

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

2

3 **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

4 Rocheleau^{1,2}, Daniel M. Jordan^{1,2*}, Ron Do^{1,2*}

5

6 **Affiliations:**

7 1. The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai,
8 New York, NY, USA

9 2. Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York,
10 NY, USA

11 3. Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA

12 4. Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

13 5. Department of Biomedical Informatics, Harvard, Medical School, Boston, MA, USA

14 6. Medical Scientist Training Program, Icahn School of Medicine at Mount Sinai, New York, NY, USA

15 * These authors jointly supervised this work.

16

17 **Correspondence:**

18 Ron Do, PhD

19 Annenberg Building, Floor 18 Room 80A

20 1468 Madison Ave

21 New York, NY-10029

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

23 Email: ron.do@mssm.edu

24

25

26

27 **Abstract**

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

42

43

44

45

46

47

48

49

50

51

52

53 **Introduction**

54 Computational methods to predict the effect of coding variants have numerous applications, such
55 as the diagnosis of genetic diseases¹⁻⁴, genetic association studies⁵⁻⁸, and drug design^{9,10}. Currently
56 available methods perform very well at discriminating pathogenic and benign missense variants, typically
57 reporting accuracy in the range of 58-86%¹¹⁻¹³. Each prediction uses a unique set of variant characteristics
58 and shows different performances across datasets making ensemble approaches more accurate¹⁴⁻¹⁶. On the
59 same basis, current guidelines recommend considering multiple prediction tools to inform decision
60 making¹³. While these methods may perform very well, they do not consider granularity of a variant's
61 effect on disease. The vast majority of these methods make a simple binary prediction: is the variant
62 pathogenic, or is it benign? Some methods make more specific predictions about whether variants cause
63 particular phenotypes¹⁷, but even these are generally still binary predictions about a single phenotype. The
64 true shape of the genotype-to-phenotype map is much more complex and highly dimensional. A full
65 assessment of a variant's effect on phenotype would include potentially pleiotropic effects on a variety of
66 different phenotypes, from the molecular level to the systems level, as well as features that modify its
67 genetic impact, such as penetrance and mode of inheritance. Large-scale computational approaches that
68 incorporate different axes of genomic information can potentially be used to inform various aspects of
69 variant function¹⁸.

70 Here, we focus on mode of inheritance as the next level of granularity to include in computational
71 prediction of variant effect. The concept behind mode of inheritance is foundational to the field of
72 genetics and in classical medical genetics it is considered one of the most important features to report
73 about a pathogenic variant^{13,19,20}. Studies have shown that disease diagnosis can be largely improved by
74 incorporating pedigree information²¹⁻²⁴. In spite of this, mode of inheritance has practically no role in
75 current variant annotation pipelines. Efforts to resolve mode of inheritance mechanisms have fallen
76 behind the gene discovery rate²⁵, limiting the availability of such information in databases of validated
77 clinically relevant variants. Even among databases that do provide mode of inheritance information, most
78 notably Online Mendelian Inheritance in Man (OMIM)²⁶, these annotations are present only for a small

79 fraction of 4,417 genes. They are also not necessarily reliable, since they derive almost entirely from
80 anecdotal case reports with small pedigrees, and very few have been replicated across studies. Currently,
81 35.5% of variants reported as “Pathogenic” or “Likely Pathogenic” in ClinVar²⁷ have no annotated mode
82 of inheritance and cannot confidently be assigned one based on existing annotations. While some
83 molecular and evolutionary features are known to be enriched in genes implicated in autosomal recessive
84 (AR) disease²⁸⁻³³, these features are not widely used at the variant level to distinguish variants causing AR
85 disease from variants causing autosomal dominant (AD) disease or benign variants. Additionally, there is
86 some evidence that current variant effect prediction methods may be fundamentally insensitive to AR
87 disease^{34,35}, reinforcing the need for new methods specifically aimed at predicting variants causing AR
88 disease³⁶. Previous efforts at developing such methods underperform binary prediction tools, lack robust
89 validation and have not achieved widespread use in the field^{16,37-39}.

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

100

101 **Results**

102 *Clinical variants missing mode of inheritance information*

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

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

113

114 *Model Training*

115 We collected a training set of 2,481 Recessive and 1,248 Dominant pathogenic missense variants
116 from ExoVar¹⁶ and 3,729 Benign missense variants from gnomAD⁴¹, annotated with a wide range of
117 features capturing functional and biological aspects of mode of inheritance. We fitted a random forest
118 model on this training set, using 10-fold cross-validation and 100 different random train-test splits to
119 assess performance. Feature selection was performed independently on each iteration, reducing the
120 number of features to a minimum of 10 and a maximum of 18 (median across 100 models is 13 features
121 (**Supplementary Figure 1**). In total, 19 unique features were selected across all 100 iterations for
122 training, incorporating a range of functional, evolutionary and combined information (**Methods,**
123 **Supplementary Table 1**).

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

131

132 *External Validation*

133 To assess the model's performance on external data sources, we collected an external Validation
134 set containing 261 Recessive and 255 Dominant pathogenic missense variants from ClinVar and 1,010
135 Benign missense variants from the Japan Whole Genome Aggregation panel v.1 (GEM)⁴², in addition to
136 the internal Train/Test set. Performance on this external Validation set was similar to performance
137 measured on the Test set, with AUROC=0.99/0.99/0.96 (SD=2.7 x 10⁻³/1.8 x 10⁻³/6.9 x 10⁻³), sensitivity
138 0.85/0.87/0.93 (SD 1.8 x 10⁻²/2.8 x 10⁻²/1.7 x 10⁻²), and specificity 0.98/0.99/0.86 (SD 7.4 x 10⁻³/4.8 x 10⁻³/
139 3/1.5 x 10⁻²) for Recessive/Dominant/Benign classes (**Figure 2**). This suggests that the model is not
140 overfitting the data sources used for training and testing (ExoVar and gnomAD), which would be
141 indicated by a significant drop in performance between the internal blind Test set and external validation.
142 Indeed, the model appears to perform better on the external validation set. This may reflect the fact that
143 the ClinVar datasets used for external validation contain more confident annotations and less noise than
144 the ExoVar datasets used for training and testing.

