HGGA, Volume *5*

Supplemental information

The performance of AlphaMissense to identify genes influencing disease

Yiheng Chen, Guillaume Butler-Laporte, Kevin Y.H. Liang, Yann Ilboudo, Summaira Yasmeen, Takayoshi Sasako, Claudia Langenberg, Celia M.T. Greenwood, and J. Brent Richards

Supplementary Figures

Figure S1. Comparison of the significant and positive control gene-trait and gene-disease pairs identified by ExWAS using different masks. (a) Upset plot shows the degree of overlap between identified significant genes and those designated as positive control genes (b) Venn diagram for significant gene-trait or gene-disease associations identified by different masks (c) Venn diagram for positive control genes identified by different masks.

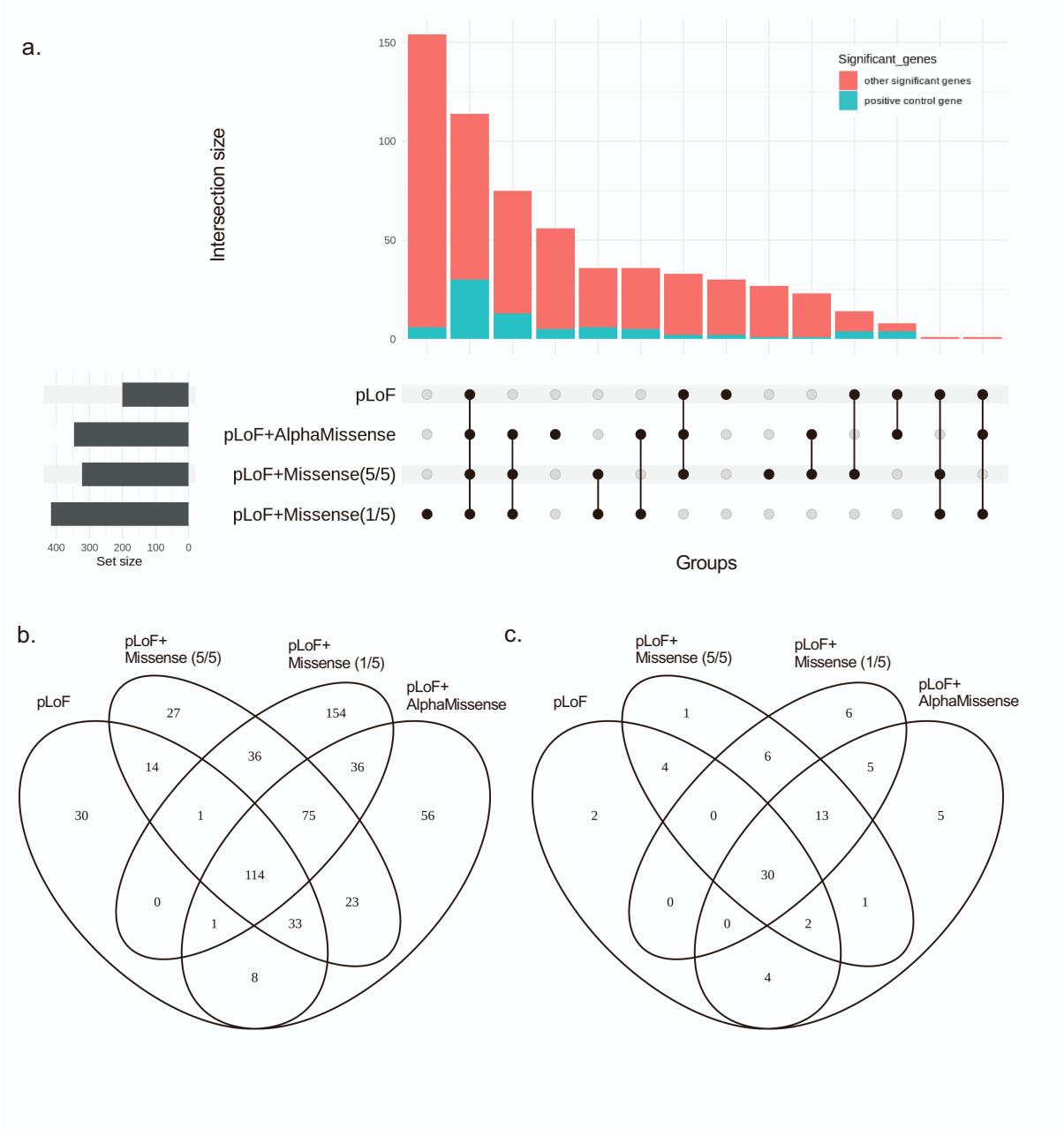


Figure S2. Comparison of the proportion of the positive control genes of each mask across three alternative allele frequency categories.

