Genetic variant pathogenicity prediction trained using large-scale disease specific clinical sequencing datasets

Short title: Clinical sequencing datasets enable disease specific variant pathogenicity prediction

Perry Evans[1], Chao Wu[2], Amanda Lindy[3], Dianalee A. McKnight[3], Matthew Lebo[4,5], Mahdi Sarmady[2,6], Ahmad N. Abou Tayoun[2,6,7]*

[1]Department of Biomedical and Health Informatics, The Children's Hospital of Philadelphia, PA, USA
[2]Division of Genomic Diagnostics, The Children's Hospital of Philadelphia, PA, USA
[3]GeneDx, Gaithersburg, MD, USA
[4]Laboratory for Molecular Medicine, Partners Healthcare Personalized Medicine, MA, USA
[5]Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, MA, USA
[6]Department of Pathology and Laboratory Medicine, The University of Pennsylvania Perelman School of Medicine, PA, USA
[7]Al Jalila Children's Specialty Hospital, Dubai, UAE
*Correspondence to: Ahmad N. Abou Tayoun (Ahmad.Tayoun@ajch.ae)

Abstract

Background
Recent advances in high-throughput DNA sequencing technologies have expanded our understanding of the molecular underpinnings of various genetic disorders and have led to increased utilization of genomic tests by clinicians. However, each test can generate thousands of variants, and given the paucity of functional studies assessing each one of them, experimental validation of a variant's clinical significance is not feasible for clinical laboratories. Therefore, many variants are reported as variants of unknown clinical significance due to this gap. However, the creation of large variant databases like the Genome Aggregation Database has significantly improved the interpretation of novel variants. Specifically, pathogenicity prediction for novel missense variants can now utilize features describing regional variant constraint. Constrained genomic regions are those that have an unusually low variant counts in the general population. Earlier pathogenicity classifiers tried to capture these regions using protein domains.

Methods and Findings
Here we introduce one of the largest variant datasets derived from clinical sequencing panels to assess the utility of using old and new concepts of regional features as pathogenicity scores. This dataset is compiled from 17,071 patients surveyed with clinical genomic sequencing for cardiomyopathy, epilepsy, or rasopathies. We use this dataset to justify the necessity of disease specific classifiers, and train PathoPredictor, a disease specific ensemble classifier of pathogenicity based on regional constraint and variant level features.

Conclusion
Disease specific features improve missense variant pathogenicity prediction. As such, PathoPredictor achieves an average precision greater than 90% for variants from all 112 tested disease genes while approaching 100% accuracy for some genes, making it superior to existing generic pathogenicity metrics it uses as features.

Introduction

Comprehensive sequencing has become the cornerstone of genomic medicine and research. However, unlike previous targeted or single gene testing, multigene sequencing can yield thousands of rare variants often requiring manual clinical correlation and interpretation. Unlike synonymous (or silent) and loss of function (mainly nonsense, frameshift, and canonical splice site) variants where the impact on the protein can be relatively easily predicted, novel missense variants are the the most challenging to interpret, often leading to inconclusive genomic reports and leaving clinicians and families with daunting uncertainties and anxieties. On the other hand, researchers are currently incapable of studying the impact of every possible missense variant in the ~20,000 genes or so of the human genome. Therefore, novel clinical-grade approaches are needed to assist clinicians and researchers in determining the pathogenicity of missense variants.

Machine learning has yielded several pathogenicity prediction tools built with variant features and previously assigned pathogenic and benign labels. Collections of variant labels available for classifier training and testing, like the Human Gene Mutation Database (HGMD) [1] and ClinVar [2], are rare, and often require augmentation using frequently occurring variants from databases like the Genome Aggregation Database (gnomAD) [3] as a substitute for benign variants. Variant features can describe single positions, like genomic sequence context and amino acid conservation, or regions that contain the variant, like protein domains and variation constraint.

Two uses of simple region features are seen in the Functional Analysis through Hidden Markov Models (FATHMM) [4] and the variant effect scoring tool (VEST) [5]. FATHMM and VEST were found to be the most important features for determining pathogenicity in an ensemble of 18 prediction scores called REVEL [6]. VEST distinguished disease missense variants in HGMD from high frequency (allele frequency >1%) missense variants from the Exome Sequencing Project (ESP) [7] using a random forest with 86 features from the SNVBox database [8]. These features describe amino acid substitutions, regional amino acid composition, conservation scores, local protein structure, and annotations of functional protein sites. FATHMM scored variants by their conservation in homologous sequences, weighted by the tolerance of each variant's protein family (Pfam) domain or SUPERFAMILY [9] to mutations observed in HGMD and the set of functionally neutral UniProt variants [10]. VEST's inclusion of functional protein sites, and FATHMM's Pfam domain tolerance consideration enabled them to capture regional protein features such as structure and conservation, but did not capture regional tolerance to genetic variation.

Large variant collections like the Exome Aggregation Consortium (ExAC) dataset [3] and gnomAD have enabled characterization of more sophisticated region attributes. Features that describe genomic regions as constrained (having less population variation than expected), or unconstrained (having more variation than expected). With this knowledge, classifiers might flag a variant as pathogenic if it lies within a genomic region that selects against variants [11]. One such constraint metric is the constrained coding region (CCR) percentile, which compared observed variant counts to those predicted by CpG density [12]. A similar feature called missense depletion was constructed for the MPC (for missense badness, PolyPhen-2, and constraint) pathogenicity classifier of *de novo* missense variants [13]. MPC's missense depletion feature was measured as the fraction of expected ExAC variation that was observed in exons. Only ExAC variants with minor allele frequencies below 0.01% were considered. The expected rate of rare missense variants was based on a model that utilized both gene and sequence context specific mutation rates [14]. An additional pathogenicity feature introduced by MPC was missense badness, which accounted for an amino acid substitution's increase in deleteriousness when it occurs in a missense-constrained region.

Now that a few region based pathogenicity predictors have been suggested, it is useful to evaluate their effectiveness in a clinical setting. We introduce three patient variant training datasets gathered from clinical

sequencing panels for cardiomyopathy, epilepsy, and rasopathies. These datasets cover 17,071 patients. All variants have been manually classified by two main clinical laboratories, whose members significantly contributed to the development of the American College of Medical Genetics and Genomics (ACMG) sequence variant interpretation guidelines [15]. We use each dataset to compare CCR, FATHMM, missense badness, missense depletion, and VEST, and to train disease specific predictors. These clinical variant sets have the advantage of being consistently reviewed in a clinically sound manner, and they originate from focused disease studies. This allows us to explore the hypothesis that disease specific classifiers are better than general genome wide classifiers. We also introduce PathoPredictor, a superior disease specific pathogenicity score trained with clinical sequencing panel variants to combine the pathogenicity scores compared here.

