Improving the informativeness of Mendelian disease-derived 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-derived 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-derived pathogenicity scores across 41 independent common diseases and complex traits (average $N = 320$K). 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, implying pervasive variant-level overlap between Mendelian disease and common disease. The boosted scores also produced significant improvements in heritability model fit and 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 disease-derived 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-derived 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-derived 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 producing improvements in heritability model fit and in predicting disease-associated, fine-mapped SNPs.

Results

Overview of methods

We define a binary annotation as an assignment of a binary value to each of low-frequency ($0.5\% \leq \text{MAF} < 5\%$) and common ($\text{MAF} \geq 5\%$) SNP in a 1000 Genomes Project European reference panel[28], as in our previous work[25,27]. We define a pathogenicity score as an assignment of a numeric value quantifying predicted pathogenicity, deleteriousness, and/or protein function to some or all of these SNPs; we refer to theses score as Mendelian disease-derived pathogenicity scores, as these scores have predominantly been developed and assessed in the context of Mendelian disease (e.g. using pathogenic variants from ClinVar[29] and HGMD[30]). We analyze 11 Mendelian disease-derived missense scores, 6 genome-wide Mendelian disease-derived 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-derived 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 75 coding, conserved, regulatory and LD-related annotations from the baseline-LD model[26,27] (Figure S1; 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[31] (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. We note that Mendelian disease-derived pathogenicity scores may score a subset of SNPs, but every baseline-LD model annotation scores all SNPs. 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[32] (average $N = 320\text{K}$; 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; annotations with significantly positive or negative τ^* are uniquely informative relative to all other annotations in the model, whereas annotations with $\tau^* = 0$ contain no unique information, even if they are enriched for heritability (see Methods). While S-LDSC models linear combinations of functional annotations, AnnotBoost constructs (linear and) non-linear combinations of baseline-LD model annotations to provide unique information.

Informativeness of Mendelian disease-derived missense scores for common disease

We assessed the informativeness for common disease of binary annotations derived from 11 Mendelian disease-derived pathogenicity scores for missense variants[1,5–8,33–37] (see Table 1). These scores reflect the predicted impact of missense mutations on Mendelian disease; we note that our analyses of common disease are focused on common and low-frequency variants, but these scores were primarily trained using very rare variants from ClinVar[29] and Human Gene Mutation Database (HGMD)[30]. 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[36] (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[36] 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 disease-derived 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[7] score, which spanned too few SNPs to be included in the S-LDSC analysis).

We constructed *boosted* scores from the 11 Mendelian disease-derived 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 disease-derived 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 disease-derived 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 disease-derived missense score (using all 5 thresholds) (baseline-LD+5). We identified conditionally informative binary annotations derived from boosted versions of 10 Mendelian disease-derived missense scores, including 8 previously uninformative scores and the 2 previously informative scores (Figure 1, Table 2 and

Table S2). Letting ↑ denote boosted scores, examples include the top 0.1% of SNPs from M-CAP↑[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↑[8], 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[30] vs. benign variants from ExAC[38] using 9 pathogenicity likelihood scores as features (including PolyPhen-2[1], MetaLR[34], 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, resulting in significantly negative τ^*. These annotations are 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); as noted above, annotations with significantly positive or negative τ^* are uniquely informative. The boosted version of the remaining Mendelian disease-derived 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 5 secondary analyses. First, we restricted the 10 significant binary annotations derived from our boosted Mendelian disease-derived missense scores to non-coding regions, which were previously unscored by the Mendelian disease-derived 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 S2). Third, we examined the trait-specific S-LDSC results, instead of meta-analyzing results across 41 traits. We identified 2 (resp. 32) annotation-trait pairs with significant τ^* values for annotations derived from published (resp. boosted) scores (Table S7); significance was assessed using FDR<5%, as Bonferroni correction would be overly conservative in this case. These annotation-trait pairs spanned 1 (resp. 10) different annotations, all of which were also conditionally significant in the meta-analysis across traits (Figure 1B). Fourth, we assessed the heterogeneity of heritability enrichment and conditional informativeness (τ^*) across the 41 traits (as in ref. 23; see Methods). We determined that 10/20 annotations tested had significant heterogeneity in enrichment and 14/20 annotations tested had significant heterogeneity in τ^*, implying substantial heterogeneity across traits (Table S8). Fifth, we investigated the overlap between genes linked to each of our 12 conditionally significant annotations and 165 gene sets of biological importance, including high-pLI genes[38] and known Mendelian genes[19] (see Methods; Table S9 and Table S10). We linked SNPs to the nearest gene, scored genes based on the maximum pathogenicity score of linked SNPs, and assessed overlap between top gene score quintiles and each of 165 reference gene sets. We consistently observed excess overlap for genes for which homozygous knockout in mice results in lethality[43,44] and high-pLI genes[38], and depleted overlap for olfactory receptor genes[45] (Table S11A-B). In addition, top gene score quintiles derived from our boosted scores often had significantly more overlap (based on Fisher's exact test) than corresponding top gene quintiles derived from published scores (Table S12; results for PolyPhen-2[1,33] shown in Figure S3); gene scores derived from published and boosted pathogenicity scores were moderately correlated (Table S11C-D). This implies that our new annotations can help identify biologically important genes.

We conclude that 2 Mendelian disease-derived missense annotations and 10 boosted annotations are uniquely informative for common disease, relative to baseline-LD model annotations.

Informativeness of genome-wide Mendelian disease-derived pathogenicity scores for common disease

We assessed the informativeness for common disease of binary annotations derived from 6 genome-wide Mendelian disease-derived pathogenicity scores[2–4,9,10] (see Table 1). These scores reflect the predicted impact of (coding and) non-coding variants on Mendelian disease; as above, these scores were primarily trained using very rare variants from ClinVar[29] and HGMD[30]. 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), τ^* = 0.82 (s.e. 0.09)), the top 0.5% of SNPs from CADD[2,46] (enrichment = 18x (s.e. 1.3), τ^* = 0.71 (s.e. 0.10)), and the top 0.1% of SNPs from Eigen[3] (enrichment = 24x (s.e. 2.1), τ^* = 0.40 (s.e. 0.06)) (Figure 2, Table 2 and Table S13). The CADD (Combined Annotation Dependent Depletion) score[2,46] 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 S13).

We applied AnnotBoost to the 6 genome-wide Mendelian disease-derived scores. We observed moderate correlations between input genome-wide Mendelian disease-derived 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 $\geq$ 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 disease-derived score (using all 5 thresholds). We identified conditionally informative binary annotations derived from boosted versions of all 6 genome-wide Mendelian disease-derived 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), τ^* = 0.74 (s.e. 0.10)) and the top 0.5% of SNPs from boosted Eigen-PC↑[3] (enrichment = 16x (s.e. 1.1), τ^* = 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), τ^* = 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[29] and HGMD[30] 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 5 secondary analyses. First, for the 4 genome-wide Mendelian disease-derived scores with <100% of SNPs scored (Table 1), we restricted the binary annotations derived from our boosted genome-wide Mendelian disease-derived 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 S14), 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 S4). Third, we examined the trait-specific S-LDSC results, instead of meta-analyzing results across 41 traits. We identified 11 (resp. 20) annotation-trait pairs with significant τ^* values for annotations derived from published (resp. boosted) scores (FDR<5%; Table S7). These annotation-trait pairs spanned 2 (resp. 4) different annotations, all of which were also conditionally significant in the meta-analysis across traits (Figure 2B). Fourth, we assessed the heterogeneity of heritability enrichment and τ^* across the 41 traits (see Methods). We determined that 10/12 annotations tested had significant heterogeneity in enrichment and 9/12 annotations tested had significant heterogeneity in τ^*, implying substantial heterogeneity across traits (Table S8). Fifth, we investigated the overlap between genes linked to each of our 9 conditionally significant annotations and 165 gene sets of biological importance (see Methods; Table S9). As above, we consistently observed excess overlap for genes for which homozygous knockout in mice results in lethality[43,44] and high-pLI genes[38], and depleted overlap for olfactory receptor genes[45] (Table S11), implying that our new annotations can help identify biologically important genes. For example, the top (bottom) quintile of genes linked to the boosted CADD annotation had 1.7x (0.34x) excess overlap with high-pLI genes[38] vs. excess overlap of 0.8x (0.5x) for the top (bottom) quintile of genes linked to the published CADD[2,46] annotation (OR = 2.4, $P <$ 6e-49) (Figure S3, Table S12).

We conclude that 3 genome-wide Mendelian disease-derived annotations and 6 boosted annotations are uniquely informative for common disease, relative to baseline-LD model annotations.

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 related to Mendelian disease, including 2 constraint-related scores[47,48], 9 scores based on deep learning predictions of epigenetic marks[49–51], and 7 gene-based scores[38,52–54] (see Table S15). 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[55] (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[47] (enrichment = 9.3x (s.e. 0.75), $\tau^* = 0.35$ (s.e. 0.06)) and the top 5% of SNPs from DeepSEA-H3K4me3[49,50] (enrichment = 3.9x (s.e. 0.23), $\tau^* = 0.21$ (s.e. 0.04)) (Figure 3, Table 2 and Table S16). CDTS (Context-Dependent Tolerance Score)[47] is a constraint score based on observed vs. expected variation in whole-genome sequence data; DeepSEA-H3K4me3 scores[49,50] 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 S16).

We applied AnnotBoost to the 18 additional scores, and to the 47 main annotations of the baseline-LD model (Table S15). 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 $\geq$ 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 S16). Examples include the top 10% of SNPs from DeepSEA-DNase↑[49,50] (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↑[48] (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↑[56] (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[48] 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 S16).

We performed 6 secondary analyses. First, for the 31 additional boosted scores which were conditionally 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 S17), implying that Annot-Boost 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 a broad set features contributed to the predictions, including conservation-related features and LD-related features (as above), but also including regulatory features (e.g. H3K4me1, DGF, H3K9ac for boosted DeepSEA↑) (Figure S5). Third, we examined the trait-specific S-LDSC results, instead of meta-analyzing results across 41 traits. We identified 9 (resp. 78) annotation-trait pairs with significant tau^* values for annotations derived from published (resp. boosted) scores (FDR<5%; Table S7). These annotation-trait pairs spanned 5 (resp. 32) different annotations, most of which (3/5 published, 23/32 boosted) were also conditionally significant in the meta-analysis across traits (Figure 3B). Fourth, we assessed the heterogeneity of heritability enrichment and τ^* across the 41 traits (see Methods). We determined that 64/81 annotations tested had significant heterogeneity in enrichment and 40/81 annotations tested had significant heterogeneity in τ^*, implying substantial heterogeneity across traits (Table S8). Fifth, we repeated the analyses of Figure 3 without including the 8 Roadmap annotations. We determined that the number of significant binary annotations increased (Table S18), 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 S19). Sixth, we investigated the overlap between genes linked to each of our 43 conditionally significant annotations and 165 gene sets of biological importance (see Methods; Table S9). As above, we consistently observed excess overlap for genes for which homozygous knockout in mice results in lethality[43,44] and high-pLI genes[38], and depleted overlap for olfactory receptor genes[45] (Figure S3, Table S11, Table S12), implying that our new annotations can help identify biologically important genes.

We conclude that 6 additional genome-wide annotations and 37 boosted annotations are uniquely informative for common disease, relative to baseline-LD model annotations.

Constructing and evaluating combined heritability models

We constructed and evaluated two heritability models incorporating our new functional annotations: (i) a combined marginal model incorporating all binary annotations that were conditionally significant (conditional on the baseline-LD model), and (ii) a combined joint model incorporating only a subset of binary annotations that were jointly and conditionally significant (conditional on each other and the baseline-LD model). We constructed the combined marginal model by merging the baseline-LD model with the 64 conditionally significant annotations (derived from 11 published scores and 53 boosted scores; Table 2) (baseline-LD+marginal). We constructed the combined joint model 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 $\tau^* < 0.25$ (see Methods) (baseline-LD+joint). We have prioritized this type of combined joint model in previous work[26,32,39,40,53], hypothesizing that there would be little value in retaining a large number of annotations containing redundant information; we note the substantial correlations between annotations in this study (Table S20).

The combined joint model included 11 binary annotations derived from 3 published scores and 8 boosted scores (Figure 4, Table 2 and Table S21). 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; as noted above, annotations with significantly positive or negative τ^* are uniquely informative.

We performed two analyses to evaluate the combined joint model and the combined marginal model, compared to the baseline-LD model; these analyses evaluated aggregate heritability models, rather than individual functional annotations. First, we computed the average $logl_{SS}$[57] (an approximate likelihood metric) of each model, relative to a model with no functional annotations, across 30 common diseases and complex traits from the UK Biobank[58] (subset of 41 traits; Table S1) ($\Delta logl_{SS}$; see Methods). Results are reported in Figure 5A and Table S22. The combined joint model attained a +7.4% larger $\Delta logl_{SS}$ than the baseline-LD model ($P <$ 6e-38); the combined marginal model attained a +23.9% larger $\Delta logl_{SS}$ than the baseline-LD model ($P <$ 2e-99), including a significantly larger improvement for 30/30 traits analyzed (Figure 5B and Table S22). The improvements were only slightly smaller when using Akaike Information Criterion (AIC) to account for increases in model complexity[57] (+7.0% and +20.3%; Table S22).

Second, we assessed each model's accuracy of classifying three different sets of fine-mapped SNPs (from 10 LD-, MAF-, and genomic-element-matched control SNPs in the reference panel[28]): 7,333 fine-mapped for 21 autoimmune diseases from Farh et al.[59], 3,768 fine-mapped SNPs for inflammatory bowel disease from Huang et al.[60], and 1,851 fine-mapped SNPs for 49 UK Biobank traits from Weissbrod et al.[61]. We note that with the exception of Weissbrod et al. fine-mapped SNPs (stringently defined by causal posterior probability ≥ 0.95; FDR < 0.05), 95% credible fine-mapped SNPs likely include a large fraction of non-causal variants. We computed the AUPRC attained by the combined joint model and the combined marginal model, relative to a model with no functional annotations (ΔAUPRC), aggregated by training a gradient boosting model (multi-score analysis); 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; specifically, our boosted scores do not use fine-mapped SNPs. The combined joint model attained a +2.5% to 6.9% larger ΔAUPRC than the baseline-LD model (each $P <$ 3e-28); the combined marginal model attained a +4.9% to 21.3% larger ΔAUPRC than the baseline-LD model (each $P <$ 7e-100); we obtained similar results using AUROC (Figure S6, Table S23). This improvement likely comes from non-linear interactions involving the boosted annotations, published annotations, and the baseline-LD model.

