

Pathogenic paralogous variants can be used to apply the ACMG PS1 and PM5 variant interpretation criteria

Tobias Br nger^{1*}, Alina Ivaniuk^{2*}, Eduardo P rez-Palma³, Ludovica Montanucci^{2,4}, Stacey Cohen^{5,6,7}, Lacey Smith⁸, Shridhar Parthasarathy^{5,6,7}, Ingo Helbig^{5,6,7,9}, Michael Nothnagel¹, Patrick May¹⁰, Dennis Lal^{1,2,4,11,12,13*}

¹Cologne Center for Genomics (CCG), University of Cologne, Cologne, 50931, Germany

²Genomic Medicine Institute, Lerner Research Institute, Cleveland Clinic, Cleveland, OH 44106, USA.

³Universidad del Desarrollo, Centro de Gen tica y Gen mica, Facultad de Medicina Cl nica Alemana. Santiago, Chile.

⁴Department of Neurology, The University of Texas Health Science Center at Houston, Houston, TX, OH 44106, USA.

⁵Division of Neurology, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA.

⁶The Epilepsy NeuroGenetics Initiative (ENGIN), Children's Hospital of Philadelphia, Philadelphia, PA, USA.

⁷Department of Biomedical and Health Informatics (DBHi), Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA.

⁸Epilepsy Genetics Program, Division of Epilepsy and Clinical Neurophysiology, Department of Neurology, Boston Children's Hospital, Boston, MA, USA.

⁹Department of Neurology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, 19104, USA.

¹⁰Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg

¹¹Stanley Center of Psychiatric Research, Broad Institute of Harvard and MIT, 75 Ames Street, Cambridge, 02142, MA, USA

¹²Epilepsy Center, Neurological Institute, Cleveland Clinic, Cleveland, OH 44106, USA

¹³Center for Neurogenetics, The University of Texas Health Science Center at Houston, Houston, TX.

*Authors contributed equally to this work

*Corresponding author. Email: dennis.lal@uth.tmc.edu

Abstract

Purpose

The majority of missense variants in clinical genetic tests are classified as variants of uncertain significance. Broadening the evidence of the PS1 and PM5 criteria has the potential to increase conclusive variant interpretation.

Methods

We hypothesized that incorporation of pathogenic missense variants in conserved residues across paralogous genes can increase the number of variants where ACMG PS1/PM5 criteria can be applied. We mapped over 2.5 million pathogenic and general population variants from ClinVar, HGMD, and gnomAD databases onto 9,990 genes and aligned these by gene families. Subsequently, we developed a novel framework to extend PS1/PM5 by incorporating pathogenic paralogous variants annotations (para-PS1/PM5).

Results

We demonstrate that para-PS1/PM5 criteria increase the number of classifiable amino acids 3.6-fold compared to PS1 and PM5. Across all gene families with at least two disease-associated genes, the calculated likelihood ratios suggest moderate evidence for pathogenicity. Moreover, for 36 genes, the extended para-PS1/PM5 criteria reach strong evidence level.

Conclusion

We show that single pathogenic paralogous variants incorporation at paralogous protein positions increases the applicability of the PS1 and PM5 criteria, likely leading to a reduction of variants of uncertain significance across many monogenic disorders. Future iterations of the ACMG guidelines may consider para-PS1 and para-PM5.

Introduction

Large gene panels, exome, and genome sequencing have led to the identification of novel variants at an exponential rate¹. Up to 80% of pathogenic variants are located within protein-coding regions of the gene², 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, which further complicates variant interpretation³⁻⁵. Despite these challenges, variant classification is necessary for diagnosing rare and genetically heterogeneous disorders, and for the development of personalized medicine.

To standardize variant interpretation, the American College of Medical Genetics and Genomics (ACMG) published recommendations for evaluating the pathogenicity of variants⁶. However, >45% of single nucleotide variants reported in the ClinVar database⁷ (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 genetic test results 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 pathogenicity⁶. 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^{8,9}, the application of these evidence criteria is limited.

About 80% of genes associated with monogenic disorders are paralogs¹⁰. These paralogous genes can be grouped into 2871 gene families as defined by the Human Gene Nomenclature Consortium (HGNC)¹¹ with >80% sequence similarity¹². 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 domains¹³. We and others have demonstrated that functional domains that are conserved within a gene family are enriched for pathogenic variants and depleted for population variants¹⁴⁻¹⁶. In molecular studies, it has been shown that a protein's biophysical function at these protein domains is mechanistically conserved within a gene family and therefore, the exchange of a single residue has similar molecular effects across gene family members¹⁶⁻¹⁸. Therefore, the current PS1/PM5 criteria could be informed not only by prior reported pathogenic variants at the same position in the gene of interest, but also by observed pathogenic variants at conserved positions in related genes of the same family.

Aside from the identification of conservation patterns within gene families, we previously identified differential distribution of missense variants from the general population vs. pathogenic missense variants^{16,18}. We showed that evolutionarily conserved protein regions across paralogous genes are enriched for pathogenic variants and depleted for population variants¹². Further, we identified regions across paralogous genes that are enriched for pathogenic versus control variants, which can be annotated as pathogenic variant enriched regions (PERs)¹⁴. We demonstrated that new missense variants within PERs are more likely to be pathogenic compared to non-PER regions in the same gene¹⁴. Although the PER-annotation has the potential to identify disease-sensitive regions across protein sequences, the approach has the disadvantage that it currently has low sensitivity since we were limited in power to identify PERs, as the vast majority of new missense variants are located outside of PERs.

In this new study, we aim to build on previous gene family conservation and PER research to overcome some of the current limitations in clinical variant interpretation. Specifically, we hypothesize that variants in conserved residues in paralogous genes can increase the number of variants classifiable using the ACMG PS1 and PM5 criteria. If our hypothesis is true, the PS1/PM5 ACMG criteria could be extended to include paralogous variants at conserved amino acid positions and increase the number of clinically resolved variants significantly. However, to the best of our knowledge, it has not been tested across the whole protein-coding exome whether single missense variants from paralogous genes can serve as a proxy for variant pathogenicity. As proof of concept, we show that in 519 gene families (1,459 genes) with high sequence similarity, the presence of a single pathogenic variant in a gene family member at a corresponding protein position increases the likelihood that a novel genetic variant at a conserved paralogous position in the target gene is pathogenic.

Methods

Gene family definition

We obtained the paralogous genes that belong to a gene family from Pérez-Palma et al. 2020¹⁴, as originally described in Lal et al., 2020¹². Briefly, the human paralog definitions were taken from Ensembl BioMart¹⁹ and filtered for those with an HGNC symbol¹¹. 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 MUSCLE²⁰ 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 Figure 1).

