ACIBADEM MEHMET ALI AYDINLAR UNIVERSITY
INSTITUTE OF HEALTH SCIENCES

MULTIVARIATE ANALYSIS OF GENOMIC IN-SILICO PATHOGENICITY PREDICTORS

EYLÜL AYDIN

M.Sc. THESIS

DEPARTMENT OF GENOME STUDIES

SUPERVISOR
Prof. Özden Hatırnaz Ng

SECONDARY SUPERVISOR
Assist. Prof. Özkan Özdemir

ISTANBUL-2024

ACIBADEM MEHMET ALI AYDINLAR UNIVERSITY
INSTITUTE OF HEALTH SCIENCES

MULTIVARIATE ANALYSIS OF GENOMIC IN-SILICO PATHOGENICITY PREDICTORS

EYLÜL AYDIN

M.Sc. THESIS

DEPARTMENT OF GENOME STUDIES

SUPERVISOR
Prof. Özden Hatırnaz Ng

SECONDARY SUPERVISOR
Assist. Prof. Özkan Özdemir

ISTANBUL-2024

Department: Genome Studies
Program: Genome Studies
Thesis Title: Multivariate Analysis of Genomic in-silico Pathogenicity Predictors
Student's name and Surname: Eylül Aydın
Date of Defence: 27/05/2024

This is to certify that I have examined this copy of the master thesis. I confirm that she/he prepared after fulfilling the specified requirements in the associated legislations before the final examining committee whose signatures are below.

Jury Member (Supervisor of the thesis)	Title, Name Surname, University	Signature
Jury Member (Secondary supervisor of the thesis)	Title, Name Surname, University	Signature
Jury Member	Title, Name Surname University	Signature
Jury Member	Title, Name Surname University	Signature
Jury Member	Title, Name Surname University	Signature

DECLARATION

I declare that this thesis work is my own work, I had no unethical behavior at any stages from the planning to the writing of the thesis, I obtained all the information in this thesis in accordance with academic and ethicel rules, I cited all the information and comments that were not obtained with this thesis work, and I provided resources in the list of references. I also declare that there was no violation of any patents and copyrights during the study and writing of this thesis.

27/05/2024

Eylül Aydın

PREFACE AND ACKNOWLEDGEMENT

This journey has been both challenging and enlightening, teaching me valuable lessons to make use of in academia, and life. A few thanks and gratitude are in order, to the people who have been supporting me throughout this journey.

First and foremost, I would like to express my deepest gratitude to my supervisors Prof. Özden Hatırnaz Ng and Özkan Özdemir, Ph.D., for their invaluable guidance. I would like to add a special thanks to my supervisor, Özkan. Thank you for your support over the past two years. Our journey, filled with its fair share of challenges and triumphs, has been instrumental in my growth. Your unique blend of annoying persistence and unwavering belief in my potential has pushed me to excel. And Özden Hocam, thank you for your lectures on "Communication with Özkan Özdemir-101".

I am also grateful to my colleague and friend, Berk Ergun, whose discussions, and perspectives have greatly enriched my work. A special thanks goes to the members of "accumulation of deadlines". I also would like to thank ACURARE for financial support, and Prof. Dr. Uğur Özbek, for the encouragement when I had the chance to present my work abroad. In addition, I would like to thank Prof. Dr. Yasemin Alanay and Assoc. Prof. Özlem Akgün Doğan for their invaluable contributions and all the support, and all of my professors in Genome Studies.

I would like to thank Asst. Prof. Ezgi Karaca, as I do every chance I get, for all her support and encouragement that enabled me to enter the computational world during the pandemic while I was an undergraduate. That was a milestone for me and I will forever be grateful for the opportunity.

I thank Sena and her family for opening their home to me in such a critical period, giving me a place to live. Last but not least, I would like to thank my dearest family, my mother, my father, and my favorite brother. My mother, especially, has been listening to me complain through this journey and deserves her name written on the cover. Thank you for your unconditional love and support.

TABLE OF CONTENTS

DECLARATION iii

PREFACE AND ACKNOWLEDGEMENT iv

TABLE OF CONTENTS v

LIST OF SYMBOLS AND ABBREVIATIONS vii

LIST OF FIGURES ix

LIST OF TABLES x

ÖZET 1

ABSTRACT 2

1 INTRODUCTION AND AIM 3

2 BACKGROUND 6

2.1 Human Genome 6
2.1.1 Human Genome Project 8
2.1.2 Complete Assembly of the Human Genome 9
2.1.3 Reference Genome 9
2.2 Genetic Variation Types 10
2.2.1 Variations in the coding regions 10
2.2.1.1 Synonymous variants 10
2.2.1.2 Non-synonymous variants 11
2.2.2 Variations in the non-coding regions 12
2.2.3 Other variations 12
2.3 Sequencing Technologies 13
2.3.1 Sequencing Technologies 13
2.3.2 Short-read sequencing 14
2.3.3 Long-read sequencing 15
2.4 Handling Sequencing Data 16
2.4.1 Raw data processing in NGS 16
2.4.2 Variant annotation 18
2.4.2.1 Functional annotation 19
2.4.2.2 Population allele frequency-based annotation 19
2.4.2.3 In silico pathogenicity predictors 20
2.4.2.4 Clinical significance 25
2.5 Variant Interpretation and Prioritization 26
2.5.1 ACMG/AMP standards and guidelines for Mendelian disorders 27
2.6 Variants with Uncertain Significance 28
2.7 Conflicting Computational Prediction in Variant Interpretation 29
2.8 ClinVar 30

2.9 dbNSFP 32

3 MATERIALS AND METHODS 33

3.1 Material 33
3.1.1 In silico pathogenicity predictors (ISPPs) considered 33
3.1.2 Dataset curation 35
3.1.2.1 The dbNSFP database 35
3.1.2.2 Canonical transcript mapping 36
3.1.2.3 The up-to-date ClinVar variant summary data 37
3.1.2.4 Final variant dataset 37
3.2 Methods 40
3.2.1 VAMPP-score calculation 40
3.2.1.1 Weighted gene coefficient (*M*) 40
3.2.2 Section 1. Statistical framework 41
3.2.3 Section 2. Variant group distribution metrics 42
3.2.3.1 Proximity (*d*) 42
3.2.3.2 Mean difference 44
3.2.4 VAMPP-score annotation 45
3.2.5 Multidimensional clustering and correlation analyses 45

4 RESULTS 47

4.1 Overview 47
4.2 Multidimensional Analysis 50
4.2.1 ISPP significance on Mendeliome 50
4.2.2 Scoring correlation on Mendeliome 55
4.2.3 VAMPP-score distribution 60

5 DISCUSSION 63

6 CONCLUSION 71

7 REFERENCES 73

8 CURRICULUM VITAE 81

LIST OF SYMBOLS AND ABBREVIATIONS

ACMG	American College of Medical Genetics and Genomics
AF	Allele frequency
AMP	Association for Molecular Pathology
B	Benign
D	Proximity
DBMS	Database Management System
DBNSFP	Database for Nonsynonymous SNPs' Functional Prediction
DMAX	Maximum distance
DNA	Deoxyribonucleic acid
ENSEMBL	Ensembl genome database project
ESM1B	Evolutionary Scale Modeling - 1b
FATHMM	Functional Analysis through Hidden Markov Models
FATHMM-MKL	FATHMM with multiple kernel learning
FATHMM-XF	FATHMM with eXtended Features
FITCONS	Fitness consequences score
470WAY-M	470 mammalian species
GM12878	Human lymphoblastoid cells
GRCH	Genome Reference Consortium
H1-HESC	Human embryonic stem cells
HGP	Human Genome Project
HPO	Human Phenotype Ontology
HUVEC	Human umbilical vein epithelial cells
100WAY-V	100 vertebrate species
ISPP	In-silico pathogenicity predictor
LB	Likely benign
LP	Likely pathogenic
LR	Logistic regression
M	Weighted gene coefficient
MSA	Multiple Sequence Alignment
NGS	Next Generation Sequencing

nsSNV	Nonsynonymous single nucleotide variation
OLAP	On-Line Analytical Processing
P	Pathogenic
PCA	Principal Component Analysis
PEF	Proximity Estimation Function
P-VALUE	Probability value
RNN	Recurrent Neural Network
17WAY-P	17 primate species
SNV	Single Nucleotide Variation
SQL	Structured Query Language
SVM	Support vector machine
U	Unknown
VAMPP	Variant Analysis with Multivariate Pathogenicity Prediction
VARITY-ER	VARITY for Extremely Rare
VARITY-ER-LOO	VARITY for Extremely Rare-Leave One Out
VARITY-R	VARITY for Rare
VARITY-R-LOO	VARITY for Rare-Leave One Out
VUS	Variant with Uncertain Significance

LIST OF FIGURES

Figure 1. DNA structure. 6

Figure 2. Gene structure and central dogma. 7

Figure 3. History of the human genome. 8

Figure 4. Illustration of variant types. 11

Figure 5. Sequencing technologies. 14

Figure 6. The workflow of NGS data processing. 17

Figure 7. ACMG variant classification evidences 28

Figure 8. ISPP categorization. 34

Figure 9. Schematic representation of the database structure. 35

Figure 10. Schematic representation of the list of variants used as input data in the analysis. 39

Figure 11. Function for the weighted gene coefficient (*M*). 40

Figure 12. Ideal variant group distribution. 43

Figure 13. The calculation of the elements of the proximity estimation function (PEF). 43

Figure 14. Proximity estimation function (PEF). 44

Figure 15. Function for mean difference. 44

Figure 16. Function for VAMPP-score. 45

Figure 17. Significant ISPPs throughout the exome and gene counts. 48

Figure 18. Principal component analysis on ISPPs for Mendeliome. 50

Figure 19. The ISPP correlation heatmap of multiple comparison p-values for the Mendeliome. 52

Figure 20. ISPP similarity matrix for the Mendeliome. 54

Figure 21. ISPP similarity matrix for benign variant assessment. 56

Figure 22. ISPP similarity matrix for pathogenic variant assessment. 58

Figure 23. VAMPP-score distribution. 60

LIST OF TABLES

Table 1. In-silico pathogenicity predictors (ISPP) table.. 20

Table 2. The review status and star assignment for ClinVar variants..... 31

ÖZET

Genomik *in-siliko* Patojenite Tahmin Araçlarının Çok Değişkenli Analizi

Yanlış anlamlı *(missense)* varyantların klinik olarak yorumlanması klinik tanı sürecinde kritik öneme sahiptir. Varyant patojenitesinin değerlendirilmesi için birçok yöntem ve algoritma geliştirilmiş olsa da bu araçların performansları sınırlıdır ve hatalı ve/veya çelişkili tahmin yapabilirler. Varyantların patojenitesinin belirli tahmin araçları dikkate alınarak değerlendirilmesi yanlış pozitif ve negatif sonuçlara yol açabilir ve bu nedenle bu konuda bir standardizasyon gereksinimi vardır. Gün geçtikte veri tabanlarında biriken veriler sayesinde ve varyantların patojenitesi hakkında daha fazla bilgi sahibi oldukça, varyant sınıflandırmasını daha hassas hale getirmek için doğru sonuçlar veren algoritmaları istatistiksel yaklaşımlarla tanımlamak mümkün hale gelmektedir. Bu tez kapsamında, *missense* varyantlar için istatistik temelli bir akış kullanılarak tasarlanan yeni bir *in-siliko* patojenite tahmin (ISPT) skoru olan VAMPP-skor geliştirilmiştir. Skor, literatürde en sık kullanılan ISPT'lerin çoklu ve ikili karşılaştırmalarının yanı sıra ISPT'lerin varyant grupları arasındaki ayrım gücüne ve ortalama fark hesaplamalarına dayanmaktadır. Bu kapsamda, ISPT skorlarının tahmin doğruluklarına göre en iyi gen-ISPT skoru eşleşmeleri kullanılmış ve kombinatoryal ağırlıklı bir puan olan VAMPP-skor geliştirilmiştir. VAMPP-skor, klinikte *missense* varyantların yorumlanmasında önemli bir ilerleme sağlayabileceğine inanılmaktadır. İstatistiksel temeli göz önüne alındığında, VAMPP-skor bir varyantın patojenik olup olmadığına ilişkin karar vermede belirleyici bir rol oynayabilir. Özellikle, nadir ve tanısız vakaların yeniden değerlendirilmesinde genetik tanının doğruluğunu artırma ve tanı sürecini iyileştirme potansiyeline sahip olduğu düşünülmektedir.

Anahtar Sözcükler: varyant yorumlama ve önceliklendirme, ACMG/AMP önerileri, *in-siliko* patojenite tahmin araçları, nadir ve tanısız vakalar, klinik önemi belirsiz varyantlar

ABSTRACT

Multivariate Analysis of Genomic in-silico Pathogenicity Predictors

Clinical interpretation of the missense variants is critically important in diagnostics. While many methods and algorithms for pathogenicity assessment have been developed, their performances are limited, and their predictions can be either inaccurate or conflicting. Evaluating the pathogenicity variants based on certain predictors may lead to false positives and negatives and therefore standardization is needed. As data accumulates in databases and we gather more information about the pathogenicity of the known variants, it becomes possible to identify algorithms that yield accurate results by statistical approaches to make variant classification more precise. This thesis introduces VAMPP-score, a novel in-silico pathogenicity prediction (ISPP) score for missense variants developed using a statistical framework. This score is based on multiple and pairwise comparison tests, as well as proximity and mean difference calculations among variant groups. It utilizes the best gene-ISPP matches, according to ISPP accuracies for each gene, offering a combinatorial weighted score. VAMPP-score introduces an advancement in missense variant interpretation. Given the statistical basis, it can be a decision-maker for a variant's pathogenicity. Its application, particularly in the reanalysis of rare and undiagnosed cases, promises to enhance the accuracy of genetic diagnostics and improve patient outcomes.

Keywords: variant interpretation and prioritization, ACMG/AMP recommendations, in-silico pathogenicity predictors, rare and undiagnosed cases, variants of uncertain significance

1 INTRODUCTION AND AIM

Due to advances in sequencing technologies, the field of genetic diagnosis has undergone a major transformation. Next-generation sequencing (NGS) has improved our ability to understand and analyze genetic information, offering a more comprehensive, rapid, and cost-effective approach in clinics. The significant reduction in sequencing costs has democratized access to genetic testing, making it a more viable option for a broader range of patients (1).

Mendelian disorders, characterized by their inheritance patterns, represent a significant portion of genetic diseases diagnosed with NGS (2, 3). These disorders, often caused by single-gene disruptions, have been pivotal in understanding the genetic basis of diseases. NGS has enabled the identification of novel variations in known disease-causing genes and the discovery of new genes associated (3). However, the complexity of genetic interactions and the presence of phenotypic heterogeneity even within the same disorder pose challenges in diagnosis and management (4). In addition, despite the developments in NGS, a considerable number of genetic disorders remain undiagnosed. This is particularly true for rare and undiagnosed diseases, where the genetic basis is often not well understood (2, 5).

The challenge is further compounded by one of the most encountered variations in the genome, missense variants. Missense variants cause an amino acid (AA) change through single nucleotide substitution (6). Determining the pathogenicity of these variants is a complex task, as they can have subtle effects on protein function that are difficult to predict or prove. Even though they are suspected to cause phenotypic effects due to protein structure/function disruption (7) and such variants are reported as “causal” if they influence the biological phenotype (8) associating a missense variant with a disease without functional studies is not feasible in real-life. The clinician’s expert opinion is also not enough to classify the variant without more evidence.

To address this, *in-silico* pathogenicity predictors have been developed. These computational tools aim to assess the potential impact of genetic variants on phenotypic outcomes. However, their performance has been limited, often leading to conflicting and inaccurate predictions, hence false positives, and negatives (9). This comes as a significant concern, as misinterpreting the pathogenicity of a variant can have serious implications for patient diagnosis and treatment.

In response to these challenges, the American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) have presented recommendations and guidelines to create a standardized approach for the interpretation of genetic variants (10). One of the key criteria in these guidelines is the PP3, which relies on the agreement of multiple computational evidence supporting the pathogenicity of a variant. The ACMG/AMP guidelines have been instrumental in improving the accuracy and consistency of genetic diagnoses, but they are not without their limitations, particularly in the context of rare genetic disorders (2, 5), which are clinical cases that require special care and attention.

Another particular challenge in the world of genetic diagnostics is the interpretation of Variants with Uncertain Significance (VUS). These genetic alterations' impacts on health and disease are not clearly understood, presenting a significant hurdle in clinical settings. The *VUS problem* is exacerbated by the fact that many genes and genetic variants, especially in the context of rare diseases, have not been adequately studied or documented, leaving clinicians and geneticists in a quandary regarding their clinical significance (10). This uncertainty can lead to significant difficulties in clinical decision-making, as the clinical implications of these variants are not clearly defined. Furthermore, current in-silico tools often struggle to accurately predict the pathogenicity of these variants, leading to potential misdiagnosis or inappropriate management strategies (9, 11, 12). The issue with these variants is not just a technical problem but also has ethical implications, as it raises questions about how to communicate these uncertainties to patients and how to manage their expectations and concerns (12-14).

Here we present VAMPP-score (Variant Analysis with Multivariate Pathogenicity Prediction-score), a novel *in-silico* pathogenicity prediction (ISPP) score for missense variants developed using a statistical framework. This score is based on multiple and pairwise comparison tests, as well as proximity estimations and mean difference calculations among variant groups. It utilizes the best gene-ISPP matches, according to ISPP accuracies for each gene, offering a combinatorial weighted score. VAMPP-score provides an advancement in missense variant interpretation. Given its statistical basis, VAMPP-score can be a decision-maker and assist the clinician in diagnostics. Its application, particularly in the reanalysis of rare and undiagnosed cases, promises to increase the success rate of diagnoses, enhance the accuracy of genetic diagnostics, and improve patient outcomes.

2 BACKGROUND

2.1 Human Genome

The human genome is composed of ~3.2 billion base pairs that carry the genetic information encoded in the form of DNA (deoxyribonucleic acid). The DNA is found within the 23 chromosome pairs in cell nuclei and as a small additional genetic material found in mitochondria (15).

The genetic material DNA carries the required elements for an organism's survival and proper functioning. It is composed of two strands that coil around each other to form a double helix. The genetic information is encrypted in the monomer building blocks of nucleic acids and called the nucleotides. The nucleotides, paired through hydrogen bonds, are composed of a nitrogenous organic base element (adenine (A), thymine (T), guanine (G), or cytosine(C)), a sugar, and a phosphate molecule (16) (Figure 1).

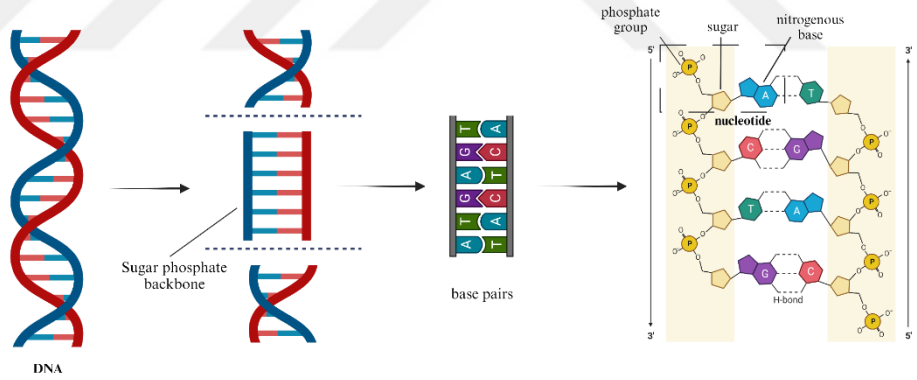

Figure 1. DNA structure.

Illustration shows the structure of the DNA and its sugar phosphate backbone, emphasizing on the double helix, formed by the base pairing of the nucleotides. (Created with BioRender.com)

Even though 99.9% of the genome is the same among individuals (17), the remaining 0.1% difference makes us who we are, genetically unique and different. Despite this small portion may seem insignificant, considering the total size of the genome it still corresponds to ~3 million base pairs and changes in the sequence within.

The human genome consists of ~20,000-25,000 genes (18) coded in two main regions: DNA sequences that encode proteins and various sequences that are not responsible for protein-coding. The latter makes up the 98-99% of the whole genome and is a diverse category that includes the coding sequences for untranslated RNA (ribonucleic acid), promoters and their associated gene-regulatory elements, DNA playing structural and replicatory roles (19). In addition, introns, the non-coding sections of an RNA transcript that are spliced (20) out before the RNA molecule is translated into a protein, make up a large percentage of non-coding DNA (21).

The coding region (CDS, coding sequence) of the human genome rests in the 1-2% of the whole sequence and it is responsible for preserving the protein information (22). The flow of genetic information is defined as the central dogma of biology and genetic information flows from where it is coded, DNA, to RNA by transcription, and to protein by translation (23) (Figure 2).