Materials and methods

Classifier

We used Python's scikit-learn machine learning library to train a logistic regression model to predict the pathogenicity of missense variants from clinical panels. Variants were classified by two well known clinical laboratories, GeneDx (Gaithersburg, MD) and the Laboratory for Molecular Medicine (Harvard Medical School, MA), using variant interpretation protocols that are well within the most recent 2015 ACMG guidelines. Pathogenic and likely pathogenic variants were assigned values of one, while benign and likely benign variants were assigned values of zero. During training, we used L2 regularization with a regularization strength of one. Our model included five features corresponding to measures of pathogenicity, one Pfam domain indicator, and all pairwise combinations of features. All model terms were standardized by removing the mean and scaling by the standard deviation with scikit-learn.

Pathogenicity score importance

The importance of each pathogenicity score as a feature in an ensemble classifier was determined using scikit-learn's ExtraTreesClassifier with 250 estimators, and the Gini impurity index as a node splitting criterion. Scikit-learn defines feature importance as the as the total decrease in node impurity, weighted by the proportion of variants reaching that node, and averaged over all 250 trees of the ensemble. Feature importances were normalized so that they sum to one.

Pathogenicity scores as classifier features

We used six features in our classifier: CCR, FATHMM, missense depletion, missense badness, VEST, and Pfam protein domains. FATHMM and VEST can provide multiple scores for one variant, depending on isoforms. VEST scores were taken as the minimum VEST v3.0 score provided by dbNSFP v2.9 [16]. FATHMM scores were taken as the negative minimum FATHMM v2.3 score provided by dbNSFP v2.9. FATHMM scores were negated so that their interpretation would match the other features. CCR scores were taken as the CCR percentile (ccr_pct) from the CCR bed file v1.20171112 [12]. Missense depletion and badness scores were taken from the constraint MPC vcf file v2 as obs_exp and mis_badness, respectively [13]. A bed file of PFam domain locations was downloaded from the UCSC genome browser. We assigned each variant a Pfam score of one if its position fell within a domain, and zero otherwise. Feature values were assigned to variants as described below in the variant annotation and filtering pipeline.

For each disease dataset we used Python's scikit-learn library to standardize each feature by removing its mean and scaling by its standard deviation. Panel and ClinVar variants for the same disease were processed together so that their features would be on the same scale.

Variant sets

GeneDx provided clinical sequencing panel results for epilepsy, while the Laboratory for Molecular Medicine provided their clinically curated data for cardiomyopathy and rasopathies. The number of patients investigated differed by gene. The maximum number of patients observed was 5466 for cardiomyopathy, 8583 for epilepsy, and 3022 for rasopathies. No gene was shared between the three datasets. Variants were provided in Human Genome Variation Society (HGVS) c. notation [17], and were converted to vcf files of hg19 based variants using Mutalyzer [18] and custom scripts. To construct variant sets for our classifier, we discarded variants of uncertain significance. We formed a benign set of variants using “benign” and “likely benign” variants. Similarly, our pathogenic variant set consisted of “pathogenic” and “likely pathogenic” variants. These labeled variants were used for training and leave-one-gene-out cross validation (see below). Diseases panel variants and labels were deposited into S1 Table.

Variants from ClinVar were chosen as a test variant set. We restricted ClinVar genes to those found in the disease panels, and removed any ClinVar variant found in the disease panels, producing an independent variant set. The hg19 ClinVar vcf file was downloaded on February 25, 2018, and limited to unambiguously pathogenic or likely pathogenic and benign or likely benign variants with no conflicts according to CLINSIG. We considered ClinVar variants with any review status as one test set, and consulted CLNREVSTAT to produce a second ClinVar test variant set restricted to reviewed variants.

Disease specific classifier evaluation

We compared estimated pathogenicity probabilities produced by our trained models and each pathogenicity score used as a model feature via average precision, as implemented in scikit-learn. Average precision is useful here due to the possibility of imbalances between pathogenic and benign variant counts. This measure is an approximation of the area under the method’s precision-recall curve. Average precision comparisons were conducted in both a leave one gene out manner, and on an independent set of ClinVar variants. We iterated over all genes in a disease, holding out the gene of interest, and training a model using variants from all remaining genes. This model was applied to disease panel variants from the gene of interest (cross validation), as well as to the gene’s ClinVar variant set (independent test). ClinVar variant datasets were restricted to variants not found in the disease panel results. Precision-recall curves were made for each pathogenicity score, and used to determine average precision.

The DeLong test as implemented in R’s pROC package [19] was used to compare areas under receiver operating characteristics curves produced by predictors. We used this test to gauge the significance of differences between classifiers.

Comparison with REVEL

As in the CCR paper, we compared PathoPredictor and REVEL using *de novo* missense variants from 5,620 neurodevelopmental disorder patients and 2,078 unaffected siblings of autism spectrum disorder patients [12,13]. *De novo* variants from patients were considered pathogenic, and *de novo* variants from unaffected siblings were considered benign. HGVS formatted variants were uploaded to VariantValidator [20], and a vcf file was constructed from the results. This file was normalized with vt [21]. We applied our epilepsy variant trained PathoPredictor (using all or dominant epilepsy genes) to *de novo* variants located in genes from the epilepsy panel. For the evaluation, we used 66 pathogenic and 6 benign variants for the epilepsy specific PathoPredictor. 35 pathogenic and 3 benign variants were used to evaluate the dominant epilepsy

PathoPredictor model. REVEL and PathoPredictor scores were compared using the area under their precision-recall curves, and the DeLong test.

Variant annotation and filtering pipeline

Our pipeline began with vcf files containing 345,849 ClinVar and 7,840 disease panel variants labeled as benign, pathogenic, or VUS. Snpeff v4.3.1T [22] was used to determine variant effects in GRCh37.75. SnpSift v4.3 [22] was used to annotate variants with allele frequencies from the ESP, FATHMM scores, and VEST scores from dbNSFP. We annotated variants with values from bed (CCR and Pfam) and vcf (missense badness and depletion and calculated ESP frequencies from ESP6500SI-V2) files using vcfanno v0.2.8 [23]. The ESP frequencies are needed next when remove the training data used for VEST.

We next removed variants that had been used to train FATHMM (~49,500 disease variants and ~37,000 putatively neutral variants) or VEST (~45,000 disease variants and ~45,000 putatively neutral variants), ensuring that these features would not have an advantage when evaluating feature importance, and that our ClinVar test dataset would not overlap with any variants used for training. Both FATHMM and VEST were trained with damaging mutations from HGMD, but they differed in their choice of neutral missense variant set. FATHMM was trained with neutral variants from UniProt [10] and VEST was trained with missense variants from ESP achieving an population frequency of 1% or higher.

We then removed variants found in the set of 154,257 DM (damaging mutation) in HGMD Professional 2016.1. To address frequent ESP variants, we took the variant frequency as the maximum of dbNSFP fields ESP6500_EA_AF, ESP6500_AA_AF, and the total ESP allele frequency determined using vcfanno. We discarded variants (68 for ClinVar, and 59 disease panel) when this maximum value reached at least 0.01. To remove neutral UniProt variants, we used “Polymorphism” annotations to build a list of neutral codons relative to hg38. Polymorphism annotations were downloaded from www.uniprot.org/docs/humsavar.txt, and joined with hg38 codon coordinates from UniProt. Both were downloaded on April 5th, 2018. We used liftOver [24] to convert these to hg19, and removed any variant found in any of 921,722 neutral codons.