We further establish an expanded set of criteria, termed para-PS1/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 Figure 1). While considering the definition of the criteria (see above) we counted for each gene *i*) the number of pathogenic variants (for details about variant collection see Supplementary Methods) 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. We calculated the positive likelihood ratios following the methodology outlined by Tavtigian et al., 2018²¹ and Pejaver et al., 2022²², to determine the level of evidence each criterion provides for pathogenicity. Briefly, we calculated the likelihood ratios in the presence of a pathogenic paralogous variant with either the same amino acid substitution (para-PS1) or a different amino acid substitution (para-PM5). The positive likelihood ratio was computed using the sensitivity and specificity of the test:

Equation 1:

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

where LR+ represents the positive likelihood ratio, TP denotes true positives, TN signifies true negatives, FP stands for false positives, and FN represents false negatives. To derive a single estimate across all genes, we combined the counts of pathogenic and general population variants and repeated the calculation. All analyses were performed using R v.4.2.1.

Results

Incorporating pathogenic paralogous variants increases the number of PS1/PM5 classifiable amino acids

The goal of this study was, through a bioinformatic framework, to test and expand the application of the PS1 and PM5 to paralogous evidence and thus increase the resolution of the current clinical genetic variant interpretation guidelines of the ACMG. Specifically, we investigated whether the presence of pathogenic variants in paralogous genes at a corresponding position can be used to extend the existing strong PS1 ('same amino acid change as a previously established pathogenic variant regardless of nucleotide change') and moderate PM5 (novel amino acid substitution where a different amino acid substitution determined to be pathogenic has been seen before) variant pathogenicity classification criteria (Figure 1). We defined a 'paralogous variant', as a variant that fulfills two criteria: (1) it is located in a paralogous gene at the corresponding residue index position, as determined by multiple sequence alignment (see Methods for details), and (2) with the same reference amino acid as in the target gene.

The inclusion of paralogous variants could decrease the number of variants of uncertain significance. Before establishing the validity of the incorporation of paralogous variants in the PS1/PM5 criteria, we quantified the number of additional amino acids residues where the updated PS1/PM5, further referred to as para-PS1 & para-PM5, could be applied. We aggregated a total of 60,486 pathogenic variants from ClinVar⁷ and HGMD²³ and mapped them to 2,871 different gene family alignments, consisting of 9,990 genes (Figure 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 application of the extended 'para-PS1' and 'para-PM5' criteria 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, Figure 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 Figure 2).

Presence of a single pathogenic paralogous variant can provide supportive to strong evidence for variant pathogenicity

Next, we tested the validity of the inclusion of pathogenic paralogous variants for the application of the para-PS1 & para-PM5 criteria. In addition to the aforementioned pathogenic variants, we included 2,478,899 general population variants from the gnomAD database²⁴, that served as controls in our study. To estimate the evidence of the para-PS1 and para-PM5 criteria we positive likelihood ratio (LR+) following the Bayesian framework proposed by Tavtigian et al. in 2018²¹. Across the 519 selected gene families, we observed an LR+ of 8.32 (8.02-8.62, 95% confidence interval (CI)) pathogenic versus control variants when a pathogenic paralogous variant with the same amino acid exchange was present at a corresponding alignment index position (para-PS1, for details on the approach, see Methods). 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). Applying proposed cutoffs to translate the LR+ to weights of evidence considering the suggested²¹ prior probability of 0.1, we can consider an LR+ > 18.7 as strong evidence, LR+ > 4.3 as moderate, and LR+ > 2.08 as supporting evidence for variant pathogenicity²¹. Consequently, across all assessed genes the para-PS1 criterium can be considered as moderate evidence. Notably, even for paralogous variants with a different substitution at the same alignment index position (para-PM5), 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% CI). The para-PM5 criterium could also be considered as moderate evidence if the point estimate is considered and as supporting evidence if the lower bound of the 95% interval ($LR+_{\text{lower bound}} = 4.24$) is considered.

Next, we calculated for a subset of genes for which at least 10 pathogenic and control variants could be mapped on the gene LR+ on a gene-wise level. We observed a wide range of LR+ across different genes (Figure 2A, B). Pathogenic paralogous variants applied as para-PS1 criterium (same amino acid exchange), provide strong evidence for variant

pathogenicity in 36 genes, moderate evidence in 87 genes, and supporting evidence in 40 genes (Figure 2D, Supplementary Table 2). After assessing the same amino acid exchange, we moved on to evaluating the evidence for pathogenic paralogous variants at alignment index positions with different substitutions, applying the para-PM5 criteria. We observed that while the para-PS1 same substitution criteria provided strong evidence in 24.5% of genes with an LR+ >2.08 (N = 163 genes), whereas the para-PM5 criteria provided strong evidence in only in 5.8 % of genes (N genes_{with para-PM5 LR+ >2.08} = 187 genes). Overall, genes where paralogous pathogenic variants can provide strong evidence under either para-PS1 and/or para-PM5 criteria typically contained fewer genes and also tended to contain fewer genes where pathogenic variants are established (Supplementary Figure 3 A-D). Notably, genes with higher constraints for missense variants did not show significantly higher positive likelihood ratios (Supplementary Figure 3 E, F).

The para-PS1/para-PM5 improves previous family-based variant interpretation approaches

Previously, we developed two methods for variant classification, both of which harnessed the power of gene family information on an exome-wide scale^{12,14}. Initially, we established a paralog conservation score (referred to as 'Parazscore'¹²) and demonstrated that amino acids conserved within a gene family are significantly enriched for pathogenic variants. Notably, a fundamental prerequisite for the implementation of para-PS1 & para-PM5 criteria is the conservation of amino acid residues between the target gene and its paralogous gene, in which a paralogous pathogenic variant is detected. Hence, whenever para-PS1 or para-PM5 criteria become applicable, a certain degree of conservation within the genes of the same gene family becomes inevitable. This conservation potentially explains a portion of the elevated LR+ we observed. To further elucidate the potential impact of paralog conservation on the calculated LR+, we repeated our previous analysis allocating every amino acid to subgroups depending on their paralog conservation. This segmentation was done using the Parazscore as a metric to quantify paralog conservation (see Supplementary 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, Figure 3A) as well as the para-PM5 criterium (Parazscore <0; LR+_{para-PM5} = 5.21, 95% CI = 4.87-5.59, Figure 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 respectively. This observation suggests that the existence of pathogenic paralogous variants provides additional information beyond the level

of conservation between paralogous genes.

