NAR Genomics and Bioinformatics, 2024, **6**, lqae154
https://doi.org/10.1093/nargab/lqae154
Advance access publication date: 21 November 2024
Standard Article

There will always be variants of uncertain significance. Analysis of VUSs

Haoyang Zhang, Muhammad Kabir, Saeed Ahmed and Mauno Vihinen*

Department of Experimental Medical Science, BMC B13, Lund University, SE-22184 Lund, Sweden

*To whom correspondence should be addressed. Tel: +46 72 5260022; Email: mauno.vihinen@med.lu.se
Present address: Saeed Ahmed, Department of Computer Science, University of Swabi, Swabi, Khyber Pakhtunkhwa 94640, Pakistan.

Abstract

The ACMG/AMP guidelines include five categories of which variants of uncertain significance (VUSs) have received increasing attention. Recently, Fowler and Rehm claimed that all or most VUSs could be reclassified as pathogenic or benign within few years. To test this claim, we collected validated benign, pathogenic, VUS and conflicting variants from ClinVar and LOVD and investigated differences at gene, protein, structure, and variant levels. The gene and protein features included inheritance patterns, actionability, functional categories for housekeeping, essential, complete knockout, lethality and haploinsufficient proteins, Gene Ontology annotations, and protein network properties. Structural properties included the location at secondary structural elements, intrinsically disordered regions, transmembrane regions, repeats, conservation, and accessibility. Gene features were distributions of nucleotides, their groupings, codons, and location to CpG islands. The distributions of amino acids and their groups were investigated. VUSs did not markedly differ from other variants. The only major differences were the accessibility and conservation of pathogenic variants, and reduced ratio of repeat-locating variants in VUSs. Thus, all VUSs cannot be distinguished from other types of variants. They display one form of natural biological heterogeneity. Instead of concentrating on eradicating VUSs, the community would benefit from investigating and understanding factors that contribute to phenotypic heterogeneity.

Introduction

Variation interpretation refers to the explanation of the impact and health relevance of genetic variations, either inherited or somatic. Interpretation guidelines from the American College of Genetics and Genomics and the American Cancer Society (ACMG/AMP) (1) provide a systematic scheme that summarizes eight types of information. Variants are described with a five-tier classification, based on the strength of the information. The tiers are benign, likely benign, pathogenic, likely pathogenic, and of uncertain significance. This scheme is widely used in clinical diagnosis in many countries, but there may be local changes and refinements, such as those from the Association for Clinical Genomic Science in the United Kingdom (2), or schemes specific for certain diseases, e.g. for breast cancer (https://enigmaconsortium.org/enigma-classification-criteria/).

Recently, Fowler and Rehm published a Perspective piece with the provocative title 'Will variants of uncertain significance still exist in 2030?' (3), where the authors claimed that most if not all variants of uncertain significance (VUSs) can be grouped into the four other categories within a few years. The article was based on the 'bold predictions' made by the National Human Gene Research Institute (NHGRI) (4). Fowler and Rehm used a mechanistic and methods-based approach and angle. They discuss how standardized variation interpretation, novel and improved computational tools, multiplexed functional assays, and improved data sharing together will reduce the number of VUSs. All these factors will contribute to more reliable classification of variations. However, the authors missed biological bases for VUSs, including the origin and relevance of biological heterogeneity. These factors will always contribute to variation outcomes and phenotypes.

Here, we discuss why the eradication of VUSs is not possible, identify problems related to VUSs, and report an extensive analysis of VUSs in comparison to pathogenic (P), benign (B) and conflicting variants in terms of various gene, protein, structure, and variant parameters. The results indicated that VUSs are very similar to variants in other categories, thus it will not be possible to reclassify all VUSs as benign or pathogenic variants. Our analysis focused on amino acid substitutions and nucleotide variations leading to them; however, we are confident that similar observations can be made with other types of variants.

Misconceptions and problems with VUSs and variation terminology

Definition of VUS

The ACMG/AMP definition of VUS is 'if a variant does not fulfill criteria using either of these sets (pathogenic or benign), or the evidence for benign and pathogenic is conflicting, the variant defaults to Uncertain Significance' (1).

There are thus two types of VUSs; those for which there is not enough evidence to distinguish between benign and pathogenic cases and those for which there is conflicting evidence for opposite annotations. By collecting more

Received: Auguat 5, 2024. **Revised:** October 2, 2024. **Editorial Decision:** October 28, 2024. **Accepted:** October 29, 2024

information, it will be possible to resolve the classification for many variants. When different individuals with the same variant display different phenotypes the cases are conflicting. This is normal and originates from several reasons and will never disappear. The penetrance of variants and conditions vary, and some factors may protect and/or diminish the effects of a variation. The dosage of the allele, modifier genes, variation type, environmental effects, lifestyle and epigenetic modifications are further factors that contribute to individual heterogeneity (5).

VUSs overlap with pathogenic and benign variants

It is not uncommon to hear or read that VUSs form a category between benign and pathogenic variants, see for example (6–9). Numbering of tiers from 1 to 5 is common, on this scale VUSs are presented in number 3. However, this practice is wrong. The ACMG/AMP classification does not number the tiers or state that VUSs are an intermediate class.

As the definition of VUSs indicates, these variants do not fulfill the criteria for being pathogenic or benign (1); but they are not outside the pathogenic-benign dichotomy. Figure 1A shows the relationships among the five tiers. The benign and pathogenic variants are at the two ends, and the classification of VUSs in this range is unknown; therefore, they are of uncertain significance.

ACMG/AMP interpretation is for individuals, not for variants

A common source of problems is the use of interpretation guidelines to describe variants in general. The guidelines are intended for the interpretation of variants in individuals carrying them, based on many features. Clinical signs, symptoms and parameters vary from individual to individual. Among the eight data types in the interpretation scheme, segregation, *de novo*, allelic and other data are for an individual, and functional data may also be for an individual.

Interpretations have often been extended beyond describing variant of an individual. It can work for clearcut cases, i.e. those that are pathogenic or benign in (almost) all individuals. In VUSs, individual heterogeneity affects the phenotype and general variation interpretation cannot be made. Therefore, such efforts will involve large numbers of VUSs.

An example of a successful variation-level interpretation is the Variation Interpretation Committee (VIC) of the International Society for Gastrointestinal Hereditary Tumours (InSiGHT), which provides interpretations for variations in four mismatch repair system genes/proteins (MLH1, MSH2, MSH6, PSM2) and some other genes and proteins involved in gastrointestinal hereditary tumors (10). Although the InSiGHT VIC members represent the major centers in the world and have worked for more than 10 years, the four genes included 28 109 variants (April 2024), 31.2% of which were VUSs. In the case of unique variants, 35.4% out of 2619 variants were classified as VUS. Another example is the Dutch variation classification by nine laboratories based on standardized interpretation procedures (11). Even this scheme classifies large numbers of VUSs, almost 80 000 at the moment.

Problematic wording

Language about VUSs is, in our opinion, often problematic. Many authors have described the burden of VUSs and some other variant types, e.g. (12,13). Since the disease relevance of VUSs is not known, they cannot be used, e.g. for diagnosis. It is evident that there is a certain burden for healthcare (14); however, the situation is similar to many other health related aspects. It is problematic to call this natural phenomenon a burden.

Another example is conflicting classification. For example, ClinVar calls variants for which the submitters do not agree on the classification as conflicting. Conflict means disagreement and can be considered negative. Differences in the phenotypes of individuals carrying the same variant can be widely different because of normal biological variation. It would be more neutral to call such cases, e.g. as having different classifications or leading to different phenotypes. Other authors call these cases discordant variants (9,15,16), which again has a negative connotation. Despite our criticism, we use the term ‘conflicting variant’ in this paper to be consistent with ClinVar, the source of the variants.

In addition to the issues mentioned above, there are other problematic practices in variant naming, including missense, nonsense and frameshift variants; indels; truncations; gain of function; loss of function and synonymous variants (17,18).

VUSs are due to normal biological heterogeneity

Currently (April 2024), ClinVar contains 2 808 943 germline variation records. VUSs account for 1 251 444 variants, 207 683 are pathogenic variants, 113 104 likely pathogenic variants, 255 441 benign variants, and 917 396 likely benign variants. VUSs were clearly the largest group (44.6%). The numbers for benign and pathogenic cases include variants with all the review statuses. There were 121 847 conflicting classifications, 951 for P/LP versus LB/B, 18 999 for VUS versus P/LP, and 102 911 for VUS versus LB/B. The largest group of conflicting variants is for VUS versus LB/B.

VUSs do not fulfill the criteria for pathogenic or benign classification and display biological heterogeneity. The signs and symptoms of patients with the same disease vary in practically every condition (with embryonic lethality and other extreme cases being exceptions).

