Title: Rare penetrant mutations confer severe risk of common diseases

Authors: Petko Fiziev[1]†, Jeremy McRae[1]†, Jacob C. Ulirsch[1], Jacqueline S. Dron[2,3], Tobias Hamp[1], Yanshen Yang[1], Pierrick Wainschtein[1], Zijian Ni[4], Joshua G. Schraiber[1], Hong Gao[1], Dylan Cable[5], Yair Field[1], Francois Aguet[1], Marc Fasnacht[1], Ahmed Metwally[1], Jeffrey Rogers[6,7], Tomas Marques-Bonet[8,9,10,11], Heidi L. Rehm[2,3,12], Anne O'Donnell-Luria[3,12,13], Amit V. Khera[2,3,14], Kyle Kai-How Farh[1]*

Affiliations:

[1]Artificial Intelligence Laboratory, Illumina, Inc.; San Diego, California 92122, USA

[2]Center for Genomic Medicine, Massachusetts General Hospital, Boston, Massachusetts 02114, USA

[3]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard; Cambridge, Massachusetts 02142, USA

[4]Department of Statistics, UW Madison; Madison, Wisconsin 53706, USA

[5]Department of Electrical Engineering and Computer Science, MIT; Cambridge, Massachusetts 02142, USA

[6]Human Genome Sequencing Center and Department of Molecular and Human Genetics, Baylor College of Medicine; Houston, Texas 77030, USA

[7]Wisconsin National Primate Research Center, University of Wisconsin; Madison 53715, USA

[8]Institute of Evolutionary Biology (UPF-CSIC); 08003 Barcelona, Spain

[9]Catalan Institution of Research and Advanced Studies (ICREA); 08010 Barcelona, Spain

[10]CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST); 08003 Barcelona, Spain

[11]Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona; 08193 Barcelona, Spain

[12]Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital; Boston, Massachusetts 02114, USA

[13] Division of Genetics and Genomics, Boston Children's Hospital; Boston, Massachusetts 02115, USA

[14] Verve Therapeutics, Cambridge, Massachusetts 02215, USA

† These authors contributed equally to this work

*Corresponding author. Email: kfarh@illumina.com

Abstract:
We examined 454,712 exomes for genes associated with a wide spectrum of complex traits and common diseases and observed that rare, penetrant mutations in genes implicated by genome-wide association studies confer ~10-fold larger effects than common variants in the same genes. Consequently, an individual at the phenotypic extreme and at the greatest risk for severe, early-onset disease is better identified by a few rare penetrant variants than by the collective action of many common variants with weak effects. By combining rare variants across phenotype-associated genes into a unified genetic risk model, we demonstrate superior portability across diverse global populations compared to common variant polygenic risk scores, greatly improving the clinical utility of genetic-based risk prediction.

One sentence summary:
Rare variant polygenic risk scores identify individuals with outlier phenotypes in common human diseases and complex traits.

Main text:
Genome-wide association studies (GWASs) have convincingly identified tens of thousands of common variants underlying complex human traits and diseases (*1*), although several key challenges remain. First, pinpointing which genes these predominately non-coding variants affect is non-trivial, hindering biological insight into disease mechanisms. Second, individual common variants have modest effects on disease risk, resulting in weak aggregate predictors with limited clinical utility and portability between populations (*2–4*). In contrast to GWASs, rare coding variant studies directly link perturbed gene function to specific phenotypes. For individuals with cancer or rare genetic diseases, analysis of whole exome sequencing (WES) routinely uncovers rare, highly penetrant variants that can dramatically alter the course of clinical management (*5–8*) and drive treatment decisions (*9*, *10*). However, in the context of common diseases, the role of rare coding variants has not been established to the same extent due to lack of methods for accurately predicting variant function and insufficient cohort sizes.

Recent large-scale genome and exome sequencing studies of the general population have revealed that the average person carries dozens of potentially deleterious rare variants that have arisen through recent germline mutation (*11*). These studies provide the opportunity to move beyond rare genetic disease and examine the impact of medium-to-large effect rare coding variants on a comprehensive set of complex human traits and diseases. In practice, individually rare variants are often combined into burden tests to more powerfully discover genes underlying these phenotypes, but these tests are limited by our ability to distinguish pathogenic from benign variants. Here, we show that our recently developed method PrimateAI-3D (*12*), a 3-D convolutional neural network trained on common genetic variants from 233 primate species, accurately quantifies missense variant pathogenicity, resulting in improved gene discovery across 454,712 individuals in the UK Biobank (*13–15*). We then show how rare variants in these genes can be combined into a unified genetic risk score which has distinct advantages over common variant polygenic risk scores, offering a glimpse into the potential utility of personal genome sequencing for the general population.

PrimateAI-3D empowers gene discovery in rare variant association tests

To identify genes underlying complex human traits and diseases, we performed rare variant burden tests for 90 well-powered, non-redundant clinical and quantitative phenotypes, including both medical diagnoses and commonly measured laboratory tests, for 454,712 individuals in the UK Biobank who underwent WES (Table S1-S3) (*16*). Using an allele frequency (AF) threshold of 0.1%, we detected 1,841 gene-phenotype associations with loss-of-function (LoF) variants, 1,510 associations with missense variants, and 3,035 associations combining missense and LoF variants (average of 33.7 per phenotype) at a false discovery rate (FDR) of 5% (**Fig. 1A**). When we applied PrimateAI-3D (*12*) to classify pathogenic and benign missense variants, we improved gene discovery by 73%, identifying 1,285 more gene-phenotype associations at the same FDR (**Fig. 1A**, **Fig. S1**, and Table S4). As a negative control, we repeated the test considering rare synonymous variants but detected only 28 gene-phenotype associations. Taken together, these results show that our rare variant tests are well calibrated and that PrimateAI-3D pathogenicity predictions improve gene discovery.

We undertook several additional approaches to validate our gene-phenotype associations and to compare to prior efforts. First, we investigated the strength of support from common variant studies for the gene-phenotype pairs identified by our approach. After performing matched GWASs for the 90 phenotypes (Table S5) (*16*), we observed that 70% of the 3,035 gene-phenotype pairs had a significant GWAS variant ($P < 5\times10^{-8}$) within 1 megabase of the transcription start site. Next, we compared our results to a recent rare variant association study in the same UK Biobank cohort (*17*) (**Fig. 1B**). Backman et al. used a burden test which included all LoF variants but permitted only missense variants predicted to be deleterious by five commonly used missense pathogenicity classifiers (*18*). For matched phenotypes and significance thresholds (*16*), we identified 23% more gene-phenotype pairs (Table S6). Gene-phenotype pairs identified exclusively in the present study were more enriched for genes implicated by matching GWASs and overlapped more with genes in related Mendelian diseases (**Fig. 1C**, Table S7), which supports their relevance to complex trait biology. Third, we benchmarked PrimateAI-3D against 15 other pathogenicity classifiers by integrating them into our burden testing pipeline. Again, gene-phenotype pairs detected exclusively by PrimateAI-3D had consistently higher enrichments for GWAS genes for the same trait compared to any other method (**Fig. S2**). Finally, we assessed how well each classifier could predict the effect size of individual variants on phenotype across 62 gene-phenotype pairs detected without variant prioritization (Table S8) (*16*) and again observed that PrimateAI-3D outperformed all other methods (median Wilcoxon $P=8\times10^{-7}$, **Fig. 1D, Fig. S3**).

Having comprehensively validated our use of PrimateAI-3D for rare variant burden testing, we explored the correlations we observed between PrimateAI-3D scores, clinical laboratory measurements, and ages of onset for common diseases. In general, we observed a linear relationship with the quantitative measurements and an inverse correlation with age of disease onset (Table S9). We focus on the examples of *LDLR* and *PCSK9* with low-density lipoprotein (LDL) cholesterol levels and *GCK* with glycated hemoglobin A1c (HbA1c) to illustrate these general findings (**Fig 1E-G**). Overall, 1,307 individuals (0.3%) carried rare, potentially deleterious missense variants in the *LDLR* gene in which pathogenic mutations can cause familial hypercholesterolemia and early-onset cardiovascular disease (*19*, *20*). PrimateAI-3D scores of missense variants in *LDLR* were significantly correlated with LDL levels (Spearman $\rho = 0.50$, $P=8\times10^{-38}$) (*16*). Individuals with variants that had scores near 0 had LDL cholesterol levels indistinguishable from non-carriers, whereas those with scores near 1 had elevated LDL cholesterol levels similar to LoF variant carriers (**Fig. 1E**, upper panel). Among individuals who received a clinical diagnosis of dyslipidemia, PrimateAI-3D scores correlated inversely with age of diagnosis (Spearman $\rho = -0.35$, $P = 3\times10^{-12}$). The most deleterious missense variants advanced age of disease onset by ~15 years, similar to that observed for LoF carriers (**Fig. 1E**, lower panel).

We next examined rare variants in the *PCSK9* gene, a target of cholesterol-lowering medications (*21*). Rare missense variants with high PrimateAI-3D scores in *PCSK9* were correlated with decreased LDL cholesterol levels (Spearman $\rho = -0.32$, $P = 3\times10^{-13}$), and acted in the opposite direction of deleterious *LDLR* variants (**Fig. 1F**, upper panel). LDL cholesterol levels increased with age at a similar rate (0.2 mmol/L per decade of normal aging) regardless of *PCSK9* carrier status, but individuals carrying prioritized rare variants in *PCSK9* had an average of 0.6 mmol/L lower LDL cholesterol levels at any given age (**Fig. 1F**, lower panel). As a consequence, fewer of these carriers had moderate-to-severe hypercholesterolemia (LDL cholesterol > 4.1 mmol/L or 160 mg/dL) or elevated cardiovascular disease risk (*22*), while those that did manifested these symptoms later in life.

Finally, we observed similar relationships between rare deleterious variants in *GCK* and HbA1c, a proxy for blood glucose levels and a diagnostic laboratory marker for type 2 diabetes (pre-diabetes HbA1c > 42 mmol/mol; diabetes HbA1c > 48 mmol/mol) (**Fig. 1G**) (*23*). Analogous to LDL cholesterol, HbA1c levels increased with age, matching the steep rise of diabetes prevalence with age observed in epidemiological studies (*24*). Rare deleterious variants in *GCK* elevated HbA1c levels by an average of 5.1 mmol/mol relative to benign variant carriers and non-carriers, 4.6-fold higher than the average rise in HbA1c levels per decade of normal aging. Correspondingly, this increased the fraction of individuals with diabetes between ages 40-50 from 3.8% to 24.8% (6.6-fold increase) for carriers of rare deleterious variants. In summary, our results across clinically relevant phenotypes such as LDL cholesterol and HbA1c demonstrate the utility of PrimateAI-3D to distinguish pathogenic from benign variants and highlight the capacity of rare high-penetrance variants to accelerate or delay the age of onset of common diseases by decades.

Rare variant polygenic risk scores identify individuals most at risk for common diseases

Recent exponential human population growth has created an abundance of rare variants via naturally-occurring mutations without providing adequate time for selection to remove those with deleterious consequences (*25*, *26*). In the UK Biobank cohort, we observed that each person carries an average of 2.96 rare deleterious missense variants and 0.97 rare LoF variants within one or more of the genes identified from our burden test. Consistent with models of negative selection (*16*, *27*, *28*), we find that rare variants exerted far greater per allele effects on human phenotypes than common variants across a subset of 893 genes implicated by both rare and common variant studies, with rare deleterious variants having on average an 11.2-fold larger effect than common GWAS variants at the same loci (**Fig. 2A**, **Fig. S4**). Within each allele frequency bin, LoF variants had the highest per allele effects followed by missense variants (PrimateAI-3D > 0.8) and cryptic splice variants (SpliceAI score > 0.2) (*29*). Benign missense (PrimateAI-3D < 0.2) and synonymous variants had nearly null per-allele effects on phenotype, even as singletons. Given the high overall prevalence and strong effect sizes of rare deleterious variants in the predominately healthy UK Biobank cohort, we reasoned that a single polygenic score combining these variants may effectively identify individuals at high risk for complex disease.

Existing polygenic risk score (PRS) models of common disease largely omit rare variants due to challenges in interpreting variants of uncertain significance and estimating the magnitude of variant effects (*30*). Here we propose a complementary, rare variant PRS model, based on a weighted sum of rare deleterious variants from multiple phenotype-associated genes, utilizing PrimateAI-3D for variant effect estimation. To construct the model, we first split the UK Biobank cohort into training and testing subsets and then fit a linear model to each phenotype on the rare variants (AF < 0.1%) in associated genes, weighted by PrimateAI-3D predicted effect size (Table S10) (*16*). For comparison, we also constructed common variant (AF >1%) PRS models by performing GWAS on the training dataset and applying the method of clumping and thresholding (Table S11) (*31*).

We illustrate the components of the rare variant PRS model using total cholesterol levels as a representative example and show that it identifies the complex network of genes, cell types, and pathways that underpin lipid metabolism (**Fig. 2B**). Rare deleterious variants in the 31 associated genes that contribute to the rare variant PRS model shifted cholesterol levels by ~0.38 mmol/L on average, 10-fold the average effect size of the 563 variants in the common variant PRS model (0.040 mmol/L) (Table 1). Out of these 31 genes, 25 were previously known to play central roles in lipid homeostasis (*32*): from absorption of cholesterol via intestinal enterocytes (*ABCG5*) (*33*), to regulation of serum LDL concentrations (*PCSK9*) (*34*), to comprising key components of lipoproteins (*APOB*) (*35*), to lipid scavenging in macrophages (*STAB1*) (*36*). Beyond identifying genes pertinent to cholesterol metabolism, the direction of effect for these rare deleterious variants was consistent with each gene's known role in the pathway. Notably, many of the genes that produce downregulatory effects on cholesterol levels are therapeutic targets that offer alternatives to statin-based cholesterol reduction for cardiovascular disease, such as *PCSK9* and *NPC1L1* inhibitors (*37*, *38*). While the average chance of an individual carrying a rare deleterious variant for any given gene was only 0.4%, when summed across all 31 genes, 1 in 8 individuals carried a rare, high-penetrance variant for cholesterol.