We then compared the para-PS1/PM5 criteria with a second approach¹⁴ which utilizes pathogenic variants from paralogous genes of the same gene family. By analyzing the distribution of pathogenic and control variants across a gene family the approach identifies pathogenic variant enriched regions ('PERs', on average 33 consecutive amino acids¹⁴) across paralogous genes that are significantly enriched for pathogenic versus control variants. Due to the sliding window approach to identifying regional enrichments of 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 para-PS1/PM5 with the PER approach. We used an independent set of pathogenic and control variants that were not utilized in the PER generation or the generation or validation of the para-PS1 and para-PM5 criteria (see Supplementary Methods for details). We found that the para-PS1 & para-PM5 approaches captured 2.2 times more residues compared to PERs (Figure 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, both suggesting moderate evidence for pathogenicity ($LR_{+PER} = 5.28$, $LR_{+para-PS1} = 5.63$, Figure 3D). In contrast, para-PM5 with an LR+ of 3.02 could provide supporting evidence for pathogenicity.

Application of para-PS1/para-PM5: Two case examples

To demonstrate how the ACMG criteria extension para-PS1 & para-PM5 could lead to a change in variant classification, we provide two examples where our para-PS1 and para-PM5 have been applied (Figure 4).

Case 1 is a young boy assessed for bilateral sensory hearing loss with a presumably infantile onset. Genetic testing revealed a heterozygous variant in *COL11A1* NM_001854.4:c.3232G>C (p.Gly1078Arg). The variant is absent in the gnomAD database²⁴ (moderate criterion PM2) and multiple lines of computational evidence support its pathogenicity (REVEL²⁵ = 0.997, MutPred2²⁶ = 0.975 – strong evidence PP3_Strong²²). The available evidence (Supplementary Table 3) would allow a classification of the variant as a variant of uncertain significance (VUS) and leave the diagnosis uncertain⁶. However, several paralog genes have pathogenic variants reported at the same alignment position (1469) (Supplementary Table 1). Variants in two paralog genes have the same exchange: *COL3A1* (p.Gly729Arg) and *COL1A2* (p.Gly643Arg). According to our findings (Supplementary Table

2), this can be applied as a criterion para-PS1_Moderate, which would allow us to classify the variant as likely pathogenic scoring 2 moderate (para-PS1_Moderate, PM2) and one strong (PP3_Strong) criteria.

Case 2 is a woman in her 40's with syncopal episodes since childhood and a recent diagnosis of catecholaminergic polymorphic ventricular tachycardia. She underwent a genetic panel testing revealing a heterozygous variant in *RYR2*, NM_001035.3:c.1870G>A (p.Ala624Thr). No other direct family members were available for testing. *In-silico* prediction tools provide supporting evidence for variant pathogenicity (REVEL²⁵ = 0.573, MutPred2²⁶ = 0.799; PP3²²). The variant affects a missense-constraint gene²⁷ (PP2) and is not present in gnomAD²⁴ database (PM2). Taken together, available evidence (Supplementary Table 3) classifies the variant as VUS. However, we observed a pathogenic variants with the same exchange in the paralogous gene *RYR1* p.Ala612Thr (Supplementary Table 1), which allows us to apply a para-PS1_Moderate (Supplementary Table 2). Considering the application of the criterion, the variant can be classified as likely pathogenic given 2 supporting (PP2, PP3) and 2 moderate (PM2, para-PS1_Moderate) criteria.

Discussion

Many paralogs are highly conserved in sequence and have similar biophysical molecular functions. However, the current variant interpretation guidelines consider only pathogenic missense variants in the gene of interest. Here, we developed and validated a bioinformatic framework to include missense variants from paralogous genes at alignment index positions as additional source for the application of the PS1 and PM5 criteria ('para-PS1' and 'para-PM5'). Our analysis revealed that the extended para-PS1 (pathogenic paralogous variant with the same amino acid exchange as the target variant) and para-PM5 (pathogenic paralogous variant with a different amino acid exchange as the target variant) criteria can increase the number of residues where the criteria can be applied compared to original PS1 and PM5 criteria by 3.6 fold. We show that the para-PS1 and para-PM5 criteria can provide supporting to strong evidence of pathogenicity and complement and extend existing methods^{12,14}.

Despite efforts to standardize criteria for pathogenicity assignment⁶ and many improvements in variant interpretation, about 75% of missense variants in ClinVar⁷ (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^{22,28–31}. We demonstrated that the PS1 and PM5 criteria of the ACMG guidelines, which consider knowledge about an established pathogenic variant with the same amino acid change as a

strong evidence criterion for variant pathogenicity, can be extended by including paralogous pathogenic variants.

According to the current guidelines⁶, the presence of a pathogenic variant with the same amino acid substitution in the same gene (PS1) provides strong evidence of pathogenicity, whereas the presence of a pathogenic variant with a different amino acid substitution in the same gene (PM5) provides moderate evidence⁶. This is consistent with our observation that the para-PS1 enrichment for pathogenic variants was higher (OP = 8.32) compared to the enrichment that we observed when the para-PM5 criterion was applied (OP = 4.32). To estimate the level of evidence the extended para-PS1 criteria could provide we followed the Bayesian framework as proposed by Tavtigian et al., 2018²¹ and previously applied in several other studies^{32–35}. In contrast to the established PS1 and PM5 criteria, which provide strong and moderate evidence respectively, both para-PS1 and para-PM5 can provide moderate evidence across genes in our study. We speculate that the difference between the moderate evidence provided by para-PS1 criterion compared to the strong evidence by PS1 may be due to potentially misclassified variants in large patient variant databases such as ClinVar³⁶ which affects the quantification of the reported effect. A problem was observed for novel variants submitted before the availability of large reference databases. Importantly, while the level of evidence for variant pathogenicity that PS1 and PM5 can provide is considered uniform across all genes⁶ we showed that for the extended criteria, the level can be calculated specifically for each gene family. For example, while the positive likelihood ratio across all gene families meets the evidence threshold for moderate strength, for 36 genes the LR+ is >18.7, which can be translated into strong evidence²¹.

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 function³⁷. These pathogenic variant-enriched regions have proven valuable for variant classification through established guidelines for variant interpretation⁶ and the use of *in-silico* prediction algorithms³⁸. 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 genes¹⁴. 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 alignments¹⁶, 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.