145 To expand on this observation, we grouped variants by ClinVar review status (level of evidence
146 for pathogenicity) and assessed how the confidence of variant annotations affects the sensitivity of our
147 predictions. We found that sensitivity improved with higher review status: sensitivity for the
148 Recessive/Dominant classes was 0.58/0.49 (SD 0.03/0.00) for 0-star review status, 0.68/0.59 (SD
149 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
150 (**Supplementary Figure 2**). This is as expected if lower-confidence variants are less likely to be truly
151 pathogenic. In this case, an accurate predictor would predict a smaller fraction of 0-star variants to be
152 pathogenic, because the fraction of those variants that are truly pathogenic is smaller.

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

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

161

162 *Model interpretation*

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

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

177

178 *Clinical validation using EHR*

179 To test the performance of the model on real-world clinical data, we collected a total of 1,845,623
180 variants present in patients from the BioMe biobank⁴³. Of these, 56,706 were missense variants present in
181 ClinVar (2,301 Pathogenic, 9,865 Benign, 35,629 Uncertain significance and 8,911 Conflicting
182 interpretation), and 19,134 remain after restricting to 2-star or higher in ClinVar review status (1,047

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

193 The contingency table analysis showed that our model is highly predictive, with all ClinVar
194 categories showing associations in the expected directions. We found that a Recessive prediction
195 significantly increases the association between homozygous carrier status and disease status for all
196 ClinVar Pathogenic (OR=4.30 [95% CI=4.07 to 4.55] and OR=1.07 [95% CI=1.04 to 1.09]), Uncertain
197 Significance (OR=5.45 [95% CI=5.13 to 5.77] and OR=0.31 [95% CI=0.30 to 0.32]) and Conflicting
198 Interpretation (OR=4.11 [95% CI=3.94 to 4.28] and OR=1.11 [95% CI=1.09 to 1.13]) annotations. A
199 Dominant prediction significantly increases the association between homozygous or heterozygous carrier
200 status and disease status for ClinVar Pathogenic (Odds ratio (OR)=1.98 for MOI-Pred's prediction [95%
201 confidence interval (CI)=1.96 to 2.00] and OR=1.56 [95% CI=1.55 to 1.57] for all other variants) and
202 Uncertain Significance (OR=1.40 [95% CI=1.39 to 1.41] and OR=0.87 [95% CI=0.87 to 0.87])
203 annotations. And as expected, a Benign prediction significantly decreases the association between
204 homozygous or heterozygous carrier status and disease status for ClinVar Pathogenic (OR=1.23 [95%
205 CI=1.22 to 1.24] and OR=2.97 [95% CI=2.94 to 2.99]) and Uncertain Significance (OR=0.87 [95%
206 CI=0.87 to 0.88] and OR=1.02 [95% CI=1.01 to 1.02]) annotations (**Figure 4, Supplementary Table 5-**
207 **7**). Restricting to ClinVar variants with 2-star or higher review status showed similar results
208 (**Supplementary Figure 9, Supplementary Table 8-10**). Notably, we observed a particularly strong

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

211 To compare MOI-Pred with a previously developed mode of inheritance prediction tool
212 (MAPPIN), we also performed the same EHR-based clinical validation analyses using MAPPIN
213 predictions (**Supplementary Figure 10, Supplementary Table 11-13**). MAPPIN shows weaker
214 enrichment for recessive association with disease of ClinVar Pathogenic predicted recessive (OR=3.61
215 [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
216 enrichment for recessive association with disease of ClinVar Uncertain Significance and Conflicting
217 Interpretation variants (OR=0.35 [95% CI=0.33 to 0.36] for MAPPIN vs. OR=5.45 [95% CI=5.13 to
218 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
219 4.28] for MOI-Pred, respectively) (**Supplementary Figure 10, Supplementary Table 11**). As expected,
220 MAPPIN shows similar or stronger enrichment for dominant association with disease for various ClinVar
221 classes (**Supplementary Figure 10, Supplementary Table 12**).

222

223 *Discovery of individual variants*

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

234

235 Discussion

236 Here we present MOI-Pred, a computational tool that jointly predicts pathogenicity and mode of
237 inheritance for missense variants. Our tool uses a random forest classifier trained on known variants to
238 combine multiple sources of annotation into a single prediction. Compared to other existing methods,
239 MOI-Pred benefits from several key innovations. First and foremost, where most methods produce a
240 binary prediction of pathogenic or benign, our method produces a three-way prediction, classifying each
241 variant as pathogenic for AR disease, pathogenic for AD disease, or benign. In particular, while many
242 existing methods perform well at predicting pathogenic variants in AD disease (e.g. O/E⁴¹, CADD⁴⁴,
243 phyloP⁴⁵, etc.), MOI-Pred specifically targets the problem of discriminating AR pathogenic variants from
244 benign, a long-lasting issue in genetics unattended by current annotation pipelines. Only one pre-existing
245 method makes three-class predictions, MAPPIN³⁸. MOI-Pred performs substantially better than MAPPIN
246 at identifying variants associated with recessive-acting diseases in both prediction performance
247 (Precision=0.79 on the training set for MAPPIN recessive predictions based on Gosalia et al. 2017³⁸) and
248 when testing on real-world clinical data in the current study.

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

261 Third, MOI-Pred is trained with a population-derived list of benign variants, and validated on
262 novel population-derived benign variants unknown to its constituent scores. Because the most difficult
263 classification task is distinguishing AR from benign, the choice of benign training data is vitally
264 important. Since we use clinically validated pathogenic variants for training, it is tempting to use
265 clinically validated benign variants as well, but this can bias the training set. These were suspected being
266 pathogenic at some point and therefore may have features that are not typical of benign variants⁴⁸. The
267 ideal source of benign variants should be found at sufficiently high frequency in a healthy human
268 population^{13,49}. We used frequency-matched variants from a large population database (gnomAD) as
269 presumed benign controls in our training set, an approach that has been used by previous methods such as
270 CADD⁴⁴ and VEST3^{50,51}. However, using these variants introduces an additional problem: population
271 genetics scores used as components in our prediction model are often themselves derived from the same
272 populations, introducing bias and the risk of overfitting⁵². We addressed this problem by using common
273 variants from a recently published cohort of healthy Japanese adults as a validation set. At the time of
274 analysis, this population had not yet been incorporated into widely used population databases, and all
275 genetics scores were therefore naïve to it. Our classifier performed better on these population-derived
276 benign variants than an equivalent classifier trained on clinically validated benign variants, and also
277 performed well on clinically validated benign variants.

