

# 1 The performance of AlphaMissense to identify genes causing disease

2 Yiheng Chen<sup>1,2</sup>, Guillaume Butler-Laporte<sup>2,3</sup>, Kevin Y. H. Liang<sup>2,4</sup>, Yann Ilboudo<sup>2</sup>, Summaira  
3 Yasmeen<sup>5</sup>, Takayoshi Sasako<sup>2,6</sup>, Claudia Langenberg<sup>7,5,8</sup>, Celia M.T. Greenwood<sup>2,3,4,9</sup>, J Brent  
4 Richards<sup>1,2,3,10,11,12,\*</sup>

5 1. Department of Human Genetics, McGill University, Montréal, QC, Canada

6 2. Lady Davis Institute, Jewish General Hospital, McGill University, Montréal, QC, Canada

7 3. Department of Epidemiology, Biostatistics and Occupational Health, McGill University,  
8 Montréal, QC, Canada

9 4. Quantitative Life Sciences Program, McGill University, Montreal, Quebec, Canada

10 5. Computational Medicine, Berlin Institute of Health at Charité—Universitätsmedizin Berlin,  
11 Berlin, Germany

12 6. Tanaka Diabetes Clinic Omiya, Saitama, Japan

13 7. Precision Healthcare University Research Institute, Queen Mary University of London,  
14 London, UK

15 8. MRC Epidemiology Unit, University of Cambridge, Cambridge, UK

16 9. Gerald Bronfman Department of Oncology, McGill University, Montreal, Quebec, Canada

17 10. 5 Prime Sciences Inc, Montréal, Quebec, Canada

18 11. Department of Medicine, McGill University, Montréal, Quebec, Canada

19 12. Department of Twin Research, King's College London, London, UK

20 \*Corresponding author: J Brent Richards, email address: [brent.richards@mcgill.ca](mailto:brent.richards@mcgill.ca)

21

22

23

24

25

## 26 **Abstract**

27 A novel algorithm, AlphaMissense, has been shown to have an improved ability to predict  
28 the pathogenicity of rare missense genetic variants. However, it is not known whether  
29 AlphaMissense improves the ability of gene-based testing to identify disease-causing  
30 genes. Using whole-exome sequencing data from the UK Biobank, we compared gene-  
31 based association analysis strategies including sets of deleterious variants: predicted  
32 loss-of-function (pLoF) variants only, pLoF plus AlphaMissense pathogenic variants, pLoF  
33 with missense variants predicted to be deleterious by any of five commonly utilized  
34 annotation methods (Missense (1/5)) or only variants predicted to be deleterious by all  
35 five methods (Missense (5/5)). We measured performance to identify 519 previously  
36 identified positive control genes, which can cause Mendelian diseases, or are the targets  
37 of successfully developed medicines. These strategies identified 850k pLoF variants and  
38 5 million deleterious missense variants, including 22k likely pathogenic missense variants  
39 identified exclusively by AlphaMissense. The gene-based association tests found 608  
40 significant gene associations (at  $P < 1.25 \times 10^{-7}$ ) across 24 common traits and diseases.  
41 Compared to pLOFs plus Missense (5/5), tests using pLOFs and AlphaMissense variants  
42 found slightly more significant gene-disease and gene-trait associations, albeit with a  
43 marginally lower proportion of positive control genes. Nevertheless, their overall  
44 performance was similar. Merging AlphaMissense with Missense (5/5), whether through  
45 their intersection or union, did not yield any further enhancement in performance. In  
46 summary, employing AlphaMissense to select deleterious variants for gene-based testing  
47 did not improve the ability to identify genes that are known to cause disease.

48

## 49 Introduction

50 Rare genetic variants are important contributors to human diseases. They contribute to  
51 most Mendelian disorders, and their effect sizes upon common diseases are larger than  
52 those attributed to common variants (1–4). Importantly, associated rare genetic variants  
53 are often coding and can therefore be directly attributed to a gene. Loss of function rare  
54 variants can offer insights into the direction of genetic effect on disease outcome.  
55 However, studying rare causal variants is challenging, since most of the genetic variation  
56 in the genome is both rare and benign. Thus, gene-based analysis is usually employed  
57 to improve statistical power by aggregating multiple rare variants across a gene into one  
58 test to improve statistical power to detect disease associations (5).

59  
60 Previous gene-based multi-variant tests like exome-wide association studies (ExWAS)  
61 have successfully identified disease-causal genes, like *WNT1* for osteoporosis (6), and  
62 drug-targeting genes, such as *PCSK9* for low-density lipoprotein (LDL)-cholesterol levels  
63 (7). Nevertheless, the power of ExWAS relies heavily on the prior identification of variants  
64 with a likely functional impact (5) to reduce the number of irrelevant genetic variants  
65 included in the tests. While predicted loss-of-function (pLoFs) rare variants are most likely  
66 to contribute to gene-based tests, deleterious missense variants can also increase  
67 statistical power as they tend to be more common. However, to use deleterious missense  
68 variants, one must understand which of the missense variants is most likely to influence  
69 protein function—a process referred to as variant annotation. Moreover, all deleterious  
70 missense variant annotation strategies must strike a balance between false positive and  
71 false negative identification of such variants (8,9).

72

73 Recent advances in missense variant effect prediction have made progress towards  
74 resolving this problem. AlphaMissense, a recently described method based on an  
75 unsupervised language model, combines protein structural context with evolutionary  
76 conservation and has claimed to achieve over 90% precision when predicting the known  
77 clinical impact of missense variants (9). Additionally, their variant pathogenicity  
78 annotations improved the prediction of gene essentiality for cell survival and fitness.

