Book Chapter

Whole Exome Sequencing Data Analysis Algorithms in Cancer Diagnostics

Áron Bartha[1,2] and Balázs Gyorffy[1,2]*

[1]Department of Bioinformatics and 2nd Department of Pediatrics, Semmelweis University, Hungary
[2]TTK Cancer Biomarker Research Group, Institute of Enzymology, Hungary

***Corresponding Author:** Balázs Gyorffy, Department of Bioinformatics and 2nd Department of Pediatrics, Semmelweis University, H-1094 Budapest, Hungary

Published **March 27, 2020**

This Book Chapter is a republication of an article published by Áron Bartha and Balázs Gyorffy. at Cancers in November 2019. (Bartha, Á.; Győrffy, B. Comprehensive Outline of Whole Exome Sequencing Data Analysis Tools Available in Clinical Oncology. Cancers 2019, 11, 1725.)

How to cite this book chapter: Áron Bartha, Balázs Gyorffy. Whole Exome Sequencing Data Analysis Algorithms in Cancer Diagnostics. In: Heidari A, editor. Prime Archives in Cancer Research. Hyderabad, India: Vide Leaf. 2020.

Author Contributions: Conceptualization, Á.B. and B.G.; methodology, Á.B. and B.G.; investigation, Á.B. and B.G.; writing—original draft preparation, Á.B.; writing—review and

editing, B.G.; visualization, Á.B. and B.G.; supervision, B.G.; project administration, Á.B.; funding acquisition, B.G.

Funding:
National Research, Development and Innovation Office of Hungary: NVKP_16-1-2016-0037;
National Research, Development and Innovation Office of Hungary: 2018-1.3.1-VKE-2018-00032;
National Research,Development and Innovation Office of Hungary: KH-129581.

Acknowledgments: Testing and evaluation of tools was performed using infrastructure and support provided by ELIXIR.

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

Abstract

Whole exome sequencing (WES) enables the analysis of all protein coding sequences in the human genome. This technology enables the investigation of cancer-related genetic aberrations that are predominantly located in the exonic regions. WES delivers high-throughput results at a reasonable price. Here, we review analysis tools enabling utilization of WES data in clinical and research settings. Technically, WES initially allows the detection of single nucleotide variants (SNVs) and copy number variations (CNVs), and data obtained through these methods can be combined and further utilized. Variant calling algorithms for SNVs range from standalone tools to machine learning-based combined pipelines. Tools for CNV detection compare the number of reads aligned to a dedicated segment. Both SNVs and CNVs help to identify mutations resulting in pharmacologically druggable alterations. The identification of homologous recombination deficiency enables the use of PARP inhibitors. Determining microsatellite instability and tumor mutation burden helps to select patients eligible for immunotherapy. To pave the way for clinical applications, we have to recognize some limitations of WES, including its restricted ability to detect CNVs, low coverage compared to targeted sequencing, and the missing consensus

regarding references and minimal application requirements. Recently, Galaxy became the leading platform in non-command line-based WES data processing. The maturation of next-generation sequencing is reinforced by Food and Drug Administration (FDA)-approved methods for cancer screening, detection, and follow-up. WES is on the verge of becoming an affordable and sufficiently evolved technology for everyday clinical use.

Keywords

Whole Exome Sequencing; Cancer; Bioinformatics

Introduction

In the last decade, the price of genome sequencing has shrunk significantly, most of the work has become automated, and preparation guidelines have evolved. Due to these achievements, sequencing a whole genome has become a readily available possibility. Sequencing only targeting regions or the exome, however, implies a significantly smaller financial burden. In whole exome sequencing (WES), we primarily target specific fragments of the genome, the protein-coding part, and we therefore are able to identify genetic variants that will affect proteins. Since most of the known disease-causing mutations fall into this category, exome sequencing is a method that significantly reduces sequencing costs and therefore represents a clinically feasible approach for patient diagnostics. In this paper, we provide a summary of bioinformatic methods addressing the detection of the most frequent genetic aberrations influencing the development and progression of cancer.

Cancer is characterized by a set of essential steps that each renegade cell has to master before it can evolve to cancer [1]. The multitude of experimental methods that are at hand to investigate these cancer hallmarks have been systematically reviewed recently [2]. Whole exome sequencing provides a versatile tool to simultaneously monitor multiple different genomic changes in the tumor tissue. Mutations in both coding and noncoding DNA sequence regions have proven to be

influential in the development of cancer [3,4]. Nucleic acid changes in the exome can result in amino acid changes in protein sequences. Amino acid changes lead to weakened activity of tumor suppressors, such as APC in colorectal cancer, VHL in renal cell cancer, or BRCA in breast cancer [5–7]. Copy number changes in cell cycle regulators, such as TP53 and RB1 [8], as well as malfunctions in repair mechanisms including the homologous recombination and DNA mismatch repair systems, predispose cells to cancer development. The activity of these repair systems can be monitored by measuring tumor mutational burden or microsatellite instability [9,10].

First Steps of Whole Exome Sequencing

At present, there are two main categories of next-generation sequencing (NGS) methods, consisting of DNA amplification-based sequencing (Illumina, Ion Torrent) and single molecule real-time sequencing (Pacific Biosciences, Oxford Nanopore). The investigated tissue samples can be freshly frozen, formalin-fixed and paraffin-embedded (FFPE), or liquid-based (blood sample); typically, each of these samples has its own isolation kits.

A critical initial step of NGS is adequate pathological examination, as a properly selected and dissected tissue sample is a necessity for any further investigation [11]. Samples should contain a sufficient proportion of tumor cells to differentiate germline and somatic mutations. DNA from an adjacent normal tissue or from a blood sample is needed to identify all germ-line mutations. DNA quality deteriorates with time and after FFPE conservation, which has a degrading effect on the DNA. As the fragmentation of the DNA increases, the genome assembly following sequencing becomes more challenging [12]. During library construction, the exons are captured after an initial fragmentation step. Exome capture can be microarray-based or magnetic-bead based. In this second case, specific probes are hybridized to the sample, which are then pulled out using the magnetic beads. Then, the intronic sequences are discarded, and sequencing is performed using all the exonic sequences. The magnetic-bead-based capture methods are more widespread due

to their simplicity [13]. To reach sufficient depth of coverage, properly capturing the targeted regions is necessary. Overall, currently used technologies deliver high efficiency [14]. Actual sequencing comes following exome capture and PCR amplification. The overall process of WES, including data processing and utilization, is summarized in Figure 1.

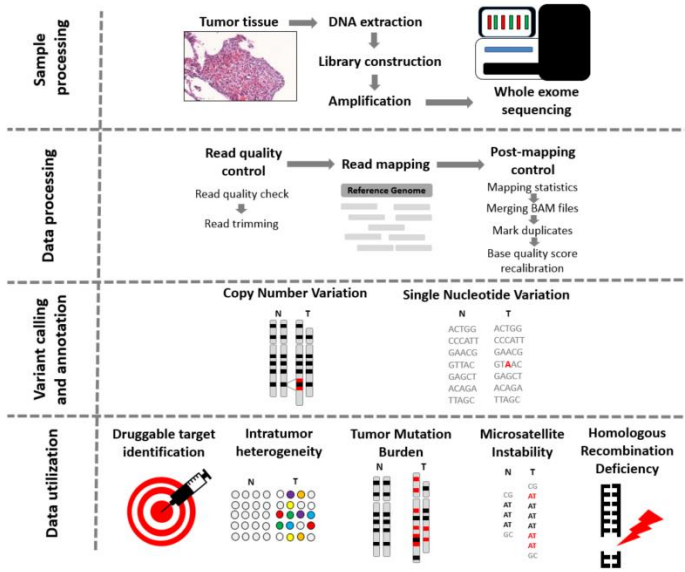

Figure 1: From tissue to data—steps of whole exome sequencing. Tissue preprocessing starts with the identification of tumor regions by an experienced pathologist, followed by DNA extraction, library construction, and amplification. Data procession commences with the quality check of reads. If the quality of trimmed reads is sufficient, the alignment of the reads to a reference genome is launched. When Binary Alignment Map (BAM) files are processed, the calling of single nucleotide variants, insertions and deletions, and copy number variants comes next, using one or more of the numerous existing algorithms. The data can be further utilized to detect microsatellite instability status, intratumor heterogeneity, tumor mutational burden, and homologous recombination deficiency.

