Improving the informativeness of Mendelian disease pathogenicity scores for common disease

Samuel S. Kim[1,2,*], Kushal K. Dey[2], Omer Weissbrod[2], Carla Marquez-Luna[3], Steven Gazal[2], and Alkes L. Price [*2,4,5,*]

[1]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, 02142

[2]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, 02115

[3]Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY 10029

[4]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, 02142

[5]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, 02115

[*]Corresponding author: sungil@mit.edu, aprice@hsph.harvard.edu

Abstract

Despite considerable progress on pathogenicity scores prioritizing both coding and non-coding variants for Mendelian disease, little is known about the utility of these pathogenicity scores for common disease. Here, we sought to assess the informativeness of Mendelian disease pathogenicity scores for common disease, and to improve upon existing scores. We first applied stratified LD score regression to assess the informativeness of annotations defined by top variants from published Mendelian disease pathogenicity scores across 41 independent common diseases and complex traits (average N = 320K). Several of the resulting annotations were informative for common disease, even after conditioning on a broad set of coding, conserved, regulatory and LD-related annotations from the baseline-LD model. We then improved upon the published pathogenicity scores by developing AnnotBoost, a gradient boosting-based framework to impute and denoise pathogenicity scores using functional annotations from the baseline-LD model. AnnotBoost substantially increased the informativeness for common disease of both previously uninformative and previously informative pathogenicity scores; our combined joint model included 3 published and 8 boosted scores. The boosted scores also significantly outperformed the corresponding published scores in classifying disease-associated, fine-mapped SNPs. Our boosted scores have high potential to improve candidate gene discovery and fine-mapping for common disease.

Introduction

Despite considerable progress on pathogenicity scores prioritizing both coding and non-coding variants for Mendelian disease[1–10] (reviewed in ref. 11), little is known about the utility of these pathogenicity scores for common disease. The shared genetic architecture between Mendelian disease and common disease has been implicated in studies reporting the impact of genes underlying monogenic forms of common diseases on the corresponding common diseases[12], significant cormobidities among Mendelian and complex diseases[13], and gene-level overlap between Mendelian diseases and cardiovascular diseases[14–16], neurodevelopmental traits[17,18], and other complex traits[19]. However, variant-level assessment of shared genetic architecture using Mendelian pathogenicity scores has not been explored. Thus, our current understanding of the genetic relationship between Mendelian disease and common disease remains limited.

Here, we sought to assess the informativeness of Mendelian disease pathogenicity scores for common disease, and to improve upon existing scores. We focused our attention on polygenic common and low-frequency variant architectures, which explain the bulk of common disease heritability[20–24]. We assessed the informativeness of annotations defined by top variants from published Mendelian disease pathogenicity scores by applying stratified LD score regression[25] (S-LDSC) with the baseline-LD model[26,27] to 41 independent common diseases and complex traits (average $N =$ 320K). We assessed informativeness conditional on the baseline-LD model, which includes a broad set of coding, conserved, regulatory and LD-related annotations.

We improved upon the published pathogenicity scores by developing AnnotBoost, a gradient boosting-based machine learning framework to impute and denoise pathogenicity scores using functional annotations from the baseline-LD model. We assessed the informativeness of annotations defined by top variants from the boosted scores by applying S-LDSC and assessing informativeness conditional on annotations from the baseline-LD model as well as annotations derived from the corresponding published scores. We also assessed the informativeness of the published and boosted pathogenicity scores in predicting disease-associated, fine-mapped SNPs.

Results

Overview of methods

We define a binary annotation as an assignment of a binary value (0 or 1) to each SNP with minor allele count ≥ 5 in a 1000 Genomes Project European reference panel[28], as in our previous work[25]. We define a pathogenicity score as an assignment of a numeric value quantifying predicted pathogenicity to some or all of these SNPs; we analyze 11 Mendelian missense scores, 6 genome-wide Mendelian scores, and 18 additional scores. Our primary focus is on binary annotations defined either using top variants from published (missense or genome-wide) Mendelian disease pathogenicity scores, or using top variants from boosted scores that we constructed from those pathogenicity scores using AnnotBoost, a gradient boosting-based framework that we developed to impute and denoise pathogenicity scores using coding, conserved, regulatory and LD-related annotations from the baseline-LD model[26,27] (see Methods). AnnotBoost uses decision trees to distinguish pathogenic variants (defined using the input pathogenicity score) from benign variants; the AnnotBoost model is trained using the XGBoost gradient boosting software[29] (see URLs). AnnotBoost uses odd (resp. even) chromosomes as training data to make predictions for even (resp. odd) chromosomes; the output of AnnotBoost is the predicted probability of being pathogenic. Further details are provided in the Methods section; we have publicly released open-source software implementing AnnotBoost, as well as all pathogenicity scores and binary annotations analyzed in this work (see URLs).

We assessed the informativeness of the resulting binary annotations for common disease heritability by applying S-LDSC[25] to 41 independent common diseases and complex traits[30] (average $N = 320$K; Table S1; see URLs), conditioned on coding, conserved, regulatory and LD-related annotations from the baseline-LD model[26,27] and meta-analyzing results across traits. We assessed informativeness for common disease using standardized effect size (τ^*), defined as the proportionate change in per-SNP heritability associated to a one standard deviation increase in the value of the annotation, conditional on other annotations[26] (see Methods). We also computed the heritability enrichment, defined as the proportion of heritability divided by the proportion of SNPs. Unlike enrichment, τ^* quantifies effects that are unique to the focal annotation (see Methods). We also assessed the informativeness of the underlying (published and/or boosted) pathogenicity scores in

predicting disease-associated, fine-mapped SNPs.

Informativeness of Mendelian missense scores for common disease

We assessed the informativeness for common disease of binary annotations derived from 11 Mendelian disease pathogenicity scores for missense variants[1,5–8,31–35] (see Table 1). These scores reflect the predicted impact of missense mutations on Mendelian disease and were trained using rare variants, primarily from ClinVar[36] and Human Gene Mutation Database (HGMD)[37]. For each of the 11 missense scores, we constructed binary annotations based on top missense variants using 5 different thresholds (from top 50% to top 10% of missense variants) and applied S-LDSC[25,26] to 41 independent common diseases and complex traits (Table S1), conditioning on coding, conserved, regulatory and LD-related annotations from the baseline-LD model[26,27] and meta-analyzing results across traits; proportions of top SNPs were optimized to maximize informativeness (see Methods). We incorporated the 5 different thresholds into the number of hypotheses tested when assessing statistical significance (Bonferroni $P < 0.05/500 = 0.0001$, based on a total of ≈ 500 hypotheses tested in this study; see Methods). We identified (Bonferroni-significant) conditionally informative binary annotations derived from 2 published missense scores: the top 30% of SNPs from MPC[34] (enrichment = 27x (s.e. 2.5), $\tau^* = 0.60$ (s.e. 0.07)) and the top 50% of SNPs from PrimateAI[8] (enrichment = 17x (s.e. 2.0), $\tau^* = 0.42$ (s.e. 0.09) (Figure 1, Table 2 and Table S2). The MPC (Missense badness, PolyPhen-2, and Constraint) score[34] is computed by identifying regions within genes that are depleted for missense variants in ExAC data[38] and incorporating variant-level metrics to predict the impact of missense variants; the PrimateAI score[8] is computed by eliminating common missense variants identified in other primate species (which are presumed to be benign in humans), incorporating a deep learning model trained on the amino acid sequence flanking the variant of interest and the orthologous sequence alignments in other species. The remaining published Mendelian missense scores all had derived binary annotations that were significantly enriched for disease heritability (after Bonferroni correction) but not conditionally informative (except for the published M-CAP score, which spanned too few SNPs to be included in the S-LDSC analysis).

We constructed *boosted* scores from the 11 Mendelian missense scores using AnnotBoost, a gradient boosting-based machine learning framework that we developed to impute and denoise pathogenicity scores using functional annotations from the baseline-LD model[26] (see Methods).

We note that AnnotBoost scores genome-wide (missense and non-missense) variants, implying low genome-wide correlations between input Mendelian missense scores and corresponding genome-wide boosted scores (0.02-0.24; Table S3A). AnnotBoost attained high predictive accuracy in out-of-sample predictions of input missense scores (AUROC= 0.76-0.94, AUPRC= 0.43-0.82; Table S4), although we caution that high predictive accuracy does not necessarily translate into conditional informativeness for common disease[39]. We further note that out-of-sample AUROCs closely tracked the genome-wide correlations between input Mendelian missense scores and corresponding genome-wide boosted scores ($r = 0.65$), implying that accurately predicting the input pathogenicity scores results in more correlated boosted scores.

For each missense pathogenicity score, after running AnnotBoost, we constructed binary annotations based on top genome-wide variants, using 6 different thresholds (ranging from top 10% to top 0.1% of genome-wide variants, as well as variants with boosted scores ≥ 0.5; see Methods). We assessed the informativeness for common disease of binary annotations derived from each of the 11 boosted scores using S-LDSC, conditioning on annotations from the baseline-LD model and 5 binary annotations derived from the corresponding published Mendelian missense score (using all 5 thresholds) (baseline-LD+5). We identified conditionally informative binary annotations derived from boosted versions of 10 Mendelian missense scores, including 8 previously uninformative scores and the 2 previously informative scores (Figure 1, Table 2 and Table S2). Letting $\uparrow$ denote boosted scores, examples include the top 0.1% of SNPs from M-CAP$\uparrow$[7], a previously uninformative score (enrichment = 23x (s.e. 2.6), $\tau^* = 0.43$ (s.e. 0.08); the published M-CAP pathogenicity score spanned too few SNPs to be included in the S-LDSC analysis of Figure 1) and the top 0.1% of SNPs from PrimateAI$\uparrow$[7], a previously informative score (enrichment = 35x (s.e. 2.7), $\tau^* = 0.83$ (s.e. 0.08)). The M-CAP (Mendelian Clinically Applicable Pathogenicity) score[7] is computed by training a gradient boosting tree classifier to distinguish pathogenic variants from HGMD[37] vs. benign variants from ExAC[38] using 9 pathogenicity likelihood scores as features (including PolyPhen-2[1], MetaLR[32], CADD[2]; see Table 1); the PrimateAI score is described above. Interestingly, binary annotations derived from 7 boosted scores had significantly negative τ^* (-0.72 (s.e. 0.07)) to -0.13 (s.e. 0.01)). All but one of these binary annotations were significantly enriched for disease heritability, but less enriched than expected based on annotations from the baseline-LD+5 model (Table S5; see ref. 40 and Methods), and thus uniquely informative for disease heritability

(analogous to transposable element annotations in ref. 40 that were significantly depleted, but less depleted than expected and thus uniquely informative). The boosted version of the remaining Mendelian missense score (MVP↑; not included in Figure 1) had a derived binary annotation that was significantly enriched for disease heritability (after Bonferroni correction) but not conditionally informative (Table S2).

