









Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases

Received: 14 August 2020

Accepted: 9 June 2023

Published online: 13 July 2023

 Check for updates

Elle M. Weeks ^{1,37} ✉, Jacob C. Ulirsch ^{1,2,32,37}, Nathan Y. Cheng¹, Brian L. Trippe^{3,4}, Rebecca S. Fine^{1,5,6,33}, Jenkai Miao^{1,5}, Tejal A. Patwardhan^{1,7}, Masahiro Kanai ^{1,8,9,10}, Joseph Nasser¹, Charles P. Fulco^{1,34}, Katherine C. Tashman¹, Francois Aguet ¹, Taibo Li ^{1,11}, Jose Ordoñas-Montanes^{1,12,13,14}, Christopher S. Smillie^{1,3}, Moshe Biton^{1,15,35}, Alex K. Shalek^{1,16,17,18,19}, Ashwin N. Ananthakrishnan²⁰, Ramnik J. Xavier^{1,15,20,21}, Aviv Regev ^{1,18,22,36}, Rajat M. Gupta^{1,23,24}, Kasper Lage^{1,25}, Kristin G. Ardlie^{1,26}, Joel N. Hirschhorn^{1,5,6,27}, Eric S. Lander^{1,28,29}, Jesse M. Engreitz ^{1,30,31} & Hilary K. Finucane ^{1,8} ✉

Genome-wide association studies (GWASs) are a valuable tool for understanding the biology of complex human traits and diseases, but associated variants rarely point directly to causal genes. In the present study, we introduce a new method, polygenic priority score (PoPS), that learns trait-relevant gene features, such as cell-type-specific expression, to prioritize genes at GWAS loci. Using a large evaluation set of genes with fine-mapped coding variants, we show that PoPS and the closest gene individually outperform other gene prioritization methods, but observe the best overall performance by combining PoPS with orthogonal methods. Using this combined approach, we prioritize 10,642 unique gene–trait pairs across 113 complex traits and diseases with high precision, finding not only well-established gene–trait relationships but nominating new genes at unresolved loci, such as *LGR4* for estimated glomerular filtration rate and *CCR7* for deep vein thrombosis. Overall, we demonstrate that PoPS provides a powerful addition to the gene prioritization toolbox.

GWASs have identified thousands of genetic loci associated with common complex traits and diseases¹. Nevertheless, for the vast majority of significant GWAS loci, the identity of the causal gene(s) underlying the association remains unknown, limiting the biological insight gained into common disease mechanisms^{2,3}. There are several major challenges to pinpointing the causal gene. First, linkage disequilibrium (LD) between variants masks the identity of the causal variant⁴. Second, most associated loci do not contain protein-coding variants. Instead, the causal variant acts through gene-regulatory mechanisms³, but incomplete maps from regulatory element to gene hinder causal gene

identification⁵. Many computational approaches try to resolve these challenges^{6–10}, yet methods in the field of gene prioritization often fail to nominate causal genes with high confidence.

Gene-prioritization strategies can be placed into two broad categories: first, locus-based methods that leverage local GWAS data by connecting the causal variants to the causal gene(s) using protein-coding variants, genomic distance, enhancer–gene maps^{11–16} or expression quantitative trait loci (eQTLs)^{7,8}; and second, similarity-based methods that search for global patterns in associated genes and nominate those with similar functions, pathways or network connections^{6,10,17}.

A full list of affiliations appears at the end of the paper. ✉ e-mail: eweeks@broadinstitute.org; finucane@broadinstitute.org

Across both categories, existing methods lack consensus and have high false-positive rates¹⁸. At the same time, related work suggests that combining results from different methods can yield better predictions¹⁹. Among similarity-based approaches, most methods focus on SNPs that meet genome-wide significance, ignoring information from sub-significant variants that explain most of the narrow sense heritability^{17,20}. Moreover, recently generated single-cell RNA sequencing (scRNA-seq) datasets hold promise for more accurately characterizing shared functions among genes and, thus, improving the accuracy of similarity-based gene prioritization.

In the present study, we propose a new similarity-based gene prioritization method, a gene-level Polygenic Priority Score (PoPS), that leverages the full polygenic signal and incorporates data about genes from a variety of sources, including 73 publicly available scRNA-seq datasets. PoPS is computationally efficient and requires only summary statistics and an LD reference panel. Across 113 complex traits and diseases, we show that PoPS outperforms other similarity-based and locus-based gene prioritization methods using a unique evaluation gene set. We further show that, by combining PoPS with locus-based gene prioritization methods, we can prioritize genes with higher confidence than PoPS or any locus-based method alone, ultimately prioritizing genes at 10,642 GWAS loci with high confidence.

Results

Overview of PoPS

Our method, PoPS, is predicated on the assumption that causal genes share functional characteristics. Specifically, we assume that genes with physical locations on the genome near significantly associated SNPs and that share similar biological annotations are most likely to be causal. PoPS uses gene-level associations computed from GWAS summary statistics to learn joint polygenic enrichments of gene features derived from cell-type-specific gene expression, biological pathways and protein–protein interactions (PPIs). To nominate causal genes, PoPS then assigns a priority score to every protein-coding gene according to these enrichments (Fig. 1).

First, PoPS applies MAGMA⁹ to compute gene-level association statistics and their correlations using GWAS summary statistics and LD information from an ancestry-matched reference panel (Methods). Gene-level covariates are then projected out of the computed gene-level associations from MAGMA to control for variables such as gene length and density of SNPs (Methods). Next, PoPS performs marginal feature selection by using MAGMA to perform enrichment analysis for each gene feature separately. MAGMA tests a gene feature, f , for enrichment by modeling the gene-level associations, y , by:

$$y = X^f \beta^f + \epsilon, \epsilon \approx \text{MVN}(0, R) \tag{1}$$

where X^f is a column vector corresponding to gene feature f (for example, a binary indicator of membership in a pathway), and the error terms, ϵ , are assumed to follow a multivariate normal (MVN) distribution with covariance matrix, R , that accounts for the LD between nearby genes computed from a reference panel. The model is fit by generalized least squares (GLS) and MAGMA reports both $\hat{\beta}^f$ and a P value for the hypothesis that $\beta^f \neq 0$. We retain features that pass a nominal significance threshold ($P < 0.05$) to reduce the noise and computational complexity of fitting the joint model.

Second, PoPS fits a joint model by replacing the vector X^f in equation (1) with a matrix X that includes all of the selected features (see Extended Data Fig. 1 for comparison of model-fitting choices).

$$y = X\beta + \epsilon, \epsilon \approx \text{MVN}(0, R) \tag{2}$$

We extend the GLS method used by MAGMA to incorporate L_2 regularization to account for a large number of features and improve test set prediction, obtaining an estimated $\hat{\beta}$.

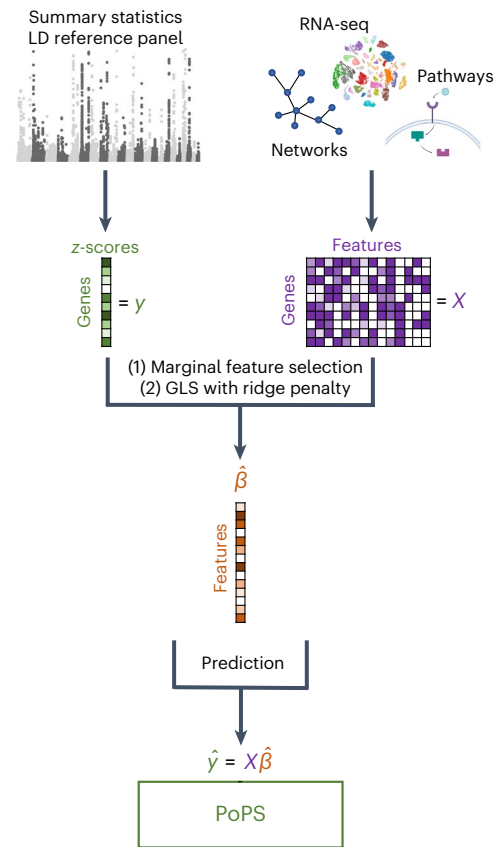


Fig. 1 | Overview of PoPS. We computed gene-level z-scores from GWAS summary statistics with an LD reference panel using MAGMA. We created gene features from gene expression data, biological pathways and predicted PPI networks, and used marginal feature selection to limit features included as those most likely to be relevant. We then fitted a linear model for the dependence of gene-level associations on gene features using GLS to account for LD and add an L_2 penalty to account for the large number of features. This results in a vector of joint polygenic enrichments of gene features, $\hat{\beta}$, which we use to assign gene priority scores.

Finally, PoPS computes polygenic priority scores for each gene, g , by multiplying its row vector of gene features, X_g , by $\hat{\beta}$.

$$y_g = X_g \hat{\beta} \tag{3}$$

We refer to y_g as the polygenic priority score (PoP score) for gene g . We say PoPS prioritizes a gene if it is in a 1-Mb locus centered on a genome-wide significant variant and has the highest PoP score among genes in the locus (Methods). In addition, we include an option for a user to compute scores in a leave-one-chromosome-out (LOCO) fashion, obtaining estimates $\hat{\beta}_{-chr_i}$ for $i = 1, \dots, 22$ and scoring, for example, a gene on chromosome 1 by using $\hat{\beta}_{-chr1}$.

Application of PoPS to 113 complex traits

We applied PoPS to 18 diseases with publicly available GWAS summary statistics and 95 complex traits from the UK Biobank²¹ (Supplementary Table 1) using EUR individuals from the 1000 Genomes Project²² as a reference panel (Methods). The full set of gene features used in these analyses included 57,543 total features: 40,546 derived from gene expression data, 8,718 extracted from a PPI network²³ and 8,479 based on pathway membership^{24–27} (Supplementary Table 2 and Methods). After marginal feature selection, 2,512–26,155 features per trait were included in the predictive model (Extended Data Fig. 1d). For each trait,

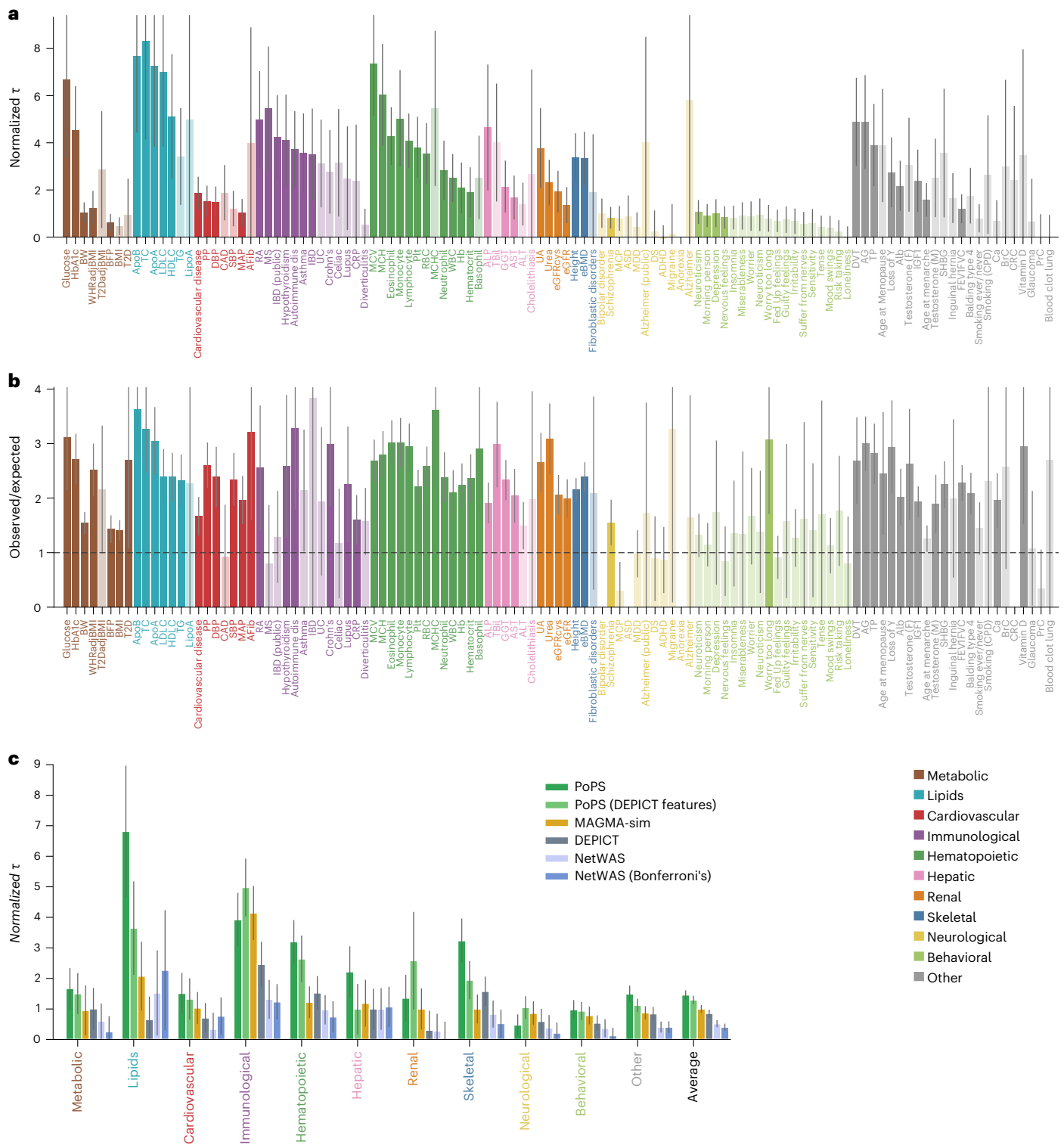


Fig. 2 | Evaluation of PoPS and comparison to other similarity-based methods. **a**, Results using Benchmarker to evaluate PoPS, grouped by trait domain and sorted by the lower bound of the 95% confidence interval (CI) of normalized τ . Normalized τ provides an estimate for the average contribution of SNPs near genes with high priority scores to per-SNP heritability, normalized by average to per-SNP heritability. Error bars represent 95% CIs around the point estimate. One-sided *P* values were computed using the z-score test for heritability enrichment in S-LDSC. Opaque bars passed Bonferroni's significance threshold. For inflammatory bowel disease (IBD) and Alzheimer's disease, we retained

summary statistics from both the UK Biobank and other publicly available sources with a greater sample size. **b**, Results using closest gene enrichment to evaluate PoPS ordered as in **a**. Error bars represent 95% CIs around the point estimate. One-sided *P* values were computed using a normal approximation to the null distribution and opaque bars passed Bonferroni's significance threshold. **c**, Results using Benchmarker to compare similarity-based gene-prioritization methods, meta-analyzed within each trait domain across independent traits ($n = 46$ independent traits). Error bars represent 95% CIs around the meta-analyzed point estimate.

