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

5,800
Open access books available

144,000
International authors and editors

180M
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

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

Chapter

Integrating Evolutionary Genetics to Medical Genomics: Evolutionary Approaches to Investigate Disease-Causing Variants

Ugur Sezerman, Tugce Bozkurt and Fatma Sadife Isleyen

Abstract

In recent years, next-generation sequencing (NGS) platforms that facilitate generation of a vast amount of genomic variation data have become widely used for diagnostic purposes in medicine. However, identifying the potential effects of the variations and their association with a particular disease phenotype is the main challenge in this field. Several strategies are used to discover the causative mutations among hundreds of variants of uncertain significance. Incorporating information from healthy population databases, other organisms' databases, and computational prediction tools are evolution-based strategies that give valuable insight to interpret the variant pathogenicity. In this chapter, we first provide an overview of NGS analysis workflow. Then, we review how evolutionary principles can be integrated into the prioritization schemes of analyzed variants. Finally, we present an example of a real-life case where the use of evolutionary genetics information facilitated the discovery of disease-causing variants in medical genomics.

Keywords: genomics, evolution, variant discovery

1. Introduction

NGS technologies can be integrated into medical diagnostics in several ways which vary in the number and type of sequenced regions. While targeted tests include sequencing particular disease-specific genes, sequencing all ~20,000 protein-coding genes by Whole-exome sequencing (WES) and entire genomes by Whole-genome Sequencing (WGS) are non-targeted approaches. These sequencing approaches are precise ways to detect genetic variation of a patient and in relation to a healthy population or healthy reference genome. However, sequencing-based diagnostic methods generate large amounts of genomic data. Approximately, 60,000–100,000 single nucleotide variations (SNV) and small insertions and deletions (indel) in each patient's personal genome can be detected on WES [1]. Translating these high numbers of genomic variants into useful clinical information is a crucial task. Although several methods have been introduced to help reduce the vast number of possible genes to clinically causative ones, this process still remains challenging.

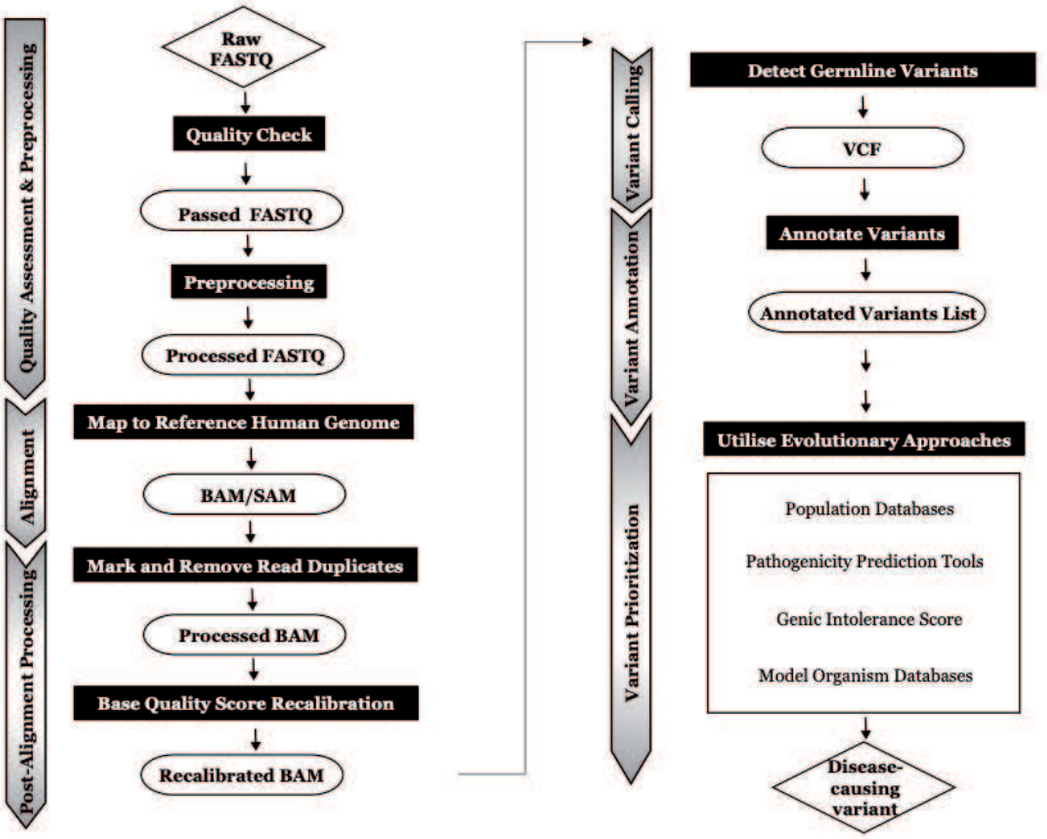

Figure 1.
A general workflow for WES data analysis. Six main steps, quality assessment & preprocessing, alignment, post-alignment processing, variant calling, variant annotation, and variant prioritization integrated with evolutionary approaches, are shown.

Disease-related genes show non-random distribution characteristics in the genome with the majority of them being already present in the eukaryotic ancestor [2]. Mendelian disease genes that underlie single-gene disorders tend to have a more ancient evolutionary origin [3]. Considering disease-related genes have evolved under the effect of natural selection like other genes, evolutionary approaches can provide powerful insight not only to understand human genetic diseases but also to detect genomic variants that cause them.

Here, we briefly describe the analysis workflow from raw data to genomic variants as the first step of the translation to the clinical outcome. We primarily focus on WES analysis because most variations that are responsible for Mendelian disorders disrupt protein-coding regions [4]. Then we give an insight into how evolutionary principles are integrated into the prioritization of detected variants. The framework of the chapter can be found in **Figure 1.**

2. From raw data to genomic variations

The common file format for the storage of data produced by sequencers is FASTQ [5]. FASTQ format stores both nucleotide sequence and its corresponding Phred quality scores [6, 7]. The Phred score related to the base-calling error probabilities indicates the quality of each nucleotide within a read. In a FASTQ file, each read is shown by four lines: The first line begins with a "@" and continues with a sequence identifier and an optional description. The second line consists of the raw sequence letters: A, T, G, C, and N (unknown). The third line starts with a "+"

character and can be followed by the same sequence identifier again. The "+" sign specifies the end of the sequence. The fourth line includes the quality scores for the sequence in the second line.

Here, we give an overview of the data analysis workflow from a FASTQ file to obtain annotated genomic variants.

2.1 Quality assessment and preprocessing

Although NGS platforms are capable of generating massively parallel sequences even in a single run, the quality of sequencing reads may not be perfect due to some reasons such as the failure in experimental processing and technical machine errors. The quality of raw FASTQ data should be assessed in the first step of the workflow since these errors affect downstream analysis.

A number of tools have been developed to evaluate raw FASTQ data. These tools generally take FASTQ files as input and generate summary statistics and graphs for a quick overview of the raw read quality. In addition to the most commonly used one FASTQC [8], developed by Simon Andrews at Babraham Institute, other tools are also available such as FQStat [9], Quack [10], SeqAssist [11], QC-Chain [12]. Based on the result of the quality check step, if there is a need, preprocessing is necessary before alignment.

