AUTOMATED SEQUENCE VARIANT CLASSIFICATION TOOL FOR DNA DIAGNOSTICS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF
MASTER OF SCIENCE
IN
MOLECULAR BIOLOGY AND GENETICS

By
R. Arda İnan
July 2024

Automated Sequence Variant Classification Tool for DNA Diagnostics

By R. Arda İnan

July 2024

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

_____

H. Tayfun Özçelik (Advisor)

_____

Özlen Konu Karakayalı

_____

Ayşe Nazlı Başak

Approved for the Graduate School of Engineering and Science:

_____

Orhan Arıkan

Director of the Graduate School

ABSTRACT

AUTOMATED SEQUENCE VARIANT CLASSIFICATION TOOL FOR DNA DIAGNOSTICS

R. Arda İnan
M.Sc. in Molecular Biology and Genetics
Advisor: H. Tayfun Özçelik
July 2024

Advancements in DNA sequencing rapidly improved our understanding of the genome in recent years. Today, these advances are revealing thousands of genetic variants that are still waiting to be deciphered. Establishing the association between genetic variation and diseases enables us to better appreciate the biology of diseases and to develop effective therapeutical solutions. To this end, clinical organizations such as American College of Medical Genetics and Genomics (ACMG) developed standards and guidelines to interpret sequence variants. Clinical Genomics Resource (ClinGen) provided further specifications to the guidelines to improve the interpretations. However, implementation of the guidelines takes considerable time and requires substantial expertise in clinical genetics. Available computational tools to automate the process (i) do not comprehensively describe how their frameworks function, (ii) fail to completely follow the latest specifications and (iii) lack high consistency with variant classifications manually performed by experts. Here, this work presents automated ACMG-based variant classifier (AAVC), which computationally interprets sequence variants based on the ACMG Guidelines and the ClinGen Specifications by aggregating information from large public databases and *in silico* prediction tools including BayesDel, ClinVar, Ensembl, gnomAD, PhyloP, RepeatMasker, SpliceAI and UniProt. The tool demonstrates a high concordance (99.67%) with FDA-approved variant classification database, reveals more than two hundred novel variants in clinically actionable genes in the Turkish Variome and reclassifies at least 57,000 inconclusive variants in ClinVar as pathogenic or likely pathogenic. The work provides a comprehensive framework to enable rapid and accurate interpretation of sequence variants by the ACMG Standards.

Keywords: ACMG, AAVC, clinical interpretation, clinical genomics, genetic diagnosis, sequence variation, variant interpretation

ÖZET

GENETİK TANI İÇİN OTOMATİK DİZİ DEĞİŞKENİ SINIFLANDIRMA ARACI

R. Arda İnan
Moleküler Biyoloji ve Genetik, Yüksek Lisans
Danışman: H. Tayfun Özçelik
Temmuz 2024

DNA dizilemesindeki ilerlemeler, genom hakkındaki bilgilerimizi son yıllarda hızla geliştirdi. Bugün bu gelişmeler, hala çözümlenmeyi bekleyen binlerce genetik değişkeni (varyant) ortaya çıkarıyor. Bu değişkenler ile hastalıklar arasındaki ilişkileri tespit etmek, hastalıkların biyolojisini daha iyi anlamamızı ve etkin tedavi yöntemleri geliştirmemizi sağlamaktadır. Bu amaçla American College of Medical Genetics and Genomics (ACMG) gibi klinik kuruluşlar, DNA değişkenlerini yorumlamak için standartlar ve kılavuzlar geliştirmiştir. Clinical Genomics Resource (ClinGen) ise bu kılavuzlara ekstra açıklamalar getirerek katkıda bulunmuştur. Ancak kılavuzların uygulanması oldukça zaman almakta ve klinik genetik alanında önemli ölçüde uzmanlık gerektirmektedir. Süreci otomatikleştirmek için geliştirilmiş bilgisayımsal araçlar, sistemlerinin işleyişine yönelik detaylı anlatımlar sunmamakta, en son yayınlanan spesifikasyonları tam olarak takip edememekte ve uzmanlarca yapılan değişken sınıflamaları ile yüksek tutarlılık gösterememektedir. Bu çalışmada, BayesDel, ClinVar, Ensembl, gnomAD, PhyloP, RepeatMasker, SpliceAI ve UniProt gibi *in silico* tahmin araçları ve büyük veritabanları gibi kaynaklardan gelen bilgileri bir araya getirerek ACMG Standartları ve ClinGen spesifikasyonlarına dayalı olarak DNA değişkenlerini otomatik olarak yorumlayan ACMG tabanlı bir değişken sınıflandırma aracı (AAVC) sunulmaktadır. Bu araç, FDA onaylı değişken sınıflandırma veritabanıyla yüksek bir uyum (%99.67) göstermekte, Türk Varyomunda klinik olarak harekete geçilebilir genlerde iki yüzden fazla yeni değişkeni ortaya çıkarmakta ve ClinVar'da sınıflandırılamayan en az 57.000 değişkeni patojenik veya olası patojenik olarak yeniden sınıflandırabilmektedir. Bu çalışma, ACMG standartlarına göre DNA değişkenlerinin hızlı ve doğru bir şekilde yorumlanmasını sağlayan bir çerçeve sunmaktadır.

Anahtar kelimeler: ACMG, AAVC, klinik değerlendirme, klinik genomik, genetik tanı, sekans varyasyonu, varyant değerlendirmesi

Acknowledgment

Before anything, I wish to thank my advisor, Tayfun Özçelik, for his constant support, insightful guidance, and valuable feedback. His expertise and encouragement have been instrumental in the completion of this work.

I am also grateful to the members of my thesis committee, for Prof. Ayşe Nazlı Başak and Assoc. Prof. Özlen Konu Karakayalı, for their time, perceptive comments and willingness to serve on my committee.

A special thanks to my friends in Bilkent University, whose companionship has been my anchor throughout my journey. I would like to specifically acknowledge Barış Kayaalp and Fatimah Safieh for their precious assistance and discussions.

I am indebted to Nezahat Doğan and İclal Özçelik for their generosity and kindness in helping me gain the strength I needed in challenging times.

I acknowledge that I am the recipient of the fellowship of 2210-A National MSc/MA Scholarship Program of Directorate of Science Fellowships and Grant Programmes (BİDEB) of Scientific and Technological Research Institution of Türkiye (TÜBİTAK).

Lastly, I am sincerely grateful to my family for their continuous support and encouragement. My parents, Erdinç İnan and Selda İnan, have always believed in me and provided the foundation upon which I could build my academic pursuits. My siblings, Bengisu and Baturay, have been a constant source of joy and motivation.

R. Arda İnan

Contents

List of Figures

List of Tables

Abbreviations

1000GP	1000 Genomes Projects
AAVC	Automated ACMG-based Variant Classifier
ACGS	The Association for Clinical Genomic Sciencea
ACMG	American College of Medical Genetics and Genomics
AI	Artificial Intelligence
alt	Alternative allele
alt_start	Alternative start codon
alt_stop	Alternative stop codon
AMP	Association for Molecular Pathology
ANNOVAR	Annotate Variation
API	Application programming interface
B	Benign
B/LB	Benign/Likely benign
BA	Benign – Stand Alone
BERT	Bidirectional Encoder Representations from Transformers
BMI	Body Mass Index
bp	Basepair
BP	Benign – Supporting
BS	Benign – Strong
CADD	Combined Annotation Dependent Depletion
cds	Coding sequence
chr	Chromosome
ClinGen	Clinical Genome Resource
ClinVar	Clinical Variants
ddNTP	Dideoxynucleosidetriphosphates
DNA	Deoxyribonucleic acid
ENCODE	Encyclopedia of DNA Elements
ENSG	Ensembl Gene
ENST	Ensembl Transcript
EP	Expert Panel
ERepo	Evidence Repository

ESP6500	Exome Sequencing Project v. 6500
ExAC	Exome Aggregation Consortium
FAF	Filtering allele frequency
FASTQ	Fast Quality Scores
FATHMM	Functional Analysis through Hidden Markov Models
FDA	Food and Drug Administration
FISH	Fluorescent *in situ* hybridization
freq	Frequency
FTP	File Transfer Protocol
GATK	Genome Analysis Toolkit
GB	Gigabytes
GERP	Genomic Evolutionary Rate Profiling
gnomAD	Genome Aggregation Database
GoF	Gain-of-function
GRC	Genome Reference Consortium
GWAS	Genome-wide Association Study
hg19	*Homo sapiens* genome assembly GRCh19
hg38	*Homo sapiens* genome assembly GRCh38
HGNC	HUGO Gene Nomenclature Committee
HGP	Human Genome Project
HGVS	Human Genome Variation Society
HI	Haploinsufficiency
HL	Hearing Loss
HPRC	Human Pangenome Reference Consortium
HTS	High-throughput Sequencing
HUGO	Human Genome Organisation
ID	Identifier
IHGS	International Human Genome Sequencing Consortium
INCONC	Inconclusive
LOEUF	Loss-of-function Observed/Expected Upper bound Fraction
LoF	Loss-of-function
LRT	Likelihood ratio test
MetaRNN	Meta-analytic Recurrent Neural Network
MetaSVM	Meta-analytic Support Vector Machine

mis_z	Missense Z-score
MitoTIP	Mitochondrial tRNA Informatics Predictor
mocopav	Most common pathogenic variant in the gene
mRNA	Messenger RNA
N/A	Not applicable
ncRNA	Non-coding RNA
NMD	Nonsense-mediated decay
NSD	Non-stop decay
nt	Nucleotide
o/e	Observed versus expected
ORF	Open-reading frame
P	Pathogenic
P/LP	Pathogenic/Likely pathogenic
pext	Proportion expressed across transcripts
phastCons	Phylogenetic Analysis with Space/Time Models – Conservation
phyloP	Phylogenetic p-values
pLI	Probability of Loss Intolerance
pLoF	Putative loss-of-function
PM	Pathogenic – Moderate
PMID	PubMed ID
PolyPhen	Polymorphism Phenotyping
pos	Position
PP	Pathogenic – Supporting
PROVEAN	Protein Variation Effect Analyzer
PS	Pathogenic – Strong
PVS	Pathogenic – Very Strong
RAM	Random-access Memory
ref	Reference allele
RefSeq	Reference Sequence
REST	Representational State Transfer
REVEL	Rare Exome Variant Ensemble Learner
RNA	Ribonucleic Acid
RNAi	RNA interference
RNA-seq	RNA-sequencing

SAM	Sequence Alignment Map
SF	Secondary Finding
SIFT	Sorting Intolerant From Tolerant
SiPhy	Site-specific Phylogenetic analysis
SMRT	Single-molecule Real-time
SNP	Single-nucleotide Polymorphism
SnpEff	SNP Effect
SV	Structural Variation
SVI	Sequence Variant Interpretation
T2T	Telomere-to-telomere Consortium
tcpt	Transcript
tRNA	Transfer RNA
TSV	Tab-separated Values
UCSC	University of California, Santa Cruz
UniProt	Universal Protein resource
UTR	Untranslated Region
VCEP	Variant Curation Expert Panel
VCF	Variant Call Format
VEP	Variant Effect Predictor
VIP-HL	Variant Interpretation Platform for Genetic Hearing Loss
VQSR	Variant Quality Score Recalibration
VUS	Variant of Uncertain Significance
VUS-H	VUS-High
VUS-L	VUS-Low
VUS-M	VUS-Medium
VUS-P	VUS/Pathogenic
wt	Wild type

Chapter 1

Introduction

1.1 Human Genome

Appreciation of the human genome is the key step to understanding the biology underlying health and disease states. Since the initial efforts began more than seven decades ago to solve the mystery of human DNA, new avenues in medicine have emerged to offer genetic counseling for disease risk assessment and prevention strategies for patients. Targeting the uniqueness of an individual's genetic makeup, we are now capable of tailoring treatments and interventions with much more improved efficacy. Understanding differences and similarities in the sequence, it has become possible to trace genetic ancestry of populations and to gain insights into evolutionary processes across species. Nonetheless, deciphering the true code of the human genome remains a challenging task due to the complex and dynamic nature of genetic interactions. Ongoing research holds promise for leveraging the full potential of the human genome in medicine.

1.1.1 Large-scale efforts aimed at deciphering human DNA

The first large-scale initiative with the aim of documenting all human DNA, Human Genome Project (HGP), was established in 1990. It brought together researchers from diverse backgrounds for not only generating a genetic map but also enable formulating strategies for data analysis (Green et al., 2015), propelling advancements in sequencing technologies (Mardis, 2011) and enacting new policies regarding maximized data sharing (Maxson Jones et al., 2018). The project took thirteen years to complete, and the final release was published in 2003 with a gap of 8% in the total sequence (McPherson et al., 2001; IHGS, 2004; Venter et al., 2001). After the completion of HGP, the Genome Reference Consortium (GRC) assumed the mission of maintaining and improving human reference genome assembly in 2007 and continued updating the assembly to date. While the GRC genome builds remain incomplete due to problematic regions of the genome, a gapless assembly was completed by Telomere-to-Telomere (T2T) Consortium (Nurk et al., 2022), The latest release, GRCh38 (Schneider et al., 2017), is currently the most widely used build among the scientific community. In addition, the Human Pangenome Reference Consortium (HPRC) recently published a draft pangenome sequence covering forty-seven individuals, which more accurately represents the diversity in the human genome (Liao et al., 2023).

Genome databases have been proven to be incredibly useful, since the beginning of the era of genomics, by providing functional annotation and classification of DNA elements. Ensembl (Hubbard et al., 2002), the University of California, Santa Cruz (UCSC) Genome Browser (Karolchick et al., 2003), RefSeq (Pruitt et al., 2006) and the Encyclopedia of DNA Elements (ENCODE) (ENCODE, 2012) are the most notable databases that comprehensively provide access to curated genomic data at multiple levels today.

1.1.2 Structure and organization of the human genome

Thanks to global collaborative efforts to decode the human genome, it is now known to us that the human genome consists of approximately 3.1 billion base pairs spread across 23 pairs of chromosomes, whereas only about one percent of it codes for proteins, which approximately corresponds to 20,000 protein-coding genes. Alternative splicing extends the complexity to 150,000 protein-coding transcripts (Frankish et al., 2021). Introns, repeat elements, pseudogenes, and non-coding RNA genes make up the remaining portion of the human genome. Of these, ncRNAs are involved in notable functions within the cell, from gene regulation to immunological pathways.

Humans inherit one copy of each gene on autosomal chromosomes from each of their parents, however expression of those genes are regulated by complex interactions at multiple levels. While most genes are required to be biallelically expressed, some are sufficient to be expressed from only one parent for proper functioning. Therefore, solving the mystery of the human genome and its dynamic interplay between expression of this information is critical in understanding human biology.

1.1.3 Types of genetic variation and their functional impact

A variant, or a mutation, is a change in the nucleotide sequence from the reference sequence. Variants can be classified into two major groups by their effects on structure: chromosomal aberrations and gene mutations. Most common form of variation in the human genome are gene mutations including single nucleotide polymorphisms (SNPs) and insertion/deletions (indels). SNPs may cause synonymous (silent), nonsense (stop-gain), start-loss, stop-loss, and missense mutations at protein level. SNPs can alter splicing at splice regions or interfere with gene regulation at regulatory regions or untranslated regions (UTRs). Indels can cause a shift in the open-reading frame (ORF) or a change protein sequence by inserting or deleting amino acids. All possible variant effects are shown in Figure 1.

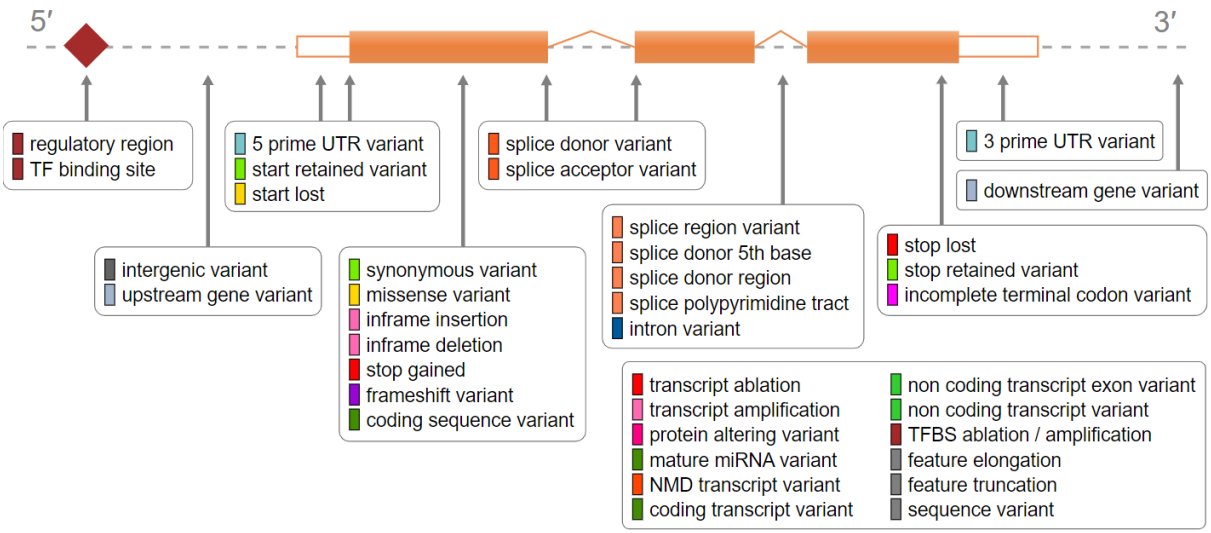


Figure 1: Variant consequence terms. The diagram demonstrates all variant consequence terms described by Sequence Ontology (Eilbeck, 2005). Adapted from Ensembl (Harrison et al., 2024).

If a protein acquires novel function or becomes overexpressed due to a variant, it is called gain-of-function (GoF). A partial or total loss in the protein abundance or function is termed loss-of-function (LoF). Canonical splice site variants (splice donor and acceptor variants), stop-gains and frameshift variants are considered as putative loss-of-function (pLoF) variants and extensively implicated in disease biology.

Canonical splice variants might cause a truncation via mechanisms of exon skipping, exon inclusion, intron retention or frameshift. Premature stop gained variants and frameshift indels produce a truncated transcript and usually targeted by nonsense-mediated decay (NMD). Start-loss variants might be rescued by an alternative in-frame start codon with a partial loss in protein sequence. The functional importance of the removed region determines the deleteriousness of start-loss.

Stop-loss variants activate non-stop decay (NSD) pathway, which eliminates aberrant transcripts similar to NMD (Klauer & van Hoof, 2012), if there is no alternative stop codon within 3'UTR. Otherwise, protein may be rescued with an increase in protein length.

Missense variants result in a change in amino acid sequence at the mutation site and the functional impact is vastly dependent on the molecular and cellular context. On the other hand, synonymous variants are generally neutral, since they cause

no change in protein sequence, unless the mutation site is a part of a conserved motif. The functional impact of the non-essential splice site variants (splice region variants) is based on the presence of a cryptic splice acceptor/donor site in proximity. Lastly, in-frame indels may impact protein function relying upon the site and sequence of insertion/deletion.

Variants in UTRs, non-splice altering intronic regions or intergenic regions are considered low impact. However, there are accumulating data suggesting variants in those regions might be contributing to disease phenotypes as well (Hong et al., 2023; Luo et al., 2023; Whiffin et al., 2020). Nevertheless, available data is still insufficient to make conclusive interpretations for pathogenicity of these variants based on sequence context.

1.2 Human Genetic Diversity

Although humans share a high degree of similarity, a small fraction of variance contributes to a wide range of traits from physical appearance to disease susceptibility. Following the HGP, international collaborations such as the 1000 Genomes Project (1000GP, 2010), the Exome Sequencing Project (ESP6500) (Fu et al., 2013) and the Exome Aggregation Consortium (ExAC) (Lek et al., 2016) provided vital insights into characterization of human genetic variation. To date, these databases have been an essential part of the study of human genetics.

1.2.1 Factors contributing to genetic diversity

While *de novo* mutations in individuals may arise spontaneously during DNA replication or be induced by external factors, most of the variation is still inherited. In addition to genetic drift, the flow of variants between generations is also influenced by non-random factors including natural selection, isolation, and mate selection.

Environmental factors such as climate, lifestyle, exposure to toxic substances and infectious diseases impose a selective pressure on the population. Individuals with favorable traits against these factors are selected and they pass on their genes to next generations by increasing the representation of those traits in the gene pool. Consanguineous marriages, in a similar manner, increases the likelihood of co-segregation of rare alleles. In contrast, if a population is isolated due to geographic barriers or cultural preferences, genetic heterogeneity will be much more reduced.

1.2.2 Genetic variation between/among populations

Each population has their genetic footprint characterized by factors influencing their gene pool. These signatures provide us with a unique opportunity for studying distinct aspects of human genetics. Within the scope of molecular genetics, abundance of variants among or between populations are determined by their effect on fitness of individuals. Variants with beneficial effects increase their frequency within populations and become fixed with time, whereas deleterious ones undergo purifying selection and remain as rare variants. Meanwhile, most variation is neutral or nearly neutral, which is driven by genetic drift, and it is the ultimate source of genetic diversity within or between populations, as proposed by Kimura (1983).

1.2.3 Implications of genetic diversity in health and disease

Although most variation is caused by neutral mutations, genetic drift alone fails to explain why diseases have different genetic backgrounds across different populations. Here, two important concepts come into play: founder effect, where a small group of founders pass on their genes throughout generations by causing a bottleneck, and heterozygote advantage, where some deleterious mutations are found more frequently due to selective advantage for environmental factors (Nussbaum et al., 2016). These factors create inter-population diversity that powers the study of medical genetics. To this end, large population databases such as the Genome Aggregation Da-

tabase (gnomAD), which is sequel to ExAC, provide ancestry-specific allele frequencies for subcontinental populations (Chen et al., 2024). This data is ubiquitously integrated in the decision-making process for determining clinical relevance of a variant.

1.3 Next-generation Sequencing

The race to sequence the human genome has led to an explosion in high-throughput sequencing technologies. Initial attempts were limited to sequencing targeted regions and the coverage was later expanded to whole-genome scale. At the same time, cost-effectiveness and accuracy have been dramatically improved. Dealing with the large data generated by sequencing was not easy at first and computational programs addressed this problem, giving birth to bioinformatics. All these advancements made it possible to use sequencing for first-line diagnostics.

1.3.1 Initial development of sequencing techniques

The first-generation sequencing exploited nucleotide analogues, dideoxynucleoside triphosphates (ddNTPs), which won Frederick Sanger the 1980 Nobel Prize in Chemistry. Sanger sequencing was based on the principle of chain termination caused by ddNTPs during DNA synthesis. Complementary radiolabeled fragments produced by these reactions are then assembled into complete DNA segments based on overlapping sequences. Despite the high accuracy of Sanger method in HGP, it is quite expensive and time-consuming for large-scale use.

Starting with pyrosequencing, sequencing by synthesis approach, in which nucleotides added to the strand are simultaneously identified, have become widespread, addressing the limitations of Sanger sequencing. In pyrosequencing, a light signal is produced by luciferase whenever a nucleotide is integrated, and this signal is read by machine. High-throughput sequencing (HTS) was introduced for the first time by 454

pyrosequencing platforms (Heather & Chain, 2016), yet the process was still expensive to utilize.

1.3.2 Principles of next-generation sequencing platforms

Although next-generation technologies principally follow sequencing-by-synthesis approach, they differ in sample preparation and signal detection methods. PacBio Single-molecule real-time (SMRT) relies upon reading fluorescently labeled nucleotides from circular DNA fragments by immobilized polymerases in porous chips (Hu et al., 2021). Oxford Nanopore sequencing is based on measurement of current change generated by nucleotides passing through nanopores embedded within an electrically resistant membrane (Wang et al., 2021). Illumina dye sequencing employs fluorescently labeled nucleotides incorporated into hybridized DNA fragments by modified polymerases in flow cells (Bentley et al., 2008). Ion Torrent sequencing detects the unique patterns of proton release during DNA polymerization in semiconductor chips (van Dijk et al., 2014).

Next generation sequencing massively parallelized the process of sequencing, bringing the cost of sequencing a whole genome under 1000$. These advancements made it possible the use of sequencing in forensics, medicine, and anthropology. However, these technologies are not still error-free and analyzing the sequencing data requires careful interpretation.

1.3.3 Analysis of next-generation sequencing data

Refinement and appreciation of the data generated by sequencing combine practices of both computer science and genetics. Typical sequencing yields millions of fragmented raw reads with corresponding quality scores, in FASTQ (Fast Quality Scores) format. These high-quality read clusters are mapped to a reference genome by alignment algorithms such as Burrows-Wheeler Alignment (Li & Durbin, 2009), and then stored in SAM (sequence alignment map) format. Finally, variants are identified from

aligned reads and written into VCF (variant call format) files by variant-calling frameworks such as Genome Analysis Toolkit (GATK) (DePristo et al., 2011).

1.4 Molecular Basis of Diseases

Alterations to genetic code can result in an increase or decrease in mRNA or protein abundance, or a change in protein function or structure, leading to disease phenotypes. As molecular dynamics resemble a web of interactions within a cell, understanding the primary cause of a disease is of utmost importance to develop effective therapies. In this context, sequencing technologies precisely empower the identification of molecular basis of diseases.

1.4.1 Molecular mechanisms underlying diseases

Given the complexity of molecular interactions, it is not surprising that single-gene disorders are rarely observed, while most diseases have a polygenic background influenced by environmental factors (Fahed et al., 2020). The mode of inheritance of diseases with genetic basis is strongly associated with gene dosage and effect of causal mutation on protein function. When LoF mutations in a single copy are not tolerated or GoF mutations in a single copy interfere with the residual activity, diseases are dominantly inherited. If both copies of a gene need to be altered to exert a disease phenotype, then diseases are recessively inherited. In contrast, complex diseases originate from accumulation of variants with small effects and are highly influenced by individual's genetic makeup and environmental factors, resulting in a variable penetrance and expressivity among patients with the same disease.

Genetic abnormalities cause diseases at two basic levels: gene expression and protein function. Dysregulation of gene expression due to a malfunction in transcriptional regulation or post-transcriptional modification affects the abundance or stability of mRNA, producing increased or decreased amount of protein product. Alterna-

tively, mutations in coding region can give rise to abnormal protein products with loss-of-function or gain-of-function. These alterations cause a change in cellular homeostasis and eventually diseases. Thus, identification of disease-causing variants is pivotal to the understanding of such abnormalities.

1.4.2 Identification of genes/variants responsible for diseases

Before widespread access to sequencing technologies, linkage analysis, in which genetic markers are tested to see whether they co-segregate with phenotype of interest, was the main methodology for identifying genes or variants underlying diseases. Linkage studies revealed many rare causal variants of large effect for diseases with Mendelian inheritance patterns including cystic fibrosis and Huntington's disease, while this method was not quite applicable for complex diseases, where relatively common variants of small effects are prevalent (Bailey-Wilson, 2011).

Having access to large-scale sequencing data, association studies, which rely upon the observation of disease-causing alleles more abundantly in cases compared to controls, have been extended to genome-wide scale and become the predominant approach in gene identification studies, since genome-wide association studies (GWAS) have not only provided a hypothesis-free approach but also enabled identification of many common variants involving in complex diseases including psychiatric traits, autoimmune diseases and obesity (Marees et al., 2018; Uffelmann et al., 2021).