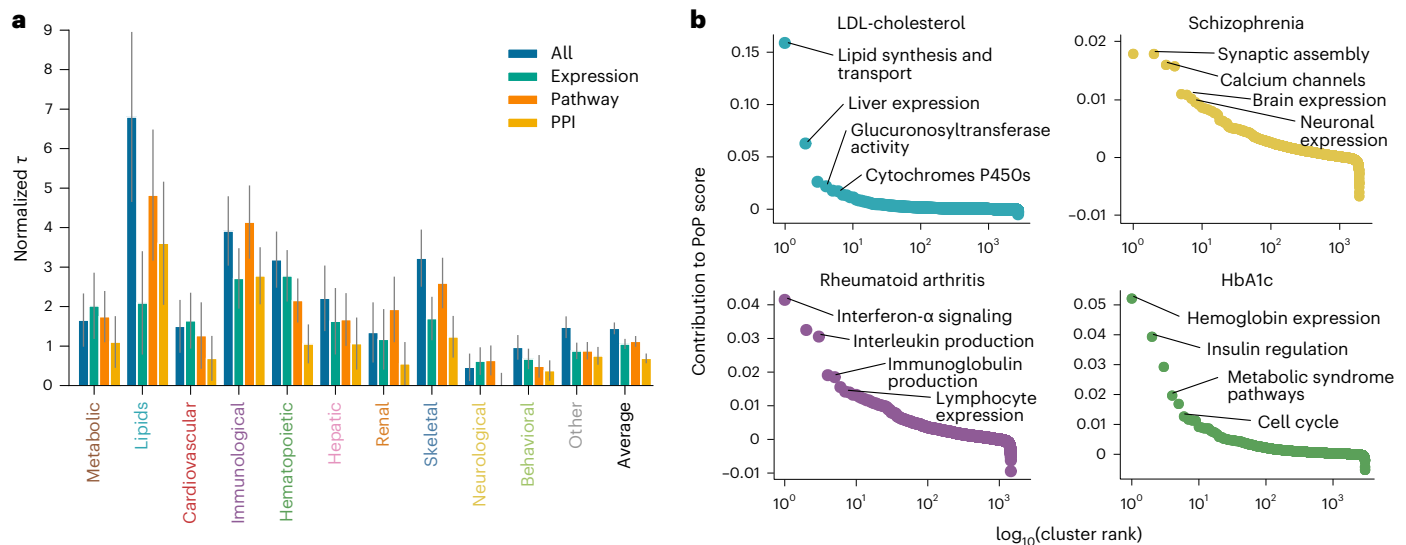


Fig. 3 | Most informative gene features used by PoPS. a, Results using Benchmarker to compare PoPS using different feature sets, meta-analyzed within each trait domain across independent traits ($n = 46$ independent traits).

Error bars represent 95% CIs around the meta-analyzed point estimate. **b**, Rank-order plots for selected traits highlighting the feature clusters with the greatest contribution to the PoP scores of prioritized genes.

we score 18,383 protein-coding genes and prioritize one gene in each genome-wide significant locus. In total, PoPS prioritized 17,906 unique gene–trait pairs in 25,342 loci across 113 complex traits.

Enrichment-based evaluation of PoPS

To evaluate the performance of PoPS for prioritizing probable causal genes, we avoided using curated sets of gold-standard genes that may be biased toward well-studied genes or genes in well-characterized pathways. We instead evaluated PoPS with two metrics unaffected by prior knowledge of trait etiology by taking advantage of the fact that PoPS can be run using a LOCO approach, allowing us to use held-out locus association data for validation. Both metrics quantify the extent to which genes with high PoP scores overlap with sets of variants or genes that we expect to be highly enriched for causal signals. First, we applied the Benchmarker method¹⁷, which evaluates methods by estimating the average contribution of SNPs near top-scoring genes to per-SNP heritability (τ). A value of normalized τ significantly greater than zero indicates that genes with high PoP scores are enriched for heritability, even after accounting for the contributions of 53 other genomic annotations. After correction for multiple testing, we found that our estimates for normalized τ were significantly greater than zero for 54 of 113 traits tested (Fig. 2a and Supplementary Table 3). As a second evaluation metric, we focused on the performance of PoPS in GWAS significant loci. Following previous work demonstrating that the causal gene is often the closest gene to the lead variant in the locus¹⁹, we tested whether PoPS-prioritized genes were the closest gene to the lead variant more often than expected by chance (Methods). Although this test is underpowered for traits with a small number of significant loci, we found that PoPS-prioritized genes were significantly enriched for being the closest gene for 64 of 113 traits tested after Bonferroni's correction (Fig. 2b and Supplementary Table 3). Thus, both Benchmarker and our closest gene metric indicate that PoPS-prioritized genes are enriched for being causal.

Comparison to similarity-based methods

After evaluating PoPS on its own, we investigated how PoPS compares with existing similarity-based methods: DEPICT⁶, NetWAS¹⁰ and a method that we call MAGMA-sim¹⁷ (Methods). Using the same set of 113 traits, we applied: (1) PoPS using the full set of 57,543 features; (2) PoPS using only the 14,461 reconstituted gene sets used by DEPICT; (3)

MAGMA-sim using the 14,461 reconstituted gene sets; (4) DEPICT using the 14,461 reconstituted gene sets; (5) NetWAS using a significance cutoff of $P < 0.01$; and (6) NetWAS using Bonferroni's significance cutoff. We found that PoPS using the full feature set showed the strongest performance compared with other similarity-based methods for 31 of 46 independent traits using Benchmarker and 33 of 46 traits using the closest gene test (Supplementary Table 3). After meta-analyzing estimates within 11 trait domains across 46 traits chosen to have low genetic correlation (Methods), we again found that PoPS significantly outperformed all other similarity-based methods tested (Fig. 2c and Extended Data Fig. 2a). Importantly, when PoPS, DEPICT and MAGMA-sim were run using the same features, PoPS significantly outperformed the other two methods by both metrics; giving PoPS access to the full set of features further increased its performance. Thus, we attribute the superior performance of PoPS compared with other similarity-based methods to both the large set of gene features and the joint modeling used to integrate signal across those gene features.

Interpreting gene features in the PoPS model

We next evaluated the relevance of each category of the features included in the PoPS model: gene expression, pathways and PPI networks (Fig. 3 and Extended Data Figs. 2 and 3). We created three alternative versions of PoPS, training on features from each category separately, to produce three new sets of results for each phenotype. In a meta-analysis of 46 independent traits, we found that inclusion of all features yielded the strongest performance in both Benchmarker and closest gene evaluations, followed in order by pathways, gene expression and, last, by PPI networks (Fig. 3a, Extended Data Fig. 3b and Supplementary Table 3).

To better understand the relevant tissues, cell types and pathways learned by PoPS, we investigated which features were most informative for prioritized genes. As many highly correlated features were included in the joint model for PoPS, the individual coefficients, $\hat{\beta}$, lacked direct interpretability. We instead grouped related features for a trait by performing hierarchical clustering on the selected features (Methods) and computed the total contribution of the features in the cluster to the PoP scores of prioritized genes. Gene features in the top clusters recapitulated known trait biology and included examples from each type of feature (Fig. 3b and Supplementary Table 4). For low-density lipoprotein (LDL)-cholesterol, we observed clusters composed of lipid

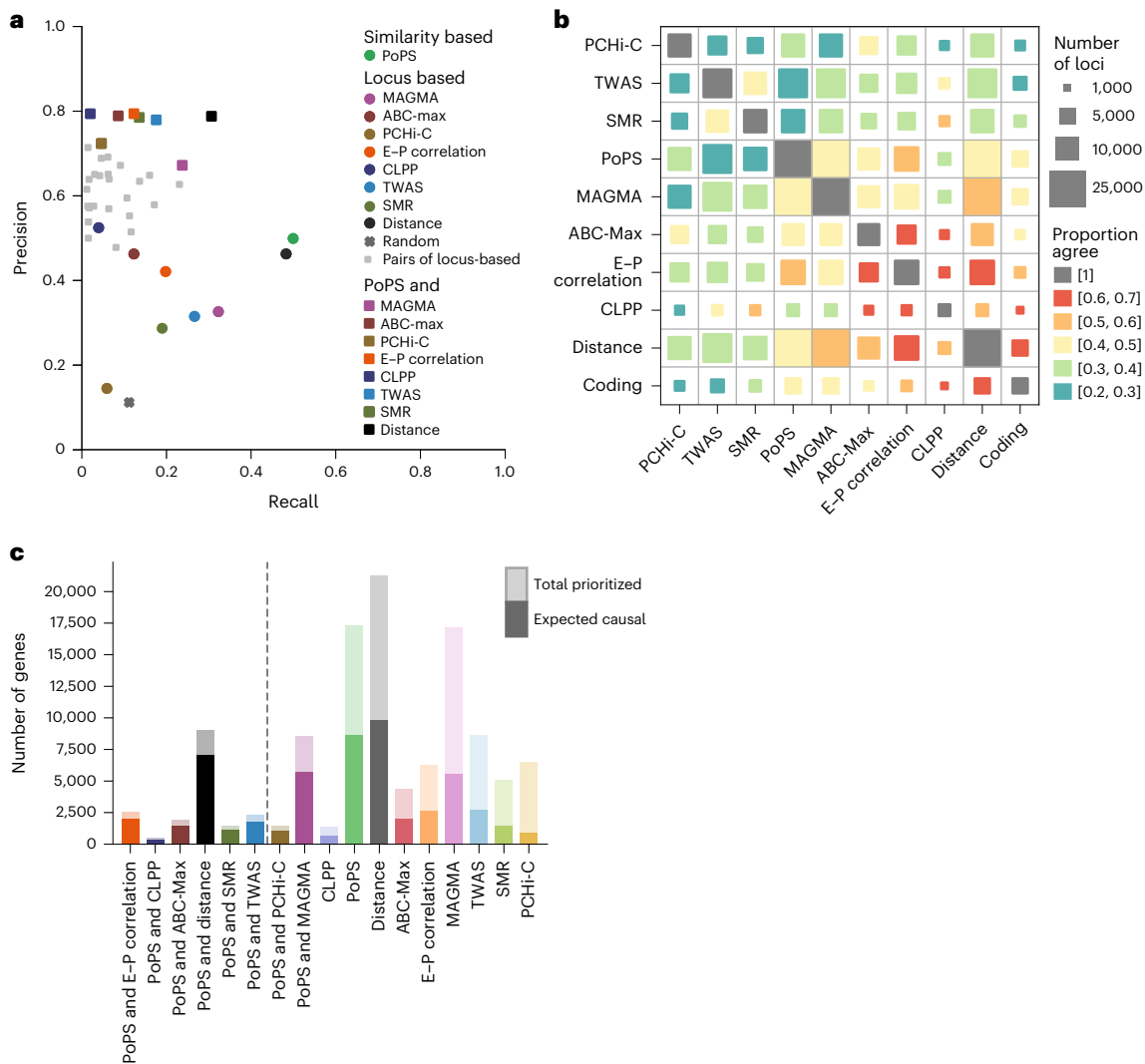


Fig. 4 | Comparing and combining PoPS with locus-based methods. **a**, Precision-recall plot showing performance of locus-based methods, PoPS, intersections of pairs of locus-based methods and intersections of PoPS with locus-based methods, using the evaluation gene set of 589 genes with fine-mapped protein-coding variants in 1,348 noncoding loci containing genes that harbor fine-mapped protein-coding variants. **b**, Overlap and agreement among methods across all genome-wide significant loci. Each square represents a pair of methods; the size corresponds to the number of loci where both methods

prioritize a gene, and the color corresponds to the proportion of these loci where both methods prioritize the same gene. **c**, Number of unique gene-trait pairs prioritized across all genome-wide significant loci by PoPS, locus-based gene prioritization methods and intersections of PoPS with locus-based methods, sorted by estimated precision. The full height of each bar represents the total number of genes prioritized. The opaque portion of each bar represents the expected number of true causal genes prioritized. Methods to the left of the dashed line achieve precision >75%.

synthesis and transport pathways in addition to liver gene expression features²⁸. For glycated hemoglobin, a test that measures average blood sugar levels but is also affected by red blood cell levels²⁹, we observed both glucose- and hemoglobin-related clusters of features. For rheumatoid arthritis (RA), an autoimmune disease³⁰, we observed a range of immune features describing expression, signaling and production of immune cells and their function. Finally, for schizophrenia, we observed clusters corresponding to regional brain expression features and mechanisms previously implicated in schizophrenia, including calcium channel dysfunction^{31,32}. Taken together, these results suggest that the PoPS can prioritize the causal genes underlying complex traits and diseases by learning biologically relevant properties from multiple types of gene features.

Development of a new evaluation gene set

We next sought to compare the PoPS with existing locus-based methods. However, the approaches we used to compare similarity-based

prioritization methods, Benchmarker and the closest gene metric, are not applicable to locus-based methods. Curated gold-standard gene sets, in addition to being small and often unavailable for complex traits and common diseases, are often biased toward well-studied genes, potentially introducing bias when evaluating the PoPS, which uses existing pathway databases. Thus, we constructed a new evaluation set of approximate gold-standard gene-trait pairs that do not suffer from the same biases.