79

80 However, it is not known whether the improvements observed in AlphaMissense's ability  
81 to predict the deleteriousness of missense variants results in improved association testing  
82 between genes and diseases. If this improvement were striking, it could help to identify  
83 new causes of disease and consequently drug targets for needed drug development.  
84 Using the UK Biobank whole exome sequencing (WES) data, we tested the ability of  
85 AlphaMissense variant annotation to improve the ability to identify positive control genes  
86 (known to cause disease) through collapsing gene-based tests on 12 continuous traits  
87 and 12 diseases. We compared its performance to other leading algorithms. The results  
88 empirically test the ability of AlphaMissense to improve the identification of genes causing  
89 disease.

90

91

92

93

94

## 95 **Methods**

### 96 **UK Biobank cohort**

97 The UK Biobank is a cohort study that has recruited over 500,000 participants between  
98 40 and 69 years of age at 22 testing centers across the United Kingdom and collected a  
99 large set of phenotypes and biological samples. We included in our analyses a total of  
100 444,072 genetically predicted European genetic ancestry individuals with available WES  
101 data generated following the OQFE protocol (10) and with measurements of selected  
102 phenotypes and diseases. The detailed steps for the sample preparation, sequencing,  
103 filtering, and calling of UK Biobank WES data have been previously described (10,11).

104

### 105 **Phenotype definitions**

106 From the UK Biobank, we selected 12 continuous traits and 12 diseases for analysis  
107 based on the trait sample sizes and whether there were known disease causal genes or  
108 drug target genes for each trait. The continuous traits included estimated bone mineral  
109 density, serum triglyceride levels, systolic blood pressure, diastolic blood pressure,  
110 standing height, serum low-density lipoproteins, serum bilirubin, serum glucose, red blood  
111 cell counts, and serum calcium level, body mass index, and waist-hip circumference ratio  
112 and the 12 diseases included hypertension, hypercholesterolemia, diaphragmatic hernia,  
113 osteoarthritis (localized), cataract, type 2 diabetes, major depressive disorder,  
114 hypothyroidism, acute renal failure, atrial fibrillation, cancer of prostate (males only) and  
115 breast cancer (females only). The sample sizes for the analysis of each trait and disease  
116 can be found in [Supplementary table 1](#). ICD-10 codes were grouped to construct diseases

117 following the phecodes system (12). The list of used ICD-10 codes for each phecode can  
118 be found in [Supplementary table 2](#).

119

## 120 **Variant annotation**

121 We annotated the variants from exome sequencing after alignment using the Ensembl  
122 Variant Effect Predictor (VEP) (v.110). Variant annotations among transcript ablation,  
123 splice acceptor, splice donor, stop gained, frameshift, stop lost, start lost, transcript  
124 amplification, feature elongation, and feature truncation, were considered as predicted  
125 loss-of-function (pLoF) variants (10). Missense variants were classified with two  
126 strategies. The first strategy used AlphaMissense (9), and missense variants were  
127 included in our analyses if AlphaMissense predicted them to be “likely pathogenic”. We  
128 built a second strategy by combining results from five commonly used annotation  
129 methods (i.e., SIFT (13), PolyPhen2 (HDIV) (14), PolyPhen2 (HVAR) (15),  
130 MutationTaster (16), and LRT (17)). We classified a missense variant as “likely  
131 deleterious” if all five algorithms predicted it to be deleterious (i.e., Missense (5/5)), and  
132 “possibly deleterious” if at least one of the five algorithms predicted it to be deleterious  
133 (i.e., Missense (1/5)), similar to methods used before (6,10).

134

## 135 **Gene-based disease and trait association test**

136 For each gene, variant annotations and alternative allele frequency (AAF) categorized the  
137 inclusion of variants into 20 gene burden exposures, created by the combination of four  
138 annotation mask definitions and five AAF thresholds and statistical testing method  
139 combinations. The four masks categories included: (1) pLoF variants; (2) pLoF or “likely

140 pathogenic” variants by AlphaMissense (pLoF with AlphaMissense); (3) pLoF or “likely  
141 deleterious” missense variants by the five commonly used methods (pLoF with Missense  
142 (5/5)); (4) pLoF or “possibly deleterious” missense variants by the five commonly used  
143 methods (pLoF with Missense (1/5)). The five AAF and statistical test method  
144 combinations included (1) standard burden test with AAF < 1%; (2) standard burden test  
145 with AAF < 0.1%; (3) standard burden test with singletons; (4) SKAT variance-component  
146 test with AAF <1%; (5) SKAT-O combined test with AAF <1%). The smallest p-value of  
147 the five AAF and test combinations for each gene under different masks were retained  
148 for subsequent significance and classification testing. For our primary method, we built  
149 masks for burden tests using the maximum number of alternative alleles found across all  
150 selected variant sites of a gene. As a sensitivity analysis, we also tested whether building  
151 masks by total number of alternative alleles across these sites, a approach assuming  
152 these sites have cumulative effect, would impact the results of association analyses.

153  
154 All analyses were performed using Regenie software (18). The regression analyses  
155 included age, age<sup>2</sup>, sex, sex\*age, sex\*age<sup>2</sup>, 10 genetic principal components (PC)  
156 obtained from common genetic variants (MAF>1%), and 20 genetic PCs obtained from  
157 rare genetic variants (MAF<1%) as covariates. The statistical significance threshold was  
158  $P < 1.25 \times 10^{-7}$  (0.05 / (approximately 20,000 genes \* 20 gene-burden exposures)).

