Title: Prediction of recessive inheritance for missense variants in human disease

Authors: Ben O. Petrazzini[1,2], Daniel J. Balick[1,2,3,4,5], Iain S. Forrest[1,2,6], Judy Cho[1,2,4], Ghislain Rocheleau[1,2], Daniel M. Jordan[1,2*], Ron Do[1,2*]

Affiliations:

1. The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA
2. Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA
3. Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA
4. Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA
5. Department of Biomedical Informatics, Harvard, Medical School, Boston, MA, USA
6. Medical Scientist Training Program, Icahn School of Medicine at Mount Sinai, New York, NY, USA

* These authors jointly supervised this work.

Correspondence:

Ron Do, PhD

Annenberg Building, Floor 18 Room 80A

1468 Madison Ave

New York, NY-10029

Phone Number: 212-241-6206 | Fax Number: 212-849-2643

Email: ron.do@mssm.edu

Abstract

The prediction of pathogenic human missense variants has improved in recent years, but a more granular level of variant characterization is required. Further axes of information need to be incorporated in order to advance the genotype-to-phenotype map. Recent efforts have developed mode of inheritance prediction tools; however, these lack robust validation and their discrimination performance does not support clinical utility, with evidence of them being fundamentally insensitive to recessive acting diseases. Here, we present MOI-Pred, a three-way variant-level mode of inheritance prediction tool aimed at recessive identification for missense variants. MOI-Pred shows strong ability to discriminate missense variants causing autosomal recessive disease (area under the receiver operating characteristic (AUROC)=0.99 and sensitivity=0.85) in an external validation set. Additionally, we introduce an electronic health record (EHR)-based validation approach using real-world clinical data and show that our recessive predictions are enriched for recessive associations with human diseases, demonstrating utility of our method. Mode of inheritance predictions - pathogenic for autosomal recessive (AR) disease, pathogenic for autosomal dominant (AD) disease, or benign – for all possible missense variants in the human genome are available at https://github.com/rondolab/MOI-Pred/.

Introduction

Computational methods to predict the effect of coding variants have numerous applications, such as the diagnosis of genetic diseases[1-4], genetic association studies[5-8], and drug design[9,10]. Currently available methods perform very well at discriminating pathogenic and benign missense variants, typically reporting accuracy in the range of 58-86%[11-13]. Each prediction uses a unique set of variant characteristics and shows different performances across datasets making ensemble approaches more accurate[14-16]. On the same basis, current guidelines recommend considering multiple prediction tools to inform decision making[13]. While these methods may perform very well, they do not consider granularity of a variant's effect on disease. The vast majority of these methods make a simple binary prediction: is the variant pathogenic, or is it benign? Some methods make more specific predictions about whether variants cause particular phenotypes[17], but even these are generally still binary predictions about a single phenotype. The true shape of the genotype-to-phenotype map is much more complex and highly dimensional. A full assessment of a variant's effect on phenotype would include potentially pleiotropic effects on a variety of different phenotypes, from the molecular level to the systems level, as well as features that modify its genetic impact, such as penetrance and mode of inheritance. Large-scale computational approaches that incorporate different axes of genomic information can potentially be used to inform various aspects of variant function[18].

Here, we focus on mode of inheritance as the next level of granularity to include in computational prediction of variant effect. The concept behind mode of inheritance is foundational to the field of genetics and in classical medical genetics it is considered one of the most important features to report about a pathogenic variant[13,19,20]. Studies have shown that disease diagnosis can be largely improved by incorporating pedigree information[21-24]. In spite of this, mode of inheritance has practically no role in current variant annotation pipelines. Efforts to resolve mode of inheritance mechanisms have fallen behind the gene discovery rate[25], limiting the availability of such information in databases of validated clinically relevant variants. Even among databases that do provide mode of inheritance information, most notably Online Mendelian Inheritance in Man (OMIM)[26], these annotations are present only for a small

fraction of 4,417 genes. They are also not necessarily reliable, since they derive almost entirely from anecdotal case reports with small pedigrees, and very few have been replicated across studies. Currently, 35.5% of variants reported as "Pathogenic" or "Likely Pathogenic" in ClinVar[27] have no annotated mode of inheritance and cannot confidently be assigned one based on existing annotations. While some molecular and evolutionary features are known to be enriched in genes implicated in autosomal recessive (AR) disease[28-33], these features are not widely used at the variant level to distinguish variants causing AR disease from variants causing autosomal dominant (AD) disease or benign variants. Additionally, there is some evidence that current variant effect prediction methods may be fundamentally insensitive to AR disease[34,35], reinforcing the need for new methods specifically aimed at predicting variants causing AR disease[36]. Previous efforts at developing such methods underperform binary prediction tools, lack robust validation and have not achieved widespread use in the field[16,37-39].

Here we present MOI-Pred, a three-way prediction method that labels missense variants as pathogenic for AR disease, pathogenic for AD disease, or benign. The method uses a random forest classifier to combine variant effect estimations with gene-level features that are predictive of AR or AD disease. The resulting predictor identifies pathogenic mutations with performance comparable to state-of-the-art binary prediction methods and distinguishes mode of inheritance at the variant level. Moreover, the tool accurately predicts disease case-control status for the three classes of mutations in an external validation using real-world electronic health record (EHR)-based clinical data. MOI-Pred addresses a shortcoming in current annotation pipelines by accurately predicting mode of inheritance, especially differentiating AR pathogenic variants from benign variants, while simultaneously improving granular predictions of variant effect crucial to achieve clinically relevant levels of accuracy.

Results

Clinical variants missing mode of inheritance information

ClinVar does not explicitly annotate mode of inheritance. Instead, this information is extracted from external resources such as OMIM or the Human Gene Mutation Database (HGMD)[40]. These

databases provide mainly gene-level information and only for a subset of diseases. Thus, most variants in ClinVar either lack a mode of inheritance annotation entirely or are simply labelled with the annotation of their corresponding gene. Only 4,126 genes in ClinVar have inheritance information, resulting in 37.63% of variants with undetermined mode of inheritance (**Supplementary Table 2**). Out of a total of 307,800 unlabelled variants, 49,745 are Pathogenic (35.51% of all Pathogenic), 119,532 are Benign (41.38% of all Benign), 122,600 have Uncertain significance (35.30% of all Uncertain significance) and 15,923 have Conflicting interpretation of pathogenicity (38.19% of all Conflicting interpretation) (**Supplementary Table 2**).

Model Training

We collected a training set of 2,481 Recessive and 1,248 Dominant pathogenic missense variants from ExoVar[16] and 3,729 Benign missense variants from gnomAD[41], annotated with a wide range of features capturing functional and biological aspects of mode of inheritance. We fitted a random forest model on this training set, using 10-fold cross-validation and 100 different random train-test splits to assess performance. Feature selection was performed independently on each iteration, reducing the number of features to a minimum of 10 and a maximum of 18 (median across 100 models is 13 features (**Supplementary Figure 1**). In total, 19 unique features were selected across all 100 iterations for training, incorporating a range of functional, evolutionary and combined information (**Methods, Supplementary Table 1**).

The prediction models performed well in the Test set, with a mean area under the receiver operator characteristic (AUROC)=0.94/0.96/0.95 (standard deviation (SD)=1.2 x 10^{-2}/6.8 x 10^{-3} /1.3 x 10^{-2}), sensitivity=0.75/0.76/0.92 (SD=3.8 x 10^{-2}/3.0 x 10^{-2}/2.8 x 10^{-2}), and specificity=0.94/0.95/0.82 (SD=1.3 x 10^{-2}/1.4 x 10^{-2}/2.2 x 10^{-2}) for Recessive/Dominant/Benign classes, respectively (**Figure 2**). This represents good overall performance with similar discrimination power across classes. Benign has higher sensitivity and lower specificity than the other two classes, representing a higher rate of false positives for Benign variants and a higher rate of false negatives for both classes of pathogenic variants.