To create a new evaluation set, we first used the results of our recent statistical fine-mapping of 95 traits from the UK Biobank²¹ to identify probable causal genes as those harboring a fine-mapped (PIP > 0.5) protein-coding variant. Matching by trait, we then identified independent, noncoding credible sets within 500 kb of these protein-coding genes. This approach identified 1,348 noncoding credible sets with physically proximal, but independent, coding variant signals. We created an evaluation set from these 1,348 loci, consisting of all genes within 500 kb of the locus-defining, noncoding credible

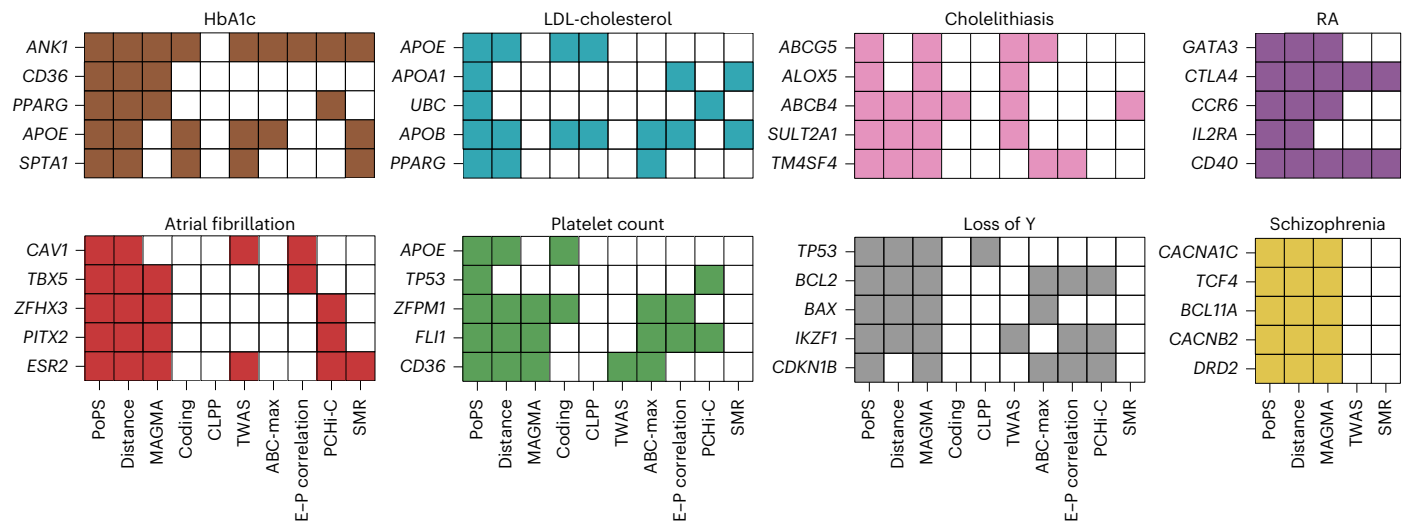


Fig. 5 | High-confidence genes for selected traits. Top five genes prioritized by PoPS + local, ranked by PoP score, for selected traits. Shaded boxes indicate if a method prioritized the gene.

set (median of 13 genes per locus). Genes at loci with an independent, fine-mapped coding variant for the same trait were labeled positive and genes without were labeled negative. This assignment directly encodes our assumptions that: (1) genes harboring a fine-mapped coding variant are trait relevant and (2) multiple independent associations in a locus are most likely to act through the same gene (Discussion). This evaluation set allows us to directly estimate the precision (the proportion of prioritized genes that are labeled positive) and recall (the proportion of positively labeled genes that are correctly prioritized) of any similarity-based or locus-based gene prioritization method.

Before considering locus-based methods, we used this new evaluation set to again compare PoPS with the other similarity-based methods (Extended Data Fig. 4 and Supplementary Table 5). We found that PoPS had both higher recall and higher precision than DEPICT, NetWAS and MAGMA-sim using noncoding genetic signal (Methods). In addition we showed that PoPS using the full noncoding genetic data had both higher recall and higher precision than PoPS run in a LOCO framework.

Combing PoPS with locus-based methods

With these metrics, we evaluated existing locus-based methods applied to the set of 95 traits from the UK Biobank (Methods), where we had not only summary statistics, but fine-mapping results. We evaluated a nonexhaustive but wide range of methods:

- (1) We overlapped fine-mapped ($PIP > 0.1$), noncoding variants with predicted enhancer–gene connections from (i) correlating enhancer and promoter activity (E–P correlation)^{15,16}, (ii) three-dimensional loops from promoter capture Hi–C (PCHi–C)^{12–14} and (iii) Activity-by-Contact (ABC)¹¹ maps to identify genes regulated by fine-mapped variants.
- (2) We incorporated eQTL data and (i) applied a transcription-wide association study (TWAS)⁸ with Genotype Tissue Expression (GTEx) v.7 (ref. 33) weights to identify significantly associated genes, (ii) applied SMR³⁴ with GTEx v.7 (ref. 33) weights to identify significantly associated genes and (iii) computed colocalization posterior probabilities (CLPPs)⁷ with fine-mapping results from GTEx v.8 (refs. 21,35,36) to identify genes where the causal variant is shared between the complex trait and the gene expression trait.
- (3) We used the raw gene scores from MAGMA, derived from local association statistics without any gene features.
- (4) We identified the closest gene to the lead variant.

Before directly comparing methods, we evaluated multiple prioritization criteria, using both absolute thresholds and relative rank within a locus (Extended Data Fig. 5, Supplementary Table 6 and Methods). Across all methods, we found that prioritizing the single best ranked gene in a locus had higher precision than including all genes passing a global score threshold, consistent with the idea that a regulatory variant can affect the expression of multiple genes^{33,35}, yet only a select few, perhaps often the most strongly regulated, have a direct effect on the complex trait of interest. Thus, our primary evaluation of gene-prioritization methods compares only the top-ranked gene per locus (Methods).

All locus-based methods for prioritizing genes from noncoding signal showed precision $< 50\%$ except CLPP, which had both the highest precision, 52%, and the lowest recall, 4% (Fig. 4a, Extended Data Fig. 6 and Supplementary Tables 5 and 7). Distance had the next highest precision, 46%, and the highest recall, 48%. The other locus-based methods yielded variable precision (14–46%) and recall (5–32%). The low recall of most of these methods can be attributed in part to limited power to isolate the causal variant because of LD, limited eQTL overlap with complex traits at current sample sizes and missing trait-relevant cell types in the variant-to-gene regulatory maps. Our method, PoPS, showed precision and recall of 50%, which was consistent when restricting to only 46 independent traits (51%) and when further restricting to only allow unique validation genes rather than unique validation gene–trait-credible set triplets (50%).

Noting previous work on the utility of combining locus-based and similarity-based gene prioritization methods^{19,37}, we investigated agreement among methods, including genes with fine-mapped ($PIP > 0.1$) coding variants. For each pair of methods, we computed the number of loci in which both methods prioritized a gene and the proportion of those loci where they prioritized the same gene (Fig. 4b and Supplementary Table 8). Overall, we found low concordance among methods. For example, PCHi–C prioritized a gene in 8,777 loci, whereas ABC–Max prioritized a gene in 7,913 loci; when both methods prioritized any gene, they agreed only 42% of the time. PoPS had mild agreement with most other methods, prioritizing the same gene in up to 52% of loci, suggesting that PoPS and locus-based methods contain independent information and may improve prioritization when combined.

Returning to our evaluation gene set to evaluate this hypothesis, we found that combining PoPS with locus-based methods improved precision, while maintaining appreciable recall (Fig. 4a, Extended Data Fig. 7 and Supplementary Table 5). Specifically, for each locus-based

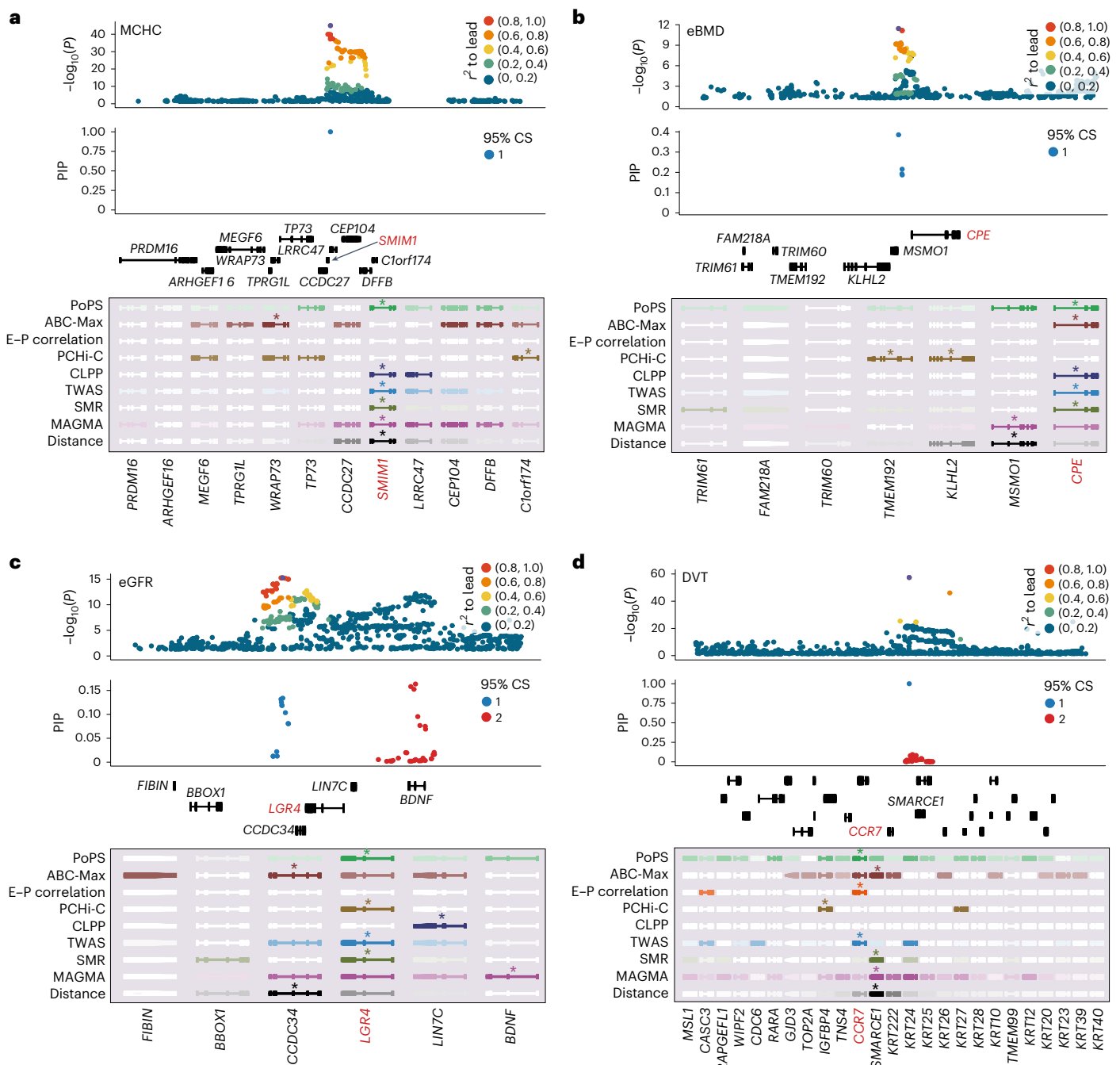


Fig. 6 | Known and new biological examples. a–d. Top, summary statistics colored by LD to the lead variant and fine-mapping results for variants in the locus colored by credible set. Bottom, results from PoPS and locus-based methods for all genes in the locus. Genes are colored by strength of prediction for each

method with a star denoting the prioritized gene. **a**, Variant rs1175550, *SMIM1* for MCH concentration (MCHC). **b**, Variant rs1550270, *CPE* for BMD (eBMD). **c**, Variant rs11029928, *LGR4* for eGFR. **d**, Variant rs112401631, *CCR7* for DVT.

method, intersecting the set of genes prioritized by the locus-based method with the set of genes prioritized by PoPS led to precision of at least 67% and up to 79%, depending on the locus-based method. In contrast, when intersecting pairs of locus-based methods, no method with recall >1% achieved precision >72%. Intersecting PoPS with distance increased precision from 46% for the closest gene to 79%, while achieving 31% recall.

High-confidence prioritized genes

We used PoPS, together with locus-based methods (PoPS + local), to prioritize genes across all genome-wide significant loci for 95 UK

Biobank traits and 18 additional complex diseases for which we only had summary statistics (Fig. 4c, Extended Data Fig. 8 and Supplementary Tables 9 and 10). Using the same evaluation gene set described above, we find that PoPS + local has an expected precision of 74% and should be considered the primary results from and used case for PoPS. In total, we prioritized 10,642 unique gene–trait pairs at 57% of loci in this analysis (Supplementary Table 11). Top genes by PoP score include many well-known causal genes (Fig. 5). For example, the lipid metabolism genes³⁸ *APOE*, *APOA1*, *APOB* and *PPARG* were four of the top five genes for LDL-cholesterol. For mosaic loss of Y chromosome in circulating blood³⁹, a phenotype with genetic relevance to multiple malignancies,

the top genes are involved in the DNA-damage response (*TP53*) and apoptosis (*BCL2*, *BAX*). For schizophrenia, the top genes by PoP score are the well-known dopamine receptor (*DRD2*)⁴⁰, calcium channel genes (*CACNA1C*, *CACNB2*)³², an important transcriptional regulator underlying developmental delay (*BCL11A*)^{40,41} and a transcription factor (*TCF4*) that is well studied in the context of schizophrenia⁴².

In addition we found that, when distinct locus-based methods nominate multiple different genes in a single locus, PoPS can be particularly useful to prioritize a single candidate causal gene. For example, when we identified genes supported by PoPS + local but locus-based methods disagreed, we estimated that PoPS + local nominates the correct gene 70% of the time. In comparison, nominating all genes supported by at least one of the locus-based methods results in a precision of 22%. We highlight three specific cases where previous experiments and local methods have shown that the causal variant regulates multiple genes, but only a single causal gene has been determined. For example, rs1175550 was fine-mapped with PIP > 0.9 for multiple red blood cell traits and has been demonstrated experimentally to affect the expression of *SMIM1*, *LRRCA47* and *CEP104* (ref. 43). Locus-based methods prioritized *SMIM1*, *WRAP73* and *C1orf74*. PoPS correctly prioritized *SMIM1*, which encodes for the rare Vel blood group protein involved in red blood cell production⁴⁴, for mean corpuscular hemoglobin concentration (MCH) (Fig. 6a). In another example, the variant rs737092 was fine-mapped with PIP = 0.72 for MCH, and experimental evidence shows that the expression of both *RBM38* and *RAE1* is affected by this variant⁴³ (Extended Data Fig. 9). Locus-based methods prioritized *RBM38* and *CTCF1*, but PoPS correctly prioritized *RBM38*, which has been shown to play a role in splicing key erythroid transcripts during erythropoiesis⁴³. As a final example, locus-based methods prioritized *TMEM192*, *MSMO1*, *KLHL2* and *CPE* in a locus associated with bone mineral density (BMD). PoPS correctly prioritized *CPE* at this locus (Fig. 6b), the knockout of which resulted in increased bone turnover and low BMD in mice⁴⁵.