We performed two secondary analyses. First, we restricted the 10 significant binary annotations derived from our boosted Mendelian missense scores to non-coding regions, which were previously unscored by the Mendelian missense scores, and assessed the informativeness of the resulting non-coding binary annotations using S-LDSC. We determined that the non-coding annotations retained the bulk of the overall signals (85%-110% of absolute τ^*; Table S6), implying that AnnotBoost leverages information about pathogenic missense variants to usefully impute scores for non-missense variants. Second, we investigated which features of the baseline-LD model contributed the most to the informativeness of the boosted annotations by applying Shapley Additive Explanation (SHAP)[41], a widely used tool for interpreting machine-learning models. We determined that conservation-related features drove the predictions of the boosted annotations, particularly (binary and continuous) GERP scores[42] (Figure S1).

Informativeness of genome-wide Mendelian pathogenicity scores for common disease

We assessed the informativeness for common disease of binary annotations derived from 6 genome-wide Mendelian disease pathogenicity scores[2–4,9,10] (see Table 1). These scores reflect the predicted impact of (coding and) non-coding variants on Mendelian disease and were also primarily trained using rare variants from ClinVar[36] and HGMD[37]. For each of the 6 genome-wide scores, we constructed binary annotations based on top genome-wide variants using 5 different thresholds (from top 0.1% to top 10% of genome-wide variants) and applied S-LDSC to the 41 traits, conditioning on the baseline-LD model[26] and meta-analyzing results across traits; proportions of top SNPs were optimized to maximize informativeness (see Methods). We identified (Bonferroni-significant) conditionally informative binary annotations derived from 3 genome-wide scores: the top 0.5% of SNPs from ReMM[4] (enrichment = 19x (s.e. 1.2), $\tau^* = 0.82$ (s.e. 0.09)), the top 0.5% of SNPs from CADD[2,43] (enrichment = 18x (s.e. 1.3), $\tau^* = 0.71$ (s.e. 0.10)), and the top 0.1% of SNPs from

Eigen[3] (enrichment = 24x (s.e. 2.1), $\tau^* = 0.40$ (s.e. 0.06)) (Figure 2, Table 2 and Table S7). The CADD (Combined Annotation Dependent Depletion) score[2,43] is computed by training a support vector machine to distinguish deleterious vs. neutral variants using functional annotations as features; the Eigen score[3] is computed from 29 input functional annotations by using an unsupervised machine learning method (leveraging blockwise conditional independence between annotations) to differentiate functional vs. non-functional variants; the ReMM (Regulatory Mendelian Mutation) score[4] is computed by training a random forest classifier to distinguish 406 hand-curated Mendelian mutations from neutral variants using conservation scores and functional annotations as features. The remaining 3 genome-wide scores all had derived binary annotations that were significantly enriched for disease heritability (after Bonferroni correction) but not conditionally informative (Table S7).

We applied AnnotBoost to the 6 genome-wide Mendelian scores. We observed moderate correlations between input genome-wide Mendelian scores and corresponding boosted scores ($r = 0.35$-0.66; Table S3B). AnnotBoost again attained high predictive accuracy in out-of-sample predictions of input genome-wide scores (AUROC = 0.83-1.00, AUPRC = 0.70-1.00; Table S4); however, out-of-sample AUROCs did not closely track the correlations between input genome-wide scores and corresponding boosted scores ($r = 0.05$).

We again constructed binary annotations based on top genome-wide variants, using 6 different thresholds (ranging from top 0.1% to top 10% of genome-wide variants, as well as variants with boosted scores ≥ 0.5; see Methods). We assessed the informativeness for common disease of binary annotations derived from each of the 6 boosted scores using S-LDSC, conditioning on annotations from the baseline-LD model and 5 binary annotations derived from the corresponding published genome-wide Mendelian score (using all 5 thresholds). We identified conditionally informative binary annotations derived from boosted versions of all 6 genome-wide Mendelian scores, including the 3 previously uninformative scores and the 3 previously informative scores (Figure 2, Table 2 and Table S2). Examples include the top 5% of SNPs from ncER↑[9] (enrichment = 6.2x (s.e. 0.30), $\tau^* = 0.74$ (s.e. 0.10)) and the top 0.5% of SNPs from boosted Eigen-PC↑[3] (enrichment = 16x (s.e. 1.1), $\tau^* = 0.62$ (s.e. 0.12)), both of which were previously uninformative scores, and the top 1% of SNPs from ReMM↑[4] (enrichment = 17x (s.e. 0.8), $\tau^* = 1.17$ (s.e. 0.12)), a previously informative score. The ncER (non-coding Essential Regulation) score[9] is computed by training a gradient

boosting tree classifier to distinguish non-coding pathogenic variants from ClinVar[36] and HGMD[37] vs. benign variants using 38 functional and structural features; the Eigen-PC score[3] (related to the Eigen score) is computed from 29 input functional annotations by using the lead eigenvector of the annotation covariance matrix to weight the annotations; the ReMM score is described above.

We performed two secondary analyses. First, for the 4 genome-wide Mendelian scores with <100% of SNPs scored (Table 1), we restricted the binary annotations derived from our boosted genome-wide Mendelian scores to previously unscored variants and assessed the informativeness of the resulting binary annotations using S-LDSC. We determined that these annotations retained only a minority of the overall signals (17%-54% of absolute τ^*; Table S8), implying that AnnotBoost usefully denoises previously scored variants. Second, we again investigated which features of the baseline-LD model contributed the most to the informativeness of the boosted annotations by applying SHAP[41]. We determined that both conservation-related features (e.g. GERP scores) and LD-related features (e.g. LLD-AFR; level of LD in Africans) drove the predictions of the boosted annotations (Figure S2).

Informativeness of additional genome-wide scores for common disease

For completeness, we assessed the informativeness for common disease of 18 additional genome-wide scores not directly related to Mendelian disease, including 2 constraint-related scores[44,45], 9 scores based on deep learning predictions of epigenetic marks[46–48], and 7 gene-based scores[38,49–51] (see Table S9). For each of the 18 additional scores, we constructed binary annotations based on top variants using 5 different thresholds and applied S-LDSC to the 41 traits, conditioning on the baseline-LD model[26] and meta-analyzing results across traits; in this analysis, we also conditioned on 8 Roadmap annotations[52] (4 annotations based on the union across cell types and 4 annotations based on the average across cell types, as in ref. 39), as many of the additional scores pertain to regulatory elements, making this an appropriate conservative step.

We identified (Bonferroni-significant) conditionally binary annotations derived from 6 informative scores, including the top 1% of SNPs from CDTS[44] (enrichment = 9.3x (s.e. 0.75), $\tau^* = 0.35$ (s.e. 0.06)) and the top 5% of SNPs from DeepSEA-H3K4me3[46,47] (enrichment = 3.9x (s.e. 0.23), $\tau^* = 0.21$ (s.e. 0.04)) (Figure 3, Table 2 and Table S10). CDTS (Context-Dependent Tolerance Score)[44] is a constraint score based on observed vs. expected variation in whole-genome sequence

data; DeepSEA-H3K4me3 scores[46,47] are computed by training a deep learning model to predict chromatin marks using proximal reference genome sequence as features and aggregated across different cell types[39] (The DeepSEA annotations in Figure 3 were more significant than those analyzed in ref. 39, because we optimized binary annotations based on top variants; however, no DeepSEA annotations were included in our combined joint model (see below)). 9 of the remaining 10 scores (excluding two that were not analyzed due to small annotation size) had derived binary annotations that were significantly enriched for disease heritability (after Bonferroni correction) but not conditionally informative (Table S10).

We applied AnnotBoost to the 18 additional scores, and to the 47 main annotations of the baseline-LD model (Table S9). Correlations between input scores and corresponding boosted scores varied widely (r = 0.005-0.93; Table S3C). AnnotBoost again attained high predictive accuracy in out-of-sample predictions of the input scores (AUROC = 0.55-1.00, AUPRC = 0.23-0.98; Table S4); out-of-sample AUROCs closely tracked the correlations between input scores and corresponding boosted scores (r = 0.65).

We again constructed binary annotations based on top genome-wide variants, using 6 different thresholds (ranging from top 0.1% to top 10% of genome-wide variants, as well as variants with boosted scores ≥ 0.5; see Methods). We assessed the informativeness for common disease of binary annotations derived from each of the 65 boosted scores using S-LDSC, conditioning on annotations from the baseline-LD model, the 8 Roadmap annotations, and (for the first 18 additional scores only) 5 binary annotations derived from the corresponding input scores (using all 5 thresholds). We identified conditionally informative binary annotations derived from boosted versions of 13/18 additional scores (including 11 previously uninformative scores and 2 previously informative scores) and 24/47 baseline-LD model annotations (Figure 3, Table 2 and Table S10). Examples include the top 10% of SNPs from DeepSEA-DNase↑[46,47] (enrichment = 3.7x (s.e. 0.27), τ^* = 0.69 (s.e. 0.11)), a previously uninformative score, the top 1% of SNPs from CCR↑[45] (enrichment = 7.9x (s.e. 0.65), τ^* = 0.51 (s.e. 0.09)), a previously uninformative score, and the top 5% of SNPs from H3K9ac↑[53] (enrichment = 5.4x (s.e. 0.31), τ^* = 0.76 (s.e. 0.09)), a baseline-LD model annotation. The CCR (Constrained Coding Regions) score[45] is a constraint score based on observed vs. expected variation in whole-exome sequence data; DeepSEA scores are described above. We note that the 18 additional scores included 7 gene-based scores, which did not perform well; 3 published gene-based

scores and 4 boosted gene-based scores yielded conditionally significant binary annotations, but their τ^* were small (-0.02 to 0.09). Boosted versions of 3 of the remaining 5 additional scores and 20 of the remaining 23 baseline-LD model annotations had derived binary annotations that were significantly enriched for disease heritability (after Bonferroni correction) but not conditionally informative (Table S10).

We performed 3 secondary analyses. First, for the 31 additional boosted scores which were marginally significant and for which the underlying published scores had $< 100\%$ of SNPs scored, we restricted the boosted scores to previously unscored variants and assessed the informativeness of the resulting binary annotations using S-LDSC. We determined that these annotations retained over half of the overall signals (average of 55% of absolute τ^*; Table S11), implying that AnnotBoost both imputes and denoises existing scores. Second, we again investigated which features of the baseline-LD model contributed the most to the informativeness of the boosted annotations by applying SHAP[41]. We determined that both conservation-related features (e.g. GERP scores) and LD-related features (e.g. LLD-AFR; level of LD in Africans) often drove the predictions of the boosted annotations, although results varied with the type of annotation (Figure S3). Third, we repeated the analyses of Figure 3 without including the 8 Roadmap annotations. We determined that the number of significant binary annotations increased (Table S12), confirming the importance of conditioning on the 8 Roadmap annotations as an appropriate conservative step[39]. We further verified that including the 8 Roadmap annotations did not impact results from previous sections (Table S13).

Classification of fine-mapped disease SNPs

We assessed the accuracy of the published and boosted scores listed in Table 2 in classifying 8,294 fine-mapped SNPs for 21 autoimmune diseases from Farh et al.[54] (defined by including all SNPs in 95% credible sets) vs. all other $\sim$10 million common and low-frequency SNPs (see Methods). We performed two analyses. In the first analysis (single-score analysis), we computed the AUROC individually attained by each of the 82 published and 82 boosted scores, comparing results for boosted scores vs. the corresponding published scores. In the second analysis (multi-score analysis), we computed the AUROC jointly attained by the baseline-LD model, 11 marginally significant published scores, and/or 53 marginally significant boosted scores (see Table 2), aggregated by

training a gradient boosting model; we used odd (resp. even) chromosomes as training data to make predictions for even (resp. odd) chromosomes (see Methods). We note that this gradient boosting model uses disease data (fine-mapped SNPs), whereas AnnotBoost does not use disease data to construct boosted pathogenicity scores.