Evolution constantly generates new natural variations. VUSs are one component of individual genetic heterogeneity (5), which is a form of pervasive biological variation known as poikilosis (19). Differences in the phenotypes of carriers of the same variants display a spectrum. These differences are associated with differential penetrance, presence or absence of modifier variants and molecules, gene and allele dosage, differential gene expression, combined effects with other variants and genes, and many other factors. Thus, VUSs are due to normal biological variation and therefore there will always be some VUSs.

In summary, variants display heterogeneous phenotypes. In some individuals, a certain variant is harmful and diagnosed as pathogenic. Other individuals having the same variation may have milder phenotype or be classified healthy. When biological heterogeneity is forgotten enters the fallacy of pathogenic-benign dichotomy. To address the issue if all VUSs can be classified as pathogenic or benign, we performed detailed statistical analysis, which showed that variants in benign, pathogenic, VUS and conflicting classes had largely similar properties over a wide range of characteristics. Each parameter showed a continuum on the variant classes. It is possible to define cutoff to classify cases. But such cutoffs are not

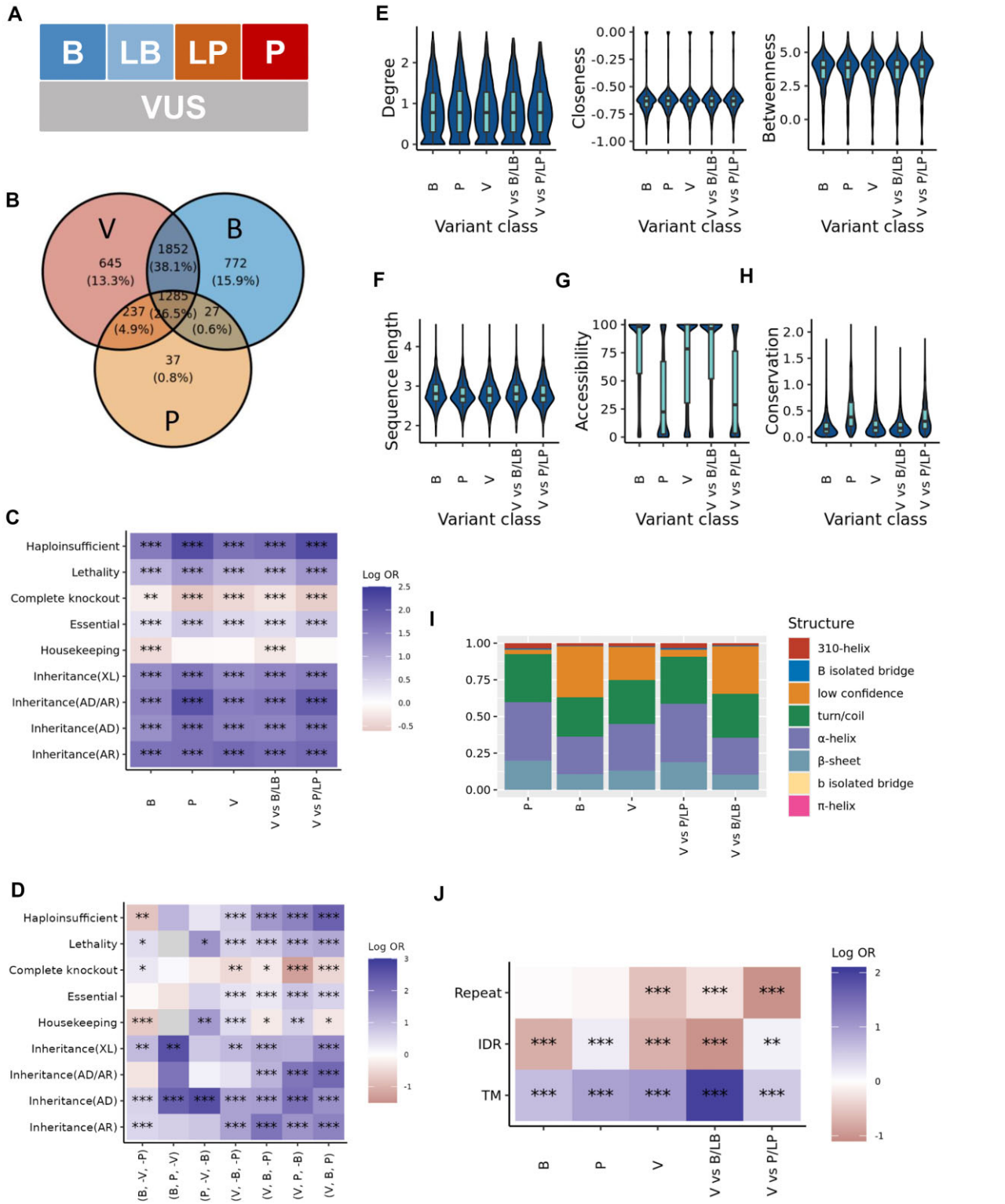

Figure 1. Analysis of variant classes. (**A**) Five tiers and their relationships in the ACMG/AMP classification. VUSs are either (likely) pathogenic or (likely) benign, however, the classification is not known. (**B**) Distribution of the investigated genes that contained benign (B), pathogenic (P) and uncertain (VUS) variants and their overlaps. The numbers are for genes, and percentages are given in brackets. (**C**) Enrichment analysis of protein functional and genetic properties to variant classes. (**D**) Enrichment analysis of unique and shared proteins according to the Venn diagram in B. The scale to the right indicates ORs from Fisher's test. (**E**) Logarithmic distribution of network parameters degree, closeness, and betweenness for proteins in the variant classes. (**F**) Distribution of amino acid sequence lengths in the proteins in the variant classes. (**G**) Accessibilities of the original amino acids in variation positions. Defined from protein three-dimensional structures obtained with AlphaFold. (**H**) Conservation score for the original amino acids. (**I**) Distribution of protein secondary structural element classifications for the original amino acid positions determined with STRIDE. (**J**) Enrichment analysis of IDRs, and transmembrane and repeat regions. The IDRs and TM regions were obtained from the DisProt and Human Transmembrane Proteome, respectively. Repeats were identified with T-REKS from protein sequences. *P* value significance: $*P \leq 0.05$; $**P \leq 0.01$; $***P \leq 0.001$. In E to H, the pale blue box plot indicates the median (in the middle of the bar) and the interquantile range.

perfect and do not correctly classify all cases, since they show heterogeneity. Therefore, there will always be VUSs.

Data and methods

Variant collection

The data were collected from ClinVar (20) and LOVD (21). Five datasets were obtained from ClinVar: VUSs, two types of conflicting cases (VUS versus P/LP and VUS versus LB/B), pathogenic and benign variants. First, we searched for each of the categories on the ClinVar website by using 'missense' as a keyword. The results were then filtered with 'germline' as Classification type, one of the four Germline classifications (conflicting classifications, benign or pathogenic), 'missense' as Molecular consequence, and 'single nucleotide' as Variation type. For pathogenic and benign variants, the Review status was set to two stars or higher. The data were downloaded and further filtered to remove duplicates and cases with insertions, deletions, or undefined variations. Then, the variants were matched to MANE transcripts (22). There were 146 336 VUSs, 12 016 VUS versus P/LP variants, and 52 103 VUS versus B/LB variants. We found 21 563 benign variants and 11 343 pathogenic variants.

To increase the number of pathogenic variants, we searched for additional cases from LOVD. We used the LOVD shared website, searched for all variants affecting transcripts, and then selected '+/+' in Effect, Clinical classification as 'pathogenic', cDNA change without symbols '+' or '-' (which indicate variation locations in introns), deletion 'del', duplication 'dup' or insertion 'ins', and Protein change without '?' for unsure classification, 'p.0' and 'p.(0)' for missing protein, termination '*', frameshift 'fs' or synonymous/silent '='. The data were further filtered by removing cases with ambiguity codes either in nucleotide or protein descriptions. We obtained 19 832 variants. The reference sequences for most of the LOVD-based variants were not in MANE. Next, we used VEP (23) to annotate the variants. Then, we merged the variants with the ClinVar data. Variants with missing cDNA, protein, or genomic details after the data enrichment steps were excluded. Duplicates with the ClinVar-mined variants were excluded. Variants with different classifications from different sources were excluded.

Overall, there were 146 186 VUSs, 11 656 VUS versus P/LP variants, 51 751 VUS versus B/LB variants, 21 466 benign variants and 14 338 pathogenic variants. The dataset is available in VariBench (24).

Gene, protein and variant properties

The variants in the five categories were analyzed for several properties. Nucleotide distributions in the variant groups were supplemented with combined two-base groups for purine and pyrimidine nucleotides (A and G versus C and T), weak and strong binding nucleotides (A and T versus C and G), and keto and imino nucleotides (G and T versus A and C). In addition to amino acid distributions, residues were classified into six categories (G1 to G6): hydrophobic (C, F, I, L, M, V, W, Y), negatively charged (D, E), positively charged (H, K, R), conformational (G, P), polar (N, Q, S) and others (A, T), respectively (25).