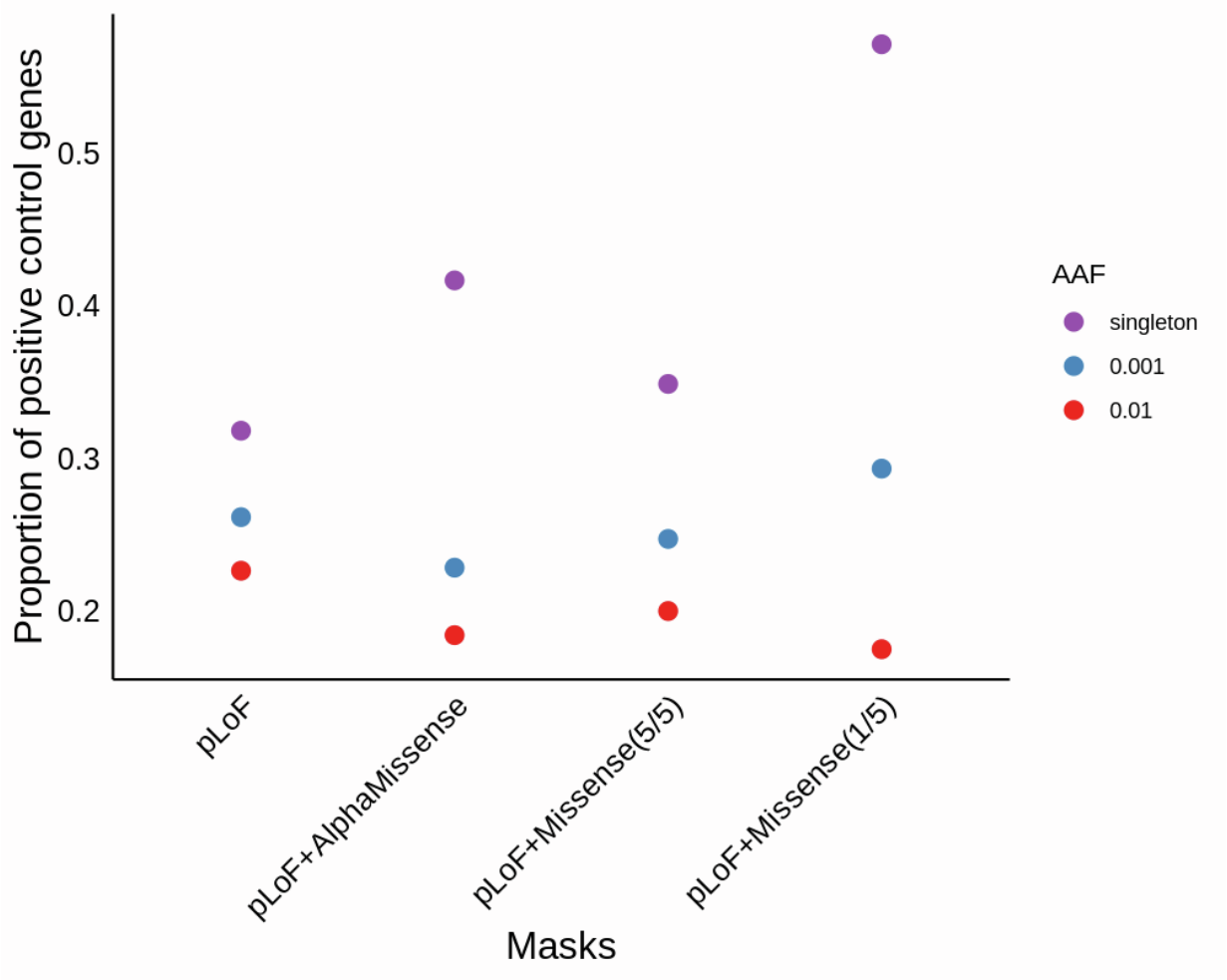


Figure S3. Comparison of the median absolute estimated effect of the 114 significant associations using different masks across three alternative allele frequency categories.