Results of the single-score analysis are reported in Figure 4A and Table S14. We determined that the boosted scores significantly outperformed the corresponding published scores, with an average improvement in AUROC of 0.05. AUROC results for published and boosted scores were moderately correlated with S-LDSC results for binary annotations derived from these scores, validating the S-LDSC results (Table S15). We also computed values of AUPRC, which were close to 0 for both SNP sets (Table S15), as expected since false discovery rate is much higher than false positive rate in highly skewed data sets; this underscores the challenges of accurately classifying fine-mapped disease SNPs without directly using disease data.

Results of the multi-score analysis are reported in Figure 4B and Table S16. For Farh et al. SNPs, although the baseline-LD model (AUROC=0.844) was not significantly improved by adding the 11 (marginally significant) published annotations (ΔAUROC = 0.001, s.e. = 0.002, $p = 0.36$), further adding the 53 (marginally significant) boosted annotations produced a slight but significant improvement (ΔAUROC = 0.011, s.e. = 0.002, $p <$ 1e-6), further validating the added value of boosted annotations, above and beyond non-linear interactions involving published annotations and the baseline-LD model only. This improvement likely comes from non-linear interactions involving the boosted annotations, published annotations, and the baseline-LD model. We note that the high classification accuracy attained by the baseline-LD model could potentially be due to the LD- and MAF-related annotations in the baseline-LD model, as Farh et al. fine-mapped SNPs are more common (29% vs. 15%) with higher LD scores (132 vs. 120); we repeated the multi-score analysis by matching the LD and MAF of positive and control sets of SNPs and obtained similar results (Table S17). As in the single-score analysis, values of AUPRC were close to 0 for both SNP sets (Table S16).

We performed 3 secondary analyses. First, we repeated the single-score analysis using 3 additional sets of fine-mapped or disease-associated SNPs: 1,851 fine-mapped SNPs for 47 UK Biobank traits from Weissbrod et al.[55] (stringently defined by causal posterior probability ≥ 0.95); 21,296 NHGRI GWAS SNPs[56,57]; and 1,591 de novo SNPs from the sequenced whole genomes of 1,790

autism spectrum disorder simple families[48] which also appeared as common or low-frequency SNPs in the 1000 Genomes reference panel (see URLs). The boosted scores again significantly outperformed the corresponding published scores in each case, with highest AUROC values for the Weissbrod et al. fine-mapped SNPs (up to 0.875) (Figure S4 and Table S16). Second, we repeated the multi-score analysis using the 3 additional sets of fine-mapped or disease-associated SNPs. Adding the 53 boosted annotations to the baseline-LD model plus 11 published annotations produced a significant improvement in AUROC for the de novo SNPs, but not for the Weissbrod et al. fine-mapped SNPs or NHGRI GWAS SNPs, perhaps because the AUROC was already very high for the Weissbrod et al. fine-mapped SNPs (0.936) and because most NHGRI GWAS SNPs are not causal SNPs (Figure S5 and Table S16). We repeated the multi-score analysis of Weissbrod et al. fine-mapped SNPs using 1,853 SNPs that were fine-mapped without using functional information[55] (to ensure that results were not circular), and obtained similar results (Table S16). Third, we repeated the multi-score analysis using all 35 published scores (excluding baseline-LD model annotations) and 82 boosted scores from Table 2 (not just those that were marginally significant), and obtained similar results for all 4 SNP sets (Table S16).

Combined joint model

We jointly analyzed the 64 binary annotations (derived from 11 published scores and 53 boosted scores) that were marginally significant in our marginal analyses (Table 2) by performing forward stepwise elimination to iteratively remove annotations that had conditionally non-significant τ^* values after Bonferroni correction ($P \geq 0.05/500 = 0.0001$) or $\tau^* < 0.25$, conditioned on the baseline-LD model, the 8 Roadmap annotations, and each other (see ref. 26 and Methods). The resulting combined joint model included 11 binary annotations derived from 3 published scores and 8 boosted scores (Figure 5, Table 2 and Table S18). These 11 annotations are each substantially uniquely informative for common disease and include 5 boosted annotations with $\tau^* > 0.5$ (e.g. boosted ReMM: $\tau^* = 1.33$ (s.e. 0.12)); annotations with $\tau^* > 0.5$ are unusual, and considered to be very important[40]. We note that the top 0.5% of SNPs from REVEL↑[6] had significantly negative τ^* (-0.95 (s.e. 0.08)), as the annotation was significantly enriched for disease heritability but less enriched than expected based on annotations from the combined joint model.

We performed 3 secondary analyses. First, we computed genome-wide correlations between

the 11 jointly significant annotations and baseline-LD model annotations (Table S19). Several of the jointly significant annotations were strongly correlated (up to 0.73) with conservation-related annotations from the baseline-LD model, particularly binary GERP scores, consistent with our SHAP results (Figure S1, Figure S2 and Figure S3). Second, we compared the informativeness of the baseline-LD model and the combined joint model. We identified the addition of 11 jointly significant annotations greatly reduced the informativeness of several existing baseline-LD annotations, including conservation-related annotations (e.g. conserved primate, binary GERP scores) and other annotations (e.g. coding, CpG content; see Figure S6 and Table S20), recapitulating the informativeness of 11 jointly significant annotations. Third, we repeated our multi-score AUROC analysis (Figure 4B and Table S16) using only the 3 published and 8 boosted scores that were jointly significant, and obtained similar results for all 4 SNP sets (Table S16).

Discussion

We analyzed the informativeness of a broad set of Mendelian pathogenicity scores across 41 independent common diseases and complex traits to show that several annotations derived from published Mendelian pathogenicity scores were conditionally informative for common disease after conditioning on the baseline-LD model. We further developed AnnotBoost, a gradient boosting-based machine learning framework to impute and denoise existing pathogenicty scores. We determined that annotations derived from boosted pathogenicity scores were even more informative for common disease, resulting in 11 jointly significant annotations in our combined joint model. Our boosted pathogenicity scores also outperformed the corresponding published scores in classifying disease-associated, fine-mapped SNPs, even when conditioning on the baseline-LD model. These variant-level results are substantially different from previous studies of gene-level overlap between Mendelian diseases and complex traits[12–19].

We note three key differences between AnnotBoost and previous approaches that utilized gradient boosting to identify pathogenic missense[7] and non-coding variants[9,10]. First, AnnotBoost uses a pathogenicity score as the only input and does not use disease data (e.g. ClinVar[36] or HGMD[37]). Second, AnnotBoost produces genome-wide scores, even when some SNPs are unscored by the input pathogenicity score. Third, AnnotBoost leverages 75 diverse features from the baseline-LD model[26,27], significantly more than previous approaches[7,9,10]. Indeed, we determined that AnnotBoost produces strong signals even when conditioned on those approaches.

Our findings have several ramifications for improving our understanding of common disease. First, it is of interest to assess the informativeness for common disease of Mendelian disease pathogenicity scores that may be developed in the future, particularly after imputing and denoising these scores using AnnotBoost. Second, annotations derived from published and boosted Mendelian pathogenicity scores can be used to improve functionally informed fine-mapping[55,58,59], as well as polygenic risk prediction[60,61] and association mapping[62]. Third, elucidating specific mechanistic links between Mendelian disease and common disease may yield important biological insights.

We note several imitations of our work. First, while we showed that annotations derived from Mendelian disease pathogenicity scores are informative for common disease, we focused on common

and low-frequency variants and did not analyze Mendelian diseases. Second, S-LDSC is not well-suited to analysis of annotations spanning a very small proportion of the genome, preventing the analysis of a subset of published pathogenicity scores; nonetheless, our main results attained high statistical significance. Third, the gene-based scores that we analyzed did not perform well, perhaps because they were defined using 100kb windows, a crude strategy employed in previous work[30,50,63]; better strategies for linking regulatory variants to genes[64,65] could potentially improve upon those results. Despite these limitations, the imputed and denoised pathogenicity scores produced by our AnnotBoost framework have high potential to improve gene discovery and fine-mapping for common disease.

Methods

Genomic annotations and the baseline-LD model

We define a genomic annotation as an assignment of a numeric value to each SNP above a specified minor allele frequency (e.g. minor allele count ≥ 5) in a predefined reference panel (e.g. 1000 Genomes[28]; see URLs). Continuous-valued annotations can have any real value. Probabilistic annotations can have any real value between 0 and 1. Binary annotations can have value 0 or 1 only. A binary annotation can be viewed as a subset of SNPs (the set of SNPs with annotation value 1); we note all annotations analyzed in this work are binary annotations. Annotations that correspond to known or predicted function are referred to as functional annotations.

The baseline-LD model[26] (v2.1) contains 86 functional annotations (see URLs). We use these annotations as features of AnnotBoost (see below). These annotations include genomic elements (e.g. coding, enhancer, promoter), conservation (e.g. GERP, PhastCon), regulatory elements (e.g. histone marks, DNaseI-hypersensitive sites (DHS), transcription factor (TF) binding sites), and linkage disequilibrium (LD)-related annotations (e.g. predicted allele age, recombination rate, SNPs with low levels of LD).

Enrichment and τ^* metrics

We used stratified LD score regression (S-LDSC[25,26]) to assess the contribution of an annotation to disease heritability by estimating the enrichment and the standardized effect size (τ^*) of an annotation.

Let a_{cj} represent the (binary or probabilistic) annotation value of the SNP j for the annotation c. S-LDSC assumes the variance of per normalized genotype effect sizes is a linear additive contribution to the annotation c:

$$\mathrm{Var}(\beta_j) = \sum_c a_{cj}\tau_c \tag{1}$$

where τ_c is the per-SNP contribution of the annotation c. S-LDSC estimates τ_c using the following

equation:

$$\mathrm{E}[\chi_j^2] = N \sum_c \ell(j,c)\tau_c + 1 \tag{2}$$

where N is the sample size of the GWAS and $\ell(j,c)$ is the LD score of the SNP j to the annotation c. The LD score is computed as follow $\ell(j,c) = \sum_k a_{ck} r_{jk}^2$ where r_{jk} is the correlation between the SNPs j and k.

We used two metrics to assess the informativeness of an annotation. First, the standardized effect size (τ^*), the proportionate change in per-SNP heritability associated with a one standard deviation increase in the value of the annotation (conditional on all the other annotations in the model), is defined as follows:

$$\tau_c* = \frac{\tau_c sd(C)}{h_g^2/M} \tag{3}$$

where $sd(C)$ is the standard deviation of the annotation c, h_g^2 is the estimated SNP-heritability, and M is the number of variants used to compute h_g^2 (in our experiment, M is equal to 5,961,159, the number of common SNPs in the reference panel). The significance for the effect size for each annotation, as mentioned in previous studies[26,30,50], is computed as ($\frac{\tau^*}{\mathrm{se}(\tau^*)} \sim N(0,1)$), assuming that $\frac{\tau^*}{\mathrm{se}(\tau^*)}$ follows a normal distribution with zero mean and unit variance.