Genes and proteins were grouped into several categories. A total of 2833 housekeeping proteins were obtained from https://housekeeping.unicamp.br. A total of 6559 essential/indispensable proteins were from (26).

Complete knockout genes were obtained by combining gene lists from published studies. We downloaded 781 genes from (27) and 1317 genes from (28) and removed duplicates. For the genes in (29), the following steps were performed. We removed all rows with a sequence MAF >2% after which we had 6275 genes. We included 'Number of compound heterozygous carriers' in a variant pair where the MAF was <2% for both variants and obtained 462 samples excluding 0 and NA cases. We included 'Observed number of imputed homozygotes' (excluding zero and NA) and obtained 1299 samples. Then, we combined the observed data items and removed duplicates. Finally, we obtained 1156 unique genes. In total, there were 2633 unique genes in the three datasets.

Lethality-related proteins were obtained from the Mouse Genome Database (30). The human orthologs of murine genes were considered as essential, when the murine gene was annotated with one of the following phenotypes: embryonic, prenatal, or perinatal lethality. There were 1786 unique genes.

Haploinsufficient proteins were obtained from the ClinGen Dosage Sensitivity Map (https://www.ncbi.nlm.nih.gov/projects/dbvar/clingen/). By using the online system, we turned Genes 'on' and Region 'off' and obtained 1 525 genes. We excluded cases with 0 (no evidence), 1 (little evidence) or 2 (emerging evidence) in the HI or TS columns. After filtering there were 1 499 genes.

Inheritance patterns were obtained from the Clinical Genomic Database at https://research.nhgri.nih.gov/CGD/ (31). We collected data on autosomal dominant (AD), autosomal recessive (AR), mixed AD/AR and X-linked (XL) inheritance. These details were available for 1 078 genes.

Protein structures predicted by AlphaFold2 (32) were obtained from the AlphaFold Protein Structure Database (33), from AlphaFold 3 (34), or from CHESS 3 (35). Secondary structural elements were defined by STRIDE (36), and the solvent-accessible surface areas (SASAs) of the original amino acids were calculated using the FreeSASA Python module (37) from the structure files.

The data for human transmembrane proteins were downloaded from the Human Transmembrane Proteome (38). We obtained 5 467 human transmembrane (TM) regions, which were mapped to 19 352 MANE reference sequences to identify positions within TMs. Intrinsically disordered protein regions (IDRs) were obtained from the DisProt website (39). Duplicates were removed and directly adjacent IDRs were merged. Then the sequences were mapped to the MANE reference sequences. A total of 1706 proteins contained IDRs.

Repeated segments in protein sequences were identified with T-REKS (40) using default parameters.

Protein-protein interaction (PPI) data were downloaded from the STRING database (41) by using 500 as the experimental score threshold. We utilized the igraph Python package from https://igraph.org to determine the degree (number of interactions), closeness and betweenness of each protein.

The presence of variations in CpG islands was downloaded from a track in the UCSC genome browser at http://hgdownload.cse.ucsc.edu/goldenpath/hg38/database/cpgIslandExt.txt.gz.

The data for 81 actionable genes were obtained from ACMG Recommendations for Reporting of Secondary Findings in Clinical Exome and Genome Sequencing v. 3.2 (42) at https://www.ncbi.nlm.nih.gov/clinvar/docs/acmg/.

Sequence conservation was determined with a position-specific scoring matrix (PSSM). We downloaded all sequences for mammals, rodents and vertebrates from UniProt_T and saved as a database. Each MANE sequence was compared to the database by running Blast version 2.12.0+ (43). The maximum number of target sequences was set to 20 000, and the e-value was set to 0.001. If there were fewer than three hits, the protein was excluded. The sequence identity threshold was 30 for minimum and 80 for maximum, and the alignment length had to be greater than seq_length * 0.8. The obtained sequences were clustered with CD-HIT (44) with parameters of -c 0.8 and -aL 0.8 for the sequence identity threshold and alignment coverage for the longest sequence, respectively. Representatives were identified for each cluster from the CD-HIT outputs. If there were ≤4 sequences in a cluster, the sequence was removed. When a MANE protein was identified as a cluster representative, another protein was chosen from that cluster. *makeblast* operation was used to collect sequences from the databases. PSIBLAST (43) with num_alignments = 50 000 and num_iterations = 3 was used to calculate conservation scores for all the MANE protein sequences. We obtained PSSM files for 19 034 MANE IDs.

Statistical analysis

Numerous statistical tests were made for several parameters to test if the variants classified to different categories differed from each other.

To evaluate the differences in structural and functional elements in different variant classes, we used Fisher's test to measure the enrichment of variants in TM regions, IDRs, repeats, inheritance patterns and gene functional properties, including housekeeping, essential, complete knockout, lethality, actionability, and haploinsufficient proteins. Enrichment/depletion in the variant classes was calculated in comparison to the human proteome. For example, when analyzing localization of VUSs to TMs, we counted how many VUS variants appeared in TM regions and outside these regions. We obtained also the total number of variants in TM regions and total number of variants not in TM regions, i.e. the total length of MANE proteins – the total number of protein variants in TM regions. Then, a cross table was made for Fisher's test.

The expected proportion of TM regions and IDRs was calculated by dividing the total number of variants present in transmembrane regions and IDRs by the length of all MANE sequences (11 271 977). The expected proportion of mode of inheritance and functional properties were calculated by dividing the number of proteins with these characteristics by the total number of proteins (19 352).

To measure substitution patterns of nucleotide and protein variants, we used connectivity matrix to illustrate the frequencies of changes. Each entry in the matrix represents the frequency of variants observed from the corresponding base/categories to the target base/categories. We utilized Fisher's exact test to determine whether any substitution types were significantly enriched or depleted in VUS, VUS versus P/LP or VUS versus B/LB, in comparison to pathogenic and benign variant categories. For example, when we assessed whether a substitution type was more or less prevalent in VUSs than in benign variants, we counted the number of VUSs with this substitution, the number of benign variants with this substitution, the number of VUSs without this substitution, and the number of benign variants without this substitution. A cross table was made to perform Fisher's test. This was repeated for every comparison. Fisher's log odds ratio (OR) values >0 and <0 indicate enrichment and depletion, respectively, in comparison to the other datasets. Fisher's log OR values were considered to be NA when there was 0 in the cross table and the test could not be applied.

ANOVA was used to compare accessibility and PPI features between variant classes.

Gene Ontology (GO) enrichment analysis was performed using the clusterProfiler R package (45). Holm-Bonferroni correction was applied to adjust for multiple comparisons in enrichment analyses. For ANOVA, Tukey's test was applied for multiple comparisons. All analyses were performed with R4.3.2.

Results

We investigated at different levels whether VUSs could be differentiated from variants in other categories. Our aim was to test whether the claim that VUSs can be classified in other categories was true. We investigated five variant classes: VUS, VUS versus P/LP, VUS versus B/LB, benign and pathogenic. Statistical tests were employed at gene, protein, structural and variant levels. Overall, the VUS versus P/LP class comparison to benign variants was similar to pathogenic variants, and the VUS versus B/LB class was similar to that of benign variants. The results indicate that VUSs do not largely differ from other variant classes. Although many results are statistically significant, the differences are marginal and biologically likely insignificant. The statistical significance is due to the large sample sizes.

Gene and protein level analysis

We identified the largest possible sets of verified variants from two high quality databases, ClinVar (20) and LOVD (21), and used them to investigate the properties of VUSs, pathogenic, benign and conflicting variations. Despite large numbers of original variations, filtering for reliable cases substantially reduced the sample size. After filtering, we identified 146 186 VUSs, 11 656 P/LP versus VUS variants, 51 751 VUS versus B/LB cases, 21 466 benign variants and 14 338 pathogenic variants. The numbers were large enough to facilitate reliable statistical analyses.

All types of variations were obtained from ClinVar. In the case of VUSs, benign and pathogenic variants, Review status was at least two stars (multiple submitters). For conflicting cases, this criterion could not be applied due to the small sample size. Therefore, all conflicting variants were included in the two categories. Pathogenic variants were supplemented with data from LOVD. All variants were mapped to MANE transcripts (22) to facilitate systematic studies.

LOVD contains >1 000 000 variants, 437 000 of which cause amino acid substitutions. Since LOVD does not allow programmatic linking of variations to information for individuals and diseases or downloading all contained data, we were left with effect and clinical classification as parameters for the quality of the data. The application of these filters substantially reduced the number of cases. Many variants were identical to those obtained from ClinVar and were thus eliminated. In the end, LOVD added only 2995 pathogenic variations to those in ClinVar. We converted all variants to

be MANE-based. All problematic cases and duplicates were removed.