External Validation

To assess the model's performance on external data sources, we collected an external Validation set containing 261 Recessive and 255 Dominant pathogenic missense variants from ClinVar and 1,010 Benign missense variants from the Japan Whole Genome Aggregation panel v.1 (GEM)[42], in addition to the internal Train/Test set. Performance on this external Validation set was similar to performance measured on the Test set, with AUROC=0.99/0.99/0.96 (SD=2.7×10^{-3}/1.8×10^{-3}/6.9×10^{-3}), sensitivity 0.85/0.87/0.93 (SD 1.8×10^{-2}/2.8×10^{-2}/1.7×10^{-2}), and specificity 0.98/0.99/0.86 (SD 7.4×10^{-3}/4.8×10^{-3}/1.5×10^{-2}) for Recessive/Dominant/Benign classes (**Figure 2**). This suggests that the model is not overfitting the data sources used for training and testing (ExoVar and gnomAD), which would be indicated by a significant drop in performance between the internal blind Test set and external validation. Indeed, the model appears to perform better on the external validation set. This may reflect the fact that the ClinVar datasets used for external validation contain more confident annotations and less noise than the ExoVar datasets used for training and testing.

To expand on this observation, we grouped variants by ClinVar review status (level of evidence for pathogenicity) and assessed how the confidence of variant annotations affects the sensitivity of our predictions. We found that sensitivity improved with higher review status: sensitivity for the Recessive/Dominant classes was 0.58/0.49 (SD 0.03/0.00) for 0-star review status, 0.68/0.59 (SD 0.04/0.00) for 1-star review status, and 0.86/0.89 (SD 0.04/0.00) for 2-star or higher review status (**Supplementary Figure 2**). This is as expected if lower-confidence variants are less likely to be truly pathogenic. In this case, an accurate predictor would predict a smaller fraction of 0-star variants to be pathogenic, because the fraction of those variants that are truly pathogenic is smaller.

We also tested the inheritance prediction model on variants unique to a single ancestry group (European American, African American, or Hispanic American) to evaluate whether performance is consistent across ancestries. We found that sensitivity was uniformly high across all three ancestries, with no specific ancestry having substantially higher power (**Supplementary Figure 3, 4 and 5**). We also

found that ancestry-specific variants across all three ancestries showed the same trend as the full dataset, with sensitivity improving in higher confidence annotations. This demonstrates that MOI-Pred is not primarily powered to detect variants observed in Europeans, but has similar performance regardless of ancestry.

Model interpretation

Examining the importance of different features in the model shows the union of functional, evolutionary and combined information that are driving the inheritance prediction. One functional feature (AD.rank with 23.8%), two combined features (MutPred and MCAP with 14.2% and 13.6%, respectively) and two evolutionary features (OE and FATHMM with 11.2% and 11% respectively) carry 73.8% of the models' weight (**Figure 3. A**).

These feature weights represent the overall importance of features to the three-way classifier. To examine which features are important to identify each individual class, we trained three two-way classifiers to distinguish Dominant from Benign, Recessive from Benign, and Dominant from Recessive. Both Benign-Pathogenic binary prediction models (Benign-Dominant and Benign-Recessive) are dominated by features carrying combined functional and evolutionary information, namely MCAP, MutPred and VEST3, in addition to FATHMM which primarily carries evolutionary information (**Figure 3. B, Supplementary Figure 6 and 7**). In contrast, the Dominant-Recessive prediction is mainly driven by gene-level features carrying either functional or evolutionary information like AD.rank and O/E (**Figure 3. B, Supplementary Figure 8**).

Clinical validation using EHR

To test the performance of the model on real-world clinical data, we collected a total of 1,845,623 variants present in patients from the Bio*Me* biobank[43]. Of these, 56,706 were missense variants present in ClinVar (2,301 Pathogenic, 9,865 Benign, 35,629 Uncertain significance and 8,911 Conflicting interpretation), and 19,134 remain after restricting to 2-star or higher in ClinVar review status (1,047

Pathogenic, 6,303 Benign and 11,784 Uncertain significance) (**Supplementary Table 3**). The model used to predict all variants shows good performance in the Train/Test and Validation sets (**Supplementary Table 4**). For each variant, we marked each patient as positive if the EHR included a diagnosis reported for the variant in ClinVar, and negative otherwise. We then used a Cochran-Mantel-Haenszel (CMH) stratified contingency test to assess the association between homozygous or heterozygous carriers of ClinVar-annotated variants and actual diagnoses, stratified by disease. An association between carrier status and disease status indicates that the variants being tested are, in aggregate, associated with disease with the specified mode of inheritance. By separating variants that receive different predictions from our model, we can test whether our model's prediction is actually predictive of carrier disease status in a real clinical population.

The contingency table analysis showed that our model is highly predictive, with all ClinVar categories showing associations in the expected directions. We found that a Recessive prediction significantly increases the association between homozygous carrier status and disease status for all ClinVar Pathogenic (OR=4.30 [95% CI=4.07 to 4.55] and OR=1.07 [95% CI=1.04 to 1.09]), Uncertain Significance (OR=5.45 [95% CI=5.13 to 5.77] and OR=0.31 [95% CI=0.30 to 0.32]) and Conflicting Interpretation (OR=4.11 [95% CI=3.94 to 4.28] and OR=1.11 [95% CI=1.09 to 1.13]) annotations. A Dominant prediction significantly increases the association between homozygous or heterozygous carrier status and disease status for ClinVar Pathogenic (Odds ratio (OR)=1.98 for MOI-Pred's prediction [95% confidence interval (CI)=1.96 to 2.00] and OR=1.56 [95% CI=1.55 to 1.57] for all other variants) and Uncertain Significance (OR=1.40 [95% CI=1.39 to 1.41 and OR=0.87 [95% CI=0.87 to 0.87]) annotations. And as expected, a Benign prediction significantly decreases the association between homozygous or heterozygous carrier status and disease status for ClinVar Pathogenic (OR=1.23 [95% CI=1.22 to 1.24] and OR=2.97 [95% CI=2.94 to 2.99]) and Uncertain Significance (OR=0.87 [95% CI=0.87 to 0.88] and OR=1.02 [95% CI=1.01 to 1.02]) annotations (**Figure 4, Supplementary Table 5-7**). Restricting to ClinVar variants with 2-star or higher review status showed similar results (**Supplementary Figure 9, Supplementary Table 8-10**). Notably, we observed a particularly strong

protective association of variants on disease that are not predicted recessive for "Uncertain Significance" variants (**Figure 4C**).

To compare MOI-Pred with a previously developed mode of inheritance prediction tool (MAPPIN), we also performed the same EHR-based clinical validation analyses using MAPPIN predictions (**Supplementary Figure 10, Supplementary Table 11-13**). MAPPIN shows weaker enrichment for recessive association with disease of ClinVar Pathogenic predicted recessive (OR=3.61 [95% CI=3.41 to 3.80] for MAPPIN vs. OR=4.31 [95% CI=4.07 to 4.55] for MOI-Pred) and no enrichment for recessive association with disease of ClinVar Uncertain Significance and Conflicting Interpretation variants (OR=0.35 [95% CI=0.33 to 0.36] for MAPPIN vs. OR=5.45 [95% CI=5.13 to 5.77] for MOI-Pred and OR=1.15 [95% CI=1.11 to 1.18] for MAPPIN vs. OR=4.11 [95% CI=3.94 to 4.28] for MOI-Pred, respectively) (**Supplementary Figure 10, Supplementary Table 11**). As expected, MAPPIN shows similar or stronger enrichment for dominant association with disease for various ClinVar classes (**Supplementary Figure 10, Supplementary Table 12**).

Discovery of individual variants

To test the utility of MOI-Pred for clinical assessment of individual variants, we performed single variant association tests in the Bio*Me* biobank and identified 18 variants showing significant associations with a single phenotype (p-value corrected for 455 recessive association tests=1.09×10^{-4}; p-value corrected for 455 recessive association tests for 6,382 dominant association tests=7.83×10^{-6}) (**Table 1 and Supplementary Table 14-17**). Three variants were found to have significant recessive associations with disease. Interestingly, none of these are labelled as Pathogenic in ClinVar as two are labelled Benign and one is labelled Conflicting Interpretation of Pathogenicity. Moreover, 15 variants showed dominant association with disease, 12 of which do not correspond with their clinical annotation using ClinVar/OMIM, showing the potential utility of MOI-Pred for discovery of novel associations with disease (**Table 1**).

