Màster Universitari
Anàlisi de Dades Òmiques /
Omics Data Analysis
FACULTAT DE CIÈNCIES I TECNOLOGIA
UVIC | UVIC·UCC

Master of Science in Omics Data Analysis

Master Thesis

It is possible to extrapolate pathogenicity in mutations in the same position in homologous proteins

by

Gabriel Ruiz Alías

Supervisor: Dr. Mireia Olivella García, Bioinformatics and Medical Statistics, UVic

Co-supervisor: Dr. Arnau Cordomí Montoya, Bioinformatics, ESCI-UPF

Biosciences Department

University of Vic – Central University of Catalonia (UVic – UCC)

The 10th of September 2023

It is possible to extrapolate pathogenicity in mutations in the same position in homologous proteins

Gabriel Ruiz Alías[1]

Supervisors: Mireia Olivella[2], Arnau Cordomí[3]

[1]MSc in Omics Data Analysis student, UVic-UCC, Barcelona, Spain; [2]Bioinformatics and Medical Statistics, UVic-UCC, Barcelona, Spain; [3]Bioinformatics, ESCI-UPF, Barcelona, Spain.

Abstract

When we compare an individual's genome with the reference, several mutations are encountered. Most of these mutations are neutral, but some others can lead to pathogenic consequences. Given the rapid increase in the amount of generated sequencing data, there is an urgent need to accurately determine whether genetic variants detected in patients are disease causing or not. While numerous computational predictive tools exist, their ability to make accurate predictions is still limited. In this study, we focus on missense variants, those that modify the coding amino acid, and our aim is to determine if the pathogenicity of these variants can be extrapolated to homologous variants, i.e., variants affecting the same position in homologous proteins and exhibiting the same or similar amino acid change. With this purpose, we extracted homologous variants in a dataset composed of all reported disease-causing (ClinVar) and neutral (gnomAD) human missense variants from proteins with autosomal dominant (AD) inheritance. We collected 63,192 pairs of homologous variants from which 60,822 were disease-causing, 1,799 were neutral and 571 of them disagreed in pathogenicity annotation, achieving an error rate of 0.9%. Thus, our data supports that pathogenicity can be extrapolated, with a high accuracy and reliability, between homologous variants. This approach expands the number of variants for which pathogenicity can be annotated with a high precision.

Contact: gabriel.ruiz@uvic.cat

GitHub link (Contact for access)**:** https://github.com/ruizzgabriel/Homologous-mutations/

Supplementary data are available at: https://ja.cat/gruiz_Supp_mat_FMP_MSc_Omics_Data_Analysis_UVic-UCC

Introduction

Genetic variation has driven evolution since the appearance of the first simplest organism to today's most complex organisms. It allows diversity in the organisms, which has the potential to improve the species. Individuals presenting a differential trait that provides them some adaptive advantage to their environment, are more likely to survive and reproduce, passing the adaptive traits onto their offspring. Therefore, species can evolve thanks to these variations that can be consolidated in a process called natural selection. Yet, individuals also present different genetic variations in different populations, due to the differences in their environments. Although some variants represent an adaptive trait, others are neutral, and some others prevent the adaptation of the organism to the environment by hindering its survival. The latter variants are less likely to be passed to the offspring because they present a survival disadvantage in the environment, as in the case of disease-causing variants, so they are rarely present in the current population. There are different mechanisms of genetic variation that allow organisms to develop differences in their genomes such as single nucleotide mutations. A mutation is a variation in the DNA sequence of an organism that can occur from errors in DNA replication during cell division, exposure to mutagens or a viral infection. Variants that occur in body cells are called somatic variants whereas variants that occur in eggs and sperm are called germline mutations and can be passed onto offspring.

Depending on the frequency of appearing in a population, variants are classified as: (1) rare variants, that present a frequency <0.5%, (2) low-frequency variants, that present a frequency between 0.5% and 5%, and (3) common variants that present a frequency higher than 5%.[1] From the total number of human variants annotated, only 10% are common. In contrast, a human genome presents approximately between 1-4% of rare variants, while the vast majority of the variants are common.[1]

The variants can be of type synonymous (the change in the nucleotide sequence does not alter the amino acid sequence), truncating variant (resulting in a premature stop coding and changing or not changing the amino acid sequence between the introduction of the stop codon), splice site variants (that affect splice site sequences), UTR variants (nucleotide changes that affect the 3'UTR and the 5'UTR sequences) and missense variants. Among the different types of single nucleotide variants, missense variants are the most common.[2] A missense variant is a DNA change that entails an amino acid encodement variation at a particular protein position. Thus, missense variants cause the change of an amino acid sequence, and it is estimated that in a typical human genome, 10,000 to 12,000 missense variants altering protein sequence can be found.[1] Depending on the position of the amino acid and the change in its physicochemical properties, missense variants can affect protein structure and function or not, i.e., conserved regions and drastic amino acid changes are more prompt

to affect protein structure and function, and thus may be involved in pathogenesis (or, more rarely, provide an advantage).

Missense variants can lead to monogenic diseases, diseases that can be caused by variants in a single gene, or to complex disease, caused by the contributions of multiple genes. Mendelian diseases, those presenting an inheritance type that follow the principles proposed by Gregor Mendel describing the presence of two alleles for each gene, are monogenic diseases. The most common interaction between alleles is a dominant/recessive relationship, where the dominant allele imposes its phenotype overruling the recessive. Mendel proposed an inheritance model in which the combination of the alleles of both parents would define the probabilities of occurrence in their offspring. Mendelian diseases typically correspond to rare diseases, because they occur infrequently in the general population, in less than 1 in 2,000 individuals.[3] Also, depending on if the gene is present in a sex chromosome or not, it will present a sex-linked or an autosomal pattern. In contrast with sex-linked, autosomal inheritance allows the transmission of traits regardless of the sex of the parent or the child. Thus, autosomal dominant Mendelian diseases allow a direct approach to study them with respect to the recessive ones because functional data can be related with the variation, because a single mutated allele is enough to cause the disease, but also because it can be studied regardless of the sex.[2] In fact, our understanding of the relationship between gene function and human phenotypes is mostly based on the study of rare genetic variations caused by Mendelian phenotypes with a dominant autosomal inheritance.[4]