We sought to evaluate the predictive power of the rare variant PRS and the corresponding common variant PRS, as well as a combination of the two methods, on the 10% of UK Biobank individuals that had been withheld for testing. Across 78 quantitative phenotypes, the unified PRS performed best with an average Pearson correlation of 0.307 (**Fig. 2C**, **Fig. S5**), compared to 0.058 and 0.303 for the rare variant PRS and common variant PRS, respectively. Consistent with the correlations, the average phenotypic variance explained was 10.4%, 0.4%, and 10.1%, respectively. We also evaluated rare variant PRS models constructed using 15 other variant pathogenicity classifiers and observed that PRSs based on PrimateAI-3D outperformed all other methods (**Fig. 2D**), underscoring the importance of accurate pathogenicity prediction to rare variant PRS performance. Overall, these observations are consistent with previous studies that have demonstrated that, in aggregate, rare variants explain less genetic heritability than common variants (*39*).

Although rare variant PRSs underperformed for average phenotype predictions, we reasoned that they may outperform common variant PRSs for identifying individuals at phenotypic extremes, which is more relevant for clinical screening and risk management. Indeed, individuals with an outlier phenotype (z-score ≥3) were 10-fold more likely than the overall population to have a rare variant PRS score in the 0.1^{st} or 99.9^{th} percentile, compared to 3-fold for common variant PRS (P=0.0026, **Fig. 2E**, **Fig. S6**). Across 78 phenotypes, rare variant PRSs significantly outperformed common variant PRSs at identifying individuals with outlier phenotypes at the 99.9% percentile (P=0.0032), had comparable performance at the 99% percentile (difference not significant), and underperformed at the 90% percentile ($P=5.2\times10^{-7}$) (**Fig. 2F, Fig. S7)**. Empirically, the prevalence of many complex human diseases is below 1%, including Parkinson's disease (0.3%) (*40*), multiple sclerosis (0.3%) (*41*), myocardial infarction before age 40 (0.6%) (*42*), and type 1 diabetes (0.2%) (*43*), supporting the relevance of these outlier phenotype thresholds for evaluating clinical risk prediction models.

For two diseases, type 2 diabetes and dyslipidemia, we evaluated the ability of common and rare PRS models to identify individuals exceeding pre-defined diagnostic clinical thresholds (HbA1c > 42 mmol/mol and LDL cholesterol > 4.9 mmol/L, respectively) (**Fig. 2G**). Up until approximately 4-fold increased odds of disease, the common variant PRS identified more at-risk individuals, whereas after this threshold the rare variant PRS overtook the common variant PRS. Because the rare and common variant PRS models use non-overlapping sets of variants, combining them into a unified model enables the identification of significantly more individuals at high disease risk (odds ratio ≥4X) than common variant PRSs alone (type 2 diabetes, 1912 vs 542, $P=1.4\times10^{-178}$, dyslipidemia 7858 vs 6306, P=1.2×10-

[39]). Taken together, these findings suggest that incorporating rare variants into PRSs can outperform common variant PRSs for identifying outlier individuals (*30*, *44*) who are most likely to require treatment or to suffer severe, early-onset manifestations of disease and for whom preventative screening would be most impactful (*45*, *46*). Moreover, the ability to point to a single penetrant variant as the primary cause of the phenotype may increase the potential clinical actionability of rare deleterious variants with respect to prognosis, management, and therapeutic interventions (*47*).

Portability of rare variant polygenic risk scores, and validation in an independent, multi-ancestry cohort

Common variant PRS models derived from European populations have poor portability in non-European populations, which may contribute to future health disparities once adopted into clinical practice (*4*). Even when applied to populations with similar ancestry, common variant PRSs have decreased performance owing to differences between the cohorts used for training and testing (*48*, *49*). We thus set out to evaluate the robustness of our rare variant PRSs across independent cohorts and ancestries. We first applied 16 rare variant PRS models, which had been trained on UK Biobank European-ancestry individuals, to predict quantitative phenotypes in 20,708 European individuals from the Massachusetts General Brigham Biobank (MGB, Table S12) (*50*). Across 16 phenotypes, the average predictive performance of the rare variant PRS model was similar in the two cohorts (Pearson R = 0.53), with a median phenotype correlation of 0.078 between the rare variant PRS and the UK Biobank withheld test cohort, compared to 0.084 for the MGB cohort (**Fig. 3A**). Notably, the rare variant PRS models achieve approximately equal performance in the two cohorts despite 43% of the rare deleterious variants in the MGB cohort never appearing in the UK Biobank cohort that was used for model training. Thus, unlike common variant PRSs, rare variant PRSs appear largely portable across cohorts with similar ancestry.

We next evaluated the performance of our rare and common variant PRS models, which had been trained only on individuals of European ancestry, in individuals of non-European ancestry from the UK Biobank and MGB. As a control, we ensured that the number of variants used per person in the rare variant PRS was closely matched for different ancestries by applying ancestry-specific allele frequency filters (AF < 0.1%) (**Fig. S8**) and verified that the resulting PRS distributions were similar across ancestries (**Fig. S9**). Consistent with previous reports, the median common variant PRS correlation with phenotype was 84% lower in individuals with African ancestry ($P=2.1\times10^{-61}$), 62% lower in individuals with East Asian ancestry ($P=3.4\times10^{-26}$), and 51% lower in individuals with South Asian ancestry ($P=2.5\times10^{-33}$) relative to the correlation in individuals with European ancestry (**Fig. 3A**). In contrast, the rare variant PRS correlation was substantially more portable with smaller reductions in median correlation of 54%, 14%, and 23%, respectively. To assess the portability of the rare variant PRS on a more clinically relevant task, we selected individuals with PRS scores at the upper and lower ends of the phenotype distribution (top or bottom 0.5%) and observed that the average phenotype differences between the two groups were similar for Europeans and non-Europeans in both the UK Biobank withheld test cohort (Pearson R = 0.85, **Fig. 3B**) and the MGB cohort (Pearson R = 0.88, **Fig. S10**). Overall, rare variant PRS models trained in Europeans performed better when tested in non-Europeans than Europeans for 14 out of 52 phenotypes, compared to the common variant PRS models which performed worse when tested in non-Europeans for all 52 phenotypes (**Fig. 3C**).

While rare variant PRSs appear to generalize better across ancestries than common variant PRSs, their average performance still decreases in non-European populations. However, this appears to be distinct from the portability issues experienced by the common variant PRS, where causal variant identification remains difficult due to linkage disequilibrium. We hypothesized that the current European bias is due primarily to more accurate allele frequency estimates within the more numerous

European individuals in the cohort and in current population databases, resulting in the inadvertent inclusion of common non-European variants into the rare variant PRS that dilute its performance. To test this hypothesis, we restricted our evaluation to ultra-rare variants (seen only once in the UK Biobank and absent from the TOPMed allele frequency database) to minimize common variant leakage. We found that PrimateAI-3D variant effect size predictions were equally accurate in European and non-European ultra-rare variants (difference not significant, **Fig. 3D**) but were significantly less accurate for non-European variants at the default allele frequency threshold of 0.1% ($P=1.5\times10^{-4}$ using PrimateAI-3D). As further indication that these issues are independent of variant effect prediction, we show that rare variant PRSs derived using only LoF variants (without PrimateAI-3D) displayed similarly decreased performance in non-European individuals (**Fig. S11A**) and that the European bias could be reduced by using L_1 regularization to limit overfitting (**Fig. S11B**). Similar challenges have been reported for rare genetic disease diagnosis in non-European populations (*51*, *52*), where inaccurate allele frequency estimates make it difficult to preclude ancestry-specific common variants as potential causes of disease. Therefore, as population allele frequency panels become more accurate and globally inclusive, we expect that the portability of rare variant PRSs will continue to improve.

The convergence of common and rare variant genes forecasts future improvements in rare variant PRSs

Looking forward, we explored how much the performance of rare variant PRS approaches is expected to improve as exome sample sizes increase, focusing first on our ability to identify additional exome-wide significant genes (FDR < 5%). We performed association tests in down-sampled subsets of the UK Biobank cohort and observed that the number of significant associations increased linearly with sample size for both rare variant burden tests (FDR < 5%) and common variant GWAS loci ($P < 5\times10^{-8}$) (**Fig. 4A, Fig. S12**). On average, PrimateAI-3D enabled discovery of the same number of exome-wide significant genes using 1.8-fold smaller cohort sizes compared to when missense prioritization was not applied. Consistent with the improved detection of phenotype-associated genes, we observed a linear increase in the number of variants carried by each individual that could be included in the rare variant PRS model (**Fig. 4B**). At the full cohort size, we found that 97% of individuals carried a rare penetrant variant in one or more of the associated genes for the 90 clinical and quantitative phenotypes in the study (**Fig. S13**). Although effect sizes were lower in newly identified genes (**Fig. 4C**), rare variant PRS performance improved steadily, with each doubling of discovery cohort size corresponding to an 88% improvement in variance explained (**Fig. 4D, Fig. S14**).

Our forecasting analyses suggest that rare variant PRSs will continue to meaningfully improve as cohort sizes increase, with newly discovered genes preferentially enriched at GWAS loci (**Fig. 4E**), consistent with recent work showing convergent biological pathways behind both rare and common variant heritability (*39*). The observed overlap of common variant GWAS hits and rare variant burden test genes was highly phenotype-specific (**Fig. 4F, Fig. S15, Fig. S16**), and was not explained by linkage disequilibrium, as we regressed out the effects of significant GWAS variants and population structure before applying the rare variant burden tests. Focusing on a subset of well-powered GWAS loci that could be unambiguously mapped to a single protein-coding gene (*16*), we found that 64% of common variant GWAS genes showed significant association in the rare variant burden test ($P < 0.05$, **Fig. 4G**). The fraction of genes with rare variant signal declined for weaker GWAS hits ($P = 3\times10^{-37}$), as well as for genes under strong evolutionary selection ($P = 5\times10^{-4}$) (*53*), reflecting reduced statistical power to detect enrichments in genes that either have weak phenotypic effects, or that have been depleted of deleterious variants by selective constraint. Similarly, we observed that shorter genes, with consequently fewer variants, were also less likely to be significant in the rare variant burden test ($P = 7\times10^{-6}$). Although we found that only 186 (6%) out of 3,097 unambiguously GWAS-implicated genes

reached the stringent exome-wide significance threshold for inclusion in the rare variant PRS (FDR < 5%), 625 (20%) were nominally significant on the burden test at a P-value threshold < 0.05, indicating that rare variant associations are likely to be discovered at these genes with larger cohort sizes. In summary, our empirical studies of the convergence of common and rare variant associations suggest that allelic series underlie most of the genes implicated in human pathophysiology and can be leveraged in ever growing sequencing cohorts to improve rare variant PRS performance.

Discussion

Understanding the role of rare penetrant variants in common diseases is of prime interest to both precision medicine (*5–7*) and targeted drug development (*21*, *54*, *55*). In this study, we leverage PrimateAI-3D's state-of-the-art predictions to model the quantitative effects of each variant on multiple phenotypes, uncovering the role played by rare penetrant variants in common human diseases and complex traits. We demonstrate the complementary utility of common and rare variants for predicting the risk of human diseases, observing that common variants explain a higher proportion of total population variance, whereas rare variants more readily identify outlier individuals at the greatest risk for severe, early-onset disease (*45*, *46*). Our results establish that the personal genome of an otherwise healthy individual is not quiescent with limited actionable potential (*56*) but instead carries a substantial burden of rare consequential variants, the clinical utility of which will be more fully realized as variant interpretation improves and discovery cohort sizes increase.

At present, the two greatest barriers to the clinical adoption of common variant PRS models for use in precision medicine are their limited generalizability between populations with different ancestries and their weak discriminatory capability to identify individuals at high risk for disease (*57*). Specifically, the inclusion of predominately non-coding variants with small effects that are non-causal, but disease-associated due to linkage disequilibrium, substantially impairs common variant PRS performance (*58*, *59*). In comparison, our rare variant PRS models are anchored on PrimateAI-3D's predictions of missense variant effect size and are largely uninfluenced by the effects of ancestry, since the PrimateAI-3D model was derived from common variants in 236 species of non-human primates. This gives rare variant PRS models an advantage over common variant PRS models at generalizing to cohorts and human populations that were not seen during training, providing more globally equitable health outcomes than current genetic studies that are predominantly European. Ultimately, rare variant PRSs can be combined with common variant PRSs into a unified risk model to significantly improve the identification of individuals from the general population who are at increased risk for common diseases.

