REVEL: An ensemble method for predicting the pathogenicity of rare missense variants

Nilah M Ioannidis,[1,2,34] Joseph H Rothstein,[2-4,34] Vikas Pejaver,[5] Sumit Middha,[6] Shannon K McDonnell,[7] Saurabh Baheti,[7] Anthony Musolf,[8] Qing Li,[8] Emily Holzinger,[8] Danielle Karyadi,[9] Lisa A Cannon-Albright,[10] Craig C Teerlink,[10] Janet L Stanford,[11] William B Isaacs,[12] Jianfeng Xu,[13] Kathleen A Cooney,[14] Ethan M Lange,[15] Johanna Schleutker,[16,17] John D Carpten,[18] Isaac J Powell,[19] Olivier Cussenot,[20] Geraldine Cancel-Tassin,[20] Graham G Giles,[21,22] Robert J MacInnis,[21,22] Christiane Maier,[23,24] Chih-Lin Hsieh,[25] Fredrik Wiklund,[26] William J Catalona,[27] William D Foulkes,[28] Diptasri Mandal,[29] Rosalind A Eeles,[30] Zsofia Kote-Jarai,[30] Carlos D Bustamante,[1,31] Daniel J Schaid,[7] Trevor Hastie,[31,32] Elaine A Ostrander,[9] Joan E Bailey-Wilson,[8] Predrag Radivojac,[5] Stephen N Thibodeau,[33] Alice S Whittemore,[2,31] and Weiva Sieh[2-4]*

1. Department of Genetics, Stanford University, Stanford, CA 94305, USA
2. Department of Health Research and Policy, Stanford University, Stanford, CA 94305, USA
3. Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
4. Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
5. Department of Computer Science and Informatics, Indiana University, Bloomington, IN 47405, USA
6. Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA
7. Department of Health Sciences Research, Mayo Clinic, Rochester, MN 55905, USA
8. Computational and Statistical Genomics Branch, National Human Genome Research Institute, Baltimore, MD 21224, USA
9. Cancer Genetics and Comparative Genomics Branch, National Human Genome Research Institute, Bethesda, MD 20892, USA
10. Department of Internal Medicine, University of Utah School of Medicine, Salt Lake City, UT 84108, USA
11. Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA
12. Brady Urological Institute, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA
13. NorthShore University HealthSystem Research Institute, Evanston, IL 60201, USA
14. Departments of Internal Medicine and Urology, University of Michigan Medical School, Ann Arbor, MI 48109, USA
15. Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
16. Department of Medical Biochemistry and Genetics, University of Turku, Turku, FI-20014, Finland
17. Department of Medical Genetics, Tyks Microbiology and Genetics, Turku University Hospital, Turku, FI-20520, Finland
18. Integrated Cancer Genomics Division, Translational Genomics Research Institute, Phoenix, AZ 85004, USA
19. Department of Urology, Wayne State University, Detroit, MI 48201, USA
20. CeRePP, Universite Paris, Paris, 75013, France
21. Cancer Epidemiology Centre, Cancer Council Victoria, Melbourne, VIC 3004, Australia
22. Centre for Epidemiology and Biostatistics, University of Melbourne, Melbourne, VIC 3010, Australia
23. Institute of Human Genetics, University Hospital of Ulm, Ulm, 89075, Germany
24. Department of Urology, University Hospital of Ulm, Ulm, 89075, Germany
25. Department of Urology, University of Southern California, Los Angeles, CA 90033, USA
26. Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, SE-171 77, Sweden
27. Department of Urology, Northwestern University Feinberg School of Medicine, Chicago, IL 60611, USA
28. Departments of Oncology and Human Genetics, Montreal General Hospital, Montreal, QC H3G 1A4, Canada
29. Department of Genetics, LSU Health Sciences Center, New Orleans, LA 70112, USA
30. Division of Genetics and Epidemiology, Institute of Cancer Research, London, SM2 5NG, UK

31. Department of Biomedical Data Science, Stanford University, Stanford, CA 94305, USA
32. Department of Statistics, Stanford University, Stanford, CA 94305, USA
33. Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN 55905, USA
34. These authors contributed equally and are listed alphabetically

* Correspondence: weiva.sieh@mssm.edu

ABSTRACT

The vast majority of coding variants are rare, and assessment of the contribution of rare variants to complex traits is hampered by low statistical power and limited functional data. Improved methods for predicting the pathogenicity of rare coding variants are needed to facilitate the discovery of disease variants from exome sequencing studies. We developed REVEL (Rare Exome Variant Ensemble Learner), an ensemble method for predicting the pathogenicity of missense variants based on individual tools: MutPred, FATHMM, VEST, Polyphen, SIFT, PROVEAN, MutationAssessor, MutationTaster, LRT, GERP, SiPhy, phyloP, and phastCons. REVEL was trained using recently discovered pathogenic and rare neutral missense variants, excluding those previously used to train its constituent tools. When applied to two independent test sets, REVEL had the best overall performance ($p<10^{-12}$) compared with any individual tool and seven ensemble methods: MetaSVM, MetaLR, KGGSeq, Condel, CADD, DANN, and Eigen. Importantly, REVEL also had the best performance for distinguishing pathogenic from rare neutral variants with allele frequencies <0.5%. Compared with other ensemble methods, the area under the receiver operating characteristic curve (AUC) for REVEL was 0.046-0.182 higher in an independent test set of 935 recent SwissVar disease variants and 123,935 putatively neutral exome sequencing variants, and 0.027-0.143 higher in an independent test set of 1,953 pathogenic and 2,406 benign variants recently reported in ClinVar. We provide pre-computed REVEL scores for all possible human missense variants to facilitate the identification of pathogenic variants in the sea of rare variants discovered as sequencing studies expand in scale.

INTRODUCTION

Interpreting genetic variation from next generation sequencing (NGS) datasets is essential for advancing personalized medicine.[1; 2] The vast majority of variants discovered by NGS are rare.[3; 4] Recent exome and genome sequencing studies have found that roughly 85% of nonsynonymous variants have alternate allele frequencies (AF) less than 0.5%, and roughly 100-400 rare nonsynonymous variants are discovered per sequenced individual.[3; 4] Rare coding variants play major roles in disease causation and may contribute to the missing heritability from genome-wide association studies.[5; 6] However, the majority of nonsynonymous variants discovered by NGS have unknown significance because experimental validation of large numbers of rare variants is infeasible and association studies require prohibitively large sample sizes to detect rare variants with modest effect sizes with high statistical power. Therefore, computational tools that can accurately predict the pathogenicity of rare variants are needed to help identify those variants that are most likely to cause disease.

Many tools for predicting the pathogenicity of missense variants have been developed based on features such as amino acid or nucleotide conservation and biochemical properties of the amino acid substitutions.[7-18] However, individual tools often disagree, in part because they utilize different predictive features. Ensemble methods that combine the results of multiple individual predictors can improve performance.[19-28] However, few existing pathogenicity prediction tools have targeted the interpretation of

rare variants.[24] Current tools are often trained on predominantly common neutral variants and some explicitly impose a minimum AF threshold for defining neutral training variants[15; 21; 25]. In contrast, most disease training variants are rare. As a result of this AF imbalance between disease and neutral training variants, tools that rely on AF as a predictive feature may have lower ability to distinguish disease variants from rare neutral variants than from common ones.[24] Biological differences such as higher conservation scores for rare versus common variants may also make rare neutral variants more difficult to distinguish from disease variants.[24; 29] Despite the fact that the vast majority of nonsynonymous variants discovered by NGS are rare, the performance of existing prediction tools on rare variants is not well known.[30] Thus, there is a growing need for the development and evaluation of tools for predicting the pathogenicity of rare variants.