We performed 8 secondary analyses. First, we repeated $logl_{SS}$ analysis on the model with 19 new annotations with conditional $\tau^* > 0.5$; we determined this model attained a +10.6% larger $\Delta logl_{SS}$ and +9.7% larger AIC than the baseline-LD model ($P <$ 2e-50) (Table S22). Second, we applied SHAP[41] to investigate which features contributed the most to classification of fine-mapped SNPs; we determined that boosted scores often drove the predictions, validating the potential utility of boosted scores in functionally-informed fine-mapping (e.g. H3K9ac↑, CADD) (Figure S7, Figure S8). Third, we repeated the classification of fine-mapped SNPs using a single LD-, MAF-, and genomic-element-matched control variant (instead of 10 control variants) for each fine-mapped SNP, and obtained similar results (Table S23). Fourth, we repeated the classification of fine-mapped disease SNPs analysis of Weissbrod et al. fine-mapped SNPs using 1,379 SNPs that were fine-mapped without using functional information[61] (to ensure that results were not circular), and obtained similar results (Figure S6, Table S23). Fifth, we computed the AUPRCs for classifying fine-mapped SNPs individually attained by each of 82 published and 82 boosted scores (single-score analysis), comparing results for boosted scores vs. the corresponding published scores. The boosted scores significantly outperformed the corresponding published scores in each case (66/82 to 80/82 scores; Figure S9, Table S24, and Table S25). We also found that AUPRC and AUROC results for published and boosted scores were moderately correlated with S-LDSC results (up to $r = 0.67$) for binary annotations derived from these scores, validating the S-LDSC results (Table S26). Sixth, we repeated the single-score and multi-score analysis using 14,807 NHGRI GWAS SNPs[62,63]; we obtained similar results (Figure S9, Figure S6, Table S24, Table S23). Seventh, we computed genome-wide correlations between all annotations analyzed including baseline-LD model annotations (Table S20). 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 S2, Figure S4 and Figure S5). Eighth, 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 S10 and Table S27), recapitulating the informativeness of 11 jointly significant annotations.

We conclude that the combined joint model and the combined marginal model both significantly outperformed the baseline-LD model, validating the informativeness of our new annotations for common disease. The improvement was much larger for the combined marginal model; this finding was surprising, in light of our previous work advocating for conservatively restricting to jointly significant annotations when expanding heritability models[26,32,39,40,53]. However, we caution that due to the much larger number of new annotations in the combined marginal model, the combined joint model may still be preferred in some settings.

Discussion

We analyzed the informativeness of a broad set of Mendelian disease-derived pathogenicity scores across 41 independent common diseases and complex traits to show that several annotations derived from published Mendelian disease-derived 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 64 marginally significant annotations and 11 jointly significant annotations and implying pervasive variant-level overlap between Mendelian disease and common disease. These variant-level results are substantially different from previous studies of gene-level overlap between Mendelian diseases and complex traits[12–19]. Notably, our new annotations produced significant improvements in heritability model fit and in classifying disease-associated, fine-mapped SNPs. We also detected significant excess overlap between genes linked to our new annotations and biologically important gene sets.

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[29] or HGMD[30]). 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, elucidating specific mechanistic links between Mendelian disease and common disease may yield important biological insights. Second, 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; this would further elucidate the variant-level overlap between Mendelian disease and common disease. Third, annotations derived from published and boosted Mendelian pathogenicity scores can be used to improve functionally informed fine-mapping[61,64–67], motivating their inclusion in future large-scale fine-mapping studies. (On the other hand, we anticipate that our new annotations will be less useful for improving functionally informed polygenic risk prediction[68,69] and association mapping[70], because there is pervasive LD between SNPs in an annotation and SNPs outside of an annotation, such that these annotations do not distinguish which LD blocks contain causal signal.) Fourth, the larger improvement for our combined marginal model versus our combined joint model (Figure 5A) advocates for a more inclusive approach to expanding heritability models, as compared to our previous work advocating for conservatively restricting to jointly significant annotations[26,32,39,40,53]. However, the combined marginal model suffers a cost of reduced interpretability (it contains a much larger number of new annotations, and it is unclear which of these annotations are providing the improvement), thus the combined joint model may still be preferred in some settings. Fifth, gene scores derived from published and boosted Mendelian pathogenicity scores can be used to help identify biologically important genes; we constructed gene scores by linking SNPs to their nearest gene, but better strategies for linking regulatory variants to genes[71–73] could potentially improve upon our results.

We note several limitations of our work. First, we focused our analyses on common disease (which are driven by common and low-frequency variants) and did not analyze Mendelian diseases (which are driven by very rare variants); the application of AnnotBoost to impute and denoise very rare pathogenic variants for Mendelian disease is a direction for future work. Second, we primarily report results that are meta-analyzed across 41 traits (analogous to previous studies[25,26,32,39,40,53]), but results and their interpretation may vary substantially across traits. Nonetheless, our combined marginal model produced a significant improvement in heritability model fit for 30/30 UK Biobank traits analyzed (Figure 5B). Third, 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. Fourth, we restricted all of our analyses to European populations, which have the largest available GWAS sample size. However, we expect our results to be generalizable to other populations, as functional enrichments have been shown to be highly consistent across ancestries[65,74,75]; we note that assessing functional enrichments in admixed populations[76] would require the application of an unpublished extension of S-LDSC[77]. Fifth, the gene-based SNP scores that we analyzed did not perform well, perhaps because they were defined using 100kb windows, a crude strategy employed in previous work[32,53,78]; better strategies for linking regulatory variants to genes[71–73] (as shown in above gene scores) 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. MAF≥0.5%) 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. 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:

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

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[26,32,53], 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:

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

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[25,32,53,78].). 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[79] (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 disease-derived pathogenicity scores

We considered a total 35 published scores: 11 Mendelian disease-derived missense pathogenicity scores, 6 genome-wide Mendelian disease-derived pathogenicity scores, and 18 additional scores (see Table 1 and Table S15). Here, we provide a short description for Mendelian missense and genome-wide Mendelian disease-derived pathogenicity scores. Details for 18 additional scores and the baseline-LD annotations are provided in Table S15. Our curated pathogenicity scores are available online (see URLs).

For all scores, we constructed annotations using GRCh37 (hg19) assembly limited to all 9,997,231 low-frequency and common SNPs (with MAF $\geq 0.5\%$) found in 1000 Genomes[28] European Phase 3 reference genome individuals (see URLs). Mendelian missense scores were readily available from dbNSFP database[80,81] using a rankscore (a converted score based on the rank among scored SNPs); genome-wide Mendelian disease-derived 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 S15) are provided below:

Mendelian disease-derived missense pathogenicity scores:

PolyPhen-2[1,33] *(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[34]: 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[35,82]: 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[30], 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[30] 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[36] *(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[37]: A deep-learning-based score trained on 32,074 pathogenic variants from ClinVar[29], HGMD[30], and UniProt[83], using 38 local context, constraint, conservation, protein structure, gene-based, and existing pathogenicity scores as features.

Genome-wide Mendelian disease-derived pathogenicity scores:

CADD[2,46]: 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[29] and HGMD[30], 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[31] (see URLs). AnnotBoost requires only one input, a pathogenicity score to boost, and generates a genome-wide (probabilistic) pathogenicity score (as described in Figure S1). 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 S11). We note that the baseline-LD annotations considered all low-frequency and common SNPs thus do not have unscored SNPs. 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 S12). As described in ref. 31, 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 because of SHAP's property of satisfying symmetry, dummy player, and additivity axioms. 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 disease-derived 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 disease-derived 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. 53), because S-LDSC does not perform well for small annotations[25]. We analyzed 155 annotations derived from published scores (31 published scores (Table 2), 5 thresholds for top x% of variants, 31*5 = 155), such that 500 hypotheses is a conservative correction in the analysis of published scores. We also analyzed 492 annotations derived from boosted scores (82 underlying published scores including 47 baseline-LD model annotations (Table 2), 6 thresholds for top x% of variants, 82*6 = 492), such that 500 hypotheses is a roughly appropriate correction in the analysis of boosted scores. For simplicity, we corrected for max(155,492) ≈ 500 hypotheses throughout. We note that, in the meta-analysis τ^* p-values, a global FDR $< 5\%$ corresponds to $P < 0.0305$; thus, our choice of $P < 0.05/500 = 0.0001$ is conservative.

In all primary analyses, we analyzed only binary annotations. However, we verified in a secondary analysis of the CDTS score[47] that probabilistic annotations produced results similar to binary annotations (see Figure S13).

Heterogeneity of enrichment and τ^* across traits

For a given annotation, we assessed the heterogeneity of enrichment and τ^* (across 41 independent traits) by estimating the standard deviation of the true parameter value across traits, as described in ref. 23. We calculated the cross-trait τ^* as the inverse variance weighted mean across the traits. Then, we compared $\sum_{i=1}^{n}(\hat{\tau}_i - \hat{\tau}_{across-trait})^2/(std.error_i^2)$ to a χ^2_n null statistic, where n = 47 (41 independent traits; 47 summary statistics; see Table S1). We repeated analysis for heritability enrichment by using enrichments and standard errors of enrichment estimates from S-LDSC.

Overlap between gene score quintiles informed by input pathogenicity scores and 165 reference gene sets

For a given pathogenicity score, we scored genes based on the maximum pathogenicity score of linked SNPs, where SNPs were linked to a unique nearest gene using ANNOVAR[84]: 9,997,231 SNP-gene links, decreasing to 5,059,740 S2G links after restricting to 18,117 genes with a protein product (according to HGNC[85]) that have an Ensembl gene identifier (ENSG ID). Gene scores are reported in Table S9. We constructed quintiles of genes scores and assessed gene-level excess fold overlap with 165 reference gene sets of biological importance (see below; summarized in Table S10). We note that this analysis used continuous-valued pathogenicity scores, instead of binary annotations.

The 165 reference gene sets (Table S10) reflected a broad range of gene essentiality[86] metrics, as outlined in ref. 53. They included known phenotype-specific Mendelian disease genes[19], constrained genes[38,87–89], essential genes[43,44,90], specifically expressed genes across GTEx tissues[78], dosage outlier genes across GTEx tissues[91], genes with a ClinVar pathogenic or likely pathogenic variants[29], genes in the Online Mendelian Inheritance in Man (OMIM[92]), high network connectivity genes in different gene networks[53,93], genes with more independent SNPs[53], known drug targets[94], human targets of FDA-approved drugs[95], eQTL-deficient genes[96,97], and housekeeping genes[98]; a subset of these gene sets were previously analyzed in ref. 53.

As defined in our previous study[53], the excess fold overlap of gene set 1 and gene set 2 is defined as follows:

$$\text{excess overlap(gene set 1, gene set2)} = P_d/P_{tot} \tag{6}$$

where $P_d = \frac{|\text{gene set 1} \cap \text{gene set 2}|}{|\text{gene set 2}|}$ and $P_{tot} = \frac{|\text{gene set 1} \cap \text{all protein-coding genes}|}{|\text{all protein-coding genes}|}$. The standard error for the excess overlap is similarly scaled:

$$SE = \sqrt{\frac{P_d(1-P_d)}{|\text{gene set 2}|}}/P_{tot} \tag{7}$$

When there is excess overlap, the excess fold overlap is > 1; when there is depletion, the excess fold overlap is < 1. We assessed the odds ratio and significance in the difference between the excess overlap between the boosted gene quintile and the published gene quintile by the Fisher's exact test.

Evaluating heritability model fit using $logl_{SS}$

Given a heritability model (e.g. the baseline-LD model, combined joint model, or combined marginal model), we define the $\Delta logl_{SS}$ of that heritability model as the $logl_{SS}$ of that heritability model minus the $logl_{SS}$ of a model with no functional annotations (baseline-LD-nofunct; 17 LD and MAF annotations from the baseline-LD model[26]), where $logl_{SS}$[57] is an approximate likelihood metric that has been shown to be consistent with the exact likelihood from restricted maximum likelihood (REML; see URLs). We compute p-values for $\Delta logl_{SS}$ using the asymptotic distribution of the Likelihood Ratio Test (LRT) statistic: $-2\ logl_{SS}$ follows a χ^2 distribution with degrees of freedom equal to the number of annotations in the focal model, so that $-2\Delta logl_{SS}$ follows a χ^2 distribution with degrees of freedom equal to the difference in number of annotations between the focal model and the baseline-LD-nofunct model. We used UK10K as the LD reference panel and analyzed 4,631,901 HRC (haplotype reference panel[99]) well-imputed SNPs with MAF ≥ 0.01 and INFO ≥ 0.99 in the reference panel; We removed SNPs in the MHC region, SNPs explaining $> 1\%$ of phenotypic variance and SNPs in LD with these SNPs.

We computed $\Delta logl_{SS}$ for 4 heritability models:

- baseline-LD: annotations from the baseline-LD model[25,26] (86 annotations)
- baseline-LD+joint: baseline-LD model + 11 jointly significant annotations (3 published, 8 boosted; 97 annotations)
- baseline-LD+marginal-stringent: baseline-LD model + 19 marginally significant annotations with conditional $\tau^* > 0.5$ (105 annotations)
- baseline-LD+marginal: baseline-LD model + 64 marginally significant annotations (11 published, 53 boosted; 150 annotations)

Classification of fine-mapped disease SNPs

We assessed the classification accuracy of fine-mapped disease SNPs. Here, we consider only low-frequency and common SNPs (MAF $\geq 0.5\%$) and report the total number of unique SNPs (regardless MAF). we assessed the accuracy of classifying five different SNP sets (summarized in Table S24): (1) 7,333 fine-mapped for 21 autoimmune diseases from Farh et al.[59] (of 7,747 total SNPs; 95% credible sets), (2) 3,768 fine-mapped SNPs for inflammatory bowel disease from Huang et al.[60] (of 4,311 total SNPs; 95% credible sets), (3) 1,851 SNPs (of 2,225 SNPs, spanning 3,025 SNP-trait pairs; stringently defined by causal posterior probability ≥ 0.95) functionally-informed fine-mapped for 49 UK Biobank traits from Weissbrod et al.[61], (4) 1,379 (of 1,853 total SNPs with causal posterior probability ≥ 0.95) non-functionally-informed fine-mapped SNPs for 49 UK Biobank traits from Weissbrod et al.[61], and (5) 14,807 SNPs from the NHGRI GWAS catalog[62,63] (2019-07-12 version; p-value $<$ 5e-8; we note only about 5% of GWAS SNPs are expected to be causal[59]).

For each of these five SNP sets, we matched 10 control SNPs for each positive fine-mapped SNP by matching LD, MAF, and genomic element, as in previous studies[9,10,47]; we note that these studies emphasized the need for matching the relative genomic region distribution in performance evaluation. MAF was based on the same reference panel (European samples from 1000 Genomes Phase 3[28]), and LD was estimated by applying S-LDSC on the all SNPs annotation ('base'). To identify the genomic element of each SNPs and the nearest gene, we annotated these five sets of SNPs using ANNOVAR[84] using the gene-based annotation. For assigning the genomic element to each SNP, we used the default ANNOVAR prioritization rule for gene-based annotations: exonic = splicing (defined by 10bp of a splicing junction) > ncRNA > UTR5 = UTR3 > intronic > upstream = downstream > intergenic (see URLs for more detailed definition of each genomic element). When SNP (in the intergenic or intronic region) is associated with overlapping genes, the nearest protein-coding gene (based on the distance to the TSS or TSE) is retained. To obtain 10 control SNPs for each positive fine-mapped SNP, we first searched the control SNPs within the same genomic element and the same chromosome of that positive SNP; then kept the 10 control SNPs with the most similar LD and MAF (based on the average of rank(LD difference from the positive SNP) and rank(MAF difference from the positive SNP)). In secondary analyses, we instead retained a unique most closely matched control SNP.

Given positive and control SNP sets, we computed the AUPRCs (and AUROCs) by an individual score (each of 82 published and 82 boosted scores). We refer this as a single-score analysis. We used AUPRC as a primary metric, as AUPRC is more robust for imbalanced data[100]. We assessed the significance of the difference between two AUPRCs using 1,000 samples bootstrapped standard errors then performed 2-sample t-test; variance of AUPRCs (and AUROCs) from 1,000 samples was sufficiently small. We note this single-score analysis measures an improvement between two scores, where one score is derived from the other (e.g. our boosted score from published score). Also, we performed a multi-score analysis. For each heritability model, we aggregated scores by training a gradient boosting model (features: aggregated scores, positive label: each of five sets of SNPs, control label: LD-, MAF-, and genomic-element-matched control 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 AUPRC and AUROC of odd and even chromosome classifiers. We also computed ΔAUPRC as AUPRC of a given model minus AUPRC of baseline-LD-nofunct model. We note that no disease data (five 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 as described above. To identify which feature(s) drives the prediction output, we applied SHAP[41] to generate a signed impact of each baseline-LD, published, and boosted score features on classifying fine-mapped disease SNPs.