Usually, the data processing part starts with quality control and trimming at which low-quality reads are removed. This step is followed by the alignment of reads to a chosen reference genome followed by a second quality check step and removal of the duplicate reads. After these data processing steps, the variant calling splits, and at this point, a plethora of tools are

available, depending on the clinical question one is attempting to answer.

Short Nucleotide Variants

Whole exome sequencing is capable of delivering information for all protein-coding regions of the genome, which makes it a useful tool to identify germline and somatic mutations from a tumor sample (Figure 2). Compared to targeted sequencing, WES has the advantage of being able to elucidate the whole exome profile of a sample and to provide information on those low-frequency mutations that can collectively ground a complex phenotypic appearance [15]. Single nucleotide variants are able to increase the expression of key druggable targets, as has been suggested in lung [16], breast [17], colon [18], and gastric cancer [19].

Accurate variant calling is a crucial component in the identification of such short variants. Currently, the most common variant caller tools in use include MuTect [20], VarScan2 [21], SomaticSniper [22], Strelka [23], and FreeBayes [24]. In addition, several clinical studies used a combination of these applications for variant calling [25–35]. A comprehensive list of all available tools is presented in Table 1, and the most common tools are presented in Figure 3.

According to a comparative analysis [57], selection of the right variant caller algorithm depends on the interest of variants. Some tools excel when dealing with low-coverage data (SomaticSniper [22], FaSD-somatic [54], and SNVSniffer [52]), while others perform better in regard to analyze low-frequency variants from high-coverage data (Strelka [23], MuTect [20], LoFreq [37], EBCall [41], deepSNV [45], LoLoPicker [56], and MuSE [42]). Other investigations also supported the approach of using specific variant callers: VarScan identified more high-quality single nucleotide variants (SNVs), while MuTect showed better performance in low-quality detection; therefore, the combinational usage of these can provide improved accuracy [58]. Examination of data from five breast cancer patients with nine variant caller algorithms affirmed the

discrepant effect of coverage variability on the results [59]. Comparison of the four most frequently used applications (MuTect2, Strelka, VarsScan2, and SomaticSniper) lead to comparable results [60]. Each caller delivered a divergent outcome, although MuTect2 and Strelka outperformed VarScan and SomaticSniper in some cases. At the end, the authors conclude that the combination of tools could increase performance but with the sacrifice of a vast amount of detected calls [60]. Similar conclusions of complementary algorithms were drawn in another study evaluating four variant callers using whole exome sequencing and simulated data [61]. These researchers also noted differences based on different aligner tools. A further study also underlined the importance of the adequate mixture of aligner and variant caller selection and recommended the combination of the BWA-MEM aligner and SAMtools for SNP calling and the BWA-MEM GATK-Haplotype caller for indel detection [62]. It is important to note that in most comparative studies, the authors used the default settings of the tools; thus, for several methods, the performance might be improved by fine tuning and customization of filters.

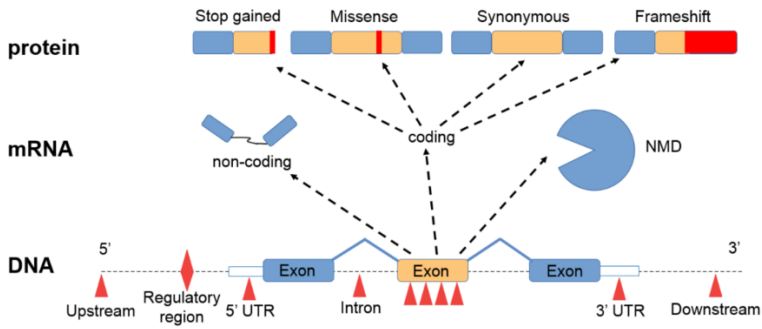

Figure 2: Effects of sequence alterations. Sequence variants in regulatory regions can activate or inhibit transcription. Mutations in exons result in an altered mRNA. Repair mechanisms, such as nonsense-mediated mRNA decay (NMD), can eliminate such abnormal mRNAs. As a result, missense mutations cause amino acid changes, while synonymous mutations result in the original amino acid sequence. Premature stop codons result in terminated amino acid sequences. Base insertions or deletions lead to frameshift mutations resulting in completely different proteins.

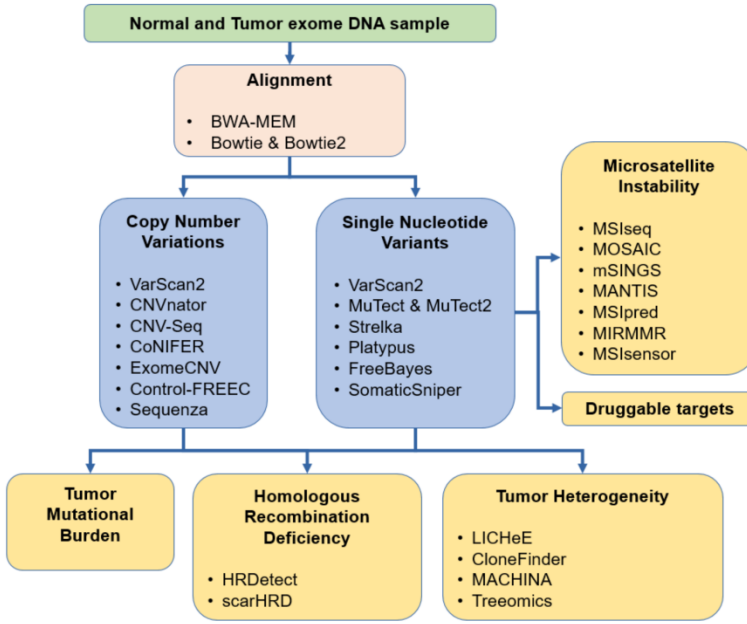

Figure 3: Overview of the most common methods for aberration detection useful in cancer diagnostics.

Table 1: Bioinformatic methods available for single nucleotide variant calling. Tools marked with an asterisk (*) are suitable for both whole genome sequencing (WGS) and whole exome sequencing (WES) data analysis.

Name	Published	Cited in 2018	Control Needed	InDel detection	Contamination Correction	Trained on Cancer Data	Environment	Ref
Varscan2	2012	2229	+	+	−	+	Java, Perl, R, Galaxy	[21]
MuTect2 *	2013	2005	+	−	+	+	Java, R	[20]
FreeBayes	2012	1121	−	+	−	+	C, C++, Galaxy	[24]
Strelka *	2012	759	+	+	−	+	C++, Perl	[23]
Platypus *	2014	462	−	+	−	+	C, Cython, Python	[36]
SomaticSniper *	2012	373	+	−	−	+	C, Galaxy	[22]
LoFreq *	2012	349	−	+	+	+	Python	[37]
VarDict *	2016	171	−	+	−	+	Perl	[38]
JointSNVMix *	2012	160	+	−	−	+	C, C++, Python, Galaxy	[39]
MutationSeq *	2012	108	+	−	−	+	C++, Python	[40]
EBCall *	2013	85	+	+	−	+	C++, Perl, R, Shell	[41]
MuSE *	2016	65	+	−	+	+	C, C++	[42]
RADIA	2014	53	+	−	+	+	Python	[43]
Virmid	2013	49	+	−	+	+	Java	[44]
deepSNV *	2014	47	+	−	−	+	R	[45]
Shimmer *	2013	45	+	−	+	+	C, Perl, R	[46]
qSNP *	2013	40	+	−	+	−	Java	[47]
BAYSIC	2014	39	+	−	−	+	R	[48]
SomaticSeq *	2015	38	+	+	−	+	Python, R	[49]
CaVEMan *	2016	31	+	−	+	+	C	[50]
SNooPer *	2016	26	−	+	+	+	Perl	[51]
SNVSniffer *	2016	17	−	+	−	+	C++	[52]
HapMuC	2014	15	−	+	−	+	C++, Python, Ruby	[53]
FaSD-somatic	2014	13	−	−	−	+	C, C++	[54]
LocHap *	2016	8	+	+	+	+	g++ complier, GNU Make	[55]
LoLoPicker *	2017	6	+	−	+	+	Python	[56]

Integrated Tools

Overall, different algorithms produce divergent output results. The utilization of combined pipelines can successfully filter the false positive hits and provide a platform for the customization of variant calling pipelines for the designated research objective. Such applications developed to deliver consensus Variant Call Format (VCF) files include VCFtools [63] NGS-pipe [64], VariantTools [65], vcfr [66], and myVCF [67]. These tools are notably useful when one aims to build pipelines that analyze VCF files generated in other tools (listed in the previous chapter). Other algorithms, such as Cake, can use BAM files as inputs. Cake runs all the variant caller tools separately and then unites the SNVs confirmed by at least two of the caller tools. Cake also offers numerous post processing filtering options [68]. Isma, an R package for the integrative analysis of mutations detected by multiple pipelines, provides a common platform for Strelka, MuTect/MuTect2, MuSE, SomaticSniper, and VarScan2. Isma provides qualification for the used calling algorithms and highlights outlier results [69].