278 Finally, we validated our method by using it to predict disease case-control status in EHR data
279 from the BioMe biobank⁴³. We demonstrated that our predictions of mode of inheritance are significantly
280 associated with the likelihood of carriers developing Mendelian disease in a real-world clinical setting.
281 This is true for variants annotated as pathogenic, variants of unknown significance, as well as novel and
282 ancestry-specific variants. This analysis also revealed that some variants with a Benign prediction appear
283 to protect against disease, particularly in variants with “Uncertain significance” (Figure 4C). Since the
284 “Uncertain significance” ClinVar class necessarily contains variants without clear evidence for or against
285 pathogenicity, it is possible that disease modifier variants or variants with protective effects in
286 heterozygous or homozygous form (underdominance or overdominance) may frequently be classified as

287 “Uncertain significance.” Similarly, these variants lack normal signatures of natural selection and so may
288 be likely to receive a Benign prediction in MOI-Pred. In general, the properties of these variants are not
289 well understood, including whether they can be predicted by computational methods, and further
290 investigation is warranted⁵³. We also found individual variants where the prediction made by MOI-Pred
291 differed from their clinical annotations using ClinVar/OMIM, and we verified using the same EHR
292 database that both the pathogenicity and the mode of inheritance predicted by our method is likely to be
293 correct. These analyses demonstrate the applicability of our method to real clinical data and decision-
294 making particularly with respect to large-scale electronic health systems. Such clinical validation is only
295 possible thanks to increasingly available EHR-linked biobanks, and we anticipate it being applied more
296 broadly to variant prediction tools in the future.

297 Our method has several limitations and areas for future work. First, it remains uncertain whether
298 the performance we observe in the test and validation sets will hold in real applications. Many existing
299 tools have reported similarly high performance in their authors’ internal testing and lower performance in
300 unbiased replication analyses¹². Many have also failed to find clinical utility despite numerically high
301 performance^{54,55}. Our EHR-based clinical validation suggests that results will hold⁵⁶ in real-world clinical
302 data, but validation in other clinical datasets and by other groups is needed. Second, our three-way
303 predictions, though more complete than typical binary predictions, do not completely account for all
304 forms of mode of inheritance. Phenomena such as incomplete dominance, overdominance, and
305 heterozygote advantage, all of which are well documented in human disease⁵⁶⁻⁵⁸, are unaccounted for in
306 our simple recessive-dominant-benign classification. Likewise, mode of inheritance itself is far from the
307 only refinement that can be added to variant effect predictions. The field would benefit enormously from
308 methods to predict gain-of-function variants, disease suppressor variants, or uniparental imprinted
309 variants, to name just a few. Third, there is room for innovation and improvement in the method. We used
310 only a subset of applicable ML methods and available features. Furthermore, we focus only on missense
311 variants, primarily because this is the largest class of variation for clinical variants and most applicable to
312 mode of inheritance prediction. Most coding annotations are available for missense variants, while a

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

320

321 **Methods**

322 *Variant collection*

323 Missense variants from publicly available resources were used to generate all datasets. For the
324 training set, pathogenic variants were obtained from ExoVar¹⁶. Presumed non-pathogenic variants were
325 selected from the Genome Aggregation Database (gnomAD)⁴¹ v2.1.1. GnomAD variants were chosen to
326 match the allele frequency (AF) of pathogenic variants to within 0.1%, based on minor allele frequency in
327 the entire gnomAD population; singletons were chosen to match variants not present in gnomAD. For the
328 validation set, pathogenic variants with review status “reviewed by expert panel” were selected from
329 ClinVar²⁷ (release June 2020), and presumed non-pathogenic variants were selected from GEM⁴², defined
330 as variants with AF \geq 1% in GEM and absent or singleton in gnomAD. Gene-level mode of inheritance
331 information for pathogenic variants was obtained from OMIM²⁶ (release May 2020).

332

333 *Variant annotation*

334 Variants were characterized using functional and evolutionary information. ANNOVAR⁵⁹ was
335 used to annotate variant-level features. We included all available features from ANNOVAR that could be
336 applied to all or nearly all missense variants. This includes 15 features built on evolutionary information
337 (e.g. phyloP⁴⁵, FATHMM⁶⁰, GERP⁶¹, PROVEAN⁶², etc.), 42 features built on both evolutionary and
338 functional information (e.g. M-CAP⁶³, CADD⁴⁴, VEST3⁵⁰, MutationTaster⁶⁴, etc.) and 14 population

339 frequency features (e.g. cg69⁶⁵, Kaviar⁶⁶, GME⁶⁷, etc.). We added to this, 7 gene-level features which
340 were retrieved manually from their original sources. This includes 2 gene level features built on
341 evolutionary information (OE score⁴¹ and s_het⁶⁸), 4 gene level features built on functional information
342 (Episcore⁶⁹, AD rank⁷⁰, StringAD and StringAR⁷¹) and 1 gene level feature combining the two
343 annotations (HI⁷²). The full list of features and their source of information can be found in
344 **Supplementary Table 1.**

345

346 *Data trimming and imputation*

347 Features with more than 60% missing values and/or high correlation (Pearson's $r \geq 0.8$) in the
348 training set were removed. In two correlated features, the one with higher mean absolute correlation
349 across all features was removed. Variants with more than 60% missing values in the remaining set of
350 features were removed from both the training and external validation sets. Missing values were imputed
351 first on variant-level features. The resulting dataset was then used to impute gene-level features, ensuring
352 low intra-gene variation in these annotations. A random forest-based algorithm (missForest v1.4⁷³) was
353 used for both imputations. The final dataset was comprised of 30 features on 5,872 and 1,526 variants
354 from the training and external validation sets respectively.

355

356 *Workflow to train inheritance prediction models*

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

365 Train set using the caret package v6.0.84⁷⁴ available in R, then applied accordingly to the Test and
366 Validation sets. A three-class (Recessive, Dominant, Benign) random forest algorithm⁷⁵ was then fitted to
367 the Train set using 10-fold cross validation to optimize parameter tuning and limit overfitting. Three two-
368 class random forest algorithms (Dominant vs. Recessive, Dominant vs. Benign, Recessive vs. Benign)
369 were fitted in parallel for subsequent feature importance analyses. Mode of inheritance label was then
370 predicted on the Test and Validation sets to compute performance metrics. This entire procedure was
371 repeated 100 times; reported performance statistics (see Results) correspond to the mean and standard
372 deviation (SD) across all 100 runs.

