Conserved missense variant pathogenicity and correlated phenotypes across paralogous genes

Tobias Bruenger
Department of Neurology, The University of Texas Health Science Center at Houston, Houston, TX
Alina Ivanuk
Department of Neurology, Mayo Clinic Florida, Jacksonville, FL
Eduardo Pérez-Palma
Universidad del Desarrollo, Centro de Genética y Genómica, Facultad de Medicina Clínica Alemana. Santiago
Ludovica Montanucci
Department of Neurology, The University of Texas Health Science Center at Houston, Houston, TX
Stacey Cohen
Division of Neurology, Children's Hospital of Philadelphia, Philadelphia, PA
Lacey Smith
Epilepsy Genetics Program, Division of Epilepsy and Clinical Neurophysiology, Department of Neurology, Boston Children's Hospital, Boston, MA
Shridhar Parthasarathy
Division of Neurology, Children's Hospital of Philadelphia, Philadelphia, PA
Ingo Helbig
Division of Neurology, Children's Hospital of Philadelphia, Philadelphia, PA
Michael Nothnagel
Cologne Center for Genomics (CCG), University of Cologne, Cologne
Patrick May
Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette
Dennis Lal
Dennis.Lal@uth.tmc.edu

Department of Neurology, The University of Texas Health Science Center at Houston, Houston, TX

Research Article

Keywords: Variant classification, Paralogs, Genetics, Missense variants

Posted Date: November 29th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-5434140/v1

Additional Declarations: No competing interests reported.

Abstract

Background

The majority of missense variants in clinical genetic tests are classified as variants of uncertain significance. Prior research has shown that the deleterious effects and the subsequent molecular consequence of variants are often conserved among paralogous protein sequences within a gene family. Here, we systematically quantified on an exome-wide scale if the existence of pathogenic variants in paralogous genes at a conserved position could serve as evidence for the pathogenicity of a new variant. For the gene family of voltage-gated sodium channels where variants and expert-curated clinical phenotypes were available, we also assessed whether phenotype patterns of multiple disorders for each gene were also conserved across variant positions within the gene family.

Methods

We developed a framework that assesses the presence of pathogenic missense variants located in conserved residues across paralogous genes. We systematically mapped 2.5 million pathogenic and general population variants from the ClinVar, HGMD, and gnomAD databases onto a total of 9,990 genes and aligned them by gene families. We evaluated the quantity of classifiable amino acids by utilizing pathogenic variants identified in databases alone and then compared this assessment to the inclusion of paralogous pathogenic variants. We validated and quantified the evidence of conserved pathogenic paralogous variants in variant pathogenicity classification.

Results

Considering conserved pathogenic variants in paralogous genes, increased the number of classifiable variants 2.8-fold across the exome, compared to pathogenic variants in the gene of interest alone. The presence of a pathogenic variant in a paralogous gene is associated with a positive likelihood ratio of 8.32 for variant pathogenicity. The likelihood ratio was gene family-specific. Across ten genes encoding voltage-gated sodium channels and 22 expert-curated disorders, we identified cross-paralog correlated phenotypes based on 3D structure spatial position. For example, the established loss-of-function disorders *SCN1A*-associated Dravet syndrome, *SCN2A*-associated autism, *SCN5A*-associated Brugarda Syndrome, and *SCN8A*-associated neurodevelopmental disorder without seizures were correlated in their spatial variant position on structure. Finally, we show that phenotype integration in paralog variant selection improves variant classification.

Conclusion

Our results show that paralogous variants, in particular with phenotype information can enhance our understanding of variant effects.

Background

Large gene panels, exome, and genome sequencing have led to the identification of novel variants at an exponential rate(1). Up to 80% of pathogenic variants are located within protein-coding regions of the gene(2), with missense variants being particularly challenging to interpret due to the variety of different molecular mechanisms through which they can cause disease. Furthermore, several disease-associated genes are pleiotropic, further complicating variant interpretation(3–5). Despite these challenges, variant classification is necessary for diagnosing rare and genetically heterogeneous disorders, and for the development of personalized medicine.

About 80% of genes associated with monogenic disorders are paralogs(6). These paralogous genes can be grouped into 2871 gene families as defined by the Human Gene Nomenclature Consortium (HGNC)(7) with > 80% sequence similarity(8). Genes within a gene family arise from gene duplication events of common ancestral genes and can share > 90% amino acid sequence similarity at functionally essential protein domains(9). We and others have shown that quantifying conservation across these paralogous genes and homologous domains is an effective strategy to distinguish between pathogenic and benign variants(8, 10–12). Molecular studies further indicate that the biophysical function of domains is conserved within a gene family. As a result, a single amino acid substitution in the same position of a homologous domain often leads to similar molecular effects across members of the same gene family(11, 13, 14). This suggests that a comprehensive understanding of variants in one gene can provide, through a form of knowledge "transfer", insights into the pathogenicity and also into the biological disease mechanisms of unstudied variants in its paralogs when these variants are located at identical positions.

Within the same gene family, proteins show similar patterns of population variant-constrained and pathogenic variant clustering. In addition to identifying conservation patterns within gene families, previous research has highlighted the differential distribution of missense variants between the general population and pathogenic missense variants which was consistent across a subset of paralogous genes(11, 14). Furthermore, our previous findings indicate that this regional clustering is prevalent across paralogous genes and enables a systematic identification of regions enriched with pathogenic variants, termed Pathogenic Variant Enriched Regions (PERs)(15). Our study showed that novel missense variants located within PERs have a higher likelihood of being pathogenic compared to those in non-PER regions of the same gene(15). However, this method currently has limited sensitivity, since many newly discovered variants are located outside of PERs. Moreover, as PERs typically define a larger protein region, interpretations regarding disease mechanisms are constrained to a regional context, preventing insights at the individual amino acid level.

To standardize variant interpretation, the American College of Medical Genetics and Genomics (ACMG) published recommendations for evaluating the pathogenicity of variants(16). However, > 45% of single nucleotide variants reported in the ClinVar database(17) (accessed March 2023) are classified as variants of uncertain significance (VUS), due to the absence of sufficient evidence for or against variant pathogenicity. The guidelines include criteria that utilize information from previous variant classifications e.g., the presence of an established pathogenic variant with the same amino acid exchange (PS1) or a different amino acid exchange (PM5) at the same position in the same gene that can provide strong to moderate evidence for pathogenicity(16). However, since the vast majority of rare monogenic disorders are genetically heterogeneous and about half of the identified pathogenic variants have not yet been observed in other individuals(18, 19), the application of these evidence criteria is limited.

In the present study, we extend prior work on gene family conservation to provide access to a paralog-based annotation that could improve the assessment of variant pathogenicity. We postulate that variants previously classified in conserved residues of paralogous genes can provide evidence for the pathogenicity of novel variants located at corresponding amino acid positions in these genes. The use of pre-classified variants in paralogs as evidence of pathogenicity has been previously suggested for a select group of genes e.g., by the RASopathy ClinGen Expert Panel(20). However, the broad applicability of this approach across the entire protein-coding exome - particularly, the potential of single missense variants from paralogs as a feature to inform variant pathogenicity - remains unquantified and untested.

In this proof of concept study, our findings reveal that for 519 gene families (comprising 1,459 genes) with high sequence similarity, the presence of a pathogenic variant in one gene family member at an equivalent protein position is associated with a significant increase in the likelihood of pathogenicity for a novel variant at a conserved paralogous site in the target gene. Additionally, we illustrate in a case study that integrating expert-curated clinical data across sodium channels can refine variant selection, which not only enhances variant pathogenicity classification but also identifies disorders across paralogs that likely share similar disease mechanisms.

