Please cite this article in press as: Fine et al., Benchmarker: An Unbiased, Association-Data-Driven Strategy to Evaluate Gene Prioritization Algorithms, The American Journal of Human Genetics (2019), https://doi.org/10.1016/j.ajhg.2019.03.027

Benchmarker: An Unbiased, Association-Data-Driven Strategy to Evaluate Gene Prioritization Algorithms

Rebecca S. Fine,[1,2,3,4] Tune H. Pers,[5,6] Tiffany Amariuta,[1,3,7,8,9,10,11] Soumya Raychaudhuri,[3,7,8,9,10,12] and Joel N. Hirschhorn[1,2,3,13,*]

Genome-wide association studies (GWASs) are valuable for understanding human biology, but associated loci typically contain multiple associated variants and genes. Thus, algorithms that prioritize likely causal genes and variants for a given phenotype can provide biological interpretations of association data. However, a critical, currently missing capability is to objectively compare performance of such algorithms. Typical comparisons rely on "gold standard" genes harboring causal coding variants, but such gold standards may be biased and incomplete. To address this issue, we developed Benchmarker, an unbiased, data-driven benchmarking method that compares performance of similarity-based prioritization strategies to each other (and to random chance) by leave-one-chromosome-out cross-validation with stratified linkage disequilibrium (LD) score regression. We first applied Benchmarker to 20 well-powered GWASs and compared gene prioritization based on strategies employing three different data sources, including annotated gene sets and gene expression; genes prioritized based on gene sets had higher per-SNP heritability than those prioritized based on gene expression. Additionally, in a direct comparison of three methods, DEPICT and MAGMA outperformed NetWAS. We also evaluated combinations of methods; our results indicated that combining data sources and algorithms can help prioritize higher-quality genes for follow-up. Benchmarker provides an unbiased approach to evaluate any similarity-based method that provides genome-wide prioritization of genes, variants, or gene sets and can determine the best such method for any particular GWAS. Our method addresses an important unmet need for rigorous tool assessment and can assist in mapping genetic associations to causal function.

Introduction

Genome-wide association studies (GWASs) have successfully identified thousands of loci genetically associated with a wide range of human diseases and traits.[1] However, determining the causal variants and genes within these loci remains challenging: the true identities of the causal variants are often obfuscated by linkage disequilibrium (LD) between neighboring variants, and assigning noncoding variants to the genes they regulate has proven difficult. To address these issues, numerous types of algorithms to prioritize the most likely causal variants and genes have been developed.[2–8] Many of these algorithms are based on a simple intuition: when all potentially causal genes or variants are pooled, we expect that those that are actually causal should share more genomic features and/or annotations in common with other causal genes and variants than with non-causal genes and variants. In other words, genomic features or annotations that are shared more strongly than expected by chance in the pool of potential causal genes can be used to prioritize genes and variants. For example, algorithms have been developed that prioritize genes or variants that share similar profiles within gene sets,[9] protein-protein interaction (PPI) or co-expression networks,[6,7,10–16] PubMed abstracts,[17] and regulatory features such as DNase hypersensitivity sites.[18]

Gene prioritization is a critical step in translating genetic discoveries into biological insights, so many methods for gene prioritization have been developed. However, it is not straightforward to compare, or "benchmark," the performance of these methods and assess which of them produces the most accurate results. Most published prioritization algorithms contain a validation component, but each study takes its own approach to do this, making comparison between algorithms difficult. A common benchmarking approach is to use "gold standard" genes (i.e., genes with a known link to the trait of interest)[9,19,20] to calculate a receiver operating characteristic or similar metric. Unfortunately, this strategy relies heavily on prior knowledge of disease etiology and is biased toward well-studied genes in well-characterized biological pathways. In fact, using gold standard genes may actually penalize a method that successfully discovers novel biology (and of course the accuracy of this strategy will suffer if any of the genes classified as "gold standards" turn out not to be truly causal). Another common approach is prospective

[1]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA; [2]Division of Endocrinology and Center for Basic and Translational Obesity Research, Boston Children's Hospital, Boston, MA 02115, USA; [3]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA; [4]Ph.D. Program in Biological and Biomedical Sciences, Graduate School of Arts and Sciences, Harvard University, Cambridge, MA 02138, USA; [5]The Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, 2200 Copenhagen, Denmark; [6]Department of Epidemiology Research, Statens Serum Institut, 2300 Copenhagen, Denmark; [7]Center for Data Sciences, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA; [8]Division of Genetics, Brigham and Women's Hospital, Boston, MA 02115, USA; [9]Division of Rheumatology, Immunology, and Allergy, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA; [10]Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA; [11]Ph.D. Program in Bioinformatics and Integrative Genomics, Graduate School of Arts and Sciences, Harvard University, Cambridge, MA 02138, USA; [12]Arthritis Research UK Centre for Genetics and Genomics, Centre for Musculoskeletal Research, Manchester Academic Health Science Centre, The University of Manchester, Manchester M13 9PL, UK; [13]Department of Pediatrics, Harvard Medical School, Boston, MA 02115, USA

*Correspondence: joel.hirschhorn@childrens.harvard.edu

https://doi.org/10.1016/j.ajhg.2019.03.027.

validation, using a newer (and generally better-powered) dataset to benchmark the prioritization results from an older one. One study conducted a large benchmarking effort with this strategy by collecting 42 trait-gene associations over 6 months to use for benchmarking purposes.[21] This methodology, however, requires the existence of multiple independent well-powered GWASs, which may not always be available. Another study used Gene Ontology (GO) annotations[22] and the FunCoup network[23] to benchmark several network-based prioritization strategies using cross-validation.[24] However, this strategy assumes that a method's ability to use PPI network connectivity to recover withheld members of a GO gene set is equivalent to its ability to use that connectivity to prioritize causal genes from a GWAS, which may not be the case. Causal genes for a trait likely relate to one another in ways more complex than membership in one gene set, so this analysis may measure the relationship between GO and the FunCoup network rather than the effectiveness of the prioritization methods per se.

An ideal strategy would combine the best features of the previously described large-scale benchmarking efforts: (1) cross-validation and (2) the use of GWAS data (rather than external sources of information that can be biased or in many cases nonexistent) as a benchmark. To that end, we propose a "leave-one-chromosome-out" strategy for benchmarking, in which the full set of GWAS data is used for both prediction and validation. Specifically, to benchmark one or more similarity-based prioritization methods, we use those methods to prioritize genes on each chromosome in turn, using GWAS data for all the other chromosomes. Next, for each method being benchmarked, we assemble all of the prioritized genes on each chromosome into a single group. Finally, we apply stratified LD score regression[25] to each group of prioritized genes to determine whether the prioritized group significantly contributes to trait heritability. To compare multiple methods within a given trait, we compare the contribution to trait heritability by genes prioritized by each method. In this way, the GWAS data itself serves as its own control, without the need for incorporating additional data sources, and the use of the leave-one-chromosome-out approach prevents overfitting because association signals are not correlated across chromosomes. This strategy is highly generalizable because it can be applied to any method that prioritizes genes or variants based on their similarity to each other with respect to some feature(s) of interest (e.g., similar patterns of gene set membership, similar epigenetic marks).

Material and Methods

We first describe the GWAS data we have used to test our method. Then, we describe our approach, which we refer to as Benchmarker; we include an overview of stratified LD score regression. Finally, we discuss the specific prioritization approaches tested.

GWAS Data

We obtained GWAS summary statistics from publicly available resources (Table S1). Specifically, we used published summary statistics for height,[26] schizophrenia,[27] inflammatory bowel disease (IBD),[28] and several lipid measures (low-density lipoprotein [LDL] cholesterol, high-density lipoprotein [HDL] cholesterol, triglyceride level, and total cholesterol).[29] We also used published UK Biobank summary statistics for body-mass index (BMI), waist-hip ratio adjusted for BMI (WHRadjBMI), skin pigment, red blood cell count, white blood cell count, diastolic and systolic blood pressure, years of education, smoking status, diagnosis of allergy or eczema, age of menarche, and age of menopause.[30] The UK Biobank GWASs each consist of an average of 448,690 European samples analyzed with BOLT-LMM (with the exception of menarche and menopause, which comprise 242,278 and 143,025 female-only samples, respectively).

Benchmarker Approach

To evaluate a given prioritization method, we assume that variants near and within the set of truly causal genes will be, on average, enriched for heritability. We first apply the prioritization method of interest to a GWAS from which one chromosome has been removed (Figure 1). Methods compatible with Benchmarker will have the following general framework: (1) being able to take as input a set *S* of trait-associated genes (or variants) where all GWAS data from one or more chromosomes have been withheld and (2) for each gene/variant in the genome, including on the withheld chromosome(s), highly ranking genes/variants that are similar to those in *S*. Based on an iterative implementation of this basic strategy, withholding each chromosome in turn, a Benchmarker-compatible method can produce a similarity-based ranking of all genes or variants in the genome. We consider the top-ranked 10% of genes from the withheld chromosome to be "prioritized." We repeat this for all 22 autosomal chromosomes and combine all prioritized genes together, which represents the set to be tested.

Comparing Methods

To evaluate method performance, we turned to a well-established method: stratified LD score regression.[25] LD score regression is based on the intuition that SNPs with more LD to other SNPs are more likely to tag a truly causal variant (and therefore to have a higher χ^2 statistic).[31] An "LD score" for each SNP is calculated by summing its squared Pearson correlation coefficient (r^2) with all nearby SNPs. Specifically, the LD score of index SNP *j* is given by:

$$LD\ score\ of\ index\ SNP\ j = \sum_k r_{j,k}^2$$

where $r_{j,k}$ is the correlation between SNPs *j* and *k* calculated from a reference panel. The slope of the relationship between LD scores and observed chi-square values, computed in a weighted regression, provides a reliable estimate of overall heritability for a given trait.[25]

Stratified LD score regression (S-LDSC) is an extension of LD score regression that allows for the estimation of the heritability explained by a particular genomic annotation (e.g., coding SNPs, SNPs predicted to be enhancers).[25] In stratified LD score regression, LD scores are calculated as described above, but only consider the LD of the index SNP with other SNPs in the category of interest *C* (rather than across all SNPs). In our case, *C* represents the set of

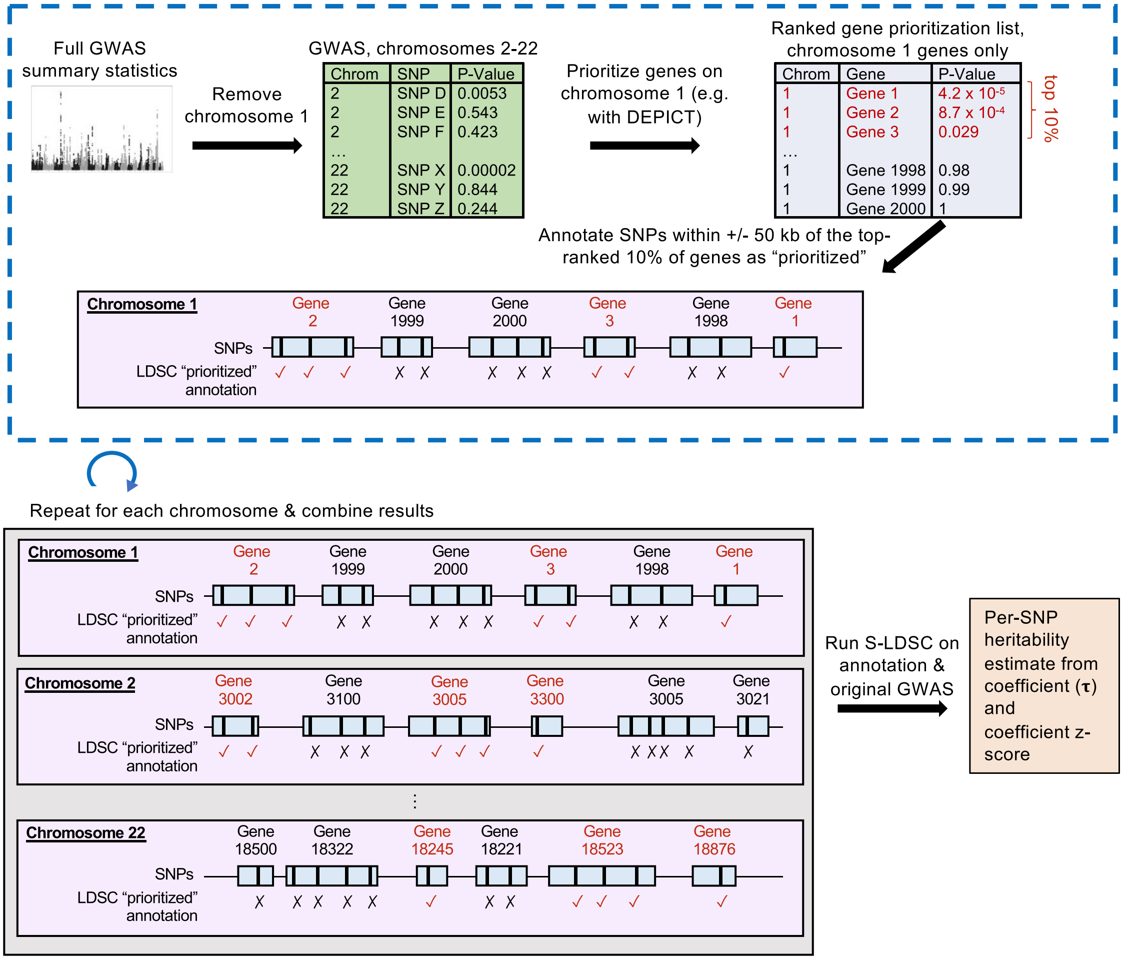


Figure 1. Schematic of the Benchmarker Strategy

prioritized SNPs (or SNPs in prioritized genes). For robust and more accurate estimation of the heritability captured by an annotation of interest, it is recommended that a set of annotations of known genomic importance be included in the S-LDSC regression model as conditional covariates. These 53 such annotations are referred to as the "baseline model"[25] and include annotations such as sequence characteristics (e.g., exon, intron) and cell-type-nonspecific regulatory marks (e.g., histone modifications). We therefore included the baseline model in each iteration of S-LDSC, as well as an additional category for SNPs that lie within 50 kb of any gene in our prioritization method as a control. Our reference SNPs for European LD score estimation were the set of 9,997,231 SNPs with a minor allele count ≥ 5 from 489 unrelated European individuals in Phase 3 of 1000 Genomes.[32] Heritability was partitioned for the set of 5,961,159 SNPs with MAF $\geq$ 0.05; regression coefficient estimation was performed with 1,217,312 HapMap3 SNPs (SNPs in HapMap3 are used because they are generally well imputed).