Here, we present an ensemble method for predicting the pathogenicity of missense variants that outperforms existing approaches overall and when applied to rare variants. The Rare Exome Variant Ensemble Learner (REVEL) method incorporates recently developed individual prediction tools as features and was trained on recently discovered disease and rare neutral missense variants that did not overlap with the training data for its constituent predictors. We also assembled two large independent test sets of recently discovered pathogenic and benign variants that parallel the likely application of REVEL to newly discovered variants from NGS studies. We benchmark the performance of REVEL and existing ensemble predictors for distinguishing disease mutations from neutral variants across a broad range of allele frequencies. To make our method easily accessible for research and clinical use, we provide pre-computed REVEL scores for all possible human missense variants[31].

METHODS

Random forest. We trained a random forest on the set of variants described below using the R 'randomForest' package[32] with 1000 binary classification trees[33; 34]. We selected the number of trees to be sufficiently large for the out-of-bag (OOB) error rate to plateau; sensitivity analyses showed that increasing the number of trees to 3000 did not improve performance on the training dataset. The OOB prediction for a given training variant is the proportion of trees that classified the variant as pathogenic across only those trees in the forest that excluded the variant from their bootstrapped training sample.[33] Four features were selected at random as candidates for each split in the random forest trees, which was the default value for 18 features described below. To address the imbalance in the numbers of available disease and neutral training variants, we sampled the same number (n = 6,182) of disease and neutral variants when generating the bootstrapped training set for each tree in the forest. The importance of each predictive feature was measured by the total decrease in the Gini index[33] (improvement in node purity) for all splits on that feature, averaged over all trees in the forest.

Training variants. REVEL was trained using putative disease and rare neutral missense variants. Disease variants were obtained from the Human Gene Mutation Database (HGMD)[35] version 2015.2 and were restricted to the set of missense Disease Mutations (DMs) added to HGMD since

August 1, 2012 to minimize overlap with variants previously used to train component features in the REVEL random forest. Exome sequencing variants (ESVs) were obtained from the Exome Sequencing Project (ESP)[4] European-American and African-American populations; Atherosclerosis Risk in Communities (ARIC) study[36] European-American and African-American populations; and 1000 Genomes Project (KGP)[3] European, Yoruban, and Asian populations, as recorded in dbNSFP[31] version 2.7. After excluding all disease variants in HGMD and the data sources for test sets 1 and 2 described below, the remaining ESVs were considered putatively neutral. For both the disease and neutral training variants, we also excluded all variants that had previously been used to train individual component features in the REVEL random forest; specifically, MutPred[8], Polyphen-2[10], MutationTaster[11], FATHMM v2.3[14], and VEST 3.0[15]. Finally, when a given genetic variant corresponded to multiple amino acid substitutions (AASs) at the protein level, only one AAS was selected at random. After applying all exclusion criteria, a total of 6,182 disease variants and 281,972 putatively neutral ESVs remained. We randomly selected approximately half (n=140,921) of the putatively neutral ESVs, of which 123,706 rare ESVs (with a maximum alternate AF between 0.1% and 1% across the seven study populations) were used for training, and 17,215 ESVs with AF >1% were used for initial evaluation of performance across a range of AFs. The remaining half of ESVs were held out for use as independent test variants as described below. Thus, the final training set consisted of 6,182 HGMD disease variants and 123,706 rare neutral ESVs.

Features. REVEL incorporates a total of 18 individual pathogenicity prediction scores from 13 tools as predictive features. MutPred scores were newly computed for this study using the UniProt[37] canonical protein sequence when available and the Ensembl[38] canonical transcript otherwise. PROVEAN[13] scores were obtained from dbNSFP v2.9 (February 3, 2015). Sixteen additional scores were obtained from dbNSFP v2.7 (September 12, 2014), including eight functional prediction scores (SIFT[7]; Polyphen-2 HVAR and HDIV; LRT[9]; MutationTaster; MutationAssessor[12]; FATHMM v2.3; and VEST 3.0) and eight conservation scores (GERP++[39]; SiPhy[40]; three phyloP[41] scores for primates, placental mammals, and vertebrates; and three phastCons[42] scores for primates, placental mammals, and vertebrates). For PolyPhen-2, FATHMM, and PROVEAN, when multiple protein isoforms were associated with a given variant, we used the average score across all isoforms. Missing features were imputed using the *k*-nearest neighbors method implemented in the R 'impute' package[43]. Missing feature values for a given variant were assigned the average value of the non-missing elements of its $k = 40$ nearest neighboring variants; when more than 50% of features were missing for a given variant, we assigned the overall mean across all variants.

Test sets. We assembled two independent test sets that did not overlap with either the REVEL training data or the training data for the component features of REVEL. **Test set 1** consisted of 935 disease variants added to SwissVar[44] (release 2015_10) since August 1, 2012 and approximately half (n = 141,051) of the putatively neutral nonsynonymous ESVs described above that had not been included in the REVEL training set or initial evaluation. **Test set 2** consisted of 1,953 pathogenic or likely pathogenic and 2,406 benign or likely benign variants recently deposited into ClinVar[45; 46] by submitters following

variant classification guidelines similar to the American College of Medical Genetics and Genomics (ACMG) guidelines[47; 48]. Specifically, all single nucleotide missense variants submitted to ClinVar by GeneDx, Emory Genetics Laboratory, Partners HealthCare Laboratory for Molecular Medicine[49], University of Chicago Genetic Services Laboratory, Ambry Genetics, and Invitae were downloaded on October 13, 2015. We excluded the following from both test sets 1 and 2: all REVEL training variants, all DM variants added to HGMD prior to August 1, 2012, and all variants that had previously been used to train individual component features in REVEL. Finally, to eliminate overlap between the two test sets, we excluded any variants that were present in both SwissVar and ClinVar from test set 1 if benign (n=9) and from test set 2 if pathogenic (n=12).

Comparators. We compared the performance of REVEL to seven ensemble prediction tools that were recently developed, widely used, and readily implemented: MetaLR[28], MetaSVM[28], Eigen[50], CADD[16] v1.3, DANN[17], Condel[19], and KGGSeq[23; 24] v0.8. We ran KGGSeq using the default model selection option that chooses an optimized set of features for each variant[24]. We plotted receiver operating characteristic (ROC) curves and compared the area under the ROC curve (AUC) estimates for different tools using Delong's test[51] implemented in the R 'pROC' package[52]. We also computed the area under precision-recall (PR) curve using the R 'ROCR' package[53]. For the training variants, REVEL scores were computed using only the OOB predictions, which have been shown to provide performance estimates that are as accurate as for an independent test set of equal size consisting of variants with similar characteristics[33].

RESULTS

Characterization of REVEL features. The REVEL ensemble score combines pathogenicity predictions from 18 individual scores (features), including eight conservation scores and 10 functional scores. **Figure 1A** shows the correlation among individual features. The conservation scores, as well as LRT and Mutation Taster, were almost all highly (Spearman rank correlation coefficient, $R > 0.6$) to moderately correlated ($0.4 < R < 0.6$). Five functional scores (MutationAssessor, PROVEAN, VEST, Polyphen-2 HDIV and HVAR) were almost all highly correlated. VEST was also highly correlated with several conservation scores, LRT and MutationTaster. In contrast, FATHMM had low correlation ($R < 0.4$) with all other scores, and MutPred and SIFT had low to moderate correlation with other scores. The five most important features in the REVEL random forest were: FATHMM, VEST, MutationAssessor, MutPred, and Polyphen-2 HVAR (**Figure 1B**). The importance measure for an individual feature reflects correlations with other features as well as its intrinsic predictive ability, because importance may be shared among correlated features[34].

Overall performance of REVEL compared with other methods. The REVEL ensemble score discriminated well between HGMD disease mutations and putatively neutral ESVs, with an overall AUC of 0.908 estimated using OOB predictions for the training set (**Figure 2A**). The AUC for REVEL was significantly better than any of its constituent features (maximum $p < 10^{-12}$ for any pairwise comparison),

among which VEST (AUC = 0.844) and FATHMM (AUC = 0.824) had the highest AUCs (**Table S1**). AUCs for the other individual prediction tools ranged from 0.589 to 0.809, and tended to be higher for functional predictors (0.717-0.844) than for conservation scores (0.589-0.791). The AUC for REVEL was also significantly better than the other ensemble methods (maximum $p < 10^{-12}$ for any pairwise comparison), among which MetaLR (AUC = 0.883) and MetaSVM (AUC = 0.879) had the next highest AUCs (**Figure 2A; Table S2**).