Methods

Annotation of missense variants from public repositories

Missense variants from patients

Missense variants associated with the disease were collected from the ClinVar database(17) (ClinVar, release October 2019) and the Human Gene Mutation Database(21) (HGMD®) Professional release 2019.2. Similarly, we gathered an updated version of the variants from ClinVar (released December 2022) and HGMD (Professional release 2023.1), processed them as described before, and extracted all previously unreported pathogenic variants not observed in the previous dataset to obtain an independent set of variants. The ClinVar missense variants were obtained in a tabular format from the FTP site (ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/) and only those variants exclusively classified as "Pathogenic" and/or "Likely Pathogenic" in their final consensus interpretation were considered to ensure high stringency. The HGMD dataset was filtered for "missense variants," "High Confidence" calls (hgmd_confidence = "HIGH" flag), and "Disease causing" state (hgmd_variantType = "DM" flag). All annotations refer to the human reference genome version GRCh37.p13/hg19, and variants belonging to non canonical transcripts as defined by Ensembl were excluded(22). Since ClinVar and HGMD are not mutually exclusive, we used the union of both resources and removed duplicate entries by comparing HGVS annotations. We further refer to the combined set of variants classified as likely-pathogenic, pathogenic, or "Disease-causing" as ‘pathogenic variants’.

Missense variants from the population

Missense variants present in the Genome Aggregation Database(23) (gnomAD, public release 2.0.2) were obtained in the Variant Call Format(24) (VCFs). We extracted the high-quality missense variants by filtering the VCF files to the "CSQ" field and “PASS” flag. The annotations were based on the human reference genome version GRCh37.p13/hg19. We extracted only entries annotated to the canonical gene transcripts, as defined by Ensembl(22). The aggregated population variants serve as control variants in our study and are further referred to as controls.

Similarly, we gathered an updated version of the variants from gnomAD (public release 2.1.1, processed them as described above, and extracted all novel variants not observed in the previous set of gnomAD variants to obtain an independent set of control variants.

Annotation of missense variants and associated phenotypes for the voltage-gated sodium channels

Brain-related phenotypes

We aggregated published patient missense variants in voltage-gated sodium channel genes (VGSC) genes from the literature. All patient variants for *SCN1A* were obtained from Brunklaus et al.,2022 and Brunklaus et al., 2022(25, 26). Variants for *SCN2A* were obtained from Wolff et al, 2017(27) and Crawford et al., 2021(28). Variants for *SCN3A* were obtained from Zaman et al., 2018(29). All *SCN8A* variants are taken from Johannesen et al, 2021(4). Affected individuals were recruited through a network of collaborating clinicians, as well as GeneMatcher(30), using a standardized phenotyping sheet to assess clinical characteristics cognition), EEG, neuroimaging, and retrospective data on antiepileptic treatment.

Non-brain phenotypes

SCN5A variants were obtained from the studies conducted by Milman et al., 2021(3), and Walsh et al., 2021(31). Data from SCN4A, SCN9A, SCN10A, and SCN11A variants were collected from various publications listed in Supplementary Table 2. Variants in the voltage-gated sodium channels (VGSC) encoding genes that were not missense-constrained were filtered for the maximum population frequency (MAF). We inferred the MAF thresholds by using the approach described by Whiffin et al., 2017(32), via the authors' app (https://www.cardiodb.org/allelefrequencyapp), based on the phenotype's estimated prevalence, mode of inheritance, and penetrance of the phenotype. We categorized SCN4A variants related to myotonia congenita and paramyotonia congenita and SCN9A variants related to primary erythromelalgia and paroxysmal episodic pain disorder into single categories (Relaxation Impairment Disorders and Paroxysmal Pain Disorders, respectively) based on their shared molecular pathology and pathophysiology after applying the MAF filter.

We mapped all variants to their Ensemble canonical transcript[24] (*SCN1A*: ENST00000303395, *SCN2A*: ENST00000283256, *SCN3A*: ENST00000283254, *SCN4A*: ENST00000435607, *SCN5A*: ENST00000423572, *SCN8A*: ENST00000283254, *SCN9A*: ENST00000409672, *SCN10A*: ENST00000449082, *SCN11A*: ENST00000302328). Only phenotypes associated with variants at more than five different protein positions were considered. The original and harmonized phenotype annotations for each phenotype are listed in Supplementary Table 2.

Gene family definition

We obtained the paralogous genes that belong to a gene family from Pérez-Palma et al. 2020(15), as originally described in Lal et al., 2020(8). Briefly, the human paralog definitions were taken from Ensembl BioMart (33) and filtered for those with an HGNC symbol(7). For each gene, the canonical transcript as defined by Ensembl was considered. To avoid aligning highly diverged sequences, families with less than 80% similarity on the full protein sequence were removed.

Definition of paralogous variants

For all the protein sequences within the same gene family, we performed a multiple sequence alignment using the MUSCLE(34) software. We then mapped pathogenic and general population variants onto these multiple sequence alignments. Given two variants on two different genes of the same gene family, we considered them as paralogous variants if they satisfied the two following conditions: *1)* they are located at the same position in the multiple protein sequence alignment of the gene family, and *2)* the reference amino acid in the target gene and the paralogous gene is the same (Supplementary Fig. 1).

We further establish an expanded set of criteria, termed para-PS1 and para-PM5, which is defined as follows:

para-PS1

This refers to a pathogenic paralogous variant that exhibits the same amino acid substitution as the investigated variant.

para-PM5

This denotes a pathogenic paralogous variant that exhibits a different amino acid substitution compared to the investigated variant.

Calculation of the positive likelihood ratio when a pathogenic paralogous variant is found

For each gene, we calculated the positive likelihood ratio using our aggregated set of pathogenic and general population variants for the para-PS1/PM5 criteria (Supplementary Fig. 1). While considering the definition of the criteria (see above) we counted for each gene *i)* the number of pathogenic variants for which at least one pathogenic paralogous variant was observed and *ii)* the number of pathogenic variants for which no pathogenic paralogous variant was observed. For the same gene we also counted *i)* the number of control variants for which at least one pathogenic paralogous variant was observed and *ii)* the number of control variants for which no pathogenic paralogous variant was observed. To determine the level of evidence each criterion can define we calculated the positive likelihood ratios for the two cases: A) Presence of a pathogenic paralogous variant with either the same amino acid substitution (para-PS1) and B) Presence of a pathogenic paralogous variant with a different amino acid substitution (para-PM5). The positive likelihood ratio was computed using the sensitivity and specificity of the test:

Equation 1:

$$Positive\ Likelihood\,ratio\ (LR+) = \frac{Sensitivity}{(1 - Specificity)} = \frac{\left(\frac{TP}{TP+FN}\right)}{1 - \left(\frac{TN}{TN+FP}\right)}$$

where LR + represents the positive likelihood ratio, TP (true positives) denotes the number of pathogenic variants, for which a pathogenic variant is observed at a conserved corresponding paralogous residue position, TN (true negatives) indicates the number of variants from the general population, for which no pathogenic variants is observed at a corresponding paralogous residue position, FP (false positives) represents the number of population variants, for which a pathogenic variant is observed at a conserved corresponding paralogous residue position, and FN (false negative) denotes the number of pathogenic variants observed, for which no pathogenic variant is observed at a corresponding paralogous residue position. We calculated the LR + both individually for each gene as well as combined across all genes. For the gene-wise metric, we counted the variants denoting TP, FP, TN, and FN for each gene separately. For the combined metric we assessed the numbers for TP, FP, TN, and FN across all disease-associated genes within a gene family together to end up with a single LR+. All analyses were performed using R v.4.2.1.

Comparison to established gene-family-based approaches

To compare our results to an established gene-family-based approach which identified pathogenic enriched regions (PERs) across paralogous genes on an exome-wide scale14, we gathered an independent set of variants (see Annotation of missense variants from public repositories) which was not previously used nor in the PER approach nor the enrichment analysis of this study, and we repeated the calculation outlined above.

To estimate LR + that are not mediated by paralog conservation we repeated the analysis described above for three paralog conservation sub-groups using the Parazscore(8). The groups we considered are alignment positions with gene family wise 1) maximum Parazscore, indicating full paralog conservation across the gene family at the alignment position 2) Parazscore > 0 & not maximum Parazscore, indicating high paralog conservation at this alignment position but not full conservation and 3) Parazscore < 0, indicating low levels of conservation between paralogous genes at the alignment position.