1.4.3 Incorporation of sequencing data into medical genetics

Low-cost sequencing is taking the place of conventional gene mapping approaches in clinics day by day. Revealing the allelic heterogeneity, sequencing enables comprehensive diagnosis of genetic diseases (Gilissen et al., 2011). It saves lives through early detection by identifying cancer susceptibility and tailoring personalized treatment strategies based on cancer signatures (Esplin et al., 2014). Despite their multifactori-

al nature, sequencing is also employed in determining the genetic basis of complex diseases such as obesity and diabetes (Akbari et al., 2021; Mahajan et al., 2022).

1.5 Variant Prioritization and Classification

Identification of likely causal variants in a dataset is frequently analogized to looking for a needle in a haystack. Once initial sequencing initiatives started detecting many variants, it quickly became clear that mutations are common within our genomes and a thorough investigation is required to identify true causal variants. Human geneticists and bioinformaticians developed variant quality control workflows, prioritization strategies, prediction programs and classification frameworks to overcome this challenge.

1.5.1 Functional annotation of sequence variants

Variant annotation, which basically involves retrieving information from large databases and assigning them to variants, is an integral component of variant calling. This information is later used in prioritization to identify likely disease-causing variants. ANNOVAR (Wang et al., 2010), SnpEff (Cingolani et al., 2012) and Variant Effect Predictor (McLaren et al., 2016) are leading annotation software integrated into variant calling pipelines.

Functional annotation typically labels variants as exonic, intronic or splice-altering (see variant consequences are listed in section 1.1.3) and identifies amino acid change in corresponding gene, transcript, and exon. Depending on the database employed, information about gene function, affected domains, associated diseases, or pathways involved can be obtained as well. Prediction programs are also heavily exploited to determine pathogenicity of variants.

Frequency annotation enables filtering out excessively common variants in populations, which are mostly neutral. Variants with frequencies greater than 1% are

usually considered as common polymorphisms. However, filtration threshold may vary depending on the disease under assessment, given that common variants were previously shown to be involved in multifactorial diseases (Helgadottir et al., 2007; Pardiñas et al., 2018).

1.5.2 Variant quality control and filtration strategies

After variant calling, the first step is to perform quality filtration to eliminate sequencing artifacts and low-quality variants based on read depth, sequencing quality, mapping quality, position bias, strand bias or heterozygosity. Another approach, namely variant quality score recalibration (VQSR), employs machine-learning models trained on known variants to estimate the confidence of suspected variants (Van der Auwera et al., 2013).

Once high-quality variants are compiled, a prioritization pipeline is established in compliance with the research questions of the study. Population filtration is a crucial step of variant interpretation, since disease-causing variants for monogenic disorders are expected to be rare in populations due to negative selection. Next, it is a common practice to focus on variants in coding regions. If a candidate gene approach is adopted, the dataset is narrowed down to variants in the gene panel. Gene discovery studies though usually apply more stringent criteria by analyzing only pLoF variants. Gene constraint and phylogenetic conservation are other metrics incorporated in variant prioritization workflows (Eilbeck et al., 2017).

1.5.3 *In silico* pathogenicity prediction tools

While functional annotation provides information regarding the genomic context of the variation, they do not inform about its clinical significance, given the complex nature of genetic interactions. It is where computational prediction programs come into play, utilizing analytical or machine learning algorithms to predict deleterious-

ness of a variant based on similarities with known pathogenic variants in terms of genomic context and functional relevance (Katsonis et al., 2022).

First examples of functional prediction programs are PolyPhen, which relies on protein structure and function (Ramensky et al., 2002), and SIFT, which is based on sequence homology (Ng & Henikoff, 2003). While many other single nucleotide variant prediction tools followed Polyphen and SIFT, PROVEAN was the first to came up with an algorithm that provides predictions for indels (Choi et al., 2012).

Combining multiple prediction tools following distinct methodologies with their weaknesses and strengths, meta-predictors started being developed to achieve higher predictive values. CADD (Kircher et al., 2014), a widely used ensemble framework, integrates at least 50 features including prediction scores, conservation metrics, structural and functional context, population frequencies. BayesDel (Feng, 2017), MetaSVM (Dong et al., 2015) and REVEL (Ioannidis et al., 2016) are other notable ensemble prediction tools and they aggregate scores, for known pathogenic variants, from more than a dozen predictors by various machine learning algorithms.

To address distinct aspects of genomic variation, prediction programs for specific purposes were also developed. For instance, MitoTip (Sonney et al., 2017) specifically evaluates mitochondrial tRNA variants, SpliceAI (Jaganathan et al., 2019) provide predictions for splice variations, and CardioBoost (Zhang et al., 2021) classifies variants in genes associated with cardiovascular diseases. Evolutionary conservation tools including phastCons (Siepel et al., 2005) and phyloP (Pollard et al., 2010) approach the problem from a phylogenetic perspective, comparing base-level changes in genomes across primates, mammals, and vertebrates. While distinguishing GoF and LoF variants by prediction tools such as LoGoFunc (Stein et al., 2023) may give a better idea about the molecular context.

1.6 ACMG/AMP Standards and Guidelines

Although clinical laboratories heavily integrate computational programs into their variant classification frameworks, *in silico* predictions only stand as supporting evidence for pathogenicity, rather than sole determinant. These frameworks, whose history predates prediction tools, were born out of a need for a standardized approach for interpretation of ever-growing number of sequence variants.

American College of Medical Genetics and Genomics (ACMG) first time provided a standardized framework for variant classification in 2000 (Kazazian et al., 2000). The ACMG Guidelines were revised twice in 2008 (Richards et al., 2008) and 2015 (Richards et al., 2015), with another revision expected in 2024. Unlike prediction tools that make inferences based on patterns, the ACMG Standards and Guidelines provide an evidence-based analytical framework to estimate the probability that a variant is pathogenic.

1.6.1 The 2015 ACMG/AMP Criteria

The guideline published by American College of Medical Genetics and Genomics (ACMG) and Association for Molecular Pathology (AMP) in 2015 (Richards et al., 2015) consists of 16 pathogenic (P) and 12 benign (B) criteria described in Table 1 and 2, respectively. The criteria basically depend on population frequencies, *in silico* predictions, functional assays and case-control studies. Each criterion has a corresponding evidence level as follows: supporting (P), moderate (M), strong (S), very strong (VS), stand-alone (A). Once a variant is evaluated for each criterion with available evidence, the criteria is combined into one of 5 possible classifications: benign (B), likely benign (LB), variant of uncertain significance (VUS), likely pathogenic (LP), pathogenic (P). The combination rules in the 2015 Guidelines are shown in Figure 2.

Table 1: Pathogenic criteria in the ACMG/AMP Standards and Guidelines (Richards et al., 2015).

Very strong evidence of pathogenicity	
PVS1	Null variant (nonsense, frameshift, canonical +/−1 or 2 splice sites, initiation codon, single or multi-exon deletion) in a gene where loss of function (LoF) is a known mechanism of disease
Strong evidence of pathogenicity	
PS1	Same amino acid change as a previously established pathogenic variant regardless of nucleotide change
PS2	*De novo* (both maternity and paternity confirmed) in a patient with the disease and no family history
PS3	Well-established *in vitro* or *in vivo* functional studies supportive of a damaging effect on the gene or gene product
PS4	The prevalence of the variant in affected individuals is significantly increased compared to the prevalence in controls
Moderate evidence of pathogenicity	
PM1	Located in a mutational hot spot and/or critical and well-established functional domain (*e.g*. active site of an enzyme) without benign variation
PM2	Absent from controls (or at extremely low frequency if recessive) in Exome Sequencing Project, 1000 Genomes or ExAC
PM3	For recessive disorders, detected in *trans* with a pathogenic variant
PM4	Protein length changes due to in-frame deletions/insertions in a non-repeat region or stop-loss variants
PM5	Novel missense change at an amino acid residue where a different missense change determined to be pathogenic has been seen before
PM6	Assumed *de novo*, but without confirmation of paternity and maternity
Supporting evidence of pathogenicity	
PP1	Co-segregation with disease in multiple affected family members in a gene definitively known to cause the disease
PP2	Missense variant in a gene that has a low rate of benign missense variation and where missense variants are a common mechanism of disease
PP3	Multiple lines of computational evidence support a deleterious effect on the gene or gene product (conservation, evolutionary, splicing impact, etc)
PP4	Patient's phenotype or family history is highly specific for a disease with a single genetic etiology
PP5	Reputable source recently reports variant as pathogenic but the evidence is not available to the laboratory to perform an independent evaluation

Table 2: **Benign criteria in the ACMG/AMP Standards and Guidelines** (Richards et al., 2015).

Stand-Alone evidence of benign impact	
BA1	Allele frequency is above 5% in Exome Sequencing Project, 1000 Genomes, or ExAC
Strong evidence of benign impact	
BS1	Allele frequency is greater than expected for disorder
BS2	Observed in a healthy adult individual for a recessive (homozygous), dominant (heterozygous), or X-linked (hemizygous) disorder with full penetrance expected at an early age
BS3	Well-established *in vitro* or *in vivo* functional studies shows no damaging effect on protein function or splicing
BS4	Lack of segregation in affected members of a family
Supporting evidence of benign impact	
BP1	Missense variant in a gene for which primarily truncating variants are known to cause disease
BP2	Observed in *trans* with a pathogenic variant for a fully penetrant dominant gene/disorder; or observed in *cis* with a pathogenic variant in any inheritance pattern
BP3	In-frame deletions/insertions in a repetitive region without a known function
BP4	Multiple lines of computational evidence suggest no impact on gene or gene product (conservation, evolutionary, splicing impact, etc)
BP5	Variant found in a case with an alternate molecular basis for disease
BP6	Reputable source recently reports variant as benign but the evidence is not available to the laboratory to perform an independent evaluation
BP7	A synonymous (silent) variant for which splicing prediction algorithms predict no impact to the splice consensus sequence nor the creation of a new splice site AND the nucleotide is not highly conserved

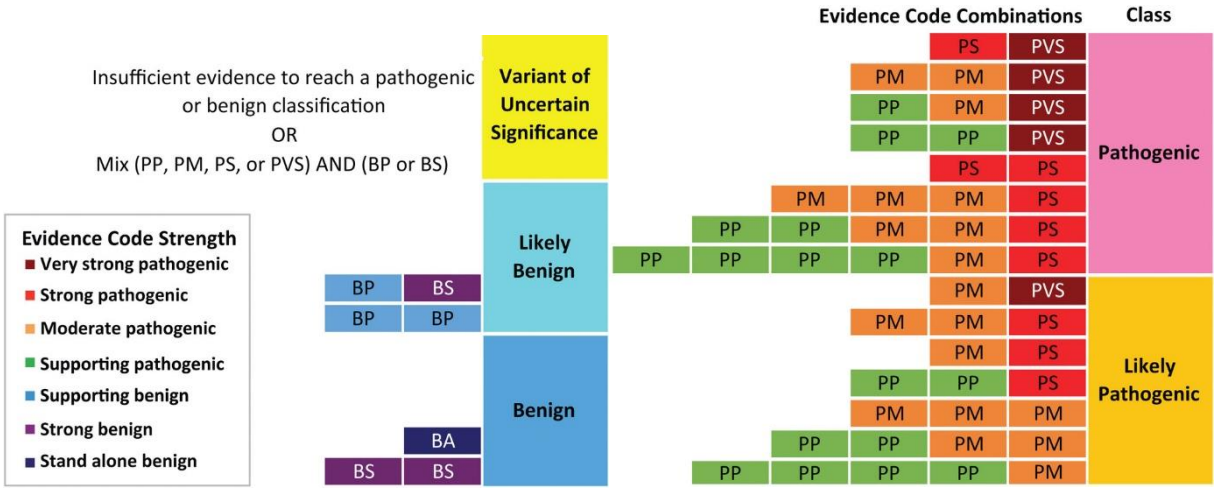


Figure 2: The ACMG Guidelines criteria combination. Letters P and B in the first letter of evidence codes represent pathogenic and benign, respectively. The following letters indicate the strength of evidence such as stand-alone (A), very strong (VS), strong (S), moderate (M) and supporting (P). Adapted from Brnich et al. (2018).

1.6.2 Limitations of the original framework

In 2016, nine independent clinical laboratories were asked to assess the pathogenicity of ninety-nine variants by the ACMG/AMP Standards (Amendola et al., 2016). The initial concordance across these diagnostics laboratories for these variants was found to be only 34%. After consensus discussions, the concordance increased to 71%. The systematic analysis revealed the primary causes of discordance as (i) subjective interpretation of the rules, (ii) improper use of the criteria, (iii) lack of experience/training in medical genetics and (iv) unfamiliarity with the guidelines. Therefore, Amendola et al. provided a set of recommendations and rule clarifications to increase consistency.

Harrison et al. (2017) reported that most of the discrepancy between classifications stems from the application of functional data (PS3, PM1, PP2, BS3) and population data (BS1, BS2, PS4). Harrison et al. (2019) further investigates the self-consistency of the rules and emphasized the problems with criteria concerning population data (BA1, BS1, PM2), functional data (PVS1, PM1, PP2), case-level data (PS2, PS4, PM3, PM6) and rule combinations.

Since the original guidelines were intended to be a suggestive guide rather than a prescriptive framework, most rules are broadly defined to allow clinical geneticists a degree of freedom. For example, BS1 rule (variant is too common to be pathogenic for disease) can be interpreted differently based on the data and such data might not always be available for every disease. Moreover, reliability of variant curations in ClinVar, a database commonly used for PP5 and BP6 rules, is another point of discussion (Biesecker et al., 2018; Richards et al., 2018). Another example is the misuse of high-level evidence such as PVS1 (null variant in LoF gene), which can dramatically affect the final verdict. In addition, the classification of variants with conflicting evidence, where pathogenic and benign evidence are activated together, was not addressed, leading to an increase in the number of variants of unknown significance. All these scenarios and the evidence from the studies mentioned above demonstrate a need to refine and improve the rules to maintain consistency.

1.6.3 ClinGen rule specifications

Although ACMG/AMP essentially standardized the variant curation and terminology, some rules remained ambiguous by leading to different interpretations. To improve the guidelines, the Clinical Genome Resource (ClinGen) published numerous recommendations for the ACMG/AMP Criteria (Abou Tayoun et al., 2018; Brnich et al., 2019; Ghosh et al., 2018; Pejaver et al., 2022; Walker et al., 2023) and disease-based specifications for variant classification (Gelb et al., 2018; Kelly et al., 2018; Lee et al., 2018; Mester et al., 2018; Zastrow et al., 2018; Luo et al., 2019; Oza et al., 2018).

1.6.3.1 PVS1 – Null variant rule

ClinGen has published their first recommendations to prevent inappropriate use of the single pathogenic very strong rule (Abou Tayoun et al., 2018). PVS1 was supposed to be assigned for null variants (stop-gain, frameshift, canonical splice variant, start-loss, exon deletion) in diseases with known LoF mechanism. Although null variants were shown to be extensively involved in LoF mutations, the ACMG/AMP Guidelines did not explicitly address specific cases that require a reduction in evidence weight. Abou Tayoun et al. provides a detailed decision tree to overcome these uncertainties.

The decision tree consists of a different branch for each consequence. It considers NMD prediction, functional relevance of altered region, protein function, percentage of removed region and population frequency for nonsense, frameshift, canonical splice site and exon deleting variants. Depending on the tree, these variants can activate PVS1 rule at moderate, strong, or very strong levels of evidence. If a variant is absent from a biologically relevant transcript, then it is not considered for PVS1. Claiming that start-altering variants are abundantly present within our genome, they recommend applying PVS1 as supporting or strong evidence based on the presence of an alternative start in a different transcript and clinical relevance of the removed region.

1.6.3.2 PS3/BS3 – Functional assay rule

Given the variability of experimental approaches to study a disease mechanism, the reliability and resemblance of these assays have always been questioned. To resolve these disputes, ClinGen published specifications for determining the validity of functional evidence, providing a decision tree that evaluates disease mechanism, applicability of the assay, use of controls and statistical metrics (Brnich et al., 2019).

1.6.3.3 BA1 – Common variant rule

The BA1 criterion rule in the ACMG Guidelines is the only rule that is sufficient to classify a variant with a frequency above 5% in the general population as benign. However, variants of low penetrance might be commonly present in populations. ClinGen analyzed variants that fit this definition on ClinVar and identified nine common low-penetrance pathogenic variants in seven genes (Ghosh et al., 2018). It was proposed to exempt these variants from the application of BA1 rule. Although they also called on laboratories to make nominations to the exception list, there have been no updates since then.

1.6.3.4 PP3/BP4 – *In silico* prediction rule

The increasing number of computational predictors has raised the question of which algorithm is better and how reliable they are. Ghosh et al. (2017) revealed that the community continues using legacy tools, even though they perform worse than newer counterparts. Thus, ClinGen calibrated thirteen prediction tools by evaluating their accuracy for known pathogenic variants (Pejaver et al., 2022). The study recommends the use of BayesDel, MutPred2, REVEL or VEST4 with defined scores ranges corresponding to levels of evidence from supporting to strong.

1.6.3.5 Splice variants

Considering the significant contribution of splice defects to diseases (Danis et al., 2021; Singh & Cooper, 2021), the accurate interpretation of splice variants is profoundly important. While the 2015 ACMG Guidelines consider the canonical splice variants as null variants, no detailed specifications were provided for non-canonical splice variants despite their potential contribution in unsolved cases (Blakes et al., 2022). Therefore, ClinGen provided further guidance for the classification splice-altering variants (Walker et al., 2023). The recommendations include a decision tree combining tools predicting splice alterations, previous clinical reports and PVS1 flowchart by Abou Tayoun et al. (2018) for variants inside (canonical) or outside donor/acceptor ±1,2 dinucleotide sites (non-canonical).

1.6.3.6 Point-based scoring system

After evaluating a variant for each criterion, it is classified into a pathogenicity class based on rule combinations described in Figure 2. However, this heuristic classification system was limited in considering different rule combinations and interpreting variants with conflicting evidence. Thus, Tavtigian et al. (2018) developed a Bayesian framework, a statistical inference method which enables calculating posterior probabilities using available information, for aggregating criteria into classification groups. Later, this statistical framework was updated to a simpler score-based system (Tavtigian et al., 2020). In this system, each weight of evidence has a corresponding score, and a variant is classified according to the total scores accumulated.

1.6.3.7 Variant Curation Expert Panels

Along with Sequence Variant Interpretation (SVI) subgroup aimed at improving the ACMG/AMP Guidelines, ClinGen established VCEPs, which consists of medical geneticists specialized in different diseases, to provide disease-specific recommendations for the interpretation of variants in clinically actionable genes. There are more

than twenty approved expert panels that have published their specifications to date (https://clinicalgenome.org/affiliation/vcep).

VCEP specifications primarily differ in the application of population frequency rules (PM2, BA1, BS1, BS2) from the ACMG/AMG Guidelines, introducing disease-specific thresholds. Claiming the thresholds in the original guidelines are too conservative, the panels almost collectively utilize Whiffin-Ware calculator (Whiffin et al., 2017) to calculate maximum credible allele frequency based on prevalence, penetrance, and heterogeneity. Additionally, VCEPs provide further specifications for the use of *in silico* predictors, functional assays, and case-control studies in variant curations. VCEPs do not only improve the high-confidence variant curations for actionable diseases, but also inspire SVI for updates to the ACMG/AMP Guidelines in terms of aspects that can be generalized to any diseases.

1.7 Automated Variant Classification Platforms

One of the concerns raised by Amendola et al. (2016) was the methodological differences in the application of the ACMG Rules. Thus, they encouraged to computationally automate the process of variant curation to minimize the discordance. Li & Wang (2017) have been first to answer this call with their tool, InterVar. There have been other solutions by clinical diagnostic companies. Of them, Franklin (https://franklin.genoox.com) and VarSome (Kopanos et al., 2019) are frequently used platforms and provide limited access to their classification tools. Currently available classification platforms and tools are listed in Table 3.

Table 3: Currently available variant classification tools.

Tool	Developer	Release Year	Guidelines	Paper
AAVC	Bilkent	2024	ACMG/ClinGen+	This study
AION	Nostos	?	ACMG	N/A
Alamut	Sophia	?	ACMG/?	N/A
Alissa Interpret	Agilent	?	ACMG/?	N/A
CharGer	WUSM	2019	ACMG	Scott et al. (2019)
Classifi	Ambry	?	ACMG/?	N/A
Emedgene	Illumina	?	ACMG/?	N/A

eVai	EnGenome	2022	ACMG/ClinGen	Nicora et al. (2022)
Franklin	Genoox	?	ACMG/ClinGen	N/A
InterVar	USC	2017	ACMG	Li & Wang (2017)
MAGI-ACMG	MAGI	2023	ACMG/ClinGen	Cristofoli et al. (2023)
Moon	Invitae	?	Sherloc	N/A
SEQ Platform	Genomize	?	ACMG/?	N/A
Sunquest Mitogen	Sunquest	?	ACMG/?	N/A
VarSeq	Golden Helix	?	ACMG/ClinGen	N/A
VarSome	Sephator	2019	ACMG/ClinGen	Kopanos et al. (2019)

ACMG: The ACMG Guidelines are followed, ACMG/?: The ACMG Guidelines are followed but it is unclear whether the ClinGen Modifications were also implemented. ACMG/ClinGen: The ACMG Guidelines are followed, and the ClinGen Guidelines are implemented to some extent. ACMG/ClinGen+: The ACMG Guidelines and the ClinGen Modifications are fully implemented including latest splice variant specifications (Walker et al., 2023). Sherloc: Invitae's ACMG-derived variant classification framework (Nykamp et al., 2017), USC: University of Southern California.

1.7.1 InterVar, Franklin and VarSome

InterVar (Li & Wang, 2017) is a command-line tool that takes VCF as input and annotates it with ACMG classifications and criteria met. If the VCF input is unannotated, InterVar first calls ANNOVAR to obtain necessary information for the interpretation. InterVar basically employs VCF annotations and pre-generated tables and checks if variant under assessment satisfies the criteria, automating 18/28 of the ACMG Criteria. Users can manually provide input for the remaining criteria. The software can also be accessed via a web server (https://wintervar.wglab.org).

Even though InterVar was a pioneering work, it has certain limitations today. Firstly, it is not quite accessible since it requires third-party software for annotation and wInterVar is not a rapid and flexible solution. Next, the pre-generated tables employed in the automated interpretation are outdatednow and we have better alternatives and improved solutions today. Most importantly, InterVar was released long before the ClinGen Specifications emerged. Thus, the software makes its interpretations by the 2015 ACMG Standards. Therefore, these limitations make it a less attractive solution for automating variant classification nowadays.

Franklin and VarSome provide solutions for variant interpretation powered by artificial intelligence (AI). They integrate information from databases including gnomAD

(Chen et al., 2024), ClinVar (Landrum et al., 2014), UniProt (Apweiler et al., 2004) and Ensembl (Hubbard et al., 2002) in their classification algorithms. Because Franklin has still not published a peer-reviewed paper, the methodology is not totally clear. It is known that either tool follows the ACMG Guidelines and some ClinGen specifications. However, they fail to implement the recent ClinGen specifications for splice variants (Walker et al., 2023). Whereas Franklin and VarSome are effective solutions, neither tool provides a detailed explanation on their methodology, raising questions about their reliability and consistency.

1.7.2 Other classification programs

There are other approaches with different uses that are promising. CharGer (Scott et al., 2019) interprets variants in a pre-annotated VCF file for the 2015 ACMG Criteria and some custom criteria via the command-line. AutoPVS1 (Xiang et al., 2020) automates the null variant rule (PVS1) based upon the ClinGen Specifications (Abou Tayoun et al., 2018). VIP-HL (Peng et al., 2021) classifies hearing loss (HL) associated variants according to the ClinGen HL VCEP Specifications (Oza et al., 2018). MAGI-ACMG (Cristofoli et al., 2023) is another classification algorithm built upon VarSome and assesses variants following the ACMG and the Association for Clinical Genomic Science Guidelines (Ellard et al., 2020).

1.7.3 Shortcomings of classification programs

Currently available programs for automating the variant classification lack detailed explanation for their methodology. The ACMG Guidelines are not a set of strictly defined rules, but a broad set of recommendations for clinical geneticists. It is today known that many clinical diagnostics labs follow their own approaches to practice the ACMG Guidelines. Moreover, VCEPs implement their own disease- or gene-specific modifications to the guidelines. These require applying adjustments to the framework for the interpretation of different genes. Available platforms fail to provide necessary

details on how their algorithms function. For the very same reason, it is not possible to assess the consistency of these tools since they are not well-documented. The way they evaluate the pathogenicity slightly deviates from the ACMG guidelines and the ClinGen Specifications as summarized in Table 4. Data privacy is another implication of these online platforms. Institutional, organizational, or governmental regulations might not allow the sharing and storage of sensitive genetic data with third parties. It is thus understandable that labs may prefer to run such algorithms in their own devices.

Generally, available programs either fail to follow the latest recommendations or lack practicality of use. None of the tools listed in Table 4, except AAVC, completely incorporates the ClinGen recommendations together with the ACMG Guidelines. Apart from that, they require third party software to run.

Table 4: **The differences between AAVC, Franklin, VarSome**. The differences in the application of frequency rules (PM2, BA1, BS1, BS2), reported variant rules (PP5, BP6), in silico prediction rules (PP3, BP4), splice variants and VUS subclassifications based on the ACMG Guidelines are described.

	AAVC	**Franklin**	**VarSome**
Frequency rules	gene-specific cut-offs	gene-specific cut-offs	universal cut-offs
Reported variants	adjusts evidence level by submission confidence	marks submissions before 2015 as low confidence	adjusts evidence level by submission confidence
***In silico* predictions**	follows ClinGen specifications (Pejaver et al., 2021)	aggregates prediction tools	uses MetaRNN
Splice variants	follows ClinGen specifications (Walker et al., 2023)	unknown	unknown
VUS subclassification	implements	does not implement	does not implement

1.8 Rationale and Motivations of the Work

Manual interpretation of sequence variants by the ACMG/AMP Guidelines is a highly subjective and an error-prone exercise due to open-ended definitions of the rules. Available platforms that automate the process come with their restrictions. They prefer not to share their methodology in a peer-reviewed journal due to commercial considerations and do not guarantee data privacy and confidentiality. While alternative solutions exist, they fail to offer convincing frameworks since they either address to specific aspects of the classification process or fail to provide comprehensive, consistent and user-friendly frameworks supported with up-to-date methodologies. Therefore, the scientific community is in need of an alternative platform, which is highly accurate, inherently consistent and up to date. Here, the work presented is a framework that enables interpretation of sequence variants by the ACMG/AMP Standards, the ACGS Best Practices, the ClinGen Specifications and other recommendations from the literature.

Chapter 2

Materials and Methods

2.1 Incorporated Databases

AAVC retrieves information regarding genomic features, *in silico* predictions and population frequencies by maintaining access to sixteen different databases. The tool executes a query algorithm to communicate with Ensembl and gnomAD web servers via their application programming interfaces (APIs). In addition, further information is obtained from pre-generated local databases. The raw data generated from these sources is later wrangled and assigned to variants for the upcoming steps.

2.1.1 Genome databases

Ensembl (Harrison et al., 2024) is the primary data source integrated in AAVC, providing variant consequences, genomic features, and protein sequences. The database, which is one of the cornerstones of genomics research, has been maintained by the European Bioinformatics Institute since 2002 (Hubbard et al., 2002). Ensembl VEP (McLaren et al., 2016) is an extensively integrated algorithm into the prioritiza-

tion pipelines to predict variant consequences. BioMart (Smedley et al., 2009) is a query tool that enables easy access to information stored in Ensembl databases via web browsers. Ensembl REST API (Yates et al., 2015) also provides programmatic access, allowing users to directly interact with the databases for large queries.