Discussion

Here we present MOI-Pred, a computational tool that jointly predicts pathogenicity and mode of inheritance for missense variants. Our tool uses a random forest classifier trained on known variants to combine multiple sources of annotation into a single prediction. Compared to other existing methods, MOI-Pred benefits from several key innovations. First and foremost, where most methods produce a binary prediction of pathogenic or benign, our method produces a three-way prediction, classifying each variant as pathogenic for AR disease, pathogenic for AD disease, or benign. In particular, while many existing methods perform well at predicting pathogenic variants in AD disease (e.g. O/E [41], CADD [44], phyloP [45], etc.), MOI-Pred specifically targets the problem of discriminating AR pathogenic variants from benign, a long-lasting issue in genetics unattended by current annotation pipelines. Only one pre-existing method makes three-class predictions, MAPPIN[38]. MOI-Pred performs substantially better than MAPPIN at identifying variants associated with recessive-acting diseases in both prediction performance (Precision=0.79 on the training set for MAPPIN recessive predictions based on Gosalia et al. 2017[38]) and when testing on real-world clinical data in the current study.

Second, MOI-Pred combines evolutionary and functional annotations on both the gene and variant level to produce a combined variant-level prediction. This gives an important advantage in predicting mode of inheritance, since different annotation sources are known to have different error profiles. In particular, it has recently been shown that evolutionary scores are primarily sensitive to heterozygote effects, making these methods very likely to misclassify AR pathogenic variants as benign[34,35]. Most predictors of pathogenicity rely primarily on these scores, and therefore may be systematically insensitive to AR pathogenic variants. By combining multiple different scores in a random forest framework, MOI-Pred is able to learn which scores are most sensitive to each mode of inheritance. For example, O/E, which relies exclusively on evolutionary constraint, is very likely to confuse AR with benign, while MutPred, which incorporates biophysical properties of proteins[46,47], is more likely to categorize AR variants as pathogenic. Accordingly, in our method, O/E is an important feature separating AD from benign, while MutPred is an important feature separating AR from benign.

Third, MOI-Pred is trained with a population-derived list of benign variants, and validated on novel population-derived benign variants unknown to its constituent scores. Because the most difficult classification task is distinguishing AR from benign, the choice of benign training data is vitally important. Since we use clinically validated pathogenic variants for training, it is tempting to use clinically validated benign variants as well, but this can bias the training set. These were suspected being pathogenic at some point and therefore may have features that are not typical of benign variants[48]. The ideal source of benign variants should be found at sufficiently high frequency in a healthy human population[13,49]. We used frequency-matched variants from a large population database (gnomAD) as presumed benign controls in our training set, an approach that has been used by previous methods such as CADD[44] and VEST3[50,51]. However, using these variants introduces an additional problem: population genetics scores used as components in our prediction model are often themselves derived from the same populations, introducing bias and the risk of overfitting[52]. We addressed this problem by using common variants from a recently published cohort of healthy Japanese adults as a validation set. At the time of analysis, this population had not yet been incorporated into widely used population databases, and all genetics scores were therefore naïve to it. Our classifier performed better on these population-derived benign variants than an equivalent classifier trained on clinically validated benign variants, and also performed well on clinically validated benign variants.

Finally, we validated our method by using it to predict disease case-control status in EHR data from the Bio*Me* biobank[43]. We demonstrated that our predictions of mode of inheritance are significantly associated with the likelihood of carriers developing Mendelian disease in a real-world clinical setting. This is true for variants annotated as pathogenic, variants of unknown significance, as well as novel and ancestry-specific variants. This analysis also revealed that some variants with a Benign prediction appear to protect against disease, particularly in variants with “Uncertain significance” (Figure 4C). Since the “Uncertain significance” ClinVar class necessarily contains variants without clear evidence for or against pathogenicity, it is possible that disease modifier variants or variants with protective effects in heterozygous or homozygous form (underdominance or overdominance) may frequently be classified as

"Uncertain significance." Similarly, these variants lack normal signatures of natural selection and so may be likely to receive a Benign prediction in MOI-Pred. In general, the properties of these variants are not well understood, including whether they can be predicted by computational methods, and further investigation is warranted[53]. We also found individual variants where the prediction made by MOI-Pred differed from their clinical annotations using ClinVar/OMIM, and we verified using the same EHR database that both the pathogenicity and the mode of inheritance predicted by our method is likely to be correct. These analyses demonstrate the applicability of our method to real clinical data and decision-making particularly with respect to large-scale electronic health systems. Such clinical validation is only possible thanks to increasingly available EHR-linked biobanks, and we anticipate it being applied more broadly to variant prediction tools in the future.

Our method has several limitations and areas for future work. First, it remains uncertain whether the performance we observe in the test and validation sets will hold in real applications. Many existing tools have reported similarly high performance in their authors' internal testing and lower performance in unbiased replication analyses[12]. Many have also failed to find clinical utility despite numerically high performance[54,55]. Our EHR-based clinical validation suggests that results will hold[56] in real-world clinical data, but validation in other clinical datasets and by other groups is needed. Second, our three-way predictions, though more complete than typical binary predictions, do not completely account for all forms of mode of inheritance. Phenomena such as incomplete dominance, overdominance, and heterozygote advantage, all of which are well documented in human disease[56-58], are unaccounted for in our simple recessive-dominant-benign classification. Likewise, mode of inheritance itself is far from the only refinement that can be added to variant effect predictions. The field would benefit enormously from methods to predict gain-of-function variants, disease suppressor variants, or uniparental imprinted variants, to name just a few. Third, there is room for innovation and improvement in the method. We used only a subset of applicable ML methods and available features. Furthermore, we focus only on missense variants, primarily because this is the largest class of variation for clinical variants and most applicable to mode of inheritance prediction. Most coding annotations are available for missense variants, while a

much smaller number are available for other forms of variation including synonymous, loss-of-function, non-coding, or multi-nucleotide variants. Ultimately, MOI-Pred is meant to be used in combination with other methods to form a holistic picture of the effects of variants. This is true for even the most widely-used prediction methods, which are rarely relied on individually. We believe that this method, together with MAPPIN and others, will enable variant function prediction to go beyond a binary prediction of pathogenicity so that the picture of variant effects formed by computational annotation begins to resemble the true complexity of actual phenotypes.

Methods

Variant collection

Missense variants from publicly available resources were used to generate all datasets. For the training set, pathogenic variants were obtained from ExoVar[16]. Presumed non-pathogenic variants were selected from the Genome Aggregation Database (gnomAD)[41] v2.1.1. GnomAD variants were chosen to match the allele frequency (AF) of pathogenic variants to within 0.1%, based on minor allele frequency in the entire gnomAD population; singletons were chosen to match variants not present in gnomAD. For the validation set, pathogenic variants with review status “reviewed by expert panel” were selected from ClinVar[27] (release June 2020), and presumed non-pathogenic variants were selected from GEM[42], defined as variants with AF >= 1% in GEM and absent or singleton in gnomAD. Gene-level mode of inheritance information for pathogenic variants was obtained from OMIM[26] (release May 2020).

Variant annotation

Variants were characterized using functional and evolutionary information. ANNOVAR[59] was used to annotate variant-level features. We included all available features from ANNOVAR that could be applied to all or nearly all missense variants. This includes 15 features built on evolutionary information (e.g. phyloP[45], FATHMM[60], GERP[61], PROVEAN[62], etc.), 42 features built on both evolutionary and functional information (e.g. M-CAP[63], CADD[44], VEST3[50], MutationTaster[64], etc.) and 14 population

frequency features (e.g. cg69[65], Kaviar[66], GME[67], etc.). We added to this, 7 gene-level features which were retrieved manually from their original sources. This includes 2 gene level features built on evolutionary information (OE score[41] and s_het[68]), 4 gene level features built on functional information (Episcore[69], AD rank[70], StringAD and StringAR[71]) and 1 gene level feature combining the two annotations (HI[72]). The full list of features and their source of information can be found in **Supplementary Table 1**.