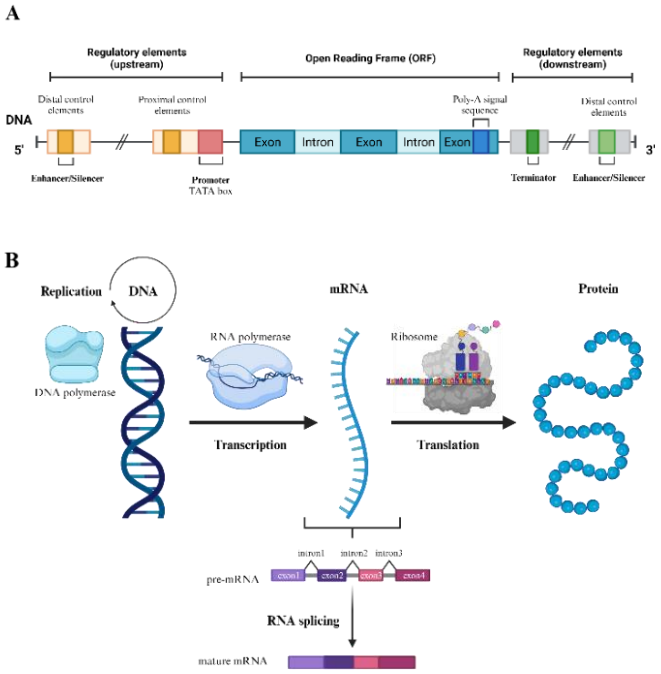

Figure 2. Gene structure and central dogma.

Panel A shows the structure of a eukaryotic gene. It has three main elements as upstream and downstream regulatory elements, and the open reading frame, which is the target region for the transcription. Panel B shows the central dogma, which displays the flow of the genetic information from DNA to protein. The transcribed mRNA

(called the pre-mature-mRNA) from the target DNA template, consists of exons and introns, as illustrated. Introns are spliced-out via RNA-splicing, creating the mature RNA that is ready for translation, in which the end-product is the protein. (Created with BioRender.com)

2.1.1 Human Genome Project

The Human Genome Project (HGP) was initiated in 1990 as an international effort and has become a milestone in genomics (24, 25). This groundbreaking project, which took more than a decade to "complete" -successful completion in 2003- (Figure 3), aimed to create an in-depth map of human genetic structure, providing an understanding (25, 26).

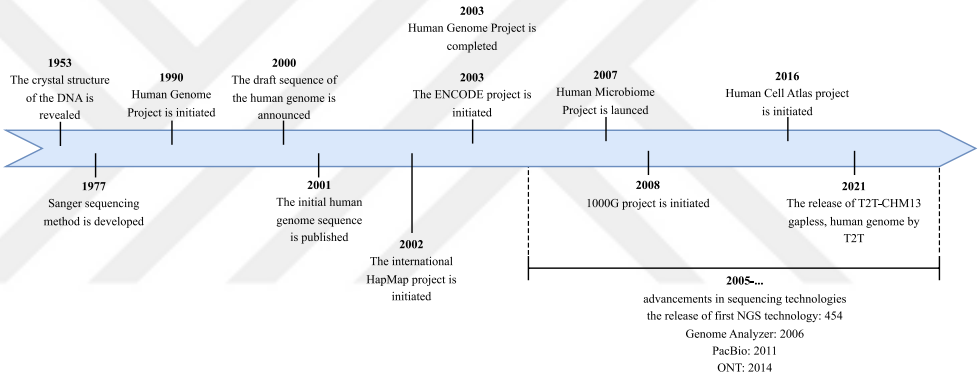

Figure 3. History of the human genome.

Timeline shows the milestones in genomics, genome projects that are initiated over the years. Starting from the discovery of the DNA's crystal structure, currently there is a human genome that is sequenced from telomere to telomere, due to the advancement in sequencing technologies**.**

The outcomes of the HGP have a broad impact in genetics and genomics. This extensive map of the human genome sequence has led the scientific community in understanding genetic diseases, developing new diagnosis and treatment strategies, and has broadened our perspective on personalized medicine (26, 27). The project has also been the driving force in the growth of bioinformatics, a field that is critical in analyzing and interpreting the large amounts of data produced in genomics (28).

2.1.2 Complete Assembly of the Human Genome

The Telomere-to-Telomere (T2T) (29) project pushes the boundaries of our understanding of the human genome as it is. Adding on to the Human Genome Project (24, 25), the T2T was initiated with the aim to create the entirely complete human genome, referring to the gaps remained by the previous initiatives.

The completion of a telomere-to-telomere sequence has become the beginning of a new era in genomics, offering a more comprehensive map of the human genome. This new extensive blueprint as the first telomere-to-telomere (T2T) human genome assembly (T2T-CHM13, latest version: v2, released on 24 January 2022 (29) release (Figure 3) enhances our understanding of the complex regions of the genome, such as telomeres and centromeres (30).

2.1.3 Reference Genome

The reference genome refers to the complete set of genes for a typical human. It's not the map of a specific individual, but rather a united assembly that shows a more general human genetic makeup (31, 32). Since its initial release, the reference genome has undergone continuous updates and refinements.

The current reference genome for humans (GRCh38) has been widely used in studies. It was released first by the Genome Reference Consortium (GRCh) as GRCh38 in 2017 (33) In February 2022, its latest version with additions, the last patch as GRCh38.p14 (34), was published. Even though there is a fully complete, gapless human genome sequence released by T2T consortium, the large-scale genomic differences between these two important genome assemblies are not characterized in detail yet (30).

2.2 Genetic Variation Types

All the changes in the genome sequence can be defined as variations. Rather than using a separate term for the more common ones or describing mutations as harmful, beneficial, or neutral, the usage of the term variation would ease the process of defining and understanding DNA sequence changes (35). Genetic variations can be divided into two main groups considering where in the genome they occur.

2.2.1 Variations in the coding regions

Variations in the coding regions of DNA, which are directly transcribed and translated into proteins, can have profound implications on protein structure and function. Any alteration can lead to changes in the amino acid (AA) composition of the protein product, which can affect the protein's conformation, stability, and interaction with other molecules (36).

2.2.1.1 Synonymous variants

Synonymous variants are single nucleotide alterations that do not interfere with the AA sequence (Figure 4A.1). They are often called "silent mutations" because the amino acid sequence remains the same. However, recent studies show that synonymous variations can also have effects on the protein structure (37-40).

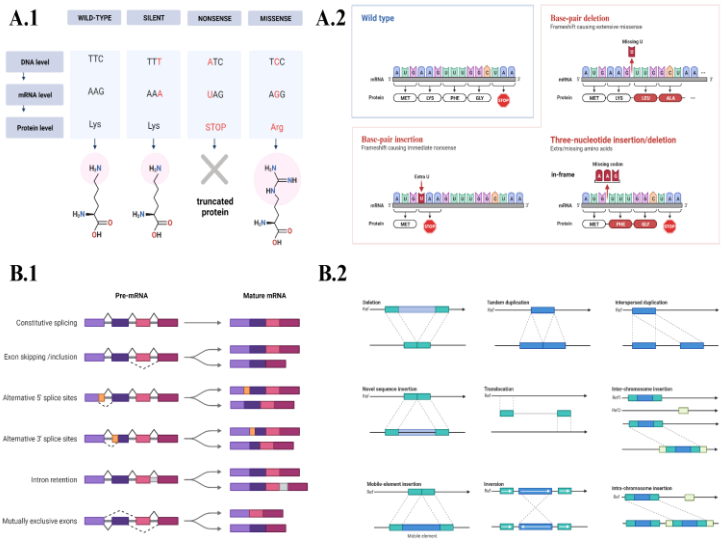

Figure 4. Illustration of variant types.

Figure 4 represents several types of genomic variation mentioned. In Panel A.1 and A.2 nonsynonymous variations which occur in the coding regions that directly interfere with protein sequence are shown, single nucleotide variations causing single AA changes, and single/short nucleotide variations causing reading frame alterations, respectively. Panel B.1 illustrates the variations that are causing abnormal splicing and splicing products, mature mRNA for translation therefore causing a potential abnormality in the protein. In Panel B.2, larger variations that can involve both coding and non-coding regions are illustrated. (Created with BioRender.com, adapted from templates).

2.2.1.2 Non-synonymous variants

Non-synonymous variants are defined as the genetic variations that cause a change in the amino acid sequence of a protein. Missense variants are the single nucleotide substitutions that result in a codon that codes for a different AA than the reference (Figure 4A.1). This change can affect the protein's function, depending on where the change occurs and which amino acids are involved, considering their biophysical and biochemical properties (36).

Nonsense variants cause a codon to change into a stop codon (Figure 4A.1), leading to premature termination during protein synthesis. This event can often result in a truncated, therefore a nonfunctional protein (41).

Frameshift variants (Figure 4A.2) occur when nucleotide insertions/deletions are not in multiples of three, causing a shift in the reading frame during translation (42, 43). This shift changes the following amino acid sequence, often resulting in an inoperative protein. In contrast, when insertions/deletions are in multiples of three, they do not disrupt the reading frame, hence named in-frame variants (Figure 4A.2). They can still maintain the overall protein structure and function; however studies show that they might occur with pathogenic effects as well (44).

2.2.2 Variations in the non-coding regions

Non-coding regions of the genome regulate gene expression rather than coding for proteins. These areas influence gene expression by modifying the transcription factor binding and the rate of transcription. Variations in these regions can result in altered levels of gene expression, consequently affecting the ultimate protein production which might lead to disease (45-47).

Intronic variants occur in the non-coding segments of the genome interspersed among exons (introns) (21). They are intra-genic variants and can be detected by whole genome sequencing (WGS). Inter-genic variations, on the other hand, occur in the sequences between genes. They might not affect protein sequences directly but can influence gene expression and regulation (48).

Introns are removed from transcripts at splice sites that are evolutionarily conserved, found at the 5′ and 3′ ends of the introns, usually with nucleotide pairs of GU and AG, known as essential splice sites. Variations of these sites can lead to the formation of abnormal splice sites (see Figure 4B.1) which may be disease-causing (49-51).

2.2.3 Other variations

Structural variations (SVs) are large-scale genome alterations like copy number variations (CNV), inversions, insertions, deletions, duplications, and translocations,

that can cover both coding and non-coding regions (see Figure 4B.2) (52-55). Microsatellites and minisatellites, short repetitive sequences found across the genome in both coding and non-coding areas, can influence gene function and regulation through repeat number changes (56).

2.3 Sequencing Technologies

Advancements and ongoing enhancements in sequencing technologies have made it possible to decode the human genome. These technologies provide in-depth understanding of genetic variations, crucial for comprehending diseases and genetic disorders (57, 58).

The sequencing technologies also raise challenges including high costs, substantial data management requirements, and ethical concerns regarding confidentiality of the genetic information (59). Latest advancements aim to reinforce accuracy, speed, affordability, and accessibility, with a focus on portable sequencing devices, improved bioinformatics tools and data management strategies (58, 59).

2.3.1 Sequencing Technologies

Sanger sequencing was developed by Frederick Sanger in 1977 as a method of DNA sequencing based on the selective incorporation of chain-terminating dideoxynucleotides by DNA polymerase during in vitro DNA replication. It begins with denaturing the DNA molecule, followed by primer annealing and extension using a mix of deoxynucleotides and dideoxynucleotides. The dideoxynucleotides, lacking a 3' hydroxyl group, terminate DNA strand elongation, resulting in fragments of varying lengths (60). These fragments are then separated by size through electrophoresis, allowing the sequence to be read.

Advantages of Sanger sequencing include its high accuracy, reliability, and ability to read longer continuous stretches of DNA compared to some other methods (57). However, it is relatively costly and time-consuming for large-scale genomic

sequencing projects and has lower throughput compared to next-generation sequencing technologies (59).

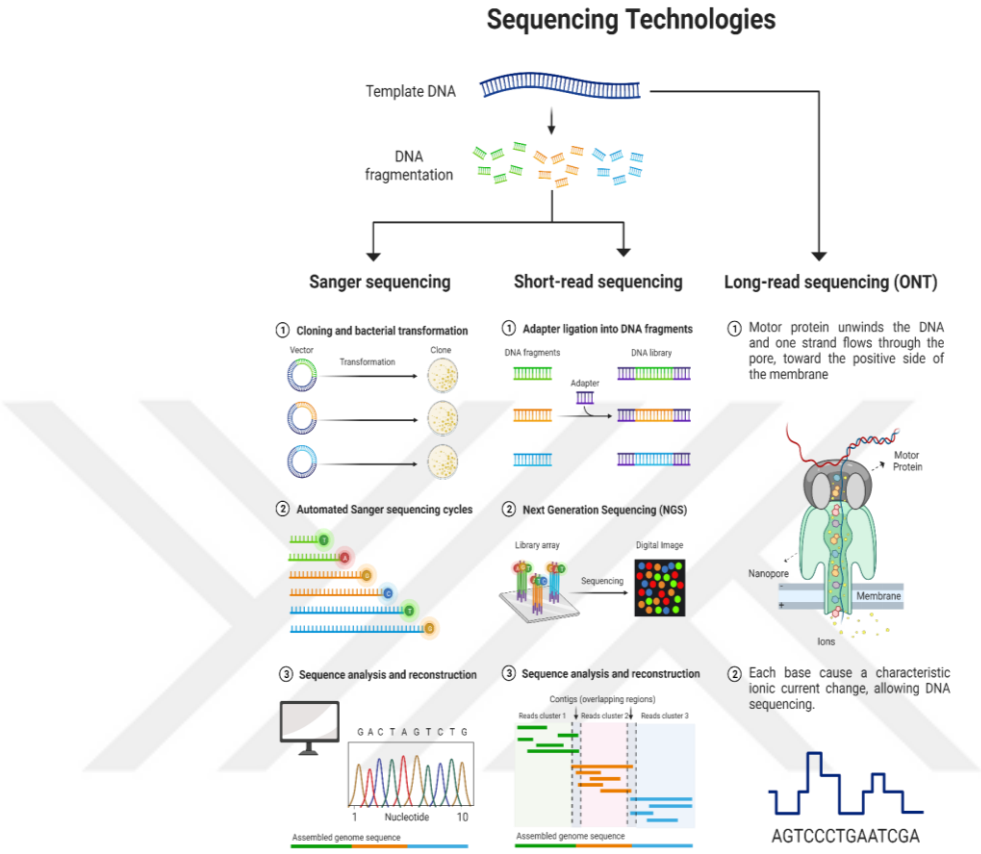

Figure 5. Sequencing technologies.

The illustration shows the sequencing technologies from Sanger to short- and long-read sequencing, which may be referred to as the generations of sequencing. Here, Oxford Nanopore technologies is shown as an example of long-read, single molecule sequencing. (Created with BioRender.com, adapted from templates; Sanger and Short-read sequencing, adapted from (61) Long-read sequencing: adapted from the template by Flo Glencross (Creator) Daid Ahmad Khan (62).

2.3.2 Short-read sequencing

Short-read sequencing, also can be referred to as second or next-generation sequencing (NGS), emerged as a rapid and cost-effective method for sequencing both DNA and RNA in the early 21st century. Unlike Sanger sequencing, NGS offers a high-throughput alternative by simultaneously sequencing millions of fragments (63).

This process involves DNA/RNA fragmentation, adapter attachment, and amplification on a solid surface (Illumina, *https://www.illumina.com/*) or in emulsion droplets (Roche, *https://sequencing.roche.com/*), depending on the sequencing technology used. Subsequently, massively parallel sequencing is conducted, producing an extensive volume of data (64) which is then computationally compiled into a sequence.

NGS can process large volumes of data quickly and cost-effectively, making whole-genome sequencing more accessible (59). However, NGS has limitations, such as shorter read lengths compared to Sanger sequencing, which can complicate data assembly and interpretation. Furthermore, the equipment needed to perform the procedure and the need for enhanced computational resources to manage data are considerable (58).

2.3.3 Long-read sequencing

Long-read sequencing, also known as the third-generation sequencing, emerged as the latest advancement to addressing the limitations of the earlier technologies. Represented as single-molecule sequencing, DNA is sequenced without requiring amplification, thereby reducing biases and errors due to PCR amplification used in earlier methods (65, 66). It offers real-time sequencing with much longer reads, which is a significant augmentation for complex and *de novo* genome assemblies, and SV detections (67). It also provides a more direct approach to detecting DNA modifications, which is vital for epigenetic studies (68).

Third-generation sequencing represents a significant leap forward in sequencing technology, offering longer reads and real-time sequencing capabilities, but it is still evolving in terms of accuracy, cost, and data management (69).

2.4 Handling Sequencing Data

Advancements in genomic technologies have expanded our knowledge in genetic diseases. These innovative techniques offer profound insights into genetic makeup, facilitating personalized medicine and sophisticated genetic research. Depending on the focus, various sequencing tests can generate data, each selected for its specific target (70).

Targeted panels and clinical exome sequencing (CES) data enrich specific regions of the genome and are used to sequence the protein-coding regions that are most likely to be associated with the phenotypic outcome. CES is typically used especially when a patient presents with a rare genetic disorder (71).

Similar to CES, whole exome sequencing (WES) targets the exome, which constitutes about 1-2% of the human genome, the majority of known disease-related variants (72). WES is broader than CES and is used for research purposes in addition to clinical purposes.

Whole genome sequencing (WGS) is the most comprehensive form of genetic testing that is used to sequence an individual's entire genome. WGS covers both the exons and the non-coding region (73).This method is useful for detecting a wide range of genetic abnormalities, including small mutations and larger chromosomal changes. WGS is valuable in research for understanding complex genetic diseases and is recommended in clinical diagnostics (74).

2.4.1 Raw data processing in NGS

The most used NGS method, Illumina, generates optical files in BCL (base call file) format, which is specific to Illumina's base calling. BCL format files consist of data files and metadata related to the flow cell used in sequencing. The built-in Illumina software converts these files into FASTQ files. These files are more

manageable and can be the input for numerous bioinformatics workflows, expediting the data analysis (Figure 6).

FASTQ is a text-based file that stores both raw sequence data and corresponding quality scores for each nucleotide. These scores provide an estimate of the probability of an error in the sequencing (75). A quality-control step is preferred before processing with further analysis using tools for quality check such as FASTQC (76). FASTQ files also involve the information on sequencing adapters used in library construction compatible with the sequencing device. Prior to further processing the sequencing data, these adapters should be trimmed to avoid misinterpretation of the sequencing data with tools like Trimmomatic (77).

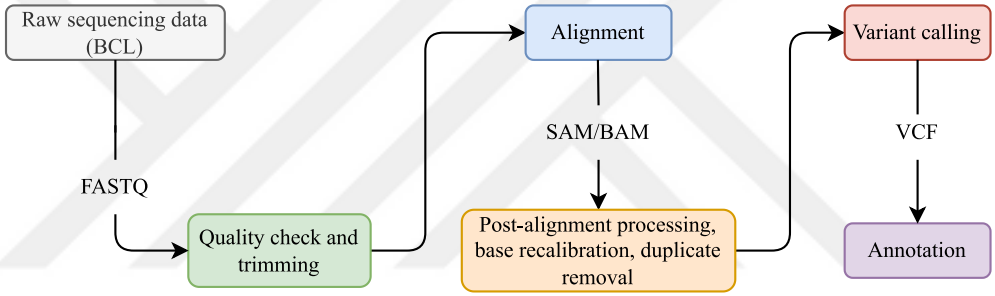

Figure 6. The workflow of NGS data processing.
The illustration above summarizes a basic bioinformatics workflow employed to process NGS data.

Trimmed and ready-to-analyze sequencing files are used for aligning the reads to the reference genome assembly. These alignment processes are done via genomic analysis tools called aligners such as bwa / bwa-mem2 (Burrows-Wheeler Alignment-maximum exact matches) (78-80), SNAP (Scalable Nucleotide Alignment Program) (81, 82) etc. After the alignment, a SAM (sequence alignment map) and BAM (binary alignment map) files are created. BAM is the machine-readable, binary format of the SAM file and preferred for downstream analysis (83).

Alignment data is checked for the base quality score recalibration (BQSR), duplicate marking, and removal. These steps can be performed via various

computational tools available, such as elPrep (84, 85). In BQSR, bases receive recalibrated quality scores that are more precise, achieved by identifying systematic errors in the sequencing machine's estimations of each base call's accuracy. In duplicate removal, the possible PCR duplicates created during library preparation are located and tagged in a SAM/BAM file (86). These are then removed to avoid overrepresentation of biases for an original molecule due to amplification steps.

Then the data is used in variant calling. With variant calling, the genomic positions where the aligned reads differ from the reference genome are identified. There are several variant callers available to use in this step, such as HaplotypeCaller, which is also preferred by GATK (Genome Analysis Tool Kit) BestPractice (86), FreeBayes (87), BCFTools (88), DeepVariant (89) etc.

The variant caller algorithms give an output file called VCF (variant call format), which is a tab-limited text format file composed of a list of variants. This list consists of each position and allelic change detected in the processed genomic data, with respect to the reference genome (90).

2.4.2 Variant annotation

Once the genomic data has been processed, a VCF file is generated, containing all detected genomic variations. To understand the impact of these variants, an additional process known as "annotation" must be carried out.

Annotation is a critical process in associating variants with phenotypic outcomes where the variants are endowed with functional information (91). This information can vary from associating the variant based on its evolutionary conservation (92), its located gene & the gene’s biological function (93, 94), related protein’s structural properties (95, 96), the variant’s prevalence in the population (97, 98), drug associations (99), computational predictions, etc. (91, 100, 101). The resources and databases utilized for annotation can be categorized into four groups.