Most identified Mendelian phenotypes result from altered function, localization, or presence of the encoded proteins, even though the protein-coding regions are only around 1% of the human genome. A typical genome differs from the reference human genome at 4.1 million to 5.0 million sites, and 24-30 variants per genome are implicated in rare diseases.[1] Over the past twenty years, there have been significant advancements in the field of genomics, resulting in a considerable decrease in the cost of genome sequencing.[5] Despite these advancements, there remains a challenge in determining which rare variants cause a monogenic disease by affecting the structure and function of a protein. There is still a lack of understanding about the relationship between genetic variability between patients and possible pathology. It has been shown that missense variants causing rare Mendelian diseases are more common than previously believed: around 50% of genes underlying the known Mendelian phenotypes were still unknown in 2015.[6] A significant portion of known human diseases, about 0.4% of live births, are made up of clinically recognized Mendelian phenotypes and, if we consider all congenital anomalies, approximately 8% of live births have a genetic disorder that can be identified by early adulthood. This means that every year, around 8 million children are born with a serious genetic condition that is either life-threatening or could result in disability. Birth defects, which include a significant proportion of Mendelian phenotypes, are the leading cause of death in infants during their first year of life. Every year, more than three million children under the age of five die from a birth defect, and a similar number survive with significant health problems. In addition, the diagnosis time is often extensive.[4] According to a European survey of eight rare diseases, including several Mendelian diseases, 25% of patients had to wait between 5 and 30 years from early symptoms to confirmatory diagnosis of their disease, and 40% of them first received an erroneous diagnosis.[6]

While high-throughput data is being generated and new mutations are being identified and detected, we are still not able to check the pathogenic/non-pathogenic effects of the variants *pari passu*. Due to the high costs derived from experimentally annotating the pathogenicity of a variant, i.e., introducing the variant and observing how the protein structure and function changes, the variant consequences must be studied computationally to save resources and time. If the protein function is altered as a consequence of the amino acid change, the variant can be disease-causing, but they can also be neutral as they are not benign and do not alter the protein function. Accordingly, variants that are (potentially) disease-causing are known as pathogenic variants. Still, there are some missense variants affecting the protein structure and function that are not disease-causing *per se*, as are involved in complex disease or because the protein function is not essential for the organism. It is vital to early detect the variant implications in the patient to predict its possible pathogenicity and provide higher accuracy on diagnosis with the aim of improving patients' quality of life.

Numerous efforts are being performed to detect the consequences of missense variants and easily differentiate whether they are pathogenic or not. Several prediction tools have therefore been developed to help this process, called variant prioritization. They use different approaches with the aim of providing approximations on the pathogenic or neutral effects of the discovered variants, so that we could improve our understanding of certain diseases and be able to detect them early on patients.

Many of the main tools used to predict the impact of genetic variations rely on analyzing phylogenetic conservation of the region and how changes may affect the structure and function of a protein, such as SIFT[7], Provean[8] or Mutation Assessor[9]. While other tools incorporate structural parameters for variant classification such as Poly-Phen-2[10]. But they still have some limitations.[2,11]

The sensitivity of SIFT and PolyPhen tools were assessed in a set of 141 missense variants and they presented 69% and 68% respectively, whereas their specificity was 13% and 16%.[11] In another study, SIFT and PolyPhen-2 accuracies were checked in GPCRs involved in neuroendocrine regulation of reproduction, showing 83% and 85%, respectively.[12]

Despite numerous tools with different approaches having been designed, they are still not accurate enough. Thus, there is a great interest in predicting the pathogenicity of variants, since identifying the disease-causing variant is the first step in designing a therapeutic strategy.

A recent study presented by our group has recently identified that for GRIN-related disorders, a monogenic rare disease, the pathogenicity of missense variants can be extrapolated between homologous variants, (variants that affect the same equivalent position in homologous proteins)[13]. This is based on the hypothesis that similar amino acid changes in similar proteins may result in the

same effect in both protein structure and functions in a dominant autosomal monogenic disease, where there is a clear correlation between the effect of a single variant in protein structure and function and the clinical phenotype. Taking advantage of this recent study, we want to explore if pathogenicity of missense variants can be extrapolated between homologous variants for all genes in the human genome involved in dominant autosomal monogenic diseases, thus contributing to expanding the pathogenicity annotations of newly identified variants, and speeding up the identification of pathogenic variants in patients that are sequenced under the suspicions of a rare genetic disease.

Objectives

This study seeks to elucidate whether it is possible to use the available information regarding the pathogenicity annotations in human missense variants to extrapolate pathogenicity annotations to homologous variants, i.e., variants presenting the same or similar amino acid changes in equivalent positions between homologous proteins. We hypothesize that similar changes in equivalent positions of homologous proteins may result in the same effect in protein structure and function. For this purpose, we will compare pathogenicity annotations between all homologous missense variants in the human genome.

Methods

High-throughput data was collected from UniProt, gnomAD and ClinVar databases and combined using the pipeline illustrated in Figure 1.

Specifically, we retrieved pathogenic and non-pathogenic variants details from ClinVar and gnomAD. UniProtKb was consulted to retrieve all reviewed protein information for Homo sapiens. For each of the 20,422 human curated proteins, the essential information was retrieved: gene and protein names, Pfam codes, and references to other databases (ENSEMBL, RefSeq and InterPro). Once UniProtKb human proteins were obtained, they were sought in gnomAD to retrieve all available (non-pathogenic or neutral) variants based on the Ensembl gene ID. Thereby, information regarding missense, truncations, splice sites, UTRs, frameshift and stop gained molecular consequences were acquired. Additionally, this information was searched in two gnomAD datasets: v2.1 and v3. For each variant, several details were taken including chromosome, position in the genome, the SNPs ID 'rsID', transcript ID, gene name, molecular consequence, variant ID, and genome and exome features. Non-pathogenic variants files from gnomAD required a processing step. We removed the non-desired columns to maintain only the desired information and to decrease the size of the files. Accordingly, the conserved features of each variant were gene name, ENSEMBL transcript ID, RefSeq transcript ID, rsID, protein change, molecular consequence, clinical significance, and genome and exome allele frequencies (AF).

