

# We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,900

Open access books available

185,000

International authors and editors

200M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index  
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?  
Contact [book.department@intechopen.com](mailto:book.department@intechopen.com)

Numbers displayed above are based on latest data collected.  
For more information visit [www.intechopen.com](http://www.intechopen.com)



# Bioinformatics Workflows for Genomic Variant Discovery, Interpretation and Prioritization

*Osman Ugur Sezerman, Ege Ulgen, Nogayhan Seymen  
and Ilknur Melis Durasi*

## Abstract

Next-generation sequencing (NGS) techniques allow high-throughput detection of a vast amount of variations in a cost-efficient manner. However, there still are inconsistencies and debates about how to process and analyse this 'big data'. To accurately extract clinically relevant information from genomics data, choosing appropriate tools, knowing how to best utilize them and interpreting the results correctly is crucial. This chapter reviews state-of-the-art bioinformatics approaches in clinically relevant genomic variant detection. Best practices of reads-to-variant discovery workflows for germline and somatic short genomic variants are presented along with the most commonly utilized tools for each step. Additionally, methods for detecting structural variations are overviewed. Finally, approaches and current guidelines for clinical interpretation of genomic variants are discussed. As emphasized in this chapter, data processing and variant discovery steps are relatively well-understood. The differences in prioritization algorithms on the other hand can be perplexing, thus creating a bottleneck during interpretation. This review aims to shed light on the pros and cons of these differences to help experts give more informed decisions.

**Keywords:** genomics, NGS, variant discovery

## 1. Introduction

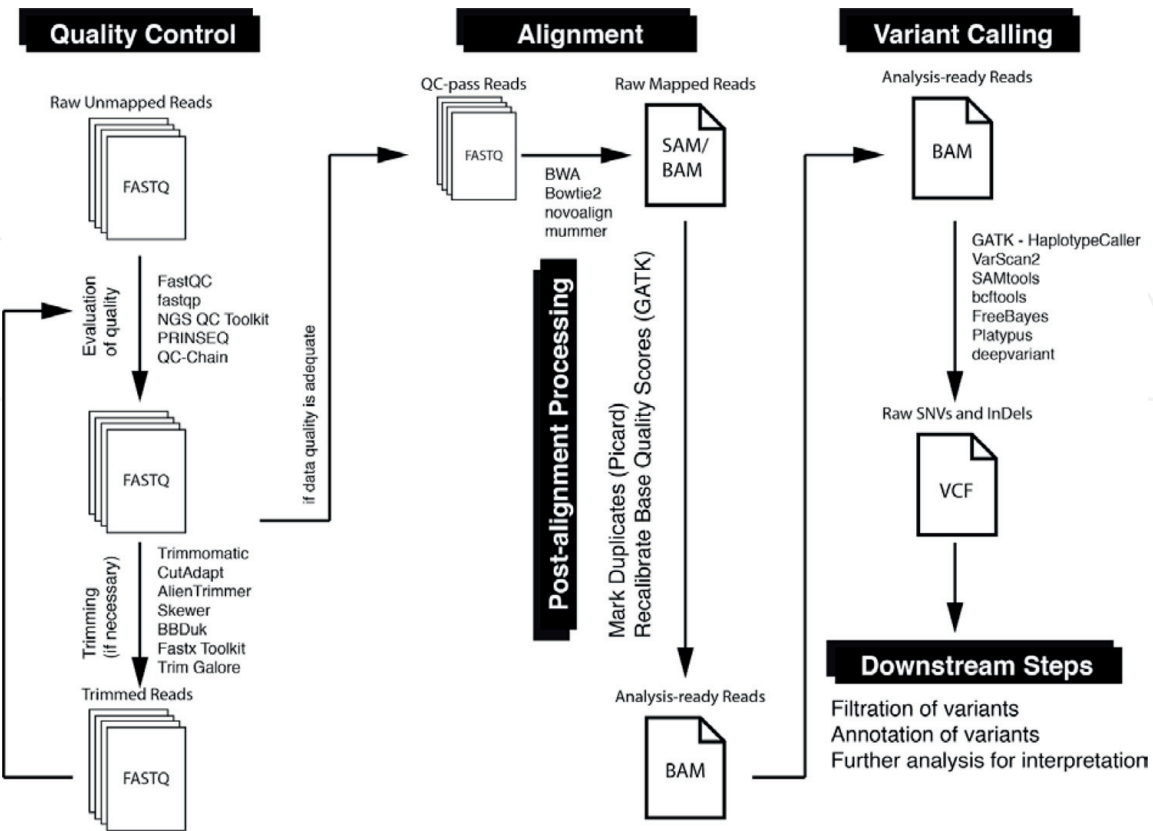
Whole genome sequencing (WGS) and whole exome sequencing are next-generation sequencing (NGS) technologies that determine the full and protein-coding genomic sequence of an organism, respectively. Deep sequencing of genomes improves understanding of clinical interpretation of genomic variations. Analyzing NGS data with the aim of understanding the impact and the importance of genomic variations in health and disease conditions is crucial for carrying the personalized medicine applications one step further.

One of the main obstacles for reaching the full potential of WES/WGS in personalized medicine is bioinformatics analysis, which mostly requires strong computational power. Analysis of WES/WGS data with publicly or commercially available algorithms and tools require a proper computational infrastructure in addition to an at least basic understanding of NGS technologies. Second, almost all publicly available algorithms and tools focus on a single aspect of the entire process

and do not provide a workflow that can aid the researcher from start to finish. Lastly, there are no gold standards for translating WES/WGS into clinical knowledge, since different diseases need different strategies for the basic analysis to obtain the genomic variants as well as further analyses, including disease-specific interpretation and prioritization of the variants.

A comprehensive workflow that can be applied for WES/WGS data analysis is composed of the following steps:

- a. Quality control
  - Evaluation of the quality of FASTQ data
  - Trimming of the low-quality reads and removal of adaptors (if necessary)
- b. Sequence alignment
- c. Post-alignment processing
  - Marking PCR duplicates
  - Base quality score recalibration (BQSR)
- d. Variant discovery
- e. Downstream analyses
  - Filtration of genomic variations



**Figure 1.** An example single-sample variant discovery workflow. Each step is labelled in the black rectangles. The most widely used tools for each operation are also presented.

- Annotation via a variant annotation tool
- Interpretation/prioritization of genomic variations

An example reads-to-variants workflow is visualized in **Figure 1**, highlighting the input and output, a brief description, and the tools that can be utilized in each step. While we present the most widely used tools, we would like to emphasize that there are a great variety of tools/algorithms that can be utilized for each process.

Through the rest of this chapter, we give a brief outline of the purpose of each step and try to provide a basic understanding of a state-of-the-art workflow for the detection and interpretation of genomic variations. While there are countless experimental designs, including WES/WGS and targeted (gene panel) sequencing, the workflow presented here is applicable for all designs, occasionally requiring slight modifications. We particularly focus on the detection and interpretation of germline short variants, namely, single nucleotide variations (SNVs) and germline short insertion or deletion events (indels). However, outlines of analyses for somatic variants and for structural variations (SVs) are also presented. Finally, current approaches and tools for clinical interpretation of genomic variations are discussed.