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

SumHer software for computing $logl_{SS}$: http://dougspeed.com/sumher/

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: https://github.com/RausellLab/NCBoost

ncER scores: https://github.com/TelentiLab/ncER_datasets

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. 51.

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

LIMBR scores: Table S1 in ref. 52.

Saha, Greene, InWeb, Sonawane network annotations:

https://data.broadinstitute.org/alkesgroup/LDSCORE/Kim_pathwaynetwork/

EDS scores: Table S1 in ref. 54.

165 reference gene sets: https://github.com/samskim/networkconnectivity

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

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

ANNOVAR: `http://annovar.openbioinformatics.org/`

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

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, Armin Schoech, and Huwenbo Shi for helpful discussions. This research was funded by NIH grants U01 HG009379, U01 MH119509, R01 MH101244, R37 MH107649, and R01 MH109978. S.S.K. was supported by the National Human Genome Research Institute of the NIH under award number F31HG010818. 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. 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.

30. 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.

31. 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.

32. 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.

33. 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.

34. 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.

35. 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.

36. 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.

37. 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.

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. (2020). 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. Blake, J. A., Bult, C. J., Kadin, J. A., Richardson, J. E., Eppig, J. T., and Group, M. G. D. (2011). The mouse genome database (mgd): premier model organism resource for mammalian genomics and genetics. Nucleic Acids Research *39*, D842–D848.

44. Georgi, B., Voight, B. F., and Bućan, M. (2013). From mouse to human: evolutionary genomics analysis of human orthologs of essential genes. PLoS Genetics *9*, e1003484.

45. Mainland, J. D., Li, Y. R., Zhou, T., Liu, W. L. L., and Matsunami, H. (2015). Human olfactory receptor responses to odorants. Scientific Data *2*, 150002.

46. 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.

47. 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.

48. 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.

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

50. 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.

51. 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.

52. 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.

53. 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.

54. Wang, X. and Goldstein, D. B. (2020). Enhancer domains predict gene pathogenicity and inform gene discovery in complex disease. The American Journal of Human Genetics *106*, 215–233.

55. 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.

56. 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.

57. Speed, D., Holmes, J., and Balding, D. J. (2020). Evaluating and improving heritability models using summary statistics. Nature Genetics *52*, 458–462.

58. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The uk biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209.

59. 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.

60. Huang, H., Fang, M., Jostins, L., Mirkov, M. U., Boucher, G., Anderson, C. A., Andersen, V., Cleynen, I., Cortes, A., Crins, F., et al. (2017). Fine-mapping inflammatory bowel disease loci to single-variant resolution. Nature *547*, 173.

61. 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. (2020). Functionally-informed fine-mapping and polygenic localization of complex trait heritability. BioRxiv pp. 807792.

62. 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.

63. 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.

64. 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.

65. Kichaev, G. and Pasaniuc, B. (2015). Leveraging functional-annotation data in trans-ethnic fine-mapping studies. The American Journal of Human Genetics *97*, 260–271.

66. 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.

67. Kichaev, G., Roytman, M., Johnson, R., Eskin, E., Lindstroem, S., Kraft, P., and Pasaniuc, B. (2017). Improved methods for multi-trait fine mapping of pleiotropic risk loci. Bioinformatics *33*, 248–255.

68. 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.

69. Marquez-Luna, C., Gazal, S., Loh, P.-R., Kim, S. S., Furlotte, N., Auton, A., Price, A. L., 23andMe Research Team, et al. (2020). Ldpred-funct: incorporating functional priors improves polygenic prediction accuracy in uk biobank and 23andme data sets. bioRxiv.

70. 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.

71. Zeggini, E., Gloyn, A. L., Barton, A. C., and Wain, L. V. (2019). Translational genomics and precision medicine: Moving from the lab to the clinic. Science *365*, 1409–1413.

72. 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.

73. 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.

74. Kanai, M., Akiyama, M., Takahashi, A., Matoba, N., Momozawa, Y., Ikeda, M., Iwata, N., Ikegawa, S., Hirata, M., Matsuda, K., et al. (2018). Genetic analysis of quantitative traits in the japanese population links cell types to complex human diseases. Nature genetics *50*, 390–400.

75. Lam, M., Chen, C.-Y., Li, Z., Martin, A. R., Bryois, J., Ma, X., Gaspar, H., Ikeda, M., Benyamin, B., Brown, B. C., et al. (2019). Comparative genetic architectures of schizophrenia in east asian and european populations. Nature genetics *51*, 1670–1678.

76. Seldin, M. F., Pasaniuc, B., and Price, A. L. (2011). New approaches to disease mapping in admixed populations. Nature Reviews Genetics *12*, 523–528.

77. Luo, Y., Li, X., Wang, X., Gazal, S., Mercader, J. M., Neale, B. M., Florez, J. C., Auton, A., Price, A. L., Finucane, H. K., et al. (2019). Estimating heritability and its enrichment in tissue-specific gene sets in admixed populations. bioRxiv pp. 503144.

78. 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.

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

80. 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.

81. 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.

82. 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.

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

84. Wang, K., Li, M., and Hakonarson, H. (2010). Annovar: functional annotation of genetic variants from high-throughput sequencing data. Nucleic acids research *38*, e164–e164.

85. Povey, S., Lovering, R., Bruford, E., Wright, M., Lush, M., and Wain, H. (2001). The hugo gene nomenclature committee (hgnc). Human genetics *109*, 678–680.

86. Bartha, I., di Iulio, J., Venter, J. C., and Telenti, A. (2018). Human gene essentiality. Nature Reviews Genetics *19*, 51.

87. Samocha, K. E., Robinson, E. B., Sanders, S. J., Stevens, C., Sabo, A., McGrath, L. M., Kosmicki, J. A., Rehnström, K., Mallick, S., Kirby, A., et al. (2014). A framework for the interpretation of de novo mutation in human disease. Nature Genetics *46*, 944.

88. Bartha, I., Rausell, A., McLaren, P. J., Mohammadi, P., Tardaguila, M., Chaturvedi, N., Fellay, J., and Telenti, A. (2015). The characteristics of heterozygous protein truncating variants in the human genome. PLoS Computational Biology *11*, e1004647.

89. Cassa, C. A., Weghorn, D., Balick, D. J., Jordan, D. M., Nusinow, D., Samocha, K. E., O'Donnell-Luria, A., MacArthur, D. G., Daly, M. J., Beier, D. R., et al. (2017). Estimating the selective effects of heterozygous protein-truncating variants from human exome data. Nature Genetics *49*, 806.

90. Hart, T., Chandrashekhar, M., Aregger, M., Steinhart, Z., Brown, K. R., MacLeod, G., Mis, M., Zimmermann, M., Fradet-Turcotte, A., Sun, S., et al. (2015). High-resolution crispr screens reveal fitness genes and genotype-specific cancer liabilities. Cell *163*, 1515–1526.

91. Mohammadi, P., Castel, S. E., Cummings, B. B., Einson, J., Sousa, C., Hoffman, P., Donkervoort, S., Jiang, Z., Mohassel, P., Foley, A. R., et al. (2019). Genetic regulatory variation in populations informs transcriptome analysis in rare disease. Science *366*, 351–356.

92. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., and McKusick, V. A. (2005). Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. Nucleic Acids Research *33*, D514–D517.

93. Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., et al. (2019). String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic acids research *47*, D607–D613.

94. Nelson, M. R., Wegmann, D., Ehm, M. G., Kessner, D., Jean, P. S., Verzilli, C., Shen, J., Tang, Z., Bacanu, S.-A., Fraser, D., et al. (2012). An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science *337*, 100–104.

95. Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., et al. (2017). Drugbank 5.0: a major update to the drugbank database for 2018. Nucleic Acids Research *46*, D1074–D1082.

96. Consortium, G. et al. (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204.

97. Aguet, F., Barbeira, A. N., Bonazzola, R., Brown, A., Castel, S. E., Jo, B., Kasela, S., Kim-Hellmuth, S., Liang, Y., Oliva, M., et al. (2019). The gtex consortium atlas of genetic regulatory effects across human tissues. BioRxiv pp. 787903.

98. Eisenberg, E. and Levanon, E. Y. (2013). Human housekeeping genes, revisited. TRENDS in Genetics *29*, 569–574.

99. McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., Kang, H. M., Fuchsberger, C., Danecek, P., Sharp, K., et al. (2016). A reference panel of 64,976 haplotypes for genotype imputation. Nature genetics *48*, 1279–1283.

100. Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and roc curves. In Proceedings of the 23rd international conference on Machine learning pp. 233–240.

101. 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.

102. 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.

103. 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, 33
PolyPhen-2-HVAR	Impact of missense variants using protein sequence and structure using HumVar	0.28%	1, 33
MetaLR	Deleterious missense mutations using ensemble scoring (logistic regression)	0.32%	34
MetaSVM	Deleterious missense mutations using ensemble scoring (support vector machine)	0.32%	34
PROVEAN	Impact of an amino acid change on protein function	0.31%	35, 82
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%	36
MVP	Impact of missense variants using deep neural networks	0.29%	37
CADD	Predicted deleterious variants using ensemble scoring	100%	2, 46
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 disease-derived missense and 6 genome-wide Mendelian disease-derived pathogenicity scores. For each of 17 Mendelian disease-derived 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) (baseline-LD+marginal 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) (baseline-LD+joint model). *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

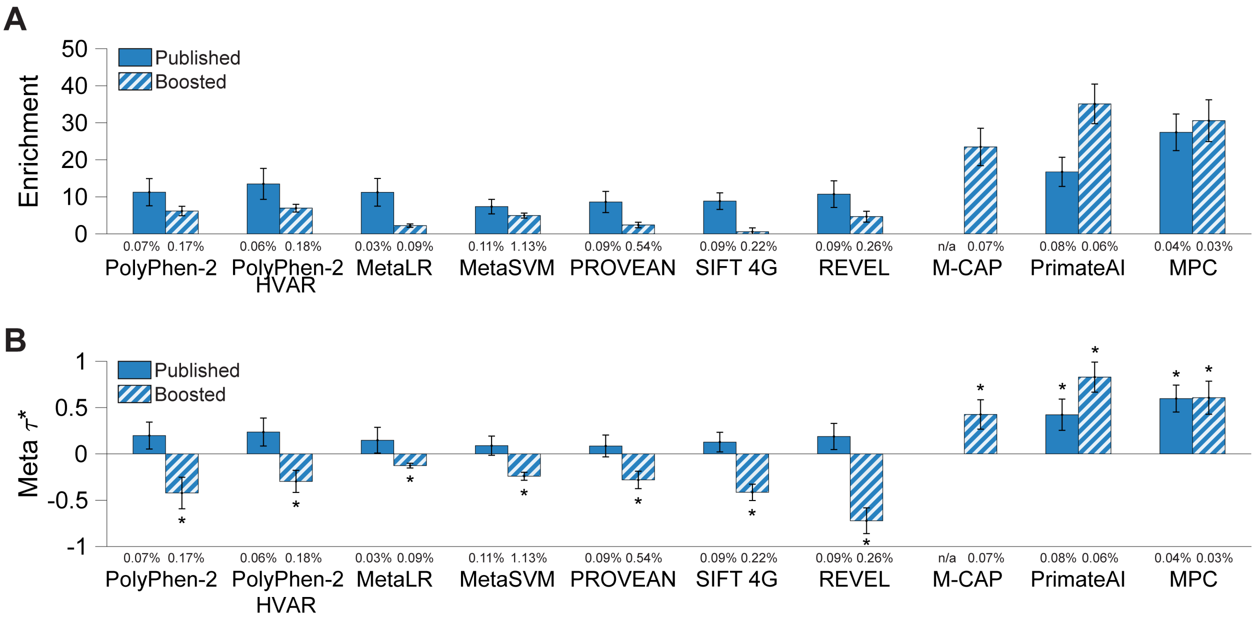


Figure 1. Informativeness for common disease of binary annotations derived from 11 Mendelian disease-derived missense scores and corresponding boosted scores. We report (A) heritability enrichment of binary annotations derived from published and boosted Mendelian disease-derived missense scores, meta-analyzed across 41 independent traits; (B) conditional τ^* values, conditioning 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 disease-derived missense scores (of 11 analyzed) for which annotations derived from published and/or boosted scores were conditionally 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 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 S25.

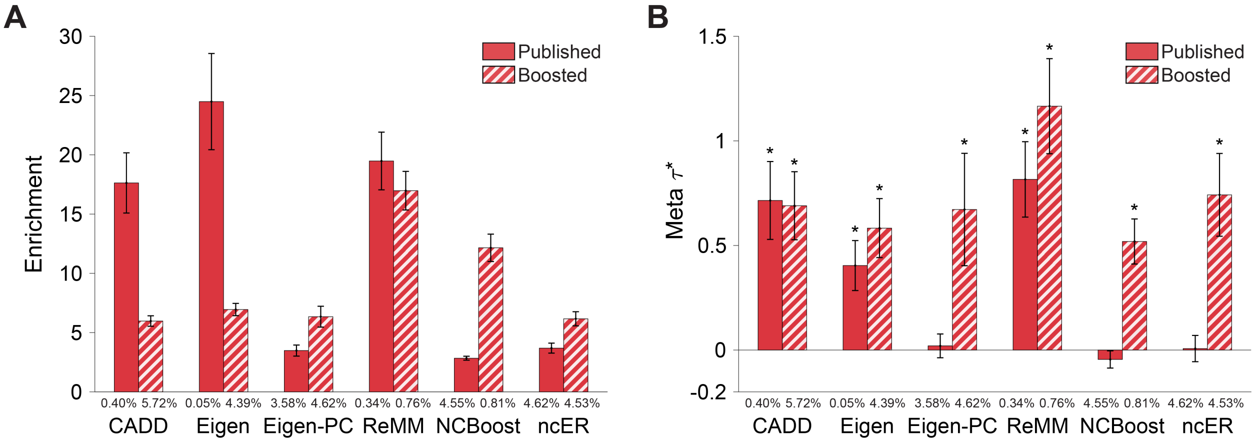


Figure 2. Informativeness for common disease of binary annotations derived from 6 genome-wide Mendelian disease-derived scores and corresponding boosted scores. We report (A) heritability enrichment of binary annotations derived from published and boosted genome-wide Mendelian disease-derived scores, meta-analyzed across 41 independent traits; (B) conditional τ^* values, conditioning 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 disease-derived scores (of 6 analyzed) for which annotations derived from published and/or boosted scores were conditionally 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 S13. Results for standardized enrichment, defined as enrichment times the standard deviation of annotation value (to adjust for annotation size), are reported in Table S25.