PanelApp by Genomics England (Martin et al., 2019) is an expert-curated database for gene-disease associations. The mode of inheritance and pathogenicity of genes from PanelApp are used for variant interpretation by AAVC. UniProt database (UniProt, 2023) is integrated into AAVC to locate functional domains for a given gene. RepeatMasker (Smit et al., 1996-2010) annotates repeat elements, satellite regions and low complexity sequences and enables AAVC to identify error-prone regions.

2.1.2 Variant databases

ClinVar (Landrum et al., 2014) is a large public database of variant submissions and associated phenotypes. As of June 2024, ClinVar contains 2.8 million variants with their classification summaries, providing a retrospective perspective for understanding the clinical relevance of human genetic variation. gnomAD (Chen et al., 2024) is a browser that aggregates genetic variation of populations. The latest version, gnomAD v4, empowers genetic studies by providing variation data from approximately 800,000 individuals across diverse populations.

2.1.3 Prediction tools

Given the high accuracy and specificity of BayesDel compared to other tools (Tian et al., 2019), it is the preferred meta-predictor for missense variants in AAVC. Feng (2017) developed BayesDel by combining PolyPhen-2 (Adzhubei et al., 2010), SIFT (Ng & Henikoff, 2003), FATHMM (Shihab et al., 2013), LRT (Chun & Fay, 2009), MutationTaster (Schwarz et al., 2010), MutationAssessor (Reva et al., 2011), phyloP (Pollard et al., 2010), GERP++ (Davydov et al., 2010), and SiPhy (Garber et al.,

2009). Moreover, BayesDel was one of the four tools recommended to be used for *in silico* evidence by ClinGen (Pejaver et al., 2022).

SpliceAI (Jaganathan et al., 2019) is a deep learning-based prediction tool for splice alterations. In the ClinGen's guide for splice variant interpretation, SpliceAI was selected for its sensitivity and specificity that outperform other splice predictors (Walker et al., 2023).

SpliceVault (Dawes et al., 2023) is a mis-splicing events database constructed on RNA sequences of 330,000 samples. The rationale behind the SpliceVault is that the consequence of a variation affecting splicing can be inferred from naturally observed mis-splicing events.

phyloP100way (Pollard et al., 2010) scores the conservation for each base for ninety-nine vertebrates in comparison to human genome. ClinGen VCEPs recommend integration of phyloP100way into the classification pipelines, since it may give an idea on functional importance of the region (Luo et al., 2019; Mester et al., 2018).

2.2 AAVC Framework Design

AAVC is a variant classification framework that automates the entire process from aggregating evidence from public resources to assessing the clinical relevance of a variant based on the ACMG Guidelines and the ClinGen Specifications. It is developed as a Python script, which is a general-purpose programming language (https://www.python.org) and built on pandas (https://pandas.pydata.org) and PostgreSQL (https://www.postgresql.org) packages that enable handling large datasets. In the generation and manipulation of in-house tables, R environment (https://www.r-project.org) was exploited. AAVC requires variant position, reference and alternative allele information as input and generates a tab-separated output file with variant classification and assertion criteria.

2.2.1 Data Query

An application programming interface (API) is software that facilitates the communication between a client and a server. AAVC retrieves essential information from Ensemble VEP and gnomAD via their APIs. Once the user provides input, the variants are split into batches of twenty-five variants due to quotas imposed by the databases. Then a query is made in these databases through their APIs for each batch.

2.2.1.1 Variant consequence

Ensembl VEP provides the most critical information in variant prioritization, which is the consequence of a variant at mRNA and protein level. Gene, transcript, strand, nucleotide change, amino acid change, canonical status are other annotations retrieved via the Ensembl REST API. Pre-calculated BayesDel, SpliceAI and phyloP100way scores are also accessible via the Ensembl API services. However, functional annotation is much more complicated, considering a typical variant affects more than once transcript due to alternative splicing. Therefore, all the information retrieved is supposed to be transcript dependent. To overcome this problem, AAVC runs the Ensembl REST API for VEP with pick parameter (--pick), which prioritizes the affected transcripts based on biological relevance of the transcript, transcript biotype, functional impact of the variant, and transcript length. Then the variant is annotated and curated for the selected transcript.

2.2.1.2 Variant frequency

Distinguishing between common and rare variations is an important aspect of variant classification, and gnomAD stands out as a unique resource documenting human variation. Nonetheless, the pattern of variation is greatly confounded by the genetic structure of populations. Therefore, the observation of a variant at low frequency in the general population does not necessarily mean that it is under selective pressure, as it can be found at higher frequencies in another population, indicating a neutral

effect. To this end, gnomAD v4 provides allele frequencies for ten distinct genetic ancestry groups: Admixed American, African/African American, Amish, Ashkenazi Jewish, East Asian, Finnish, Non-Finnish European, Middle Eastern, Remaining and South Asian. Filtering allele frequency (FAF), at the lower limit of the 95% confidence interval of the observed frequency, is calculated for each non-bottlenecked group. Then based on the ancestry group with maximum FAF, each variant is assigned with its GroupMax FAF value. In AAVC, these GroupMax FAF annotations are obtained together with number of individuals with homozygous alleles for each variant via gnomAD API services.

2.2.1.3 Gene-disease associations

For inheritance patterns and disease associations of genes, AAVC refers to Genomics England PanelApp, which guides medical geneticists with their expert-curated panels for over five thousand genes. Gene-level annotation is especially important for complex diseases, where pleiotropic effects are not uncommon. Plus, the mode of inheritance of a gene modifies the implementation of population data rules. AAVC retrieves that information locally using annotated gene identifiers.

2.2.2 Local Tables

Although the essential annotation comes through web services, AAVC uses local databases to retrieve consequence-specific annotation, reducing the processing times. These tables are derived from large public databases including Ensembl BioMart, gnomAD and ClinVar. Given the large raw file sizes, all the manipulation, including pruning, grouping, cross-annotating, and deriving the raw data, was performed via Linux terminal and R environment by data.table package (https://rstudio.github.io/DT) that offers a rapid manipulation for large datasets by multithreading. All local tables are stored in PostgreSQL with appropriate indices and interacted via SQLAlchemy (https://www.sqlalchemy.org) when required.

2.2.2.1 Genomic features

exons.txt: It is a BioMart-derived tab-delimited table that was compiled to complement the information in Ensemble VEP. The table consists of 21 columns described in Table 5. It contains gene and transcript identifiers, protein length, transcript and exon positions, exon ranks and counts, and other exon-specific annotations.

Table 5: Structure and description of exons.txt file.

#	Column name	Description
1	gene	HGVS gene symbol
2	synonyms	HGNC gene synonyms
3	ensg_id	Ensembl gene ID (ENSG…)
4	biotype	Transcript class (protein coding, ncRNA, pseudogene, etc.)
5	enst_id	Ensembl transcript ID (ENST…)
6	refseq_id	RefSeq transcript ID (NM_…)
7	is_canonical	Ensembl canonical transcript status (1 or 0)
8	prot_length	Protein length in amino acids
9	alt_start	Position of the second methionine in the transcript
10	alt_stop	Position of the first stop codon in the 3' UTR
11	chr	Chromosome name
12	strand	Forward (1) or reverse (-1) strand
13	tcpt_start	Transcript start position
14	tcpt_end	Transcript end position
15	cds_start	Coding sequence start position
16	cds_end	Coding sequence end position
17	exon_rank	Exon order in the transcript
18	exon_count	Total number of exons in the transcript
19	exon_start	Exon start position
20	exon_end	Exon end position
21	highest_freq	Highest pathogenic variant frequency within the exon

The primary information in the features table was provided by Ensembl Biomart, however it was also annotated by additional resources. Gene synonyms were acquired from HUGO Gene Nomenclature Committee (HGNC) web services via their BioMart (https://biomart.genenames.org). To identify the alternative start and stop positions, the corresponding sequences (coding DNA and 3' UTR sequences) were retrieved via Ensembl Biomart (https://www.ensembl.org/biomart/martview), then positions were computed via R and matched with the transcript IDs in the table. Of variants in gnomAD v4.1, ClinVar variants classified as pathogenic or likely pathogenic were listed. Then, the highest pathogenic variant frequency for each exon was determined.

domains.txt: UniProt (UniProt, 2023) provides access to expert curated protein features including processing sites, domains, interaction sites, protein modifications and secondary structures. Functionally relevant protein features and class-specific secondary structures based on the characterization of Iqbal et al. (2020). The list of features that domains.txt includes is shown in Table 6.

Table 6: **Protein features and secondary structures from UniProt**.
All protein features included in the domain dataset in the AAVC are listed below.

Active site	Domain	Non-standard amino acids
Beta strand*	Functional region	Peptide
Binding site	Functional site	Sequence motif
Coiled coil	Helix*	Topological domain
Disulfide bond	Intramembrane	Transmembrane
DNA-binding site	Modified residue	Zinc finger

* Includes only specific protein classes for which these features are functionally relevant as reported by Iqbal et al. (2020).

repeats.txt: RepeatMasker annotations are available on UCSC Genome Browser (https://hgdownload.soe.ucsc.edu/goldenPath/hg38/database). Using ClinVar data, each repeat range in the RepeatMasker database was annotated with pathogenic variant burden. In addition to variants with pathogenic, likely pathogenic and risk factor classifications, VUS with only likely pathogenic and/or pathogenic submissions were also included.

homopolymers.txt: Homopolymeric regions are mostly misread by sequencing platforms due to polymerase slippage during amplification (Singer-Berk et al., 2023). This table contains all homopolymers with at least five nucleotide repeats for each chromosome.

pext.txt: The expression patterns of alternative transcripts across tissues differ significantly, contributing complexity of the variant interpretation. Therefore, the expression data should be integrated into variant prioritization workflows for accurate interpretations. Cummings et al. (2020) introduced a metric called pext (proportion expressed across transcripts) to account for the base-level expressivity across transcripts. gnomAD provides mean pext values across tissues on their website (https://gnomad.broadinstitute.org/downloads). However, positions in the provided data were calculated for gnomAD v2.1.1 and aligned to hg19. Hence it was converted

to hg38 positions via LiftOver software from the UCSC (https://hgdownload.soe.ucsc.edu/downloads.html). Deriving this data, the pext table was generated by matching the values to genes and grouping gene segments by pext values (Table 7).

Table 7: A sample from pext.txt.

gene_id	chr	start_pos	end_pos	mean_pext
ENSG00000237683	1	138530	139309	1.00000
ENSG00000235249	1	450738	451676	1.00000
ENSG00000185097	1	685716	686654	1.00000
ENSG00000187634	1	925942	930311	0.16993
ENSG00000187634	1	930312	935793	0.88016
ENSG00000187634	1	935794	939129	0.87175
ENSG00000187634	1	939275	939291	0.93715
ENSG00000187634	1	939292	939412	0.81618
ENSG00000187634	1	939413	939460	0.10595
ENSG00000187634	1	941144	943808	0.81618

regional_constraints.txt: Analyzing the genomic regions that do not tolerate the missense variation, gnomAD provided regional missense constraints based on population data for coding segments. The constraints are basically derived by comparing the observed and expected number of missense variants (o/e) in a region (Samocha et al., 2017). As the data is calculated for gnomAD v2.1.1 dataset, it was provided in hg19. Therefore, it was converted to hg38 positions by LiftOver. Unaligned positions were omitted. An o/e threshold of 0.4 ($p \leq 0.001$) was used to define the constrained regions.

vault.txt: SpliceVault (Dawes et al., 2023) is a compilation of naturally occurring mis-splicing events derived from RNA-seq data. Processing the information in the raw dataset, the functional consequence of the most common mis-splicing event of each splice site was annotated by characterizing the change in frame, alternative stop site for intron retention, lost protein percentage for exon deletion and skipping, truncated range for mis-splicing events causing a stop gain.

2.2.2.2 Gene constraints

disease_assoc.txt: PanelApp (Martin et al., 2019) is a gene-disease database curated by expert reviewers. The database extends the gene-disease associations to mode of pathogenicity and mode of inheritance annotations, which are two valuable information pieces integrated into the AAVC.

gene_metrics.txt: It stores gene based LoF metrics, haploinsufficiency curations, ClinVar statistics and frequency cutoffs. gnomAD provides statistical metrics based on population data: the probability of loss intolerance (pLI), the loss-of-function observed/expected upper bound fraction (LOEUF) and missense Z scores (mis_z). A pLI value greater than or equal to 0.90 for the pLI is considered LoF intolerance, while a LOEUF value less than 0.6 indicates intolerance. Although they slightly differ in handling the uncertainty, both parameters are widely accepted in the evaluation of intolerance to LoF. A mis_z value greater than 3.09 suggests an intolerance to missense variation based on the deviation between expected and observed variants.

ClinGen Dosage Sensitivity Working Group provides curations for over than 1500 genes on their website (https://search.clinicalgenome.org/kb/gene-dosage). Their gene dosage curations follow an evidence-based workflow (Riggs et al., 2012). Haploinsufficiency data from these curations is integrated to this table.

Based on ClinVar submissions, mutation spectrums of genes are characterized by two metrics: LoF percent and missense percent. LoF percent is computed as the abundance of pLoF variants among pathogenic or likely pathogenic (P/LP) variants for genes with at least 10 P/LP variants. Missense percent is calculated as the percentage of pathogenic variants among all missense variants for a given gene. These two metrics are later assessed for BP1 and PP2 rules.

Lastly, this table contains gene-based frequency cutoffs based on population data from gnomAD. This cutoff is derived from the most common pathogenic variant in the gene. Along with this metric, the number of individuals homozygous for the

given variant for each gene is stored in the table. The use of the two metrics is explained in detail in Section 2.2.4.2.

2.2.2.3 ClinVar Submissions

clinvar.txt: ClinVar is weekly updated, and submissions can be downloaded as VCF or TSV files via their FTP servers (https://ftp.ncbi.nlm.nih.gov/pub/clinvar). Annotations in the INFO column in VCF file were extracted using regular expressions (regex) in R and organized into separate columns. VUS with multiple submissions were resolved if they had at least one conclusive classification (LP, P, LB, or B). The schema for the table can be found in Table 8.

The allelic origin of a variant is reported in ClinVar VCF files with numerical labels. This information was processed to annotate *de novo* variants in the database. Variants reported to be *in trans* with known pathogenic variants were also extracted by expression matching from submissions summary file.

Table 8: **Structure and description of clinvar.txt file.**

#	Column name	Description
1	chr	Chromosome name
2	pos	Variant position in chromosome
3	id	ClinVar Variation ID
4	ref	Reference allele
5	alt	Alternative allele
6	gene	Gene symbol
7	effect	HGVS expression (ex: NM_001005484:c.A44G:p.E15G)
8	tcpt	RefSeq transcript ID (ex: NM_001005484)
9	nt_change	Nucleotide change (ex: c.A44G)
10	nt_pos	Nucleotide change position in the transcript (ex: 44)
11	aa_change	Amino acid change (ex: p.E15G)
12	aa_pos	Amino acid change position in the protein (ex: 15)
13	is_splicing	Located in splice site or region (1 or 0)
14	is_missense	Results in amino acid change (1 or 0)
15	is_indel	Insertion/deletion (1 or 0)
16	is_denovo	Reported as *de novo* (1 or 0)
17	is_intrans	Detected as *in trans* with a pathogenic variant (1 or 0)
18	cls	ClinVar classification (P, LP, B, LB)
19	star	Review status (1-4)

2.2.3 Variant Annotation

Once the basic variant annotation from APIs is obtained, the algorithm annotates all variants with transcript information, functional domains, LoF metrics, gene-based frequency metrics, genomic context and ClinVar status from local databases described in the section 2.2.2. Additional annotations are called by the algorithm in a variant-specific manner.

2.2.3.1 Accessing transcript-specific information

While multiple endpoints are provided with Ensembl REST API for several types of annotations, retrieving all required information via APIs is costly. Thus, some information is accessed locally, minimizing the dependency on the web servers. Coding strand, transcript length, protein length, canonical status, transcript start and end positions, exon start and end positions, highest pathogenic variant frequency with the exon are fetched from local databases by matching the Ensembl transcript ID provided by Ensembl API.

2.2.3.2 Sourcing ClinVar classifications

In the AAVC, the ClinVar database (see Section 2.2.2.3) consists of over 1 million variants classified as B, LB, LP, or P. The algorithm first checks whether the variant under assessment is reported previously. Depending on the rule, it seeks out reported variants within the same gene, domain, exon, splice site, removed region or a specific range. The query is primarily made by genomic positions.

2.2.3.3 Retrieving genomic features

Understanding the genomic context of variation is made possible via multiple annotations in the AAVC. RepeatMasker and Uniprot databases supply the algorithm with annotations for repeat elements and functional domains, respectively (see Section

2.2.2.1). Pext scores and homopolymeric regions are accessed for null variants (see Sections 2.2.2.1 & 2.2.2.2).

2.2.4 Criteria Interpretation

Evaluation of certain criteria requires some advanced calculation, and this is achieved by either running scripts during processing or using pre-calculating values for the assessment. Next, deciding whether a variant is rare or common in a gene-specific manner involves a set of calculations. Finally, identifying alternative start codons in the transcript for start-loss variants and alternative stop codons for stop-loss variants requires sequence-level information.

2.2.4.1 Nonsense-mediated decay prediction

Nonsense-mediated decay (NMD) is a cellular pathway triggered by truncated variants and eliminates aberrant transcripts, resulting in no protein product. Based on functional studies, today we know mechanisms that can lead to NMD and predict it. Ensembl provides an NMD prediction algorithm (https://github.com/Ensembl/VEP_plugins/blob/release/111/NMD.pm) based on findings by Coban-Akdemir et al. (2018) and Lindeboom et al. (2016). Adopting the principles provided here, it was re-produced and embedded in the AAVC. All stop-gain and frameshift indels subjected to the algorithm are evaluated whether they satisfy one of the following features that leads to NMD escape:

i. Variant located within first 100 bp in the transcript
ii. Variant in a single-exon transcript
iii. Variant in last exon
iv. Variant in the last 55 bp of the penultimate exon

2.2.4.2 Gene-based frequency cut-offs

Many VCEPs adopted disease-based approaches for population data rules (Gelb et al., 2018; Kelly et al., 2018; Mester et al., 2018; Luo et al., 2019; Oza et al., 2018). Given the unavailability of established metrics such as prevalence and penetrance for most diseases, these approaches fail to be extended to all diseases. Alternatively, implementation of gene-based allele frequency cut-offs for accurate interpretation of population data was first time suggested by Kobayashi et al. (2017). Inspired by this study, AAVC takes a gene-based approach to evaluating population data similar to ClinGen VCEPs.

Genes (n=1954) with at least one two-star pathogenic (or likely pathogenic) variants from submissions with independent confirmation, in ClinVar were first determined. Then the most common pathogenic variant (mocopav) for each gene based on gnomAD v4 Exomes popmax values was identified. Rounding up these frequencies, a threshold of rarity is set for each gene. Similarly, the number of homozygotes for the mocopav is also determined for each gene. The processing was done in Python through gnomAD API. These values are accessible for the algorithm from the file gene_metrics.txt.

2.2.4.3 Text mining functional assays from ClinVar

The ACMG Guidelines consider *in vitro* or *in vivo* functional assay results as strong evidence for or against pathogenicity (PS3/BS3). While most clinical laboratories do not have access to wet lab experiments, reviewing the literature for functional studies performed for the variant under assessment might be time-consuming. Thus, a language model was trained for text classification to extract functional assays by their findings, whether disrupting or neutral, from ClinVar submission summaries. The trained model enables AAVC assigning PS3 or BS3 for variants reported on ClinVar.

Firstly, submissions with terms demonstrated in Table 9 were extracted from ClinVar. Then, approximately three thousand entries were manually evaluated to de-

termine whether the assays produced mutant or wild type phenotypes or inconclusive results. Then, a biomedical language model, BioBERT (biobert-large-cased-v1.1) (Lee et al., 2020), was employed to train a classification model on the labeled data. As a result, about 250,000 submissions that contained at least one of keywords in Table 9 were classified by the model with an accuracy of 97% for the test set after several steps of fine-tuning. A database of functionally studied variants was compiled using classified data.

Table 9: **Keywords used to extract submissions with reported functional assays**.

activity	drosophila	loxP	pombe	southern
affinity	elegans	mice	precipitation	transcription(al)
animal	enzym(e/atic)	minigene	ps3	transfect(ion)
assay	experiment	model	pull down	transgen(e/ic)
binding	FISH	morpholino	rerio	voltage
bs3	fluorescence	mouse	rescue	western
cell (culture/line)	functional	murine	RNA stud(y/ies)	wild
cerevisiae	histochemistry	musculus	RNAi	wt
cybrid	in vivo	mutagenesis	saccharomyces	yeast

In addition, multiplexed functional assays from the literature for x genes were manually curated and added into the dataset (Table 10). The final database consists of 25,800 functionally damaging and 21,857 functionally neutral alterations. The integration of the database into the classification is further detailed in Section 2.3.2.

Table 10: **List of genes with multiplexed assays of variant effect**. Approximately 30,000 functionally studied variants across 30 genes were manually compiled from the literature.

Gene(s)	PMID	Paper
ALPL	32160374	Del Angel et al., 2020
BRCA1	31907386	Kim et al., 2020
BRCA1	30209399	Findlay et al., 2018
BRCA1	32546644	Bouwman et al., 2020
BRCA1	23867111	Bouwman et al., 2013
BRCA1	18992264	Carvalho et al., 2009
BRCA1	17308087	Carvalho et al., 2007
BRCA1, BRCA2	29884841	Hart et al., 2019
BRCA2	33609447	Richardson et al., 2021
BRCA2	29394989	Guidugli et al., 2018
BRCA2	35736817	Hu et al., 2022
CHEK2	30851065	Delimitsou et al., 2019

CYP2C9	34314704	Amorosi et al., 2021
GALM	30910422	Iwasawa et al., 2019
GATM	27233232	DesRoches et al., 2016
HMBS	37729906	van Loggerenberg et al., 2023
HNF1A, HNF4A, HNF1B	33950347	Çubuk & Yalçın Çapan, 2021
IDS	30639582	Zhang et al., 2019
KCNQ4	35760561	Zheng et al., 2022
LDLR	35568682	Graça et al., 2022
LDLR	34167030	Alves et al., 2021
LDLR	29874871	Benito-Vicente et al., 2018
LDLR	23021490	Silva et al., 2012
LEPR, PCSK1, POMC	36864747	Shah et al., 2023
MLH1	17510385	Takahashi et al., 2007
MLH1, MSH2	19697156	Christensen et al., 2009
MSH2	33357406	Jia et al., 2021
MSH2	36550560	Scott et al., 2022
MSH6	34445333	Frederiksen et al., 2021
NPC1	35190686	Erwood et al., 2022
NUDT15	32094176	Suiter et al., 2020
OTC	37146589	Lo et al., 2023
PALB2	31757951	Boonen et al., 2019
PTEN	29706350	Mighell et al., 2018
PTEN	32350270	Post et al., 2020
RAD51C	36099300	Prakash et al., 2022
RAD51C	37253112	Hu et al., 2023
SCN1A	31782251	Brunklaus et al., 2020
SHH	32939873	Hong et al., 2020
UBE3A	34815418	Weston et al., 2021
UBE3A	33607653	Bossuyt et al., 2021
VKOR	32870157	Chiasson et al., 2020

2.3 Rule Implementation

AAVC offers an evidence-based variant classification methodology following the ACMG Standards and the ClinGen Recommendations. In addition, the ACGS Best Practices and the ClinGen VCEP Specifications are partially incorporated in the algorithm. With the aim of maintaining consistency, all evidence is sourced from well recognized large public databases, described Sections 2.2.1 & 2.2.2, by the scientific community. Furthermore, the custom modifications, discussed in Section 2.2.4, are provided with appropriate justification to improve the classification. Adhering to

these principles, AAVC automatically evaluates 21 of the 28 ACMG Criteria as followed: PVS1, PS1, PS2, PS3, PM1, PM2, PM3, PM4, PM5, PP2, PP3, PP5, BA1, BS1, BS2, BS3 BP1, BP3, BP4, BP6, BP7.

2.3.1 Population Frequency Rules

A variant commonly observed in populations is considered neutral, while a rare variant is under selective pressure and likely to be associated with a disease phenotype. The original guidelines consider rare variants pathogenic based on their absence or low frequency in population databases, but the definition of rare or common may be variable for genes. Thus, AAVC employs gene-based thresholds described in Section 2.2.4.2.

The frequency of the most common pathogenic variant (mocopav) is set as cut-off **BS1** (too common to be pathogenic for the disease) and an order of magnitude is used as cut-off for **PM2** (rare in population databases). The PM2 rule is applied at supporting level of evidence, as recommended by EPs. The number of homozygous individuals for the same variant is the cut-off for **BS2** (observed in healthy individuals). When a mocopav is not identified for the gene, 1% and 1.5% are used as thresholds for PM2 and BS1, respectively. If the gene was previously associated with an autosomal dominant or X-linked disease, the threshold is lowered to 0.5% for PM2. Finally, the original cut-off (5%) in the 2015 ACMG Standards continues to be used for **BA1** (too common to be pathogenic) rule.

For example, frequency rule cut-offs for insulin receptor (*INSR*) gene are shown in Table 11. *INSR*, whose protein is an integral component of the plasma membrane (Payankaulam et al., 2019), is assigned with a LOEUF constraint of 0.56, which indicates its essentialness and intolerance to LoF. Therefore, pathogenic variants for *INSR* are expected to be quite rare. The mocopav of *INSR*, INSR:c.1573C>T (rs1599937180), is a stop gain mutation in exon 7. Its grpmax frequency (maximum allele frequency observed in an ancestry group) in gnomAD v4.1 is 0.0000031. It is

rounded up to allow a degree of freedom and the value 0.000004 is set cut-off for BS1. Variant being interpreted with a frequency of greater than this value is assumed to be common, activating BS1. Next, an order of magnitude of this value is set cut-off for PM2. Therefore, variants rarer than 0.0000004 assigned with PM2. Finally, INSR:c.1573C>T was never observed in homozygous state in gnomAD v4.1, therefore the variant under assessment should be observed at least in two homozygous individuals to satisfy the BS2.

Table 11: **Thresholds for the frequency rules for *INSR* gene.**

Rule	Threshold	Incidence
PM2	≤ 0.00004%	≤ 1 in 2,500,000
BS1	> 0.0004%	> 1 in 250,000
BS2	≥ 2 individuals	≥ 2 individuals
BA1	≥ 5%	≥ 1 in 20

2.3.2 Reported Variant Rules

If the same variant or a similar variant was previously reported in reputable sources, the findings can be used as evidence for pathogenic or benign impact. In AAVC, the evidence for reported variant rules PS1, PS2, PS3, PM3, PM5, PP5, BS3, BP6 is obtained from ClinVar database.

PS1 criterion is applied if another nucleotide change in the same amino acid residue is reported pathogenic or likely pathogenic (P/LP), while **PM5** criterion is applied if a different change in the same amino acid residue is reported P/LP. As recommended by hearing loss and myeloid malignancy EPs (Luo et al., 2019; Oza et al., 2019), the PM5 is applied at strong level if more than one P/LP variant is reported for the same residue. The ClinVar database in AAVC also records *de novo* and *in trans* P/LP variants as described in Section 2.2.3.2, variants with these labels assigned with **PS2** and **PM3** criterion, respectively. **PS3** or **BS3** is assigned to variants with ClinVar reports whose functional studies show pathogenic or benign impact, respectively. This information is extracted via text classification using machine learning as described in Section 2.2.4.3. Reputable source criteria (**PP5/BP6**) are also applied according to the ClinVar reports. Considering the concerns regarding the re-

liability of ClinVar curations discussed in Section 1.6.2, the evidence level for these rules is adjusted based on the review status as shown in Table 12.