Besides neutral variants, the study also required disease-causing variants. Pathogenic and likely pathogenic variants were retrieved from ClinVar. A dataset of 208,426 disease-causing human variants was obtained. The retrieved information from ClinVar comprised variant name, transcript name, gene name and protein change, clinical significance, review status, chromosome location, variation ID, allele ID, among others.

Once the raw data was collected, we processed it to retrieve the information of interest. We extracted the information related to the annotated NM NCBI accession code and protein change into different columns. Also, the protein change was split into different columns to separate the position and initial and final amino acids in different columns. In addition, Pfam codes and UniProt accession names were added, linking ClinVar and UniProt files by the RefSeq transcript code (NM). Despite that, NM codes present in UniProt obtained from RefSeq were only the canonical ones, and since each gene can be associated with different NM codes, a high number of entries would have not been related with UniProt. To address this and avoid losing information, the BioMart database was used to relate those NM codes in ClinVar that could not have been related with its corresponding entries in UniProt.

As a final step, several restrictions were applied to the large datasets of protein variants to retain only entries that met the desired conditions. Consequently, after applying all filters, the numbers of pathogenic and neutral variants were reduced (See Table 2). First, pathogenic and non-pathogenic variants were filtered to keep only genes following an autosomal dominant (AD) inheritance pattern (see introduction). This information was taken from the OMIM dataset and resulted in 2,224 AD genes out of 20,422 human genes. Second, we discarded variants with types other than missense. ClinVar was also filtered to exclude somatic variants, which are DNA alterations that occur in non-germ cells and therefore are not inherited in the offspring. Although gnomAD is meant to be a non-pathogenic database, it is possible that it may include pathogenic variants that have not yet been characterized. To prevent this, we have discarded variants with an AF$< 5 * 10^{-4}$.

Homologous positions using sequence alignments

We seek for homologous variants, i.e., those that occur in members of the same family and modify amino acids in the same position. We used Pfam (as integrated into the InterPro database) both for family classification and for the family alignments that allowed normalizing the protein change positions annotated in ClinVar and gnomAD to check for homologous variants.

From the different alignments available, full alignments were chosen. Protein change positions were compared with the alignments from Pfam, and a standard position was duly extracted. In this process we checked that the residue in the alignment matched the initial amino acid described in the protein change at the corresponding position. We only kept cases that met this requirement since not all NM amino acid numbering are equal as in Pfam.

In the retrieved Pfam alignments, upper and lower case residues were found, as well as different gap characters: points and slashes. Only upper case residues were considered to obtain equivalent positions, but both different gap types were treated uniformly as gaps. Upper case residues are the ones that are confidently aligned to the profile HMM (Hidden Markov Model), which is the

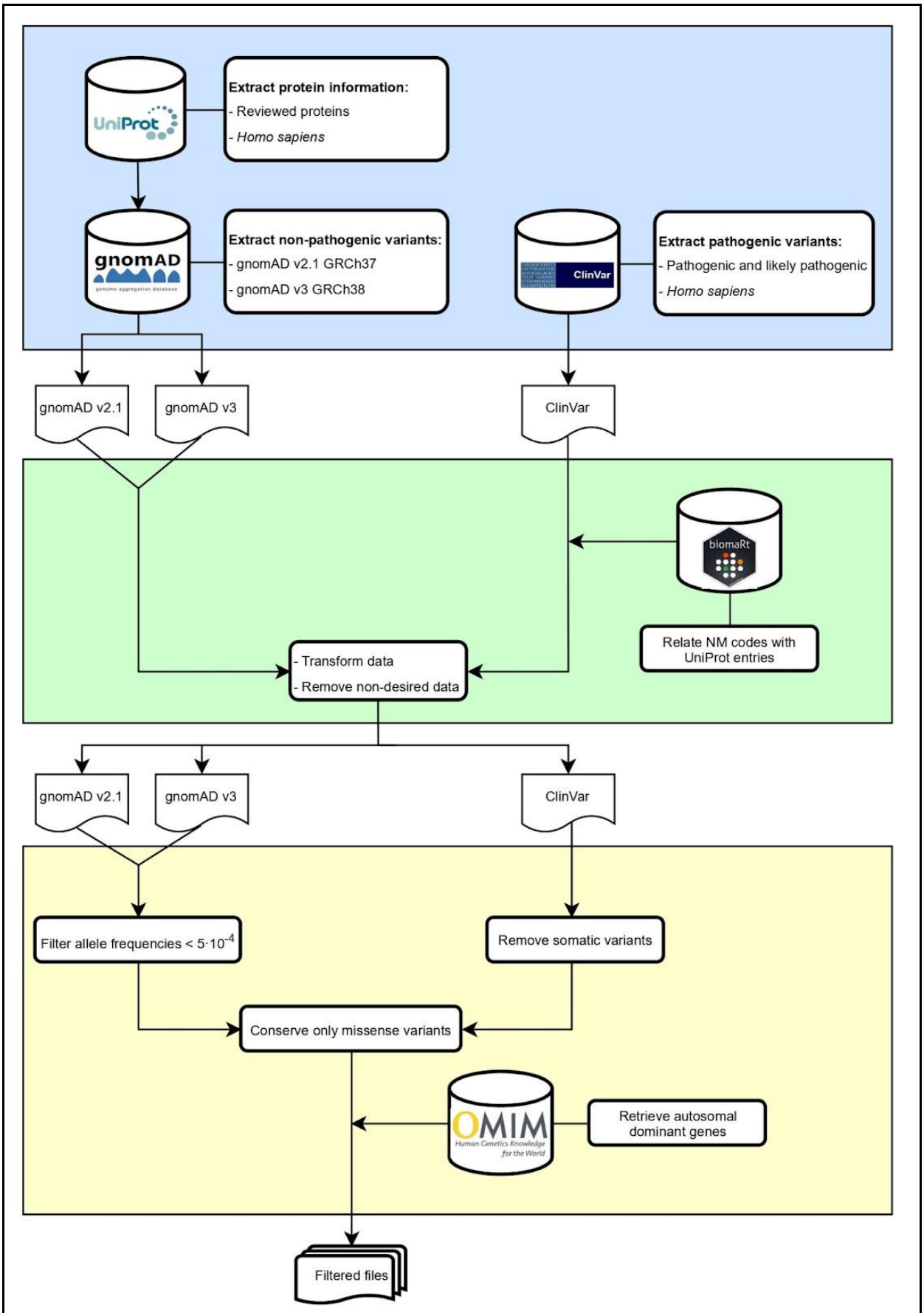

Fig. 1. Workflow representing the different steps of data collection (blue), data processing (green) and data filtering (yellow).