Using machine learning methods might further improve the specificity, sensitivity and comparability of these applications. BAYSIC integrates, among others, FreeBayes, SamTools, and GATK, and it can accept input from any variant caller algorithm [48]. SomaticSeq merges five algorithms (MuTect, VarScan2, SomaticSniper, JointSNVMix2, and VarDict), providing another machine learning-based ensembled application for SNV and indel identification [49]. SMuRF is another machine learning-based pipeline combining MuTect2, Freebayes, VarDict, and VarScan. SMuRF had the advantage of faster computing speed than other machine learning tools. While SMuRF outperformed several methods, it showed slightly poorer results than SomaticSeq; however, the time needed for SMuRF to compute the results was unsubstantial compared to SomaticSeq (10 min vs. 24 h) [70]. NeoMutate, a recently developed framework, also has the advantages of a mixture of separate tools and a machine learning-based perspective [71].

The application of machine learning ensemble methods has become increasingly accepted and shows a possible path for the development of future variant calling methods. However, currently implemented tools have an important drawback, as their sensitivity depends on that of the included algorithms.

Galaxy—An Open Source, Web-Based Platform

To use the applications discussed above, one has to possess advanced or at least intermediate programing skills, not to mention that many of these algorithms require different programming languages. Numerous user-friendly platforms have been established in the past years to overcome this obstacle. Generally, these are capable to give a platform in which users can build workflows made of genomic analysis tools. Researchers can use local workflow management systems like Taverna [72] or KNIME [73]. However, computing power is limited by the performance of the local computer. Cloud computing can serve as a possible solution for this issue [74]. Platforms like Cancer Genomics Cloud (CGC) [75], GenePattern [76], or Galaxy [77] are becoming more and more popular amongst scientists. Additional platforms available are listed in Table 2. Of these tools, Galaxy is the most widespread, due to the wide range of tools included and free availability. Users can utilize publicly available Galaxy servers or can set up their own private server.

Table 2: Platforms available for bioinformatic analysis.

Name	Description	Year	Citation	License	System type	Ref.
Galaxy	Open-source web-platform with several analysis tools	2005	1977	free	cloud-based	[77]
GenePattern	Workflow management system, provides access to multiple genomic analysis tools	2006	1573	free	cloud-based	[76]
KNIME	Software enabling creation, analysis, and visualization of data	2008	1476	free	local installation needed	[73]
UGENE	Workflow management system installed on a local computer	2012	876	free	local installation needed	[78]
Taverna	Open source software tool for designing and executing workflows	2013	643	free	local installation needed	[72]
Cancer Genomics Cloud	Provides access to data, tools, and computing resources	2017	32	commercial	cloud-based	[75]
SciApps	Platform for building, running, and sharing scientific workflows	2018	5	free	cloud-based	[79]
Terra	Bioinformatic workspace, including a repository of public best practices, methods, and public data sets	–	–	commercial	cloud-based	–

When setting up a private server, one can include any of more than 5500 tools and algorithms from the Galaxy toolshed, which serves as an "AppStore" of applications [80]. However, establishing a private server requires constant maintenance and a skilled system administrator. Using a publicly available server, on the other hand, requires only a registration to the designated server, and the leading Galaxy servers already contain most commonly used tools. In addition to accessible research, Galaxy also has two additional important advantages: it makes it easier to reproduce analyses and provides a platform for users to communicate.

In regard to variant calling, Galaxy ToolShed provides numerous algorithms. The Galaxy training materials suggest a few recommended tools: VarScan for the identification of germline and somatic variants from tumor-normal sample pairs and FreeBayes for germ line variant calling.

As the clinical significance of variant caller methods expands, demands are increasing to solve specific problems. These problems include the detection of low-frequency variants—one possible solution could be utilization of unique molecule identifiers—and the accommodation of non-Illumina platforms. The perpetual improvement of the algorithmic tools is foreseeable if they want to compete with deep learning algorithms [57]. On the other hand, it is important to note that even the most well-established pipelines can be inefficient if the quality of utilized data is poor, e.g., inadequate exome capture, low coverage or modest sequencing quality [62].

Copy Number Variations

Copy number variations (CNVs) are structural changes of DNA, sized between a couple of hundred base changes and amplification or deletion of millions of base pairs [81]. The clinical relevance of CNVs in oncology has risen in the past several years, and CNVs have been indicated to be important in several types of cancer, such as adenomatous polyposis coli, familiar breast cancer, and ovarian cancer [8].

The clinically used gold standards for CNV detection are array Comparative Genome Hybridization (aCGH), Fluorescent In Situ Hybridization (FISH), and qPCR [82]. Current Food and Drug Administration (FDA)-approved methods for CNV detection are mainly FISH-based such as the "Dako TOP2A FISH PharmDX kit" for the detection of Topoisomerase 2-alpha aberrations or targeted sequencing based on the "FoundationOneCDx" NGS panel, which is capable of measuring the copy number changes in 324 genes.

Each of the gold standard techniques is relatively inexpensive and provides reliable clinical data. Nonetheless, the opportunity to use sequencing can provide a robust amount of additional data with versatile further utility. Using whole genome sequencing (WGS) data for CNV detection has already been demonstrated to be useful [83]. However, due to financial issues, WGS is unlikely to become a clinical tool in the near future. WES, on the other hand, is a more affordable option to identify CNV changes.

Currently, dozens of algorithms and pipelines exist to detect CNVs from WES data; we have summarized these in Table 3, and the most common tools are listed in Figure 3. Most of the algorithms are based on the Read Depth approach, and they attempt to measure the CNV changes based on the number of reads aligned to a dedicated segment [84]. Although these algorithms can be relatively precise, normalization problems and other biases present as limitations of NGS technology. These limitations include contamination with normal cells, multiple types of clones among one sample and other experimental noises [85]. Only a few of the methods are capable of detecting CNV from cancer data, and substantial discrepancies can be observed when paralleling these tools. Although several studies have been conducted to compare these applications, only a few have focused on patients suffering from cancer as the study population.

Table 3: Computational methods available for copy number variation estimation from whole exome sequencing data. Tools marked with an asterisk are suitable for both WGS and WES data analysis.

Name	Published	Control Needed	Contamination Correction	GC-Content Correction	Trained on Cancer Data	Cited in 2018	Environment	Ref.
Varscan2	2012	+	−	−	+	2229	Java, Perl, R, Galaxy	[21]
CNVnator	2011	+	−	+	−	767	C++	[86]
CNV-Seq	2009	+	−	−	−	463	Perl, R	[87]
CoNIFER	2012	−	+	−	−	378	Python	[88]
Control-FREEC *	2012	−	+	+	+	342	C, C++, R	[89]
ExomeCNV	2011	+	+	−	+	338	R	[90]
XHMM	2012	−	+	+	+	322	C++	[91]
ExomeDepth	2012	+	−	+	−	264	R	[92]
cn.MOPS	2012	−	+	+	−	249	R	[93]
Cnvkit *	2016	+	+	+	+	219	Python, Galaxy	[94]
CONTRA	2012	−	−	+	−	194	Python, R	[95]
Sequenza *	2015	+	−	+	+	167	Python, R	[96]
EXCAVATOR	2013	+	+	+	+	155	Perl	[97]
CODEX	2015	−	+	+	+	72	R	[98]
ADTEx	2014	+	+	−	+	57	Python, R	[99]
Seqgene	2011	+	−	−	+	43	R	[100]
FishingCNV	2013	−	−	−	−	41	Java, R	[101]
HMZDelFinder	2017	−	−	−	−	33	R	[102]
ExoCNVTest	2012	+	−	−	−	27	Java, R	[103]
CLAMMS	2016	−	−	+	−	23	C	[104]
falcon	2015	+	+	−	+	22	C	[105]
saasCNV *	2015	+	+	−	+	17	R	[106]
WISExome	2017	−	−	−	−	1	C, C++	[107]

Zare et. al. examined five algorithms on tumorous samples and concluded that some applications have achieved relatively good results on specificity and sensitivity [108]. In particular, ExomeCNV [90] showed high specificity and sensitivity with a moderate false discovery rate. SAAS-CNV [106] might be a useful tool for CNV detection; however, the specificity and sensitivity of the algorithm are inferior compared to the array methods [109].