373 The AUROC was calculated using the pROC package v1.14.0⁷⁶ available in R v3.5.3⁷⁷. To obtain
374 a per-class discrimination metric the remaining two labels were treated as negative classes. Accuracy,
375 sensitivity, specificity and positive/negative predictive values (PPV/NPV) as well as the ML framework
376 was implemented using the caret package.

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

380

381 *Clinical validation of inheritance prediction models in electronic health records*

382 A single three-class random forest algorithm was fitted, tested and validated as described above
383 to predict mode of inheritance in genotype data from 29,981 individuals in the BioMe biobank⁴³. BioMe is
384 a multiethnic, EHR-linked, clinical care biobank of more than 60,000 samples from individuals recruited
385 at the Mount Sinai Health System between 2007 and 2015. Participants were genotyped using the
386 Illumina Global Screening Array, imputation was performed using the 1000 Genomes Phase 3 reference
387 panel, and genetic ancestry was determined through k-means clustering of principal components.
388 Longitudinal biomedical traits including diagnostic codes and laboratory test results were obtained mainly
389 through ambulatory care practices resulting in a high median number of encounters per patient⁷⁸. Only
390 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 Orphanet⁸⁰.

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 BioMe 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 v3.6.2⁷⁷ was used to perform the CMH test and the metafor package v3.0.2⁸¹ was used for
416 the Q-test.

417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442

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 v1.9⁸². Whole exome sequencing data and EHR from the BioMe biobank were used in the analysis; 10 principal components were used as covariates to account for population stratification.

443 References

- 444 1. Yang, Y. *et al.* Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders. *New*
445 *England Journal of Medicine* **369**, 1502-1511 (2013).
- 446 2. Posey, J.E. *et al.* Resolution of Disease Phenotypes Resulting from Multilocus Genomic Variation.
447 *New England Journal of Medicine* **376**, 21-31 (2016).
- 448 3. Adams, D.R. & Eng, C.M. Next-Generation Sequencing to Diagnose Suspected Genetic Disorders.
449 *The New England Journal of Medicine* **379**, 1353-1362 (2018).
- 450 4. Monies, D. *et al.* Lessons Learned from Large-Scale, First-Tier Clinical Exome Sequencing in a
451 Highly Consanguineous Population. *The American Journal of Human Genetics* **104**, 1182-1201
452 (2019).
- 453 5. Akawi, N. *et al.* Discovery of four recessive developmental disorders using probabilistic genotype
454 and phenotype matching among 4,125 families. *Nature Genetics* **47**, 1363-1369 (2015).
- 455 6. Turro, E. *et al.* Whole-genome sequencing of patients with rare diseases in a national health
456 system. *Nature* **583**, 96-102 (2020).
- 457 7. Van Hout, C.V. *et al.* Exome sequencing and characterization of 49,960 individuals in the UK
458 Biobank. *Nature* **586**, 749-756 (2020).
- 459 8. Do, R. *et al.* Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for
460 myocardial infarction. *Nature* **518**, 102-106 (2015).
- 461 9. Spreafico, R., Soriaga, L.B., Grosse, J., Virgin, H.W. & Telenti, A. Advances in Genomics for Drug
462 Development. *Genes* **11**(2020).
- 463 10. Plenge, R.M., Scolnick, E.M. & Altshuler, D. Validating therapeutic targets through human
464 genetics. *Nature Reviews Drug Discovery* **12**, 581-594 (2013).
- 465 11. Dong, C. *et al.* Comparison and integration of deleteriousness prediction methods for
466 nonsynonymous SNVs in whole exome sequencing studies. *Human Molecular Genetics* **24**, 2125-
467 2137 (2015).
- 468 12. Li, J. *et al.* Performance evaluation of pathogenicity-computation methods for missense variants.
469 *Nucleic acids research* **46**, 7793-7804 (2018).
- 470 13. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint
471 consensus recommendation of the American College of Medical Genetics and Genomics and the
472 Association for Molecular Pathology. *Genetics in Medicine* **17**, 405-423 (2015).
- 473 14. Ioannidis, N.M. *et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare
474 Missense Variants. *The American Journal of Human Genetics* **99**, 877-885 (2016).
- 475 15. Alirezaie, N., Kernohan, K.D., Hartley, T., Majewski, J. & Hocking, T.D. ClinPred: Prediction Tool
476 to Identify Disease-Relevant Nonsynonymous Single-Nucleotide Variants. *The American Journal*
477 *of Human Genetics* **103**, 474-483 (2018).
- 478 16. Li, M.-X. *et al.* Predicting Mendelian Disease-Causing Non-Synonymous Single Nucleotide
479 Variants in Exome Sequencing Studies. *PLOS Genetics* **9**, e1003143 (2013).
- 480 17. Zhang, X. *et al.* Disease-specific variant pathogenicity prediction significantly improves variant
481 interpretation in inherited cardiac conditions. *Genetics in Medicine* **23**, 69-79 (2021).
- 482 18. Lappalainen, T. & MacArthur Daniel, G. From variant to function in human disease genetics.
483 *Science* **373**, 1464-1468 (2021).
- 484 19. Claustres, M. *et al.* Recommendations for reporting results of diagnostic genetic testing
485 (biochemical, cytogenetic and molecular genetic). *European Journal of Human Genetics* **22**, 160-
486 170 (2014).
- 487 20. MacArthur, D.G. *et al.* Guidelines for investigating causality of sequence variants in human
488 disease. *Nature* **508**, 469-476 (2014).