2.4.2.1 Functional annotation

Functional annotation of genomic variants utilizes several resources and databases to identify variations and associate them with biological implications. Gene Ontology (GO) terms are provided in a list of words as a vocabulary designated for gene product attributes across species, grouping the terms into three main categories biological processes, cellular components, and molecular functions (93, 94). KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways offers a collection of databases, that map genomic variants to specific pathways that curate genomic data, biological pathways, diseases, drugs, and chemical substances (102). STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) database, on the other hand, provides protein-protein interaction information for understanding the network context of genomic variants that might affect protein interactions and signaling pathways (103).

Similar to GO (93, 94) terms, Human Phenotype Ontology (HPO) provides phenotypic abnormalities in human diseases dictionary, useful in associating phenotypic abnormalities to genomic variants (104). The HPO enables the annotation of human diseases with specific phenotypic descriptors, facilitating the identification of phenotypic similarities between diseases and their associated genetic causes. This is especially valuable in the field of rare diseases, where genomic variants often lead to unique and specific phenotypes.

One other important annotation is Online Mendelian Inheritance in Man (OMIM) (105), which is an essential database focusing on the association between human genes and genetic disorders, similar to HPO. Managed by the McKusick-Nathans Institute of Genetic Medicine at Johns Hopkins University, OMIM provides details on genotypes-phenotypes associations, and the genetic background of numerous diseases (106).

2.4.2.2 Population allele frequency-based annotation

Evaluating a variant's population allele frequency information relies on its prevalence within different populations. Frequency of each variant in a given

population is crucial for determining the variant's potential impact on health and disease (97, 98, 107). In this context, the variants are commonly observed in the population often considered benign and/or low-impact, while the variants that are not can be associated with pathogenic effects (108).

Databases like the Exome Aggregation Consortium (ExAC) (109), 1000 Genome Project (1KGP) (98), and the Genome Aggregation Database (gnomAD) (97) provide extensive data on allele frequencies across diverse populations, enhancing the understanding of genetic diversity and its implications for health and disease (110). In addition, there are also resources that represent specific ethnicities/populations, which can specifically be used when dealing with sequencing data coming from a population of interest. Great Middle Eastern (GME) Project (111), Iranome Project (112), and Chinese Millionome Database (CMDB) (113) can be given as examples for these databases.

2.4.2.3 In-silico pathogenicity predictors

In silico pathogenicity predictors (ISPPs) are computational tools used to assess the pathogenicity of the variants. These ISPPs are crucial, as they provide a rapid and cost-effective evaluation of the potential impact of a variant. In consideration of various aspects, such as sequence conservation, protein structure, and functional impact, these predictors are designed to aid in distinguishing pathogenic from benign (91-96, 100, 101).

ISPPs that are most frequently used in practice (Table 1) involve predictors with algorithms that focus on sequence-based evolutionary conservation, protein structure and function, and ensemble ones with integrated approaches.

Table1. In-silico pathogenicity predictors (ISPP) table. The table shows the most frequently used ISPPs with rank scores available in dbNSFP (114, 115) as of March 2024. Definitions of ISPP acronyms were written below the name of the ISPP, if available. (continued)

Predictor	Subtype, if any	Basic Information	Reference
GERP **G**enomic **E**volutionary **R**ate **P**rofiling	_RS: rejected substitution	identifies evolutionary constrained elements in multiple-species alignments by comparing rates of nucleotide substitutions	(92)
phyloP	17way: primate species list 100way: vertebrate species list 470way: mammalian species list	measures evolutionary conservation at individual alignment sites by assessing the likelihood of observing the given number of substitutions under a neutral model	(116)
phastCons	17way: primate species list 100way: vertebrate species list 470way: mammalian species list	estimates the probability of each nucleotide being conserved based on a phylogenetic hidden Markov model	(117)
SiPhy	29way: mammalian species list	utilizes patterns of nucleotide substitution and insertion/deletion events to identify regions under evolutionary constraint	(118)
bStatistic		quantifies the degree of background selection at a given site in the human genome	(119)
LIST-S2 **L**ocal **I**dentity and **S**hared **T**axa-**S2**		assesses local sequence conservation and shared taxa to predict pathogenicity	(121)
FATHMM **F**unctional **A**nalysis **t**hrough **H**idden **M**arkov **M**odels	_MKL: multiple kernel learning (122) _XF: eXtended features (123)	predicts the functional effects of missense variants using hidden Markov models	(124)

Table1. In-silico pathogenicity predictors (ISPP) table. The table shows the most frequently used ISPPs with rank scores available in dbNSFP (114, 115) as of March 2024. Definitions of ISPP acronyms were written below the name of the ISPP, if available. (continued)

Predictor	Subtype, if any	Basic Information	Reference
SIFT **S**orting **I**ntolerant **f**rom **T**olerant	_4G: for genomes	predicts whether an amino acid substitution affects protein function based on sequence homology and the physical properties of amino acids	(101)
EVE **E**volutionary Model of **V**ariant **E**ffec**t**		uses a machine learning approach to predict the functional impact of coding variants based on evolutionary information	(125)
LRT **L**ikelihood **R**atio **T**est		assesses whether a nonsynonymous (amino acid changing) variant is likely to be pathogenic based on a likelihood ratio test	(126)
VARITY	_R:rare _ER:extremely rare _R/ER-LOO:leave-one-out	predicts the impact of amino acid substitutions on protein function, has subtypes on rare and extremely rare variants, with LOO approach	(127)
MutationAssessor		predicts the functional impact of amino acid substitutions in proteins based on evolutionary conservation	(128)
MutationTaster2		predicts disease-causing potential of genetic variants including SNPs and indels	(129)
VEST4 **V**ariant **E**ffect **S**coring **T**ool		uses a machine learning approach to score the pathogenicity of missense mutations	(130)

Table1. In-silico pathogenicity predictors (ISPP) table. The table shows the most frequently used ISPPs with rank scores available in dbNSFP (114, 115) as of March 2024. Definitions of ISPP acronyms were written below the name of the ISPP, if available. (continued)

Predictor	Subtype, if any	Basic Information	Reference
PROVEAN **Pro**tein **V**ariation **E**ffect **An**alyzer		predicts whether a protein sequence variation is likely to be deleterious	(131)
MutPred		predicts the potential pathogenic effects of AA substitutions in humans	(132)
PrimateAI		predicts disease-causing variants based on a deep learning model trained on primate data; the human+primate deep learning network	(133)
DEOGEN2		incorporates heterogeneous information about the molecular effects of the variants, the protein domains, and gene interaction information	(134)
AlphaMissense		makes variant effect prediction proteome-wide, focuses on missense variants; a deep learning model that builds on AlphaFold2 for protein structure predictions	(135)
PolyPhen2 **Poly**morphism **Phen**otyping v**2**	_HDIV & _HVAR: two data sets	makes predictions for AA substitutions on the protein structure and function using physical and comparative considerations	(100)
fitCons **fit**ness **Cons**equences	integrated_fitCons GM12878(human lymphoblastoid cells)_fitCons H1-hESC(human embryonic stem cells)_fitCons HUVEC(human umbilical vein epithelial cells)_fitCons	assesses the selective constraint on individual nucleotides within genomic regions	(136)

Table1. In-silico pathogenicity predictors (ISPP) table. The table shows the most frequently used ISPPs with rank scores available in dbNSFP (114, 115) as of March 2024. Definitions of ISPP acronyms were written below the name of the ISPP, if available. (continued)

Predictor	Subtype, if any	Basic Information	Reference
MVP **M**issense **V**ariant **P**athogenicity prediction		prediction method that uses deep residual network to predict the effect of missense variants	(137)
BayesDel	_noAF: no allele frequency information added _addAF: allele frequency information added	combines multiple predictors using a Bayesian framework, utilizes the approach PERCH (Polymorphism Evaluation, Ranking, and Classification for a Heritable trait)	(138)
CADD **C**ombine **A**nnotation **D**ependent **D**epletion	_19: considers hg19 as the reference genome	integrates multiple annotations into a single metric to predict the deleteriousness	(139)
DANN		a deep learning approach that utilizes a neural network for prediction, trained on CADD data	(140)
Eigen_coding	_PC: principal component, uses the lead eigenvector for weighting annotations	uses a machine learning model to predict the functional effect of coding variant	(141)
GenoCanyon		statistical method to predict functional regions in the human genome	(142)
ClinPred		predicts the pathogenicity of missense variants based on clinical data	(143)
MetaSVM		support vector machine-based method integrating multiple predictors to assess variant pathogenicity.	(144)

Table1. In-silico pathogenicity predictors (ISPP) table. The table shows the most frequently used ISPPs with rank scores available in dbNSFP (114, 115) as of March 2024. Definitions of ISPP acronyms were written below the name of the ISPP, if available. (continued)

Predictor	Subtype, if any	Basic Information	Reference
MetaLR		uses logistic regression to integrate multiple variant effect predictors.	(144)
MetaRNN		recurrent neural network approach integrating various predictors	(145)
M-CAP **M**endelian **C**linically **A**pplicable **P**athogenicity		predicts the pathogenicity of missense variants in Mendelian disorders	(146)
REVEL **R**are **E**xome **V**ariant **E**nsemble **L**earner		ensemble method combining multiple tools to predict the pathogenicity of rare missense variants	(147)
MPC **M**issense badness, **P**olyPhen-2, **C**onstraint		combines missense deleteriousness, PolyPhen-2, and constraint scores	(148)
ESM1b **E**volutionary **S**cale **M**odeling		uses evolutionary modeling to predict the effects of protein mutations	(149)
LINSIGHT **L**inear **I**nference of **N**atural **S**election from **I**nterspersed **G**enomically co**H**erent elemen**T**s		predicts functional significance of genomic variants based on evolutionary conservation	(150)

2.4.2.4 Clinical significance

Several databases and annotators are used in interpreting the implications of the variant on human health. ClinVar is the most frequently accessed resource for clinical significance assessment, offering an archive for relationships between variations and phenotypic outcomes. Through ClinVar, interpretation of clinical significance and

submission information can be reached, especially when the variant was previously reported to the database with similar phenotypic effects (151). Likewise, the Human Gene Mutation Database (HGMD) provides a comprehensive variant set associated with inherited disease, which can be pivotal in clinics (152). Over and above, a comprehensive variant archive for genetic variations with their functional annotations is presented as dbSNP (Database of Single Nucleotide Polymorphisms) (153)

Concerning cancer, somatic variations are recorded and annotated in COSMIC (Catalogue Of Somatic Mutations In Cancer). COSMIC's data comes from broad-scale projects and systematic literature review, offering information on molecular genetics of cancer (154). Aside from functional annotators that focus on variant-associated disease pairs, genotype-phenotype correlations and drug interactions can be investigated with such databases like PharmGKB (99).

2.5 Variant Interpretation and Prioritization

After completing the processing and annotation of raw genomic data, analysts handle a list containing tens of thousands of variants from WES and millions from WGS. The critical task for variant analysts is to interpret and prioritize these variants to identify the causal variant(s) linked to the phenotypic outcome, thereby reaching a definitive conclusion (155).

With the functional association by annotation, the variants can be ranked in terms of clinical importance and/or relevance such as their genomic location, sequence ontology, scores from ISPPs and other methods, including assessing the variant's frequency in the healthy population (91).

The criteria mentioned above for variant prioritization were published by the ACMG (American College of Medical Genetics and Genomics)/AMP (American Association of Molecular Pathology) in 2015 (10) in a consolidated form as a guideline for the interpretation of sequence variants.

2.5.1 ACMG/AMP standards and guidelines for Mendelian disorders

The era of genomics has been evolving rapidly year by year and the clinical laboratories were taking advantage of these opportunities and performing various types of genetic tests for understanding the backgrounds of genetic diseases.

This expansion and improvement in genetic testing have resulted in new challenges. In this context, in 2013 (years after publishing their first for the interpretation of sequence variants guidance in 2007 (156), ACMG (American College of Medical Genetics and Genomics) and representatives from AMP (Association for Molecular Pathology) and the College of American Pathologists (CAP) gathered as a group which also included clinical laboratory directors and clinicians to revise the previously recommended criteria for the variant interpretation. They published a new recommendation report, the standards, and guidelines for the interpretation of sequence variants as a joint consensus in May 2015 (10).

In this report, they introduced a standardized terminology for variant classification: “pathogenic," "likely pathogenic," "uncertain significance," "likely benign," and "benign" to describe variants located on genes that are associated with Mendelian disorders. Furthermore, this guideline outlines a method for categorizing variants into these five groups, based on variant evidence such as population, computational, functional, and segregation data (Figure 7).

Figure 7. ACMG variant classification evidences (10)

2.6 Variants with Uncertain Significance

The interpretation of the variants with uncertain significance (VUS), particularly when they are missense variants, is presented as a complex issue in diagnostics, due to their potentially subtle impact on gene function (6-8, 10, 36). The primary challenge is the lack of definitive evidence to classify missense VUS as either pathogenic or benign, when there is insufficient or conflicting data (10).

The uncertainty in VUS interpretation can arise from limited population data, inadequate functional studies, and a nascent understanding of genetic variations (157). This complexity often leads to challenges in risk assessment and management in clinical settings, making it difficult for healthcare providers to offer clear guidance to patients (12-14).

To address these challenges, computational predictive models are used to assess the pathogenicity of genetic variants, although they have limitations in prediction accuracy (9, 11, 12). Assays such as MAVE (Multiple Assays for Variant Effects) are mainly used to detect a variant's functional impact (158), but they are not feasible for all possible variants, they are labor-intensive and can be challenging to interpret (36). The information reached through population frequency databases can also be useful to interpret upon a variation's prevalence, although this is limited by the representation of the gene/variant of interest and of the population in the database (10, 110-113).

Interpretation of missense VUS involves a coordinated method with computational modeling, functional insights, and population frequency data. The everlasting challenge in the field lies in developing more expansive databases, enhancing predictive algorithms, and increasing the scope of functional studies all together to improve the pathogenicity prediction and interpretation accuracy (155, 158-160).

2.7 Conflicting Computational Prediction in Variant Interpretation

In genetics and genomics, computational predictors are developed to mainly predict the potential implications of the variants on a disease and/or condition (91, 100, 101). In spite of their common usage, they have limited prediction ability, and it is crucial to understand how they work, and one needs to be cautious when interpreting the results (9-11).

The limitations of these tools can arise from several reasons. One of which is the input data quality used in the construction of the model. Prediction errors can be resulted from a deficient or inaccurate dataset (161). Also, as the scope of the predictor expands, the required data to process would also grow, hence the operation would need more computational power and memory capacity.

Pathogenicity predictors might yield inconsistent results due to the different input datasets that they use or their methodological approaches (162). Such controversial

results can affect practical applications in clinical genetics and diagnosis drastically as they are accepted to be decision-makers (9, 11-14).

The ACMG/AMP 2015 guideline for sequence variant interpretation accepts the evidence from computational predictors in PP3 criterion as "supporting" level by default. PP3 criterion is recommended to be used in the final classification only if there is an agreement among the pathogenicity prediction tool results (10).There have been initiatives in this regard to compare these tools performance to evaluate their prediction accuracy (163).

The exponential data accumulation and regular updates in datasets gives promising results in the development of such approaches. As data accumulates and knowledge on known variants' clinical significance expands (151), integrating information and combining multiple methodologies together assisted with statistical approaches has a potential to strengthen these assessments (163).

2.8 ClinVar

ClinVar (151) is an open-access human genetic variation archive, managed by National Institutes of Health (NIH). ClinVar not only consists of the variants but also their clinical significance on human diseases.

The clinical significance interpretations and assertions for the variants are submitted to ClinVar by clinical testing and research laboratories, expert panels, and other groups. Variant submission includes the description of the variant, and various information related to the variant such as the associated condition or disease, and its interpreted clinical significance. Additionally, submitters can provide the observed inheritance pattern and any evidence on the variant interpretation to support their submission (151).

Each variant submission in ClinVar is appointed with a review status. The review status is to aid the user in understanding how much the interpretation is supported. The

review status assigned to the variant record is constructed on the submission criteria (151).

Table 1. The review status and star assignment for ClinVar (151) variants.

Number of gold stars	Review status	Description
****	practice guideline	practice guideline
***	reviewed by expert panel	reviewed by expert panel
**	criteria provided, multiple submitters, no conflicts	Two or more submitters with assertion criteria and evidence (or a public contact) provided the same interpretation.
*	criteria provided, conflicting interpretations	Multiple submitters provided assertion criteria and evidence (or a public contact) but there are conflicting interpretations. The independent values are enumerated for clinical significance.
*	criteria provided, single submitter	One submitter provided an interpretation with assertion criteria and evidence (or a public contact).
-	no assertion for the individual variant	The allele was not interpreted directly in any submission; it was submitted to ClinVar only as a component of a haplotype or a genotype.
-	no assertion criteria provided	The allele was included in a submission with an interpretation but without assertion criteria and evidence (or a public contact).
-	no assertion provided	The allele was included in a submission that did not provide an interpretation.

All the data in ClinVar, weekly-updated on Mondays, can be reached through the ClinVar FTP site and is available in several formats. ClinVar, as of 9 December 2023, has more than 3.5 million submissions, approximately 2.5 million of them being unique variation records coming from a total of 2696 submitters (151).

2.9 dbNSFP

dbNSFP (database for nonsynonymous SNPs' functional prediction) is a database constructed for a comprehensive annotation that includes all potential single nucleotide variations (SNVs) in the human genome (114, 115).This extended variant list includes in total 84013490 non-synonymous and splice-site SNVs.

The current version of dbNSFP (version 4.5) is available as of 2 November 2023 and it gathers 52 in-silico pathogenicity predictors, 9 of them being evolutionary conservation scores and 43 of them being prediction algorithms. Three (144, 145) out of these 43 are developed by the dbNSFP researcher team. dbNSFP not only compiles prediction scores but also additional relevant information on the variants from various allele frequency databases, functional annotation on corresponding genes, expression and interaction data, etc., which enables a comprehensive variant interpretation process by dbNSFP (114, 115).

Here, we introduce the VAMPP-score (Variant Analysis with Multivariate Pathogenicity Prediction-score), an innovative in-silico pathogenicity prediction (ISPP) score specifically designed for missense variants. This score is crafted using a statistical framework that incorporates multiple and pairwise comparison tests, with functions to estimate ISPP performances on variant group distribution. It leverages the best gene-ISPP matches. With its statistical foundation, the VAMPP-score serves as a valuable tool for clinicians in making diagnostic decisions.

3 MATERIALS AND METHODS

VAMPP-score is based on multiple and pairwise comparison tests, as well as proximity and mean difference calculations, among variant groups defined under three as Pathogenic, Benign, and Unknown. The methodology is detailed below.

3.1 Material

3.1.1 In-silico pathogenicity predictors (ISPPs) considered

ISPPs for variant interpretation in this study were selected based on their pre-calculated score availability in the database of human nonsynonymous SNPs (dbNSFP, version 4.7A) and their functional predictions (114, 115). In this context, a total of 52 pathogenicity scores were included. LINSIGHT (150) & GenoCanyon (142) and CADD_19 (139) scores were eliminated since they target the non-coding regions of the genome and are based on the genome assembly hg19, respectively. The predictors included in the analysis were further categorized into four groups (Figure 8, see also Table1 for ISPP details).

Figure 8. ISPP categorization.

The ISPPs available in dbNSFP4.7A (114, 115)were categorized into four groups based on their prediction approaches. **Category A** represents the ISPPs that make their prediction based on evolutionary conservation on the nucleotide sequence level. The ISPPs that are listed under **Category B** make their prediction based on evolutionary conservation on amino acid sequence level and are designed to focus on missense variant pathogenicity prediction mainly. **Category C** represents the ISPPs that can be defined as meta-predictors that integrate multiple parameters into their prediction algorithm, such as evolutionary conservation, biological function, protein structure, and/or population frequency data. The final category, **Category D**, is composed of ISPPs that integrate multiple parameters into their prediction algorithm, as well as using multiple prediction scores and/or algorithms for their final scoring.

In downstream analysis, to maintain consistency and ease of analysis, the ranked scores for each ISPP were chosen as 0-to-1. The higher scores were indicative for pathogenicity, and lower scores indicated benign characteristics.

3.1.2 Dataset curation

3.1.2.1 The dbNSFP database

All possible single nucleotide variations in the human genome, over 84 million variants, with their corresponding information provided by dbNSFP4.7A (114, 115) was used to create a DuckDB (164) database (Figure 9), which allows fast and efficient querying and extraction of the initial dataset based on specified criteria of choice for downstream analysis.

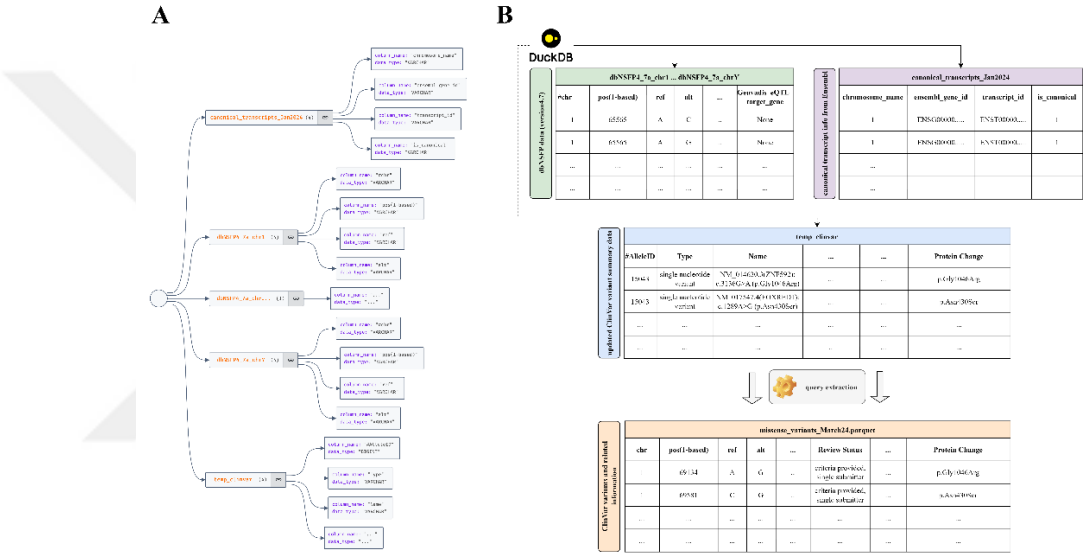

Figure 9. Schematic representation of the database structure.