Regarding overall specificity and sensitivity using simulated data [110], ADTEx [99] produced the best results followed by ControlFREEC [89], VarScan2 and ExomeCNV, but ExomeCNV and VarScan2 missed several homozygous deletions. Using breast cancer data in the same comparative study, ExomeCNV [90] showed the best results, while it produced moderate concordance with SNP arrays. Overall, ControlFREEC presented the best algorithm due to the balanced performance on both simulated and cancer data [110].

Based on the study examining six methods (ADTEx, CONTRA [95], ControlFREEC, EXCAVATOR, ExomeCNV, and VarScan2), these can identify homozygous deletions or large gains from WES data, but heterozygous deletions or low-level amplifications cannot be detected with sufficient consistency [111]. The results provided by ADTEx and EXCAVATOR were the most reliable [111].

Taken together, all the cited studies compare algorithms that were designed for somatic CNV detection from cancer-related data, and each came to a similar conclusion. At present, neither sensitivity nor specificity is precise enough to compete with the existing non-WES methods. Furthermore, multiple studies highlighted that using these algorithms on stimulated data shows better performance than on patient data, which indicates that the tools are not sufficiently fine-tuned to address tumor complexity, although some of them, such as ADTEx and ExomeCNV, have a built-in tool to tackle this issue.

Each application has different strengths and weaknesses; for instance, ADTEx can detect medium-sized CNVs, while

EXCAVATOR is suitable for the identification of larger CNVs. Similar to SNVs, merging, fine tuning and recalibration of these tools could be a means of improving the specificity and sensitivity [112,113]. It is important to mention, however, that these discrepancies are not specific to somatic mutation detection, as similar issues appeared in germline mutation-based comparison [84].

Dealing with NGS data demands well-trained bioinformaticians because most of the algorithms can only be used in command line-based platforms. The availability of the aforementioned applications in Galaxy is slightly limited—to date, VarScan2 and a CNV caller part of the bcftools package are available in the basic Galaxy setup. Several further algorithms can be installed in the case of a private Galaxy server.

Homologous Recombination Deficiency

DNA double-strand breaks are one of the most mutagenic forms of DNA damage [114,115]. Cells have developed multiple solutions to confront these effects, such as homologous recombination and nonhomologous end-joining [116]. Germline mutations of the BRCA genes have been described as reliable markers to identify homologous recombination deficiency (HRD). Currently, one FDA-approved clinical tool is available to detect germline BRCA mutations, the BRACAnalysisCDx platform (Myriad Genetics; Salt Lake City, UT, USA), which is used to identify BRCA status in patients with ovarian cancer. The presence of a BRCA mutation enables treatment with a PARP inhibitor. PARP repairs single strand breaks, and the loss of both double-strand and single-strand break repair renders the tumor highly vulnerable to chemotherapy.

HRDetect is a WGS-based method to identify the presence of homologous recombination repair mechanism mutations; this tool has proven to be effective and reliable regardless of germline and somatic mutation or tissue type. However, using this tool on WES data revealed a considerable decrease in the detection sensitivity [117]. Another recent WES-based tool promises comparable results with SNP array examinations

based on genomic scar analysis and might be a useful tool to detect BRCA status [118]. Since HRD detection mainly focuses on BRCA status, we currently have a lack of application capable of measuring overall HRD status involving all related genetic aberrations. Meanwhile, several other genes have also been shown to play important roles in HRD [119]. An improved future WES-based algorithm could enable the simultaneous investigation of all involved genes.

Response to Immunotherapy

Immune checkpoint inhibitors and immunomodulatory agents have become standard treatments for solid tumors, including renal-cell carcinoma, melanoma and NSCLC [120]. The number of mutations per coding sequence in the tumor genome is a reliable predictive biomarker of immunotherapy response [121]. At present, the application of WES to detect tumor mutational burden (TMB) is a widely accepted gold standard. In addition, multiple targeted panels have also been accepted as targeted sequencing show comparable results in the detection of TMB status as exome sequencing [122].

Although TMB bears strong potential as a predictive biomarker, there is a lack of unambiguous consensus on the correct determination, definition, and cut-off values. The Friends of Cancer Research established a working group to create a universal reference and harmonize these methods to address this issue [123]. Because of the lack of solid guidelines, various studies have used numerous methods and computational techniques for TMB status determination. We evaluated eleven phase II and III clinical studies, and MuTect was the most frequently used tool for somatic variant detection, while the applications applied for InDel detection showed a wide variety [25–35]. A significant set of publications use the pipeline proposed by the Genome Analysis Toolkit—supplementing it with additional tools—which recommends GATK-Mutect2, which is based on MuTect and the GATK-HaplotypeCaller.

Another concept recently gaining attention is the examination of mutational signatures. Mutations in cancer can originate in

different mutagenic effects or defects in repair mechanisms. Each genetic aberration has its unique mutational signature which can include base substitutions, small insertions and deletions, CNV changes, or genomic rearrangements [124]. As the quantity of explored signatures is growing, a systematic and curated archive of genetic patterns is needed. The Catalogue of Somatic Mutations in Cancer (COSMIC) provides such a repository for mutational signatures and specific summary vignettes. Deciphering characteristic mutational patterns in a chosen cancer type requires bioinformatic analysis as well. Currently, there are several algorithms designed for mutational landscape identification, such as SigProfiler [125], deconstrutSigs [126], and mutationalPatterns [127]. HRDetect, a tool developed as a kind of mutational signature detecting algorithm designed for the identification of homologue recombination repair deficiency, has been already discussed in a separate paragraph. Accepted analysis standards for these methods are still missing [128]. Clinical cancer diagnostics might benefit from the application of mutational signature detection, as aberration patterns can be useful for targeted treatment selection [129].

A different predictive biomarker for immune modulatory response is the evaluation of Microsatellite Instability (MSI). From the time when the FDA approved pembrolizumab for the treatment of adult and pediatric microsatellite instability high (MSI-H) or mismatch repair-deficient (dMMR) solid tumors, MSI detection gathered significant clinical attention [29]. A recent study suggests that impaired mismatch repair activity might result in higher mutational burden resulting in augmented response to immunomodulatory agents [130]. The currently existing method for MSI detection, known as the combination of PCR with fluorescent primers and capillary electrophoresis, is becoming obsolete with the introduction of WES and targeted gene panel sequencing [131].

At present, the number of applications for MSI identification from exome sequencing data is not as high as the number of those for CNV or short variant detection. Comparing some of these tools in six cancer types revealed that MANTIS produces

better sensitivity and specificity than MSIsensor and mSINGS [132]. MSIseq show results comparable to MSISensor and mSINGS, while the MSIseq R package runs much faster than the two other [133]. MSIseq and MSIpred have the advantage that these algorithms can measure MSI from tumor data only. Based on data comparison using TCGA data, MSIpred exhibited higher accuracy and sensitivity than MSIseq [134]. MIRMMR also displayed similarities in accuracy and sensitivity with MSIsensor and mSINGS [135]. A recently implemented tool based on the examination of 5930 tumor exomes across 18 cancer types, called MOSAIC, produced remarkable sensitivity and specificity [136]. Overall, out of the seven algorithms available, MOSAIC has the strongest and most well-established analytical background, while MSIpred shows better performance than others with the advantage that it can operate without a normal reference sample.

Unfortunately, no specific tool has been developed for MSI detection from exome sequencing data for those who have less experience in command line coding. This finding indicates that the Galaxy platform is the only alternative.