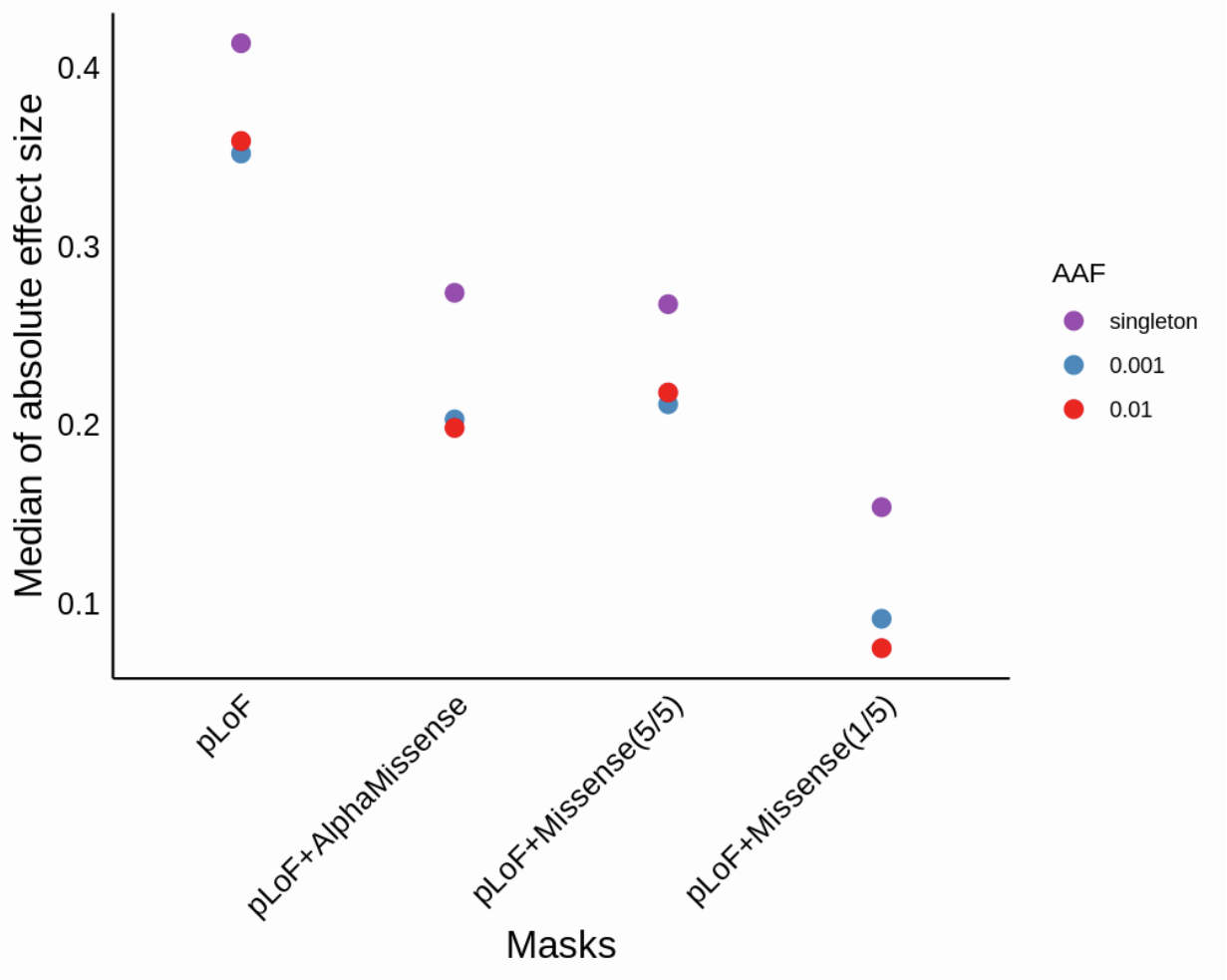


Figure S4. Performance by AUROC and AUPRC for all four masks on identifying positive controls genes in each of the tested traits and diseases. Abbreviations: estimated bone mineral density (eBMD), body mass index (BMI), waist-hip circumference ratio (WHR), serum low-density lipoproteins (LDL), type 2 diabetes (T2D).