Identification of phenotype correlation based on 3D-variant positions

To identify phenotypes associated with variants located at corresponding positions across voltage-gated sodium channels (VGSCs), we evaluated the spatial distribution of sodium channel protein structures for variants associated with each phenotype. We tallied the number of patients reported for each variant in every phenotype. Since not all sodium channels had available protein structures, we mapped the patient variants and their corresponding number of patients on the Nav1.2 protein structure (PDB-ID: 6j8e) using the multiple protein sequence alignment. We only considered patient variants that could be mapped to the protein structure for downstream analysis. For every residue in the Nav1.2 protein structure, we counted the number of patients with a variant in the residue or its local 3D neighborhood using a 5-angstrom radius cutoff, as previously introduced in Iqbal et al., 2022(35). The number of patients with variants at a certain residue position was evaluated independently for each phenotype. To identify phenotypes associated with variants at similar 3D-variant positions we calculated the Pearson correlation between the 3D-variant distribution associated with each phenotype.

Integrating variant similarity between phenotypes for the assessment of paralogous variant-based pathogenicity

We next explored whether utilizing phenotype correlation could refine the selection of variants for our paralogous patient variant approach. To test our hypothesis we first extracted the variants of the most common phenotypes in each sodium channel with > 40 different variants (*SCN1A*: Dravet Syndrome, *SCN2A*: Early onset developmental epileptic encephalopathy (DEE), *SCN5A*: Brugada Syndrome, *SCN8A*: DEE). We divided these cohorts randomly into four subsets of patient variants, each containing 25% of the variants. We then combined three of the four subsets (representing 75% of variants for each phenotype) with our remaining patient cohort containing all variants associated with other phenotypes. Following the approach outlined in the previous section we then identified 3D-variant position-based phenotype correlations. Finally, using the independent test cohort (the fourth subset), we calculated the LR + of patient vs control variants a) using paralogous pathogenic variants associated with non-correlated phenotypes and b) using paralogous pathogenic variants with significant (Bonferroni adjusted P < 0.05) 3D-position-based phenotype correlation. We repeated this approach three times, such that each set of variants was used as part of the training set three times and once as the test set, and calculated the LR + by summing up the individual TP, FP, TN, and FN values of each iteration.

Results

Incorporating pathogenic paralogous variants triples classifiable amino acid residues

The guidelines of the ACMG suggest that for determining the pathogenicity of novel variants, two scenarios can be considered: (1) the presence of a variant in the same gene with an identical amino acid change, irrespective of the nucleotide alteration and (2) a novel amino acid substitution at a position where another substitution was previously been considered pathogenic, named PS1 and PM5 criteria respecitvely(16). In this study, our objective was to explore whether this principle could be extrapolated to encompass pathogenic variants in paralogous genes. We specifically assessed if the existence of pathogenic variants in paralogous genes at a conserved, corresponding position could serve as evidence for the pathogenicity of a new variant. For our study, we termed a 'paralogous variant' as a variant that meets two conditions: (1) it's positioned in a paralogous gene at the analogous residue index position, as delineated by multiple sequence alignment (refer to Methods for details), and (2) it shares the same reference amino acid as the target gene.

First, we assessed the number of amino acid residues not overlapping with pathogenic variants within the same gene at equivalent paralogous amino acid positions, but yet overlapping with pathogenic variants in paralogous genes. We aggregated a total of 60,486 pathogenic variants from ClinVar(17) and HGMD(21) and mapped them to 2,871 different gene family alignments, consisting of 9,990 genes (Fig. 1). Our paralog variant analysis integrates pathogenic variants from multiple genes in the same gene family (see Methods for details). We, therefore, restricted the dataset to gene families harboring pathogenic variants in at least two genes and identified 1,459 genes from 519 gene families. Within these genes, 41,223 pathogenic missense variants and 171,690 pathogenic paralogous variants were found that covered 32,137 and 91,259 amino acid residues respectively (Supplementary Table 1). Of these 91,259 residues that are covered by a paralogous pathogenic variant 92.6% (N = 84,553 residues) were not covered by a pathogenic variant in the same gene. Therefore, the integration of paralogous pathogenic variants would increase the number of amino acids in these gene families were the criteria can be applied by about 3.6-fold (N = 116,690 residues, Fig. 2A). The increase in the number of classifiable amino acids in each gene family is highly correlated with the number of disease-associated genes in a gene family (R = 0.97, P = < 1e-300, Supplementary Fig. 2).

Presence of single pathogenic paralogous variants can be used to assess variant pathogenicity

Next, we quantified the value of incorporating pathogenic variants at paralogous positions to assess the variant pathogenicity of novel variants. Therefore, in addition to the aforementioned pathogenic variants, we included 2,478,899 variants from the gnomAD database(23) which served as controls in our study. When a pathogenic paralogous variant with the same amino acid exchange was present at a corresponding alignment index position (termed para-PS1 criterium, for details on the approach, see Methods) we observed across 519 gene families an average LR + of 8.32 (8.02–8.62, 95% confidence interval (CI), Fig. 2B). Restricting the analysis to missense variant-constrained genes (Missense-z score > 3.09(1)), increased the LR + to 8.91 (8.03–9.88, 95% CI, Fig. 2C). Notably, even for paralogous variants with a different substitution at the same alignment index position (termed para-PM5 criterium), we observed an increased LR+ (All genes: LR + = 4.32, (4.24–4.48, 95% CI), Missense constraint genes: LR + = 6.48, (6.05–6.94, 95% C). Overall we observed a wide range of LR + across different genes (Fig. 2B, C).

The presence of pathogenic paralogous variants provides evidence for pathogenicity beyond evolutionary conservation

Variant mapping across paralogous residues requires residue conservation. Next, we investigated the added value of mapping beyond conservation. Previously, we developed a 'parazscore(8)' to measure the conservation across paralog genes, showing that amino acids conserved within a gene family are significantly enriched for pathogenic variants. Notably, a fundamental prerequisite for the incorporation of pathogenic paralogous variants into the variant is assessment is the conservation of amino acid residues between the target gene and its paralogous gene. Hence, whenever pathogenic paralogous variants criteria are incorporated, a certain degree of conservation within the genes of the same gene family becomes inevitable. This conservation likely explains a portion of the elevated LR + we observed. Notably, while many methods(36–38) employ evolutionary conservation as a predictor of variant pathogenicity, it is crucial to discern the added value our approach provides beyond solely relying on conservation-based evidence. To achieve this, we reconsidered our previous analysis, segmenting amino acids based on their paralog conservation and grouping amino acid residues with similar conservation across paralogs together (see Methods for details). Interestingly, within these subgroups, the highest LR + were observed for residues exhibiting the least paralog conservation for both the para-PS1 criterium (Parazscore < 0; $LR +_{para-PS1}$ = 10.49, 95% CI = 9.60-11.45, Fig. 3A) as well as the para-PM5 criterium (Parazscore < 0; $LR +_{para-PM5}$ = 5.21, 95% CI = 4.87–5.59, Fig. 3B). Yet, even within the subgroup demonstrating the least increase in LR+, where maximum conservation across all paralogous genes of the same gene family was noted, we still detected an increased LR + of 4.88 and 2.69, for para-PS1 and para-PM5 criteria respectively. This observation suggests that the existence of pathogenic paralogous variants provides additional information beyond the level of conservation between paralogous genes.

Integrating single pathogenic paralogous variants improves a previous family-based variant interpretation approach

We compared our approach, using paralogous pathogenic variants located at corresponding amino acids to a previously published method(15). In contrast to our new approach, the published approach identifies ‘pathogenic variant enriched regions’ (‘PERs’, on average 33 consecutive amino acids(15)) across a gene family that is consistently enriched for pathogenic variants while depleted for control variants. Due to the sliding window approach the identified regions that are enriched for pathogenic variants, PERs can span amino acid residues without an established pathogenic variant across paralogs, and the regional association is derived from adjacent variants. However, identifying PERs within a gene family alignment requires a large number of pathogenic variants, limiting its applicability. First, we compared the number of exome-wide classifiable variants using single paralogous pathogenic variants with the PER approach. We used an independent set of pathogenic and control variants that were not utilized in the PER generation or the application of the para-PS1/PM5 criteria (see Methods for details). We found that the approach based on single paralogous pathogenic variants captured 2.2 times more residues compared to PERs (Fig. 3C). In the second comparison, we compared the LR + for each approach and observed similar LR + for the PER approach and for the para-PS1 approach ($LR+_{PER} = 5.28$, $LR+_{para-PS1} = 5.63$, Fig. 3D).