To evaluate model performance, we focused on two metrics: the regression coefficient τ and its p value (derived from a block jackknife) for our annotation. τ measures the average per-SNP contribution of the annotation to heritability after accounting for the other categories in the model. We note that, since we used the top 10% of genes for each method, the number of prioritized SNPs for each trait will vary mainly by the average gene length of prioritized genes. To make τ comparable across traits, we normalized by the average per-SNP heritability for each trait (i.e., we divided each estimate by the total trait heritability / the number of SNPs used to compute total trait heritability); we refer to this as "normalized τ." In previous work,[33] normalized τ may be multiplied by the standard deviation of the annotation; this quantity, τ^*, measures the increase in per-SNP heritability per standard deviation increase in the annotation value. For a binary annotation, the standard deviation corresponds directly to the proportion of SNPs assigned to that annotation. For our purposes, we were less interested in the change in heritability associated with a one-standard-deviation increase in the annotation value (which, for a binary annotation, does not have an intuitive interpretation) and more interested in the change in heritability associated with the value of the annotation increasing from 0 to 1, which is captured by normalized τ rather than τ^*. We therefore computed normalized τ as an evaluation metric throughout our analyses (this normalization has also been used in Finucane et al.[34]).

To compute p values for pairwise comparisons of prioritization methods, we used a block jackknife to compute the standard error around the difference between two τ estimates within a trait (i.e., $\tau_A - \tau_B$). We also calculated random-effects meta-analysis p values from the normalized τ values to determine whether one annotation generally outperformed another across multiple traits, using the R package rmeta. To do this, we selected only one GWAS from groups of obviously overlapping sets of traits; specifically, from the lipid trait group we retained LDL cholesterol and excluded HDL cholesterol, triglycerides, and total cholesterol, and from the blood pressure traits we included systolic blood pressure and excluded diastolic blood pressure. We also report overall meta-analyzed τ along with standard error and p value estimates from each analysis (Table S2). Results were visualized in R version 3.5, using ggplot2[35] and ComplexHeatmap.[36]

Choice of Parameters

Benchmarker relies on two major parameters: (1) the percentage cutoff to use for top-ranked prioritized genes and (2) the size of the window used to map SNPs to genes. We used 10% and 50 kb, respectively, but our results do not differ substantially with respect to comparative method performance with alternative values of 5%, 15%, 25 kb, and 100 kb (Figure S1). There is also precedence in the literature for both of these parameters;[34,37] we chose 10% largely to strike a balance between minimizing standard error and obtaining a number of genes that would not be too large to be useful in prioritization. However, users can vary these parameters using the provided scripts.

Assessment of Type 1 Error

For each SNP-based type 1 error simulation, we randomly selected 10% of the SNPs that were within ±50 kb of any gene on each chromosome. These SNPs were then used as an annotation for input to S-LDSC. We tested each simulation against summary statistics from 10 different well-powered GWAS (for a total of 10,000 runs: 1,000 null simulations each using 10 GWASs), including BMI, diastolic blood pressure, age of menopause, height, schizophrenia, years of education, age of menarche, IBD, total cholesterol, and allergy/eczema. For the gene-based type 1 error simulations, we randomly selected 10% of genes and prioritized all SNPs within ±50 kb of those genes; these were then tested the same way as in the SNP-based type 1 error analysis (10,000 runs: 1,000 null simulations each tested with 10 GWAS).

Prioritization Methods

We began by evaluating DEPICT (release 194),[9] a method for gene prioritization, gene set enrichment analysis, and tissue enrichment analysis. DEPICT's primary innovation is the use of "reconstituted" gene sets, which consist of 14,462 gene sets downloaded from multiple databases that have been extended based on 77,840 publicly available expression microarrays.[38] The reconstituted gene sets contain z-scores for each gene in the genome for each of the 14,462 gene sets, representing how strongly each gene is predicted to be a member of each gene set. DEPICT's gene prioritization algorithm involves (1) identifying all genes in trait-associated loci (referred to as *S*), which in DEPICT are defined as all genes overlapping any SNP with $r^2 > 0.5$ to an index variant and (2) for each of the genes identified in *S*, assessing its correlation with the rest of the genes in *S* across the reconstituted gene sets. The stronger the overall correlation a gene has with the rest of the genes in *S*, the more highly it will be prioritized. We adapted this method for Benchmarker by forcing the prioritization to be genome-wide rather than across only the genes in *S* (that is, each gene in the genome is compared to the genes in *S* across the reconstituted gene sets). DEPICT requires a GWAS p value threshold to define "trait-associated loci." We used $p < 1 \times 10^{-5}$ for our analyses here, except for a few GWASs for which this threshold caused DEPICT to exceed its maximum number of loci. For these GWASs (height, BMI, WHRadjBMI, red blood cell count, white blood cell count, diastolic blood pressure, and systolic blood pressure), we used a cutoff of $p < 5 \times 10^{-8}$.

As described, DEPICT's default behavior is to prioritize genes based on correlation across the reconstituted gene sets. The implicit assumption in this method is that the genes most likely to be truly causal are the ones with the most similar profile across these gene sets. As a first question for Benchmarker, we asked: how would correlating across tissue expression (rather than gene set membership) fare in comparison? To answer this question, we applied DEPICT's prioritization algorithm, exchanging the matrix of z-scores of reconstituted gene sets for a matrix of z-scores of expression data (gene expression for each gene across a range of tissues). We used two different expression data sources. The first was the matrix DEPICT typically uses to perform tissue enrichment analysis, which was derived from 37,427 publicly available human microarrays representing 209 different tissues (based on Medical Subject Heading annotations). We note that this is a subset of the microarray data from the Gene Expression Omnibus (GEO)[39] used in the process of "reconstituting" the gene sets. The second source of expression data was a matrix based on RNA sequencing data from the Genotype Tissue Expression project (GTEx, v6),[40] including 53 human tissues with an average of 161.32 samples per tissue (processed as in Finucane et al.[34]). For each tissue, we calculated the mean expression across all samples. To make the expression data as comparable as possible, we normalized the GTEx expression matrix in the same way as the GEO matrix:[9] z-score normalizing across all tissues, then across genes. For all DEPICT analyses, we considered the top-ranked 10% of genes "prioritized." All analyses were done on a set of 16,876 genes that were (1) outside the major histocompatibility complex region (chromosome 6: 25-35 Mb in hg19 genome build) and (2) present in both the DEPICT and GTEx data.

The second method we evaluated was MAGMA (v1.06b).[41] MAGMA works in two steps. First, a gene-based p value is computed as the mean association of SNPs in the gene, corrected for LD. Then, competitive gene set and/or continuous covariate p values are calculated based on the association of the gene-based p values with the category of interest. We ran MAGMA with default parameters. For gene set enrichment analysis, we treated the reconstituted gene sets as a continuous covariate and calculated one-tailed p values (alternative hypothesis = genes with high gene set membership z-scores have a stronger trait association than those with low gene set membership z-scores).

Comparing DEPICT to MAGMA presents an immediate challenge: MAGMA does not explicitly prioritize genes based on the gene set enrichment analysis, as DEPICT does. Therefore, we needed to establish a framework for deriving gene prioritization from gene set enrichment analysis results (Figure S2). Specifically, we first generated gene-based p values from MAGMA. Then, we removed all genes from each chromosome in turn and applied the gene set enrichment analysis function to the remaining genes. From the gene set enrichment analysis results, we reasoned that we could prioritize genes that were members of the most highly enriched gene sets. However, the reconstituted gene sets used for the gene set enrichment do not have "members" per se; rather, they have z-scores for gene set membership prediction. To address this issue, we created several binarized forms of the gene sets in which we used z-score cutoffs ($Z > 1.96$, 2.58, or 3.29, which correspond to two-tailed p values of 0.05, 0.01, and 0.001, respectively) or rankings (top 50, 100, or 200 genes per gene set) to define the gene set "members" (Figures S2 and S3). (For the $Z > 3.29$ condition, we removed 14 gene sets that contained fewer than

10 genes.) Then, for each withhold-one-chromosome gene set enrichment analysis result, we (1) ranked the gene sets and (2) for each gene set, annotated each member gene on the withheld chromosome as “prioritized” until we reached 10% of genes on the withheld chromosome (where gene set members were based on one of the six versions of the binarized gene sets). For the purposes of comparison, we applied DEPICT in exactly the same way, performing the prioritization based directly on the leave-one-chromosome-out gene set enrichment results and the binarized gene sets. We note that this basic strategy is itself a useful technique for converting the results of any gene set enrichment analysis to gene prioritization; this may be helpful in generalizing the types of data that can be used with Benchmarker.

The third method we evaluated was NetWAS,[12,42] which prioritizes genes based on PPI network connectivity. A total of 144 tissue-specific PPI networks are available for use with the algorithm, in addition to a tissue-naive global network. NetWAS takes gene-level p values from a GWAS as input and uses genes below a specified p value threshold (e.g., $p < 0.01$) as “positive examples” of trait association. The positive examples are used in a support vector machine classifier, which learns the patterns of network connectivity in a user-specified tissue and uses them to reprioritize all genes in the genome.

For our NetWAS analysis, we first generated gene-level p values for each of our 20 GWASs with MAGMA. We removed all genes that were not present in the previously defined DEPICT-GTEx overlapping set of 16,876 genes. The recommended p value threshold cutoff for setting “positive examples” is nominal significance ($p < 0.01$).[12,42] However, our GWASs are so well powered that using $p < 0.01$ could define too many genes as positive examples for the method to be successful. We therefore tested three different p value cutoffs for each trait: $p < 0.01$, $p < 0.0001$, and a Bonferroni-corrected threshold for the number of genes tested (roughly $p < 3 \times 10^{-6}$, differing slightly from trait to trait). We tested these three thresholds with the “global” (i.e., tissue-nonspecific) PPI network for all 20 GWASs. Then, for nine of the GWASs, we also tested one to four relevant tissue-specific networks, chosen based on a combination of DEPICT tissue enrichment results and published S-LDSC analyses,[34] and compared their performance to the global network. For each trait, we used the p value threshold that was most successful from the global analyses.

eQTL Analyses

We used data from a recent analysis of blood expression quantitative trait loci (eQTLs) from 31,684 individuals,[43] including all significant *cis*-eQTLs (a total of 3,699,823 SNPs regulating 16,989 genes). For all S-LDSC eQTL analyses, an additional control annotation was included that consisted of either (1) all annotated significant *cis*-eQTLs (for the analysis in which we split all prioritized SNPs into eQTLs and non-eQTLs) or (2) all annotated *cis*-eQTLs regulating at least one gene in our dataset (for the analysis in which we mapped each prioritized gene to eQTLs found to regulate it).