The standard preprocessing step consists of trimming of low-quality bases and adapter sequence removal at the end of the reads. Adapter sequences can be ligated to 3′ and 5′ ends of reads depending on the used library preparation protocol during the sequencing. These adapter fragments should be removed correctly because of leading to either missed alignments or wrong genotyping in further downstream analyses. Many tools with different principles of implementation have been developed to perform preprocessing. Ktrim [13], PE-Trimmer [14], SeqPurge [15], AdapterRemoval [16], PEAT [17], Skewer [18], Trimmomatic [19], QcReads [20], AlienTrimmer [21], and Btrim [22] are tools can be used for adapter and quality trimming depending on the study design. In addition to these, some tools such as FastqCleaner [23], FastProNGS [24], EasyQC [25], fastp [26], TrimGalore, FASTX-Toolkit, afterQC, ClinQC, NGS QC Toolkit, PRINSEQ, fastQ_brew carry out both quality check and preprocessing functions.

2.2 Alignment of reads

After quality check and preprocessing of raw data, processed reads must be aligned to the reference genome. Both GRCh37 (hg19) and GRCh38 (hg38) are widely used as a reference for the human genome. Optimal alignment to reference sequences is not easy computational task and requires a fast and tolerant algorithm to obtain an imperfect alignment due to genomic variations. Several tools have been developed to align short reads. They mainly use the Burrows-Wheeler Transformation (BWT) algorithm, the Smith-Waterman (SW) dynamic programming algorithm or a combination of both of them. Bowtie2 [27] and BWA [28], which implement the BWT algorithm, are widely used for short reads alignment. Novoalign [29], MOSAIK [30], and SHRiMP2 [31] implement SW algorithm. For a comprehensive review of these methods and their differences, benchmark studies can be found in the literature [32, 33].

The output of the alignment step is the Sequence Alignment Map (SAM) file which contains mapped reads. BAM stands for Binary Alignment Map and is the binary version of a SAM file. Both BAM files and SAM files have the same information which include a header and an alignment section. The header section

provides some information such as reference sequence, read group, sequencing platform details and applied process information to the reads. The alignment section includes the genomic position with relevant descriptive information of each sequence.

SAMtools [34] and Integrative Genomics Viewer (IGV) [35] are also commonly used programs to view BAM/SAM files for further confirmation analysis of detected variants.

2.3 Post-alignment processing

Processing of aligned reads is recommended to improve the quality of downstream variant calling analysis. The processing step generally consists of marking read duplicates and base quality score recalibration (BQSR) to minimize technical biases.

During the sequencing, a library of DNA fragments from a particular genomic region is prepared using PCR amplification to provide adequate DNA fragments for the sequencing process. Therefore, some amplified fragments could share the same sequence and the same corresponding alignment position leading to bias in variant detection. These duplicates should be removed to eliminate PCR-introduced bias. MarkDuplicates available in the Picard [36] and SAMtools [34] are widely-used tools to detect read duplicates based on their identical 5′ region and position on the genome.

In addition to marking duplicates, base quality is also an important factor for variant detection. As mentioned in the section "Quality check and preprocessing", each sequence read has a Phred quality score generated by the sequencing machine. However, the machine could generate systematically biased scores. On the contrary, BQSR patterns errors empirically to recalibrate the base quality scores using a machine learning approach. Thus, technical bias is significantly minimized. The key point in this process is to exclude known variants before BQSR since they are true genomic variations. So, they should not be considered as sequencing errors. The most widely used tool for recalibration of base qualities is BaseRecalibrator available in Genome Analysis Toolkit (GATK) [37].

2.4 Variant calling

After the post-alignment processing step, variant analysis can be started on an analysis-ready BAM file. In the variant calling step, the differences between the reference genome and genome of interest are calculated. Variants can be categorized as germline and somatic variants while dealing with variant calling. Germline variants are inherited variations present in the germ cells. Somatic variants are present only in somatic cells and can be specific to a tissue. In this chapter, we focus on the identification of germline SNV and indels. Several tools based on different algorithms have been developed to call germline short variants. Tools such as HaplotypeCaller available in GATK [38], SAMtools [34], FreeBayes [39], and Platypus [40] are based on Bayesian approaches. VarScan [41] relies on a heuristic approach to identify variants, while SNVer [42] uses a frequentist approach. The performance of different tools has been evaluated by recent studies [43–45], yet, these tools mostly generate an analysis-ready VCF (Variant Call Format) file. A VCF file is a text file that contains header lines and data lines. The header lines begin with "##" symbol. The first header line is always the VCF format version and continues with lines defining the name, length, value type, and description of each item in relevant fields of each data line.

2.5 Variant annotation and prioritization

After variants are detected, biologically important features such as gene symbols, genomic position, amino acid change, and consequences of variants add to a VCF file in the annotation step. In addition to the basic annotation, several tools can be used to integrate the annotations from countless sources including information of known variants with minor allele frequency (MAF) found in public databases and pathogenicity prediction of variants. There are numerous variant annotation tools that implement different methods and most widely used ones are AnnoVar [46], VEP [47], SnpEff [48], GEMINI [49], VarAFT [50], AnnTools [51], SVA [52], NGS-SNP [53]. These annotation tools enable to filtering and prioritizing potential disease-causing mutations. The prioritization of clinically causative mutation among a vast amount of annotated variations is the most challenging part of the analysis and is not a fully automatized. In the next section, we are going to discuss how evolutionary approaches can be used to prioritize genomic variants.

3. Utilizing evolutionary information in variant prioritization

We have described the process of obtaining annotated variations from raw FASTQ data. Experimentally evaluation of each variant at a genomic scale would be an impractical process, but evolutionary principles can provide us a valuable set of an experiment from nature. Integrating evolutionary approaches into the prioritization step have the potential to distinguish the variant responsible for a particular disease among all annotated variants. Indeed, the association between disease and evolution has been attributed to natural selection [54, 55]. During evolution, variations at highly conserved genomic regions are exposed to natural selection because of their negative impact on fitness that make these conserved genes intolerant to variations [56]. On the contrary, at the faster-evolving regions of the genome, many variations have been tolerated over evolutionary time and accumulate in the population with high MAF. However, there is a predisposition for Mendelian disease genes to be more intolerant than the other genes [57]. These genes are also more conserved across species allowing us to compare the phenotypes of different mutant genes on a multispecies level [58].

In this part, we discuss the role of evolutionary approaches in variant prioritization. The first prioritization method aims to filter variants using information from allele frequencies in population databases. Then we introduce several pathogenicity prediction tools to interpret the rest of the variants, especially the ones with uncertain significance. Following that, we describe the usage of gene intolerance information while making inference the variant pathogenicity. Finally, we list commonly used model organism databases that can be used for the comparison of mutant gene phenotypes in several species.