## 2. Detection of genomic variations

Detection of genomic variations beginning from raw read data is a multistep task that can be executed using numerous tools and resources. The workflow outlined in the introduction section is laid out in detail in this section, including the best practice recommendations and common pitfalls.

### 2.1 Acquisition of raw read data: the FASTQ file format

The raw data from a sequencing machine are most widely provided as FASTQ files, which include sequence information, similar to FASTA files, but additionally contain further information, including sequence quality information.

A FASTQ file consists of blocks, corresponding to reads, and each block consists of four elements in four lines (**Figure 2**).

The first line contains a sequence identifier and includes an optional description of sequencing information (such as machine ID, lane, tile, etc.). The raw sequence letters are presented in line 2. The third line begins with a “+” sign and optionally contains the same sequence identifier. The last line encodes the quality score for the sequence in line 2 in the form of ASCII characters. While specific scoring measures might differ among platforms, Phred Score ( $Q_{\text{phred}} = -10\log_{10}P$ , where P being the probability of misreading any given base) is the most widely used.

### 2.2 Quality control

In general, the raw sequence data acquired from a sequencing provider is not immediately ready to be used for variant discovery. The first and most important

```
@EAS100R:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCAG
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*(((((***+))%%%+))%%%) .1***-+*') **55CCF>>>>>CCCCCCC65
```

**Figure 2.**  
*Example FASTQ file format.*

step of the WES/WGS analysis workflow following data acquisition is the quality control (QC) step. QC is the process of improving raw data by removing any identifiable errors from it. By performing QC at the beginning of the analysis, chances encountering any contamination, bias, error, and missing data are minimized.

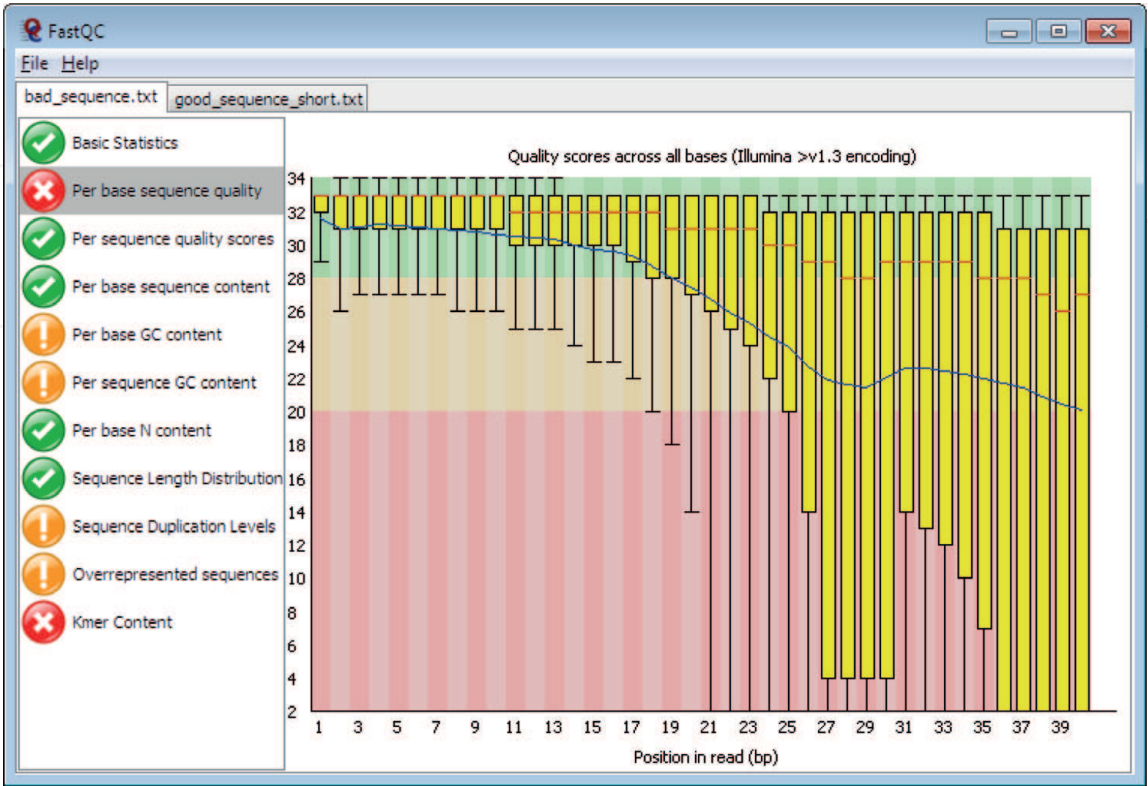
The QC process is a cyclical process, in which (i) the quality is evaluated, (ii) QC is stopped if the quality is adequate, and (iii) a data altering step (e.g., trimming of low-quality reads, removal of adapters, etc.) is performed, and then the QC is repeated beginning from step (i).

The most commonly used tool for evaluating and visualizing the quality of FASTQ data is FastQC (Babraham Bioinformatics, n.d.), which provides comprehensive information about data quality, including but not limited to per base sequence quality scores, GC content information, sequence duplication levels, and overrepresented sequences (**Figure 3**). Alternatives to FastQC include, but are not limited to, fastq, NGS QC Toolkit, PRINSEQ, and QC-Chain.

Below, QC approaches for the most commonly encountered data quality issues are discussed: adapter contamination and low-quality measurements toward the 5' and 3' ends of reads.

Adapters are ligated to the 5' and 3' ends of each single DNA molecule during sequencing. These adapter sequences hold barcoding sequences, forward/reverse primers, and the binding sequences to immobilize the fragments to the flow cell and allow bridge amplification. Since the adapter sequences are synthetic and are not seen in any genomic sequence, adapter contamination often leads to NGS alignment errors and an increased number of unaligned reads. Hence, any adapter sequences need to be removed before mapping. In addition to adapter removal, trimming can be performed to discard any low-quality reads, which generally occur at the 5' and 3' ends.

There is an abundance of tools for QC, namely, Trimmomatic [1], CutAdapt [2], AlienTrimmer [3], Skewer [4], BBDuk [5], Fastx Toolkit [6], and Trim Galore [7].



**Figure 3.**  
An example FastQC result.



In addition to these stand-alone tools, R packages for QC, such as PIQA and ShortRead, are also available.

While QC is the most important step of NGS analysis, one must keep in mind that once basic corrections (such as the ones described above) are made, no amount of further QC can produce a radically better outcome. QC cannot simply turn bad data into good data. Moreover, it is also important to remember that because QC may also introduce error that can affect the analysis, it is vital never to perform error correction on data that does not need it.

### 2.3 Sequence alignment

In order to find the exact locations of reads, each must be aligned to a reference genome. Efficiency and accuracy are crucial in this step because large quantities of reads could take days to align and a low-accuracy alignment would cause inadequate analyses. For humans, the most current and widely used reference sequences are GRCh37 (hg19) and GRCh38 (hg38). Similar to any bioinformatics problem, there are a great number of tools for alignment of sequences to the reference genome, to name a few, BWA [8], Bowtie2 [9], novoalign [10], and mummer [11].