Finally, predicting response to immunotherapy has an additional option as—according to a state-of-the-art paper—elevated DNA damage might be a possible biomarker of response [137].

Tumor Heterogeneity

Tumor heterogeneity stands for diversity within one tumor population, where several different populations coexist. These cancerous populations coexist with normal cells and infiltrating immune-related cells in a special microenvironment. The subclonal populations can cooperatively evolve and are even capable of adapting to altered circumstances, including the emergence of therapy-resistant clones following systemic anticancer treatments [138]. Currently, there is no broadly accepted consensus method for the estimation of tumor heterogeneity. Identification of the clonal subpopulations is possible by all three sequencing methods—WGS, WES, and targeted panel—and by single-cell approaches.

A widely accepted way to measure tumor diversity is the use of WES to measure the genetic heterogeneity of a tumor sample by counting Shannon's diversity index of the estimated SNVs [139]. The determination of tumor clonality and evolutionary background from bulk sequencing data is a multistep process. This method begins with the cancer cell fraction estimation, then the identification of tumor subclones followed by the construction of a phylogenic tree based on the distribution of somatic variants and/or CNV status. Finally, temporal differentiation can assist in distinguishing between passenger and driver mutations [140].

In addition to the aforementioned approach, numerous algorithms have been developed to illuminate subclone phylogenesis. Unfortunately, due to the scarcity of comparative studies, we have only limited guidance on proper algorithm selection at this time. In a study of nine methods, LICHeE and CloneFinder produced decent accuracy compared to the others [141]. In a recent comparative study currently available in a preprint server only, the authors examined seven clonality prediction methods. CloneFinder, MACHINA, Treeomics, and LICHeE showed the best performance, but it is important to mention that none of the applications showed impaired overall performance on all the stimulated datasets [142].

Overall, the examination of tumor heterogeneity by NGS-based methods has a limited history, and because of this reason, many of the currently existing methods require further fine-tuning. In vitro experiments might serve as guidance for adequate algorithm calibration and could provide further information on the detection threshold and coverage cut-off value selection. Recently, we have shown that cellular movement can also lead to a significant technical bias when using NGS to determine the clonal composition of a tumor [143]. With the technical development of both bulk sequencing and single cell methods, we will soon be able to confidently obtain an accurate picture of a cancer population in its complete heterogeneity.

Discussion

The first U.S. Food and Drug Administration (FDA) approval for NGS technology was issued in 2013, and a few years later, the approval of the first tests for diagnostic and screening was granted. We provide an overview of NGS-based tests approved for somatic or germline mutation detection in Table 4.

We are now in the big data era borne by the vast amount of data delivered by new sequencing methods. Deciphering this information requires complex bioinformatical analytical tools. At the same time, we have to account for the unquestionable weaknesses of exome sequencing [144]. These disadvantages include the limited power to detect structural gene fusions and the limited ability to delineate tumor purity and differentiate from normal cell contamination. The previously discussed machine learning algorithms in short variant detection can improve the accuracy of TMB and MSI detection, as punctual short variant identification is a crucial part of both. Improved detection of copy number changes can lead to more accurate HRD and tumor heterogeneity analysis [145].

The final outcome of our paper is that, due to discrepancies amongst tools used during sample preparation and data preprocessing and processing, it is almost impossible to define a gold standard guideline of the most handy algorithms. Of note, anyone can customize the selected algorithms specifically for their own experiment rather than using it on default settings.

The clinical significance of NGS-based methods is consistently expanding. Although discrepancies can be observed among the currently available tools, the continuous fine-tuning and the merged utilization of these applications paves the way for clinically reliable applications in the coming years. Overall, WES is emerging as a future “Swiss army knife” of cancer genome profiling. After as bioinformatic processes have evolved to trustworthy pipelines, WES will be an affordable and mature technology for everyday clinical use.

Table 4: Food and Drug Administration (FDA)-approved next-generation sequencing (NGS)-based methods suitable for cancer predisposition identification, cancer detection, or follow-up.

Tradename	Description	Year	Target	Tumor	Utility
Illumina MiSeqDX platform	High throughput DNA sequence analyzer	2013	-	-	technology
FoundationFocus CDxBRCA	NGS oncology panel, somatic or germline variant detection system	2016	BRCA	ovarian	diagnosis
MSK-IMPACT	NGS-based tumor profiling test	2017	468 genes	various	predisposition, diagnosis
FoundationOne CDx	NGS oncology panel, somatic or germline variant detection system	2017	324 genes	various	predisposition, diagnosis
Oncomine Dx Target Test	NGS oncology panel, somatic or germline variant detection system	2017	24 genes	lung	diagnosis
Praxis Extended RAS Panel	NGS oncology panel, somatic or germline variant detection system	2017	RAS	colon	diagnosis
Adaptive Biotechnologies clonoSEQ	DNA-based test for minimal residual disease for hematologic malignancies	2018	BCL1, BCL2	leukemia, myeloma	follow-up

References

1. Hanahan D, Weinberg RA. Hallmarks of cancer: The next generation. Cell. 2011; 144: 646–674.
2. Menyhart O, Harami-Papp H, Sukumar S, Schafer R, Magnani L, et al. Guidelines for the selection of functional assays to evaluate the hallmarks of cancer. Biochim. Biophys. Acta. 2016; 1866: 300–319.
3. Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. Linking disease associations with regulatory information in the human genome. Genome Res. 2012; 22: 1748–1759.
4. Li G, Pan T, Guo D, Li LC. Regulatory Variants and Disease: The E-Cadherin -160C/A SNP as an Example. Mol. Biol. Int. 2014; 2014: 967565.
5. Minde DP, Anvarian Z, Rudiger SG, Maurice MM. Messing up disorder: How do missense mutations in the tumor suppressor protein APC lead to cancer? Mol. Cancer. 2011; 10: 101.
6. Gnarra JR, Tory K, Weng Y, Schmidt L, Wei MH, et al. Mutations of the VHL tumour suppressor gene in renal carcinoma. Nat. Genet. 1994; 7: 85–90.
7. Farmer H, McCabe N, Lord CJ, Tutt AN, Johnson DA, et al. Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. Nature. 2005; 434: 917–921.
8. Shlien A, Malkin D. Copy number variations and cancer. Genome Med. 2009; 1: 62.
9. Torgovnick A, Schumacher B. DNA repair mechanisms in cancer development and therapy. Front. Genet. 2015; 6: 157.
10. Luchini C, Bibeau F, Ligtenberg MJL, Singh N, Nottegar A, et al. ESMO recommendations on microsatellite instability testing for immunotherapy in cancer, and its relationship with PD-1/PD-L1 expression and tumour mutational burden: A systematic review-based approach. Ann. Oncol. O. J. Eur. Soc. Med Oncol. 2019; 30: 1232–1243.
11. Morlote D, Janowski KM, Siniard RC, Guo RJ, Winokur T, et al. Effects of Improved DNA Integrity by Punch from Tissue Blocks as Compared to Pinpoint Extraction from Unstained Slides on Next-Generation Sequencing Quality Metrics. Am. J. Clin. Pathol. 2019; 152: 27–35.
12. McDonough SJ, Bhagwate A, Sun Z, Wang C, Zschunke M,

et al. Use of FFPE-derived DNA in next generation sequencing: DNA extraction methods. PLoS ONE. 2019; 14: e0211400.
13. Warr A, Robert C, Hume D, Archibald A, Deeb N, et al. Exome Sequencing: Current and Future Perspectives. G3 Genes Genomes Genet. 2015; 5: 1543–1550.
14. Chilamakuri CS, Lorenz S, Madoui MA, Vodak D, Sun J, et al. Performance comparison of four exome capture systems for deep sequencing. BMC Genom. 2014; 15: 449.
15. Pongor L, Kormos M, Hatzis C, Pusztai L, Szabo A, et al. A genome-wide approach to link genotype to clinical outcome by utilizing next generation sequencing and gene chip data of 6697 breast cancer patients. Genome Med. 2015; 7: 104.
16. Nagy A, Pongor LS, Szabo A, Santarpia M, Gyor yB. KRAS driven expression signature has prognostic power superior to mutation status in non-small cell lung cancer. Int. J. Cancer. 2017; 140: 930–937.
17. Gyor yB, Pongor L, Bottai G, Li X, Budczies J, et al. An integrative bioinformatics approach reveals coding and non-coding gene variants associated with gene expression profiles and outcome in breast cancer molecular subtypes. Br. J. Cancer. 2018; 118: 1107–1114.
18. Menyhart O, Kakisaka T, Pongor LS, Uetake H, Goel A, et al. Uncovering Potential Therapeutic Targets in Colorectal Cancer by Deciphering Mutational Status and Expression of Druggable Oncogenes. Cancers. 2019; 11: 983.
19. Menyhart O, Pongor LS, Gyor yB. Mutations Defining Patient Cohorts with Elevated PD-L1 Expression in Gastric Cancer. Front. Pharmacol. 2018; 9: 1522.
20. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jae D, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol. 2013; 31: 213–219.
21. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, et al. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012; 22: 568–576.
22. Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, et al. SomaticSniper: Identification of somatic point mutations in whole genome sequencing data. Bioinformatics. 2012;