Final variant sets were taken as missense variants with CCR and missense depletion scores to avoid missing data issues. After applying all the above filters, 671 variants remained to train PathoPredictor, and 2421 ClinVar variants were used as a test set.

Results

Variants and genes studied

We focus on patient variants from three disease panels: cardiomyopathy, epilepsy, and rasopathies (Figure 1). We also investigate the subset of epilepsy dominant genes: *SCN1A*, *SCN2A*, *SCN8A*, *SCN1B*, *KCNQ2*, *KCNQ3*, *CDKL5*, *PCDH19*, *SLC2A1*, *SPTAN1*, *STXBP1*, and *TSC1*. Variants in these genes are relatively more common and, since they follow a dominant inheritance pattern, might have distinct characteristics impacting variant prediction relative to all other epilepsy genes (see below). For each disease variant set, we compare the performance of CCR, FATHMM, missense badness, missense depletion, and VEST using panel and ClinVar variants. We also build PathoPredictor, an ensemble classifier of pathogenicity, and test it with variants from ClinVar that are not found in disease panel genes. To ensure the reliability of ClinVar variants, we examine only unambiguously pathogenic or benign variants, and split ClinVar into two variant groups: all ClinVar variants and those that have been reviewed. Note that few variant collections have an equal amount of pathogenic and benign variants, with a drastic imbalance for cardiomyopathy panel variants.

Feature classification importance and correlations differ by disease

The hypothesis that each disease variant set needs its own classifier is tested by examining feature correlations and relative feature importance by disease. If the relationships between features change by disease, then it is useful to capture these relationships in different disease models.

For each disease panel, we determine the importance of each feature for splitting pathogenic and benign variants in an extra-trees classifier (Figure 2). CCR is consistently the most important feature, and Pfam domain annotations are the least helpful. The importance of VEST is nearly equal to that of CCR for all epilepsy and rasopathy variants, indicating that the single position features from VEST are equally as important as the CCR region feature for these diseases. Interestingly, VEST is not as close to CCR in importance for dominant epilepsy genes as it is when using all epilepsy genes, suggesting an expectedly higher emphasis on constrained regions for this subset of dominant genes. The classifier for the latter genes also makes a greater use of FATHMM. The higher relative importance of missense depletion, and the lower relative importance of VEST and FATHMM for cardiomyopathy suggests that region features are also more important for the cardiomyopathy disease. Finally, rasopathies yield a classifier with more importance placed on missense badness after CCR and VEST.

Highly correlated features can make it difficult to interpret relative feature importance from tree based classifiers, so we also investigated the correlations between features for each disease variant set (Figure 3). CCR and VEST have a Pearson correlation coefficient of 0.65 for epilepsy, but 0.28 and 0.32 for cardiomyopathy and rasopathies, respectively, so it is unlikely that VEST could replace CCR. Viewed along with the feature importances, the feature correlation differences observed between diseases, like those for the missense depletion and FATHMM correlation, further highlight the necessity of disease specific classifiers.

Disease specific classifier evaluation

We use disease panel and ClinVar datasets to compare pathogenicity classifiers, and to train and test PathoPredictor (Figure 4). For each disease, we used panel and ClinVar variants to build precision-recall curves using pathogenicity scores from CCR, FATHMM, missense badness, missense depletion, and VEST. These curves were summarized using average precision. To evaluate PathoPredictor, we examined each disease panel gene, trained a model using disease panel variants not found in the selected gene, and tested the model using panel and ClinVar variants from the gene. By training PathoPredictor with disease specific variants, we collect variants that belong to genes that are more likely to share a common biological pathway, and might behave similarly in terms of tolerance.

PathoPredictor predicts the pathogenicity of disease panel variants with an average precision higher than that obtained with any single feature (Figure 5). This performance is significantly better than most features for all panel variants (15 of the total 20 features in all four panels, $p < 0.03$). Of the five exceptions, three were in the rasopathies panel, which had the lowest variant numbers (Figure 1). When evaluating with more variants for rasopathies and cardiomyopathy, a significant advantage of PathoPredictor over all single features was observed (Figure 6). PathoPredictor was also not significantly better than CCR for the dominant epilepsy genes, where it is expected that regional constraint is most critical (see above). PathoPredictor utilizes CCR the most (Figure 2). A similar relation between feature importance and average precision was seen with the missense depletion feature and cardiomyopathy (see above and Figure 2). The average precision of PathoPredictor is greater than 90% for all disease panels. Consistent with our feature importance analysis, CCR has the highest average precision among the single features.

To get a better assessment of PathoPredictor, its performance was evaluated using a larger independent variant set from ClinVar (Figure 6). For each disease panel, we find that PathoPredictor performs significantly better than any single feature ($p < 0.014$). PathoPredictor has similar performance when using all of ClinVar and the reviewed subset of ClinVar. The poor average precision obtained when using VEST, FATHMM, and missense badness to predict cardiomyopathy panel variants was not replicated using ClinVar variants in cardiomyopathy panel genes. This discrepancy can be attributed to the lower number of cardiomyopathy panel variants, especially pathogenic (Figure 5). PathoPredictor showed consistent results for rasopathy ClinVar and panel variants, however, given the larger number of ClinVar variants, the improved performance of PathoPredictor was now statistically significant in ClinVar ($p < 0.014$).

The performance of PathoPredictor is challenging to assess by gene because most genes have a small variant sample size. However, some genes with large numbers of variants were found to best demonstrate the utility of PathoPredictor (Figure 7a). For example, PathoPredictor has an accuracy of 95% for panel variants and 96% for ClinVar variants for *KCNQ2.*

Comparison with REVEL

REVEL is a state-of-art ensemble classifier of pathogenicity. It was built using more features (18 prediction scores), but does not contain recent genomic constraint features like CCR and missense depletion. REVEL used a complex benign variant dataset as training data, and removing it from our ClinVar variant set would be arduous. Instead, we utilized the set of *de novo* variants used to compare CCR and REVEL. We focused on PathoPredictor epilepsy classifiers, as these were expected to be most relevant to the neurodevelopmental and autism disorder variants from the testing *de novo* dataset. We found that PathoPredictor had a higher average precision than REVEL (Figure 7b), however, additional testing variants are needed to better demonstrate this improvement.