After aligning, a Sequence Alignment Map (SAM) file is produced. This file contains the reads aligned to the reference. The binary version of a SAM file is termed a Binary Alignment Map (BAM) file, and BAM files are utilized for random-access purposes. The SAM/BAM file consists of a header and an alignment section. The header section contains contigs of aligned reference sequence, read groups (carrying platform, library, and sample information), and (optionally) data processing tools applied to the reads. The alignment section includes information on the alignments of reads.

### 2.4 Post-alignment processing

One of the key steps in any reads-to-variants workflow is post-alignment data processing to produce analysis-ready BAM files. This step includes data clean-up operations to correct for technical biases: marking duplicates and recalibration of base quality scores.

During the preparation of samples for sequencing, PCR duplicates arise at the step of PCR amplification of fragments. Since they share the same sequence and the same alignment position, they can lead to problems in variant detection. For example, during SNV calling, false-positive variants may arise as some alleles may be overrepresented due to amplification biases. To overcome this issue, PCR duplicates are marked with a certain tag using an algorithm (MarkDuplicates) available in the tool Picard [12]. Marking duplicates constitutes a major bottleneck since it involves making a large number of comparisons between all the read pairs. Thus, the majority of the effort in optimizing the runtime of reads-to-variants workflows is focused on this step.

As aforementioned, NGS platforms provide information on the quality of each base that they measure in the Phred Score format. The relationship of a Phred Score with accuracy is straightforward: a Phred Score of 10 represents 90% accuracy, 20 equals 99%, 30 equals 99.9%, and so on. The raw scores produced by the sequencing machine are prone to technical errors, leading to over- or underestimated base quality scores. Base quality score recalibration (BQSR) is a machine learning approach that models these errors empirically and readjusts the base quality scores accordingly. Through this recalibration, more accurate and reliable base quality scores are achieved, which in turn improves the reliability of the downstream steps

in further analyses. The most widely used tool for BQSR is provided by the Genome Analysis Toolkit (GATK) [13].

After these post-alignment data processing operations, an analysis-ready BAM file is obtained.

## **2.5 Short variant discovery**

In this section, approaches for the discovery of germline SNV and indels are discussed. In the following sections, approaches for the discovery of somatic short variants and of structural variations are outlined.

Following data processing steps, the reads are ready for downstream analyses, and the following step is most frequently variant calling. Variant calling is the process of identifying differences between the sequencing reads, resulting from NGS experiments and a reference genome. Countless variant callers have been and are being developed for accomplishing this challenging task as alignment and sequencing artifacts complicate the process of variant calling. For recent studies comparing different variant callers, see [14–16]. Methods for detecting short variants can be broadly categorized into “probabilistic methods” and “heuristic-based algorithms.” In probabilistic methods, the distribution of the observed data is modeled, and then Bayesian statistics is utilized to calculate genotype probabilities. In contrast, in heuristic-based algorithms, variant calls are made based on a number of heuristic factors, such as read quality cutoffs, minimum allele counts, and bounds on read depth. Whereas heuristic-based algorithms are not as widely used, they can be robust to outlying data that violate the assumptions of probabilistic models.

The most widely used state-of-the-art variant callers include, but are not limited to, GATK-HaplotypeCaller [13], SOAPsnp [17], SAMTools [18], bcftools [18], Strelka [19], FreeBayes [20], Platypus [21], and DeepVariant [22]. We would like to emphasize that for WES/WGS, a combination of different variant callers outperforms any single method [23].

## **2.6 Filtration of variants**

Following the variant calling step, raw SNV and indels in the Variant Call Format (VCF) are obtained. These should then be filtered either through applying hard filters to the data or through a more complex approach such as GATK’s Variant Quality Score Recalibration (VQSR).

Hard filtering is applied by filtering via thresholds for metrics such as QualByDepth, FisherStrand, RMSMappingQuality, MappingQualityRankSumTest, ReadPosRankSumTest, and StrandOddsRatio.

VQSR, on the other hand, relies on machine learning to identify annotation profiles of variants that are likely to be real. It requires a large training dataset (minimum 30 WES data, at least one WGS data if possible) and well-curated sets of known variants. The aim is to assign a well-calibrated probability to each variant call to create accurate variant quality scores that are then used for filtering.

The accuracy of variant calling is also affected by coverage. Coverage can be broadly defined as the number of unique reads that include a given nucleotide. Coverage is affected by the accuracy of alignment algorithms and by the “mappability” of reads. Coverage can be utilized for both the filtration of variants and for a general evaluation of the sequencing experiment. Tools for assessing coverage information include GATK [13], BEDTools [24], Sambamba [25], and RefCov [26].

For validation of variants and for detecting sequencing artifacts, Integrative Genomics Viewer (IGV) [27] can be used to visualize the processed reads. In addition to this in silico evaluation, Sanger sequencing can be performed.

## 2.7 Variant annotation

Variant annotation is yet another critical step in the WES/WGS analysis workflow. The aim of all functional annotation tools is to annotate information of the variant effects/consequences, including but not limited to (i) listing which gene(s)/transcript(s) are affected, (ii) determination of the consequence on protein sequence, (iii) correlation of the variant with known genomic annotations (e.g., coding sequence, intronic sequence, noncoding RNA, regulatory regions, etc.), and (iv) matching known variants found in variant databases (e.g., dbSNP [28], 1000 Genomes Project [29], ExAc [30], gnomAD [31], COSMIC [32], ClinVar [33], etc.). The consequence of each variant is expressed through Sequence Ontology (SO) terms. The severity and impact of these consequences are often indicated using qualifiers (e.g., low, moderate, high).

Many annotation tools utilize the predictions of SNV/indel deleteriousness prediction methods, to name a few, SIFT [34], PolyPhen-2 [35], LRT [36], MutationTaster [37], MutationAssessor [38], FATHMM [39], GERP++ [40], PhyloP [41], SiPhy [42], PANTHER-PSEP [43], CONDEL [44], CADD [45], CHASM [46], CanDrA [47], and VEST [48].

Annotation can have a strong influence on the ultimate conclusions during interpretation of genomic variations as incomplete or incorrect annotation information will result in the researcher/clinician to overlook potentially relevant findings.

Once the analysis-ready VCF is produced, the genomic variants can then be annotated using a variety of tools and a variety of transcript sets. Both the choice of annotation software and transcript set (e.g., RefSeq transcript set [49], Ensembl transcript set [50]) have been shown to be important for variant annotation [51]. The most widely used functional annotation tools include but are not limited to AnnoVar [52], SnpEff [53], Variant Effect Predictor (VEP) [54], GEMINI [55], VarAFT [56], VAAST [57], TransVar [58], MAGI [59], SNPnexus [60], and VarMatch [61]. Below some of the popular tools are briefly described:

**AnnoVar:** AnnoVar is one of the most popular tools for annotation of SNV and indels. AnnoVar takes a simple text-based format that includes chr, start, end, ref, alt, and optional field(s) as an input. To use AnnoVar, one must convert VCF file format to the AnnoVar input file format. The tool returns a single annotation for each variant. If there exists more than one transcript for a specific variant resulting in different consequences, AnnoVar chooses the transcript according to the gene definition set by the user.

**SnpEff:** SnpEff is an open-source tool that annotates variants and predicts their effects on genes by using an interval forest approach. SnpEff annotates variants based on their genomic locations such as intronic, untranslated region, upstream, downstream, splice site, or intergenic regions and predicts coding effects. snpEff also generates extensive report files and is easily customizable.