Performance for rare versus common neutral variants. We next compared the performance of REVEL to that of other ensemble methods for discriminating between HGMD disease mutations, which are predominantly rare, and putatively neutral ESVs with AFs ranging from very rare (0.1-0.3%) to common (>5%). We found that all of the ensemble methods tended to have worse ability to discriminate disease mutations from rare neutral variants than from common neutral variants (**Figure 2B**; **Table S2**). However, compared to other ensemble methods, REVEL had superior discriminatory ability for neutral variants within all AF ranges up to 3%, with the greatest improvements in AUC for rare variants with AF < 0.5% (**Figure 2B**; **Table S2**). For neutral variants with AF > 3%, REVEL had the second highest AUC after MetaLR. In addition, the performance of REVEL appeared to be less sensitive to neutral variant AF than other methods. The AUC range for very rare to common variants was narrowest for REVEL (0.897 to 0.957) and widest for DANN (0.703 to 0.897), which appeared to be most sensitive to AF (**Table S2**).

Performance evaluation in two independent test sets. In test set 1, consisting of 935 independent disease mutations from SwissVar and 141,051 putatively neutral ESVs, the relative performance of all eight ensemble predictors (**Figure 3**; **Table S3**) was similar to that observed in the training set. REVEL had the best performance both overall ($p < 10^{-12}$) and for neutral variants within all AF ranges up to 5%. For common neutral variants with AF > 5%, REVEL was again surpassed only by MetaLR. The improvement in AUC obtained using REVEL versus the other ensemble methods was again greatest for rare neutral variants. In test set 2, consisting of 1,953 pathogenic and 2,406 benign variants from ClinVar, we confirmed that REVEL had the best performance among the ensemble methods both overall ($p < 10^{-12}$) and for neutral variants within all AF ranges up to 3%, and that the improvement in AUC was greatest for rare neutral variants (**Figure 4**; **Table S4**). All of the ensemble methods had better overall ability to distinguish benign vs. pathogenic variants from ClinVar, than putatively neutral ESVs vs. disease variants from SwissVar or HGMD, which may be a consequence of the more stringent definition of benign variants from ClinVar. REVEL also had the best overall performance measured by the area under the PR curve (**Table S5**) across a wide range of proportions of disease variants represented in the training set (4.8%), and test sets 1 (0.7%) and 2 (44.8%).

Interpretation of REVEL scores. The REVEL score for an individual variant can range from zero to one, reflecting the proportion of trees in the random forest that classified the variant as pathogenic. REVEL score distributions for the 6,182 HGMD disease and 123,706 putatively neutral ESV training variants, and for all 1,125,160 ESVs reported by ESP, ARIC and KGP, are shown in **Figure 5a**. The distributions of REVEL scores were very similar for all reported ESVs and the subset of putatively neutral

ESV training variants, with only a small shift towards higher scores for all ESVs. **Figure 5b** shows the percentiles of the REVEL scores separately for disease and neutral training variants or all ESVs. **Figure S1** shows the sensitivity and specificity corresponding to different REVEL score thresholds, above which a variant would be classified as pathogenic. For example, 75.4% of disease mutations but only 10.9% of neutral variants (and 12.4% of all ESVs) have a REVEL score above 0.5, corresponding to a sensitivity of 0.754 and specificity of 0.891. Selecting a more stringent REVEL score threshold of 0.75 would result in higher specificity but lower sensitivity, with 52.1% of disease mutations, 3.3% of neutral variants, and 4.1% of all ESVs being classified as pathogenic.

DISCUSSION

REVEL is an ensemble method for predicting the pathogenicity of rare missense variants. Rare variants are likely to comprise the vast majority of variants of unknown significance discovered in future sequencing studies. We have shown that REVEL consistently has the best overall performance compared to existing methods, particularly for distinguishing disease mutations from uncommon neutral missense variants with an AF below 3%. To facilitate use by clinicians and researchers, we have pre-computed REVEL scores for all missense variants in dbNSFP 2.7, a database of all potential nonsynonymous single nucleotide variants in the human genome. REVEL thus addresses the need for a pathogenicity prediction tool with improved accuracy for interpreting rare genetic variants.

The REVEL method has several strengths. First, REVEL was trained and tested on recently identified disease and neutral variants that may closely resemble novel variants discovered by future NGS studies, which are likely to include variants with lower allele frequencies and more modest effects than previously discovered variants. The REVEL neutral training variants were specifically restricted to AFs between 0.1% and 1% to improve performance when interpreting rare variants. Second, REVEL incorporates a larger number of individual predictors than prior ensemble methods, including both MutPred and VEST, which were among the most important features in the REVEL random forest. MutPred scores, in particular, were not previously widely available and have now been computed for all missense variants in dbNSFP 2.7 as part of this study. Finally, we carefully removed from the training and test sets all variants used to train any of the component predictors in REVEL to reduce overfitting and inflated performance estimates.

A key limitation of this study and others is the reliance on pathogenicity assertions from existing databases, which may be inaccurate and incomplete. Misclassification of training and test variants as disease or neutral would limit both the accuracy of the prediction method and the resulting performance estimates. Nonetheless, we expect that the putative disease variants used to train REVEL are enriched for true disease variants compared to the putative neutral variants, allowing identification of key predictive features of pathogenic variants. An additional complication is that existing pathogenicity assertions for some variants may have been based in part on predictions from popular tools, such as SIFT and Polyphen-2, potentially resulting in inflated performance of these predictors and ensemble scores that use

them. Finally, the performance of REVEL and other ensemble methods is limited by the accuracy of the component predictors and could benefit from inclusion of additional predictors as they become available in the future.

REVEL had the highest overall performance of any method in independent test sets, although its performance on common variants with AF > 3-5% was slightly worse than MetaLR or MetaSVM. The strong overall performance of REVEL reflects the fact that the majority of neutral variants in the training and test datasets were rare, as expected for novel variants discovered by NGS. Furthermore, while we carefully removed all variants used to train REVEL and its constituent features from the two test sets, we did not systematically exclude training variants for the comparator ensemble scores; thus the performance estimates for the comparators could be overly optimistic. MetaLR and MetaSVM had the next highest overall performance. Compared to REVEL, these two ensemble methods included many of the same predictive features, except for VEST, MutPred, and PROVEAN, and also included the AF, which could contribute to their greater sensitivity to the neutral variant AF. Condel is a weighted average of FATHMM and MutationAssessor, and its lower performance relative to some ensemble methods may be due to the inclusion of fewer predictive features. Eigen employs an unsupervised approach to separate variants into two classes, and also uses fewer predictive features than REVEL, MetaLR, and MetaSVM. CADD and DANN differ from the other ensemble methods in their use of many basic genomic and protein annotations from ENCODE and Ensembl as features in addition to functional predictions from Polyphen-2 and SIFT. Although CADD and DANN did not perform as well as the other ensemble methods for missense variants, they have important advantages for genome-wide NGS applications because they provide scores for noncoding and regulatory variants that are on the same scale as for coding variants.

The improved performance of REVEL relative to other ensemble methods was greatest for discriminating between disease and rare neutral variants. This result may be partly explained by the fact that REVEL was trained on rare neutral variants with AF <1% and did not rely on AF as a predictive feature. To our knowledge, one other ensemble predictor, KGGSeq[24], was similarly trained on rare neutral variants. KGGSeq uses many of the same component predictors as REVEL, except for MutPred, and also includes CADD as a predictive feature. However, KGGSeq adaptively selects an optimal subset of features rather than using all features to predict the pathogenicity of each variant, in part to allow exclusion of features with missing data. Possible explanations for the improved performance of REVEL over KGGSeq include: use of all features for all variants by first imputing missing scores, importance of MutPred as a predictive feature, and use of a random forest approach rather than logistic regression.