Our proposed inclusion of the paralogous variants as an extension of the PS1 and PM5 ACMG criteria in future iterations comes with a few limitations. One limitation is the focus on paralog conserved regions. Our approach does not capture pathogenic variants in gene-specific regions that are not conserved across paralogous genes. Within these regions, the original PS1/PM5 criteria can be applied whereas the extended para-PS1/PM5 criteria can not. Our approach is an extension of the PS1/PM5 criteria and does not limit their application in such regions. Furthermore, our gene families were required to have >80% sequence similarity and previous studies have shown that the majority of functionally important regions are conserved across paralogs¹², and *de novo* domains are rare and could be annotated separately. Secondly, 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. Another limitation is the partially small number of established pathogenic variants and paralogous pathogenic variants. Many gene families lack a sufficiently large collection of identified pathogenic variants that serve as a proxy for pathogenicity which limits statistical power to confidently quantify the gene-specific evidence provided by paralogous pathogenic variants. The number of individual genes where paralogous pathogenic variants provide significant supportive or higher evidence for pathogenicity was associated with the number of paralogous pathogenic variants ($P_{\text{para-PS1}} = 1.6\text{e-}24$, $P_{\text{para-PM5}} = 4.6\text{e-}09$, Supplementary Figure 4). Consequently, although we have demonstrated that our approach can be powerful to extend the current ACMG criteria, there are individual genes where our approach lacks sufficient power to validate evidence strength on the gene level, as indicated by 169 genes that have increased LR+ when a paralogous pathogenic variant with the same exchange is present, however without statistical significance.

In conclusion, our findings suggest that the extended criteria para-PS1 and para-PM5 have 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 the applicability of the PS1 and PM5 criteria will, as a consequence, will grow the applicability of the para-PS1/PM5 criteria will increase at an even faster rate. 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.

References

1. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
2. Choi, M. *et al.* Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci U S A* **106**, 19096–19101 (2009).
3. Milman, A. *et al.* Genotype-Phenotype Correlation of SCN5A Genotype in Patients With Brugada Syndrome and Arrhythmic Events: Insights From the SABRUS in 392 Proband. *Circ Genom Precis Med* **14**, e003222 (2021).
4. Johannesen, K. M. *et al.* Genotype-phenotype correlations in SCN8A-related disorders reveal prognostic and therapeutic implications. *Brain* **145**, 2991–3009 (2022).
5. Kamada, F. *et al.* A novel KCNQ4 one-base deletion in a large pedigree with hearing loss: implication for the genotype-phenotype correlation. *J Hum Genet* **51**, 455–460 (2006).
6. Richards, S. *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* **17**, 405–424 (2015).
7. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).
8. Marinakis, N. M. *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* **185**, 2561–2571 (2021).
9. Zech, M. *et al.* Monogenic variants in dystonia: an exome-wide sequencing study. *Lancet Neurol* **19**, 908–918 (2020).

10. Dickerson, J. E. & Robertson, D. L. On the Origins of Mendelian Disease Genes in Man: The Impact of Gene Duplication. *Mol Biol Evol* **29**, 61–69 (2012).
11. Yates, B., Gray, K. A., Jones, T. E. M. & Bruford, E. A. Updates to HCOP: the HGNC comparison of orthology predictions tool. *Briefings in Bioinformatics* (2021) doi:10.1093/bib/bbab155.
12. Lal, D. *et al.* Gene family information facilitates variant interpretation and identification of disease-associated genes in neurodevelopmental disorders. *Genome Med* **12**, 28 (2020).
13. Chen, W.-H., Zhao, X.-M., van Noort, V. & Bork, P. Human Monogenic Disease Genes Have Frequently Functionally Redundant Paralogs. *PLoS Comput Biol* **9**, e1003073 (2013).
14. Pérez-Palma, E. *et al.* Identification of pathogenic variant enriched regions across genes and gene families. *Genome Res* **30**, 62–71 (2020).
15. Wiel, L., Venselaar, H., Veltman, J. A., Vriend, G. & Gilissen, C. Aggregation of population-based genetic variation over protein domain homologues and its potential use in genetic diagnostics. *Hum Mutat* **38**, 1454–1463 (2017).
16. Ware, J. S., Walsh, R., Cunningham, F., Birney, E. & Cook, S. A. Paralogous annotation of disease-causing variants in long QT syndrome genes. *Hum Mutat* **33**, 1188–1191 (2012).
17. Brunklaus, A. *et al.* Gene variant effects across sodium channelopathies predict function and guide precision therapy. *Brain* awac006 (2022) doi:10.1093/brain/awac006.
18. Walsh, R., Peters, N. S., Cook, S. A. & Ware, J. S. Paralogue annotation identifies novel pathogenic variants in patients with Brugada syndrome and catecholaminergic polymorphic ventricular tachycardia. *J Med Genet* **51**, 35–44 (2014).
19. Kinsella, R. J. *et al.* Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database (Oxford)* **2011**, bar030 (2011).
20. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792–1797 (2004).
21. Tavtigian, S. V. *et al.* Modeling the ACMG/AMP Variant Classification Guidelines as a Bayesian Classification Framework. *Genet Med* **20**, 1054–1060 (2018).

22. Pejaver, V. *et al.* Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. *Am J Hum Genet* **109**, 2163–2177 (2022).
23. Stenson, P. D. *et al.* Human Gene Mutation Database (HGMD): 2003 update. *Hum. Mutat.* **21**, 577–581 (2003).
24. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
25. Ioannidis, N. M. *et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am J Hum Genet* **99**, 877–885 (2016).
26. Pejaver, V. *et al.* Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. *Nature Communications* **11**, 5918 (2020).
27. Samocha, K. E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46**, 944–950 (2014).
28. Gelb, B. D. *et al.* ClinGen’s RASopathy Expert Panel consensus methods for variant interpretation. *Genet Med* **20**, 1334–1345 (2018).
29. Kelly, M. A. *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* **20**, 351–359 (2018).
30. Frazer, J. *et al.* Disease variant prediction with deep generative models of evolutionary data. *Nature* **599**, 91–95 (2021).
31. Patel, M. J. *et al.* Disease-specific ACMG/AMP guidelines improve sequence variant interpretation for hearing loss. *Genet Med* **23**, 2208–2212 (2021).
32. Hsieh, T.-C. *et al.* GestaltMatcher facilitates rare disease matching using facial phenotype descriptors. *Nat Genet* **54**, 349–357 (2022).
33. Fortuno, C. *et al.* Specifications of the ACMG/AMP variant interpretation guidelines for germline TP53 variants. *Human Mutation* **42**, 223–236 (2021).