Second, enrichment of the binary and probabilistic annotation is the fraction of heritability explained by SNPs in the annotation divided by the proportion of SNPs in the annotation, as shown below:

$$\text{Enrichment} = \frac{\%h_g^2(C)}{\%\mathrm{SNP}(C)} = \frac{\frac{h_g^2(C)}{h_g^2}}{\frac{\sum_j a_{jc}}{M}} \tag{4}$$

where $h_g^2(C)$ is the heritability captured by the cth annotation. When the annotation is enriched for trait heritability, the enrichment is > 1; the overlap is greater than one would expect given the trait heritablity and the size of the annotation. The significance for enrichment is computed using the block jackknife as mentioned in previous studies[25,30,50,63].). The key difference between enrichment and τ^* is that τ^* quantifies effects that are unique to the focal annotation after conditioning on

all the other annotations in the model, while enrichment quantifies effects that are unique and/or non-unique to the focal annotation.

In all our analyses, we used the European samples in 1000G[28] (see URLs) as reference SNPs. Regression SNPs were obtained from HapMap 3[66] (see URLs). SNPs with marginal association statistics > 80 and SNPs in the major histocompatibility complex (MHC) region were excluded. Unless stated otherwise, we included the baseline-LD model[26] in all primary analyses using S-LDSC, both to minimize the risk of bias in enrichment estimates due to model mis-specification[25,26] and to estimate effect sizes (τ^*) conditional on known functional annotations.

Published Mendelian pathogenicity scores

We considered a total 35 published scores: 11 Mendelian missense pathogenicity scores, 6 genome-wide Mendelian pathogenicity scores, and 18 additional scores (see Table 1 and Table S9). Here, we provide a short description for Mendelian missense and genome-wide Mendelian pathogenicity scores. Details for 18 additional scores and the baseline-LD annotations are provided in Table S9. Our curated pathogenicity scores are available online (see URLs).

For all scores, we constructed annotations using GRCh37 assembly limited to all 9,997,231 low-frequency and common SNPs (with minor allele frequency (MAF) >= 0.05%) found in 1000 Genomes[28] European Phase 3 reference genome individuals (see URLs). Mendelian missense scores were readily available from dbNSFP database[67,68] using a rankscore (a converted score based on the rank among scored SNPs); genome-wide Mendealian scores were individually downloaded and used with no modification to original scores (see URLs). For each pathogenicity score, we constructed a binary annotation based on optimized threshold (See below). Short descriptions for each pathogenicity score (excluding 18 additional scores and the baseline-LD annotations; provided in Table S9) are provided below:

Mendelian missense pathogenicity scores:

PolyPhen-2[1,31] *(HDIV and HVAR)*: Higher scores indicate higher probability of the missense mutation being damaging on the protein function and structure. The default predictor is based on a naive Bayes classifier using HumDiv (HDIV), and the other is trained using HumVar (HVAR), using 8 sequence-based and 3 structure-based features.

MetaLR/MetaSVM[32]: An ensemble prediction score based on logistic regression (LR) or support vector machine (SVM) to classify pathogenic mutations from background SNPs in whole exome sequencing, combining 9 prediction scores and one additional feature (maximum minor allele frequency).

PROVEAN[33,69]: An alignment-based score to predict the damaging single amino acid substitutions.

SIFT 4G[5]: Predicted deleterious effects of an amino acid substition to protein function based on sequence homology and physical properties of amino acids.

REVEL[6]: An ensemble prediction score based on a random forest classifier trained on 6,182 missense disease mutations from HGMD[37], using 18 pathogenicity scores as features.

M-CAP[7]: An ensemble prediction score based on a gradient boosting classifier trained on pathogneic variants from HGMD[37] and benign variants from ExAC data set[38], using 9 existing pathogenicity scores, 7 base-pair, amino acid, genomic region, and gene-based features, and 4 features from multiple sequence alignments across 99 species.

PrimateAI[8]: A deep-learning-based score trained on the amino acid sequence flanking the variant of interest and the orthologous sequence alignments in other species and eliminating common missense variants identified in 6 non-human primate species.

MPC[34] *(missense badness, PolyPhen-2, and constraint)*: Logistic regression-based score to identify regions within genes that are depleted for missense variants in ExAC data[38] and incorporating variant-level metrics to predict the impact of missense variants. Higher MPC score indicates increased deleteriousness of amino acid substitutions once occured in missense-constrained regions.

MVP[35]: A deep-learning-based score trained on 32,074 pathogenic variants from ClinVar[36], HGMD[37], and UniProt[70], using 38 local context, constraint, conservation, protein structure, gene-based, and existing pathogenicity scores as features.

Genome-wide Mendelian pathogenicity scores:

CADD[2,43]: An ensemble prediction score based on a support vector machine classifier trained to differentiate 14.7 million high-frequency human-derived alleles from 14.7 million simulated variants, using 63 conservation, regulatory, protein-level, and existing pathogenicity scores as features. We used PHRED-scaled CADD score for all possible SNVs of GRCh37.

Eigen/Eigen-PC[3]: Unsupervised machine learning score based on 29 functional annotations and leveraging blockwise conditional independence between annotations to differentiate functional vs. non-functional variants. Eigen-PC uses the lead eigenvector of the annotation covariance matrix to weight the annotations. For both Eigen and Eigen-PC, we used PHRED-scaled scores and combined coding and non-coding regions to make it as a single genome-wide score. Higher score indicates more important (predicted) functional roles.

ReMM[4] *(regulatory Mendelian mutation)*: An ensemble prediction score based on a random forest classifier to to distinguish 406 hand-curated Mendelian mutations from neutral variants using conservation scores and functional annotations. Higher ReMM score indicate greater potential to cause a Mendelian disease if mutated.

NCBoost[10]: An ensemble prediction score based on a gradient boosting classifier trained on 283 pathogenic non-coding SNPs associated with Mendelian disease genes and 2830 common SNPs, using 53 conservation, natural selection, gene-based, sequence context, and epigenetic features.

ncER[9] *(non-coding essential regulation)*: An ensemble prediction score based on a gradient boosting classifier trained on 782 non-coding pathogenic variants from ClinVar[36] and HGMD[37], using 38 gene essentiality, 3D chromatin structure, regulatory, and existing pathogenicity scores as features.

AnnotBoost framework

AnnotBoost is based on gradient boosting, a machine learning method for classification; the AnnotBoost model is trained using the XGBoost gradient boosting software[29] (see URLs). AnnotBoost requires only one input, a pathogenicity score to boost, and generates a genome-wide (probabilistic) pathogenicity score. During the training, AnnotBoost uses decision trees, where each node in a tree splits SNPs into two classes (pathogenic and benign) using 75 coding, conserved, regulatory, and LD-related features from the baseline-LD model[26] (excluding 10 MAF bins features; we obtained similar results with or without MAF bins features; see Figure S7). The method generates training data from the input pathogenicity scores without using external variant data; top 10% SNPs from the input pathogenicity score are labeled as a positive training set, and bottom 40% SNPs are labeled as a control training set; we obtained similar results with other training data ratios (see Figure S8). As described in ref. 29, the prediction is based on T additive estimators (we use $T =$

200 to 300; see below), minimizing the following loss objective function L^t at the t-th iteration:

$$L^t = \sum_{i=1}^{n} l(y_i, \hat{y_i}^{t-1} + f_t(x_i)) + \gamma(f_t) \tag{5}$$

where l is a differentiable convex loss function (which measures the difference between the prediction ($\hat{y_i}$) and the target y_i at the i-th instance), f_t is an independent tree structure, and last term $\gamma(f_t)$ penalizes the complexity of the model, helping to avoid over-fitting. The prediction ($\hat{y_i}$) is made by $\sum_{t=1}^{T} f_t(x_i)$ by ensembling outputs of multiple weak-learner trees. Odd (resp. even) chromosome SNPs are used for training to score even (resp. odd) chromosome SNPs. The output of the classifier is the probability of being similar to the positive training SNPs and dissimilar to the control training SNPs.

We used the following model parameters: the number of estimators (200, 250, 300), depth of the tree (25, 30, 35), learning rate (0.05), gamma (minimum loss reduction required before additional partitioning on a leaf node; 10), minimum child weight (6, 8 ,10), and subsample (0.6, 0.8, 1); we optimized parameters with hyperparamters tuning (a randomized search) with five-fold cross-validation. Two important parameters to avoid over-fitting are gamma and learning rate; we chose these values consistent with previous studies[9,10]. The model with the highest AUROCs on the held-out data was selected and used to make a prediction.

To identify which feature(s) drives the prediction output with less bias, AnnotBoost uses Shapley Addictive Explanation (SHAP[41]), a widely used tool to interpret complex non-linear models, instead of built-in feature importance tool. SHAP uses the training matrix (features x SNP labels) and the trained model to generate a signed impact of each baseline-LD features on the AnnotBoost prediction.

To evaluate the performance of classifiers, we plotted receiver operating characteristic (ROC) and precision-recall (PR) curves. As we train AnnotBoost by splitting SNPs into odd and even chromosomes, we report the average out-of-sample area under the curve (AUC) of the odd and even chromosomes classifier. We used the threshold of 0.5 to define a class; that is, class 1 includes SNPs with the output probability > 0.5. We caution that high classification accuracy does not necessarily translate into conditional informativeness for common disease[39].

Constructing binary annotations using top variants from published and boosted scores

For published Mendelian missense pathogenicity scores, we considered five different thresholds to construct binary annotations: top 50%, 40%, 30%, 20% or 10% of scored variants. For published scores that produce Bonferroni-significant binary annotations, we report results for the binary annotation with largest $|\tau^*|$ among those that are Bonferroni-significant. For published scores that do not produce Bonferroni-significant binary annotations, we report results for the threshold with most significant τ^* (even though not Bonferroni-significant).

For all other published pathogenicity scores, we considered the top 10%, 5%, 1%, 0.5% or 0.1% of scored variants to construct binary annotations; we used more inclusive thresholds for published Mendelian missense pathogenicity scores due to the small proportion of variants scored ($\sim 0.3\%$; see Table 1). For published scores that produce Bonferroni-significant binary annotations, we report results for the binary annotation with largest $|\tau^*|$ among those that are Bonferroni-significant. For published scores that do not produce Bonferroni-significant binary annotations, we report results for the top 5% of variants (the average optimized proportion among Bonferroni-significant binary annotations); we made this choice because (in contrast to published Mendelian missense scores) for many other published scores the most significant τ^* was not even weakly significant.

For boosted pathogenicity scores, we considered the top 10%, 5%, 1%, 0.5% or 0.1% of scored variants, as well as variants with boosted scores ≥ 0.5; we note that top 10% of SNPs does not necessarily translate to 10% of SNPs, as some SNPs share the same score, and some genomic regions (e.g. MHC) are excluded when running S-LDSC (see below). For boosted scores that produce Bonferroni-significant binary annotations, we report results for the binary annotation with largest $|\tau^*|$ among those that are Bonferroni-significant. For boosted scores that do not produce Bonferroni-significant binary annotations, we report results for the top 5% of variants.