Data trimming and imputation

Features with more than 60% missing values and/or high correlation (Pearson's r >= 0.8) in the training set were removed. In two correlated features, the one with higher mean absolute correlation across all features was removed. Variants with more than 60% missing values in the remaining set of features were removed from both the training and external validation sets. Missing values were imputed first on variant-level features. The resulting dataset was then used to impute gene-level features, ensuring low intra-gene variation in these annotations. A random forest-based algorithm (missForest v1.4[73]) was used for both imputations. The final dataset was comprised of 30 features on 5,872 and 1,526 variants from the training and external validation sets respectively.

Workflow to train inheritance prediction models

A machine learning (ML) approach was used to develop mode of inheritance prediction models. To minimize sampling biases, 100 models were trained, tested and validated using random sets of variants. The workflow is described below for a single iteration. A random sample of 90% of available Dominant variants from the training set plus equal numbers of Recessive and Benign variants constituted a balanced Train set. The remaining variants from the training set were used to sample a balanced 10% Test set. The Validation set consisted of all Dominant variants available in the external validation dataset plus equal numbers of randomly sampled Recessive and Benign variants. Scaling and feature selection (using a wrapper random forest-based approach, recursive feature elimination) were performed on the

Train set using the caret package v6.0.84[74] available in R, then applied accordingly to the Test and Validation sets. A three-class (Recessive, Dominant, Benign) random forest algorithm[75] was then fitted to the Train set using 10-fold cross validation to optimize parameter tuning and limit overfitting. Three two-class random forest algorithms (Dominant vs. Recessive, Dominant vs. Benign, Recessive vs. Benign) were fitted in parallel for subsequent feature importance analyses. Mode of inheritance label was then predicted on the Test and Validation sets to compute performance metrics. This entire procedure was repeated 100 times; reported performance statistics (see Results) correspond to the mean and standard deviation (SD) across all 100 runs.

The AUROC was calculated using the pROC package v1.14.0[76] available in R v3.5.3[77]. To obtain a per-class discrimination metric the remaining two labels were treated as negative classes. Accuracy, sensitivity, specificity and positive/negative predictive values (PPV/NPV) as well as the ML framework was implemented using the caret package.

Three-way variant-level mode of inheritance predictions (pathogenic for autosomal recessive (AR) disease, pathogenic for autosomal dominant (AD) disease, or benign) for all possible missense variants in the human genome build hg38 are available at https://github.com/rondolab/MOI-Pred/.

Clinical validation of inheritance prediction models in electronic health records

A single three-class random forest algorithm was fitted, tested and validated as described above to predict mode of inheritance in genotype data from 29,981 individuals in the Bio*Me* biobank[43]. Bio*Me* is a multiethnic, EHR-linked, clinical care biobank of more than 60,000 samples from individuals recruited at the Mount Sinai Health System between 2007 and 2015. Participants were genotyped using the Illumina Global Screening Array, imputation was performed using the 1000 Genomes Phase 3 reference panel, and genetic ancestry was determined through k-means clustering of principal components. Longitudinal biomedical traits including diagnostic codes and laboratory test results were obtained mainly through ambulatory care practices resulting in a high median number of encounters per patient[78]. Only variants present in ClinVar (release June 2020) were considered for posterior analyses. ClinVar's

391 phenotype information was mapped to 456 categories of International Classification of Disease 10 (ICD-
392 10) diagnostic codes using information from the Systematized Nomenclature of Medicine Clinical Terms
393 (SNOMED-CT)[79] and Orphanet[80].

394 Contingency table analyses were performed to test recessive, dominant and benign models on
395 variants predicted with the corresponding mode of inheritance. Each table evaluates a subset of variants
396 having the same inheritance prediction, same clinical significance label in ClinVar, and mapped to the
397 same set of billing codes. An individual was considered "affected" if diagnosed with an ICD-10 code
398 mapped to the above-mentioned subset of variants. Likewise, an individual was considered a "carrier" if
399 homozygous for the pathogenic allele for the recessive model, homozygous for the pathogenic allele or
400 heterozygous for the dominant model, and homozygous for the pathogenic allele or heterozygous for the
401 benign model. An individual was considered a 'non-carrier' if heterozygous for the recessive model,
402 homozygous for the non-pathogenic allele for the dominant model, and homozygous for the non-
403 pathogenic allele for the benign model. Each 2x2 table of carrier status vs. phenotype case/control status
404 was restricted to individuals from independent ancestries and weighted by the prevalence of the
405 corresponding set of ICD-10 codes in each of the ancestries in the Bio*Me* EHR data. The analysis was
406 repeated twice, once for variants with predicted mode of inheritance corresponding to the model (e.g.
407 variants predicted Recessive when evaluating on the recessive model) and once for all variants not
408 predicted in the respective model (e.g. variants predicted Benign and Dominant when evaluating on the
409 recessive model). Furthermore, a secondary analysis restricting to variants with ClinVar review status of
410 two stars or higher was performed.

411 A CMH test was applied to the ancestry-specific tables from the same predicted mode of
412 inheritance and clinical significance to obtain OR, 95% CI and corresponding p-values. The results were
413 then aggregated across ancestries using an inverse variance meta-analysis. A Q-test was performed to
414 evaluate heterogeneity between ORs of variants predicted and not predicted in the respective models. The
415 stats package v3.6.2[77] was used to perform the CMH test and the metafor package v3.0.2[81] was used for
416 the Q-test.

Single nucleotide variant association discovery

A total of 433 groups of ICD-10 codes were tested for dominant and recessive association with 6,382 variants present in ClinVar, having an ICD-10 code mapping and MOI-Pred prediction. The analysis was performed in individual ancestries (European-American, African-American, Hispanic-American and other ancestries) and meta-analyzed using plink v1.9[82]. Whole exome sequencing data and EHR from the Bio*Me* biobank were used in the analysis; 10 principal components were used as covariates to account for population stratification.

References

1. Yang, Y. *et al.* Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders. *New England Journal of Medicine* **369**, 1502-1511 (2013).
2. Posey, J.E. *et al.* Resolution of Disease Phenotypes Resulting from Multilocus Genomic Variation. *New England Journal of Medicine* **376**, 21-31 (2016).
3. Adams, D.R. & Eng, C.M. Next-Generation Sequencing to Diagnose Suspected Genetic Disorders. *The New England Journal of Medicine* **379**, 1353-1362 (2018).
4. Monies, D. *et al.* Lessons Learned from Large-Scale, First-Tier Clinical Exome Sequencing in a Highly Consanguineous Population. *The American Journal of Human Genetics* **104**, 1182-1201 (2019).
5. Akawi, N. *et al.* Discovery of four recessive developmental disorders using probabilistic genotype and phenotype matching among 4,125 families. *Nature Genetics* **47**, 1363-1369 (2015).
6. Turro, E. *et al.* Whole-genome sequencing of patients with rare diseases in a national health system. *Nature* **583**, 96-102 (2020).
7. Van Hout, C.V. *et al.* Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature* **586**, 749-756 (2020).
8. Do, R. *et al.* Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction. *Nature* **518**, 102-106 (2015).
9. Spreafico, R., Soriaga, L.B., Grosse, J., Virgin, H.W. & Telenti, A. Advances in Genomics for Drug Development. *Genes* **11**(2020).
10. Plenge, R.M., Scolnick, E.M. & Altshuler, D. Validating therapeutic targets through human genetics. *Nature Reviews Drug Discovery* **12**, 581-594 (2013).
11. Dong, C. *et al.* Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Human Molecular Genetics* **24**, 2125-2137 (2015).
12. Li, J. *et al.* Performance evaluation of pathogenicity-computation methods for missense variants. *Nucleic acids research* **46**, 7793-7804 (2018).
13. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine* **17**, 405-423 (2015).
14. Ioannidis, N.M. *et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *The American Journal of Human Genetics* **99**, 877-885 (2016).
15. Alirezaie, N., Kernohan, K.D., Hartley, T., Majewski, J. & Hocking, T.D. ClinPred: Prediction Tool to Identify Disease-Relevant Nonsynonymous Single-Nucleotide Variants. *The American Journal of Human Genetics* **103**, 474-483 (2018).
16. Li, M.-X. *et al.* Predicting Mendelian Disease-Causing Non-Synonymous Single Nucleotide Variants in Exome Sequencing Studies. *PLOS Genetics* **9**, e1003143 (2013).
17. Zhang, X. *et al.* Disease-specific variant pathogenicity prediction significantly improves variant interpretation in inherited cardiac conditions. *Genetics in Medicine* **23**, 69-79 (2021).
18. Lappalainen, T. & MacArthur Daniel, G. From variant to function in human disease genetics. *Science* **373**, 1464-1468 (2021).
19. Claustres, M. *et al.* Recommendations for reporting results of diagnostic genetic testing (biochemical, cytogenetic and molecular genetic). *European Journal of Human Genetics* **22**, 160-170 (2014).
20. MacArthur, D.G. *et al.* Guidelines for investigating causality of sequence variants in human disease. *Nature* **508**, 469-476 (2014).