Finally, we probed 2,004 loci where PoPS and distance to gene disagreed (Supplementary Table 12). When PoPS + local identified a gene that was not the closest to the sentinel variant, we found that PoPS + local had 60% precision. On the other hand, the closest gene, when similarly combined with a local method, had only 27% precision. Genes prioritized by PoPS + local that are not the closest gene probably represent a set enriched for new candidate genes. For example, our analysis nominated *LGR4* for estimated glomerular filtration rate (eGFR) from cystatin C (eGFRcys), a marker of kidney function (Fig. 6c). *LGR4* is near two credible sets that together define a locus containing eight genes. Three of these eight genes, including *LGR4*, have support from at least one locus-based method. *LGR4* is a G-coupled protein receptor that activates the Wnt signaling pathway and has been shown to be essential for kidney development in mice⁴⁶. For deep vein thrombosis (DVT), our analysis nominated *CCR7* at a locus with 2 credible sets and 41 genes, including 28 genes in the keratin family and 3 genes supported by locus-based methods (Fig. 6d). The top PoPS features supporting the relevance of *CCR7* were abnormal thymus medulla morphology, increased immunoglobulin (Ig)E level and response to prostaglandin. *CCR7*, chemokine receptor 7, is a regulator of inflammation that is involved in the development of DVT⁴⁷ and may be involved in thrombogenesis through platelet activation⁴⁸.

Discussion

We developed a new computational method, PoPS, for prioritizing causal genes from GWASs that predicts polygenic genetic association from gene expression profiles, PPI networks and pathway databases. We applied PoPS to summary statistics from 113 traits and showed that PoPS outperforms other similarity-based and locus-based methods. Combining PoPS with locus-based methods greatly increased precision while maintaining an adequate recall and is the approach we recommend. Using this combined approach, we nominated several genes

at unresolved GWAS loci, highlighting the utility of our approach for gene prioritization. In addition to developing PoPS, we created a large evaluation set of noncoding associations near fine-mapped variants in protein-coding genes. This set serves as a powerful tool that, unlike gene sets comprising mendelian disease genes or drug targets, allows us to evaluate both similarity-based and locus-based methods in a framework that is unbiased by previous trait-specific knowledge.

Although our similarity-based approach to gene prioritization allows for confident prediction of causal genes, it has several limitations. First, our approach assumes that causal genes share biological characteristics captured by the gene features included in the model. Causal genes that act through unrelated mechanisms or genes with shared functions that are not described by our features would not be identified by PoPS. We note a similar limitation when interpreting locus-based methods, where we cannot distinguish whether the performance of a method is limited by the methodology or the availability of the necessary data (for example, missing cell type specific data). Second, to leverage the polygenic signal, we assume that the causal mechanisms are shared between top loci and sub-significant loci. Third, although informative for ranking genes, the PoPS lacks interpretable units, is not comparable across traits and does not quantify uncertainty in the predictions. Fourth, PoPS does not directly link causal genes with their relevant cell types. Fifth, the joint linear model includes many highly correlated features, requiring ad hoc methods to interpret model fits and predictions. Finally, the large evaluation set has several limitations. As the number of noncoding associations increases, our simplistic assignment of noncoding associations to positive evaluation genes within a predefined genomic distance will probably become less accurate (Extended Data Fig. 10), suggesting that new methods are needed as sample sizes increase.

In conclusion, PoPS is a powerful tool for identifying causal genes from GWAS summary statistics and marks an important step toward building functional understanding from genetic associations. The ability to prioritize causal genes more confidently will aid in understanding the underlying trait biology and nominate genes that are strong candidates for experimental follow-up.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41588-023-01443-6>.

References

- Visscher, P. M. et al. 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
- Donnelly, P. Progress and challenges in genome-wide association studies in humans. *Nature* **456**, 728–731 (2008).
- Gallagher, M. D. & Chen-Plotkin, A. S. The post-GWAS era: from association to function. *Am. J. Hum. Genet.* **102**, 717–730 (2018).
- Reich, D. E. et al. Linkage disequilibrium in the human genome. *Nature* **411**, 199–204 (2001).
- van Arensbergen, J., van Steensel, B. & Bussemaker, H. J. In search of the determinants of enhancer-promoter interaction specificity. *Trends Cell Biol.* **24**, 695–702 (2014).
- Pers, T. H. et al. Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.* **6**, 5890 (2015).
- Hormozdiari, F. et al. Colocalization of GWAS and eQTL signals detects target genes. *Am. J. Hum. Genet.* **99**, 1245–1260 (2016).
- Gusev, A. et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).

9. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11**, e1004219 (2015).
10. Greene, C. S. et al. Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.* **47**, 569–576 (2015).
11. Fulco, C. P. et al. Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* **51**, 1664–1669 (2019).
12. Jung, I. et al. A compendium of promoter-centered long-range chromatin interactions in the human genome. *Nat. Genet.* **51**, 1442–1449 (2019).
13. Ulirsch, J. C. et al. Interrogation of human hematopoiesis at single-cell and single-variant resolution. *Nat. Genet.* **51**, 683–693 (2019).
14. Javierre, B. M. et al. Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell* **167**, 1369–1384 (2016).
15. Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
16. Liu, Y., Sarkar, A., Kheradpour, P., Ernst, J. & Kellis, M. Evidence of reduced recombination rate in human regulatory domains. *Genome Biol.* **18**, 193 (2017).
17. Fine, R. S., Pers, T. H., Amariuta, T., Raychaudhuri, S. & Hirschhorn, J. N. Benchmark: an unbiased, association-data-driven strategy to evaluate gene prioritization algorithms. *Am. J. Hum. Genet.* **104**, 1025–1039 (2019).
18. Barbeira, A. N. et al. Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. *Genome Biol.* **22**, 49 (2021).
19. Stacey, D. et al. ProGeM: a framework for the prioritization of candidate causal genes at molecular quantitative trait loci. *Nucleic Acids Res.* **47**, e3 (2019).
20. Yang, J. et al. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
21. Kanai, M. et al. Insights from complex trait fine-mapping across diverse populations. Preprint at *medRxiv* <https://doi.org/2021.09.03.21262975> (2021).
22. The 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
23. Li, T. et al. A scored human protein–protein interaction network to catalyze genomic interpretation. *Nat. Methods* **14**, 61–64 (2017).
24. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
25. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **40**, D109–D114 (2012).
26. Croft, D. et al. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.* **39**, D691–D697 (2011).
27. Blake, J. A. et al. The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse. *Nucleic Acids Res.* **42**, D810–D817 (2014).
28. Teslovich, T. M. et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
29. Wheeler, E. et al. Impact of common genetic determinants of hemoglobin A1c on type 2 diabetes risk and diagnosis in ancestrally diverse populations: a transethnic genome-wide meta-analysis. *PLoS Med.* **14**, e1002383 (2017).
30. Kurkó, J. et al. Genetics of rheumatoid arthritis—a comprehensive review. *Clin. Rev. Allergy Immunol.* **45**, 170–179 (2013).
31. Gejman, P. V., Sanders, A. R. & Duan, J. The role of genetics in the etiology of schizophrenia. *Psychiatr. Clin. North Am.* **33**, 35–66 (2010).
32. Heyes, S. et al. Genetic disruption of voltage-gated calcium channels in psychiatric and neurological disorders. *Prog. Neurobiol.* **134**, 36–54 (2015).
33. GTEx Consortium et al. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
34. Zhu, Z. et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
35. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
36. Wang, Q. S. et al. Leveraging supervised learning for functionally informed fine-mapping of cis-eQTLs identifies an additional 20,913 putative causal eQTLs. *Nat. Commun.* **12**, 3394 (2021).
37. Mountjoy, E. et al. An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. *Nat. Genet.* **53**, 1527–1533 (2021).
38. Dron, J. S. & Hegele, R. A. Genetics of lipid and lipoprotein disorders and traits. *Curr. Genet. Med. Rep.* **4**, 130–141 (2016).
39. Thompson, D. J. et al. Genetic predisposition to mosaic Y chromosome loss in blood. *Nature* **575**, 652–657 (2019).
40. Brisch, R. et al. The role of dopamine in schizophrenia from a neurobiological and evolutionary perspective: old fashioned, but still in vogue. *Front. Psychiatry* **5**, 47 (2014).
41. Basak, A. et al. *BCL11A* deletions result in fetal hemoglobin persistence and neurodevelopmental alterations. *J. Clin. Invest.* **125**, 2363–2368 (2015).
42. Quednow, B. B., Brzózka, M. M. & Rossner, M. J. Transcription factor 4 (TCF4) and schizophrenia: integrating the animal and the human perspective. *Cell. Mol. Life Sci.* **71**, 2815–2835 (2014).
43. Ulirsch, J. C. et al. Systematic functional dissection of common genetic variation affecting red blood cell traits. *Cell* **165**, 1530–1545 (2016).
44. Cvejic, A. et al. SMIM1 underlies the Vel blood group and influences red blood cell traits. *Nat. Genet.* **45**, 542–545 (2013).
45. Cawley, N. X. et al. Obese carboxypeptidase E knockout mice exhibit multiple defects in peptide hormone processing contributing to low bone mineral density. *Am. J. Physiol. Endocrinol. Metab.* **299**, E189–E197 (2010).
46. Kato, S. et al. Leucine-rich repeat-containing G protein-coupled receptor-4 (LGR4, Gpr48) is essential for renal development in mice. *Nephron Exp. Nephrol.* **104**, e63–e75 (2006).
47. Budnik, I. & Brill, A. Immune factors in deep vein thrombosis initiation. *Trends Immunol.* **39**, 610–623 (2018).
48. Lambert, M. P., Sachais, B. S. & Kowalska, M. A. Chemokines and thrombogenicity. *Thromb. Haemost.* **97**, 722–729 (2007).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2023

¹Broad Institute of MIT and Harvard, Cambridge, MA, USA. ²Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, MA, USA. ³Program in Computational & Systems Biology, MIT, Cambridge, MA, USA. ⁴Computer Science & Artificial Intelligence Lab, MIT, Cambridge, MA, USA. ⁵Division of Endocrinology and Center for Basic and Translational Obesity Research, Boston Children's Hospital, Boston, MA, USA. ⁶Department of Genetics, Harvard Medical School, Boston, MA, USA. ⁷Department of Statistics, Harvard University, Cambridge, MA, USA. ⁸Analytic and Translational Genetics Unit, MGH, Boston, MA, USA. ⁹Program in Bioinformatics and Integrative Genomics, Harvard Medical School, Boston, MA, USA. ¹⁰Department of Statistical Genetics, Osaka University Graduate School of Medicine, Suita, Japan. ¹¹MD-PhD Program, Johns Hopkins University School of Medicine, Baltimore, MD, USA. ¹²Division of Gastroenterology, Hepatology, and Nutrition, Boston Children's Hospital, Boston, MA, USA. ¹³Program in Immunology, Harvard Medical School, Boston, MA, USA. ¹⁴Harvard Stem Cell Institute, Cambridge, MA, USA. ¹⁵Department of Molecular Biology, MGH, Boston, MA, USA. ¹⁶Institute for Medical Engineering and Science, MIT, Cambridge, MA, USA. ¹⁷Department of Chemistry, MIT, Cambridge, MA, USA. ¹⁸Koch Institute for Integrative Cancer Research, MIT, Cambridge, MA, USA. ¹⁹Ragon Institute of MGH, MMIT, Cambridge, MA, USA. ²⁰Gastrointestinal Unit and Center for the Study of Inflammatory Bowel Disease, MGH, Boston, MA, USA. ²¹Center for Computational and Integrative Biology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA. ²²Howard Hughes Medical Institute, MIT, Cambridge, MA, USA. ²³Division of Cardiovascular Medicine and Genetics, Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA. ²⁴Harvard Medical School, Boston, MA, USA. ²⁵Department of Surgery, MGH, Boston, MA, USA. ²⁶Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA. ²⁷Department of Pediatrics, Harvard Medical School, Boston, MA, USA. ²⁸Department of Biology, MIT, Cambridge, MA, USA. ²⁹Department of Systems Biology, Harvard Medical School, Boston, MA, USA. ³⁰Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA. ³¹BASE Initiative, Betty Irene Moore Children's Heart Center, Lucile Packard Children's Hospital, Stanford University School of Medicine, Stanford, CA, USA. ³²Present address: Artificial Intelligence Laboratory, Illumina, Inc., San Diego, CA, USA. ³³Present address: Vertex Pharmaceuticals Incorporated, Boston, MA, USA. ³⁴Present address: Bristol Myers Squibb, Cambridge, MA, USA. ³⁵Present address: Department of Biological Regulation, Weizmann Institute of Science, Rehovot, Israel. ³⁶Present address: Genentech, San Francisco, CA, USA. ³⁷These authors contributed equally: Elle M. Weeks, Jacob C. Ulirsch. ✉e-mail: eweeks@broadinstitute.org; finucane@broadinstitute.org

Methods

MAGMA gene z-scores

We applied MAGMA⁹ to the summary statistics for each trait using EUR individuals from the 1000 Genomes Project reference panel²² to compute gene-level association statistics and gene–gene correlations using the SNP-wise mean gene analysis and a 0-kb window around the gene body for mapping SNPs to genes. For each gene, MAGMA computes a gene *P* value from the mean χ^2 statistic of SNPs in the gene body and its approximate sampling distribution. The gene *P* value is converted to a z-score using the probit function. The resulting z-score reflects the gene–trait association after correcting for LD among SNPs within the gene body. MAGMA approximates the gene–gene correlation matrix, *R*, using the correlations between the model sum of squares of each gene pair under the joint null hypothesis of no association. These correlations are a function of the LD between SNPs in each pair of genes and represent the LD at a gene level. To ensure a well-conditioned positive definite correlation matrix, we add a small value to the entries of *R* along the diagonal. Specifically, we add $|\min(\lambda_{\min}, 0)| + 0.05 + 0.9 \times \max(0, \text{var}(Y) - 1)$ to each element along the diagonal, where λ_{\min} is the minimum eigenvalue of *R* and *Y* is the MAGMA z-score.

PoPS covariates

We included covariates corresponding to gene density, effective gene size and inverse of the mean minor allele count (MAC) of SNPs in the gene, as well as the log of these variables as computed by the MAGMA⁹ software. MAGMA defines gene density as the ratio of the effective number of independent SNPs in the gene to the total number of SNPs in the gene, and defines effective gene size as the effective number of independent SNPs in the gene. In addition, we include a covariate corresponding to gene size and the log of this variable, defined as the length of the gene in base pairs.