REVEL also outperformed its individual constituent prediction tools as expected for ensemble methods[19-28]. The top performing individual tools on our training dataset were VEST[15], FATHMM[14], and MutPred[8], consistent with their high importance in the REVEL random forest. VEST predictions are based on a particularly large set of 86 basic genomic and protein annotations and had the best performance among the individual tools. FATHMM uses a hidden Markov modeling approach to analyze multiple sequence alignments and alignments of conserved protein domain families to compute position-specific

amino acid probabilities. The uniqueness of this method may contribute to the low correlation between FATHMM and other prediction tools and high importance in REVEL[28]. Finally, the strong performance of MutPred may be because its predictions are based on a particularly detailed model of protein structural and functional properties, including secondary structure, solvent accessibility, functional domains, methylation, phosphorylation, and glycosylation, with quantitative estimates of the probability of losing each property due to a particular amino acid change.

In conclusion, REVEL is an ensemble method that outperforms existing tools for distinguishing disease variants from rare neutral variants. REVEL can be used to prioritize the most likely clinically or functionally relevant variants among the sea of rare variants that are increasingly discovered as sequencing studies expand in scale. For example, REVEL scores have been used by the International Consortium of Prostate Cancer Genetics as weights for combining variants discovered by exome sequencing in gene-level case-control studies. Pre-computed REVEL pathogenicity scores for all possible human missense variants, based on GENCODE v9 gene annotations[54] for hg19, are available for download (see URLs). To aid interpretation, we also provide estimates of REVEL sensitivity and specificity for different score thresholds, and the quantiles of the REVEL score in over 1 million ESVs observed in KGP, ESP, and ARIC. Future studies may explore the application of REVEL to specific genes to evaluate its clinical utility for interpreting variants of unknown significance for a broad spectrum of clinical conditions.

SUPPLEMENTAL DATA

Supplemental Data include one figure and four tables.

ACKNOWLEDGEMENTS

This research was funded by the National Institutes of Health (NIH) grants: U01CA089600, R01CA094069, R01LM009722, R01MH105524, K07CA143047, and F32HG008330; and by the Intramural Research Program of the National Human Genome Research Institute, NIH.

WEB RESOURCES

The URLs for data presented herein are as follows:

ClinVar, http://www.ncbi.nlm.nih.gov/clinvar/

dbNSFP, https://sites.google.com/site/jpopgen/dbNSFP/

HGMD, http://www.hgmd.cf.ac.uk/

REVEL, https://sites.google.com/site/revelgenomics/

SwissVar, http://swissvar.expasy.org/

REFERENCES

1. Peterson, T.A., Doughty, E., and Kann, M.G. (2013). Towards precision medicine: advances in computational approaches for the analysis of human variants. J Mol Biol 425, 4047-4063.
2. Yang, Y., Muzny, D.M., Reid, J.G., Bainbridge, M.N., Willis, A., Ward, P.A., Braxton, A., Beuten, J., Xia, F., Niu, Z., et al. (2013). Clinical whole-exome sequencing for the diagnosis of mendelian disorders. N Engl J Med 369, 1502-1511.
3. 1000 Genomes Project Consortium, Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56-65.
4. Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., et al. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337, 64-69.
5. Cirulli, E.T., and Goldstein, D.B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet 11, 415-425.
6. Bamshad, M.J., Ng, S.B., Bigham, A.W., Tabor, H.K., Emond, M.J., Nickerson, D.A., and Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet 12, 745-755.
7. Kumar, P., Henikoff, S., and Ng, P.C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 4, 1073-1081.
8. Li, B., Krishnan, V.G., Mort, M.E., Xin, F., Kamati, K.K., Cooper, D.N., Mooney, S.D., and Radivojac, P. (2009). Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics 25, 2744-2750.
9. Chun, S., and Fay, J.C. (2009). Identification of deleterious mutations within three human genomes. Genome Res 19, 1553-1561.
10. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat Methods 7, 248-249.
11. Schwarz, J.M., Rodelsperger, C., Schuelke, M., and Seelow, D. (2010). MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods 7, 575-576.
12. Reva, B., Antipin, Y., and Sander, C. (2011). Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res 39, e118.
13. Choi, Y., Sims, G.E., Murphy, S., Miller, J.R., and Chan, A.P. (2012). Predicting the functional effect of amino acid substitutions and indels. PLoS One 7, e46688.
14. Shihab, H.A., Gough, J., Cooper, D.N., Stenson, P.D., Barker, G.L., Edwards, K.J., Day, I.N., and Gaunt, T.R. (2013). Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum Mutat 34, 57-65.
15. Carter, H., Douville, C., Stenson, P.D., Cooper, D.N., and Karchin, R. (2013). Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics 14 Suppl 3, S3.
16. Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46, 310-315.
17. Quang, D., Chen, Y., and Xie, X. (2015). DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics 31, 761-763.
18. Niroula, A., Urolagin, S., and Vihinen, M. (2015). PON-P2: prediction method for fast and reliable identification of harmful variants. PLoS One 10, e0117380.
19. Gonzalez-Perez, A., and Lopez-Bigas, N. (2011). Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am J Hum Genet 88, 440-449.
20. Crockett, D.K., Ridge, P.G., Wilson, A.R., Lyon, E., Williams, M.S., Narus, S.P., Facelli, J.C., and Mitchell, J.A. (2012). Consensus: a framework for evaluation of uncertain gene variants in laboratory test reporting. Genome Med 4, 48.
21. Lopes, M.C., Joyce, C., Ritchie, G.R., John, S.L., Cunningham, F., Asimit, J., and Zeggini, E. (2012). A combined functional annotation score for non-synonymous variants. Hum Hered 73, 47-51.
22. Olatubosun, A., Valiaho, J., Harkonen, J., Thusberg, J., and Vihinen, M. (2012). PON-P: integrated predictor for pathogenicity of missense variants. Hum Mutat 33, 1166-1174.
23. Li, M.X., Gui, H.S., Kwan, J.S., Bao, S.Y., and Sham, P.C. (2012). A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. Nucleic Acids Res 40, e53.

24. Li, M.X., Kwan, J.S., Bao, S.Y., Yang, W., Ho, S.L., Song, Y.Q., and Sham, P.C. (2013). Predicting mendelian disease-causing non-synonymous single nucleotide variants in exome sequencing studies. PLoS Genet 9, e1003143.
25. Frousios, K., Iliopoulos, C.S., Schlitt, T., and Simpson, M.A. (2013). Predicting the functional consequences of non-synonymous DNA sequence variants--evaluation of bioinformatics tools and development of a consensus strategy. Genomics 102, 223-228.
26. Capriotti, E., Altman, R.B., and Bromberg, Y. (2013). Collective judgment predicts disease-associated single nucleotide variants. BMC Genomics 14 Suppl 3, S2.
27. Bendl, J., Stourac, J., Salanda, O., Pavelka, A., Wieben, E.D., Zendulka, J., Brezovsky, J., and Damborsky, J. (2014). PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. Plos Computational Biology 10, e1003440.
28. Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K., and Liu, X. (2015). Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum Mol Genet 24, 2125-2137.
29. Hodgkinson, A., Casals, F., Idaghdour, Y., Grenier, J.C., Hernandez, R.D., and Awadalla, P. (2013). Selective constraint, background selection, and mutation accumulation variability within and between human populations. BMC Genomics 14, 495.
30. Grimm, D.G., Azencott, C.A., Aicheler, F., Gieraths, U., MacArthur, D.G., Samocha, K.E., Cooper, D.N., Stenson, P.D., Daly, M.J., Smoller, J.W., et al. (2015). The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. Hum Mutat 36, 513-523.
31. Liu, X., Jian, X., and Boerwinkle, E. (2013). dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. Hum Mutat 34, E2393-2402.
32. Liaw, A., and Wiener, M. (2002). Classification and Regression by randomForest. R News 2, 18-22.
33. Breiman, L. (2001). Random forests. Machine Learning 45, 5-32.
34. Hastie, T., Tibshirani, R., and Friedman, J.H. (2009). The elements of statistical learning : data mining, inference, and prediction.(New York: Springer).
35. Stenson, P.D., Mort, M., Ball, E.V., Shaw, K., Phillips, A., and Cooper, D.N. (2014). The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet 133, 1-9.
36. The ARIC investigators (1989). The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. Am J Epidemiol 129, 687-702.
37. Magrane, M., and UniProt Consortium (2011). UniProt Knowledgebase: a hub of integrated protein data. Database (Oxford) 2011, bar009.
38. Flicek, P., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., et al. (2014). Ensembl 2014. Nucleic Acids Res 42, D749-755.
39. Davydov, E.V., Goode, D.L., Sirota, M., Cooper, G.M., Sidow, A., and Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. Plos Computational Biology 6, e1001025.
40. Garber, M., Guttman, M., Clamp, M., Zody, M.C., Friedman, N., and Xie, X. (2009). Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics 25, i54-62.
41. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res 20, 110-121.
42. Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15, 1034-1050.
43. Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R.B. (2001). Missing value estimation methods for DNA microarrays. Bioinformatics 17, 520-525.
44. Mottaz, A., David, F.P., Veuthey, A.L., and Yip, Y.L. (2010). Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar. Bioinformatics 26, 851-852.
45. Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M., and Maglott, D.R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res 42, D980-985.
46. Landrum, M.J., Lee, J.M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., et al. (2015). ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res.