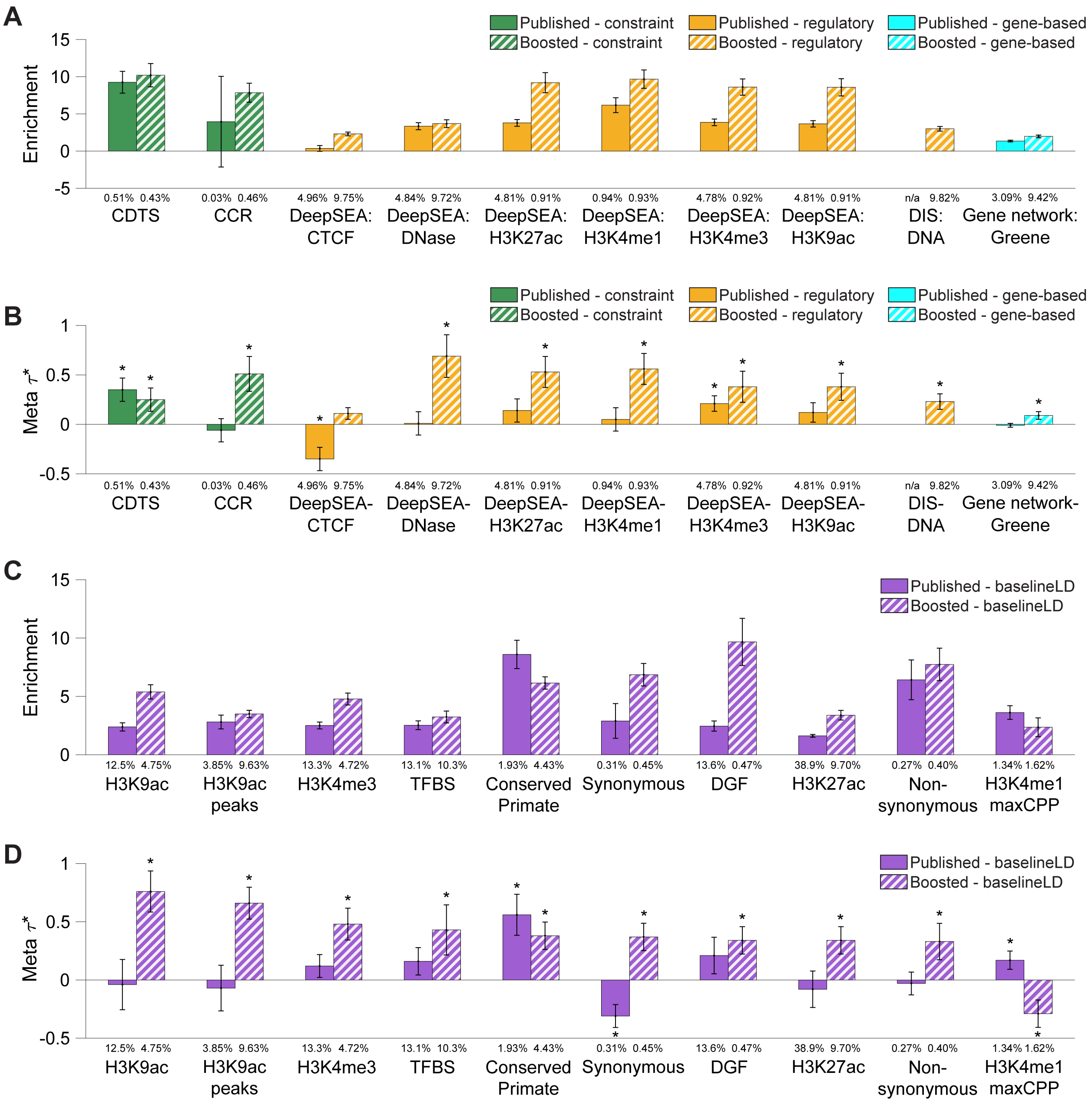


Figure 3. Informativeness for common disease of binary annotations derived from 18 additional genome-wide scores + 47 baseline-LD model annotations and corresponding boosted scores. We report (A) heritability enrichments of binary annotations derived from published and boosted additional genome-wide scores, meta-analyzed across 41 independent traits; (B) conditional τ^* values, conditioning 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) conditional 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 conditionally significant annotations. Numerical results are reported in Table S16. Results for standardized enrichment, defined as enrichment times the standard deviation of annotation value (to adjust for annotation size), are reported in Table S25.

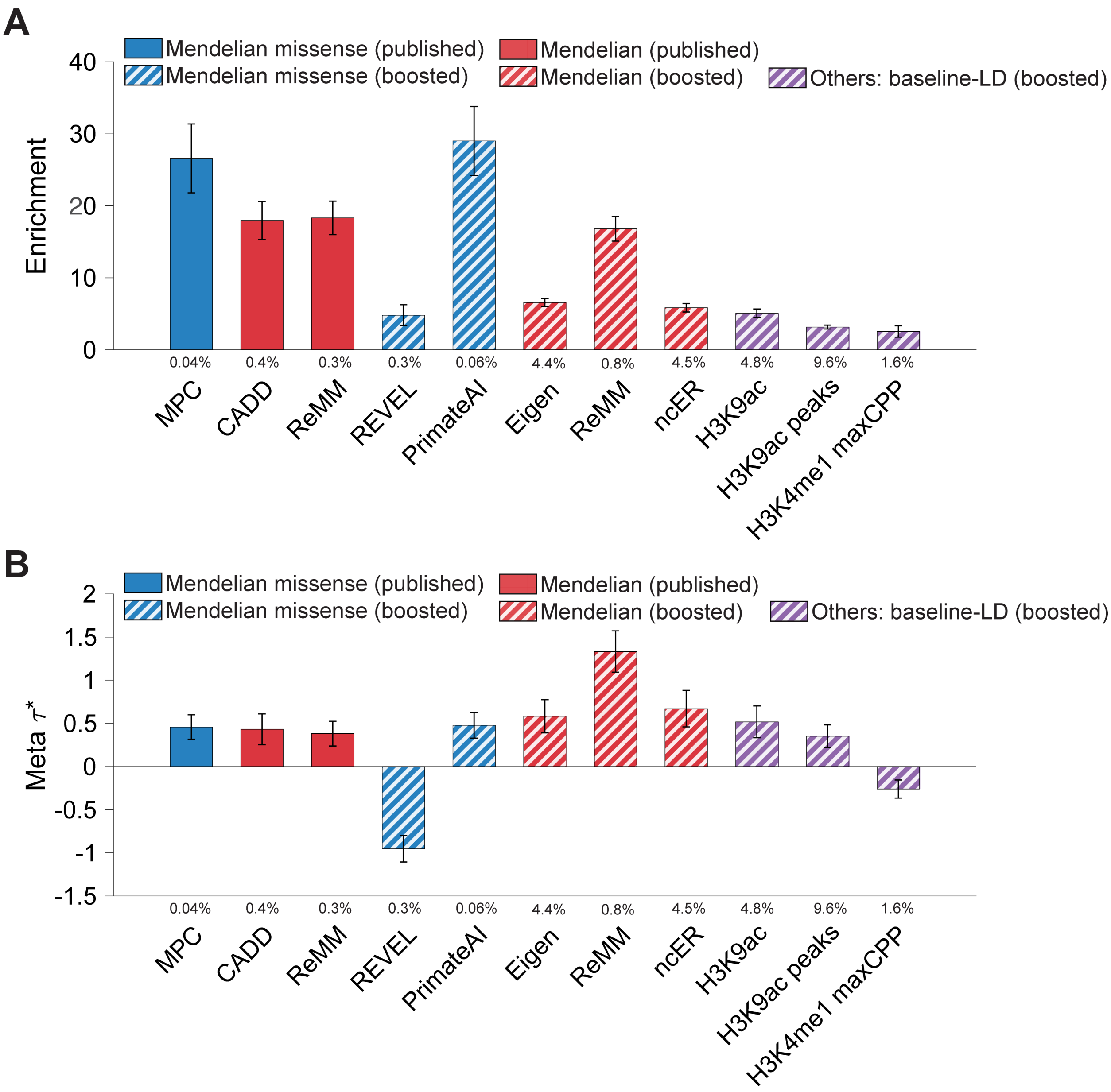


Figure 4. Informativeness for common disease of 11 jointly significant binary annotations from combined joint model. We report (A) heritability enrichment of 11 jointly significant binary annotations, meta-analyzed across 41 independent traits; (B) joint τ^* values, conditioned on the baseline-LD model, 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 S21. Results for standardized enrichment, defined as enrichment times the standard deviation of annotation value (to adjust for annotation size), are reported in Table S25.

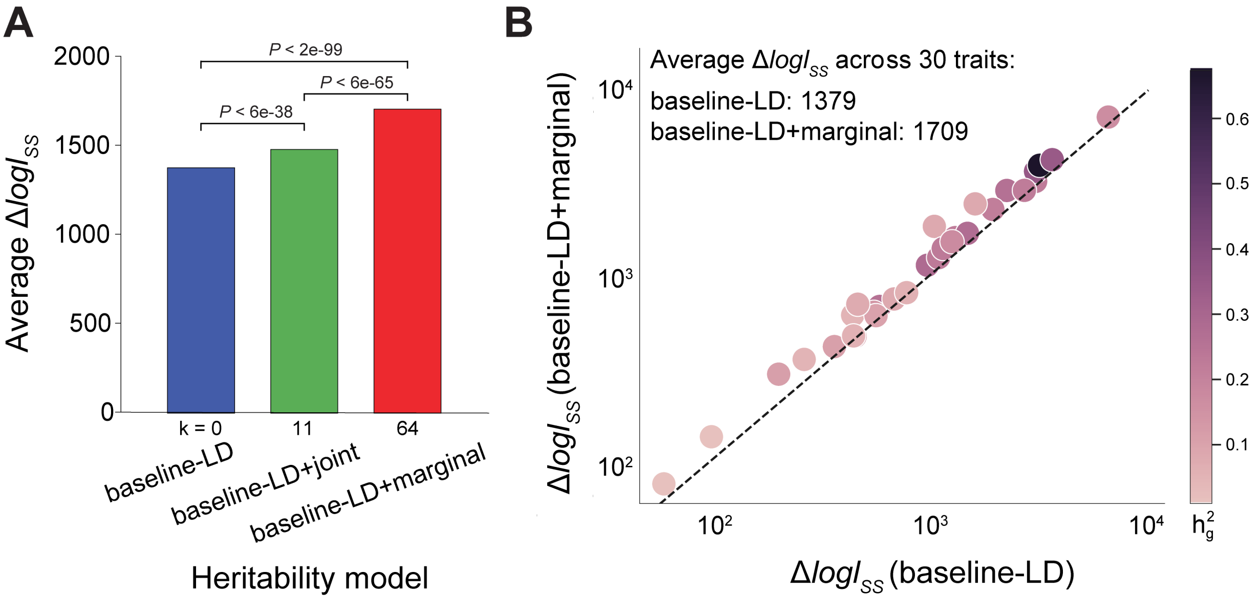


Figure 5. Evaluation of improvement in heritability model fit. We report (A) average $\Delta logl_{SS}$ (an approximate model likelihood metric[57]) across 30 UKBB traits; (B) $\Delta logl_{SS}$ of the baseline-LD and baseline-LD+marginal models for each trait. $\Delta logl_{SS}$ is computed as $logl_{SS}$ of a given model - $logl_{SS}$ of a model with no functional annotations (baseline-LD-nofunct model: MAF/LD annotations only). In panel (A), k denotes the number of new annotations beyond the baseline-LD model. Numerical results are reported in Table S22.

Supplementary tables

See Excel file for all supplementary tables. Titles and captions are provided below.

Table S1. List of 41 independent diseases and complex traits analyzed. Analogous to a previous study[32], we considered 89 GWAS summary association statistics, including 34 traits from publicly available sources and 55 traits from the UK Biobank (up to N = 459K); summary association statistics were computed using BOLT-LMM v2.3[101,102]. We obtained 41 independent traits (average N = 320K) with genetic correlation less than 0.9 (computed using cross-trait LDSC[103]). For 6 traits, we analyzed two different sources (both publicly available and UK Biobank), resulting in total 47 summary statistics analyzed. For each trait, we report a trait identifier, trait description, reference, sample size, and heritability z-score.

Table S2. Informativeness for common disease of binary annotations derived from 11 Mendelian disease-derived missense scores and corresponding boosted scores. We applied S-LDSC, 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) and meta-analyzed results across 41 independent traits. We report meta-analyzed enrichments and τ^*.

Table S3. Correlations between published scores and genome-wide boosted scores We report Pearson correlation (r) between published scores and genome-wide boosted scores for (A) Mendelian disease-derived missense scores, (B) genome-wide Mendelian disease-derived scores, (C) other scores, and (D) all scores. We regard originally unscored SNPs as having a score of zero.

Table S4. Predictive accuracy of AnnotBoost in out-of-sample predictions of input published scores. We evaluated the AnnotBoost model classification performance based on AUROC and AUPRC on a 20% held-out testing set (consisted of odd (resp. even) chromosome SNPs) through the 5-fold cross-validation. For each boosted score, we reported the average AUROCs of odd-chr and even-chr model.

Table S5. Expected heritability enrichments of binary annotations derived from boosted Mendelian disease-derived missense scores with significantly negative τ^*. We report the heritability enrichment that is expected based on the boosted annotation's overlap with the baseline-LD model and corresponding published annotations, by assuming that the τ of the annotation is zero. In general, expected enrichments were significantly less than observed enrichments.

Table S6. Informativeness for common disease of binary annotations derived from boosted Mendelian disease-derived missense scores, restricted to non-coding regions. We restricted boosted Mendelian missense annotations (whose top X% variants are already optimized) to non-coding SNPs and applied S-LDSC (conditioning on the baseline-LD model and corresponding published annotations). We report meta-analyzed enrichments, τ, and τ^* across 41 independent traits; for meta-analysis, we used the same weights from the meta-analysis of the unrestricted boosted annotations.

Table S7. Trait-specific informativeness for common disease of binary annotations. Instead of meta-analyzing S-LDSC results across traits, we report the trait-specific results for 5,311 annotation-trait pairs (113 annotations x 47 traits; 113 annotations whose proportions of top SNPs were optimized). We indicated 152 annotation-trait pairs that were significant at FDR $< 5\%$ on τ^* p-value.

Table S8. Heterogeneity of heritability enrichment and τ^* across traits. We report the heterogeneity p-value of heritability enrichment and τ^* across 41 traits, for each of annotations analyzed.

Table S9. Gene scores derived from published and boosted pathogenicity scores. We report genes scores (maximum pathogenicity score of linked SNPs) for genes linked to each of 203 scores analyzed (consisting of 35 published, 82 boosted, 86 baseline-LD annotations).

Table S10. 165 reference gene sets of biological importance. We report (A) gene set's short description, number of genes, and reference; (B) ENSG ID for each of 165 gene sets (limited to protein-coding genes).

Table S11. Excess overlap of 165 reference gene sets in each quintile bin of gene scores derived from input pathogenicity scores. We report (A) the excess fold overlap of genes in each quintile bin of both published and boosted gene scores; (B) the standard error of the excess overlap; (C) the Pearson correlation (r) among the gene scores; and (D) the Spearman correlation (ρ) among the gene scores.

Table S12. Odds ratios for fold overlaps by gene quintiles from boosted scores and corresponding gene quintiles from published scores in 165 reference gene sets. We computed the significance of the association between gene score quintiles from each of 82 published-boosted score pairs using Fisher's exact test. We report the odds ratios and p-value for (A) fifth quintile (top 20%) gene sets, (B) fourth quintile gene sets, (C) third quintile gene sets, (D) second quintile gene sets, and (E) first quintile (bottom 20%) gene sets.