Our data included 5120 unique genes. There were 1586 genes with pathogenic variants, 3936 genes with benign variants and 4 019 genes with VUS variants (Figure 1B). There were genes in all the domains of the Venn diagram, and a substantial number of genes (1285) contained variants in all three categories. The number of genes was the smallest for the pathogenic genes. A total of 26.5% genes were shared among the three functional effect categories. The number of group-specific genes was the smallest for pathogenic variants, only 37 genes. These ratios are likely to change in the future when more reliably annotated cases are identified. All genes are not expected to contain (m)any disease-causing variations, for example, nondisease nonhousekeeping genes are tolerant to most variations (46).

The types of genes were further investigated by grouping them into categories. Statistical significance for enrichment/depletion was obtained for nine categories, including haploinsufficiency, lethality, complete knockout, essential, and housekeeping genes and proteins. In addition, we had four categories for inheritance: X-linked (XL), autosomal dominant (AD), autosomal recessive (AR) and a combined class for the last two groups (AD/AR)).

In haploinsufficiency, expression of both gene copies is needed. If one allele contains a disease-related variation, the expression of the other allele is not sufficient for the required biological activity. Lethality-related genes were defined based on experiments in mice. Human genes homologous to lethality genes in the rodent model were identified. In the case of complete knockout genes, both alleles can contain harmful variants in some healthy individuals. Housekeeping genes are expressed in most cells and in most situations, they code for essential cellular functions. Inheritance patterns indicate how the phenotype of a gene is transmitted to offspring.

The investigated groups represent widely different gene and protein properties and could be considered to have different distributions to variant categories, e.g. if VUSs were different from other variant types. However, the results were very similar across the groups (Figure 1C and Supplementary Table S1). The gene groups were enriched in all the categories, apart from complete knockout and housekeeping genes which were depleted in the variant classes. The findings were highly statistically significant in almost all the analyses. We observed the same pattern throughout the study; the differences were usually very small and may not have biological relevance. The significant differences were due to the large sample sizes, and therefore, even minor differences were statistically significant. We mainly discuss differences (if any), and the significance of the observations is shown in the figures and Supplementary Tables. Housekeeping genes did not significantly differ, except for those in the benign and V versus B/LB classes. In all other cases, all variant classes showed enrichment of all the properties.

The depletion of housekeeping proteins is understandable because they are essential for many cellular functions. However, somewhat surprisingly, variations in these proteins were depleted only in the benign and V versus B/LB categories.

As an additional functional group, we investigated actionable genes for which secondary findings should be reported. There were only 81 genes, almost all of which were present in all the variant classes and were highly enriched. The reason for overrepresentation in every class is likely because these genes have been widely studied.

Figure 1D shows the same analysis for the unique and combined categories according to the Venn diagram in Figure 1B. Each section of the Venn diagram was investigated separately. Note that the numbers of unique pathogenic variants containing genes and genes shared by pathogenic and benign variants were very small, 37 and 27, respectively. Thus, the statistical data for these categories may be less reliable.

The enrichments and depletions were statistically more pronounced in Figure 1C and Supplementary Table S1 than in Figure 1D and Supplementary Table S2; however, the results were mainly in line. In this analysis, enrichment/depletion was similar for most of the functional categories. Only complete knockout and housekeeping genes were involved in both enrichment and depletion, and haploinsufficiency was different for benign-only genes than for other genes. Unique VUSs containing genes and categories combined with other types of variants contained the largest numbers of significant observations. VUSs differed from unique benign and pathogenic genes in a few functional categories; however, these results were not reliable due to the very small number of proteins. Otherwise, VUSs were very similar to categories that contain VUSs and other variant classes.

To further investigate the types of proteins and genes, we performed GO term (47) enrichment analysis. GO annotations were separately investigated for biological process, molecular function, and cellular component. All the variant classes showed large numbers of enriched terms (see Supplementary Tables S3–S5). The most statistically important GO terms were shared by genes that contained benign, pathogenic or VUS variants. Among the terms with the highest biological process enrichment were sensory perception, development, morphogenesis, and muscle-related concepts, among others. At the molecular function level, there were terms, e.g. for extracellular matrix, binding to various compounds, and several enzymatic activities, transporter and channel activities. Enriched cellular components included various membranous structures, the mitochondrial matrix, the sarcolemma and the transporter complex.

Next, we examined the protein-protein interactions in the variant classes. The interactions for each protein were obtained from the STRING database (41). We included only high-quality experimentally validated interactions. Three measures for the interaction networks were calculated. The degree indicates the number of interactions a protein has. Highly connected proteins have high degree. Closeness centrality and betweenness are two measures of the centrality of a node in a network.

The results were practically identical for all the variant classes (Figure 1E and Supplementary Table S6), indicating that the proteins containing the different types of variants had very similar overall network characteristics. The pairwise Tukey's post hoc test demonstrated statistically significant differences in the degrees of the variant classes. However, the distributions were very similar. The results for closeness and betweenness metrics did not exhibit significant differences across the gene categories (Supplementary Table S7). Since only approximately one-fourth of the proteins were shared by benign, pathogenic and VUS variant groups (Figure 1B), the results cannot be explained based only on these proteins; the shared properties were observed in the entire dataset.

Figure 1F and Supplementary Table S8 show the analysis of another central protein characteristic, polypeptide chain length. The mean length of the MANE transcript-based protein chains was 583 and the median length was 431, the range was from 12 to 35 991. The pairwise Tukey's test showed universally significant differences, except between VUS and benign variant-containing proteins (Supplementary Table S9). Since this difference was so small, it likely has no biological significance.

Structure level analysis

When proteins fold to their characteristic three dimensional structures, some amino acids are located to the protein core where they cannot interact with other molecules, while other residues are on the surface and accessible. The degree of variant position accessibility was determined from the 3D structures with the program FreeSASA (37), which uses a spherical probe to map the protein surface using the algorithm of Lee and Richards (48).

Figure 1G and Supplementary Table S10 show the results of the amino acid accessibility analysis. Tukey's pairwise test revealed significant differences for VUSs, pathogenic, and benign variants (Supplementary Table S11). All the classes contained variants of all states of solvent exposure but had different distributions. One explanation for the large portions of fully or almost fully accessible residues is their location within IDRs and/or low confidence regions. AlphaFold predicts these regions as elongated strands without any connections that would reduce accessibility. Thus, such residues display very high accessibility.

We analyzed the conservation of the variant positions based on position specific scoring matrices generated for each protein, via Blast searches (Altschul *et al.* 1997) against animal sequences. Sequence conservation is typically the most important feature used in computational variation interpretation, see e.g. (König *et al.* 2016; Niroula *et al.* 2015), and many predictors are based solely on this information. The results in Figure 1H and Supplementary Table S12 indicated that the positions were quite conserved and that there were no major differences except for the pathogenic and VUS versus P/LP groups, which were close to each other but were clearly different from the others because of their somewhat higher conservation. The pairwise Tukey's test showed universally significant differences among VUSs, pathogenic and benign variants (Supplementary Table S13).

Figure 1I and Supplementary Table S14 show the distributions of the variation positions in seven categories of secondary structural elements determined with STRIDE (36) based on backbone torsion angles. The structural elements were α-helices, β-sheets, π-helices, 3_{10} helices, isolated β-bridges in two categories and turns/coils. The eighth group included low confidence regions, which may be intrinsically disordered, for example. The structures were obtained from the AlphaFold Protein Structure Database (33) and contained experimentally determined structures, when available; otherwise AlphaFold2 or AlphaFold 3 predictions were made (32,34).

The positions of the variants within the secondary structural elements displayed some differences (Figure 1I). Variants within α-helices were the most pronounced in the pathogenic and VUS versus P/LP groups, while low confidence structures were the most common for the benign and VUS versus B/LB groups, and the least common for the pathogenic and VUS versus P/LP groups. The benign variants contained the largest proportion of low condifence positions along with the VUS versus B/LB class.

Low confidence regions do not have reliable predictions, e.g. due to being located within IDRs. IDRs are structurally and functionally special regions that are known to contain pathogenic variants. IDRs can adopt various structures and bind to several partners. They are involved in many important functions (49).

Verified IDRs were obtained from DisProt (39). IDRs were significantly depleted in all categories, except for pathogenic and VUS versus P/LP variants, which were somewhat enriched (Figure 1J). A similar analysis was performed for transmembrane regions based on data from Human Transmembrane Proteome (38). All the variant classes were highly biased toward variants in membrane proteins (Figure 1J and Supplementary Table S15). A substantial number of proteins are attached to membranes and function especially as receptors or transporters.

Some proteins are vulnerable for sequence repeat variants, especially single amino acid repeats. We identified the locations of all types of repeats at variant positions. Analysis of all the human proteins with the T-REKS program revealed 25 736 repeats in 6344 proteins. Figure 1H and Supplementary Table S15 indicate that variants were depleted in repeats in the variant classes, except for benign and pathogenic variants.