computational algorithm used by Pfam to identify similarities between protein sequences. Otherwise, a lower case residue is represented in the alignment when an insertion occurs after an envelope is detected for a profile HMM match. An envelope is considered the region on the sequence where the match has been probabilistically determined to lie.[14] Moreover, points are used to align sequences that have extra amino acids that are not part of the match in the profile HMM, and slash characters are used when the HMM profile expects a residue to be present in the sequence but is missing.[15]

After aligning the protein change positions, pathogenic and neutral variant datasets were merged conserving their database of origin, resulting in a single variants dataset. After that, this whole variant's dataset was split into Pfam families.

Furthermore, since we were interested in matching homologous variants, apart from matching exact amino acid changes, we were interested in detecting variants involving amino acids with similar physicochemical properties. To accomplish that, the BLOSUM62 substitution matrix was computed to extend the previous results matching pairs with positive amino acid changes. Since the aim of the experiment was to find variants with equivalent physicochemical properties to extrapolate the possible pathology, we were interested in the protein change result.

General purpose databases

In addition to the database specifically dealing with mutation data, we have used additional biological databases to get complementary information for the characterization of the genes and its families. Some of the following ones contain protein information, sequence alignment information or relations between different accession codes.

The Universal Protein Resource (**UniProt**, https://www.uniprot.org/) is a freely accessible, comprehensive resource for protein sequence and functional information. It is maintained by the UniProt Consortium, which consists of the European Bioinformatics Institute (EMBL-EBI), the Swiss Institute of Bioinformatics (SIB), and the Protein Information Resource (PIR). The UniProt Knowledgebase (UniProtKB) is the central hub for the collection of functional information on proteins, with accurate, consistent and rich annotation. It includes both manually reviewed (UniProtKB/Swiss-Prot) and unreviewed (UniProtKB/TrEMBL) records.[16] It contains sequences from 14,403 different species where the most represented one is *Homo sapiens* with 20,422 reviewed entries (UniProtKB/Swiss-Prot release 2023_01).[17]

Pfam(https://www.ebi.ac.uk/interpro/entry/pfam/#table) is a database of protein families and domains integrated in the InterPro database. It is used to analyze novel genomes and metagenomes, as well as to guide experimental work on particular proteins and systems. Pfam entries are manually annotated with functional information from the literature where available.[18] Current version (Pfam release 35.0) contains 19,632 families but only 6,680 of them are found in human.

BioMart(http://www.ensembl.org/info/data/biomart/index.html) is a community-driven platform supported by Ensembl, that provides access to over 800 different biological datasets from various fields such as genomics, proteomics, cancer data, and more. Can be accessed by a web-based tool that provides an easy way to extract data, where different parameters can be modified to access the specific requirements.[19]

NCBI's Reference Sequence (**RefSeq**, https://www.ncbi.nlm.nih.gov/refseq/) database integrates curated and non-redundant DNA, RNA and protein sequences from a wide range of species. In addition, it also provides genetic and functional information with the sequence data. Can be accessed by a web-based tool or by available links in other NCBI resources.[20]

Variants Databases

The large amount of mutation data available is conveniently stored and curated in specialized databases. In the present report we have used the following resources, which provide valuable information about the variants pathogenicity:

ClinVar(https://www.ncbi.nlm.nih.gov/clinvar/) is a freely available, public archive of human genetic variants and interpretations of their relationships to diseases and other conditions, maintained at the National Institutes of Health (NIH). More than 1300 organizations have contributed their interpretations of variants to the ClinVar database, including clinical testing laboratories, research laboratories, locus-specific databases, clinicians, patient registries, expert panels, and other organizations.[21]

The Genome Aggregation Database (**gnomAD**, https://gnomad.broadinstitute.org/) is the largest publicly available resource that aggregates and harmonizes both exome and genome sequencing data from a wide variety of large-scale sequencing projects. It provides summary data on genetic variants observed in tens of thousands of individuals, with the goal of facilitating the interpretation of genetic variation in both clinical and research settings.[22] In this project, two versions of gnomAD were used: v2.1 and v3. v2.1 release is composed of 125,748 exomes and 15,708 genomes (GRCh37) from 15 population subgroups[23], whereas v3 includes 71,102 genomes (GRCh38) from 9 population groups.[24]

The Online Mendelian Inheritance in Man (**OMIM**, https://www.omim.org/) is a comprehensive and authoritative database of human genes and genetic phenotypes that is updated daily. All its entries are curated and contain full-text overviews of genes and genetic phenotypes that can be used by students, researchers, and clinicians.[25]

Results and discussion

Description of the Human variants

We extracted all pathogenic and non-pathogenic human missense variants available in ClinVar and gnomAD, respectively (see Methods). A summary of the data obtained is presented in Table 1. Within the pathogenic variants, we found a total number of 208,426 variants. Among them, truncation/frameshift/stop gained variants was the largest group with 124,945 variants (58.9%), followed by missense variants with 52,190 variants (24.6%). Also, we found 25,140 splice site variants (11.9%) and 9,740 UTR variants (4.6%). The distribution is displayed in Figure 2. In the case of non-pathogenic variant datasets, we obtained 18,017,079 and 14,569,514 variants in v2.1 (annotated on reference genome GRCh37) and v3 (annotated on GRCh38 reference genome), respectively. In total we compiled 32,450,901 unique variants (only 135,692 variants were present in both datasets). Among them, missense variants were the largest group found in both versions (35.9% in v2.1 and 31.6% in v3), followed by synonymous variants, 3,003,140 variants in v2.1 (16.7%) and 2,200,104 in v3 (15.1%). In addition, we found 929,699 UTR variants in v2.1 (5.2%) and 920,608 in v3 (6.3%), and 903,882 splice site variants in v2.1 (5%) and 653,326 in v3 (4.5%). Truncation/frameshift/stop gained group was the smallest group (3% in v2.1 and 2.8% in v3). (See Fig. 2 for pathogenic and non-pathogenic variant distributions). Variants from ClinVar occur in only 6,567 unique genes whereas variants from gnomAD v2.1 and gnomAD v3 occur in 18,656 and 19,033 unique genes, respectively. Interestingly, from the 20,423 coding genes in the human genome according to the UniProt, only 32% present pathogenic variants.