Nearest-Gene Analysis

For each GWAS, we used the default clumping and loci generation procedure from DEPICT to define genes in significant loci (which we refer to as *S*), as well as the closest gene to each index SNP. The p value threshold for “significant” loci was defined as described above for the DEPICT analyses ($p < 1 \times 10^{-5}$ or $p < 5 \times 10^{-8}$, depending on the number of loci for the trait). For each of our seven analyses (comparison of DEPICT-gene-sets/DEPICT-GEO/DEPICT-GTEx and comparison of DEPICT and MAGMA for all six binarizations), we then restricted genes prioritized by at least one method to those in *S* (which we refer to as *S*-prioritized) and did the same for the intersect (*S*-intersect) and the outersect (*S*-outersect) (i.e., *S*-prioritized is the union of *S*-intersect and *S*-outersect). We performed three Fisher’s exact tests for each trait, comparing the fraction of nearest-genes in *S*-prioritized, *S*-intersect, and *S*-outersect to the fraction of nearest-genes in *S* overall.

Results

The Benchmarker framework is outlined in Figure 1. First, we remove one chromosome from a set of GWAS summary statistics. Then, we apply a gene prioritization method of interest to this partial GWAS; this produces prioritization p values for each gene in the genome, including on the withheld chromosome. We rank the genes on the withheld chromosome by prioritization p value and take the top 10% as “prioritized.” Then, we annotate all SNPs within ±50 kb of these genes as prioritized SNPs. We repeat this process for each chromosome, successively withholding each chromosome and annotating prioritized SNPs. We then combine all the prioritized SNPs for each chromosome into a single “prioritized” annotation. Finally, we apply stratified LD score regression (S-LDSC), which produces an estimate of the average per-SNP heritability (τ) of the prioritized annotation. For each of our applications, we tested GWASs of 20 different traits; to improve comparability across traits, we normalize our estimates of τ by the average genome-wide per-SNP heritability of each trait. We note that this method can easily be used for variant prioritization strategies in addition to gene prioritization strategies, with the only difference being that prioritized variants on the withheld chromosome can be directly annotated for stratified LD score regression without the step of mapping variants to genes.

We first assessed whether the type 1 error rate of our method was well controlled by conducting 1,000 null simulations using randomly prioritized SNPs, each tested against ten different sets of actual GWAS summary statistics, for a total of 10,000 runs (see Material and Methods). The coefficient (τ) z-scores were well controlled, with a one-tailed type 1 error rate of 0.0456 (95% confidence interval = 0.0415, 0.050) at $p = 0.05$ (Figure S4A; we report one-tailed p values because we were concerned about type 1 error that overestimates rather than underestimates heritability). This result indicates that S-LDSC is well calibrated and correctly determines that a group of randomly selected SNPs does not significantly explain any heritability in any of the tested GWAS (i.e., in actual GWAS results). To more explicitly mimic our experimental setup, we also tested type 1 error by randomly sampling 10% of genes rather than variants (Figure S4B; again, 1,000 null simulations each tested against 10 different GWASs, for a total of 10,000 runs). Then, we annotated all SNPs within ±50 kb of these genes as "prioritized." For these simulations, we observed a slightly conservative one-tailed type 1 error rate of 0.0441 (95% confidence interval = 0.040, 0.048) at p = 0.05; we therefore proceeded without additional correction.

Next, we applied Benchmarker to compare three different methods of running DEPICT's prioritization algorithm: (1) prioritizing based on shared patterns of gene set membership (DEPICT's standard approach, using a gene × gene set matrix of z-scores) and (2) prioritizing based on shared patterns of tissue expression, using either of two different gene × tissue expression matrices of z-scores (one microarray-based from GEO[39] and one RNA-seq-based from GTEx;[40] see Material and Methods). We refer to these approaches as DEPICT-gene-sets, DEPICT-GEO, and DEPICT-GTEx, respectively. We observed that the prioritized variants had coefficients significantly greater than zero for nearly all traits and all three input matrices, indicating that prioritization by DEPICT using any of these data sources captures meaningful additional heritability beyond the effects of the baseline model (Figure S5; Table S3). To more directly compare these annotations to each other, we performed a conditional analysis in which all three annotations were modeled jointly (i.e., in the same S-LDSC model). This indicated that over all traits, DEPICT-gene-sets performed better than both DEPICT-GEO (random-effects meta-analysis p value calculated over 16 traits = 2.40×10^{-3}; we excluded four phenotypically overlapping traits, see Material and Methods) and DEPICT-GTEx (meta-analysis p value = 8.00×10^{-6}) (Table S4).

All three of these methods use gene expression information as a data source, either explicitly (for the two gene × tissue expression matrices) or implicitly (for the gene × gene set matrix, where z-scores for gene set membership are derived from gene expression data). We therefore considered whether the three methods prioritize similar genes. However, when we compared the overlap of prioritized genes, we found that in fact the approaches prioritized relatively distinct groups of genes. Specifically, the average number of genes prioritized by all three methods for a given trait was 374.1; in contrast, the average number of genes prioritized by DEPICT-gene-sets only, DEPICT-GEO only, and DEPICT-GTEx only were 796.7, 732.4, and 757.7, respectively (Figure 2). Therefore, the use of three different data sources produced three substantially different groups of genes that each significantly contribute to heritability, suggesting that each set of prioritized genes contains different and useful information.

Because of this partial overlap in prioritized genes across the three methods, we tested whether we could combine results to create a smaller set of prioritized genes that still captured most of the heritability in the genes prioritized across different methods. Specifically, we created a new annotation consisting of genes prioritized by at least two of the three input matrices, which we refer to as the "intersect" set (average number of genes = 1,203.05; average proportion of SNPs = 0.072) (Figures S6 and S7). We also created a separate annotation consisting of the remaining prioritized genes, i.e., genes prioritized by only one of those three methods; for simplicity, we refer to this group as the "outersect" set (average number of genes = 2,286.80; average proportion of SNPs = 0.120). We first tested the performance of the "intersect" and "outersect" annotations in separate LD score regression models (Figure S5; Table S3) and found that across most traits, the intersect performed better than the outersect, in some cases significantly. We then directly compared the intersect and outersect in a joint S-LDSC model (Figure 3; Table S5). We observed a clear difference between the two annotations: across most traits, the intersect group of genes substantially outperformed the outersect (random-effects meta-analysis p value for 16 traits = 1.39×10^{-4}). These results indicate that the majority of the heritability explained by all prioritized genes could actually be localized to SNPs near or within the genes prioritized by more than one method. To analyze this further, we also tested intersect and outersect sets for additional combinations of the DEPICT implementations (Figure S8). All three methods showed the same pattern of intersect genes outperforming outersect genes. Additionally, consistent with

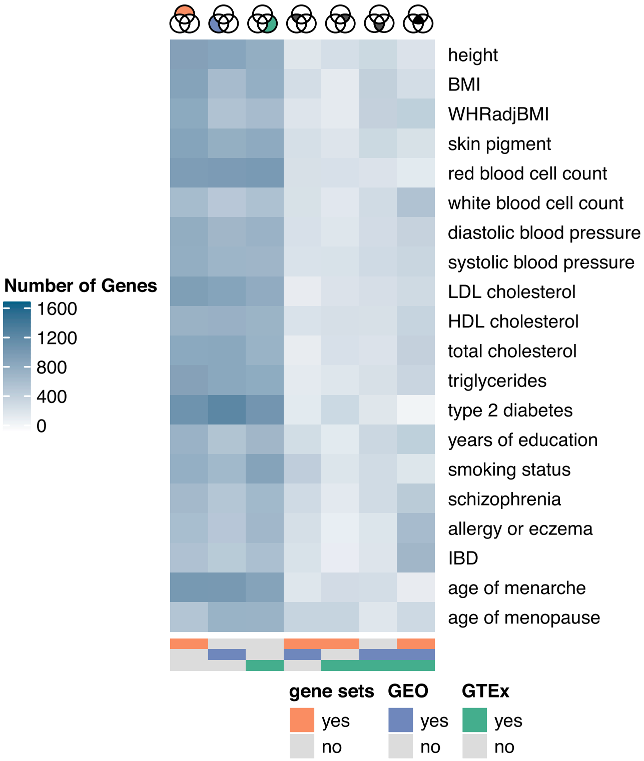


Figure 2. Overlap in Prioritized Genes for DEPICT-Gene-Sets, DEPICT-GEO, and DEPICT-GTEx
Columns of the heatmap represent all possible categories of overlap, also illustrated by the Venn diagrams on top and the annotation bar below (e.g., prioritized by DEPICT-gene-sets only, prioritized by DEPICT-gene-sets and DEPICT-GEO, prioritized by all three methods, etc.). Darker blues indicate more genes in the category.

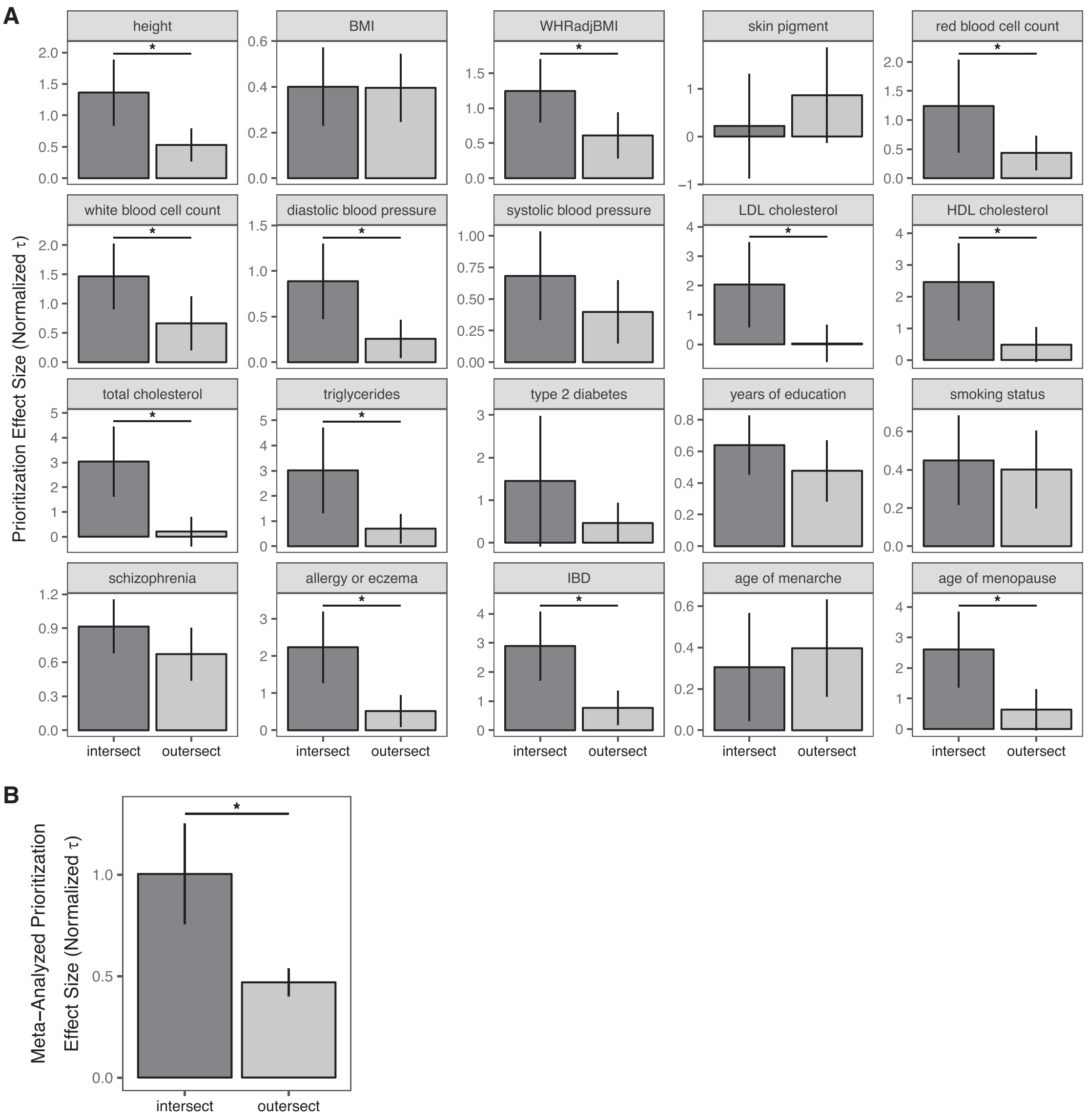


Figure 3. Effect Sizes (Normalized τ) for the Joint LD Score Regression Model Comparing Intersect and Outersect Genes for 20 Different GWASs

Here, the intersect represents genes prioritized by at least two of (1) DEPICT-gene-sets, (2) DEPICT-GEO, and (3) DEPICT-GTEx. The outersect represents genes prioritized by only one of those three methods. Asterisks mark comparisons for which the difference between the intersect and outersect achieved nominal significance ($p < 0.05$). Error bars represent 95% confidence intervals.

(A) Results for each trait; note that y-axis scales differ for each panel.

(B) Results meta-analyzed over 16 traits.

the observation that DEPICT-gene-sets outperformed the other two DEPICT approaches, the intersect of genes prioritized by DEPICT-gene-sets and at least one other implementation performed best (in fact, somewhat better than the intersect containing genes prioritized by all three implementations).