Variation level analysis

We investigated the distributions of variations in the six variant classes at nucleotides, nucleotide groups, codons, encoded amino acids, and amino acid classes. Comparisons of the distributions in Figure 2A and Supplementary Table S16 showed statistically significant differences, but the distributions were largely similar. VUS versus P/LP cases were the closest to pathogenic variants and VUS versus B/LB to benign variant group, as expected. V versus P/LP was more different in comparison to benign than pathogenic variants, similarly, V versus B/LB was more different from pathogenic than benign variants. In the VUS versus pathogenic comparison, variations from A and C were enriched, and those from G and T were depleted. When comparing VUSs to benign, only two types of variants were depleted, and the others were either enriched or not statistically significant.

A similar analysis was performed for groupings of nucleotides into three sets of two-base categories. DNA nucleotides are either purines (A, G) or pyrimidines (C, T). Purines are two-carbon nitrogen ring bases, while pyrimidines have one-carbon nucleotides. Weak nucleotides (A, T) form two hydrogen bonds when base pairing, while strong nucleotides (C, G) form three hydrogen bonds. The third classification refers to keto (G, T) and imino nucleotide (A, C) according to the major tautomeric forms.

Figure 2B and Supplementary Table S17 show comparisons of two-base groups in the variant classes. Again, many of the comparisons showed statistically significant results, however the log OR values showed a larger range than in Figure 2A. Most of the differences were relatively small. In summary, amino/keto and strong/weak base analyses revealed significant differences in all the fields in both the VUS versus pathogenic and VUS versus benign comparisons. Despite statistically significant observations, base or two-base compositions would not work as features separating the different

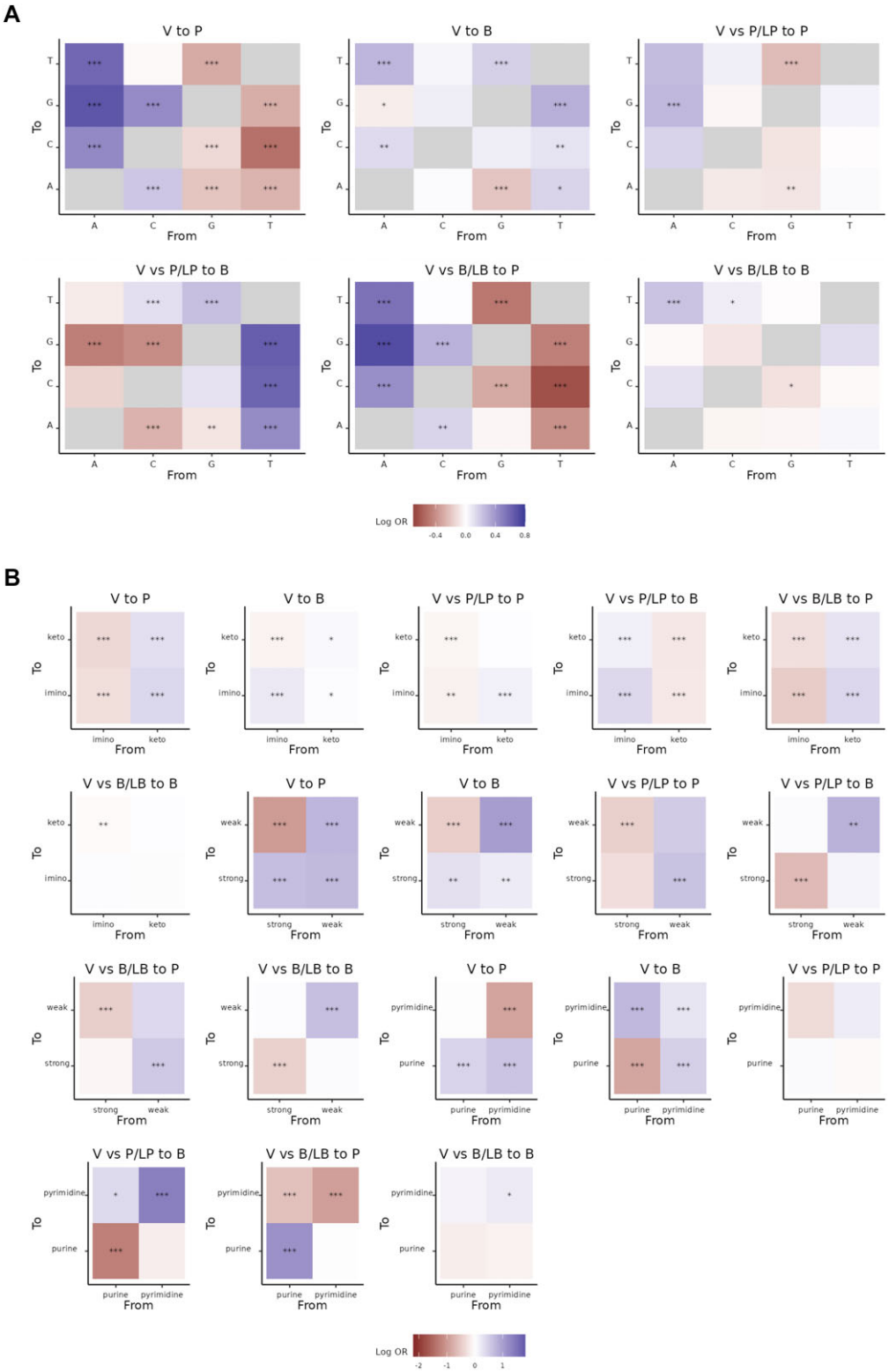

Figure 2. (**A**) Nucleotide level comparison of variant positions. Pairwise comparison of nucleotide substitutions in the five variant classes. (**B**) Pairwise comparisons for two-base groups in the variant classes. *P* value significance: $*P \leq 0.05$; $**P \leq 0.01$; *** $P \leq 0.001$. Gray indicates cases where there were not enough variants for statistical analysis or substitution by the same type of base.

variant classes, as they would be too simplistic since there are only 4 of 16 categories.

Similar analyses were performed for amino acids. Figure 3A and Supplementary Table S18 show comparisons for amino acids and Figure 3B and Supplementary Table S19 show comparisons for amino acid groups. The six amino acid groups combine physicochemically related residues. Many of the substitutions differed significantly. The biological effect of these differences is likely very small or negligible. Amino acid usage in VUSs and benign variants were close to each other, and pathogenic variants had more different distributions. Interestingly, almost all amino acid substitutions that were enriched in VUS compared to benign variants were also significant in comparison to pathogenic variants. The sign of the log OR might, however, be different. The conflicting classes were closer to pathogenic or benign variants, similar to previous analyses.

Some amino acids were enriched in variations. Arginine was the most frequent amino acid among pathogenic variants despite its overall low frequency. Arginine variants were also enriched in VUS versus benign and VUS versus P/LP comparisons to benign and in VUS versus B/LB comparisons to pathogenic. Arginine is encoded by a total of six codons, four of which contain CpG dinucleotide, which is the most variable dinucleotide. This observation was also apparent from Figure 4 for codons. Despite the enrichment or depletion of certain amino acids in the six comparisons, amino acid substitution types cannot reliably distinguish between variant categories. Note that only 150 of the 380 amino acid substitutions are possible by single nucleotide alterations; therefore, many cells in Figure 3 are gray.

The results of the analysis of amino acids in the six groups are shown in Figure 3B. The outcome was a simplification of that in Figure 3A. Group G4 did not show significant differences compared to itself. This is because the two structural amino acids G and P in G4 cannot be replaced by each other with only one substitution at gene level. Since our analysis was for single nucleotide variants, such rare multiple variants were not included. Comparison of VUS to pathogenic, VUS versus P/LP, and VUS versus B/LB revealed the greatest numbers of significant alterations. In comparison to pathogenic variants from G1 and G4 were depleted, whereas benign depletions were common in variants changing to G5 or G6 amino acids. Other variants in these comparisons were mainly enriched or not significantly different.

Are the differences in nucleotide and amino acid usage biologically relevant? It is not straightforward to answer. There were some statistically significant differences, but each substitution type appeared in every dataset. It has been known for a long time that certain amino acids frequently exhibit disease-related variations, for example, changes from arginine. No variants of any amino acid type are always disease related, the effect is largely context dependent regarding both sequence and structure.

Next, we investigated the distribution of codons. The data in Figure 3 for the amino acid distribution analysis already indicated that certain amino acids were enriched or depleted in certain variation classes. As we concentrated on amino acid substitutions, there were no data for the stop codons, which are indicated in gray.

At the codon level, VUSs were close to benign cases. VUS vs pathogenic, VUS vs B/LB comparison to pathogenic, and VUS vs P/LP versus benign cases had the greatest numbers of significant differences. The two first mentioned comparisons were almost identical and different from the third analysis. Interestingly, all the codons for those amino acids that are coded by several codons behaved similarly within comparisons. The enrichment of codon depletion followed the distibution of the corresponding amino acids, see Figure 4 and Supplementary Table S20.