In all analyses, we excluded binary annotations with proportion of SNPs $< 0.02\%$ (the same threshold used in ref. 50), because S-LDSC does not perform well for small annotations[25].

In all primary analyses, we analyzed only binary annotations. However, we verified in a secondary analysis of the CDTS score[44] that probabilistic annotations produced results similar to binary annotations (see Figure S9).

Classification of fine-mapped disease SNPs: single-score analysis

As a primary analysis, we analyzed 8,294 common and low-frequency SNPs (of 8,741 total SNPs) fine-mapped for 21 autoimmune diseases from Farh et al.[54]. As a secondary analysis, we considered 3 other sets of SNPs: 1,851 common and low-frequency SNPs (of 2,225 SNPs, spanning 3,025 SNP-trait pairs) fine-mapped for 47 UK Biobank traits from Weissbrod et al.[55]; 21,296 common and low-frequency SNPs (of 23,205 total SNPs) from the NHGRI GWAS catalog[56,57] (2019-07-12 version; p-value $<$ 5e-8); and 1,591 de novo SNPs (out of 127,140 non-repeat-region single nucleotide variants) from the sequenced whole genomes of 1,790 autism spectrum disorder simple families[48] which also appeared as common or low-frequency SNPs in the 1000 Genomes reference panel (see URLs). We computed AUROCs for classifying each set of SNPs vs. all other $\sim$10 million common and low-frequency SNPs in the reference panel (European samples from 100 Genomes Phase 3[28]).

Classification of fine-mapped disease SNPs: multi-score analysis

We considered same 4 sets of SNPs as described above. We computed the AUROCs and AUPRCs jointly attained by:

- Annotations from the baseline-LD model
- 11 marginally significant published scores
- 53 marginally significant boosted scores
- 11 marginally significant published score + 53 marginally significant boosted scores
- 3 jointly significant published scores
- 8 jointly significant boosted scores
- 3 jointly significant published scores + 8 jointly significant boosted scores
- baseline-LD model + 11 marginally significant published scores
- baseline-LD model + 53 marginally significant boosted scores
- baseline-LD model + 11 marginally significant published scores + 53 marginally significant boosted scores
- baseline-LD model + 3 jointly significant published scores
- baseline-LD model + 8 jointly significant boosted scores
- baseline-LD model + 3 jointly significant published scores + 8 jointly significant boosted scores

We aggregated these scores by training a gradient boosting model (features: aggregated scores, labels: each of four sets of SNPs); we used odd (resp. even) chromosomes as training data to make predictions for even (resp. odd) chromosomes. We used the same training parameters as AnnotBoost (carefully selected to avoid over-fitting, consistent with the previous study[9,10]) with hyperparameters tuned using a randomized search method with five-fold cross-validation. We report the average AUROC and AUPRC of odd and even chromosome classifiers. We note that no disease data (four sets of SNPs used as labels) was re-used in these analyses, as AnnotBoost uses only the input pathogenicity scores to generate positive and negative sets of training data. We assessed the significance of the difference between two AUROCs as in ref. 71 (see URLs). For the LD-matched and MAF-matched analysis, we subsampled control SNPs to match the LD and MAF distribution of the positive SNP set.

URLs

AnnotBoost source code, published and boosted pathogenicity scores and binary annotations, and SHAP results: https://data.broadinstitute.org/alkesgroup/LDSCORE/Kim_annotboost

S-LDSC software: https://github.com/bulik/ldsc

XGBoost: https://github.com/dmlc/xgboost

SHAP (SHapley Additive exPlanations) feature importance: https://github.com/slundberg/shap

dbNSFP database: https://sites.google.com/site/jpopgen/dbNSFP

CADD scores: https://cadd.gs.washington.edu/download

Eigen/Eigen-PC scores: https://xioniti01.u.hpc.mssm.edu/v1.1/

ReMM scores: https://charite.github.io/software-remm-score.html

NCBoost scores: http://www.hli-opendata.com/noncoding/

ncER scores: http://www.hli-opendata.com/noncoding/

CDTS scores: http://www.hli-opendata.com/noncoding/

CCR scores: https://s3.us-east-2.amazonaws.com/ccrs/ccr.html/

DeepSEA (2018 version) scores: https://github.com/FunctionLab/ExPecto

DIS scores: Table S1. in ref. 48.

pLI scores: https://gnomad.broadinstitute.org/downloads

LIMBR scores: Table S1 in ref. 49.

Saha, Greene, InWeb, Sonawane network annotations:
https://data.broadinstitute.org/alkesgroup/LDSCORE/Kim_pathwaynetwork/

EDS scores: Table S1 in ref. 51.

baseline-LD (v.2.1) annotations: https://data.broadinstitute.org/alkesgroup/LDSCORE/

Significance of the difference in AUROCs calculator[71]: http://vassarstats.net/roc_comp.html

Ensembl biomart: https://www.ensembl.org/biomart

HapMap: ftp://ftp.ncbi.nlm.nih.gov/hapmap/

GWAS Catalog (Release v1.0): https://www.ebi.ac.uk/gwas.

1000 Genomes Project Phase 3 data: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502

PLINK software: https://www.cog-genomics.org/plink2

BOLT-LMM software: https://data.broadinstitute.org/alkesgroup/BOLT-LMM

BOLT-LMM summary statistics for UK Biobank traits: https://data.broadinstitute.org/alkesgroup/UKBB

UK Biobank: http://www.ukbiobank.ac.uk/

UK Biobank Genotyping and QC Documentation: http://www.ukbiobank.ac.uk/wp-content/uploads/2014/04/UKBiobank_genotyping_QC_documentation-web.pdf

Acknowledgements

We are grateful to Bryce van de Geijn, Farhad Hormozdiari, and Huwenbo Shi for helpful discussions. This research was funded by NIH grants U01 HG009379, R01 MH101244, R01 MH107649, and R01 MH109978.. This research was conducted using the UK Biobank Resource under Application 16549.

Contributions

S.S.K. and A.L.P. designed experiments. S.S.K. performed experiments. S.S.K., K.D., O.W., C.M-L., and S.G. analyzed data. S.S.K. and A.L.P. wrote the manuscript with the assistance from K.D., O.W., C.M-L., and S.G.

Declaration of Interests

The authors declare no competing interests.

References

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

2. Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nature Genetics *46*, 310.

3. Ionita-Laza, I., McCallum, K., Xu, B., and Buxbaum, J. D. (2016). A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nature Genetics *48*, 214.

4. Smedley, D., Schubach, M., Jacobsen, J. O., Köhler, S., Zemojtel, T., Spielmann, M., Jäger, M., Hochheiser, H., Washington, N. L., McMurry, J. A., et al. (2016). A whole-genome analysis framework for effective identification of pathogenic regulatory variants in mendelian disease. The American Journal of Human Genetics *99*, 595–606.

5. Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M., and Ng, P. C. (2016). Sift missense predictions for genomes. Nature protocols *11*, 1.

6. Ioannidis, N. M., Rothstein, J. H., Pejaver, V., Middha, S., McDonnell, S. K., Baheti, S., Musolf, A., Li, Q., Holzinger, E., Karyadi, D., et al. (2016). Revel: an ensemble method for predicting the pathogenicity of rare missense variants. The American Journal of Human Genetics *99*, 877–885.

7. Jagadeesh, K. A., Wenger, A. M., Berger, M. J., Guturu, H., Stenson, P. D., Cooper, D. N., Bernstein, J. A., and Bejerano, G. (2016). M-cap eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. Nature Genetics *48*, 1581.

8. Sundaram, L., Gao, H., Padigepati, S. R., McRae, J. F., Li, Y., Kosmicki, J. A., Fritzilas, N., Hakenberg, J., Dutta, A., Shon, J., et al. (2018). Predicting the clinical impact of human mutation with deep neural networks. Nature Genetics *50*, 1161.

9. Wells, A., Heckerman, D., Torkamani, A., Yin, L., Sebat, J., Ren, B., Telenti, A., and di Iulio, J. (2019). Ranking of non-coding pathogenic variants and putative essential regions of the human genome. Nature communications *10*, 1–9.

10. Caron, B., Luo, Y., and Rausell, A. (2019). Ncboost classifies pathogenic non-coding variants in mendelian diseases through supervised learning on purifying selection signals in humans. Genome biology *20*, 32.

11. Eilbeck, K., Quinlan, A., and Yandell, M. (2017). Settling the score: variant prioritization and mendelian disease. Nature Reviews Genetics *18*, 599.

12. Peltonen, L., Perola, M., Naukkarinen, J., and Palotie, A. (2006). Lessons from studying monogenic disease for common disease. Human molecular genetics *15*, R67–R74.

13. Blair, D. R., Lyttle, C. S., Mortensen, J. M., Bearden, C. F., Jensen, A. B., Khiabanian, H., Melamed, R., Rabadan, R., Bernstam, E. V., Brunak, S., et al. (2013). A nondegenerate code of deleterious variants in mendelian loci contributes to complex disease risk. Cell *155*, 70–80.

14. Teslovich, T. M., Musunuru, K., Smith, A. V., Edmondson, A. C., Stylianou, I. M., Koseki, M., Pirruccello, J. P., Ripatti, S., Chasman, D. I., Willer, C. J., et al. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. Nature *466*, 707.

15. Kathiresan, S. and Srivastava, D. (2012). Genetics of human cardiovascular disease. Cell *148*, 1242–1257.

16. Chong, J. X., Buckingham, K. J., Jhangiani, S. N., Boehm, C., Sobreira, N., Smith, J. D., Harrell, T. M., McMillin, M. J., Wiszniewski, W., Gambin, T., et al. (2015). The genetic basis of mendelian phenotypes: discoveries, challenges, and opportunities. The American Journal of Human Genetics *97*, 199–215.

17. Zhu, X., Need, A. C., Petrovski, S., and Goldstein, D. B. (2014). One gene, many neuropsychiatric disorders: lessons from mendelian diseases. Nature neuroscience *17*, 773.

18. Katsanis, N. (2016). The continuum of causality in human genetic disorders. Genome biology *17*, 233.

19. Freund, M. K., Burch, K. S., Shi, H., Mancuso, N., Kichaev, G., Garske, K. M., Pan, D. Z., Miao, Z., Mohlke, K. L., Laakso, M., et al. (2018). Phenotype-specific enrichment of mendelian disorder genes near gwas regions across 62 complex traits. The American Journal of Human Genetics *103*, 535–552.

20. Zeng, J., De Vlaming, R., Wu, Y., Robinson, M. R., Lloyd-Jones, L. R., Yengo, L., Yap, C. X., Xue, A., Sidorenko, J., McRae, A. F., et al. (2018). Signatures of negative selection in the genetic architecture of human complex traits. Nature genetics *50*, 746.

21. Zhang, Y., Qi, G., Park, J.-H., and Chatterjee, N. (2018). Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. Nature genetics *50*, 1318.

22. Zhu, X. and Stephens, M. (2018). Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes. Nature communications *9*, 4361.

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

24. O'Connor, L. J., Schoech, A. P., Hormozdiari, F., Gazal, S., Patterson, N., and Price, A. L. (2019). Extreme polygenicity of complex traits is explained by negative selection. The American Journal of Human Genetics *105*, 456–476.

25. Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K., et al. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nature Genetics *47*, 1228.

26. Gazal, S., Finucane, H. K., Furlotte, N. A., Loh, P.-R., Palamara, P. F., Liu, X., Schoech, A., Bulik-Sullivan, B., Neale, B. M., Gusev, A., et al. (2017). Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. Nature Genetics *49*, 1421.

27. Gazal, S., Marquez-Luna, C., Finucane, H. K., and Price, A. L. (2019). Reconciling s-ldsc and ldak functional enrichment estimates. Nature Genetics *51*, 1202–1204.

28. 1000 Genomes Project Consortium et al. (2015). A global reference for human genetic variation. Nature *526*, 68.

29. Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining ACM pp. 785–794.

30. Hormozdiari, F., Gazal, S., van de Geijn, B., Finucane, H. K., Ju, C. J.-T., Loh, P.-R., Schoech, A., Reshef, Y., Liu, X., O'Connor, L., et al. (2018). Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits. Nature Genetics *50*, 1041–1047.

31. Adzhubei, I., Jordan, D. M., and Sunyaev, S. R. (2013). Predicting functional effect of human missense mutations using polyphen-2. Current protocols in human genetics *76*, 7–20.

32. Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K., and Liu, X. (2014). Comparison and integration of deleteriousness prediction methods for nonsynonymous snvs in whole exome sequencing studies. Human molecular genetics *24*, 2125–2137.

33. Choi, Y., Sims, G. E., Murphy, S., Miller, J. R., and Chan, A. P. (2012). Predicting the functional effect of amino acid substitutions and indels. PloS one *7*, e46688.

34. Samocha, K. E., Kosmicki, J. A., Karczewski, K. J., O'Donnell-Luria, A. H., Pierce-Hoffman, E., MacArthur, D. G., Neale, B. M., and Daly, M. J. (2017). Regional missense constraint improves variant deleteriousness prediction. BioRxiv pp. 148353.

35. Qi, H., Chen, C., Zhang, H., Long, J. J., Chung, W. K., Guan, Y., and Shen, Y. (2018). Mvp: predicting pathogenicity of missense variants by deep learning. bioRxiv pp. 259390.

36. Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., et al. (2015). Clinvar: public archive of interpretations of clinically relevant variants. Nucleic acids research *44*, D862–D868.

37. Stenson, P. D., Mort, M., Ball, E. V., Evans, K., Hayden, M., Heywood, S., Hussain, M., Phillips, A. D., and Cooper, D. N. (2017). The human gene mutation database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Human genetics *136*, 665–677.

38. Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285.

39. Dey, K. K., Van de Geijn, B., Kim, S. S., Hormozdiari, F., Kelley, D. R., and Price, A. L. (2019). Evaluating the informativeness of deep learning annotations for human complex diseases. bioRxiv pp. 784439.

40. Hormozdiari, F., van de Geijn, B., Nasser, J., Weissbrod, O., Gazal, S., Ju, C. J.-T., O'Connor, L., Hujoel, M. L., Engreitz, J., Hormozdiari, F., et al. (2019). Functional disease architectures reveal unique biological role of transposable elements. Nature communications *10*.

41. Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems pp. 4765–4774.

42. Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., and Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using gerp++. PLoS computational biology *6*, e1001025.

43. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., and Kircher, M. (2018). Cadd: predicting the deleteriousness of variants throughout the human genome. Nucleic acids research *47*, D886–D894.

44. Di Iulio, J., Bartha, I., Wong, E. H., Yu, H.-C., Lavrenko, V., Yang, D., Jung, I., Hicks, M. A., Shah, N., Kirkness, E. F., et al. (2018). The human noncoding genome defined by genetic diversity. Nature Genetics *50*, 333.

45. Havrilla, J. M., Pedersen, B. S., Layer, R. M., and Quinlan, A. R. (2019). A map of constrained coding regions in the human genome. Nature Genetics *51*, 88–95.

46. Zhou, J. and Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. Nature methods *12*, 931.

47. Zhou, J., Theesfeld, C. L., Yao, K., Chen, K. M., Wong, A. K., and Troyanskaya, O. G. (2018). Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. Nature Genetics *50*, 1171.

48. Zhou, J., Park, C. Y., Theesfeld, C. L., Wong, A. K., Yuan, Y., Scheckel, C., Fak, J. J., Funk, J., Yao, K., Tajima, Y., et al. (2019). Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. Nature Genetics *51*, 973.

49. Hayeck, T. J., Stong, N., Wolock, C. J., Copeland, B., Kamalakaran, S., Goldstein, D. B., and Allen, A. S. (2019). Improved pathogenic variant localization via a hierarchical model of sub-regional intolerance. The American Journal of Human Genetics *104*, 299–309.

50. Kim, S. S., Dai, C., Hormozdiari, F., van de Geijn, B., Gazal, S., Park, Y., O'Connor, L., Amariuta, T., Loh, P.-R., Finucane, H., et al. (2019). Genes with high network connectivity are enriched for disease heritability. The American Journal of Human Genetics *104*, 896–913.

51. Wang, X. and Goldstein, D. B. (2018). Enhancer redundancy predicts gene pathogenicity and informs complex disease gene discovery. bioRxiv pp. 459123.

52. Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317.

53. Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B. E., Liu, X. S., and Raychaudhuri, S. (2013). Chromatin marks identify critical cell types for fine mapping complex trait variants. Nature Genetics *45*, 124.

54. Farh, K. K.-H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W. J., Beik, S., Shoresh, N., Whitton, H., Ryan, R. J., Shishkin, A. A., et al. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature *518*, 337.

55. Weissbrod, O., Hormozdiari, F., Benner, C., Cui, R., Ulirsch, J., Gazal, S., Schoech, A. P., Van De Geijn, B., Reshef, Y., Marquez-Luna, C., et al. (2019). Functionally-informed fine-mapping and polygenic localization of complex trait heritability. BioRxiv pp. 807792.

56. MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., et al. (2016). The new nhgri-ebi catalog of published genome-wide association studies (gwas catalog). Nucleic Acids Research *45*, D896–D901.

57. Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2018). The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic acids research *47*, D1005–D1012.

58. Kichaev, G., Yang, W.-Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A. L., Kraft, P., and Pasaniuc, B. (2014). Integrating functional data to prioritize causal variants in statistical fine-mapping studies. PLoS genetics *10*, e1004722.

59. Chen, W., McDonnell, S. K., Thibodeau, S. N., Tillmans, L. S., and Schaid, D. J. (2016). Incorporating functional annotations for fine-mapping causal variants in a bayesian framework using summary statistics. Genetics *204*, 933–958.

60. Hu, Y., Lu, Q., Powles, R., Yao, X., Yang, C., Fang, F., Xu, X., and Zhao, H. (2017). Leveraging functional annotations in genetic risk prediction for human complex diseases. PLoS computational biology *13*, e1005589.

61. Marquez-Luna, C., Gazal, S., Loh, P.-R., Kim, S. S., Furlotte, N., Auton, A., Price, A. L., 23andMe Research Team, et al. (2019). Modeling functional enrichment improves polygenic prediction accuracy in uk biobank and 23andme data sets. bioRxiv.

62. Kichaev, G., Bhatia, G., Loh, P.-R., Gazal, S., Burch, K., Freund, M. K., Schoech, A., Pasaniuc, B., and Price, A. L. (2019). Leveraging polygenic functional enrichment to improve gwas power. The American Journal of Human Genetics *104*, 65–75.

63. Finucane, H. K., Reshef, Y. A., Anttila, V., Slowikowski, K., Gusev, A., Byrnes, A., Gazal,

S., Loh, P.-R., Lareau, C., Shoresh, N., et al. (2018). Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. Nature Genetics *50*, 621.

64. Jung, I., Schmitt, A., Diao, Y., Lee, A. J., Liu, T., Yang, D., Tan, C., Eom, J., Chan, M., Chee, S., et al. (2019). A compendium of promoter-centered long-range chromatin interactions in the human genome. Nature genetics *51*, 1442–1449.

65. Fulco, C. P., Nasser, J., Jones, T. R., Munson, G., Bergman, D. T., Subramanian, V., Grossman, S. R., Anyoha, R., Doughty, B. R., Patwardhan, T. A., et al. (2019). Activity-by-contact model of enhancer–promoter regulation from thousands of crispr perturbations. Nature Genetics *51*, 1664–1669.

66. International HapMap 3 Consortium et al. (2010). Integrating common and rare genetic variation in diverse human populations. Nature *467*, 52.

67. Liu, X., Jian, X., and Boerwinkle, E. (2011). dbnsfp: a lightweight database of human nonsynonymous snps and their functional predictions. Human mutation *32*, 894–899.

68. Liu, X., Wu, C., Li, C., and Boerwinkle, E. (2016). dbnsfp v3. 0: A one-stop database of functional predictions and annotations for human nonsynonymous and splice-site snvs. Human mutation *37*, 235–241.

69. Choi, Y. and Chan, A. P. (2015). Provean web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics *31*, 2745–2747.

70. UniProt Consortium. (2018). Uniprot: a worldwide hub of protein knowledge. Nucleic acids research *47*, D506–D515.

71. Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. Radiology *143*, 29–36.

72. Loh, P.-R., Bhatia, G., Gusev, A., Finucane, H. K., Bulik-Sullivan, B. K., Pollack, S. J., de Candia, T. R., Lee, S. H., Wray, N. R., Kendler, K. S., et al. (2015). Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. Nature Genetics *47*, 1385.

73. Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P., and Price, A. L. (2018). Mixed-model association for biobank-scale datasets. Nature Genetics *50*, 906–908.

74. Bulik-Sullivan, B., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P.-R., Duncan, L., Perry, J. R., Patterson, N., Robinson, E. B., et al. (2015). An atlas of genetic correlations across human diseases and traits. Nature Genetics *47*, 1236.

Tables

Score	Description	Coverage (% SNPs scored)	Ref.
PolyPhen-2	Impact of missense variants using protein sequence and structure using HumDiv	0.28%	1, 31
PolyPhen-2-HVAR	Impact of missense variants using protein sequence and structure using HumVar	0.28%	1, 31
MetaLR	Deleterious missense mutations using ensemble scoring (logistic regression)	0.32%	32
MetaSVM	Deleterious missense mutations using ensemble scoring (support vector machine)	0.32%	32
PROVEAN	Impact of an amino acid change on protein function	0.31%	33, 69
SIFT 4G	Impact of an amino acid change on protein function	0.31%	5
REVEL	Pathogenic missense variants using ensemble scoring	0.32%	6
M-CAP	Pathogenic rare missense variants	0.03%	7
PrimateAI	Impact of missense variants using deep neural networks	0.26%	8
MPC	Regional missense constraint	0.10%	34
MVP	Impact of missense variants using deep neural networks	0.29%	35
CADD	Predicted deleterious variants using ensemble scoring	100%	2, 43
Eigen	Putatively causal variants using unsupervised learning	83.79%	3
Eigen-PC	Putatively causal variants using unsupervised learning using the lead eigenvector	83.79%	3
ReMM	Pathogenic regulatory variants using ensemble scoring	100%	4
NCBoost	Pathogenic non-coding variants using ensemble scoring	28.55%	10
ncER	Essential regulatory variants using ensemble scoring	61.94%	9

Table 1. 11 Mendelian missense and 6 genome-wide Mendelian pathogenicity scores. For each of 17 Mendelian disease pathogenicity scores analyzed, we provide a description and report the coverage (% of SNPs scored) and corresponding reference. The first 11 annotations are scores for missense variants, and the last 6 annotations are genome-wide scores. Annotations are ordered first by type and then by the year of publication.