34. Foreman, J. *et al.* DECIPHER: Supporting the interpretation and sharing of rare disease phenotype-linked variant data to advance diagnosis and research. *Human Mutation* **43**, 682–697 (2022).
35. Nicora, G., Zucca, S., Limongelli, I., Bellazzi, R. & Magni, P. A machine learning approach based on ACMG/AMP guidelines for genomic variant classification and prioritization. *Sci Rep* **12**, 2517 (2022).
36. Shah, N. *et al.* Identification of Misclassified ClinVar Variants via Disease Population Prevalence. *Am J Hum Genet* **102**, 609–619 (2018).
37. Tokheim, C. *et al.* Exome-scale discovery of hotspot mutation regions in human cancer using 3D protein structure. *Cancer Res* **76**, 3719–3731 (2016).
38. Quinodoz, M. *et al.* Analysis of missense variants in the human genome reveals widespread gene-specific clustering and improves prediction of pathogenicity. *Am J Hum Genet* **109**, 457–470 (2022).

Figure legends

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 alignment 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). D) Barplot showing the number of genes where pathogenic variants with the same (Para-PS1, purple) or different amino acid exchange (para-PM5 pink) provide supporting-strong evidence for variant pathogenicity. *Following the Bayesian framework proposed by Tavtigian et al., 2018.

Figure 3: Comparison to established gene family-based methods. A) The forest plot illustrates the enrichment of pathogenic versus control variants applying para-PS1 for residues with similar paralog conservation levels, as defined in Lal et al., 2020¹². 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., 2019¹⁴) 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 Supplementary Methods for details).

Figure 4: Case examples to demonstrate the utility of the para-PS1/PM5 criteria. Upper panel: Case 1, a boy with bilateral sensory hearing loss and a heterozygous variant in *COL11A1* that is initially classified as a variant of uncertain significance (VUS). By considering a pathogenic paralogous variant found in *COL3A1* and *COL1A2* and consequently applying the para-PS1 criterium (moderate evidence) the variant could be reclassified to likely-pathogenic (more details in the results section). Lower panel: Case 2, a woman in her 40's with syncopal episodes, a recent diagnosis of catecholaminergic polymorphic ventricular tachycardia and a heterozygous variant in *RYR2* gene, NM_001035.3:c.1870G>A (p.Ala624Thr) that is initially classified as VUS. By considering a pathogenic paralogous variant with the same amino acid exchange found in *RYR1* we could apply the para-PS1 criterium and reclassify the variant to likely-pathogenic. Abbreviations: B = benign, LB = likely-benign, VUS = variant of uncertain significance, LP = likely-pathogenic, P = Pathogenic.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used CHAT-GTP-4 to improve readability and language. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Data availability

Data is available in the Supplementary Tables.

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 Famille 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 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.

Ethics Declaration

This study was approved by Cleveland Clinic IRB, approval ID 22-147. Informed consent was waived considering the retrospective nature of the study.

Conflict of Interest

The authors report no conflicts of interests.

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

Current

	PS1/PM5									
Target gene	A	-	G	F	E	Q	G	F	T	R
→ PS1/PM5 applicable for 1/10 cases										

Current criteria

Pathogenic missense variant in same gene &

PS1: The same amino acid exchange

PM5: A different amino acid exchange

Method

Proposed extension

	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/para-PM5 applicable for 3/10 cases										

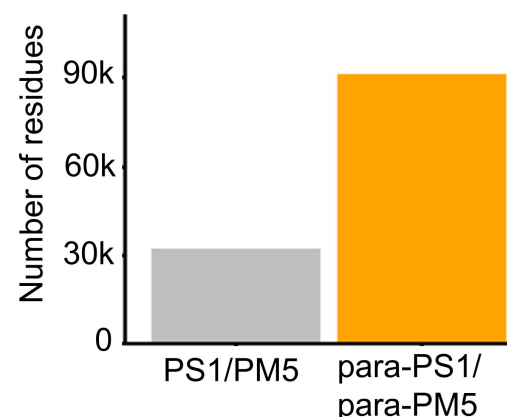
Extended Criteria

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

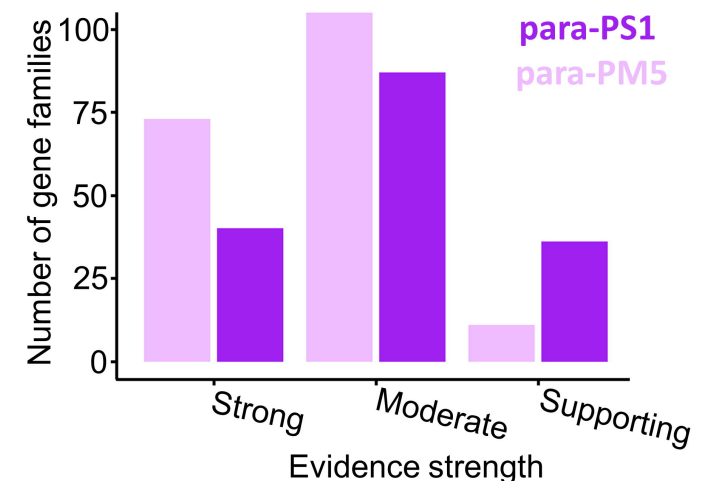
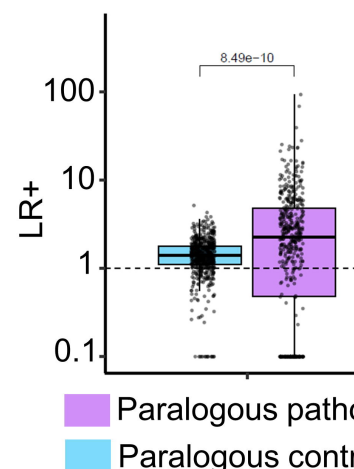
para-PS1: The same amino acid exchange