CpG islands are important regulatory elements that are often located in front of genes, although they can also appear in coding regions. CpG dinucleotides within these C + G rich regions are often linked to gene expression (50). Methylations in islands are often associated with gene silencing, including genomic imprinting, which causes monoallelic gene expression. Analysis of coding variant locations within CpG islands indicated only minor differences between the variant classes. The percentages of original amino acids within CpG islands ranged from 9.2% to 10.8% (Supplementary Table S21), and there were no major differences between the groups.

Discussion

The idea that all VUSs could be reclassified either as benign or pathogenic is simplistic and presents a mechanistic and technological viewpoint that ignores natural biological heterogeneity. Extensive analysis of substitutions at genes/proteins, protein structures, nucleotides, and amino acid sequences in which the variants appear indicated that there were no major differences between the classes. The only considerable differences were in the accessibility and conservation of the original amino acid, distribution of the original positions to some secondary structural elements, and differences in codon and amino acid substitution frequencies. These differences were mainly for pathogenic variants in comparison to other classes, not for VUSs. Variations were less common in repeats in VUSs and conflicting categories than in benign or pathogenic variants. Some differences appeared in the distribution to secondary structural elements. As all variant classes displayed distributions throughout the full range, these characteristics are not sufficient to distinguish VUSs from other variants. They could possibly be used as features, e.g. in variant pathogenicity/tolerance predictors. Evolutionary conservation is utilized practically for all pathogenicity predictors, some of which are based solely on it. Pathogenic and VUS versus P/LP classes had somewhat different distributions for conservation in comparison to the other classes. Accessibility and structural characteristics are less common predictive features, as most of the predictors are sequence-based.

Our analysis concentrated on single nucleotide variants leading to amino acid substitutions. This is by far the largest group of known variations. It is highly likely that the other types of variants behave similarly.

None of the investigated features clearly distinguished VUSs from other types of variants or the other categories from others. When this fact is combined with pervasive heterogeneity, it is evident that all VUSs can never be distinguished from variants with other phenotypes. The generation of additional information will reduce the number of VUSs to some extent in the future. Those variants that display wide phenotypic differences so that some individuals with a variant have a disease and others do not (or have other forms of disease) will always be present. Thus, there will always be VUSs.

Functional studies have been presented as a solution for classifying VUSs. This is a way forward; however, there are issues to consider. Multiplexed assays of variant effects

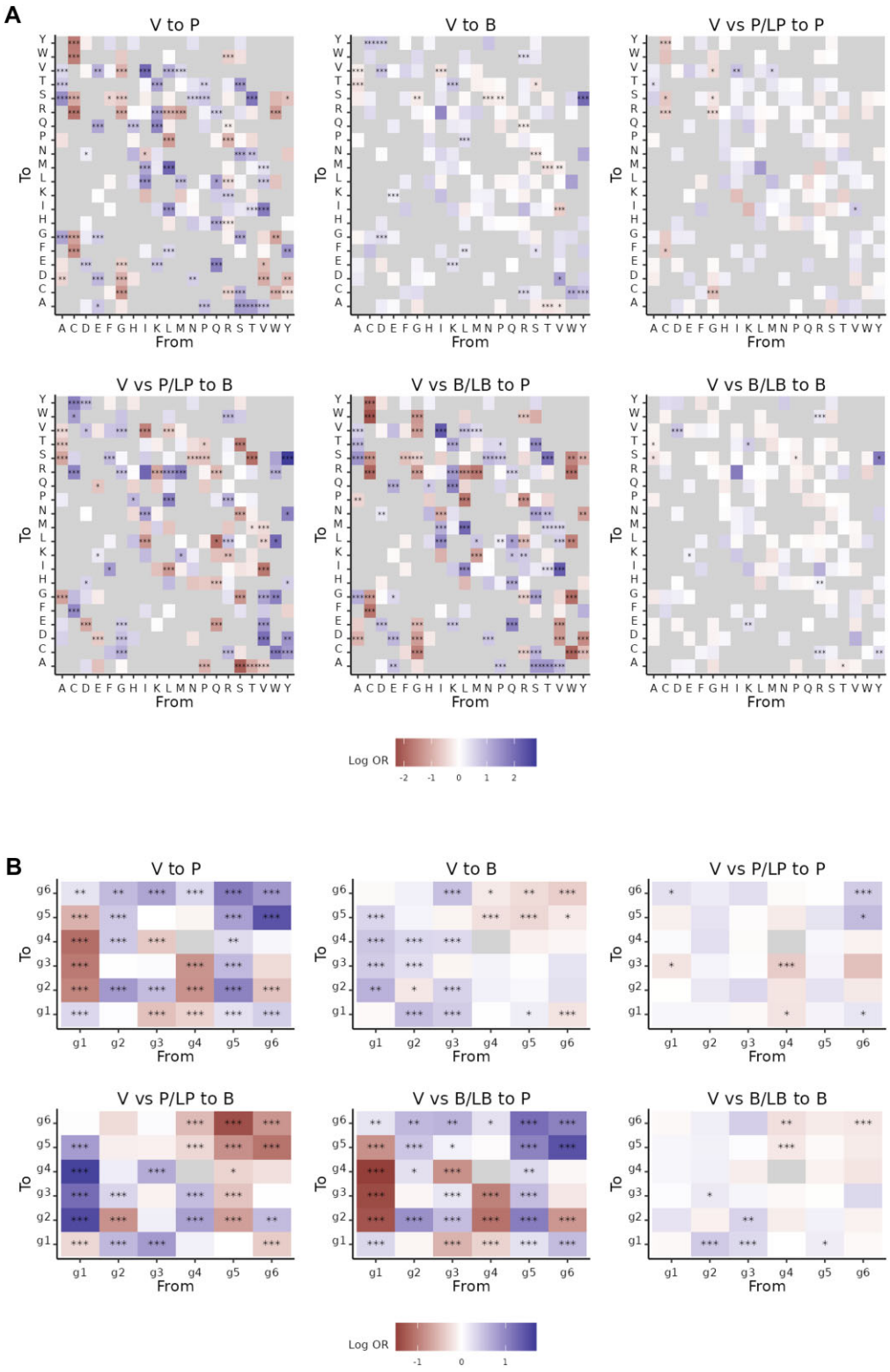

Figure 3. (**A**) Pairwise comparison of amino acid substitutions. (**B**) Pairwise comparison of amino acid groups in the six variant classes. *P* value significance: $*P \leq 0.05$; $**P \leq 0.01$; *** $P \leq 0.001$. Gray indicates alterations that are not possible by single nucleotide substitution at the nucleotide level or when there were not enough cases for statistical analysis.