3.1 Population databases

During human evolution, present and novel variations have been evaluated in terms of their biological impact. Population databases record the outcomes of genetic variations providing an extensive catalog that include thousands of individuals' genomic variations to researchers. At the end of the 1990s, the establishment of dbSNP has led to record genotype-phenotype associations via variant databases [59]. Latterly, large-scale projects such as gnomAD and 1000 Genome Project Databases that actively collect genomic data from various populations have become available MAF at population level found in these databases is one of the

primary guides to interpret that variant pathogenicity. Because causative variants related to most Mendelian disorders have deleterious effects on reproductive fitness. Generally, causative alleles are less likely to reside in these databases or are present with low frequencies. In any global population database, except for the well-known founder alleles, >5% MAF can be considered as benign [60]. Therefore, a subset of the total number of variants inside these databases can be used for variant filtration. This is often achieved according to three different approaches. The first approach, called discrete filtering, assumes that a disease-causing variant should not found in these databases [61, 62]. This approach can be useful for very rare Mendelian disorders, but it can be problematic in some cases. Excluding observed alleles, independent from their MAF, can lead to the elimination of truly pathogenic alleles found in the general population at low frequencies because of the increasing number of genomes in databases. Especially, elucidating autosomal recessive disorders are affected by this risk. The second approach, called 1%-approach, is based on allele frequency thresholds that change according to the inheritance model of variants. While the analysis of autosomal recessive variants MAF threshold can be set at 1%, MAF cutoff of 0.1% can be useful for autosomal dominant variants [62]. Alternatively, the third approach, called the quantile-based approach, employs frequency thresholds as in the previous method. However, the thresholds in the quantile-based method are variable and depend on disease prevalence, mode of inheritance, database size, and database characteristics [63].

Depending on the case, different approaches can be employed using population databases with different scopes and data collection. Here, we summarize the widely used population databases. 1000 Genome Project (1KGP) Database.

3.1.1 1000 Genome Project (1KGP) database

1KGP database provides a comprehensive set of human genetic variations from a diverse set of individuals of multiple populations. The database includes the reconstructed genomes of 2504 individuals from 26 populations obtained by combining low-coverage whole-genome sequencing, deep exome sequencing, and dense microarray genotyping. The database contains over 88 million variants, which consist of around 84.7 million SNPs, 3.6 million indels, and 60,000 structural variants [64, 65].

3.1.2 The Genome Aggregation Database (gnomAD)

gnomAD is an extensive collection of exome and genome sequencing data from several large-scale sequencing projects. The first release of gnomAD is also known as the Exome Aggregation Consortium (ExAC) dataset. gnomAD short variant v2 release contains 125,748 exomes, and 15,708 whole genomes mapped to the GRCh37/hg19 reference sequence. In contrast, the short variant v3 release contains 71,702 whole genomes, including most of the whole genomes from v2 release mapped to the GRCh38 reference sequence. Therefore, gnomAD v2 provides higher power for the analysis of the coding regions, while v3 offers a valuable resource for the analysis of non-coding regions. For the analysis of structural variants, gnomAD SV v2.1 data set grants access to a total of 10,847 genomes aligned against the GRCh37 reference sequence [66].

3.1.3 Database of short genetic variations (dbSNP) and the database of genomic structural variations (dbVar)

The National Center for Biotechnology Information (NCBI) maintains dbSNP and dbVar databases which together contain almost 2 billion submitted human

variants. Although dbVar does not have a reference structural variant database since the current technology cannot detect the precise breakpoints in the genome, dbSNP presents the reference variants as rs identifiers. Other contents of the dataset include population frequency, geographic origin of the population, population-specific genotype and allele frequencies as well as population-specific heterozygosity estimates. Besides serving as a human population database, dbSNP and dbVar also contain a variety of organisms´ genomic variations that can be a valuable resource for evolutionary studies [67, 68].

3.1.4 ClinVar

ClinVar is a public database that archives genetic variances of any type and the interpretations of their clinical significance for reported conditions. Unlike dbSNP and dbVar that are also maintained by NCBI, ClinVar only focuses on the medically relevant variations. Although ClinVar reviews the submissions of variants for validation, the clinical significance of the variants is reported directly from submitters. ClinVar displays any conflict between the interpretations for the same variant from different submitters or the consensus. In the strict comparison approaches, the algorithm evaluates submissions for a variant to be pathogenic and likely pathogenic as conflicting. In the more relaxed approach, the variants can be categorized as pathogenic/likely pathogenic, benign/likely benign, or uncertain significance [69].

3.1.5 Database of chromosomal imbalance and phenotype in humans using Ensembl resources (DECIPHER)

DECIPHER provides a catalog of common copy-number changes in healthy populations as well as chromosome rearrangements of patients and their phenotype record submitted by clinical researchers upon informed consent [70]. Therefore, DECIPHER can serve as a valuable platform during variant prioritization. Users can check both the healthy population database and the previously submitted clinical records within DECIPHER to understand the effect of the variant of interests better and to identify novel and potentially pathogenic variants.

3.2 Pathogenicity prediction tools

Even population MAF-based filtering, individuals generally have many variants that are not present in databases. Most of these variants do not classify definitively as benign or pathogenic according to criteria proposed by some clinical guidelines such as the American College of Medical Genetics and Genomics (ACMG) [60]. These types of alterations termed variants of uncertain significance (VUS). Further filtering approaches must use to reduce the number of VUS. For this purpose, numerous pathogenicity prediction tools based on different principles have been developed to evaluate the variant effect. ACMG and the European Society of Human Genetics (ESHG) [71] guidelines also recommend these in-silico methods to interpret variant pathogenicity.

The first methods were proposed to predict computationally whether an amino acid substitution will disturb the protein function. These methods, now part of the PolyPhen algorithm [72], use physical properties of the mutational change along with a multispecies alignment as a basis to evaluate mutations. Many methods have been derived from this idea and are based on different principles. Evolutionary conservation is among the most useful features for such predictions. Some methods such as SIFT [73], PROVEAN [74] and PANTHER [75] rely on sequence conservation. For example, SIFT, as the most widely used algorithm, compares the

alignments of related sequences by performing a PSI-BLAST search to check if the variant is tolerated in an evolutionary aspect. In addition to sequence conservation, another group of methods which take into account several features such as amino acid physicochemical properties, the context of variation position, protein structural features through machine learning algorithms are also available. CADD [76], MutationTaster2 [77], PolyPhen-2 [72], DANN [78] and VEST3 [79] are well-known examples of such tools.

The predicted impact of a variation obtained from different tools may not be the same. This problem led to researchers making efforts to develop meta predictors that combine the results from existing tools by using several approaches such as logistic regression, decision trees, random forests, and support vector machines to make their own decisions. MetaSVM and MetaLR [80], M-CAP [81] and REVEL [82] are well-known examples of meta-predictors.

Below, several useful tools are explained without a performance comparison. However, various benchmark studies that have extensively examined the accuracy of these tools can be found in the literature [83–85].