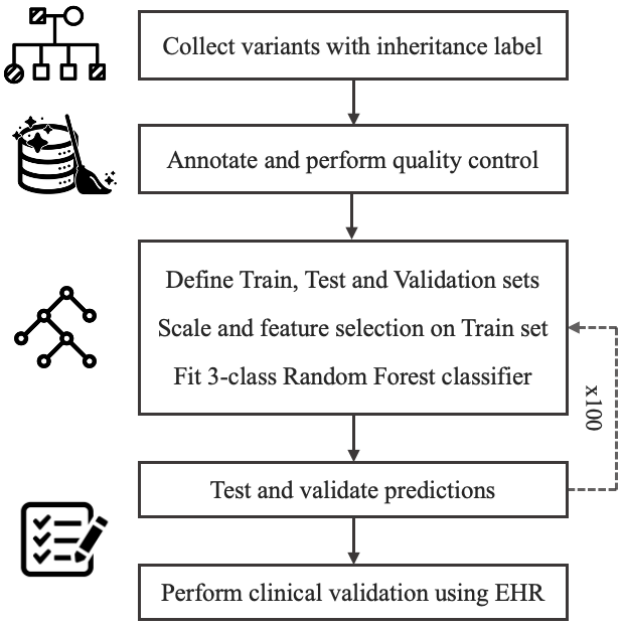
21. Eldomery, M.K. *et al.* Lessons learned from additional research analyses of unsolved clinical exome cases. *Genome Medicine* **9**, 26 (2017).
22. Ewans, L.J. *et al.* Whole-exome sequencing reanalysis at 12 months boosts diagnosis and is cost-effective when applied early in Mendelian disorders. *Genetics in Medicine* **20**, 1564-1574 (2018).
23. Lee, H. *et al.* Clinical Exome Sequencing for Genetic Identification of Rare Mendelian Disorders. *JAMA* **312**, 1880-1887 (2014).
24. Retterer, K. *et al.* Clinical application of whole-exome sequencing across clinical indications. *Genetics in Medicine* **18**, 696-704 (2016).
25. Chong, Jessica X. *et al.* The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *The American Journal of Human Genetics* **97**, 199-215 (2015).
26. Online Mendelian Inheritance in Man, O.M.-N.I.o.G.M., Johns Hopkins University (Baltimore, MD), May 2020. World Wide Web URL: https://omim.org/.
27. Landrum, M.J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* **46**, D1062-d1067 (2018).
28. Furney, S.J., Albà, M.M. & López-Bigas, N. Differences in the evolutionary history of disease genes affected by dominant or recessive mutations. *BMC Genomics* **7**, 165 (2006).
29. Jimenez-Sanchez, G., Childs, B. & Valle, D. Human disease genes. *Nature* **409**, 853-855 (2001).
30. Kondrashov, F.A. & Koonin, E.V. A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications. *Trends in Genetics* **20**, 287-290 (2004).
31. López-Bigas, N., Blencowe, B.J. & Ouzounis, C.A. Highly consistent patterns for inherited human diseases at the molecular level. *Bioinformatics* **22**, 269-277 (2006).
32. Blekhman, R. *et al.* Natural selection on genes that underlie human disease susceptibility. *Curr Biol* **18**, 883-9 (2008).
33. Rapaport, F. *et al.* Negative selection on human genes underlying inborn errors depends on disease outcome and both the mode and mechanism of inheritance. *Proceedings of the National Academy of Sciences* **118**, e2001248118 (2021).
34. Fuller, Z.L., Berg, J.J., Mostafavi, H., Sella, G. & Przeworski, M. Measuring intolerance to mutation in human genetics. *Nature Genetics* **51**, 772-776 (2019).
35. Balick, D.J., Jordan, D.M., Sunyaev, S. & Do, R. Overcoming constraints on the detection of recessive selection in human genes from population frequency data. *bioRxiv*, 2021.05.06.443024 (2021).
36. Antonarakis, S.E. Carrier screening for recessive disorders. *Nature Reviews Genetics* **20**, 549-561 (2019).
37. Furney, S.J., Albà, M.M. & López-Bigas, N. Differences in the evolutionary history of disease genes affected by dominant or recessive mutations. *BMC Genomics* **7**, 165 (2006).
38. Gosalia, N., Economides, A.N., Dewey, F.E. & Balasubramanian, S. MAPPIN: a method for annotating, predicting pathogenicity and mode of inheritance for nonsynonymous variants. *Nucleic Acids Research* **45**, 10393-10402 (2017).
39. Quinodoz, M. *et al.* DOMINO: Using Machine Learning to Predict Genes Associated with Dominant Disorders. *The American Journal of Human Genetics* **101**, 623-629 (2017).
40. Stenson, P.D. *et al.* Human Gene Mutation Database (HGMD): 2003 update. *Hum Mutat* **21**, 577-81 (2003).
41. Karczewski, K.J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434-443 (2020).
42. https://togovar.biosciencedbc.jp/doc/datasets/gem_j_wga., G.J.W.G.A.G.-J.W.P.J.G.M.a.J.P.G.-J.A.f.
43. BioMeTM BioBank Program. https://icahn.mssm.edu/research/ipm/programs/biome-biobank. Accessed June.

44. Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res* **47**, D886-d894 (2019).
45. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20**, 110-21 (2010).
46. Li, B. *et al.* Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics* **25**, 2744-50 (2009).
47. Pejaver, V. *et al.* Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. *Nat Commun* **11**, 5918 (2020).
48. Shah, N. *et al.* Identification of Misclassified ClinVar Variants via Disease Population Prevalence. *American journal of human genetics* **102**, 609-619 (2018).
49. Ghosh, R. *et al.* Updated recommendation for the benign stand-alone ACMG/AMP criterion. *Human mutation* **39**, 1525-1530 (2018).
50. Carter, H., Douville, C., Stenson, P.D., Cooper, D.N. & Karchin, R. Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics* **14 Suppl 3**, S3 (2013).
51. Carter, H., Douville, C., Stenson, P.D., Cooper, D.N. & Karchin, R. Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics* **14 Suppl 3**, S3 (2013).
52. Grimm, D.G. *et al.* The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Hum Mutat* **36**, 513-23 (2015).
53. Rahit, K. & Tarailo-Graovac, M. Genetic Modifiers and Rare Mendelian Disease. *Genes (Basel)* **11**(2020).
54. Ghosh, R., Oak, N. & Plon, S.E. Evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines. *Genome Biol* **18**, 225 (2017).
55. Mahmood, K. *et al.* Variant effect prediction tools assessed using independent, functional assay-based datasets: implications for discovery and diagnostics. *Hum Genomics* **11**, 10 (2017).
56. Wexler, N.S. *et al.* Homozygotes for Huntington's disease. *Nature* **326**, 194-7 (1987).
57. Hughes, A.L. & Nei, M. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335**, 167-170 (1988).
58. Schroeder, S.A., Gaughan, D.M. & Swift, M. Protection against bronchial asthma by CFTR ΔF508 mutation: A heterozygote advantage in cystic fibrosis. *Nature Medicine* **1**, 703-705 (1995).
59. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research* **38**, e164-e164 (2010).
60. Shihab, H.A. *et al.* Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models. *Human Mutation* **34**, 57-65 (2013).
61. Davydov, E.V. *et al.* Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. *PLOS Computational Biology* **6**, e1001025 (2010).
62. Choi, Y., Sims, G.E., Murphy, S., Miller, J.R. & Chan, A.P. Predicting the functional effect of amino acid substitutions and indels. *PLoS One* **7**, e46688 (2012).
63. Jagadeesh, K.A. *et al.* M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nature Genetics* **48**, 1581-1586 (2016).
64. Schwarz, J.M., Cooper, D.N., Schuelke, M. & Seelow, D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods* **11**, 361-2 (2014).
65. Drmanac, R. *et al.* Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays. *Science* **327**, 78 (2010).
66. Glusman, G., Caballero, J., Mauldin, D.E., Hood, L. & Roach, J.C. Kaviar: an accessible system for testing SNV novelty. *Bioinformatics* **27**, 3216-7 (2011).
67. Scott, E.M. *et al.* Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery. *Nature genetics* **48**, 1071-1076 (2016).