para-PM5: A different amino acid exchange

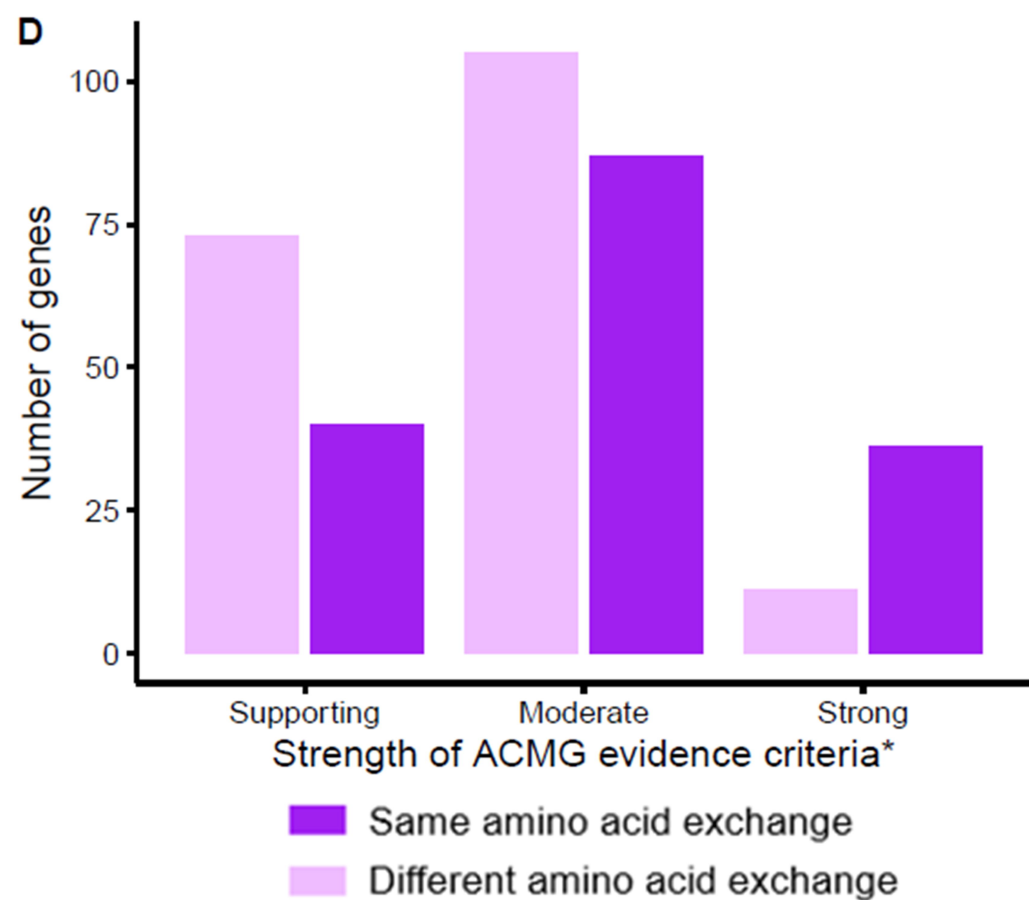
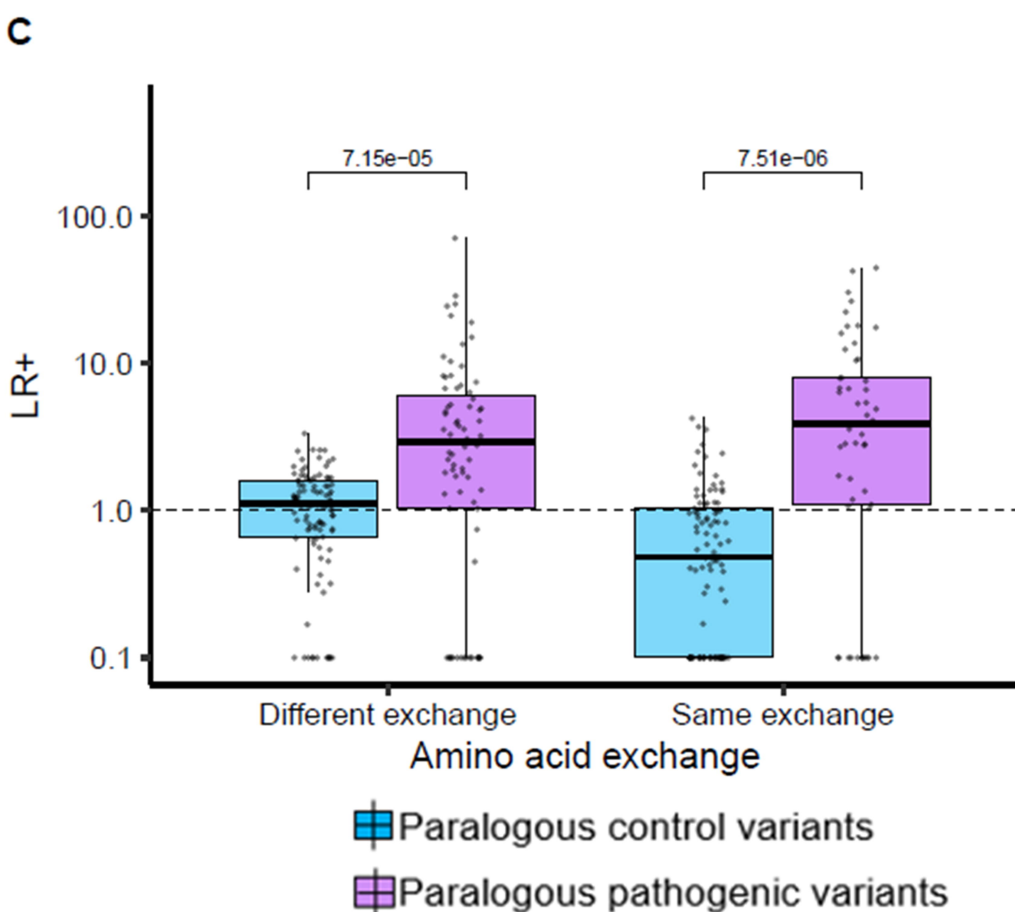