Discussion

We have shown that the efficacy of variant pathogenicity prediction varies by disease, whereby each disease dictates a unique combination of classifier features. We have also presented PathoPredictor, a new missense variant pathogenicity predictor trained with variants from clinical sequencing results to produce pathogenicity scores from disease specific combinations of regional constraint and variant features. PathoPredictor achieves an average precision greater than its components: CCR, FATHMM, missense depletion, missense badness, Pfam domain status, and VEST. FATHMM, Pfam domain status, and VEST capture regional constraint by using domains and protein families, while CCR, missense depletion, and missense badness locate genomic regions with less natural population variants than expected by null models of variation.

CCR was determined to be the most useful feature for classification, replicating results from the CCR manuscript [12]. A variant's presence in a Pfam domain was found to be the worst performing feature. The poor performance of Pfam domains is not surprising. The Pfam domain binary status does not capture the functional differences between domains, and Pfam domains were also utilized by FATHMM and VEST, possibly making the feature redundant. VEST used annotations of functional protein regions, and FATHMM incorporated Pfam domain annotations when deciding how to weight changes in amino acid conservation with respect to domain variant tolerance. The inclusion of Pfam domains in these features might mitigate its usefulness as a standalone feature. Furthermore, one goal of missense depletion was to provide a better alternative to Pfam domains. Missense depletion and badness might capture functionally similar regions better than Pfam domain annotations.

To construct PathoPredictor, we introduced a unique variant dataset derived from clinical panel sequencing results for cardiomyopathy, epilepsy, and rasopathy patients. We also used this variant dataset to show that a feature's importance to classification, and its correlations with other features were observed to change by disease. A benefit of this dataset compared to ClinVar is that the variants are labeled and obtained in a more homogeneous way, which helps remove data biases. Furthermore, the variant labels followed clinical interpretation standards similar to the ACMG guidelines, making the dataset more similar to real world clinical use cases. To avoid any biases in our training and test datasets, we removed all variants previously used to train any of the component features. However, this significantly reduced the number of variants to optimize PathoPreditor, which despite its superior performance, can still improve with more unique training datasets.

We demonstrated the utility of PathoPredictor using missense variants from ClinVar and a variant set of *de novo* variants previously used to compare REVEL and CCR. PathoPredictor performs significantly better than its constituent features when evaluated with ClinVar. Furthermore, PathoPredictor has a higher average precision than REVEL when testing with the *de novo* variants, which is consistent with CCR's improvement over REVEL using this same dataset [12].

In conclusion, we recommend using PathoPredictor scores to predict missense variant pathogenicity for cardiomyopathy, epilepsy, and rasopathies. Source code for this manuscript is available at https://github.com/samesense/pathopredictor. Predictions for all possible missense variants for disease panel genes are located in S2 Table.

Figure legends

Figure 1. Study datasets. Missense variant and gene counts are shown by disease panel and ClinVar variant set. We only use ClinVar variants from panel genes, and consider either any ClinVar variant (Total ClinVar), or ClinVar variants that have been reviewed (ClinVar w/ Evidence). ClinVar pathogenicity assignments are only used for variants with no conflicting labels. Any variant found in both ClinVar and the panel is not considered as a ClinVar variant.

Figure 2. Feature importance by disease. For each disease panel, we trained a six feature extra-trees classifier to predict pathogenicity. The relative importance of each feature in each classifier is shown, with error bars describing the importance variance across 250 trees. While CCR is the best feature across all disease panels, and the Pfam domain feature is the least important feature, other features change importance based on disease.

Figure 3. Feature correlations by disease. For all features except the Pfam domain feature, the Pearson correlation coefficient is plotted for feature pair values observed across panel variants grouped by disease. Like the feature importances, relationships between features also differ by disease.

Figure 4. Method description. Our goal is to build disease specific classifiers of missense variant pathogenicity using variants from clinical panels. For all genes in a disease panel, we trained a model using variants from all other genes except the gene in question, and tested the model using variants from that gene of interest. We then used ClinVar variants from the gene of interest as an independent test set. Test results were summarized as average precision scores.

Figure 5. Disease specific classifier performance using disease panel data. For each disease panel, we used a hold-one-gene-out approach to evaluate a logistic regression model's ability to predict pathogenicity. For all

genes in a disease panel, we trained PathoPredictor using variants from all other genes, and tested the model using variants from the gene of interest. Using the held-out gene variant prediction scores, we summarized the classifier's precision-recall curve as the average precision from all prediction scores, and compared this average precision to those obtained from each feature's precision-recall curve obtained using untransformed scores. The numbers of pathogenic and benign variants investigated are shown at the bottom left of each panel. For epilepsy, PathoPredictor performs significantly better than any single feature ($p < 10^{-4}$), and PathoPredictor only failed to be significantly better in 5 of the 20 total feature comparisons (CCR, VEST, and missense depletion for rasopathies, CCR for limited epilepsy genes, and missense depletion for cardiomyopathy).

Figure 6. Disease specific classifier testing using ClinVar data. For each disease panel, we applied the hold-one-gene-out models to ClinVar variants from the held-out gene to obtain pathogenicity prediction scores. We summarize each disease classifier's precision-recall curve (PathoPredictor) as the average precision from all prediction scores, and compared its average precision to those obtained from each feature's precision-recall curve. We used either all ClinVar variants (Total ClinVar), or ClinVar variants with a review status that included at least one submitter, or an expert panel (Clinvar w/ Evidence). The numbers of pathogenic and benign variants investigated are shown at the bottom left of each panel. PathoPredictor performs significantly better than any single feature ($p < 0.014$).

Figure 7. PathoPredictor performance. a) ROC curves are shown for select genes evaluated during cross-validation with the disease panel dataset and testing with ClinVar variants. The curve for *RAF1* closely follows, and is obscured by that of *SCN2A*. PathoPredictor has an accuracy of 95% for panel variants and 96% for ClinVar variants for *KCNQ2*. b) PathoPredictor epilepsy specific classifiers were compared to REVEL. *De novo* missense variants from neurodevelopmental disorder patients were used as pathogenic variants (66 for total epilepsy genes, and 35 for dominant). Variants from unaffected siblings of autism spectrum disorder patients were used as benign variants (6 for total epilepsy genes, and 3 for dominant). For each comparison, variants were limited to genes from PathoPredictor's training set, but variants were filtered using the same methods applied to ClinVar variants. We summarized each classifier's precision-recall curve as the average precision. PathoPredictor achieved a greater average precision than REVEL, but this improvement was not found to be significant.

Supporting information

S1 Table. Missense disease panel variants. All missense clinical sequencing variants from cardiomyopathy, epilepsy, and rasopathies are included with hg19 coordinates. Each variant has an associated disease and pathogenicity (V for VUS, B for benign, and P for pathogenic).

S2 Table. Disease panel gene predictions. Every possible missense variant for disease panel genes is listed with predictions from each disease panel classifier. The pathopredictor_class column holds the predicted pathogenicity (1 for pathogenic, and 0 for benign). The pathopredictor_score column holds the prediction score yielded by the model. This score ranges from 0 (most benign) to 1 (most pathogenic).

Acknowledgments
We thank all members of GeneDx and the Laboratory for Molecular Medicine (LMM) who manually curated and classified the disease variants used in our study.