We also noted that the intersect outperformed the outersect most strongly for the lipid traits (LDL and HDL cholesterol, total cholesterol, and triglycerides), immune traits (allergy/eczema and IBD), and height. In contrast, we observed that the most brain-related traits we tested (BMI, years of education, smoking status, schizophrenia, and age of menarche) failed to show a nominally significant difference between the intersect and outersect ($p > 0.05$) (meta-analysis p value for brain-related traits = 0.210; meta-analysis p value for all other traits = 3.72×10^{-6}). (We consider these traits to be brain related based on empirical evidence from both S-LDSC analyses on tissue-specific expression and DEPICT analyses on general tissue expression enrichment.[25,34,44,45]) This suggests that brain-related traits may not benefit as much as other traits from combining information across these particular data sources (i.e., the reconstituted gene sets and tissue expression matrices).

We next wanted to compare DEPICT with another popular gene set enrichment analysis algorithm, MAGMA.[41]

However, MAGMA does not perform gene prioritization based on its gene set enrichment, so we needed a way to convert prioritized gene sets to prioritized genes. We accomplished this by (1) ranking the prioritized gene sets, (2) using binarized versions of the reconstituted gene sets to assign genes to gene sets, and (3) prioritizing genes in the most enriched gene sets on the withheld chromosome (Figure S2). For step 2, we created six different binarized versions of the reconstituted gene sets, three based on z-score ($Z > 1.96$, 2.58, or 3.29) and three based on ranking (top 50, 100, or 200 genes per gene set). For the purposes of comparison, we used the same strategy for DEPICT (i.e., using the binarized gene sets to identify and prioritize genes on the withheld chromosome within enriched gene sets). This basic approach can be used for any method that prioritizes genomic features (e.g., gene sets, tissue expression, epigenomic annotations) but does not necessarily explicitly prioritize genes or variants; it also illustrates that the Benchmarker strategy can be used to evaluate a wide variety of algorithm types.

DEPICT and MAGMA performed similarly: we observed no strongly significant differences between the two methods for each trait, either modeled separately or jointly (Figure S9; Tables S6 and S7). However, as we observed for the comparison across different data sources, we again noted that DEPICT and MAGMA prioritized fairly different groups of genes (average number of genes prioritized by both methods = 931.18, average number of genes prioritized by one method only = 757.82) (Figures 4 and S10–S12). Using the same logic as before, we again asked whether the genes found by both methods (the intersect) outperformed the genes found by either method individually (the outersect). (Note that the outersect includes the union of genes prioritized by only DEPICT and only MAGMA, so it is on average slightly larger than the intersect [Figure S12].) Interestingly, we observed an even stronger trend toward the overlapping set outperforming the individual sets than for the analysis of different data sources, and for several traits this difference was nominally significant (Figure S9; Table S6). When we modeled the intersect and outersect jointly, this difference became even more apparent, with the intersecting group of genes outperforming the outersect for nearly every trait (Tables 1 and S8; Figure 5). The differences were particularly pronounced for immune traits (IBD, allergy/eczema) and total cholesterol (which, interestingly, also were some of the best-performing traits in the previous analysis). For these traits, not only did the intersect significantly outperform the outersect for every gene set binarization, but outersect per-SNP heritability often did not significantly differ from zero. This implies that the majority of the heritability originally explained by both sets of prioritized genes (i.e., the union of DEPICT and MAGMA) was captured by the intersect genes only, with almost none remaining in the outersect.

We also observed that the top 50 genes, top 100 genes, and $Z > 3.29$ conditions showed the most significant difference between intersect and outersect performance (Table 1), and the $Z > 3.29$ intersect also had the highest overall per-SNP heritability (Table S2). These conditions also represent the gene sets with the smallest number of genes (Figure S3), suggesting that for our choice of method, more stringent cutoffs for gene set membership produce more power. Given our prioritization schema (Figure S2; see Material and Methods), this is not surprising: when gene sets have a very large number of members, a few gene sets will dominate the gene prioritization results. In contrast, when smaller gene sets are used, more gene sets will end up contributing prioritized genes, likely beneficially increasing diversity.

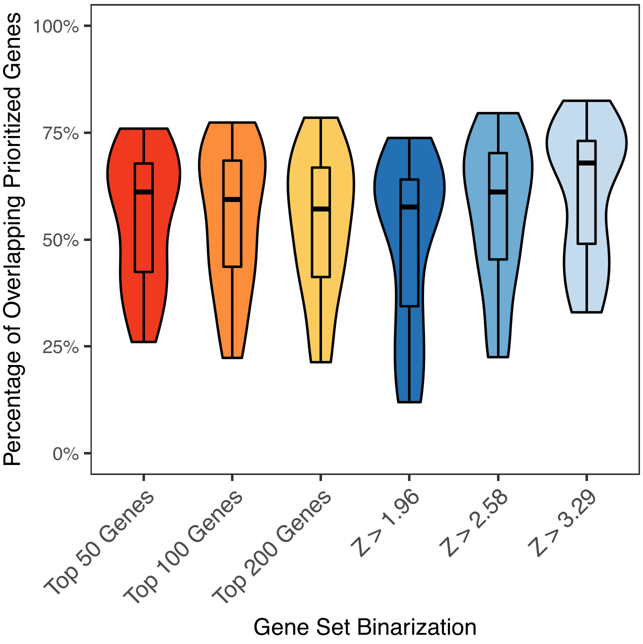


Figure 4. Overlap in Prioritized Genes from DEPICT and MAGMA
For each version of the binarized gene sets (x-axis), the distribution of the percentage of overlapping genes across all 20 traits is represented as a violin plot overlaid with a boxplot.

We next performed some additional validation for these sets of prioritized genes. In the following sections, for brevity, we will refer to the DEPICT-gene-sets versus DEPICT-GEO versus DEPICT-GTEx analysis as “DEPICT-datatype” and the DEPICT versus MAGMA analysis as “DEPICT-MAGMA.” We considered whether the intersects from these analyses are enriched for the genes nearest to GWAS index SNPs, as such genes are slightly more likely to be causal than the other genes in associated loci. We observed that for most traits, the set of prioritized genes was enriched for nearest-genes and that the intersect was much more strongly enriched than the outersect (Table S9; see Material and Methods). This suggests that at least some of the information captured by nearest-genes is also captured by the different prioritization methods and that this is largely driven by genes in the intersects.

In addition, to validate prioritized genes, we used a previously curated list of 277 genes from Online Mendelian

Table 1. Comparison of Intersect and Outersect Genes for DEPICT and MAGMA

Gene Set Binarization	Meta-analysis p Value for $\tau_{\text{intersect}}$ versus $\tau_{\text{outersect}}$	Height OMIM Enrichment Odds Ratio (Intersect versus Outersect)	Height OMIM Enrichment p Value (Intersect versus Outersect)
Top 50 genes	8.546×10^{-7}	2.531	1.541×10^{-4}
Top 100 genes	1.558×10^{-5}	2.421	2.101×10^{-4}
Top 200 genes	1.084×10^{-4}	1.364	0.185
$Z > 1.96$	2.537×10^{-4}	0.832	0.437
$Z > 2.58$	8.489×10^{-5}	1.298	0.243
$Z > 3.29$	9.110×10^{-6}	1.869	6.180×10^{-3}

Each row includes data based on one of the six gene set binarizations. The columns represent the overall p value for comparing $\tau_{\text{intersect}}$ and $\tau_{\text{outersect}}$ and results from the height OMIM enrichment analysis (odds ratio and p value from Fisher's exact test). OMIM, Online Mendelian Inheritance in Man.

Inheritance in Man (OMIM) known to harbor variants that cause disorders of skeletal growth.[26] We note that in general, we do not endorse gold standards, but we use them here as an orthogonal source of validation data for height, where there are a large number of well-established and well-validated Mendelian disease-associated genes. We performed a Fisher's exact test to determine whether the intersect genes for height were enriched for these genes relative to the outersect genes. We found that in the DEPICT-datatype analysis, the intersect was indeed enriched for OMIM genes (odds ratio = 2.004, $p = 4.72 \times 10^{-4}$). Similarly, for the DEPICT-MAGMA analysis, in the three versions of the binarized gene sets with the most significant meta-analysis p value for intersect versus outersect, we also observed an enrichment of OMIM genes (Table 1).

Our prioritization annotations, which encompass the entire gene body and a 50-kb window for each prioritized gene, will include a fair amount of noise, since many of the SNPs assigned to each gene will not be in LD with variants that have functional effects on that gene. Therefore, we next investigated whether we could improve our heritability enrichment at a variant level by incorporating expression quantitative loci (eQTL) information, using a recently published set of blood *cis*-eQTL data from more than 31,000 individuals.[43] Based on these data, we first performed S-LDSC on a single annotation consisting of all significant eQTL variants regulating at least one gene in our dataset. We observed that this annotation generally explained a small but statistically significant amount of per-SNP heritability (Table S2, Figure S13, meta-analysis p value = 1.02×10^{-3}). Unsurprisingly, given that the eQTLs were derived from blood, the annotation was the most significant for three of our most immune-relevant traits: white blood cell count, IBD, and allergy/eczema. Furthermore, by separating out the subset of eQTL variants that are associated with expression of prioritized genes in our intersect sets and testing them jointly with the all-eQTL annotation, we found that the heritability explained for IBD and allergy/eczema by the all-eQTL annotation was largely driven specifically by this subset of eQTLs (Figure S13).

In general, prioritization of variants that were *cis*-eQTLs of our prioritized intersect genes (which can include SNPs up to 1 Mb away from a prioritized gene) resulted in worse performance than simply including all SNPs within 50 kb of prioritized genes, with the possible exception of the most brain-relevant traits (for DEPICT-datatype intersect, meta-analysis p value = 0.0612; for DEPICT-MAGMA intersect with $Z > 3.29$, $p = 3.63 \times 10^{-4}$; Figure S14). We also performed an additional analysis in which we split our intersect sets of SNPs (i.e., all SNPs within 50 kb of each prioritized gene) into two categories each: eQTL (i.e., listed as a significant *cis*-eQTL for any gene in the genome) and non-eQTL. For many, though not all, traits, these eQTL variants performed better than the non-eQTL variants in a joint model (for DEPICT-datatype intersect, p = 0.0212; for DEPICT-MAGMA intersect, $p = 4.24 \times 10^{-4}$; Figure S15). Thus, eQTL information likely helps with prioritization, but the prioritization may be particularly helpful at the variant level rather than at the gene level. Indeed, the overall meta-analyzed normalized τ values for eQTLs within or near prioritized genes from our two intersect sets were the highest observed for any of our analyses (Table S2).

Finally, we evaluated NetWAS, an algorithm that uses a different approach than either DEPICT or MAGMA: prioritizing genes from a GWAS based on patterns of PPI network connectivity.[12,42] We used Benchmarker to test different parameter options with NetWAS. NetWAS uses gene-level p values from a GWAS as input, and the user provides a p value threshold below which NetWAS will consider a gene a "positive" example for trait association in model training. We tested three p value thresholds: 0.01, 0.0001, and a Bonferroni-corrected p value for the number of genes tested (roughly 3×10^{-6}, depending on the trait). Using these thresholds, we evaluated NetWAS performance on our 20 GWASs using the provided global (tissue-nonspecific) network. We observed that in general, the Bonferroni-corrected p value threshold performed best; for most traits, NetWAS with at least one of the p value thresholds performed above chance (Figure S16, Table S10). However, with the parameters we tested, NetWAS generally did not perform as well as our DEPICT-gene-sets or MAGMA analyses (in a joint model including NetWAS, DEPICT-gene sets, and MAGMA, DEPICT-gene-sets versus NetWAS p = 0.0178, MAGMA versus NetWAS $p = 2.28 \times 10^{-3}$, DEPICT-gene-sets versus MAGMA