Leveraging phenotype correlations across paralogs can enhance pathogenicity assessment

A single gene can be associated with different disorders. The number of disorders associated with variants in the same gene frequently correlates with the number of different molecular functional defects. Given that structure determines function, the molecular consequences of variants often relate to their specific position within the protein structure(39). Thus, pinpointing phenotype correlations based on analogous variant distributions might reveal paralogous variants with consistent molecular effects. In the context of voltage-gated sodium channels (VGSCs), past research has underscored not only the conservation of pathogenicity but also the consistent functional effects among paralogous variants(13). Building on this, we hypothesized that uncovering phenotype correlations across VGSCs could fine-tune the application of pathogenic paralogous variants for variant pathogenicity assessment. We hypothesize that within gene family phenotype correlations could identify correlated phenotypes based on substitution position, subsequently enhancing the likelihood of conserved pathogenicity for variants at equivalent positions. To test this hypothesis, we curated a comprehensive dataset featuring 1,346 affected individuals, associated with 22 diverse phenotypes and possessing 886 unique missense variants in VGSC-encoding genes (detailed in Supplementary Table 2). Performing alignment position-based mapping onto the same structure combined with spatial-based phenotype proximity correlation analysis (see Methods for details), we identified within gene family position correlated phenotypes (Fig. 4A). For example, *SCN1A*-associated Dravet syndrome variants exhibited 3D positional correlations with *SCN2A* variants associated with autism (R = 0.31, P = 2.1e-35), and Brugada syndrome variants in *SCN5A* (R = 0.29, P = 2.8e-40).

For genes associated with several related disorders, such as the VGSC, variant classification is challenging since phenotype specificity is not high. Therefore, not all pathogenic classified variants might be correctly classified. Next, we tested whether variants from spatially correlated phenotypes across different paralogous genes could increase variant pathogenicity classification accuracy. We selected the most frequently reported phenotypes for VGSC genes with at least 40 patients. The four genes *SCN1A*, *SCN2A*, *SCN5A*, and *SCN8A* fulfilled this criterion. We dissected the associated variants into four subsets and calculated the evidence for variant pathogenicity (see methods for details). We observed an increased positive likelihood ratio by a factor of 3–8 for paralogous variants associated with 3D-position correlated phenotypes, in contrast to those paralogous variants without a significant 3D-position correlation (Fig. 4B). For example, for SCN8A DEE cases pathogenic paralogous variants whose phenotype correlate with the DEE in SCN8A (LR + = 34.7, CI 16.3) showed an 8.6- fold higher strength to asses variant pathogenicity compared to pathogenic paralogous variants found in cases with non-correlating phenotypes (LR + = 4.0, CI 1.8–8.9).

Discussion

Many paralogs are highly conserved in sequence and have similar biophysical molecular functions. Current variant interpretation guidelines only consider previously classified pathogenic missense variants in the gene of interest as evidence for pathogenicity. Here, we developed and validated a bioinformatic framework to integrate pathogenic missense variants in paralogous genes at corresponding alignment index positions as evidence for the pathogenicity of novel variants. We demonstrated that integrating paralogous pathogenic variants located at a corresponding protein position can provide evidence for pathogenicity even if the amino acid exchange is not conserved. Compared to approaches, such as the PS1 and PM5 criteria of the ACMG guidelines(16) which consider pathogenic variants in the same gene at the same position as evidence, our approach can be applied to 3.6 fold more protein residues where novel variants of unknown pathogenicity could be observed.

Pathogenic missense variants in paralogous genes can serve as a proxy for pathogenicity. Within a protein sequence, pathogenic variants are unevenly distributed and tend to accumulate in certain regions that are critical for protein function(40). These pathogenic variant-enriched regions have proven valuable for variant classification through established guidelines for variant interpretation(16) and the use of *in-silico* prediction algorithms(41). Moreover, the observation that critical protein regions tend to be evolutionarily conserved between paralogous genes can be harnessed to enhance statistical robustness by incorporating pathogenic variants across these paralogous genes(15). Still, about 70%, of pathogenic variants are located outside the regions identified as essential. As a result, individual pathogenic variants in paralogous genes outside these regions were not considered for variant interpretation. In a study examining long QT syndrome, it was observed that individual pathogenic variants in paralogous genes are often located at paralogous positions as determined from multiple sequence alignments(11), suggesting that the presence of a pathogenic variant at a particular position may serve as a proxy of pathogenicity at that alignment position in other paralogs. Our data test this hypothesis across a wide range of gene families and suggests that individual pathogenic paralogous variants can indeed serve as proxies for pathogenicity on a broad scale, thereby augmenting the efficacy of established variants in variant interpretation frameworks.

Pathogenic variants in voltage-gated sodium channel (VGSC) genes are associated with a broad spectrum of clinical phenotypes, even within the same gene(4, 25, 27, 29). Prior research demonstrated a strong correlation between different molecular variant effects, such as the gain or loss of a protein function, and the clinical phenotype(42). We identified phenotypes across VGSC genes with different organ or cellular gene expressions that are caused by corresponding paralogous variants located at the same alignment index position. The location of a variant in the protein structure in VGSC, particularly in critical regions like the selectivity filter or the inactivation gate, is often associated with conserved molecular function(13). Our findings of 3D-position-based phenotype correlations across VGSC genes likely identify phenotypes caused by variants in paralogous genes with similar molecular effects. The framework we developed assumes that both pathogenicity and the molecular impact of a variant are generally conserved. We confirmed that pathogenicity is often preserved across paralogous genes at conserved residues. Nonetheless, our results suggest that applying correlations derived from the 3D positioning of these variants can potentially identify cases where this conservation does not hold or where variants previously classified as pathogenic were misclassified.

Despite efforts to standardize criteria for pathogenicity assignment(16) and many improvements in variant interpretation, about 75% of missense variants in ClinVar(17) (accessed 12/2022) are annotated as variants of uncertain significance (VUS). Extending or modifying existing ACMG criteria has been demonstrated as a promising approach to reclassifying VUSs (20, 38, 43–45). We demonstrated that the PS1 and PM5 criteria of the ACMG guidelines could, in principle, be extended by considering already classified pathogenic variants with corresponding amino acid positions as evidence for pathogenicity. This approach was previously suggested by Clingen Expert curated guidelines for a small set of genes associated with Rasopathies(20). However, here we have demonstrated the generalizability of the approach across a large set of 519 gene families and quantified the evidence gained from this approach.

Our proposed inclusion of the paralogous variants as biologically interpretable evidence of variant pathogenicity has several limitations. First, incorporating pathogenic variants at paralogous positions into the established ACMG/AMP variant classification guidelines requires careful evaluation. This is due to the potential overlap between the basic data supporting an extension of PS1/PM5 criteria to paralogous genes and those already covered by the existing guidelines. Notably, the *in silico* scores, PP3/BP4, overlap, given that many predictive models, such as REVEL(36) or Bayesdel(46), incorporate evolutionary conservation as a fundamental training feature. On the other hand, the para-PS1/PM5 criteria we defined require conservation across paralogous genes at the specified position, thus also considering evidence derived from evolutionary conservation across paralogous genes. We demonstrated that orthologous conservation commonly harnessed in most in silico predictive scores, differs from the evolutionary insights acquired from paralogous gene analyses, albeit they are correlated(15). Furthermore, we demonstrated in this study that even for residues similarly conserved across paralogs, the presence of a pathogenic variant at a conserved paralogous residue provides additional evidence supporting pathogenicity. Nevertheless, enabling the implementation of criteria based on pathogenic variants at paralogous positions along with PP3/BP4 requires a rigorous analysis to determine the discrete evidence provided by integrating pathogenic variants at paralogous positions beyond that provided by the selected PP3, to ensure that information is not considered redundantly. Therefore, incorporating evidence from pathogenic variants at paralogous positions—especially when concurrently considering other related criteria for final classification—introduces a potential risk of inadvertently over-representing shared basic elements. This could lead to an inflated assessment of evidence either supporting or contesting pathogenicity. Second, variants integrated in our framework of pathogenic variants at paralogous positions could be inflated by spliceogenic exonic variants(47). Although previous results suggest that their impact might be minor on our approach, an exclusion of variants with a predicted high splicing impact could resolve this concern. Third, a limitation of our study is the inclusion of control variants aggregated from the gnomAD database, some of which may be pathogenic despite their presence in the general population. In instances where these control variants are indeed pathogenic, the likelihood ratios calculated in our study may represent underestimations, maintaining the conservative nature of our findings.