Table S13. Informativeness for common disease of binary annotations derived from 6 genome-wide Mendelian disease-derived scores and corresponding boosted scores. We applied S-LDSC, 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) and meta-analyzed results across 41 independent traits. We report meta-analyzed enrichments and τ^*.

Table S14. Informativeness for common disease of binary annotations derived from boosted genome-wide Mendelian pathogenicity scores, restricted to previously unscored variants. We restricted boosted Mendelian annotations (whose coverage is $< 100\%$; see Table 1) to previously unscored variants and applied S-LDSC (conditioning on the baseline-LD model and corresponding published annotations). We report meta-analyzed enrichments, τ, and τ^* across 41 independent traits; for meta-analysis, we used the same weights from the meta-analysis of the unrestricted boosted annotations.

Table S15. Summary of 18 additional genome-wide scores and 47 baseline-LD model annotations. We provide a description of 18 additional genome-wide scores and 47 baseline-LD model annotations. These include 2 constraint-related annotations, 9 predicted regulatory annotations (7 epigenetic states and 2 predicted transcriptional/post-transcriptional regulatory effects), 7 gene-based annotations, and 47 annotations from the baseline-LD model. We report the coverage, out of 9,997,231 common SNPs we considered; we note that the baseline-LD model annotations consider genome-wide SNPs and do not have any missingness (see Methods). For binary annotations, the proportion of SNPs with value 1 is considered as coverage. Annotations are first ordered by the type and then by the year of publication. base (all SNPs), 10 MAF bins, and 6 LD-related annotations constitute baseline-LD-nofunct model (model with no functional annotations).

Table S16. Informativeness for common disease of binary annotations derived from 18 additional genome-wide scores + 47 baseline-LD model annotations and corresponding boosted scores. We applied S-LDSC, 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) and meta-analyzed results across 41 independent traits. We report meta-analyzed enrichments and τ^*.

Table S17. Informativeness for common disease of binary annotations derived from boosted versions of additional genome-wide scores and baseline-LD model annotations, restricted to previously unscored variants. We restricted boosted additional annotations (whose coverage is < 100%; see Table S15) to previously unscored variants and applied S-LDSC (conditioning on the baseline-LD model and corresponding published annotations). We report meta-analyzed enrichments, τ, and τ^* across 41 independent traits; for meta-analysis, we used the same weights from the meta-analysis of the unrestricted boosted annotations.

Table S18. Informativeness for common disease of binary annotations derived from 18 additional genome-wide scores + 47 baseline-LD model annotations and corresponding boosted scores without conditioning on 8 Roadmap annotations. We applied S-LDSC, 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) and meta-analyzed results across 41 independent traits. We report meta-analyzed enrichments and τ^*.

Table S19. Informativeness for common disease of marginally conditionally significant annotations derived from Mendelian disease-derived missense scores and genome-wide Mendelian disease-derived scores and corresponding boosted scores, when additionally conditioning on 8 Roadmap annotations. We applied S-LDSC to marginally significant Mendelian missense and genome-wide Mendelian annotations, conditioning on the baseline-LD model, 8 Roadmap annotations, and published annotations (for boosted annotations), and meta-analyzed results across 41 independent traits. We report meta-analyzed enrichments and τ^*.

Table S20. Genome-wide correlations among functional annotations. We report Pearson correlation (r) (A) among 11 jointly significant annotations, (B) between 11 jointly significant annotations and functional annotations from the baseline-LD model, and (C) among all 198 annotations (85 baseline-LD annotations (excluding 'base' annotation; all SNPs annotated with 1), 31 annotations derived from published scores, and 82 boosted annotations).

Table S21. Informativeness for common disease of 11 jointly significant binary annotations in a combined joint model. We applied S-LDSC, conditioned on the baseline-LD model, 8 Roadmap annotations, and each other, and meta-analyzed results across 41 independent traits. We report proportion of SNPs, meta-analyzed enrichments, and τ^*.

Table S22. Evaluation of heritability model fit using $logl_{SS}$. We report $logl_{SS}$, $\Delta logl_{SS}$, and AIC for GCTA (number of annotations = 1), baseline-LD-nofunct (17), baseline-LD (86), baseline-LD+joint (97), baseline-LD+marginal$\tau^* \geq$0.5 (105), and baseline-LD+marginal model (150), after applying SumHer[57] across 30 UKBB summary statistics. $\Delta logl_{SS}$ is computed as $logl_{SS}$ of a given model - $logl_{SS}$ of baseline-LD-nofunct model. AIC (to account for increase in model complexity) is computed as 2*number of annotations - $2logl_{SS}$, as described in ref. 57. We also report the p-value on the differences in $\Delta logl_{SS}$.

Table S23. Classification of fine-mapped disease SNPs: multi-score analysis. We report AUROCs and AUPRCs of different combined scores in classifying 5 different independent SNP sets: 7,333 fine-mapped for 21 autoimmune diseases from Farh et al.[59], 3,768 fine-mapped SNPs for inflammatory bowel disease from Huang et al.[60], 1,851 fine-mapped SNPs for 47 traits from UK Biobank[61] (stringently defined by causal posterior probability $\geq$ 0.95; both functionally-informed and non functionally-informed), and 14,807 GWAS significant SNPs[62,63] (from 10 LD-, MAF-, and genomic element-matched control SNPs as well as the most matched low frequency and common SNPs in the reference panel[28]).

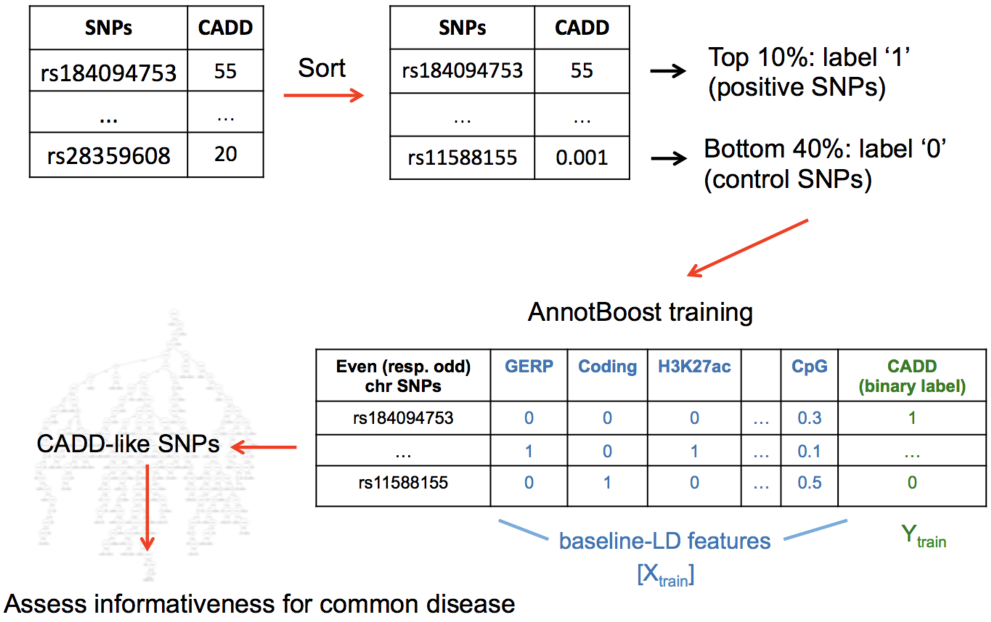
Table S24. Summary of fine-mapped SNPs analyzed and results of single-score classification analysis. We report the reference SNP ID (RSID) of four fine-mapped SNP sets and GWAS significant SNPs and their corresponding control SNPs that are found in the reference panel (MAF $\geq$ 0.5%). We also report their short description and the number of unique SNPs of 10 matched control SNPs and most matched control SNPs. We also report the summary of single-score classification analysis, including the number of scores, where boosted scores perform better than the corresponding published scores, and the average AUROC and AUPRC improvement. Numeric results for the single-score analysis can be found in Table S25.

Table S25. Classification of fine-mapped disease SNPs: single-score analysis. We report AUROCs and AUPRCs of published and boosted scores in classifying 5 different independent SNP sets: 7,333 fine-mapped for 21 autoimmune diseases from Farh et al.[59], 3,768 fine-mapped SNPs for inflammatory bowel disease from Huang et al.[60], 1,851 fine-mapped SNPs for 47 traits from UK Biobank[61] (stringently defined by causal posterior probability $\geq$ 0.95; both functionally-informed and non functionally-informed), and 14,807 GWAS significant SNPs[62,63] ((A) from 10 LD-, MAF-, and genomic element-matched control SNPs as well as (B) the most matched low frequency and common SNPs in the reference panel[28]). We report the S-LDSC results (including standardized enrichments) of derived annotations based on top prioritized SNPs side by side. The summary of single-score analysis can be found in Table S24.

Table S26. Correlation between S-LDSC metrics and AUROCs from single-score analysis of classifying fine-mapped SNPs. For (A) 75 published annotations (excluding small annotations with proportion of SNPs $< 0.02\%$) and (B) 82 boosted annotations, we computed Pearson correlation (r) between different S-LDSC metrics (standardized enrichment, τ^*) and AUROCs on classifying fine-mapped SNPs. We note AUROCs are computed on all SNPs, and S-LDSC results correspond to results on top prioritized SNPs. We further note it is 75 published annotations, not 82, because 7 scores were excluded in S-LDSC analysis due to small annotation size for 4 scores and not comparable enrichments for 3 continuous-valued baseline-LD annotations.

Table S27. Informativeness of the baseline-LD model before and after adding 11 jointly significant binary annotations We applied S-LDSC to (A) the baseline-LD model + 8 Roadmap annotations and (B) the baseline-LD model + 8 Roadmap annotations + 11 jointly significant annotations and meta-analyzed results across 41 independent traits. We report proportion of SNPs, meta-analyzed enrichments, and τ^*.

Supplementary figures



SNPs	CADD
rs184094753	55
...	...
rs28359608	20

SNPs	CADD
rs184094753	55
...	...
rs11588155	0.001

Even (resp. odd) chr SNPs	GERP	Coding	H3K27ac		CpG	CADD (binary label)
rs184094753	0	0	0	...	0.3	1
...	1	0	1	...	0.1	...
rs11588155	0	1	0	...	0.5	0

Sort
Top 10%: label '1' (positive SNPs)
Bottom 40%: label '0' (control SNPs)
AnnotBoost training
CADD-like SNPs
baseline-LD features
$[X_{train}]$
Y_{train}
Assess informativeness for common disease

Figure S1. Overview of AnnotBoost framework. We describe the AnnotBoost model training. AnnotBoost requires only one input, a pathogenicity score to boost, and generates a genome-wide (probabilistic) boosted pathogenicity score. From the input pathogenicity score (e.g. CADD as shown here), we built a classification model, each for even and odd chromosome SNPs using 75 baseline-LD annotations[25,26] as features. We assessed informativeness of annotations derived from published scores (input) and boosted scores (output) for common disease using S-LDSC[78].

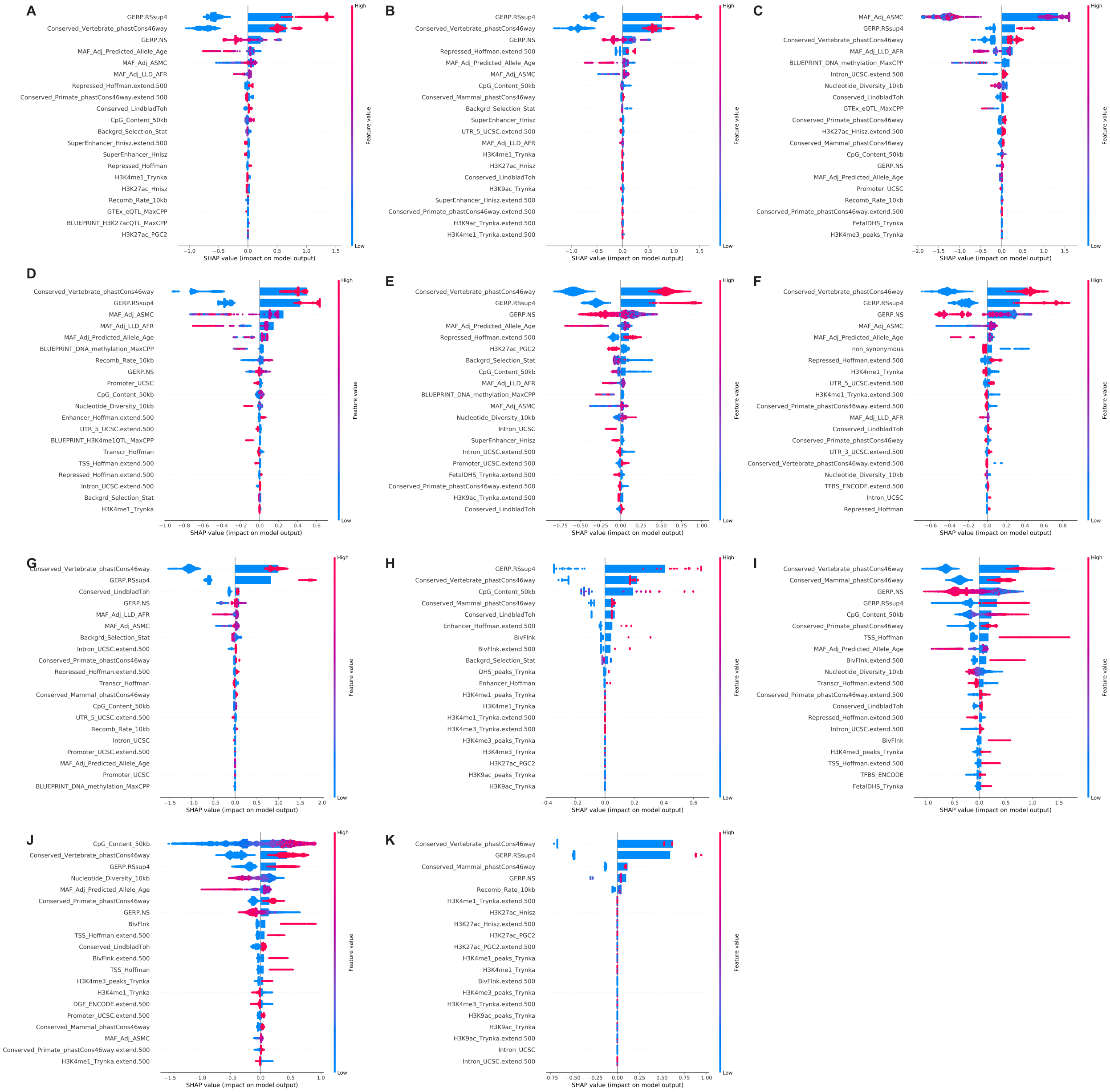


Figure S2. Feature importance of boosted Mendelian disease-derived missense pathogenicity scores. We applied SHAP[41] to assess which features from the baseline-LD drives the prediction of 11 boosted missense scores by AnnotBoost. We report the signed impact of top 20 features for each of 11 predictive models: (A) PolyPhen-2, (B) PolyPhen-2-HVAR, (C) MetaLR, (D) MetaSVM, (E) PROVEAN, (F) SIFT 4G, (G) REVEL, (H) M-CAP, (I) Primate-AI, (J) MPC, and (K) MVP. We obtained similar results for even/odd chromosome classifiers; we report odd chromosome results here (see full results online; see URLs).