References

1. Stenson PD, Mort M, Ball EV, Howells K, Phillips AD, Thomas NS, et al. The Human Gene Mutation Database: 2008 update. Genome Med. 2009;1: 13.

2. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 2016;44: D862–8.

3. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536: 285–291.

4. Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GLA, Edwards KJ, et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum Mutat. 2013;34: 57–65.

5. Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics. 2013;14 Suppl 3: S3.

6. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet. 2016;99: 877–885.

7. Project NES, Others. Exome variant server, NHLBI GO exome sequencing project (ESP), Seattle, WA. 2013.

8. Wong WC, Kim D, Carter H, Diekhans M, Ryan MC, Karchin R. CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer. Bioinformatics. 2011;27: 2147–2148.

9. Gough J, Karplus K, Hughey R, Chothia C. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J Mol Biol. 2001;313: 903–919.

10. The UniProt Consortium. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2017;45: D158–D169.

11. Amr SS, Al Turki SH, Lebo M, Sarmady M, Rehm HL, Abou Tayoun AN. Using large sequencing data sets to refine intragenic disease regions and prioritize clinical variant interpretation. Genet Med. 2017;19: 496–504.

12. bioRxiv; doi:10.1101/220814

13. Samocha KE, Kosmicki JA, Karczewski KJ, O'Donnell-Luria AH, Pierce-Hoffman E, MacArthur DG, et al. Regional missense constraint improves variant deleteriousness prediction [Internet]. bioRxiv. 2017. p. 148353. doi:10.1101/148353

14. Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM, et al. A framework for the interpretation of de novo mutation in human disease. Nat Genet. 2014;46: 944–950.

15. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;17: 405–424.

16. Liu X, Jian X, Boerwinkle E. dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. Hum Mutat. 2013;34: E2393–402.

17. Dunnen JT, Dalgleish R, Maglott DR, Hart RK, Greenblatt MS, McGowan-Jordan J, et al. HGVS recommendations for the description of sequence variants: 2016 update. Hum Mutat. Wiley Online Library; 2016;37: 564–569.

18. Wildeman M, van Ophuizen E, den Dunnen JT, Taschner PEM. Improving sequence variant descriptions

in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. Hum Mutat. 2008;29: 6–13.

19. Robin X, Turck N, Hainard A, Sanchez JC, Muller M, Others. Package “pROC”: Display and analyze ROC curves. Version 1.5. 4. BMC Bioinformatics. 2012;12: 77.

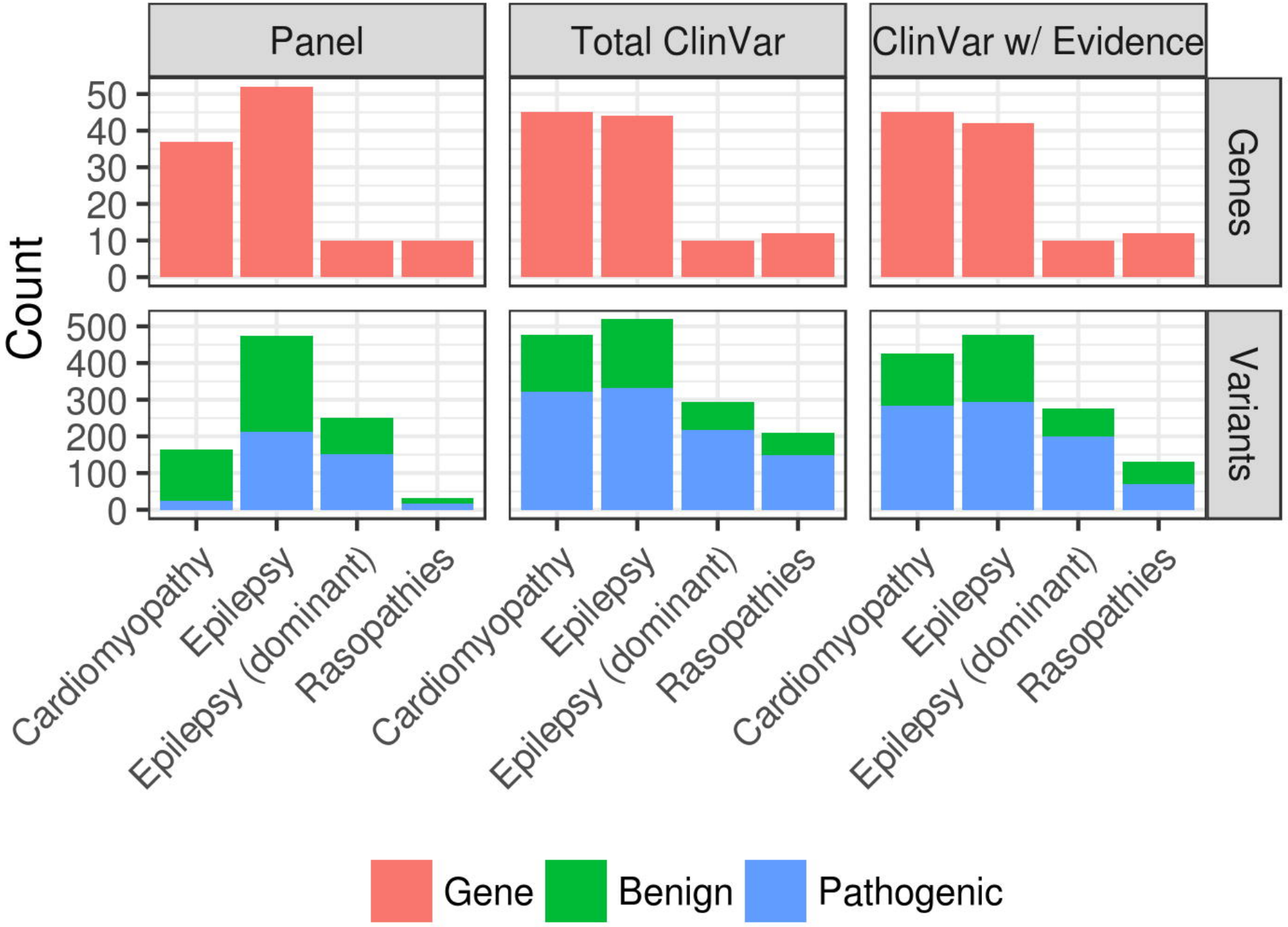
20. Freeman PJ, Hart RK, Gretton LJ, Brookes AJ, Dalgleish R. VariantValidator: Accurate validation, mapping, and formatting of sequence variation descriptions. Hum Mutat. 2018;39: 61–68.