Although the rare variant PRS models presented in this work show promise for accurate identification of high-risk individuals across diverse human populations, our study has several limitations. At present, rare variant PRS models have limited power; we are only capable of robustly estimating variant effects for well-powered genes, finding 217 GWAS loci but only 34 rare variant genes on average per trait. We empirically forecast that the exact causal genes underlying most of these GWAS loci will be uncovered by rare variant studies with larger cohort sizes and advances in variant interpretation algorithms (*60*). Second, while interpretation of variants of uncertain significance (VUS) remains a challenge, recent advances applying deep learning (*12*, *61*), high-throughput experimental assays (*62*), and variant information from closely related primate species (*63*) have each demonstrated promise towards solving variant interpretation on a genome-wide scale. Third, although we observed improved portability across ancestries for rare variant polygenic prediction, more accurate allele frequency resources for global populations will further shrink the discrepancies in performance across populations. Indeed, systematic efforts to catalog rare variation in non-European populations are ongoing (*64*, *65*) and will likely precede well-powered common variant GWAS studies in diverse

global populations (*66*). Finally, although we only evaluate rare coding or splice altering variants, improved non-coding variant prediction coupled with larger sample sizes would likely reveal the pervasive phenotypic impacts of rare penetrant variants in each person, with transformative implications for the utility of clinical whole-genome sequencing in the general population.

Methods summary

Datasets

We analyzed data from unrelated individuals in the UK Biobank, all of whom had genotypes obtained from microarrays and 454,712 of whom had genotypes available from exome-sequencing. The work described in this manuscript was approved by the UK Biobank under application no. 33751. In addition, we performed validation experiments with 20,708 individuals from the MGB Biobank.

Phenotype processing

Quantitative traits were standardized by inverse rank normal-transformation and adjusted for medication usage and further covariates including age, sex, ancestry, diet and others. Binary traits were adjusted for age, age^2, sex, age × sex, age^2 × sex and ancestry.

Common variant associations

GWAS were performed with common variants (AF > 1%) in individuals of European ancestry in the UK Biobank and causal gene sets were derived by linkage disequilibrium between independent GWAS significant variants ($P < 5\times10^{-8}$) and coding variants, splicing variants or eQTLs in nearby genes or by proximity with local transcription start sites.

Rare variant associations

Burden tests were performed with rare variants (AF < 0.1%) on individuals from all ethnicities by searching for combinations of allele frequencies and missense pathogenicity scores per gene and further calibrating via permutations to maximize significance prior to FDR correction. Significant gene-phenotype pairs were reported at 5% FDR after correction for multiple hypothesis testing across all autosomal protein coding genes in the human genome and across all tested traits. Rare variant results generated in this study were compared to results from a recent well-powered rare variant analysis in the UK Biobank (*17*) by examining the overlap of significant genes, along with the enrichment of GWAS and clinically relevant genes. Multiple missense classifiers were considered for pathogenicity prediction in the burden tests, including BayesDel (*67*), CADD (*68*), ClinPred (*69*), DEOGEN2 , EVE* (*61*), FATHMM-XF (*70*), M-CAP (*71*), MetaLR (*72*), MetaSVM (*72*), MutationAssessor (*73*), Polyphen-2 (*74*), PrimateAI-3D (*12*), PROVEAN (*75*), REVEL (*76*), SIFT (*77*), and VEST4 (*78*). Scores for the EVE-style variational autoencoder (EVE*) were generated by reimplementing the method. The different classifiers were compared via Spearman correlation with the average phenotype values of the carriers of each qualifying missense variant in high-confidence associated gene-phenotype pairs.

Polygenic risk scores

PRS models were constructed from GWAS and burden test results from training datasets. Common variant (AF >1%) PRS models were constructed by applying the method of clumping and thresholding (*31*). In contrast, rare variant PRS models were constructed by fitting linear models to each phenotype on the rare variants (AF < 0.1%) in significantly associated genes, weighted by predicted missense

pathogenicity. A unified PRS model was also constructed, which summed the rare and common variant PRS models per individual. As with the burden test results, rare variant PRS performance was evaluated using PrimateAI-3D and other classifiers across 78 traits. The overlap of individuals at phenotypic and PRS extremes was examined to further elucidate PRS performance. For two traits, HbA1c and LDL cholesterol, clinical risk prediction was assessed, since clinically diagnostic thresholds could distinguish cases from controls. PRS portability was assessed in two ways - firstly between cohorts, by applying models constructed in the UK Biobank to the MGB Biobank, and secondly between ancestries, by comparing the performance between different ancestry groups in the UK Biobank.

Forecasting analysis

Growth projections of rare and common variant associations, PRS performance, and overlap of significantly associated genes from rare and common variants were made from randomly down-sampled data ranging from 20% to 100% of the whole UK Biobank exome cohort with 20% increments.

Full materials and methods are available in the supplementary materials (*16*).

References and Notes

1. A. Buniello, J. A. L. MacArthur, M. Cerezo, L. W. Harris, J. Hayhurst, C. Malangone, A. McMahon, J. Morales, E. Mountjoy, E. Sollis, D. Suveges, O. Vrousgou, P. L. Whetzel, R. Amode, J. A. Guillen, H. S. Riat, S. J. Trevanion, P. Hall, H. Junkins, P. Flicek, T. Burdett, L. A. Hindorff, F. Cunningham, H. Parkinson, The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).

2. T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. C. Mackay, S. A. McCarroll, P. M. Visscher, Finding the missing heritability of complex diseases. *Nature*. **461**, 747–753 (2009).

3. Y. Ding, K. Hou, K. S. Burch, S. Lapinska, F. Privé, B. Vilhjálmsson, S. Sankararaman, B. Pasaniuc, Large uncertainty in individual polygenic risk score estimation impacts PRS-based risk stratification. *Nat Genet*. **54**, 30–39 (2022).

4. A. R. Martin, M. Kanai, Y. Kamatani, Y. Okada, B. M. Neale, M. J. Daly, Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics*. **51**, 584–591 (2019).

5. R. Henderson, M. O'Kane, V. McGilligan, S. Watterson, The genetics and screening of familial hypercholesterolaemia. *J. Biomed. Sci.* **23**, 1–12 (2016).

6. K. B. Kuchenbaecker, J. L. Hopper, D. R. Barnes, K. A. Phillips, T. M. Mooij, M. J. Roos-Blom, S. Jervis, F. E. van Leeuwen, R. L. Milne, N. Andrieu, D. E. Goldgar, M. B. Terry, M. A. Rookus, D. F. Easton, A. C. Antoniou, L. McGuffog, D. G. Evans, D. Barrowdale, D. Frost, J. Adlard, K. R. Ong, L. Izatt, M. Tischkowitz, R. Eeles, R. Davidson, S. Hodgson, S. Ellis, C. Nogues, C. Lasset, D. Stoppa-Lyonnet, J. P. Fricker, L. Faivre, P. Berthet, M. J. Hooning, L. E. van der Kolk, C. M.

Kets, M. A. Adank, E. M. John, W. K. Chung, I. L. Andrulis, M. Southey, M. B. Daly, S. S. Buys, A. Osorio, C. Engel, K. Kast, R. K. Schmutzler, T. Caldes, A. Jakubowska, J. Simard, M. L. Friedlander, S. A. McLachlan, E. Machackova, L. Foretova, Y. Y. Tan, C. F. Singer, E. Olah, A. M. Gerdes, B. Arver, H. Olsson, Risks of Breast, Ovarian, and Contralateral Breast Cancer for BRCA1 and BRCA2 Mutation Carriers. *JAMA*. **317** (2017), doi:10.1001/jama.2017.7112.

7. S. A. Cohen, C. C. Pritchard, G. P. Jarvik, Lynch Syndrome: From Screening to Diagnosis to Treatment in the Era of Modern Molecular Oncology (2019), doi:10.1146/annurev-genom-083118-015406.

8. A. R. Kim, J. C. Ulirsch, S. Wilmes, E. Unal, I. Moraga, M. Karakukcu, D. Yuan, S. Kazerounian, N. J. Abdulhay, D. S. King, N. Gupta, S. B. Gabriel, E. S. Lander, T. Patiroglu, A. Ozcan, M. A. Ozdemir, K. C. Garcia, J. Piehler, H. T. Gazda, D. E. Klein, V. G. Sankaran, Functional Selectivity in Cytokine Signaling Revealed Through a Pathogenic EPO Mutation. *Cell*. **168**, 1053-1064.e15 (2017).

9. C. I. Karen Lisa Smith, BRCA Mutation Testing in Determining Breast Cancer Therapy. *Cancer J.* **17**, 492 (2011).

10. M. Delvecchio, C. Pastore, P. Giordano, Treatment Options for MODY Patients: A Systematic Review of Literature. *Diabetes Ther.* **11** (2020), doi:10.1007/s13300-020-00864-4.

11. K. J. Karczewski, L. C. Francioli, G. Tiao, B. B. Cummings, J. Alföldi, Q. Wang, R. L. Collins, K. M. Laricchia, A. Ganna, D. P. Birnbaum, L. D. Gauthier, H. Brand, M. Solomonson, N. A. Watts, D. Rhodes, M. Singer-Berk, E. M. England, E. G. Seaby, J. A. Kosmicki, R. K. Walters, K. Tashman, Y. Farjoun, E. Banks, T. Poterba, A. Wang, C. Seed, N. Whiffin, J. X. Chong, K. E. Samocha, E. Pierce-Hoffman, Z. Zappala, A. H. O'Donnell-Luria, E. V. Minikel, B. Weisburd, M. Lek, J. S. Ware, C. Vittal, I. M. Armean, L. Bergelson, K. Cibulskis, K. M. Connolly, M. Covarrubias, S. Donnelly, S. Ferriera, S. Gabriel, J. Gentry, N. Gupta, T. Jeandet, D. Kaplan, C. Llanwarne, R. Munshi, S. Novod, N. Petrillo, D. Roazen, V. Ruano-Rubio, A. Saltzman, M. Schleicher, J. Soto, K. Tibbetts, C. Tolonen, G. Wade, M. E. Talkowski, Genome Aggregation Database Consortium, B. M. Neale, M. J. Daly, D. G. MacArthur, The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. **581**, 434–443 (2020).

12. H. Gao, T. Hamp, J. Ede, J. G. Schraiber, J. McRae, M. Singer-Berk, Y. Yang, A. Dietrich, P. Fiziev, L. Kuderna, L. Sundaram, Y. Wu, A. Adhikari, Y. Field, S. Chen, S. Batzoglou, F. Aguet, G. Lemire, R. Reimers, D. Balick, M. C. Janiak, M. Kuhlwilm, J. D. Orkin, S. Manu, A. Valenzuela, J. Bergman, M. Rouselle, F. E. Silva, L. Agueda, J. Blanc, M. Gut, D. de Vries, I. Goodhead, R. A. Harris, M. Raveendran, A. Jensen, I. S. Chuma, J. Horvath, C. Hvilsom, D. Juan, P. Frandsen, F. R. de Melo, F. Bertuol, H. Byrne, I. Sampaio, I. Farias, J. V. do Amaral, M. Messias, M. N. F. da Silva, M. Trivedi, R. Rossi, T. Hrbek, N. Andriaholinirina, C. J. Rabarivola, A. Zaramody, C. J. Jolly, J. Phillips-Conroy, G. Wilkerson, C. Abee, J. H. Simmons, E. Fernandez-Duque, S. Kanthaswamy, F. Shiferaw, D. Wu, L. Zhou, Y. Shao, G. Zhang, J. D. Keyyu, S. Knauf, M. D. Le, E. Lizano, S. Merker, A. Navarro, T. Batallion, T. Nadler, C. C. Khor, J. Lee, P. Tan, W. K. Lim, A. C. Kitchener, D. Zinner, I. Gut, A. Melin, K. Guschanski, M. H. Schierup, R. M. D. Beck, G. Umapathy, C. Roos, J. P. Boubli, M. Lek, S. Sunyaev, A. O'Donnell, H. Rehm, J. Xu, J. Rogers, T. Marques-Bonet, K. K.-H. Farh, The landscape of tolerated genetic variation in humans and primates. *Science*. **In press**.

13. C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O'Connell, A. Cortes, S. Welsh, A. Young, M. Effingham, G. McVean, S. Leslie, N.

Allen, P. Donnelly, J. Marchini, The UK Biobank resource with deep phenotyping and genomic data. *Nature*. **562**, 203–209 (2018).