A. The figure shows the DuckDB (164) database structure. The database mainly consists of twenty-five tables: *canonical_transcripts_Jan2024* and *dbNSFP4_7a* data for each chromosome 1-22, X, and Y. The *canonical_transcripts_Jan2024* table has the canonical transcripts available in the Ensembl.release 111 (165) for the genes in human genome chromosomes 1-22, X, and Y. *dbNSFP4_7a* chromosome tables include all the data provided in dbNSFP v4.7A (114, 115). *temp_clinvar* represents the ClinVar (151) variant summary data that is updated weekly, which can be reached through the ClinVar FTP site. All the data available in the ClinVar variant summary is integrated as a temporary table, hence named temp_clinvar, which is then used for the variant list querying and matching variant information among the tables. B. The figure shows the general structures of the input and output datasets. The three tables located on the top of the figure, before query extraction, correspond to the database tables shown in A. The final variant list was exported as a parquet file, which is a column-oriented format for efficient storage and retrieval. It provides efficient data compression and encoding schemes with enhanced performance to handle complex data in bulk (166). The extracted file consists of the matching variants from ClinVar

(151) with their corresponding ISPP rank scores for the statistical analysis and further downstream applications.

DuckDB (version 0.9.3-dev) uses an architecture storing data in columns instead of rows. This columnar storage in the DuckDB database management system allows for the efficient writing and reading of data, accelerating the query return. Considering the volume and complexity of the genomic data, this architecture becomes crucial for handling and manipulating the data (167). DuckDB (164) is also compatible with Python and R languages, which enhances its accessibility and handling for researchers in genomics without profound data science and data management expertise. DuckDB also supports fast schema changes and updates, which are important for databases requiring frequent modifications.

The utilization of DuckDB (164) was preferred for the reasons mentioned above, especially because of the need for an updatable database. The datasets integrated into the structure are composed of genomic variation data (Figure 9), at least one of which is known to get updated weekly (151) (Section 3.1.2.3, variant summary). In addition, the dbNSFP (114, 115) data has been getting updates more frequently for the past two years (168), due to advancements in the field, development of new computational predictors, updates, and error fixes.

3.1.2.2 Canonical transcript mapping

Multiple sequence ontology annotations for variants chosen to be included in the analysis were aimed to be avoided. For this, canonical transcript IDs for each gene based on Ensembl Release 111 (January 2024) (165) were integrated into our DuckDB (164) database as a new table. During query extraction, variants that have matching transcript IDs within dbNSFP (114, 115) and the reference variant list in pathogenicity assessment from ClinVar (151) were chosen upon Ensembl (165) canonical transcripts (Figure 9).

3.1.2.3 The up-to-date ClinVar variant summary data

The updated ClinVar (151) variant summary data was incorporated as a reference variant list for pathogenicity assessment (March 2024). The variant summary data was accessed through the ClinVar FTP site, and it was pre-filtered based on the following criteria: missense variants, annotated according to the genome assembly GRCh38 annotations with unspecified review status. The variants were further refined upon their clinical significance to avoid the involvement of conflicting interpretations and to ensure that more evident clinical outcomes were included in the analysis. Accordingly, the variants with the following clinical submissions were kept: Pathogenic (P), Likely pathogenic (LP), Pathogenic/Likely pathogenic (P/LP), Uncertain significance (VUS), Likely benign (LB), Benign/Likely benign (B/LB), and Benign (B).

It was strategically chosen to utilize the ClinVar variant list without imposing restrictions on the review status of the variants, a decision primarily influenced by the need to increase statistical power of the analysis. Since the restriction of the variants according to their review status significantly reduces the number of variants in the list, this inclusive approach of no-restrictions allows for a more comprehensive examination of the variants with known clinical significance with at least a submission level to a database.

3.1.2.4 Final variant dataset

VAMPP-score utilizes the best gene-ISPP matches by integrating several parameters into the calculation, including the results from statistical tests and proximity metrics. In order to conduct these tests and calculate these metrics, a final variant dataset with the required information was created. This dataset (ClinVar_variant_summary_March24) was generated based on canonical transcript ID (section Materials 3.1.2.2) and refined ClinVar (151) variant data (section Materials 3.1.2.3) transcript ID-match status, querying a list of variants with the necessary information available in dbNSFP (114, 115) (section Materials 3.1.2.1). Due to the

size of the dbNSFP dataset, only specific columns, including general information on the variants and ISPP scores and rank scores, were extracted to avoid irrelevant information handling and computational exceeding.

The final variant dataset consisted of 1162673 variants of three main variant groups as Pathogenic, Benign, and Unknown on 17712 unique genes (Figure 10).

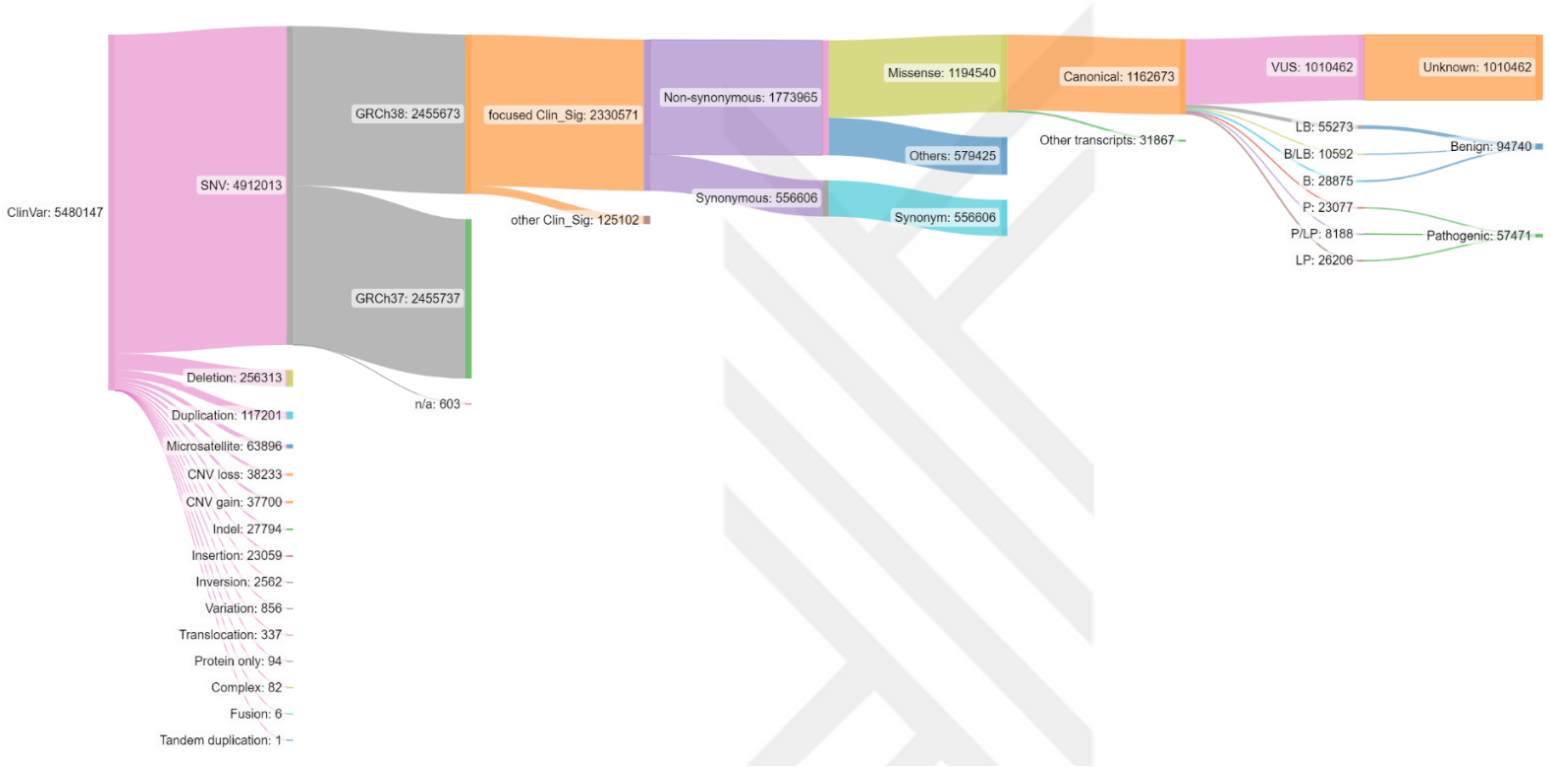

Figure 10. Schematic representation of the list of variants used as input data in the analysis.

Sankey diagram illustrates the flow and relationships between the variants submitted to ClinVar as of March 2024. The width of each flow in the diagram is proportional to the volume of each represented variant type, genome assembly, and reported clinical significance. The diagram starts with the initial variant summary data accessed through ClinVar FTP site and ends with the variant grouping as Pathogenic, Benign, and Unknown used in the analysis in the scope of this study. **focused_Clin_Sig* is described as the reported clinical significance of the variants submitted to ClinVar and the specific classes that are focused in the analysis as pathogenic (P), likely pathogenic (LP), pathogenic/likely pathogenic (P/LP), uncertain significance (VUS), benign (B), likely benign (LB), benign/likely benign(B/LB), whereas *other_Clin_Sig* class represents all the other clinical significance submissions. Sankey diagram was created utilizing a freely accessible tool SankeyMatic, available at *http://sankeymatic.com/build/*.

All the data used for curation in this thesis is publicly available in dbNSFP (114, 115), Ensembl (165), and ClinVar (151).

3.2 Methods

3.2.1 VAMPP-score calculation

The VAMPP-score is designed as a pathogenicity score calculated based on the gene-based prediction accuracy of the ISPPs combined with the ISPP rank scores available in dbNSFP (114, 115) for each variant. It consists of four main components that can be divided into two main sections (Section 3.2.2 Statistical Framework and Section 3.2.3 Variant Group Distribution Metrics) that combine within a weighted gene coefficient (*M*) assigned to each gene for each ISPP (gene-ISPP pair). The VAMPP-score also has a fifth component, which is the individual ISPP rank scores for each variant.

3.2.1.1 Weighted gene coefficient (*M*)

Weighted gene coefficient (*M*) is defined as the multiplier specific to each gene-ISPP duo based on ISPP accuracy for the gene of interest according to the multiple and pairwise comparison test results, along with variant group distribution metrics (Figure 11), explained in detail in the following sections. In short, *M* is the combined value assigned to each gene-ISPP pair, which was later used to calculate the VAMPP-score for each variant.

$$\mathbf{M} = (1 - P_k)(1 - P_{pb})(d)(s)$$

Figure 11. Function for the weighted gene coefficient (*M*).

The function shows the elements used in weighted gene coefficient (*M*) calculation. In this formula, 'P_k' represents the p-value (probability value) from Kruskal-Wallis and 'P_{pb}' represents the p-value from the Wilcoxon-rank-sum test for pathogenic-benign comparison of each gene-ISPP pair. Both p-values are used after subtracting the values from 1 to reflect significance with decreased p-value. '*d*' represents the proximity value, and '*s*' represents the mean difference, both indicating ISPP performance on variant group distribution.

M in VAMPP-score calculation meets different criteria for prediction accuracy, hence aiming to weaken the effect of conflicting results by ISPPs as much as possible. With gene-specific ISPP performance values, VAMPP-score decreases the contribution of the inaccurate predictors to the final score, while increasing it for more accurate ISPPs.

3.2.2 Section 1. Statistical framework

ISPP scores do not often follow a normal distribution, which is represented with a bell-shaped curve symmetrical around the mean. In normally distributed data, there is an accumulation of the data near the mean, and there are less data points far from the mean (169), hence creating a bell-shape as a graph. However, in the case of genomic data, including pathogenicity scores, there are generally several extreme values that might be lower or higher than the average, creating outliers in the data. Due to these outliers which are less represented in the data set, the distribution is more likely to get skewed, and have a tail; following a non-normal distribution (170, 171). Several factors, such as the categorical nature of the data, various methodologies, contradictory interpretations, outliers, and others cause the non-normal distribution of the data. The resulting scores are variable and may not reflect the "actual" consequence of the variant but just indicate the algorithm's prediction (171-173) .

As a part of the VAMPP-score, a statistical framework of multiple and pairwise comparisons were created. Firstly, for multiple comparisons, Kruskal-Wallis test (174) was chosen based on the properties of the input dataset. It was used for each gene-ISPP pair to determine if there is a significant pathogenicity score difference ($p<0.05$) among the three independent variant groups, Pathogenic, Benign, and Unknown.

Kruskal-Wallis is a non-parametric, distribution-free test used to determine if there are any significant differences between three or more groups of variables to be analyzed.

Subsequently, the Wilcoxon-Rank-Sum test (175) was applied for each gene-ISPP pair to detect a significant difference in scoring ($p<0.05$) among variant groups. These groups were “Pathogenic vs. Benign”, “Pathogenic vs. Unknown", and “Unknown vs. Benign”. Similar to Kruskal-Wallis (174) , Wilcoxon rank sum is a non-parametric, distribution-free statistical test. Due to the non-normal distribution of the input data used, it was chosen to be the test to compare variant groups in pairs. By utilizing Wilcoxon-Rank-Sum, it was aimed to determine how the variant groups differ from each other and/or whether or not they come from the same distribution.

The ISPPs with p-values (probability values) less than 0.05 were integrated into the VAMPP-score as a part of the weighted gene coefficient calculated for each gene-ISPP pair. In this context, the ISPPs with p-values less than 0.05 from multiple comparisons and ISPPs with p-values less than 0.05 from Pathogenic vs. Benign pairwise comparisons were included in the calculation with their p-values subtracted from 1 to reflect the power of the ISPP’s significance.

3.2.3 Section 2. Variant group distribution metrics

3.2.3.1 Proximity (*d*)

The proximity (*d*) estimation function (PEF) is based on the assumption of how the three predefined variant groups should be distributed by means of assigned pathogenicity scores by each ISPP. In this context, the means of each variant group for each gene-ISPP pair were calculated and integrated with the VAMPP-score, considering their ideal distribution (Figure 12).

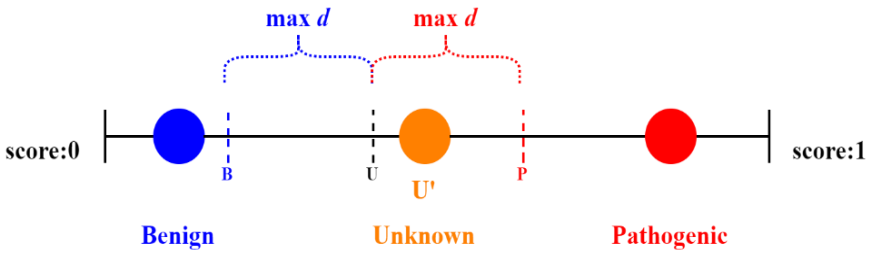

Figure 12. Ideal variant group distribution.

The figure shows the ideal variant group distribution for a gene based on the assigned pathogenicity scores from each ISPP. Filled circles represent the *ideal* means of the defined variant groups as Benign (blue), Unknown (orange), and Pathogenic (red). Dashed lines represent any variant group mean for Benign (B), Unknown (U), and Pathogenic (P) for an ISPP, respectively. "max *d*" represents the maximum distance, which is the numerical expression of how much the mean of the Unknown variant group diverges from and serves as an indicator of how it is positioned among the other two variant groups for each gene-ISPP pair.

To calculate *d*, a theoretical Unknown variant group mean (U') is defined as the median of pathogenic and benign variant group means for each gene-ISPP pair. Then, relying on the assumption that U' would be equidistant from the Pathogenic or Benign variant groups, the maximum distance (d_{max}) is calculated by subtracting the U' from the Pathogenic variant group mean (Figure 13).

$$\bar{U}' = \left| \frac{\bar{P}+\bar{B}}{2} \right| \qquad \mathbf{d_{max}} = \left| \bar{P} - \bar{U}' \right|$$

Figure 13. The calculation of the elements of the proximity estimation function (PEF).

The figure shows the main components of PEF.U' represents the theoretical mean of the Unknown group, with regard to actual Pathogenic and Benign group means (P' and B', respectively) considering the ideal variant group distribution (illustrated in Figure 12). 'd_{max}' represents the difference between the P' and U', indicating the maximum distance between the groups.

d_{max} is defined as the numerical expression of how the Unknown variant group mean is positioned among the Pathogenic and Benign variant groups for each gene-ISPP pair.

Finally, U' is subtracted from the actual mean of the unknown variant group (U) for each gene-ISPP pair and divided by d_{max}. The absolute resulting value is subtracted from 1 to be referenced as a significant prediction accuracy value and included in the final VAMPP-score calculation as *d* (Figure 14).

$$\mathbf{Proximity(d)} = 1 - \left| \frac{\bar{U} - \bar{U}'}{d_{max}} \right|$$

Figure 14. Proximity estimation function (PEF).

The figure shows the calculation of the proximity value, *d*. U' represents the actual mean of the Unknown group for a gene-ISPP pair, and the metric indicates the degree of the ISPP on how accurately positioning the Unknown variant group in between Pathogenic and Benign, regarding the ideal variant group distribution (Figure 12).

3.2.3.2 Mean difference

A score difference multiplier is another element of VAMPP-score, the final element included in M, described as *s*. *s* ultimately determines the sign of M for each gene-ISPP pair (Figure 15).

$$\mathbf{s} = \bar{P} - \bar{B}$$

Figure 15. Function for mean difference.

Figure shows the calculation of the *s*, where P and B represent Pathogenic, and Benign variant groups, respectively. *s* reflects the ISPPs performance on distinguishing Pathogenic and Benign variants and positioning them accordingly to the ideal distribution (Figure 12). *s* defines the sign of the final VAMPP-score since it checks whether the Pathogenic variants were scored higher than Benign considering the rank score range of 0-to-1.

In this context, the mean of the benign variant group is subtracted from the mean of the pathogenic variant group in consideration of the previously defined variant group distribution assumption (Figure 12).

3.2.4 VAMPP-score annotation

After calculating all the parametric values of *M* (Figure 11), VAMPP-score was calculated (Figure 16) for each variant available in dbNSFP v4.7A (114, 115) ,then scaled between 0 and 1 for ease of comparison.

$$\mathbf{VAMPP_{score}} = \frac{(1-P_{k_1})(1-P_{pb_1})(d_1)(s_1)(x_1)+(1-P_{k_1})(1-P_{pb_1})(d_1)(s_1)(x_1)+\cdots\cdots\cdots\cdots+(1-P_{k_n})(1-P_{pb_n})(d_n)(s_n)(x_n)}{n}$$

Figure 16. Function for VAMPP-score.

Figure shows the calculation of VAMPP-score for each variant. Here, *M* for the gene of interest (Figure 11) where the variant is located is used as an impactor, and is multiplied by the rank score of the ISPP of interest (x_1, x_2,,..x_n) for that variant. Addition is performed for the number of ISPPs where there is an available M for the gene-ISPP pair, and the sum is then divided by the number of the ISPPs included in the calculation for the gene where the variant of interest is located.

This operation was performed for each genomic variation in a total of over 84 million, available in the dbNSFP v4.7A (114, 115). Then, the calculated VAMPP-scores were scaled between 0 and 1 by stretching the values as the minimum value was made equal to 0, whereas the maximum value became 1 using a linear normalization (176).

3.2.5 Multidimensional clustering and correlation analyses

The results obtained from the multiple comparison tests were also used to determine gene-based ISPP consistency. In this context, for multidimensional clustering, principal component analysis (PCA) (177, 178) was performed for categorized ISPPs utilizing ClustVis v1.0 (179) . PCA helps to reduce the dimensionality of the data with a large set of variables, hence giving a generalized representation of the data by still keeping all the variables (177, 178) PCA was used to cluster ISPPs on their significance, and the unit variance scaling method was applied to rows (Mendeliome genes) for scaling, along with SVD with imputation as the PCA method. The ISPP performance on the Mendeliome gene list was also visualized as a heatmap ClustVis v1.0 (179), no scaling and transformation were applied to the initial

data, and raw p-values were observed in the original 0-1 scale. The ISPPs were hierarchically clustered utilizing Euclidean distance (180) , based on their predictions to observe the relationship between the prior categorization and prediction level.