Locus definition

From the set of associated variants with $P < 5 \times 10^{-8}$, we designated independent lead variants from which to define loci. For the 18 traits where we used publicly available summary statistics, we performed PLINK⁴⁹ clumping using EUR individuals in the 1000 Genomes Project reference panel with a *P*-value threshold of 5×10^{-8} and r^2 threshold of 0.1. Within each clump, we defined the variant with the most significant *P* value as the lead variant. For the 95 traits from the UK Biobank, where we had fine-mapping results for regions containing genome-wide significant variants, we defined one locus for every independent credible set (CS). For each fine-mapped CS, we defined the variant with the highest posterior inclusion probability (PIP) as the lead variant. We then defined the locus boundaries as 500 kb on either side of the lead variant and included all genes that fell within or overlapped with the locus boundaries (see Extended Data Fig. 10 for sensitivity to boundary size).

Complex traits and disease associations

GWAS analyses for 95 heritable traits in the UK Biobank were performed as part of a fine-mapping study²¹. Up to 361,194 individuals of white British ancestry with available phenotypes and variants, with INFO > 0.8, minor allele frequency > 0.01% and Hardy–Weinberg equilibrium *P* value > 1×10^{-10} , were included in the GWASs. Covariates for the top 20 principal components (PCs), sex, age, age², sex × age, sex × age² and dilution factor, where applicable, were controlled for in the association studies. Quantitative traits were inverse rank transformed and associations were estimated using BOLT-LMM⁵⁰. For binary traits, associations were estimated using SAIGE⁵¹. Publicly available summary statistics were downloaded for an additional 18 diseases (Supplementary Table 1).

Gene features

We created gene features from three main data types: (1) bulk and single-cell gene expression datasets, (2) curated biological pathways and (3) predicted PPI networks.

Bulk and single-cell gene expression datasets. For each of the 77 gene expression datasets (Supplementary Table 2), we uniformly reprocessed the raw count (or normalized count when raw counts were not provided) matrices using Seurat v.3 (ref. 52). First, cells with total counts outside of the 5th to 95th quantiles were removed, and only the 18,383 protein-coding genes used in the PoPS analysis were included. Counts were then scaled to counts per million, log(normalized) and scaled such that each gene had a mean of 0 and variance of 1 across cells. PCs and gene loadings were computed on scaled expression values for the top 1,000–3,000 variable genes using truncated singular value decomposition (SVD)⁵³. Independent components and gene loadings were computed using fastICA⁵⁴. A *k*-nearest neighbor graph was created using the top PCs (based on inspection of elbow plot) and clusters were identified using the Louvain algorithm. The uniform manifold approximation and projection (UMAP) algorithm⁵⁵ was used to visualize clusters and investigate batch effects. When batch effects were visually apparent and predefined batch annotations were provided, we attempted to remove batch effects using the anchor approach in Seurat. Finally, we performed differential expression between clusters using a one-versus-all approach with a two-sided Welch's *t*-test. We provide code to reproduce these analyses, a repository of processed features and visualizations of the top derived features at https://github.com/FinucaneLab/gene_features.

We then derived features for PoPS: (1) on the whole dataset, (2) within clusters representing different cell populations and (3) between clusters. (1) On the whole dataset, we derived features of gene loadings from PC analysis and gene loadings from independent component analysis. (2) Within each cluster, either predefined (when available) or identified in our analysis, we derived features of average scaled gene expression and gene loadings from the top 10 PCs. (3) Comparing across clusters (one versus all), we derived features of a *t*-test statistic for differential expression and a binary indicator for differentially up- and downregulated genes (Benjamini–Hochberg's false discovery rate (FDR) < 0.05 and $|\log_2(\text{fold-change})| > 2$).

Curated biological pathways. We created features from biological pathways curated for DEPICT from *Kyoto Encyclopedia of Genes and Genomes*²⁵, gene ontology²⁴, Reactome²⁶ and the Mouse Genome databases²⁷. Each feature was encoded as a binary indicator for membership to a pathway.

Predicted PPI networks. We created features using the predicted InWeb_IM PPI network²³. For each gene, we included as a feature a binary indicator for the set of genes that were its first-degree neighbors.

Finally, for each distinct dataset, we included a control feature as a binary indicator for the set of genes that were reported in that dataset. All features were centered and scaled to have a mean of 0 and a variance of 1 across genes.

DEPICT

We ran DEPICT⁶ with default parameters on the summary statistics for each trait and DEPICT's 14,461 reconstituted gene sets to prioritize genes in genome-wide significant loci. First, we performed PLINK clumping with a *P*-value threshold of 5×10^{-8} , r^2 threshold of 0.05 and distance threshold of 500 kb, as recommended by the DEPICT software. Loci are defined by taking all genes that reside within boundaries defined by the most distal variants in either direction with LD > 0.5 to the lead variant, identified by PLINK clumping. To make running DEPICT computationally tractable for traits with large numbers of genome-wide significant loci, we restricted the input to the top 1,000 loci by the *P* value of the index variant. DEPICT then scores genes by correlating their membership to reconstituted gene sets to those of other genes in genome-wide significant loci and performs a bias adjustment for the scores. Finally, to prioritize genes in each locus, we prioritized the single gene in each genome-wide significant locus

with the most significant P value. For Benchmarker and closest gene enrichment analyses, DEPICT was run in a LOCO framework where, all variants on the chromosome for which gene P values were computed were removed from the summary statistics before running DEPICT.

NetWAS

NetWAS¹⁰ trains a support vector machine classifier constructed using a gene network. We applied NetWAS using the global network and MAGMA gene P values generated from the summary statistics for each trait and the 1000 Genomes Project reference panel²². We applied NetWAS using both the default threshold of $P < 0.01$ and Bonferroni's significance threshold, which was shown in previous work to have better performance for well-powered GWASs¹⁷. In cases where fewer than 15 genes passed the significance threshold, we relaxed the P -value threshold until there were 15 passing genes. For Benchmarker and closest gene enrichment analyses, NetWAS was run in a LOCO framework where, all genes on the chromosome for which gene scores were being computed were removed from the MAGMA gene-level associations before running NetWAS.

MAGMA-sim

MAGMA-sim¹⁷, described by Fine et al. and referred to in that manuscript as MAGMA, is an approach for leveraging individual gene-set enrichments to prioritize genes, by prioritizing all genes that are members of the most highly enriched gene sets. To run MAGMA-sim, we computed gene set enrichment P values for the 14,461 reconstituted gene sets from DEPICT using MAGMA. Using the best performing approach from Fine et al., we binarized the reconstituted gene sets using a z -score threshold of $z > 2.58$ on the reconstituted gene sets. We then constructed a set of prioritized genes by: (1) ranking the gene sets by enrichment P value and (2) adding member genes of the most significant gene sets until we reached 500 prioritized genes. If the last added gene set contained more genes than were necessary to reach 500 prioritized genes, we selected the required number of genes from that last gene set at random. For precision-recall analyses, MAGMA-sim was run leaving out the coding signal.

To validate MAGMA-sim in a LOCO framework, we ranked gene sets according to enrichment P values that were computed after removing genes from the test chromosome. We then added member genes of the most significant gene sets until we reached K prioritized genes, where $K = 500 \times (\text{percentage of genes on test chromosome})$, thus prioritizing 500 total genes across all chromosomes. MAGMA-sim is similar to PoPS in that it leverages genome-wide gene set enrichments from MAGMA for gene prioritization. However, MAGMA-sim does not perform any joint modeling of these gene sets, instead ranking them and prioritizing all genes in any of the top-ranked gene sets. Moreover, MAGMA-sim provides only a binary result for each gene—prioritized or not prioritized—without any ranking or quantitative score, and so can prioritize zero or multiple genes in a locus. We thus could not apply our closest gene evaluation metric to MAGMA-sim.

Benchmarker

Benchmarker¹⁷ is an unbiased, data-driven approach to evaluate gene prioritization methods. Based on the assumption that SNPs near causal genes should be enriched for trait heritability, Benchmarker uses stratified LD-score regression (S-LDSC)⁵⁶ to estimate the average contribution of SNPs near prioritized genes to per-SNP heritability. Using S-LDSC, Benchmarker jointly models an SNP annotation corresponding to prioritized genes, along with 53 annotations in the 'baseline model', which include genic, regulatory and conserved regions. To evaluate performance, we use the regression coefficient, τ , and its P value for the hypothesis $\tau > 0$; τ measures the contribution of SNPs near prioritized genes to per-SNP heritability after controlling for the baseline annotations. To make τ comparable across traits, we normalized τ by the average per-SNP heritability for each trait and refer to this quantity as normalized τ .

For our analyses, we selected the 500 genes with the highest PoP scores for each trait as the set of prioritized genes and used a 100-kb window on either side of the transcription start site of each gene for mapping SNPs to genes.

Closest gene enrichment

We used a normal approximation to the null distribution for our test statistic, c , the number of genes that are PoPS prioritized and the closest gene to the lead variant in a locus. Under the null, PoPS prioritizes the closest gene in a locus at random with probability $\frac{1}{n_l}$, where n_l is the number of genes in a locus, l . Across all L loci, the distribution of c under this null is a sum of independent Bernoulli's with different biases. For computational tractability when L is large, we approximate this by a normal (N) distribution with matched moments:

$$c \sim N\left(\sum_{l:1}^L \frac{1}{n_l}, \sqrt{\sum_{l:1}^L \frac{1}{n_l} \left(1 - \frac{1}{n_l}\right)}\right).$$

We performed a one-sided test for $c > \sum_{l:1}^L \frac{1}{n_l}$ under the null. In addition, we computed the enrichment of the number of PoPS-prioritized genes that are the closest to the ratio of the observed to the expected, $\frac{c}{\sum_{l:1}^L \frac{1}{n_l}}$, and estimated the s.e. of the enrichment. We used the bootstrap to estimate the s.e. of the enrichments, not assuming a null distribution, and performed 1,024 bootstrap repetitions resampling the L loci for each trait.

Independent traits

To identify independent traits, we first computed genetic correlations between all pairs of traits using cross-trait LD-score regression⁵⁷ with LD scores from UK10K⁵⁸. Next, we created an adjacency matrix of traits with edge weights corresponding to whitened ($|r_g| < 0.2$ was set to 0), absolute genetic correlations. We then identified the maximum independent set of vertices (traits) such that no two were adjacent using the igraph package⁵⁹ in R v.3.5. The resulting set contained 46 independent traits (Supplementary Table 1).

Feature clustering

For each trait, 50 PCs were derived from the scaled gene by feature matrix using truncated SVD. A feature by feature distance matrix was then created as the dissimilarity between features using $1 - r^2$ (the squared Pearson correlation) between PCs. Complete linkage hierarchical clustering was then performed on this distance matrix. Clusters were determined such that Pearson $r^2 > 0.1^2$ for all features within a cluster. This inclusive threshold was chosen to reduce the impact of multicollinearity when interpreting the contribution of top clusters to PoP scores and was validated by manual investigation of within-dataset composition of large clusters as well as biological interpretability of the top clusters.

Fine-mapping

Fine-mapping was performed²¹ for 95 complex traits in the UK Biobank and 49 tissues in GTEx v.8 using the Sum of Single Effects (SuSiE) method⁶⁰, allowing for up to 10 causal variants in each region. Prior variance and residual variance were estimated using the default options and single effects (potential 95% CSs) were pruned using the standard purity filter such that no pair of variants in a CS could have $r^2 > 0.25$. Regions were defined for each trait as ± 1.5 Mb around the most significantly associated variant and overlapping regions were merged. As inputs to SuSiE, summary statistics for each region were obtained using BOLT-LMM⁵⁰ for quantitative traits and SAIGE⁵¹ for binary traits, in sample dosage LD was computed using LDStore⁶¹ and phenotypic variance was computed empirically. Variants in the major histocompatibility complex region (chr6: 25–36 Mb) were excluded, as were 95% CSs containing variants with < 100 MACs. Coding (missense and

predicted loss of function) variants were annotated using the Variant Effect Predictor v.85 (ref. 62). Fine-mapping data used in the present study are available at <https://www.finuanelab.org/data>.

Precision and recall

We used our evaluation gene set to estimate the precision and recall for each method, evaluating the following two questions: first, if a gene is prioritized, how confident should we be that it is truly relevant? Second, what proportion of all truly relevant genes does the method prioritize? To answer these questions for a given method, we applied the method to the 1,348 loci with fine-mapped coding variants, excluding the nearby coding signal where relevant (see below). A true positive (TP) is a prioritized gene that is condition positive, a false positive (FP) is a prioritized gene that is condition negative, a true negative (TN) is a gene that is not prioritized and is condition negative and a false negative (FN) is a gene that is not prioritized and is condition positive. The answers to our two questions are given, respectively, by precision, no. of TPs/(no. of TPs + no. of FPs), and recall, no. of TPs/(no. of TPs + no. of FNs).

ABC-Max

We used the ABC model¹¹ to predict enhancer–gene connections in 131 biosamples from 74 distinct cell types and tissues, based on measurements of chromatin accessibility (assay for transposase-accessible chromatin with sequencing or DNase-seq) and histone modifications (H3K27ac chromatin immunoprecipitation sequencing). For each trait, we included only predicted enhancer–gene connections where the enhancer contained a fine-mapped variant (PIP > 0.1) in a credible set that did not contain any coding or splice-site variants. We assigned each gene in a locus a single score for the corresponding fine-mapped CS by taking the highest ABC score of predicted enhancers for that gene–CS pair across all biosamples that are enriched for overlapping fine-mapped variants for that trait. Finally, to predict a single gene for each credible set, ABC-Max prioritizes the gene with the highest ABC score in the locus.

E–P correlation

We downloaded predicted E–P maps based on the correlation of biochemical marks at regulatory regions and expression of nearby genes across cell types for 808 tissues and cell lines from the FANTOM5 project¹⁵, 127 tissues and cell lines from the ROADMAP Epigenomics project¹⁶ and 16 primary blood cell types¹³. For the FANTOM5 dataset, we filtered for interactions with Benjamini–Hochberg’s FDR < 10^{−5} for a nonzero Pearson’s correlation. For the ROADMAP dataset, we filtered for interactions with a confidence score > 2.5. For the Ulirsch et al. dataset, we filtered for interactions with Pearson’s correlation > 0.7 and a Storey FDR < 10^{−4}. Finally, for each trait, we included only predicted interactions where the enhancer contained a fine-mapped variant (PIP > 0.1). We assigned each gene in a locus a single score for each corresponding fine-mapped CS by taking the highest confidence score or correlation of predicted enhancers for that gene–CS pair across all tissues and cell lines.