68. Cassa, C.A. *et al.* Estimating the selective effects of heterozygous protein-truncating variants from human exome data. *Nature Genetics* **49**, 806-810 (2017).
69. Han, X. *et al.* Distinct epigenomic patterns are associated with haploinsufficiency and predict risk genes of developmental disorders. *Nature Communications* **9**, 2138 (2018).
70. Hsu, J.S. *et al.* Inheritance-mode specific pathogenicity prioritization (ISPP) for human protein coding genes. *Bioinformatics* **32**, 3065-3071 (2016).
71. Szklarczyk, D. *et al.* STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* **47**, D607-d613 (2019).
72. Huang, N., Lee, I., Marcotte, E.M. & Hurles, M.E. Characterising and Predicting Haploinsufficiency in the Human Genome. *PLOS Genetics* **6**, e1001154 (2010).
73. Stekhoven, D.J. & Bühlmann, P. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28**, 112-118 (2012).
74. Kuhn, M. Building Predictive Models in R Using the caret Package. *Journal Of Statistical Software* **28**(2008).
75. Liaw, A. & Wiener, M. Classification and Regression by randomForest. *R News* **2**, 18-22 (2002).
76. Robin, X. *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**(2011).
77. R Core Team. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria.* (2019).
78. Tayo, B.O. *et al.* Genetic Background of Patients from a University Medical Center in Manhattan: Implications for Personalized Medicine. *PLOS ONE* **6**, e19166 (2011).
79. International Health Terminology Standards Development Organisation. SNOMED CT Starter Guide, Accessed June 2020. https://confluence.ihtsdotools.org/display/DOCSTART/SNOMED+CT+Starter+Guide.
80. Pavan, S. *et al.* Clinical Practice Guidelines for Rare Diseases: The Orphanet Database. *PLoS One* **12**, e0170365 (2017).
81. Viechtbauer, W. Conducting Meta-Analyses in R with the metafor Package. *2010* **36**, 48 (2010).
82. Chang, C.C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**(2015).

624 **Author Contributions**: Dr. Do, Dr. Jordan and Mr. Petrazzini had full access to all of the data in the
625 study and take responsibility for the integrity of the data and accuracy of the data analysis.
626
627 *Concept and design:* Petrazzini, Jordan, Do.
628 *Acquisition, analysis, or interpretation of the data:* All authors.
629 *Drafting of the manuscript:* Petrazzini, Balick, Forrest, Rocheleau, Jordan and Do.
630 *Critical revision of the manuscript for important intellectual concept:* All authors.
631 *Statistical analysis:* Petrazzini, Rocheleau, Jordan, Do.
632 *Obtained funding:* Do.
633 *Administrative, technical, or material support:* Cho, Do.
634 *Supervision:* Jordan and Do.
635
636 **Conflict of Interest Disclosures**: Dr. Do reported receiving grants from AstraZeneca, grants and non-
637 financial support from Goldfinch Bio, being a scientific co-founder, consultant and equity holder for
638 Pensieve Health, and being a consultant for Variant Bio.
639
640 **Funding/Support**: Mr. Forrest is supported by the National Institute of General Medical Sciences of the
641 National Institutes of Health (NIH) (T32-GM007280). Dr. Do is supported by the National Institute of
642 General Medical Sciences of the NIH (R35-GM124836) and the National Heart, Lung, and Blood
643 Institute of the NIH (R01-HL139865 and R01-HL155915).
644
645 **Disclaimer**: The content is solely the responsibility of the authors and does not necessarily represent the
646 official views of the National Institutes of Health.
647
648 **Data sharing statement**: Not applicable.
649

Figure 1. Study design and machine learning workflow.

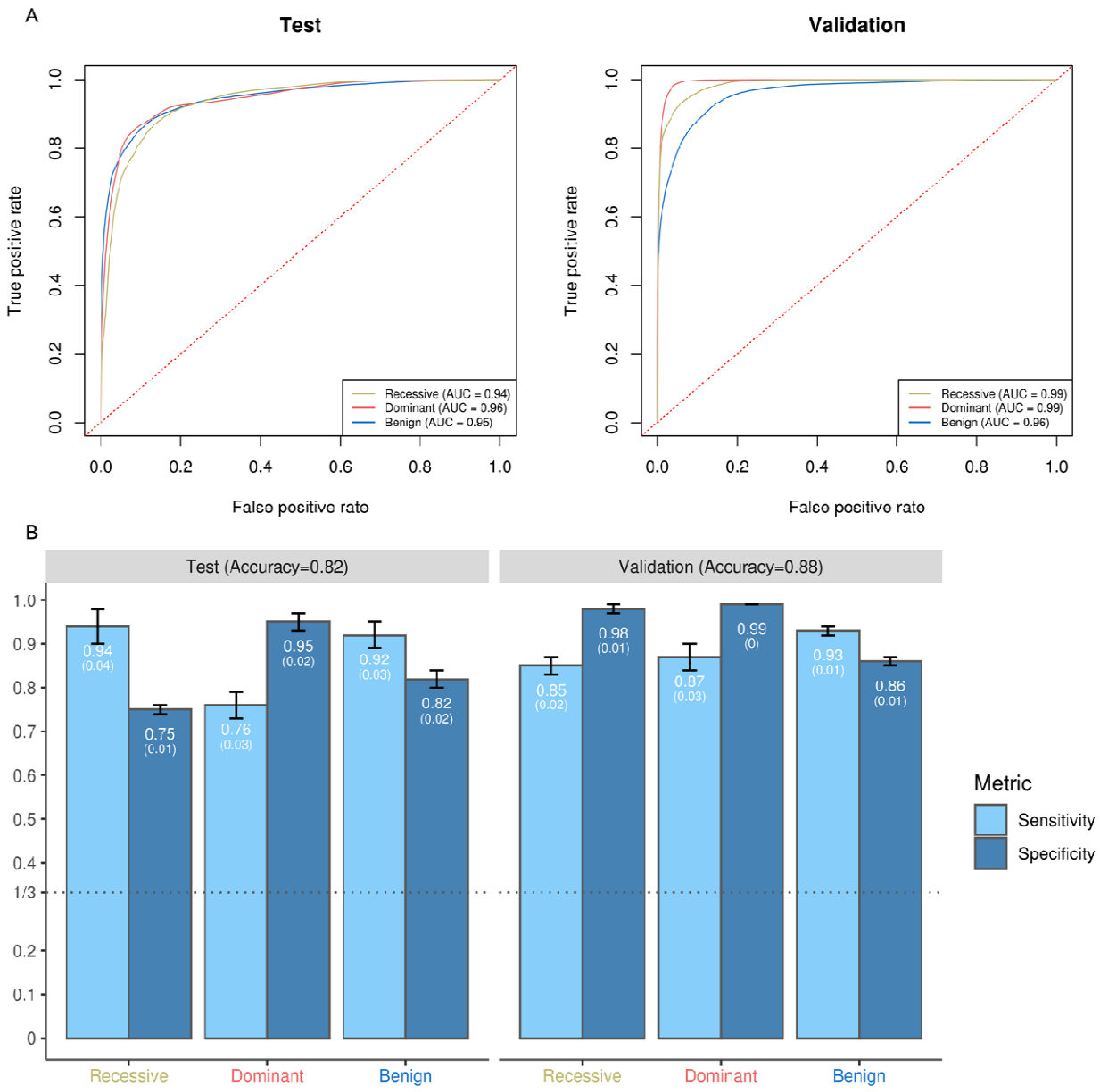


Train and Test sets correspond to the 90% and 10% balanced datasets, built from ExoVar and gnomAD variants, used for training and testing respectively. Validation set corresponds to the balanced dataset, built from ClinVar and GEM variants, used for external validation. EHR corresponds to electronic health records.