- 489 21. Eldomery, M.K. *et al.* Lessons learned from additional research analyses of unsolved clinical
490 exome cases. *Genome Medicine* **9**, 26 (2017).
- 491 22. Ewans, L.J. *et al.* Whole-exome sequencing reanalysis at 12 months boosts diagnosis and is cost-
492 effective when applied early in Mendelian disorders. *Genetics in Medicine* **20**, 1564-1574 (2018).
- 493 23. Lee, H. *et al.* Clinical Exome Sequencing for Genetic Identification of Rare Mendelian Disorders.
494 *JAMA* **312**, 1880-1887 (2014).
- 495 24. Retterer, K. *et al.* Clinical application of whole-exome sequencing across clinical indications.
496 *Genetics in Medicine* **18**, 696-704 (2016).
- 497 25. Chong, Jessica X. *et al.* The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and
498 Opportunities. *The American Journal of Human Genetics* **97**, 199-215 (2015).
- 499 26. Online Mendelian Inheritance in Man, O.M.-N.I.o.G.M., Johns Hopkins University (Baltimore,
500 MD), May 2020. World Wide Web URL: <https://omim.org/>.
- 501 27. Landrum, M.J. *et al.* ClinVar: improving access to variant interpretations and supporting
502 evidence. *Nucleic Acids Res* **46**, D1062-d1067 (2018).
- 503 28. Furney, S.J., Albà, M.M. & López-Bigas, N. Differences in the evolutionary history of disease
504 genes affected by dominant or recessive mutations. *BMC Genomics* **7**, 165 (2006).
- 505 29. Jimenez-Sanchez, G., Childs, B. & Valle, D. Human disease genes. *Nature* **409**, 853-855 (2001).
- 506 30. Kondrashov, F.A. & Koonin, E.V. A common framework for understanding the origin of genetic
507 dominance and evolutionary fates of gene duplications. *Trends in Genetics* **20**, 287-290 (2004).
- 508 31. López-Bigas, N., Blencowe, B.J. & Ouzounis, C.A. Highly consistent patterns for inherited human
509 diseases at the molecular level. *Bioinformatics* **22**, 269-277 (2006).
- 510 32. Blekhman, R. *et al.* Natural selection on genes that underlie human disease susceptibility. *Curr*
511 *Biol* **18**, 883-9 (2008).
- 512 33. Rapaport, F. *et al.* Negative selection on human genes underlying inborn errors depends on
513 disease outcome and both the mode and mechanism of inheritance. *Proceedings of the National*
514 *Academy of Sciences* **118**, e2001248118 (2021).
- 515 34. Fuller, Z.L., Berg, J.J., Mostafavi, H., Sella, G. & Przeworski, M. Measuring intolerance to
516 mutation in human genetics. *Nature Genetics* **51**, 772-776 (2019).
- 517 35. Balick, D.J., Jordan, D.M., Sunyaev, S. & Do, R. Overcoming constraints on the detection of
518 recessive selection in human genes from population frequency data. *bioRxiv*, 2021.05.06.443024
519 (2021).
- 520 36. Antonarakis, S.E. Carrier screening for recessive disorders. *Nature Reviews Genetics* **20**, 549-561
521 (2019).
- 522 37. Furney, S.J., Albà, M.M. & López-Bigas, N. Differences in the evolutionary history of disease
523 genes affected by dominant or recessive mutations. *BMC Genomics* **7**, 165 (2006).
- 524 38. Gosalia, N., Economides, A.N., Dewey, F.E. & Balasubramanian, S. MAPPIN: a method for
525 annotating, predicting pathogenicity and mode of inheritance for nonsynonymous variants.
526 *Nucleic Acids Research* **45**, 10393-10402 (2017).
- 527 39. Quinodoz, M. *et al.* DOMINO: Using Machine Learning to Predict Genes Associated with
528 Dominant Disorders. *The American Journal of Human Genetics* **101**, 623-629 (2017).
- 529 40. Stenson, P.D. *et al.* Human Gene Mutation Database (HGMD): 2003 update. *Hum Mutat* **21**, 577-
530 81 (2003).
- 531 41. Karczewski, K.J. *et al.* The mutational constraint spectrum quantified from variation in 141,456
532 humans. *Nature* **581**, 434-443 (2020).
- 533 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.-
534 J.A.f.
- 535 43. BioMeTM BioBank Program. <https://icahn.mssm.edu/research/ipm/programs/biome-biobank>.
536 Accessed June.

- 537 44. Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J. & Kircher, M. CADD: predicting the
538 deleteriousness of variants throughout the human genome. *Nucleic Acids Res* **47**, D886-d894
539 (2019).
- 540 45. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R. & Siepel, A. Detection of nonneutral substitution
541 rates on mammalian phylogenies. *Genome Res* **20**, 110-21 (2010).
- 542 46. Li, B. *et al.* Automated inference of molecular mechanisms of disease from amino acid
543 substitutions. *Bioinformatics* **25**, 2744-50 (2009).
- 544 47. Pejaver, V. *et al.* Inferring the molecular and phenotypic impact of amino acid variants with
545 MutPred2. *Nat Commun* **11**, 5918 (2020).
- 546 48. Shah, N. *et al.* Identification of Misclassified ClinVar Variants via Disease Population Prevalence.
547 *American journal of human genetics* **102**, 609-619 (2018).
- 548 49. Ghosh, R. *et al.* Updated recommendation for the benign stand-alone ACMG/AMP criterion.
549 *Human mutation* **39**, 1525-1530 (2018).
- 550 50. Carter, H., Douville, C., Stenson, P.D., Cooper, D.N. & Karchin, R. Identifying Mendelian disease
551 genes with the variant effect scoring tool. *BMC Genomics* **14 Suppl 3**, S3 (2013).
- 552 51. Carter, H., Douville, C., Stenson, P.D., Cooper, D.N. & Karchin, R. Identifying Mendelian disease
553 genes with the variant effect scoring tool. *BMC Genomics* **14 Suppl 3**, S3 (2013).
- 554 52. Grimm, D.G. *et al.* The evaluation of tools used to predict the impact of missense variants is
555 hindered by two types of circularity. *Hum Mutat* **36**, 513-23 (2015).
- 556 53. Rahit, K. & Tarailo-Graovac, M. Genetic Modifiers and Rare Mendelian Disease. *Genes (Basel)*
557 **11**(2020).
- 558 54. Ghosh, R., Oak, N. & Plon, S.E. Evaluation of in silico algorithms for use with ACMG/AMP clinical
559 variant interpretation guidelines. *Genome Biol* **18**, 225 (2017).
- 560 55. Mahmood, K. *et al.* Variant effect prediction tools assessed using independent, functional assay-
561 based datasets: implications for discovery and diagnostics. *Hum Genomics* **11**, 10 (2017).
- 562 56. Wexler, N.S. *et al.* Homozygotes for Huntington's disease. *Nature* **326**, 194-7 (1987).
- 563 57. Hughes, A.L. & Nei, M. Pattern of nucleotide substitution at major histocompatibility complex
564 class I loci reveals overdominant selection. *Nature* **335**, 167-170 (1988).
- 565 58. Schroeder, S.A., Gaughan, D.M. & Swift, M. Protection against bronchial asthma by CFTR $\Delta F508$
566 mutation: A heterozygote advantage in cystic fibrosis. *Nature Medicine* **1**, 703-705 (1995).
- 567 59. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from
568 high-throughput sequencing data. *Nucleic Acids Research* **38**, e164-e164 (2010).
- 569 60. Shihab, H.A. *et al.* Predicting the Functional, Molecular, and Phenotypic Consequences of Amino
570 Acid Substitutions using Hidden Markov Models. *Human Mutation* **34**, 57-65 (2013).
- 571 61. Davydov, E.V. *et al.* Identifying a High Fraction of the Human Genome to be under Selective
572 Constraint Using GERP++. *PLoS Computational Biology* **6**, e1001025 (2010).
- 573 62. Choi, Y., Sims, G.E., Murphy, S., Miller, J.R. & Chan, A.P. Predicting the functional effect of amino
574 acid substitutions and indels. *PLoS One* **7**, e46688 (2012).
- 575 63. Jagadeesh, K.A. *et al.* M-CAP eliminates a majority of variants of uncertain significance in clinical
576 exomes at high sensitivity. *Nature Genetics* **48**, 1581-1586 (2016).
- 577 64. Schwarz, J.M., Cooper, D.N., Schuelke, M. & Seelow, D. MutationTaster2: mutation prediction
578 for the deep-sequencing age. *Nat Methods* **11**, 361-2 (2014).
- 579 65. Drmanac, R. *et al.* Human Genome Sequencing Using Unchained Base Reads on Self-Assembling
580 DNA Nanoarrays. *Science* **327**, 78 (2010).
- 581 66. Glusman, G., Caballero, J., Mauldin, D.E., Hood, L. & Roach, J.C. Kaviar: an accessible system for
582 testing SNV novelty. *Bioinformatics* **27**, 3216-7 (2011).
- 583 67. Scott, E.M. *et al.* Characterization of Greater Middle Eastern genetic variation for enhanced
584 disease gene discovery. *Nature genetics* **48**, 1071-1076 (2016).