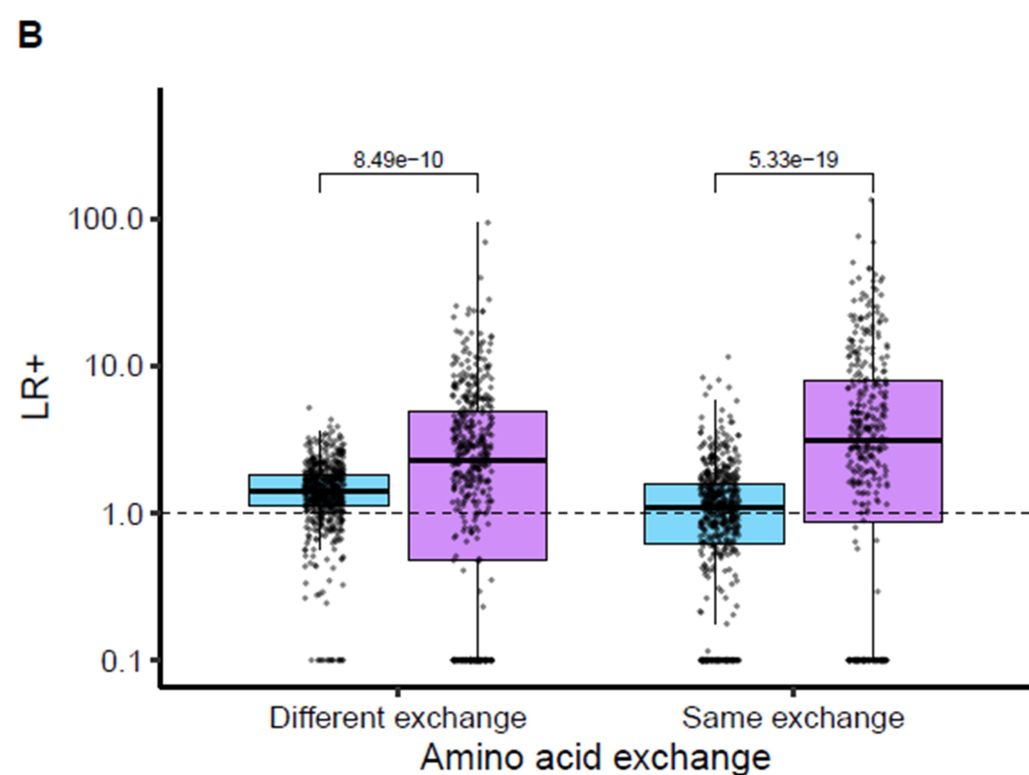
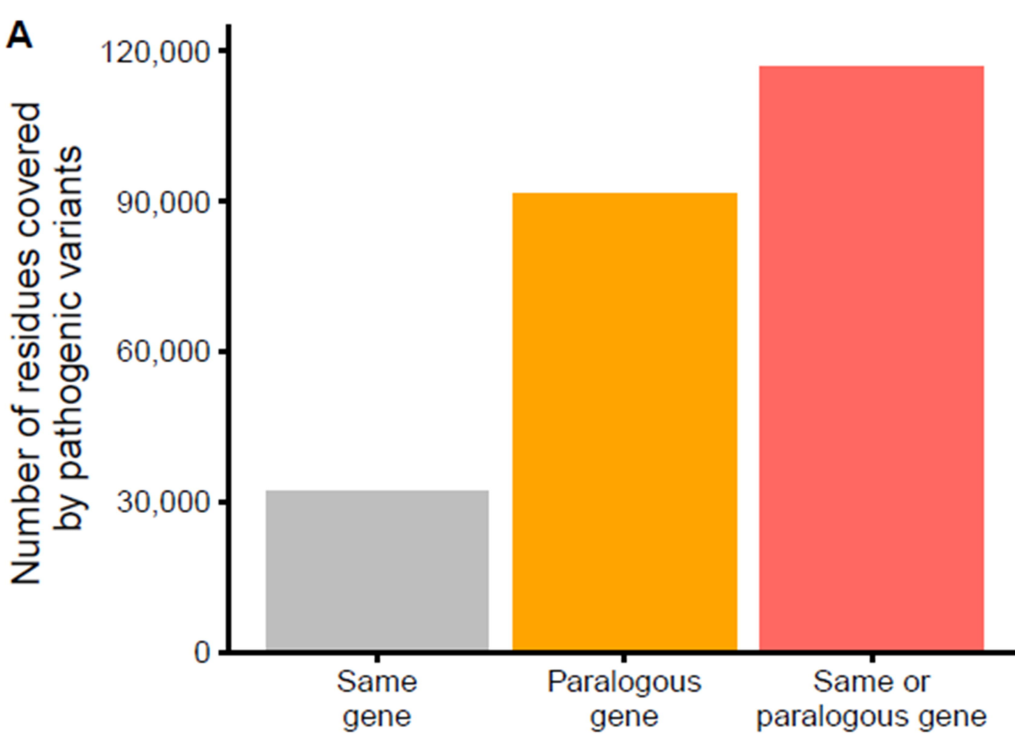
PS1/PM5 extension increase applicability of criteria