3.2.1 MutationTaster2

MutationTaster2, using a naive Bayes classifier, predicts the functional consequences of variants that are both in exonic and intronic regions by incorporating a scoring system for the evolutionary conservation around DNA variants. MutationTaster uses information from several variant databases, including 1KGP and ClinVar. The tool automatically predicts a variant as neutral if it is found more than four times in the homozygous state in these databases and as disease-causing if it is reported as pathogenic in ClinVar by listing the associated disease phenotypes [77].

3.2.2 Combined annotation-dependent depletion (CADD)

CADD combines 63 genomic features derived from evolutionary constraint, surrounding sequence context, and functional predictions to evaluate SNVs and short indels. The tool integrates all of these features into a single CADD score using a machine learning approach trained on a binary distinction between simulated variants and variants that have become fixed in human populations since the split between humans and chimpanzees. C scores correlate with pathogenicity of a variant and disease severity [76].

3.2.3 The Mendelian clinically applicable pathogenicity (M-CAP)

M-CAP uses a supervised learning classifier to interpret genomic variants and focus especially on coding mutations for Mendelian diseases. As a meta-predictor, it uses nine existing tools SIFT, PolyPhen-2, CADD, MutationTaster, MutationAssessor [86], FATHMM [87], LRT [88], MetaLR and MetaSVM. It also combines information of base-pair, amino acid, genomic region, and gene conservation from RVIS [89], PhyloP [90], PhastCons [91], SIPHY [92], GERP [93], PAM250 and BLOSUM62 [94]. Additionally, M-CAP establishes multiple-sequence alignments of 99 primate, mammalian, and vertebrate genomes to the human genome as a new feature [81].

3.2.4 PrimateAI

PrimateAI [95] is a deep neural network trained by a comprehensive dataset that includes around 380,000 common missense variants from humans and six

non-human primate species. PrimateAI categorizes the common missense mutations from other primate species as non-pathogenic for humans. Thus, it enables the identification of the pathogenic variants. PrimateAI has previously shown 88% accuracy in disease-causing variant identification and allowed the discovery of 14 novel candidate genes related to intellectual disability. PrimateAI also incorporates protein structure information as it learns to predict the secondary structure and solvent accessibility from amino acid sequences. PrimateAI provides a score to the user in which a threshold of >0.8 is for likely pathogenic classification, <0.6 is for likely benign, and 0.6–0.8 is as intermediate in genes with dominant modes of inheritance, and a threshold of >0.7 is for likely pathogenic and <0.5 for likely benign in genes with recessive modes of inheritance.

3.3 Genic intolerance

Genic intolerance is a gene-level assessment that has a potential to prioritize genomic variants. It has been developed as a scoring system to calculate tolerance of genes to a functional genetic variation on a genome-wide scale and rank them using 6503 WES data available in the National Heart, Lung, and Blood Institute-NHLBI Exome Sequencing Project [89]. This system predicts the expected common functional variation in the gene and compares them to apparently neutral variation found in the gene. The deviation from this prediction is attributed to the intolerance score, namely the Residual Variation Intolerance Score (RVIS). While genes with a positive RVIS score have more common functional variation than expected, genes with negative RVIS scores have less. A negative RVIS score indicates that the gene is intolerant. The scoring system also shows that the genes that cause Mendelian diseases are significantly more intolerant to functional variation than genes that do not cause any known disease.

3.4 Model organism databases

The evolutionary conservation of many biological processes among species allows the usage of several different model organisms to study human diseases. Although not all the human genes are conserved in invertebrate models such as worms and fruit flies, vertebrate models such as zebrafish and mouse provide valuable resources to study such genes. When evaluating the function of a conserved gene in model organisms, it is critical to keep in mind that orthologous genes usually cause different phenotypes in different species, although the gene products have a similar molecular function. The model organism databases listed below provide the related information on the molecular function of query genes so that they serve as a valuable resource during the variant prioritization process.

3.4.1 Mouse genome informatics (MGI)

MGI is the primary database that integrates genetic, genomic, and biological data for the laboratory Mouse. Mouse Genome Database (MGD) and Mouse Gene Expression Database (GXD) are the two largest contributors to MGI, both serving as valuable resources for the studies of human disease. MGD provides curated phenotypes and functional annotations for mouse genes and alleles, while GXD contains mouse gene expression data with an emphasis on endogenous gene expression during mouse development [96, 97]. The Human-Mouse Disease Connection tool within MGI is another important feature that facilitates exploring gene-phenotype-disease relationships between human and mouse. By simply searching the list of human genes on MGI, the algorithm finds matching mouse genes and their

homologs and displays the both human and mouse phenotypes associated with the genes of interest. MGI is updated once every week by adding new annotations from the literature.

3.4.2 International Mouse Phenotyping Consortium (IMPC)

IMPC aims to establish a comprehensive dataset of mouse genome and phenome by knocking out each gene individually and characterizing the physical and chemical changes, thus providing the foundations for the functional analysis of human genetic variation [98]. The project also aims to generate putative human pathogenic variants in both coding and non-coding regions of the mouse genome.

IMPC uses an algorithm that has been developed to detect phenotypic similarities between the mouse strains of IMPC and more than 7000 rare diseases. The algorithm evaluates a very diverse set of phenotyping parameters that comprise neurological, behavioral, metabolic, cardiovascular, pulmonary, reproductive, respiratory, sensory, musculoskeletal, and immunological parameters and provides a quantitative measure on how well a mouse model recapitulates disease features.

So far, over 3000 genes have already been cataloged and revealed models for 360 diseases, with 90% of the annotated phenotypes being novel [99]. By 2021, IMPC plans to analyze more than 9000 mouse genes to facilitate the prioritization and validation of variations obtained from clinical sequencing efforts.

3.4.3 Rat Genome Database (RGD)

RGD provides genetic, genomic, phenotypic, and disease-related data for the laboratory rat, *Rattus norvegicus*. Rats have been one of the most commonly used model organisms for human disease research. RGD catalogs the rat data and also serves as a comparative data analysis platform between species such as rat, mouse, and human by validating the orthologous relationships. The database currently contains more than 1300 rat strains with disease/phenotype annotations [100]. RGD contains several tools that facilitate the analysis of data in disease-related content. PhenoMiner is such a tool that standardizes the phenotype data obtained from different rat studies by using a variety of ontologies developed at RGD [101]. Users can select one of the Phenominer search categories that include rat strains, experimental conditions, clinical measurements, and measurement methods to begin their search. Then, the algorithm filters the data according to the selected conditions and displays the results.

3.4.4 FlyBase

FlyBase is the central resource for integrated Drosophila genetic and genomic data, including but not limited to sequence-level gene models, mutant phenotypes, mutant lesions and chromosome aberrations, as well as gene expression patterns [102]. The fruit fly—*Drosophila melanogaster*—is a member of the Drosophila family widely used as a model for human disease research.