Table 12: **Evidence levels for PP5/BP6 criteria adjusted by ClinVar review status.**

Review Status	Gold Stars	Evidence Level
Practice guideline	★ ★ ★ ★	Very strong
Reviewed by expert panel	★ ★ ★ ☆	Very strong
Criteria provided, multiple submitters, no conflicts	★ ★ ☆ ☆	Strong
Criteria provided, single submitter or conflicting classifications	★ ☆ ☆ ☆	Moderate
No assertion criteria provided	☆ ☆ ☆ ☆	Supporting

2.3.3 *In silico* Prediction Rules

Following the PP3/BP4 rule specifications by ClinGen (Pejaver et al., 2022), BayesDel scores for missense variants are interpreted based on the recalibrated threshold ranges for strong, moderate, and supporting evidence levels as demonstrated in Figure 3.

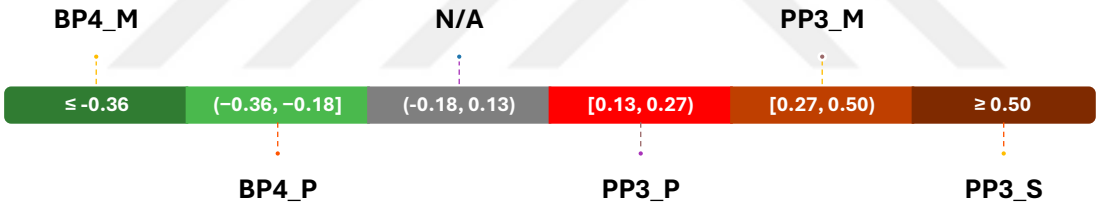


Figure 3: **Recalibrated thresholds for BayesDel predictions.** The thresholds provided by Pejaver et al. (2022) correspond to strong (S), moderate (M) and supporting (P) levels of evidence for predicted pathogenic (PP3) and benign impact (BP4) criteria, demonstrated as shades of green and red, respectively.

While ClinGen splice variant specifications (Walker et al., 2023) recommends the use of SpliceAI, there is no such recalibration based on levels of evidence. Adhering to the recommended moderate-level cut-offs, SpliceAI is used as *in silico* evidence in AAVC. The score ranges for SpliceAI are shown in Table 13.

Table 13: **The cut-offs for SpliceAI and corresponding evidence levels.** SpliceAI delta scores indicate the probability of a mis-splicing event. Delta scores are given between 0 and 1.

Evidence Level	Cut-off	Author description
BP3_Moderate	$\Delta < 0.10$	no predicted splice impact
PP3_Moderate	$\Delta \geq 0.20$	high sensitivity of splicing alterations

2.3.4 Protein Effect Rules

The ACMG Criteria employs several protein effect rules to assess the genomic context of the variation, which enables make inferences regarding the impact on the protein product. Of these, the **PVS1** rule (null variant in LoF gene) dramatically affects the classification and is highly susceptible misinterpretation as discussed in Section 1.6.3. Thus, AAVC follows the decision tree provided by ClinGen (Abou Tayoun et al., 2018). The provided decision tree evaluates stop-gain, frameshift, canonical splice site, start-loss, exonic deletion and duplication variants. While AAVC does not interpret exonic deletions and duplication events, it extends the framework to stop-loss variants based on the ACGS Best Practices (Durkie et al., 2024). Although splice variants in essential splice sites are considered for the PVS1 criterion, the algorithm followed for splice variants will be discussed in Section 2.3.5.

In the original guidelines, the PVS1 rule is designed for truncating variant in genes that do not tolerate haploinsufficiency. Thus, the first step of the PVS1 rule is to decide whether loss-of-function is a mechanism of disease for the gene. AAVC considers PVS1 for genes that meet at least one of these criteria below:

i. The ClinGen Haploinsufficiency (HI) score is 2 or 3
ii. gnomAD pLI score is at least 0.9 and HI score is not 40
iii. gnomAD LOEUF score is less than 0.6 and HI score is not 40
iv. PanelApp reported LoF as mode of pathogenicity for the gene

Once the LoF-intolerance for the gene is established, the variant is subjected to the PVS1 algorithm based on its effect. (It should be noted that LoF-intolerance check described above is disabled in AAVC by default. But users can easily activate it.) Stop-gain and frameshift variants are first checked for NMD (see Section 2.2.4.1 for NMD

prediction). If the transcript is predicted to be targeted for NMD, then the biological relevance is determined by regional pext scores (calculation of regional pext scores are described in Section 2.2.2.2). A high pext score leads to classification as PVS1_VS. When NMD is not expected, PVS1_S is applied if the removed region is functionally important (more than two-third of variants reported within the exon are expected to be P/LP). In case the functional relevance is unknown, a significant pext score is required to proceed further. If the LoF with highest frequency in the same gene reported in ClinVar is common (allele frequency ≥ 0.1), the percentage removed from the protein is calculated. A loss greater than 10% in the protein length classifies as PVS1_S, while a smaller loss classifies as PVS1_M. The decision tree for stop-gain and frameshift variants is summarized in Figure 4.

For start-loss variants, the first in-frame alternative start codon is identified. Next, the variants reported as P/LP in ClinVar in the removed region between the original and the rescue start codon is quantified. If at least one P/LP variant is reported, then PVS1_M is assigned, otherwise PVS1_P is assigned.

Similarly, the algorithm searches for in-frame stop codons within the 3' UTR for stop-loss variants. If there is no stop codon, then non-stop decay pathway is expected to be activated. Therefore, PVS1_VS is assigned. In case there is an alternative stop codon within the 3'UTR, only PM4 (change in protein length) is activated. The identification of alternative start and stop codons is detailed in Section 2.2.2.1.

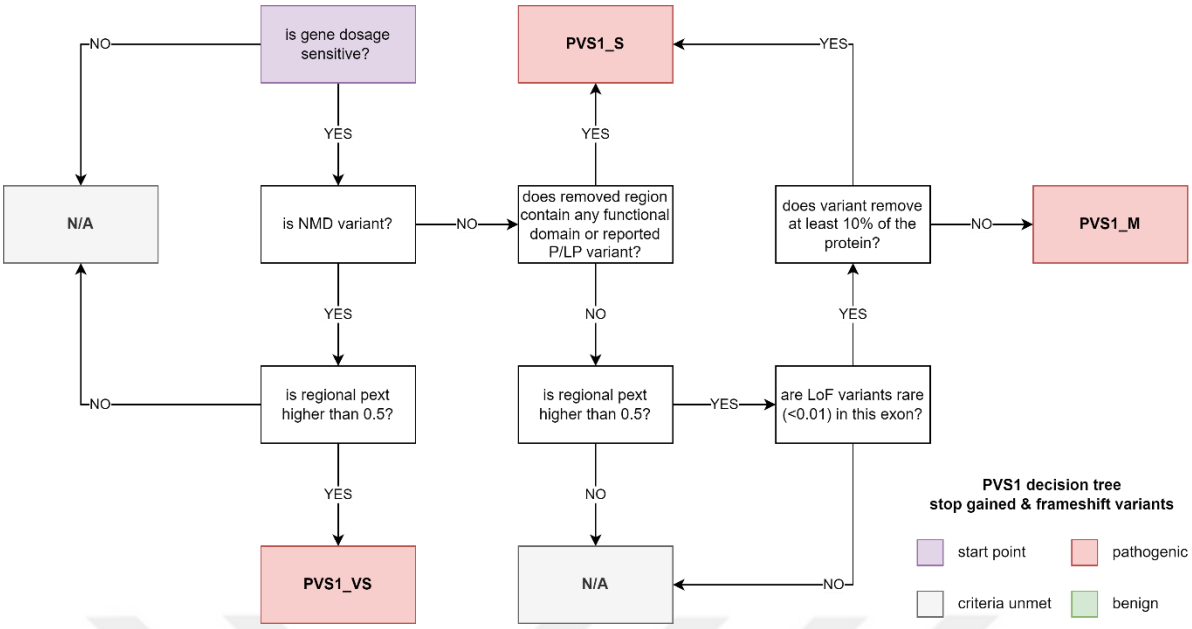


Figure 4: **PVS1 framework for stop gained and frameshift variants**. The decision tree is based on the ACMG Standards (Richards et al., 2015) and the ClinGen PVS1 Specifications (Abou Tayoun et al., 2021).

The **PM1** rule (mutational hotspot) is particularly considered for missense variants. To check whether the variant lies upon a functional domain, AAVC queries Pfam database. If it is found within a domain, which is absent from B/LB variants reported in ClinVar, then PM1 is applied. For variants that are out of a functional domain, the regional constraint is evaluated. If the variant is found within a constrained region and its pext is greater than 0.5, PM1 is assigned. Otherwise, the aggregation of pathogenic and benign reports in ClinVar for the exon is evaluated. If the exon contains at least two pathogenic variants and no benign variant reports, then PM1 is applied.

The **PM4** (change in protein length) and **BP3** (in-frame indels in repeat region) are evaluated together in AAVC. Stop-loss variants activate PM4. In-frame indels are checked with RepeatMasker database to see if they localize to a repeat element. In-frame indels in non-repeat regions activate PM4, while the ones in repeat regions are further assessed whether the repeat domain contains any P/LP variant reported in ClinVar. If the repeat is absent from P/LP variants, BP3_P is applied.

The **PP2** criterion addresses missense variants in genes where benign missense variations are uncommon. AAVC employs gnomAD v4 missense Z (mis_z) scores, which quantify the difference between observed and expected missense variants in a gene (Samocha et al., 2014; Lek et al., 2016), as recommended by ClinGen Standard Operating Procedure (ClinGen, 2022). A missense variant in a gene with mis_z score greater than 3.09, which suggests intolerance to missense variation, activates PP2. When the gene has a lower mis_z score than this threshold, the proportion of P/LP variants for the gene in ClinVar is determined. If 85% of variants in the gene are P/LP, then PP2 is applied.

Missense variants in LoF-intolerant genes satisfy the **BP1** rule. As described in the PVS1 section above, AAVC evaluates three metrics to classify a gene as LoF-intolerant: HI = 2 or 3, pLI ≥ 0.9 or LOEUF < 0.6. A missense variant in a gene that meets at least one of these criteria activates BP1.

Although the **BP7** rule is originally intended for synonymous variants with no splice impact, some VCEPs (Kelly et al., 2018; Mester et al., 2018; Luo et al.; 2019) apply this rule for intronic variants in non-conserved regions. Correspondingly, AAVC assesses phyloP100way scores to determine the conservation for synonymous, UTR, flanking region and intronic variants. A threshold of > 0.1 is used to apply BP7.

2.3.5 Splice Variant Interpretation

Interpretation of splice variants requires special consideration, but the original guidelines lack sufficient depth to address all aspects of these variations. Thus, ClinGen recently approved additional specifications for splice variant interpretation (Walker et al., 2023). AAVC follows these specifications tree to evaluate PVS1, PS1, PP3, BP4 and BP7 rules for canonical and non-canonical splice site variants.

Splice variant evaluation commences with the identification of the variant if it is reported in SpliceVault database. Both canonical and non-canonical variants are searched in the database. Walker et al. (2023) defines the canonical (or essential)

splice site as ±1/2 sites at the boundaries of an exon. The -1 and -2 dinucleotides located at the upstream exon are called splice acceptor sites, +1 and +2 dinucleotides located at the downstream exon are called splice donor sites. In addition, the positions from -3 to -1 and from +3 to +6 at the downstream exon are considered donor splice region and the positions from -20 to -3 and +1 positions at the upstream exon are called acceptor splice region. All splice associated regions are illustrated in Figure 5.

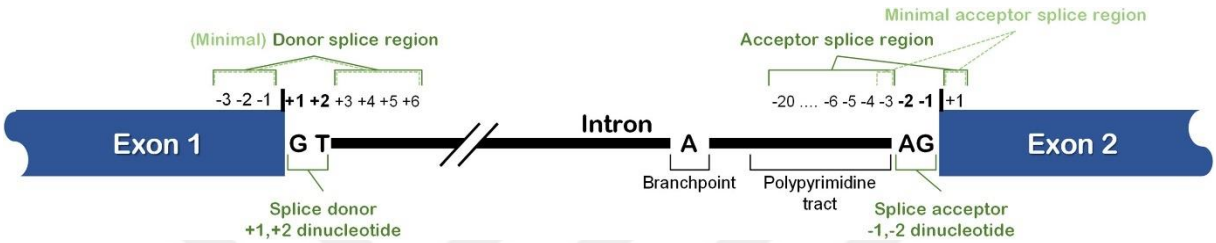


Figure 5: **The illustration of splice sites and splice regions**. Canonical donor and acceptor site refer to dinucleotides (±1,2) at the 5' and 3' ends of the intron, respectively. Donor splice region encompasses nucleotides between -3 to +6 in relative to exon-intron junction. Acceptor splice region encompasses nucleotides between -20 to +1 in relative to exon-intron junction. Polypyrimidine tract and branchpoint occur at the 3' end of the intron. Reprinted from Walker et al. (2023) with permission from Elsevier, Copyright @2024.

If variant is annotated, it is first evaluated whether the alteration results in a frameshift and the transcript is targeted for NMD. If they are, it is assigned with PVS1_VS. If there is a reported P/LP variant in the same acceptor/donor site or splice region, PS1_P is also applied. If NMD is not predicted or the variant does not result in a frameshift, the functional importance of the truncated region is assessed. PVS1_S is assigned is the removed region is functionally important. If it is not, and if LoF variants are uncommon in the transcript, the percentage lost due to truncation is calculated. A loss larger than %10 in the protein length assigns it PVS1_M. Variants that satisfied PVS1_S or PVS1_M are further assessed for PS1. If a P splice variant was previously reported at the same nucleotide or the same canonical splice site, PVS1_S assigned. If a P or LP variant was reported at the same splice region, PS1_M or PS1_P is applied, respectively. The canonical splice algorithm tree is shown in Figure 6.

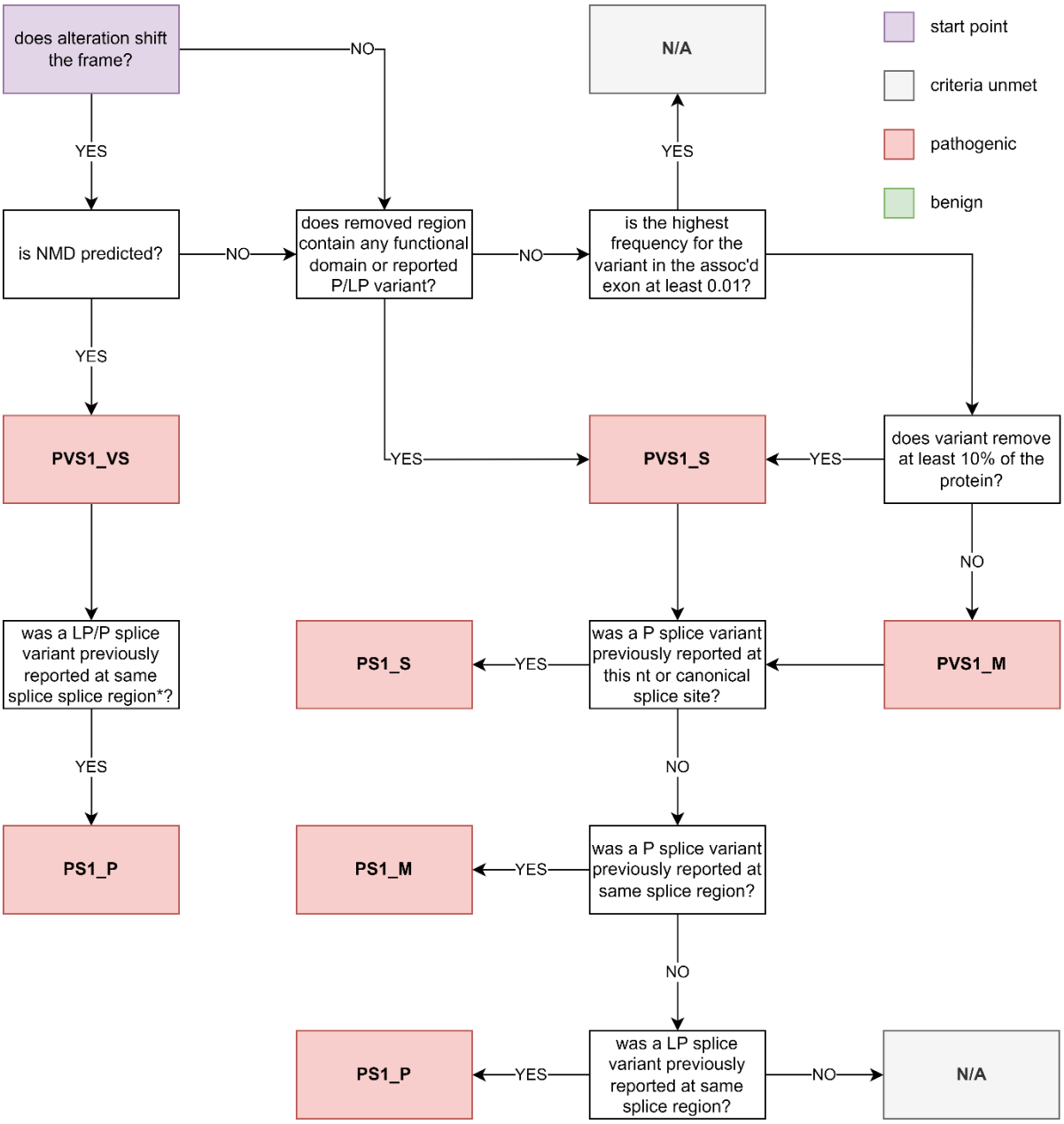


Figure 6: **The annotated splice variant algorithm tree**. The tree evaluates splice variants with reported consequences in SpliceVault for PVS1 and PS1 rules based on the ACMG Standards (Richards et al., 2015) and the ClinGen PVS1 Specifications (Walker et al., 2023).

To assess the impact of variants not reported in the SpliceVault splice variants, SpliceAI predictions are assessed. If the SpliceAI Δ score ≥ 0.2, PP3 (pathogenic *in silico* prediction) with corresponding level of evidence is assigned as described in Section 2.3.3. Then it is checked whether a P or LP variant was reported in ClinVar at the same nucleotide to assign PS1_S or PS1_M, respectively. If not, the same splice re-

gion is searched for a P or LP variants to assign PS1_M or PS1_P, respectively. In case the SpliceAI Δ score ≤ 0.1, which means no splice impact is predicted, BP4_M (benign *in silico* prediction) is assigned. If the same variant is synonymous or intronic in a non-conserved nucleotide (see Section 2.3.4 for conservation prediction), BP7 is applied as well. The algorithm tree followed by AAVC for non-canonical site splice variants is described in Figure 7.

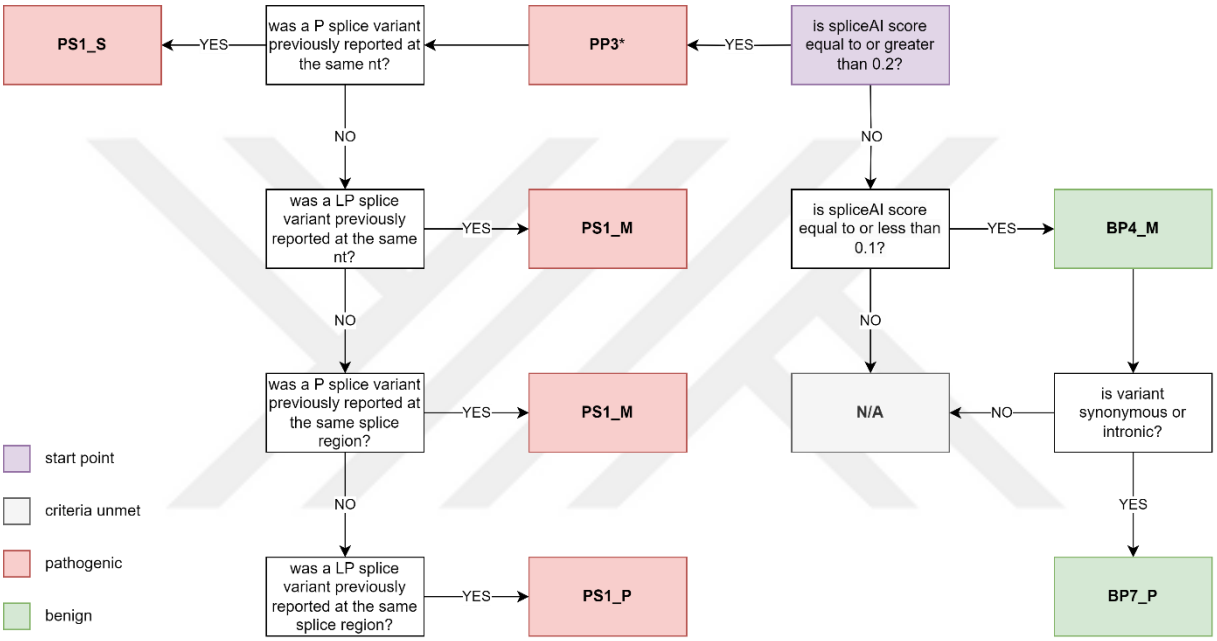


Figure 7: The unannotated splice variant algorithm tree. for PVS1 and PS1 rules based on the ACMG Standards (Richards et al., 2015) and the ClinGen PVS1 Specifications (Walker et al., 2023).

2.3.6 Final Classification

Once a variant is assessed for each criterion in AAVC, the levels of evidence are combined according to ClinGen's point-based scoring system to reach a final verdict. Each score is derived from Bayesian probabilities of pathogenicity (Tavtigian et al., 2020) and corresponds to a classification category as demonstrated in Table 14. Then combining scores for the evidence criteria, the variant is classified into one of five categories: benign (B), likely benign (LB), variant of unknown significance (VUS), likely pathogenic (LP) and pathogenic (P). Tavtigian et al. (2020) also calculated the score

ranges for these criteria (Table 15). AAVC particularly provides posterior probability of being pathogenic for each variant based on this model.

Table 14: **The levels of evidence and corresponding point scores**.
The scaling is adjusted based on Bayesian probabilities, as recommended by Tavtigian et al. (2020).

Evidence	Point scale	
Strength	**Pathogenic**	**Benign**
Supporting	1	-1
Moderate	2	-2
Strong	4	-4
Very Strong	8	-8

Table 15: **The score range for classification categories**.
The ranges and their posterior probabilities of pathogenicity were derived by Tavtigian et al. (2020).

Category	Score ranges	Probabilities
Benign	≤ -7	< 0.1%
Likely benign	[-1, -6]	0.1% to 10%
VUS	[0, 5]	10% to 90%
Likely pathogenic	[6, 9]	90% to 99%
Pathogenic	≥ 10	> 99%

AAVC further classifies VUS by exploiting the point-based classification system to reduce the rate of inconclusive results. Harrison & Biesecker (2021), on behalf of the ACMG, readily provided scores ranges compatible with Bayesian system for three VUS subcategories: VUS-high (4-5), VUS-middle (2-3), and VUS-low (0-1). Unpublished data from Baylor College of Medicine, Mass General Brigham and Quest Diagnostics suggests that VUS-high and VUS-low subclassifications prospectively align with LP and LB classifications, respectively (Figure 8) (Rehm, 2023).

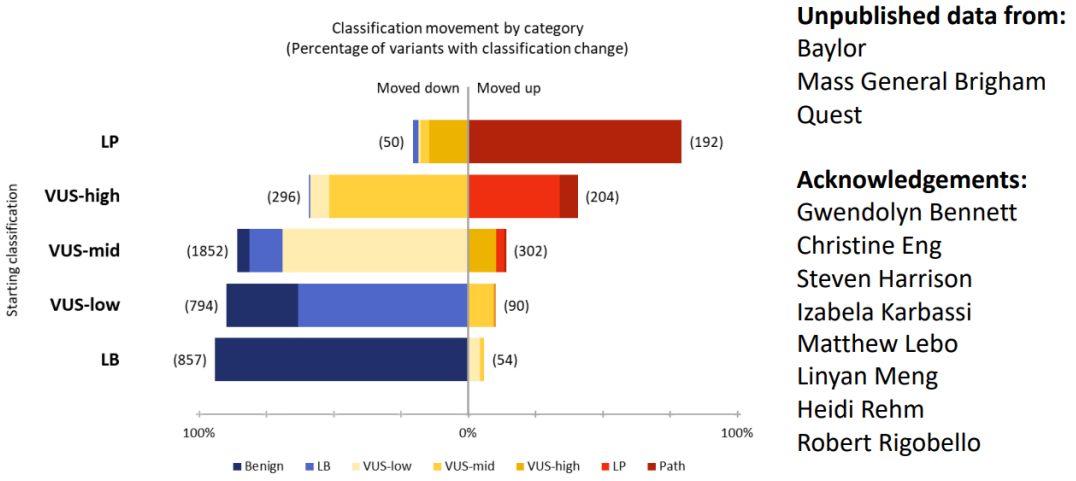


Figure 8: **The change in the classification of VUS by time**. The vertical axis shows the starting classifications, and the horizontal axis shows the final classifications. The figure illustrates the classification movements of likely pathogenic (LP), VUS-high, VUS-mid, VUS-low and likely benign (LB) after re-classification by additional data. Reprinted by permission from Rehm (2023).

Chapter 3

Results

3.1 Concordance with ClinGen Repository

ClinGen Evidence Repository (ERepo) (https://erepo.clinicalgenome.org/evrepo/) is an FDA-recognized collection of variant curations asserted by VCEPs. The database contains 7229 curated variants in 110 genes as of June 2024. The curations are based on instructions specified by 32 Expert Panels (EP) (https://cspec.genome.network/cspec/ui/svi/). It is a unique resource to assess the accuracy of AAVC classifications.

3.1.1 AAVC classifications highly correlate with ClinGen ERepo

After omitting structural alterations, retracted submissions and VUS, the remaining 4915 variants were classified by AAVC. The analysis reveals a concordance of 99.67% between AAVC and ERepo classifications for P/LP and B/LB variants (Table-16a). It increases to 99.88% by incorporating VUS-H and VUS-L into the calculations. The sensitivity (true pathogenic variants among all variants classified pathogenic) and specificity (true benign variants among all variants classified benign) are both 100%.

When the same set of variants is classified by Franklin, it demonstrates a concordance of 94.16% (Table 16-b). The performance for the classification of P/LP variants is highly similar (99.71% for AAVC, 99.30% for Franklin), but Franklin misses out 17.63% of the ERepo B/LB (n=1492) variants by classifying them either VUS (n=254) or LP (n=9). While AAVC does not misclassify any B/LB ERepo variants as P/LP, it classifies 0.40% (n=6) of ERepo B/LB variants either VUS-M (n=1) or VUS-L (n=5). These comparisons do not only imply that AAVC can access the same evidence lines as human curators, but also suggest that its classifications are almost completely concordant with ClinGen ERepo curations, performing better than Franklin.

Table 16: **The concordance of AAVC and Franklin for ClinGen ERepo curations**. AAVC **(a)** and Franklin **(b)** classifications are shown for variants pathogenic (P), likely pathogenic (LP), likely benign (LB) and benign (B) in ClinGen.

a.

AAVC CLASSIFICATIONS \ EREPO CLASSIFICATIONS	P	LP	LB	B	TOTAL
P	1835	1362	0	0	**3197**
LP	72	144	0	0	**216**
VUS-H	1	4	0	0	**5**
VUS-M	0	2	1	0	**3**
VUS-L	1	2	4	1	**8**
LB	0	0	83	40	**123**
B	0	0	616	747	**1363**
TOTAL	**1909**	**1514**	**704**	**788**	**4915**

b.