- 585 68. Cassa, C.A. *et al.* Estimating the selective effects of heterozygous protein-truncating variants
586 from human exome data. *Nature Genetics* **49**, 806-810 (2017).
- 587 69. Han, X. *et al.* Distinct epigenomic patterns are associated with haploinsufficiency and predict risk
588 genes of developmental disorders. *Nature Communications* **9**, 2138 (2018).
- 589 70. Hsu, J.S. *et al.* Inheritance-mode specific pathogenicity prioritization (ISPP) for human protein
590 coding genes. *Bioinformatics* **32**, 3065-3071 (2016).
- 591 71. Szklarczyk, D. *et al.* STRING v11: protein-protein association networks with increased coverage,
592 supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* **47**,
593 D607-d613 (2019).
- 594 72. Huang, N., Lee, I., Marcotte, E.M. & Hurler, M.E. Characterising and Predicting
595 Haploinsufficiency in the Human Genome. *PLOS Genetics* **6**, e1001154 (2010).
- 596 73. Stekhoven, D.J. & Bühlmann, P. MissForest—non-parametric missing value imputation for
597 mixed-type data. *Bioinformatics* **28**, 112-118 (2012).
- 598 74. Kuhn, M. Building Predictive Models in R Using the caret Package. *Journal Of Statistical Software*
599 **28**(2008).
- 600 75. Liaw, A. & Wiener, M. Classification and Regression by randomForest. *R News* **2**, 18-22 (2002).
- 601 76. Robin, X. *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves.
602 *BMC Bioinformatics* **12**(2011).
- 603 77. R Core Team. R: A language and environment for statistical computing. *R Foundation for*
604 *Statistical Computing, Vienna, Austria.* (2019).
- 605 78. Tayo, B.O. *et al.* Genetic Background of Patients from a University Medical Center in Manhattan:
606 Implications for Personalized Medicine. *PLOS ONE* **6**, e19166 (2011).
- 607 79. International Health Terminology Standards Development Organisation. SNOMED CT Starter
608 Guide, Accessed June 2020.
609 <https://confluence.ihtsdotools.org/display/DOCSTART/SNOMED+CT+Starter+Guide>.
- 610 80. Pavan, S. *et al.* Clinical Practice Guidelines for Rare Diseases: The Orphanet Database. *PLoS One*
611 **12**, e0170365 (2017).
- 612 81. Viechtbauer, W. Conducting Meta-Analyses in R with the metafor Package. *2010* **36**, 48 (2010).
- 613 82. Chang, C.C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets.
614 *GigaScience* **4**(2015).

615

616

617

618

619

620

621

622

623

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

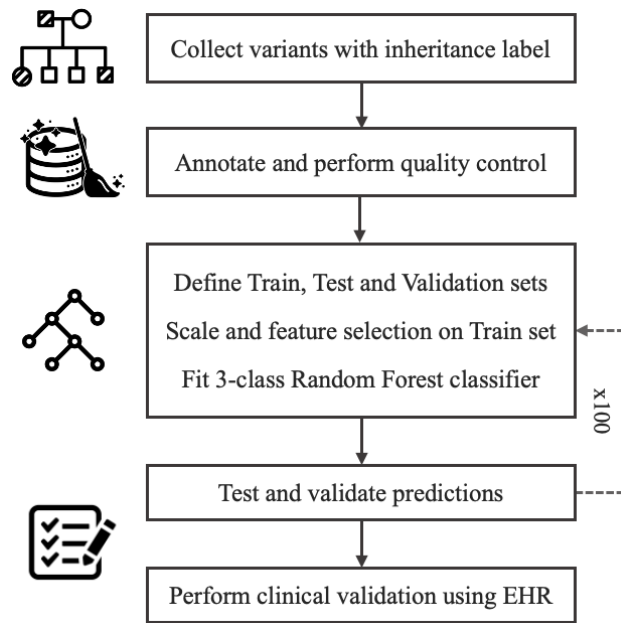
645 **Disclaimer:** The content is solely the responsibility of the authors and does not necessarily represent the
646 official views of the National Institutes of Health.

647

648 **Data sharing statement:** Not applicable.

649

650 **Figure 1.** Study design and machine learning workflow.