Pathogenic and non-pathogenic missense variants in autosomal dominant (AD) inheritance

We aimed to check if pathogenicity could be extrapolated between homologous variants, defined as variants presenting the exact amino acid change or when resulting amino acids exhibit similar physicochemical properties. In order to compare the pathogenicity between variants able to cause a disease due to a change in one single amino acid, only missense variants from proteins involved in AD inheritance type were kept. This inheritance model implies that the gene in question is located on non-sex chromosomes and only a single copy of the mutated gene is enough to cause the disorder. In contrast with sex-linked variants, autosomal variants have an inheritance and presentation pattern that does not depend on the sex of the parent or the child so that we can study them regardless of the sex. We found 2,224 AD genes in the OMIM database. When considering only variants in AD genes, the pathogenic dataset reduces from 52,190 to 28,587 missense variants. For non-pathogenic variants, this implied a reduction from 6,465,390 and 4,606,664 variants to 875,383 and 601,823. (See Table 2)

Overall, as we can see in Table 2, our pathogenic variant dataset suffered a reduction of 7.3-fold mainly due to excluding variants other than missense type, but also to conserving only AD inheritance type genes. In contrast, when referring to non-pathogenic variants, a more drastic drop can be noticed. From the original 6,47 million missense variants in gnomAD v2.1, only 875,383 of these variants accomplished these requirements, while v3 only conserved 601,823 missense variants from the original 4,61 million.

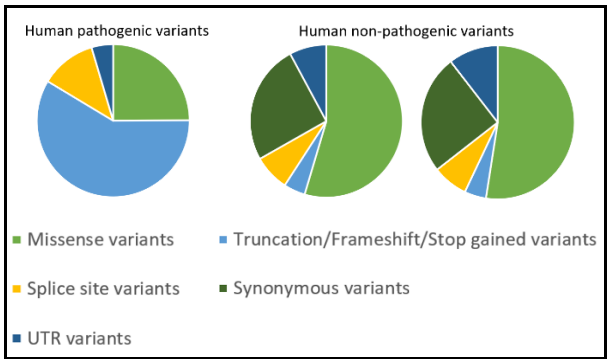

Fig. 2. Frequencies of the different human pathogenic and non-pathogenic variant types. From left to right, variants obtained from ClinVar, gnomAD v2.1 and gnomAD v3.

Table 1. The number of variants and genes after the processing step.

	PATHOGENIC (ClinVar)	NON-PATHOGENIC (gnomAD v2.1)	NON-PATHOGENIC (gnomAD v3)
Missense variants	52,190	6,465,390	4,606,664
Truncation/Frameshift/Stop gained variants	124,945	541,008	400,856
Splice site variants	25,140	903,882	653,326
Synonymous variants	-	3,003,140	2,200,104
UTR variants	9,740	929,699	920,608
Total variants	208,426	18,017,079	14,569,514
(Unique genes)	(6,563)	(18,656)	(19,033)

Table 2. Number of variants before and after each filter is applied.

	PATHOGENIC (ClinVar)	NON-PATHOGENIC (gnomAD v2.1)	NON-PATHOGENIC (gnomAD v3)
Total variants in the Human genome	208,426	18,017,079	14,569,514
Missense variants	52,190	6,465,390	4,606,664
Missense variants in AD genes	28,587	875,383	601,823

Identification of equivalent positions between homologous variants

Our goal is to extrapolate pathogenicity between homologous variants, i.e., compare the pathogenicity between variants able to cause a disease due to a change in one single amino acid. Two variants were considered homologous if they occur in members of the same Pfam family and modify amino acids in the same position in the family alignment (see Methods).

Our filtered dataset contained variants located in 1,612 different Pfam families. This is a small fraction of the 6,680 Pfam families that contain human proteins. Moreover, not all variants could be mapped in the Pfam alignments. Consequently, pathogenic variants were reduced from 28,587 to 19,318, while non-pathogenic variants were reduced by 3-fold in both gnomAD versions (see Table 3). The number of variants was reduced due to the fact that no equivalent position was found in Pfam for all of them. Also, we only conserved the variants that matched the initial amino acid of the protein change with the amino acid present in the Pfam alignment (see Methods). A combined dataset of 492,433 variants was finally considered, from which 96.1% were non-pathogenic variants and 3.9% were pathogenic.

Comparison of pathogenicity between homologous variants

We seek for pairs of homologous variants. That is, variants in members of the same family that occur in the same position according to the family multiple sequence alignment file (see Methods). Two analyses were performed considering variants presenting the exact amino acid change or when resulting amino acids exhibit similar physicochemical properties.

a) Assessing allele frequencies

Despite gnomAD being a non-pathogenic variant database, it may include some (yet unknown) pathogenic variants. If present, these variants should have very low allele frequencies (AF), and thus, we set different AF thresholds at $10^{-6}, 10^{-5}, 10^{-4}, 5*10^{-4}$ and 10^{-3} to overcome this potential contamination.

Those variants with the lowest AF are more susceptible to be pathogenic. In table 4 we can observe the non-pathogenic variants number at each AF threshold, before seeking variants with Pfam alignment. On the one hand, setting a permissive threshold such as AF $>10^{-6}$, sample data is reduced in 2.54-fold in v2.1 and 2.3-fold in v3 with respect to non-filtered condition, but we may be including some unknown pathogenic variants. On the other hand, setting a strict threshold, such as AF $>10^{-3}$, we assure a smaller susceptibility to keep unknown pathogenic variants, but the sample data is largely reduced, maintaining 4,764 and 4,564 non-pathogenic variants in v2.1 and v3, respectively. Due to the absence of more data about the unknown pathogenic variants among our non-pathogenic variants' dataset, it was not possible to firmly set a threshold, so we followed our analyses maintaining all different AF thresholds.

b) Strict pair analysis

In a first analysis, we observed pairs corresponding to homologous variants which present the same protein position and the same initial and final amino acids. These pairs were classified into three groups: i) pathogenic/pathogenic (P-P) when both homologous variants were labelled as pathogenic); ii) non-pathogenic/non-pathogenic (N-N) when both homologous variants were non-pathogenic); and iii) non-pathogenic/pathogenic (N-P) when the variants disagreed in pathogenicity. As mentioned before, we kept different variants' datasets considering different AF thresholds on the non-pathogenic variant datasets, so we sought for homologous variant pairs in all of them.