Conclusion

In Conclusion, our findings suggest that utilizing pathogenic paralogous variants provides significant potential to improve variant interpretation and aid in the diagnosis of pathogenic variants in clinical practice. Reference databases continue to grow and include well-classified pathogenic variants. While we have demonstrated that pathogenic variants in paralogous genes at the same alignment position provide evidence for pathogenicity across all disease-associated gene families, the potential integration of these criteria into the ACMG classification framework would require a careful approach to avoid double counting due to correlation with other criteria that your evolutionary conservation as a feature (e.g., *in silico* prediction scores). Future iterations of variant interpretation guidelines that consider the presence of paralogous pathogenic variants as evidence of pathogenicity could thus significantly increase the application of criteria based on already established pathogenic variants.

Abbreviations

HGNC: Human Gene Nomenclature Consortium; PER: Pathogenic Variant Enriched Regions ACMG: American College of Medical Genetics and Genomics; VUS: Variants of Uncertain Significance; HGMD: Human Gene Mutation Database; gnomAD: Genome Aggregation Database; MAF: Maximum Population Frequency; VGSC: Voltage-Gated Sodium Channels; LR+: Positive Likelihood Ratio; FN: False Negative; FP: False Positive; TP: True Positive; TN: True Negative; DEE: Developmental Epileptic Encephalopathy

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

Data is available in the Supplementary Tables. Code to annotate para-PS1/PM5 criteria with a custom variant dataset for any gene family can be obtained from https://github.com/TobiasBruenger/paraPS1-PM5-annotation.

Competing interests

The authors report no conflicts of interest.

Funding

Funding for this work was provided by the German Federal Ministry for Education and Research (BMBF, Treat-ION, 01GM1907D) to D.L., T.B., and P.M., by the BMBF (Treat-Ion2, 01GM2210B) to P.M, the Fonds Nationale de la Recherche in Luxembourg (FNR, Research Unit FOR-2715, INTER/DFG/21/16394868 MechEPI2) to P.M., the Agencia Nacional de Investigación y Desarrollo de Chile (ANID, Fondecyt 1221464 grant) to E.P., the Familie SCN2A foundation 2020 Action Potential Grant to E.P., the Dravet Syndrome Foundation (grant number, 272016) to D.L, and the NIH NINDS (Channelopathy-Associated Epilepsy Research Center, 5-U54-NS108874) to D.L.

Author’s Contributions

Conceptualization: T.B., A.I., D.L. I.H. ; Data curation: T.B., A.I., E.P; Analysis: T.B, A.I., Supervision: D.L.; P.M.; M.N.; Writing-original draft: T.B., A.I., L.M.; Writing- editing: D.L., P.M., S.C, L.S., S.P., L.M.

Acknowledgement

Not applicable.

References

1. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016 Aug 18;536(7616):285–91.
2. Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci U S A. 2009 Nov 10;106(45):19096–101.
3. Milman A, Behr ER, Gray B, Johnson DC, Andorin A, Hochstadt A, et al. Genotype-Phenotype Correlation of SCN5A Genotype in Patients With Brugada Syndrome and Arrhythmic Events: Insights From the SABRUS in 392 Probands. Circ Genom Precis Med. 2021 Oct;14(5):e003222.
4. Johannesen KM, Liu Y, Koko M, Gjerulfsen CE, Sonnenberg L, Schubert J, et al. Genotype-phenotype correlations in SCN8A-related disorders reveal prognostic and therapeutic implications. Brain. 2022

Sep 14;145(9):2991–3009.

5. Kamada F, Kure S, Kudo T, Suzuki Y, Oshima T, Ichinohe A, et al. A novel KCNQ4 one-base deletion in a large pedigree with hearing loss: implication for the genotype-phenotype correlation. J Hum Genet. 2006;51(5):455–60.
6. Dickerson JE, Robertson DL. On the Origins of Mendelian Disease Genes in Man: The Impact of Gene Duplication. Mol Biol Evol. 2012 Jan;29(1):61–9.
7. Yates B, Gray KA, Jones TEM, Bruford EA. Updates to HCOP: the HGNC comparison of orthology predictions tool. Briefings in Bioinformatics [Internet]. 2021 May 6 [cited 2021 Jul 23];(bbab155). Available from: https://doi.org/10.1093/bib/bbab155
8. Lal D, May P, Perez-Palma E, Samocha KE, Kosmicki JA, Robinson EB, et al. Gene family information facilitates variant interpretation and identification of disease-associated genes in neurodevelopmental disorders. Genome Med. 2020 17;12(1):28.
9. Chen WH, Zhao XM, van Noort V, Bork P. Human Monogenic Disease Genes Have Frequently Functionally Redundant Paralogs. PLoS Comput Biol. 2013 May 16;9(5):e1003073.
10. Wiel L, Venselaar H, Veltman JA, Vriend G, Gilissen C. Aggregation of population-based genetic variation over protein domain homologues and its potential use in genetic diagnostics. Hum Mutat. 2017 Nov;38(11):1454–63.
11. Ware JS, Walsh R, Cunningham F, Birney E, Cook SA. Paralogous annotation of disease-causing variants in long QT syndrome genes. Hum Mutat. 2012 Aug;33(8):1188–91.
12. Zhang X, Theotokis PI, Li N, Investigators the Sh, Wright CF, Samocha KE, et al. Genetic constraint at single amino acid resolution improves missense variant prioritisation and gene discovery [Internet]. medRxiv; 2022 [cited 2023 Oct 19]. p. 2022.02.16.22271023. Available from: https://www.medrxiv.org/content/10.1101/2022.02.16.22271023v1
13. Brunklaus A, Feng T, Brünger T, Perez-Palma E, Heyne H, Matthews E, et al. Gene variant effects across sodium channelopathies predict function and guide precision therapy. Brain. 2022 Jan 17;awac006.
14. Walsh R, Peters NS, Cook SA, Ware JS. Paralogue annotation identifies novel pathogenic variants in patients with Brugada syndrome and catecholaminergic polymorphic ventricular tachycardia. J Med Genet. 2014 Jan;51(1):35–44.
15. Pérez-Palma E, May P, Iqbal S, Niestroj LM, Du J, Heyne HO, et al. Identification of pathogenic variant enriched regions across genes and gene families. Genome Res. 2020;30(1):62–71.
16. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015 May;17(5):405–24.
17. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018 04;46(D1):D1062–7.

18. Marinakis NM, Svingou M, Veltra D, Kekou K, Sofocleous C, Tilemis FN, et al. Phenotype-driven variant filtration strategy in exome sequencing toward a high diagnostic yield and identification of 85 novel variants in 400 patients with rare Mendelian disorders. Am J Med Genet A. 2021 Aug;185(8):2561–71.

19. Zech M, Jech R, Boesch S, Škorvánek M, Weber S, Wagner M, et al. Monogenic variants in dystonia: an exome-wide sequencing study. Lancet Neurol. 2020 Nov;19(11):908–18.

20. Gelb BD, Cavé H, Dillon MW, Gripp KW, Lee JA, Mason-Suares H, et al. ClinGen's RASopathy Expert Panel consensus methods for variant interpretation. Genet Med. 2018 Nov;20(11):1334–45.

21. Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NST, et al. Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat. 2003 Jun;21(6):577–81.

22. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, et al. Ensembl 2018. Nucleic Acids Res. 2018 Jan 4;46(Database issue):D754–61.

23. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434–43.

24. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011 Aug 1;27(15):2156–8.

25. Brunklaus A, Brünger T, Feng T, Fons C, Lehikoinen A, Panagiotakaki E, et al. The gain of function SCN1A disorder spectrum: novel epilepsy phenotypes and therapeutic implications. Brain. 2022 Jun 13;awac210.