PCHi-C

We downloaded PCHi-C datasets containing observed physical interactions between fragmented DNA and targeted genic promoters for 28 diverse human tissues and cell lines¹² and 15 primary blood cell types¹⁴. For the Jung et al.¹² dataset, we filtered to interactions with *P* values for interaction < 0.01 and raw frequency counts > 5. For the Javierre et al.¹⁴ dataset, we filtered to interactions with CHiCAGO⁶³ scores > 5. In both cases, we defined a variant–gene interaction as a variant with PIP > 0.10 overlapping with a relevant region of accessible chromatin, based on 39 ROADMAP tissues⁶⁴ for Jung et al. and 44 primary blood cell types^{13,65} for Javierre et al. Finally, for each trait, we included only predicted interactions where the enhancer contained a fine-mapped variant (PIP > 0.1). We assigned each gene in a locus a single score for each corresponding fine-mapped CS by taking the highest connection strength of predicted enhancers for that gene–CS pair across all tissues and cell lines.

TWAS

We applied TWAS⁸ using the FUSION software package and precomputed expression reference weights for 48 tissues from GTEx v.7 (ref. 33). To avoid leveraging the coding signal for the precision-recall analysis, we excluded all variants in LD ($r^2 > 0.2$ to a coding variant with PIP > 0.1). For all other analyses, we included all variants in the GWAS summary statistics. In both cases, we took the most significant association across tissues for each gene. For precision-recall analyses, a TWAS was run leaving out the coding signal. TWAS weights were obtained from <http://gusevlab.org/projects/fusion/weights/GTEX7.txt>.

SMR

We applied SMR³⁴ using the SMR software tool and precomputed *cis*-eQTL summary data across 48 human tissues from GTEx v.7 (ref. 33). The *cis*-eQTL summary data were prefiltered to SNPs within 1 Mb of the transcription start site for each gene. For precision-recall analyses, SMR was run leaving out the coding signal.

Colocalization posterior probability

Using fine-mapping results for 95 complex traits from the UK Biobank and for eQTLs in 49 tissues from GTEx v.8 (ref. 35), we computed CLPPs, analogous to those reported by the eCAVIAR software⁷. For each variant, *i*, fine-mapped for a complex trait, *g*, and an eQTL trait, *e*, the CLPP was computed as $P(C_{ig}, C_{ie}) = P(C_{ig})P(C_{ie})$, where $P(C_{ig})$ is the PIP of variant *i* in complex trait *g* and $P(C_{ie})$ is the PIP of variant *i* in eQTL trait *e*. This quantity is an estimate of the probability that the variant is causal for both the complex trait and the gene expression trait. Within each fine-mapped CS and for each gene, we took the maximum CLPP across all variants and GTEx tissues.

Gene-prioritization criteria

To prevent information leakage from coding variant associations, which are used as part of the evaluation set, into noncoding variant-gene prioritizations, all variants in LD ($r^2 > 0.2$) with a trait-associated coding variant (PIP > 0.1) were removed before running PoPS, DEPICT, NETWAS, MAGMA-sim, TWAS and SMR for evaluations on the set of 1,348 loci containing a fine-mapped, protein-coding variant used as a positive label. We evaluated multiple prioritization criteria for each locus-based method and PoPS including various absolute thresholds and the relative rank of genes within a locus (Extended Data Fig. 5 and Supplementary Table 6). We chose the following prioritization criteria to maximize precision:

- (1) E–P correlation, PCHi-C, ABC-Max: for each locus such that at least one gene has a predicted connection with an enhancer containing a variant with PIP > 0.1, the gene that has the highest correlation or connection score. To combine across datasets for E–P correlation and PCHi-C, we included any gene prioritized in at least one dataset.
- (2) TWAS: for each locus such that at least one gene is significantly associated after Bonferroni’s correction, the gene with the most significant *P* value.
- (3) SMR: for each locus such that at least one gene is significantly associated after Bonferroni’s correction, the gene with the most significant *P* value.
- (4) CLPP: for each locus such that at least one gene has a variant with CLPP > 0.1, the gene with the highest CLPP.
- (5) Distance: for each locus, the gene that is closest to the lead variant by distance to the gene body.
- (6) PoPS: for each locus, the gene that has the highest PoP score.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

A repository of processed gene features, visualizations of top derived features and code to reproduce these analyses are available on GitHub at https://github.com/FinucaneLab/gene_features. Complete PoPS results for 95 complex traits in the UK Biobank and 18 additional disease traits, as well as results for PoPS and locus-based methods in genome-wide significant loci, are available at <https://www.finucanelab.org/data>.

Code availability

PoPS is available as an open-source Python package at <https://github.com/FinucaneLab/pops>. A static version of the PoPS method used in the present study is available at <https://doi.org/10.5281/zenodo.8002379>.

References

49. Purcell, S. et al. PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
50. Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association for biobank-scale datasets. *Nat. Genet.* **50**, 906–908 (2018).
51. Zhou, W. et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341 (2018).
52. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).
53. Baglama, J. & Reichel, L. Restarted block Lanczos bidiagonalization methods. *Numer. Algorithms* **43**, 251–272 (2007).
54. Hyvärinen, A. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Netw.* **10**, 626–634 (1999).
55. McInnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at <https://doi.org/10.48550/arXiv.1802.03426> (2018).
56. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
57. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
58. UK10K Consortium et al. The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).
59. Csárdi, G. & Nepusz, T. The igraph software package for complex network research. *Int. J. complex syst.* **1695**, 1–9 (2006).
60. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Series B Stat. Methodol.* **82**, 1273–1300 (2020).
61. Benner, C. et al. Prospects of fine-mapping trait-associated genomic regions by using summary statistics from genome-wide association studies. *Am. J. Hum. Genet.* **101**, 539–551 (2017).
62. McLaren, W. et al. The Ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
63. Cairns, J. et al. CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. *Genome Biol.* **17**, 127 (2016).
64. Roadmap Epigenomics Consortium et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
65. Calderon, D. et al. Landscape of stimulation-responsive chromatin across diverse human immune cells. *Nat. Genet.* **51**, 1494–1505 (2019).

Acknowledgements

We thank K. Aragam, A. Butterworth, M. Daly, N. Artomov, Y. Reshef and all members of the Finucane lab for helpful discussions. This

research was conducted using the UK Biobank Resource under project 31063. H.K.F. was funded by a National Institutes of Health (NIH) grant (no. DP5 OD024582) and by Eric and Wendy Schmidt. J.M.E. was supported by a Pathway to Independence Award (grant nos. K99HG00917 and R00HG009917), the Harvard Society of Fellows and the Base Research Initiative at Stanford University. J.M. and J.N.H. were supported by an NIH grant (no. R01DK075787). R.S.F. was supported by National Human Genome Research Institute, NIH (grant no. F31HG009850). J.O.-M. was supported by the Richard and Susan Smith Family Foundation, the HHMI Damon Runyon Cancer Research Foundation Fellowship (no. DRG-2274-16), the AGA Research Foundation's AGA-Takeda Pharmaceuticals Research Scholar Award in Inflammatory Bowel Disease (grant no. AGA2020-13-01), the HDDC Pilot and Feasibility (grant no. P30 DK034854) and the Food Allergy Science Initiative.

Author contributions

E.M.W. and H.K.F. conceived of the study. E.M.W., J.C.U., N.Y.C. and H.K.F. designed the research, performed the experiments, analyzed the data and interpreted the results. B.L.T. and R.S.F. designed and performed the enrichment-based validations. J.M., T.A.P., M.K., J.N., C.P.F., K.C.T., F.A., T.L., J.O.-M., C.S.S., M.B., A.K.S., A.N.A., R.J.X., A.R., R.M.G., K.L., K.G.A., J.N.H. and J.M.E. provided data or analysis tools used by PoPS or other gene prioritization methods. E.S.L. helped advise the project. E.M.W., J.C.U. and H.K.F. wrote the manuscript with input from all authors. H.K.F. supervised the project.

Competing interests

J.C.U. reports compensation from consulting services with Goldfinch Bio and is an employee of Illumina. R.S.F. is an employee of Vertex Pharmaceuticals Incorporated. C.P.F. is an employee of Bristol Myers Squibb. J.O.-M. reports compensation for consulting services with Cellarity. A.R. is a cofounder and equity holder of Celsius Therapeutics and an equity holder in Immunitas, and was an SAB member of Thermo Fisher Scientific, Syros Pharmaceuticals, Neogene Therapeutics and Asimov until 31 July 2020. From 1 August 2020, A.R. is an employee of Genentech. J.N.H. served on the Scientific Advisory Board of and consults for Camp4 Therapeutics. E.S.L. serves on the Board of Directors for Codiak BioSciences and Neon Therapeutics, and serves on the Scientific Advisory Board of F-Prime Capital Partners and Third Rock Ventures; he is also affiliated with several nonprofit organizations including serving on the Board of Directors of the Innocence Project, Count Me In and Biden Cancer Initiative, and the Board of Trustees for the Parker Institute for Cancer Immunotherapy. He has served and continues to serve on various federal advisory committees. The remaining authors declare no competing interests.

Additional information

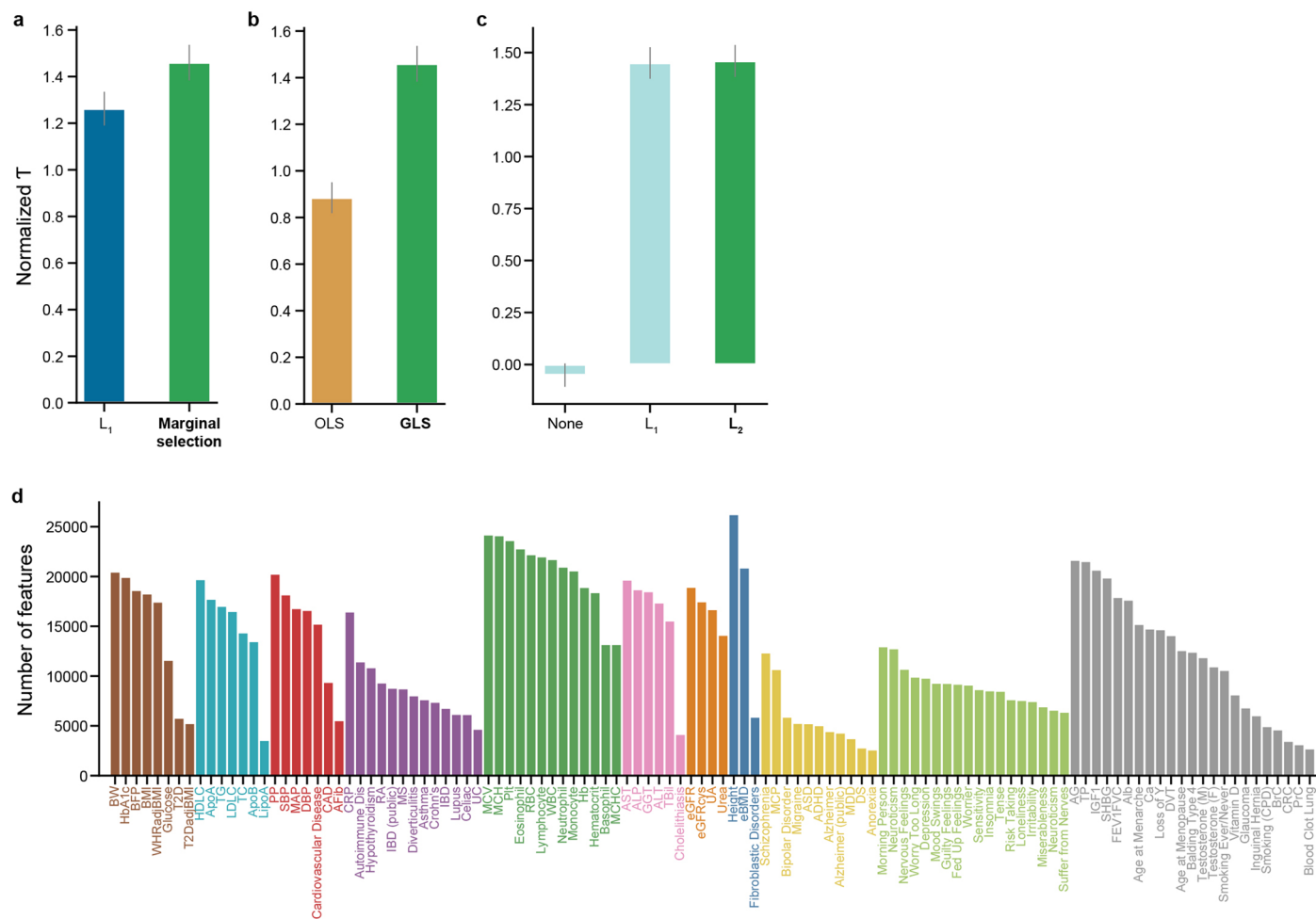
Extended data is available for this paper at <https://doi.org/10.1038/s41588-023-01443-6>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41588-023-01443-6>.

Correspondence and requests for materials should be addressed to Elle M. Weeks or Hilary K. Finucane.

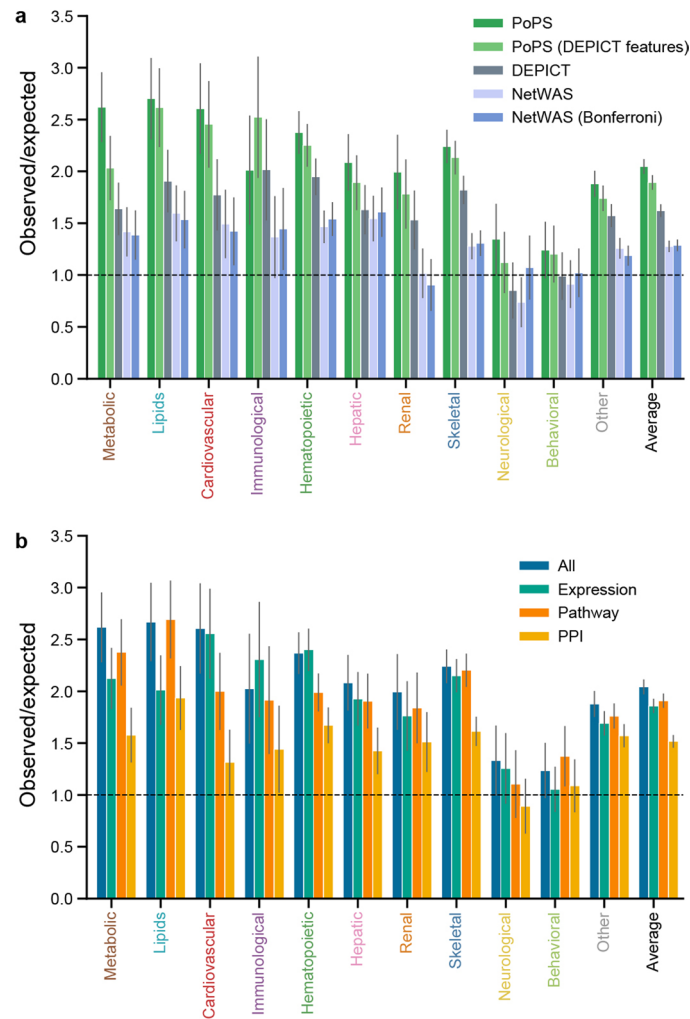
Peer review information *Nature Genetics* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.