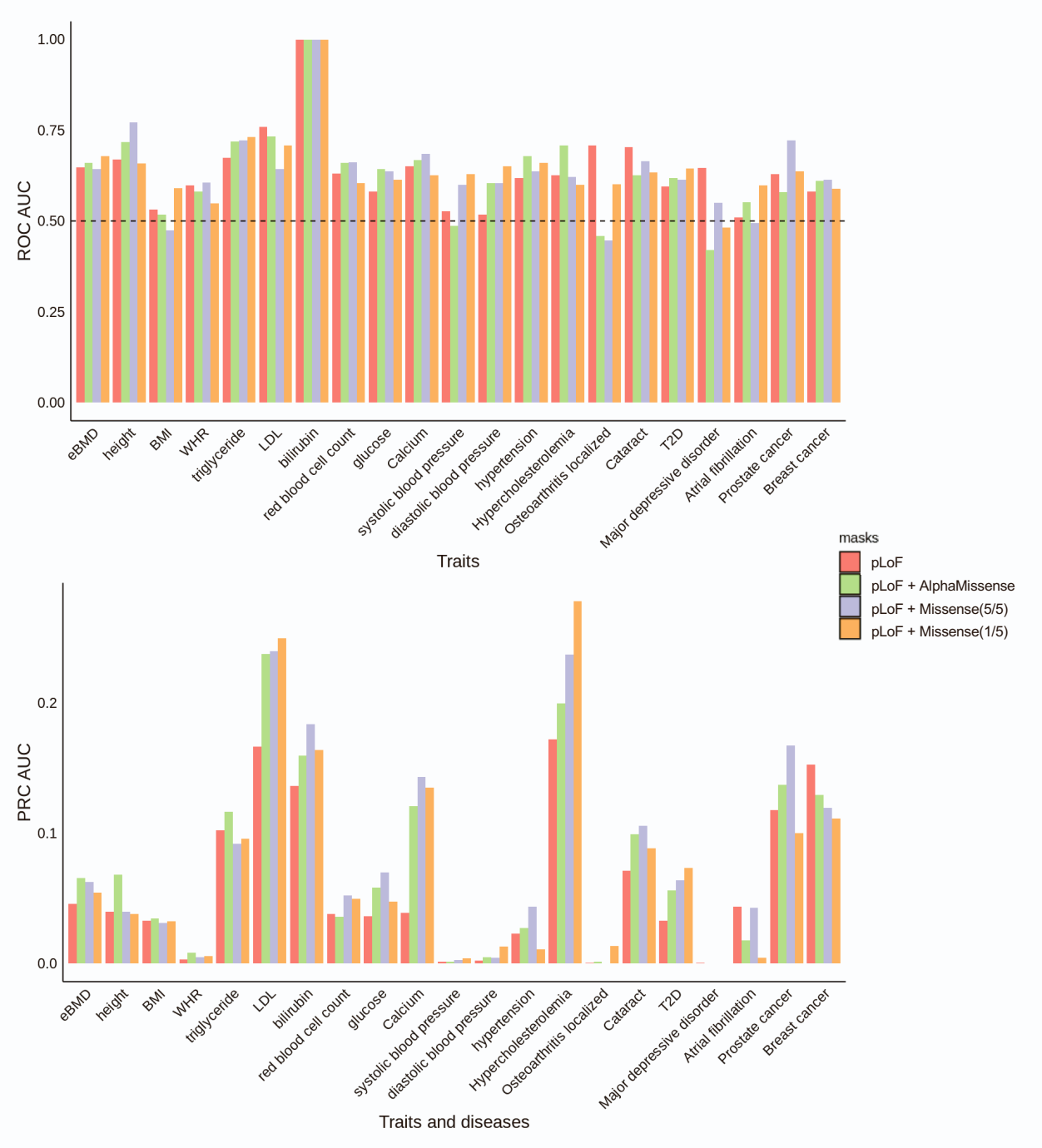
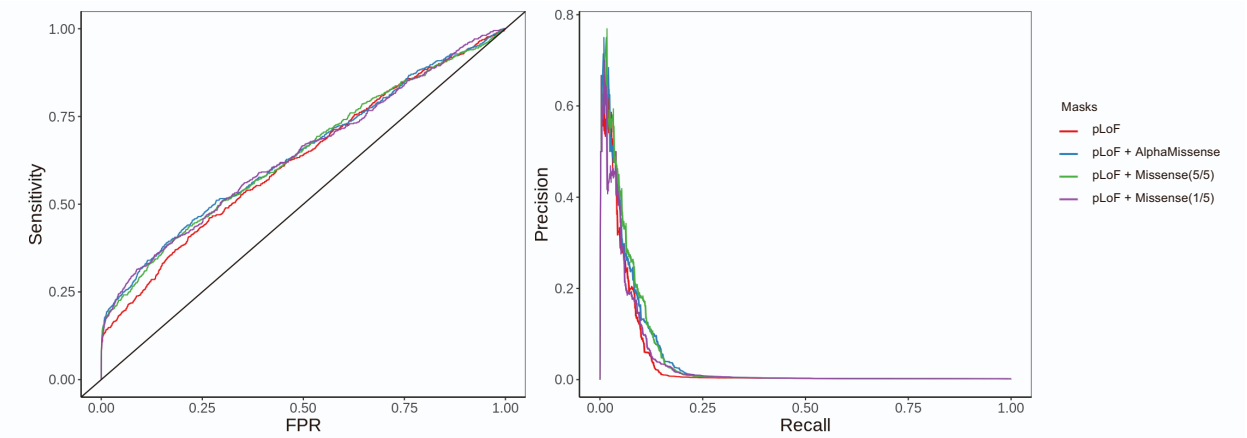