47. Richards, C.S., Bale, S., Bellissimo, D.B., Das, S., Grody, W.W., Hegde, M.R., Lyon, E., Ward, B.E., and Molecular Subcommittee of the ACMG Laboratory Quality Assurance Committee. (2008). ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007. Genet Med 10, 294-300.
48. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 17, 405-424.
49. Duzkale, H., Shen, J., McLaughlin, H., Alfares, A., Kelly, M.A., Pugh, T.J., Funke, B.H., Rehm, H.L., and Lebo, M.S. (2013). A systematic approach to assessing the clinical significance of genetic variants. Clin Genet 84, 453-463.
50. Ionita-Laza, I., McCallum, K., Xu, B., and Buxbaum, J.D. (2016). A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat Genet 48, 214-220.
51. DeLong, E.R., DeLong, D.M., and Clarke-Pearson, D.L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44, 837-845.
52. Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.C., and Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 12, 77.
53. Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: visualizing classifier performance in R. Bioinformatics 21, 3940-3941.
54. Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., et al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res 22, 1760-1774.

FIGURE LEGENDS

Figure 1. Individual prediction tools included as features in the REVEL random forest.

(A) Correlation among the individual features ordered by hierarchical clustering. The heatmap illustrates the Spearman rank correlation coefficients between features computed for the REVEL training variants.

(B) Relative importance of individual features. Gini importance estimates were normalized to sum to one.

Figure 2. Performance of ensemble methods for discrimination of disease training variants from putatively neutral exome sequencing variants (ESVs).

(A) Receiver operating characteristic (ROC) curves for 6182 HGMD disease mutations and 123,706 rare (AF 0.001-0.01**)** neutral ESVs used to train REVEL. REVEL scores were computed using only the out-of-bag predictions for its training variants.

(B) Area under the ROC curve (AUC) for 6182 HGMD disease mutations and 140,921 neutral ESVs, including REVEL training variants, stratified by neutral variant allele frequency.

Figure 3. Performance of ensemble methods in an independent test set of SwissVar disease mutations and putatively neutral exome sequencing variants (ESVs).

(A) Receiver operating characteristic (ROC) curves for 935 SwissVar disease mutations and 123,935 rare (AF 0.001-0.01) neutral ESVs that did not overlap with the training set.

(B) Area under the ROC curve (AUC) for 935 SwissVar disease mutations and 141,051 neutral ESVs, excluding REVEL training variants, stratified by neutral variant allele frequency.

Figure 4. Performance of ensemble methods in an independent test set of 1953 pathogenic and 2406 benign variants from ClinVar.

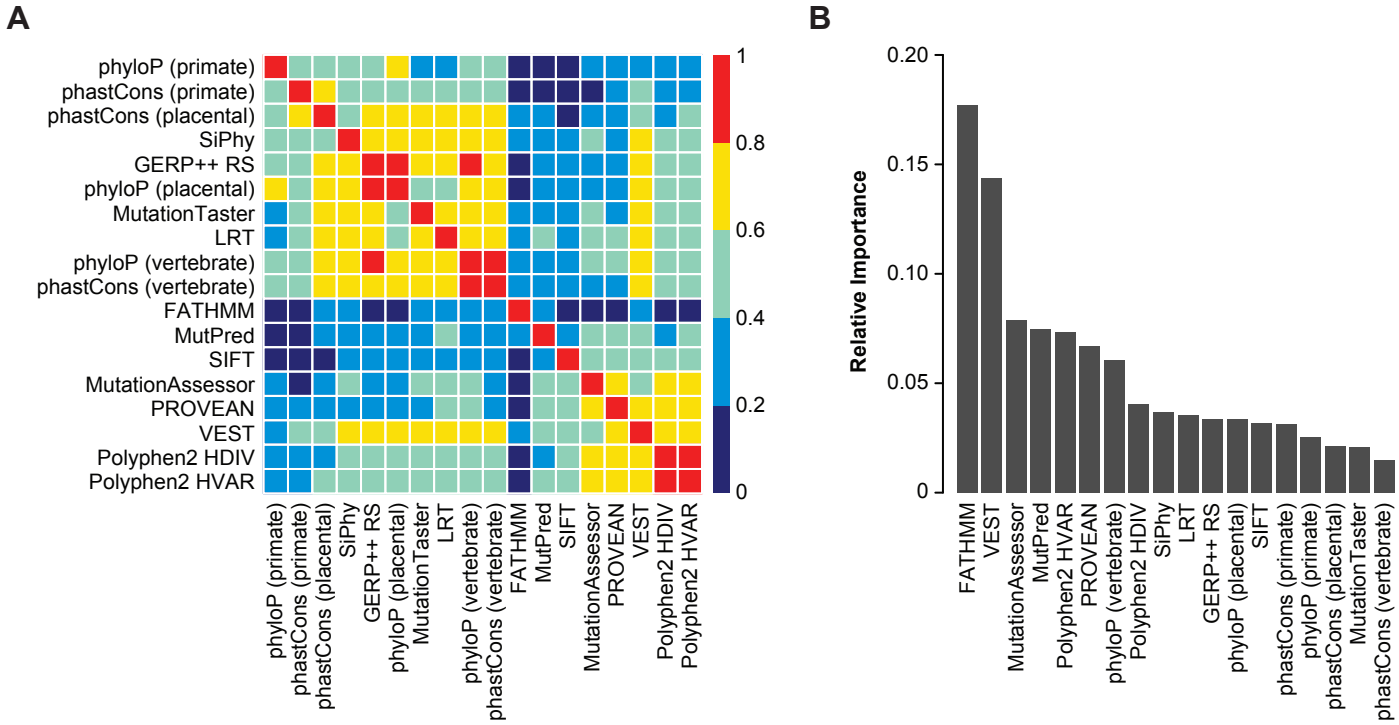
(A) Receiver operating characteristic (ROC) curves and the area under the ROC curve (AUC) for all variants.