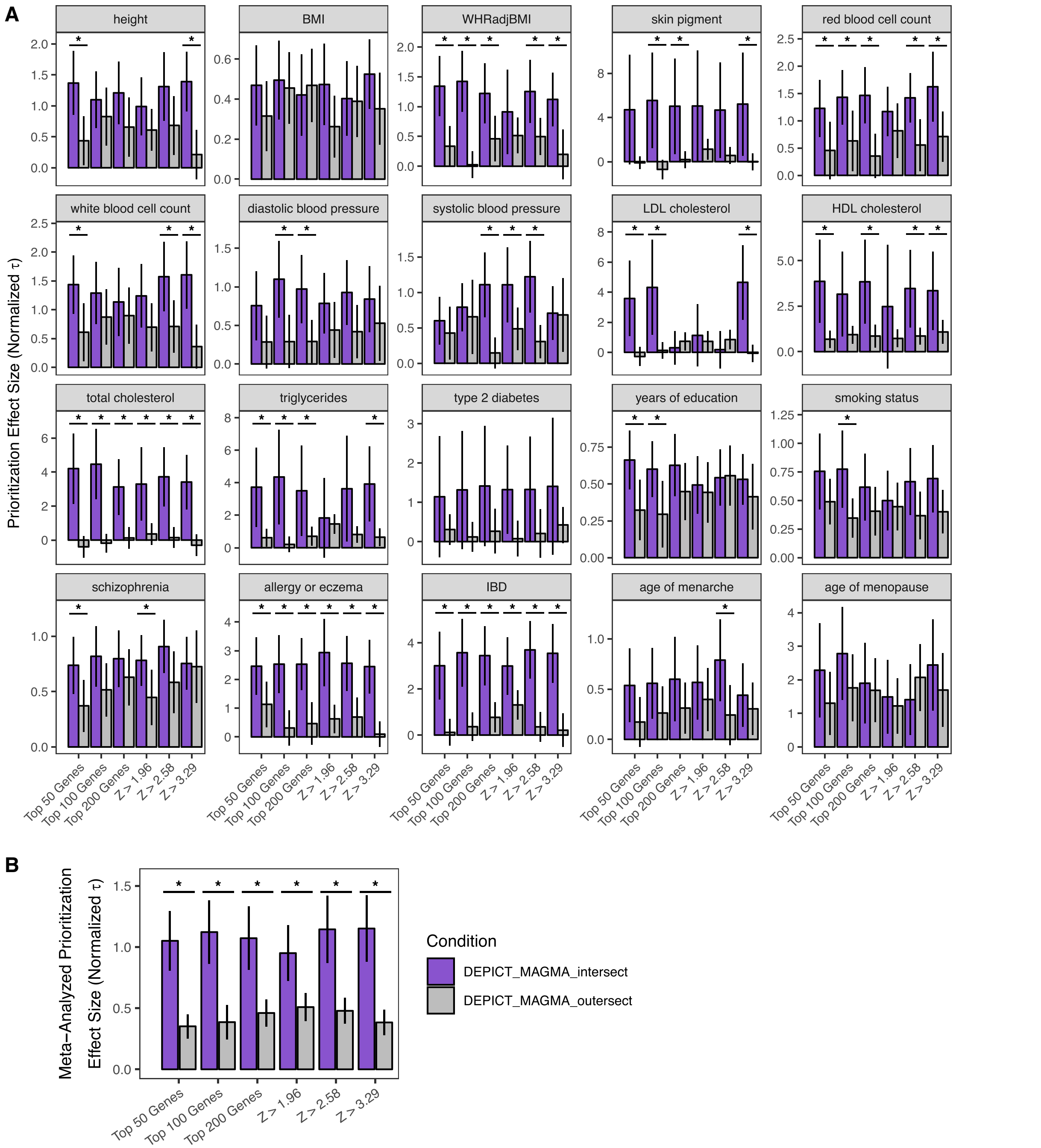


Figure 5. Effect Sizes (Normalized τ) for the Joint LD Score Regression Model Comparing Intersect and Outersect Genes for 20 Different GWASs

Here, the intersect represents genes prioritized by both DEPICT and MAGMA. The outersect represents genes prioritized by only DEPICT or only MAGMA. Asterisks mark comparisons for which the difference between the intersect and outersect achieved nominal significance ($p < 0.05$). Error bars represent 95% confidence intervals.

(A) Results for each trait; note that y-axis scales differ for each panel.

(B) Results meta-analyzed over 16 traits.

$p = 0.635$; Figure 6; Table S11). We also carried out an additional analysis of NetWAS using several tissue-specific PPI networks; for each of nine traits, we chose one to four relevant tissues based on a combination of evidence from S-LDSC[34] and DEPICT. For the majority of these analyses, the global network outperformed all of the tested tissue-specific networks (Figure S17, Table S10). In general, the brain-related tissues ("brain," "cerebellum," "cerebral cortex," and "hippocampus") and the blood/immune-related tissues ("blood," "spleen," "leukocyte," "lymphocyte," and "bone marrow") came closest to the global network performance (when used with relevant phenotypes).

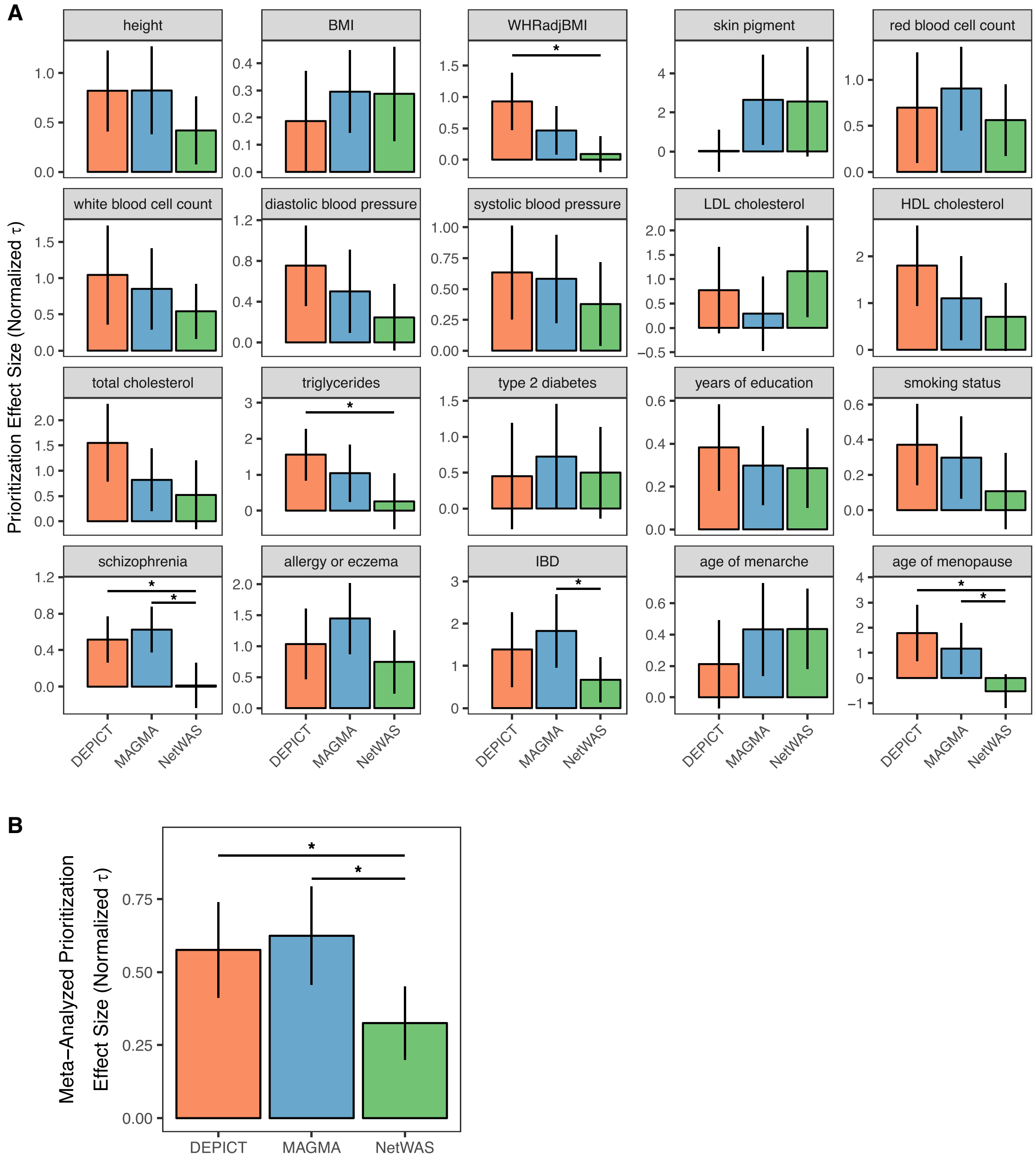


Figure 6. Effect Sizes (Normalized τ) for the Joint LD Score Regression Model Comparing DEPICT-Gene-Sets, MAGMA (with the Z > 2.58 Gene Set Binarization), and NetWAS (Bonferroni-Corrected p Value Threshold with the Global Network)
Asterisks mark comparisons for which the difference between any pair of conditions achieved nominal significance ($p < 0.05$). Error bars represent 95% confidence intervals.
(A) Results for each trait; note that y-axis scales differ for each panel.
(B) Results meta-analyzed over 16 traits.

Discussion

We have developed and implemented a leave-one-chromosome-out approach for benchmarking similarity-based gene prioritization algorithms. The use of heritability explained as a metric has several advantages over methods that have been used elsewhere.[21,24] First, it does not rely on the idea of "true positive" or "gold standard" genes, which is inherent in commonly used area-under-the-curve benchmarks; true positives are only as good as the existing knowledge base on a given trait and are likely biased toward a subset of relevant biology. Second, heritability explained directly measures an actual quantity of interest, namely, how well the method can independently identify genes that colocalize with GWAS association signals. Our leave-one-chromosome-out strategy is also an obviously unbiased way of measuring method performance, as it enables the use of the GWAS data itself as its own control rather than external sources of data. We recommend that this idea should be used in benchmarking new and existing prioritization methods; that is, testing should be done on genes or variants prioritized on a chromosome (or a set of chromosomes) that has been withheld from the input data. We also recommend leaving out at least an entire chromosome to avoid overfitting from correlation of association signals from neighboring genes. In this way, it is possible to both have the advantages of cross-validation without the disadvantages of relying on gold standards. With this approach, Benchmarker can assess any method that prioritizes genes or variants based on common features between the associated genes/variants and the left-out genes/variants, as long as the features do not depend on the GWAS association results themselves (because those are used as the benchmark). We also note that some existing prioritization methods that theoretically are capable of prioritizing genome-wide are not implemented to do so (i.e., they prioritize only genes in trait-associated loci rather than across the genome); we recommend that future developers of such methods include this capability, at a minimum for benchmarking purposes.

In two different sets of comparisons, one using different data sources and one using different prioritization algorithms, we showed that selecting genes prioritized by multiple approaches outperforms the use of genes prioritized by exactly one approach. Supporting this observation, genes associated with Mendelian skeletal growth disorders are more enriched in "intersect" than "outersect" genes for height; similarly, nearest-genes from GWAS data are generally more enriched in intersect than outersect genes across most tested traits. This finding has important implications for translating genetic associations into biological insights, as it empirically demonstrates that combining prioritization approaches is superior to relying on the somewhat arbitrary choice of a single approach. We also observed that intersect performance was particularly strong for immune and lipid traits. In contrast, brain-related traits generally showed fewer significant differences between intersect and outersect gene performance. One possible explanation is the heterogeneity of the brain, which has extremely high regional and cell-type specificity. Gene prioritization for these traits might therefore improve with a different approach, such as restricting tissue expression data to brain regions only, or, conversely, including only a single representative brain region in the tissue expression analysis rather than many. Another possible reason for this finding is the importance of brain-region-specific isoform expression, which is not accounted for in our general tissue expression data.

We also explored the possibility of combining gene prioritization with additional variant-level information, using eQTLs as an example. First, we showed that SNPs within 50 kb of prioritized genes that are annotated as blood *cis*-eQTLs outperform similar variants not annotated as eQTLs for some (but not all) traits, and that this combination of *cis*-eQTL information and prioritized genes from the intersect sets yielded a set of variants with particularly high per-variant heritability explained. We speculate that this enrichment of heritability arises in part because, regardless of the tissue they affect, SNPs that are eQTLs are more likely to be in functional elements; therefore, partitioning the prioritized SNPs into eQTLs and non-eQTLs may help separate SNPs enriched for function from SNPs enriched for inactivity. It is interesting that this finding was particularly pronounced for brain-related traits, where our prioritization methods produced lower per-SNP heritability estimates in general; this could reflect general differences in genetic architecture of brain-related traits compared to others, an idea for which there is some support from other S-LDSC analyses.[46,47]

In addition, we used Benchmarker to analyze a PPI-based method, NetWAS, and determined that it also prioritized genes better than random chance but did not perform quite as well as our other tested approaches. We emphasize, however, that many parameters in NetWAS could be changed and optimized beyond those we tested, and Benchmarker could be used to evaluate such potential improvements (for example, different methods could be used to generate the gene-based p values used as input). In addition, the NetWAS analysis suggests another useful application of Benchmarker: determining the best combination of prioritization approaches and tissue-specific datasets to use for any trait of interest, in a manner complementary to that used in Finucane et al.[34]

We also note that, because Benchmarker relies on a cross-validation strategy, it can be used to fairly determine the best prioritization method for any given trait. This strategy provides a route to best practices for gene prioritization for the field: by benchmarking multiple approaches (and particularly the intersection of multiple approaches), it should be possible to objectively improve the gene prioritization of any given trait. For example, we observed that age of menarche was one of the few traits for which combining information across gene-set- and tissue-matrix-based prioritization did not improve per-SNP heritability; one possible explanation based on our original analysis is that the GTEx matrix alone did not significantly contribute to heritability, so using information from that analysis may have actually worsened the signal-to-noise ratio. Such observations have the potential to inform specific decisions for performing gene prioritization for GWASs from individual traits. In addition, with our analysis of MAGMA, we have shown that this benchmarking strategy can be extended to methods that evaluate enrichment of genomic features (e.g., pathways or tissue expression) but do not explicitly assign prioritization p values to genes. This high generalizability will allow for the comparison of a wide variety of approaches for different traits.