14. D. Taliun, D. N. Harris, M. D. Kessler, J. Carlson, Z. A. Szpiech, R. Torres, S. A. G. Taliun, A. Corvelo, S. M. Gogarten, H. M. Kang, A. N. Pitsillides, J. LeFaive, S.-B. Lee, X. Tian, B. L. Browning, S. Das, A.-K. Emde, W. E. Clarke, D. P. Loesch, A. C. Shetty, T. W. Blackwell, A. V. Smith, Q. Wong, X. Liu, M. P. Conomos, D. M. Bobo, F. Aguet, C. Albert, A. Alonso, K. G. Ardlie, D. E. Arking, S. Aslibekyan, P. L. Auer, J. Barnard, R. G. Barr, L. Barwick, L. C. Becker, R. L. Beer, E. J. Benjamin, L. F. Bielak, J. Blangero, M. Boehnke, D. W. Bowden, J. A. Brody, E. G. Burchard, B. E. Cade, J. F. Casella, B. Chalazan, D. I. Chasman, Y.-D. I. Chen, M. H. Cho, S. H. Choi, M. K. Chung, C. B. Clish, A. Correa, J. E. Curran, B. Custer, D. Darbar, M. Daya, M. de Andrade, D. L. DeMeo, S. K. Dutcher, P. T. Ellinor, L. S. Emery, C. Eng, D. Fatkin, T. Fingerlin, L. Forer, M. Fornage, N. Franceschini, C. Fuchsberger, S. M. Fullerton, S. Germer, M. T. Gladwin, D. J. Gottlieb, X. Guo, M. E. Hall, J. He, N. L. Heard-Costa, S. R. Heckbert, M. R. Irvin, J. M. Johnsen, A. D. Johnson, R. Kaplan, S. L. R. Kardia, T. Kelly, S. Kelly, E. E. Kenny, D. P. Kiel, R. Klemmer, B. A. Konkle, C. Kooperberg, A. Köttgen, L. A. Lange, J. Lasky-Su, D. Levy, X. Lin, K.-H. Lin, C. Liu, R. J. F. Loos, L. Garman, R. Gerszten, S. A. Lubitz, K. L. Lunetta, A. C. Y. Mak, A. Manichaikul, A. K. Manning, R. A. Mathias, D. D. McManus, S. T. McGarvey, J. B. Meigs, D. A. Meyers, J. L. Mikulla, M. A. Minear, B. D. Mitchell, S. Mohanty, M. E. Montasser, C. Montgomery, A. C. Morrison, J. M. Murabito, A. Natale, P. Natarajan, S. C. Nelson, K. E. North, J. R. O'Connell, N. D. Palmer, N. Pankratz, G. M. Peloso, P. A. Peyser, J. Pleiness, W. S. Post, B. M. Psaty, D. C. Rao, S. Redline, A. P. Reiner, D. Roden, J. I. Rotter, I. Ruczinski, C. Sarnowski, S. Schoenherr, D. A. Schwartz, J.-S. Seo, S. Seshadri, V. A. Sheehan, W. H. Sheu, M. B. Shoemaker, N. L. Smith, J. A. Smith, N. Sotoodehnia, A. M. Stilp, W. Tang, K. D. Taylor, M. Telen, T. A. Thornton, R. P. Tracy, D. J. Van Den Berg, R. S. Vasan, K. A. Viaud-Martinez, S. Vrieze, D. E. Weeks, B. S. Weir, S. T. Weiss, L.-C. Weng, C. J. Willer, Y. Zhang, X. Zhao, D. K. Arnett, A. E. Ashley-Koch, K. C. Barnes, E. Boerwinkle, S. Gabriel, R. Gibbs, K. M. Rice, S. S. Rich, E. K. Silverman, P. Qasba, W. Gan, NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, G. J. Papanicolaou, D. A. Nickerson, S. R. Browning, M. C. Zody, S. Zöllner, J. G. Wilson, L. A. Cupples, C. C. Laurie, C. E. Jaquish, R. D. Hernandez, T. D. O'Connor, G. R. Abecasis, Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*. **590**, 290–299 (2021).

15. All of Us Research Program Investigators, J. C. Denny, J. L. Rutter, D. B. Goldstein, A. Philippakis, J. W. Smoller, G. Jenkins, E. Dishman, The "All of Us" Research Program. *N. Engl. J. Med.* **381**, 668–676 (2019).

16. See supplementary materials.

17. J. D. Backman, A. H. Li, A. Marcketta, D. Sun, J. Mbatchou, M. D. Kessler, C. Benner, D. Liu, A. E. Locke, S. Balasubramanian, A. Yadav, N. Banerjee, C. E. Gillies, A. Damask, S. Liu, X. Bai, A. Hawes, E. Maxwell, L. Gurski, K. Watanabe, J. A. Kosmicki, V. Rajagopal, J. Mighty, Regeneron Genetics Center, DiscovEHR, M. Jones, L. Mitnaul, E. Stahl, G. Coppola, E. Jorgenson, L. Habegger, W. J. Salerno, A. R. Shuldiner, L. A. Lotta, J. D. Overton, M. N. Cantor, J. G. Reid, G. Yancopoulos, H. M. Kang, J. Marchini, A. Baras, G. R. Abecasis, M. A. R. Ferreira, Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature*. **599**, 628–634 (2021).

18. J. Mbatchou, L. Barnard, J. Backman, A. Marcketta, J. A. Kosmicki, A. Ziyatdinov, C. Benner, C. O'Dushlaine, M. Barber, B. Boutkov, L. Habegger, M. Ferreira, A. Baras, J. Reid, G. Abecasis, E. Maxwell, J. Marchini, Computationally efficient whole-genome regression for quantitative and binary traits. *Nat Genet*. **53**, 1097–1103 (2021).

19. J. L. Goldstein, M. S. Brown, Binding and degradation of low density lipoproteins by cultured human fibroblasts. Comparison of cells from a normal subject and from a patient with homozygous familial hypercholesterolemia. *J Biol Chem*. **249**, 5153–5162 (1974).

20. M. S. Brown, J. L. Goldstein, Expression of the Familial Hypercholesterolemia Gene in Heterozygotes: Mechanism for a Dominant Disorder in Man. *Science*. **185**, 61–63 (1974).

21. M. S. Sabatine, PCSK9 inhibitors: clinical evidence and implementation. *Nat. Rev. Cardiol.* **16**, 155–165 (2019).

22. S. M. Grundy, N. J. Stone, A. L. Bailey, C. Beam, K. K. Birtcher, R. S. Blumenthal, L. T. Braun, S. de Ferranti, J. Faiella-Tommasino, D. E. Forman, R. Goldberg, P. A. Heidenreich, M. A. Hlatky, D. W. Jones, D. Lloyd-Jones, N. Lopez-Pajares, C. E. Ndumele, C. E. Orringer, C. A. Peralta, J. J. Saseen, S. C. Smith Jr, L. Sperling, S. S. Virani, J. Yeboah, 2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA Guideline on the Management of Blood Cholesterol: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Circulation*. **139**, e1082–e1143 (2019).

23. American Diabetes Association, 2. Classification and Diagnosis of Diabetes: *Diabetes Care*. **42**, S13–S28 (2019).

24. C. C. Cowie, S. S. Casagrande, L. S. Geiss, "Prevalence and Incidence of Type 2 Diabetes and Prediabetes" in *Diabetes in America. 3rd edition*, C. C. Cowie, S. S. Casagrande, A. Menke, M. A. Cissell, M. S. Eberhardt, J. B. Meig, E. W. Gregg, W. C. Knowler, E. Barrett-Connor, D. J. Becker, F. L. Brancati, E. J. Boyko, W. H. Herman, B. V. Howard, K. M. V. Narayan, M. Rewers, J. E. Fradkin, Eds. (2018), p. Chapter 3.

25. W. Fu, T. D. O'Connor, G. Jun, H. M. Kang, G. Abecasis, S. M. Leal, S. Gabriel, M. J. Rieder, D. Altshuler, J. Shendure, D. A. Nickerson, M. J. Bamshad, NHLBI Exome Sequencing Project, J. M. Akey, Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*. **493**, 216–220 (2013).

26. G. V. Kryukov, L. A. Pennacchio, S. R. Sunyaev, Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am. J. Hum. Genet.* **80**, 727–739 (2007).

27. J. Zeng, R. de Vlaming, Y. Wu, M. R. Robinson, L. R. Lloyd-Jones, L. Yengo, C. X. Yap, A. Xue, J. Sidorenko, A. F. McRae, J. E. Powell, G. W. Montgomery, A. Metspalu, T. Esko, G. Gibson, N. R. Wray, P. M. Visscher, J. Yang, Signatures of negative selection in the genetic architecture of human complex traits. *Nat. Genet.* **50**, 746–753 (2018).

28. A. P. Schoech, D. M. Jordan, P.-R. Loh, S. Gazal, L. J. O'Connor, D. J. Balick, P. F. Palamara, H. K. Finucane, S. R. Sunyaev, A. L. Price, Quantification of frequency-dependent genetic architectures in 25 UK Biobank traits reveals action of negative selection. *Nat. Commun.* **10**, 790 (2019).

29. K. Jaganathan, S. Kyriazopoulou Panagiotopoulou, J. F. McRae, S. F. Darbandi, D. Knowles, Y. I. Li, J. A. Kosmicki, J. Arbelaez, W. Cui, G. B. Schwartz, E. D. Chow, E. Kanterakis, H. Gao, A. Kia, S. Batzoglou, S. J. Sanders, K. K.-H. Farh, Predicting Splicing from Primary Sequence with Deep Learning. *Cell*. **176**, 535-548.e24 (2019).

30. A. V. Khera, M. Chaffin, K. G. Aragam, M. E. Haas, C. Roselli, S. H. Choi, P. Natarajan, E. S. Lander, S. A. Lubitz, P. T. Ellinor, S. Kathiresan, Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, 1 (2018).

31. International Schizophrenia Consortium, S. M. Purcell, N. R. Wray, J. L. Stone, P. M. Visscher, M. C. O'Donovan, P. F. Sullivan, P. Sklar, Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*. **460**, 748–752 (2009).

32. J. Luo, H. Yang, B.-L. Song, Mechanisms and regulation of cholesterol homeostasis. *Nat. Rev. Mol. Cell Biol.* **21**, 225–245 (2020).

33. K. E. Berge, H. Tian, G. A. Graf, L. Yu, N. V. Grishin, J. Schultz, P. Kwiterovich, B. Shan, R. Barnes, H. H. Hobbs, Accumulation of dietary cholesterol in sitosterolemia caused by mutations in adjacent ABC transporters. *Science*. **290**, 1771–1775 (2000).

34. J. D. Horton, J. C. Cohen, H. H. Hobbs, Molecular biology of PCSK9: its role in LDL metabolism. *Trends Biochem. Sci.* **32**, 71–77 (2007).

35. J. Behbodikhah, S. Ahmed, A. Elyasi, L. J. Kasselman, J. De Leon, A. D. Glass, A. B. Reiss, Apolipoprotein B and Cardiovascular Disease: Biomarker and Potential Therapeutic Target. *Metabolites*. **11** (2021), doi:10.3390/metabo11100690.

36. J. E. Nahon, M. Hoekstra, S. van Hulst, C. Manta, S. Goerdt, J. J. Geerling, C. Géraud, M. Van Eck, Hematopoietic Stabilin-1 deficiency does not influence atherosclerosis susceptibility in LDL receptor knockout mice. *Atherosclerosis*. **281**, 47–55 (2019).

37. J. J. P. Kastelein, H. N. Ginsberg, G. Langslet, G. K. Hovingh, R. Ceska, R. Dufour, D. Blom, F. Civeira, M. Krempf, C. Lorenzato, J. Zhao, R. Pordy, M. T. Baccara-Dinet, D. A. Gipe, M. J. Geiger, M. Farnier, ODYSSEY FH I and FH II: 78 week results with alirocumab treatment in 735 patients with heterozygous familial hypercholesterolaemia. *Eur. Heart J.* **36**, 2996–3003 (2015).

38. M. Van Heek, C. F. France, D. S. Compton, R. L. McLeod, N. P. Yumibe, K. B. Alton, E. J. Sybertz, H. R. Davis Jr, In vivo metabolism-based discovery of a potent cholesterol absorption inhibitor, SCH58235, in the rat and rhesus monkey through the identification of the active metabolites of SCH48461. *J. Pharmacol. Exp. Ther.* **283**, 157–163 (1997).

39. D. J. Weiner, A. Nadig, K. A. Jagadeesh, K. K. Dey, B. M. Neale, E. B. Robinson, K. J. Karczewski, L. J. O'Connor, Polygenic architecture of rare coding variation across 400,000 exomes (2022), p. 2022.07.06.22277335, , doi:10.1101/2022.07.06.22277335.

40. C. Marras, J. C. Beck, J. H. Bower, E. Roberts, B. Ritz, G. W. Ross, R. D. Abbott, R. Savica, S. K. Van Den Eeden, A. W. Willis, C. M. Tanner, Parkinson's Foundation P4 Group, Prevalence of Parkinson's disease across North America. *NPJ Parkinsons Dis*. **4**, 21 (2018).

41. M. T. Wallin, W. J. Culpepper, J. D. Campbell, L. M. Nelson, A. Langer-Gould, R. A. Marrie, G. R. Cutter, W. E. Kaye, L. Wagner, H. Tremlett, S. L. Buka, P. Dilokthornsakul, B. Topol, L. H. Chen, N. G. LaRocca, US Multiple Sclerosis Prevalence Workgroup, The prevalence of MS in the United States: A population-based estimate using health claims data. *Neurology*. **92**, e1029–e1040 (2019).

42. A. Gupta, Y. Wang, J. A. Spertus, M. Geda, N. Lorenze, C. Nkonde-Price, G. D'Onofrio, J. H. Lichtman, H. M. Krumholz, Trends in acute myocardial infarction in young patients and differences by sex and race, 2001 to 2010. *J Am Coll Cardiol*. **64**, 337–345 (2014).

43. J. M. Lawrence, J. Divers, S. Isom, S. Saydah, G. Imperatore, C. Pihoker, S. M. Marcovina, E. J. Mayer-Davis, R. F. Hamman, L. Dolan, D. Dabelea, D. J. Pettitt, A. D. Liese, SEARCH for Diabetes in Youth Study Group, Trends in Prevalence of Type 1 and Type 2 Diabetes in Children and Adolescents in the US, 2001-2017. *JAMA*. **326**, 717–727 (2021).

44. A. V. Khera, M. Chaffin, K. H. Wade, S. Zahid, J. Brancale, R. Xia, M. Distefano, O. Senol-Cosar, M. E. Haas, A. Bick, K. G. Aragam, E. S. Lander, G. D. Smith, H. Mason-Suares, M. Fornage, M. Lebo, N. J. Timpson, L. M. Kaplan, S. Kathiresan, Polygenic Prediction of Weight and Obesity Trajectories from Birth to Adulthood. *Cell*. **177**, 587-596.e9 (2019).