In Figure 3A, we see the addition of P-P and N-N pairs versus the number of N-P pairs. As the filter becomes more restrictive and less non-pathogenic variants are conserved, fewer N-P pairs are encountered. This tendency is also followed in Figure 4A, where we can observe the total number of pairs seen for each AF threshold filter condition, with respect to an error percentage. This error percentage was calculated dividing the number of pairs N-P by the total number of pairs, and was plotted in respect of the total number of pairs. As we can see, the error percentage becomes

Table 3. Number of variants having an equivalent position found in Pfam. When the initial amino acid of the protein change matches with the amino acid present at the same site in a Pfam alignment, we refer to equivalent positions.

	PATHOGENIC (ClinVar)	NON-PATHOGENIC (gnomAD v2.1)	NON-PATHOGENIC (gnomAD v3)
Total variants (Unique genes)	28,587 (1,449)	875,383 (1,850)	601,823 (1,858)
Number of variants with Pfam alignment	19,318	273,909	199,206

Table 4. Total variants at different allele frequency (AF) thresholds set in non-pathogenic variant datasets.

		NON-PATHOGENIC (gnomAD v2.1)	NON-PATHOGENIC (gnomAD v3)
Missense Variants with AF	No frequency filter	875,383	601,823
	10^{-6}	344,517	265,545
	10^{-5}	55,037	43,230
	10^{-4}	14,598	13,040
	$5 * 10^{-4}$	6,748	6,330
	10^{-3}	4,764	4,564

lower when the AF threshold is stricter, becoming an error of 0.6% when the AF>10^{-3}, in contrast with the 8.5% when any AF filter is applied. (See numbers in Supp. Tables S9 and S13) We observe a high reduction of the error percentage from 8.8% when the filter AF>10^{-6} is applied, to 1.8% when setting AF>10^{-4} and 0.8% when AF>$5 * 10^{-4}$. These results propose that our approach is more efficient when the AF threshold is more restrictive. The possible presence of unknown pathogenic variants among the non-pathogenic ones, can be increasing the number of N-P pairs, which increases the error percentage in less restrictive AF thresholds.

c) Similar pairs analysis

As a second step, we asked whether extrapolation to homologous variations could also be possible for strict and similar variations, i.e., adding similar initial and/or final amino acids. Amino acids were considered similar if the score between the two amino acids in a BLOSUM62 substitution matrix was positive, indicating similar physicochemical properties (see Methods). In this approach, homologous variant pairs were also sought in datasets presenting the different AF thresholds applied in non-pathogenic variants.

When similar final amino acids were searched, we noticed a high increase in the addition of P-P and N-N variants encountered, compared to the strict pairs analysis. When no AF filter was applied, we observed a 4-fold increase in the total number of pairs, and a close 5-fold increase in AF>10^{-3} and >$5 * 10^{-4}$ (Fig. 3B. See numbers in Supp. Tables S10 and S14). Interestingly, the number of N-P pairs presented a higher increase than the total number, a 5-fold increase when no filter was applied and 5.4-fold in AF>10^{-3}. Despite that, the tendency is equal to the strict pairs condition: the stricter the AF filter is, the less N-P pairs are detected. In Fig. 4B, we see that the total number of pairs increases in all AF filters as the number of P-P + N-N. As in the strict pairs condition, the error suffers a high reduction from the AF>10^{-6} to the AF>$5 * 10^{-4}$, but in this case, it drops from 13.3% to 1.8%, noticing a value stagnation in the latter. Surprisingly, in AF>$5 * 10^{-4}$ increases only a tenth the error to 0.9%, a slight increase considering that the sample size presents a 6-fold increase.

In addition, we checked the results when searching for pairs with strict and similar initial amino acids. We do not observe an intense increase of total pairs as in the final amino acid condition, if we compare with the strict pairs condition, we do obtain a very few increase both on total variants in all AF filters, and in P-P + N-N addition (Fig. 3C and 4C, see numbers in Supp. Tables S11 and S15). The tendency is maintained, N-P pairs are reduced as the AF filter is more restrictive. In Figure 4C, we see that error percentages are maintained with respect to strict pairs condition, where the highest difference is noticed when no AF filters are applied, presenting 7.1% error compared to the 8.5% in the strict pairs condition.

Regarding the search for pairs with similar and strict initial amino acids, we did not reveal noteworthy findings as expected. When an amino acid change occurs, the importance relies on the conservation of the amino acid introduced, meaning that the change is tolerated because it has similar physicochemical properties or not, due to an alteration in the function of the protein. In contrast, the initial amino acid does not play a role due to its replacement.

Comparing with the strict pairs condition, the search for similar and strict initial amino acids did not show upgrades in our results. For that reason, we do not consider it worth to include and discuss the condition where pairs with strict and similar initial and final amino acids were sought (See Supp. Tables S12 and S16).

Given that, compared to the strict pairs condition, the search of similar and strict final amino acids showed a considerable increase in the sample size without increasing the error rate equivalently, we consider it the best approach for searching homologous variant pairs. More specifically, we consider that variants overstepping an AF>$5 * 10^{-4}$ showed a good non-pathogenic variant representation in the dataset, and presented an

Table 5. Number of variants having an equivalent position found in Pfam. When the initial amino acid of the protein change matches with the amino acid present at the same site in a Pfam alignment, we refer to equivalent positions.

