HHS Public Access

Author manuscript

Nat Genet. Author manuscript; available in PMC 2024 February 02.

Published in final edited form as:

Nat Genet. 2023 August ; 55(8): 1267–1276. doi:10.1038/s41588-023-01443-6.

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

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

[1]Broad Institute of MIT and Harvard, Cambridge, MA, USA

[2]Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, MA, USA

[3]Current address: Artificial Intelligence Laboratory, Illumina, Inc.; San Diego, California 92122, USA

[4]Program in Computational & Systems Biology, MIT, Cambridge, MA, USA

[5]Computer Science & Artificial Intelligence Lab, MIT, Cambridge, MA, USA

[6]Division of Endocrinology and Center for Basic and Translational Obesity Research, Boston Children's Hospital, Boston, MA, USA

[7]Department of Genetics, Harvard Medical School, Boston, MA, USA

[8]Current address: Vertex Pharmaceuticals Incorporated, Boston, MA, USA

[9]Department of Statistics, Harvard University, Cambridge, MA, USA

[†]Correspondence to eweeks@broadinstitute.org or finucane@broadinstitute.org.
[*]Co-first author

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, J.O-M.E, M.B., A.K.S, A.V., R.M.G, K.L., K.G.A, and J.E.M provided data used by PoPS or other gene prioritization methods. 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 co-founder and equity holder of Celsius Therapeutics, an equity holder in Immunitas, and was an SAB member of ThermoFisher Scientific, Syros Pharmaceuticals, Neogene Therapeutics and Asimov until July 31, 2020. From August 1, 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 non-profit 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.

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 this study is available at https://doi.org/10.5281/zenodo.8002379.

[10]Analytic and Translational Genetics Unit, MGH, Boston, MA, USA

[11]Program in Bioinformatics and Integrative Genomics, Harvard Medical School, Boston, MA, USA

[12]Department of Statistical Genetics, Osaka University Graduate School of Medicine, Suita, Japan

[13]Current address: Bristol Myers Squibb, Cambridge, MA, USA

[14]MD-PhD Program, Johns Hopkins University School of Medicine, Baltimore, MD, USA

[15]Division of Gastroenterology, Hepatology, and Nutrition, Boston Children's Hospital, Boston, MA, USA

[16]Program in Immunology, Harvard Medical School, Boston, MA, USA

[17]Harvard Stem Cell Institute, Cambridge, MA, USA

[18]Department of Molecular Biology, MGH, Boston, MA, USA

[19]Current address: Department of Biological Regulation, Weizmann Institute of Science, Rehovot 7610001, Israel

[20]Institute for Medical Engineering and Science, MIT, Cambridge, MA, USA

[21]Department of Chemistry, MIT, Cambridge, MA, USA

[22]Koch Institute for Integrative Cancer Research, MIT, Cambridge, MA, USA

[23]Ragon Institute of MGH, MIT, Cambridge, MA, USA

[24]Gastrointestinal Unit and Center for the Study of Inflammatory Bowel Disease, MGH, Boston, MA, USA

[25]Center for Computational and Integrative Biology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

[26]Howard Hughes Medical Institute, MIT, Cambridge, MA, USA

[27]Current address: Genetec, San Francisco, CA, USA

[28]Division of Cardiovascular Medicine and Genetics, Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA

[29]Harvard Medical School, Boston, MA, USA

[30]Department of Surgery, MGH, Boston, MA, USA

[31]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA

[32]Department of Pediatrics, Harvard Medical School, Boston, MA, USA

[33]Department of Biology, MIT, Cambridge, MA, USA

[34]Department of Systems Biology, Harvard Medical School, Boston, MA, USA

[35]Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA

[36]BASE Initiative, Betty Irene Moore Children's Heart Center, Lucile Packard Children's Hospital, Stanford University School of Medicine, Stanford, CA, USA

Abstract

Genome-wide association studies (GWAS) are a valuable tool for understanding the biology of complex human traits and diseases, but associated variants rarely point directly to causal genes. Here, 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 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.

INTRODUCTION

Genome-wide association studies (GWAS) have identified thousands of genetic loci associated with common complex traits and diseases[1]. Nonetheless, 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 variant[4]. Second, most associated loci do not contain protein-coding variants. Instead, the causal variant acts through gene regulatory mechanisms[3], but incomplete maps from regulatory element to gene hinder causal gene identification[5]. 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 eQTLs[7,8]; 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]. Across both categories, existing methods lack consensus and have high false positive rates[18]. At the same time, related work suggests that combining results from different methods can yield better predictions[19]. Among similarity-based approaches, most methods focus on single-nucleotide polymorphisms (SNPs) that meet genome-wide significance, ignoring information from sub-significant variants that explain the majority of narrow sense heritability[17,20]. Moreover, recently-generated single-cell RNA-seq datasets hold promise for more accurately characterizing shared functions among genes and, thus, improving the accuracy of similarity-based gene prioritization.

Here, 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 single-cell RNA-seq data sets. 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 genes whose physical locations on the genome are near associated SNPs and who 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 (PPI). To nominate causal genes, PoPS then assigns a priority score to every protein coding gene according to these enrichments (Fig. 1).

First, PoPS applies MAGMA[9] to compute gene-level association statistics and their correlations using GWAS summary statistics and LD information from an ancestry-matched reference panel (see 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 (see 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

$$y = X^f \beta^f + \varepsilon,\ \varepsilon \sim MVN(0, R) \quad [1]$$

where X^f is a column vector corresponding to gene feature f (*e.g.* a binary indicator of membership in a pathway), and R is a covariance matrix designed to account 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 with 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.)

$$y = X\beta + \varepsilon,\ \varepsilon \sim MVN(0, R) \quad [2]$$

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 estimate $\hat{\beta}$.

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

$$\hat{y}_g = X_g\hat{\beta} \quad [3]$$

We refer to $\hat{y}_g$ 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 (see Methods). We additionally include an option for a user to compute scores in a leave-one-chromosome-out fashion, obtaining estimates $\hat{\beta}_{-chr\,i}$ for $i = 1, \ldots, 22$ and scoring, for example, a gene on chromosome 1 by using $\hat{\beta}_{-chr\,1}$.

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 Biobank[21] (Supplementary Table 1) using EUR individuals from the 1000 Genomes Project[22] as a reference panel (see 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 protein-protein interaction network[23], and 8,479 based on pathway membership[24–27] (Supplementary Table 2; see Methods). After marginal feature selection, 2,512 to 26,155 features per trait were included in the predictive model (Extended Data Fig. 1d). For each trait, 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 likely causal genes, we avoided using curated sets of gold standard genes that may be biased towards 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 leave-one-chromosome-out 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 we expect to be highly enriched for causal signals. First, we applied the Benchmarker method[17], which evaluates methods by estimating the average contribution of SNPs near top scoring genes to per SNP heritability (τ). Here, 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 our estimates for normalized τ were significantly greater than zero for 54 of 113 traits tested (Fig. 2a, Supplementary Table 3). As a second evaluation metric, we focused on the performance of PoPS in GWAS significant loci. Following prior work demonstrating that the causal gene is often the closest gene to the lead variant in the locus[19], we tested whether PoPS prioritized genes were the closest gene to the lead variant more often than expected by chance (see 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 correction (Fig. 2b, 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 to existing similarity-based methods: DEPICT[6], NetWAS[10], and a method that we call MAGMA-sim[17] (see 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 a Bonferroni significance cutoff. We found that PoPS using the full feature set showed the strongest performance compared to 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 (see Methods), we again found that PoPS significantly outperformed all other similarity-based methods tested (Fig. 2c, 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 to 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 features included in the PoPS model: gene expression, pathways, and PPI networks (Fig. 3, Extended Data Fig. 2, Extended Data Fig. 3). We created three alternate versions of PoPS, training on features from each category separately, to produce three new sets of results for each phenotype. In a meta-analysis over 46 independent traits, we found that including all features yielded the strongest performance in both Benchmarker and closest gene evaluations, followed in order by pathways, gene expression, and lastly by PPI networks (Fig. 3a, Extended Data Fig. 3b, 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. Because 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 (see 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, Supplementary Table 4). For LDL cholesterol, we observed clusters composed of lipid synthesis and transport pathways in addition to liver gene expression features[28]. For HbA1c, a test that measures average blood sugar levels but is also affected by red blood cell levels[29], we observed both glucose and hemoglobin related clusters of features. For rheumatoid arthritis, an autoimmune disease[30], 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 PoPS is able to 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 PoPS to 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 towards well-studied genes, potentially introducing bias when evaluating 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 Biobank[21] to identify likely causal genes as those harboring a fine-mapped (PIP > 0.5) protein-coding variant. Matching by trait, we then identified independent, non-coding credible sets within 500kb of these protein-coding genes. This approach identified 1,348 non-coding 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 500kb of the locus-defining non-coding credible 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 (see 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 to the other similarity-based methods (Extended Data Fig. 4, Supplementary Table 5). We found that PoPS had both higher recall and higher precision than DEPICT, NetWAS, and

MAGMA-sim using non-coding genetic signal (see Methods). We additionally show that PoPS using the full non-coding genetic data had both higher recall and higher precision than PoPS run in a leave-one-chromosome-out (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 (see Methods), where we had not only summary statistics, but fine-mapping results. We evaluated a non-exhaustive but wide range of methods:

1. We overlapped fine-mapped (PIP > 0.1) non-coding variants with predicted enhancer-gene connections from (**a**) correlating enhancer and promoter activity (E-P correlation)[15,16], (**b**) 3-D loops from promoter capture Hi-C (PCHi-C)[12–14], and (**c**) activity-by-contact (ABC)[11] maps to identify genes regulated by fine-mapped variants.
2. We incorporated eQTL data and (**a**) applied TWAS[8] with GTEx v7[33] weights to identify significantly associated genes, (**b**) applied SMR[34] with GTEx v7[33] weights to identify significantly associated genes, and (**c**) computed co-localization posterior probabilities (CLPP)[7] with fine-mapping results from GTEx v8[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; see 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 (see Methods).