Score	# scores	# marginally significant annotations		# significant annotations in a combined joint model	
		published	boosted	published	boosted
Mendelian missense	11	2*	10	1*	2
Genome-wide Mendelian	6	3	6	2	3
Additional scores	18	6**	13	0**	0
Baseline-LD model annotations	47	n/a	24	n/a	3

Table 2. Summary of informativeness for common disease of annotations derived from 82 published scores and corresponding boosted scores For each category of scores, we report the number of scores; the number of marginally conditionally informative annotations (S-LDSC τ^* $p < 0.0001$, conditional on the baseline-LD model; and the number of jointly conditionally informative annotations in a combined joint model (S-LDSC τ^* $p < 0.0001$ and $|\tau^*| \geq 0.25$, conditional on the baseline-LD model and each other). *Based on 9/11 published Mendelian missense scores analyzed, as binarized annotations were too small to analyze for the remaining 2 published Mendelian missense scores. **Based on 16/18 published additional scores analyzed, as binarized annotations were too small to analyzed for the remaining 2 published additional scores.

Figures

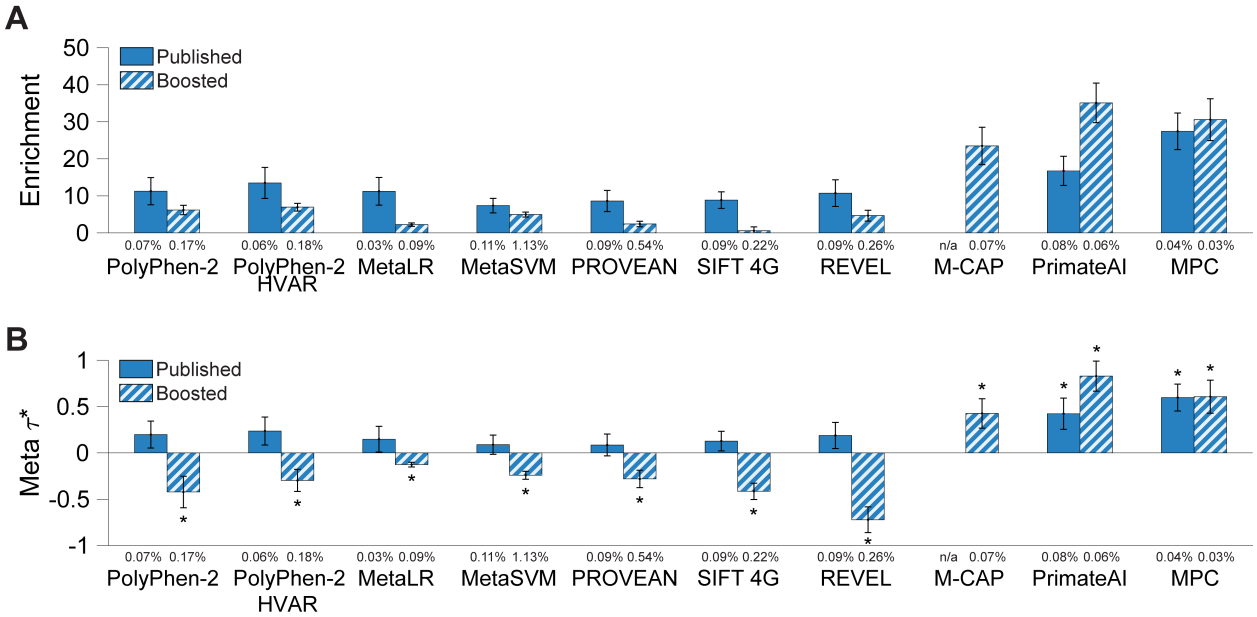

Figure 1. Informativeness for common disease of binary annotations derived from 11 Mendelian missense scores and corresponding boosted scores. We report (A) heritability enrichment of binary annotations derived from published and boosted Mendelian missense scores, meta-analyzed across 41 independent traits; (B) marginal τ^* values, conditional on the baseline-LD model (for annotations derived from published scores) or the baseline-LD model and corresponding published annotations (for annotations derived from boosted scores). We report results for 10 Mendelian missense scores (of 11 analyzed) for which annotations derived from published and/or boosted scores were marginally significant; the published M-CAP score spanned too few SNPs to be included in the S-LDSC analysis. The percentage under each bar denotes the proportion of SNPs in the annotation; the proportion of top SNPs included in each annotation was optimized to maximize informativeness (largest $|\tau^*|$ among Bonferroni-significant annotations, or most significant p-value if no annotation was Bonferroni-significant). Error bars denote 95% confidence intervals. In panel (B), * denotes marginally conditionally significant annotations. Numerical results are reported in Table S2. Results for standardized enrichment, defined as enrichment times the standard deviation of annotation value (to adjust for annotation size), are reported in Table S4.

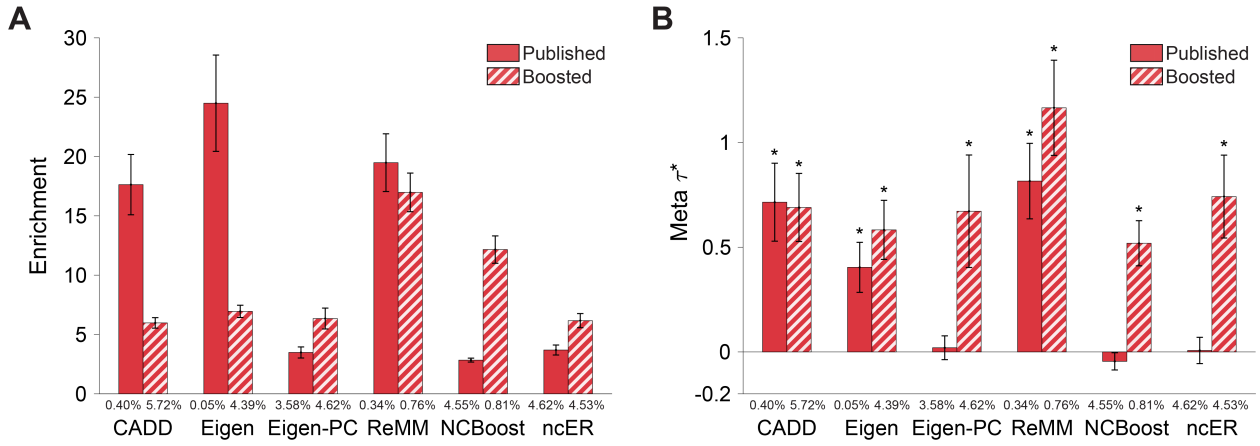

Figure 2. Informativeness for common disease of binary annotations derived from 6 genome-wide Mendelian scores and corresponding boosted scores. We report (A) heritability enrichment of binary annotations derived from published and boosted genome-wide Mendelian scores, meta-analyzed across 41 independent traits; (B) marginal τ^* values, conditional on the baseline-LD model (for annotations derived from published scores) or the baseline-LD model and corresponding published annotations (for annotations derived from boosted scores). We report results for 6 genome-wide Mendelian scores (of 6 analyzed) for which annotations derived from published and/or boosted scores were marginally significant. The percentage under each bar denotes the proportion of SNPs in the annotation; the proportion of top SNPs included in each annotation was optimized to maximize informativeness (largest $|\tau^*|$ among Bonferroni-significant annotations, or top 5% if no annotation was Bonferroni-significant; top 5% was the average optimized proportion among significant annotations). Error bars denote 95% confidence intervals. In panel (B), * denotes marginally conditionally significant annotations. Numerical results are reported in Table S7. Results for standardized enrichment, defined as enrichment times the standard deviation of annotation value (to adjust for annotation size), are reported in Table S4.