**VEP:** VEP is an open-source, free-to-use toolset for the analysis, annotation, and prioritization of genomic variants in coding and noncoding regions. VEP is one of the few annotation tools that annotates variants in regulatory regions.

**GEneome MINIng (GEMINI):** GEMINI is a flexible software package for exploring all forms of human genetic variations. Different from most other annotation tools, GEMINI integrates genetic variation with a set of genome annotations.



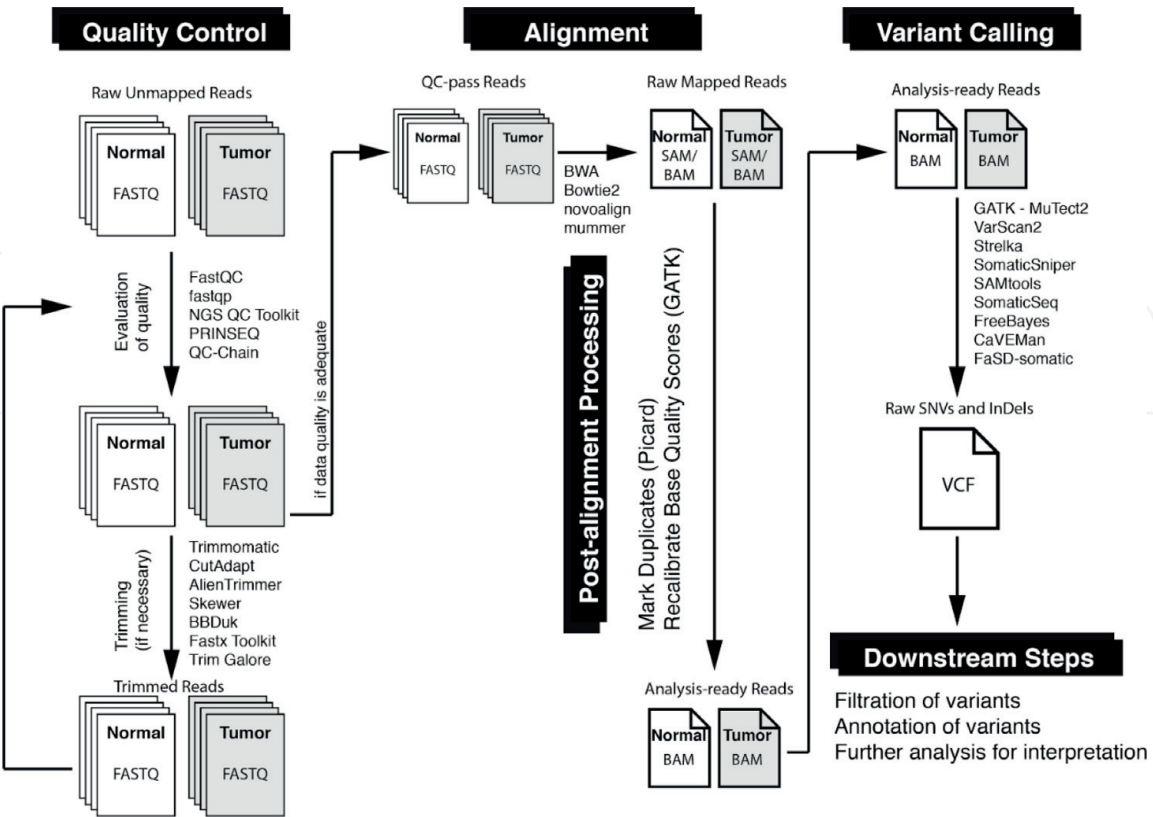
While the abovementioned are all variant annotation tools, it might be wise to put GEMINI in a different category as it has other built-in tools to make further analysis of the variants easier.

2.8 Somatic genomic variations

The workflow for identifying somatic short variants (somatic SNV/indels) is nearly identical to the germline short variant discovery workflow (Figure 4). However, several differences exist. Firstly, for the discovery of somatic genomic variations, sequencing both tumor tissue and a matched normal sample (blood, adjacent normal tissue, etc.) is mostly (but not necessarily) preferred. The QC, alignment, and post-alignment data processing steps are identical and are performed for both the tumor and normal data, separately. The main difference is the variant calling step, where both the tumor and normal processed read data are utilized to identify somatic SNV/indels, i.e., short variants that are present in the tumor but not in the normal. Some tools (such as GATK-MuTect2) can utilize additional information from a panel of normals, a collection of normal samples (typically larger than 40) that are believed to have no somatic variants, processed in the same manner at each step and the purpose of which is to capture recurrent technical artifacts.

Several tools exist for tumor-normal somatic variant calling, to name a few, GATK - MuTect2 [13], VarScan2 [62], Strelka [19], SomaticSniper [63], SAMtools [18], SomaticSeq [64], FreeBayes [65], CaVEMan [66], and FaSD-somatic [67]. For further information on somatic variant calling, we encourage the reader to refer to a recent and comprehensive review on somatic variant calling algorithms [16].

Filtration of the raw SNV and indels also differs for the somatic workflow. Several different approaches exist for the filtration of raw somatic variants. Most



**Figure 4.** An example somatic variant discovery workflow. Each step is labelled in the black rectangles. Most widely used tools for each operation are also presented. As can be seen in the diagram, the processing steps until the variant calling step are performed for both the normal and tumor data, separately.

frequently, tumor-specific metrics including the estimation of tumor heterogeneity and cross-sample contamination are used in addition to the aforementioned metrics for detection of sequencing/alignment artifacts.

While all the annotation tools, presented in the “variant annotation” section, can be used to annotate somatic variants, a number of tools that provide cancer-specific annotation in addition to general annotation are available. Two popular examples are Oncotator [68] and CRAVAT [69]. Oncotator, a widely used cancer-specific annotation tool, is often preferred for the annotation of somatic short variants. Oncotator provides variant- and gene-centric information relevant to cancer researchers, utilizing resources including but not limited to the Catalogue of Somatic Mutations in Cancer (COSMIC) [70], the Cancer Gene Census [71], Cancer Cell Line Encyclopedia [72], The Cancer Genome Atlas (TCGA), and Familial Cancer Database [73].

## 2.9 Structural variations

So far, we focused only on the discovery of small-scale genomic variations (SNVs and indels). There also exist large-scale (1 kb and larger) genomic variations, which either be copy number variations (CNV) or chromosomal rearrangement events (including translocations, inversions, and duplications).