All locus-based methods for prioritizing genes from non-coding signal showed precision less than 50% except CLPP, which had both the highest precision, 52%, and the lowest recall, 4% (Fig. 4a, Extended Data Fig. 6, Supplementary Table 5, Supplementary Table 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 the majority 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 to 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, Supplementary Table 8). Overall, we found low concordance among methods. For example, PCHi-C prioritized a gene in 8,777 loci, while 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, Supplementary Table 5). Specifically, for each locus-based 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 above 1% achieved precision greater than 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, Supplementary Table 9, Supplementary Table 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 use 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 genes[38], *APOE, APOA1, APOB,* and *PPARG*, were four of the top five genes for LDL cholesterol. For mosaic loss of Y (LOY) chromosome in circulating blood[39], 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 schizophrenia[42].

We additionally find 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 when locus-based methods disagreed, we estimate 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*, *LRRC47*, and *CEP104*[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 production[44], for mean corpuscular hemoglobin concentration (Fig. 6a). In another example, the variant rs737092 was fine-mapped with PIP = 0.72 for mean corpuscular hemoglobin (MCH), and experimental evidence shows that the expression of both *RBM38* and *RAE1* are affected by this variant[43] (Extended Data Fig. 9). Locus-based methods prioritized *RBM38* and *CTCFL*, but PoPS correctly prioritized *RBM38*, which has been shown to play a role in splicing key erythroid transcripts during erythropoiesis[43]. 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 mice[45].

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, closest gene, when similarly combined with a local method, had only 27% precision. Genes prioritized by PoPS+local that are not the closest gene likely represent a set enriched for novel candidate genes. For example, our analysis nominated *LGR4* for estimated glomerular filtration rate 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 mice[46]. For deep vein thrombosis, our analysis nominated *CCR7* at a locus with two credible sets and 41 genes, including 28 genes in the Keratin family and three genes supported by locus-based methods (Fig. 6d). The top PoPS features supporting the relevance of *CCR7* were *abnormal thymus medulla morphology*, *increased IgE level*, and *response to prostaglandin*. CCR7, chemokine receptor 7, is a regulator of inflammation which is involved in the development of DVT[47] and may be involved in thrombogenesis through platelet activation[48].

DISCUSSION

We developed a new computational method, PoPS, for prioritizing causal genes from GWAS that predicts polygenic genetic association from gene expression profiles, protein-protein interaction 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 non-coding associations near fine-mapped variants in protein-coding genes. This set serves as a powerful tool that, unlike gene sets comprised of 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 whose shared functions 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 (*i.e.* cell-type). Second, to leverage the polygenic signal, we assume the causal mechanisms are shared between top loci and sub-significant loci. Third, while informative for ranking genes, the polygenic priority score 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 non-coding associations increases our simplistic assignment of non-coding associations to positive evaluation genes within a pre-defined genomic distance will likely become less accurate (Extended Data Fig. 10), suggesting 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 towards 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.

METHODS

MAGMA gene z-scores

We applied MAGMA[9] to the summary statistics for each trait using EUR individuals from 1000 Genomes Project reference panel[22] 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 chi-square 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 linkage disequilibrium (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 on 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

$|\min(\lambda_{min}, 0)| + 0.05 + 0.9 \times \max(0, var(Y) - 1)$ to each element along the diagonal, where λ_{min} is the minimum eigenvalue of R, and Y are the MAGMA z-scores.

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 MAGMA[9] 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. We additionally 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 $P < 5\times10^{-8}$, we designated independent lead variants from which to define loci. For the 18 traits where we used publicly available summary statistics, we performed PLINK[49] 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 for 95 heritable traits in the UK Biobank were performed as part of a fine-mapping study[21]. Up to 361,194 individuals of white British ancestry with available phenotypes and variants with INFO > 0.8, minor allele frequency (MAF) > 0.01%, and Hardy-Weinberg equilibrium (HWE) p-value > 1e-10 were included in the GWAS. Covariates for the top 20 PCs, sex, age, age^2, sex*age, sex*age^2, and dilution factor where applicable were controlled for in the association studies. Quantitative traits were inverse rank transformed and associations were estimated using BOLT-LMM[50]. For binary traits, associations were estimated using SAIGE[51]. 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 protein-protein interaction networks.

(2) For each of the 77 gene expression datasets (Supplementary Table 2), we uniformly re-processed the raw count (or normalized count when raw counts were not provided)

matrices using Seurat v3[52]. First, cells with total counts outside of the 5–95th 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 (CPM), log normalized, and scaled such that each gene had a mean of 0 and variance of 1 across cells. Principal components and gene loadings were computed on scaled expression values for the top 1,000–3,000 variable genes using truncated SVD[53]. Independent components and gene loadings were computed using fastICA[54]. A k-nearest neighbor graph was created using the top principal components (PCs, based upon inspection of elbow plot) and clusters were identified using the Louvain algorithm. The uniform manifold approximation and projection (UMAP) algorithm[55] was used to visualize clusters and investigate batch effects. When batch effects were visually apparent and pre-defined 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-vs-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 top derived features at https://github.com/FinucaneLab/gene_features.

We then derived features for PoPS (**a**) on the whole dataset, (**b**) within clusters representing different cell populations, and (**c**) between clusters. (**a**) On the whole dataset, we derived features of gene loadings from principal component analysis (PCA) and gene loadings from independent component analysis (ICA). (**b**) 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. Comparing across clusters (1-vs-all), we derived features of a *t*-statistic for differential expression and a binary indicator for differentially up- and down-regulated genes (Benjamini–Hochberg FDR < 0.05 and $|\log_2(\text{fold-change})| > 2$).

(**2**) We created features from biological pathways curated for DEPICT from KEGG[25], Gene Ontology[24], Reactome[26], and the Mouse Genome databases[27]. Each feature was encoded as a binary indicator for membership to a pathway. (**3**) We created features using the predicted InWeb_IM protein-protein interaction (PPI) network[23]. 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 mean of 0 and variance of 1 across genes.

DEPICT

We ran DEPICT[6] 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 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 leave-one-chromosome-out framework. Here, all variants on the chromosome for which gene p-values were computed were removed from the summary statistics before running DEPICT.

NetWAS

NetWAS[10] 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 panel[22]. We applied NetWAS using both the default threshold of $P<0.01$ and the Bonferroni significance threshold, which was shown in previous work to have better performance for well-powered GWAS[17]. In cases where fewer than 15 genes passed the significance threshold, we relax the p-value threshold until there are 15 passing genes. For Benchmarker and closest gene enrichment analyses, NetWAS was run in a leave-one-chromosome-out framework. Here, 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-sim[17], 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 leave-one-chromosome-out 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 $K = 500 \times (\%\text{ of genes on test chromosome})$, thus prioritizing 500 total genes across all chromosomes. MAGMA-sim is similar to PoPS in that in 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

Benchmarker[17] 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 a 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, $\boldsymbol{\tau}$, and its p-value for the hypothesis $\boldsymbol{\tau} > 0$. $\boldsymbol{\tau}$ measures the contribution of SNPs near prioritized genes to per SNP heritability after controlling for the baseline annotations. To make $\boldsymbol{\tau}$ comparable across traits, we normalized $\boldsymbol{\tau}$ by the average per-SNP heritability for each trait and refer to this quantity as normalized $\boldsymbol{\tau}$.

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 Bernoullis with different biases. For computational tractability when L is large, we approximate this by a normal with matched moments.

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

We performed a one-sided test for $c > \sum_{1:L} \frac{1}{n_l}$ under the null. We additionally computed the enrichment of the number of PoPS prioritized genes that are the closest as the ratio of the observed to the expected, $\frac{c}{\sum_{1:L} \frac{1}{n_l}}$, and estimated the standard error of the enrichment. We used the bootstrap to estimate the standard error of the enrichments, not assuming a null distribution, and performed 1024 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 regression[57] with LD scores from UK10K[58]. Next, we created an adjacency matrix of traits with edge weights corresponding to whitened ($|r_g| < 0.2$ were 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 package[59] in R 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 one minus the squared Pearson correlation (r^2) 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 in order 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 performed[21] for 95 complex traits in the UK Biobank and for 49 tissues in GTEx v8 using the Sum of Single Effects (SuSiE) method[60], 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-LMM[50] for quantitative traits and SAIGE[51] for binary traits, in sample dosage LD was computed using LDStore[61], and phenotypic variance was computed empirically. Variants in the MHC region (chr6: 25–36 Mb) were excluded as were 95% CSs containing variants with fewer than 100 minor allele counts. Coding (missense and predicted loss of function) variants were annotated using the Variant Effect Predictor (VEP) version 85[62]. Fine-mapping data used in this study is available at https://www.finucanelab.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? And 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 that is condition negative, and a false negative (FN) is a gene that is not prioritized that is condition positive. The answers to our two questions are given, respectively, by precision, #TP/(#TP+#FP), and recall, #TP/(#TP+#FN).

ABC-Max

We used the Activity-by-Contact (ABC) model[11] to predict enhancer-gene connections in 131 biosamples from 74 distinct cell-types and tissues based on measurements of chromatin accessibility (ATAC-seq or DNase-seq) and histone modifications (H3K27ac ChIP-seq). 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.

Enhancer-promoter correlation

We downloaded predicted enhancer-promoter maps based upon 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 project[15], 127 tissues and cell-lines from the ROADMAP Epigenomics project[16], and also for 16 primary blood cell types[13]. For the FANTOM5 dataset, we filtered to interactions with Benjamini–Hochberg FDR < 10^{-5} for a non-zero Pearson correlation. For the ROADMAP dataset, we filtered to interactions with a confidence score > 2.5. For the Ulirsch et al. dataset, we filtered to interactions with a Pearson 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.