FlyBase allows different approaches for data presentation to facilitate Drosophila translational research as the two main methods being the gene-centric and disease-centric ones. The Gene Report displays information on individual genes. The report also lists the mutant alleles of the gene and the expression pattern of the gene products. The Human Disease Model Report provides background information on a specific disease and presents summaries of the experimental data and results from previous fruit fly studies.

FlyBase also incorporates orthology prediction tools such as OrthoDB and DIOPT that have been developed to identify orthologs of fly genes in multiple organisms [103, 104]. Integrating the results of these tools to the Gene Reports provides users the identification of orthologs in up to 5000 species. The predicted orthologs serve as a valuable resource for the human disease gene variants prediction as FlyBase also indicates whether the human ortholog functionally complements the fly mutant upon transfer into the Drosophila genome.

3.4.5 WormBase

WormBase serves as the main database for genetic, genomic, and biological information on C.elegans and related nematodes. *C. elegans* is a widely used model for human disease variant research as over 40% of human genes have a C.elegans ortholog. WormBase catalogs the available mutant strains for each gene as well as related nematode studies. WS273 release of WormBase contains over 160,000 gene summaries for 10 nematode species. The gene summaries also include human ortholog diseases and phenotypes to aid the detection of human disease-causing variants [105].

3.4.6 Zebrafish Information Network (ZFin)

ZFIN is the main database that provides genetic, genomic, and phenotypic data from zebrafish studies [106]. Zebrafish—*Danio reriro*—is a model organism extensively used in biomedical research, especially for developmental and genomic studies. Powerful approaches are available to model human diseases using zebrafish. Genetic manipulation of zebrafish orthologs of human disease genes is a common strategy to model genetic disorders such as Duchenne muscular dystrophy [107] and Rett Syndrome [108]. Another strategy of disease modeling is generating transgenic zebrafish lines that express human genes. This approach allows testing the function of the potential disease-causative variant in disease pathology. For example, a transgenic zebrafish model confirmed the pathogenicity of two novel XPNPEP3 gene mutations predicted to be ciliopathy-causing in the clinic [109]. Users can easily search ZFIN to reach information on disease models, including the transgenic lines and mutant phenotypes related to their query.

4. Real-life case

4.1 Variation in the frizzled class receptor 6 (FZD6) protein found in individuals with the nail disorder

Nonsyndromic congenital nail disorder 1 (OMIM #1161050) is a condition affecting the fingernails and toenails characterized by extremely thick nails, onycholysis, hyponychia and claw-like appearance. Autosomal recessive mutations in the FZD6 gene (OMIM *603409) were found to be associated with this disorder [110]. FZD6 is a member of the highly conserved WNT receptors family crucial for developmental processes and differentiation. The study conducted on mice demonstrated that FZD6-mediated Wnt signaling has a regulatory role in the differentiation process of claw/nail formation [111].

In a previous study from our group, a Turkish family with three affected individuals reported. After performing WES on the index case, 96 de novo heterozygous, 421 homozygous, and 185 compound heterozygous variants were obtained

from data analysis. Employing population MAF frequency filtering according to the mode of inheritance has decreased the number of variants to 19, 46, 3 for de novo heterozygous, homozygous, compound heterozygous variants respectively. Further prioritization approaches were applied by integrating pathogenicity prediction scores provided by PrimateAI and other tools, model organism phenotypes, and gene intolerance scores. Ultimately, the FZD6 gene was found to be the most prominent gene even though the gene does not have a high intolerance score. However, the potential functional impact of the mutation was supported by the examination of the evolutionary conservation of the disturbed amino acid region. The region was found to be evolutionarily conserved in other FZD6 orthologues including *Pan troglodytes*, *Macaca mulatta*, *Pongo abelii*, *Bos taurus*, *Canis lupus familiaris*, *Rattus norvegicus*, *Mus musculus*, *Xenopus laevis*. The index case had a homozygous 8 bp deletion on the FZD6 gene caused p.Gly559Aspfs*16. Additionally, this mutation has previously been reported in two other Turkish families. It is also reported that all three families have a common ancestor. In this study, the pathogenicity mechanism for this mutation in nail dysplasia is provided for the first time. The mutation causes a frameshift and creates a premature stop codon at position 16 of the new reading frame [112].

This case study demonstrated that the promising applications of evolutionary approaches assist the clinical diagnosis.

5. Conclusion

Associating genomic variants with diseases is a multistep process. The early steps of this process are highly automated through the usage of several bioinformatics tools. However, the final prioritization step, which is the most critical step, is not completely automated. It requires a comprehensive interpretation together with integrative approaches. In this chapter, we aimed to explain the potential of integrating evolutionary principles into variant prioritization toward clinical utility. This chapter provides sufficient basic information to understand the required bioinformatics tools, various databases with increasing sequence data from individuals as well as model organism research. Finally, we conclude that the pre-evaluation of individual variations with evolutionary approaches can help shorten the diagnostic odyssey, hence saving time and resources. This chapter aims to contribute to the integration of evolutionary genetics to medical genomics. Further studies that combine theoretical and analytical approaches are needed to improve the field of precision medicine via the use of evolutionary insight.

Conflict of interest

The authors declare no conflict of interest.

DOI: http://dx.doi.org/10.5772/intechopen.92738

IntechOpen

IntechOpen

Author details

Ugur Sezerman*, Tugce Bozkurt and Fatma Sadife Isleyen
Graduate School of Health Sciences, Biostatistics and Bioinformatics Program,
Acibadem Mehmet Ali Aydinlar University, Istanbul, Turkey

*Address all correspondence to: sezermanu@gmail.com

IntechOpen

© 2020 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. CC BY

References

[1] Jamuar SS, Tan E-C. Clinical application of next-generation sequencing for Mendelian diseases. Human Genomics. 2015;**9**:10. DOI: 10.1186/s40246-015-0031-5

[2] Domazet-Loso T, Tautz D. An ancient evolutionary origin of genes associated with human genetic diseases. Molecular Biology and Evolution. 2008;**25**:2699-2707. DOI: 10.1093/molbev/msn214

[3] Cai JJ, Borenstein E, Chen R, Petrov DA. Similarly strong purifying selection acts on human disease genes of all evolutionary ages. Genome Biology and Evolution. 2009;**1**:131-144. DOI: 10.1093/gbe/evp013

[4] Stenson PD, Ball EV, Howells K, Phillips AD, Mort M, Cooper DN. The human gene mutation database: Providing a comprehensive central mutation database for molecular diagnostics and personalised genomics. Human Genomics. 2009;**4**:69. DOI: 10.1186/1479-7364-4-2-69

[5] Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. The sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Research. 2010;**38**:1767-1771. DOI: 10.1093/nar/gkp1137

[6] Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using *Phred*. I. Accuracy assessment. Genome Research. 1998;**8**:175-185. DOI: 10.1101/gr.8.3.175

[7] Ewing B, Green P. Base-calling of automated sequencer traces using *Phred*. II. Error probabilities. Genome Research. 1998;**8**:186-194. DOI: 10.1101/gr.8.3.186