Figure S5. Performance curves (ROC and PRC) for all four masks on identifying positive controls genes across tested traits and diseases using the sum of alternative alleles to pool deleterious variants.



	pLoF	pLoF+AlphaMissense	pLoF+Missense (5/5)	pLoF+Missense (1/5)
AUROC (95% CI)	0.62 (0.60, 0.65)	0.64 (0.62, 0.67)	0.64 (0.62, 0.67)	0.64 (0.62, 0.66)
AUPRC (95% CI)	0.040 (0.026, 0.059)	0.048 (0.030, 0.064)	0.050 (0.031, 0.070)	0.041 (0.026, 0.063)

Figure S6. Significant gene-trait and gene-disease associations identified in exome-wide gene burden analysis across 21 traits and diseases with at least one positive control gene using pLoF with the intersection or union of predicted deleterious variants by AlphaMissense and Missense (5/5). The inset figure shows the total number of significant genes and positive controls identified by each mask across all the tested traits and diseases. Abbreviations: estimated bone mineral density (eBMD), body mass index (BMI), waist-hip circumference ratio (WHR), serum low-density lipoproteins (LDL), type 2 diabetes (T2D).

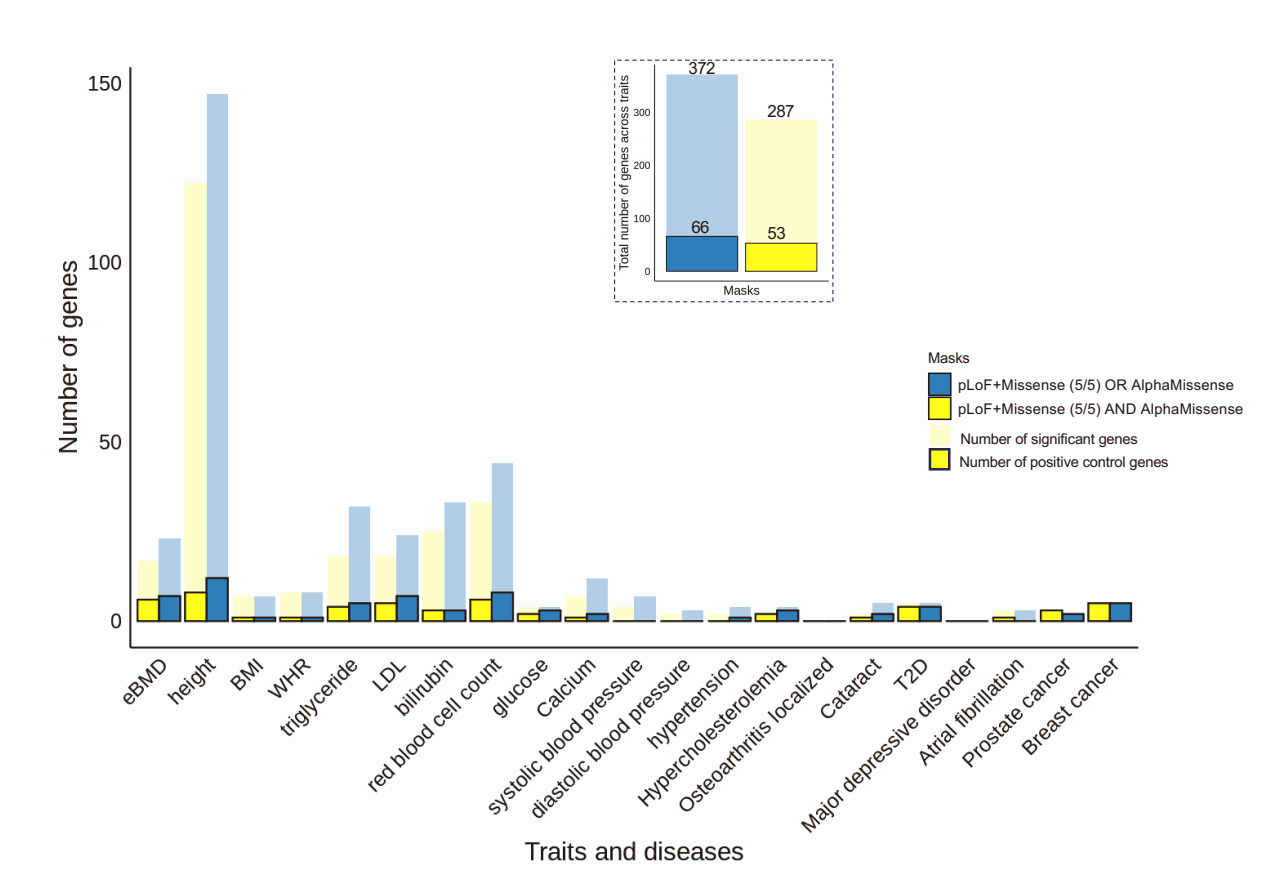
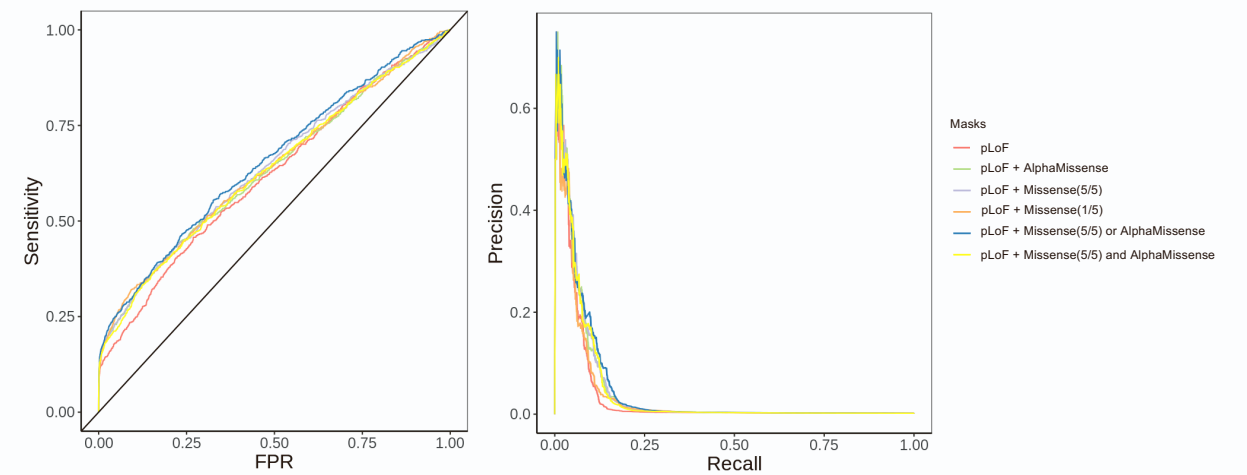
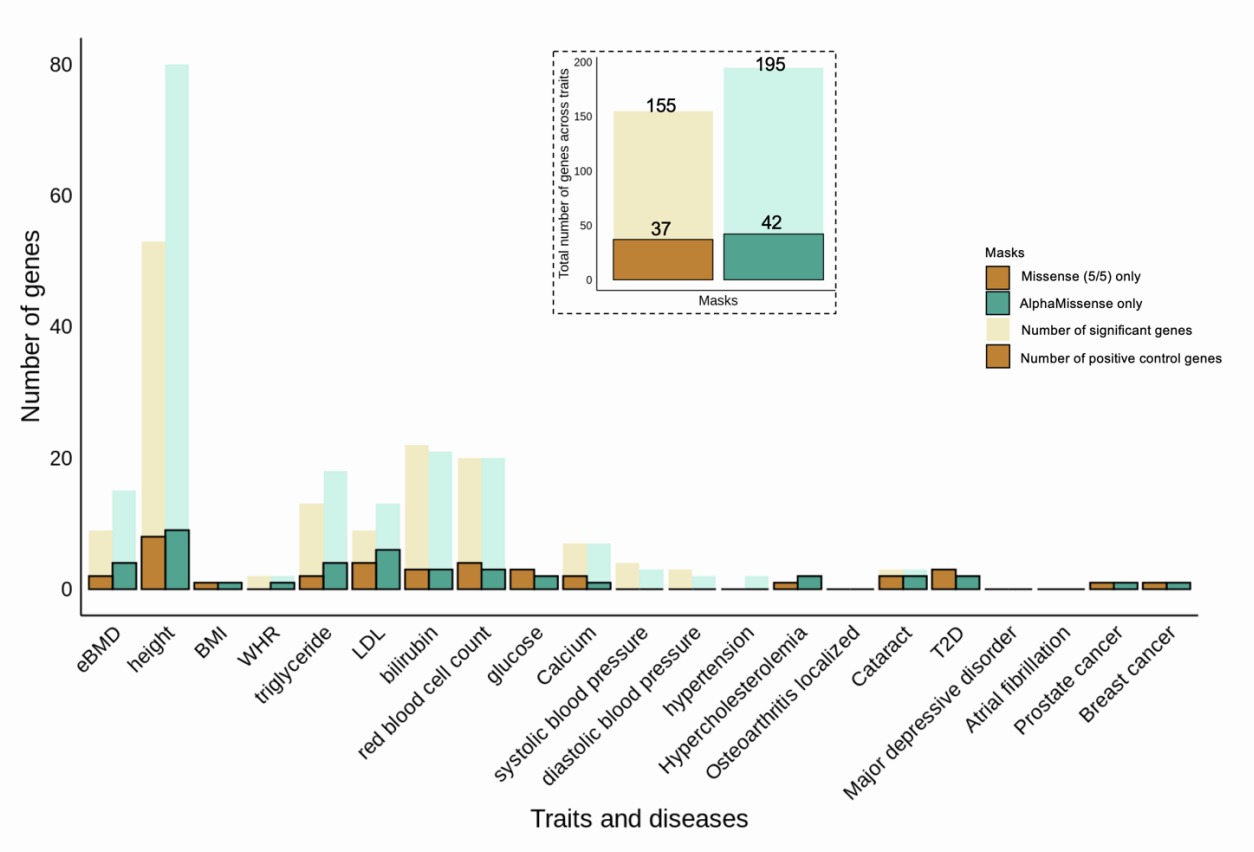