ISPP performance and their variant-based pathogenicity prediction score correlation analyses were carried out. Three different similarity matrices for ISPPs were created and visualized utilizing Broad Institute's Morpheus (*https://software.broadinstitute.org/morpheus*): a similarity matrix created with ISPP p-values from multiple comparisons, and two other similarity matrices created with the ISPP rank scores for Pathogenic and Benign group variants. The Euclidean distance metric was used in similarity matrices. It uses the basis of Pythagoras theorem and measures the similarity between two recorded observations (180) by detecting the shortest distance between them. Here, it was used to detect the similarity between ISPPs upon their p-values from multiple comparisons and within Pathogenic and Benign variant groups upon rank scores from ISPPs.

The general distribution of the VAMPP-score throughout the genome was visualized using a histogram by Python v.3.10.12's Matplotlib (version 3.8.4) (181). In addition, VAMPP-score distribution for Pathogenic, Benign, and Unknown variant groups was plotted as a histogram, showing the variant count for each group utilizing Matplotlib.

4 RESULTS

4.1 Overview

As of March 2024, a total of 1162673 variants from canonical transcripts of 17712 unique genes were submitted to ClinVar (151) and analyzed in this study. Out of 1162673 variants, 57741 of them were classified as **Pathogenic** (23077 Pathogenic; 26206 Likely pathogenic; 8188 Pathogenic/Likely pathogenic), 94740 as **Benign** (28875 Benign; 55273 Likely benign; 10592 Benign/Likely benign), and 1010462 as **Unknown** (uncertain significance).

Out of 17712 genes analyzed, 3955 genes had at least one ISPP, which yielded statistically significant results according to the Kruskal-Wallis test (174) ($p<0.05$). 13757 genes yielded no results due to limited data availability.

Among the ISPPs, MetaRNN (145) yielded significant results for the highest number of genes (3718 genes) throughout the exome. MetaRNN was followed by ClinPred (143) with significant results for 3615 genes, and BayesDel_addAF (BayesDel with allele frequency added) (138) for 3547 genes, making them the top three ISPPs with the best performance (Figure 17.A).

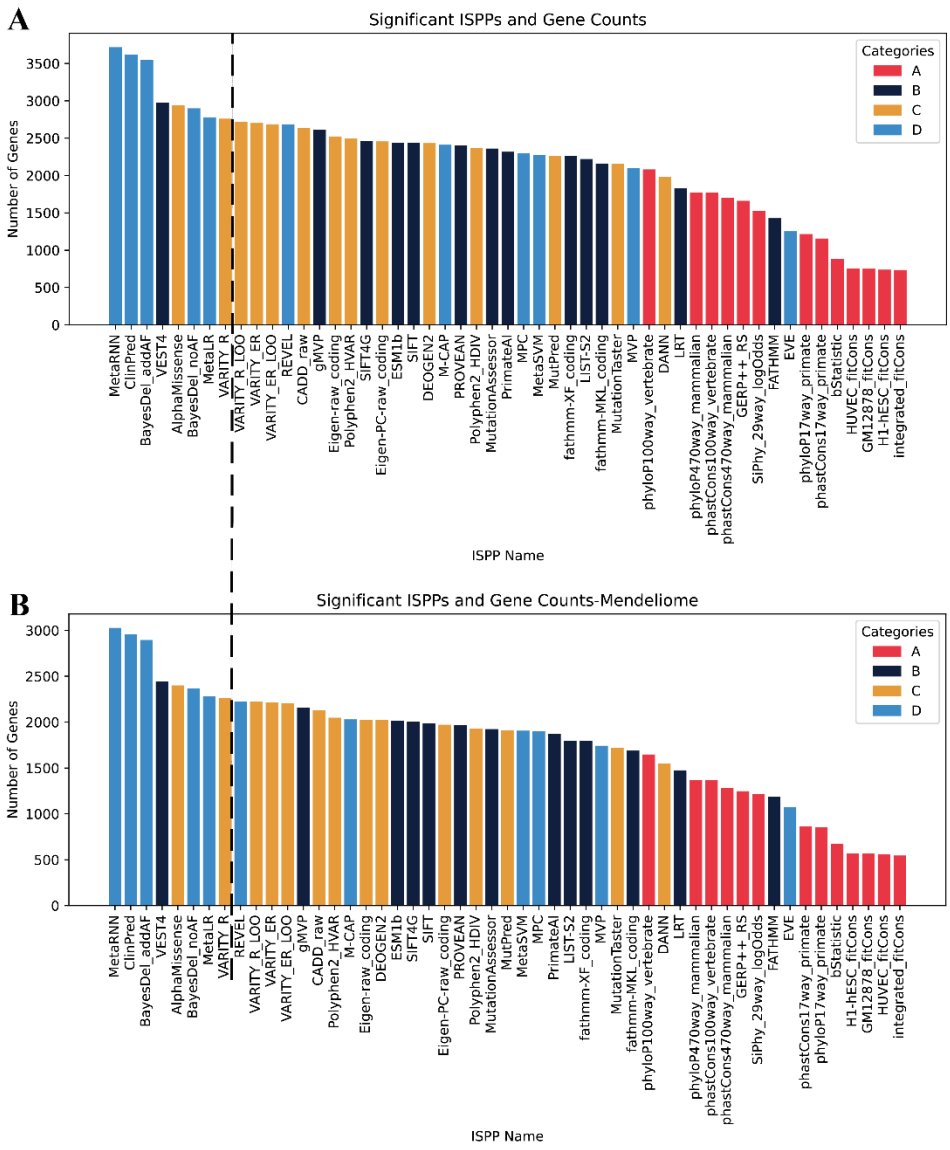

Figure 17. Significant ISPPs throughout the exome and gene counts.

The bar plot shows the ISPPs analyzed on the x-axis and gene counts on the y-axis. Gene counts represent the number of genes each ISPP found to be significant in multiple comparisons A. throughout the exome B. for the Mendeliome gene list. The bars in the plot are colored based on ISPP categorization, shown in the color legend. The top 8 ISPPs that are found significant for the most number of genes are the same for both gene lists, separated from the rest with the dashed black line.

In addition, using the total final variant dataset, a “Mendeliome variant list” was constructed by filtering the variants on disease-causing genes with Mendelian inheritance according to the HPO (104-106). In this context, 841674 variants on 4703 Mendelian inheritance-associated genes were analyzed. Out of 841674 variants, 55253

of them were classified as **Pathogenic** (22236 Pathogenic; 25085 Likely pathogenic; 7932 Pathogenic/Likely pathogenic), 67347 as **Benign** (20340 Benign; 37099 Likely benign; 9908 Benign/Likely benign), and 719074 as **Unknown** (uncertain significance).

Out of 4703 genes analyzed, 3193 genes had at least one ISPP, which yielded statistically significant results according to the Kruskal-Wallis test (174) ($p<0.05$). 1510 genes yielded no results due to limited data availability for genes of interest.

The ISPP performance on pathogenicity predictions was observed to be the same as throughout the exome (Figure 17.A) for the top 8 in the list (Figure 17.B). The top 8 ISPPs were listed as follows with the gene counts that they were found significant for the exome & Mendeliome: MetaRNN (145) with 3718 & 3027genes, ClinPred (143) with 3615 & 2960 genes, BayesDel_addAF (138) with 3547 & 2895 genes, VEST4 (130) with 2979 & 2441 genes, AlphaMissense (135) with 2944 & 2400 genes, BayesDel_noAF (138) with 2900 & 2368 genes, MetaLR (144) with 2779 & 2279 genes, and VARITY_R (VARITY for Rare) (127) with 2761 & 2264 genes.

Five out of eight ISPPs (MetaRNN, ClinPred, BayesDel_addAF, BayesDel_noAF, MetaLR belong to category D, 2 ISPPs (AlphaMissense and VARITY_R to category C, and 1 ISPP (VEST4) to category B (Figure 8).

The ISPPs evolved from the fitCons (136) algorithm were found to be significant for the least number of genes throughout the exome and Mendeliome for an average of 740 and 560 genes, respectively.

It was also shown that two ISPPs were separated from their fellow ISPPs in their categories. FATHMM (124) from category B was found to be significant for 1429 and 1183, in addition to EVE (125) from category D with 1254 and 1073 genes for exome and Mendeliome, respectively.

4.2 Multidimensional Analysis

4.2.1 ISPP significance on Mendeliome

The principal component analysis (PCA), performed with ClustVis v1.0 (179), showed the ISPP clustering for Mendeliome (Figure 18).

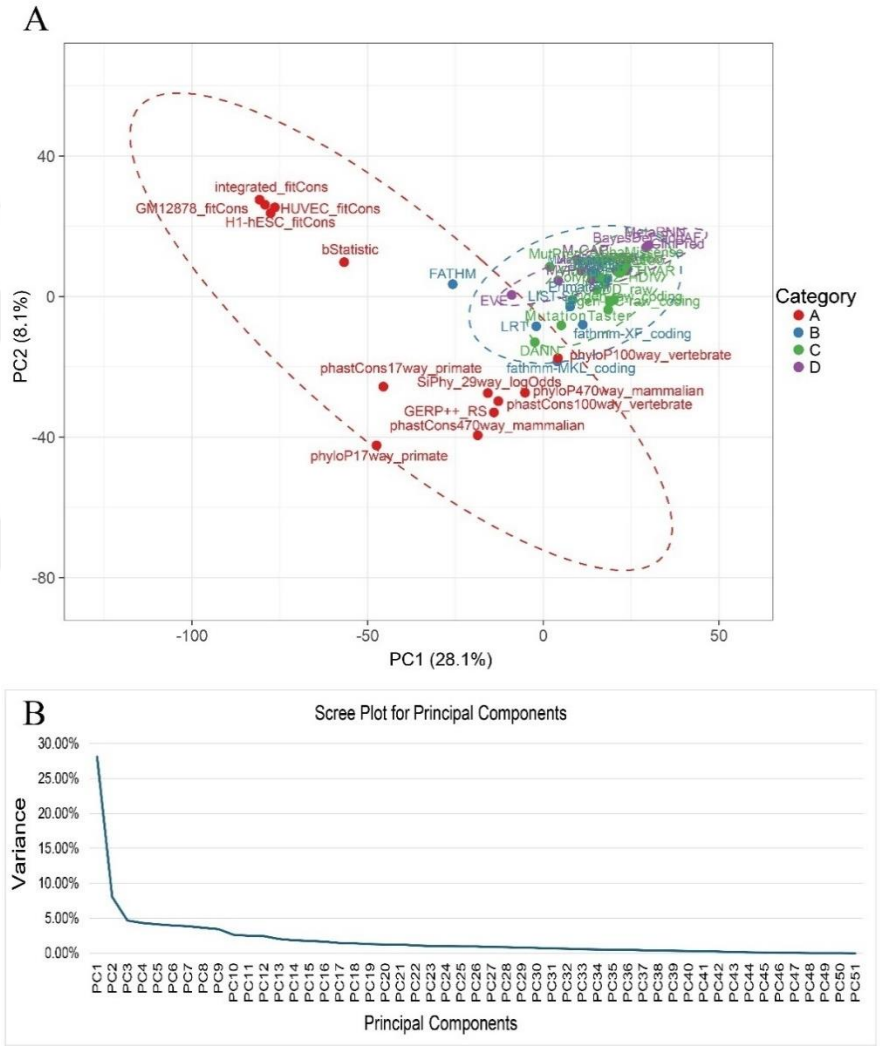

Figure 18. Principal component analysis on ISPPs for Mendeliome.

A. Principal component analysis (PCA) shows clustering of the ISPPs. Dotted ellipses in different colors represent the areas where ISPPs in the same category reside; colors represent ISPP categorization, as shown in the legend. The plot was created based on the first two principal components (PC1 and PC2). B. Scree plot shows the explained variances from PCA. Variance percentages in the y-axis show the volume of the data represented in the corresponding component.

ISPPs in Category A were the most dispersed. In contrast, Categories B, C and D were clustered together (Figure 18, Figure 8). In Category B, the FATHMM (124) was located further away and separately from its category (Figure 8, Figure 18). The mean p-value for FATHMM was calculated as 0.249, whereas the rest of the ISPPs in category B had p-values in the range of 0.067-0.249 (FATHMM with the p-value-maximum), with a mean of 0.135.

The heatmap (Figure 19) showed that MetaRNN (145), ClinPred (143) and BayesDel_addAF (138) showed the highest consistency. The variances among p-values for these ISPPs were calculated as 0.004, 0.005, and 0.006, respectively. The variance was shown to be decreasing from category A to category D ISPPs (Figure 8) with the means of 0.079, 0.045, 0.036, and 0.031, respectively. As the variances decrease, the variability among p-values for the ISPPs in each category decreases. This indicates the consistency between ISPPs that are grouped together.

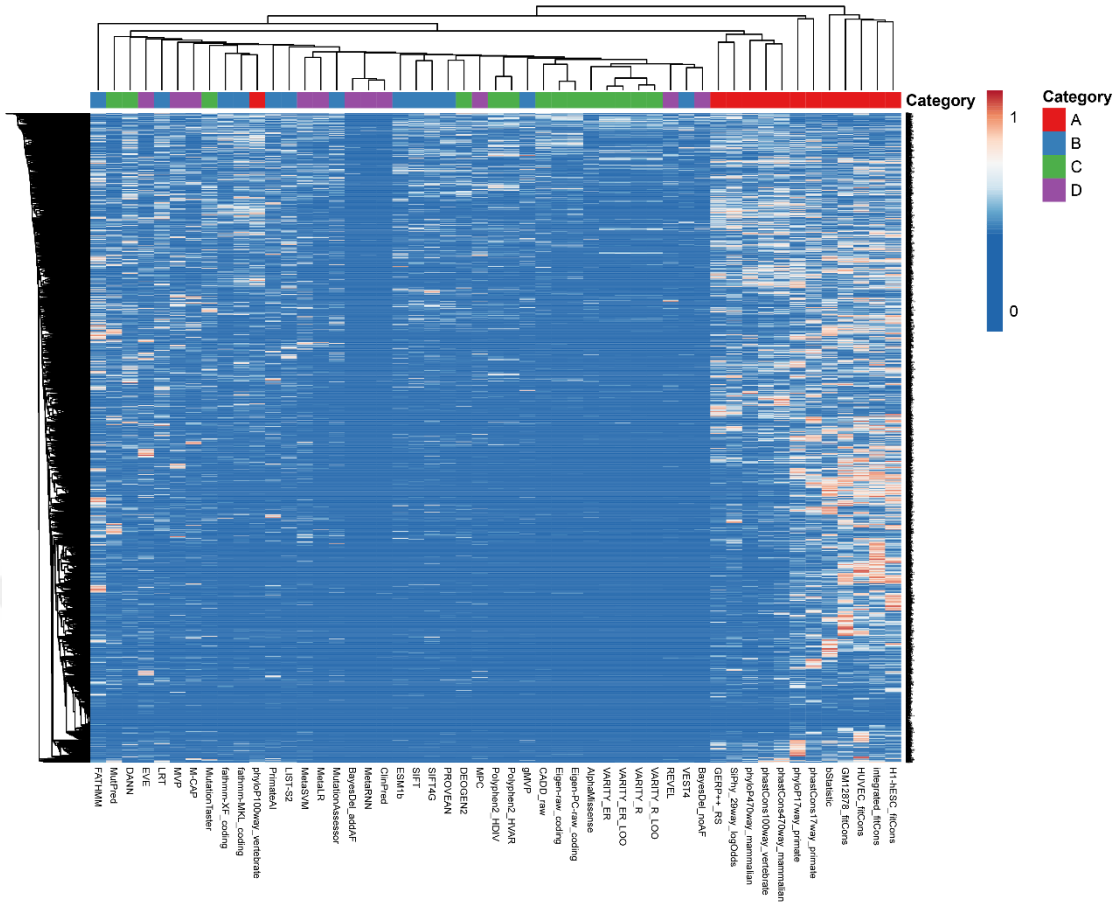

Figure 19. The ISPP correlation heatmap of multiple comparison p-values for the Mendeliome.

The correlation heatmap for Mendeliome of the p-values from multiple comparisons for ISPPs. The dendrogram on the top of the figure shows the hierarchical clustering using the Euclidean distance of the ISPPs, whereas the dendrogram on the left shows the hierarchical clustering result of the Mendeliome genes. ISPP categories are represented by four different colors: red, blue, green, and purple for Category A, B, C, and D, respectively. The color scale for the heatmap was set to be between 0 and 1, ranging from blue to red, which represents the p-values for each gene-ISPP pair. No transformation or scaling was applied to the p-values.

In addition, the dendrogram showed that the ISPPs that were categorized in the same group were clustered together, indicating that prior ISPP categorization executed in this study (Figure 8) corresponds to the prediction performance and ISPP similarity, starting from the tightest clusters with a couple of exceptions (Figure 19). The exceptions represent the clustering of three ISPP pairs that belong to different categories. Two of the three pairs were clusters of ISPPs from adjacent categories: phyloP100way-V (116) and fathmm-MKL (FATHMM with multiple kernel

learning)_coding (122) from categories A and B; DEOGEN2 (134) and PROVEAN (131) from categories B and C, respectively. The last pair was VEST4 (130) from category B and BayesDel_noAF (138) from category D.

FATHMM (Functional Analysis through Hidden Markov Models) (124) from the category B showed variability (Figure 18, Figure 19) with a variance of 0.077 from both category B ISPPs and gave more inconsistent results than the other scores that evolved from itself with variances of 0.056 and 0.044 for fathmm-MKL (122) and fathmm-XF (FATHMM with eXtended Features) (123), respectively.

Within category A ISPPs (Figure 8), pathogenicity scores evolved from phyloP (116) whereas scores evolved from the fitCons (fitness consequences score) (136) algorithm showed the most variability. GERP++_RS (Genomic Evolutionary Rate Profiling-rejected substitution) (92) had a smaller variance than 17_way_primate but showed more variability than 100-way_vertebrate and 470way-M approaches.

The variability within phyloP (116) & phastCons (117) algorithms was shown to follow the same pattern as most variable to least: 17_way_primate: 0.083 & 0.080, 470way-M: 0.064 & 0.079, and 100-way_vertebrate: 0.055 & 0.070. bStatistic (119) was shown to rank with having the second most variability with 0.088, after fitCons algorithms. The individual scores from fitCons (136) algorithm showed slight differences in the variability of their p-values among Mendeliome (integrated_fitCons: 0.0923; GM12878(human lymphoblastoid cells)_fitCons:0.0917, H1-hESC(human embryonic stem cells)_fitCons: 0.0936, and HUVEC(human umbilical vein epithelial cells)_fitCons: 0.0918).

A similarity matrix for the ISPPs was created with Broad Institute's Morpheus (*https://software.broadinstitute.org/morpheus*) by utilizing Euclidean distance to assess the relation over the proximity between p-values from multiple comparisons (Figure 20).

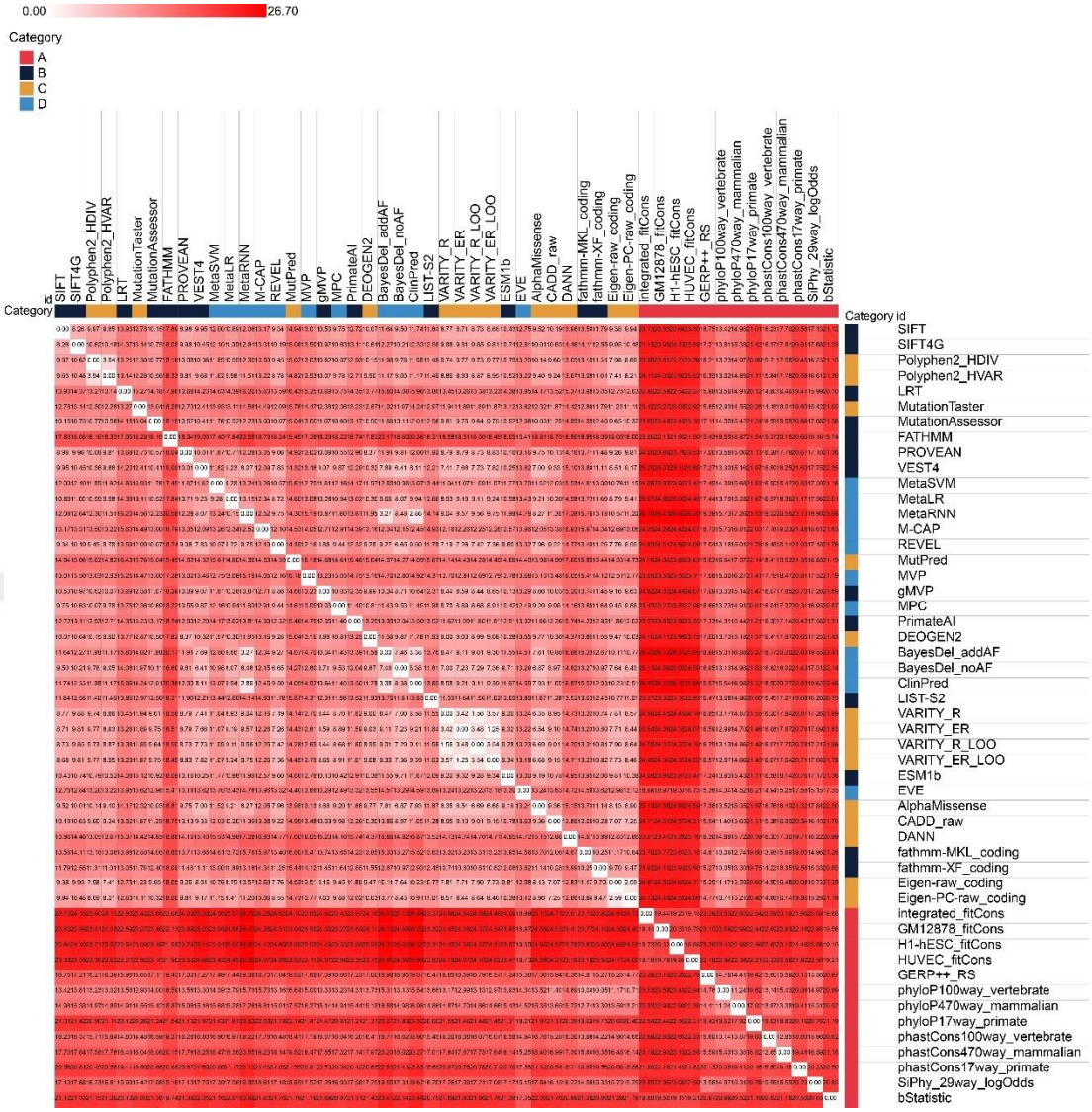

Figure 20. ISPP similarity matrix for the Mendeliome.