An important caveat to our results is that LD score regression may have insufficient power to identify small differences in explained heritability using different approaches to gene prioritization and that other metrics (such as the genomic inflation factor for prioritized variants) may be more sensitive. We note, however, that LD score regression is a widely-used method in the field and has found important and significant differences in heritability explained by a variety of other types of annotations, such as cell-type-specific expression[34] and epigenomic marks.[25] We believe that any small losses in power from our choice of method are outweighed by the benefits of the output (i.e., per-SNP heritability) being directly interpretable and extremely meaningful. Furthermore, improvements to the annotations we have used here will provide boosts in power (as well as answers to other questions about how effective such "improvements" actually are). For example, using a 50-kb window around prioritized genes, as we have done here, means that a large amount of noise is included in our annotations. A major improvement would therefore be assignment of noncoding SNPs to the genes they regulate based on expression, epigenetic, and/or chromatin conformation data; we demonstrated one such possibility with our eQTL analyses, but there are many additional approaches that could be taken.

In conclusion, we have developed a powerful and well-controlled approach for benchmarking gene prioritization strategies that relies solely on GWAS data and does not require any assumptions about the "correct" biology. Our method shows that combining prioritization strategies can improve heritability enrichment and suggests a strong recommendation that follow-up studies be focused on genes prioritized using multiple approaches. Future prioritization methods will benefit from incorporating different statistical approaches and data sources; even apparently similar data (such as two different sources of tissue expression) can provide different and complementary information. We believe that the overall cross-validation approach described and implemented here provides a better "gold standard" for benchmarking existing and future methods for gene and variant prioritization. Finally, Benchmarker can be used to determine the best algorithm and dataset for any particular trait of interest.

Supplemental Data

Supplemental Data can be found online at https://doi.org/10.1016/j.ajhg.2019.03.027.

Acknowledgments

We gratefully acknowledge Hilary Finucane, Steven Gazal, Eric Bartell, and Jack Kosmicki for helpful discussion. This research has been conducted using the UK Biobank resource. This work was supported by NHGRI NIH F31HG009850 (R.S.F.), NIDDK NIH R01DK075787 and R01DK105154 (J.N.H.), the Lundbeck Foundation R190-2014-3904 and the Novo Nordisk Foundation NNF18CC0034900 (T.H.P.), NHGRI NIH T32 HG002295 (T.A.), and NIAMS NIH 1R01AR063759-01A1 (S.R.).

Declaration of Interests

J.N.H. is a member of the scientific advisory board of Camp4 Therapeutics. All other authors declare no competing interests.

Received: December 14, 2018
Accepted: March 28, 2019
Published: May 2, 2019

Web Resources

Benchmarker, https://github.com/RebeccaFine/benchmarker
DEPICT software, https://data.broadinstitute.org/mpg/depict/ (release 194)
eQTLs, https://molgenis26.gcc.rug.nl/downloads/eqtlgen/cis-eqtl/
GEO, https://www.ncbi.nlm.nih.gov/geo/
GTEx, https://www.gtexportal.org
LDSC software, https://github.com/bulik/ldsc (version 1.0)
LD scores and other files for LDSC, https://data.broadinstitute.org/alkesgroup/LDSCORE/
MAGMA software, https://ctg.cncr.nl/software/magma (version 1.06b)
NetWAS, https://hb.flatironinstitute.org/api/, https://hb.flatironinstitute.org/netwas/

References

1. Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. Am. J. Hum. Genet. *101*, 5–22.
2. Moreau, Y., and Tranchevent, L.-C. (2012). Computational tools for prioritizing candidate genes: boosting disease gene discovery. Nat. Rev. Genet. *13*, 523–536.
3. Tranchevent, L.C., Capdevila, F.B., Nitsch, D., De Moor, B., De Causmaecker, P., and Moreau, Y. (2011). A guide to web tools to prioritize candidate genes. Brief. Bioinform. *12*, 22–32.
4. Spain, S.L., and Barrett, J.C. (2015). Strategies for fine-mapping complex traits. Hum. Mol. Genet. *24* (R1), R111–R119.
5. Ward, L.D., and Kellis, M. (2012). Interpreting noncoding genetic variation in complex traits and human disease. Nat. Biotechnol. *30*, 1095–1106.
6. Cowen, L., Ideker, T., Raphael, B.J., and Sharan, R. (2017). Network propagation: a universal amplifier of genetic associations. Nat. Rev. Genet. *18*, 551–562.
7. Jia, P., and Zhao, Z. (2014). Network.assisted analysis to prioritize GWAS results: principles, methods and perspectives. Hum. Genet. *133*, 125–138.
8. Hou, L., and Zhao, H. (2013). A review of post-GWAS prioritization approaches. Front. Genet. *4*, 280.
9. Pers, T.H., Karjalainen, J.M., Chan, Y., Westra, H.-J., Wood, A.R., Yang, J., Lui, J.C., Vedantam, S., Gustafsson, S., Esko, T., et al.; Genetic Investigation of ANthropometric Traits (GIANT) Consortium (2015). Biological interpretation of genome-wide association studies using predicted gene functions. Nat. Commun. *6*, 5890.
10. Taşan, M., Musso, G., Hao, T., Vidal, M., MacRae, C.A., and Roth, F.P. (2015). Selecting causal genes from genome-wide association studies via functionally coherent subnetworks. Nat. Methods *12*, 154–159.
11. Rossin, E.J., Lage, K., Raychaudhuri, S., Xavier, R.J., Tatar, D., Benita, Y., Cotsapas, C., Daly, M.J.; and International Inflammatory Bowel Disease Genetics Constortium (2011). Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. PLoS Genet. *7*, e1001273.

12. Greene, C.S., Krishnan, A., Wong, A.K., Ricciotti, E., Zelaya, R.A., Himmelstein, D.S., Zhang, R., Hartmann, B.M., Zaslavsky, E., Sealfon, S.C., et al. (2015). Understanding multicellular function and disease with human tissue-specific networks. Nat. Genet. *47*, 569–576.
13. Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., and Sharan, R. (2010). Associating genes and protein complexes with disease via network propagation. PLoS Comput. Biol. *6*, e1000641.
14. Lee, I., Blom, U.M., Wang, P.I., Shim, J.E., and Marcotte, E.M. (2011). Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome Res. *21*, 1109–1121.
15. Shim, J.E., Bang, C., Yang, S., Lee, T., Hwang, S., Kim, C.Y., Singh-Blom, U.M., Marcotte, E.M., and Lee, I. (2017). GWAB: a web server for the network-based boosting of human genome-wide association data. Nucleic Acids Res. *45* (W1), W154–W161.
16. Gottlieb, A., Magger, O., Berman, I., Ruppin, E., and Sharan, R. (2011). PRINCIPLE: a tool for associating genes with diseases via network propagation. Bioinformatics *27*, 3325–3326.
17. Raychaudhuri, S., Plenge, R.M., Rossin, E.J., Ng, A.C.Y., Purcell, S.M., Sklar, P., Scolnick, E.M., Xavier, R.J., Altshuler, D., Daly, M.J.; and International Schizophrenia Consortium (2009). Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. PLoS Genet. *5*, e1000534.
18. Schmidt, E.M., Zhang, J., Zhou, W., Chen, J., Mohlke, K.L., Chen, Y.E., and Willer, C.J. (2015). GREGOR: evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach. Bioinformatics *31*, 2601–2606.
19. Tranchevent, L.-C., Ardeshirdavani, A., ElShal, S., Alcaide, D., Aerts, J., Auboeuf, D., and Moreau, Y. (2016). Candidate gene prioritization with Endeavour. Nucleic Acids Res. *44* (W1), W117-21.
20. Picart-Armada, S., Barrett, S.J., Willé, D.R., Perera-Lluna, A., Gutteridge, A., and Dessailly, B.H. (2018). Benchmarking network propagation methods for disease gene identification. bioRxiv. https://doi.org/10.1101/439620.
21. Börnigen, D., Tranchevent, L.C., Bonachela-Capdevila, F., Devriendt, K., De Moor, B., De Causmaecker, P., and Moreau, Y. (2012). An unbiased evaluation of gene prioritization tools. Bioinformatics *28*, 3081–3088.
22. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al.; The Gene Ontology Consortium (2000). Gene ontology: tool for the unification of biology. Nat. Genet. *25*, 25–29.
23. Schmitt, T., Ogris, C., and Sonnhammer, E.L.L. (2014). FunCoup 3.0: database of genome-wide functional coupling networks. Nucleic Acids Res. *42*, D380–D388.
24. Guala, D., and Sonnhammer, E.L.L. (2017). A large-scale benchmark of gene prioritization methods. Sci. Rep. *7*, 46598.
25. Finucane, H.K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K., et al.; ReproGen Consortium; Schizophrenia Working Group of the Psychiatric Genomics Consortium; and RACI Consortium (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. *47*, 1228–1235.
26. Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., Chu, A.Y., Estrada, K., Luan, J., Kutalik, Z., et al.; Electronic Medical Records and Genomics (eMERGE) Consortium; MIGen Consortium; PAGE Consortium; and LifeLines Cohort Study (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. Nat. Genet. *46*, 1173–1186.
27. Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature *511*, 421–427.
28. Liu, J.Z., van Sommeren, S., Huang, H., Ng, S.C., Alberts, R., Takahashi, A., Ripke, S., Lee, J.C., Jostins, L., Shah, T., et al.; International Multiple Sclerosis Genetics Consortium; and International IBD Genetics Consortium (2015). Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat. Genet. *47*, 979–986.
29. Teslovich, T.M., Musunuru, K., Smith, A.V., Edmondson, A.C., Stylianou, I.M., Koseki, M., Pirruccello, J.P., Ripatti, S., Chasman, D.I., Willer, C.J., et al. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. Nature *466*, 707–713.
30. Loh, P.R., Kichaev, G., Gazal, S., Schoech, A.P., and Price, A.L. (2018). Mixed-model association for biobank-scale datasets. Nat. Genet. *50*, 906–908.
31. Bulik-Sullivan, B.K., Loh, P.-R., Finucane, H.K., Ripke, S., Yang, J., Patterson, N., Daly, M.J., Price, A.L., Neale, B.M.; and Schizophrenia Working Group of the Psychiatric Genomics Consortium (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. *47*, 291–295.
32. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., Abecasis, G.R.; and 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.
33. Gazal, S., Finucane, H.K., Furlotte, N.A., Loh, P.-R., Palamara, P.F., Liu, X., Schoech, A., Bulik-Sullivan, B., Neale, B.M., Gusev, A., and Price, A.L. (2017). Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. Nat. Genet. *49*, 1421–1427.
34. Finucane, H.K., Reshef, Y.A., Anttila, V., Slowikowski, K., Gusev, A., Byrnes, A., Gazal, S., Loh, P.R., Lareau, C., Shoresh, N., et al.; Brainstorm Consortium (2018). Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. Nat. Genet. *50*, 621–629.
35. Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis (Verlag New York: Springer).
36. Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics *32*, 2847–2849.
37. Lamparter, D., Marbach, D., Rueedi, R., Kutalik, Z., and Bergmann, S. (2016). Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics. PLoS Comput. Biol. *12*, e1004714.
38. Fehrmann, R.S.N., Karjalainen, J.M., Krajewska, M., Westra, H.-J., Maloney, D., Simeonov, A., Pers, T.H., Hirschhorn, J.N., Jansen, R.C., Schultes, E.A., et al. (2015). Gene expression analysis identifies global gene dosage sensitivity in cancer. Nat. Genet. *47*, 115–125.
39. Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., et al. (2013). NCBI GEO: archive for

functional genomics data sets–update. Nucleic Acids Res. *41*, D991–D995.

40. GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. Nat. Genet. *45*, 580–585.

41. de Leeuw, C.A., Mooij, J.M., Heskes, T., and Posthuma, D. (2015). MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput. Biol. *11*, e1004219.

42. Wong, A.K., Krishnan, A., and Troyanskaya, O.G. (2018). GIANT 2.0: genome-scale integrated analysis of gene networks in tissues. Nucleic Acids Res. *46* (W1), W65–W70.

43. Võsa, U., Claringbould, A., Westra, H.-J., Jan Bonder, M., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Kasela, S., et al. (2018). Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis. bioRxiv. https://doi.org/10.1101/447367.