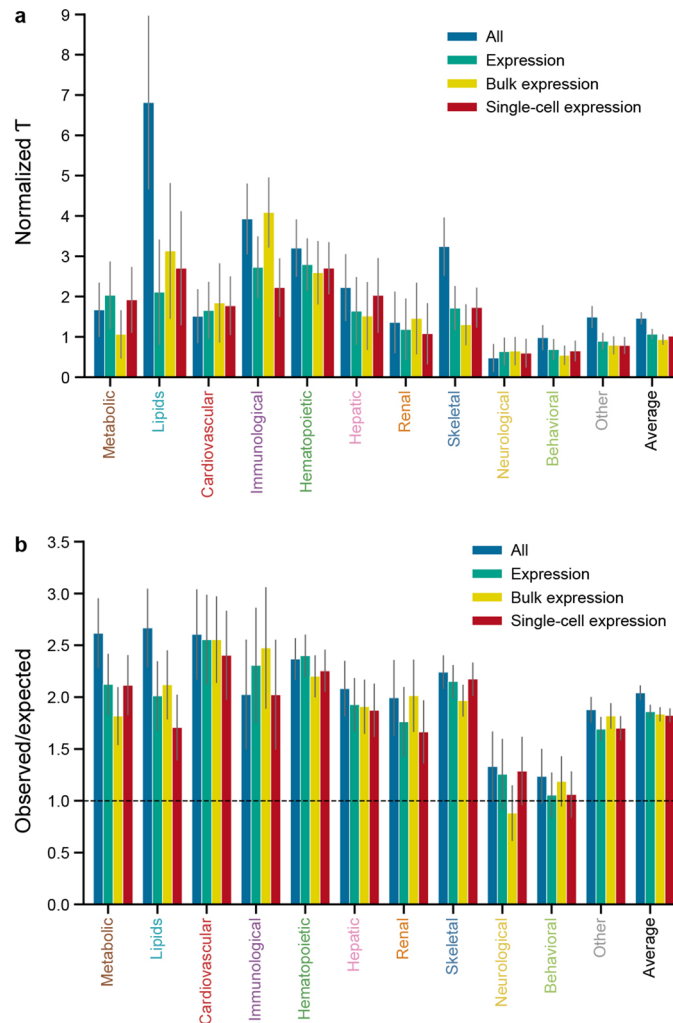
Extended Data Fig. 1 | PoPS model parameter choices and feature selection. **a-c**, Results using Benchmarker to compare different parameter choices for fitting the PoPS model, meta-analyzed across independent traits ($n = 46$). Error bars represent 95% confidence intervals around the meta-analyzed point estimate. **a**, Feature selection: GLS with an L_1 penalty on the full set of features performs less well than GLS after marginal selection using a P value < 0.05 threshold from the two-sided Wald test. **b**, Error model: ordinary least squares

(OLS) performs less well than generalized least squares (GLS) using marginal selection from **a, c**. Joint model regularization: GLS after marginal feature selection with an L_2 penalty performs better than similar models with an L_1 penalty or no penalty. **d**, Number of features selected (marginal P value < 0.05 from the two-sided Wald test) and included in the joint predictive model for PoPS for each trait. A legend for trait domain colors is provided in Fig. 2.

**Extended Data Fig. 2 | Additional comparisons using closest gene metric.**

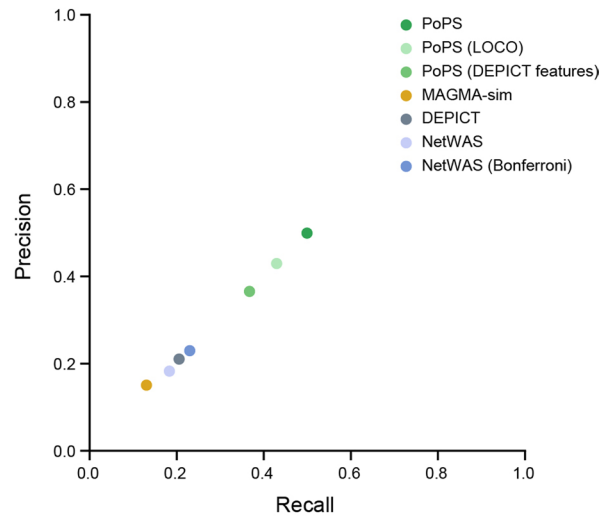
a, Results using closest gene enrichment to compare similarity-based gene prioritization methods, meta-analyzed within each trait domain across independent traits ($n = 46$). Error bars represent 95% confidence intervals around

the meta-analyzed point estimate. **b**, Results using closest gene enrichment to compare PoPS results using different feature sets, meta-analyzed within each trait domain across independent traits ($n = 46$). Error bars represent 95% confidence intervals around the meta-analyzed point estimate.

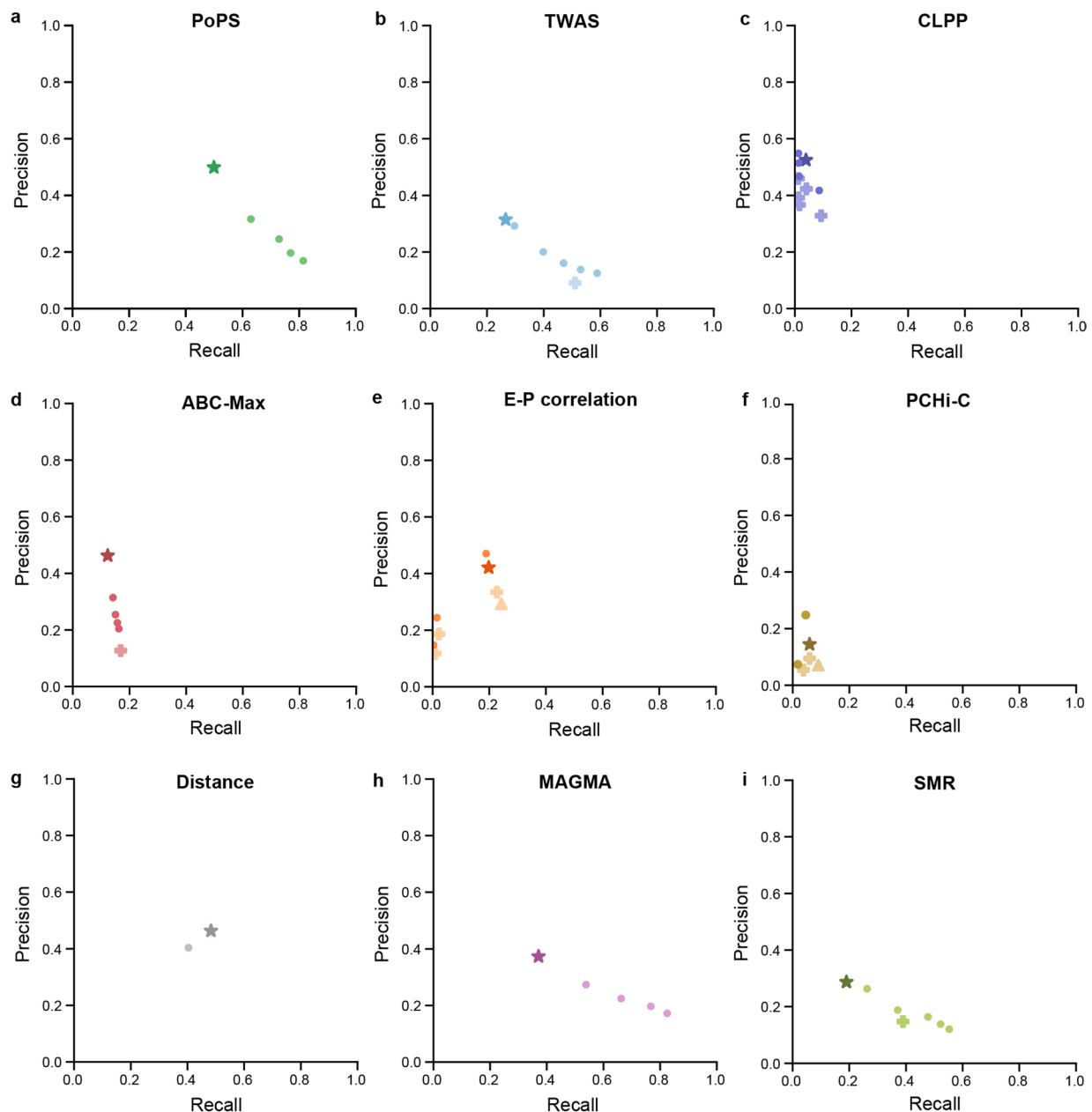


Extended Data Fig. 3 | Comparison of gene expression features derived from bulk and single-cell RNA seq datasets. a, Results using Benchmarker to compare PoPS results using different feature sets, meta-analyzed within each trait domain across independent traits ($n = 46$). Error bars represent 95% confidence intervals around the meta-analyzed point estimate. **b,** Results

using closest gene enrichment to compare PoPS results using different feature sets, meta-analyzed within each trait domain across independent traits ($n = 46$). Error bars represent 95% confidence intervals around the meta-analyzed point estimate.

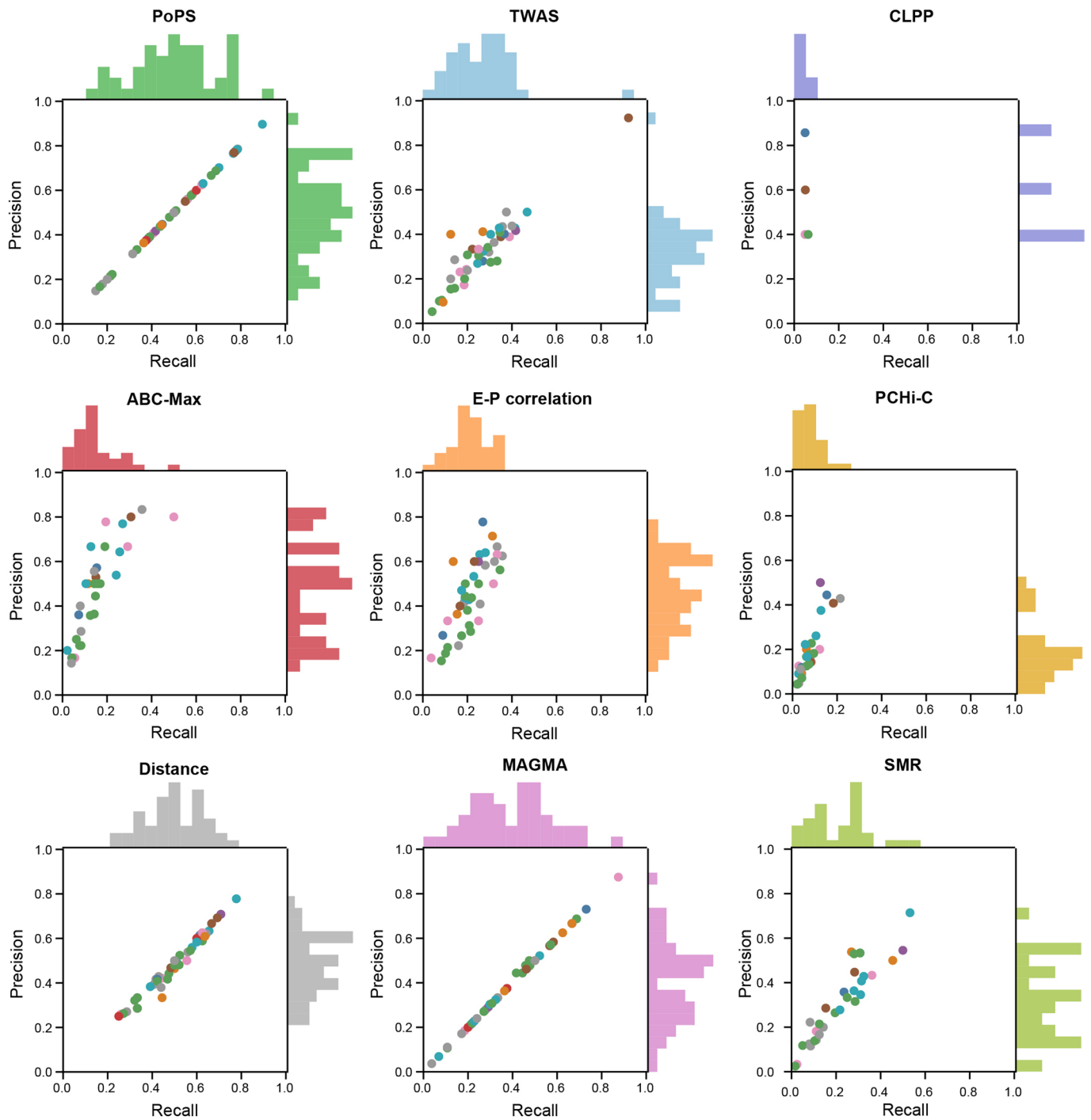


Extended Data Fig. 4 | Comparison of similarity-based methods using precision and recall. Precision-recall plot showing performance of similarity-based methods.