### 2.9.1 Copy number variations

CNV is a frequent form of critical genetic variation that results in an abnormal number of copies of large genomic regions (either gain or loss events). CNV is clinically relevant, as they may play vital roles in disease processes, especially during oncogenesis. It is possible to detect CNVs using WES/WGS data. Several different approaches exist for this purpose [74]:

- i. **Paired-end mapping** strategy detects CNVs through discordantly mapped reads. Tools utilizing this approach include BreakDancer, PEMer, VariationHunter, commonLAW, GASV, and Spanner.
- ii. **Split read-based** methods use the incompletely mapped read from each read pair to identify small CNVs. Split read-based tools include AGE, Pindel, SLOPE, and SRiC.
- iii. **Read depth-based** approach detects CNV by counting the number of reads mapped to each genomic region. Tools using this approach include GATK, SegSeq, CNV-seq, RDXplorer, BIC-seq, CNVseq, cn.MOPS, jointSLM, ReadDepth, rSW-seq, CNVnator, CNVnorm, CMDS, mrCaNaVar, CNVeM, and cnvHMM.
- iv. **Assembly-based** approach detects CNVs by mapping contigs to the reference genome. Tools using this approach include Magnolia, Cortex assembler, and TIGRA-SV.
- v. **Combinatorial** approach combines read depth and paired-end mapping information to detect CNVs. Tools using this approach include NovelSeq, HYDRA, CNVer, GASVPro, Genome STRiP, SVDetect, inGAP-sv, and SVseq.

In addition to the noise and artifacts caused by WES/WGS, tumor complexity (the strongest factor being tumor heterogeneity) makes the detection of somatic

CNVs more challenging. To overcome this challenge, numerous tools have been developed. Widely used tools for detecting specifically somatic CNVs include ADTEX [75], CONTRA [76], cn.MOPS [77], ExomeCNV [78], VarScan2 [62], SynthEx [79], Control-FREEC [80], GATK [13], and CloneCNA [81].

### *2.9.2 Chromosomal rearrangements*

Chromosomal rearrangements are variations in chromosome structure whose impact on genetic diversity and disease susceptibility has become increasingly evident [82]. Per SO, numerous types of rearrangements exist: duplication, deletion, insertion, mobile element insertion, novel sequence insertion, tandem duplication, inversion, intrachromosomal breakpoint, interchromosomal breakpoint, translocation, and complex SVs. Similar to CNV detection, there are multiple approaches for rearrangement detection: read-pair, split-read, read-depth, and assembly approaches. The underlying aims of each of these approaches are very similar to those for CNV detection. As for CNV detection, countless tools have been developed for the detection of rearrangement variations, including but not limited to Breakdancer [83], GRIMM [84], LUMPY [85], BreaKmer [86], BreakSeek [87], CREST [88], DELLY [89], HYDRA [90], MultiBreak-SV [91], Pindel [92], SoftSearch [93], SVdetect [94], and TIGRA-SV [95].

Lastly, we would like to point out that long reads (enabled by the emergence of so-called third-generation sequencing technologies) allow for more accurate and reliable determination of SVs with the development of novel algorithms that specifically exploits these long reads [96].

## **3. Clinical interpretation of genomic variations**

Perhaps the most challenging process in WES/WGS analysis is the clinical interpretation of genomic variations. While WES/WGS is rapidly becoming a routine approach for the diagnosis of monogenic and complex disorders and personalized treatment of such disorders, it is still challenging to interpret the vast amount of genomic variation data detected through WES/WGS [97].

There exist numerous standardized widely accepted guidelines for the evaluation of genomic variations obtained through NGS such as the American College of Medical Genetics and Genomics (ACMG), the EuroGentest, and the European Society of Human Genetics guidelines. These provide standards and guidelines for the interpretation of genomic variations and include evidence-based recommendations on aspects including the use of literature and database and the use of in silico predictors, criteria for variant interpretation, and reporting.

In addition to variant-dependent annotation such as allele frequency (e.g., in 1000 Genomes [29], ExAc [30], gnomAD [31]), the predicted effect on protein and evolutionary conservation, disease-dependent inquiries such as mode of inheritance, co-segregation of variant with disease within families, prior association of the variant/gene with disease, investigation of clinical actionability, and pathway-based analysis are required for the interpretation of genomic variants.

Databases such as ClinVar [33], HGVS databases [98], OMIM [99], COSMIC [100], and CIViC [101] are excellent resources that can aid interpretations of clinical significance of germline and somatic variants for reported conditions. The availability of shared genetic data in such databases makes it possible to identify patients with similar conditions and aid the clinician to make a conclusive diagnosis.

While one may perform interpretation of genomic variations completely manually after annotation and filtering of variants, there are several tools to aid in

interpretation and prioritization of these variants, including Ingenuity Variant Analysis [102], BaseSpace Variant Interpreter [103], VariantStudio [104], KGGSeq [105], PhenoTips [106], VarElect [107], PhenoVar [108], InterVar [109], VarSifter [110], eXtasy [111], VAAST [57], and Exomiser [112]. For personalized oncology purposes, numerous cancer-specific tools exist, specifically developed to determine driver genes/mutations as well as to aid in interpretation of somatic variants. Some of the most widely used somatic interpretation tools are PHIAL, PCGR, and HitWalker2.

Pathway analysis is another powerful component that can enhance the interpretation of genomic variations. Pathway analysis can be broadly defined as a group of methods incorporating biological information from public databases to simplify analysis by grouping long lists of genes into smaller sets of related genes (for a comprehensive review on pathway analysis, see [113]). Pathway analysis improves the detection of causal variants by incorporating biologic insight. The clinician can gain a better understanding of the functions of rare genetic variants of unknown clinical significance in the context of biological pathways. While the gene carrying the variant may not be related to the phenotype, its associated genes in the pathway might be causally related to the phenotype at hand. Moreover, through pathway analysis, the role of multiple variants and their interaction on disease formation can be discovered.

Countless tools for pathway analysis exist. Some of the widely used pathway analysis tools are GSEA [114], DAVID [115], IPA [116], SPIA [117], pathfindR [118], enrichr [119], reactomePA [120], MetaCore [121], and PathVisio [122]. Additionally, many different pathway resources exist, the most popular of which are Kyoto Encyclopedia of Genes and Genomes [123], Reactome [124], WikiPathways [125], MSigDb [126], STRINGDB [127], Pathway Commons [128], Ingenuity Knowledge Base [129], and Pathway Studio [130].

In silico interpretation often fails to provide conclusive evidence for pathogenicity of genomic variations [131]. Furthermore, these in silico interpretations are mostly only well-supported predictions (this is especially true for VUS). It is therefore vital to perform functional validation to understand the functional consequences of genetic variants, provide a conclusive diagnosis, and inform the patient on the disease course. Functional validation can be performed using different model systems (e.g., patient cells, model cell lines, model organisms, induced pluripotent stem cells) and performing the suitable type of assay (e.g., genetic rescue, overexpression, biomarker analysis).

## 4. Conclusion

The advancements in NGS, the increasing availability and applicability of WES/WGS analysis due to decrease in cost, and the development of countless bioinformatics methods and resources enabled the usage of WES/WGS to detect, interpret, and validate genomic variations in the clinical setting.

As we attempted to describe in this chapter, WES/WGS analysis is challenging, and there are a great number of tools for each step of variation discovery. Therefore, one must carefully evaluate the advantages and disadvantages and suitability of different tools (depending on the specific application) before adapting the “optimal” one into the variation discovery workflow. An optimal and coordinated combination of tools is required to identify the different types of genomic variants, described here. On the one hand, an efficient analysis strategy needs to adopt one or more methods for the detection of each type of variant and, on the other hand, needs to integrate results for the different types of variants into a single comprehensive solution.



We attempted to describe the best practices for variant discovery, outlining the fundamental aspects. We hope to have provided a basic understanding of WES/ WGS analysis as we believe awareness of the steps involved as well as the challenges involved at each step is important to understand how each piece may affect the downstream steps (and eventually affect interpretation). As emphasized throughout the chapter, substantial (or even minor) changes at any step can fundamentally alter the outcomes in the later stages.