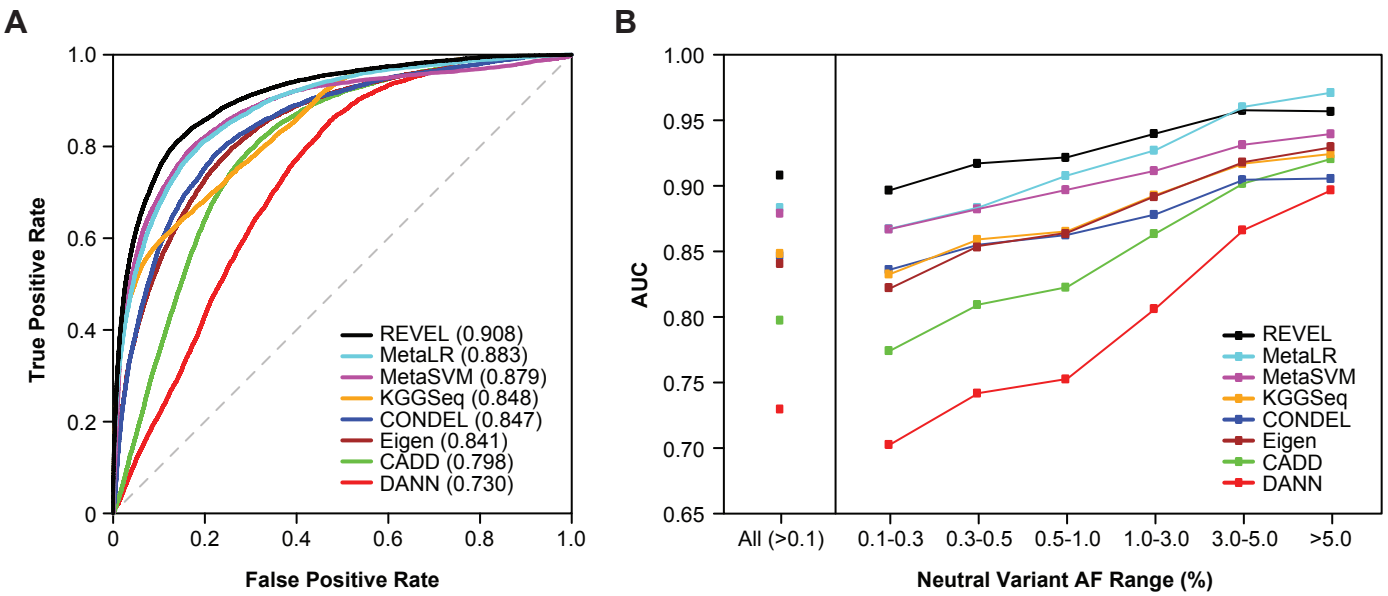
(B) AUC for each ensemble method, stratified by neutral variant allele frequency.

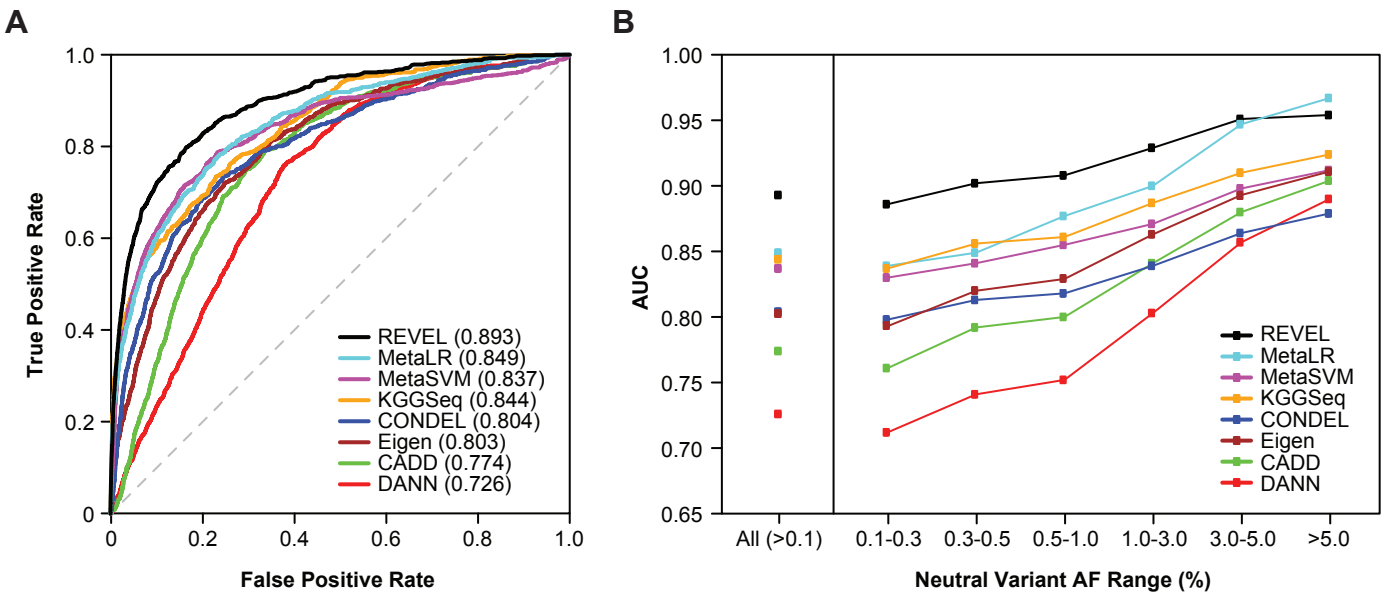
Figure 5. Interpretation of REVEL scores.

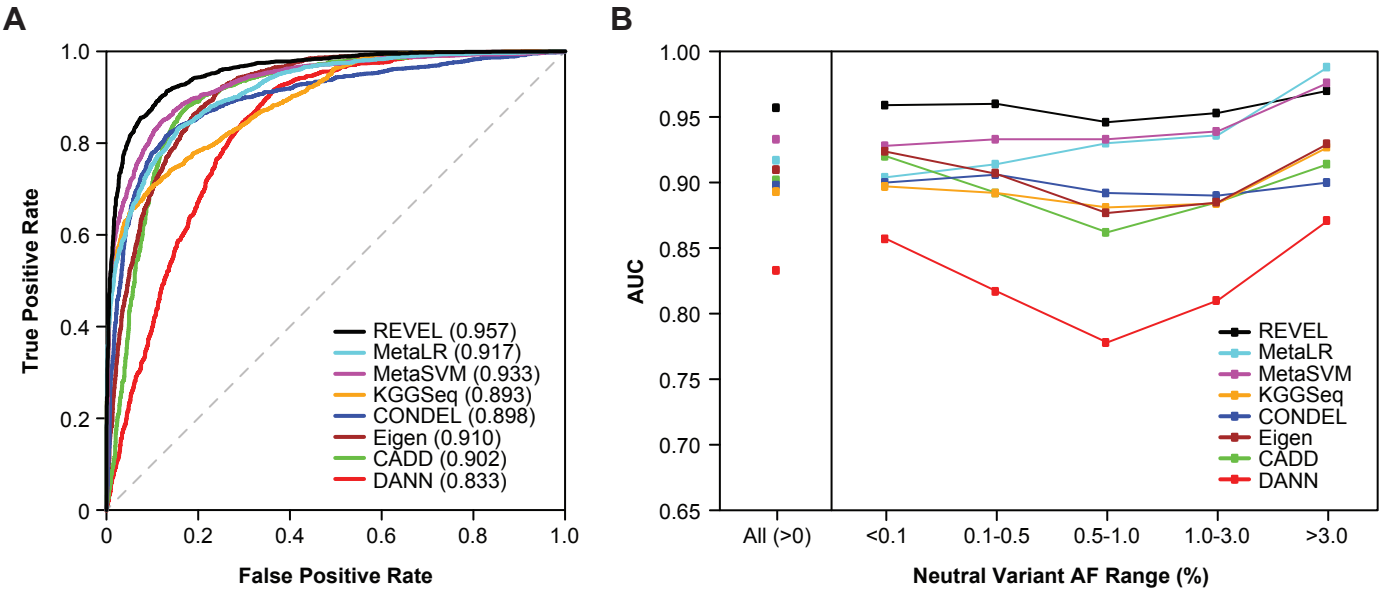
(A) Distribution of REVEL scores for 6182 disease (magenta) and 123,706 neutral (cyan) training variants, and 1,125,160 exome sequencing variants (black). REVEL scores were computed using only the out-of-bag predictions for training variants.