Promoter capture Hi-C

We downloaded promoter capture Hi-C datasets (PCHi-C) containing observed physical interactions between fragmented DNA and targeted genic promoters for 28 diverse human tissues and cell-lines[12] and 15 primary blood cell types[14]. For the Jung et al. dataset, we filtered to interactions with p-values for interaction < 0.01 and raw frequency counts > 5. For the Javierre et al. dataset, we filtered to interactions with CHiCAGO[63] 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 upon 39 ROADMAP tissues[64] 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 TWAS[8] using the FUSION software package and precomputed expression reference weights for 48 tissues from GTEx v7[33]. To avoid leveraging the coding signal for the precision-recall analysis, we excluded all variants in LD (r^2 greater than 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, TWAS was run leaving out the coding signal. TWAS weights were obtained from http://gusevlab.org/projects/fusion/weights/GTEX7.txt.

SMR

We applied SMR[34] using the SMR software tool and pre-computed cis-eQTL summary data across 48 human tissues from GTEx v7[33]. Cis-eQTL summary data was pre-filtered to SNPs within 1Mb of the transcription start site for each gene. For precision-recall analyses, SMR was run leaving out the coding signal.

Co-localization posterior probability

Using fine-mapping results for 95 complex traits from the UK Biobank and for eQTLs in 49 tissues from GTEx v8[35] we computed co-localization posterior probabilities (CLPP), analogous to those reported by the eCAVIAR software[7]. 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 non-coding 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, Supplementary Table 6). We chose the following prioritization criteria to maximize precision:

(**1a-c**) 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.

(**2a**) TWAS: for each locus such that at least one gene is significantly associated after Bonferroni correction, the gene with the most significant p-value.

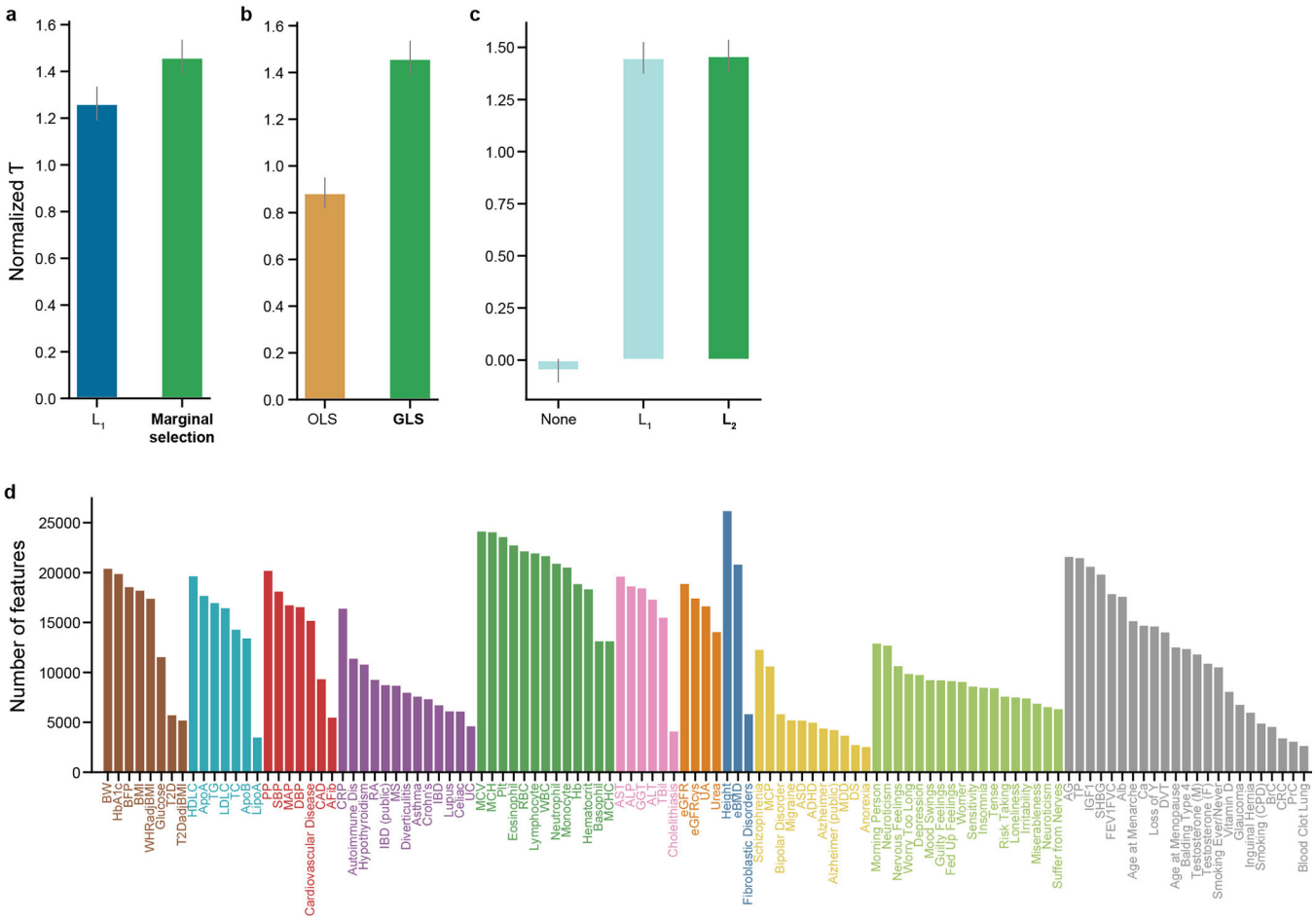
(**2b**) SMR: for each locus such that at least one gene is significantly associated after Bonferroni correction, the gene with the most significant p-value.

(**2c**) CLPP: for each locus such that at least one gene has a variant with CLPP > 0.1, the gene with the highest CLPP.

(**3**) Distance: for each locus, the gene that is closest to the lead variant by distance to the gene body.

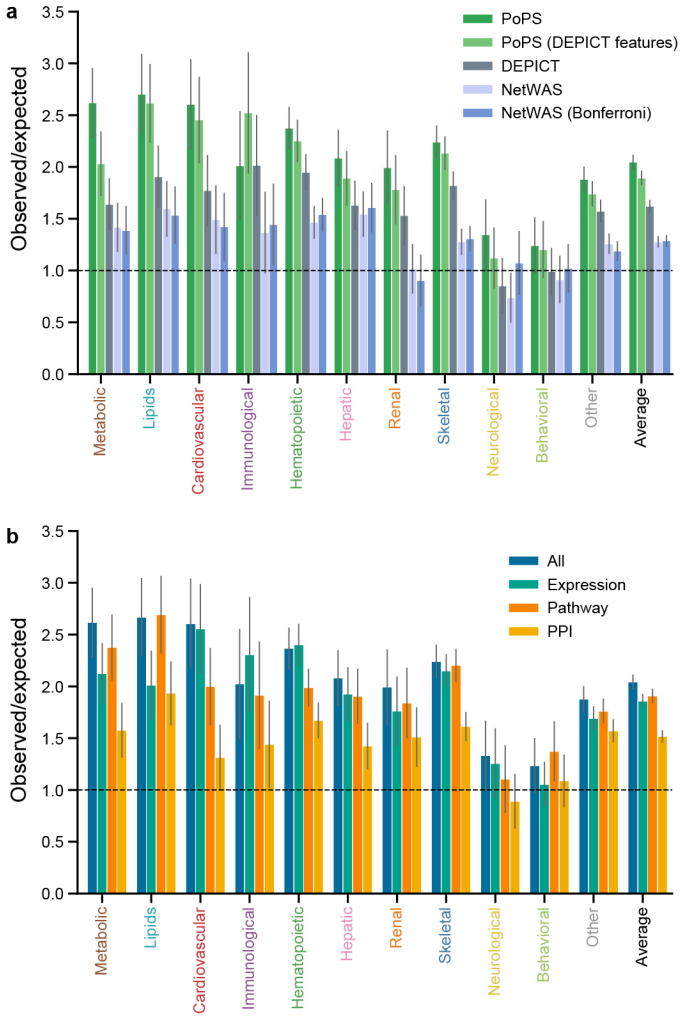
(**4**) PoPS: for each locus, the gene that has the highest PoP score.

Extended Data



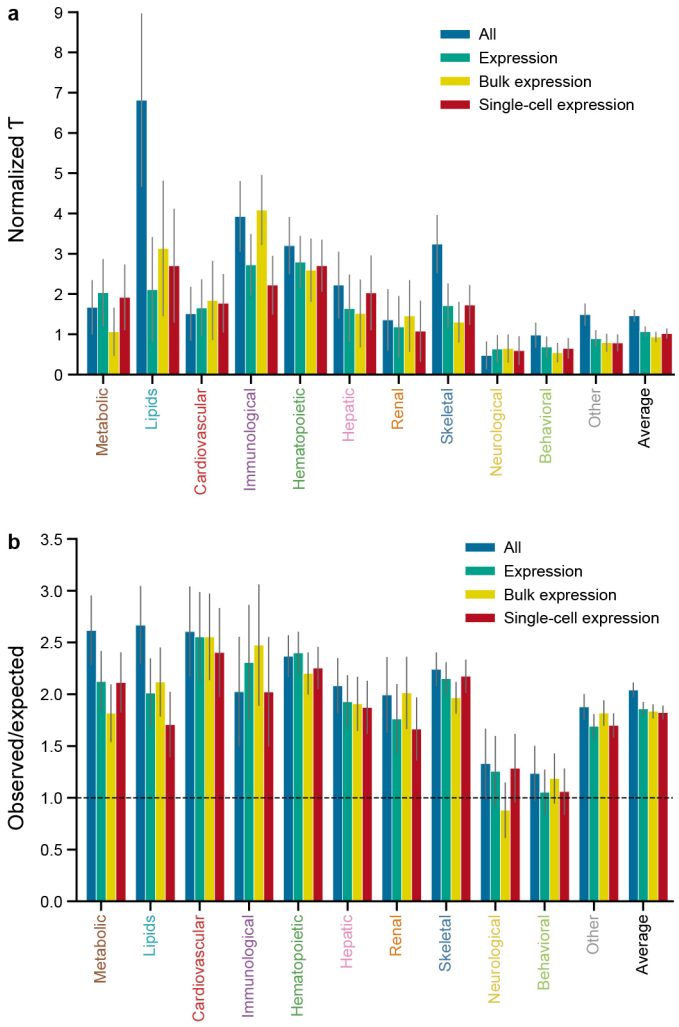
Extended Data Fig. 1. PoPS model parameter choices and feature selection.