45. B. G. Nordestgaard, M. J. Chapman, S. E. Humphries, H. N. Ginsberg, L. Masana, O. S. Descamps, O. Wiklund, R. A. Hegele, F. J. Raal, J. C. Defesche, A. Wiegman, R. D. Santos, G. F. Watts, K. G. Parhofer, G. K. Hovingh, P. T. Kovanen, C. Boileau, M. Averna, J. Borén, E. Bruckert, A. L. Catapano, J. A. Kuivenhoven, P. Pajukanta, K. Ray, A. F. H. Stalenhoef, E. Stroes, M.-R. Taskinen, A. Tybjærg-Hansen, European Atherosclerosis Society Consensus Panel, Familial hypercholesterolaemia is underdiagnosed and undertreated in the general population: guidance for clinicians to prevent coronary heart disease: consensus statement of the European Atherosclerosis Society. *Eur. Heart J.* **34**, 3478–90a (2013).

46. G. Thanabalasingham, K. R. Owen, Diagnosis and management of maturity onset diabetes of the young (MODY). *BMJ*. **343**, d6044 (2011).

47. A. Markham, Evinacumab: First Approval. *Drugs*. **81**, 1101–1105 (2021).

48. D. Curtis, Polygenic risk score for schizophrenia is more strongly associated with ancestry than with schizophrenia. *Psychiatr. Genet.* **28**, 85–89 (2018).

49. C. Márquez-Luna, S. Gazal, P.-R. Loh, S. S. Kim, N. Furlotte, A. Auton, A. L. Price, 23andMe Research Team, LDpred-funct: incorporating functional priors improves polygenic prediction accuracy in UK Biobank and 23andMe data sets. *bioRxiv* (2018), , doi:10.1101/375337.

50. E. W. Karlson, N. T. Boutin, A. G. Hoffnagle, N. L. Allen, Building the Partners HealthCare Biobank at Partners Personalized Medicine: Informed Consent, Return of Research Results, Recruitment Lessons and Operational Considerations. *J Pers Med*. **6** (2016), doi:10.3390/jpm6010002.

51. E. M. Scott, A. Halees, Y. Itan, E. G. Spencer, Y. He, M. A. Azab, S. B. Gabriel, A. Belkadi, B. Boisson, L. Abel, A. G. Clark, Greater Middle East Variome Consortium, F. S. Alkuraya, J.-L. Casanova, J. G. Gleeson, Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery. *Nat Genet*. **48**, 1071–1076 (2016).

52. N. Shah, Y.-C. C. Hou, H.-C. Yu, R. Sainger, C. T. Caskey, J. C. Venter, A. Telenti, Identification of Misclassified ClinVar Variants via Disease Population Prevalence. *Am J Hum Genet*. **102**, 609–619 (2018).

53. M. Lek, K. J. Karczewski, E. V. Minikel, K. E. Samocha, E. Banks, T. Fennell, A. H. O'Donnell-Luria, J. S. Ware, A. J. Hill, B. B. Cummings, T. Tukiainen, D. P. Birnbaum, J. A. Kosmicki, L. E.

Duncan, K. Estrada, F. Zhao, J. Zou, E. Pierce-Hoffman, J. Berghout, D. N. Cooper, N. Deflaux, M. DePristo, R. Do, J. Flannick, M. Fromer, L. Gauthier, J. Goldstein, N. Gupta, D. Howrigan, A. Kiezun, M. I. Kurki, A. L. Moonshine, P. Natarajan, L. Orozco, G. M. Peloso, R. Poplin, M. A. Rivas, V. Ruano-Rubio, S. A. Rose, D. M. Ruderfer, K. Shakir, P. D. Stenson, C. Stevens, B. P. Thomas, G. Tiao, M. T. Tusie-Luna, B. Weisburd, H.-H. Won, D. Yu, D. M. Altshuler, D. Ardissino, M. Boehnke, J. Danesh, S. Donnelly, R. Elosua, J. C. Florez, S. B. Gabriel, G. Getz, S. J. Glatt, C. M. Hultman, S. Kathiresan, M. Laakso, S. McCarroll, M. I. McCarthy, D. McGovern, R. McPherson, B. M. Neale, A. Palotie, S. M. Purcell, D. Saleheen, J. M. Scharf, P. Sklar, P. F. Sullivan, J. Tuomilehto, M. T. Tsuang, H. C. Watkins, J. G. Wilson, M. J. Daly, D. G. MacArthur, Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. **536**, 285–291 (2016).

54. B. Kaufman, R. Shapira-Frommer, R. K. Schmutzler, M. William Audeh, M. Friedlander, J. Balmaña, G. Mitchell, G. Fried, K. Bowen, A. Fielding, S. M. Domchek, Olaparib monotherapy in patients with advanced cancer and a germ-line BRCA1/2 mutation: An open-label phase II study. *Journal of Clinical Oncology*. **31**, 11024–11024 (2013).

55. C. P. Cannon, M. A. Blazing, R. P. Giugliano, A. McCagg, J. A. White, P. Theroux, H. Darius, B. S. Lewis, T. O. Ophuis, J. W. Jukema, G. M. De Ferrari, W. Ruzyllo, P. De Lucca, K. Im, E. A. Bohula, C. Reist, S. D. Wiviott, A. M. Tershakovec, T. A. Musliner, E. Braunwald, R. M. Califf, IMPROVE-IT Investigators, Ezetimibe Added to Statin Therapy after Acute Coronary Syndromes. *N. Engl. J. Med.* **372**, 2387–2397 (2015).

56. J. P. Evans, B. C. Powell, J. S. Berg, Finding the Rare Pathogenic Variants in a Human Genome. *JAMA*. **317**, 1904–1905 (2017).

57. N. J. Schork, S. S. Murray, K. A. Frazer, E. J. Topol, Common vs. rare allele hypotheses for complex diseases. *Curr. Opin. Genet. Dev.* **19**, 212–219 (2009).

58. H. Shi, S. Gazal, M. Kanai, E. M. Koch, A. P. Schoech, K. M. Siewert, S. S. Kim, Y. Luo, T. Amariuta, H. Huang, Y. Okada, S. Raychaudhuri, S. R. Sunyaev, A. L. Price, Population-specific causal disease effect sizes in functionally important regions impacted by selection. *Nat Commun*. **12**, 1098 (2021).

59. S. L. Spain, J. C. Barrett, Strategies for fine-mapping complex traits. *Hum. Mol. Genet.* **24**, R111-9 (2015).

60. P. M. Visscher, N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, M. A. Brown, J. Yang, 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).

61. J. Frazer, P. Notin, M. Dias, A. Gomez, J. K. Min, K. Brock, Y. Gal, D. S. Marks, Disease variant prediction with deep generative models of evolutionary data. *Nature*. **599**, 91–95 (2021).

62. G. M. Findlay, R. M. Daza, B. Martin, M. D. Zhang, A. P. Leith, M. Gasperini, J. D. Janizek, X. Huang, L. M. Starita, J. Shendure, Accurate classification of BRCA1 variants with saturation genome editing. *Nature*. **562**, 217–222 (2018).

63. L. Sundaram, H. Gao, S. R. Padigepati, J. F. McRae, Y. Li, J. A. Kosmicki, N. Fritzilas, J. Hakenberg, A. Dutta, J. Shon, J. Xu, S. Batzoglou, X. Li, K. K.-H. Farh, Predicting the clinical impact of human mutation with deep neural networks. *Nat. Genet.* **50**, 1161–1170 (2018).

64. H3Africa Consortium, C. Rotimi, A. Abayomi, A. Abimiku, V. M. Adabayeri, C. Adebamowo, E. Adebiyi, A. D. Ademola, A. Adeyemo, D. Adu, D. Affolabi, G. Agongo, S. Ajayi, S. Akarolo-Anthony, R. Akinyemi, A. Akpalu, M. Alberts, O. Alonso Betancourt, A. M. Alzohairy, G. Ameni, O. Amodu, G. Anabwani, K. Andersen, F. Arogundade, O. Arulogun, D. Asogun, R. Bakare, M. L. Baniecki, C. Beiswanger, A. Benkahla, L. Bethke, M. Boehnke, V. Boima, J. Brandful, A. I. Brooks, F. C. Brosius, C. Brown, B. Bucheton, D. T. Burke, B. G. Burnett, S. Carrington-Lawrence, N. Carstens, J. Chisi, A. Christoffels, R. Cooper, H. Cordell, N. Crowther, T. Croxton, J. de Vries, L. Derr, P. Donkor, S. Doumbia, A. Duncanson, I. Ekem, A. El Sayed, M. E. Engel, J. C. K. Enyaru, D. Everett, F. M. Fadlelmola, E. Fakunle, K. H. Fischbeck, A. Fischer, O. Folarin, J. Gamieldien, R. F. Garry, S. Gaseitsiwe, R. Gbadegesin, A. Ghansah, M. Giovanni, P. Goesbeck, F. X. Gomez-Olive, D. S. Grant, R. Grewal, M. Guyer, N. A. Hanchard, C. T. Happi, S. Hazelhurst, B. J. Hennig, C. Hertz-Fowler, W. Hide, F. Hilderbrandt, C. Hugo-Hamman, M. E. Ibrahim, R. James, Y. Jaufeerally-Fakim, C. Jenkins, U. Jentsch, P.-P. Jiang, M. Joloba, V. Jongeneel, F. Joubert, M. Kader, K. Kahn, P. Kaleebu, S. H. Kapiga, S. K. Kassim, I. Kasvosve, J. Kayondo, B. Keavney, A. Kekitiinwa, S. H. Khan, P. Kimmel, M.-C. King, R. Kleta, M. Koffi, J. Kopp, M. Kretzler, J. Kumuthini, S. Kyobe, C. Kyobutungi, D. T. Lackland, K. A. Lacourciere, G. Landouré, R. Lawlor, T. Lehner, M. Lesosky, N. Levitt, K. Littler, Z. Lombard, J. F. Loring, S. Lyantagaye, A. Macleod, E. B. Madden, C. R. Mahomva, J. Makani, M. Mamven, M. Marape, G. Mardon, P. Marshall, D. P. Martin, D. Masiga, R. Mason, M. Mate-Kole, E. Matovu, M. Mayige, B. M. Mayosi, J. C. Mbanya, S. A. McCurdy, M. I. McCarthy, H. McIlleron, S. O. Mc'Ligeyo, C. Merle, A. O. Mocumbi, C. Mondo, J. V. Moran, A. Motala, M. Moxey-Mims, W. S. Mpoloka, C. L. Msefula, T. Mthiyane, N. Mulder, G. her Mulugeta, D. Mumba, J. Musuku, M. Nagdee, O. Nash, D. Ndiaye, A. Q. Nguyen, M. Nicol, O. Nkomazana, S. Norris, B. Nsangi, A. Nyarko, M. Nyirenda, E. Obe, R. Obiakor, A. Oduro, S. F. Ofori-Acquah, O. Ogah, S. Ogendo, K. Ohene-Frempong, A. Ojo, T. Olanrewaju, J. Oli, C. Osafo, O. Ouwe Missi Oukem-Boyer, B. Ovbiagele, A. Owen, M. O. Owolabi, L. Owolabi, E. Owusu-Dabo, G. Pare, R. Parekh, H. G. Patterton, M. B. Penno, J. Peterson, R. Pieper, J. Plange-Rhule, M. Pollak, J. Puzak, R. S. Ramesar, M. Ramsay, R. Rasooly, S. Reddy, P. C. Sabeti, K. Sagoe, T. Salako, O. Samassékou, M. S. Sandhu, O. Sankoh, F. S. Sarfo, M. Sarr, G. Shaboodien, I. Sidibe, G. Simo, M. Simuunza, L. Smeeth, E. Sobngwi, H. Soodyall, H. Sorgho, O. Sow Bah, S. Srinivasan, D. J. Stein, E. S. Susser, C. Swanepoel, G. Tangwa, A. Tareila, O. Tastan Bishop, B. Tayo, N. Tiffin, H. Tinto, E. Tobin, S. M. Tollman, M. Traoré, M. J. Treadwell, J. Troyer, M. Tsimako-Johnstone, V. Tukei, I. Ulasi, N. Ulenga, B. van Rooyen, A. P. Wachinou, S. P. Waddy, A. Wade, M. Wayengera, J. Whitworth, L. Wideroff, C. A. Winkler, S. Winnicki, A. Wonkam, M. Yewondwos, T. sen, N. Yozwiak, H. Zar, Enabling the genomic revolution in Africa. *Science*. **344**, 1346–1348 (2014).

65. GenomeAsia100K Consortium, J. D. Wall, E. W. Stawiski, A. Ratan, H. L. Kim, C. Kim, R. Gupta, K. Suryamohan, E. S. Gusareva, R. W. Purbojati, T. Bhangale, V. Stepanov, V. Kharkov, M. S. Schröder, V. Ramprasad, J. Tom, S. Durinck, Q. Bei, J. Li, J. Guillory, S. Phalke, A. Basu, J. Stinson, S. Nair, S. Malaichamy, N. K. Biswas, J. C. Chambers, K. C. Cheng, J. T. George, S. S. Khor, J.-I. Kim, B. Cho, R. Menon, T. Sattibabu, A. Bassi, M. Deshmukh, A. Verma, V. Gopalan, J.-Y. Shin, M. Pratapneni, S. Santhosh, K. Tokunaga, B. M. Md-Zain, K. G. Chan, M. Parani, P. Natarajan, M. Hauser, R. R. Allingham, C. Santiago-Turla, A. Ghosh, S. G. K. Gadde, C. Fuchsberger, L. Forer, S. Schoenherr, H. Sudoyo, J. S. Lansing, J. Friedlaender, G. Koki, M. P. Cox, M. Hammer, T. Karafet, K. C. Ang, S. Q. Mehdi, V. Radha, V. Mohan, P. P. Majumder, S. Seshagiri, J.-S. Seo, S. C. Schuster, A. S. Peterson, The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature*. **576**, 106–111 (2019).

66. M. C. Mills, C. Rahal, The GWAS Diversity Monitor tracks diversity by disease in real time. *Nat Genet*. **52**, 242–243 (2020).