159  
160 **Selection of positive control genes**

161 To evaluate whether different masks have different abilities to identify genes that were  
162 known to cause Mendelian forms of disease, or the targets of successfully developed

163 medicines, we compiled a list of positive control genes from two sources. We first included  
164 positive control genes from two previous studies where these genes were used to train  
165 their algorithms to prioritize disease-causal or drug-targeting genes from genome-wide  
166 association study (GWAS) signals (19,20). Their positive control gene lists were  
167 generated by combining genetic evidence, drug–target–indication associations, and  
168 manual curation from board certified physicians and domain experts. Additionally, we  
169 included Mendelian diseases genes from the MendelVar database which was created by  
170 integrating functional annotations from the Online Mendelian Inheritance in Man (OMIM),  
171 Deciphering Developmental Disorders Study (DECIPHER), Orphanet and Genomics  
172 England databases (21). The full list of 509 positive control genes for the selected traits  
173 and diseases can be found in [Supplementary Table 3](#).

174

### 175 **Evaluation of classification accuracy**

176 The ability to accurately identify positive control genes using gene burden tests with  
177 different variant sets and mask settings was measured by the area under the receiver-  
178 operator curves (AUROC) and precision-recall curves (AUPRC). Specifically, PRC and  
179 ROC were generated using results from 21 traits and diseases where we could confirm  
180 at least one positive control gene. The 95% confidence intervals (CI) for AUROC and  
181 AUPRC were determined using 1000 bootstrap replicates. The baseline for AUROC is  
182 0.5, an uninformative classifier. The baseline level for AUPRC is 0.0018 which equals the  
183 proportion of positive control genes among tested genes.

184

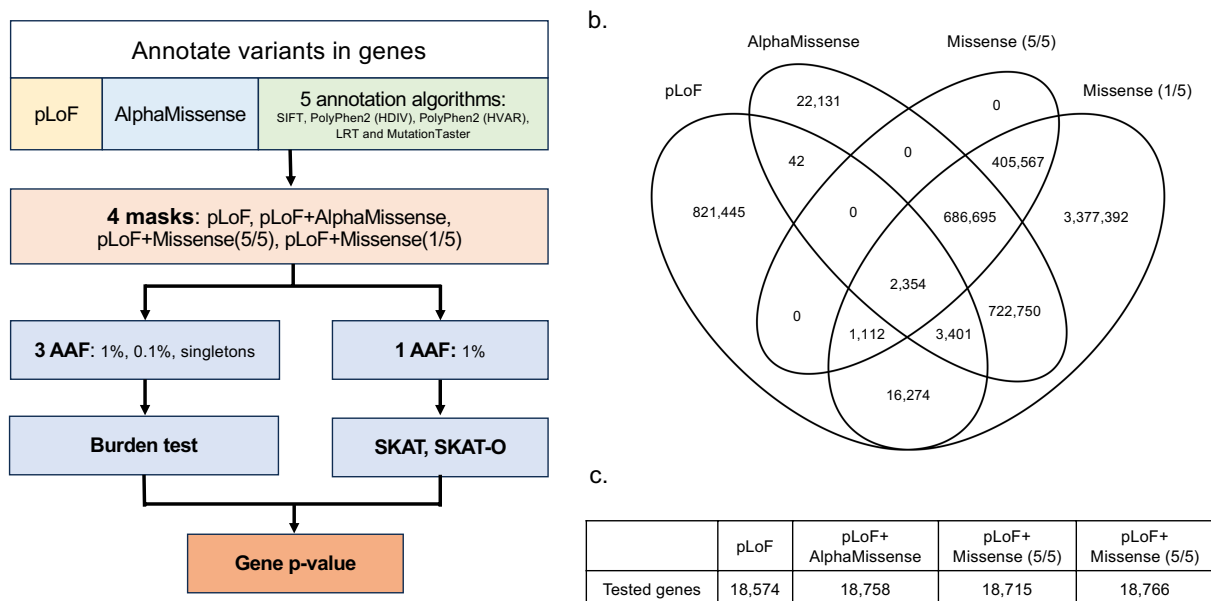
185



## 186 Results

187 Starting with 19,606 genes, for every exon, we annotated deleterious variants into four  
 188 categories: pLoF, AlphaMissense, Missense (5/5), and Missense (1/5). Then we  
 189 assembled four sets of predicted deleterious variants (i.e., masks): (1) pLoF, (2) pLoF  
 190 with AlphaMissense, (3) pLoF with Missense (5/5), and (4) pLoF with Missense (1/5).  
 191 Each mask provided a list of variants for genes in gene-based association analysis. Lastly,  
 192 we retained the smallest p-values from the five different combinations of alternative allele  
 193 frequency and statistical test method for the association between each gene and each  
 194 tested trait or disease under different masks (Figure 1a).