666 **Figure 2**. Receiver operator characteristic curves for 3-class mode of inheritance prediction models (A).

667 Bar-plots showing sensitivity and specificity for 3-class mode of inheritance prediction models (B).

668



669

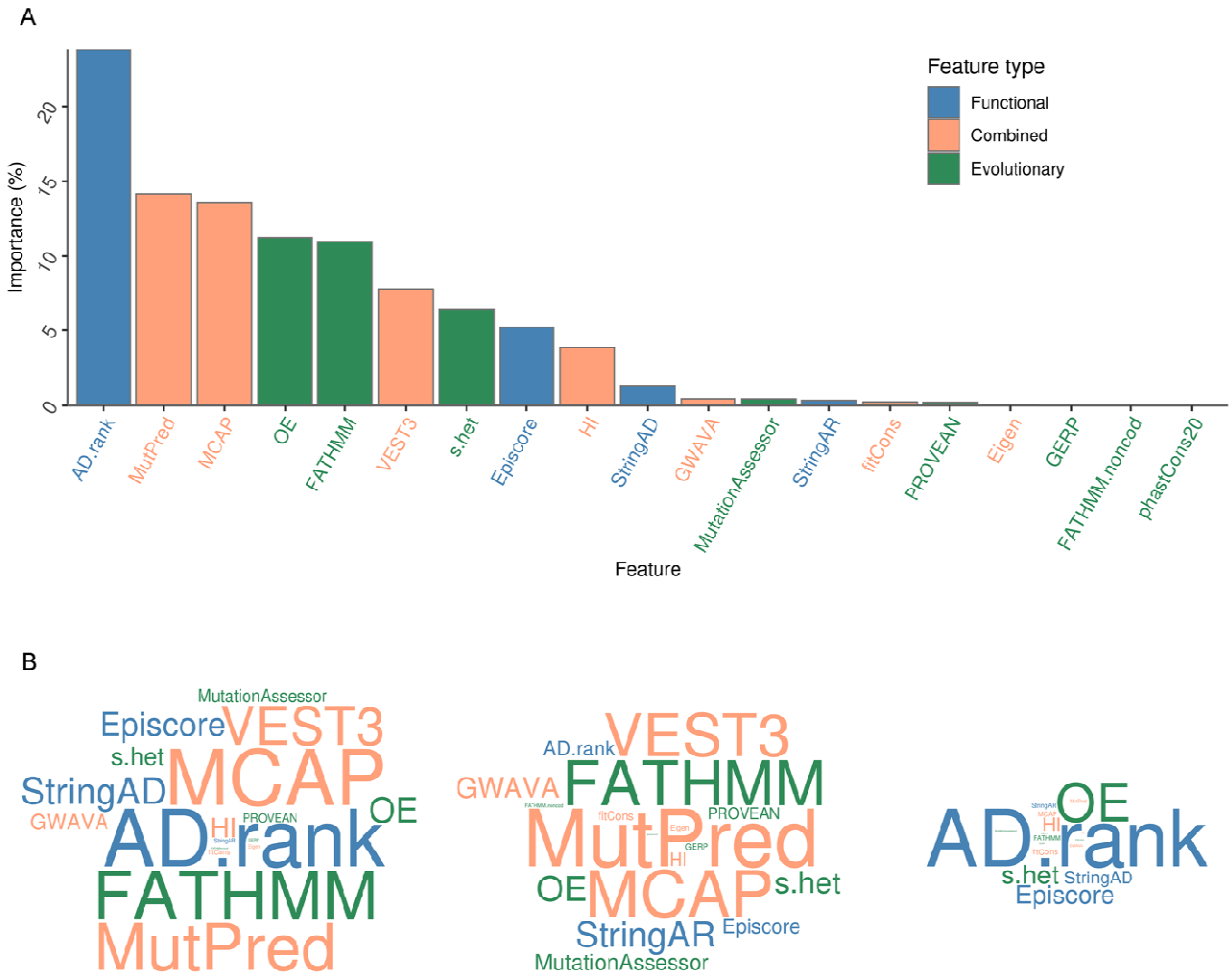
670

671

672

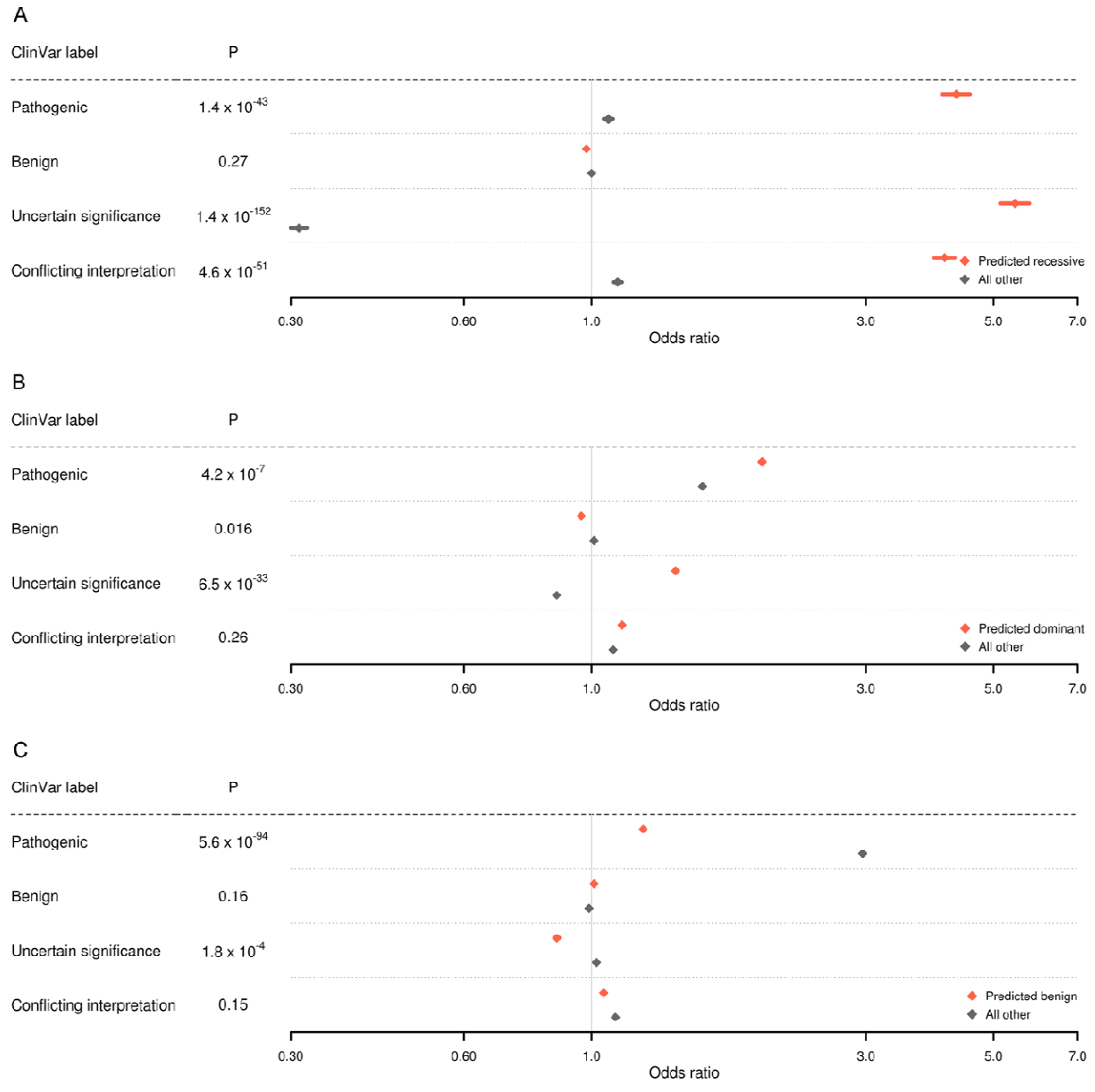
Test set corresponds to the 10% balanced dataset, built from ExoVar and gnomAD variants, used for testing. Validation set corresponds to the balanced dataset, built from ClinVar and GEM variants, used for external validation. Reported AUC corresponds to the mean across 100 models. Reported Sensitivity and Specificity corresponds to mean (standard deviation) across 100 models. AUC corresponds to area under the receiver operator characteristic curve.