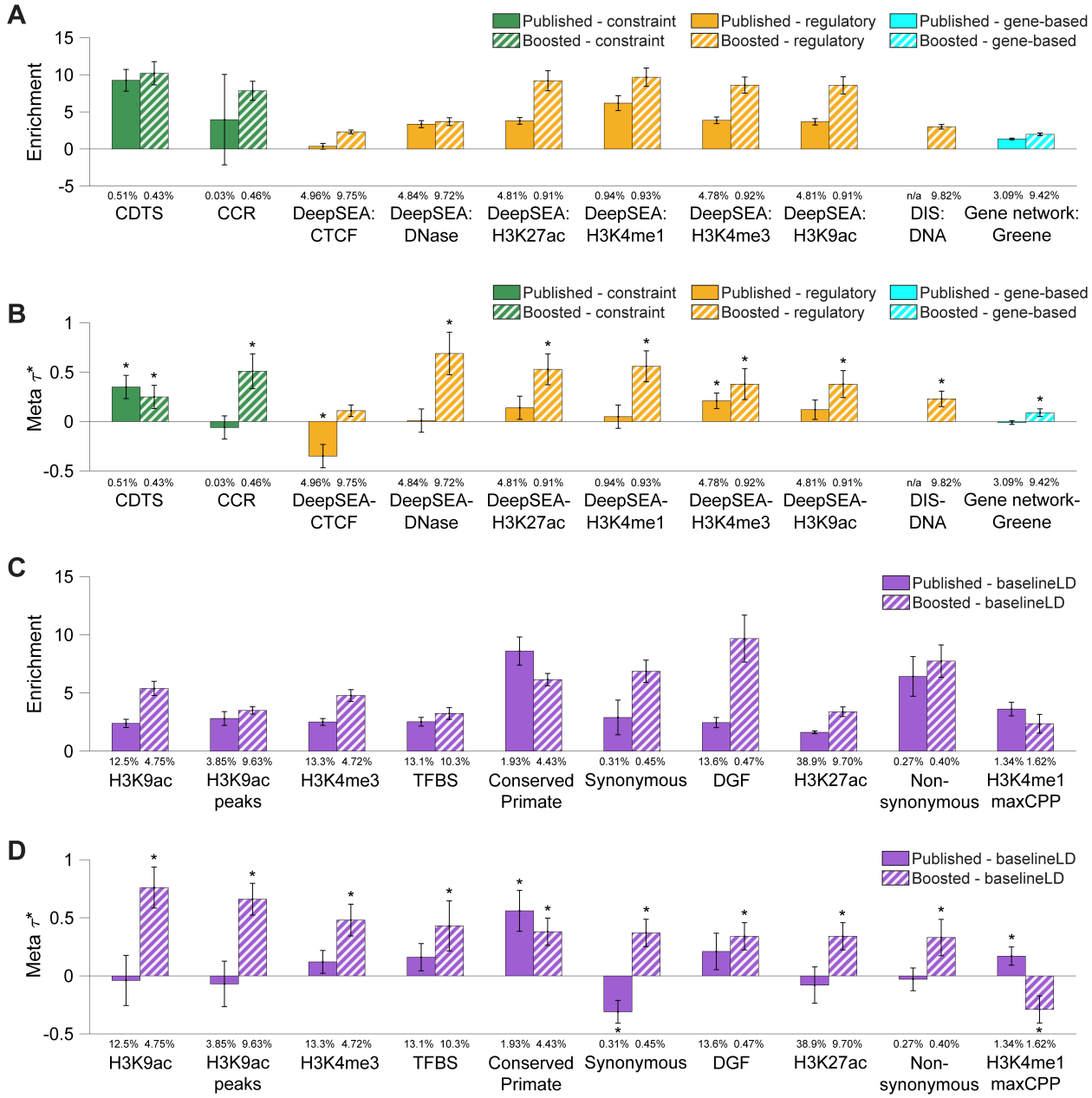

Figure 3. Informativeness for common disease of binary annotations derived from 18 additional genome-wide scores + 47 baseline-LD model annotations and corresponding boosted scores. We report (A) heritability enrichments of binary annotations derived from published and boosted additional genome-wide scores, meta-analyzed across 41 independent traits; (B) marginal τ^* values, conditional on the baseline-LD model and 8 Roadmap annotations (for annotations derived from published scores) or the baseline-LD model, 8 Roadmap annotations, and corresponding published annotations (for annotations derived from boosted scores); (C) heritability enrichments of binary annotations derived from published and boosted baseline-LD model annotations; and (D) marginal τ^* values of binary annotations derived from published and boosted baseline-LD model annotations. In (A) and (B), we report results for the 10 most informative additional genome-wide scores (of 18 analyzed). In (C) and (D), we report results for the 10 most informative baseline-LD model annotations (of 47 analyzed). The percentage under each bar denotes the proportion of SNPs in the annotation; the proportion of top SNPs included in each annotation was optimized to maximize informativeness (largest $|\tau^*|$ among Bonferroni-significant annotations, or top 5% if no annotation was Bonferroni-significant; top 5% was the average optimized proportion among significant annotations). Error bars denote 95% confidence intervals. In panels (B) and (D), * denotes marginally conditionally significant annotations. Numerical results are reported in Table S10. Results for standardized enrichment, defined as enrichment times the standard deviation of annotation value (to adjust for annotation size), are reported in Table S4.

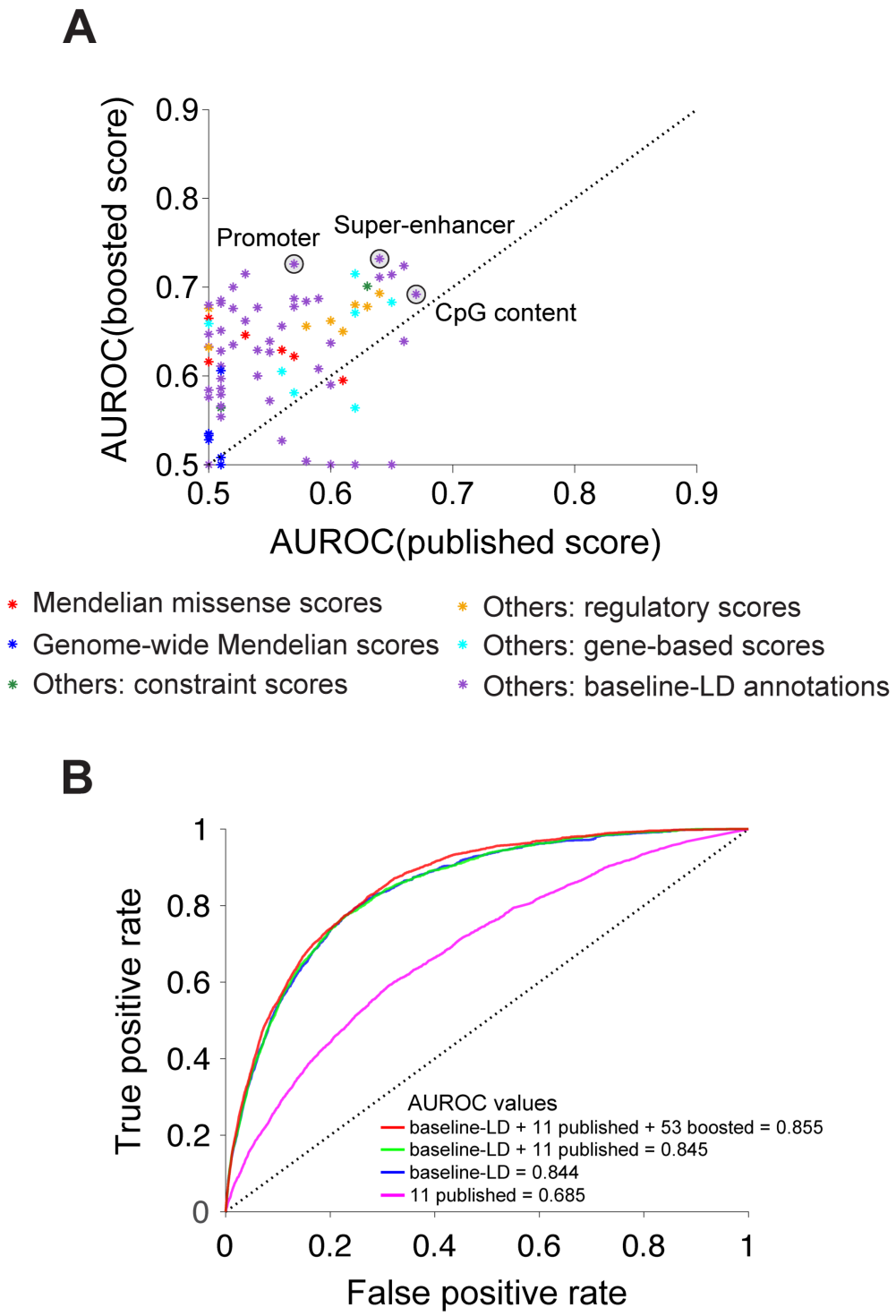

Figure 4. Classification of fine-mapped disease SNPs using published and boosted scores. (A) (single-score analysis) We report the classification accuracy (AUROC) of each of the 82 boosted scores compared to the corresponding published score. We highlighted three scores: the score with the highest AUROC(published) (CpG content), the score with the highest AUROC(boosted) (Super-enhancer), and the score with the largest difference between AUROC(boosted) and AUROC(published) (Promoter). Scores with AUROC < 0.5 are displayed as AUROC $= 0.5$. (B) (multi-score analysis) We report the true positive rate vs. false positive rate and the classification accuracy (AUROC) of four aggregated scores. Numerical results are reported in Table S14 and Table S16.

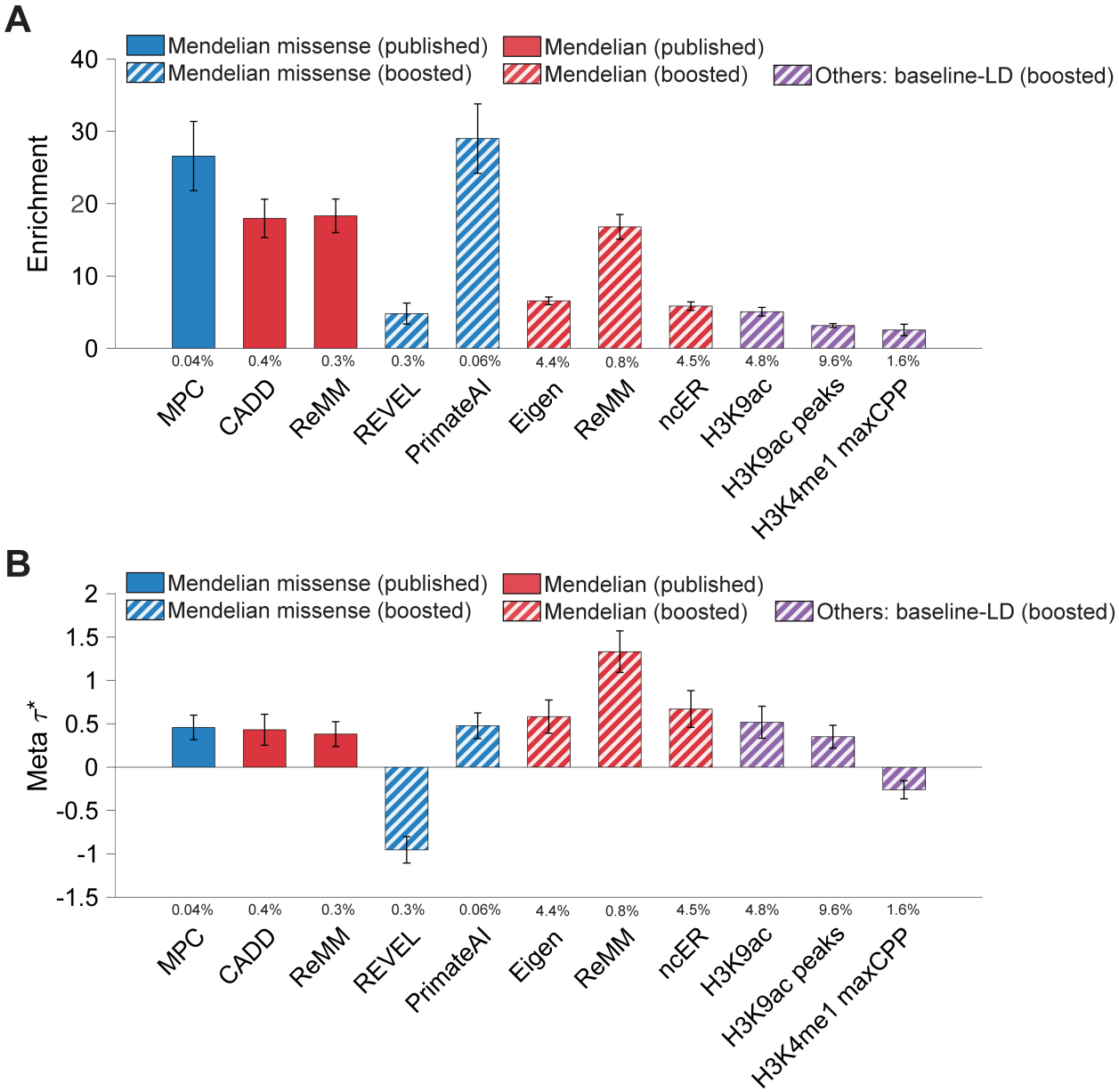

Figure 5. Informativeness for common disease of 11 jointly significant binary annotations from combined joint model. We report (A) heritability enrichment of 11 jointly significant binary annotations, meta-analyzed across 41 independent traits; (B) joint τ^* values, conditioned on the baseline-LD model, 8 Roadmap annotations, and each other. We report results for the 11 jointly conditionally informative annotations in the combined joint model (S-LDSC τ^* $p < 0.0001$ and $|\tau^*| \geq 0.25$). The percentage under each bar denotes the proportion of SNPs in the annotation. Error bars denote 95% confidence intervals. Numerical results are reported in Table S18. Results for standardized enrichment, defined as enrichment times the standard deviation of annotation value (to adjust for annotation size), are reported in Table S4.