67. B.-J. Feng, PERCH: A Unified Framework for Disease Gene Prioritization: HUMAN MUTATION. *Human Mutation*. **38**, 243–251 (2017).

68. P. Rentzsch, M. Schubach, J. Shendure, M. Kircher, CADD-Splice—improving genome-wide variant effect prediction using deep learning-derived splice scores. *Genome Med*. **13**, 31 (2021).

69. N. Alirezaie, K. D. Kernohan, T. Hartley, J. Majewski, T. D. Hocking, ClinPred: Prediction Tool to Identify Disease-Relevant Nonsynonymous Single-Nucleotide Variants. *The American Journal of Human Genetics*. **103**, 474–483 (2018).

70. M. F. Rogers, H. A. Shihab, M. Mort, D. N. Cooper, T. R. Gaunt, C. Campbell, FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. *Bioinformatics*. **34**, 511–513 (2018).

71. K. A. Jagadeesh, A. M. Wenger, M. J. Berger, H. Guturu, P. D. Stenson, D. N. Cooper, J. A. Bernstein, G. Bejerano, M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat Genet*. **48**, 1581–1586 (2016).

72. C. Dong, P. Wei, X. Jian, R. Gibbs, E. Boerwinkle, K. Wang, X. Liu, Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum. Mol. Genet.* **24**, 2125–2137 (2015).

73. B. Reva, Y. Antipin, C. Sander, Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Research*. **39**, e118–e118 (2011).

74. I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, S. R. Sunyaev, A method and server for predicting damaging missense mutations. *Nat Methods*. **7**, 248–249 (2010).

75. Y. Choi, G. E. Sims, S. Murphy, J. R. Miller, A. P. Chan, Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLoS ONE*. **7**, e46688 (2012).

76. N. M. Ioannidis, J. H. Rothstein, V. Pejaver, S. Middha, S. K. McDonnell, S. Baheti, A. Musolf, Q. Li, E. Holzinger, D. Karyadi, L. A. Cannon-Albright, C. C. Teerlink, J. L. Stanford, W. B. Isaacs, J. Xu, K. A. Cooney, E. M. Lange, J. Schleutker, J. D. Carpten, I. J. Powell, O. Cussenot, G. Cancel-Tassin, G. G. Giles, R. J. MacInnis, C. Maier, C.-L. Hsieh, F. Wiklund, W. J. Catalona, W. D. Foulkes, D. Mandal, R. A. Eeles, Z. Kote-Jarai, C. D. Bustamante, D. J. Schaid, T. Hastie, E. A. Ostrander, J. E. Bailey-Wilson, P. Radivojac, S. N. Thibodeau, A. S. Whittemore, W. Sieh, REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *The American Journal of Human Genetics*. **99**, 877–885 (2016).

77. N.-L. Sim, P. Kumar, J. Hu, S. Henikoff, G. Schneider, P. C. Ng, SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Research*. **40**, W452–W457 (2012).

78. H. Carter, C. Douville, P. D. Stenson, D. N. Cooper, R. Karchin, Identifying Mendelian disease genes with the Variant Effect Scoring Tool. *BMC Genomics*. **14**, S3 (2013).

79. C. Churchhouse, Rapid GWAS of thousands of phenotypes for 337,000 samples in the UK Biobank (2017), (available at http://www.nealelab.is/blog/2017/7/19/rapid-gwas-of-thousands-of-phenotypes-for-337000-samples-in-the-uk-biobank).

80. Y. Wu, E. M. Byrne, Z. Zheng, K. E. Kemper, L. Yengo, A. J. Mallett, J. Yang, P. M. Visscher, N. R. Wray, Genome-wide association study of medication-use and associated disease in the UK Biobank. *Nat. Commun.* **10**, 1891 (2019).

81. A. G. Barnett, Regression to the mean: what it is and how to deal with it. *International Journal of Epidemiology*. **34**, 215–220 (2004).

82. C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, J. J. Lee, Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaSci*. **4**, 7 (2015).

83. K. J. Galinsky, G. Bhatia, P.-R. Loh, S. Georgiev, S. Mukherjee, N. J. Patterson, A. L. Price, Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1B in Europe and East Asia. *The American Journal of Human Genetics*. **98**, 456–472 (2016).

84. A. Buniello, J. A. L. MacArthur, M. Cerezo, L. W. Harris, J. Hayhurst, C. Malangone, A. McMahon, J. Morales, E. Mountjoy, E. Sollis, D. Suveges, O. Vrousgou, P. L. Whetzel, R. Amode, J. A. Guillen, H. S. Riat, S. J. Trevanion, P. Hall, H. Junkins, P. Flicek, T. Burdett, L. A. Hindorff, F. Cunningham, H. Parkinson, The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).

85. F. Hormozdiari, E. Kostem, E. Y. Kang, B. Pasaniuc, E. Eskin, Identifying causal variants at loci with multiple signals of association. *Genetics*. **198**, 497–508 (2014).

86. GTEx Consortium, Genetic effects on gene expression across human tissues. *Nature*. **550**, 204–213 (2017).

87. X. Liu, X. Jian, E. Boerwinkle, dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions. *Human Mutation*. **32**, 894–899 (2011).

88. X. Liu, C. Li, C. Mou, Y. Dong, Y. Tu, dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Med.* **12**, 103 (2020).

Acknowledgments: We would like to thank Daniel MacArthur, Jonathan Pritchard, Manuel Rivas, Nicole Ersaro, and Ileena Mitra for helpful discussions, and the participants and investigators in the UK Biobank (Resource Application Number 33751) and MGB studies (protocol 2018P001236) who made this work possible.

Funding: TMB is supported by funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 864203), PID2021-126004NB-100 (MINECO/FEDER, UE), "Unidad de Excelencia María de Maeztu", funded by the AEI (CEX2018-000792-M), NIH 1R01HG010898-01A1 and Secretaria d'Universitats i Recerca and CERCA Programme del Departament d'Economia i Coneixement de la Generalitat de Catalunya (GRC 2021 SGR 00177). HR receives funding from Illumina, Inc. to support rare disease gene discovery and diagnosis.

Author contributions: PF, JM, JU, JD, TH, YY, PW, ZN, JS, HG, AM, DC, FA, and KF performed the analysis and wrote the manuscript. JR, TMB, MR, HR, AOL, AK, and KF supervised the work.

Competing interests: Employees of Illumina, Inc. are indicated in the list of author affiliations. Patents related to this work are (1) Covariate correction including drug use from temporal data, filing No.: 63/351317; (2) Optimized burden test based on nested t-tests that maximize separation between carriers and non-carriers, filing No.: 63/351283; (3) Rare variant polygenic risk scores, filing No.: 63/351299.

Data and materials availability: PrimateAI-3D prediction scores are available with a non-commercial license upon request and are displayed at https://primad.basespace.illumina.com. Source code is available open source under an academic license upon request.

Supplementary Materials

Materials and Methods

Figs. S1 to S16

Supplemental Tables S1 to S12

References (79–*88*)

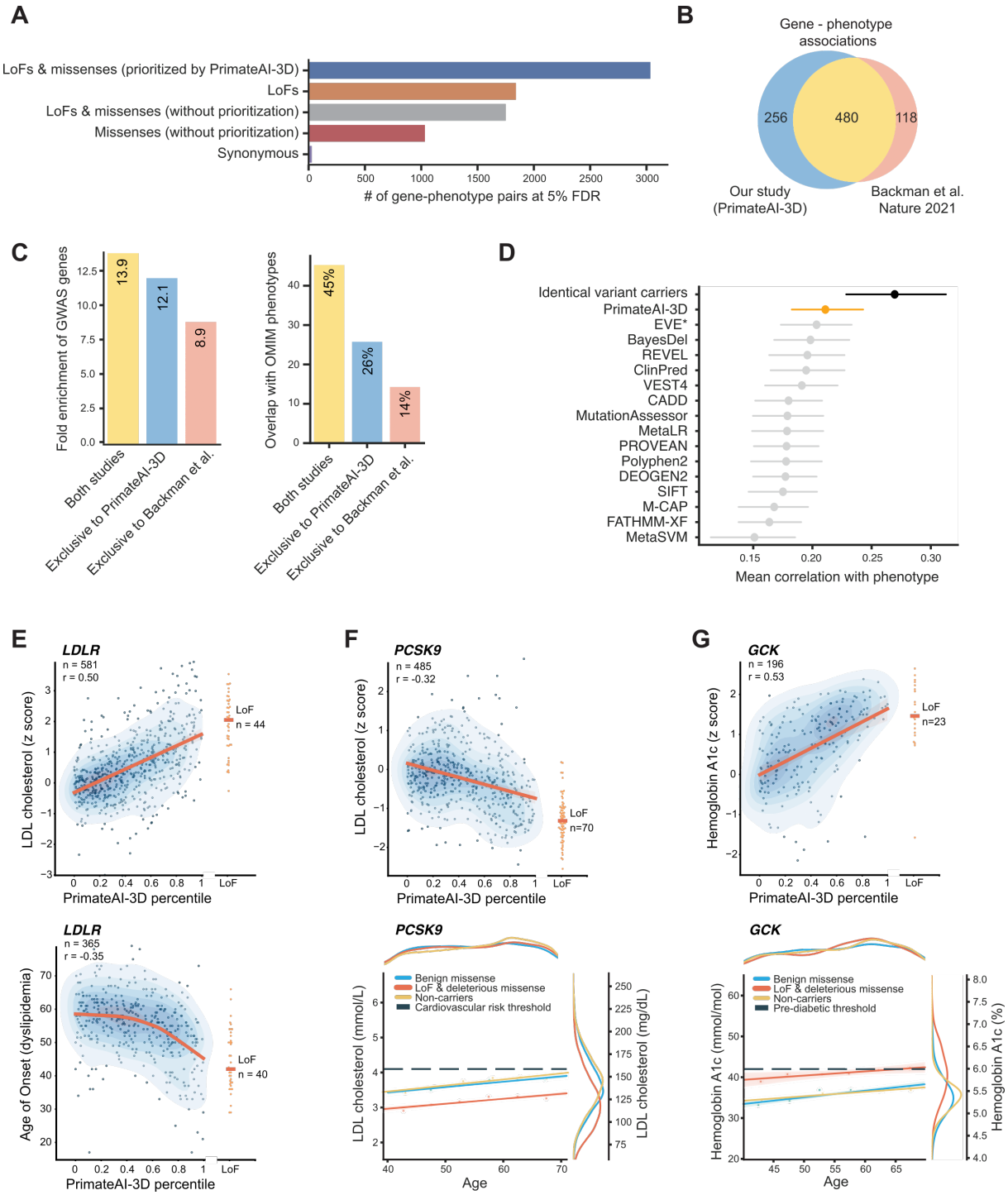
A
LoFs & missenses (prioritized by PrimateAI-3D)
LoFs
LoFs & missenses (without prioritization)
Missenses (without prioritization)
Synonymous
0
500
1000
1500
2000
2500
3000
# of gene-phenotype pairs at 5% FDR
B
Gene - phenotype associations
256
480
118
Our study (PrimateAI-3D)
Backman et al. Nature 2021
C
Fold enrichment of GWAS genes
13.9
12.1
8.9
Both studies
Exclusive to PrimateAI-3D
Exclusive to Backman et al.
Overlap with OMIM phenotypes
45%
26%
14%
D
Identical variant carriers
PrimateAI-3D
EVE*
BayesDel
REVEL
ClinPred
VEST4
CADD
MutationAssessor
MetaLR
PROVEAN
Polyphen2
DEOGEN2
SIFT
M-CAP
FATHMM-XF
MetaSVM
Mean correlation with phenotype
E
LDLR
n = 581
r = 0.50
LDL cholesterol (z score)
LoF
n = 44
PrimateAI-3D percentile
LDLR
n = 365
r = -0.35
Age of Onset (dyslipidemia)
LoF
n = 40
PrimateAI-3D percentile
F
PCSK9
n = 485
r = -0.32
LDL cholesterol (z score)
LoF
n=70
PrimateAI-3D percentile
PCSK9
Benign missense
LoF & deleterious missense
Non-carriers
Cardiovascular risk threshold
LDL cholesterol (mmol/L)
LDL cholesterol (mg/dL)
Age
G
GCK
n = 196
r = 0.53
Hemoglobin A1c (z score)
LoF
n=23
PrimateAI-3D percentile
GCK
Benign missense
LoF & deleterious missense
Non-carriers
Pre-diabetic threshold
Hemoglobin A1c (mmol/mol)
Hemoglobin A1c (%)
Age

Fig. 1. PrimateAI-3D identifies rare deleterious variants that affect disease severity and age of onset. (A) Total number of significant gene-phenotype associations (FDR < 5%) identified across 90 phenotypes for rare variant burden tests using different inclusion criteria for variants. As a negative control, the number of significant genotype-phenotype associations for a burden test using only synonymous variants is also shown. (**B**) Comparison of the current study with a recent study of rare variants in the UK Biobank (*17*) on the number of gene-phenotype associations detected exclusively by one or both studies for the same traits and matched significance thresholds. **(C)** Comparison of rare variant genes discovered in this study versus the previous study (*17*) using orthogonal genetic evidence. (**left**) Fold enrichment of rare variant genes at common variant GWAS loci, matched for the same phenotypes. (**right**) Percentage of rare variant genes overlapping with OMIM genes matched for related phenotypes. (**D**) Performance of different variant pathogenicity classifiers (see Methods) at predicting variant effects on quantitative phenotypes. Spearman correlations between pathogenicity scores and phenotype values on a set of 62 gene-phenotype pairs are shown. The phenotypic correlation between individuals carrying an identical missense variant is shown in black as an upper bound for classifier performance. Dots and error bars represent mean ± 95% confidence interval. **(E)** (**top**) Positive correlation of LDL cholesterol concentrations (y-axis) with PrimateAI-3D scores (x-axis) for rare missense variants in *LDLR*. (**bottom)** PrimateAI-3D score is predictive of age of onset for dyslipidemia in carriers of rare missense variants in *LDLR*. (**F**) (**top**) Negative correlation of LDL cholesterol concentrations with PrimateAI-3D scores for rare missense variants in *PCSK9*, a down-regulator of *LDLR*. (**bottom**) LDL cholesterol concentrations increase with age at a similar rate regardless of carrier status, but carriers of prioritized rare variants have lower LDL concentrations across all ages. (**G**) (**top**) Positive correlation of HbA1c concentrations with PrimateAI-3D scores for rare missense variants in *GCK*. (**bottom**) HbA1c concentrations increase with age at a similar rate regardless of carrier status, but carriers of rare deleterious variants reach pre-diabetic thresholds earlier in their lives on average. Deleterious and benign missense variants are defined as variants with PrimateAI-3D score greater than 0.5 and less than 0.5, respectively. For **E**, **F** and **G** red, blue or yellow lines show regression models fitted to the data.