26. Brunklaus A, Pérez-Palma E, Ghanty I, Xinge J, Brilstra E, Ceulemans B, et al. Development and Validation of a Prediction Model for Early Diagnosis of SCN1A-Related Epilepsies. Neurology. 2022 Mar 15;98(11):e1163–74.

27. Wolff M, Johannesen KM, Hedrich UBS, Masnada S, Rubboli G, Gardella E, et al. Genetic and phenotypic heterogeneity suggest therapeutic implications in SCN2A-related disorders. Brain. 2017 May 1;140(5):1316–36.

28. Crawford K, Xian J, Helbig KL, Galer PD, Parthasarathy S, Lewis-Smith D, et al. Computational analysis of 10,860 phenotypic annotations in individuals with SCN2A-related disorders. Genet Med. 2021 Jul;23(7):1263–72.

29. Zaman T, Helbig KL, Clatot J, Thompson CH, Kang SK, Stouffs K, et al. SCN3A-related neurodevelopmental disorder: A spectrum of epilepsy and brain malformation. Ann Neurol. 2020 Aug;88(2):348–62.

30. Sobreira N, Schiettecatte F, Valle D, Hamosh A. GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. Hum Mutat. 2015 Oct;36(10):928–30.

31. Walsh R, Lahrouchi N, Tadros R, Kyndt F, Glinge C, Postema PG, et al. Enhancing rare variant interpretation in inherited arrhythmias through quantitative analysis of consortium disease cohorts and population controls. Genet Med. 2021 Jan;23(1):47–58.

32. Whiffin N, Minikel E, Walsh R, O'Donnell-Luria AH, Karczewski K, Ing AY, et al. Using high-resolution variant frequencies to empower clinical genome interpretation. Genet Med. 2017 Oct;19(10):1151–

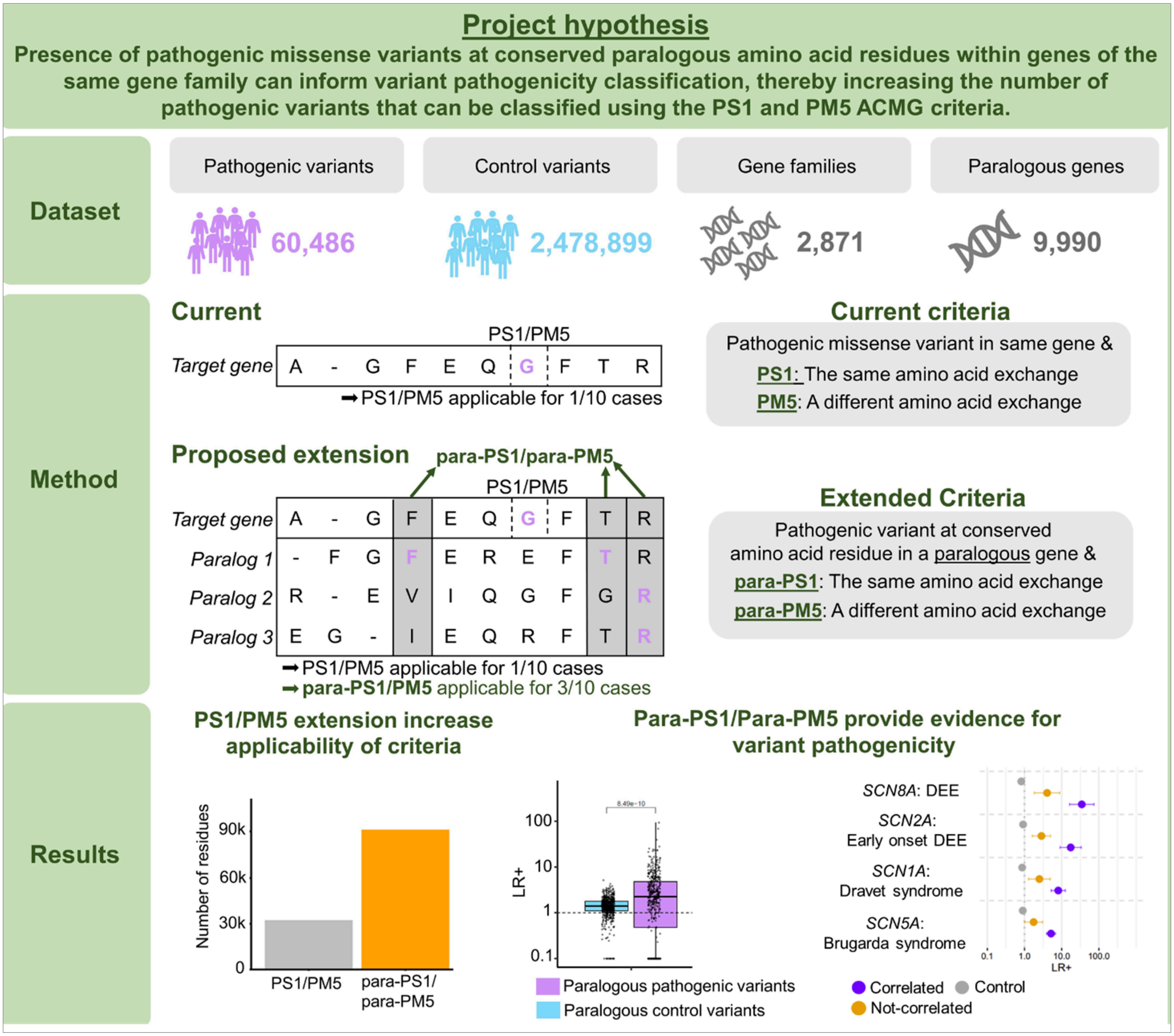
8.

33. Kinsella RJ, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G, et al. Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database (Oxford). 2011;2011:bar030.
34. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research. 2004 Mar 8;32(5):1792–7.
35. Iqbal S, Brünger T, Pérez-Palma E, Macnee M, Brunklaus A, Daly MJ, et al. Delineation of functionally essential protein regions for 242 neurodevelopmental disorders. Brain. 2022 Oct 18;awac381.
36. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet. 2016 Oct 6;99(4):877–85.
37. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. PLoS Comput Biol [Internet]. 2010 Dec 2 [cited 2019 Dec 29];6(12). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2996323/
38. Frazer J, Notin P, Dias M, Gomez A, Min JK, Brock K, et al. Disease variant prediction with deep generative models of evolutionary data. Nature. 2021 Nov;599(7883):91–5.
39. Heyne HO, Baez-Nieto D, Iqbal S, Palmer DS, Brunklaus A, May P, et al. Predicting functional effects of missense variants in voltage-gated sodium and calcium channels. Sci Transl Med. 2020 Aug 12;12(556).
40. Tokheim C, Bhattacharya R, Niknafs N, Gygax DM, Kim R, Ryan M, et al. Exome-scale discovery of hotspot mutation regions in human cancer using 3D protein structure. Cancer Res. 2016 Jul 1;76(13):3719–31.
41. Quinodoz M, Peter VG, Cisarova K, Royer-Bertrand B, Stenson PD, Cooper DN, et al. Analysis of missense variants in the human genome reveals widespread gene-specific clustering and improves prediction of pathogenicity. Am J Hum Genet. 2022 Mar 3;109(3):457–70.
42. Brunklaus A, Du J, Steckler F, Ghanty II, Johannesen KM, Fenger CD, et al. Biological concepts in human sodium channel epilepsies and their relevance in clinical practice. Epilepsia. 2020;61(3):387–99.
43. Kelly MA, Caleshu C, Morales A, Buchan J, Wolf Z, Harrison SM, et al. Adaptation and validation of the ACMG/AMP variant classification framework for MYH7-associated inherited cardiomyopathies: recommendations by ClinGen's Inherited Cardiomyopathy Expert Panel. Genet Med. 2018 Mar;20(3):351–9.
44. Patel MJ, DiStefano MT, Oza AM, Hughes MY, Wilcox EH, Hemphill SE, et al. Disease-specific ACMG/AMP guidelines improve sequence variant interpretation for hearing loss. Genet Med. 2021 Nov;23(11):2208–12.
45. Pejaver V, Byrne AB, Feng BJ, Pagel KA, Mooney SD, Karchin R, et al. Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. Am J Hum Genet. 2022 Dec 1;109(12):2163–77.