195



196

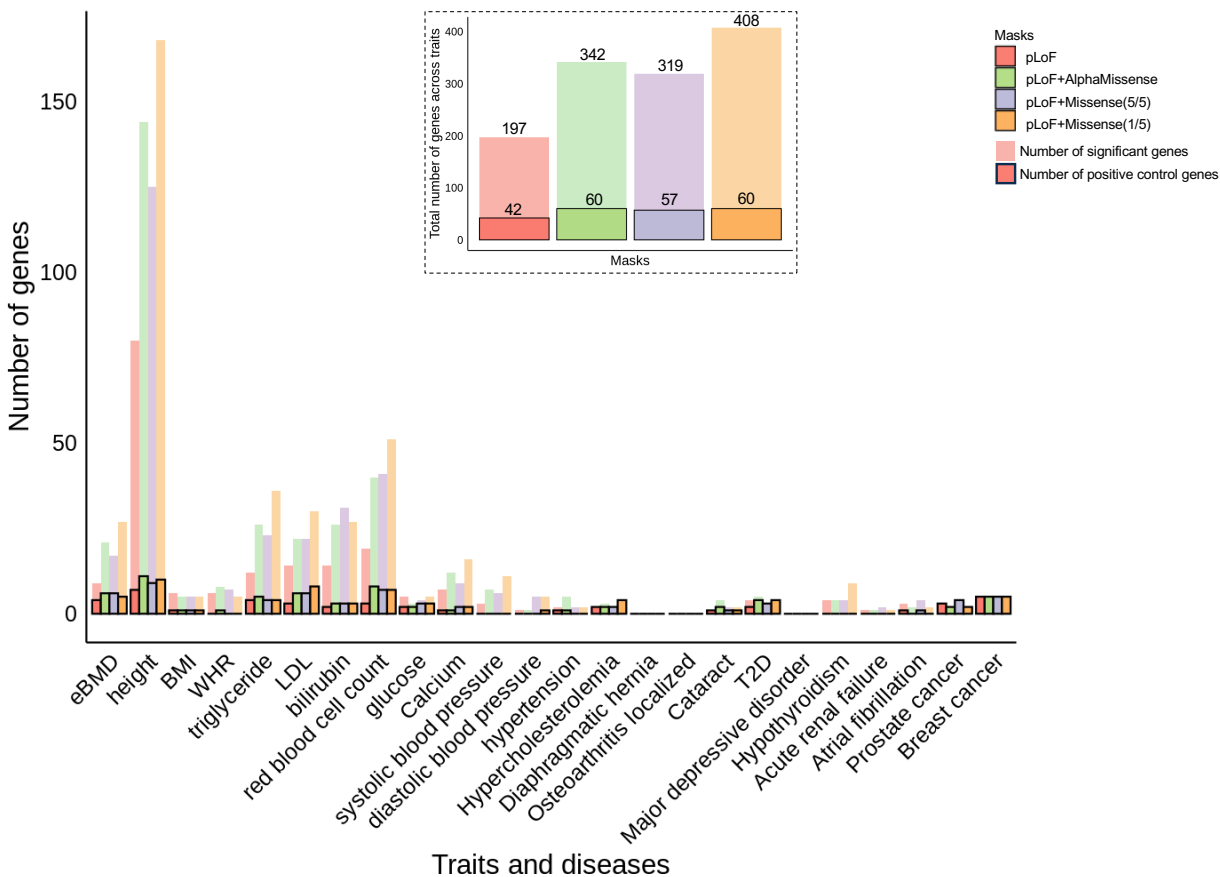
197 **Figure 1. Overview of ExWAS analysis and comparison of variants annotations by**  
 198 **different methods. (a) strategy to obtain single *P* value for each gene (b) overlap of**  
 199 **annotated likely deleterious variants by different ExWAS mask settings (c) Number of**  
 200 **tested genes with different mask settings.**

201

202 Of 26 million variants from UK Biobank WES data, we identified 850k pLoF variants and  
203 5 million predicted deleterious missense variants by AlphaMissense or any of the five  
204 commonly used annotation methods (i.e., SIFT, PolyPhen2 (HDIV), PolyPhen2 (HVAR),  
205 MutationTaster, and LRT). Specifically, AlphaMissense classified 1.4 million variants as  
206 “likely pathogenic”, including 22k identified exclusively by AlphaMissense. Missense (1/5)  
207 captured over 98% of AlphaMissense predicted “likely pathogenic” variants while  
208 Missense (5/5) covered 48% of AlphaMissense predicted “likely pathogenic” variants  
209 (Figure 1b). Moreover, our results showed that among the masks evaluated, Missense  
210 (1/5) labeled the highest number of deleterious variants per gene on average (267  
211 variants per gene), followed by AlphaMissense (74 variants per gene), Missense (5/5) (56  
212 variants per gene), and pLoF (43 variants per gene) (Supplementary Table 4). Despite  
213 the considerable variance in the number of annotated variants across different annotation  
214 categories, 99% of genes were tested in all masks (Figure 1c).

215

216 In the exome-wide gene-based analysis, we first checked the genomic inflation factors of  
217 the p-values for each mask and test method combination. In general, no strong genomic  
218 inflation was observed (value range: 0.96-1.37) except for standing height (value range:  
219 1.11-1.94) (Supplementary Table 5). This is not surprising as height is a well-known highly  
220 polygenic trait (22).