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 L1 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 L2 penalty performs better than similar models with an L1 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.



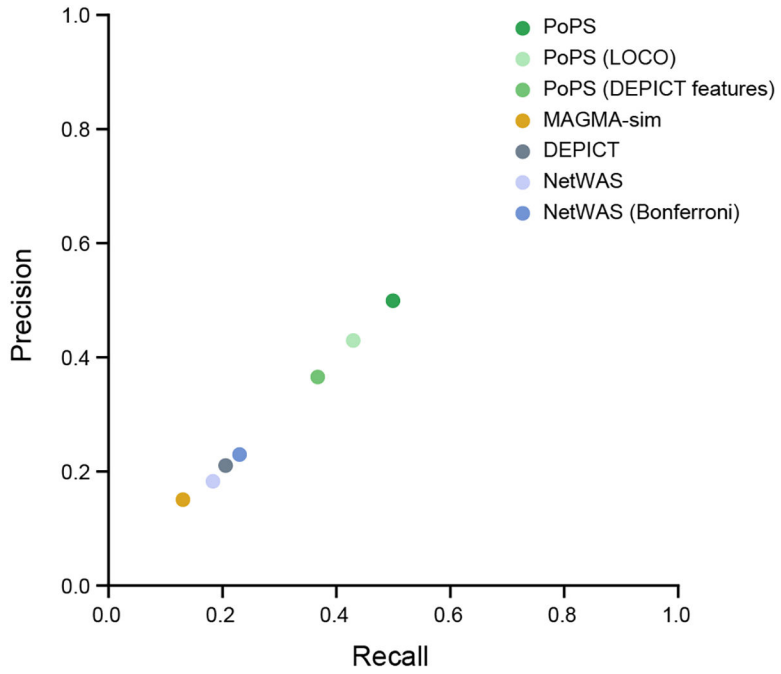
Extended Data Fig. 2. Additional comparisons using closest gene metric.

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.

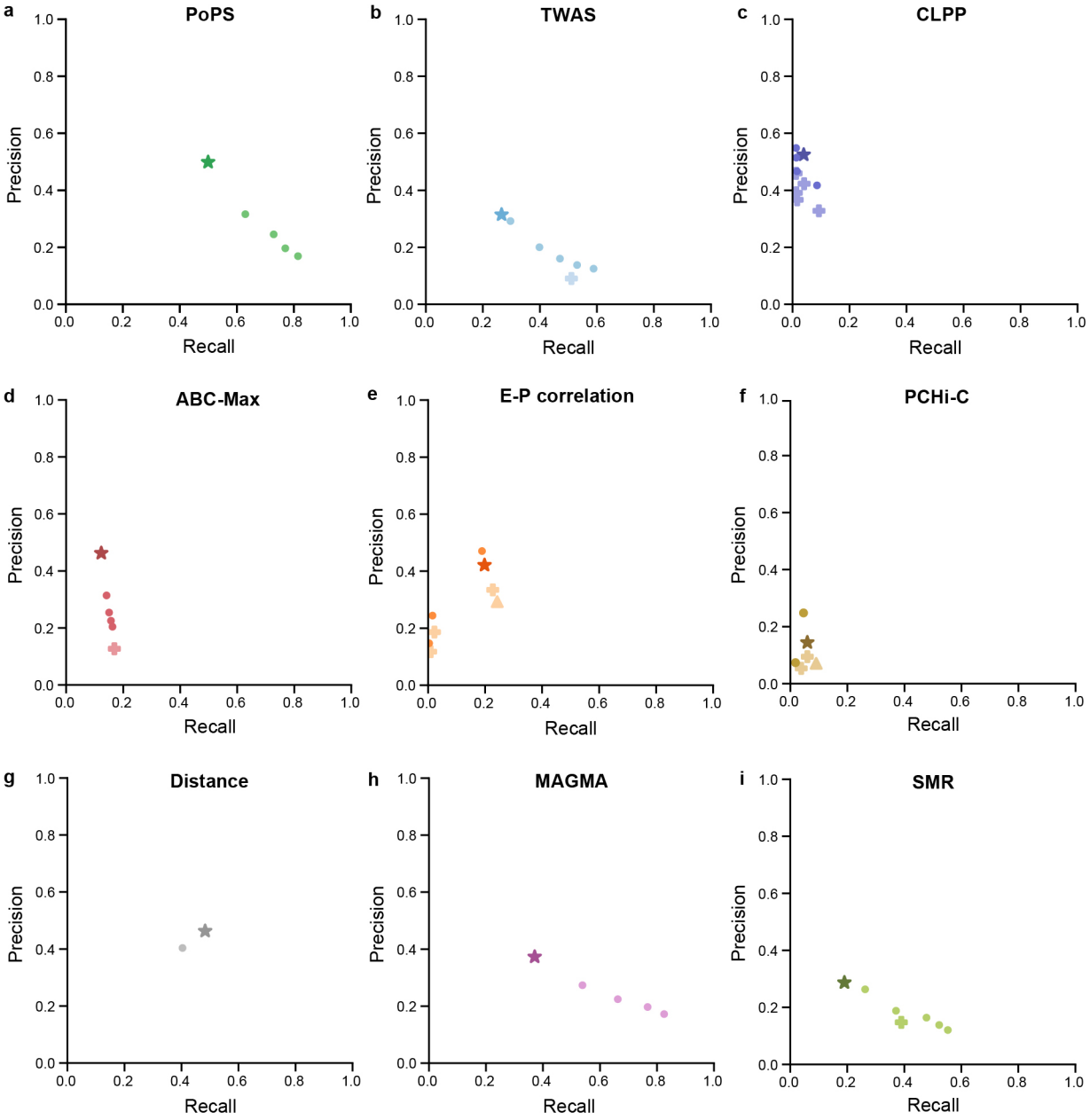


Extended Data Fig. 3. Comparison of gene expression features derived from bulk and single-cell RNA seq datasets.

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.



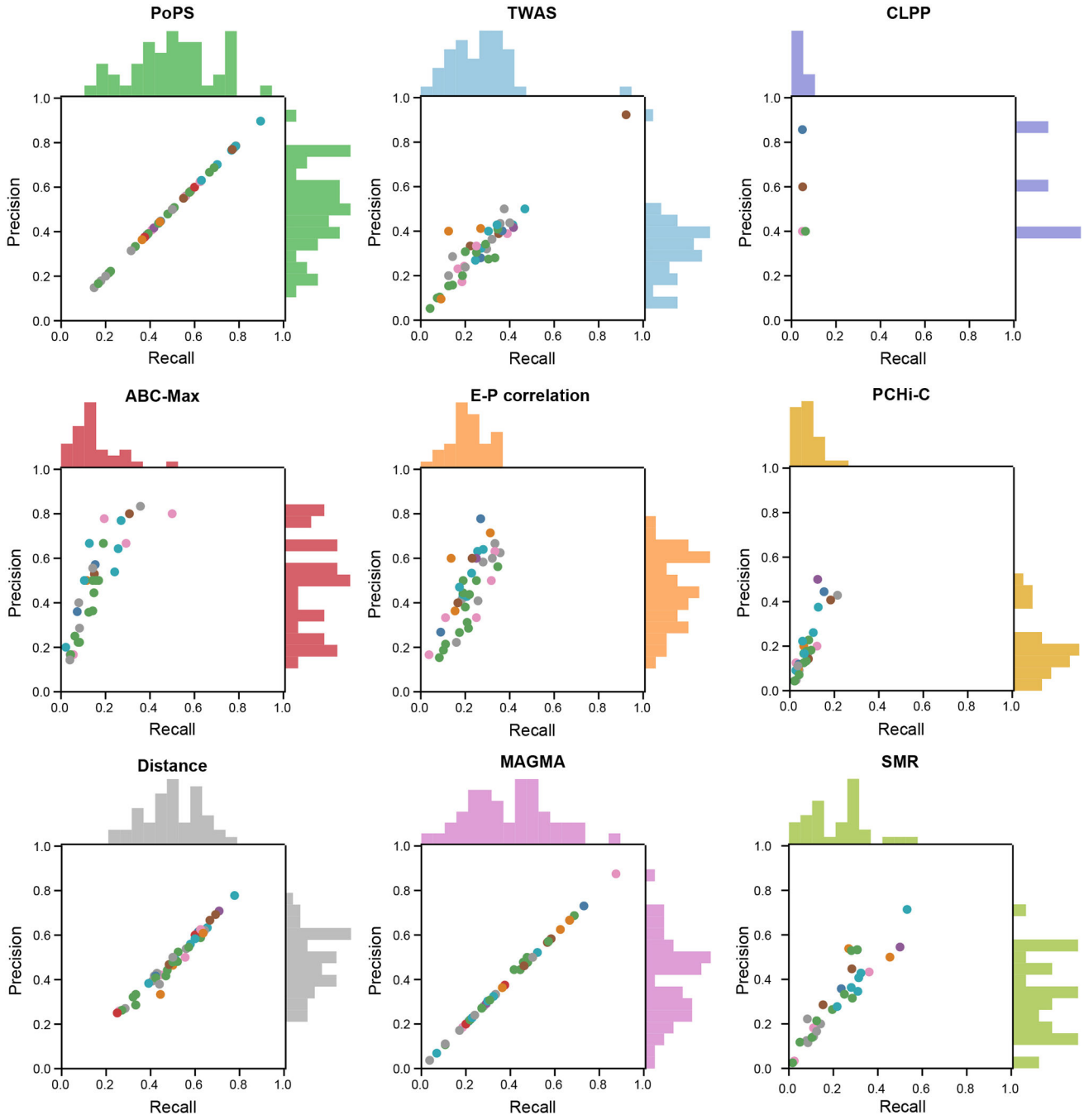
Extended Data Fig. 4. Comparison of similarity-based methods using precision and recall.
Precision-recall plot showing performance of similarity-based methods.