46. Feng BJ. PERCH: A Unified Framework for Disease Gene Prioritization. Hum Mutat. 2017 Mar;38(3):243–51.

47. Loong L, Cubuk C, Choi S, Allen S, Torr B, Garrett A, et al. Quantifying prediction of pathogenicity for within-codon concordance (PM5) using 7541 functional classifications of BRCA1 and MSH2 missense variants. Genet Med. 2022 Mar;24(3):552–63.

Figures



Project hypothesis

Presence of pathogenic missense variants at conserved paralogous amino acid residues within genes of the same gene family can inform variant pathogenicity classification, thereby increasing the number of pathogenic variants that can be classified using the PS1 and PM5 ACMG criteria.

Dataset

Pathogenic variants 60,486

Control variants 2,478,899

Gene families 2,871

Paralogous genes 9,990

Method

Current

Current criteria

Pathogenic missense variant in same gene &

PS1: The same amino acid exchange

PM5: A different amino acid exchange

Proposed extension

Extended Criteria

Pathogenic variant at conserved amino acid residue in a paralogous gene &

para-PS1: The same amino acid exchange

para-PM5: A different amino acid exchange

Results

PS1/PM5 extension increase applicability of criteria

Para-PS1/Para-PM5 provide evidence for variant pathogenicity

PS1/PM5
Target gene A - G F E Q G F T R
➡PS1/PM5 applicable for 1/10 cases
para-PS1/para-PM5
PS1/PM5
Target gene A - G F E Q G F T R
Paralog 1 - F G F E R E F T R
Paralog 2 R - E V I Q G F G R
Paralog 3 E G - I E Q R F T R
➡PS1/PM5 applicable for 1/10 cases
➡para-PS1/PM5 applicable for 3/10 cases
Number of residues
90k
60k
30k
0
PS1/PM5
para-PS1/para-PM5
LR+
100
10
1
0.1
8.49e-10
Paralogous pathogenic variants
Paralogous control variants
SCN8A: DEE
SCN2A: Early onset DEE
SCN1A: Dravet syndrome
SCN5A: Brugarda syndrome
0.1
1.0
10.0
100.0
LR+
Correlated
Control
Not-correlated

Figure 1

Graphical summary of the study.

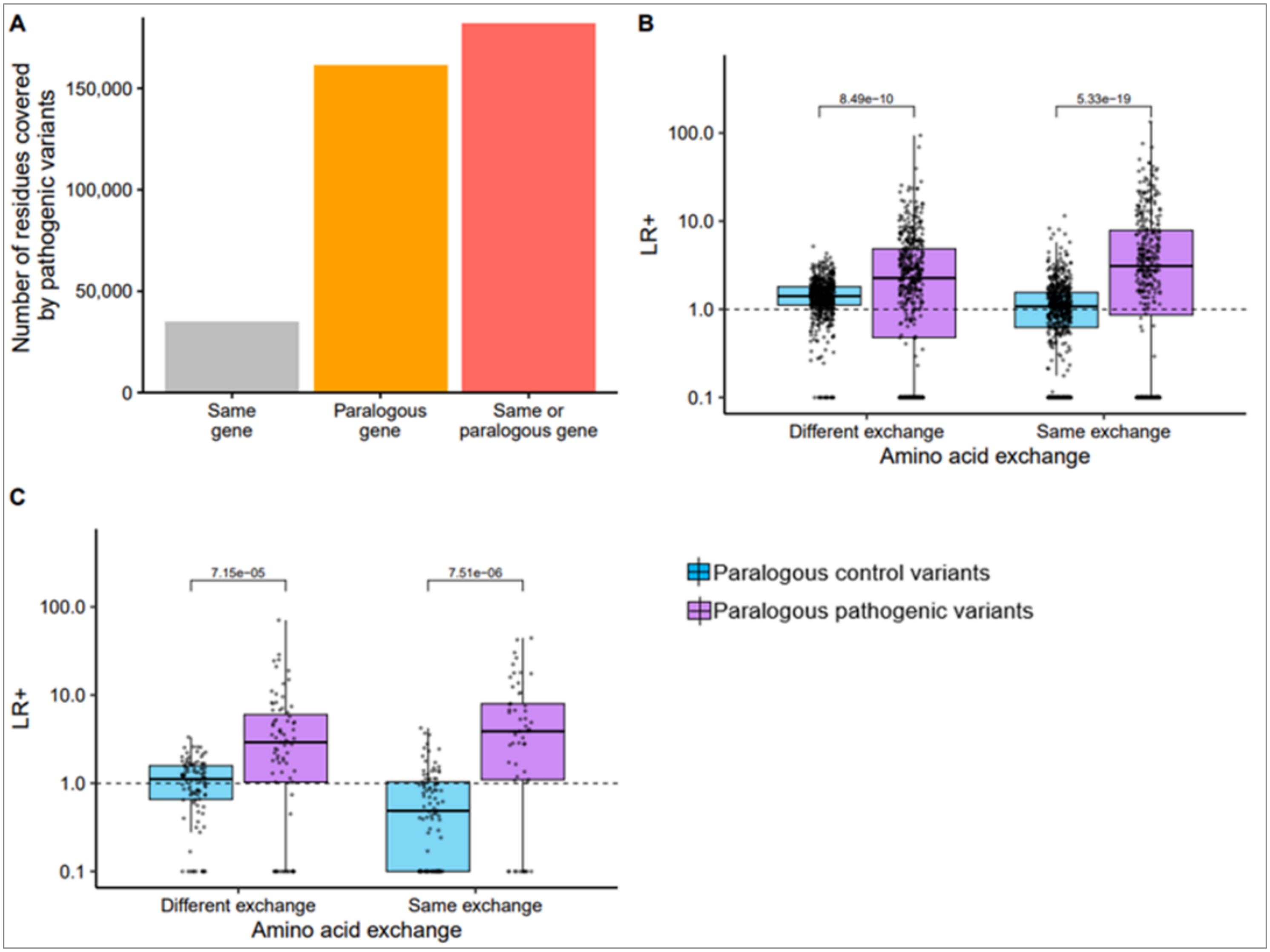


Figure 2

Individual pathogenic paralogous variants can serve as a proxy for variant pathogenicity. A) Number of amino acid residues in 519 gene families that have a pathogenic variant (ClinVar/HGMD) at the same protein position in the same gene or a corresponding protein residue in a paralogous gene. B) Amino acids with a paralogous pathogenic variant at a paralogous aliment position have an increased positive likelihood ratio (LR+ >1). In contrast, amino acids with a paralogous control variant (gnomAD) at a paralogous alignment position are not enriched for pathogenic variants. Each data point represents the gene-wise LR+. The gene-wise LR+ was calculated for genes where 10 or more pathogenic variants (ClinVar/HGMD) and control variants (gnomAD) could be mapped. C) As in (B), but limited to missense constraint genes (Missense-z score > 3.09).