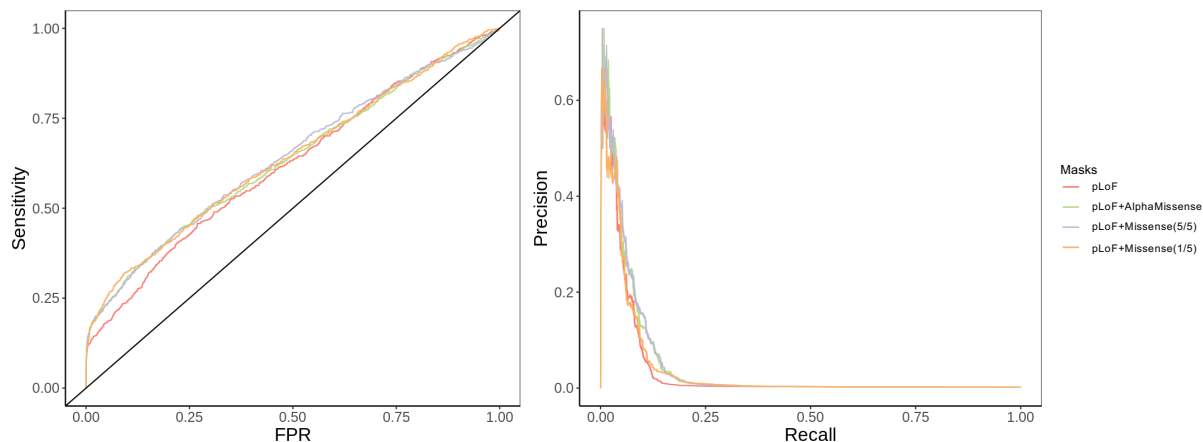


221  
 222 **Figure 2. Significant gene associations identified in exome-wide gene burden**  
 223 **analysis across 12 traits and 12 diseases.** The bars without outlines indicate the  
 224 numbers of significant genes ( $P < 1.25 \times 10^{-7}$ ) identified in each trait and disease by  
 225 different masks. The bars with outlines indicate the number of significant genes that are  
 226 also positive control genes for each trait and diseases identified by different masks. The  
 227 inset figure shows the total number of significant genes and positive controls identified by  
 228 each mask across all the tested traits and diseases. Abbreviations: estimated bone  
 229 mineral density (eBMD), body mass index (BMI), waist-hip circumference ratio (WHR),  
 230 serum low-density lipoproteins (LDL), type 2 diabetes (T2D).  
 231

232 In total, our gene-based association tests found 608 significant gene associations ( $P$   
233  $<1.25 \times 10^{-7}$ ) across 24 common traits and diseases. We found that adding predicted  
234 deleterious missense variants to masks led to the identification of at least 60% more  
235 significant gene-trait associations and about 30% more positive control genes as  
236 compared to pLoF-only mask ([Figure 2, Supplementary Figure 1a, and Supplementary](#)  
237 [Table 6](#)). Despite different numbers of associations identified, 114 significant associations  
238 and 30 positive control genes were captured using any of the masks, which accounts for  
239 between 27-57% and 50-71% of the findings, respectively, of each mask ([Supplementary](#)  
240 [Figure 1b and Supplementary Figure 1c](#)). Comparing across four masks, pLoF with  
241 AlphaMissense and pLoF with Missense (5/5) resulted in more significant associations  
242 and positive control genes than the pLoF-only mask while keeping a lower false positive  
243 rate than pLoF with Missense (1/5) mask, indicating their superiority over the other masks.  
244 Between these two preferred masks, pLoF with AlphaMissense identified largely similar  
245 or slightly higher numbers of significant gene-trait and gene-disease associations  
246 compared to pLoF with Missense (5/5). Meanwhile, these two methods demonstrated a  
247 similar sensitivity in capturing positive control genes, as indicated by the proportions of  
248 positive control genes among significant associations (17.9% for pLoF with Missense  
249 (5/5), and 17.5% for pLoF with AlphaMissense) ([Figure 2](#)). Furthermore, the pLoF with  
250 AlphaMissense and pLoF with Missense (5/5) masks shared 245 (71 and 75%) significant  
251 association findings and 46 (77 and 80%) of identified positive control genes  
252 ([Supplementary Figure 1b and 1c](#)).

253

254



|                       | pLoF                 | pLoF+AlphaMissense   | pLoF+Missense (5/5)  | pLoF+Missense (1/5)  |
|-----------------------|----------------------|----------------------|----------------------|----------------------|
| <b>AUROC (95% CI)</b> | 0.62 (0.60, 0.64)    | 0.63 (0.61, 0.65)    | 0.64 (0.62, 0.66)    | 0.64 (0.62, 0.66)    |
| <b>AUPRC (95% CI)</b> | 0.037 (0.023, 0.057) | 0.045 (0.030, 0.064) | 0.046 (0.031, 0.064) | 0.039 (0.025, 0.057) |

255  
 256 **Figure 3. Performance curves (ROC and PRC) for all four masks to identify positive**  
 257 **controls genes across all 21 tested traits and diseases with positive control genes.**