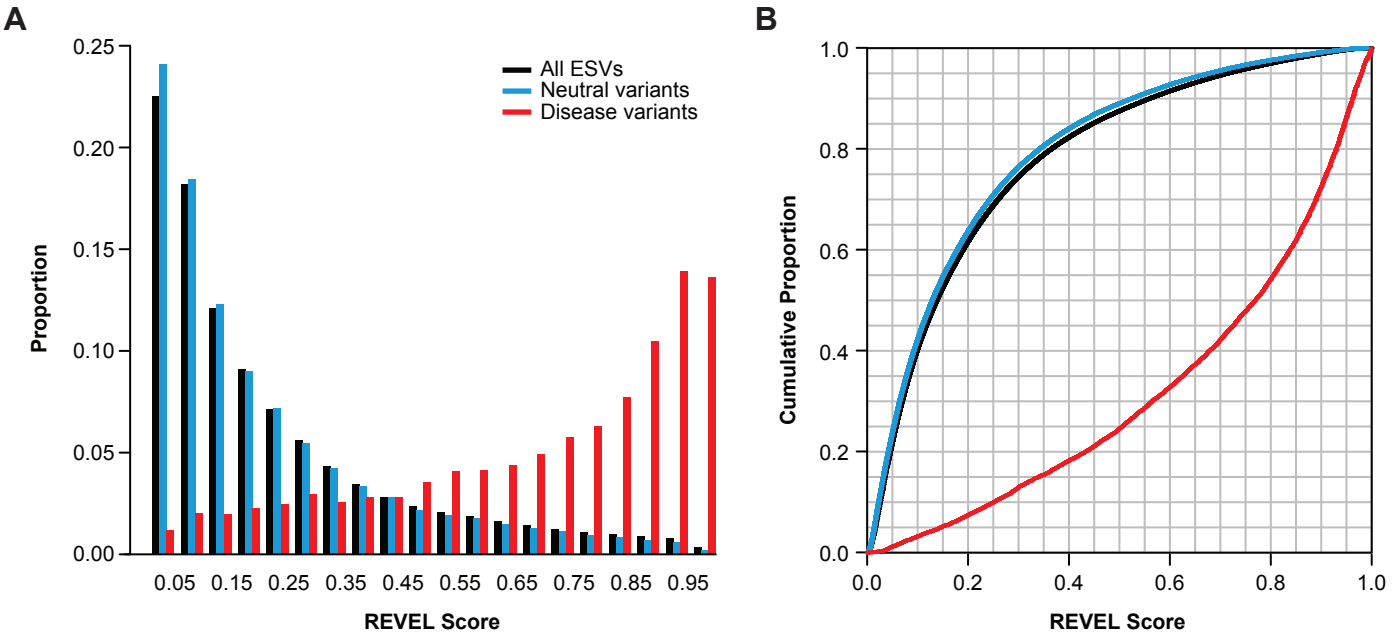
(B) Percentiles of the REVEL score distribution for 6182 disease (magenta) and 123,706 neutral (cyan) training variants, and 1,125,160 exome sequencing variants (black). REVEL scores were computed using only the out-of-bag predictions for training variants.


A
phyloP (primate)
phastCons (primate)
phastCons (placental)
SiPhy
GERP++ RS
phyloP (placental)
MutationTaster
LRT
phyloP (vertebrate)
phastCons (vertebrate)
FATHMM
MutPred
SIFT
MutationAssessor
PROVEAN
VEST
Polyphen2 HDIV
Polyphen2 HVAR
1
0.8
0.6
0.4
0.2
0
B
Relative Importance
0.20
0.15
0.10
0.05
0
FATHMM
VEST
MutationAssessor
MutPred
Polyphen2 HVAR
PROVEAN
phyloP (vertebrate)
Polyphen2 HDIV
SiPhy
LRT
GERP++ RS
phyloP (placental)
SIFT
phastCons (primate)
phyloP (primate)
phastCons (placental)
MutationTaster
phastCons (vertebrate)


A
True Positive Rate
False Positive Rate
REVEL (0.908)
MetaLR (0.883)
MetaSVM (0.879)
KGGSeq (0.848)
CONDEL (0.847)
Eigen (0.841)
CADD (0.798)
DANN (0.730)
B
AUC
All (>0.1)
0.1-0.3
0.3-0.5
0.5-1.0
1.0-3.0
3.0-5.0
>5.0
Neutral Variant AF Range (%)
REVEL
MetaLR
MetaSVM
KGGSeq
CONDEL
Eigen
CADD
DANN


A
True Positive Rate
False Positive Rate
REVEL (0.893)
MetaLR (0.849)
MetaSVM (0.837)
KGGSeq (0.844)
CONDEL (0.804)
Eigen (0.803)
CADD (0.774)
DANN (0.726)
B
AUC
Neutral Variant AF Range (%)
All (>0.1)
0.1-0.3
0.3-0.5
0.5-1.0
1.0-3.0
3.0-5.0
>5.0
REVEL
MetaLR
MetaSVM
KGGSeq
CONDEL
Eigen
CADD
DANN


A
True Positive Rate
False Positive Rate
REVEL (0.957)
MetaLR (0.917)
MetaSVM (0.933)
KGGSeq (0.893)
CONDEL (0.898)
Eigen (0.910)
CADD (0.902)
DANN (0.833)
B
AUC
All (>0)
<0.1
0.1-0.5
0.5-1.0
1.0-3.0
>3.0
Neutral Variant AF Range (%)
REVEL
MetaLR
MetaSVM
KGGSeq
CONDEL
Eigen
CADD
DANN


A
B
All ESVs
Neutral variants
Disease variants
Proportion
REVEL Score
Cumulative Proportion
REVEL Score
0.25
0.20
0.15
0.10
0.05
0.00
0.05 0.15 0.25 0.35 0.45 0.55 0.65 0.75 0.85 0.95
1.0
0.8
0.6
0.4
0.2
0.0
0.0 0.2 0.4 0.6 0.8 1.0

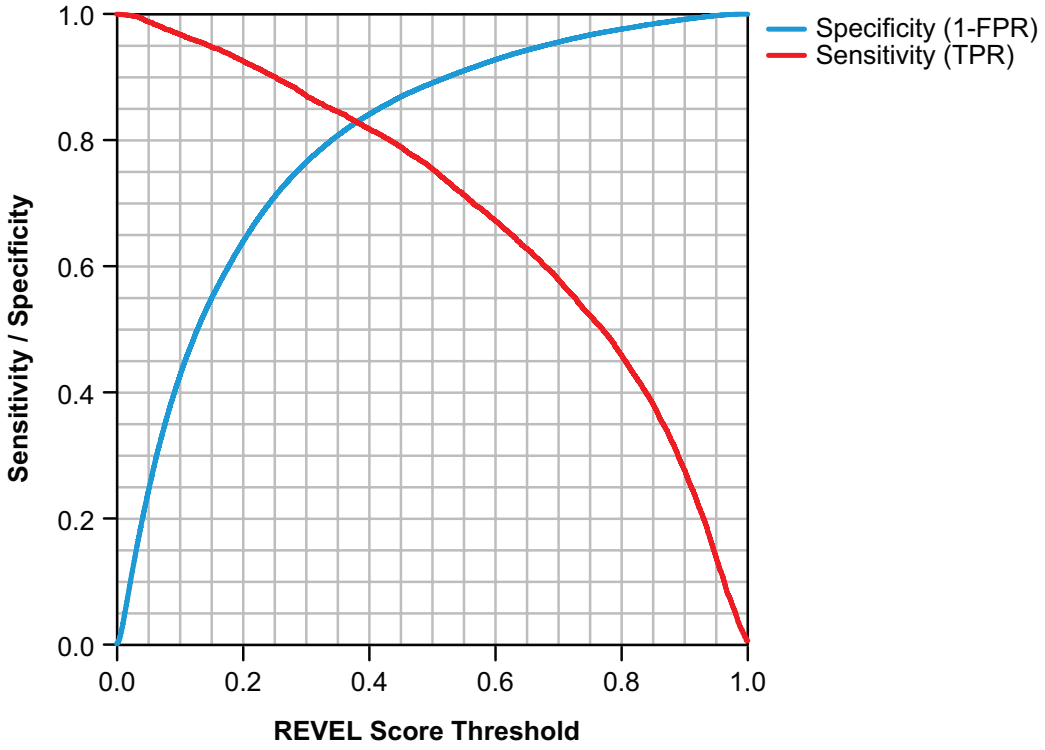


Figure S1. Sensitivity and specificity for different REVEL score thresholds.
REVEL scores were computed using the out-of-bag predictions for 6182 HGMD disease mutations and 123,706 rare (AF 0.001-0.01) putatively neutral exome sequencing variants. FPR = false positive rate, and TPR = true positive rate.