[8] Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[9] Chanumolu SK, Albahrani M, Otu HH. FQStat: A parallel architecture for very high-speed assessment of sequencing quality metrics. BMC Bioinformatics. 2019;**20**:424. DOI: 10.1186/s12859-019-3015-y

[10] Thrash A, Arick M, Peterson DG. Quack: A quality assurance tool for high throughput sequence data. Analytical Biochemistry. 2018;**548**:38-43. DOI: 10.1016/j.ab.2018.01.028

[11] Peng Y, Maxwell AS, Barker ND, Laird JG, Kennedy AJ, Wang N, et al. SeqAssist: A novel toolkit for preliminary analysis of next-generation sequencing data. BMC Bioinformatics. 2014;**15**:S10. DOI: 10.1186/1471-2105-15-S11-S10

[12] Zhou Q, Su X, Wang A, Xu J, Ning K. QC-chain: Fast and holistic quality control method for next-generation sequencing data. PLoS One. 2013;**8**:e60234. DOI: 10.1371/journal.pone.0060234

[13] Sun K. Ktrim: An extra-fast and accurate adapter- and quality-trimmer for sequencing data. Bioinformatics. 2020:btaa171. DOI: 10.1093/bioinformatics/btaa171

[14] Liao X, Li M, Zou Y, Wu F, Pan Y, Wang J. An efficient trimming algorithm based on multi-feature fusion scoring model for NGS data. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2019:1-1. DOI: 10.1109/TCBB.2019.2897558

[15] Sturm M, Schroeder C, Bauer P. SeqPurge: Highly-sensitive adapter trimming for paired-end NGS data. BMC Bioinformatics. 2016;**17**:208. DOI: 10.1186/s12859-016-1069-7

[16] Schubert M, Lindgreen S, Orlando L. AdapterRemoval v2: Rapid adapter trimming, identification,

and read merging. BMC Research Notes. 2016;**9**:88. DOI: 10.1186/s13104-016-1900-2

[17] Li Y-L, Weng J-C, Hsiao C-C, Chou M-T, Tseng C-W, Hung J-H. PEAT: An intelligent and efficient paired-end sequencing adapter trimming algorithm. BMC Bioinformatics. 2015;**16**:S2. DOI: 10.1186/1471-2105-16-S1-S2

[18] Jiang H, Lei R, Ding S-W, Zhu S. Skewer: A fast and accurate adapter trimmer for next-generation sequencing paired-end reads. BMC Bioinformatics. 2014;**15**:182. DOI: 10.1186/1471-2105-15-182

[19] Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 2014;**30**:2114-2120. DOI: 10.1093/bioinformatics/btu170

[20] Ma Y, Xie H, Han X, Irwin DM, Zhang Y-P. QcReads: An adapter and quality trimming tool for next-generation sequencing reads. Journal of Genetics and Genomics. 2013;**40**:639-642. DOI: 10.1016/j.jgg.2013.11.001

[21] 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**:500-506. DOI: 10.1016/j.ygeno.2013.07.011

[22] Kong Y. Btrim: A fast, lightweight adapter and quality trimming program for next-generation sequencing technologies. Genomics. 2011;**98**:152-153. DOI: 10.1016/j.ygeno.2011.05.009

[23] Roser LG, Agüero F, Sánchez DO. FastqCleaner: An interactive bioconductor application for quality-control, filtering and trimming of FASTQ files. BMC Bioinformatics. 2019;**20**:361. DOI: 10.1186/s12859-019-2961-8

[24] Liu X, Yan Z, Wu C, Yang Y, Li X, Zhang G. FastProNGS: Fast preprocessing of next-generation sequencing reads. BMC Bioinformatics. 2019;**20**:345. DOI: 10.1186/s12859-019-2936-9

[25] Rangamaran VR, Uppili B, Gopal D, Ramalingam K. EasyQC: Tool with interactive user Interface for efficient next-generation sequencing data quality control. Journal of Computational Biology. 2018;**25**:1301-1311. DOI: 10.1089/cmb.2017.0186

[26] Chen S, Zhou Y, Chen Y, Gu J. Fastp: An ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;**34**:i884-i890. DOI: 10.1093/bioinformatics/bty560

[27] Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. Nature Methods. 2012;**9**:357-359. DOI: 10.1038/nmeth.1923

[28] Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics. 2009;**25**:1754-1760. DOI: 10.1093/bioinformatics/btp324

[29] Available from: http://www.novocraft.com/products/novoalign/

[30] Lee W-P, Stromberg MP, Ward A, Stewart C, Garrison EP, Marth GT. MOSAIK: A hash-based algorithm for accurate next-generation sequencing short-read mapping. PLoS One. 2014;**9**:e90581. DOI: 10.1371/journal.pone.0090581

[31] David M, Dzamba M, Lister D, Ilie L, Brudno M. SHRiMP2: Sensitive yet practical short read mapping. Bioinformatics. 2011;**27**:1011-1012. DOI: 10.1093/bioinformatics/btr046

[32] Fonseca NA, Rung J, Brazma A, Marioni JC. Tools for mapping high-throughput sequencing data. Bioinformatics. 2012;**28**:3169-3177. DOI: 10.1093/bioinformatics/bts605

[33] Hatem A, Bozdag D, Toland AE, Çatalyürek ÜV. Benchmarking short sequence mapping tools. BMC Bioinformatics. 2013;**14**:184. DOI: 10.1186/1471-2105-14-184

[34] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;**25**:2078-2079. DOI: 10.1093/bioinformatics/btp352

[35] Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nature Biotechnology. 2011;**29**:24-26. DOI: 10.1038/nbt.1754

[36] picard n.d. Available from: http://broadinstitute.github.io/picard/

[37] DePristo MA, Banks E, Poplin RE, Garimella KV, Maguire JR, Hartl C. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics. 2011;**43**:491-498. DOI: 10.1038/ng.806

[38] Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From FastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. Current Protocols in Bioinformatics. 2013;**43**. DOI: 10.1002/0471250953.bi1110s43

[39] Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. ArXiv:12073907 [q-Bio]; 2012

[40] Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SRF, et al. Nature Genetics. 2014;**46**:912-918. DOI: 10.1038/ng.3036

[41] Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, et al. VarScan: Variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics. 2009;**25**:2283-2285. DOI: 10.1093/bioinformatics/btp373

[42] Wei Z, Wang W, Hu P, Lyon GJ, Hakonarson H. SNVer: A statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. Nucleic Acids Research. 2011;**39**:e132-e132. DOI: 10.1093/nar/gkr599

[43] Sandmann S, de Graaf AO, Karimi M, van der Reijden BA, Hellström-Lindberg E, Jansen JH, et al. Evaluating variant calling tools for non-matched next-generation sequencing data. Scientific Reports. 2017;**7**:43169. DOI: 10.1038/srep43169

[44] Chen J, Li X, Zhong H, Meng Y, Du H. Systematic comparison of germline variant calling pipelines cross multiple next-generation sequencers. Scientific Reports. 2019;**9**:9345. DOI: 10.1038/s41598-019-45835-3