Extended Data Fig. 5. Comparing prioritization criteria.

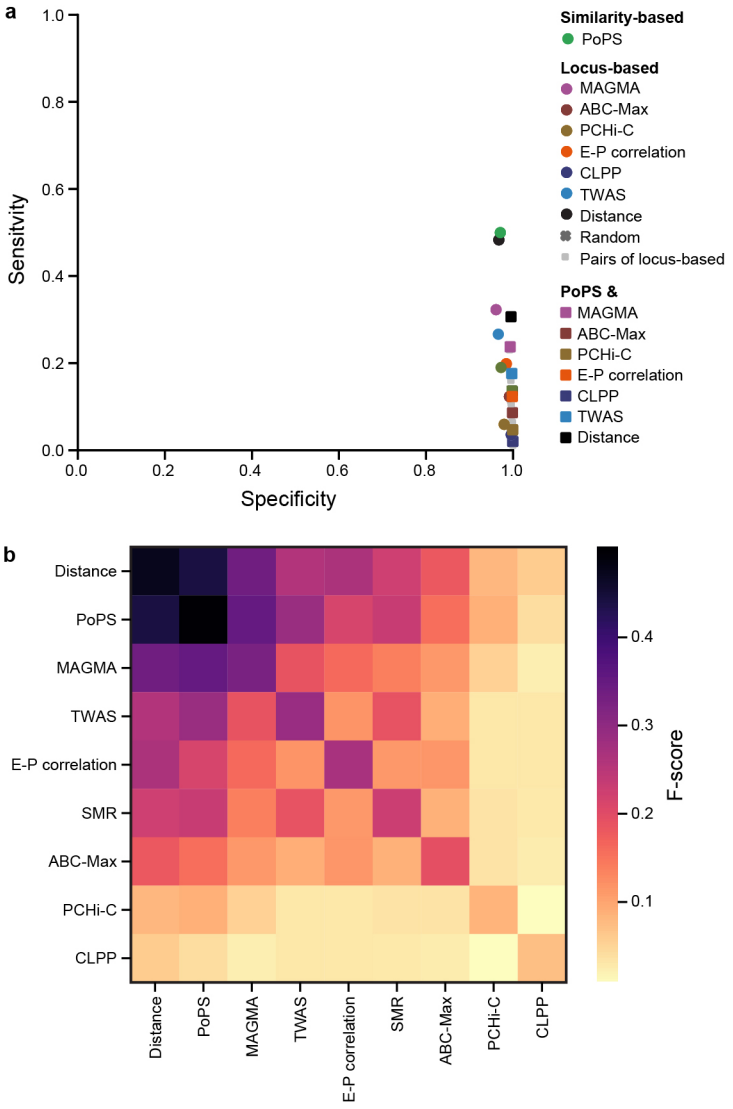
Precision-recall plots for each method with varying prioritization criteria. Each point shows the precision and recall for a set of prioritized genes selected using prioritization criteria based on absolute thresholds and/or relative rank in a locus. For all methods, the star represents the final chosen criteria. a, Circles: PoP scores ranked ≤ 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 ≤ 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 ≤ 2–5 in the locus. Star: highest ABC connection strength in the locus. e, Pluses: any predicted connection from PCHi-C 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 ≤ 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 ≤ 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.



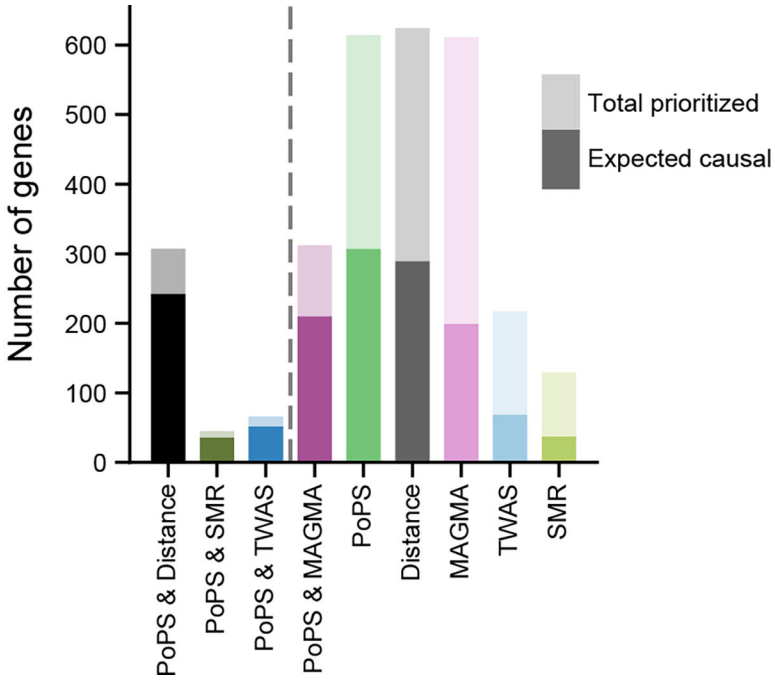
Extended Data Fig. 6. Performance of PoPS and locus-based gene prioritization methods by trait.

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.



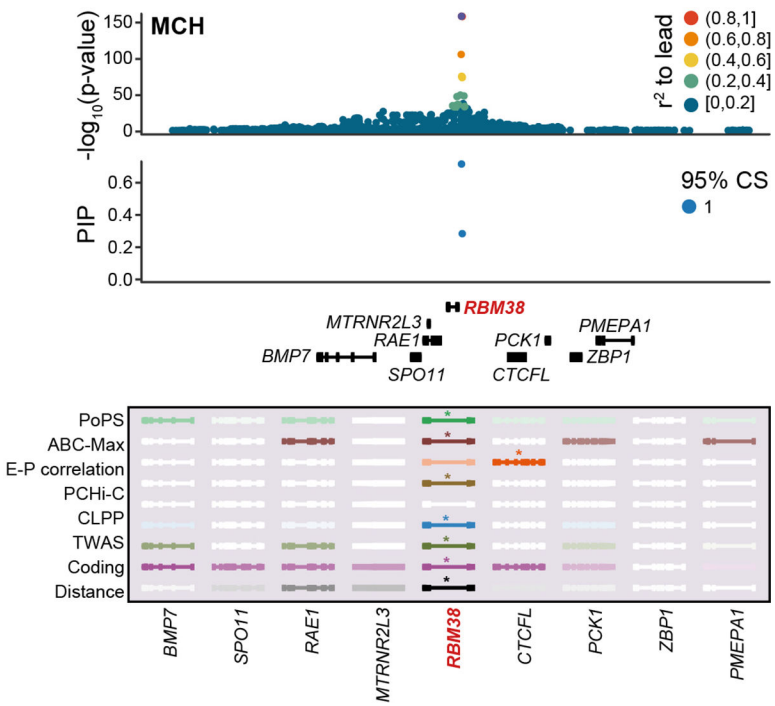
Extended Data Fig. 7. Additional performance metrics using evaluation gene set in 1,348 non-coding loci containing genes that harbor fine-mapped protein coding variants.

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.



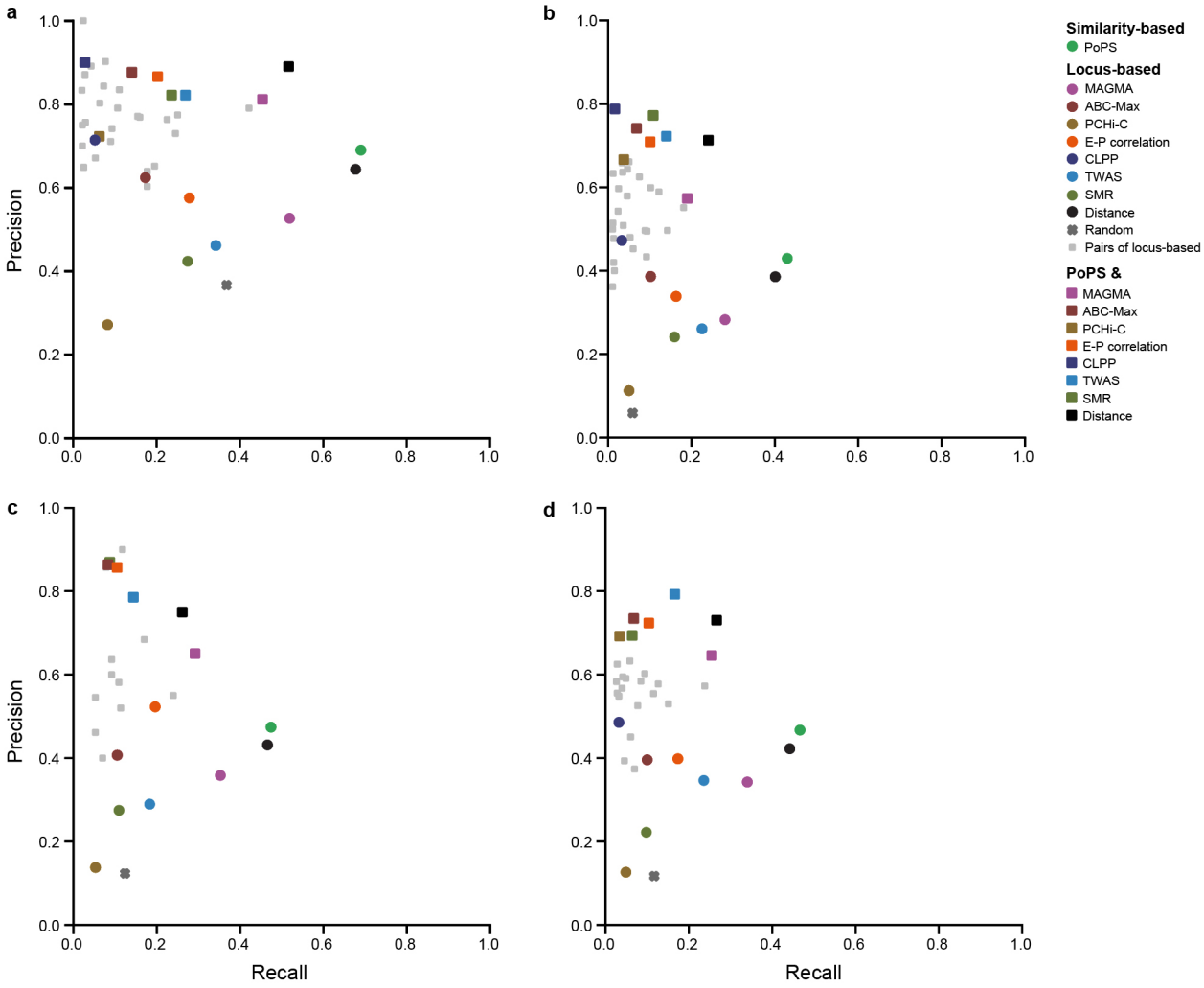
Extended Data Fig. 8. Number of prioritized genes for non-UK Biobank traits.