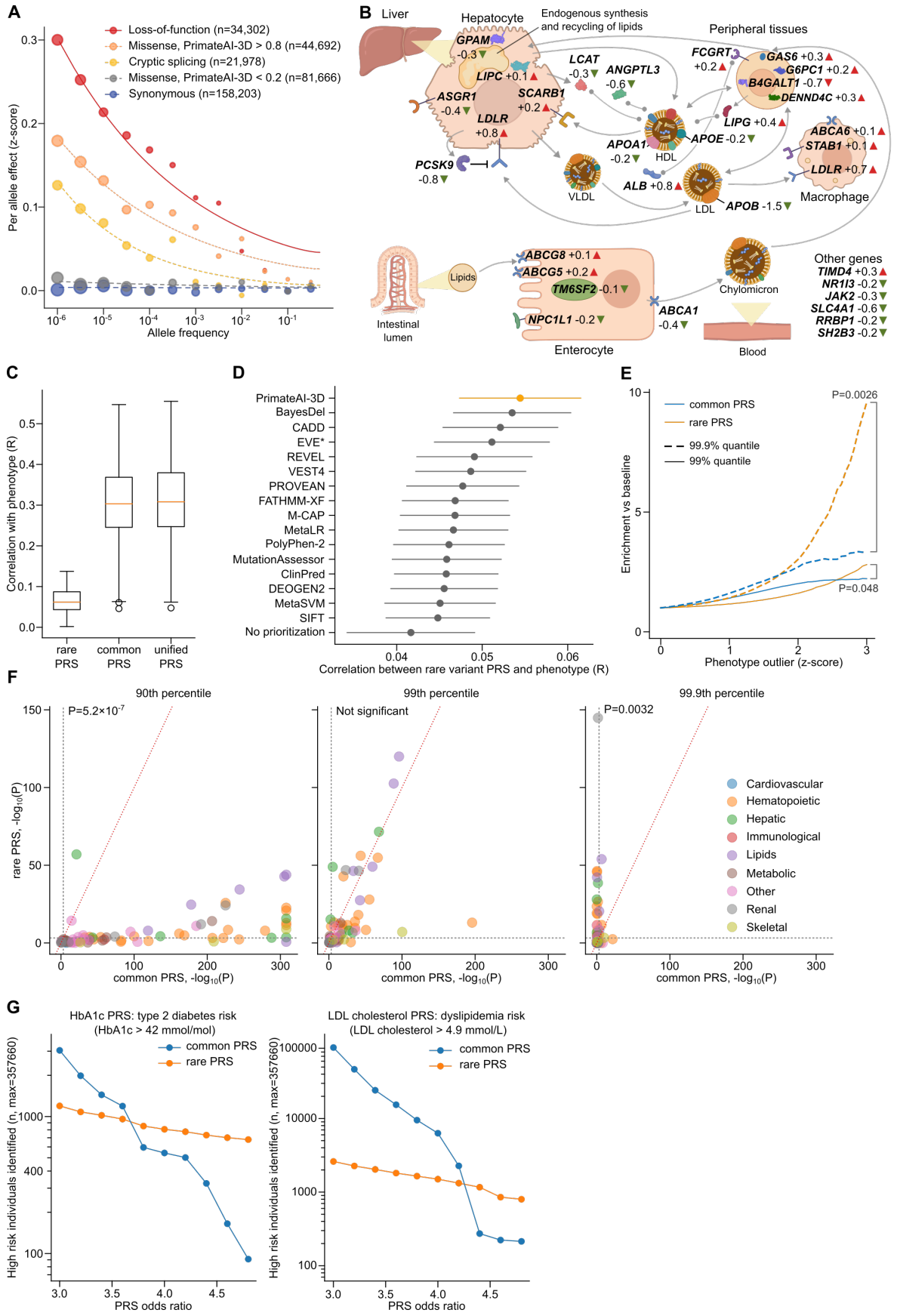
A
Loss-of-function (n=34,302)
Missense, PrimateAI-3D > 0.8 (n=44,692)
Cryptic splicing (n=21,978)
Missense, PrimateAI-3D < 0.2 (n=81,666)
Synonymous (n=158,203)
Per allele effect (z-score)
Allele frequency
B
Liver
Hepatocyte
Endogenous synthesis and recycling of lipids
Peripheral tissues
GPAM -0.3
LIPC +0.1
SCARB1 +0.2
ASGR1 -0.4
LDLR +0.8
PCSK9 -0.8
LCAT -0.3
ANGPTL3 -0.6
FCGRT +0.2
GAS6 +0.3
G6PC1 +0.2
B4GALT1 -0.7
DENND4C +0.3
LIPG +0.4
ABCA6 +0.1
STAB1 +0.1
LDLR +0.7
Macrophage
APOA1 -0.2
HDL
APOE -0.2
VLDL
ALB +0.8
LDL
APOB -1.5
Lipids
Intestinal lumen
ABCG8 +0.1
ABCG5 +0.2
TM6SF2 -0.1
NPC1L1 -0.2
Enterocyte
ABCA1 -0.4
Chylomicron
Blood
Other genes
TIMD4 +0.3
NR1I3 -0.2
JAK2 -0.3
SLC4A1 -0.6
RRBP1 -0.2
SH2B3 -0.2
C
Correlation with phenotype (R)
rare PRS
common PRS
unified PRS
D
PrimateAI-3D
BayesDel
CADD
EVE*
REVEL
VEST4
PROVEAN
FATHMM-XF
M-CAP
MetaLR
PolyPhen-2
MutationAssessor
ClinPred
DEOGEN2
MetaSVM
SIFT
No prioritization
Correlation between rare variant PRS and phenotype (R)
E
common PRS
rare PRS
99.9% quantile
99% quantile
P=0.0026
P=0.048
Enrichment vs baseline
Phenotype outlier (z-score)
F
90th percentile
99th percentile
99.9th percentile
P=5.2×10-7
Not significant
P=0.0032
rare PRS, -log10(P)
common PRS, -log10(P)
Cardiovascular
Hematopoietic
Hepatic
Immunological
Lipids
Metabolic
Other
Renal
Skeletal
G
HbA1c PRS: type 2 diabetes risk
(HbA1c > 42 mmol/mol)
LDL cholesterol PRS: dyslipidemia risk
(LDL cholesterol > 4.9 mmol/L)
common PRS
rare PRS
High risk individuals identified (n, max=357660)
PRS odds ratio

Fig. 2. Comparison of polygenic risk scores (PRSs) from common and rare variants. (**A**) Relationship between variant effect size and allele frequency for different pathogenicity classes of variants. Synonymous variants are shown as negative controls. Dot sizes are proportional to the cube root of the number of variants in each group. Regression fits between the allelic effect size and minor allele frequency are shown by curves for each pathogenicity class, calculated using the equation $\beta = \sigma[2p(1-p)]^{-a/2}$ where β is the per allele effect, p is the minor allele frequency and σ and a are parameters for selective constraint. (**B**) Illustration of the cholesterol pathway. Genes in the rare variant PRS model are superimposed. For each gene, values indicate effect sizes in standardized units (see Methods), and triangles indicate direction of effect. **(C)** Comparison of the performance of rare variant PRS, common variant PRS, and a unified PRS across 78 phenotypes in the withheld UK Biobank test set. Pearson correlations between PRS predictions and phenotypes are shown. (**D**) Comparison of rare variant PRSs constructed using different pathogenicity classifiers (see Methods). Mean absolute Pearson correlations between PRS and phenotypes are shown. Dots and error bars represent mean ± 95% confidence intervals. **(E)** Enrichment of outlier PRS scores in individuals who are phenotype outliers. Phenotype outlier individuals were defined as exceeding a certain z-score cutoff (x-axis), and the y-axis shows the enrichment of outlier PRS scores in phenotype outlier individuals versus the baseline population, aggregated across 78 phenotypes. **(F)** Comparison of the performance of common variant PRS (x-axis) versus rare variant PRS (y-axis) at identifying individuals at the 90th, 99th, and 99.9^{th} percentiles (left, middle, and right panels) for 78 quantitative phenotypes. Dashed horizontal and vertical lines show Bonferroni corrected significance thresholds. Lines of equivalence are shown by dashed diagonal red lines. (**G**) Number of individuals at high clinical risk for type 2 diabetes (**left**) and dyslipidemia (**right**), identified by rare and common variant PRSs at varying risk thresholds (x-axis). Rare variant PRSs identified more individuals at higher risk (>3.8 higher odds for type 2 diabetes, and >4.4 higher odds for dyslipidemia) than common variant PRSs.

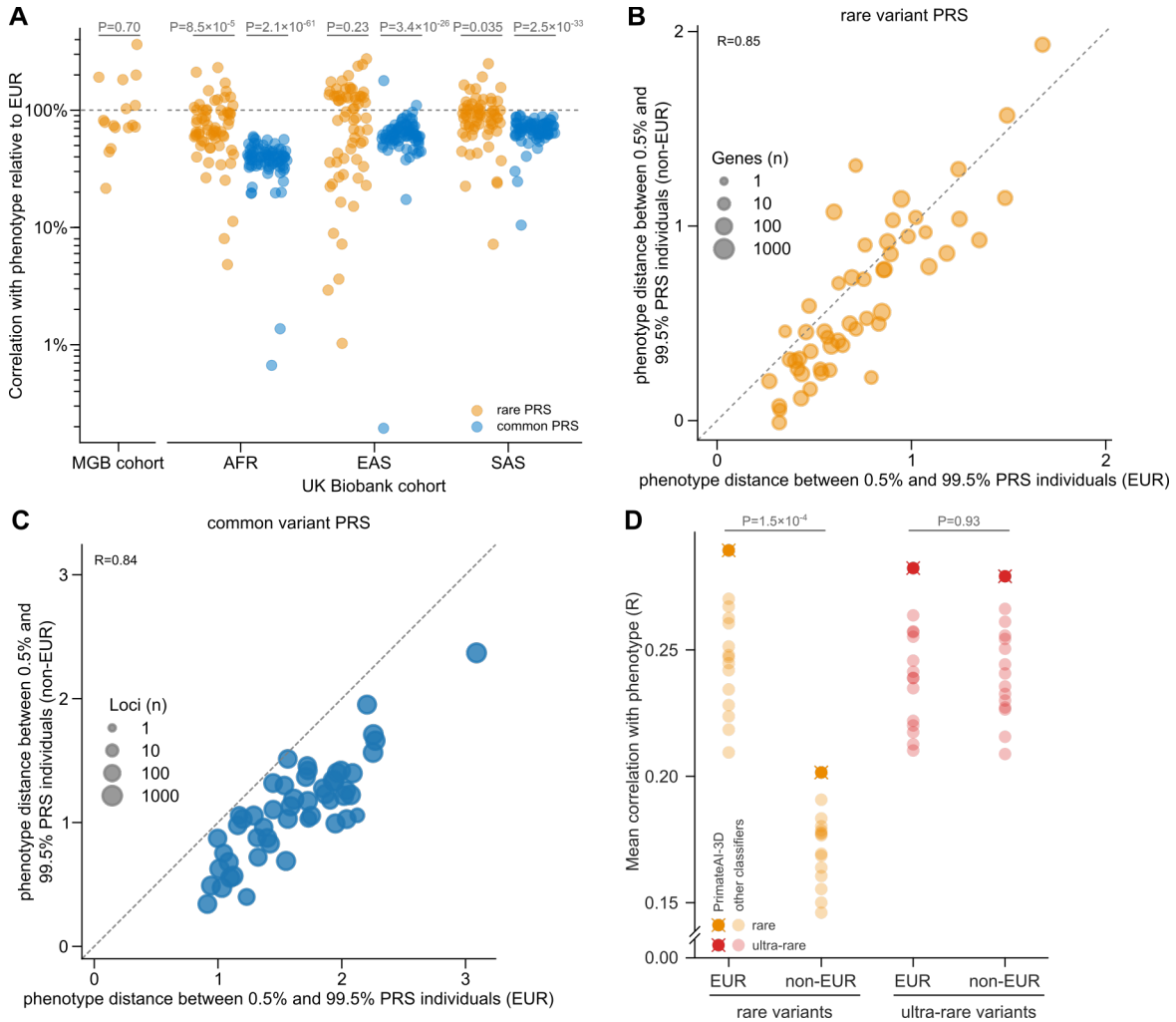

Fig. 3. Validation of rare variant PRS performance in diverse human populations. (**A**) Performance of rare and common variant PRSs derived from UK Biobank Europeans (EUR), measured in the MGB cohort (left) and in UK Biobank non-Europeans (non-EUR) stratified by ancestry (right, AFR: African, EAS: East Asian, SAS: South Asian). Performance is shown relative to held out European individuals in the UK Biobank. P-values indicate whether the difference in performance versus held out Europeans is significant. (**B**) Mean phenotype distance between UK Biobank EUR (x-axis) and UK Biobank non-EUR (y-axis) individuals is shown for 52 matching traits. The phenotypic distance is calculated by comparing individuals with low (<0.5%) and high (>99.5%) rare variant PRS percentiles. The Pearson correlation is reported. A line of equivalence is shown by the gray diagonal dashed line. (**C**) Same as (B) but showing the results for common variant PRSs. (**D**) Performance of PrimateAI-3D variant effect predictions stratified by ancestry and allele frequency for 49 gene-phenotype pairs. Correlation of predicted variant effects with observed phenotypes is shown on the y-axis. Rare variants have AF < 0.1% in each population. Ultra-rare variants are absent from TOPMed, and non-EUR ultra-rare variants are singletons (AC=1) whereas EUR ultra-rare variants have allele frequencies less than or equal to those of the non-EUR singletons. P-values are displayed for comparisons across ancestries using PrimateAI-3D. The performance of other variant classifiers is also shown for context.