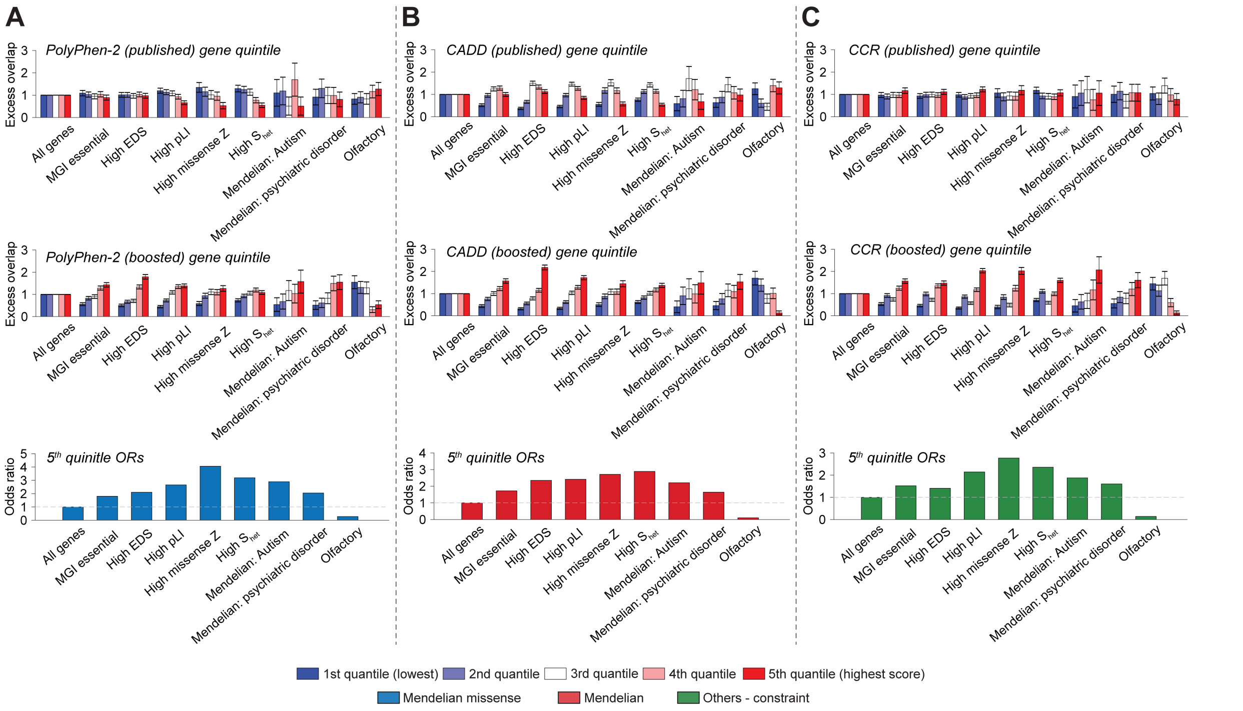


Figure S3. Excess overlap between gene scores derived from input pathogenicity scores and 165 reference gene sets of biological importance. We report the excess overlap of genes linked to published and boosted scores in existing gene sets of biological importance (summarized in Table S10): (A) PolyPhen-2[1,33] gene quintiles from published and boosted scores, (B) CADD[2,46] gene quintiles from published and boosted scores, and (C) CCR[48] gene quintiles from published and boosted scores. Error bars represent 95% confidence intervals. Numeric results for excess overlap and correlaton among gene scores are shown in Table S11. Numeric results for odds ratios and p-values from Fisher's exact test between published gene quintiles and boosted gene quintiles are reported in Table S12.

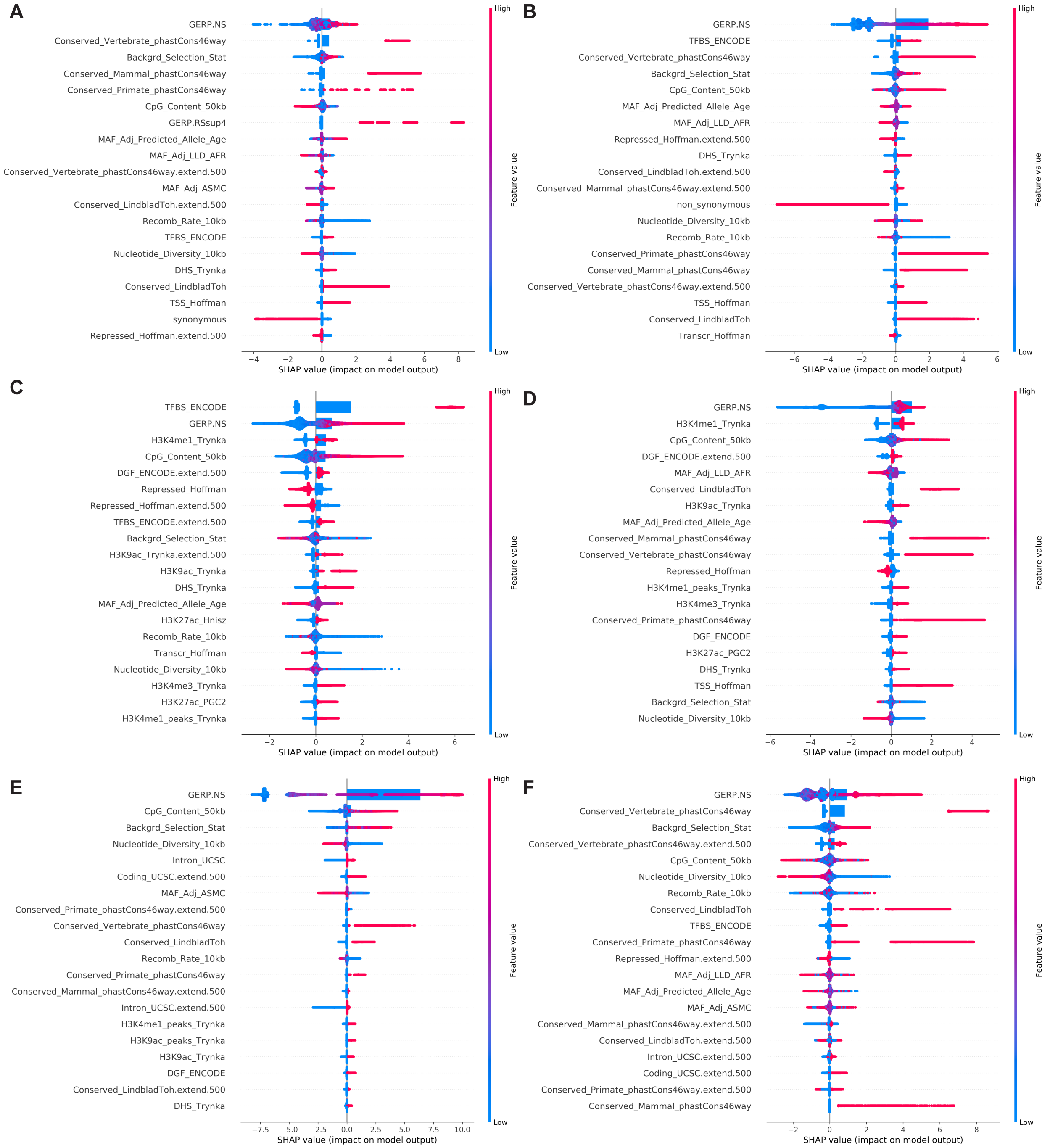


Figure S4. Feature importance of boosted genome-wide Mendelian disease-derived pathogenicity scores. We applied SHAP[41] to assess which features from the baseline-LD drives the prediction of 11 boosted missense scores by AnnotBoost. We report the signed impact of top 20 features for each of 6 genome-wide Mendelian disease-derived pathogenicity scores: (A) CADD, (B) Eigen, (C) Eigen-PC, (D) ReMM, (E) NCBoost, and (F) ncER. We obtained similar results for even/odd chromosome classifiers; we report odd chromosome results here (see full results online; see URLs).

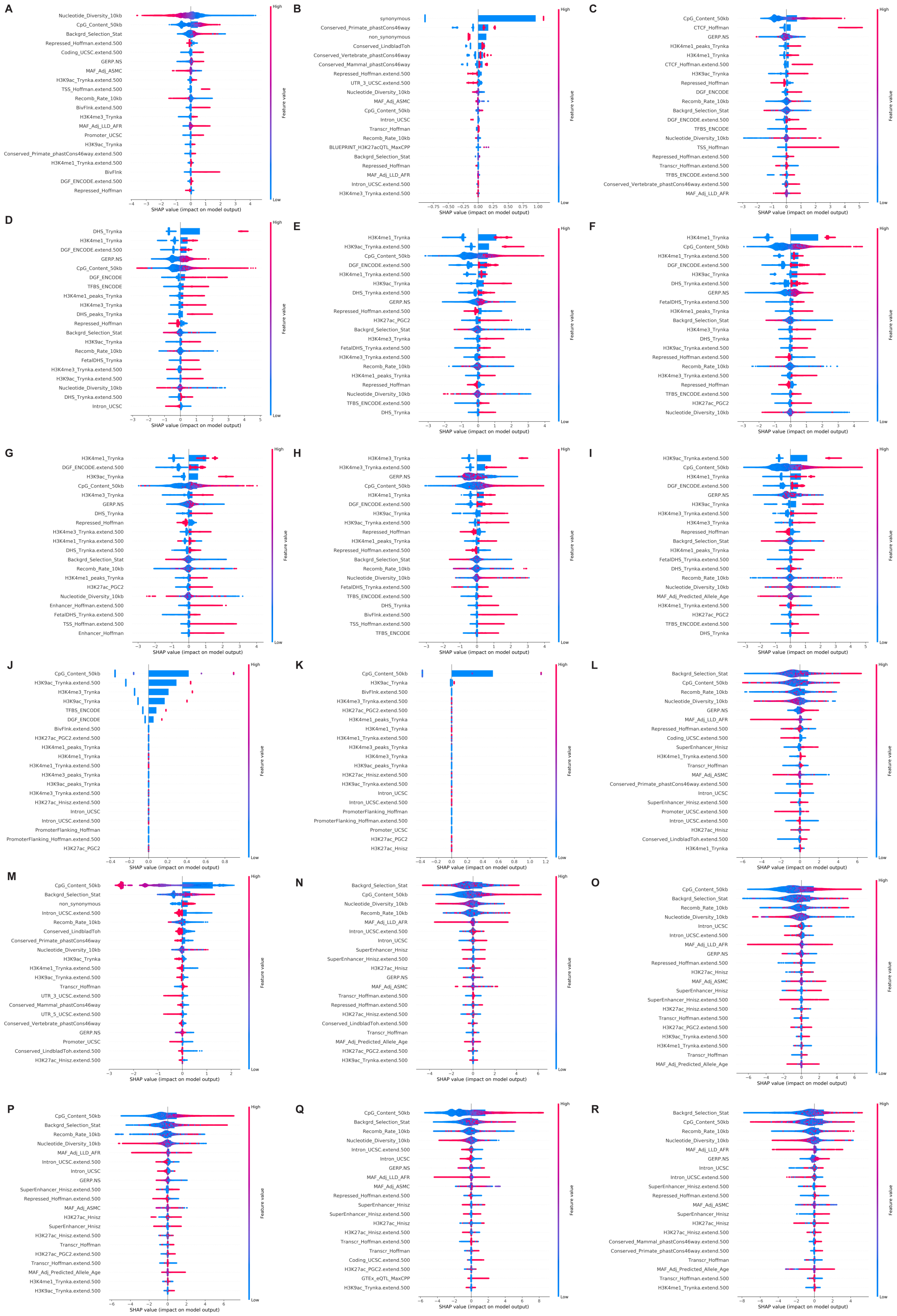


Figure S5. Feature importance of boosted scores derived from 18 additional genome-wide scores and 47 baseline-LD model annotations. We applied SHAP[41] to assess which features from the baseline-LD drives the prediction of 18 boosted additional scores by AnnotBoost. We report the signed impact of top 20 features for each of 18 additional scores: (A) CDTS, (B) CCR, (C-I) DeepSEA-CTCF, -DNase, -H3K27ac, -H3K4me1, -H3K4me2, -H3K4me3, -H3K9ac, (J-K) DIS-DNA, -RNA, (L) pLI, (M) LIMBR, (N-Q) Gene network connectivity-Saha, Greene, InWeb, Sonawane, (R) EDS. We obtained similar results for even/odd chromosome classifiers; we report odd chromosome results here (see full results online; see SHAP results of 47 boosted baseline-LD scores online; see URLs).

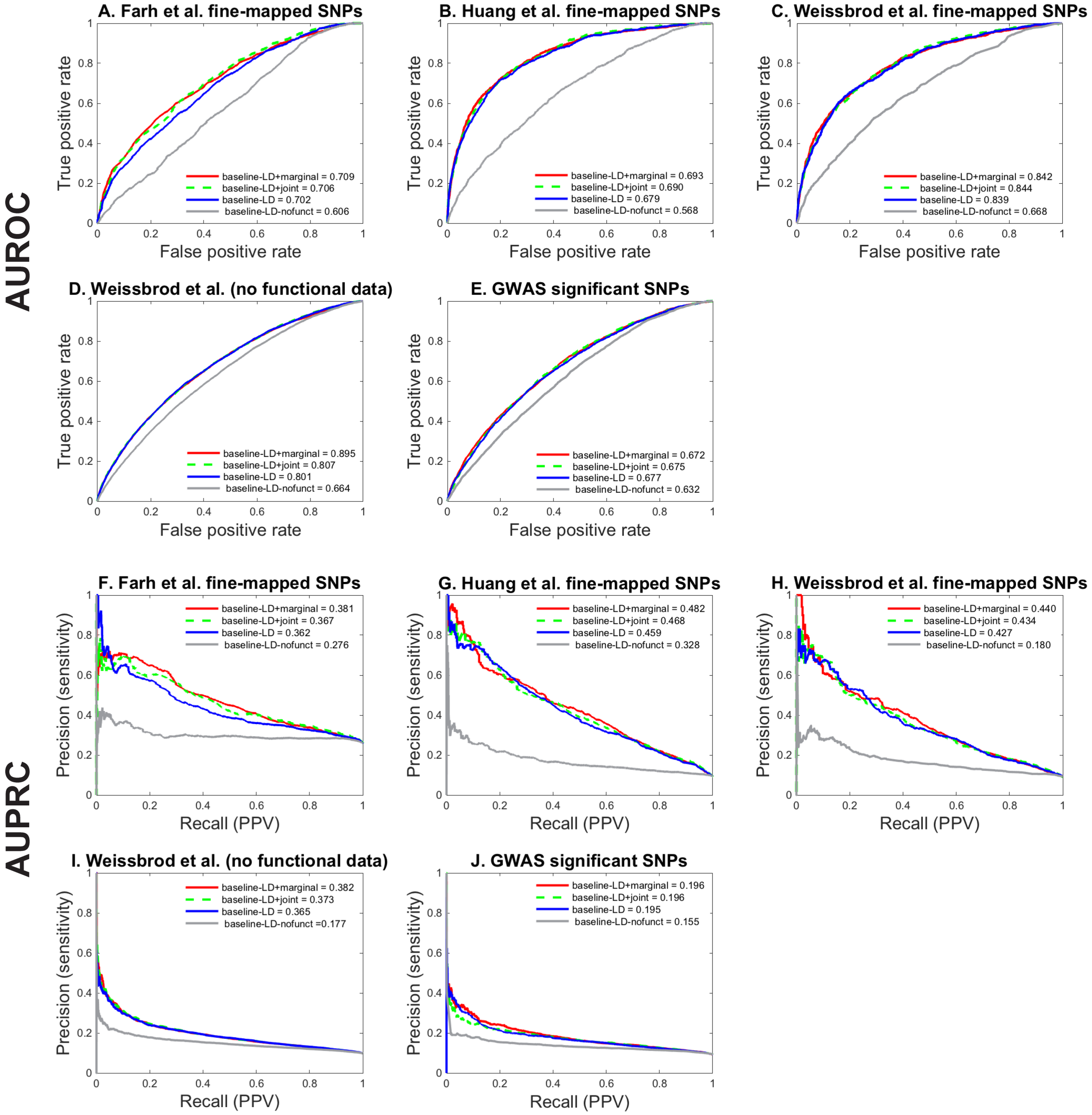


Figure S6. Classification of fine-mapped disease SNPs using aggregated scores. We report the true positive rate, false positive rate, precision, and recall along with the classification accuracy (AUPRCs and AUROCs) of four aggregated scores on classifying 5 different independent SNP sets: (A, F) 7,333 fine-mapped for 21 autoimmune diseases from Farh et al.[59], (B, G) 3,768 fine-mapped SNPs for inflammatory bowel disease from Huang et al.[60], (C, H) 1,851 fine-mapped SNPs for 47 traits from UK Biobank[61], (D, I) 1,379 fine-mapped SNPs without functional data for 47 traits from UK Biobank[61], and (E, J) 14,807 GWAS significant SNPs[62,63], from 10 LD-, MAF-, and genomic element-matched control SNPs. We report the average AUPRCs and AUROCs of even/odd-chromosome classifiers. Differences for AUROCs and AUPRCs attained between (1) baseline-LD and baseline-LD+joint model, (2) baseline-LD and baseline-LD+marginal model, and (3) baseline-LD+joint and baseline-LD+marginal were largely significant (p-val < 0.008). Numerical results, including results using the most matched control SNPs (instead of 10), are reported in Table S23.