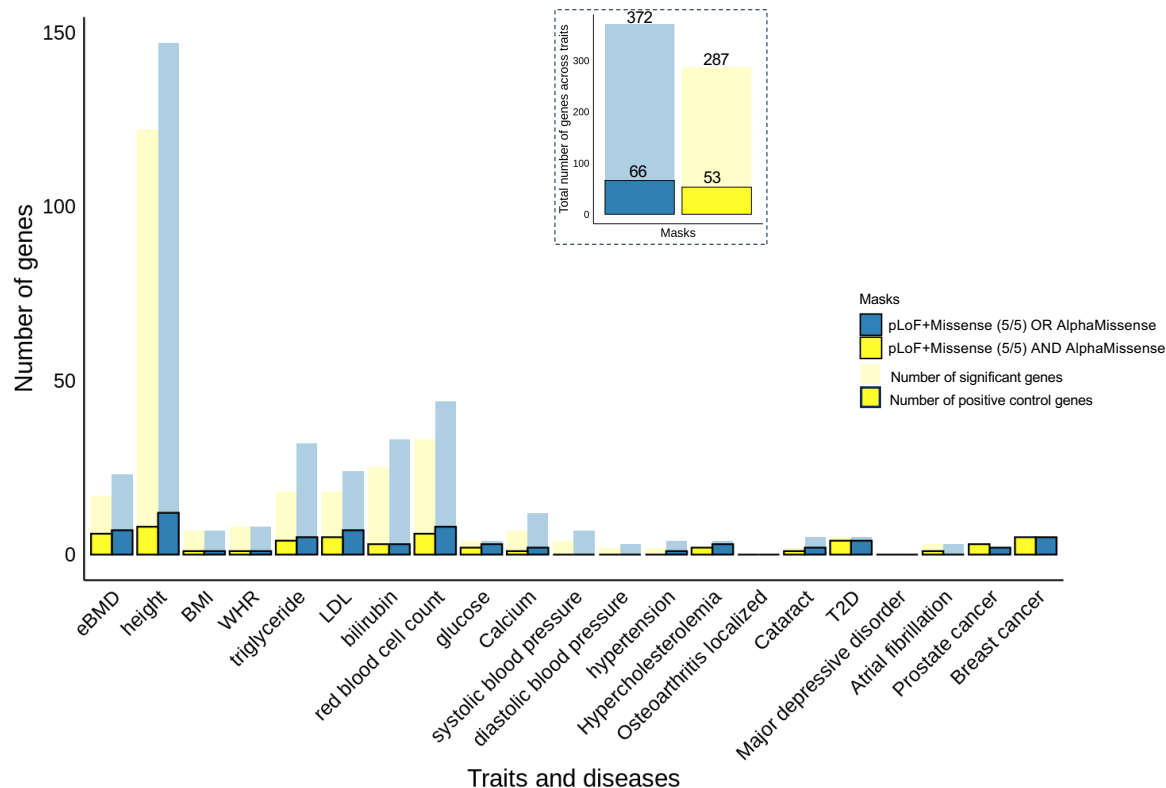
258  
 259 Next, to evaluate whether different masks enhanced the distinction between positive  
 260 control genes and non-positive control genes by offering more divergent *P* values, we  
 261 evaluated the performance of using different masks in classifying these genes by  
 262 calculating the operating characteristic curve (ROC) and precision-recall curve (PRC).  
 263 Upon comparison, we observed that all four masks have statistically indistinguishable  
 264 area under the receiver-operator curves (AUROC) (Figure 3, left panel). However, pLoF  
 265 with Missense (5/5) and pLoF with AlphaMissense have a higher estimated area under  
 266 the precision-recall curves (AUPRC) than the other two masks despite the fact that all the  
 267 95% confidence intervals of AUPRCs overlapped (Figure 3, right panel). Similar AUROC  
 268 and AUPRC patterns can be observed across tested traits, but we did observe that  
 269 specific masks could perform better for certain traits and diseases (Supplementary Figure

270 2). Additionally, we tested whether using different aggregating methods for counting  
271 alleles, in burden tests, across genetic sites within genes changed the mask performance.  
272 Using the maximum number of alternative alleles across sites (the default approach) and  
273 using the sum of the number of alternative alleles in gene-based association analyses  
274 performed similarly ([Supplementary Figure 3](#)).

275

276 Considering that performance was better when pLoF variants were combined with either  
277 Missense (5/5) or AlphaMissense annotated deleterious variants, we further investigated  
278 whether merging AlphaMissense and Missense (5/5) annotations before combining with  
279 the pLoF variants could improve their ability to classify positive control genes. We tested  
280 two designs: using pLoF variants and variants predicted to be deleterious by (1) both  
281 AlphaMissense and Missense (5/5) or by (2) either AlphaMissense and Missense (5/5).

282



283  
 284 **Figure 4. Significant gene-trait and gene-disease associations identified in exome-**  
 285 **wide gene burden analysis across 24 traits using pLoF with the intersection or**  
 286 **union of predicted deleterious variants by AlphaMissense and Missense (5/5). The**  
 287 **inset figure shows the total number of significant genes and positive controls identified by**  
 288 **each mask across all the tested traits and diseases. Abbreviations: estimated bone**  
 289 **mineral density (eBMD), body mass index (BMI), waist-hip circumference ratio (WHR),**  
 290 **serum low-density lipoproteins (LDL), type 2 diabetes (T2D).**

291  
 292 As shown in [Figure 4](#), utilizing deleterious variants predicted by either method identified  
 293 slightly more significant associations (372 pairs), although the precision remained similar  
 294 (17.7%) ([Supplementary Table 7](#)). In contrast, using the overlapping predictions led to  
 295 fewer significant associations (287 pairs) but marginally higher precision (18.5%). The

296 AUROC and AUPRC of these two new mask definitions are similar to other masks  
297 ([Supplementary Figure 4](#)). Overall, little improvement was observed by merging Missense  
298 (5/5) with AlphaMissense.

299

## 300 **Discussion**

301 Gene-based tests offer an elegant way to study the effect of rare coding variants on  
302 human traits by improving statistical power. However, the best way to combine genetic  
303 variants into gene sets is still not fully determined, simply because there are usually many  
304 irrelevant genetic variants in each gene set which may dilute any signal from the set of  
305 causal variants. Hence, such analyses usually rely on algorithms to predict which variants  
306 are likely to be loss of function or missense variants with deleterious effects. As gene-  
307 based analyses are restricted to a likely deleterious subset of variants to increase this  
308 signal to noise ratio, the success of these analyses rest partially on the performance of  
309 the predictions. The emergence of a language model-based variant effect prediction  
310 methods, AlphaMissense, has been suggested to be able to improve gene-based  
311 association. However, our results showed that AlphaMissense did not importantly  
312 outperform the current state-of-the-art masks in gene-based association analyses using  
313 whole-exome data.

314

315 There are multiple reasons why the inclusion of 'likely pathogenic' missense variants, as  
316 annotated by AlphaMissense, does not lead to significant improvements. First, the masks  
317 used in our analysis always included pLoF variants, which already contribute significantly  
318 to the associations observed between genes and traits. Furthermore, the addition of



319 AlphaMissense's predicted pathogenic missense variants expands the analyzed gene  
320 pool by only 184 genes (when added to pLOFs) or 33 genes (when added to pLOF and  
321 Missense (5/5)) beyond those tested using pLoF-only masks. This modest increase in the  
322 number of genes tested offers limited scope for enhancing the performance of gene-  
323 based association tests. Lastly, as noted earlier in this report, other missense annotation  
324 methods largely capture the same 'likely pathogenic' variants identified by AlphaMissense.  
325 Given that all gene-based tests then summarize information across all analyzed variants  
326 in a gene (in various ways), the small number of differently-prediction variants may not  
327 render a large difference in the associated genes.