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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Krishna Aragam, Adam Butterworth, Mark Daly, Nikita Artomov, Yakir 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 NIH grant DP5 OD024582 and by Eric and Wendy Schmidt. J.M.E. was supported by a Pathway to Independence Award (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 NIH grant R01DK075787. R.S.F. was supported by NHGRI NIH F31HG009850. J.O.M was supported by the Richard and Susan Smith Family Foundation, the HHMI Damon Runyon Cancer Research Foundation Fellowship (DRG-2274-16), the AGA Research Foundation's AGA-Takeda Pharmaceuticals Research Scholar Award in IBD – AGA2020-13-01, the HDDC Pilot and Feasibility P30 DK034854, and the Food Allergy Science Initiative.

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.

REFERENCES

1. Visscher PM et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. Am. J. Hum. Genet. 101, 5–22 (2017). [PubMed: 28686856]
2. Donnelly P Progress and challenges in genome-wide association studies in humans. Nature 456, 728–731 (2008). [PubMed: 19079049]
3. Gallagher MD & Chen-Plotkin AS The post-GWAS era: From association to function. Am. J. Hum. Genet. 102, 717–730 (2018). [PubMed: 29727686]
4. Reich DE et al. Linkage disequilibrium in the human genome. Nature 411, 199–204 (2001). [PubMed: 11346797]
5. van Arensbergen J, van Steensel B & Bussemaker HJ In search of the determinants of enhancer-promoter interaction specificity. Trends Cell Biol. 24, 695–702 (2014). [PubMed: 25160912]
6. Pers TH et al. Biological interpretation of genome-wide association studies using predicted gene functions. Nat. Commun. 6, 5890 (2015). [PubMed: 25597830]
7. Hormozdiari F et al. Colocalization of GWAS and eQTL Signals Detects Target Genes. Am. J. Hum. Genet. 99, 1245–1260 (2016). [PubMed: 27866706]
8. Gusev A et al. Integrative approaches for large-scale transcriptome-wide association studies. Nat. Genet. 48, 245–252 (2016). [PubMed: 26854917]
9. de Leeuw CA, Mooij JM, Heskes T & Posthuma D MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput. Biol. 11, e1004219 (2015). [PubMed: 25885710]
10. Greene CS et al. Understanding multicellular function and disease with human tissue-specific networks. Nat. Genet. 47, 569–576 (2015). [PubMed: 25915600]
11. Fulco CP et al. Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. Nat. Genet. 51, 1664–1669 (2019). [PubMed: 31784727]
12. Jung I et al. A compendium of promoter-centered long-range chromatin interactions in the human genome. Nat. Genet. 51, 1442–1449 (2019). [PubMed: 31501517]
13. Ulirsch JC et al. Interrogation of human hematopoiesis at single-cell and single-variant resolution. Nat. Genet. 51, 683–693 (2019). [PubMed: 30858613]
14. Javierre BM et al. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. Cell 167, 1369–1384.e19 (2016). [PubMed: 27863249]
15. Andersson R et al. An atlas of active enhancers across human cell types and tissues. Nature 507, 455–461 (2014). [PubMed: 24670763]

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). [PubMed: 29058599]
17. Fine RS, Pers TH, Amariuta T, Raychaudhuri S & Hirschhorn JN Benchmarker: An unbiased, association-data-driven strategy to evaluate gene prioritization algorithms. Am. J. Hum. Genet. 104, 1025–1039 (2019). [PubMed: 31056107]
18. Barbeira AN et al. Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. Genome Biol. 22, 49 (2021). [PubMed: 33499903]
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). [PubMed: 30239796]
20. Yang J et al. Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42, 565–569 (2010). [PubMed: 20562875]
21. Kanai M et al. Insights from complex trait fine-mapping across diverse populations. medRxiv 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). [PubMed: 26432245]
23. Li T et al. A scored human protein–protein interaction network to catalyze genomic interpretation. Nat. Methods 14, 61–64 (2017). [PubMed: 27892958]
24. Ashburner M et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25, 25–29 (2000). [PubMed: 10802651]
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, D109–14 (2012). [PubMed: 22080510]
26. Croft D et al. Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res. 39, D691–7 (2011). [PubMed: 21067998]
27. Blake JA et al. The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse. Nucleic Acids Res. 42, D810–7 (2014). [PubMed: 24285300]
28. Teslovich TM et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466, 707–713 (2010). [PubMed: 20686565]
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). [PubMed: 28898252]
30. Kurkó J et al. Genetics of rheumatoid arthritis - a comprehensive review. Clin. Rev. Allergy Immunol. 45, 170–179 (2013). [PubMed: 23288628]
31. Gejman PV, Sanders AR & Duan J The role of genetics in the etiology of schizophrenia. Psychiatr. Clin. North Am. 33, 35–66 (2010). [PubMed: 20159339]
32. Heyes S et al. Genetic disruption of voltage-gated calcium channels in psychiatric and neurological disorders. Prog. Neurobiol. 134, 36–54 (2015). [PubMed: 26386135]
33. Consortium GTEx. Erratum: Genetic effects on gene expression across human tissues. Nature 553, 530–530 (2018).
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). [PubMed: 27019110]
35. Consortium GTEx. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science 369, 1318–1330 (2020). [PubMed: 32913098]
36. Wang QS et al. Leveraging supervised learning for functionally informed fine-mapping of cis-eQTLs identifies an additional 20,913 putative causal eQTLs. Nat. Commun. 12, 1–11 (2021). [PubMed: 33397941]
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). [PubMed: 34711957]
38. Dron JS & Hegele RA Genetics of lipid and lipoprotein disorders and traits. Curr. Genet. Med. Rep. 4, 130–141 (2016). [PubMed: 28286704]
39. Thompson DJ et al. Genetic predisposition to mosaic Y chromosome loss in blood. Nature 575, 652–657 (2019). [PubMed: 31748747]

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). [PubMed: 24904434]
41. Basak A et al. BCL11A deletions result in fetal hemoglobin persistence and neurodevelopmental alterations. J. Clin. Invest. 125, 2363–2368 (2015). [PubMed: 25938782]
42. Quednow BB, Brzózka MM & Rossner MJ Transcription factor 4 (TCF4) and schizophrenia: integrating the animal and the human perspective. Cell. Mol. Life Sci. 71, 2815–2835 (2014). [PubMed: 24413739]
43. Ulirsch JC et al. Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits. Cell 165, 1530–1545 (2016). [PubMed: 27259154]
44. Cvejic A et al. SMIM1 underlies the Vel blood group and influences red blood cell traits. Nat. Genet. 45, 542–545 (2013). [PubMed: 23563608]
45. Cawley NX 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, E189–97 (2010). [PubMed: 20460579]
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, e63–75 (2006). [PubMed: 16785743]
47. Budnik I & Brill A Immune factors in deep vein thrombosis initiation. Trends Immunol. 39, 610–623 (2018). [PubMed: 29776849]
48. Lambert MP, Sachais BS & Kowalska MA Chemokines and thrombogenicity. Thromb. Haemost. 97, 722–729 (2007). [PubMed: 17479182]

METHODS-ONLY REFERENCES

49. Purcell S et al. PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics (2007).
50. Loh P-R, Kichaev G, Gazal S, Schoech AP & Price AL Mixed-model association for biobank-scale datasets. Nat. Genet. 50, 906–908 (2018). [PubMed: 29892013]
51. Zhou W et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nature Genetics vol. 50 1335–1341 Preprint at 10.1038/s41588-018-0184-y (2018). [PubMed: 30104761]
52. Stuart T et al. Comprehensive integration of single-cell data. Cell 177, 1888–1902.e21 (2019). [PubMed: 31178118]
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). [PubMed: 18252563]
55. McInnes L, Healy J & Melville J UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv [stat.ML] (2018).
56. Finucane HK et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. 47, 1228–1235 (2015). [PubMed: 26414678]
57. Bulik-Sullivan B et al. An atlas of genetic correlations across human diseases and traits. Nat. Genet. 47, 1236–1241 (2015). [PubMed: 26414676]
58. UK10K Consortium et al. The UK10K project identifies rare variants in health and disease. Nature 526, 82–90 (2015). [PubMed: 26367797]
59. Csárdi G & Nepusz T The igraph software package for complex network research. https://www.semanticscholar.org › paper › The-igraph-so...https://www.semanticscholar.org › paper › The-igraph-so... (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). [PubMed: 37220626]

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). [PubMed: 28942963]

62. McLaren W et al. The Ensembl Variant Effect Predictor. Genome Biol. 17, 122 (2016). [PubMed: 27268795]

63. Cairns J et al. CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. Genome Biol. 17, (2016).

64. Roadmap Epigenomics Consortium et al. Integrative analysis of 111 reference human epigenomes. Nature 518, 317–330 (2015). [PubMed: 25693563]