28: 311–317.
23. Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, et al. Strelka: Accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics. 2012; 28: 1811–1817.
24. Erik Garrison GM. Haplotype-based variant detection from short-read sequencing. 2012; arXiv: 1207.3907.
25. Carbone DP, Reck M, Paz-Ares L, Creelan B, Horn L, et al. First-Line Nivolumab in Stage IV or Recurrent Non-Small-Cell Lung Cancer. N. Engl. J. Med. 2017; 376: 2415–2426.
26. Cristescu R, Mogg R, Ayers M, Albright A, Murphy E, et al. Pan-tumor genomic biomarkers for PD-1 checkpoint blockade-based immunotherapy. Science. 2018; 362.
27. Hellmann MD, Nathanson T, Rizvi H, Creelan BC, Sanchez-Vega F, et al. Genomic Features of Response to Combination Immunotherapy in Patients with Advanced Non-Small-Cell Lung Cancer. Cancer Cell. 2018; 33: 843–852.e4.
28. Hugo W, Zaretsky JM, Sun L, Song C, Moreno BH, et al. Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. Cell. 2016; 165: 35–44.
29. Le DT, Uram JN, Wang H, Bartlett BR, Kemberling H, et al. PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. N. Engl. J. Med. 2015; 372: 2509–2520.
30. Riaz N, Havel JJ, Makarov V, Desrichard A, Urba WJ, et al. Tumor and Microenvironment Evolution during Immunotherapy with Nivolumab. Cell. 2017; 171: 934–949.e16.
31. Rizvi H, Sanchez-Vega F, La K, Chatila W, Jonsson P, et al. Molecular Determinants of Response to Anti-Programmed Cell Death (PD)-1 and Anti-Programmed Death-Ligand 1 (PD-L1) Blockade in Patients With Non-Small-Cell Lung Cancer Profiled With Targeted Next-Generation Sequencing. J. Clin. Oncol. O . J. Am. Soc. Clin. Oncol. 2018; 36: 633–641.
32. Rizvi NA, Hellmann MD, Snyder A, Kvistborg P, Makarov V, et al. Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. Science. 2015; 348: 124–128.

33. Snyder A, Makarov V, Merghoub T, Yuan J, Zaretsky JM, et al. Genetic basis for clinical response to CTLA-4 blockade in melanoma. N. Engl. J. Med. 2014; 371: 2189–2199.
34. Snyder A, Nathanson T, Funt SA, Ahuja A, Buros Novik J, et al. Contribution of systemic and somatic factors to clinical response and resistance to PD-L1 blockade in urothelial cancer: An exploratory multi-omic analysis. PLoS Med. 2017; 14: e1002309.
35. Van Allen EM, Miao D, Schilling B, Shukla SA, Blank C, et al. Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. Science. 2015; 350: 207–211.
36. Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SRF, et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nat. Genet. 2014; 46: 912–918.
37. Wilm A, Aw PP, Bertrand D, Yeo GH, Ong SH, et al. LoFreq: A sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res. 2012; 40: 11189–11201.
38. Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, et al. VarDict: A novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Res. 2016; 44: e108.
39. Roth A, Ding J, Morin R, Crisan A, Ha G, et al. JointSNVMix: A probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. Bioinformatics. 2012; 28: 907–913.
40. Ding J, Bashashati A, Roth A, Oloumi A, Tse K, et al. Feature-based classifiers for somatic mutation detection in tumour-normal paired sequencing data. Bioinformatics. 2012; 28: 167–175.
41. Shiraishi Y, Sato Y, Chiba K, Okuno Y, Nagata Y, et al. An empirical Bayesian framework for somatic mutation detection from cancer genome sequencing data. Nucleic Acids Res. 2013; 41: e89.
42. Fan Y, Xi L, Hughes DS, Zhang J, Zhang J, et al. MuSE: Accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation

calling from sequencing data. Genome Biol. 2016; 17: 178.
43. Radenbaugh AJ, Ma S, Ewing A, Stuart JM, Collisson EA, et al. RADIA: RNA and DNA integrated analysis for somatic mutation detection. PLoS ONE. 2014; 9: e111516.
44. Kim S, Jeong K, Bhutani K, Lee J, Patel A, et al. Virmid: Accurate detection of somatic mutations with sample impurity inference. Genome Biol. 2013; 14: R90.
45. Gerstung M, Papaemmanuil E, Campbell PJ. Subclonal variant calling with multiple samples and prior knowledge. Bioinformatics. 2014; 30: 1198–1204.
46. Hansen NF, Gartner JJ, Mei L, Samuels Y, Mullikin JC. Shimmer: Detection of genetic alterations in tumors using next-generation sequence data. Bioinformatics. 2013; 29: 1498–1503.
47. Kassahn KS, Holmes O, Nones K, Patch AM, Miller DK, et al. Somatic point mutation calling in low cellularity tumors. PLoS ONE. 2013; 8: e74380.
48. Cantarel BL, Weaver D, McNeill N, Zhang J, Mackey AJ, et al. BAYSIC: A Bayesian method for combining sets of genome variants with improved specificity and sensitivity. BMC Bioinform. 2014; 15: 104.
49. Fang LT, Afshar PT, Chhibber A, Mohiyuddin M, Fan Y, et al. An ensemble approach to accurately detect somatic mutations using SomaticSeq. Genome Biol. 2015; 16: 197.
50. Jones D, Raine KM, Davies H, Tarpey PS, Butler AP, et al. cgpCaVEManWrapper: Simple Execution of CaVEMan in Order to Detect Somatic Single Nucleotide Variants in NGS Data. Curr. Protoc. Bioinform. 2016; 56: 15.
51. Spinella JF, Mehanna P, Vidal R, Saillour V, Cassart P, et al. SNooPer: A machine learning-based method for somatic variant identification from low-pass next-generation sequencing. BMC Genom. 2016; 17: 912.
52. Liu Y, Loewer M, Aluru S, Schmidt B. SNVSni er: An integrated caller for germline and somatic single-nucleotide and indel mutations. BMC Syst. Biol. 2016; 10: 47.
53. Usuyama N, Shiraishi Y, Sato Y, Kume H, Homma Y, et al. HapMuC: Somatic mutation calling using heterozygous germ line variants near candidate mutations. Bioinformatics. 2014; 30: 3302–3309.
54. Wang W, Wang P, Xu F, Luo R, Wong MP, et al. FaSD-