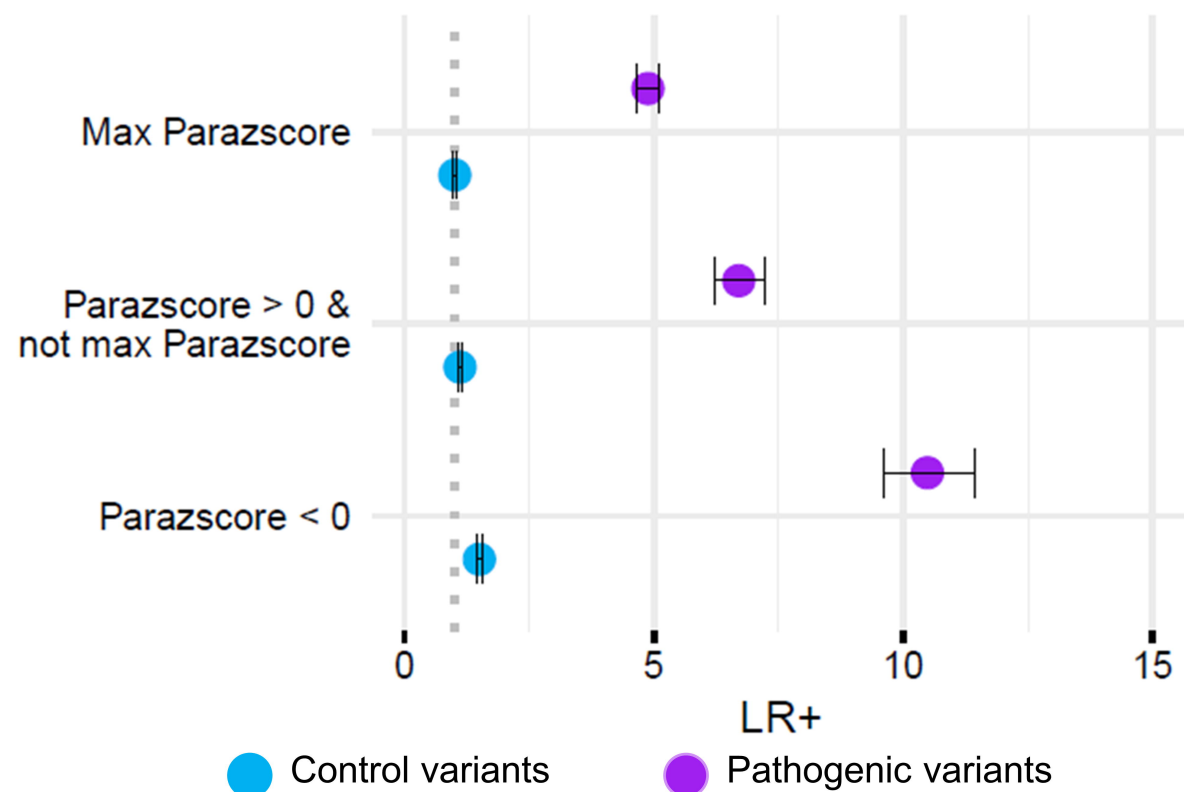
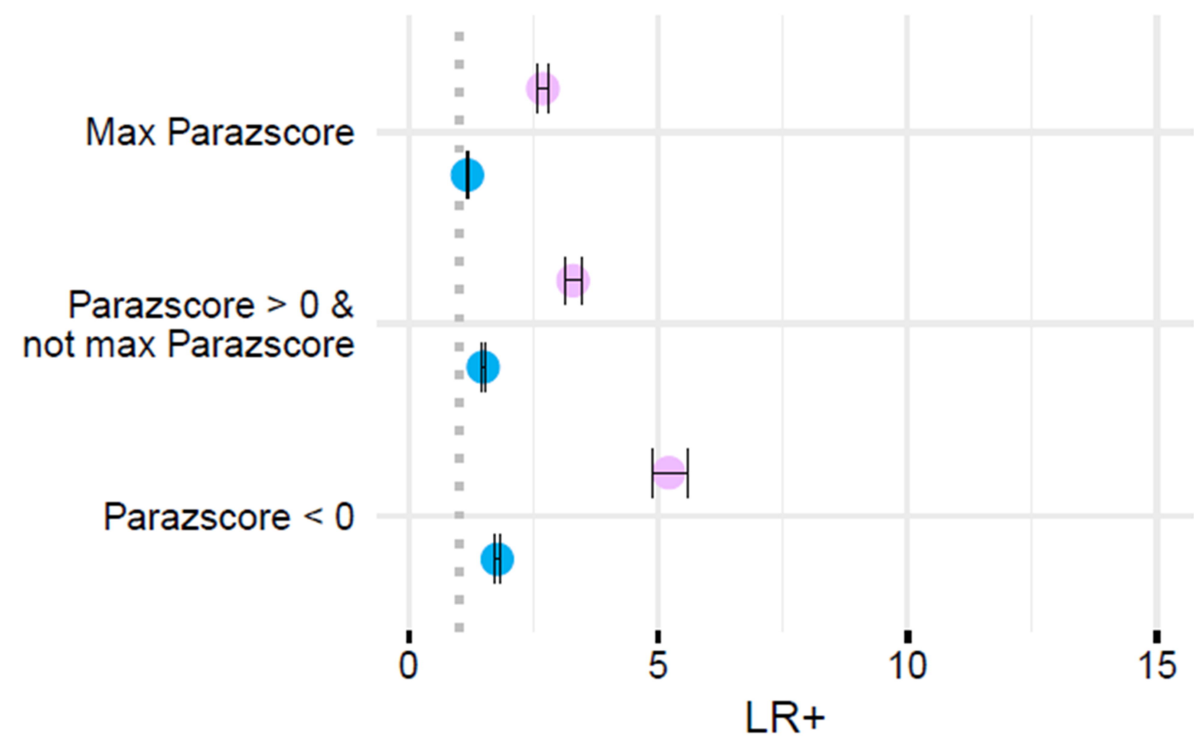
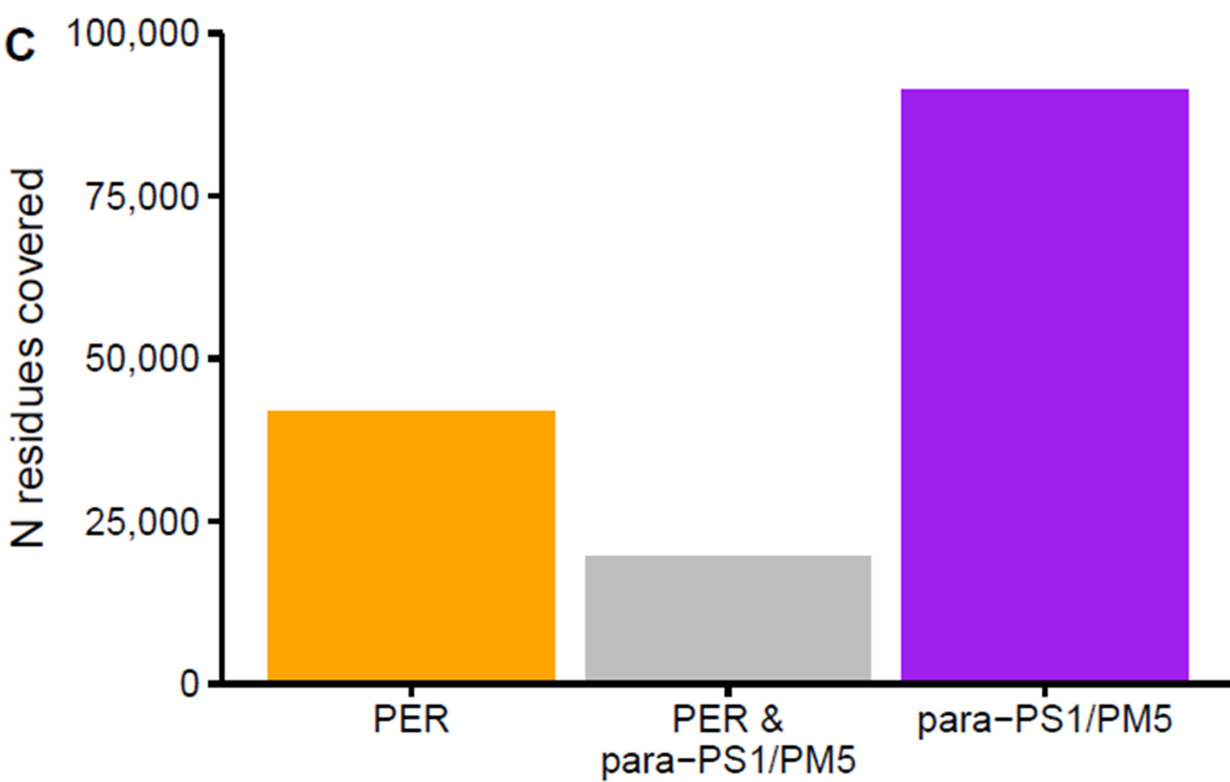


Para-PS1/Para-PM5 provide evidence for variant pathogenicity



Results



A**B****C****D**