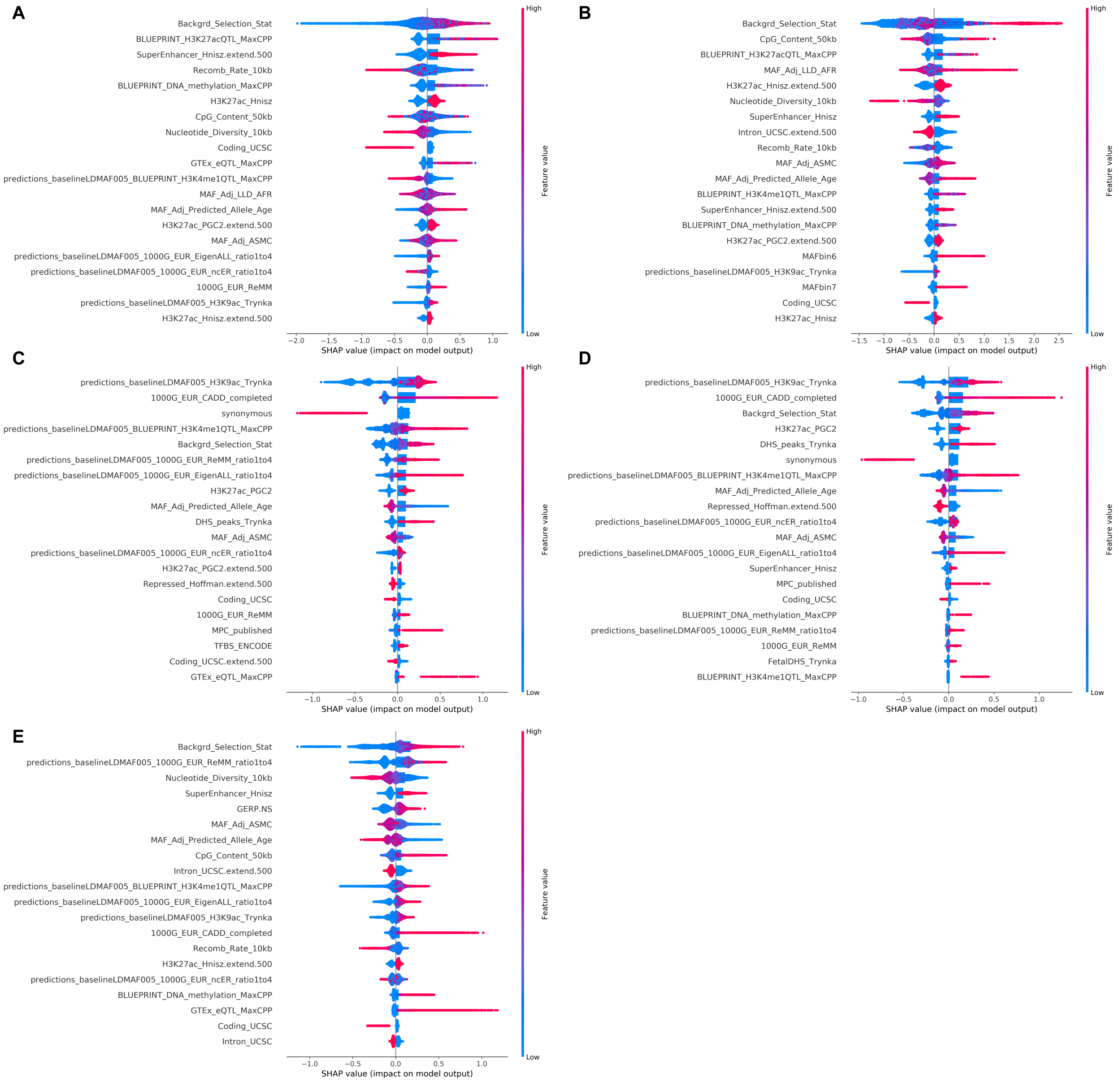


Figure S7. **Feature importance of baseline-LD+joint model in predicting fine-mapped or GWAS significant SNPs.** We applied SHAP[41] to assess which features from the baseline-LD+joint drive the prediction of fine-mapped or GWAS significant SNPs from 10 matched control SNPs for each positive SNP. We report the signed impact of top 20 feature for each of 4 fine-mapped SNPs and GWAS significant SNPs: (A) Farh et al., (B) Huang et al., (C) Weissbrod et al., (D) Weissbrod et al. (fine-mapped without functional data), (E) GWAS significant SNPs. We obtained similar results for even/odd chromosome classifiers; we report odd chromosome results here (see full results online; see URLs).

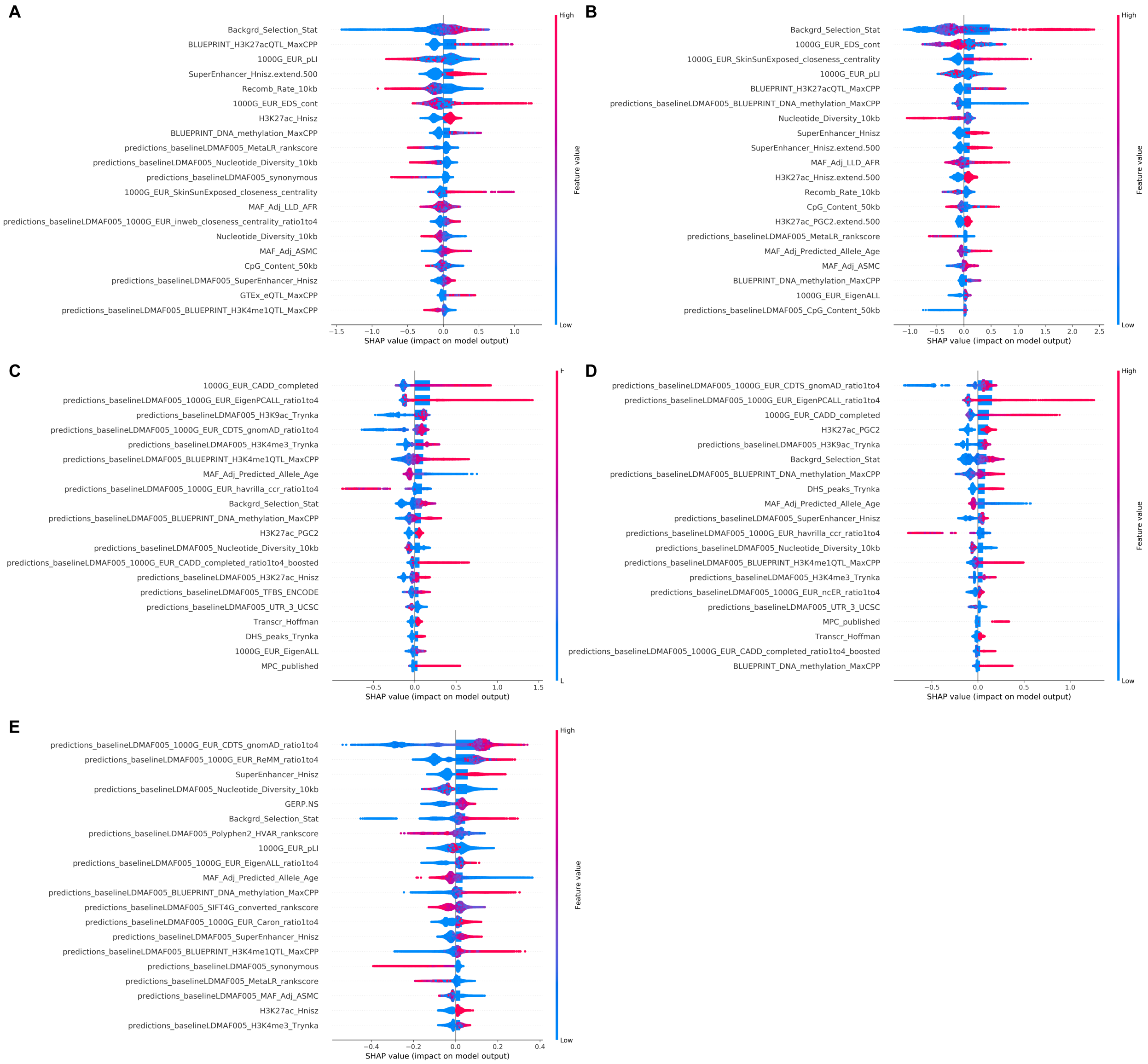


Figure S8. Feature importance of baseline-LD+marginal model in predicting fine-mapped or GWAS significant SNPs. We applied SHAP[41] to assess which features from the baseline-LD+marginal drive the prediction of fine-mapped or GWAS significant SNPs from 10 matched control SNPs for each positive SNP. We report the signed impact of top 20 feature for each of 4 fine-mapped SNPs and GWAS significant SNPs: (A) Farh et al., (B) Huang et al., (C) Weissbrod et al., (D) Weissbrod et al. (fine-mapped without functional data), (E) GWAS significant SNPs. We obtained similar results for even/odd chromosome classifiers; we report odd chromosome results here (see full results online; see URLs).

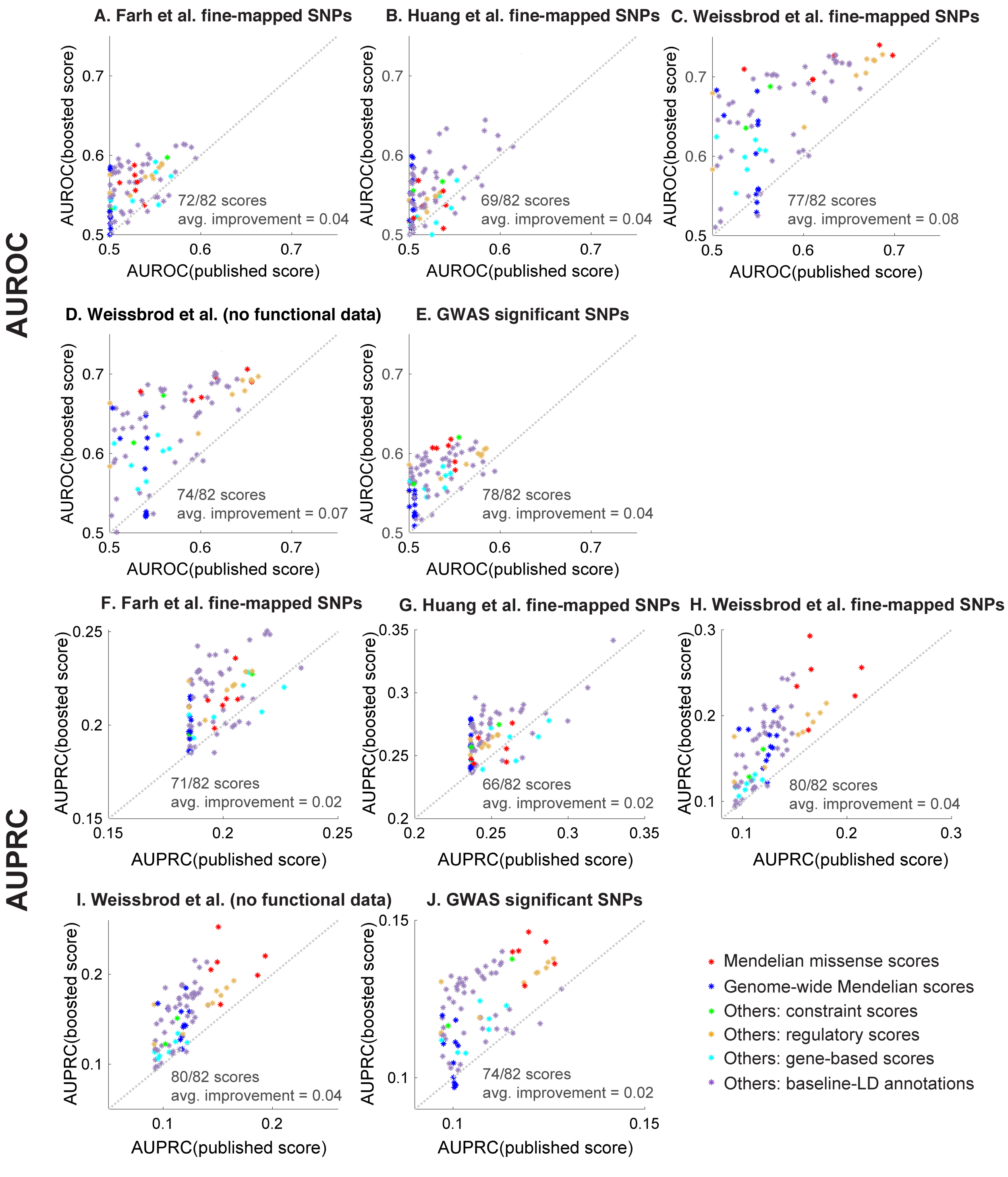


Figure S9. Classification of fine-mapped disease SNPs in single-score analysis using published and boosted scores. We report the classification accuracy (AUROC and AUPRC) of each of the 82 boosted scores compared to the corresponding published score. We report AUROC (resp. AUPRC) on (A, F) 7,333 fine-mapped for 21 autoimmune diseases from Farh et al.[59], (B, G) 3,768 fine-mapped SNPs for inflammatory bowel disease from Huang et al.[60], (C, H) 1,851 fine-mapped SNPs for 47 traits from UK Biobank[61], (D, I) 1,379 fine-mapped SNPs without functional data for 47 traits from UK Biobank[61], and (E, J) 14,807 GWAS significant SNPs[62,63] from 10 LD-, MAF-, and genomic element-matched control SNPs. Numerical results, including results using the most matched control SNPs (instead of 10), are reported in Table S25