While there is no definite gold standard for the interpretation of genomic variations, we attempted to briefly describe the currently available and widely used guidelines, tools, and resources for clinical evaluation of genomic variations.


In the following years, with the advancements in bioinformatics, increasing cooperation between the clinician and bioinformatician and large-scale efforts (such as IRDiRC [132], TCGA [133], and ICGC [134]), we expect that a greater focus will be on developing novel tools for clinical interpretation of genomic variations. Cooperation between multiple disciplines is vital to improve the existing approaches as well as to develop novel approaches and resources.

## **Author details**

Osman Ugur Sezerman\*, Ege Ulgen, Nogayhan Seymen and Ilknur Melis Durasi  
Department of Biostatistics and Medical Informatics, Acibadem Mehmet Ali  
Aydinlar University, Istanbul, Turkey

\*Address all correspondence to: [sezermanu@gmail.com](mailto:sezermanu@gmail.com)

## **IntechOpen**

© 2019 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

## References

- [1] Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for illumina sequence data. *Bioinformatics*. 2014;**30**(15):2114-2120
- [2] Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*. 2011;**17**(1):10-12
- [3] Criscuolo A, Brisse S. AlienTrimmer: A tool to quickly and accurately trim off multiple short contaminant sequences from high-throughput sequencing reads. *Genomics*. 2013;**102**(5-6): 500-506
- [4] Jiang H, Lei R, Ding SW, Zhu S. Skewer: A fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics*. 2014;**15**:182
- [5] Available from: <https://jgi.doe.gov/data-and-tools/bbtools/>
- [6] Available from: [http://hannonlab.cshl.edu/fastx\\_toolkit/](http://hannonlab.cshl.edu/fastx_toolkit/)
- [7] Available from: <https://github.com/FelixKrueger/TrimGalore>
- [8] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;**25**(14):1754-1760
- [9] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nature Methods*. 2012;**9**(4):357-359
- [10] Available from: <http://novocraft.com/>
- [11] Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: A fast and versatile genome alignment system. *PLoS Computational Biology*. 2018;**14**(1):e1005944
- [12] Available from: <http://broadinstitute.github.io/picard/>
- [13] Mckenna A, Hanna M, Banks E, et al. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*. 2010;**20**(9): 1297-1303
- [14] Sandmann S, De graaf AO, Karimi M, et al. Evaluating variant calling tools for non-matched next-generation sequencing data. *Scientific Reports*. 2017;**7**:43169
- [15] Bian X, Zhu B, Wang M, et al. Comparing the performance of selected variant callers using synthetic data and genome segmentation. *BMC Bioinformatics*. 2018;**19**(1):429
- [16] Xu C. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Computational and Structural Biotechnology Journal*. 2018;**16**:15-24. DOI: 10.1016/j.csbj.2018.01.003
- [17] Li R, Li Y, Fang X, et al. SNP detection for massively parallel whole-genome resequencing. *Genome Research*. 2009;**19**(6):1124-1132
- [18] Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009;**25**(16):2078-2079
- [19] Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: Accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics*. 2012; **28**(14):1811-1817
- [20] Available from: <https://arxiv.org/abs/1207.3907>
- [21] Rimmer A, Phan H, Mathieson I, et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing

- plications.
- Nature Genetics*
- . 2014;
- 
- 46**
- (8):912-918
- [22] Poplin R, Chang P-C, Alexander D, et al. A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology*. 2018;  
**36**(10):983-987. DOI: 10.1038/nbt.4235
- [23] Bao R, Huang L, Andrade J, et al. Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing. *Cancer Informatics*. 2014;**13**(Suppl 2):67-82
- [24] Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;  
**26**(6):841-842
- [25] Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: Fast processing of NGS alignment formats. *Bioinformatics*. 2015;**31**(12):2032-2034
- [26] Available from: <http://gmt.genome.wustl.edu/gmt-refcov>
- [27] Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative genomics viewer. *Nature Biotechnology*. 2011;  
**29**(1):24-26
- [28] Sherry ST, Ward MH, Kholodov M, et al. dbSNP: The NCBI database of genetic variation. *Nucleic Acids Research*. 2001;**29**(1):308-311
- [29] Auton A, Brooks LD, Durbin RM, et al. A global reference for human genetic variation. *Nature*. 2015;  
**526**(7571):68-74
- [30] Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2016;**536**(7616):285-291
- [31] Available from: <https://gnomad.broadinstitute.org/>
- [32] Tate JG, Bamford S, Jubb HC, et al. COSMIC: The catalogue of somatic mutations in cancer. *Nucleic Acids Research*. 2019;**47**(D1):D941-D947
- [33] Landrum MJ, Lee JM, Riley GR, et al. ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*. 2014;**42**(Database issue): D980-D985
- [34] Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Research*. 2003;  
**31**(13):3812-3814
- [35] Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. *Nature Methods*. 2010;**7**(4):248-249
- [36] Chun S, Fay JC. Identification of deleterious mutations within three human genomes. *Genome Research*. 2009;**19**(9):1553-1561
- [37] Schwarz JM, Rödelberger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nature Methods*. 2010;**7**:575-576
- [38] Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: Application to cancer genomics. *Nucleic Acids Research*. 2011;  
**39**(17):e118
- [39] Shihab HA, Gough J, Cooper DN, et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Human Mutation*. 2013;**34**(1): 57-65
- [40] Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Computational Biology*. 2010;**6**(12):e1001025
- [41] Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral

- p substitution rates on mammalian phylogenies.
- Genome Research*
- . 2010;
- 20**
- (1):110-121
- [42] Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics*. 2009; **25**(12): i54-i62
- [43] Tang H, Thomas PD. PANTHER-PSEP: Predicting disease-causing genetic variants using position-specific evolutionary preservation. *Bioinformatics*. 2016; **32**(14):2230-2232
- [44] González-pérez A, López-bigas N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *American Journal of Human Genetics*. 2011; **88**(4):440-449
- [45] Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*. 2019; **47**(D1): D886-D894
- [46] Wong WC, Kim D, Carter H, Diekhans M, Ryan MC, Karchin R. CHASM and SNVBox: Toolkit for detecting biologically important single nucleotide mutations in cancer. *Bioinformatics*. 2011; **27**(15):2147-2148
- [47] Mao Y, Chen H, Liang H, Meric-bernstam F, Mills GB, Chen K. CanDrA: Cancer-specific driver missense mutation annotation with optimized features. *PLoS One*. 2013; **8**(10):e77945
- [48] Carter H, Douville C, Yeo G, Stenson PD, Cooper DN, Karchin R. Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics*. 2013; **14**(3):1-16
- [49] O'leary NA, Wright MW, Brister JR, et al. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*. 2016; **44**(D1):D733-D745
- [50] Zerbino DR, Achuthan P, Akanni W, et al. Ensembl 2018. *Nucleic Acids Research*. 2018; **46**(D1):D754-D761
- [51] McCarthy DJ, Humburg P, Kanapin A, et al. Choice of transcripts and software has a large effect on variant annotation. *Genome Medicine*. 2014; **6**(3):26
- [52] Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from next-generation sequencing data. *Nucleic Acids Research*. 2010; **38**:e164
- [53] Cingolani P, Platts A, Wang le L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)*. 2012; **6**(2):80-92
- [54] McLaren W, Gil L, Hunt SE, et al. The Ensembl variant effect predictor. *Genome Biology*. 2016; **17**(1):122
- [55] Paila U, Chapman BA, Kirchner R, Quinlan AR. GEMINI: Integrative exploration of genetic variation and genome annotations. *PLoS Computational Biology*. 2013; **9**(7): e1003153
- [56] Desvignes JP, Bartoli M, Delague V, et al. VarAFT: A variant annotation and filtration system for human next generation sequencing data. *Nucleic Acids Research*. 2018; **46**(W1):W545-W553
- [57] Hu H, Huff CD, Moore B, Flygare S, Reese MG, Yandell M. VAAST 2.0: Improved variant classification and disease-gene identification using a conservation-controlled amino acid substitution matrix. *Genetic Epidemiology*. 2013; **37**(6):622-634