[45] Kumaran M, Subramanian U, Devarajan B. Performance assessment of variant calling pipelines using human whole exome sequencing and simulated data. BMC Bioinformatics. 2019;**20**:342. DOI: 10.1186/s12859-019-2928-9

[46] Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Research. 2010;**38**:e164-e164. DOI: 10.1093/nar/gkq603

[47] McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl variant effect predictor. Genome Biology. 2016;**17**:122. DOI: 10.1186/s13059-016-0974-4

[48] Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide

polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w^{1118} ; iso-2; iso-3. Fly. 2012;**6**: 80-92. DOI: 10.4161/fly.19695

[49] Paila U, Chapman BA, Kirchner R, Quinlan AR. GEMINI: Integrative exploration of genetic variation and genome annotations. PLoS Computational Biology. 2013;**9**:e1003153. DOI: 10.1371/journal.pcbi.1003153

[50] Desvignes J-P, Bartoli M, Delague V, Krahn M, Miltgen M, Béroud C, et al. VarAFT: A variant annotation and filtration system for human next generation sequencing data. Nucleic Acids Research. 2018;**46**:W545-W553. DOI: 10.1093/nar/gky471

[51] Makarov V, O'Grady T, Cai G, Lihm J, Buxbaum JD, Yoon S. AnnTools: A comprehensive and versatile annotation toolkit for genomic variants. Bioinformatics. 2012;**28**:724-725. DOI: 10.1093/bioinformatics/bts032

[52] Ge D, Ruzzo EK, Shianna KV, He M, Pelak K, Heinzen EL, et al. SVA: Software for annotating and visualizing sequenced human genomes. Bioinformatics. 2011;**27**:1998-2000. DOI: 10.1093/bioinformatics/btr317

[53] Grant JR, Arantes AS, Liao X, Stothard P. In-depth annotation of SNPs arising from resequencing projects using NGS-SNP. Bioinformatics. 2011;**27**:2300-2301. DOI: 10.1093/bioinformatics/btr372

[54] Miller MP. Understanding human disease mutations through the use of interspecific genetic variation. Human Molecular Genetics. 2001;**10**:2319-2328. DOI: 10.1093/hmg/10.21.2319

[55] Thomas PD, Kejariwal A. Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: Evolutionary evidence for differences in molecular effects. Proceedings of the National Academy of Sciences. 2004;**101**:15398-15403. DOI: 10.1073/pnas.0404380101

[56] Zhu Q, Ge D, Maia JM, Zhu M, Petrovski S, Dickson SP, et al. A genome-wide comparison of the functional properties of rare and common genetic variants in humans. The American Journal of Human Genetics. 2011;**88**:458-468. DOI: 10.1016/j.ajhg.2011.03.008

[57] Cooper DN, Kehrer-Sawatzki H. Exploring the potential relevance of human-specific genes to complex disease. Human Genomics. 2011;**5**:99. DOI: 10.1186/1479-7364-5-2-99

[58] Botstein D, Risch N. Discovering genotypes underlying human phenotypes: Past successes for Mendelian disease, future approaches for complex disease. Nature Genetics. 2003;**33**:228-237. DOI: 10.1038/ng1090

[59] Sherry ST, Ward M, Sirotkin K. dbSNP—Database for Single Nucleotide Polymorphisms and Other Classes of Minor Genetic Variation. n.d.:4

[60] Richards S, Aziz N, Bale S, Bick D, Das S, et al. Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;**17**:405-423. DOI: 10.1038/gim.2015.30

[61] Stitziel NO, Kiezun A, Sunyaev S. Computational and statistical approaches to analyzing variants identified by exome sequencing. Genome Biology. 2011;**12**:227. DOI: 10.1186/gb-2011-12-9-227

[62] Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, et al. Exome sequencing as a tool for

Mendelian disease gene discovery. Nature Reviews. Genetics. 2011;**12**: 745-755. DOI: 10.1038/nrg3031

[63] Broeckx BJG, Peelman L, Saunders JH, Deforce D, Clement L. Using variant databases for variant prioritization and to detect erroneous genotype-phenotype associations. BMC Bioinformatics. 2017;**18**:535. DOI: 10.1186/s12859-017-1951-y

[64] The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;**526**:68-74. DOI: 10.1038/nature15393

[65] Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015;**526**:75-81. DOI: 10.1038/nature15394

[66] Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Genomics. 2019. DOI: 10.1101/531210

[67] Sherry ST. dbSNP: The NCBI database of genetic variation. Nucleic Acids Research. 2001;**29**:308-311. DOI: 10.1093/nar/29.1.308

[68] MacDonald JR, Ziman R, Yuen RKC, Feuk L, Scherer SW. The database of genomic variants: A curated collection of structural variation in the human genome. Nucl Acids Res. 2014;**42**:D986-D992. DOI: 10.1093/nar/gkt958

[69] Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: Public archive of interpretations of clinically relevant variants. Nucleic Acids Research. 2016;**44**:D862-D868. DOI: 10.1093/nar/gkv1222

[70] Firth HV, Richards SM, Bevan AP, Clayton S, Corpas M, Rajan D, et al. DECIPHER: Database of chromosomal imbalance and phenotype in humans using Ensembl resources. The American Journal of Human Genetics. 2009;**84**:524-533. DOI: 10.1016/j.ajhg.2009.03.010

[71] Matthijs G, Souche E, Alders M, Corveleyn A, Eck S, Feenstra I, et al. Guidelines for diagnostic next-generation sequencing. European Journal of Human Genetics. 2016;**24**:2-5. DOI: 10.1038/ejhg.2015.226

[72] Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nature Methods. 2010;**7**:248-249. DOI: 10.1038/nmeth0410-248

[73] Ng PC. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Research. 2003;**31**:3812-3814. DOI: 10.1093/nar/gkg509

[74] Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and Indels. PLoS One. 2012;**7**:e46688. DOI: 10.1371/journal.pone.0046688

[75] Thomas PD. PANTHER: A library of protein families and subfamilies indexed by function. Genome Research. 2003;**13**:2129-2141. DOI: 10.1101/gr.772403

[76] Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nature Genetics. 2014;**46**:310-315. DOI: 10.1038/ng.2892

[77] Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: Mutation prediction for the deep-sequencing age. Nature Methods.