328

329 AlphaMissense may provide useful and clarifying information in scenarios where  
330 understanding single variant effects is crucial. For example, AlphaMissense could be  
331 particularly helpful in pinpointing actionable genetic sites within known disease-causing  
332 genes. This may be particularly useful for patients with Mendelian diseases without major  
333 structural disruptions in the genetic region (23,24). Additionally, since AlphaMissense  
334 integrates protein structure context into its predictions of variant effects, it should be more  
335 effective when identifying deleterious variants for diseases where protein malfunction  
336 arises from changes in protein conformation. AlphaMissense could also be advantageous  
337 in predicting pharmacogenetic effects that involve protein-drug interactions (25).

338

339 We recognize that while pLoF and missense variant annotations should not be affected  
340 by genetic ancestry, we only performed our analyses in European genetic ancestry  
341 individuals from the UK Biobank, and we only examined 24 traits. Hence, these results

342 will need replication in other populations once sample sizes allow this. Second, the UK  
343 Biobank cohort is a relatively healthy cohort. The number of disease cases is low, which  
344 can limit the statistical power to identify disease-related genes, which may make it more  
345 difficult to compare the performance of different masks in ExWAS. Lastly, there are other  
346 annotation masks that we have not tested, and which may perform differently.  
347 Nevertheless, we compared our results to the best currently available annotations (10),  
348 and we have established that any future work should make comparisons to  
349 AlphaMissense.

350

351 In summary, we found that most of the "likely pathogenic" missense variants identified by  
352 AlphaMissense were also generally predicted to be deleterious by at least one of five  
353 commonly used variant annotation methods. Using masks combining AlphaMissense with  
354 pLoF did not outperform the state-of-the-art missense annotation tools for gene-based  
355 studies.

356

### 357 **Data availability**

358 Individual-level genotype, exome sequencing, and phenotype data is available to  
359 approved researchers via UK Biobank at: <https://www.ukbiobank.ac.uk>. ExWAS  
360 summary statistics will be made available at GWAS Catalog  
361 (<https://www.ebi.ac.uk/gwas/>).

362

### 363 **Code availability**

364 VEP software can be downloaded at <https://github.com/Ensembl/ensembl-vep>. Regenie  
365 software can be found at <https://github.com/rgcgithub/regenie>. UK Biobank exome data  
366 was analyzed using Regenie 3.2.1. All other data analysis was performed using R  
367 (v.4.1.2). Additional codes can be accessed through Github upon publication.

368

## 369 **Acknowledgments**

370 We appreciate the individuals who participated in UK Biobank. This research has been  
371 conducted using UK Biobank data under application ID 27449.

372 The Richards research group is supported by the Canadian Institutes of Health Research  
373 (CIHR: 365825, 409511, 100558, 169303), the McGill Interdisciplinary Initiative in  
374 Infection and Immunity (MI4), the Lady Davis Institute of the Jewish General Hospital, the  
375 Jewish General Hospital Foundation, the Canadian Foundation for Innovation, the NIH  
376 Foundation, Cancer Research UK, Genome Québec, the Public Health Agency of  
377 Canada, McGill University, Cancer Research UK, and the Fonds de Recherche Québec  
378 Santé (FRQS). J.B.R. is supported by an FRQS Mérite Clinical Research Scholarship.  
379 Support from Calcul Québec and Compute Canada is acknowledged. TwinsUK is funded  
380 by the Wellcome Trust, Medical Research Council, European Union, the National Institute  
381 for Health Research (NIHR)-funded BioResource, Clinical Research Facility and  
382 Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust in  
383 partnership with King's College London. Y.C. is supported by an FRQS doctoral training  
384 fellowship and the Lady Davis Institute/TD Bank Studentship Award. G.B.L. is supported  
385 by scholarships from the FRQS, the CIHR, and Québec's ministry of health and social  
386 services.

387

## 388 **Author Contribution Statement**

389 YC - Writing initial draft. YC, GBL, KYHL, CMTG, JBR – Methodology. YC, GBL, KYHL -  
390 Data Analysis. YC, GBL, KYHL, YI, SY, TS, CL, CMTG, JBR - Writing review and editing  
391 draft. JBR – Supervision. All authors commented/revised the manuscript and agreed to  
392 its final submitted version.

393

## 394 **Ethical approval**

395 The UK Biobank was approved by the North West Multi-centre Research Ethics  
396 Committee and informed consent was obtained from all participants prior to participation.

397

## 398 **Competing Interests**

399 J.B.R is the CEO of 5 Prime Sciences ([www.5primesciences.com](http://www.5primesciences.com)), which provides  
400 research services for biotech, pharma, and venture capital companies for projects  
401 unrelated to this research. He has served as an advisor to GlaxoSmithKline and Deerfield  
402 Capital. J.B.R.'s institution has received investigator-initiated grant funding from Eli Lilly,  
403 GlaxoSmithKline, and Biogen for projects unrelated to this research. YC is an employee  
404 of 5 Prime Sciences.