Figure 3: Bar-plot showing feature importance on 3-class mode of inheritance prediction models (A). Word-clouds representing feature importance on 2-class mode of inheritance prediction models (B).



Feature importance is reported as the median across 100 models. Word-clouds represent feature importance on Benign-Dominant (left), Benign-Recessive (middle) and Dominant-Recessive (right) models respectively. Exact feature importance values in 2-class prediction models can be found in Supplementary Figure 6-8.

Figure 4: Forest plots showing disease association with variants predicted to be Recessive (A), Dominant (B) and Benign (C).



Effect sizes (Odds ratios) and 95% confidence intervals were obtained for individual ancestries using a Cochran-Mantel-Haenszel (CMH) test. The reported effect sizes correspond to an inverse variance meta-analysis across ancestries. P values for heterogeneity between Odds ratios are derived from a Q-test Test set.

744

745 **Table 1:** Description of significant variant associations with disease.

746

Variant ID	Gene	AF	ICD10 codes	P value	OR (95% CI)	MOI-Pred	OMIM	ClinVar
rs79985808	SUMF1	0.016	Other degenerative diseases of basal ganglia (G23) and Disorders of sphingolipid metabolism and other lipid storage disorders (E75)	4.03×10^{-5}	126.7 (118.5 to 134.9)	AR	Recessive	Benign
rs17144835	DNAH11	0.061	Other congenital malformations of respiratory system (Q34)	4.04×10^{-7}	35.9 (31.9 to 39.9)	AR	Recessive	Benign
rs1800562	HFE	0.022	Disorders of mineral metabolism (E83) and Genetic susceptibility to disease (Z15)	1.10×10^{-9}	13.7 (10.7 to 16.7)	AR	Conflicting	Conflicting interpretations of pathogenicity
rs145214720	COL10A1	3.41×10^{-4}	Osteochondrodysplasias (Q78) and osteochondrodysplasia with defects of growth of tubular bones and spine (Q77)	3.64×10^{-8}	352.3 (348.9 to 355.7)	AD	Dominant	Likely benign
rs34539681	COL10A1	3.90×10^{-3}		2.03×10^{-6}	23.4 (15.5 to 31.3)	AD	Dominant	Benign
rs140075817	EXT2	2.27×10^{-4}		2.39×10^{-6}	248.1 (239.8 to 256.4)	AD	Conflicting	Conflicting interpretations of pathogenicity
rs770821909	EXT2	1.30×10^{-4}		2.36×10^{-6}	471.8 (461.1 to 482.4)	AD	Conflicting	Uncertain significance
rs146098187	EXT2	3.57×10^{-4}		3.17×10^{-6}	262.8 (254.2 to 271.4)	AD	Conflicting	Benign
rs138495222	EXT2	3.57×10^{-4}		5.93×10^{-7}	161.0 (153.9 to 168.1)	AD	Conflicting	Conflicting interpretations of pathogenicity
rs35221558	LEMD3	1.99×10^{-3}		4.24×10^{-7}	40.3 (36.1 to 44.5)	AD	Dominant	Likely benign
rs36105360	LMNB1	9.89×10^{-3}	Degenerative diseases of basal ganglia (G23) and disorders of sphingolipid metabolism and other lipid storage disorders (E75)	7.74×10^{-7}	7.7 (5.4 to 9.9)	AD	Dominant	Benign
rs139644798	RARS1	3.40×10^{-4}		2.39×10^{-6}	98.1 (91.5 to 104.6)	AD	Recessive	Likely pathogenic
rs34637584	LRRK2	1.78×10^{-3}	Neoplasms of unspecified behavior (D49), parkinson's	1.61×10^{-6}	4.7 (2.4 to 6.9)	AD	Dominant	Pathogenic

			disease (G20) and leprosy [Hansen's disease] (A30)					
rs141230910	SDHB	5.84×10^{-4}	Genetic susceptibility to disease (Z15), phakomatoses (Q85), malignant neoplasm of other endocrine glands and related structures (C74), neoplasm of uncertain behavior of endocrine glands (D44) and malignant neoplasm of other and ill-defined digestive organs (C26)	1.98×10^{-6}	18.4 (14.9 to 21.8)	AD	Dominant	Conflicting interpretations of pathogenicity
rs372115732	TBX4	1.13×10^{-4}	Primary disorders of muscles (G71) and congenital malformations of limb(s) (Q74)	8.81×10^{-7}	228.6 (220.1 to 237.1)	AD	Conflicting	Likely benign
rs141707850	FBN2	1.29×10^{-4}	Other congenital musculoskeletal deformities (Q68) and other specified congenital malformation syndromes affecting multiple systems (Q87)	1.34×10^{-6}	153.3 (145.6 to 160.9)	AD	Dominant	Uncertain significance
rs147272790	MBD5	2.75×10^{-4}	Nonrheumatic aortic valve disorders (I35), congenital malformations of great arteries (Q25), monosomies and deletions from the autosomes, not elsewhere classified (Q93) and other congenital malformations of skin (Q82)	1.42×10^{-6}	19.2 (15.6 to 22.8)	AD	Dominant	Conflicting interpretations of pathogenicity
rs77375493	JAK2	6.00×10^{-4}	Other venous embolism and thrombosis (I82), polycythemia vera (D45), mast cell neoplasms of uncertain behavior (D47), myeloid leukemia (C92) and other and unspecified diseases of blood and blood-forming organs (D75)	3.84×10^{-13}	18.9 (16.6 to 21.2)	AD	Dominant	Pathogenic

754

755

756 The significance threshold is set to $p=1.09 \times 10^{-4}$ for the recessive association test and $p=7.83 \times 10^{-6}$ for
757 the dominant association test after a Bonferroni correction based on 455 and 6,382 tests respectively. AF
758 corresponds to allele frequency, OR corresponds to odds ratio, CI corresponds to confidence interval,
759 MOI-Pred corresponds to Mode Of Inheritance Predictor, AR corresponds to variants having autosomal
760 recessive prediction in MOI-Pred, AD corresponds to variants having autosomal dominant prediction in
761 MOI-Pred, OMIM corresponds to Online Mendelian Inheritance in Man, Conflicting corresponds to
762 genes having autosomal dominant and autosomal recessive inheritance label in OMIM, Dominant
763 corresponds to genes having autosomal dominant inheritance label in OMIM, Recessive corresponds to
764 genes having autosomal recessive inheritance label in OMIM, Conf. Int. corresponds to variants having
765 conflicting interpretation of pathogenicity label in ClinVar, Benign corresponds to variants having
766 Benign, Likely benign and/or Benign/Likely benign label in ClinVar, Pathogenic corresponds to variants
767 having Pathogenic, Likely pathogenic and/or Pathogenic/Likely pathogenic label in ClinVar.

768