somatic: A fast and accurate somatic SNV detection algorithm for cancer genome sequencing data. Bioinformatics. 2014; 30: 2498–2500.
55. Sengupta S, Gulukota K, Zhu Y, Ober C, Naughton K, et al. Ultra-fast local-haplotype variant calling using paired-end DNA-sequencing data reveals somatic mosaicism in tumor and normal blood samples. Nucleic Acids Res. 2016; 44: e25.
56. Carrot-Zhang J, Majewski J. LoLoPicker: Detecting low allelic-fraction variants from low-quality cancer samples. Oncotarget. 2017; 8: 37032–37040.
57. Xu C. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. Comput. Struct. Biotechnol. J. 2018; 16: 15–24.
58. Liu ZK, Shang YK, Chen ZN, Bian H. A three-caller pipeline for variant analysis of cancer whole-exome sequencing data. Mol. Med. Rep. 2017; 15: 2489–2494.
59. Kroigard AB, Thomassen M, Laenkholm AV, Kruse TA, Larsen MJ. Evaluation of Nine Somatic Variant Callers for Detection of Somatic Mutations in Exome and Targeted Deep Sequencing Data. PLoS ONE. 2016; 11: e0151664.
60. Cai L, Yuan W, Zhang Z, He L, Chou KC. In-depth comparison of somatic point mutation callers based on di erent tumor next-generation sequencing depth data. Sci. Rep. 2016; 6: 36540.
61. 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.
62. Hwang S, Kim E, Lee I, Marcotte EM. Systematic comparison of variant calling pipelines using gold standard personal exome variants. Sci. Rep. 2015; 5: 17875.
63. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, et al. The variant call format and VCFtools. Bioinformatics. 2011; 27: 2156–2158.
64. Singer J, Ruscheweyh HJ, Hofmann AL, Thurnherr T, Singer F, et al. NGS-pipe: A flexible, easily extendable and highly configurable framework for NGS analysis. Bioinformatics. 2018; 34: 107–108.
65. Lawrence M, Gentleman R. VariantTools: An extensible

framework for developing and testing variant callers. Bioinformatics. 2017; 33: 3311–3313.
66. Knaus BJ, Grunwald NJ. vcfr: A package to manipulate and visualize variant call format data in R. Mol. Ecol. Resour. 2017; 17: 44–53.
67. Pietrelli A, Valenti L. myVCF: A desktop application for high-throughput mutations data management. Bioinformatics. 2017; 33: 3676–3678.
68. Rashid M, Robles-Espinoza CD, Rust AG, Adams DJ. Cake: A bioinformatics pipeline for the integrated analysis of somatic variants in cancer genomes. Bioinformatics. 2013; 29: 2208–2210.
69. Di Nanni N, Moscatelli M, Gnocchi M, Milanesi L, Mosca E. isma: An R package for the integrative analysis of mutations detected by multiple pipelines. BMC Bioinform. 2019; 20: 107.
70. Huang W, Guo YA, Muthukumar K, Baruah P, Chang MM, et al. SMuRF: Portable and accurate ensemble prediction of somatic mutations. Bioinformatics. 2019.
71. Anzar I, Sverchkova A, Stratford R, Clancy T. NeoMutate: An ensemble machine learning framework for the prediction of somatic mutations in cancer. BMC Med. Genom. 2019; 12: 63.
72. Wolstencroft K, Haines R, Fellows D, Williams A, Withers D, et al. The Taverna workflow suite: Designing and executing workflows of Web Services on the desktop, web or in the cloud. Nucleic Acids Res. 2013; 41: W557–W561.
73. Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, et al. KNIME: The Konstanz Information Miner. Berlin: Springer. 2008; 319–326.
74. Langmead B, Nellore A. Cloud computing for genomic data analysis and collaboration. Nat. Rev. Genet. 2018; 19: 208–219.
75. Lau JW, Lehnert E, Sethi A, Malhotra R, Kaushik G, et al. The Cancer Genomics Cloud: Collaborative, Reproducible, and Democratized-A New Paradigm in Large-Scale Computational Research. Cancer Res. 2017; 77: e3–e6.
76. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, et al. GenePattern 2.0. Nat. Genet. 2006; 38: 500–501.
77. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L,

et al. Galaxy: A platform for interactive large-scale genome analysis. Genome Res. 2005; 15: 1451–1455.
78. Okonechnikov K, Golosova O, Fursov M, Ugene Team. Unipro UGENE: A unified bioinformatics toolkit. Bioinformatics. 2012; 28: 1166–1167.
79. Wang L, Lu Z, Van Buren P, Ware D. SciApps: A cloud-based platform for reproducible bioinformatics workflows. Bioinformatics. 2018; 34: 3917–3920.
80. Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res. 2018; 46: W537–W544.
81. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, et al. Origins and functional impact of copy number variation in the human genome. Nature. 2010; 464: 704–712.
82. Handsaker RE, Van Doren V, Berman JR, Genovese G, Kashin S, et al. Large multiallelic copy number variations in humans. Nat. Genet. 2015; 47: 296–303.
83. Zhou B, Ho SS, Zhang X, Pattni R, Haraksingh RR, et al. Whole-genome sequencing analysis of CNV using low-coverage and paired-end strategies is e cient and outperforms array-based CNV analysis. J. Med Genet. 2018; 55: 735–743.
84. Tan R, Wang Y, Kleinstein SE, Liu Y, Zhu X, et al. An evaluation of copy number variation detection tools from whole-exome sequencing data. Hum. Mutat. 2014; 35: 899–907.
85. Liu B, Morrison CD, Johnson CS, Trump DL, Qin M, et al. Computational methods for detecting copy number variations in cancer genome using next generation sequencing: Principles and challenges. Oncotarget. 2013; 4: 1868–1881.
86. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 2011; 21: 974–984.
87. Xie C, Tammi MT. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. BMC Bioinform. 2009; 10: 80.
88. Krumm N, Sudmant PH, Ko A, O'Roak BJ, Malig M, et al.

Copy number variation detection and genotyping from exome sequence data. Genome Res. 2012; 22: 1525–1532.
89. Boeva V, Popova T, Bleakley K, Chiche P, Cappo J, et al. Control-FREEC: A tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics. 2012; 28: 423–425.
90. Sathirapongsasuti JF, Lee H, Horst BA, Brunner G, Cochran AJ, et al. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. Bioinformatics. 2011; 27: 2648–2654.
91. Fromer M, Moran JL, Chambert K, Banks E, Bergen SE, et al. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. Am. J. Hum. Genet. 2012; 91: 597–607.
92. Plagnol V, Curtis J, Epstein M, Mok KY, Stebbings E, et al. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. Bioinformatics. 2012; 28: 2747–2754.
93. Klambauer G, Schwarzbauer K, Mayr A, Clevert DA, Mitterecker A, et al. cn.MOPS: Mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. Nucleic Acids Res. 2012; 40: e69.
94. Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. PLoS Comput. Biol. 2016; 12: e1004873.
95. Li J, Lupat R, Amarasinghe KC, Thompson ER, Doyle MA, et al. CONTRA: Copy number analysis for targeted resequencing. Bioinformatics. 2012; 28: 1307–1313.
96. Favero F, Joshi T, Marquard AM, Birkbak NJ, Krzystanek M, et al. Sequenza: Allele-specific copy number and mutation profiles from tumor sequencing data. Ann. Oncol. O . J. Eur. Soc. Med Oncol. 2015; 26: 64–70.
97. Magi A, Tattini L, Cifola I, D'Aurizio R, Benelli M, et al. EXCAVATOR: Detecting copy number variants from whole-exome sequencing data. Genome Biol. 2013; 14: R120.
98. Jiang Y, Oldridge DA. Diskin SJ, Zhang NR. CODEX: A normalization and copy number variation detection method

for whole exome sequencing. Nucleic Acids Res. 2015; 43: e39.
99. Amarasinghe KC, Li J, Hunter SM, Ryland GL, Cowin PA, et al. Inferring copy number and genotype in tumour exome data. BMC Genom. 2014; 15: 732.
100. Deng X. SeqGene: A comprehensive software solution for mining exome- and transcriptome- sequencing data. BMC Bioinform. 2011; 12: 267.
101. Shi Y, Majewski J. FishingCNV: A graphical software package for detecting rare copy number variations in exome-sequencing data. Bioinformatics. 2013; 29: 1461–1462.
102. Gambin T, Akdemir ZC, Yuan B, Gu S, Chiang T, et al. Homozygous and hemizygous CNV detection from exome sequencing data in a Mendelian disease cohort. Nucleic Acids Res. 2017; 45: 1633–1648.
103. Coin LJ, Cao D, Ren J, Zuo X, Sun L, et al. An exome sequencing pipeline for identifying and genotyping common CNVs associated with disease with application to psoriasis. Bioinformatics. 2012; 28: i370–i374.
104. Packer JS, Maxwell EK, O'Dushlaine C, Lopez AE, Dewey FE, et al. CLAMMS: A scalable algorithm for calling common and rare copy number variants from exome sequencing data. Bioinformatics. 2016; 32: 133–135.
105. Chen H, Bell JM, Zavala NA, Ji HP, Zhang NR. Allele-specific copy number profiling by next-generation DNA sequencing. Nucleic Acids Res. 2015; 43: e23.
106. Zhang Z, Hao K. SAAS-CNV: A Joint Segmentation Approach on Aggregated and Allele Specific Signals for the Identification of Somatic Copy Number Alterations with Next-Generation Sequencing Data. PLoS Comput. Biol. 2015; 11: e1004618.
107. Straver R, Weiss MM, Waisfisz Q, Sistermans EA, Reinders MJT. WISExome: A within-sample comparison approach to detect copy number variations in whole exome sequencing data. Eur. J. Hum. Genet. 2017; 25: 1354–1363.
108. Zare F, Dow M, Monteleone N, Hosny A, Nabavi S. An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. BMC Bioinform. 2017; 18: 286.