Figure S7. Performance curves (ROC and PRC) for identifying positive control genes across tested traits and diseases by six mask settings.



	pLoF	pLoF+ AlphaMissense	pLoF+ Missense (5/5)	pLoF+ Missense (1/5)	pLoF+ Missense (5/5) OR AlphaMissense	pLoF+ Missense (5/5) AND AlphaMissense
AUROC (95% CI)	0.62 (0.60,0.64)	0.63 (0.61, 0.65)	0.64 (0.62,0.66)	0.64 (0.62,0.66)	0.66 (0.63, 0.68)	0.63 (0.61, 0.66)
AUPRC (95% CI)	0.037 (0.023, 0.057)	0.045 (0.030, 0.064)	0.046 (0.031, 0.064)	0.039 (0.025, 0.057)	0.048 (0.033, 0.067)	0.044 (0.030, 0.063)

Figure S8. Significant gene-trait and gene-disease associations identified using exome-wide gene burden analysis across 21 traits and diseases with at least one positive control gene using predicted deleterious variants by AlphaMissense or Missense (5/5) only. The inset figure shows the total number of significant genes and positive controls identified by each mask across all the tested traits and diseases. Abbreviations: estimated bone mineral density (eBMD), body mass index (BMI), waist-hip circumference ratio (WHR), serum low-density lipoproteins (LDL), type 2 diabetes (T2D).



Supplementary methods

UK Biobank cohort

The UK Biobank is a cohort study that has recruited over 500,000 participants between 40 and 69 years of age at 22 testing centers across the United Kingdom and collected a large set of phenotypes and biological samples. We included in our analyses a total of 444,072 genetically predicted European genetic ancestry individuals with available WES data generated following the OQFE protocol (1) and with measurements of selected phenotypes and diseases. The detailed steps for the sample preparation, sequencing, filtering, and calling of UK Biobank WES data have been previously described (1,2).

Phenotype definitions

From the UK Biobank, we selected 12 continuous traits and 12 diseases for analysis based on the trait sample sizes and whether there were known disease causal genes or drug target genes for each trait. The continuous traits included estimated bone mineral density, serum triglyceride levels, systolic blood pressure, diastolic blood pressure, standing height, serum low-density lipoproteins, serum bilirubin, serum glucose, red blood cell counts, serum calcium level, body mass index, and waist-hip circumference ratio and the 12 diseases included hypertension, hypercholesterolemia, diaphragmatic hernia, osteoarthritis (localized), cataract, type 2 diabetes, major depressive disorder, hypothyroidism, acute renal failure, atrial fibrillation, cancer of prostate (males only) and breast cancer (females only). The sample sizes for the analysis of each trait and disease can be found in Table S1. ICD-10 codes were grouped to construct diseases following

the phecodes system (3). The list of used ICD-10 codes for each phecode can be found in Table S2.