- [58] Zhou W, Chen T, Chong Z, et al. TransVar: A multilevel variant annotator for precision genomics. *Nature Methods*. 2015;12(11):1002-1003
- [59] Leiserson MD, Gramazio CC, Hu J, Wu HT, Laidlaw DH, Raphael BJ. MAGI: Visualization and collaborative annotation of genomic aberrations. *Nature Methods*. 2015;12(6):483-484
- [60] Dayem Ullah AZ, Oscanoa J, Wang J, Nagano A, Lemoine NR, Chelala C. SNPnexus: Assessing the functional relevance of genetic variation to facilitate the promise of precision medicine. *Nucleic Acids Research*. 2018; 46(W1):W109-W113
- [61] Sun C, Medvedev P. VarMatch: Robust matching of small variant datasets using flexible scoring schemes. *Bioinformatics*. 2017;33(9):1301-1308
- [62] Koboldt DC, Zhang Q, Larson DE, et al. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*. 2012;22(3):568-576
- [63] Larson DE, Harris CC, Chen K, et al. SomaticSniper: Identification of somatic point mutations in whole genome sequencing data. *Bioinformatics*. 2012; 28(3):311-317
- [64] Fang LT, Afshar PT, Chhibber A, Mohiyuddin M, Fan Y, Mu JC, et al. An ensemble approach to accurately detect somatic mutations using SomaticSeq. *Genome Biology*. 2015;16:197
- [65] Available from: <http://arxiv.org/abs/1207.3907>
- [66] Stephens PJ et al. The landscape of cancer genes and mutational processes in breast cancer. *Nature*. 2012;486: 400-404
- [67] Wang W, Wang P, Xu F, Luo R, Wong MP, Lam T-W. FaSD-somatic: A fast and accurate somatic SNV detection algorithm for cancer genome sequencing data. *Bioinformatics*. 2014; 30(17):2498-2500
- [68] Ramos AH, Lichtenstein L, Gupta M, Lawrence MS, Pugh TJ, Saksena G, et al. Oncotator: Cancer variant annotation tool. *Human Mutation*. 2015; 36(4):E2423-E24E9. pmid: 25703262
- [69] Douville C, Carter H, Kim R, et al. CRAVAT: Cancer-related analysis of variants toolkit. *Bioinformatics*. 2013; 29(5):647-648
- [70] Forbes SA, Beare D, Boutselakis H, et al. COSMIC: Somatic cancer genetics at high-resolution. *Nucleic Acids Research*. 2016;45(D1):D777-D783
- [71] Futreal PA, Andrew Futreal P, Coin L, Marshall M, Down T, Hubbard T, et al. A census of human cancer genes. *Nature Reviews. Cancer*. 2004;4:177-183
- [72] Barretina J, Caponigro G, Stransky N, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*. 2012;483(7391):603-607. Published 2012 Mar 28. DOI: 10.1038/nature11003
- [73] Sijmons RH. Identifying Patients with Familial Cancer Syndromes. 2010 Feb 27 [Updated 2010 Feb 27]. In: Riegert-Johnson DL, Boardman LA, Hefferon T, et al., editors. *Cancer Syndromes* [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2009. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK45295/>
- [74] Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: Features and perspectives. *BMC Bioinformatics*. 2013;14(11):S1
- [75] Amarasinghe KC, Li J, Hunter SM, et al. Inferring copy number and