	PATHOGENIC (ClinVar)	NON-PATHOGENIC (gnomAD v2.1) variants with AF>$5 * 10^{-4}$	NON-PATHOGENIC (gnomAD v3) variants with AF>$5 * 10^{-4}$
Total variants (Unique genes)	28,587 (1,449)	6,748 (1,388)	6,330 (1,349)
Number of variants with Pfam alignment	19,318	1,729	1,723

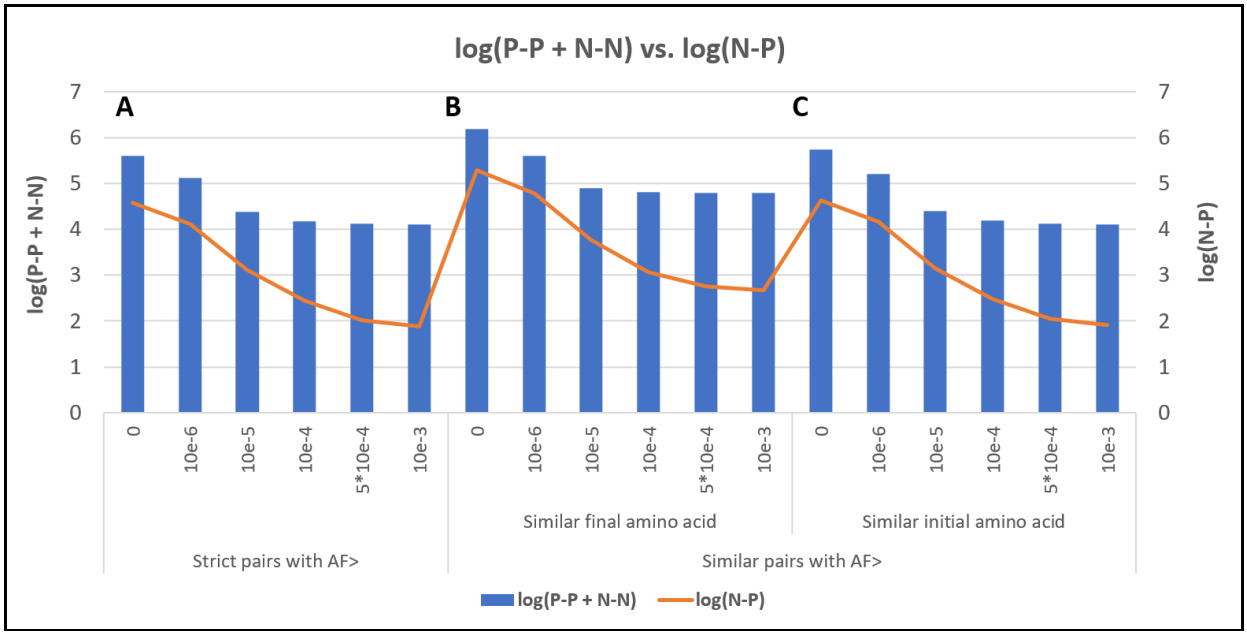

Fig. 3. Summation of P-P and N-N pairs with respect to the N-P pairs for all allele frequencies (AF) in: **A)** Strict pairs, **B)** Similar final amino acid, **C)** Similar initial amino acid. P-P refers to pairs of variants presenting pathogenic annotations, N-N to non-pathogenic annotations, and N-P to pairs of variants that disagree on their annotations. Values are log normalized.

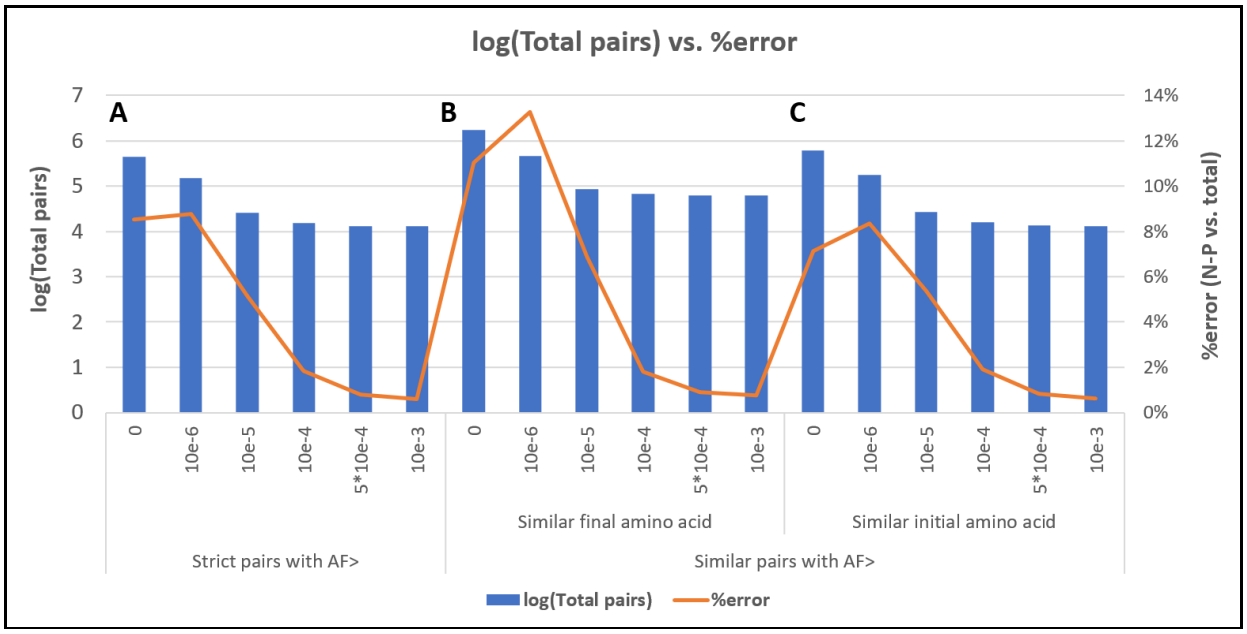

Fig. 4. Total number of pairs with the error percentage for all allele frequencies (AF) thresholds in: **A)** Strict pairs, **B)** Similar final amino acid, **C)** Similar initial amino acid. P-P refers to pairs of variants presenting pathogenic annotations, N-N to non-pathogenic annotations, and N-P to pairs of variants that disagree on their annotations. Total pair values are log normalized. %error is calculated dividing N-P pairs/total pairs and multiplied by 100.

extraordinarily low error rate.

Thus, we assume that non-pathogenic variants overstepping this threshold are not likely to be pathogenic. In Table 5, we can observe the number of variants having an equivalent position found in Pfam for this AF threshold: 1,729 variants from 6,748 in v2.1, and 1,723 variants from 6,330 variants in v3.