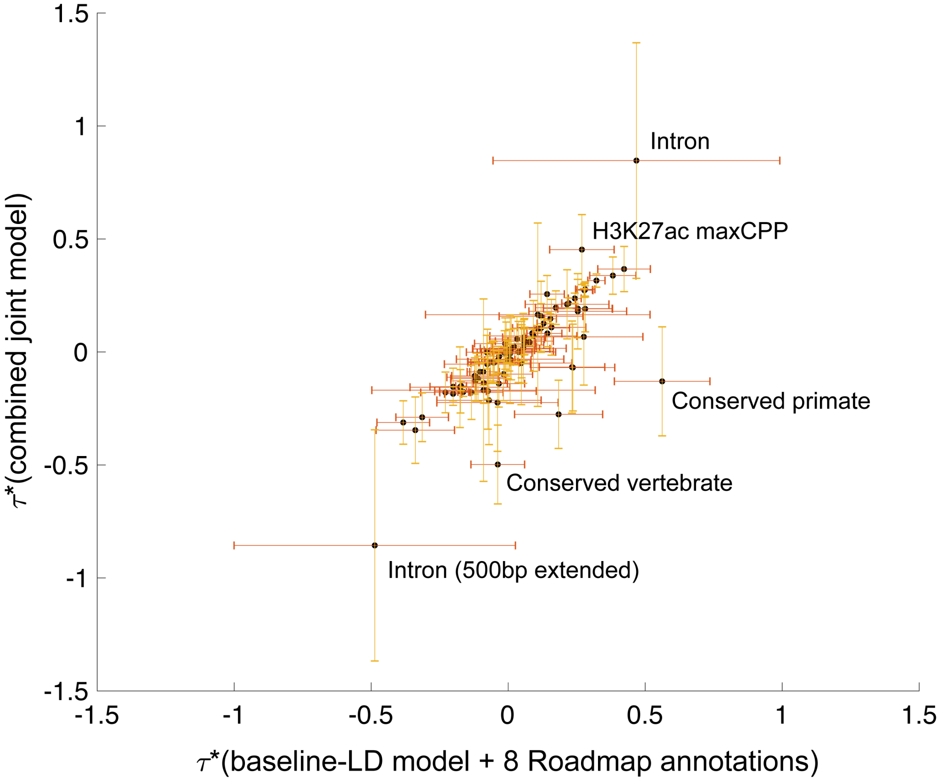


Figure S10. Informativeness of the baseline-LD model before and after adding 11 jointly significant binary annotations We report meta-analyzed τ^* of the baseline-LD model annotations, across 41 independent traits, from two different S-LDSC analyses: (1) the baseline-LD model + 8 Roadmap annotations and (2) the baseline-LD model + 8 Roadmap annotations + 11 jointly significant annotations. Error bars represent 95% confidence intervals. Numerical results are reported in Table S27.

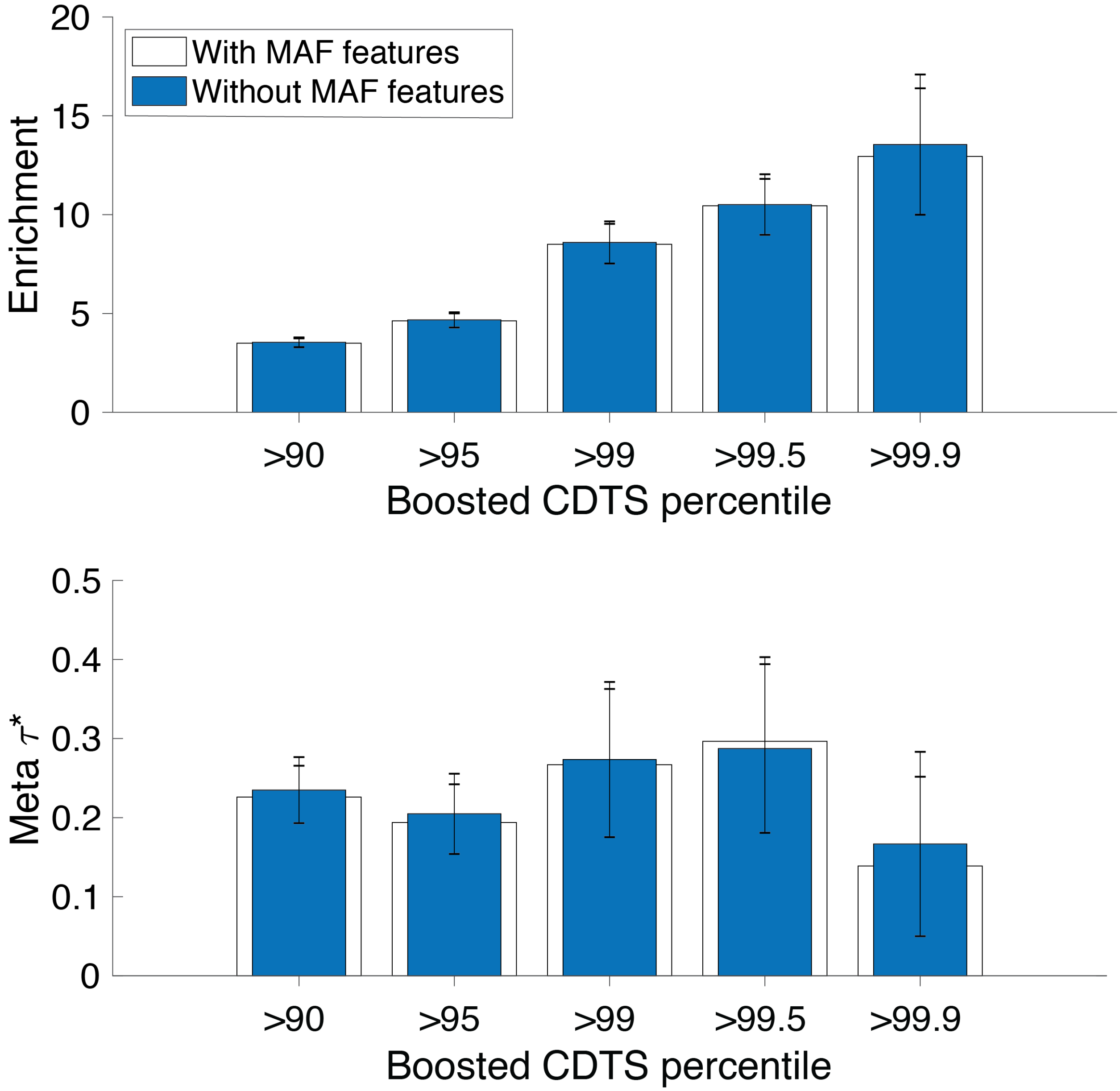


Figure S11. Informativeness for common disease of binary annotations derived from boosted CDTS scores with or without MAF features. We applied AnnotBoost to CDTS annotation using all baseline-Ld features and all features excluding MAF bins. Then, we applied S-LDSC, conditioning on published binary CDTS annotations (five thresholds from 90th percentile to 99.9th percentile) and baseline-LD model annotations; and meta-analyzed results across 41 independent traits. We report meta-analyzed enrichments and τ^*. Error bars represent 95% confidence intervals.

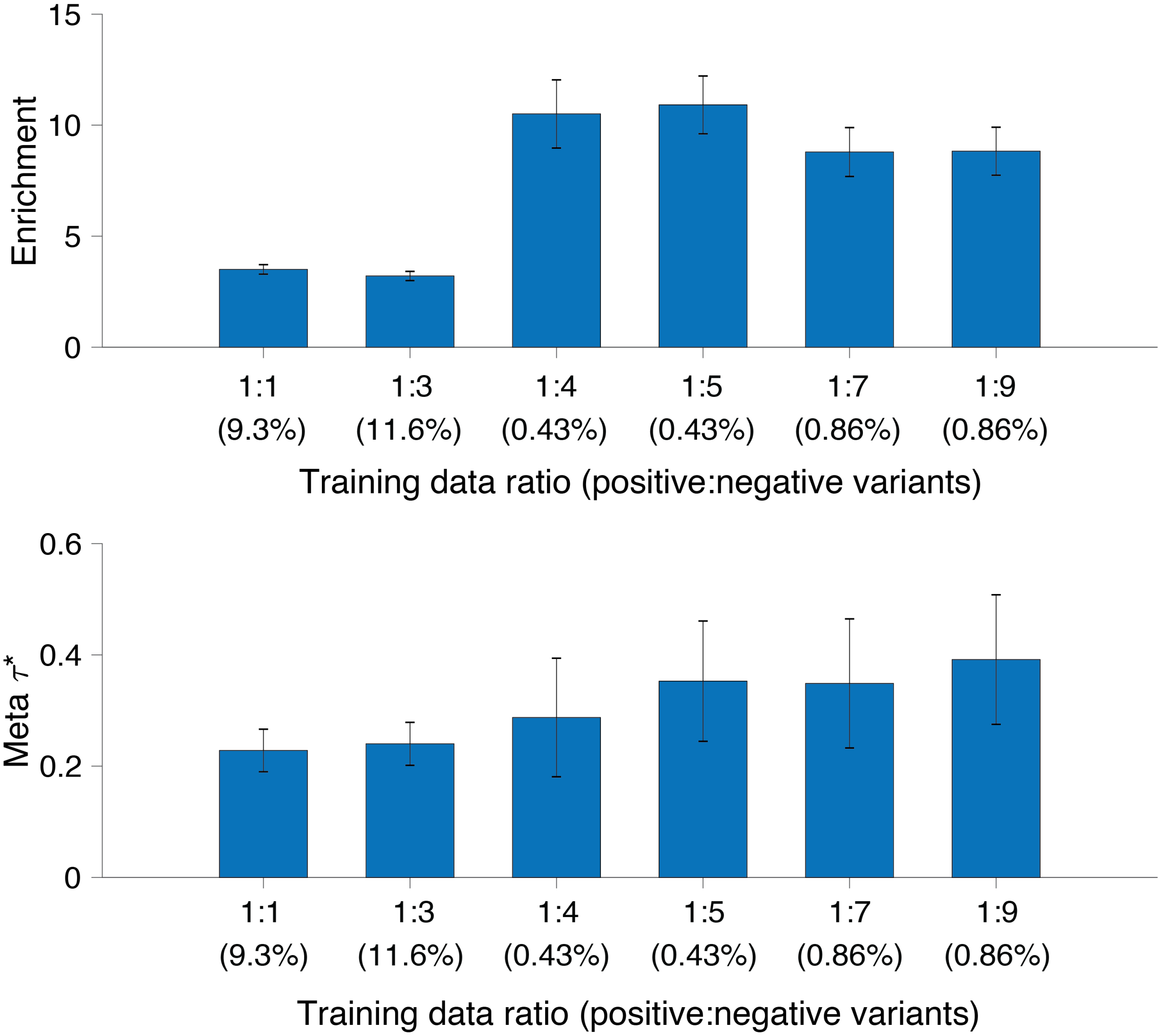


Figure S12. Informativeness for common disease of binary annotations derived from boosted CDTS scores using imbalanced data. We applied AnnotBoost to CDTS annotation of varying training data. Then, we applied S-LDSC, conditioning on published binary CDTS annotations (five thresholds from 90th percentile to 99.9th percentile) and baseline-LD model annotations; and meta-analyzed results across 41 independent traits. We report meta-analyzed enrichments and τ^*. Error bars represent 95% confidence intervals.

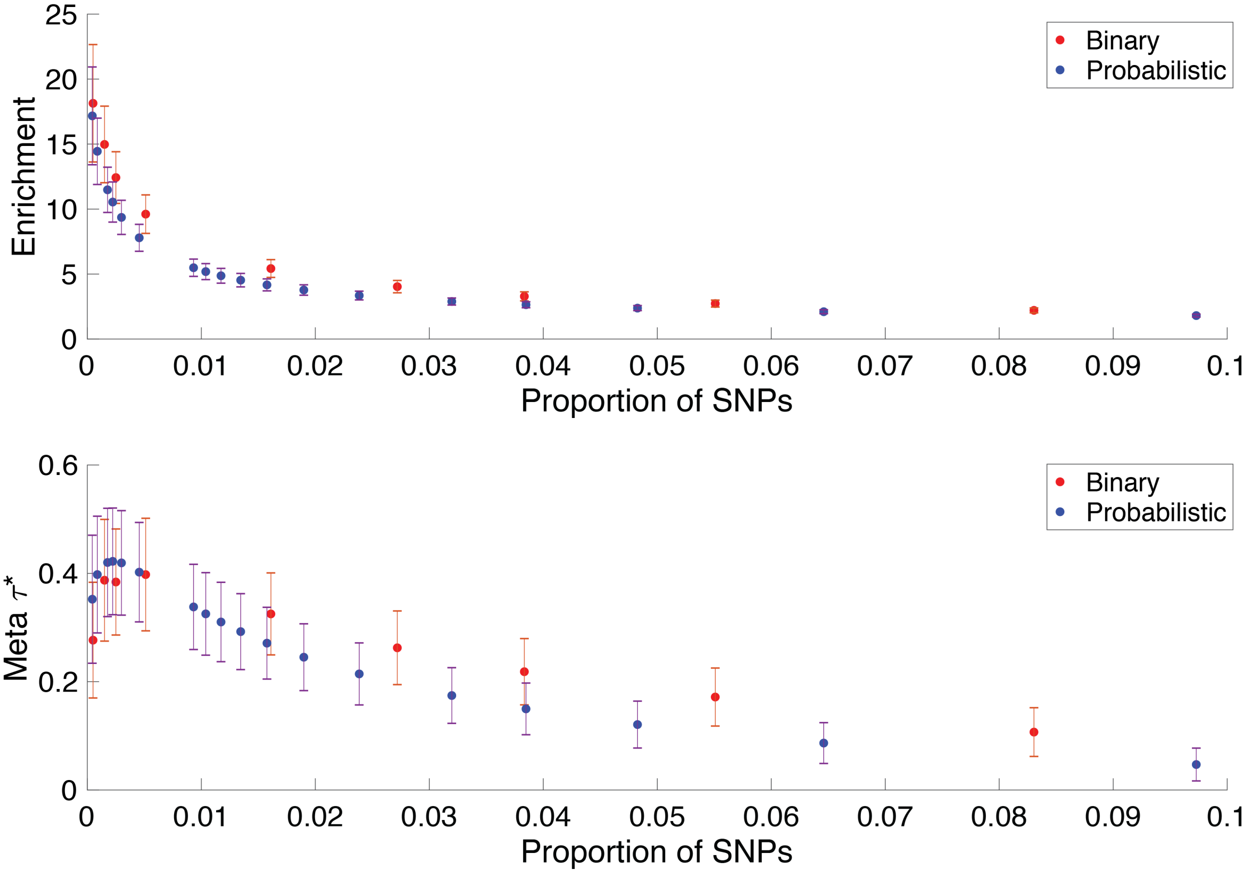


Figure S13. Informativeness for common disease of binary and probabilistic annotations derived from published CDTS scores. We constructed binary and probabilistic annotations for published CDTS[47]. We applied S-LDSC, conditional on baseline-LD model annotations and meta-analyzed results across 41 independent traits. We report meta-analyzed enrichments and τ^*. To construct probabilistic annotations of varying proportion of SNPs, we performed the following transformation to upweight the upper percentile and downweight the lower percentile SNPs: $\frac{e^{\alpha * annot}}{e^{\alpha}}$ with α varied from 3 to 2000. Error bars represent 95% confidence intervals.