FRANKLIN CLASSIFICATIONS \ EREPO CLASSIFICATIONS	P	LP	LB	B	TOTAL
P	1747	1178	0	0	**2925**
LP	154	320	7	2	**483**
VUS	8	15	188	66	**277**
LB	0	0	346	110	**456**
B	0	1	163	610	**774**
TOTAL	**1909**	**1514**	**704**	**788**	**4915**

3.2 Secondary findings in the Turkish Variome

The ACMG periodically publishes recommendations on reporting variants in actionable genes associated with cancer, cardiovascular diseases, or inborn errors of metabolism phenotypes since 2013 (Green et al., 2013). This secondary finding (SF) list has been updated in 2015 (ACMG, 2015), 2021 (Miller et al., 2021) and 2023 (Miller et al. 2023). The current version consists of 81 genes (Table 17). The ACMG encourages labs to screen these genes to help us understand better underlying disease mechanisms, develop effective management strategies, and discover novel therapeutic targets.

Table 17: ACMG Secondary Findings v3.2 list for reporting actionable genes.

Gene	Disease name
ACTA2	Aortic aneurysm
ACTC1	Familial hypertrophic cardiomyopathy
ACVRL1	Hereditary hemorrhagic telangiectasia type
APC	Adenomatous polyposis coli
APOB	Familial hypercholesterolemia
ATP7B	Wilson disease
BAG3	Dilated cardiomyopathy, Myofibrillar myopathy
BMPR1A	Juvenile polyposis syndrome
BRCA1	Breast-ovarian cancer
BRCA2	Breast-ovarian cancer
BTD	Biotinidase deficiency
CACNA1S	Malignant hyperthermia
CALM1	Catecholaminergic polymorphic ventricular tachycardia, Long QT syndrome
CALM2	Long QT syndrome
CALM3	Long QT syndrome
CASQ2	Catecholaminergic polymorphic ventricular tachycardia
COL3A1	Ehlers-Danlos syndrome
DES	Dilated cardiomyopathy, Myofibrillar myopathy
DSC2	Arrhythmogenic right ventricular cardiomyopathy
DSG2	Arrhythmogenic right ventricular cardiomyopathy
DSP	Arrhythmogenic right ventricular cardiomyopathy, Dilated cardiomyopathy
ENG	Hereditary hemorrhagic telangiectasia type
FBN1	Marfan's syndrome
FLNC	Dilated cardiomyopathy, Myofibrillar myopathy
GAA	Pompe disease
GLA	Fabry's disease
HFE	Hereditary hemochromatosis
HNF1A	Maturity-Onset of Diabetes of the Young
KCNH2	Long QT syndrome
KCNQ1	Long QT syndrome
LDLR	Familial hypercholesterolemia
LMNA	Dilated cardiomyopathy

MAX	Pheochromocytoma
MEN1	Multiple endocrine neoplasia
MLH1	Lynch syndrome
MSH2	Lynch syndrome
MSH6	Lynch syndrome
MUTYH	MYH-associated polyposis
MYBPC3	Familial hypertrophic cardiomyopathy
MYH11	Aortic aneurysm
MYH7	Dilated cardiomyopathy, Familial hypertrophic cardiomyopathy
MYL2	Familial hypertrophic cardiomyopathy
MYL3	Familial hypertrophic cardiomyopathy
NF2	Neurofibromatosis
OTC	Ornithine carbamoyltransferase deficiency
PALB2	Hereditary breast cancer
PCSK9	Hypercholesterolemia, autosomal dominant
PKP2	Arrhythmogenic right ventricular cardiomyopathy
PMS2	Lynch syndrome
PRKAG2	Familial hypertrophic cardiomyopathy
PTEN	PTEN hamartoma tumor syndrome
RB1	Retinoblastoma
RBM20	Dilated cardiomyopathy
RET	Familial medullary thyroid carcinoma, Multiple endocrine neoplasia
RPE65	RPE65-related retinopathy
RYR1	Malignant hyperthermia
RYR2	Catecholaminergic polymorphic ventricular tachycardia
SCN5A	Brugada syndrome, Dilated cardiomyopathy, Long QT syndrome
SDHAF2	Paragangliomas
SDHB	Hereditary paraganglioma-pheochromocytoma syndrome
SDHC	Paragangliomas
SDHD	Hereditary paraganglioma-pheochromocytoma syndrome
SMAD3	Loeys-Dietz syndrome
SMAD4	Juvenile polyposis syndrome
STK11	Peutz-Jeghers syndrome
TGFBR1	Loeys-Dietz syndrome
TGFBR2	Loeys-Dietz syndrome
TMEM127	Pheochromocytoma
TMEM43	Arrhythmogenic right ventricular cardiomyopathy
TNNC1	Dilated cardiomyopathy
TNNI3	Familial hypertrophic cardiomyopathy
TNNT2	Dilated cardiomyopathy, Familial hypertrophic cardiomyopathy
TP53	Li-Fraumeni syndrome
TPM1	Familial hypertrophic cardiomyopathy
TRDN	Catecholaminergic polymorphic ventricular tachycardia, Long QT syndrome
TSC1	Tuberous sclerosis
TSC2	Tuberous sclerosis
TTN	Dilated cardiomyopathy
TTR	Hereditary transthyretin-related amyloidosis
VHL	Von Hippel-Lindau syndrome
WT1	Wilms' tumor

Hosting numerous civilizations throughout the centuries in the Anatolian Peninsula, the Turkish Variome provides a unique opportunity to study rare and complex diseases (Kars et al., 2021). Responding to ACMG's call to report suspected pathogenic variants in these genes, the Turkish Variome can provide valuable insights into the pathogenic burden of those genes in Turkish population and enable us to unearth novel causal variants. The Turkish Variome consists of exome and genome sequencing data of 3824 samples. The cohort includes 935 obese individuals (BMI ≥30 for adults or ≥95th percentile for children). Mean age of the individuals with phenotype data (n=1161) is 38.8±14.5 and female-to-male ratio is 2.38.

3.2.1 Pathogenic variants activate an average of 4.25 criteria

All variants in 81 genes from ACMG SF v3.2 list (Table 10) in the Turkish Variome were assessed for their pathogenicity by AAVC. Of 11,337 variants, 9346 variants (82.4%) map to coding sequences, the remaining 1991 (17.6%) variants are found within non-coding regions. The distribution of the variants by their major consequences is demonstrated in Figure 9.

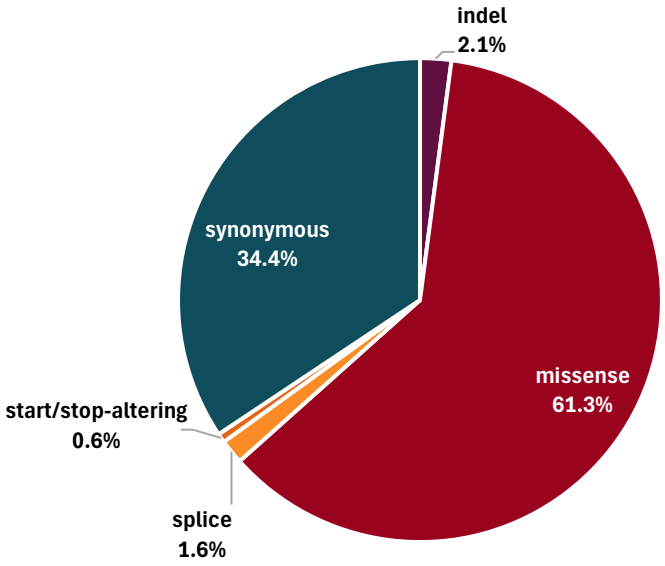


Figure 9: The distribution of the SF gene variants in the Turkish Variome. Of 11,337 variants in the 81 ACMG SF v3.2 genes in the Turkish Variome, there are 5826 missense variants, 3264 synonymous variants, 1835 non-coding variants, 198 indels, 156 splice variants and 58 start/stop-altering variants across were analyzed.

AAVC activated a combination of 32,723 criteria for 11,337 variants, with an average of 2.89 criteria activated per variant. While approximately 4.27 pathogenic criteria were activated to classify a variant P/LP, 2.49 benign criteria were activated for B/LB variants. VUS-H, VUS-M and VUS-L groups had an average of 1.58 pathogenic and 1.17 benign criteria activated per variant. Figure 10 shows the number of occurrences for each criterion. When a variant meets PM3 (observed *in trans* with a pathogenic variant), it was classified pathogenic with a chance of 87.3%. PVS1 (67.9%), PS1 (66.7%), PS3 (63.4%) and PP5 (75.4%) were other criteria that mostly assigned to pathogenic or likely pathogenic variants.

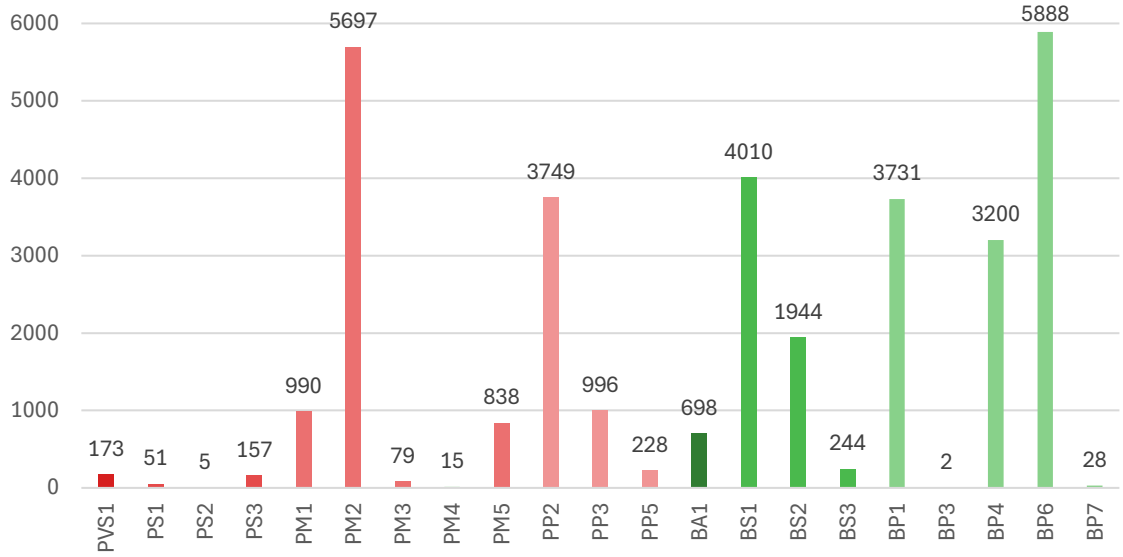
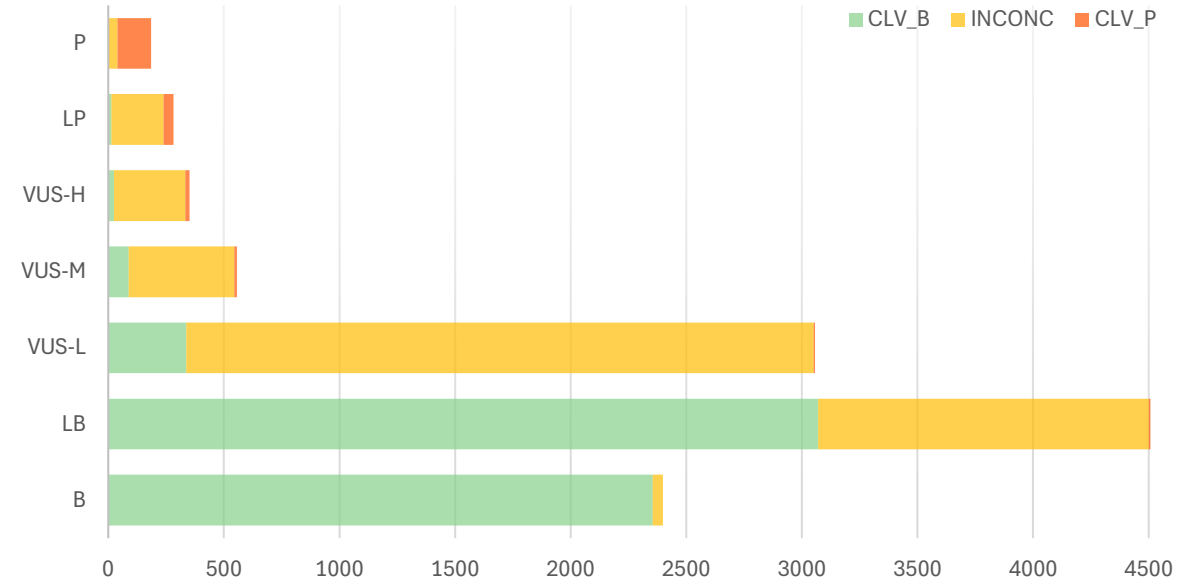


Figure 10: **Assigned ACMG criteria for variants in the SF genes**. AAVC assigned 32,723 criteria for 11,337 variants. Pathogenic and benign criteria are demonstrated in shades of red and green, respectively.

3.2.2 AAVC discovers novel pathogenic variants in the SF genes

The algorithm classified 818 (7.2%) variants P/LP/VUS-H, 9963 (87.9%) variants B/LB/VUS-L and 556 (4.9%) variants VUS-M. Whereas ClinVar classified 228 (2.0%) variants P/LP/VUS-P and 5888 (51.9%) variants B/LB/VUS-B. AAVC improved the resolution of pathogenic variants by 3.59-fold and benign variants by 1.69-fold. The results are summarized in Figure 11.

a.



b.

	CLV_B	INCONC	CLV_P	TOTAL	PERC
AAVC P	0	40	145	185	1.63%
AAVC LP	13	226	43	282	2.49%
AAVC VUS-H	24	310	17	351	3.10%
AAVC VUS-M	89	456	11	556	4.90%
AAVC VUS-L	337	2714	6	3057	26.96%
AAVC LB	3070	1431	6	4507	39.75%
AAVC B	2355	44	0	2399	21.16%
	228 (2.0%)	5221 (46.1%)	5888 (51.9%)	11337	

Figure 11: **The concordance between ClinVar and AAVC classifications**. **(a)** The concordance of ClinVar classifications with AAVC shown in the clustered bar chart. The AAVC classifications are in vertical axis. ClinVar B/LB/VUS-B and P/LP/VUS-P are highlighted with green and red in the bar chart. Inconclusive variants (conflicting/inconclusive/unreported in ClinVar) are shown in orange. **(b)** The table shows the analysis of interpretation of 11,337 actionable gene variants in the Turkish Variome. The algorithm classified all variants into seven categories: pathogenic (P) (1.63%), likely pathogenic (LP) (2.49%), VUS-high (VUS-H) (3.10%), VUS-medium (VUS-M) (4.90%), VUS-low (VUS-L) (26.96%), likely benign (LB) (39.75%) and benign (B) (21.16%).

Of 5221 unreported/inconclusive/conflicting variants in ClinVar, AAVC could resolve 266 (5.1%) variants as P/LP and 1475 (28.3%) variants as B/LB. This recommends that one in every three novel or uncertain variants can be resolved by AAVC. Considering VUS-high and VUS-low classifications, the resolution power rises to 91.3%. The classification of 19 variants were not concordant between ClinVar and AAVC. 13 ClinVar B/LB/VUS-B variants were classified P/LP, while 6 ClinVar P/LP were classified B/LB by AAVC. All of these variants were of low-confidence (12 zero-

star entries and 1 one-star entry) in ClinVar since the assertion criteria was not either provided or not confirmed by an independent study.

AAVC revealed 279 novel pathogenic or likely pathogenic variants, which were either misclassified, controversial or unreported in ClinVar, across 57 SF genes in the Turkish Variome. These variants have an aggregated allele frequency of 3.9% based on gnomAD v4.1 popmax, indicating that they are present in a considerable proportion of the population.

3.2.3 BayesDel fails to reliably classify missense variants alone

To see how reliable *in silico* predictions alone are in comparison to evidence-driven classification, 5681 missense variants in the SF genes were analyzed by BayesDel and AAVC. Using cut-offs (≥0.13 for pathogenicity, ≤-0.18 for benignity) defined in Pejaver et al. (2022), 849 and 3117 variants were classified as pathogenic and benign by BayesDel, respectively. Showing a concordance of 72.2% (2450/3392) for benign variants with AAVC, the predictive value of BayesDel was somewhat convincing. While 213 (25.1%) of 849 BayesDel pathogenic variants were classified P/LP, 217 (25.6%) of these variants were B/LB by AAVC. These figures recommend using BayesDel alone is highly misleading for finding pathogenic variants, with about a fourth of predictions being false positives. The comparison is shown in Figure 12.

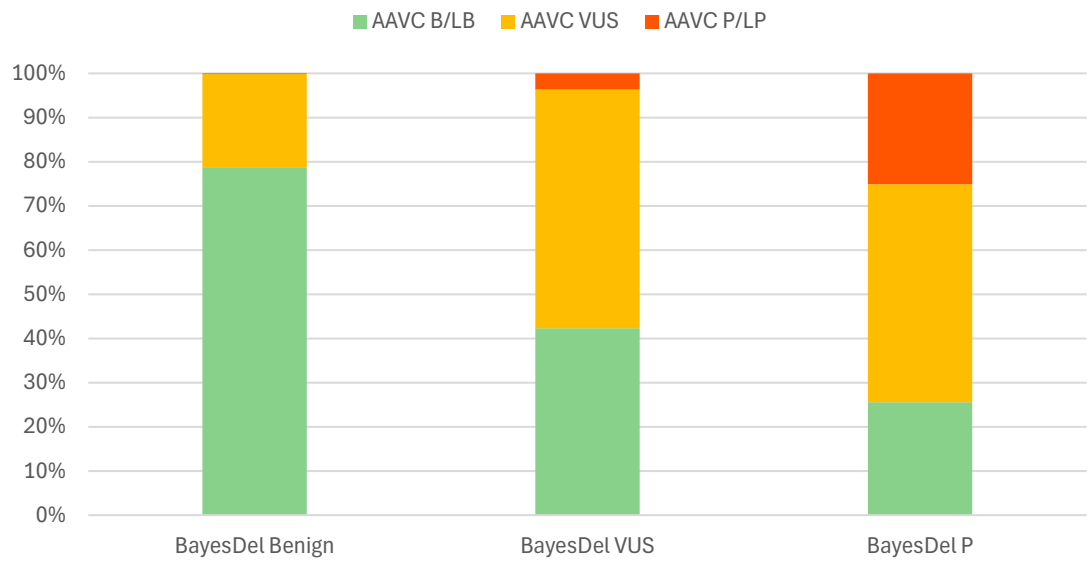


Figure 12: **The concordance between BayesDel and AAVC classifications**. Benign (B), VUS and pathogenic (P) classifications by BayesDel are shown as three separate bars in the horizontal axis. Colors indicate the AAVC classifications.

3.3 Reclassification of ClinVar variants by AAVC

ClinVar is the most comprehensive variant curation database, containing over 2.8 million variants, although its credibility is heavily criticized. In order to investigate the concordance with ClinVar, all variants in ClinVar were classified by the algorithm.

3.3.1 AAVC can resolve unclassified variants in ClinVar

The comparison showed a concordance of 95.85% between ClinVar and AAVC classifications for P/LP and B/LB groups. VUS-H and VUS-L curations increased it to 98.02%. While AAVC confirmed (P/LP/VUS-H) the pathogenicity of 89.92% of ClinVar P/LP variants, 0.16% of them were classified B/LB and 8.28% were VUS-M. On the other hand, B/LB assertions in ClinVar were highly accurate with a concordance of 99.31% (B/LB/VUS-L) with AAVC. Reclassifying the VUS variants in ClinVar by AAVC resolves 53.40% of variants into P/LP (4.02%) or B/LB (49.38%). When VUS-H and VUS-L subclassifications are incorporated, 90.09% of variants are resolved. The Sankey graph in Figure 13 depicts the differences between classifications.

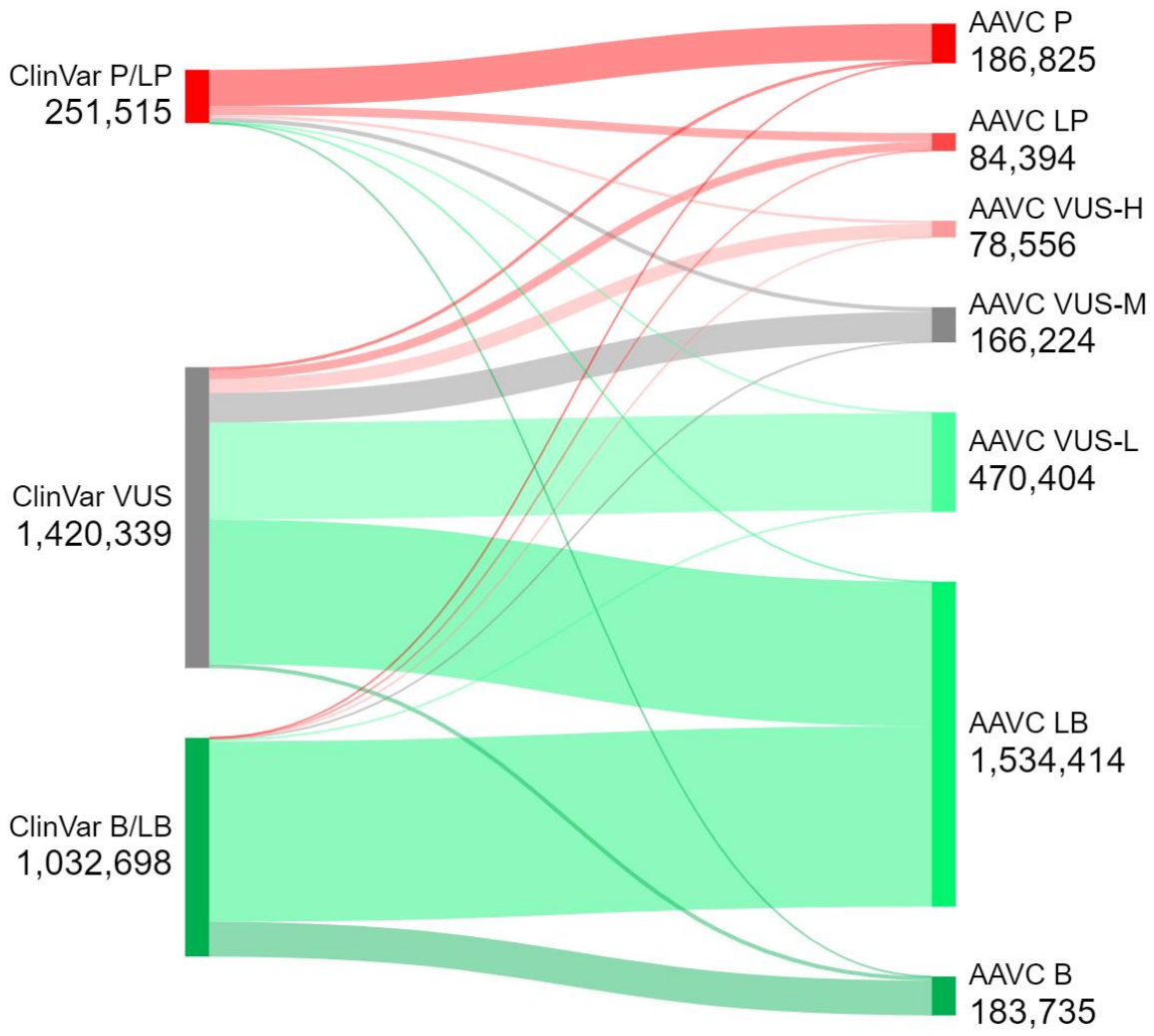


Figure 13: The re-classification of ClinVar variants by AAVC. AAVC confirms 89.92% and 99.31% of the P/LP and B/LB, respectively, assertions in ClinVar. The tool resolves 90.09% of the all ClinVar variants into P, LP, VUS-H, VUS-L, LB, or B.

When only entries with multiple submissions (two-star and above) were evaluated, the concordance was found to be 97.55% for P/LP and B/LB. Considering the VUS subclassifications, it increased to 99.83%. Variants with at least two gold stars in ClinVar are excessively used in comparison as a truth set due to their reliability. Given AAVC shows a nearly full concordance with ClinGen ERepo variants, the incorporation of VUS-H and VUS-L to the decision-making process can be reliable.

Chapter 4

Discussion

4.1 Conclusions

4.1.1 Motivation of the work

With the mainstream availability of sequencing technologies and rapid accumulation of sequence data in the genomic era, hundreds of thousands of novel variants have begun to being revealed. While discovering a novel variant was something big in the beginning of the era, today it is simply part of the standard process in genetic research. Currently, we understand that most of the variation in the genome is actually neutral, but tweezing the causal variants still remains as a challenge. Various *in silico* prediction tools were developed to sort out the pathogenic variants. However, they fell short against the complexity and context-dependency of the human genome. Thus, human geneticists came up with standardized frameworks to confidently make use of the sequence data for diagnosis, management, and treatment of diseases in the clinics. But these set of guidelines were not without problems. The frameworks required a substantial amount of time, knowledge, and experience to put into practice. Moreover, controversial assertions for the same variant by different human curators

following the same guidelines led to disagreements. These problems have paved the way for the birth of automated variant classification tools.

4.1.2 Addressing the needs

AAVC is provided as a standalone software that can run in Python environments. Aggregating data from ClinVar, Ensembl, gnomAD, PanelApp along with other online and local databases, the algorithm assess pathogenicity based on the ACMG Criteria and the ClinGen Specifications. It comes with utilities to keep local databases up to date. Providing the necessary annotations with tools like ANNOVAR, the algorithm can completely run offline. The input can be provided as a text file containing variant identifiers and the program produces a tab-separated output in return. Alternatively, it can work with VCF files. AAVC does not store or share any information with third parties. Therefore, clinical laboratories and research groups can utilize the program without any concern for data privacy. Installation and use of the tool requires basic understanding of computer science. Following the instructions, anyone with basic knowledge in bioinformatics can run the algorithm.

4.1.3 Impact and implications

Today, many bioinformatics companies provide their variant interpretation tools. However, their variant curation processes lack justifiable methodologies in peer-reviewed journals and prevent people from understanding, analyzing, and integrating the framework into their work. Their consistency and accuracy are questionable at this point. Another point of concern regarding these variant classification platforms is data privacy. They strictly run on web platforms. Moreover, these tools either fail to follow the latest updates to the guidelines or provide the necessary environment for large scale analysis. AAVC offers a highly accurate, rapid and up-to-date framework for clinical laboratories and research groups to automatically interpret sequence variants. AAVC shows almost full concordance (99.67%) with FDA-approved variant

classifications in ClinGen ERepo. It enables revealing novel pathogenic variants in actionable genes. It resolves a significant proportion of VUS classifications in ClinVar, reclassifying at least 57,000 inconclusive variants as P/LP.

4.2 Limitations

AAVC is primarily designed to classify single nucleotide variations and short indels, whereas it does not interpret copy number variations and long indels. Although the framework can interpret non-coding variants, our current understanding of non-coding genome is quite limited, therefore such variants will likely to be classified VUS. Next, AAVC is developed in Python and requires internet connection for every-day use, which might not be very feasible for large-scale analysis. For such use, pre-annotated VCF files can be used to interpret variants, boosting the processing speed. However, such preparation will require familiarity with annotation tools such as ANNOVAR and VEP. Finally, the current version of the ACMG Guidelines still possesses problems and additional specifications are required to guide clinical geneticists. It is expected new guidelines to address them.