109. Kim HY, Choi JW, Lee JY, Kong G. Gene-based comparative analysis of tools for estimating copy number alterations using whole-exome sequencing data. Oncotarget. 2017; 8: 27277–27285.
110. Alkodsi A, Louhimo R, Hautaniemi S. Comparative analysis of methods for identifying somatic copy number alterations from deep sequencing data. Brief. Bioinform. 2015; 16: 242–254.
111. Nam JY, Kim NK, Kim SC, Joung JG, Xi R, et al. Evaluation of somatic copy number estimation tools for whole-exome sequencing data. Brief. Bioinform. 2016; 17: 185–192.
112. Gao J, Wan C, Zhang H, Li A, Zang Q, et al. Anaconda: AN automated pipeline for somatic COpy Number variation Detection and Annotation from tumor exome sequencing data. BMC Bioinform. 2017; 18: 436.
113. Jiang Y, Wang R, Urrutia E, Anastopoulos IN, Nathanson KL, et al. CODEX2: Full-spectrum copy number variation detection by high-throughput DNA sequencing. Genome Biol. 2018; 19: 202.
114. Chatterjee N, Walker GC. Mechanisms of DNA damage, repair, and mutagenesis. Environ. Mol. Mutagen. 2017; 58: 235–263.
115. Shee C, Gibson JL, Rosenberg SM. Two mechanisms produce mutation hotspots at DNA breaks in Escherichia coli. Cell Rep. 2012; 2: 714–721.
116. Hoppe MM, Sundar R, Tan DSP, Jeyasekharan AD. Biomarkers for Homologous Recombination Deficiency in Cancer. J. Natl. Cancer Inst. 2018; 110: 704–713.
117. Davies H, Glodzik D, Morganella S, Yates LR, Staaf J, et al. HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. Nat. Med. 2017; 23: 517–525.
118. Sztupinszki Z, Diossy M, Krzystanek M, Reiniger L, Csabai I, et al. Migrating the SNP array-based homologous recombination deficiency measures to next generation sequencing data of breast cancer. NPJ Breast Cancer. 2018; 4: 16.
119. Riaz N, Blecua P, Lim RS, Shen R, Higginson DS, et al. Pan-cancer analysis of bi-allelic alterations in homologous

recombination DNA repair genes. Nat. Commun. 2017; 8: 857.

120. Hargadon KM, Johnson CE, Williams CJ. Immune checkpoint blockade therapy for cancer: An overview of FDA-approved immune checkpoint inhibitors. Int. Immunopharmacol. 2018; 62: 29–39.

121. Melendez B, Van Campenhout C, Rorive S, Remmelink M, Salmon I, et al. Methods of measurement for tumor mutational burden in tumor tissue. Transl. Lung Cancer Res. 2018; 7: 661–667.

122. Buttner R, Longshore JW, Lopez-Rios F, Merkelbach-Bruse S, Normanno N, et al. Implementing TMB measurement in clinical practice: Considerations on assay requirements. ESMO Open. 2019; 4: e000442.

123. FoCR. Friends of Cancer Research Announces Launch of Phase II TMB Harmonization Project. Washington: FoCR. 2018.

124. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, et al. Signatures of mutational processes in human cancer. Nature. 2013; 500: 415–421.

125. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Ng AW, et al. The Repertoire of Mutational Signatures in Human Cancer. bioRxiv. 2019; 578: 94–101.

126. Rosenthal R, McGranahan N, Herrero J, Taylor BS, Swanton C. DeconstructSigs: Delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biol. 2016; 17: 31.

127. Blokzijl F, Janssen R, van Boxtel R, Cuppen E. MutationalPatterns: Comprehensive genome-wide analysis of mutational processes. Genome Med. 2018; 10: 33.

128. Maura F, Degasperi A, Nadeu F, Leongamornlert D, Davies H, et al. A practical guide for mutational signature analysis in hematological malignancies. Nat. Commun. 2019; 10: 2969.

129. Van Hoeck A, Tjoonk NH, van Boxtel R, Cuppen E. Portrait of a cancer: Mutational signature analyses for cancer diagnostics. BMC Cancer. 2019; 19: 457.

130. Germano G, Lamba S, Rospo G, Barault L, Magri A, et al. Inactivation of DNA repair triggers neoantigen

generation and impairs tumour growth. Nature. 2017; 552: 116–120.
131. Baudrin LG, Deleuze JF, How-Kit A. Molecular and Computational Methods for the Detection of Microsatellite Instability in Cancer. Front. Oncol. 2018; 8: 621.
132. Kautto EA, Bonneville R, Miya J, Yu L, Krook MA, et al. Performance evaluation for rapid detection of pan-cancer microsatellite instability with MANTIS. Oncotarget. 2017; 8: 7452–7463.
133. Huang MN, McPherson JR, Cutcutache I, Teh BT, Tan P, et al. MSIseq: Software for Assessing Microsatellite Instability from Catalogs of Somatic Mutations. Sci. Rep. 2015; 5: 13321.
134. Wang C, Liang C. MSIpred: A python package for tumor microsatellite instability classification from tumor mutation annotation data using a support vector machine. Sci. Rep. 2018; 8: 17546.
135. Foltz SM, Liang WW, Xie M, Ding L. MIRMMR: Binary classification of microsatellite instability using methylation and mutations. Bioinformatics. 2017; 33: 3799–3801.
136. Hause RJ, Pritchard CC, Shendure J, Salipante SJ. Classification and characterization of microsatellite instability across 18 cancer types. Nat. Med. 2016; 22: 1342–1350.
137. Xia J, Chiu LY, Nehring RB, Bravo Nunez MA, Mei Q, et al. Bacteria-to-Human Protein Networks Reveal Origins of Endogenous DNA Damage. Cell. 2019; 176: 127–143.e24.
138. Santarpia L, Bottai G, Kelly CM, Gyor yB, Szekely B, et al. Deciphering and Targeting Oncogenic Mutations and Pathways in Breast Cancer. Oncologist. 2016; 21: 1063–1078.
139. Oh BY, Shin HT, Yun JW, Kim KT, Kim J, et al. Intratumor heterogeneity inferred from targeted deep sequencing as a prognostic indicator. Sci. Rep. 2019; 9: 4542.
140. Goh G, McGranahan N, Wilson GA. Computational Methods for Analysis of Tumor Clonality and Evolutionary History. Methods Mol. Biol. 2019; 1878: 217–226.

141. Miura S, Gomez K, Murillo O, Huuki LA, Vu T, et al. Predicting clone genotypes from tumor bulk sequencing of multiple samples. Bioinformatics. 2018; 34: 4017–4026.
142. Miura S, Vu T, Deng J, Buturla T, Choi J, et al. Power and pitfalls of computational methods for inferring clone phylogenies and mutation orders from bulk sequencing data. bioRxiv 2019; 10: Article number 3498.
143. Pongor L, Harami-Papp H, Mehes E, Czirok A, Gyor yB. Cell Dispersal Influences Tumor Heterogeneity and Introduces a Bias in NGS Data Interpretation. Sci. Rep. 2017; 7: 7358.
144. Yang L, Lee MS, Lu H, Oh DY, Kim YJ, et al. Analyzing Somatic Genome Rearrangements in Human Cancers by Using Whole-Exome Sequencing. Am. J. Hum. Genet. 2016; 98: 843–856.
145. D'Agaro E. Artificial intelligence used in genome analysis studies. EuroBiotech J. 2018; 2: 78–88.