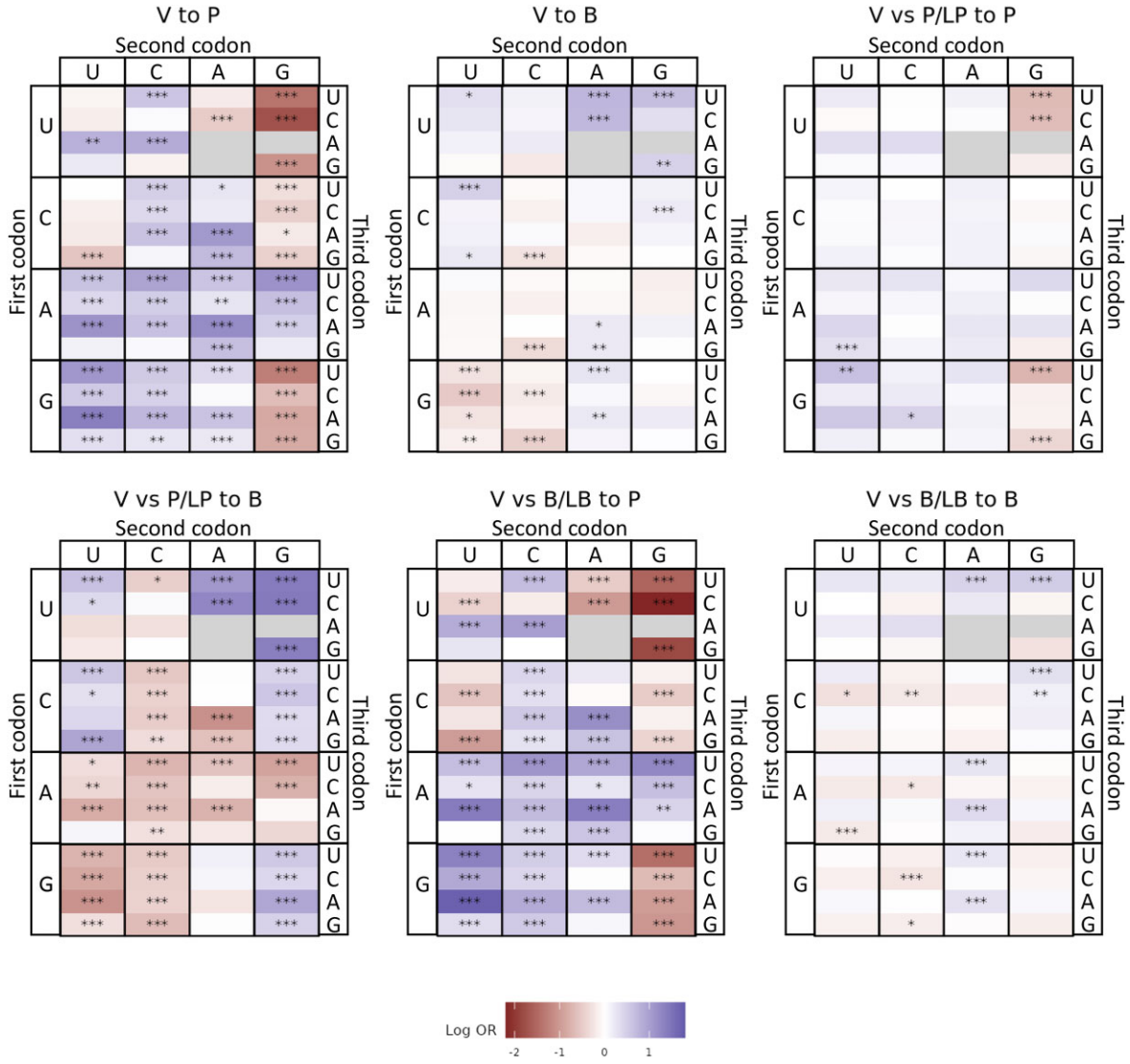

Figure 4. Pairwise comparison of codon usage. *P* value significance: $*P \leq 0.05$; $**P \leq 0.01$; $*** P \leq 0.001$. Gray indicates stop codons that were not investigated in this study.

(MAVEs) have recently been implemented for a few proteins to investigate various properties (51). The results of many of these studies are available in MaveDB (52). In these studies, all or almost all substitutions and possibly other variants were investigated within individual proteins at the functional level. The functional parameters measured depend on the protein. MAVE studies are limited to proteins for which there is an assay that allows large-scale study. The measured parameter may not always be the most relevant for the biological function; instead, the method is chosen based on availability.

MAVE datasets contain many different types of experiments that measure different properties. One may wonder whether effects on fitness, growth rate, enrichment, abundance scores, functional complementation studies, binding free energy changes, etc., are comparable and provide suitable proxies for damaging/functional effects. Do these scores measure functional, biological effect and if so, then how is that related to the actual variation? Many of the measured scores are secondary and do not describe the primary effect of the variants. For example, protein stability-reducing variants affect abundance, which indirectly affects activity.

The functional effect does not equal to the biological effect. One would likely call a reduction in activity of 90% important and disease related. However, this may not be the case. In several enzymopathies, normal activity must be reduced by >90% to achieve a disease phenotype (53). It is thus essential to understand the mechanism of each gene/protein/disease and to make protein-based adjustments to functional parameters. How is it possible to lose almost all activity without having a biological or medical effect? Due to saturation kinetics, even a substantial reduction in enzyme activity does not have a major effect on the flux of the pathway (53). Many biological systems are robust against variations.

We estimate, based on our extensive experience with variation interpretation, e.g. in benchmarking (54–56) and the development of predictors (57–60), that the ratio of VUSs will likely remain between 20 and 30% depending on the gene/protein in the future. VUSs must be accepted and considered as natural variation, not as a burden or something to eliminate. Those working on the variation interpretation must admit that and keep it in mind. In the end, variation data, as any clinical evidence, should be used only when clearly war-

ranted. Functional and other studies are needed to reclassify VUSs; however, efforts should consider individual phenotypic heterogeneity. It is essential to understand the reasons for and bases of the different forms of heterogeneity, which should be prioritized in future studies.

Data availability

The datasets generated during this study are available at VariBench https://structure-next.med.lu.se/VariBench/data/variationtype/substitutions/train/Dataset33/VUS.csv.

Supplementary data

Supplementary Data are available at NARGAB Online.

Funding

European Commission [JTC2022]; Vetenskapsrådet [2019-01403]; Cancerfonden [CAN 20 1350].

Conflict of interest statement

None declared.

References

1. Richards,S., Aziz,N., Bale,S., Bick,D., Das,S., Gastier-Foster,J., Grody,W.W., Hegde,M., Lyon,E., Spector,E., *et al.* (2015) Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.*, **17**, 405–423.
2. Ellard,S., Baple,E.L., Callaway,A., Berry,I., Forrester,N., Turnbul,C., Owens,M., Eccles,D.M., Abbs,S., Scott,R., *et al.* (2020)
3. Fowler,D.M. and Rehm,H.L. (2024) Will variants of uncertain significance still exist in 2030?*Am. J. Hum. Genet.*, **111**, 5–10.
4. Gunter,C. and Green,E.D. (2023) To boldly go: unpacking the NHGRI's bold predictions for human genomics by 2030. *Am. J. Hum. Genet.*, **110**, 1829–1831.
5. Vihinen,M. (2022) Individual genetic heterogeneity. *Genes (Basel)*, **13**, 1626.
6. Waddell-Smith,K.E., Skinner,J.R. and Bos,J.M. (2020) Pre-test probability and genes and variants of uncertain significance in familial long QT syndrome. *Heart Lung Circ*, **29**, 512–519.
7. Lin,Y., Williams,N., Wang,D., Coetzee,W., Zhou,B., Eng,L.S., Um,S.Y., Bao,R., Devinsky,O., McDonald,T.V., *et al.* (2017) Applying high-resolution variant classification to cardiac arrhythmogenic gene testing in a demographically diverse cohort of sudden unexplained deaths. *Circ. Cardiovasc. Genet.*, **10**, e001839.
8. Hoskinson,D.C., Dubuc,A.M. and Mason-Suares,H. (2017) The current state of clinical interpretation of sequence variants. *Curr. Opin. Genet. Dev.*, **42**, 33–39.
9. Walsh,N., Cooper,A., Dockery,A. and O'Byrne,J.J. (2024) Variant reclassification and clinical implications. *J. Med. Genet.*, **61**, 207–211.
10. Thompson,B.A., Spurdle,A.B., Plazzer,J.P., Greenblatt,M.S., Akagi,K., Al-Mulla,F., Bapat,B., Bernstein,I., Capella,G., den Dunnen,J.T., *et al.* (2014) Application of a 5-tiered scheme for standardized classification of 2,360 unique mismatch repair gene variants in the InSiGHT locus-specific database. *Nat. Genet.*, **46**, 107–115.
11. Fokkema,I., van der Velde,K.J., Slofstra,M.K., Ruivenkamp,C.A.L., Vogel,M.J., Pfundt,R., Blok,M.J., Lekanne Deprez,R.H., Waisfisz,Q., Abbott,K.M., *et al.* (2019) Dutch genome diagnostic laboratories accelerated and improved variant interpretation and increased accuracy by sharing data. *Hum. Mutat.*, **40**, 2230–2238.
12. Anderson,C.L., Munawar,S., Reilly,L., Kamp,T.J., January,C.T., Delisle,B.P. and Eckhardt,L.L. (2022) How functional genomics can keep pace with VUS identification. *Front. Cardiovasc. Med.*, **9**, 900431.
13. Burstein,D.S., Gaynor,J.W., Griffis,H., Ritter,A., Connor,M.J.O., Rossano,J.W., Lin,K.Y. and Ahrens-Nicklas,R.C. (2021) Genetic variant burden and adverse outcomes in pediatric cardiomyopathy. *Pediatr. Res.*, **89**, 1470–1476.
14. Rehm,H.L., Alaimo,J.T., Aradhya,S., Bayrak-Toydemir,P., Best,H., Brandon,R., Buchan,J.G., Chao,E.C., Chen,E., Clifford,J., *et al.* (2023) The landscape of reported VUS in multi-gene panel and genomic testing: time for a change. *Genet. Med.*, **25**, 100947.
15. Frone,M.N., Stewart,D.R., Savage,S.A. and Khincha,P.P. (2021) Quantification of discordant variant interpretations in a large family-based study of Li-Fraumeni syndrome. *JCO Precis. Oncol.*, **5**, 1727–1737.
16. Amendola,L.M., Muenzen,K., Biesecker,L.G., Bowling,K.M., Cooper,G.M., Dorschner,M.O., Driscoll,C., Foreman,A.K.M., Golden-Grant,K., Greally,J.M., *et al.* (2020) Variant classification concordance using the ACMG-AMP variant interpretation guidelines across nine genomic implementation research studies. *Am. J. Hum. Genet.*, **107**, 932–941.
17. Vihinen,M. (2015) Muddled genetic terms miss and mess the message. *Trends Genet.*, **31**, 423–425.
18. Vihinen,M. (2023) Systematic errors in annotations of truncations, loss-of-function and synonymous variants. *Front. Genet.*, **10**, 1015017.
19. Vihinen,M. (2020) Poikilosis – pervasive biological variation. *F1000Research*, **9**, 602.
20. Landrum,M.J., Lee,J.M., Riley,G.R., Jang,W., Rubinstein,W.S., Church,D.M. and Maglott,D.R. (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.*, **42**, D980–D985.
21. Fokkema,I., Kroon,M., López Hernández,J.A., Asscheman,D., Lugtenburg,I., Hoogenboom,J. and den Dunnen,J.T. (2021) The LOVD3 platform: efficient genome-wide sharing of genetic variants. *Eur. J. Hum. Genet.*, **29**, 1796–1803.
22. Morales,J., Pujar,S., Loveland,J.E., Astashyn,A., Bennett,R., Berry,A., Cox,E., Davidson,C., Ermolaeva,O., Farrell,C.M., *et al.* (2022) A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature*, **604**, 310–315.
23. McLaren,W., Gil,L., Hunt,S.E., Riat,H.S., Ritchie,G.R., Thormann,A., Flicek,P. and Cunningham,F. (2016) The Ensembl variant effect predictor. *Genome Biol.*, **17**, 122.
24. Nair,P.S. and Vihinen,M. (2013) VariBench: a benchmark database for variations. *Hum. Mutat.*, **34**, 42–49.
25. Shen,B. and Vihinen,M. (2004) Conservation and covariance in PH domain sequences: physicochemical profile and information theoretical analysis of XLA-causing mutations in the Btk PH domain. *Protein Eng. Des. Sel.*, **17**, 267–276.
26. Singh,A.K., Amar,I., Ramadasan,H., Kappagantula,K.S. and Chavali,S. (2023) Proteins with amino acid repeats constitute a rapidly evolvable and human-specific essentialome. *Cell Rep.*, **42**, 112811.
27. Narasimhan,V.M., Hunt,K.A., Mason,D., Baker,C.L., Karczewski,K.J., Barnes,M.R., Barnett,A.H., Bates,C., Bellary,S., Bockett,N.A., *et al.* (2016) Health and population effects of rare gene knockouts in adult humans with related parents. *Science*, **352**, 474–477.
28. Saleheen,D., Natarajan,P., Armean,I.M., Zhao,W., Rasheed,A., Khetarpal,S.A., Won,H.H., Karczewski,K.J., O'Donnell-Luria,A.H., Samocha,K.E., *et al.* (2017) Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. *Nature*, **544**, 235–239.
29. Sulem,P., Helgason,H., Oddson,A., Stefansson,H., Gudjonsson,S.A., Zink,F., Hjartarson,E., Sigurdsson,G.T., Jonasdottir,A., Jonasdottir,A., *et al.* (2015) Identification of a large set of rare complete human knockouts. *Nat. Genet.*, **47**, 448–452.

