loo, P.V.; Bossche, M.V.D.; Catthoor, K.; Sabbe, B.; et al. Optimized filtering reduces the error rate in detecting genomic variants by short-read sequencing. *Nat. Biotechnol.* **2012**, *30*, 61–68. [[CrossRef](#)] [[PubMed](#)]