405

406

## 407 Reference

408

- 409 1. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo  
410 MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human  
411 genomes. *Nature*. 2012 Nov 1;491(7422):56–65.
- 412 2. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, et  
413 al. A systematic survey of loss-of-function variants in human protein-coding  
414 genes. *Science*. 2012 Feb 17;335(6070):823–8.
- 415 3. Gibson G. Rare and common variants: twenty arguments. *Nat Rev Genet*. 2012  
416 Jan 18;13(2):135–45.
- 417 4. Weiner DJ, Nadig A, Jagadeesh KA, Dey KK, Neale BM, Robinson EB, et al.  
418 Polygenic architecture of rare coding variation across 394,783 exomes. *Nature*.  
419 2023 Feb 16;614(7948):492–9.
- 420 5. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-Variant Association Analysis: Study  
421 Designs and Statistical Tests. *The American Journal of Human Genetics*. 2014  
422 Jul;95(1):5–23.
- 423 6. Zhou S, Sosina OA, Bovijn J, Laurent L, Sharma V, Akbari P, et al. Converging  
424 evidence from exome sequencing and common variants implicates target genes  
425 for osteoporosis. *Nat Genet*. 2023 Aug;55(8):1277–87.
- 426 7. Cohen JC, Boerwinkle E, Mosley TH, Hobbs HH. Sequence variations in PCSK9,  
427 low LDL, and protection against coronary heart disease. *N Engl J Med*. 2006 Mar  
428 23;354(12):1264–72.

- 429 8. Miosge LA, Field MA, Sontani Y, Cho V, Johnson S, Palkova A, et al. Comparison  
430 of predicted and actual consequences of missense mutations. *Proc Natl Acad Sci*  
431 *U S A*. 2015 Sep 15;112(37):E5189-98.
- 432 9. Cheng J, Novati G, Pan J, Bycroft C, Žemgulytė A, Applebaum T, et al. Accurate  
433 proteome-wide missense variant effect prediction with AlphaMissense. *Science*  
434 (1979). 2023 Sep 22;381(6664).
- 435 10. Backman JD, Li AH, Marcketta A, Sun D, Mbatchou J, Kessler MD, et al. Exome  
436 sequencing and analysis of 454,787 UK Biobank participants. *Nature*. 2021 Nov  
437 25;599(7886):628–34.
- 438 11. Van Hout C V, Tachmazidou I, Backman JD, Hoffman JD, Liu D, Pandey AK, et al.  
439 Exome sequencing and characterization of 49,960 individuals in the UK Biobank.  
440 *Nature*. 2020 Oct;586(7831):749–56.
- 441 12. Wu P, Gifford A, Meng X, Li X, Campbell H, Varley T, et al. Mapping ICD-10 and  
442 ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation.  
443 *JMIR Med Inform*. 2019 Nov 29;7(4):e14325.
- 444 13. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous  
445 variants on protein function using the SIFT algorithm. *Nat Protoc*.  
446 2009;4(7):1073–81.
- 447 14. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A  
448 method and server for predicting damaging missense mutations. *Nat Methods*.  
449 2010 Apr;7(4):248–9.

- 450 15. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human  
451 missense mutations using PolyPhen-2. *Curr Protoc Hum Genet.* 2013  
452 Jan;Chapter 7:Unit7.20.
- 453 16. Schwarz JM, Rödelberger C, Schuelke M, Seelow D. MutationTaster evaluates  
454 disease-causing potential of sequence alterations. *Nat Methods.* 2010  
455 Aug;7(8):575–6.
- 456 17. Chun S, Fay JC. Identification of deleterious mutations within three human  
457 genomes. *Genome Res.* 2009 Sep;19(9):1553–61.
- 458 18. Mbatchou J, Barnard L, Backman J, Marcketta A, Kosmicki JA, Ziyatdinov A, et al.  
459 Computationally efficient whole-genome regression for quantitative and binary  
460 traits. *Nat Genet.* 2021 Jul 20;53(7):1097–103.
- 461 19. Forgetta V, Jiang L, Vulpescu NA, Hogan MS, Chen S, Morris JA, et al. An  
462 effector index to predict target genes at GWAS loci. *Hum Genet.* 2022 Aug  
463 11;141(8):1431–47.
- 464 20. Mountjoy E, Schmidt EM, Carmona M, Schwartzentruber J, Peat G, Miranda A, et  
465 al. An open approach to systematically prioritize causal variants and genes at all  
466 published human GWAS trait-associated loci. *Nat Genet.* 2021 Nov;53(11):1527–  
467 33.
- 468 21. Sobczyk MK, Gaunt TR, Paternoster L. MendelVar: gene prioritization at GWAS  
469 loci using phenotypic enrichment of Mendelian disease genes. *Bioinformatics.*  
470 2021 Apr 9;37(1):1–8.

- 471 22. Yang J, Weedon MN, Purcell S, Lettre G, Estrada K, Willer CJ, et al. Genomic  
472 inflation factors under polygenic inheritance. *Eur J Hum Genet.* 2011  
473 Jul;19(7):807–12.
- 474 23. Staklinski SJ, Scheben A, Siepel A, Kilberg MS. Utility of AlphaMissense  
475 predictions in Asparagine Synthetase deficiency variant classification. *bioRxiv.*  
476 2023 Nov 2;
- 477 24. Utsuno Y, Hamada K, Hamanaka K, Miyoshi K, Tsuchimoto K, Sunada S, et al.  
478 Novel missense variants cause intermediate phenotypes in the phenotypic  
479 spectrum of SLC5A6-related disorders. *J Hum Genet.* 2023 Nov 27;
- 480 25. Park Y, Lauschke V. Towards more accurate pharmacogenomic variant effect  
481 predictions. *Pharmacogenomics.* 2023 Nov;24(16):841–4.
- 482
- 483
- 484
- 485
- 486
- 487
- 488
- 489
- 490
- 491
- 492