30. Blake,J.A., Baldarelli,R., Kadin,J.A., Richardson,J.E., Smith,C.L. and Bult,C.J. (2021) Mouse Genome Database (MGD): knowledgebase for mouse-human comparative biology. *Nucleic Acids Res.*, **49**, D981–D987.
31. Solomon,B.D., Nguyen,A.D., Bear,K.A. and Wolfsberg,T.G. (2013) Clinical genomic database. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 9851–9855.
32. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Žídek,A., Potapenko,A., *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.
33. Varadi,M., Anyango,S., Deshpande,M., Nair,S., Natassia,C., Yordanova,G., Yuan,D., Stroe,O., Wood,G., Laydon,A., *et al.* (2022) AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.*, **50**, D439–D444.
34. Abramson,J., Adler,J., Dunger,J., Evans,R., Green,T., Pritzel,A., Ronneberger,O., Willmore,L., Ballard,A.J., Bambrick,J., *et al.* (2024) Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, **630**, 493–500.
35. Varabyou,A., Sommer,M.J., Erdogdu,B., Shinder,I., Minkin,I., Chao,K.H., Park,S., Heinz,J., Pockrandt,C., Shumate,A., *et al.* (2023) CHESS 3: an improved, comprehensive catalog of human genes and transcripts based on large-scale expression data, phylogenetic analysis, and protein structure. *Genome Biol.*, **24**, 249.
36. Heinig,M. and Frishman,D. (2004) STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Res.*, **32**, W500–W502.
37. Mitternacht,S. (2016) FreeSASA: an open source C library for solvent accessible surface area calculations. *F1000Res*, **5**, 189.
38. Dobson,L., Remenyi,I. and Tusnady,G.E. (2015) The human transmembrane proteome. *Biol. Direct*, **10**, 31.
39. Aspromonte,M.C., Nugnes,M.V., Quaglia,F., Bouharoua,A., Tosatto,S.C.E. and Piovesan,D. (2024) DisProt in 2024: improving function annotation of intrinsically disordered proteins. *Nucleic Acids Res.*, **52**, D434–D441.
40. Jorda,J. and Kajava,A.V. (2009) T-REKS: identification of Tandem REpeats in sequences with a K-meanS based algorithm. *Bioinformatics*, **25**, 2632–2638.
41. Szklarczyk,D., Gable,A.L., Nastou,K.C., Lyon,D., Kirsch,R., Pyysalo,S., Doncheva,N.T., Legeay,M., Fang,T., Bork,P., *et al.* (2021) The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.*, **49**, D605–D612.
42. Miller,D.T., Lee,K., Abul-Husn,N.S., Amendola,L.M., Brothers,K., Chung,W.K., Gollob,M.H., Gordon,A.S., Harrison,S.M., Hershberger,R.E., *et al.* (2023) ACMG SF v3.2 list for reporting of secondary findings in clinical exome and genome sequencing: a policy statement of the American College of Medical Genetics and Genomics (ACMG). *Genet. Med.*, **25**, 100866.
43. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
44. Fu,L., Niu,B., Zhu,Z., Wu,S. and Li,W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.
45. Wu,T., Hu,E., Xu,S., Chen,M., Guo,P., Dai,Z., Feng,T., Zhou,L., Tang,W., Zhan,L., *et al.* (2021) clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation (Camb)*, **2**, 100141.
46. Schaafsma,G.C.P. and Vihinen,M. (2017) Large differences in proportions of harmful and benign amino acid substitutions between proteins and diseases. *Hum. Mutat.*, **38**, 839–848.
47. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T., *et al.* (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
48. Lee,B. and Richards,F.M. (1971) The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.*, **55**, 379–400.
49. Uversky,V.N., Davé,V., Iakoucheva,L.M., Malaney,P., Metallo,S.J., Pathak,R.R. and Joerger,A.C. (2014) Pathological unfoldomics of uncontrolled chaos: intrinsically disordered proteins and human diseases. *Chem. Rev.*, **114**, 6844–6879.
50. Illingworth,R.S. and Bird,A.P. (2009) CpG islands–'a rough guide.'. *FEBS Lett.*, **583**, 1713–1720.
51. Weile,J. and Roth,F.P. (2018) Multiplexed assays of variant effects contribute to a growing genotype-phenotype atlas. *Hum. Genet.*, **137**, 665–678.
52. Esposito,D., Weile,J., Shendure,J., Starita,L.M., Papenfuss,A.T., Roth,F.P., Fowler,D.M. and Rubin,A.F. (2019) MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biol.*, **20**, 223.
53. Vihinen,M. (2021) Functional effects of protein variants. *Biochimie*, **180**, 104–120.
54. Thusberg,J., Olatubosun,A. and Vihinen,M. (2011) Performance of mutation pathogenicity prediction methods on missense variants. *Hum. Mutat.*, **32**, 358–368.
55. Khan,S. and Vihinen,M. (2010) Performance of protein stability predictors. *Hum. Mutat.*, **31**, 675–684.
56. Niroula,A. and Vihinen,M. (2019) How good are pathogenicity predictors in detecting benign variants? *PLoS Comput. Biol.*, **15**, e1006481.
57. Niroula,A. and Vihinen,M. (2017) Predicting severity of disease-causing variants. *Hum. Mutat.*, **38**, 357–364.
58. Niroula,A., Urolagin,S. and Vihinen,M. (2015) PON-P2: prediction method for fast and reliable identification of harmful variants. *PLoS One*, **10**, e0117380.
59. Olatubosun,A., Väliaho,J., Härkönen,J., Thusberg,J. and Vihinen,M. (2012) PON-P: integrated predictor for pathogenicity of missense variants. *Hum. Mutat.*, **33**, 1166–1174.
60. Yang,Y., Shao,A. and Vihinen,M. (2022) PON-all, amino acid substitution tolerance predictor for all organisms. *Front Mol. Biosci.*, **9**, 867572.

Received: Auguat 5, 2024. **Revised:** October 2, 2024. **Editorial Decision:** October 28, 2024. **Accepted:** October 29, 2024