651

652

653 Train and Test sets correspond to the 90% and 10% balanced datasets, built from ExoVar and gnomAD

654 variants, used for training and testing respectively. Validation set corresponds to the balanced dataset,

655 built from ClinVar and GEM variants, used for external validation. EHR corresponds to electronic health

656 records.

657

658

659

660

661

662

663

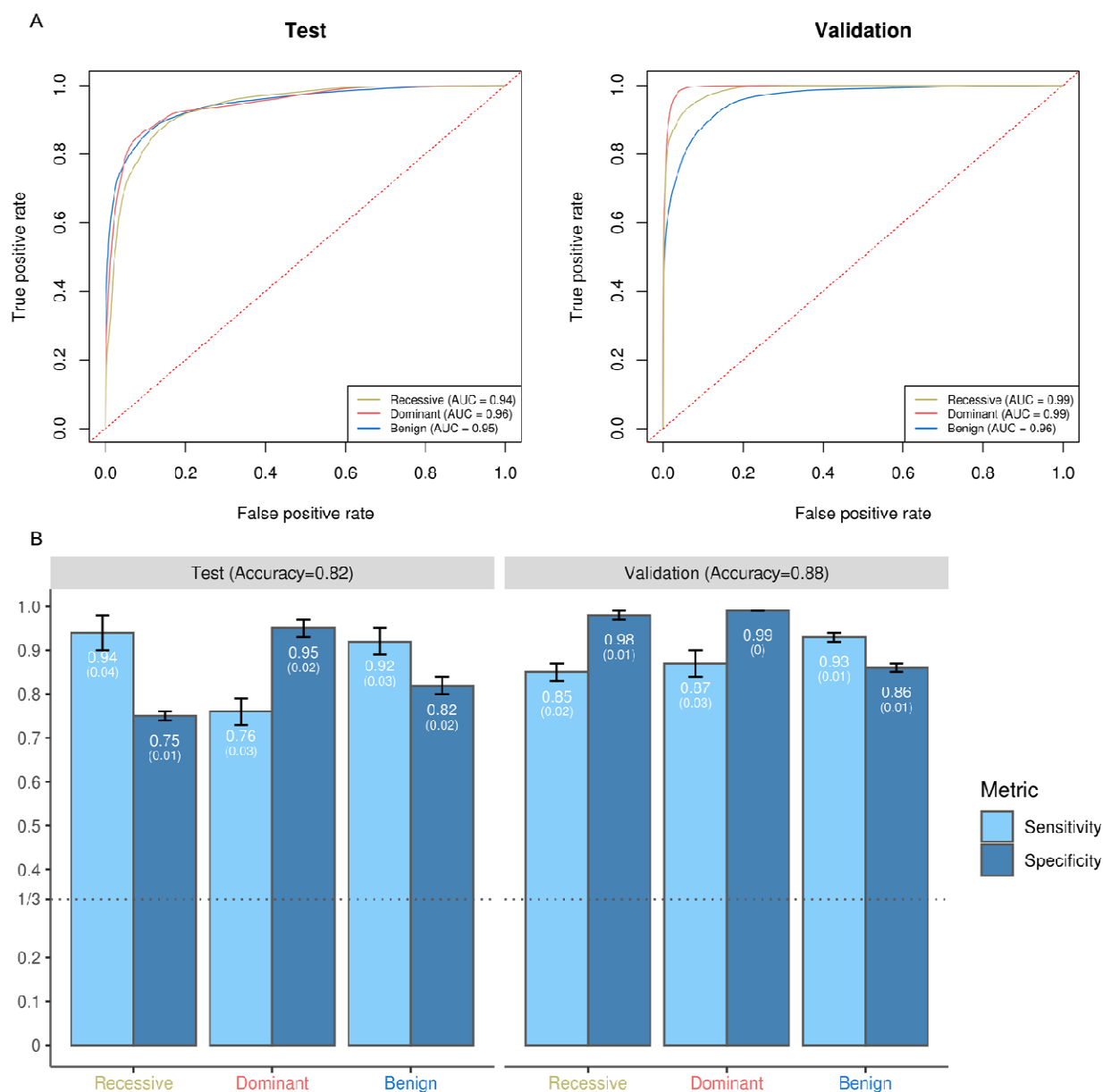
664

665

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

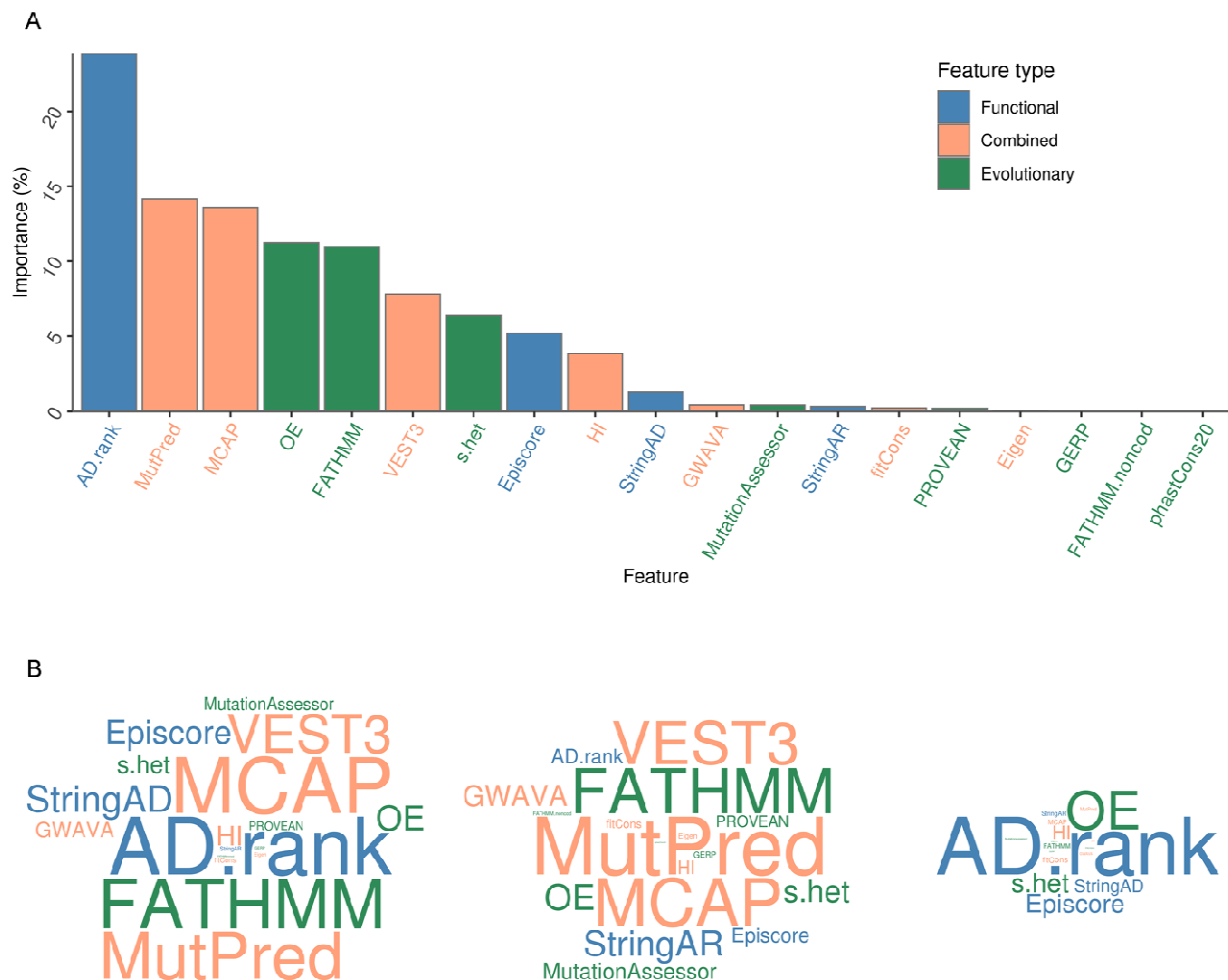
672

673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700

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.

701 **Figure 3:** Bar-plot showing feature importance on 3-class mode of inheritance prediction models (A).

702 Word-clouds representing feature importance on 2-class mode of inheritance prediction models (B).