The similarity matrix shows the relation between ISPP predictions among Mendeliome genes based on p-values from multiple comparisons utilizing Euclidean distance, shown in each square. The color scale ranges in shades of red, darker shades represent increased distance between two points. ISPP categories are represented by four different colors, shown in legend.

The matrix (Figure 20) showed that ISPPs that evolved from a similar algorithmic approach yielded clusters as their p-values are more similar, hence there is less distance between (Figure 20). These approaches are listed as follows: PolyPhen2 (100), VARITY (127), and Eigen_coding (141). Also, ISPPs in Categories B, C, and D

yielded more similar results, (represented with the largest cluster with lighter shades of red, the upper left of the matrix), than category A ISPPs.

The largest clusters in the similarity matrix (Figure 20), which is interfered with by FATHMM (124), involve ISPPs that are in the top 34 most significant ISPPs (Figure 17) from categories B, C, and D (Figure 8). None of the ISPPs from category A was found to be in this cluster, suggesting that category A ISPPs differ from the other categories. ISPPs from category A have only given similar results within, yet still not as similar as the other categories. Category A cluster yielded results corresponding to ~20 units of Euclidean distance, where the maximum distances identified in the matrix were circa ~25, which corresponds to the dissimilarity level of the fitCons (136) algorithms from the other ISPPs, and circa ~20 within.

In addition, it was observed that ISPPs in the same category with different approaches, MetaRNN (145), ClinPred (143), and BayesDel_addAF (138), have the most similar and close prediction results among the independent scores. ClinPred-MetaRNN yielded the most similar results with 2.86 distance units, followed by MetaRNN-Bayesdel_addAF with 3.27, and ClinPred-BayesDel_addAF with 3.38, which correlate with the previous results (Figure 17, Figure 19).

Moreover, FATHMM (124), was shown to yield variability and dissimilar results (Figure 20) compared to its ISPP category (Figure 8). FATHMM has given results that are variable and dissimilar in general, however comparably closer within category B. Yet, considering other in-category-ISPP pairwise distances, FATHMM was shown as an ISPP with variable prediction.

4.2.2 Scoring correlation on Mendeliome

The prediction accuracy for each ISPP on variant pathogenicity was further analyzed based on rank scores from the ISPPs. In this context, both Benign (Figure 21) and Pathogenic (Figure 22) variants on Mendeliome genes were examined utilizing Broad Institute's Morpheus (*https://software.broadinstitute.org/morpheus*).

The following similarity matrices focus on the relation between the assigned scores by ISPPs for Mendeliome genes for Benign and Pathogenic variant groups. The purpose behind this approach was to detect the similarity among ISPPs based upon their raw ranked scores rather than p-values and observe their prediction performance. In this context, although the accuracy of variant prediction and the performance of assigning the correct score to the variant cannot be directly demonstrated by examining only the similarity matrix, it can be determined which ISPPs make similar predictions for the variants with known clinical significance.

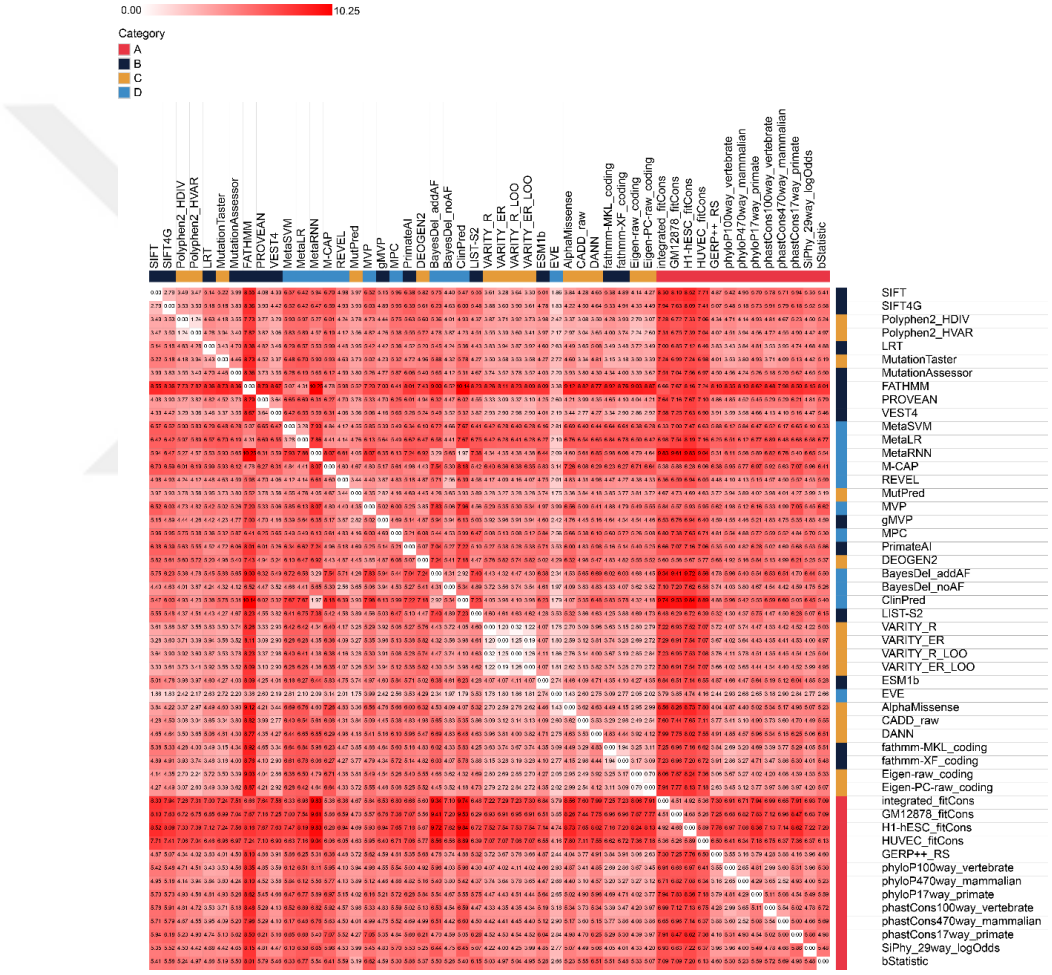

Figure 21. ISPP similarity matrix for benign variant assessment.

Heatmap shows the similarity for the ISPPs regarding the Benign variant group, obtained via Euclidean distance metric. The color scale ranges in shades of red, darker shades represent increased distance between two points. ISPP categories are represented by four different colors, shown in legend.

The heatmap (Figure 21) showed several clusters of ISPPs with similar scores assigned for the variants in the Benign group. ISPPs that evolved from the same algorithmic approaches were clustered together in lighter shades of red, due to minimal values through Euclidean distance. In this context, the VARITY_ER (for Extremely Rare-pair (127), with 0.19 units yielded the most similar results and identified as the ISPPs with the most similar results in Benign variant scoring. This pair was followed by Eigen algorithms (141) with 0.7, VARITY-R-ER (127), pair with 1.2, PolyPhen2 algorithms (100) with 1.24, and VARITY-R/ER-LOO (Leave One Out) pair (127), with 1.26.

Additionally, ISPPs that belong to the same category were mostly observed to be clustered together with decreased distance units. It was also detected that category B ISPP FATHMM (124) was the predictor with the most dissimilar results, which correlates with the previous results (Figure 20). In spite of this dissimilarity, FATHMM showed similarity with ISPPs under categories C and D, interestingly. The ISPPs are listed as follows: MetaSVM (144), MetaLR (144), M-CAP (146), REVEL (147), MVP (137), EVE (125) from category D, and MutPred (132) from Category C. It was previously shown Category B as a cluster of ISPPs was close to the clusters C & D (Figure 18).

ISPPs that evolved from fitCons algorithm (136) were shown to make relatively similar predictions within (decreased distance in units compared to other clusters) (Figure 21), but differently from its ISPP category (Figure 8) and the rest of the ISPPs. In contrast, it was shown that scores from phyloP (116) and phastCons (117) algorithms yielded more similarity in benign variant score assignment within, and among ISPPs, considering the overall distance (Figure 21).

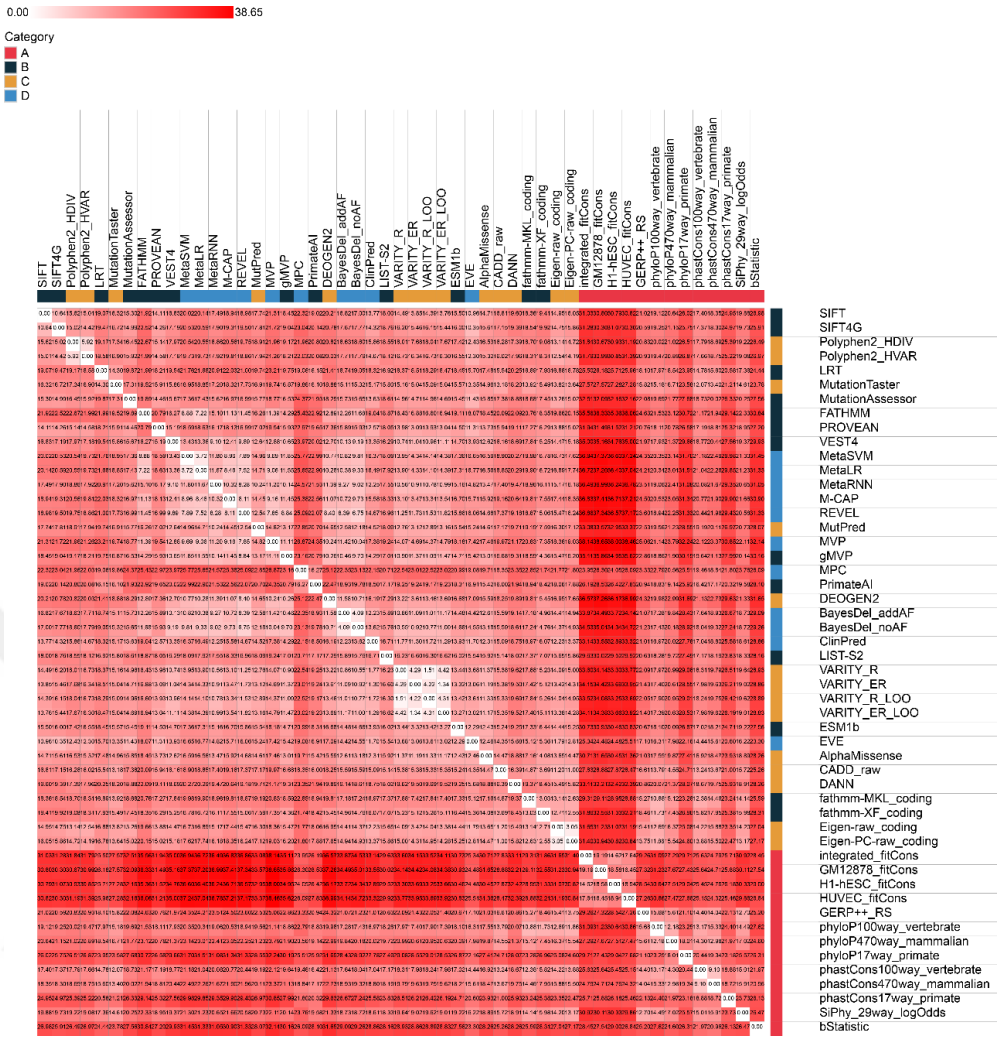

Figure 22. ISPP similarity matrix for pathogenic variant assessment.

Heatmap shows the similarity matrix for the ISPPs regarding the Pathogenic variant group, utilizing Euclidean distance. The color scale ranges in shades of red, darker shades represent increased distance between two points. ISPP categories are represented by four different colors, shown in legend.

The heatmap above (Figure 22) showed ISPP clusters that have similar results in assigned scores to the variants in the Pathogenic group. The clusters formed were wider than the ISPP clusters for the Benign variant group (Figure 21). These wider clusters were shown to involve ISPPs from various categories (Figure 8).

The heatmap (Figure 22) showed that there were no category-specific clusters of ISPPs (Figure 8). However, the ISPPs that evolved from the same algorithmic approaches, which were mentioned before several times in Results Section 4.2.1 and

4.2.2), yielded similar scores for Pathogenic variants represented with significantly less distance units.

ISPPs that evolved from fitCons algorithm (136) were shown to have relatively more similar results within (~14-18 Euclidean distance units), compared to their similarity with the rest of the ISPP (represented with ~25-34 units, closer to the maximum defined in the matrix as 38.65).

Furthermore, differing from the results of the Benign variant similarity matrix (Figure 21), MPC (148) and PrimateAI (133) yielded results that are dissimilar from their own categories (Figure 8) and overall, from the rest of the ISPPs. In addition to those, FATHMM (124) yielded results that interfere with the wide cluster of similar ISPPs on the upper left of the matrix, with the average Euclidean distance units of 30, following MPC (148) and PrimateAI (133).

fitCons algorithm (136) ISPPs were shown to cluster together similar to the results from the Benign variant group (Figure 21). Additionally, as shown in the previous figure (Figure 21), they had increased distance values within category A and the rest of the ISPPs, yielding dissimilar results for the Pathogenic variant scores.

Moreover, the clustered category A ISPPs were disrupted by phyloP (116) and phastCons (117) algorithms' 17way-P approaches, in addition to fitCons (136) group. pyhloP17way-P and phastCons17way-P yielded dissimilar results from the other ISPPs, the ISPPs in category A. To elaborate on the phyloP and phastCons algorithm results within, phyloP17way-P had the Euclidean distance unit of 23.25 for phyloP100way-V, and 18.01 for phyloP470way-M. phastCons17way-P had the Euclidean distance unit of 18.88 for phastCons100way-V, and 18.72 for phastCons470way-M. These distance units were represented as the smallest values for both 17way-P approaches across all ISPPs for Pathogenic variants.

4.2.3 VAMPP-score distribution

In total, 23464253 variants from dbNSFP v4.7A (114, 115) were able to get annotated with VAMPP-score. VAMPP-score's distribution throughout the exome for all the missense alteration possibilities (Figure 23.A). Its distributions for Pathogenic and Benign variant groups were visualized as a histogram (Figure 23.B) combined with a rug plot for the representation of individual data points using the Matplotlib library (181) in Python (version 3.10.12).

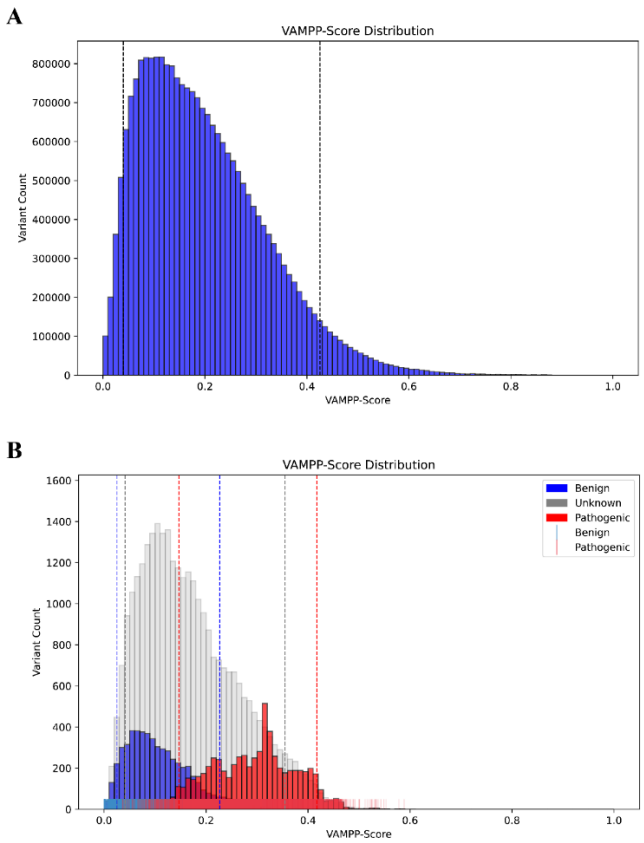

Figure 23. VAMPP-score distribution.

A. Histogram shows the distribution of VAMPP-score in a total of 23464253 variants from dbNSFP v4.7A (114, 115). The quantiles for 0.05 and 0.95 at score points of 0.04003 and 0.42583, respectively (black dashed line). B. Histogram shows the distribution of VAMPP-score throughout the human exome for Benign, Unknown, and Pathogenic group variants. The bar plot shows the VAMPP-score distribution for Benign, Unknown, and Pathogenic variant groups with y-axis representing the variant counts. The rug plot represents individual data points at the bottom of the histogram, representing each variant in Benign and Pathogenic groups with its corresponding VAMPP-score. Quantiles for Benign variants were calculated as 0.0252 for 0.05 and

0.226 for 0.95 (blue dashed line), quantiles for Unknown variants as 0.041 and 0.354 (gray dashed line), and quantiles for Pathogenic variants as 0.147 for 0.05 and 0.41 for 0.95 (red dashed line).

VAMPP-score's distribution (Figure 23.A) does not follow a normal distribution and was shown to be strongly skewed to the right, with a high proportion of the low VAMPP-scores. The long tail of the distribution is stretched towards higher values, closer to 1. The distribution peaks at the lower scores, around 0.1-0.2. The histogram indicates a non-normal, right-skewed distribution with a mode at the lower values. The presence of the long right tail suggests that while most variants are grouped towards lower scores, the values are dispersed along a wide range.

The histogram that shows the Pathogenic and Benign groups together (Figure 23.B) reveals the distinction of the groups from each other. The benign variant distribution shows accumulation on the lower scores, peaking around 0.1. This distribution is unimodal representing one peak, and it is skewed to the right, with a tail towards the higher scores. The pathogenic variant distribution was more spread throughout the scores, with a peak of around 0.3. The distributions showed that Pathogenic variants extend more to the right towards higher scores compared to the Benign variants. Even though the pathogenic variants also displayed a unimodal distribution with one peak, it was less symmetrical than the benign variant distribution. The tail of the pathogenic variant distribution extends to the right, towards high VAMPP-scores, also shown to be more prominent than the benign group distribution. indicating pathogenic variants with higher VAMPP-scores were less in number than the benign variants.

The two variant distributions were overlapping in the range approximately from 0.1 to 0.3, also shown with the rug plot at the bottom of the figure representing each variant in the data set. Out of 3245 pathogenic variants that are in the 0.1-0.3 range, which represents the overlap, 1901 variants were under the review status of criteria provided, single submitter. It was followed by no assertion criteria provided (704), criteria provided, multiple submitters, no conflicts (576), and reviewed by expert panel (64). In the case of benign variants in the same range, out of 2585 variants, 1964 had

the review status of criteria provided, single submitter. The rest followed as: criteria provided, multiple submitters, no conflicts (533), reviewed by expert panel (50), and no assertion criteria provided (38).

Comprehensive analysis on ISPP prediction performances for the variants with known clinical significance utilizing PCA and similarity matrices has identified clusters of ISPPs based on their p-values and pathogenicity scores. The ISPP clusters were shown to be aligned with the novel categorization introduced (Figure 8) and revealed insights into the predictive capabilities of the ISPPs considered. Distinct patterns of similarity showed the closely aligned and diverged ISPPs. Also, a novel pathogenicity prediction score, VAMPP-score, that incorporates the essence of these relations and predictive accuracies of the ISPPs was introduced. VAMPP-score offers a refined approach to evaluate genetic variants, integrating the gene-specific prediction performance of ISPPs.

5 DISCUSSION

In variant interpretation, various prediction algorithms are currently in use, and each is designed to aid the diagnostics by providing insights on deciding the clinical significance. Even though these algorithms differ in approach and targets, they use similar population and clinical data. As this data among algorithms ensures persistence for accurate prediction, it may also create limitations, particularly by supporting inaccurate results if they were to occur and by generalizing results across diverse populations and genetic events. This raises critical importance for users to recognize the limitations and biases that may be present in the datasets and, therefore, in the results from these predictors. Such awareness is essential, especially when dealing with rare disorders and novel variants, where the input data may not reflect the situation for the case of interest.