Variant annotation

We annotated the variants from exome sequencing after alignment using the Ensembl Variant Effect Predictor (VEP) (v.110). Variant annotations among transcript ablation, splice acceptor, splice donor, stop gained, frameshift, stop lost, start lost, transcript amplification, feature elongation, and feature truncation, were considered as predicted loss-of-function (pLoF) variants (10). Missense variants were classified with two strategies. The first strategy used AlphaMissense (5), and missense variants were included in our analyses if AlphaMissense predicted them to be “likely pathogenic”. We built a second strategy by combining results from five commonly used annotation methods (i.e., SIFT (6), PolyPhen2 (HDIV) (7), PolyPhen2 (HVAR) (8), MutationTaster (9), and LRT (10)). We classified a missense variant as “likely deleterious” if all five algorithms predicted it to be deleterious (i.e., Missense (5/5)), and “possibly deleterious” if at least one of the five algorithms predicted it to be deleterious (i.e., Missense (1/5)), similar to methods used before (1,4).

Gene-based disease and trait association test

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

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

All analyses were performed using Regenie software (11). The regression analyses included age, age^2, sex, sex*age, sex*age^2, 10 genetic principal components (PC) obtained from common genetic variants (MAF>1%), and 20 genetic PCs obtained from rare genetic variants (MAF<1%) as covariates. The statistical significance threshold was $P < 1.25 \times 10^{-7}$ (0.05 / (approximately 20,000 genes * 20 gene-burden exposures)).

Selection of positive control genes

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

medicines, we compiled a list of positive control genes from two sources. We first included positive control genes from two previous studies where these genes were used to train their algorithms to prioritize disease-causal or drug-targeting genes from genome-wide association study (GWAS) signals (12,13). Their positive control gene lists were generated by combining genetic evidence, drug–target–indication associations, and manual curation from board certified physicians and domain experts. Additionally, we included Mendelian diseases genes from the MendelVar database which was created by integrating functional annotations from the Online Mendelian Inheritance in Man (OMIM), Deciphering Developmental Disorders Study (DECIPHER), Orphanet and Genomics England databases (14). The full list of 519 positive control genes for the selected traits and diseases can be found in Table S3.

Evaluation of classification accuracy

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

Ethical approval

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

Reference

1. Backman JD, Li AH, Marcketta A, Sun D, Mbatchou J, Kessler MD, et al. Exome sequencing and analysis of 454,787 UK Biobank participants. Nature. 2021 Nov 25;599(7886):628–34.
2. Van Hout C V, Tachmazidou I, Backman JD, Hoffman JD, Liu D, Pandey AK, et al. Exome sequencing and characterization of 49,960 individuals in the UK Biobank. Nature. 2020 Oct;586(7831):749–56.
3. Wu P, Gifford A, Meng X, Li X, Campbell H, Varley T, et al. Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation. JMIR Med Inform. 2019 Nov 29;7(4):e14325.
4. Zhou S, Sosina OA, Bovijn J, Laurent L, Sharma V, Akbari P, et al. Converging evidence from exome sequencing and common variants implicates target genes for osteoporosis. Nat Genet. 2023 Aug;55(8):1277–87.
5. Cheng J, Novati G, Pan J, Bycroft C, Žemgulytė A, Applebaum T, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science. 2023 Sep 22;381(6664).
6. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4(7):1073–81.
7. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010 Apr;7(4):248–9.
8. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. 2013 Jan;Chapter 7:Unit7.20.
9. Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods. 2010 Aug;7(8):575–6.
10. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res. 2009 Sep;19(9):1553–61.
11. Mbatchou J, Barnard L, Backman J, Marcketta A, Kosmicki JA, Ziyatdinov A, et al. Computationally efficient whole-genome regression for quantitative and binary traits. Nat Genet. 2021 Jul 20;53(7):1097–103.
12. Forgetta V, Jiang L, Vulpescu NA, Hogan MS, Chen S, Morris JA, et al. An effector index to predict target genes at GWAS loci. Hum Genet. 2022 Aug 11;141(8):1431–47.

13. Mountjoy E, Schmidt EM, Carmona M, Schwartzentruber J, Peat G, Miranda A, et al. An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. Nat Genet. 2021 Nov;53(11):1527–33.
14. Sobczyk MK, Gaunt TR, Paternoster L. MendelVar: gene prioritization at GWAS loci using phenotypic enrichment of Mendelian disease genes. Bioinformatics. 2021 Apr 9;37(1):1–8.