- p>genotype in tumour exome data. BMC Genomics. 2014;
- 15**
- :732
- [76] Hooghe B, Hulpiau P, van Roy F, De Bleser P. ConTra: A promoter alignment analysis tool for identification of transcription factor binding sites across species. Nucleic Acids Research. 2008; **36**:W128-W132
- [77] Klambauer G, Schwarzbauer K, Mayr 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 Research. 2012;**40**(9):e69
- [78] Sathirapongsasuti JF, Lee H, Horst BA, et al. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. Bioinformatics. 2011;**27**(19):2648-2654
- [79] Silva GO, Siegel MB, Mose LE, et al. SynthEx: A synthetic-normal-based DNA sequencing tool for copy number alteration detection and tumor heterogeneity profiling. Genome Biology. 2017;**18**(1):66
- [80] Boeva V, Popova T, Bleakley K, et al. Control-FREEC: A tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics. 2012; **28**(3):423-425
- [81] Yu Z, Li A, Wang M. CloneCNA: Detecting subclonal somatic copy number alterations in heterogeneous tumor samples from whole-exome sequencing data. BMC Bioinformatics. 2016;**17**:310
- [82] Sedlazeck FJ, Dhroso A, Bodian DL, Paschall J, Hermes F, Zook JM. Tools for annotation and comparison of structural variation. F1000Res. 2017;**6**:1795
- [83] Fan X, Abbott TE, Larson D, Chen K. BreakDancer: Identification of genomic structural variation from paired-end read mapping. Current Protocols in Bioinformatics. 2014;**45**:15.6.1-15.611
- [84] Tesler G. GRIMM: Genome rearrangements web server. Bioinformatics. 2002;**18**(3):492-493
- [85] Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: A probabilistic framework for structural variant discovery. Genome Biology. 2014;**15**(6):R84
- [86] Abo RP, Ducar M, Garcia EP, et al. BreaKmer: Detection of structural variation in targeted massively parallel sequencing data using kmers. Nucleic Acids Research. 2014;**43**(3):e19
- [87] Zhao H, Zhao F. BreakSeek: A breakpoint-based algorithm for full spectral range INDEL detection. Nucleic Acids Research. 2015;**43**(14):6701-6713
- [88] Wang J, Mullighan CG, Easton J, et al. CREST maps somatic structural variation in cancer genomes with base-pair resolution. Nature Methods. 2011; **8**(8):652-654. Published 2011 Jun 12. DOI: 10.1038/nmeth.1628
- [89] Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: Structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012;**28**(18):i333-i339
- [90] Miller PL, Blumenfrucht SJ, Rose JR, et al. HYDRA: A knowledge acquisition tool for expert systems that critique medical workup. Medical Decision Making. 1987;**7**(1):12-21
- [91] Ritz A, Bashir A, Sindi S, Hsu D, Hajirasouliha I, Raphael BJ. Characterization of structural variants with single molecule and hybrid sequencing approaches. Bioinformatics. 2014;**30**(24):3458-3466
- [92] Ye K, Schulz MH, 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**(21):2865-2871
- [93] Hart SN, Sarangi V, Moore R, et al. SoftSearch: Integration of multiple sequence features to identify breakpoints of structural variations. *PLoS One*. 2013;**8**(12):e83356
- [94] Zeitouni B, Boeva V, Janoueix-Lerosey I, et al. SVDetect: A tool to identify genomic structural variations from paired-end and mate-pair sequencing data. *Bioinformatics*. 2010;**26**(15):1895-1896
- [95] Chen K, Chen L, Fan X, Wallis J, Ding L, Weinstock G. TIGRA: A targeted iterative graph routing assembler for breakpoint assembly. *Genome Research*. 2014;**24**(2):310-317
- [96] Guan P, Sung WK. Structural variation detection using next-generation sequencing data: A comparative technical review. *Methods*. 2016;**102**:36-49
- [97] Sayitoğlu M. Clinical interpretation of genomic variations. *Turkish Journal of Haematology*. 2016;**33**(3):172-179
- [98] Freeman PJ, Hart RK, Gretton LJ, Brookes AJ, Dalgleish R. VariantValidator: Accurate validation, mapping, and formatting of sequence variation descriptions. *Human Mutation*. 2017;**39**(1):61-68
- [99] Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*. 2002;**30**(1):52-55
- [100] Bamford S, Dawson E, Forbes S, et al. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *British Journal of Cancer*. 2004;**91**(2):355-358
- [101] Griffith M, Spies NC, Krysiak K, et al. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nature Genetics*. 2017;**49**(2):170-174
- [102] Available from: [www.ingenuity.com](http://www.ingenuity.com)
- [103] Available from: <https://www.illumina.com/products/by-type/informatics-products/basespace-variant-interpreter.html>
- [104] Available from: <https://www.bioz.com/result/VariantStudio%20variant/product/Illumina>
- [105] Li MX, Gui HS, Kwan JS, Bao SY, Sham PC. A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. *Nucleic Acids Research*. 2012;**40**(7):e53
- [106] Girdea M, Dumitriu S, Fiume M, et al. PhenoTips: Patient phenotyping software for clinical and research use. *Human Mutation*. 2013;**34**(8):1057-1065
- [107] Stelzer G, Plaschkes I, Oz-levi D, et al. VarElect: The phenotype-based variation prioritizer of the GeneCards Suite. *BMC Genomics*. 2016;**17**(Suppl 2):444
- [108] Trakadis YJ, Buote C, Therriault JF, Jacques PÉ, Larochelle H, Lévesque S. PhenoVar: A phenotype-driven approach in clinical genomics for the diagnosis of polymalformative syndromes. *BMC Medical Genomics*. 2014;**7**:22. Published 2014 May 12. DOI: 10.1186/1755-8794-7-22
- [109] Li Q, Wang K. InterVar: Clinical interpretation of genetic variants by the 2015 ACMG-AMP guidelines. *American Journal of Human Genetics*. 2017;**100**(2):267-280
- [110] Teer JK, Green ED, Mullikin JC, Biesecker LG. VarSifter: Visualizing and



analyzing exome-scale sequence variation data on a desktop computer. *Bioinformatics*. 2011;28(4):599-600

[111] Sifrim A, Popovic D, Tranchevent LC, Ardeshirdavani A, Sakai R, Konings P, et al. eXtasy: Variant prioritization by genomic data fusion. *Nature Methods*. 2013;10(11):1083-1084. DOI: 10.1038/nmeth.2656

[112] Smedley D, Jacobsen JO, Jäger M, et al. Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nature Protocols*. 2015; 10(12):2004-2015

[113] García-Campos MA, Espinal-Enríquez J, Hernández-Lemus E. Pathway analysis: State of the art. *Frontiers in Physiology*. 2015;6:383. doi: 10.3389/fphys.2015.00383

[114] Powers RK, Goodspeed A, Pielke-Lombardo H, Tan AC, Costello JC. GSEA-InContext: Identifying novel and common patterns in expression experiments. *Bioinformatics*. 2018; 34(13):i555-i564

[115] Dennis G, Sherman BT, Hosack DA, et al. DAVID: Database for annotation, visualization, and integrated discovery. *Genome Biology*. 2003;4(9): R60

[116] Yu J, Gu X, Yi S. Ingenuity pathway analysis of gene expression profiles in distal Nerve stump following nerve injury: Insights into Wallerian degeneration. *Frontiers in Cellular Neuroscience*. 2016;10:274. Published 2016 Dec 6. DOI: 10.3389/fncel.2016.00274

[117] Tarca AL, Draghici S, Khatri P, et al. A novel signaling pathway impact analysis. *Bioinformatics*. 2008;25(1): 75-82

[118] Ulgen E, Ozisik O, Sezerman OU. pathfindR: An R Package for Pathway Enrichment Analysis Utilizing Active Subnetworks. *bioRxiv*. 2018

[119] Kuleshov MV, Jones MR, Rouillard AD, et al. Enrichr: A comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*. 2016;44(W1):W90-W97

[120] Yu G, HeReactomePA Q-Y. An R/Bioconductor package for reactome pathway analysis and visualization. *Molecular BioSystems*. 2016;12:477-479

[121] Available from: <https://clarivate.com/products/metacore/>

[122] Kutmon M, Van iersel MP, Bohler A, et al. PathVisio 3: An extendable pathway analysis toolbox. *PLoS Computational Biology*. 2015;11(2): e1004085

[123] Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*. 2000;28(1): 27-30

[124] Croft D, O'Kelly G, Wu G, et al. Reactome: A database of reactions, pathways and biological processes. *Nucleic Acids Research*. 2010;39 (Database issue):D691-D697

[125] Kelder T, van Iersel MP, Hanspers K, et al. WikiPathways: Building research communities on biological pathways. *Nucleic Acids Research*. 2011; 40(Database issue):D1301-D1307

[126] Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*. 2011;27(12):1739-1740

[127] Szklarczyk D, Morris JH, Cook H, et al. The STRING database in 2017: Quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Research*. 2016;45(D1):D362-D368

[128] Cerami EG, Gross BE, Demir E, et al. Pathway commons, a web resource



for biological pathway data. Nucleic Acids Research. 2010;**39**(Database issue):D685-D690

[129] Available from: <http://www.ingenuity.com/>

[130] Nikitin A, Egorov S, Daraselia N, Mazo I. Pathway studio—The analysis and navigation of molecular networks. Bioinformatics. 2003;**19**(16):2155-2157

[131] Rodenburg RJ. The functional genomics laboratory: Functional validation of genetic variants. Journal of Inherited Metabolic Disease. 2018; **41**(3):297-307

[132] Austin CP, Cuttillo CM, Lau LPL, et al. Future of rare diseases research 2017–2027: An IRDiRC perspective. Clinical and Translational Science. 2018; **11**(1):21-27

[133] Available from: <https://cancergenome.nih.gov/>

[134] Zhang J, Baran J, Cros A, et al. International cancer genome consortium data portal—A one-stop shop for cancer genomics data. Database: The Journal of Biological Databases and Curation. 2011;**2011**: bar026