Searching for similar final amino acids with an AF>$5 * 10^{-4}$, we obtained 63,192 total pairs featuring 612 Pfam families, out of the 1,090 families that presented alignment for these proteins (Table 6). The reason why there were not found pairs in 478 families is because there were very unrepresented families: 357 families presented 4 variants or less. Comparing a few number of variants between them results in a low probability to find pairs. Despite that, we found pairs in 76 additional Pfam families compared to the strict pairs condition. From the total pairs, 60,822 correspond to P-P pairs, 1,799 are N-N pairs, and 571 are N-P pairs. Interestingly, considering the N-P pairs with respect to the total, we obtained an error of 0.9%. This is because if the final amino acid presents similar physicochemical properties, the effect of the amino acid change in the same equivalent position between homologous proteins may be similar. This small number of false positives confirms that our hypothesis is true and can be safely accepted.

Table 6. Strict and similar pairs obtained computing BLOSUM62 to final amino acid once gnomAD datasets are filtered by $5 * 10^{-4}$ allele frequencies.

	Pairs with similar final amino acid
P-P	60,822
N-N	1,799
N-P	571
Total (Pfam families)	63,192 (612)
% error	0,9%

The total number of pairs was increased by almost 5-fold in similar final amino acid condition from the strict pair condition. This computation allowed us to highly increase the sample size and observe that the error rate did not increase proportionally. Accordingly, one of the key highlights of our results is the low error percentage obtained, demonstrating the accuracy and reliability of our approach.

Conclusions and future directions

It is reasonable to consider that missense variants that occur at equivalent positions in proteins of the same family are likely to have the same effect if the original and the mutated amino acid are the same. We noticed, however, that this had never been probed and that this assumption is not employed by the available predictive tools. In the present study we have demonstrated that this assumption can be safely taken, with an error of 0.8%. Moreover, we show that this is also true when the final amino acids are not the same, but exhibit similar physicochemical properties. This allows us to safely expand the number of variants with pathogenicity annotations without performing (costly) functional experiments. The only requirement is the availability of functional data in an equivalent position in a member of the same family and a sequence alignment between the corresponding protein sequences, which in our case were the family multiple sequence alignments provided by Pfam. The results hold significant promise for enhancing the accuracy of current predictive tools, which can be very useful in the identification of disease-causing variants in rare diseases. We recommend incorporating this concept in future tool development to harness the full potential of available data and drive advancements in the field.

References

1. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, et al. A global reference for human genetic variation. Nature 2015;526(7571):68–74.

2. Karen Eilbeck, Aaron Quinlan and MY. Setting the score: variant prioritization and Mendelian disease. Nat Rev Genet 2017;18:599–612.

3. EURORDIS. Rare Diseases : understanding this Public Health Priority. 2005;(November).

4. Chong JX, Buckingham KJ, Jhangiani SN, Boehm C, Sobreira N, Smith JD, et al. The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. Am J Hum Genet 2015;97(2):199–215.

5. NIH. The Cost of Sequencing a Human Genome [Internet]. [cited 2023 Jun 28];Available from: https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

6. Eurordis. Survey of the Delay in Diagnosis for 8 Rare Diseases in Europe ('EURORDISCARE 2'). Eurordis 2007;1.

7. Sim NL, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. SIFT web server: Predicting effects of amino acid substitutions on proteins. Nucleic Acids Res 2012;40(W1):452–7.

8. Sandell L, Sharp NP. Fitness Effects of Mutations: An Assessment of PROVEAN Predictions Using Mutation Accumulation Data. Genome Biol Evol 2022;14(1):1–15.

9. Boris Reva; Yevgeniy Antipin; Chris Sander. Mutation assessor [Internet]. [cited 2023 Jul 28];Available from: http://mutationassessor.org/r3/

10. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. 2013.

11. Flanagan SE, Patch AM, Ellard S. Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations. Genet Test Mol Biomarkers 2010;14(4):533–7.

12. Min L, Nie M, Zhang A, Wen J, Noel SD, Lee V, et al. Computational Analysis of Missense Variants of G Protein-Coupled Receptors Involved in the Neuroendocrine Regulation of Reproduction. Neuroendocrinology 2016;103(3–4):230–9.

13. Santos-Gómez A, García-Recio A, Miguez-Cabello F, Soto D, Altafaj X, Olivella M. Identification of homologous GluN subunits variants accelerates GRIN variants stratification. Front Cell Neurosci 2022;16(December):1–10.

14. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, et al. The Pfam protein families database. Nucleic Acids Res 2009;38(SUPPL.1):211–22.

15. Pfam. Frequently Asked Questions (FAQs) [Internet]. [cited 2023 Jul 23];Available from: https://pfam-docs.readthedocs.io/en/latest/faq.html

16. The Uniprot Consortium. UniProt: the Universal Protein Knowledgebase in 2023. 2023;51(November 2022):523–31.

17. UniProt. UniProtKB/Swiss-Prot protein knowledgebase release 2023_01 statistics [Internet]. 2023;Available from: https://ftp.uniprot.org/pub/databases/uniprot/previous_releases/release-2023_01/knowledgebase/UniProtKB_SwissProt-relstat.html

18. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: The protein families database in 2021. Nucleic Acids Res 2021;49(D1):D412–9.

19. Smedley D, Haider S, Durinck S, Pandini L, Provero P, Allen J, et al. The BioMart community portal: An innovative alternative to large, centralized data repositories. Nucleic Acids Res 2015;43(W1):W589–98.

20. Pruitt K, Brown G, Tatusova T, Maglott D. Chapter 18 : The Reference Sequence (RefSeq) Database Database. NCBI Handb 2002;(Bethesda(MD): National Center for BIotechnology Information (US)):1–22.

21. 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):D835–44.

22. 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.

23. gnomAD. gnomAD v2.1 [Internet]. 2018 [cited 2023 Jun 27];Available from: https://gnomad.broadinstitute.org/news/2018-10-gnomad-v2-1/

24. gnomAD. gnomAD v3.0 [Internet]. 2019 [cited 2023 Jun 27];Available from: https://gnomad.broadinstitute.org/news/2019-10-gnomad-v3-0/

25. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 2005;33(Database Issue):D514.