In this study, we aimed to analyze the prediction accuracy of the most frequently used ISPPs using statistical approaches to strengthen variant interpretation in diagnostics. In this context, a statistical framework was created to evaluate ISPPs in multiple and pairwise comparisons. This statistical framework was further supported with proximity functions to assess the performance of the ISPPs on distinguishing pathogenic and benign variants, resulting in a new *in-silico* pathogenicity prediction score called VAMPP (Variant Analysis with Multivariate Pathogenicity Prediction)-score.

A database with variant information used in this thesis was created utilizing DuckDB. DuckDB is an in-process SQL on-line analytical processing (OLAP) database management system (164). Unlike traditional RDBMS (Relational Database Management System) and NoSQL databases, DuckDB excels in handling large datasets with fast querying capabilities, with an interface offering ease of use. Here, with the use of DuckDB, we created a database structure that is easy to manipulate and update, which allows high-performance data analysis (164, 182).

ClinVar (151) variant summary data get used here, get weekly updates, and new variant submissions are added. On the other hand, dbNSFP (114, 115) and Ensembl (165) are less likely to be updated that frequently. Hence, dbNSFP and Ensembl contents were implemented into the database as main tables. In contrast, the variant summary list was integrated as a temporary table, followed by query extraction and VAMPP-score calculation, which was planned to occur monthly.

ISPPs have been categorized under three groups as missense prediction, splice site prediction, and nucleotide conservation prediction by ACMG (10). ACMG also defined the targets of various ISPPs in their categorization. On the other hand, the ISPPs enrolled in this study were categorized into four groups based on prediction approaches (Figure 8) by individually investigating ISPPs. These groups were shaped after a comprehensive literature review and creating a sensible ground, with an effort not to generalize the aspects while trying to categorize under main features.

ISPPs in Category A had the least number of genes in which the ISPPs were found significant, which may indicate that the nucleotide-based predictions raise limitations. ISPPs in this category use multiple sequence alignments (MSA) to identify evolutionarily conserved genomic regions. The MSA includes non-coding areas, which are often rich in repeat sites and highly variable. If these regions can be removed and positive selection inference might be improved, which can weaken the effects of the accuracy loss (183). Also, gaps in the genomic sequences often lead to misalignments, and intronic regions might introduce gaps or be absent in some (182). Therefore, careful monitoring is essential when conducting such analyses and applying the derived scores in variant analysis. However, for the non-coding regions of the genome, the evolutionary conservation analysis still must be determined by these ISPPs.

phyloP (116) and phastCons (117) have a slight difference in algorithmic approaches. phastCons gives a score for a nucleotide upon its probability to belong to a conserved genomic element, whereas phyloP is the -log(p-value), indicating the evolution rate for a genomic region (184, 185). In addition, the differences in

phylogenetic species used can lead to variations in results. As 100way-V list (184, 185) represents a broader phylogenetic scope, it allows for more conserved elements and evolutionary patterns to be identified across a diverse set of animals, aiding conserved region detection across lineages. In contrast, other lists (17way-P and 470way-M (185, 186), despite having more species, focus on more specific groups, providing a different perspective on evolutionary conservation.

It was noted that the most consistent predictors, referred to as top 8 ISPPs (Figure 17), are of categories C and D, with one exception from category B, VEST4 (Variant Effect Scoring Tool) (130). VEST4 is presented as a tool that is supervised machine learning-based. It is trained to prioritize rare missense variants that are likely to cause diseases. The VEST relies on a gene-level approach, where variants are analyzed across exome data from diseased individuals to rank the genes upon disease-causing potential. So, even though it focuses on the classification of the missense variants hence grouped under category B, its comprehensive methodology and gene-based approach makes VEST4 yield consistent predictions along with the ensemble ISPPs of category D.

Intriguingly, 5 out of 8 (MetaRNN (145), ClinPred (143), BayesDel_addAF (138), BayesDel_noAF (138), MetaLR (144)) belong to category D, which further indicates the increased accuracy of ensemble predictors. This correlates with the findings of VarSome (186) , which recently implemented an approach for ACMG-PP3 criterion evaluation (163). MetaRNN (145) was shown to be the most accurate ISPP, among the ones that are also analyzed in this thesis, with 91% accuracy for classifying ClinVar (151) variants, by Varsome (186) using the methodology described by Pejaver et al. In addition, the rest of the top 5, except ClinPred (143), since it was not included, follows the same pattern in accuracy levels as our results (187). So, we recommend using these top 8 ISPPs for general variant interpretation practices for Mendeliome, based on our results.

The gene count was shown to be significantly decreased for the last 4 ISPPs that were grouped under category A (Figure 8, Figure 17), with all of them evolving from

the same algorithm: fitCons (136).This suggests that different approaches, such as analysis on different cell lines for variant effect prediction were not improving the performance of the algorithm. fitCons showed the lowest performance on variant prediction accuracy, both through the ISPPs analyzed and within its category. This correlates with the clustered separation of the fitCons algorithms in PCA (Figure 18).

fitCons (136) is an algorithm that relies on fitness consequence that can be interpreted as an evolution-based measure of potential genomic function. It was mainly raised to solve the problem of the identification of noncoding functional elements, even though the scores were also considered to be elevated for both coding and noncoding functional regions in the genome. fitCons scores were previously described as elevated in known coding and noncoding functional elements but with considerably better sensitivity for the noncoding regions. This correlates with our results and suggests that fitCons scores should be investigated focusing on the noncoding regions of the human genome for a more accurate performance evaluation.

As observed in the significant ISPP counts (Figure 17), category A ISPPs were the ones that showed the lowest performance with less number of genes. However, the smooth list of category A was interfered with by two ISPPs from other categories: FATHMM (124) from category B and EVE (125) from category D. These also happened to have separated from their category clusters in PCA (Figure 18).

For the Mendeliome, ISPPs in category C were tightly clustered (Figure 18) compared to other categories, showing their similarity in pathogenicity prediction accuracy. This is relevant since they consider biological function and properties for prediction, and Mendeliome consists of disease-causing genes.

EVE (125) from category D was separated from its cluster (Figure 18) and it was also found to be significant for the least number of genes throughout the whole exome and Mendeliome compared to other ISPPs that belong to the same category (Figure 17). Even though EVE integrates several parameters and uses a combined scoring algorithm, it roots from an evolutionary conservation model. This might indicate its

similarity with the algorithm that uses evolutionary conservation (Category A, Figure 8).

Moreover, as mentioned before in the case of MetaRNN (145) prediction accuracy (188), FATHMM (124) was defined with relatively low accuracy in variant classification (188), which was less than many other ISPPs in addition to the FATHMM-evolved algorithms fathmm-MKL (122) and fathmm-XF (123), in this order. As this correlates with our results (Figure 17, Figure 18), it also indicates the improvements that came with additional algorithmic features implemented to FATHMM and correlates with our results.

Also, several ISPPs in the category D integrated supervised machine learning algorithm (REVEL (147), BayesDel (138), M-CAP (146)), where the learning model has a baseline understanding of what the correct output values should be, which increases the accuracy.

Narrow clusters of ISPPs for Benign variants, and more extended clusters with many ISPPs from different categories for Pathogenic variants suggest that ISPPs can assign and distinguish benign variants better than pathogenic variants. Because the expected clustering of ISPPs upon categories was not significant for Pathogenic variants. ISPPs in general, might encounter difficulties in assigning variants pathogenicity, whereas calling them benign effortlessly. However, one should keep in mind that similarity matrices do not directly reflect accurate pathogenicity prediction of correct score assignments. They only show the relation between ISPP results utilizing a distance metric. Further analysis on the ISPPs is required for the assessment of their pathogenicity prediction performance on variant-score level.

Out of over 84 million variants in dbNSFP v4.7A (114, 115), only 23464253 variants were able to get annotated with VAMPP-score, due to limited data in ClinVar (151) which can arise from several reasons. Firstly, benign variants are not accepted as of clinical significance and are not reported as causal in diagnostics. So, they are not paid attention to as much as pathogenic variants and more often than not submitted

to variant databases. Moreover, there might be variants that are not observed in the population due to severe consequences caused by their incompatibility with life. Variants occurring in the conserved regions of the genome, causing severe effects and dysfunctions, may be eliminated through the course of evolution (92, 188). Also, random variations that happen to occur in these regions may cause prenatal or early death, leading to their unrecognition.

VAMPP-score's distribution showed accumulation in a specific range. However, this would not determine the inaccuracy of VAMPP-score, since VAMPP-score was developed using readily available ISPP scores. This only indicates that ISPPs that are frequently used in practice might have misleading results if the numerical value is considered alone. In spite of the accumulation, the wide range of scores assigned by VAMPP-score indicates potential characteristic differences for variants with differentiating scores.

VAMPP-score demonstrated a distinct scoring between the Pathogenic and Benign variant groups. Pathogenic variants expectedly had higher scores, which implies that VAMPP-score successfully integrates gene-ISPP pairs to its algorithm. Although, it is noteworthy to mention, there is a degree of uncertainty, especially in the overlapping variants for the middle range of scores. As the numbers of variants with their corresponding ClinVar (151) review status groups indicates that these variants may be considered as VUS depending on the case, even though they were submitted with specific clinical significance.

VAMPP-score's distribution reflects a critical point in variant pathogenicity interpretation. Each ISPP might have different scoring ranges and significant thresholds for pathogenicity assignment. Although the scores are ranked between 0-1, ISPPs may have different distributions and cumulative scoring in terms of numerical values. Therefore, it is very important to evaluate the performance of each ISPP specifically for each gene. In addition, the proximity estimation function and mean difference integrated into the VAMPP-score calculation are critical in minimizing the effects of the problems arising from these unequal distributions.

This study has the potential to directly impact patient care in genetic diagnostics with further improvements. By assessing the accuracy of ISPPs at a certain level, the rate of misdiagnosis and uncertainties can be reduced in genetic diagnosis, particularly when dealing with variants of uncertain significance (VUS).

The ACMG/AMP 2015 standards and guidelines for variant interpretation (10) play a crucial role, primarily appealing to the end-users – clinicians and genetic analysts. The scoring metric of ACMG (10) serves as a provider in the interpretation process, rather than being the ultimate decision-maker. This highlights the importance of integrating the scoring metrics with clinical judgment and expertise. In this context, the clinician or analyst is encouraged to consider a broader spectrum of factors beyond the computational outputs. These may include approaching each case individually focusing on the patient's phenotype and the clinical presentation, known patterns of inheritance for the gene of interest, known clinical representation of the variant in question, and possible genetic events that can occur.

Pathogenicity prediction algorithms and scoring systems are invaluable tools in the field of genetics, their use requires an understanding of principles. In the context of ACMG/AMP 2015 guidelines (10), computational evidence on pathogenicity (PP3) and benignity (BP4) have specified criteria. Both require the agreement of ISPPs, and even then, it is recommended to be used as a supporting level evidence due to known limitations of these computational predictors.

VAMPP-score highlights the best performing ISPPs, with its robust mathematical foundation. In this way, the ACMG (10) PP3 criterion can be proposed to be upgraded from supportive to moderate, relying on the VAMPP-score predictions within a subjective evaluation. Likewise, BP4 criterion can also be evaluated with VAMPP-score. Upgrading or downgrading these criteria's level can affect the final decisions on variant classification. Variants that were previously classified as VUS might be reclassified as likely pathogenic if appropriate criteria are met according to the guidelines (10). This can be particularly important for candidate variants compatible with the phenotypic outcome and thought as causal by the clinician; however, they are

still classified as VUS. The reclassification might lead to a genetic diagnosis, and/or support the treatment process, especially for rare disorders.

In spite of this VAMPP-score utilization, variants can still be classified as VUS due to limited data availability and/or insufficient evidence (10). Then, functional studies are recommended for the variant(s). However, the VAMPP score can be used as assistance in prioritizing candidate variants chosen for functional analysis supported by expert opinions.

A key decision was made to not limit the variants based on their review status when refining the variant dataset (151). This choice was driven by the aim to enhance the statistical robustness of the study. However, it is noted that this might raise some concerns upon "trusting" the variant's submitted clinical significance. As mentioned previously, a significant proportion of each pathogenic and benign variant in the overlapping region of the VAMPP score was represented with a single submitter. Since this may lead to skepticism of the clinical significance of the variant, we plan to implement a star-system to VAMPP-score. In this context, as the submitter number increases, the score will be represented with more *stars*, indicating a confidence level in the score and the variant's effect on the phenotype.

VAMPP-score represents a significant advancement in variant prioritization, yet it is important to acknowledge its limitations. In the current version of the VAMPP-score, there is no sub-setting of genes based on their related phenotypic features and their known inheritance patterns. Also, the population allele frequency information is yet to be implemented. These features together will aid to further evaluate variants upon their prevalence and representation in the population (monoallelic or biallelic). Another shortfall of VAMPP-score is the lack of protein structure and function information, considering the effects of the missense variants. Utilizing the above-mentioned specifications, we plan to construct an improved statistical framework and an extended network with more variables, which we believe would make VAMPP-score a go-to pathogenicity assessor in clinical diagnostics.

6 CONCLUSION

The advancement in sequencing technologies extended the perspective on genetic diagnostics approaches, particularly in Mendelian disorders. However, the complexity of genetic interactions and the interpretations of the missense variants have displayed challenges. The introduction of VAMPP-score, a novel *in-silico* pathogenicity prediction (ISPP) score, proposes an approach and marks an improvement in variant prioritization. By taking advantage of the statistical framework, VAMPP-score addresses the need for a more accurate interpretation of missense variants, which has been an existing dilemma in genetic diagnostics.

The ISPPs used in variant interpretation has been a double-edged sword. While these predictors assist experts in genetic diagnosis, they are often represented with conflicting results, which may lead to challenges in clinical decision-making. The inconsistency of these computational models stems from their varied methodologies and underlying databases.

The most frequently used ISPPs were assessed upon their prediction performance based on scores reached through dbNSFP (114, 115)upon variants with known clinical significance from ClinVar (151). ISPPs were categorized into four groups as regards to their prediction algorithm. Statistical analysis showed that only one-fifth of the available genes had results, due to limited data availability. Moreover, multidimensional clustering and correlation analyses demonstrated that ISPPs were clustering upon their categories, while the predictors that focus on nucleotide-level evolutionary conservation had the most unsteady clustering indicating their inconsistent results within and among other categories. Also, ISPPs had more ability to distinguish benign variants rather than pathogenic variants. In the case of benign variants, several narrower bundles of ISPPs under -mostly- same categories were observed to have similar predictions.

VAMPP-score’s distribution throughout the human genome was observed to be cumulating in a specific range due to VAMPP-score’s design that uses the available

ISPPs and their scores. This indicates that ISPPs might have misleading results on numerical values. VAMPP-score demonstrated a distinction between Pathogenic and Benign variant groups, with Pathogenic variants annotated with higher scores. This implies that VAMPP-score successfully integrates gene-ISPP pairs to its algorithm based on their variant group distinction performances.

It is noteworthy to mention that, with this thesis:

- ISPPs that are frequently used in variant interpretation were thoroughly analyzed upon their prediction performances with statistical approaches.
- Eight ISPPs with the most significant gene-specific results throughout exome and Mendeliome have been identified and evaluated further with downstream analysis. These ISPPs were recommended to assess variant pathogenicity in diagnostics and research.
- A novel ISPP was developed utilizing individual ISPPs upon their gene-specific performances, which can assist analysts and clinicians in diagnostics, especially when interpreting VUS in rare and undiagnosed diseases where every piece of information is critical.

As we continue to work on the complexities of genetic diseases, improvements like VAMPP-score will be essential to build a stronger connection and network between the technology and clinical application. With the VAMPP-score's utilization of the best gene-ISPP pairs, it is believed to be an advancement in missense variant interpretation. Given the statistical basis, it can be a decision-maker for a variant's pathogenicity, assisting clinicals and experts on using the ACMG criteria (10) more effectively and in a more adaptable way.

7 REFERENCES

1. Dixon-Salazar TJ, Silhavy JL, Udpa N, Schroth J, Bielas S, Schaffer AE, et al. Exome sequencing can improve diagnosis and alter patient management. Sci Transl Med. 2012;4(138):138ra78.
2. Fernandez G, Yubero D, Palau F, Armstrong J. Molecular Modelling Hurdle in the Next-Generation Sequencing Era. Int J Mol Sci. 2022;23(13).
3. Rabbani B, Mahdieh N, Hosomichi K, Nakaoka H, Inoue I. Next-generation sequencing: impact of exome sequencing in characterizing Mendelian disorders. J Hum Genet. 2012;57(10):621-32.
4. Capriotti E, Ozturk K, Carter H. Integrating molecular networks with genetic variant interpretation for precision medicine. Wiley Interdiscip Rev Syst Biol Med. 2019;11(3):e1443.
5. Sullivan JA, Schoch K, Spillmann RC, Shashi V. Exome/Genome Sequencing in Undiagnosed Syndromes. Annu Rev Med. 2023;74:489-502.
6. [2024-01-08]. NHGRIA. Missense Mutation. Genome.gov.: National Human Genome Research Institute. ; [Available from: https://www.genome.gov/genetics-glossary/Missense-Mutation.
7. Iqbal S, Pérez-Palma E, Jespersen JB, May P, Hoksza D, Heyne HO, et al. Comprehensive characterization of amino acid positions in protein structures reveals molecular effect of missense variants. Proc Natl Acad Sci U S A. 2020;117(45):28201-11.
8. Hormozdiari F, Kichaev G, Yang WY, Pasaniuc B, Eskin E. Identification of causal genes for complex traits. Bioinformatics. 2015;31(12):i206-13.
9. Gunning AC, Fryer V, Fasham J, Crosby AH, Ellard S, Baple EL, et al. Assessing performance of pathogenicity predictors using clinically relevant variant datasets. J Med Genet. 2021;58(8):547-55.
10. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;17(5):405-24.
11. Petrazzini BO, López-Bello F, Naya H, Spangenberg L. Clinical prediction of pathogenic variants in non-coding regions of the human genome. medRxiv. 2022:2022.02.25.22271514.
12. Clift K, Macklin S, Halverson C, McCormick JB, Abu Dabrh AM, Hines S. Patients' views on variants of uncertain significance across indications. J Community Genet. 2020;11(2):139-45.
13. Biesecker LG, Green RC. Diagnostic clinical genome and exome sequencing. N Engl J Med. 2014;370(25):2418-25.
14. Kaufman R, Timmermans S, Raz A. Genomic uncertainty and genetic counsellors' professional authority. Sociol Health Illn. 2023;45(3):485-502.
15. Brown T. Genomes. 2 ed: Oxford: Wiley-Liss
Garland Science; 2002. Chapter1, The Human Genome p.
16. Alberts B, Johnson A, Lewis J, al. e. Molecular Biology of the Cell. 4 ed: Garland Science; 2002. The Structure and Function of DNA p.
17. Institute NHGR. Genetics vs. Genomics [Available from: https://www.genome.gov/about-genomics/fact-sheets/Genetics-vs-Genomics.
18. (NIH) NIoH. Gene [cited 2024 01-08]. Available from: https://medlineplus.gov/genetics/understanding/basics/gene/.
19. (NIH) NIoH. Noncoding DNA [cited 2024 01-08]. Available from: https://medlineplus.gov/genetics/understanding/basics/noncodingdna/.
20. Ule J, Blencowe BJ. Alternative Splicing Regulatory Networks: Functions, Mechanisms, and Evolution. Mol Cell. 2019;76(2):329-45.
21. Institute NHGR. Intron [cited 2024 01-08]. Available from: https://www.genome.gov/genetics-glossary/Intron.
22. Eory L, Halligan DL, Keightley PD. Distributions of selectively constrained sites and deleterious mutation rates in the hominid and murid genomes. Mol Biol Evol. 2010;27(1):177-92.
23. Crick F. Central dogma of molecular biology. Nature. 1970;227(5258):561-3.
24. Sawicki MP, Samara G, Hurwitz M, Passaro E. Human Genome Project. Am J Surg. 1993;165(2):258-64.
25. Collins FS, Green ED, Guttmacher AE, Guyer MS, Institute UNHGR. A vision for the future of genomics research. Nature. 2003;422(6934):835-47.
26. Hood L, Rowen L. The Human Genome Project: big science transforms biology and medicine. Genome Med. 2013;5(9):79.