4.2.1 Inaccessibility of the data

One of the greatest challenges of the genomic age is variants of uncertain significance (VUS). These variants fulfill neither the criteria to be pathogenic nor the criteria to be benign. Our analysis has shown that VUS in the 81 ACMG SF genes in the Turkish Variome has an average of 1.58 pathogenic and 1.17 benign criteria met, which was lower than P/LP (4.27) and B/LB (2.49). It recommends that lack of data is the reason why these variants could not be classified into a group. This corresponds to approximately 35% of classifications. At this point, data sharing policies and accession strategies need to be improved to address the VUS challenge.

ClinVar has achieved remarkable success in documenting human genetic variation. ClinVar submitters report whether a variant is *de novo*, observed *in trans* with a pathogenic variant, or functionally assessed *in vitro* along with their assertions. All these information pieces dramatically affect the classification. AAVC implements a BioBERT-derived text-mining algorithm to extract this information from the ClinVar submissions. However, the choice of submitting a variant to ClinVar is optional and many more variants in the literature are lacking their assertions in a variant database. Thus, the missing information needs to be extracted from the literature. Although there are currently available services such as LitVar2 (Allot et al., 2023) that screens the literature, a human curator should still review the retrieved information. Therefore, it is not completely automated. Nevertheless, such automation is not far away, with the power of biomedical large language models in development such as Meditron-70b (Chen et al., 2023).

4.2.2 Programmatic constraints

AAVC is designed to be used with minimum user involvement and no prior experience in bioinformatics. It collects the most critical information from the web services of Ensembl and gnomAD. These services are intended to give users access to databases from anywhere in the world. An extraordinary load on their servers will prevent people from accessing them. Such excessive queries can also affect the integrity of databases. Thus, they apply a quota for accessions through their APIs to prevent it. AAVC is unlikely to exceed the quotas, because the processing rate for batch queries is limited to twenty-five variants per query submitted to Ensembl and gnomAD. The algorithm counts the inputs and splits them into batches. This gives necessary time to avoid the quotas without causing any load on the servers.

These constraints can be quite limiting for large-scale analyses. The VCF mode in AAVC, which totally runs offline, can overcome this limitation by accelerating the processing. But this mode requires a pre-annotated VCF file with VEP, gnomAD, REVEL, SpliceAI and phyloP. While VEP has its own command-line tool, the remain-

ing annotation can be performed through ANNOVAR, which enables annotation with user-generated tables. Running on a powerful processor fed with 64 GB of RAM, the algorithm can process about twelve variants in a second in offline mode.

4.2.3 Shortcomings of the guidelines

The original ACMG Guidelines have been notably improved by the ClinGen Specifications. The main problem with the initial version was ambiguities in the definitions and lack of in-depth explanations of the rules. ClinGen's efforts addressed these limitations with their operating procedures. However, the current state of the guidelines still has many unclear points that can lead to confusion. For example, NMD is the primary mechanism that leads to truncation of transcript. But neither the ACMG Guidelines nor the ClinGen Specifications provide clear instructions to evaluate whether a variant is likely to escape NMD. Next, the evaluation of splice variants is still challenging. Walker et al. (2023) extended the original guidelines and provided a detailed decision tree, but there are no available computational tools to infer the consequence of a splice variant whether it results in a frameshift, leads to multi-exon skipping or cause pseudoexonization. Canson et al. (2023) developed a SpliceAI-based framework that predicts splice consequences, but its technical impracticality makes it non-feasible to implement. Moreover, the ACMG framework still classifies a large fraction of missense variants VUS due to limited evidence or criteria for the evaluation of missense alterations. One approach to dealing with VUS is to use VUS-High and VUS-Low classification, as discussed in Section 2.3.6. A retrospective analysis from an unpublished study suggests that VUS-H and VUS-L align with LP and LB, respectively, by time. While the new guidelines under development are expected to address most of the limitations of the 2015 Guidelines.

4.3 Future Perspectives

4.3.1 Interpretation of the non-coding and structural variants

Although most clinical sequencing practices target coding regions of the genome, genome-wide association studies (GWAS) increasingly reveal signals that map to non-coding regions (Duan et al., 2014; Madelaine et al., 2018). Although Ellingford et al. (2022) provided some considerations for evaluation of non-coding variants, our knowledge is still quite limited to follow them. In this way, understanding the regulatory code is essential. Based on polymorphism data, Aggarwala & Voight (2016) could explain at least 81% of variation from heptanucleotide context. Following a similar methodology, the constrained regions in the non-coding genome were defined by di Iulio et al. (2018). More recently, gnomAD introduced the genomic non-coding constraint map using population data (Chen et al., 2024). In addition, chromatin marks can ideally help us distinguish critical regions in the non-coding genome, since epigenetically accessible regions are mostly regulatory elements such as enhancers and promoters (Klemm et al., 2019). With the availability of epigenomic data and these constraint metrics, the interpretation of non-coding variants can be improved.

AAVC is currently limited to SNPs and short indels, while it does not interpret large indels, inversions, copy number variation, duplication, mobile element insertions, short tandem repeats or translocation events. However, the detection of structural variation (SV) by short-read sequencing methods is technically problematic. Sanchis-Juan et al. (2023) demonstrates that complementation of short-read sequencing with long-read sequencing increases diagnostic yield by capturing SV. GATK-SV is an ensemble pipeline and enables calling SV from sequencing data (Collins et al., 2020). gnomAD reports their short-read SV data generated by this approach is almost as reliable as long-read data with a precision of 96.9%. Given the standards for the interpretation of SV was already outlined by ClinGen (Riggs et al., 2020), the incorporation of these guidelines with the availability of high-quality SV data to AAVC would reveal more causal variants.

4.3.2 Incorporation of cohort-based evidence

One of the evidence lines in the ACMG Guidelines is cohort-based evidence that involves co-segregation (PP1) or lack of segregation (BS4) of variants in related individuals with disease phenotype. Biesecker et al. (2024) recently published a quantitative framework to assess co-segregation rules PP1 and BS4. Providing family information to AAVC as input, it is possible evaluate whether variant co-segregates with the phenotype based on this framework. Based on the availability of a control group, case-control rule (PS4) can also be evaluated for the disease phenotype. This will involve comparing the enrichment in case and control groups, calculating odds ratio for the variant and determining the standard error. Again, this would require an additional input that assigned with individuals as case or control. These additional rules can potentially take the interpretation one step further, providing more insight into the clinical relevance of uncertain variants.

4.3.3 Updated standards for sequence variant interpretation

The ACMG Sequence Variant Classification working group announced that an update to the ACMG Guidelines is under review. The draft standards were introduced to the clinical genetics community during the ACMG Annual Meeting in March 2024. The final version of the guidelines is expected to arrive this year. While the new guidelines do not introduce new lines of evidence, it primarily aims at addressing the weaknesses of the current guidelines by removing ambiguities and providing quantitative parameters.

Bibliography

1000 Genomes Project Consortium, Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E., & McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. *Nature, 467*(7319), 1061–1073.

Abou Tayoun, A. N., Pesaran, T., DiStefano, M. T., Oza, A., Rehm, H. L., Biesecker, L. G., Harrison, S. M., & ClinGen Sequence Variant Interpretation Working Group (ClinGen SVI) (2018). Recommendations for interpreting the loss of function PVS1 ACMG/AMP variant criterion. *Human Mutation, 39*(11), 1517–1524.

ACMG Board of Directors (2015). ACMG policy statement: updated recommendations regarding analysis and reporting of secondary findings in clinical genome-scale sequencing. *Genetics in Medicine, 17*(1), 68–69.

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., & Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods, 7*(4), 248–249.

Aggarwala, V., & Voight, B. F. (2016). An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. *Nature Genetics, 48*(4), 349–355.

Akbari, P., Gilani, A., Sosina, O., Kosmicki, J. A., Khrimian, L., Fang, Y. Y., Persaud, T., Garcia, V., Sun, D., Li, A., Mbatchou, J., Locke, A. E., Benner, C., Verweij, N., Lin, N., Hossain, S., Agostinucci, K., Pascale, J. V., Dirice, E., Dunn, M., … Lotta, L. A. (2021). Sequencing of 640,000 exomes identifies GPR75 variants associated with protection from obesity. *Science, 373*(6550), eabf8683.

Allot, A., Wei, C. H., Phan, L., Hefferon, T., Landrum, M., Rehm, H. L., & Lu, Z. (2023). Tracking genetic variants in the biomedical literature using LitVar 2.0. *Nature Genetics, 55*(6), 901–903.

Alves, A. C., Azevedo, S., Benito-Vicente, A., Graça, R., Galicia-Garcia, U., Barros, P., Jordan, P., Martin, C., & Bourbon, M. (2021). LDLR variants functional characterization: Contribution to variant classification. *Atherosclerosis, 329*, 14–21.

Amendola, L. M., Jarvik, G. P., Leo, M. C., McLaughlin, H. M., Akkari, Y., Amaral, M. D., Berg, J. S., Biswas, S., Bowling, K. M., Conlin, L. K., Cooper, G. M., Dorschner, M. O., Dulik, M. C., Ghazani, A. A., Ghosh, R., Green, R. C., Hart, R., Horton, C., Johnston, J. J., Lebo, M. S., … Rehm, H. L. (2016). Performance of ACMG-AMP Variant-Interpretation Guidelines among Nine Laboratories in the Clinical Sequencing Exploratory Research Consortium. *American Journal of Human Genetics, 98*(6), 1067–1076.

Amorosi, C. J., Chiasson, M. A., McDonald, M. G., Wong, L. H., Sitko, K. A., Boyle, G., Kowalski, J. P., Rettie, A. E., Fowler, D. M., & Dunham, M. J. (2021). Massively parallel characterization of CYP2C9 variant enzyme activity and abundance. *American Journal of Human Genetics, 108*(9), 1735–1751.

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N., & Yeh, L. S. (2004). UniProt: the Universal Protein knowledgebase. *Nucleic Acids Research, 32*(Database issue), D115–D119.

Bailey-Wilson, J. E., & Wilson, A. F. (2011). Linkage analysis in the next-generation sequencing era. *Human Heredity, 72*(4), 228–236.

Benito-Vicente, A., Uribe, K. B., Jebari, S., Galicia-Garcia, U., Ostolaza, H., & Martin, C. (2018). Validation of LDLr Activity as a Tool to Improve Genetic Diagnosis of Familial Hypercholesterolemia: A Retrospective on Functional Characterization of LDLr Variants. *International Journal of Molecular Sciences, 19*(6), 1676.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bry-

ant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., Irving, L. J., … Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature, 456*(7218), 53–59.

Biesecker, L. G., Harrison, S. M., & ClinGen Sequence Variant Interpretation Working Group (2018). The ACMG/AMP reputable source criteria for the interpretation of sequence variants. *Genetics in Medicine, 20*(12), 1687–1688.

Biesecker, L. G., Byrne, A. B., Harrison, S. M., Pesaran, T., Schäffer, A. A., Shirts, B. H., Tavtigian, S. V., Rehm, H. L., & ClinGen Sequence Variant Interpretation Working Group (2024). ClinGen guidance for use of the PP1/BS4 co-segregation and PP4 phenotype specificity criteria for sequence variant pathogenicity classification. *American Journal of Human Genetics, 111*(1), 24–38.

Blakes, A. J. M., Wai, H. A., Davies, I., Moledina, H. E., Ruiz, A., Thomas, T., Bunyan, D., Thomas, N. S., Burren, C. P., Greenhalgh, L., Lees, M., Pichini, A., Smithson, S. F., Taylor Tavares, A. L., O'Donovan, P., Douglas, A. G. L., Genomics England Research Consortium, Splicing and Disease Working Group, Whiffin, N., Baralle, D., & Lord, J. (2022). A systematic analysis of splicing variants identifies new diagnoses in the 100,000 Genomes Project. *Genome Medicine, 14*(1), 79.

Boonen, R. A. C. M., Rodrigue, A., Stoepker, C., Wiegant, W. W., Vroling, B., Sharma, M., Rother, M. B., Celosse, N., Vreeswijk, M. P. G., Couch, F., Simard, J., Devilee, P., Masson, J. Y., & van Attikum, H. (2019). Functional analysis of genetic variants in the high-risk breast cancer susceptibility gene PALB2. *Nature Communications, 10*(1), 5296.

Bossuyt, S. N. V., Punt, A. M., de Graaf, I. J., van den Burg, J., Williams, M. G., Heussler, H., Elgersma, Y., & Distel, B. (2021). Loss of nuclear UBE3A activity is the predominant cause of Angelman syndrome in individuals carrying UBE3A missense mutations. *Human Molecular Genetics, 30*(6), 430–442.

Bouwman, P., van der Gulden, H., van der Heijden, I., Drost, R., Klijn, C. N., Prasetyanti, P., Pieterse, M., Wientjens, E., Seibler, J., Hogervorst, F. B., & Jonkers, J. (2013). A high-throughput functional complementation assay for classification of BRCA1 missense variants. *Cancer Discovery, 3*(10), 1142–1155.

Bouwman, P., van der Heijden, I., van der Gulden, H., de Bruijn, R., Braspenning, M. E., Moghadasi, S., Wessels, L. F. A., Dutch-Belgian VUS workgroup, Vreeswijk, M. P. G., & Jonkers, J. (2020). Functional Categorization of BRCA1 Variants of Uncertain Clinical Significance in Homologous Recombination Repair Complementation Assays. *Clinical Cancer Research, 26*(17), 4559–4568.

Brnich, S. E., Rivera-Muñoz, E. A., & Berg, J. S. (2018). Quantifying the potential of functional evidence to reclassify variants of uncertain significance in the categorical and Bayesian interpretation frameworks. *Human Mutation, 39*(11), 1531–1541.

Brnich, S. E., Abou Tayoun, A. N., Couch, F. J., Cutting, G. R., Greenblatt, M. S., Heinen, C. D., Kanavy, D. M., Luo, X., McNulty, S. M., Starita, L. M., Tavtigian, S. V., Wright, M. W., Harrison, S. M., Biesecker, L. G., Berg, J. S., & Clinical Genome Resource Sequence Variant Interpretation Working Group (2019). Recommendations for application of the functional evidence PS3/BS3 criterion using the ACMG/AMP sequence variant interpretation framework. *Genome Medicine, 12*(1), 3.

Brunklaus, A., Schorge, S., Smith, A. D., Ghanty, I., Stewart, K., Gardiner, S., Du, J., Pérez-Palma, E., Symonds, J. D., Collier, A. C., Lal, D., & Zuberi, S. M. (2020). SCN1A variants from bench to bedside-improved clinical prediction from functional characterization. *Human Mutation, 41*(2), 363–374.

Canson, D. M., Davidson, A. L., de la Hoya, M., Parsons, M. T., Glubb, D. M., Kondrashova, O., & Spurdle, A. B. (2023). SpliceAI-10k calculator for the prediction of pseudoexonization, intron retention, and exon deletion. *Bioinformatics 39*(4), btad179.

Carvalho, M. A., Marsillac, S. M., Karchin, R., Manoukian, S., Grist, S., Swaby, R. F., Urmenyi, T. P., Rondinelli, E., Silva, R., Gayol, L., Baumbach, L., Sutphen, R., Pickard-Brzosowicz, J. L., Nathanson, K. L., Sali, A., Goldgar, D., Couch, F. J., Radice, P., & Monteiro, A. N. (2007). Determination of cancer risk associated with germ line BRCA1 missense variants by functional analysis. *Cancer Research, 67*(4), 1494–1501.

Carvalho, M., Pino, M. A., Karchin, R., Beddor, J., Godinho-Netto, M., Mesquita, R. D., Rodarte, R. S., Vaz, D. C., Monteiro, V. A., Manoukian, S., Colombo, M., Ripamonti, C. B., Rosenquist, R., Suthers, G., Borg, A., Radice, P., Grist, S. A., Monteiro, A. N., & Billack, B. (2009). Analysis of a set of missense, frameshift, and in-frame deletion variants of BRCA1. *Mutation Research, 660*(1-2), 1–11.

Chen, Z., Cano, A. H., Romanou, A., Bonnet, A., Matoba, K., Salvi, F., ... & Bosselut, A. (2023). Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint*, arXiv:2311.16079.

Chen, S., Francioli, L. C., Goodrich, J. K., Collins, R. L., Kanai, M., Wang, Q., Alföldi, J., Watts, N. A., Vittal, C., Gauthier, L. D., Poterba, T., Wilson, M. W., Tarasova, Y., Phu, W., Grant, R., Yohannes, M. T., Koenig, Z., Farjoun, Y., Banks, E., Donnelly, S., ... Karczewski, K. J. (2024). A genomic mutational constraint map using variation in 76,156 human genomes. *Nature, 625*(7993), 92–100.

Chiasson, M. A., Rollins, N. J., Stephany, J. J., Sitko, K. A., Matreyek, K. A., Verby, M., Sun, S., Roth, F. P., DeSloover, D., Marks, D. S., Rettie, A. E., & Fowler, D. M. (2020). Multiplexed measurement of variant abundance and activity reveals VKOR topology, active site and human variant impact. *eLife, 9*, e58026.

Christensen, L. L., Kariola, R., Korhonen, M. K., Wikman, F. P., Sunde, L., Gerdes, A. M., Okkels, H., Brandt, C. A., Bernstein, I., Hansen, T. V., Hagemann-Madsen, R., Andersen, C. L., Nyström, M., & Ørntoft, T. F. (2009). Functional characterization of rare missense mutations in MLH1 and MSH2 identified in Danish colorectal cancer patients. *Familial Cancer, 8*(4), 489–500.

Chun, S., & Fay, J. C. (2009). Identification of deleterious mutations within three human genomes. *Genome Research, 19*(9), 1553–1561.

Cingolani, P., Platts, A., Wang, leL., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly, 6*(2), 80–92.

ClinGen. (2022). ClinGen General Sequence Variant Curation Process Standard Operating Procedure [PDF]. https://clinicalgenome.org/docs/variant-curation-standard-operating-procedure-version-3/

Coban-Akdemir, Z., White, J. J., Song, X., Jhangiani, S. N., Fatih, J. M., Gambin, T., Bayram, Y., Chinn, I. K., Karaca, E., Punetha, J., Poli, C., Baylor-Hopkins Center for Mendelian Genomics, Boerwinkle, E., Shaw, C. A., Orange, J. S., Gibbs, R. A., Lappalainen, T., Lupski, J. R., & Carvalho, C. M. B. (2018). Identifying Genes Whose Mutant Transcripts Cause Dominant Disease Traits by Potential Gain-of-Function Alleles. *American Journal of Human Genetics, 103*(2), 171–187.

Collins, R. L., Brand, H., Karczewski, K. J., Zhao, X., Alföldi, J., Francioli, L. C., Khera, A. V., Lowther, C., Gauthier, L. D., Wang, H., Watts, N. A., Solomonson, M., O'Donnell-Luria, A., Baumann, A., Munshi, R., Walker, M., Whelan, C. W., Huang, Y., Brookings, T., Sharpe, T., … Talkowski, M. E. (2020). A structural variation reference for medical and population genetics. *Nature, 581*(7809), 444–451.

Cristofoli, F., Daja, M., Maltese, P. E., Guerri, G., Tanzi, B., Miotto, R., Bonetti, G., Miertus, J., Chiurazzi, P., Stuppia, L., Gatta, V., Cecchin, S., Bertelli, M., & Marceddu, G. (2023). MAGI-ACMG: Algorithm for the Classification of Variants According to ACMG and ACGS Recommendations. *Genes, 14*(8), 1600.

Cummings, B. B., Karczewski, K. J., Kosmicki, J. A., Seaby, E. G., Watts, N. A., Singer-Berk, M., Mudge, J. M., Karjalainen, J., Satterstrom, F. K., O'Donnell-Luria, A. H., Poterba, T., Seed, C., Solomonson, M., Alföldi, J., Genome Aggregation

Database Production Team, Genome Aggregation Database Consortium, Daly, M. J., & MacArthur, D. G. (2020). Transcript expression-aware annotation improves rare variant interpretation. *Nature, 581*(7809), 452–458.

Çubuk, H., & Yalçın Çapan, Ö. (2021). A Review of Functional Characterization of Single Amino Acid Change Mutations in HNF Transcription Factors in MODY Pathogenesis. The *Protein Journal, 40*(3), 348–360.

Danis, D., Jacobsen, J. O. B., Carmody, L. C., Gargano, M. A., McMurry, J. A., Hegde, A., Haendel, M. A., Valentini, G., Smedley, D., & Robinson, P. N. (2021). Interpretable prioritization of splice variants in diagnostic next-generation sequencing. *American Journal of Human Genetics, 108*(9), 1564–1577.

Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., & Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Computational Biology, 6*(12), e1001025.

Dawes, R., Bournazos, A. M., Bryen, S. J., Bommireddipalli, S., Marchant, R. G., Joshi, H., & Cooper, S. T. (2023). SpliceVault predicts the precise nature of variant-associated mis-splicing. *Nature Genetics, 55*(2), 324–332.

Del Angel, G., Reynders, J., Negron, C., Steinbrecher, T., & Mornet, E. (2020). Large-scale in vitro functional testing and novel variant scoring via protein modeling provide insights into alkaline phosphatase activity in hypophosphatasia. *Human Mutation, 41*(7), 1250–1262.

Delimitsou, A., Fostira, F., Kalfakakou, D., Apostolou, P., Konstantopoulou, I., Kroupis, C., Papavassiliou, A. G., Kleibl, Z., Stratikos, E., Voutsinas, G. E., & Yannoukakos, D. (2019). Functional characterization of CHEK2 variants in a Saccharomyces cerevisiae system. *Human Mutation, 40*(5), 631–648.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., & Daly, M. J. (2011). A framework for variation discovery and

genotyping using next-generation DNA sequencing data. *Nature Genetics, 43*(5), 491–498.

DesRoches, C. L., Bruun, T., Wang, P., Marshall, C. R., & Mercimek-Mahmutoglu, S. (2016). Arginine-Glycine Amidinotransferase Deficiency and Functional Characterization of Missense Variants in GATM. *Human Mutation, 37*(9), 926–932.

di Iulio, J., Bartha, I., Wong, E. H. M., Yu, H. C., Lavrenko, V., Yang, D., Jung, I., Hicks, M. A., Shah, N., Kirkness, E. F., Fabani, M. M., Biggs, W. H., Ren, B., Venter, J. C., & Telenti, A. (2018). The human noncoding genome defined by genetic diversity. *Nature Genetics, 50*(3), 333–337.

Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K., & Liu, X. (2015). Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Human Molecular Genetics, 24*(8), 2125–2137.

Duan, J., Shi, J., Fiorentino, A., Leites, C., Chen, X., Moy, W., Chen, J., Alexandrov, B. S., Usheva, A., He, D., Freda, J., O'Brien, N. L., Molecular Genetics of Schizophrenia collaboration, Genomic Psychiatric Cohort consortium, McQuillin, A., Sanders, A. R., Gershon, E. S., DeLisi, L. E., Bishop, A. R., Gurling, H. M., … Gejman, P. V. (2014). A rare functional noncoding variant at the GWAS-implicated MIR137/MIR2682 locus might confer risk to schizophrenia and bipolar disorder. *American Journal of Human Genetics, 95*(6).

Durkie, M., Cassidy, E.-J., Berry, I., Owens, M., Turnbull, C., Scott, R. H., Taylor, R. W., Deans, Z. C., Ellard, S., Baple, E. L., & McMullan, D. J. (2024). ACGS Best Practice Guidelines for Variant Classification in Rare Disease 2024 [PDF]. https://www.acgs.uk.com/media/12533/uk-practice-guidelines-for-variant-classification-v12-2024.pdf

Eilbeck, K., Lewis, S. E., Mungall, C. J., Yandell, M., Stein, L., Durbin, R., & Ashburner, M. (2005). The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biology, 6*(5), R44.

Eilbeck, K., Quinlan, A., & Yandell, M. (2017). Settling the score: variant prioritization and Mendelian disease. *Nature Reviews Genetics, 18*(10), 599–612.

Ellard, S., Baple, E. L., Callaway, A., Berry, I., Forrester, N., Turnbull, C., Owens, M., Eccles, D. M., Abbs, S., Scott, R., & et al. (2020). ACGS Best Practice Guidelines for Variant Classification in Rare Disease 2020 [PDF]. https://www.acgs.uk.com/media/11631/uk-practice-guidelines-for-variant-classification-v4-01-2020.pdf

Ellingford, J. M., Ahn, J. W., Bagnall, R. D., Baralle, D., Barton, S., Campbell, C., Downes, K., Ellard, S., Duff-Farrier, C., FitzPatrick, D. R., Greally, J. M., Ingles, J., Krishnan, N., Lord, J., Martin, H. C., Newman, W. G., O'Donnell-Luria, A., Ramsden, S. C., Rehm, H. L., Richardson, E., … Whiffin, N. (2022). Recommendations for clinical interpretation of variants found in non-coding regions of the genome. *Genome Medicine, 14*(1), 73.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature, 489*(7414), 57–74.

Erwood, S., Bily, T. M. I., Lequyer, J., Yan, J., Gulati, N., Brewer, R. A., Zhou, L., Pelletier, L., Ivakine, E. A., & Cohn, R. D. (2022). Saturation variant interpretation using CRISPR prime editing. *Nature Biotechnology, 40*(6), 885–895.

Esplin, E. D., Oei, L., & Snyder, M. P. (2014). Personalized sequencing and the future of medicine: discovery, diagnosis and defeat of disease. *Pharmacogenomics, 15*(14), 1771–1790.

Fahed, A. C., Wang, M., Homburger, J. R., Patel, A. P., Bick, A. G., Neben, C. L., Lai, C., Brockman, D., Philippakis, A., Ellinor, P. T., Cassa, C. A., Lebo, M., Ng, K., Lander, E. S., Zhou, A. Y., Kathiresan, S., & Khera, A. V. (2020). Polygenic

background modifies penetrance of monogenic variants for tier 1 genomic conditions. *Nature Communications, 11*(1), 3635.

Feng B. J. (2017). PERCH: A Unified Framework for Disease Gene Prioritization. *Human Mutation, 38*(3), 243–251.

Findlay, G. M., Daza, R. M., Martin, B., Zhang, M. D., Leith, A. P., Gasperini, M., Janizek, J. D., Huang, X., Starita, L. M., & Shendure, J. (2018). Accurate classification of BRCA1 variants with saturation genome editing. *Nature, 562*(7726), 217–222.

Frankish, A., Diekhans, M., Jungreis, I., Lagarde, J., Loveland, J. E., Mudge, J. M., Sisu, C., Wright, J. C., Armstrong, J., Barnes, I., Berry, A., Bignell, A., Boix, C., Carbonell Sala, S., Cunningham, F., Di Domenico, T., Donaldson, S., Fiddes, I. T., García Girón, C., Gonzalez, J. M., … Flicek, P. (2021). GENCODE 2021. *Nucleic Acids Research, 49*(D1), D916–D923.

Frederiksen, J. H., Jensen, S. B., Tümer, Z., & Hansen, T. V. O. (2021). Classification of MSH6 Variants of Uncertain Significance Using Functional Assays. *International Journal of Molecular Sciences, 22*(16), 8627.

Fu, W., O'Connor, T. D., Jun, G., Kang, H. M., Abecasis, G., Leal, S. M., Gabriel, S., Rieder, M. J., Altshuler, D., Shendure, J., Nickerson, D. A., Bamshad, M. J., NHLBI Exome Sequencing Project, & Akey, J. M. (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature, 493*(7431), 216–220.