703

704

705 Feature importance is reported as the median across 100 models. Word-clouds represent feature

706 importance on Benign-Dominant (left), Benign-Recessive (middle) and Dominant-Recessive (right)

707 models respectively. Exact feature importance values in 2-class prediction models can be found in

708 Supplementary Figure 6-8.

709

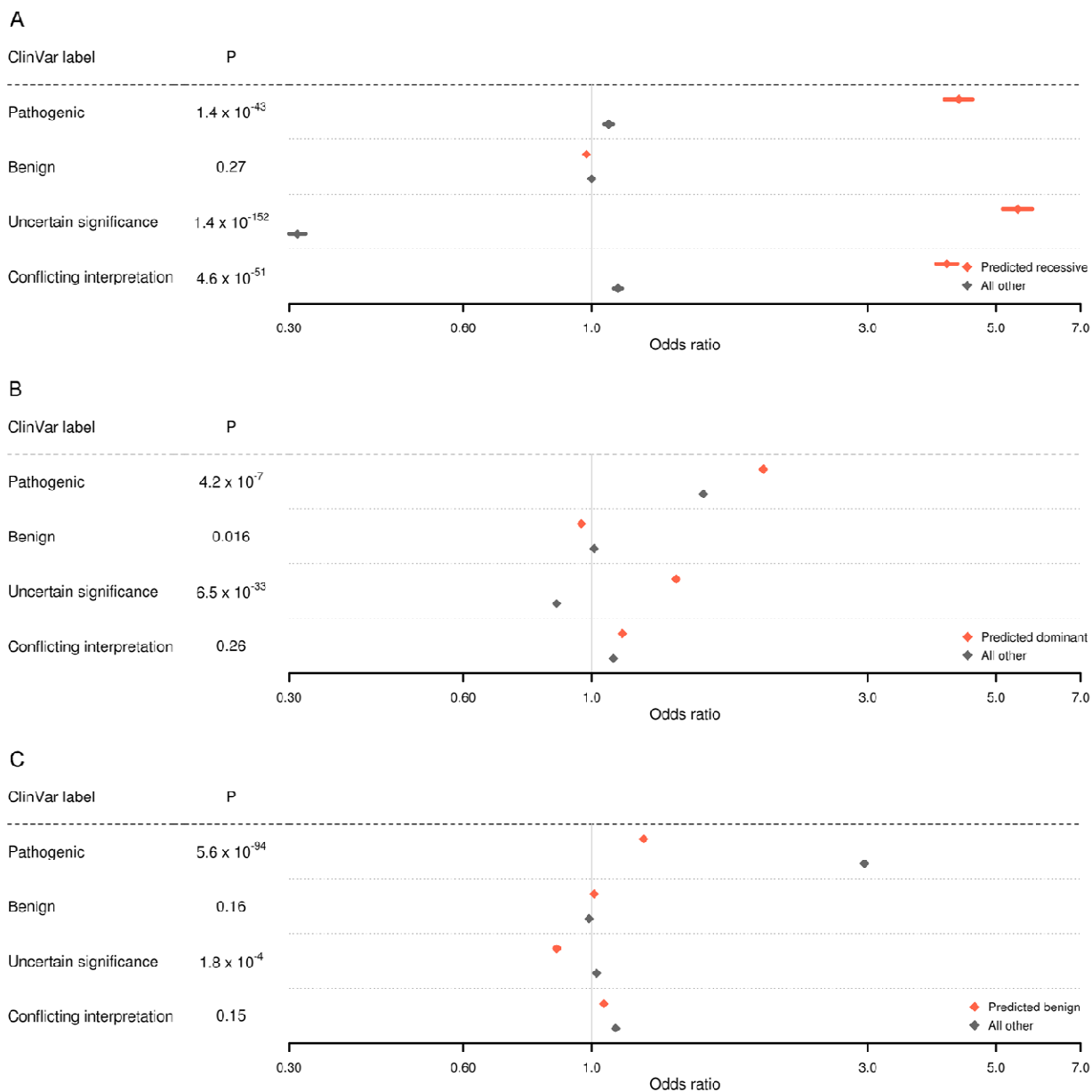
710

711

712 **Figure 4:** Forest plots showing disease association with variants predicted to be Recessive (A), Dominant

713 (B) and Benign (C).

714



715

716

717

718

719

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

724

725

726

727

728

729

730

731

732

733

734

735

736

737

738

739

740

741

742

743

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

747

748

749

750

751

752

753

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