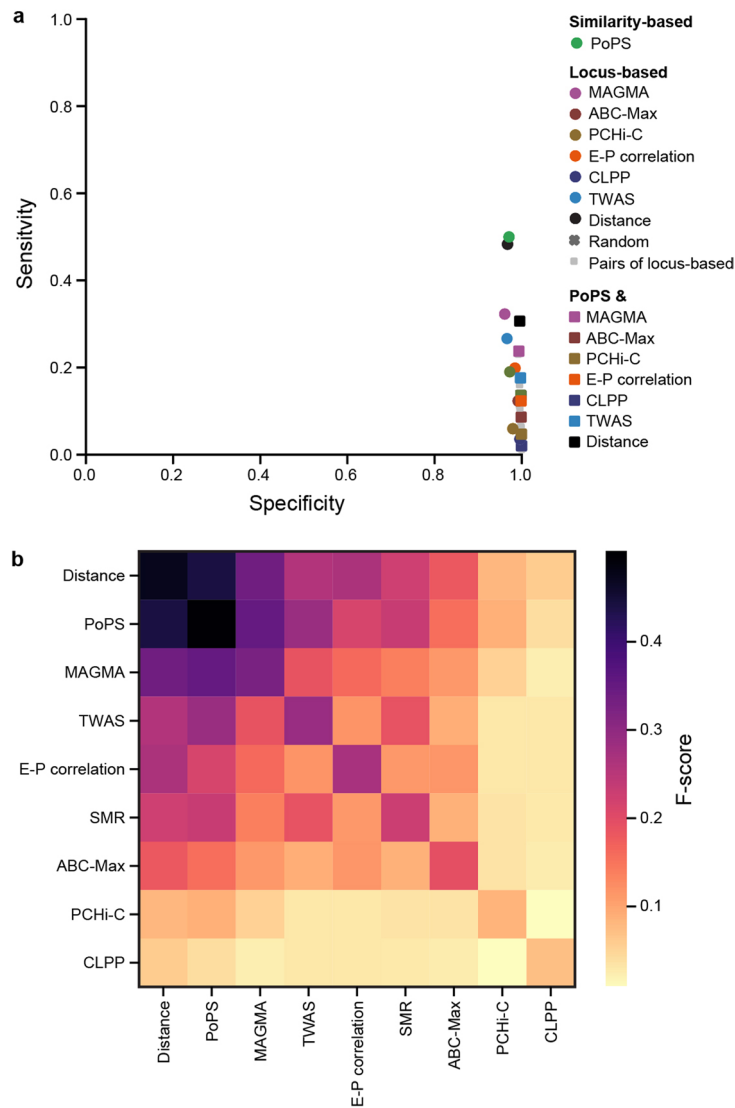


Extended Data Fig. 5 | Comparing prioritization criteria. Precision-recall plots for each method with varying prioritization criteria. Each point shows the precision and recall for a set of prioritized genes selected using prioritization criteria based on absolute thresholds and/or relative rank in a locus. For all methods, the star represents the final chosen criteria. **a**, Circles: PoPS scores ranked $\leq 2-5$ in the locus. Star: highest PoPS score in the locus. **b**, Plus: significant TWAS P value after Bonferroni correction ($P < 0.05/235,584$). Circles: TWAS P values ranked $\leq 2-5$ in the locus. Star: significant TWAS P value after Bonferroni correction ($P < 0.05/235,584$) and the most significant in the locus. **c**, Pluses: CLPP $> 0.01, 0.1, 0.5, 0.9$, and 0.99 . Circles: CLPP $> 0.01, 0.1, 0.5, 0.9$, and 0.99 and also the highest CLPP in the locus. Star: CLPP > 0.1 and also the highest CLPP in the locus. **d**, Plus: any predicted connection from ABC. Circles: ABC connection strength ranked $\leq 2-5$ in the locus. Star: highest ABC connection strength in the

locus. **e**, Pluses: any predicted connection from PCHiC for individual datasets. Triangle: any predicted connection from PCHi-C in any dataset. Circles: highest connection strength in the locus for individual datasets. Star: highest connection strength in the locus in any dataset. **f**, Pluses: any predicted connection from E-P correlation for individual datasets. Triangle: any predicted connection from E-P correlation in any dataset. Circles: highest connection strength in the locus for individual datasets. Star: highest connection strength in the locus in any dataset. **g**, Circle: closest gene by distance to the transcription start site. Star: closest gene by distance to the gene body. **h**, Circles: MAGMA z -scores ranked $\leq 2-5$ in the locus. Star: highest MAGMA score in the locus. **i**, Plus: significant SMR P value after Bonferroni correction ($P < 0.05/18,383$). Circles: SMR P values ranked $\leq 2-5$ in the locus. Star: significant SMR P value after Bonferroni correction ($P < 0.05/18,383$) and the most significant in the locus.

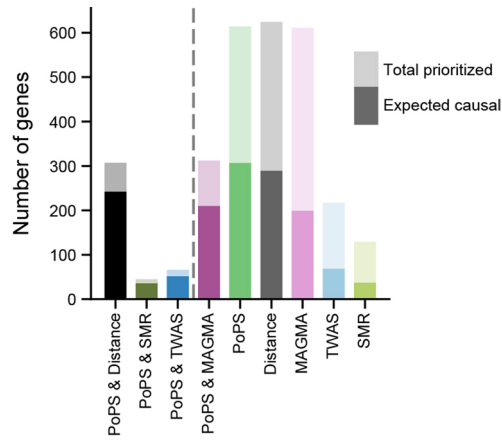


Extended Data Fig. 6 | Performance of PoPS and locus-based gene prioritization methods by trait. Precision-recall plots for each method. Each point represents a single trait colored by trait domain. Only traits for which the method prioritized at least five genes in the validation loci were included. A legend for trait domain colors is provided in Fig. 2.



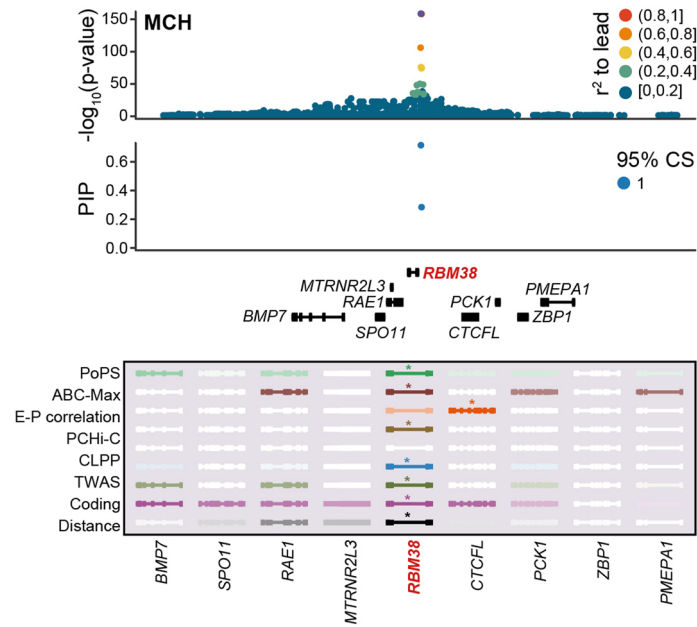
Extended Data Fig. 7 | Additional performance metrics using evaluation gene set in 1,348 non-coding loci containing genes that harbor fine-mapped protein coding variants. a, Sensitivity-specificity plot showing performance of locus-based methods, PoPS, intersections of pairs of locus-based methods, and intersections of PoPs with locus-based methods on the evaluation gene set

of 589 genes with fine-mapped protein coding variants. **b,** Heatmap showing performance using the F-score of locus-based methods, PoPS, intersections of pairs of locus-based methods, and intersections of PoPs with locus-based methods.

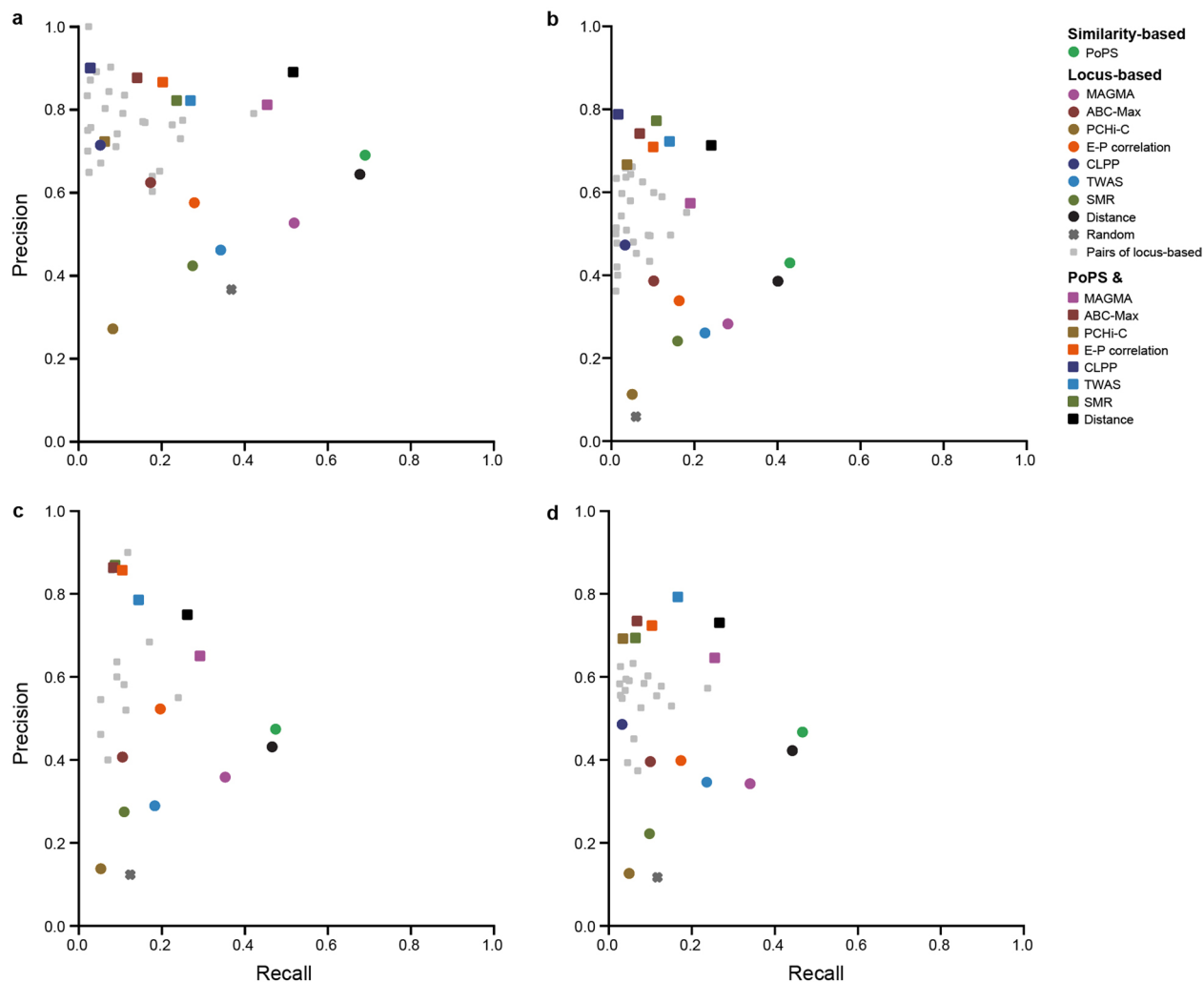


Extended Data Fig. 8 | Number of prioritized genes for non-UK Biobank traits. Number of unique gene-trait pairs prioritized by PoPS, locus-based gene prioritization methods, and their intersections, sorted by estimated precision. The full height of each bar represents the total number of genes prioritized.

The opaque portion of each bar represents the expected number of true causal genes prioritized. Methods to the left of the dashed line achieve precision greater than 75%.



Extended Data Fig. 9 | Known example *RBM38*. Top: summary statistics colored by LD to the lead variant and fine-mapping results for variants in the locus colored by credible set. Bottom: results from PoPS and locus-based methods for all genes in the locus. Genes are colored by strength of prediction for each method with a star denoting the prioritized gene. Variant rs737092, *RBM38* for mean corpuscular hemoglobin (MCH).



Extended Data Fig. 10 | Sensitivity of precision and recall estimates to locus definition. **a**, Loci defined as ± 100 kb on either side of the lead variant. **b**, Loci defined as ± 1 Mb on either side of the lead variant. **c**, Results restricted

to loci in fine-mapped regions with three or fewer independent credible sets. **d**, Results restricted to loci in fine-mapped regions with five or fewer independent credible sets.

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

Data analysis

We also used the following software packages:
 PoPS v0.2 (<https://github.com/Finucanelab/pops> & <https://doi.org/10.5281/zenodo.8002379>)
 MAGMA vl.07b (<https://ctg.cncr.nl/software/magma>)
 DEPICT release 1.94 (<https://github.com/perslab/depict>)
 NetWAS 2019 version (<https://hb.flatironinstitute.org/netwas/>)
 FUSION (https://github.com/gusevlab/fusion_twas)
 Benchmarker (<https://github.com/RebeccaFine/benchmarker>)
 LDSC vl.0 (<https://github.com/bulik/ldsc>)
 PLINK vl.90b (<https://www.cog-genomics.org/plink2/>)
 Seruat v3 (<https://github.com/satijalab/seurat>)
 ABC (<https://github.com/broadinstitute/ABC-Enhancer-Gene-Prediction>)
 SMR (<https://yanglab.westlake.edu.cn/software/smr/>)

Code to reproduce gene feature analyses is available on GitHub (https://github.com/Finucanelab/gene_features).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

A repository of processed gene features and visualizations of top derived features is available on GitHub (https://github.com/Finucanelab/gene_features). Complete PoPS results for 95 complex traits in the UK Biobank and 18 additional disease traits and results for PoPS and locus-based methods in genome-wide significant loci are available at <https://www.finucanelab.org/data>.

We also obtained data for:

Fine-mapping (<https://www.finucanelab.org/data>)

Biological pathways (KEGG, Gene Ontology, Reactome, Mouse Genome Databases) (<https://github.com/perslab/depict>)

FANTOM5 (<https://fantom.gsc.riken.jp/5/>)

ROADMAP Epigenomics (<https://ernstlab.biolchem.ucla.edu/roadmaplinking/>)

PChI-C (https://github.com/caleblareau/singlecell_bloodtraits/tree/master/data/)

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender	NA
Reporting on race, ethnicity, or other socially relevant groupings	NA
Population characteristics	NA
Recruitment	NA
Ethics oversight	NA

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We analyzed available GWAS summary statistics with a median sample size of 345,247.
Data exclusions	Data were not excluded.
Replication	There were no experimental findings. Computational predictions were validated using Benchmarker, a closest gene hypothesis test, and by estimating precision and recall on the fine-mapped coding variant evaluation set.
Randomization	There was no group allocation.
Blinding	There was no group allocation.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

- | n/a | Included in the study |
|-------------------------------------|--|
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Antibodies |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Eukaryotic cell lines |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Palaeontology and archaeology |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Animals and other organisms |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Clinical data |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Dual use research of concern |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Plants |

Methods

- | n/a | Included in the study |
|-------------------------------------|---|
| <input checked="" type="checkbox"/> | <input type="checkbox"/> ChIP-seq |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Flow cytometry |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> MRI-based neuroimaging |