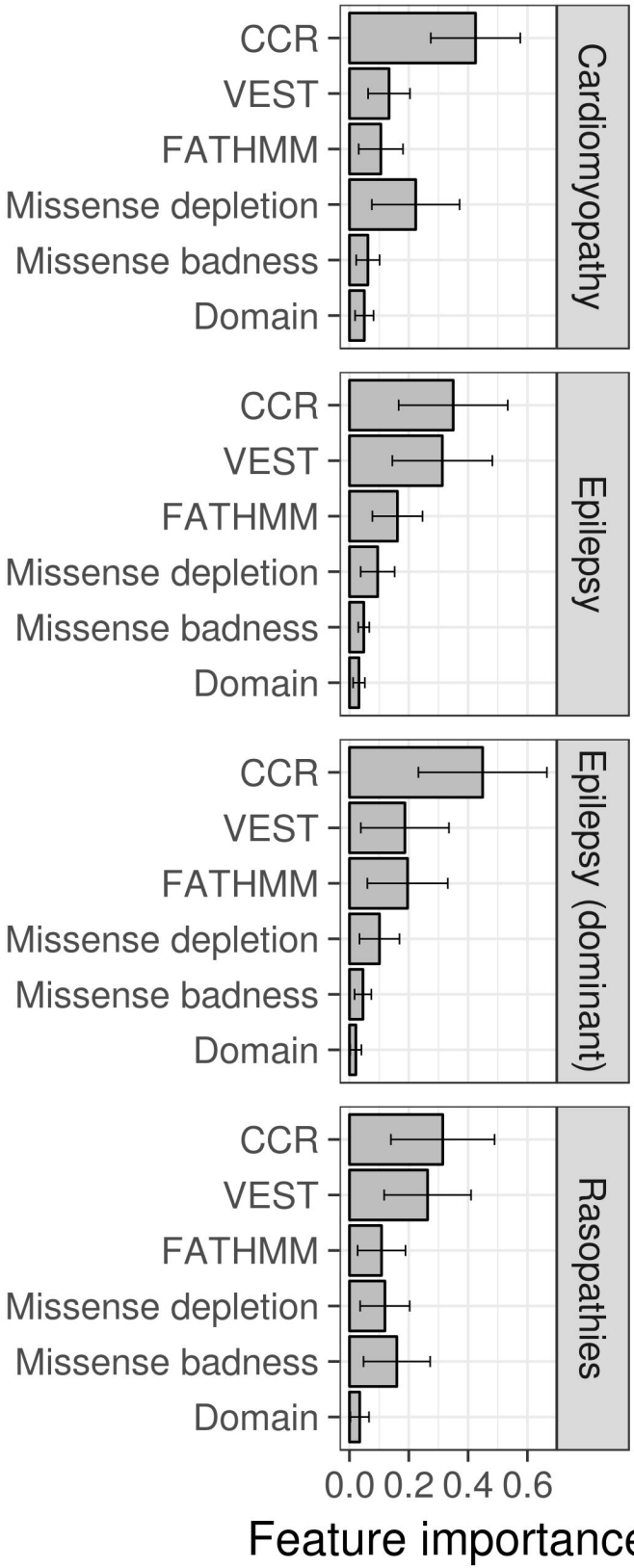
21. Tan A, Abecasis GR, Kang HM. Unified representation of genetic variants. Bioinformatics. 2015;31: 2202–2204.

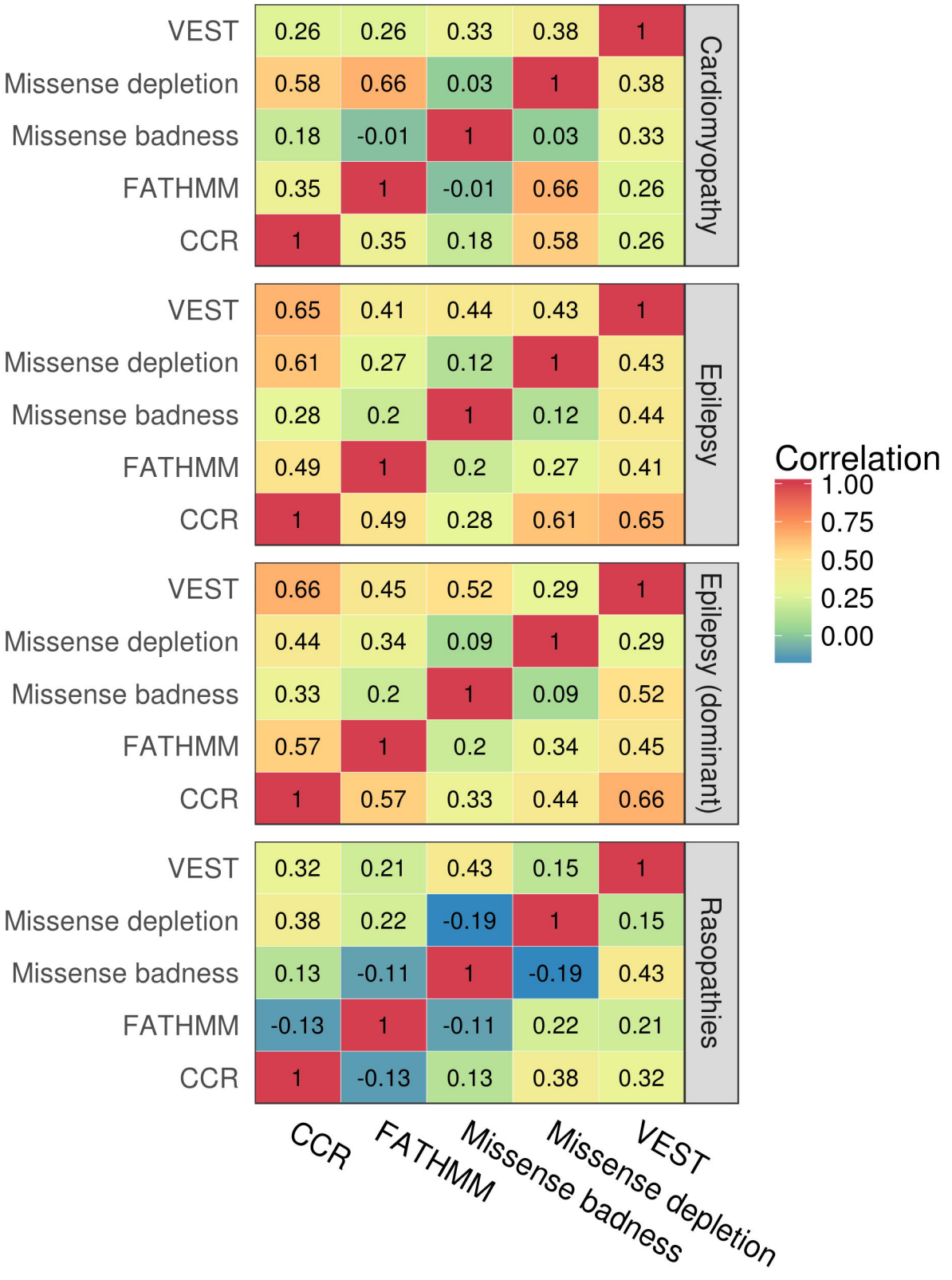
22. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly . 2012;6: 80–92.