44. Day, F.R., Thompson, D.J., Helgason, H., Chasman, D.I., Finucane, H., Sulem, P., Ruth, K.S., Whalen, S., Sarkar, A.K., Albrecht, E., et al.; LifeLines Cohort Study; InterAct Consortium; kConFab/AOCS Investigators; Endometrial Cancer Association Consortium; Ovarian Cancer Association Consortium; and PRACTICAL consortium (2017). Genomic analyses identify hundreds of variants associated with age at menarche and support a role for puberty timing in cancer risk. Nat. Genet. *49*, 834–841.

45. Locke, A.E., Kahali, B., Berndt, S.I., Justice, A.E., Pers, T.H., Day, F.R., Powell, C., Vedantam, S., Buchkovich, M.L., Yang, J., et al.; LifeLines Cohort Study; ADIPOGen Consortium; AGEN-BMI Working Group; CARDIOGRAMplusC4D Consortium; CKDGen Consortium; GLGC; ICBP; MAGIC Investigators; MuTHER Consortium; MIGen Consortium; PAGE Consortium; ReproGen Consortium; GENIE Consortium; and International Endogene Consortium (2015). Genetic studies of body mass index yield new insights for obesity biology. Nature *518*, 197–206.

46. Gazal, S., Loh, P.-R., Finucane, H.K., Ganna, A., Schoech, A., Sunyaev, S., and Price, A.L. (2018). Functional architecture of low-frequency variants highlights strength of negative selection across coding and non-coding annotations. Nat. Genet. *50*, 1600–1607.

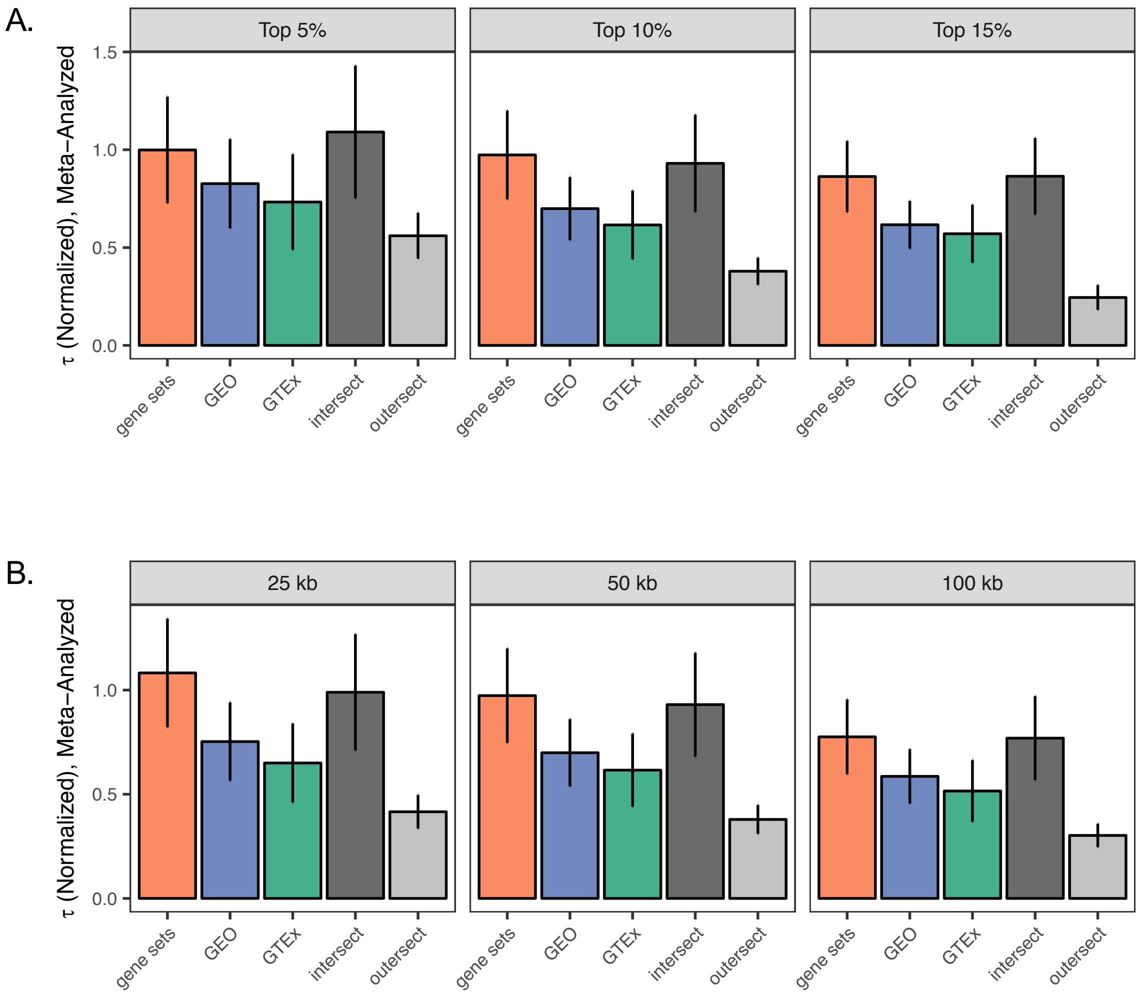
47. O'Connor, L.J., Schoech, A.P., Hormozdiari, F., Gazal, S., Patterson, N., and Price, A.L. (2018). Polygenicity of complex traits is explained by negative selection. bioRxiv. https://doi.org/10.1101/420497.

The American Journal of Human Genetics, Volume *104*

Supplemental Data

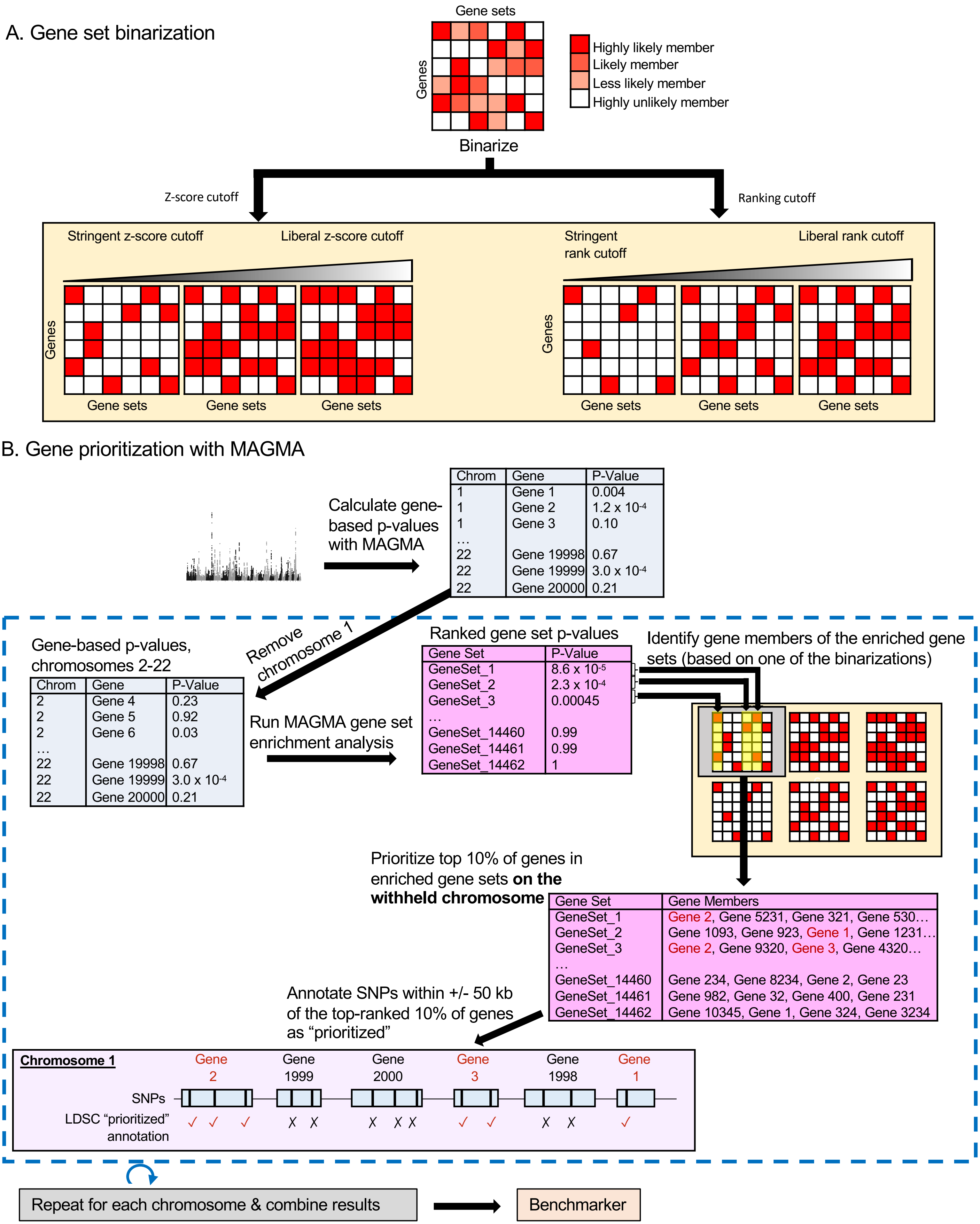
Benchmarker: An Unbiased, Association-Data-Driven Strategy to Evaluate Gene Prioritization Algorithms

Rebecca S. Fine, Tune H. Pers, Tiffany Amariuta, Soumya Raychaudhuri, and Joel N. Hirschhorn



Supplemental Figure 1. Meta-analyzed effect sizes (normalized τ) for varying parameters of Benchmarker.
a) Varying percentage cutoffs for prioritizing genes (top 5%, top 10%, and top 15%).
b) Varying kb window size for assigning SNPs to genes (25 kb, 50 kb, and 100 kb).

Figure includes data for separate LD score regression models comparing 1) DEPICT-gene-sets, 2) DEPICT-GEO, 3) DEPICT-GTEx, 4) the "intersect" (genes prioritized by at least two of the three preceding methods) and 5) the "outersect" (genes prioritized by only one method). ("Separate LD score regression models" indicates that these annotations are tested individually rather than jointly). Error bars represent 95% confidence intervals.


A. Gene set binarization
Gene sets
Genes
Highly likely member
Likely member
Less likely member
Highly unlikely member
Binarize
Z-score cutoff
Ranking cutoff
Stringent z-score cutoff
Liberal z-score cutoff
Stringent rank cutoff
Liberal rank cutoff
Genes
Gene sets
Gene sets
Gene sets
Genes
Gene sets
Gene sets
Gene sets
B. Gene prioritization with MAGMA
Calculate gene-based p-values with MAGMA
Remove chromosome 1
Gene-based p-values, chromosomes 2-22
Run MAGMA gene set enrichment analysis
Ranked gene set p-values
Identify gene members of the enriched gene sets (based on one of the binarizations)
Prioritize top 10% of genes in enriched gene sets **on the withheld chromosome**
Annotate SNPs within +/- 50 kb of the top-ranked 10% of genes as "prioritized"
Chromosome 1
Gene 2
Gene 1999
Gene 2000
Gene 3
Gene 1998
Gene 1
SNPs
LDSC "prioritized" annotation
✓ ✓ ✓ ✗ ✗ ✗ ✗ ✗ ✓ ✓ ✗ ✗ ✓
Repeat for each chromosome & combine results
Benchmarker

Chrom	Gene	P-Value
1	Gene 1	0.004
1	Gene 2	1.2 x 10^{-4}
1	Gene 3	0.10
…		
22	Gene 19998	0.67
22	Gene 19999	3.0 x 10^{-4}
22	Gene 20000	0.21

Gene-based p-values, chromosomes 2-22

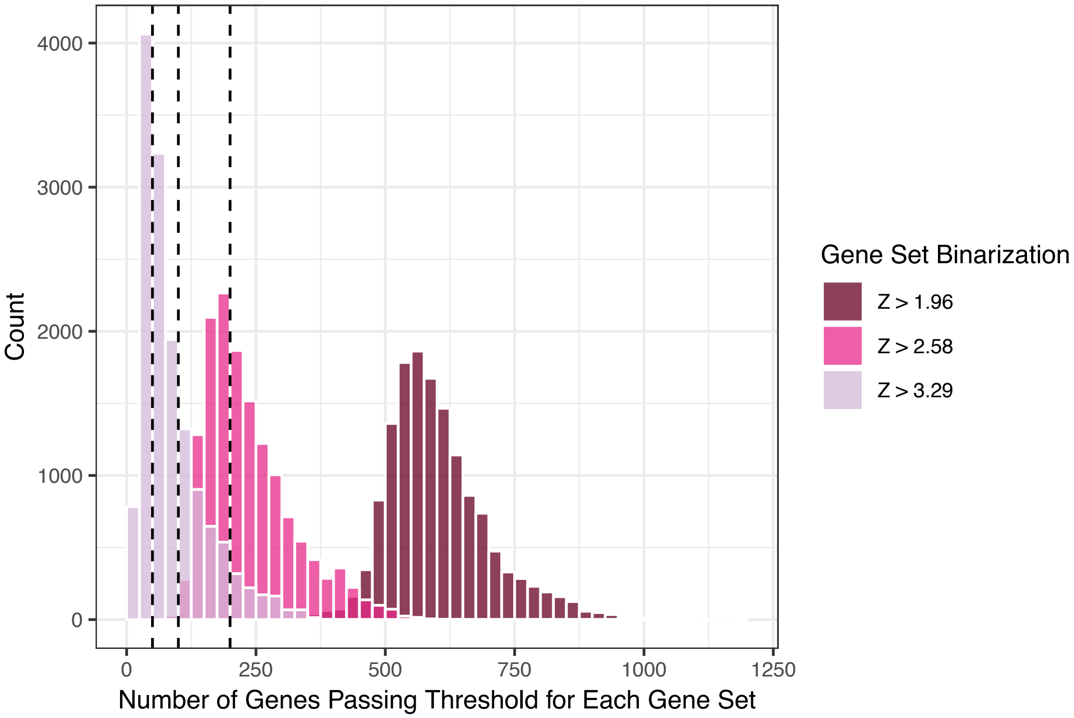
Chrom	Gene	P-Value
2	Gene 4	0.23
2	Gene 5	0.92
2	Gene 6	0.03
…		
22	Gene 19998	0.67
22	Gene 19999	3.0 x 10^{-4}
22	Gene 20000	0.21