65. Calderon D et al. Landscape of stimulation-responsive chromatin across diverse human immune cells. Nat. Genet. 51, 1494–1505 (2019). [PubMed: 31570894]

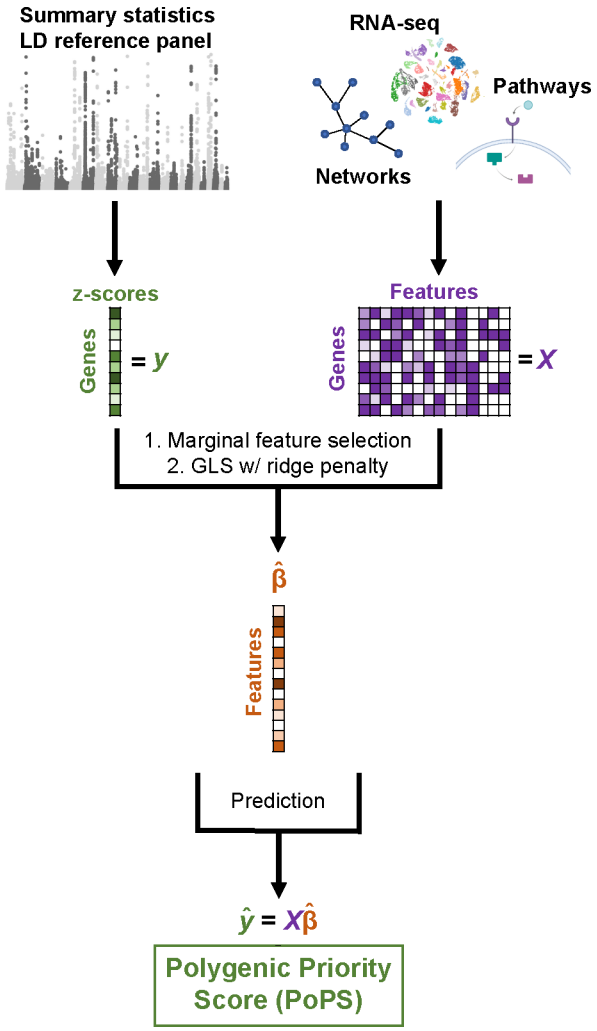


Fig. 1 |. Overview of PoPS.

We compute gene-level z-scores from GWAS summary statistics with an LD reference panel using MAGMA. We create gene features from gene expression data, biological pathways, and predicted PPI networks and use marginal feature selection to limit features included to those most likely to be relevant. We then fit a linear model for the dependence of gene-level associations on gene features using generalized least squares (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.

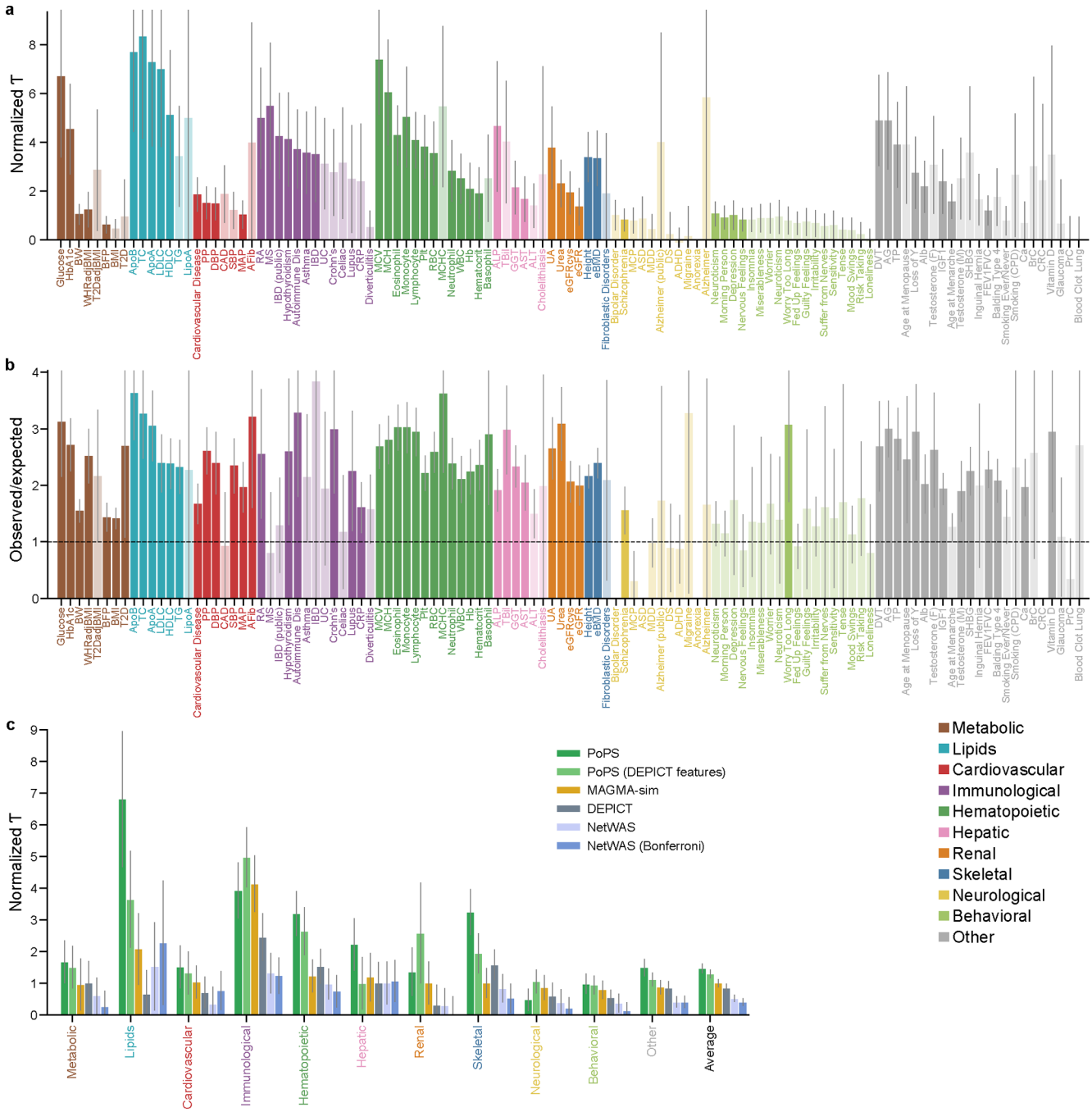


Fig. 2 |. Evaluation of PoPS and comparison to other similarity-based methods.

a, Results using Benchmarker to evaluate PoPS, grouped by trait domain and sorted by the lower bound of the 95% confidence interval 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 per SNP heritability. Error bars represent 95% confidence intervals around the point estimate. One-sided p-values were computed using the z-score test for heritability enrichment in S-LDSC. Opaque bars passed the Bonferroni significance threshold. For IBD and Alzheimer's we retained summary statistics from both

UK Biobank and other publicly available sources with a greater sample size. **b**, Results using closest gene enrichment to evaluate PoPS ordered as in panel **a**. Error bars represent 95% confidence intervals around the point estimate. One-sided p-values were computed using a normal approximation to the null distribution, and opaque bars passed the Bonferroni 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% confidence intervals around the meta-analyzed point estimate.

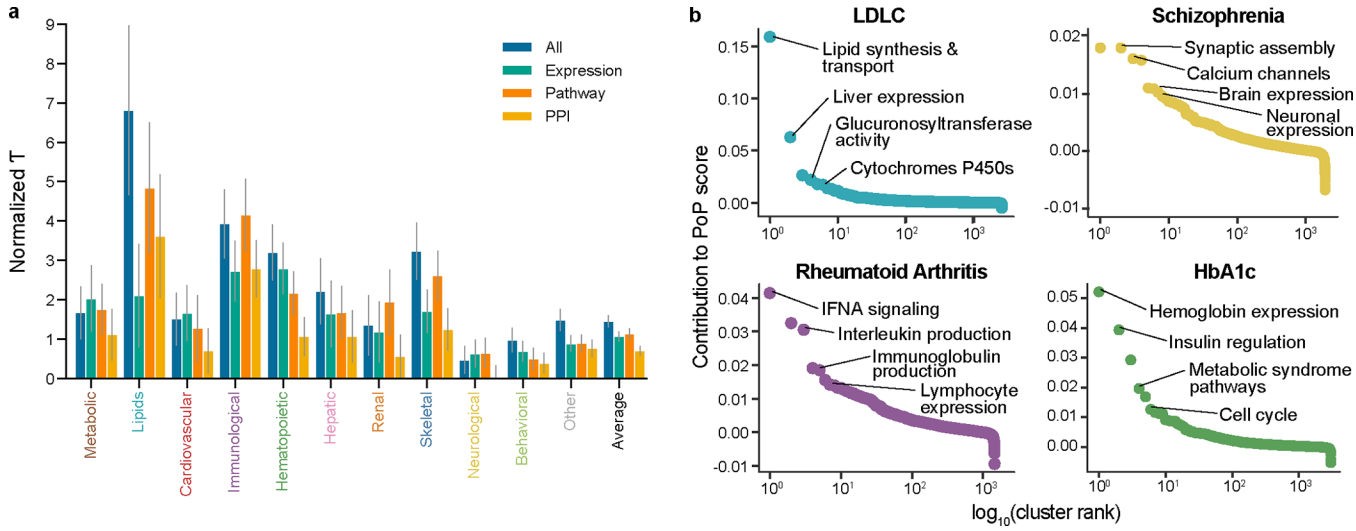


Fig. 3 |. Most informative gene features used by PoPS.

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% confidence intervals 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.