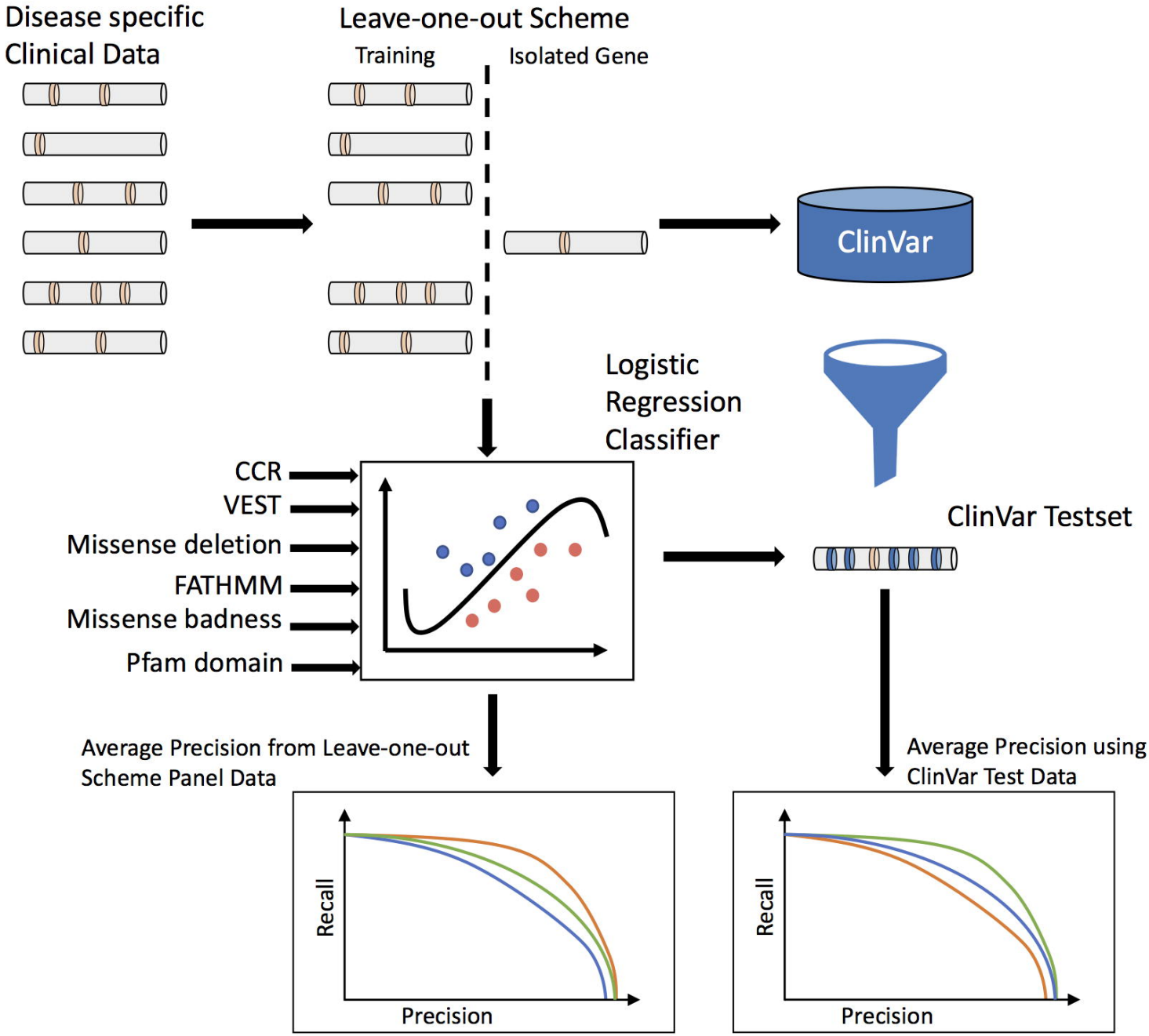
23. Pedersen BS, Layer RM, Quinlan AR. Vcfanno: fast, flexible annotation of genetic variants. Genome Biol. 2016;17: 118.

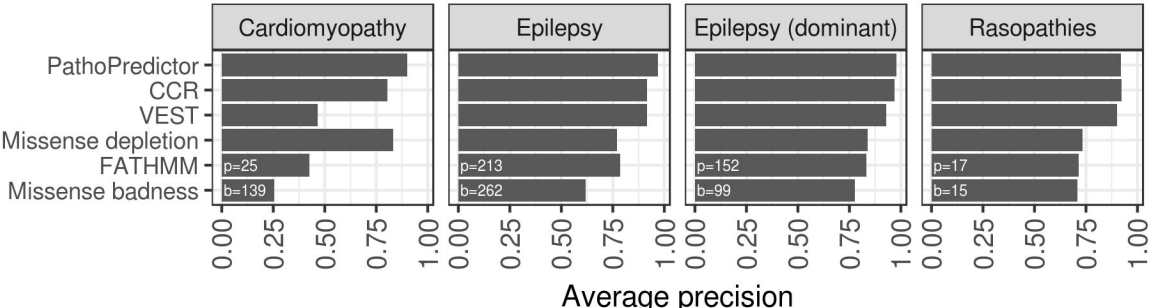
24. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, et al. The UCSC genome browser database: update 2006. Nucleic Acids Res. Oxford University Press; 2006;34: D590–D598.