Garber, M., Guttman, M., Clamp, M., Zody, M. C., Friedman, N., & Xie, X. (2009). Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics (Oxford, England), 25(12), i54–i62. https://doi.org/10.1093/bioinformatics/btp190

Gelb, B. D., Cavé, H., Dillon, M. W., Gripp, K. W., Lee, J. A., Mason-Suares, H., Rauen, K. A., Williams, B., Zenker, M., Vincent, L. M., & ClinGen RASopathy

Working Group (2018). ClinGen's RASopathy Expert Panel consensus methods for variant interpretation. *Genetics in Medicine, 20*(11), 1334–1345.

Ghosh, R., Oak, N., & Plon, S. E. (2017). Evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines. *Genome Biology, 18*(1), 225.

Ghosh, R., Harrison, S. M., Rehm, H. L., Plon, S. E., Biesecker, L. G., & ClinGen Sequence Variant Interpretation Working Group (2018). Updated recommendation for the benign stand-alone ACMG/AMP criterion. *Human Mutation, 39*(11), 1525–1530.

Gilissen, C., Hoischen, A., Brunner, H. G., & Veltman, J. A. (2011). Unlocking Mendelian disease using exome sequencing. *Genome Biology, 12*(9), 228.

Graça, R., Alves, A. C., Zimon, M., Pepperkok, R., & Bourbon, M. (2022). Functional profiling of LDLR variants: Important evidence for variant classification: Functional profiling of LDLR variants. *Journal of Clinical Lipidology, 16*(4), 516–524.

Green, R. C., Berg, J. S., Grody, W. W., Kalia, S. S., Korf, B. R., Martin, C. L., McGuire, A. L., Nussbaum, R. L., O'Daniel, J. M., Ormond, K. E., Rehm, H. L., Watson, M. S., Williams, M. S., Biesecker, L. G., & American College of Medical Genetics and Genomics (2013). ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genetics in Medicine, 15*(7), 565–574.

Green, E. D., Watson, J. D., & Collins, F. S. (2015). Human Genome Project: Twenty-five years of big biology. *Nature, 526*(7571), 29–31.

Guidugli, L., Shimelis, H., Masica, D. L., Pankratz, V. S., Lipton, G. B., Singh, N., Hu, C., Monteiro, A. N. A., Lindor, N. M., Goldgar, D. E., Karchin, R., Iversen, E. S., & Couch, F. J. (2018). Assessment of the Clinical Relevance of BRCA2 Missense Variants by Functional and Computational Approaches. *American Journal of Human Genetics, 102*(2), 233–248.

Harrison, S. M., Dolinsky, J. S., Knight Johnson, A. E., Pesaran, T., Azzariti, D. R., Bale, S., Chao, E. C., Das, S., Vincent, L., & Rehm, H. L. (2017). Clinical laboratories collaborate to resolve differences in variant interpretations submitted to ClinVar. *Genetics in Medicine, 19*(10), 1096–1104.

Harrison, S. M., Biesecker, L. G., & Rehm, H. L. (2019). Overview of Specifications to the ACMG/AMP Variant Interpretation Guidelines. *Current Protocols in Human Genetics, 103*(1), e93.

Harrison, S. M. & Biesecker L. G. (2021) *ACMG/AMP Updated Guidelines and Engaging Underrepresented Groups in ClinGen* [Video]. ClinGen Virtual Retreat 2021. https://clinicalgenome.org/tools/clingen-virtual-retreat-2021/acmg-amp-updated-guidelines-and-engaging-underrepresented-groups-in-clingen-main-session

Harrison, P. W., Amode, M. R., Austine-Orimoloye, O., Azov, A. G., Barba, M., Barnes, I., Becker, A., Bennett, R., Berry, A., Bhai, J., Bhurji, S. K., Boddu, S., Branco Lins, P. R., Brooks, L., Ramaraju, S. B., Campbell, L. I., Martinez, M. C., Charkhchi, M., Chougule, K., Cockburn, A., … Yates, A. D. (2024). Ensembl 2024. *Nucleic Acids Research, 52*(D1), D891–D899.

Hart, S. N., Hoskin, T., Shimelis, H., Moore, R. M., Feng, B., Thomas, A., Lindor, N. M., Polley, E. C., Goldgar, D. E., Iversen, E., Monteiro, A. N. A., Suman, V. J., & Couch, F. J. (2019). Comprehensive annotation of BRCA1 and BRCA2 missense variants by functionally validated sequence-based computational prediction models. *Genetics in Medicine, 21*(1), 71–80.

Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics, 107*(1), 1–8.

Helgadottir, A., Thorleifsson, G., Manolescu, A., Gretarsdottir, S., Blondal, T., Jonasdottir, A., Jonasdottir, A., Sigurdsson, A., Baker, A., Palsson, A., Masson, G., Gudbjartsson, D. F., Magnusson, K. P., Andersen, K., Levey, A. I., Backman, V. M., Matthiasdottir, S., Jonsdottir, T., Palsson, S., Einarsdottir, H., …

Stefansson, K. (2007). A common variant on chromosome 9p21 affects the risk of myocardial infarction. *Science, 316*(5830), 1491–1493.

Hong, S., Hu, P., Jang, J. H., Carrington, B., Sood, R., Berger, S. I., Roessler, E., & Muenke, M. (2020). Functional analysis of Sonic Hedgehog variants associated with holoprosencephaly in humans using a CRISPR/Cas9 zebrafish model. *Human Mutation, 41*(12), 2155–2166.

Hong, K. U., Walls, K. M., & Hein, D. W. (2023). Non-coding and intergenic genetic variants of human arylamine N-acetyltransferase 2 (NAT2) gene are associated with differential plasma lipid and cholesterol levels and cardiometabolic disorders. *Frontiers in Pharmacology, 14*, 1091976.

Hu, T., Chitnis, N., Monos, D., & Dinh, A. (2021). Next-generation sequencing technologies: An overview. *Human Immunology, 82*(11), 801–811.

Hu, C., Susswein, L. R., Roberts, M. E., Yang, H., Marshall, M. L., Hiraki, S., Berkofsky-Fessler, W., Gupta, S., Shen, W., Dunn, C. A., Huang, H., Na, J., Domchek, S. M., Yadav, S., Monteiro, A. N. A., Polley, E. C., Hart, S. N., Hruska, K. S., & Couch, F. J. (2022). Classification of BRCA2 Variants of Uncertain Significance (VUS) Using an ACMG/AMP Model Incorporating a Homology-Directed Repair (HDR) Functional Assay. *Clinical Cancer Research, 28*(17), 3742–3751.

Hu, C., Nagaraj, A. B., Shimelis, H., Montalban, G., Lee, K. Y., Huang, H., Lumby, C. A., Na, J., Susswein, L. R., Roberts, M. E., Marshall, M. L., Hiraki, S., LaDuca, H., Chao, E., Yussuf, A., Pesaran, T., Neuhausen, S. L., Haiman, C. A., Kraft, P., Lindstrom, S., … Couch, F. J. (2023). Functional and Clinical Characterization of Variants of Uncertain Significance Identifies a Hotspot for Inactivating Missense Variants in RAD51C. *Cancer Research, 83*(15), 2557–2571.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Huminiecki, L., Kasprzyk, A., Lehvaslaiho, H., Lijnzaad, P., Melsopp, C., Mongin,

E., … Clamp, M. (2002). The Ensembl genome database project. *Nucleic Acids Research, 30*(1), 38–41.

Iqbal, S., Pérez-Palma, E., Jespersen, J. B., May, P., Hoksza, D., Heyne, H. O., Ahmed, S. S., Rifat, Z. T., Rahman, M. S., Lage, K., Palotie, A., Cottrell, J. R., Wagner, F. F., Daly, M. J., Campbell, A. J., & Lal, D. (2020). Comprehensive characterization of amino acid positions in protein structures reveals molecular effect of missense variants. *Proceedings of the National Academy of Sciences of the United States of America, 117*(45), 28201–28211.

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature, 431*(7011), 931–945.

Iwasawa, S., Kikuchi, A., Wada, Y., Arai-Ichinoi, N., Sakamoto, O., Tamiya, G., & Kure, S. (2019). The prevalence of GALM mutations that cause galactosemia: A database of functionally evaluated variants. *Molecular Genetics and Metabolism, 126*(4), 362–367.

Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I., Kosmicki, J. A., Arbelaez, J., Cui, W., Schwartz, G. B., Chow, E. D., Kanterakis, E., Gao, H., Kia, A., Batzoglou, S., Sanders, S. J., & Farh, K. K. (2019). Predicting Splicing from Primary Sequence with Deep Learning. *Cell, 176*(3), 535–548.e24.

Jia, X., Burugula, B. B., Chen, V., Lemons, R. M., Jayakody, S., Maksutova, M., & Kitzman, J. O. (2021). Massively parallel functional testing of MSH2 missense variants conferring Lynch syndrome risk. *American Journal of Human Genetics, 108*(1), 163–175.

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T. S., Hinrichs, A., Lu, Y. T., Roskin, K. M., Schwartz, M., Sugnet, C. W., Thomas, D. J., Weber, R. J., Haussler, D., Kent, W. J., & University of California Santa Cruz (2003). The UCSC Genome Browser Database. *Nucleic Acids Research, 31*(1), 51–54.

Kars, M. E., Başak, A. N., Onat, O. E., Bilguvar, K., Choi, J., Itan, Y., Çağlar, C., Palvadeau, R., Casanova, J. L., Cooper, D. N., Stenson, P. D., Yavuz, A., Buluş, H., Günel, M., Friedman, J. M., & Özçelik, T. (2021). The genetic structure of the Turkish population reveals high levels of variation and admixture. *Proceedings of the National Academy of Sciences, 118*(36), e2026076118.

Katsonis, P., Wilhelm, K., Williams, A., & Lichtarge, O. (2022). Genome interpretation using in silico predictors of variant impact. *Human Genetics, 141*(10), 1549–1577.

Kazazian, J., Boehm, C. D., & Seltzer, W. K. (2000). ACMG recommendations for standards for interpretation of sequence variations. *Genetics in Medicine, 2*(5), 302-303.

Kelly, M. A., Caleshu, C., Morales, A., Buchan, J., Wolf, Z., Harrison, S. M., Cook, S., Dillon, M. W., Garcia, J., Haverfield, E., Jongbloed, J. D. H., Macaya, D., Manrai, A., Orland, K., Richard, G., Spoonamore, K., Thomas, M., Thomson, K., Vincent, L. M., Walsh, R., … Funke, B. (2018). Adaptation and validation of the ACMG/AMP variant classification framework for MYH7-associated inherited cardiomyopathies: recommendations by ClinGen's Inherited Cardiomyopathy Expert Panel. *Genetics in Medicine, 20*(3), 351–359.

Kim, H. K., Lee, E. J., Lee, Y. J., Kim, J., Kim, Y., Kim, K., Lee, S. W., Chang, S., Lee, Y. J., Lee, J. W., Lee, W., Chun, S., Son, B. H., Jung, K. H., Kim, Y. M., Min, W. K., & Ahn, S. H. (2020). Impact of proactive high-throughput functional assay data on BRCA1 variant interpretation in 3684 patients with breast or ovarian cancer. *Journal of Human Genetics, 65*(3), 209–220.

Kimura, M. (1983). The Neutral Theory of Molecular Evolution. *Cambridge: Cambridge University Press*.

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics, 46*(3), 310–315.

Klauer, A. A., & van Hoof, A. (2012). Degradation of mRNAs that lack a stop codon: a decade of nonstop progress. *Wiley Interdisciplinary Reviews: RNA, 3*(5), 649–660.

Klemm, S. L., Shipony, Z., & Greenleaf, W. J. (2019). Chromatin accessibility and the regulatory epigenome. Nature reviews. *Genetics, 20*(4), 207–220.

Kobayashi, Y., Yang, S., Nykamp, K., Garcia, J., Lincoln, S. E., & Topper, S. E. (2017). Pathogenic variant burden in the ExAC database: an empirical approach to evaluating population data for clinical variant interpretation. *Genome Medicine, 9*(1), 13.

Kopanos, C., Tsiolkas, V., Kouris, A., Chapple, C. E., Albarca Aguilera, M., Meyer, R., & Massouras, A. (2019). VarSome: the human genomic variant search engine. *Bioinformatics, 35*(11), 1978–1980.

Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., & Maglott, D. R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research, 42*(Database issue), D980–D985.

Lee, K., Krempely, K., Roberts, M. E., Anderson, M. J., Carneiro, F., Chao, E., Dixon, K., Figueiredo, J., Ghosh, R., Huntsman, D., Kaurah, P., Kesserwan, C., Landrith, T., Li, S., Mensenkamp, A. R., Oliveira, C., Pardo, C., Pesaran, T., Richardson, M., Slavin, T. P., … Karam, R. (2018). Specifications of the ACMG/AMP variant curation guidelines for the analysis of germline CDH1 sequence variants. *Human Mutation, 39*(11), 1553–1568.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics, 36*(4), 1234–1240.

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., Tukiainen, T., Birnbaum, D. P., Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F., Zou, J.,

Pierce-Hoffman, E., Berghout, J., Cooper, D. N., … Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature, 536*(7616), 285–291.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics, 25*(14), 1754–1760.

Li, Q., & Wang, K. (2017). InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. *American Journal of Human Genetics, 100*(2), 267–280.

Liao, W. W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J. K., Monlong, J., Abel, H. J., Buonaiuto, S., Chang, X. H., Cheng, H., Chu, J., Colonna, V., Eizenga, J. M., Feng, X., Fischer, C., Fulton, R. S., Garg, S., … Paten, B. (2023). A draft human pangenome reference. *Nature, 617*(7960), 312–324.

Lindeboom, R. G., Supek, F., & Lehner, B. (2016). The rules and impact of nonsense-mediated mRNA decay in human cancers. *Nature Genetics, 48*(10), 1112–1118.

Lo, R. S., Cromie, G. A., Tang, M., Teng, K., Owens, K., Sirr, A., Kutz, J. N., Morizono, H., Caldovic, L., Ah Mew, N., Gropman, A., & Dudley, A. M. (2023). The functional impact of 1,570 individual amino acid substitutions in human OTC. American Journal of Human Genetics, 110(5), 863–879.

Luo, X., Feurstein, S., Mohan, S., Porter, C. C., Jackson, S. A., Keel, S., Chicka, M., Brown, A. L., Kesserwan, C., Agarwal, A., Luo, M., Li, Z., Ross, J. E., Baliakas, P., Pineda-Alvarez, D., DiNardo, C. D., Bertuch, A. A., Mehta, N., Vulliamy, T., Wang, Y., … Godley, L. A. (2019). ClinGen Myeloid Malignancy Variant Curation Expert Panel recommendations for germline RUNX1 variants. *Blood Advances, 3*(20), 2962–2979.

Luo, X., Wang, R., Sun, Y., Qiu, W., Lu, D., Wang, Y., Gong, Z., Zhang, H., Han, L., Liang, L., Gu, X., Yu, Y., & Xiao, B. (2023). Deep Intronic PAH Variants Ex-

plain Missing Heritability in Hyperphenylalaninemia. *The Journal of Molecular Diagnostics, 25*(5), 284–294.

Madelaine, R., Notwell, J. H., Skariah, G., Halluin, C., Chen, C. C., Bejerano, G., & Mourrain, P. (2018). A screen for deeply conserved non-coding GWAS SNPs uncovers a MIR-9-2 functional mutation associated to retinal vasculature defects in human. *Nucleic Acids Research, 46*(7), 3517–3531.

Mahajan, A., Spracklen, C. N., Zhang, W., Ng, M. C. Y., Petty, L. E., Kitajima, H., Yu, G. Z., Rüeger, S., Speidel, L., Kim, Y. J., Horikoshi, M., Mercader, J. M., Taliun, D., Moon, S., Kwak, S. H., Robertson, N. R., Rayner, N. W., Loh, M., Kim, B. J., Chiou, J., … Morris, A. P. (2022). Multi-ancestry genetic study of type 2 diabetes highlights the power of diverse populations for discovery and translation. *Nature Genetics, 54*(5), 560–572.

Mardis E. R. (2011). A decade's perspective on DNA sequencing technology. *Nature, 470*(7333), 198–203.

Marees, A. T., de Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C., & Derks, E. M. (2018). A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International Journal of Methods in Psychiatric Research, 27*(2), e1608.

Martin, A. R., Williams, E., Foulger, R. E., Leigh, S., Daugherty, L. C., Niblock, O., Leong, I. U. S., Smith, K. R., Gerasimenko, O., Haraldsdottir, E., Thomas, E., Scott, R. H., Baple, E., Tucci, A., Brittain, H., de Burca, A., Ibañez, K., Kasperaviciute, D., Smedley, D., Caulfield, M., … McDonagh, E. M. (2019). PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nature Genetics, 51*(11), 1560–1565.

Maxson Jones, K., Ankeny, R. A., & Cook-Deegan, R. (2018). The Bermuda Triangle: The Pragmatics, Policies, and Principles for Data Sharing in the History of the Human Genome Project. *Journal of the History of Biology, 51*(4), 693–805.

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., Flicek, P., & Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome biology, 17*(1), 122.

McPherson, J. D., Marra, M., Hillier, L., Waterston, R. H., Chinwalla, A., Wallis, J., Sekhon, M., Wylie, K., Mardis, E. R., Wilson, R. K., Fulton, R., Kucaba, T. A., Wagner-McPherson, C., Barbazuk, W. B., Gregory, S. G., Humphray, S. J., French, L., Evans, R. S., Bethel, G., Whittaker, A., … International Human Genome Mapping Consortium (2001). A physical map of the human genome. *Nature, 409*(6822), 934–941.

Mester, J. L., Ghosh, R., Pesaran, T., Huether, R., Karam, R., Hruska, K. S., Costa, H. A., Lachlan, K., Ngeow, J., Barnholtz-Sloan, J., Sesock, K., Hernandez, F., Zhang, L., Milko, L., Plon, S. E., Hegde, M., & Eng, C. (2018). Gene-specific criteria for PTEN variant curation: Recommendations from the ClinGen PTEN Expert Panel. *Human Mutation, 39*(11), 1581–1592.

Mighell, T. L., Evans-Dutson, S., & O'Roak, B. J. (2018). A Saturation Mutagenesis Approach to Understanding PTEN Lipid Phosphatase Activity and Genotype-Phenotype Relationships. *American Journal of Human Genetics, 102*(5), 943–955.

Miller, D. T., Lee, K., Chung, W. K., Gordon, A. S., Herman, G. E., Klein, T. E., Stewart, D. R., Amendola, L. M., Adelman, K., Bale, S. J., Gollob, M. H., Harrison, S. M., Hershberger, R. E., McKelvey, K., Richards, C. S., Vlangos, C. N., Watson, M. S., Martin, C. L., & ACMG Secondary Findings Working Group (2021). ACMG SF v3.0 list for reporting of secondary findings in clinical exome and genome sequencing: a policy statement of the American College of Medical Genetics and Genomics (ACMG). *Genetics in Medicine, 23*(8), 1381–1390.

Miller, D. T., Lee, K., Abul-Husn, N. S., Amendola, L. M., Brothers, K., Chung, W. K., Gollob, M. H., Gordon, A. S., Harrison, S. M., Hershberger, R. E., Klein, T. E., Richards, C. S., Stewart, D. R., Martin, C. L., & ACMG Secondary Findings

Working Group. (2023). ACMG SF v3.2 list for reporting of secondary findings in clinical exome and genome sequencing: A policy statement of the American College of Medical Genetics and Genomics (ACMG). *Genetics in Medicine, 25*(8), 100866.

Ng, P. C., & Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Research, 31*(13), 3812–3814.

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., Caldas, G. V., … Phillippy, A. M. (2022). The complete sequence of a human genome. *Science, 376*(6588), 44–53.

Nussbaum, R. L., McInnes, R. R., & Willard, H. F. (2016). Chapter 9: Genetic Variation in Populations. *Thompson & Thompson Genetics in Medicine (8th ed.)* (pp. 165-166). Elsevier.

Nykamp, K., Anderson, M., Powers, M., Garcia, J., Herrera, B., Ho, Y. Y., Kobayashi, Y., Patil, N., Thusberg, J., Westbrook, M., Invitae Clinical Genomics Group, & Topper, S. (2017). Sherloc: a comprehensive refinement of the ACMG-AMP variant classification criteria. *Genetics in Medicine, 19*(10), 1105–1117.

Oza, A. M., DiStefano, M. T., Hemphill, S. E., Cushman, B. J., Grant, A. R., Siegert, R. K., Shen, J., Chapin, A., Boczek, N. J., Schimmenti, L. A., Murry, J. B., Hasadsri, L., Nara, K., Kenna, M., Booth, K. T., Azaiez, H., Griffith, A., Avraham, K. B., Kremer, H., Rehm, H. L., … ClinGen Hearing Loss Clinical Domain Working Group (2018). Expert specification of the ACMG/AMP variant interpretation guidelines for genetic hearing loss. *Human Mutation, 39*(11), 1593–1613.

Pardiñas, A. F., Holmans, P., Pocklington, A. J., Escott-Price, V., Ripke, S., Carrera, N., Legge, S. E., Bishop, S., Cameron, D., Hamshere, M. L., Han, J., Hubbard, L., Lynham, A., Mantripragada, K., Rees, E., MacCabe, J. H., McCarroll, S. A.,

Baune, B. T., Breen, G., Byrne, E. M., … Walters, J. T. R. (2018). Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nature Genetics, 50*(3), 381–389.

Payankaulam, S., Raicu, A. M., & Arnosti, D. N. (2019). Transcriptional Regulation of INSR, the Insulin Receptor Gene. *Genes, 10*(12), 984.

Pejaver, V., Byrne, A. B., Feng, B. J., Pagel, K. A., Mooney, S. D., Karchin, R., O'Donnell-Luria, A., Harrison, S. M., Tavtigian, S. V., Greenblatt, M. S., Biesecker, L. G., Radivojac, P., Brenner, S. E., & ClinGen Sequence Variant Interpretation Working Group (2022). Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. *American Journal of Human Genetics, 109*(12), 2163–2177.

Peng, J., Xiang, J., Jin, X., Meng, J., Song, N., Chen, L., Abou Tayoun, A., & Peng, Z. (2021). VIP-HL: Semi-automated ACMG/AMP variant interpretation platform for genetic hearing loss. *Human Mutation, 42*(12), 1567–1575.

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., & Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research, 20*(1), 110–121.

Post, K. L., Belmadani, M., Ganguly, P., Meili, F., Dingwall, R., McDiarmid, T. A., Meyers, W. M., Herrington, C., Young, B. P., Callaghan, D. B., Rogic, S., Edwards, M., Niciforovic, A., Cau, A., Rankin, C. H., O'Connor, T. P., Bamji, S. X., Loewen, C. J. R., Allan, D. W., Pavlidis, P., … Haas, K. (2020). Multi-model functionalization of disease-associated PTEN missense mutations identifies multiple molecular mechanisms underlying protein dysfunction. *Nature Communications, 11*(1), 2073.

Prakash, R., Rawal, Y., Sullivan, M. R., Grundy, M. K., Bret, H., Mihalevic, M. J., Rein, H. L., Baird, J. M., Darrah, K., Zhang, F., Wang, R., Traina, T. A., Radke, M. R., Kaufmann, S. H., Swisher, E. M., Guérois, R., Modesti, M., Sung, P., Jasin, M., & Bernstein, K. A. (2022). Homologous recombination-deficient

mutation cluster in tumor suppressor RAD51C identified by comprehensive analysis of cancer variants. *Proceedings of the National Academy of Sciences of the United States of America, 119*(38), e2202727119.

Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research, 35*(Database issue), D61–D65.

Ramensky, V., Bork, P., & Sunyaev, S. (2002). Human non-synonymous SNPs: server and survey. *Nucleic Acids Research, 30*(17), 3894–3900.

Rehm, H. (2023). Research on testing phase: what tests to recommend, how are they offered and what types of results are provided [PDF]. https://www.genome.gov/sites/default/files/media/files/2023-11/Session_7-2__Rehm.pdf

Reva, B., Antipin, Y., & Sander, C. (2011). Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Research, 39*(17), e118.

Richards, C. S., Bale, S., Bellissimo, D. B., Das, S., Grody, W. W., Hegde, M. R., Lyon, E., Ward, B. E., & Molecular Subcommittee of the ACMG Laboratory Quality Assurance Committee (2008). ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007. *Genetics in Medicine, 10*(4), 294–300.

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., Rehm, H. L., & ACMG Laboratory Quality Assurance Committee (2015). 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. *Genetics in Medicine, 17*(5), 405–424.

Richards, C. S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., Rehm, H. L., & ACMG/AMP

Interpretation of Sequence Variants Work Group 2015 (2018). Response to Biesecker and Harrison. *Genetics in Medicine, 20*(12), 1689–1690.

Richardson, M. E., Hu, C., Lee, K. Y., LaDuca, H., Fulk, K., Durda, K. M., Deckman, A. M., Goldgar, D. E., Monteiro, A. N. A., Gnanaolivu, R., Hart, S. N., Polley, E. C., Chao, E., Pesaran, T., & Couch, F. J. (2021). Strong functional data for pathogenicity or neutrality classify BRCA2 DNA-binding-domain variants of uncertain significance. *American Journal of Human Genetics, 108*(3), 458–468.

Riggs, E. R., Church, D. M., Hanson, K., Horner, V. L., Kaminsky, E. B., Kuhn, R. M., Wain, K. E., Williams, E. S., Aradhya, S., Kearney, H. M., Ledbetter, D. H., South, S. T., Thorland, E. C., & Martin, C. L. (2012). Towards an evidence-based process for the clinical interpretation of copy number variation. *Clinical Genetics, 81*(5), 403–412.

Riggs, E. R., Andersen, E. F., Cherry, A. M., Kantarci, S., Kearney, H., Patel, A., Raca, G., Ritter, D. I., South, S. T., Thorland, E. C., Pineda-Alvarez, D., Aradhya, S., & Martin, C. L. (2020). Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen). *Genetics in Medicine, 22*(2), 245–257.

Samocha, K. E., Robinson, E. B., Sanders, S. J., Stevens, C., Sabo, A., McGrath, L. M., Kosmicki, J. A., Rehnström, K., Mallick, S., Kirby, A., Wall, D. P., MacArthur, D. G., Gabriel, S. B., DePristo, M., Purcell, S. M., Palotie, A., Boerwinkle, E., Buxbaum, J. D., Cook, E. H., Jr, Gibbs, R. A., … Daly, M. J. (2014). A framework for the interpretation of de novo mutation in human disease. *Nature Genetics*, 46(9), 944–950.

Samocha, K. E., Kosmicki, J. A., Karczewski, K. J., O'Donnell-Luria, A. H., Pierce-Hoffman, E., MacArthur, D. G., Neale, B. M., & Daly, M. J. (2017). Regional

missense constraint improves variant deleteriousness prediction. *bioRxiv*, 148353.

Sanchis-Juan, A., Megy, K., Stephens, J., Armirola Ricaurte, C., Dewhurst, E., Low, K., French, C. E., Grozeva, D., Stirrups, K., Erwood, M., McTague, A., Penkett, C. J., Shamardina, O., Tuna, S., Daugherty, L. C., Gleadall, N., Duarte, S. T., Hedrera-Fernández, A., Vogt, J., Ambegaonkar, G., … Carss, K. J. (2023). Genome sequencing and comprehensive rare-variant analysis of 465 families with neurodevelopmental disorders. *American Journal of Human Genetics, 110*(8), 1343–1355.

Schneider, V. A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H. C., Kitts, P. A., Murphy, T. D., Pruitt, K. D., Thibaud-Nissen, F., Albracht, D., Fulton, R. S., Kremitzki, M., Magrini, V., Markovic, C., McGrath, S., Steinberg, K. M., Auger, K., Chow, W., Collins, J., Harden, G., … Church, D. M. (2017). Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research, 27*(5), 849–864.

