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

4,200
Open access books available

116,000
International authors and editors

125M
Downloads

154
Countries delivered to

TOP 1%
Our authors are among the 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

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com
Chapter

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 emphasised 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...
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, fastqc, 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].
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 used widely, 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.

**GEnome 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, CNASeq, 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 Magnolya, 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], HGV 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], Phenomen [106], VarElect [107], PhenoVar [108], InterVar [109], VarSifter [110], eXtsuy [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

© 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


DOI: http://dx.doi.org/10.5772/intechopen.85524
applications. Nature Genetics. 2014; 46(8):912-918


[34] Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Research. 2003; 31(13):3812-3814


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


[75] Amarasinghe KC, Li J, Hunter SM, et al. Inferring copy number and
genotype in tumour exome data. BMC Genomics. 2014;15:732


[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


[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


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


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