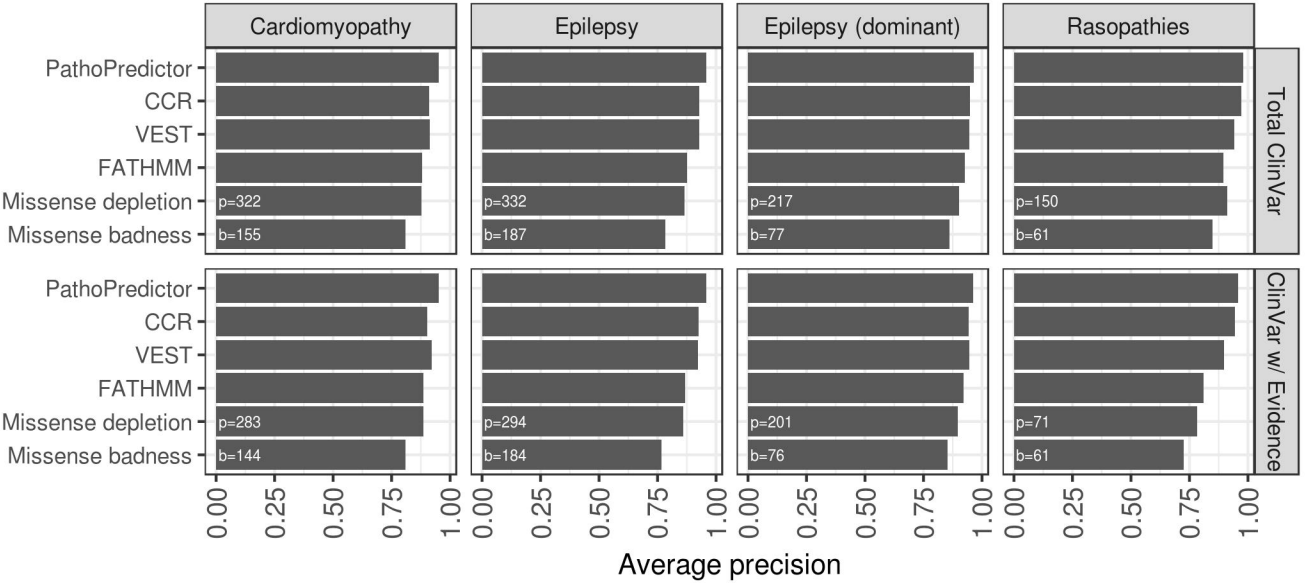

Panel
Total ClinVar
ClinVar w/ Evidence
Genes
Variants
Count
50
40
30
20
10
0
500
400
300
200
100
0
Cardiomyopathy
Epilepsy
Epilepsy (dominant)
Rasopathies
Gene
Benign
Pathogenic

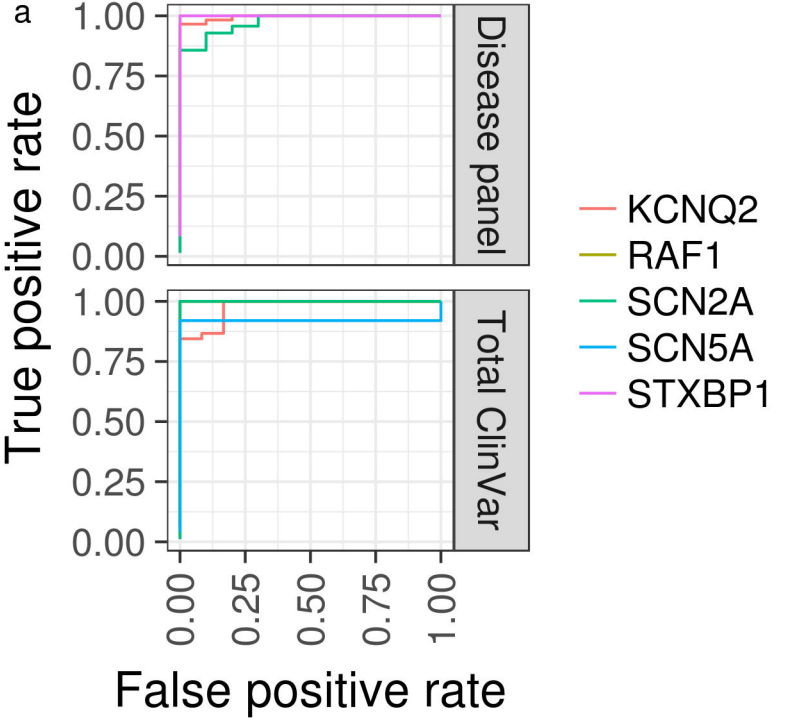

Cardiomyopathy
Epilepsy
Epilepsy (dominant)
Rasopathies
CCR
VEST
FATHMM
Missense depletion
Missense badness
Domain
0.0
0.2
0.4
0.6
Feature importance

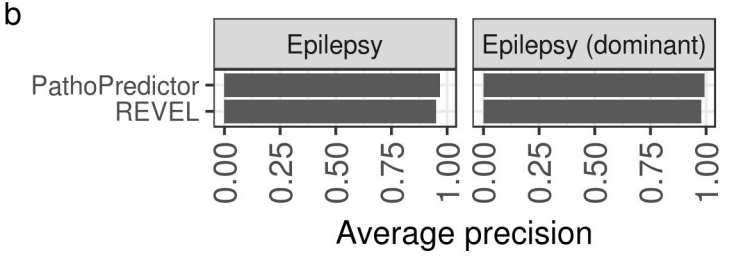

Cardiomyopathy
VEST 0.26 0.26 0.33 0.38 1
Missense depletion 0.58 0.66 0.03 1 0.38
Missense badness 0.18 -0.01 1 0.03 0.33
FATHMM 0.35 1 -0.01 0.66 0.26
CCR 1 0.35 0.18 0.58 0.26
Epilepsy
VEST 0.65 0.41 0.44 0.43 1
Missense depletion 0.61 0.27 0.12 1 0.43
Missense badness 0.28 0.2 1 0.12 0.44
FATHMM 0.49 1 0.2 0.27 0.41
CCR 1 0.49 0.28 0.61 0.65
Epilepsy (dominant)
VEST 0.66 0.45 0.52 0.29 1
Missense depletion 0.44 0.34 0.09 1 0.29
Missense badness 0.33 0.2 1 0.09 0.52
FATHMM 0.57 1 0.2 0.34 0.45
CCR 1 0.57 0.33 0.44 0.66
Rasopathies
VEST 0.32 0.21 0.43 0.15 1
Missense depletion 0.38 0.22 -0.19 1 0.15
Missense badness 0.13 -0.11 1 -0.19 0.43
FATHMM -0.13 1 -0.11 0.22 0.21
CCR 1 -0.13 0.13 0.38 0.32
CCR
FATHMM
Missense badness
Missense depletion
VEST
Correlation
1.00
0.75
0.50
0.25
0.00


Disease specific
Clinical Data
Leave-one-out Scheme
Training
Isolated Gene
ClinVar
Logistic
Regression
Classifier
CCR
VEST
Missense deletion
FATHMM
Missense badness
Pfam domain
ClinVar Testset
Average Precision from Leave-one-out
Scheme Panel Data
Average Precision using
ClinVar Test Data
Recall
Precision
Recall
Precision


Cardiomyopathy
Epilepsy
Epilepsy (dominant)
Rasopathies
PathoPredictor
CCR
VEST
Missense depletion
FATHMM
Missense badness
p=25
b=139
p=213
b=262
p=152
b=99
p=17
b=15
0.00
0.25
0.50
0.75
1.00
Average precision


Cardiomyopathy
Epilepsy
Epilepsy (dominant)
Rasopathies
Total ClinVar
ClinVar w/ Evidence
PathoPredictor
CCR
VEST
FATHMM
Missense depletion
Missense badness
p=322
b=155
p=332
b=187
p=217
b=77
p=150
b=61
p=283
b=144
p=294
b=184
p=201
b=76
p=71
b=61
0.00
0.25
0.50
0.75
1.00
Average precision


a
True positive rate
1.00
0.75
0.50
0.25
0.00
Disease panel
Total ClinVar
0.00
0.25
0.50
0.75
1.00
False positive rate
KCNQ2
RAF1
SCN2A
SCN5A
STXBP1


b
Epilepsy
Epilepsy (dominant)
PathoPredictor
REVEL
0.00
0.25
0.50
0.75
1.00
Average precision