Schwarz, J. M., Rödelsperger, C., Schuelke, M., & Seelow, D. (2010). MutationTaster evaluates disease-causing potential of sequence alterations. *Nature Methods, 7*(8), 575–576.

Scott, A. D., Huang, K. L., Weerasinghe, A., Mashl, R. J., Gao, Q., Martins Rodrigues, F., Wyczalkowski, M. A., & Ding, L. (2019). CharGer: clinical Characterization of Germline variants. *Bioinformatics, 35*(5), 865–867.

Scott, A., Hernandez, F., Chamberlin, A., Smith, C., Karam, R., & Kitzman, J. O. (2022). Saturation-scale functional evidence supports clinical variant interpretation in Lynch syndrome. *Genome Biology, 23*(1), 266.

Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L., Edwards, K. J., Day, I. N., & Gaunt, T. R. (2013). Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Human Mutation, 34*(1), 57–65.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., Weinstock, G. M., Wilson, R. K., Gibbs, R. A., Kent, W. J., Miller, W., & Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research, 15*(8), 1034–1050.

Shah, B. P., Sleiman, P. M., Mc Donald, J., Moeller, I. H., & Kleyn, P. (2023). Functional characterization of all missense variants in LEPR, PCSK1, and POMC genes arising from single-nucleotide variants. *Expert Review of Endocrinology & Metabolism, 18*(2), 209–219.

Silva, S., Alves, A. C., Patel, D., Malhó, R., Soutar, A. K., & Bourbon, M. (2012). In vitro functional characterization of missense mutations in the LDLR gene. *Atherosclerosis, 225*(1), 128–134.

Singer-Berk, M., Gudmundsson, S., Baxter, S., Seaby, E. G., England, E., Wood, J. C., Son, R. G., Watts, N. A., Karczewski, K. J., Harrison, S. M., MacArthur, D. G., Rehm, H. L., & O'Donnell-Luria, A. (2023). Advanced variant classification framework reduces the false positive rate of predicted loss-of-function variants in population sequencing data. *American Journal of Human Genetics, 110*(9), 1496–1508.

Singh, R. K., & Cooper, T. A. (2012). Pre-mRNA splicing in disease and therapeutics. *Trends in Molecular Medicine, 18*(8), 472–482.

Smedley, D., Haider, S., Ballester, B., Holland, R., London, D., Thorisson, G., & Kasprzyk, A. (2009). BioMart--biological queries made easy. *BMC Genomics, 10*, 22.

Smit, A. F. A., Hubley, R. & Green, P. (1996-2010). RepeatMasker Open-3.0. https://www.repeatmasker.org

Sonney, S., Leipzig, J., Lott, M. T., Zhang, S., Procaccio, V., Wallace, D. C., & Sondheimer, N. (2017). Predicting the pathogenicity of novel variants in mitochondrial tRNA with MitoTIP. *PLoS Computational Biology, 13*(12), e1005867.

Stein, D., Kars, M. E., Wu, Y., Bayrak, Ç. S., Stenson, P. D., Cooper, D. N., Schlessinger, A., & Itan, Y. (2023). Genome-wide prediction of pathogenic gain- and loss-of-function variants from ensemble learning of a diverse feature set. *Genome Medicine, 15*(1), Article 103.

Suiter, C. C., Moriyama, T., Matreyek, K. A., Yang, W., Scaletti, E. R., Nishii, R., Yang, W., Hoshitsuki, K., Singh, M., Trehan, A., Parish, C., Smith, C., Li, L., Bhojwani, D., Yuen, L. Y. P., Li, C. K., Li, C. H., Yang, Y. L., Walker, G. J., Goodhand, J. R., … Yang, J. J. (2020). Massively parallel variant characterization identifies NUDT15 alleles associated with thiopurine toxicity. *Proceedings of the National Academy of Sciences of the United States of America, 117*(10), 5394–5401.

Takahashi, M., Shimodaira, H., Andreutti-Zaugg, C., Iggo, R., Kolodner, R. D., & Ishioka, C. (2007). Functional analysis of human MLH1 variants using yeast and in vitro mismatch repair assays. *Cancer Research, 67*(10), 4595–4604.

Tavtigian, S. V., Greenblatt, M. S., Harrison, S. M., Nussbaum, R. L., Prabhu, S. A., Boucher, K. M., Biesecker, L. G., & ClinGen Sequence Variant Interpretation Working Group (ClinGen SVI) (2018). Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. *Genetics in Medicine, 20*(9), 1054–1060.

Tavtigian, S. V., Harrison, S. M., Boucher, K. M., & Biesecker, L. G. (2020). Fitting a naturally scaled point system to the ACMG/AMP variant classification guidelines. *Human Mutation, 41*(10), 1734–1737.

Tian, Y., Pesaran, T., Chamberlin, A., Fenwick, R. B., Li, S., Gau, C. L., Chao, E. C., Lu, H. M., Black, M. H., & Qian, D. (2019). REVEL and BayesDel outperform other in silico meta-predictors for clinical variant classification. *Scientific Reports, 9*(1), 12752.

Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., … Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers, 1*(1).

UniProt Consortium (2023). UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research, 51*(D1), D523–D531.Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., … DePristo, M. A. (2013). From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Current Protocols in Bioinformatics*, 11.10.1–11.10.33.

van Dijk, E. L., Auger, H., Jaszczyszyn, Y., & Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in Genetics, 30*(9), 418–426.

van Loggerenberg, W., Sowlati-Hashjin, S., Weile, J., Hamilton, R., Chawla, A., Sheykhkarimli, D., Gebbia, M., Kishore, N., Frésard, L., Mustajoki, S., Pischik, E., Di Pierro, E., Barbaro, M., Floderus, Y., Schmitt, C., Gouya, L., Colavin, A., Nussbaum, R., Friesema, E. C. H., Kauppinen, R., … Roth, F. P. (2023). Systematically testing human HMBS missense variants to reveal mechanism and pathogenic variation. *American Journal of Human Genetics, 110*(10), 1769–1786.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., … Zhu, X. (2001). The sequence of the human genome. *Science, 291*(5507), 1304–1351.

Walker, L. C., Hoya, M., Wiggins, G. A. R., Lindy, A., Vincent, L. M., Parsons, M. T., Canson, D. M., Bis-Brewer, D., Cass, A., Tchourbanov, A., Zimmermann, H., Byrne, A. B., Pesaran, T., Karam, R., Harrison, S. M., Spurdle, A. B., & ClinGen Sequence Variant Interpretation Working Group (2023). Using the ACMG/AMP framework to capture evidence related to predicted and observed

impact on splicing: Recommendations from the ClinGen SVI Splicing Subgroup. *American Journal of Human Genetics, 110*(7), 1046–1067.

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research, 38*(16), e164.

Wang, Y., Zhao, Y., Bollas, A., Wang, Y., & Au, K. F. (2021). Nanopore sequencing technology, bioinformatics and applications. *Nature Biotechnology, 39*(11), 1348–1365.

Weston, K. P., Gao, X., Zhao, J., Kim, K. S., Maloney, S. E., Gotoff, J., Parikh, S., Leu, Y. C., Wu, K. P., Shinawi, M., Steimel, J. P., Harrison, J. S., & Yi, J. J. (2021). Identification of disease-linked hyperactivating mutations in UBE3A through large-scale functional variant analysis. *Nature Communications, 12*(1), 6809.

Whiffin, N., Minikel, E., Walsh, R., O'Donnell-Luria, A. H., Karczewski, K., Ing, A. Y., Barton, P. J. R., Funke, B., Cook, S. A., MacArthur, D., & Ware, J. S. (2017). Using high-resolution variant frequencies to empower clinical genome interpretation. *Genetics in Medicine, 19*(10), 1151–1158.

Whiffin, N., Karczewski, K. J., Zhang, X., Chothani, S., Smith, M. J., Evans, D. G., Roberts, A. M., Quaife, N. M., Schafer, S., Rackham, O., Alföldi, J., O'Donnell-Luria, A. H., Francioli, L. C., Genome Aggregation Database Production Team, Genome Aggregation Database Consortium, Cook, S. A., Barton, P. J. R., MacArthur, D. G., & Ware, J. S. (2020). Characterising the loss-of-function impact of 5' untranslated region variants in 15,708 individuals. *Nature Communications, 11*(1), 2523.

Xiang, J., Peng, J., Baxter, S., & Peng, Z. (2020). AutoPVS1: An automatic classification tool for PVS1 interpretation of null variants. *Human Mutation, 41*(9), 1488–1498.

Yates, A., Beal, K., Keenan, S., McLaren, W., Pignatelli, M., Ritchie, G. R., Ruffier, M., Taylor, K., Vullo, A., & Flicek, P. (2015). The Ensembl REST API: Ensembl Data for Any Language. *Bioinformatics, 31*(1), 143–145.

Zastrow, D. B., Baudet, H., Shen, W., Thomas, A., Si, Y., Weaver, M. A., Lager, A. M., Liu, J., Mangels, R., Dwight, S. S., Wright, M. W., Dobrowolski, S. F., Eilbeck, K., Enns, G. M., Feigenbaum, A., Lichter-Konecki, U., Lyon, E., Pasquali, M., Watson, M., Blau, N., … ClinGen Inborn Errors of Metabolism Working Group (2018). Unique aspects of sequence variant interpretation for inborn errors of metabolism (IEM): The ClinGen IEM Working Group and the Phenylalanine Hydroxylase Gene. *Human Mutation, 39*(11), 1569–1580.

Zhang, W., Xie, T., Sheng, H., Shao, Y., Lin, Y., Jiang, M., Xu, A., Su, X., Liu, Z., Zhao, X., Liu, L., & Huang, Y. (2019). Genetic analysis of 63 Chinese patients with mucopolysaccharidosis type II: Functional characterization of seven novel IDS variants. *Clinica Chimica Acta, 491*, 114–120.

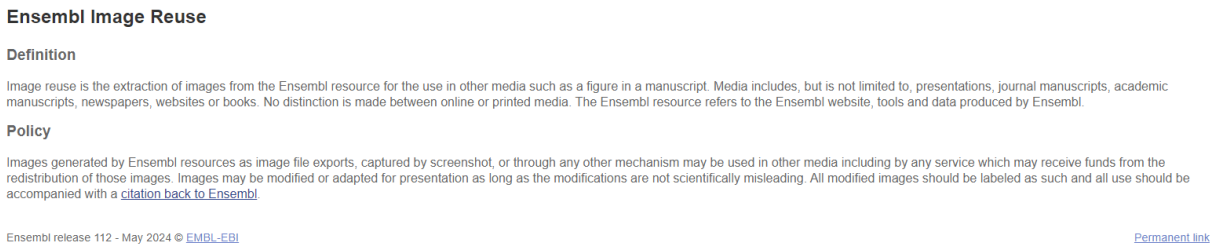
Zhang, X., Walsh, R., Whiffin, N., Buchan, R., Midwinter, W., Wilk, A., Govind, R., Li, N., Ahmad, M., Mazzarotto, F., Roberts, A., Theotokis, P. I., Mazaika, E., Allouba, M., de Marvao, A., Pua, C. J., Day, S. M., Ashley, E., Colan, S. D., Michels, M., … Ware, J. S. (2021). Disease-specific variant pathogenicity prediction significantly improves variant interpretation in inherited cardiac conditions. *Genetics in Medicine, 23*(1), 69–79.

Zheng, H., Yan, X., Li, G., Lin, H., Deng, S., Zhuang, W., Yao, F., Lu, Y., Xia, X., Yuan, H., Jin, L., & Yan, Z. (2022). Proactive functional classification of all possible missense single-nucleotide variants in KCNQ4. *Genome Research, 32*(8), 1573–1584.

Copyright Permissions

Figure 1: Harrison, P. W., Amode, M. R., Austine-Orimoloye, O., Azov, A. G., Barba, M., Barnes, I., Becker, A., Bennett, R., Berry, A., Bhai, J., Bhurji, S. K., Boddu, S., Branco Lins, P. R., Brooks, L., Ramaraju, S. B., Campbell, L. I., Martinez, M. C., Charkhchi, M., Chougule, K., Cockburn, A., … Yates, A. D. (2024). Ensembl 2024. *Nucleic Acids Research, 52*(D1), D891–D899.

Ensembl Image Reuse Policy:



Ensembl Image Reuse

Definition

Image reuse is the extraction of images from the Ensembl resource for the use in other media such as a figure in a manuscript. Media includes, but is not limited to, presentations, journal manuscripts, academic manuscripts, newspapers, websites or books. No distinction is made between online or printed media. The Ensembl resource refers to the Ensembl website, tools and data produced by Ensembl.

Policy

Images generated by Ensembl resources as image file exports, captured by screenshot, or through any other mechanism may be used in other media including by any service which may receive funds from the redistribution of those images. Images may be modified or adapted for presentation as long as the modifications are not scientifically misleading. All modified images should be labeled as such and all use should be accompanied with a citation back to Ensembl.

Ensembl release 112 - May 2024 © EMBL-EBI

Permanent link

Figure 2: Brnich, S. E., Rivera-Muñoz, E. A., & Berg, J. S. (2018). Quantifying the potential of functional evidence to reclassify variants of uncertain significance in the categorical and Bayesian interpretation frameworks. *Human Mutation, 39*(11), 1531–1541.

Figure 5: Walker, L. C., Hoya, M., Wiggins, G. A. R., Lindy, A., Vincent, L. M., Parsons, M. T., Canson, D. M., Bis-Brewer, D., Cass, A., Tchourbanov, A., Zimmermann, H., Byrne, A. B., Pesaran, T., Karam, R., Harrison, S. M., Spurdle, A. B., & ClinGen Sequence Variant Interpretation Working Group (2023). Using the ACMG/AMP framework to capture evidence related to predicted and ob-served impact on splicing: Recommendations from the ClinGen SVI Splicing Subgroup. American Journal of Human Genetics, 110(7), 1046–1067.

Figure 8: Rehm, H. (2023). Research on testing phase: what tests to recommend, how are they offered and what types of results are provided [PDF]. https://www.genome.gov/sites/default/files/media/files/2023-11/Session_7-2__Rehm.pdf

Human Mutation

Permissions

Sign up for email alerts

Enter your email to receive alerts when new articles and issues are published.

Email address*

Enter email

Continue

Submit an article

Journal Metrics

ELSEVIER LICENSE
TERMS AND CONDITIONS

Jul 03, 2024

This Agreement between Arda İnan ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

License Number	5821380378462
License date	Jul 03, 2024
Licensed Content Publisher	Elsevier
Licensed Content Publication	The American Journal of Human Genetics
Licensed Content Title	Using the ACMG/AMP framework to capture evidence related to predicted and observed impact on splicing: Recommendations from the ClinGen SVI Splicing Subgroup
Licensed Content Author	Logan C. Walker,Miguel de la Hoya,George A.R. Wiggins,Amanda Lindy,Lisa M. Vincent,Michael T. Parsons,Daffodil M. Canson,Dana Bis-Brewer,Ashley Cass,Alexander Tchourbanov,Heather Zimmermann,Alicia B. Byrne,Tina Pesaran,Rachid Karam et al.
Licensed Content Date	Jul 6, 2023
Licensed Content Volume	110
Licensed Content Issue	7
Licensed Content Pages	22
Start Page	1046

End Page	1067
Type of Use	reuse in a thesis/dissertation
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
Format	both print and electronic
Are you the author of this Elsevier article?	No
Will you be translating?	No
Title of new work	An Automated Sequence Variant Classification Tool for DNA Diagnostics
Institution name	Bilkent University
Expected presentation date	Jul 2024
Portions	Figure 1
The Requesting Person / Organization to Appear on the License	Arda İnan
Requestor Location	Bilkent University Bilkent University Dept of Molecular Biology & Genetics Bilkent Ankara 06800 TURKEY Ankara, 06800 Turkey Attn: Bilkent University
Publisher Tax ID	GB 494 6272 12
Total	0.00 USD

Terms and Conditions

INTRODUCTION

1. The publisher for this copyrighted material is Elsevier. By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your RightsLink account and that are available at any time at https://myaccount.copyright.com).

GENERAL TERMS

2. Elsevier hereby grants you permission to reproduce the aforementioned material subject to the terms and conditions indicated.

3. Acknowledgement: If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source, permission must also be sought from that source. If such permission is not obtained then that material may not be included in your publication/copies. Suitable acknowledgement to the source must be made, either as a footnote or in a reference list at the end of your publication, as follows:

"Reprinted from Publication title, Vol /edition number, Author(s), Title of article / title of chapter, Pages No., Copyright (Year), with permission from Elsevier [OR APPLICABLE SOCIETY COPYRIGHT OWNER]." Also Lancet special credit - "Reprinted from The Lancet, Vol. number, Author(s), Title of article, Pages No., Copyright (Year), with permission from Elsevier."

4. Reproduction of this material is confined to the purpose and/or media for which permission is hereby given. The material may not be reproduced or used in any other way, including use in combination with an artificial intelligence tool (including to train an algorithm, test, process, analyse, generate output and/or develop any form of artificial intelligence tool), or to create any derivative work and/or service (including resulting from the use of artificial intelligence tools).

5. Altering/Modifying Material: Not Permitted. However figures and illustrations may be altered/adapted minimally to serve your work. Any other abbreviations, additions, deletions and/or any other alterations shall be made only with prior written authorization of Elsevier Ltd. (Please contact Elsevier's permissions helpdesk here). No modifications can be made to any Lancet figures/tables and they must be reproduced in full.

6. If the permission fee for the requested use of our material is waived in this instance, please be advised that your future requests for Elsevier materials may attract a fee.

7. Reservation of Rights: Publisher reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

8. License Contingent Upon Payment: While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by publisher or by CCC) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never

granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and publisher reserves the right to take any and all action to protect its copyright in the materials.

9. Warranties: Publisher makes no representations or warranties with respect to the licensed material.

10. Indemnity: You hereby indemnify and agree to hold harmless publisher and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.

11. No Transfer of License: This license is personal to you and may not be sublicensed, assigned, or transferred by you to any other person without publisher's written permission.

12. No Amendment Except in Writing: This license may not be amended except in a writing signed by both parties (or, in the case of publisher, by CCC on publisher's behalf).

13. Objection to Contrary Terms: Publisher hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC's Billing and Payment terms and conditions. These terms and conditions, together with CCC's Billing and Payment terms and conditions (which are incorporated herein), comprise the entire agreement between you and publisher (and CCC) concerning this licensing transaction. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall control.

14. Revocation: Elsevier or Copyright Clearance Center may deny the permissions described in this License at their sole discretion, for any reason or no reason, with a full refund payable to you. Notice of such denial will be made using the contact information provided by you. Failure to receive such notice will not alter or invalidate the denial. In no event will Elsevier or Copyright Clearance Center be responsible or liable for any costs, expenses or damage incurred by you as a result of a denial of your permission request, other than a refund of the amount(s) paid by you to Elsevier and/or Copyright Clearance Center for denied permissions.

LIMITED LICENSE

The following terms and conditions apply only to specific license types:

15. **Translation**: This permission is granted for non-exclusive world **English** rights only unless your license was granted for translation rights. If you licensed translation rights you may only translate this content into the languages you requested. A professional translator must perform all translations and reproduce the content word for word preserving the integrity of the article.

16. **Posting licensed content on any Website**: The following terms and conditions apply as follows: Licensing material from an Elsevier journal: All content posted to the web site must maintain the copyright information line on the bottom of each image; A hyper-text must be included to the Homepage of the journal from which you are licensing at http://www.sciencedirect.com/science/journal/xxxxx or the Elsevier homepage for books at http://www.elsevier.com; Central Storage: This license does not include permission for a scanned version of the material to be stored in a central repository such as that provided by Heron/XanEdu.

Licensing material from an Elsevier book: A hyper-text link must be included to the Elsevier homepage at http://www.elsevier.com . All content posted to the web site must maintain the copyright information line on the bottom of each image.

Posting licensed content on Electronic reserve: In addition to the above the following clauses are applicable: The web site must be password-protected and made available only to bona fide students registered on a relevant course. This permission is granted for 1 year only. You may obtain a new license for future website posting.

17. **For journal authors:** the following clauses are applicable in addition to the above:

Preprints:

A preprint is an author's own write-up of research results and analysis, it has not been peer-reviewed, nor has it had any other value added to it by a publisher (such as formatting, copyright, technical enhancement etc.).

Authors can share their preprints anywhere at any time. Preprints should not be added to or enhanced in any way in order to appear more like, or to substitute for, the final versions of articles however authors can update their preprints on arXiv or RePEc with their Accepted Author Manuscript (see below).

If accepted for publication, we encourage authors to link from the preprint to their formal publication via its DOI. Millions of researchers have access to the formal publications on ScienceDirect, and so links will help users to find, access, cite and use the best available version. Please note that Cell Press, The Lancet and some society-owned have different preprint policies. Information on these policies is available on the journal homepage.

Accepted Author Manuscripts: An accepted author manuscript is the manuscript of an article that has been accepted for publication and which typically includes author-incorporated changes suggested during submission, peer review and editor-author communications.

Authors can share their accepted author manuscript:

- immediately
  - via their non-commercial person homepage or blog
  - by updating a preprint in arXiv or RePEc with the accepted manuscript
  - via their research institute or institutional repository for internal institutional uses or as part of an invitation-only research collaboration work-group
  - directly by providing copies to their students or to research collaborators for their personal use
  - for private scholarly sharing as part of an invitation-only work group on commercial sites with which Elsevier has an agreement
- After the embargo period
  - via non-commercial hosting platforms such as their institutional repository
  - via commercial sites with which Elsevier has an agreement

In all cases accepted manuscripts should:

- link to the formal publication via its DOI
- bear a CC-BY-NC-ND license - this is easy to do
- if aggregated with other manuscripts, for example in a repository or other site, be shared in alignment with our hosting policy not be added to or enhanced in any way to appear more like, or to substitute for, the published journal article.

Published journal article (JPA): A published journal article (PJA) is the definitive final record of published research that appears or will appear in the journal and embodies all

value-adding publishing activities including peer review co-ordination, copy-editing, formatting, (if relevant) pagination and online enrichment.

Policies for sharing publishing journal articles differ for subscription and gold open access articles:

Subscription Articles: If you are an author, please share a link to your article rather than the full-text. Millions of researchers have access to the formal publications on ScienceDirect, and so links will help your users to find, access, cite, and use the best available version.

Theses and dissertations which contain embedded PJAs as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.

If you are affiliated with a library that subscribes to ScienceDirect you have additional private sharing rights for others' research accessed under that agreement. This includes use for classroom teaching and internal training at the institution (including use in course packs and courseware programs), and inclusion of the article for grant funding purposes.

Gold Open Access Articles: May be shared according to the author-selected end-user license and should contain a CrossMark logo, the end user license, and a DOI link to the formal publication on ScienceDirect.

Please refer to Elsevier's posting policy for further information.

18. **For book authors** the following clauses are applicable in addition to the above: Authors are permitted to place a brief summary of their work online only. You are not allowed to download and post the published electronic version of your chapter, nor may you scan the printed edition to create an electronic version. **Posting to a repository:** Authors are permitted to post a summary of their chapter only in their institution's repository.

19. **Thesis/Dissertation**: If your license is for use in a thesis/dissertation your thesis may be submitted to your institution in either print or electronic form. Should your thesis be published commercially, please reapply for permission. These requirements include permission for the Library and Archives of Canada to supply single copies, on demand, of the complete thesis and include permission for Proquest/UMI to supply single copies, on demand, of the complete thesis. Should your thesis be published commercially, please reapply for permission. Theses and dissertations which contain embedded PJAs as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.

Elsevier Open Access Terms and Conditions

You can publish open access with Elsevier in hundreds of open access journals or in nearly 2000 established subscription journals that support open access publishing. Permitted third party re-use of these open access articles is defined by the author's choice of Creative Commons user license. See our open access license policy for more information.

Terms & Conditions applicable to all Open Access articles published with Elsevier:

Any reuse of the article must not represent the author as endorsing the adaptation of the article nor should the article be modified in such a way as to damage the author's honour or reputation. If any changes have been made, such changes must be clearly indicated.

The author(s) must be appropriately credited and we ask that you include the end user license and a DOI link to the formal publication on ScienceDirect.

If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source it is the responsibility of the user to ensure their reuse complies with the terms and conditions determined by the rights holder.

Additional Terms & Conditions applicable to each Creative Commons user license:

CC BY: The CC-BY license allows users to copy, to create extracts, abstracts and new works from the Article, to alter and revise the Article and to make commercial use of the Article (including reuse and/or resale of the Article by commercial entities), provided the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, indicates if changes were made and the licensor is not represented as endorsing the use made of the work. The full details of the license are available at http://creativecommons.org/licenses/by/4.0.

CC BY NC SA: The CC BY-NC-SA license allows users to copy, to create extracts, abstracts and new works from the Article, to alter and revise the Article, provided this is not done for commercial purposes, and that the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, indicates if changes were made and the licensor is not represented as endorsing the use made of the work. Further, any new works must be made available on the same conditions. The full details of the license are available at http://creativecommons.org/licenses/by-nc-sa/4.0.

CC BY NC ND: The CC BY-NC-ND license allows users to copy and distribute the Article, provided this is not done for commercial purposes and further does not permit distribution of the Article if it is changed or edited in any way, and provided the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, and that the licensor is not represented as endorsing the use made of the work. The full details of the license are available at http://creativecommons.org/licenses/by-nc-nd/4.0. Any commercial reuse of Open Access articles published with a CC BY NC SA or CC BY NC ND license requires permission from Elsevier and will be subject to a fee.

Commercial reuse includes:

- Associating advertising with the full text of the Article
- Charging fees for document delivery or access
- Article aggregation
- Systematic distribution via e-mail lists or share buttons

Posting or linking by commercial companies for use by customers of those companies.

20. **Other Conditions**:

v1.10

Questions? customercare@copyright.com.

RE: Permission to use figure

From <hrehm@broadinstitute.org>
To 'Arda İnan' <arda.inan@bilkent.edu.tr>
Date 2024-07-04 01:16

Yes, that's fine. Good luck with your defense.
Heidi

Heidi L. Rehm, PhD, FACMG
Director, Genomic Medicine Unit, Center for Genomic Medicine, MGH
Chief Genomics Officer, Department of Medicine, MGH
Co-Director, Medical and Population Genetics, Broad Institute
Medical and Clinical Lab Director, Broad Clinical Labs
Professor of Pathology, Harvard Medical School
hrehm@mgh.harvard.edu
Admin: Samie Wong tggadmin@broadinstitute.org
Lab Website: The-TGG.org

**I often send emails outside of traditional business hours because it works with my schedule. Please do not feel that a response is expected outside of your regular work schedule.

-----Original Message-----
From: Arda İnan <arda.inan@bilkent.edu.tr>
Sent: Wednesday, July 3, 2024 7:23 AM
To: hrehm@broadinstitute.org
Subject: Permission to use figure

Dear Prof. Rehm,

I hope this e-mail finds you well.

I will defend my master's thesis in the coming weeks. I would like permission to use a figure you used in one of your presentations. To my knowledge, it has not been published anywhere yet. If you permit, I plan to include it with no changes (including acknowledgements).

You can find the link to the presentation:
https://www.genome.gov/sites/default/files/media/files/2023-11/Session_7-2__Rehm.pdf

I plan to use the figure in the slide 5 with title "VUS tiering correlates with reclassification".

Sincerely,
Arda İnan
--
Arda İnan, BSc
Özçelik Lab
Bilkent University
Dept of Molecular Biology & Genetics
Bilkent Ankara 06800 TURKEY