Table S1. AUC of individual prediction tools in the training data and independent test sets.

Individual prediction tool	*Training set*[a]	*Test set 1*[b]	*Test set 2*[c]
VEST	0.844	0.843	0.916
FATHMM	0.824	0.787	0.804
MutPred	0.809	0.783	0.871
Polyphen2 HVAR	0.806	0.779	0.894
MutationAssessor	0.806	0.785	0.885
PROVEAN	0.799	0.792	0.891
phyloP (vertebrate)	0.791	0.741	0.845
Polyphen2 HDIV	0.780	0.752	0.867
LRT	0.765	0.783	0.836
SIFT	0.761	0.745	0.831
SiPhy	0.742	0.728	0.781
phastCons (vertebrate)	0.719	0.720	0.759
MutationTaster	0.717	0.729	0.744
GERP++ RS	0.717	0.687	0.735
phyloP (placental)	0.701	0.675	0.743
phastCons (placental)	0.692	0.722	0.749
phastCons (primate)	0.647	0.691	0.679
phyloP (primate)	0.589	0.593	0.591

AUC = area under the receiver operating characteristic curve.

a. The training set included 6182 novel disease mutations from HGMD, and 123,706 rare (AF 0.001-0.01) putatively neutral exome sequencing variants (ESVs).

b. Test set 1 included 935 novel SwissVar disease variants and 123,935 rare (AF 0.001-0.01) putatively neutral exome sequencing variants (ESVs) that did not overlap with the training set.

c. Test set 2 included 1953 pathogenic and 2406 benign variants recently reported in the ClinVar database.

Table S2. AUC values for discrimination of HGMD disease mutations[a] from neutral exome sequencing variants, by neutral variant allele frequency.

Ensemble method	*All* *n=140,921*	*Neutral variant AF* *0.001-0.003* *n=90,445*	*0.003-0.005* *n=15,502*	*0.005-0.01* *n=17,759*	*0.01-0.03* *n=10,687*	*0.03-0.05* *n=1925*	*>0.05* *n=4603*
REVEL[b]	0.908	0.897	0.917	0.922	0.940	0.958	0.957
MetaLR	0.883	0.867	0.884	0.908	0.927	0.960	0.971
MetaSVM	0.879	0.867	0.882	0.897	0.912	0.931	0.940
KGGSeq 0.8[c]	0.848	0.833	0.859	0.865	0.893	0.917	0.924
Condel 2.0[d]	0.847	0.836	0.855	0.863	0.878	0.905	0.906
Eigen	0.841	0.822	0.854	0. 864	0. 892	0.918	0.929
CADD 1.3	0.798	0.774	0.809	0.823	0.864	0.902	0.921
DANN	0.730	0.703	0.742	0.753	0.806	0.866	0.897

AUC = area under the receiver operating characteristic curve.

a. 6182 HGMD disease mutations with allele frequency distribution: 99.56% <0.01, 0.37% 0.01-0.05, and 0.06% ≥0.05.

b. REVEL scores were computed using out-of-bag (OOB) predictions for variants in the training set, and predictions from all trees in the random forest otherwise.

c. KGGSeq scores were missing for 13 variants.

d. Condel scores were missing for 11,491 variants.

Table S3. AUC values in an independent test set of SwissVar disease mutations[a] and neutral exome sequencing variants, by neutral variant allele frequency.

Ensemble method	***All*** *n=141,051*	***Neutral variant AF***					
		0.001-0.003 *n=90,642*	***0.003-0.005*** *n=15,273*	***0.005-0.01*** *n=18,020*	***0.01-0.03*** *n=10,458*	***0.03-0.05*** *n=1957*	***>0.05*** *n=4701*
REVEL	0.893	0.880	0.902	0.908	0.929	0.951	0.954
MetaLR[b]	0.849	0.830	0.849	0.877	0.900	0.947	0.967
MetaSVM[b]	0.837	0.823	0.841	0.855	0.871	0.898	0.912
KGGSeq 0.8[c]	0.844	0.829	0.856	0.861	0.887	0.910	0.924
Condel 2.0[d]	0.804	0.792	0.813	0.818	0.839	0.864	0.879
Eigen	0.803	0.781	0.820	0.829	0.863	0.893	0.911
CADD 1.3	0.774	0.749	0.792	0.800	0.841	0.880	0.904
DANN	0.726	0.698	0.741	0.752	0.803	0.857	0.890

AUC = area under the receiver operating characteristic curve.

a. 935 SwissVar disease mutations with allele frequency distribution: 99.68% <0.01 and 0.32% 0.01-0.05.

b. MetaLR and MetaSVM scores were missing for 1 variant.

c. KGGSeq scores were missing for 15 variants.

d. Condel scores were missing for 11,642 variants.

Table S4. AUC values in an independent test set of ClinVar disease[a] and neutral variants, by neutral variant allele frequency.

Ensemble method	*All* *n=2406*	*Neutral variant AF* *<0.001* *n=1224*	*0.001-0.005* *n=417*	*0.005-0.01* *n=364*	*0.01-0.03* *n=288*	*>0.03* *n=113*
REVEL	0.957	0.959	0.960	0.946	0.953	0.970
MetaLR	0.917	0.904	0.914	0.930	0.936	0.988
MetaSVM	0.933	0.928	0.933	0.933	0.939	0.976
KGGSeq 0.8	0.893	0.897	0.892	0.881	0.884	0.927
Condel 2.0[b]	0.898	0.900	0.906	0.892	0.890	0.900
Eigen	0.910	0.924	0.907	0.877	0.885	0.930
CADD 1.3	0.902	0.920	0.892	0.862	0.885	0.914
DANN	0.833	0.857	0.817	0.778	0.810	0.871

AUC = area under the receiver operating characteristic curve.

a. 1953 ClinVar disease mutations with allele frequency distribution: 99.95% <0.01 and 0.05% 0.01-0.05.

b. Condel scores were missing for 84 variants.

Table S5. Area under the precision-recall curves for REVEL and other ensemble methods.

Ensemble method	*Training set*[a]	*Test set 1*[b]	*Test set 2*[c]
REVEL	0.465	0.127	0.951
MetaLR	0.386	0.092	0.911
MetaSVM	0.345	0.070	0.924
KGGSeq 0.8	0.374	0.107	0.891
Condel 2.0	0.244	0.038	0.887
Eigen	0.244	0.038	0.868
CADD 1.3	0.121	0.017	0.838
DANN	0.087	0.014	0.745

a. The training set (4.8% disease variants) included 6182 novel disease mutations from HGMD, and 123,706 rare (AF 0.001-0.01) putatively neutral exome sequencing variants (ESVs).

b. Test set 1 (0.7% disease variants) included 935 novel SwissVar disease variants and 123,935 rare (AF 0.001-0.01) putatively neutral exome sequencing variants (ESVs) that did not overlap with the training set.

c. Test set 2 (44.8% disease variants) included 1953 pathogenic and 2406 benign variants recently reported in the ClinVar database.