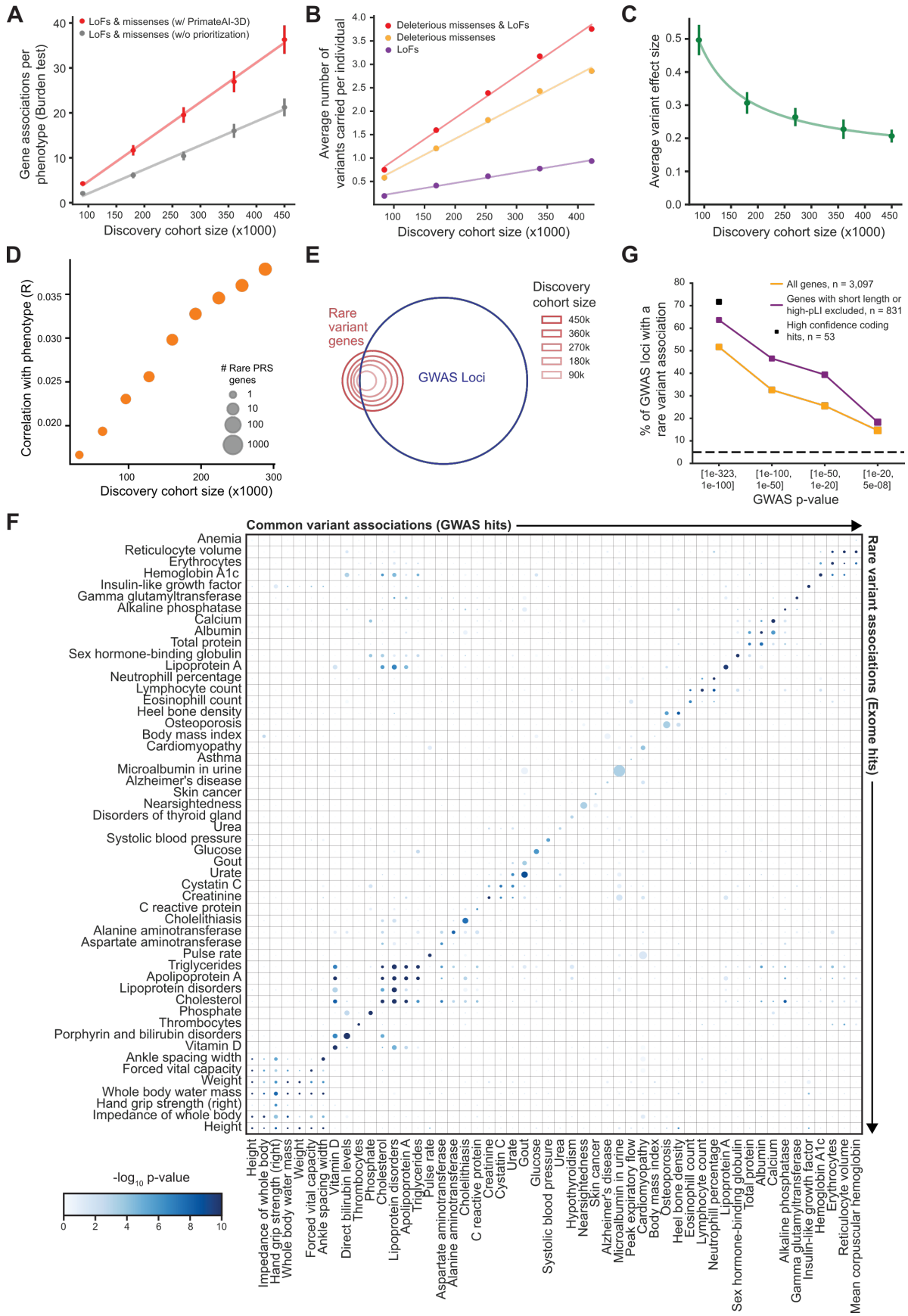
A
LoFs & missenses (w/ PrimateAI-3D)
LoFs & missenses (w/o prioritization)
Gene associations per phenotype (Burden test)
Discovery cohort size (x1000)
B
Deleterious missenses & LoFs
Deleterious missenses
LoFs
Average number of variants carried per individual
Discovery cohort size (x1000)
C
Average variant effect size
Discovery cohort size (x1000)
D
Correlation with phenotype (R)
Rare PRS genes
1
10
100
1000
Discovery cohort size (x1000)
E
Rare variant genes
GWAS Loci
Discovery cohort size
450k
360k
270k
180k
90k
G
All genes, n = 3,097
Genes with short length or high-pLI excluded, n = 831
High confidence coding hits, n = 53
% of GWAS loci with a rare variant association
[1e-323, 1e-100]
[1e-100, 1e-50]
[1e-50, 1e-20]
[1e-20, 5e-08]
GWAS p-value
F
Common variant associations (GWAS hits)
Rare variant associations (Exome hits)
Anemia
Reticulocyte volume
Erythrocytes
Hemoglobin A1c
Insulin-like growth factor
Gamma glutamyltransferase
Alkaline phosphatase
Calcium
Albumin
Total protein
Sex hormone-binding globulin
Lipoprotein A
Neutrophill percentage
Lymphocyte count
Eosinophill count
Heel bone density
Osteoporosis
Body mass index
Cardiomyopathy
Asthma
Microalbumin in urine
Alzheimer's disease
Skin cancer
Nearsightedness
Disorders of thyroid gland
Urea
Systolic blood pressure
Glucose
Gout
Urate
Cystatin C
Creatinine
C reactive protein
Cholelithiasis
Alanine aminotransferase
Aspartate aminotransferase
Pulse rate
Triglycerides
Apolipoprotein A
Lipoprotein disorders
Cholesterol
Phosphate
Thrombocytes
Porphyrin and bilirubin disorders
Vitamin D
Ankle spacing width
Forced vital capacity
Weight
Whole body water mass
Hand grip strength (right)
Impedance of whole body
Height
-log10 p-value
0
2
4
6
8
10

Fig. 4. Forecasting the growth of rare variant associations with increasing cohort size. (**A**) Number of significant (FDR < 0.05) genes identified per phenotype with rare variant burden tests as a function of the discovery cohort size in thousands of individuals. Missense prioritization with PrimateAI-3D substantially increased the number of genes detected at all cohort sizes. Dots and bars represent mean ± standard error. (**B**) Number of rare deleterious variants identified per individual as a function of the discovery cohort size. (**C**) Average per variant absolute effect size for newly associated genes (FDR < 0.05) at each discovery cohort size. The fit from the regression $y = {}^{a}/_{x} + b$ is shown. Dots and error bars represent mean ± standard error. **(D)** Rare variant PRS performance increases with increasing discovery cohort size. Median correlation between the PRSs and the phenotype is shown on the y-axis. Number of genes included in the PRS is represented by the size of each point. **(E)** Venn diagram showing the overlap of rare variant genes with common variant GWAS loci as a function of discovery cohort size. **(F)** A non-symmetrical heatmap showing the phenotype-specific overlap of common and rare variant associations. Each point shows the statistical significance of the overlap between common variant GWAS genes associated with the x-axis phenotype and rare variant genes associated with the y-axis phenotype. The size of the points represents the magnitude of the enrichment while the color represents the P-value. **(G)** Percentage of unambiguously mapped GWAS genes with rare variant associations (nominal P-value ≤ 0.05) stratified by GWAS significance thresholds. Results are shown for all genes (orange) and after excluding genes that are less likely to show rare variant signal (purple) due to short length (<2 kb coding sequence) or strong selective constraint (pLI > 0.99, probability of being loss-of-function intolerant). High-confidence coding hits are defined as having a lead variant with GWAS $P < 10^{-100}$ with strong linkage disequilibrium ($r^2 >= 0.9$) to a coding variant in the associated gene. The dashed line shows the background false discovery rate.

Table 1: Comparison of effect sizes and frequencies for common PRS variants and rare PRS genes used for normalized cholesterol concentrations.

Common PRS variants							Rare PRS genes			
Chrom	Position (GRCh37)	Major allele	Minor allele	Allele Freq	Effect size (z-score)	P-value	Gene	Aggregate Allele Freq	Effect size (z-score)	P-value
1	109415445	C	G	0.012	-0.071	1.2×10^{-10}	*ABCA1*	0.009	-0.377	6.8×10^{-110}
8	59393273	A	G	0.336	0.040	5.0×10^{-111}	*ABCA6*	0.009	0.105	3.6×10^{-8}
8	74894748	G	A	0.321	-0.016	1.3×10^{-18}	*ABCG5*	0.009	0.168	4.8×10^{-23}
14	74250100	C	T	0.275	-0.014	3.8×10^{-12}	*ABCG8*	0.006	0.137	1.6×10^{-8}
17	38244153	G	A	0.334	0.016	1.6×10^{-19}	*ALB*	0.0004	0.827	1.1×10^{-26}
7	18091019	A	G	0.199	0.015	2.3×10^{-10}	*ANGPTL3*	0.003	-0.641	1.5×10^{-119}
10	65255514	CA	C	0.440	0.012	1.0×10^{-15}	*APOA1*	0.003	-0.235	9.6×10^{-12}
19	10948031	A	G	0.178	-0.079	5.3×10^{-198}	*APOB*	0.002	-1.455	$<2.3\times10^{-308}$
6	161111700	T	C	0.015	0.185	2.2×10^{-77}	*APOE*	0.004	-0.183	1.5×10^{-10}
19	11192226	C	T	0.060	0.236	$<2.3\times10^{-308}$	*ASGR1*	0.001	-0.37	4.5×10^{-15}
7	75899085	C	T	0.144	0.021	4.5×10^{-13}	*B4GALT1*	0.0002	-0.74	6.9×10^{-7}
20	62909520	A	G	0.199	0.018	6.6×10^{-11}	*DENND4C*	0.002	0.274	1.9×10^{-11}
19	45319631	A	G	0.046	-0.417	2.3×10^{-308}	*FCGRT*	0.002	0.239	4.1×10^{-8}
16	79504057	A	G	0.245	0.017	1.4×10^{-14}	*G6PC1*	0.003	0.18	2.5×10^{-8}
20	62696024	T	C	0.495	-0.013	4.5×10^{-16}	*GAS6*	0.001	0.268	2.2×10^{-6}
19	11257169	T	C	0.228	-0.083	1.6×10^{-215}	*GPAM*	0.002	-0.25	4.0×10^{-8}
13	74735830	T	C	0.499	-0.017	2.8×10^{-33}	*JAK2*	0.002	-0.34	3.9×10^{-17}
3	69810294	G	T	0.349	-0.017	1.4×10^{-22}	*LCAT*	0.002	-0.347	2.8×10^{-26}
17	65259726	G	A	0.477	0.012	2.2×10^{-18}	*LDLR*	0.003	0.814	1.2×10^{-186}
1	109427458	G	A	0.030	-0.052	4.2×10^{-14}	*LIPC*	0.009	0.111	2.9×10^{-9}
5	156369171	C	T	0.347	-0.045	9.0×10^{-148}	*LIPG*	0.001	0.36	1.0×10^{-13}
1	220970593	T	G	0.314	-0.036	8.0×10^{-85}	*NPC1L1*	0.013	-0.153	1.6×10^{-26}
2	118535808	T	C	0.082	-0.042	1.9×10^{-25}	*NR1I3*	0.002	-0.21	2.0×10^{-7}
6	26093141	A	G	0.078	-0.060	1.0×10^{-47}	*PCSK9*	0.005	-0.812	1.9×10^{-284}
20	47724665	CA	C	0.304	0.014	3.8×10^{-14}	*RRBP1*	0.002	-0.247	2.6×10^{-7}
4	40036216	CA	C	0.071	-0.024	1.1×10^{-7}	*SCARB1*	0.007	0.19	6.9×10^{-21}
10	17255095	A	G	0.418	0.019	2.6×10^{-35}	*SH2B3*	0.004	-0.154	1.4×10^{-6}
	. . .						*SLC4A1*	0.0002	-0.606	1.9×10^{-7}
	. . .						*STAB1*	0.011	0.108	4.4×10^{-11}
536 variants omitted							*TIMD4*	0.002	0.302	9.2×10^{-13}
							TM6SF2	0.005	-0.158	2.7×10^{-11}