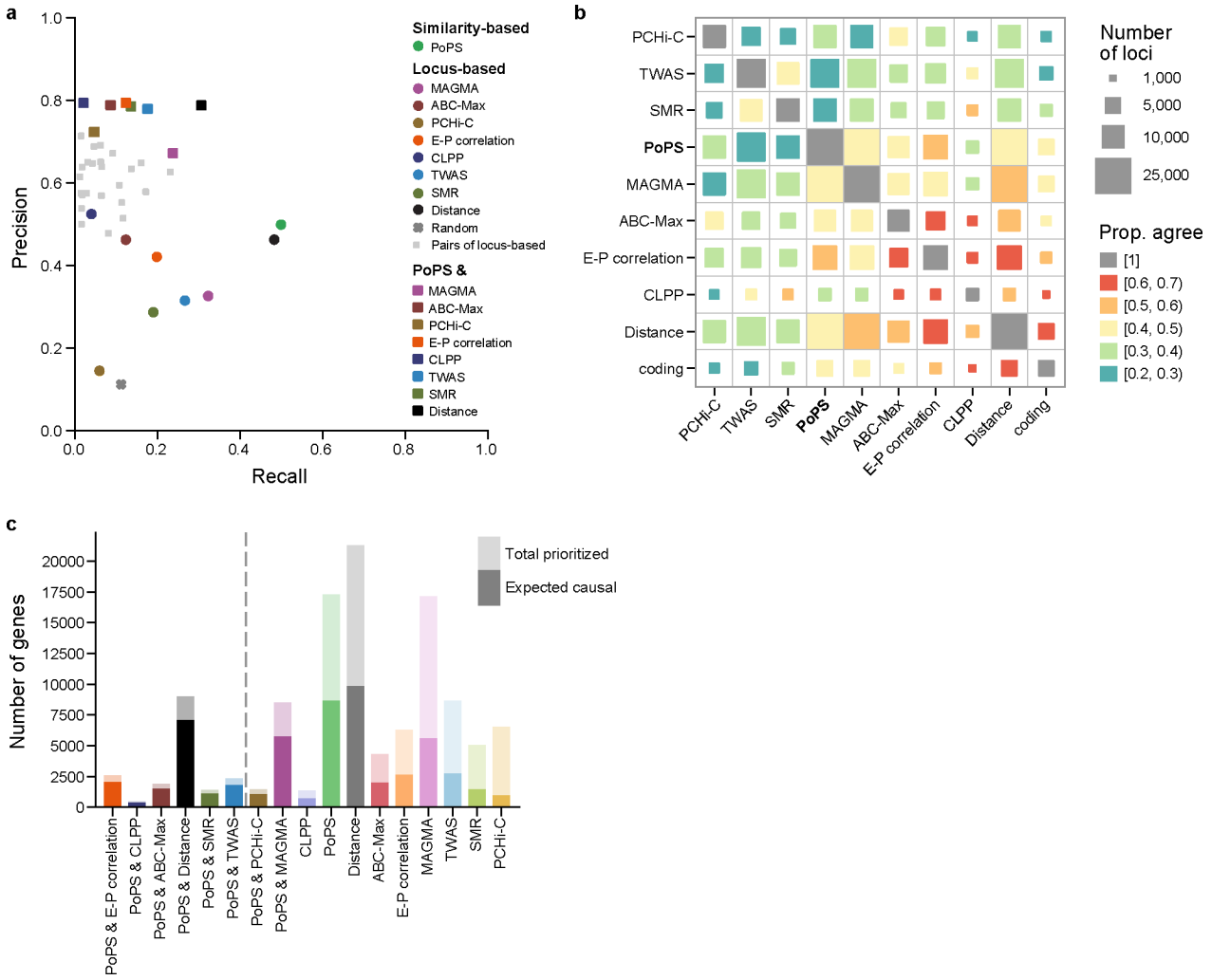


Fig. 4 |. Comparing and combining PoPS with locus-based methods.

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 non-coding 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 greater than 75%.

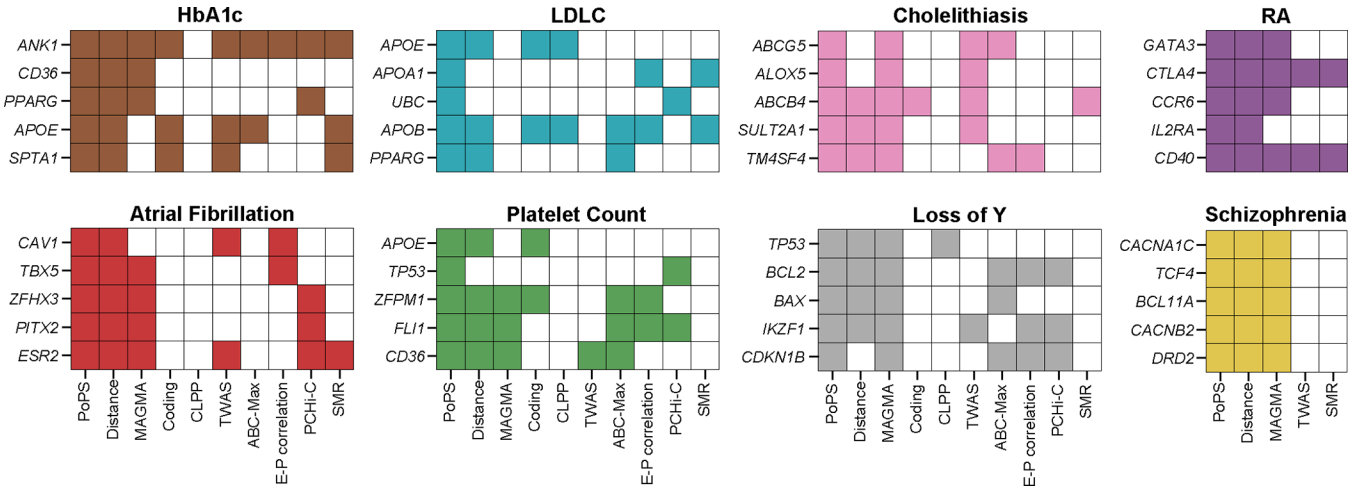


Fig. 5 |. High confidence genes for selected traits.

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

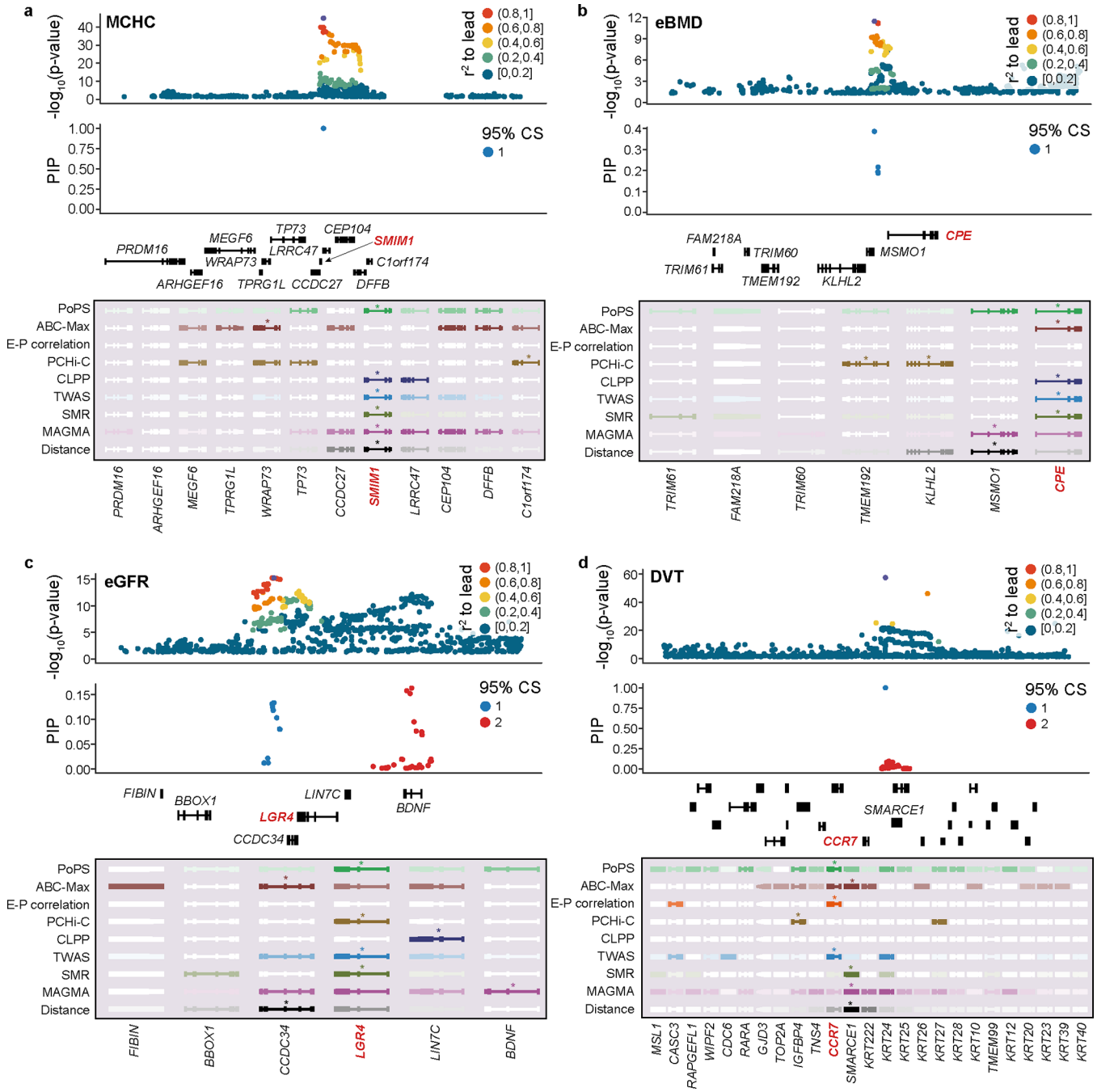


Fig. 6 |. Known and novel biological examples.

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**, rs1175550, *SMIM1* for mean corpuscular hemoglobin concentration (MCHC). **b**, rs1550270, *CPE* for bone mineral

density (eBMD). **c**, rs11029928, *LGR4* for estimated glomerular filtration rate (eGFR). **d**, rs112401631, *CCR7* for deep vein thrombosis (DVT).