27. Lyon GJ, Wang K. Identifying disease mutations in genomic medicine settings: current challenges and how to accelerate progress. Genome Med. 2012;4(7):58.
28. de la Iglesia D, García-Remesal M, de la Calle G, Kulikowski C, Sanz F, Maojo V. The impact of computer science in molecular medicine: enabling high-throughput research. Curr Top Med Chem. 2013;13(5):526-75.
29. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. Science. 2022;376(6588):44-53.
30. Yang X, Wang X, Zou Y, Zhang S, Xia M, Fu L, et al. Characterization of large-scale genomic differences in the first complete human genome. Genome Biol. 2023;24(1):157.
31. Health NIo. [cited 2024. Available from: https://www.ncbi.nlm.nih.gov/grc/help/faq/#human-reference-genome-individuals.
32. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A draft sequence of the Neandertal genome. Science. 2010;328(5979):710-22.
33. Medicine. NNLo. Genome assembly GRCh38 2013 [cited 2024 01-08]. Available from: https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.26/.
34. Medicine. NNLo. Genome assembly GRCh38.p14 2022 [cited 2024 01-08]. Available from: https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.40/.
35. den Dunnen JT, Dalgleish R, Maglott DR, Hart RK, Greenblatt MS, McGowan-Jordan J, et al. HGVS Recommendations for the Description of Sequence Variants: 2016 Update. Hum Mutat. 2016;37(6):564-9.
36. Shameer K, Tripathi LP, Kalari KR, Dudley JT, Sowdhamini R. Interpreting functional effects of coding variants: challenges in proteome-scale prediction, annotation and assessment. Brief Bioinform. 2016;17(5):841-62.
37. Giacoletto CJ, Rotter JI, Grody WW, Schiller MR. Synonymous Variants of Uncertain Silence. Int J Mol Sci. 2023;24(13).
38. Kikutake C, Suyama M. Possible involvement of silent mutations in cancer pathogenesis and evolution. Sci Rep. 2023;13(1):7593.
39. Zeng Z, Bromberg Y. Predicting Functional Effects of Synonymous Variants: A Systematic Review and Perspectives. Front Genet. 2019;10:914.
40. Vihinen M. Nonsynonymous Synonymous Variants Demand for a Paradigm Shift in Genetics. Curr Genomics. 2023;24(1):18-23.
41. Torella A, Zanobio M, Zeuli R, Del Vecchio Blanco F, Savarese M, Giugliano T, et al. The position of nonsense mutations can predict the phenotype severity: A survey on the DMD gene. PLoS One. 2020;15(8):e0237803.
42. Huang X, Chen R, Sun M, Peng Y, Pu Q, Yuan Y, et al. Frame-shifted proteins of a given gene retain the same function. Nucleic Acids Res. 2020;48(8):4396-404.
43. Zhou H, Gao M, Skolnick J. ENTPRISE-X: Predicting disease-associated frameshift and nonsense mutations. PLoS One. 2018;13(5):e0196849.
44. Sergouniotis PI, Barton SJ, Waller S, Perveen R, Ellingford JM, Campbell C, et al. The role of small in-frame insertions/deletions in inherited eye disorders and how structural modelling can help estimate their pathogenicity. Orphanet J Rare Dis. 2016;11(1):125.
45. Zhang F, Lupski JR. Non-coding genetic variants in human disease. Hum Mol Genet. 2015;24(R1):R102-10.
46. Hertzog A, Selvanathan A, Farnsworth E, Tchan M, Adams L, Lewis K, et al. Intronic variants in inborn errors of metabolism: Beyond the exome. Front Genet. 2022;13:1031495.
47. Vitsios D, Dhindsa RS, Middleton L, Gussow AB, Petrovski S. Prioritizing non-coding regions based on human genomic constraint and sequence context with deep learning. Nat Commun. 2021;12(1):1504.
48. Macintyre G, Jimeno Yepes A, Ong CS, Verspoor K. Associating disease-related genetic variants in intergenic regions to the genes they impact. PeerJ. 2014;2:e639.
49. Anna A, Monika G. Splicing mutations in human genetic disorders: examples, detection, and confirmation. J Appl Genet. 2018;59(3):253-68.
50. Sanders SJ, Schwartz GB, Farh KK. Clinical impact of splicing in neurodevelopmental disorders. Genome Med. 2020;12(1):36.
51. Blakes AJM, Wai HA, Davies I, Moledina HE, Ruiz A, Thomas T, et al. A systematic analysis of splicing variants identifies new diagnoses in the 100,000 Genomes Project. Genome Med. 2022;14(1):79.

52. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015;526(7571):75-81.
53. Sedlazeck FJ, Lee H, Darby CA, Schatz MC. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. Nat Rev Genet. 2018;19(6):329-46.
54. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nat Rev Genet. 2011;12(5):363-76.
55. Carvalho CM, Lupski JR. Mechanisms underlying structural variant formation in genomic disorders. Nat Rev Genet. 2016;17(4):224-38.
56. Richard GF, Pâques F. Mini- and microsatellite expansions: the recombination connection. EMBO Rep. 2000;1(2):122-6.
57. Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. Genomics. 2016;107(1):1-8.
58. Reuter JA, Spacek DV, Snyder MP. High-throughput sequencing technologies. Mol Cell. 2015;58(4):586-97.
59. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17(6):333-51.
60. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977;74(12):5463-7.
61. Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol. 2008;26(10):1135-45.
62. Kasianowicz JJ, Brandin E, Branton D, Deamer DW. Characterization of individual polynucleotide molecules using a membrane channel. Proc Natl Acad Sci U S A. 1996;93(24):13770-3.
63. Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010;11(1):31-46.
64. Tucker T, Marra M, Friedman JM. Massively parallel sequencing: the next big thing in genetic medicine. Am J Hum Genet. 2009;85(2):142-54.
65. Thompson JF, Milos PM. The properties and applications of single-molecule DNA sequencing. Genome Biol. 2011;12(2):217.
66. Rhoads A, Au KF. PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics. 2015;13(5):278-89.
67. Zhang H, Li H, Jain C, Cheng H, Au KF, Aluru S. Real-time mapping of nanopore raw signals. Bioinformatics. 2021;37(Suppl_1):i477-i83.
68. Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. Detecting DNA cytosine methylation using nanopore sequencing. Nat Methods. 2017;14(4):407-10.
69. Warburton PE, Sebra RP. Long-Read DNA Sequencing: Recent Advances and Remaining Challenges. Annu Rev Genomics Hum Genet. 2023;24:109-32.
70. National Academies of Sciences E, and, Medicine, Division HaM, Services BoHC, on B, Populations tHoS, et al. An Evidence Framework for Genetic Testing., 2, Genetic Testing. . Washington (DC): National Academies Press (US); 2017.
71. Lee H, Deignan JL, Dorrani N, Strom SP, Kantarci S, Quintero-Rivera F, et al. Clinical exome sequencing for genetic identification of rare Mendelian disorders. JAMA. 2014;312(18):1880-7.
72. Warr A, Robert C, Hume D, Archibald A, Deeb N, Watson M. Exome Sequencing: Current and Future Perspectives. G3 (Bethesda). 2015;5(8):1543-50.
73. Morris HR, Houlden H, Polke J. Whole-genome sequencing. Pract Neurol. 2021;21(4):322-7.
74. Souche E, Beltran S, Brosens E, Belmont JW, Fossum M, Riess O, et al. Recommendations for whole genome sequencing in diagnostics for rare diseases. Eur J Hum Genet. 2022;30(9):1017-21.
75. Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 2010;38(6):1767-71.
76. Babraham, Bioinformatics. FastQC
A Quality Control tool for High Throughput Sequence Data [Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
77. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-20.
78. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60.
79. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26(5):589-95.
80. Li H. Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly. Bioinformatics. 2012;28(14):1838-44.

81. Zaharia M, Bolosky WJ, Curtis K, Fox A, Patterson D, Shenker S, et al. Faster and More Accurate Sequence Alignment with SNAP. 2011.
82. Bolosky WJ, Subramaniyan A, Zaharia M, Pandya R, Sittler T, Patterson d. Fuzzy set intersection based paired-end short-read alignment. 2021.
83. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.
84. Herzeel C, Costanza P, Decap D, Fostier J, Reumers J. elPrep: High-Performance Preparation of Sequence Alignment/Map Files for Variant Calling. PLoS One. 2015;10(7):e0132868.
85. Herzeel C, Costanza P, Decap D, Fostier J, Verachtert W. elPrep 4: A multithreaded framework for sequence analysis. PLoS One. 2019;14(2):e0209523.
86. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013;43(1110):11.0.1-.0.33.
87. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. . 2012.
88. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27(21):2987-93.
89. Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol. 2018;36(10):983-7.
90. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156-8.
91. Halim-Fikri H, Syed-Hassan SR, Wan-Juhari WK, Assyuhada MGSN, Hernaningsih Y, Yusoff NM, et al. Central resources of variant discovery and annotation and its role in precision medicine. Asian Biomed (Res Rev News). 2022;16(6):285-98.
92. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol. 2010;6(12):e1001025.
93. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25-9.
94. Aleksander SA, Balhoff J, Carbon S, Cherry JM, Drabkin HJ, Ebert D, et al. The Gene Ontology knowledgebase in 2023. Genetics. 2023;224(1).
95. Rodrigues CHM, Pires DEV, Ascher DB. DynaMut2: Assessing changes in stability and flexibility upon single and multiple point missense mutations. Protein Sci. 2021;30(1):60-9.
96. Khanna T, Hanna G, Sternberg MJE, David A. Missense3D-DB web catalogue: an atom-based analysis and repository of 4M human protein-coding genetic variants. Hum Genet. 2021;140(5):805-12.
97. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434-43.
98. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74.
99. Whirl-Carrillo M, Huddart R, Gong L, Sangkuhl K, Thorn CF, Whaley R, et al. An Evidence-Based Framework for Evaluating Pharmacogenomics Knowledge for Personalized Medicine. Clin Pharmacol Ther. 2021;110(3):563-72.
100. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. 2013;Chapter 7:Unit7.20.
101. Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC. SIFT missense predictions for genomes. Nat Protoc. 2016;11(1):1-9.
102. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017;45(D1):D353-D61.
103. Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, et al. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Res. 2023;51(D1):D638-D46.
104. Köhler S, Gargano M, Matentzoglu N, Carmody LC, Lewis-Smith D, Vasilevsky NA, et al. The Human Phenotype Ontology in 2021. Nucleic Acids Res. 2021;49(D1):D1207-D17.
105. McKusick-Nathans, Institute of Genetic Medicine JHUB, MD). Online Mendelian Inheritance in Man, OMIM ®. [cited 2024. Available from: https://omim.org/.

106. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Res. 2015;43(Database issue):D789-98.
107. Gudmundsson S, Singer-Berk M, Watts NA, Phu W, Goodrich JK, Solomonson M, et al. Variant interpretation using population databases: Lessons from gnomAD. Hum Mutat. 2022;43(8):1012-30.
108. Gibson G. Rare and common variants: twenty arguments. Nat Rev Genet. 2012;13(2):135-45.
109. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285-91.
110. Lek M. Decoding the genetics of rare disease: an interview with Monkol Lek. Dis Model Mech. 2022;15(6).
111. Scott EM, Halees A, Itan Y, Spencer EG, He Y, Azab MA, et al. Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery. Nat Genet. 2016;48(9):1071-6.
112. Fattahi Z, Beheshtian M, Mohseni M, Poustchi H, Sellars E, Nezhadi SH, et al. Iranome: A catalog of genomic variations in the Iranian population. Hum Mutat. 2019;40(11):1968-84.
113. Li Z, Jiang X, Fang M, Bai Y, Liu S, Huang S, et al. CMDB: the comprehensive population genome variation database of China. Nucleic Acids Res. 2023;51(D1):D890-D5.
114. Liu X, Jian X, Boerwinkle E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Hum Mutat. 2011;32(8):894-9.
115. Liu X, Li C, Mou C, Dong Y, Tu Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. Genome Med. 2020;12(1):103.
116. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010;20(1):110-21.
117. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005;15(8):1034-50.
118. Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, et al. A high-resolution map of human evolutionary constraint using 29 mammals. Nature. 2011;478(7370):476-82.
119. McVicker G, Gordon D, Davis C, Green P. Widespread genomic signatures of natural selection in hominid evolution. PLoS Genet. 2009;5(5):e1000471.
120. Zhang H, Xu MS, Fan X, Chung WK, Shen Y. Predicting functional effect of missense variants using graph attention neural networks. Nat Mach Intell. 2022;4(11):1017-28.
121. Malhis N, Jacobson M, Jones SJM, Gsponer J. LIST-S2: taxonomy based sorting of deleterious missense mutations across species. Nucleic Acids Res. 2020;48(W1):W154-W61.
122. Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day IN, et al. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics. 2015;31(10):1536-43.
123. Rogers MF, Shihab HA, Mort M, Cooper DN, Gaunt TR, Campbell C. FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. Bioinformatics. 2018;34(3):511-3.
124. Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GL, Edwards KJ, et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum Mutat. 2013;34(1):57-65.
125. Frazer J, Notin P, Dias M, Gomez A, Min JK, Brock K, et al. Disease variant prediction with deep generative models of evolutionary data. Nature. 2021;599(7883):91-5.
126. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res. 2009;19(9):1553-61.
127. Wu Y, Li R, Sun S, Weile J, Roth FP. Improved pathogenicity prediction for rare human missense variants. Am J Hum Genet. 2021;108(10):1891-906.
128. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. 2011;39(17):e118.
129. Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: mutation prediction for the deep-sequencing age. Nat Methods. 2014;11(4):361-2.
130. Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics. 2013;14 Suppl 3(Suppl 3):S3.
131. Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics. 2015;31(16):2745-7.
132. Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, et al. Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics. 2009;25(21):2744-50.

133. Sundaram L, Gao H, Padigepati SR, McRae JF, Li Y, Kosmicki JA, et al. Predicting the clinical impact of human mutation with deep neural networks. Nat Genet. 2018;50(8):1161-70.
134. Raimondi D, Tanyalcin I, Ferté J, Gazzo A, Orlando G, Lenaerts T, et al. DEOGEN2: prediction and interactive visualization of single amino acid variant deleteriousness in human proteins. Nucleic Acids Res. 2017;45(W1):W201-W6.
135. Cheng J, Novati G, Pan J, Bycroft C, Žemgulytė A, Applebaum T, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science. 2023;381(6664):eadg7492.
136. Gulko B, Hubisz MJ, Gronau I, Siepel A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. Nat Genet. 2015;47(3):276-83.
137. Qi H, Zhang H, Zhao Y, Chen C, Long JJ, Chung WK, et al. MVP predicts the pathogenicity of missense variants by deep learning. Nat Commun. 2021;12(1):510.
138. Feng BJ. PERCH: A Unified Framework for Disease Gene Prioritization. Hum Mutat. 2017;38(3):243-51.
139. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019;47(D1):D886-D94.
140. Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics. 2015;31(5):761-3.
141. Ionita-Laza I, McCallum K, Xu B, Buxbaum JD. A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat Genet. 2016;48(2):214-20.
142. Lu Q, Hu Y, Sun J, Cheng Y, Cheung KH, Zhao H. A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. Sci Rep. 2015;5:10576.
143. Alirezaie N, Kernohan KD, Hartley T, Majewski J, Hocking TD. ClinPred: Prediction Tool to Identify Disease-Relevant Nonsynonymous Single-Nucleotide Variants. Am J Hum Genet. 2018;103(4):474-83.
144. Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum Mol Genet. 2015;24(8):2125-37.
145. Li C, Zhi D, Wang K, Liu X. MetaRNN: differentiating rare pathogenic and rare benign missense SNVs and InDels using deep learning. Genome Med. 2022;14(1):115.
146. Jagadeesh KA, Wenger AM, Berger MJ, Guturu H, Stenson PD, Cooper DN, et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. Nat Genet. 2016;48(12):1581-6.
147. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet. 2016;99(4):877-85.
148. Samocha KE, Kosmick JA, Karczewski KJ, O'Donnell-Luria AH, Pierce-Hoffman E, MacArthur DG, et al. Regional missense constraint improves variant deleteriousness prediction. 2017.
149. Brandes N, Goldman G, Wang CH, Ye CJ, Ntranos V. Genome-wide prediction of disease variant effects with a deep protein language model. Nat Genet. 2023;55(9):1512-22.
150. Huang YF, Gulko B, Siepel A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. Nat Genet. 2017;49(4):618-24.
151. Landrum MJ, Chitipiralla S, Brown GR, Chen C, Gu B, Hart J, et al. ClinVar: improvements to accessing data. Nucleic Acids Res. 2020;48(D1):D835-D44.
152. Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, et al. Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat. 2003;21(6):577-81.
153. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29(1):308-11.
154. Forbes SA, Bhamra G, Bamford S, Dawson E, Kok C, Clements J, et al. The Catalogue of Somatic Mutations in Cancer (COSMIC). Curr Protoc Hum Genet. 2008;Chapter 10:Unit 10.1.
155. Jalali Sefid Dashti M, Gamieldien J. A practical guide to filtering and prioritizing genetic variants. Biotechniques. 2017;62(1):18-30.
156. Richards CS, Bale S, Bellissimo DB, Das S, Grody WW, Hegde MR, et al. ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007. Genet Med. 2008;10(4):294-300.
157. Amendola LM, Jarvik GP, Leo MC, McLaughlin HM, Akkari Y, Amaral MD, et al. Performance of ACMG-AMP Variant-Interpretation Guidelines among Nine Laboratories in the Clinical Sequencing Exploratory Research Consortium. Am J Hum Genet. 2016;98(6):1067-76.

158. Starita LM, Ahituv N, Dunham MJ, Kitzman JO, Roth FP, Seelig G, et al. Variant Interpretation: Functional Assays to the Rescue. Am J Hum Genet. 2017;101(3):315-25.
159. MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, Abecasis GR, et al. Guidelines for investigating causality of sequence variants in human disease. Nature. 2014;508(7497):469-76.
160. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic intolerance to functional variation and the interpretation of personal genomes. PLoS Genet. 2013;9(8):e1003709.
161. Miller DD, Brown EW. Artificial Intelligence in Medical Practice: The Question to the Answer? Am J Med. 2018;131(2):129-33.
162. Walters-Sen LC, Hashimoto S, Thrush DL, Reshmi S, Gastier-Foster JM, Astbury C, et al. Variability in pathogenicity prediction programs: impact on clinical diagnostics. Mol Genet Genomic Med. 2015;3(2):99-110.
163. Pejaver V, Byrne AB, Feng BJ, Pagel KA, Mooney SD, Karchin R, et al. Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. Am J Hum Genet. 2022;109(12):2163-77.
164. Raasveldt M, Mühleisen H. DuckDB: an Embeddable Analytical Database. Proceedings of the 2019 International Conference on Management of Data; Amsterdam, Netherlands: Association for Computing Machinery; 2019. p. 1981–4.
165. Harrison PW, Amode MR, Austine-Orimoloye O, Azov AG, Barba M, Barnes I, et al. Ensembl 2024. Nucleic Acids Res. 2024;52(D1):D891-D9.
166. Apache-Parquet [cited 2024 01-08]. Available from: ***https://parquet.apache.org***.
167. Ahmad T, Ahmed N, Al-Ars Z, Hofstee HP. Optimizing performance of GATK workflows using Apache Arrow In-Memory data framework. BMC Genomics. 2020;21(Suppl 10):683.
168. Jpopgen-dbNSFP [cited 2024 01-08]. [Available from: https://sites.google.com/site/jpopgen/dbNSFP.
169. Krithikadatta J. Normal distribution. J Conserv Dent. 2014;17(1):96-7.
170. de Torrenté L, Zimmerman S, Suzuki M, Christopeit M, Greally JM, Mar JC. The shape of gene expression distributions matter: how incorporating distribution shape improves the interpretation of cancer transcriptomic data. BMC Bioinformatics. 2020;21(Suppl 21):562.
171. Marko NF, Weil RJ. Non-gaussian distributions affect identification of expression patterns, functional annotation, and prospective classification in human cancer genomes. PLoS One. 2012;7(10):e46935.
172. Accetturo M, Bartolomeo N, Stella A. In-silico Analysis of. Int J Mol Sci. 2020;21(3).
173. Garcia FAO, de Andrade ES, Palmero EI. Insights on variant analysis. Front Genet. 2022;13:1010327.
174. Kruskal WH, Wallis WA. Use of Ranks in One-Criterion Variance Analysis. Journal of the American Statistical Association. 1952;47:583-621.
175. Wilcoxon F. Individual Comparisons by Ranking Methods. Biometrics Bulletin. 1945;1:4.
176. Beaulieu J. Feature Scaling and Normalization 2020 [cited 2024 01-08]. Available from: https://julienbeaulieu.gitbook.io/wiki/sciences/machine-learning/linear-regression/feature-scaling-and-normalization.
177. Groth D, Hartmann S, Klie S, Selbig J. Principal components analysis. Methods Mol Biol. 2013;930:527-47.
178. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Philos Trans A Math Phys Eng Sci. 2016;374(2065):20150202.
179. Metsalu T, Vilo J. ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. Nucleic Acids Res. 2015;43(W1):W566-70.
180. Ultsch A, Lötsch J. Euclidean distance-optimized data transformation for cluster analysis in biomedical data (EDOtrans). BMC Bioinformatics. 2022;23(1):233.
181. Hunter JD. Matplotlib [Available from: https://matplotlib.org/.
182. Golubchik T, Wise MJ, Easteal S, Jermiin LS. Mind the gaps: evidence of bias in estimates of multiple sequence alignments. Mol Biol Evol. 2007;24(11):2433-42.
183. Privman E, Penn O, Pupko T. Improving the performance of positive selection inference by filtering unreliable alignment regions. Mol Biol Evol. 2012;29(1):1-5.
184. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, et al. The UCSC Genome Browser Database. Nucleic Acids Res. 2003;31(1):51-4.
185. Nassar LR, Barber GP, Benet-Pagès A, Casper J, Clawson H, Diekhans M, et al. The UCSC Genome Browser database: 2023 update. Nucleic Acids Res. 2023;51(D1):D1188-D95.

186. Kopanos C, Tsiolkas V, Kouris A, Chapple CE, Albarca Aguilera M, Meyer R, et al. VarSome: the human genomic variant search engine. Bioinformatics. 2019;35(11):1978-80.
187. VarSome-Germline Implementation [cited 2024 01-08]. Available from: *mso-list:l0 level1 lfo1"> https://varsome.com/about/resources/germline-implementation/*.
188. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56-65.

8 CURRICULUM VITAE