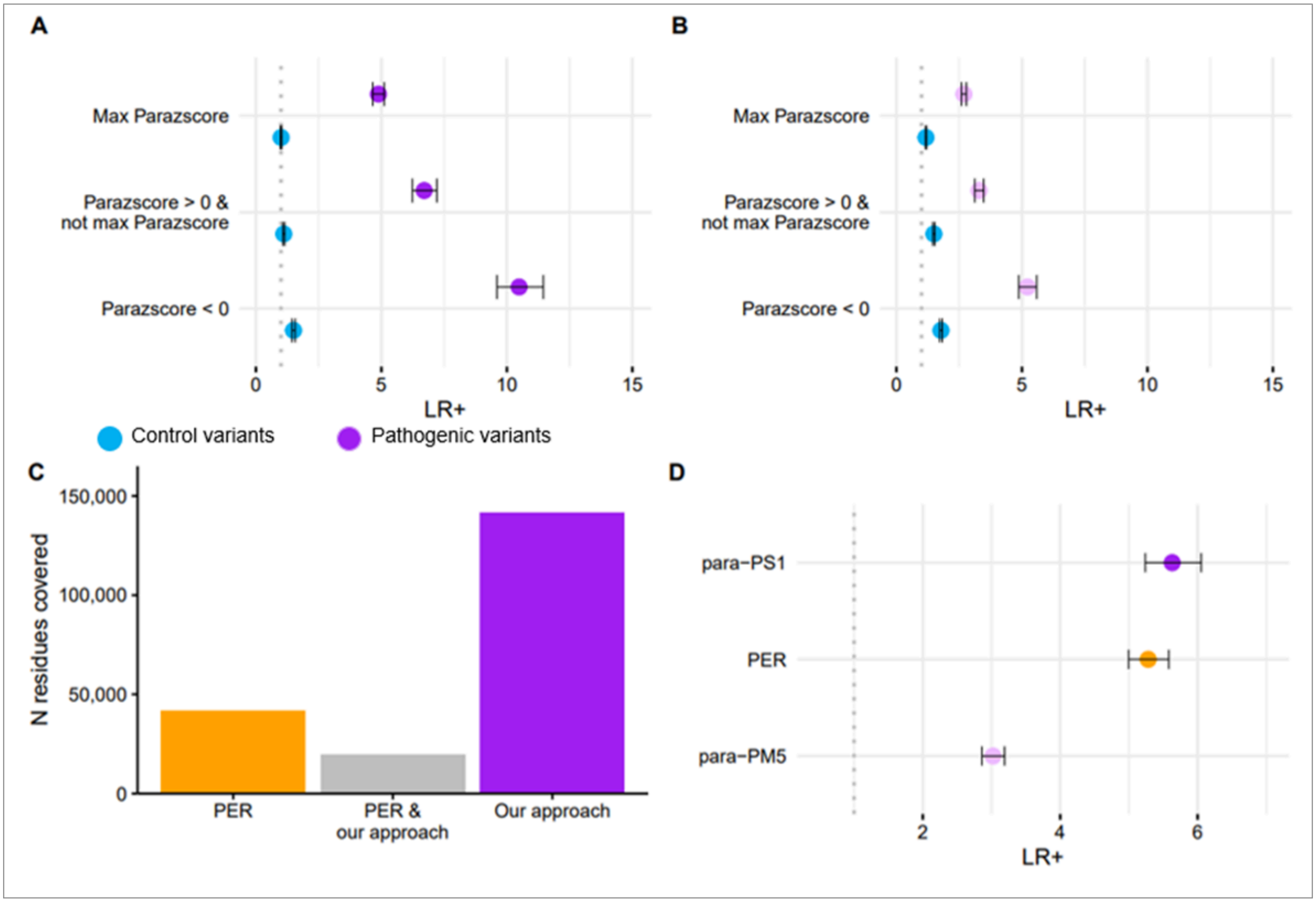


Figure 3

Comparison to established gene family-based methods. A) The forest plot illustrates the enrichment of pathogenic versus control variants applying the para-PS1 criterium for residues with similar paralog conservation levels, as defined in Lal et al., 2020(8). B) Similar to (A), but for the para-PM5 criterium. C) The bar plot shows the number (N) of amino acid residues across all genes where a previously established approach (Pathogenic Enriched Region, PER; Perez-Palma et al., 2019(15)) and/or our para-PS1/ para-PM5 ACMG criteria extension can be applied. D) The forest plot compares the likelihood ratios (LR+) for amino acid residues within a PER and amino acid residues where para-PS1/para-PM5 criteria can be applied (see Methods for details).

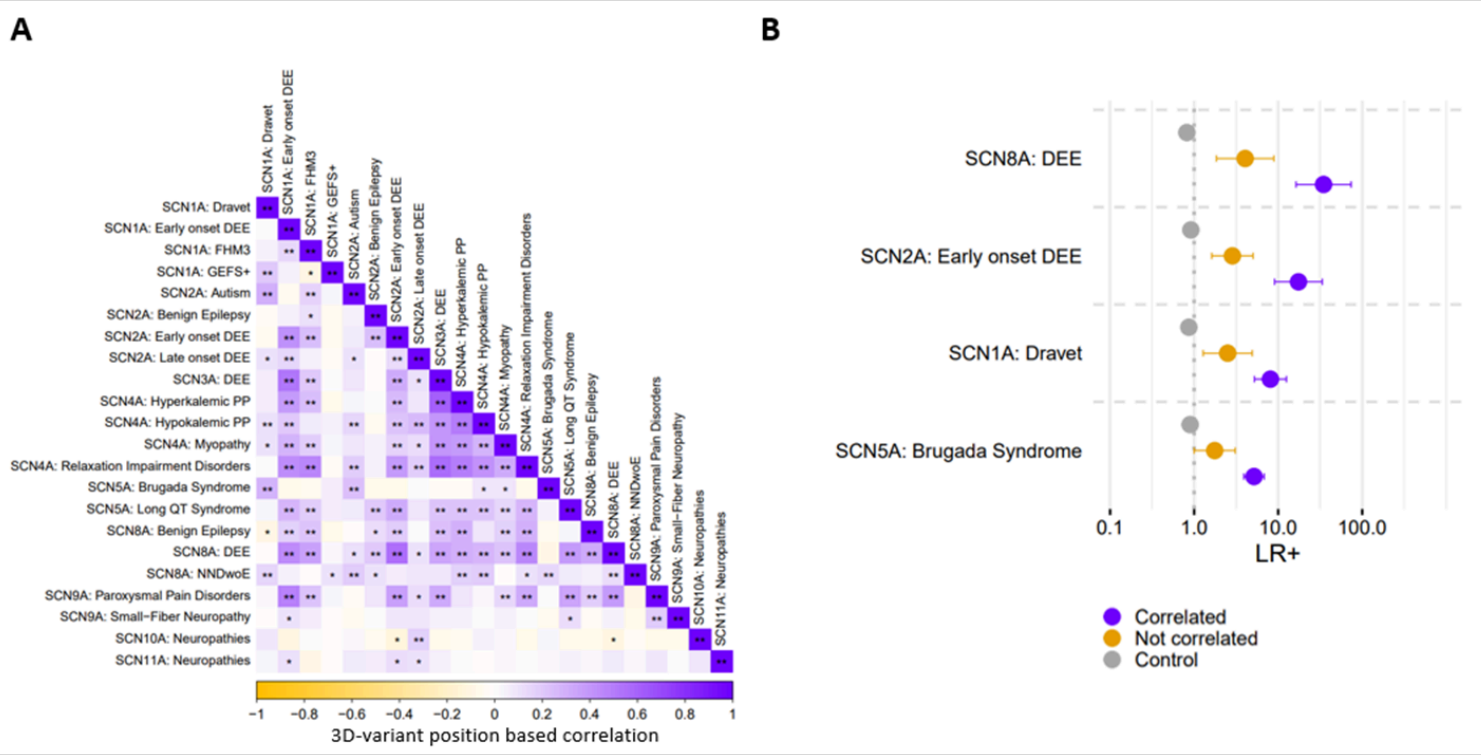


Figure 4

Leveraging phenotype correlations to enhance the application of paralogous pathogenic variants.

A) Displayed is a correlation matrix that delineates the relationships between the 3D variant distributions across various phenotypes. Phenotypes that share significantly (after Bonferroni adjustment) similar 3D-variant distributions are color-coded in purple, whereas those with significantly distinct distributions are in orange. Statistically significant correlations are marked with stars (* for Padj<0.05 and ** for P<0.001). B) Presented is a forest plot capturing the positive likelihood ratio for four pivotal phenotypes that is derived from a comparison of affected individuals and control variants sourced from gnomAD. These ratios were computed by either 1) employing paralogous variants from affected individuals that exhibited a significantly positive correlation based on 3D position (depicted in purple), 2) utilizing paralogous variants from affected individuals displaying a 3D position-based negative correlation (showcased in orange) and 3) considering paralogous control variants (represented in grey). Abbreviations: DEE – Developmental Epileptic Encephalopathy; FHM3 – Familial Hemiplegic Migraine Type 3; PP – Periodic Paralysis; NDDwoE – Neurodevelopmental Disorders Without Epilepsy; Dravet – Dravet Syndrome.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementarymaterialGenomemedicine.docx
- SupplementaryTable1.xlsx
- SupplementaryTable2.xlsx