2014;**11**:361-362. DOI: 10.1038/nmeth.2890

[78] Quang D, Chen Y, Xie X. DANN: A deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics. 2015;**31**:761-763. DOI: 10.1093/bioinformatics/btu703

[79] Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics. 2013;**14**:S3. DOI: 10.1186/1471-2164-14-S3-S3

[80] Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Human Molecular Genetics. 2015;**24**:2125-2137. DOI: 10.1093/hmg/ddu733

[81] Jagadeesh KA, Wenger AM, Berger MJ, Guturu H, Stenson PD, Cooper DN, et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. Nature Genetics. 2016;**48**:1581-1586. DOI: 10.1038/ng.3703

[82] Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: An ensemble method for predicting the pathogenicity of rare missense variants. The American Journal of Human Genetics. 2016;**99**:877-885. DOI: 10.1016/j.ajhg.2016.08.016

[83] Li J, Zhao T, Zhang Y, Zhang K, Shi L, Chen Y, et al. Performance evaluation of pathogenicity-computation methods for missense variants. Nucleic Acids Research. 2018;**46**:7793-7804. DOI: 10.1093/nar/gky678

[84] Schaafsma GCP, Vihinen M. Representativeness of variation benchmark datasets. BMC Bioinformatics. 2018;**19**:461. DOI: 10.1186/s12859-018-2478-6

[85] Niroula A, Vihinen M. How good are pathogenicity predictors in detecting benign variants? PLoS Computational Biology. 2019;**15**:e1006481. DOI: 10.1371/journal.pcbi.1006481

[86] Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: Application to cancer genomics. Nucleic Acids Research. 2011;**39**:e118-e118. DOI: 10.1093/nar/gkr407

[87] Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GLA, Edwards KJ, et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Human Mutation. 2013;**34**:57-65. DOI: 10.1002/humu.22225

[88] Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Research. 2009;**19**:1553-1561. DOI: 10.1101/gr.092619.109

[89] Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic intolerance to functional variation and the interpretation of personal genomes. PLoS Genetics. 2013;**9**:e1003709. DOI: 10.1371/journal.pgen.1003709

[90] Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Research. 2010;**20**:110-121. DOI: 10.1101/gr.097857.109

[91] Siepel A. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Research. 2005;**15**:1034-1050. DOI: 10.1101/gr.3715005

[92] Garber M, Guttman M, Clamp M, Zody MC, Friedman N,

Xie X. Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics. 2009;**25**:i54-i62. DOI: 10.1093/bioinformatics/btp190

[93] 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**:e1001025. DOI: 10.1371/journal.pcbi.1001025

[94] Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences. 1992;**89**:10915-10919. DOI: 10.1073/pnas.89.22.10915

[95] Sundaram L, Gao H, Padigepati SR, McRae JF, Li Y, Kosmicki JA, et al. Predicting the clinical impact of human mutation with deep neural networks. Nature Genetics. 2018;**50**:1161-1170. DOI: 10.1038/s41588-018-0167-z

[96] Bult CJ, Blake JA, Smith CL, Kadin JA, Richardson JE, The Mouse Genome Database Group, et al. Mouse Genome Database (MGD) 2019. Nucleic Acids Research. 2019;**47**:D801-D806. DOI: 10.1093/nar/gky1056

[97] Smith CM, Hayamizu TF, Finger JH, Bello SM, McCright IJ, Xu J, et al. The mouse gene expression database (GXD): 2019 update. Nucleic Acids Research. 2019;**47**:D774-D779. DOI: 10.1093/nar/gky922

[98] The International Mouse Phenotyping Consortium, Dickinson ME, Flenniken AM, Ji X, Teboul L, Wong MD, et al. High-throughput discovery of novel developmental phenotypes. Nature. 2016;**537**:508-514. DOI: 10.1038/nature19356.

[99] The International Mouse Phenotyping Consortium, Meehan TF, Conte N, West DB, Jacobsen JO, Mason J, et al. Disease model discovery from 3,328 gene knockouts by the International Mouse Phenotyping Consortium. Nature Genetics. 2017;**49**:1231-1238. DOI: 10.1038/ng.3901.

[100] Smith JR, Hayman GT, Wang S-J, Laulederkind SJF, Hoffman MJ, Kaldunski ML, et al. The year of the rat: The rat genome database at 20: A multi-species knowledgebase and analysis platform. Nucleic Acids Research. 2019:gkz1041. DOI: 10.1093/nar/gkz1041

[101] Laulederkind SJF, Liu W, Smith JR, Hayman GT, Wang S-J, Nigam R, et al. PhenoMiner: Quantitative phenotype curation at the rat genome database. Database. 2013;**2013**:bat015. DOI: 10.1093/database/bat015

[102] Thurmond J, Goodman JL, Strelets VB, Attrill H, Gramates LS, Marygold SJ, et al. FlyBase 2.0: The next generation. Nucleic Acids Research. 2019;**47**:D759-D765. DOI: 10.1093/nar/gky1003

[103] Zdobnov EM, Tegenfeldt F, Kuznetsov D, Waterhouse RM, Simão FA, Ioannidis P, et al. OrthoDB v9.1: Cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. Nucleic Acids Research. 2017;**45**:D744-D749. DOI: 10.1093/nar/gkw1119

[104] Hu Y, Flockhart I, Vinayagam A, Bergwitz C, Berger B, Perrimon N, et al. An integrative approach to ortholog prediction for disease-focused and other functional studies. BMC Bioinformatics. 2011;**12**:357. DOI: 10.1186/1471-2105-12-357

[105] Harris TW, Arnaboldi V, Cain S, Chan J, Chen WJ, Cho J, et al. WormBase: A modern model organism information resource. Nucleic Acids Research. 2019:gkz920. DOI: 10.1093/nar/gkz920

[106] Ruzicka L, Howe DG, Ramachandran S, Toro S, Van Slyke CE, Bradford YM, et al. The Zebrafish information network: New support for non-coding genes, richer gene ontology annotations and the Alliance of genome resources. Nucleic Acids Research. 2019;**47**:D867-D873. DOI: 10.1093/nar/gky1090

[107] Berger J, Currie PD. Zebrafish models flex their muscles to shed light on muscular dystrophies. Disease Models & Mechanisms. 2012;**5**:726-732. DOI: 10.1242/dmm.010082

[108] Pietri T, Roman A-C, Guyon N, Romano SA, Washbourne P, Moens CB, et al. The first mecp2-null zebrafish model shows altered motor behaviors. Frontiers in Neural Circuits. 2013;**7**:1-10. DOI: 10.3389/fncir.2013.00118

[109] O'Toole JF, Liu Y, Davis EE, Westlake CJ, Attanasio M, Otto EA, et al. Individuals with mutations in XPNPEP3, which encodes a mitochondrial protein, develop a nephronophthisis-like nephropathy. The Journal of Clinical Investigation. 2010;**120**:791-802. DOI: 10.1172/JCI40076

[110] Fröjmark A-S, Schuster J, Sobol M, Entesarian M, Kilander MBC, Gabrikova D, et al. Mutations in frizzled 6 cause isolated autosomal-recessive nail dysplasia. The American Journal of Human Genetics. 2011;**88**:852-860. DOI: 10.1016/j.ajhg.2011.05.013

[111] Cui C-Y, Klar J, Georgii-Heming P, Fröjmark A-S, Baig SM, Schlessinger D, et al. Frizzled6 deficiency disrupts the differentiation process of nail development. Journal of Investigative Dermatology. 2013;**133**:1990-1997. DOI: 10.1038/jid.2013.84

[112] Saygi C, Alanay Y, Sezerman U, Yenenler A, Özören N. A possible founder mutation in FZD6 gene in a Turkish family with autosomal recessive nail dysplasia. BMC Medical Genetics. 2019;**20**:15. DOI: 10.1186/s12881-019-0746-6