Ranked gene set p-values

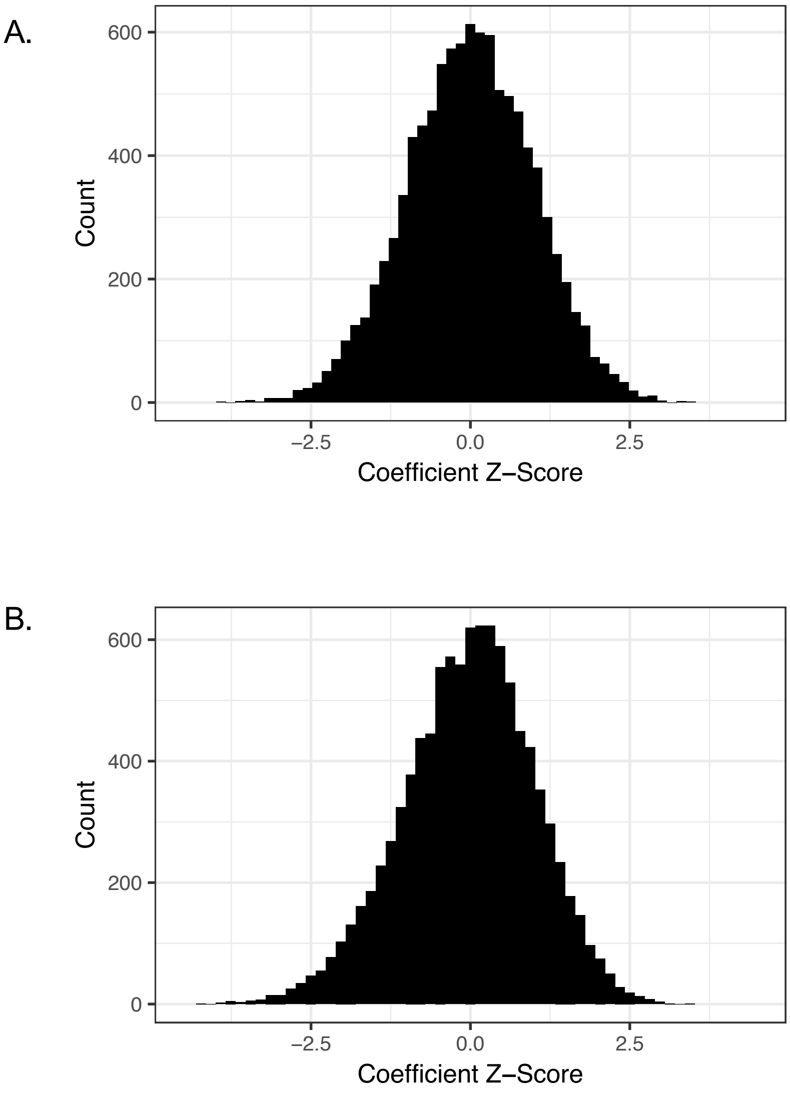
Gene Set	P-Value
GeneSet_1	8.6 x 10^{-5}
GeneSet_2	2.3 x 10^{-4}
GeneSet_3	0.00045
…	
GeneSet_14460	0.99
GeneSet_14461	0.99
GeneSet_14462	1

Gene Set	Gene Members
GeneSet_1	Gene 2, Gene 5231, Gene 321, Gene 530…
GeneSet_2	Gene 1093, Gene 923, Gene 1, Gene 1231…
GeneSet_3	Gene 2, Gene 9320, Gene 3, Gene 4320…
…	
GeneSet_14460	Gene 234, Gene 8234, Gene 2, Gene 23
GeneSet_14461	Gene 982, Gene 32, Gene 400, Gene 231
GeneSet_14462	Gene 10345, Gene 1, Gene 324, Gene 3234

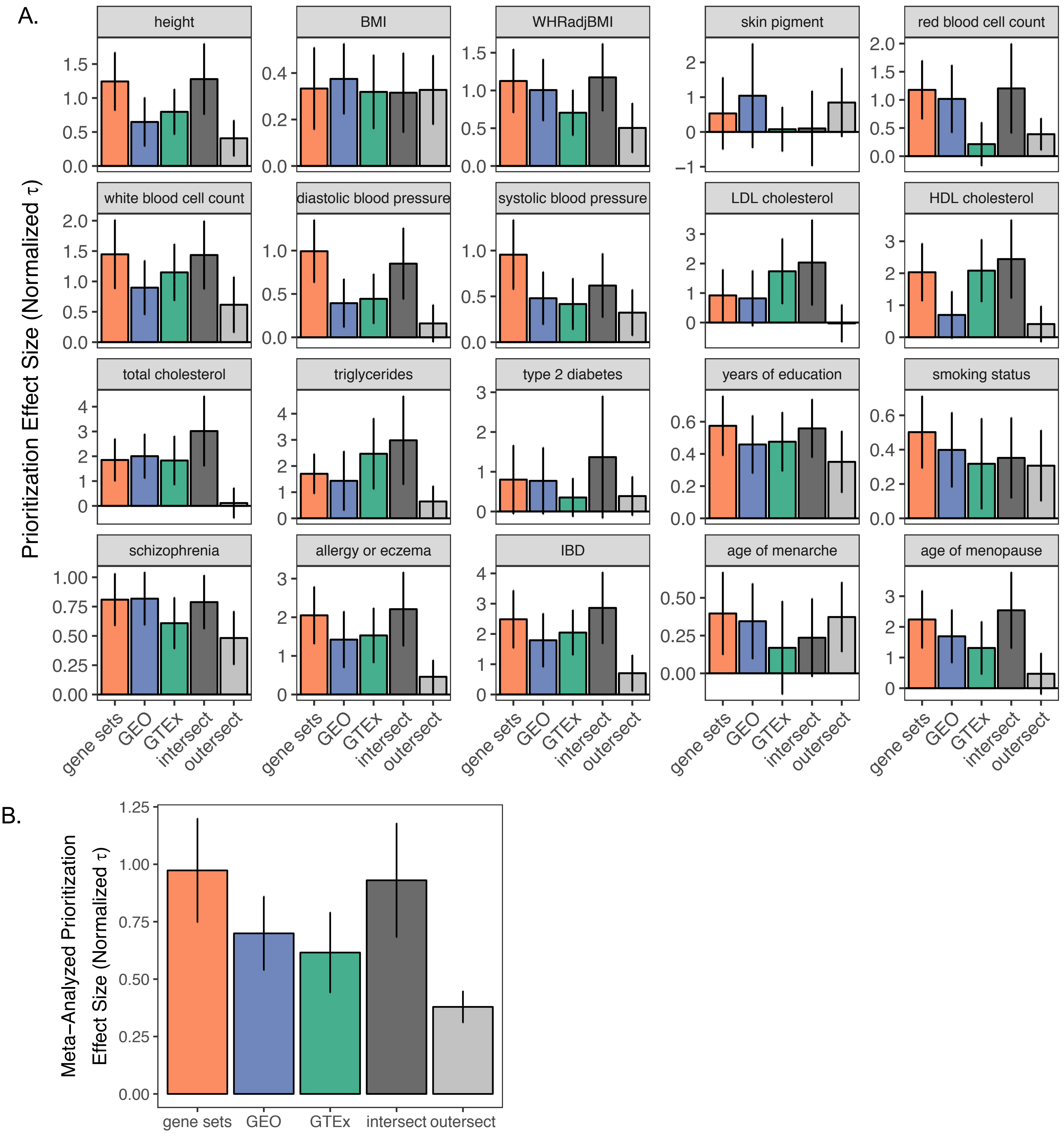
Supplemental Figure 2. Schematic of prioritization with MAGMA.



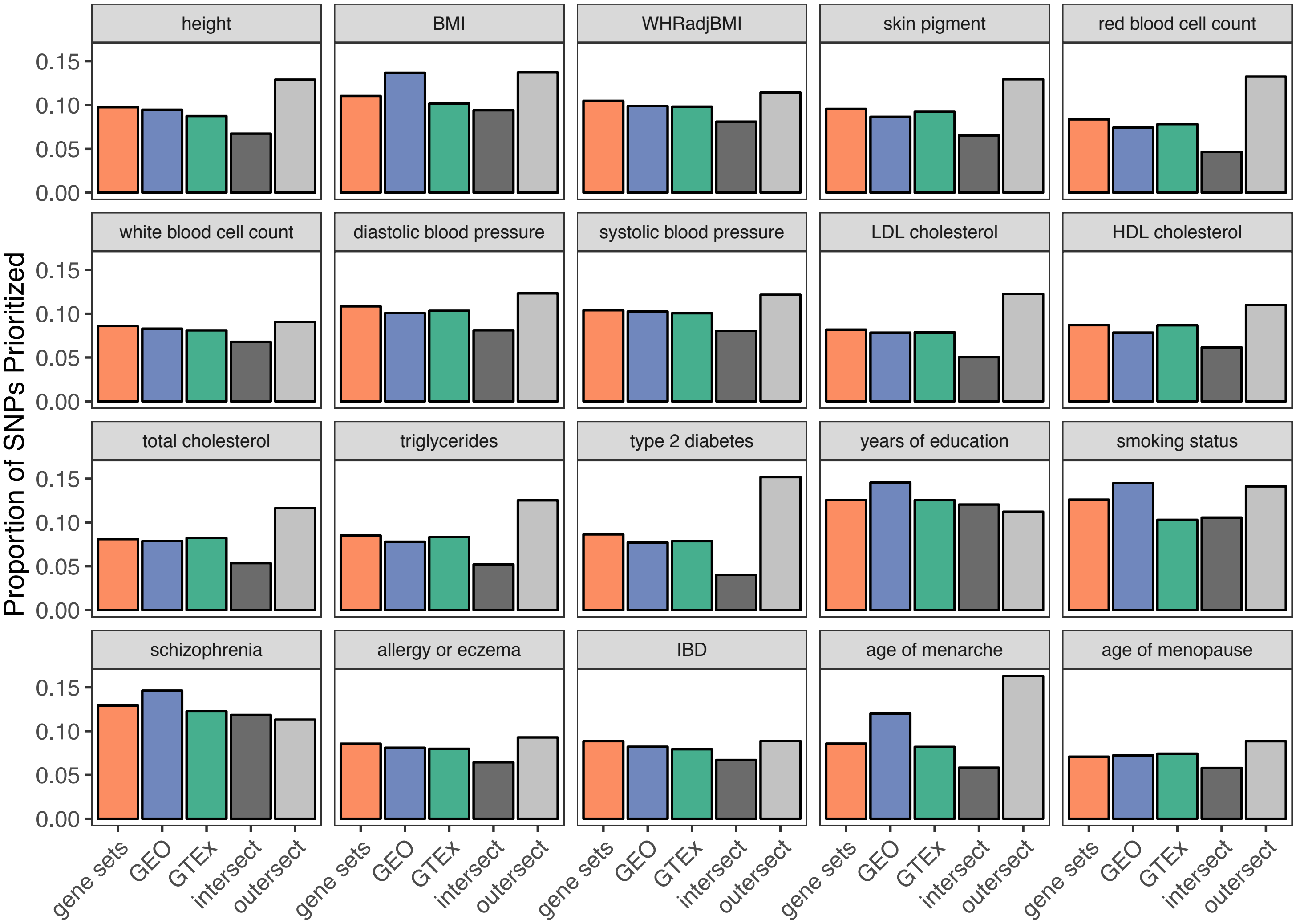
Supplemental Figure 3. Distribution of size of gene sets after binarization with different z-score cutoffs. Dashed vertical lines are at 50, 100, and 200, representing the cutoffs for the three rank-based binarization strategies.



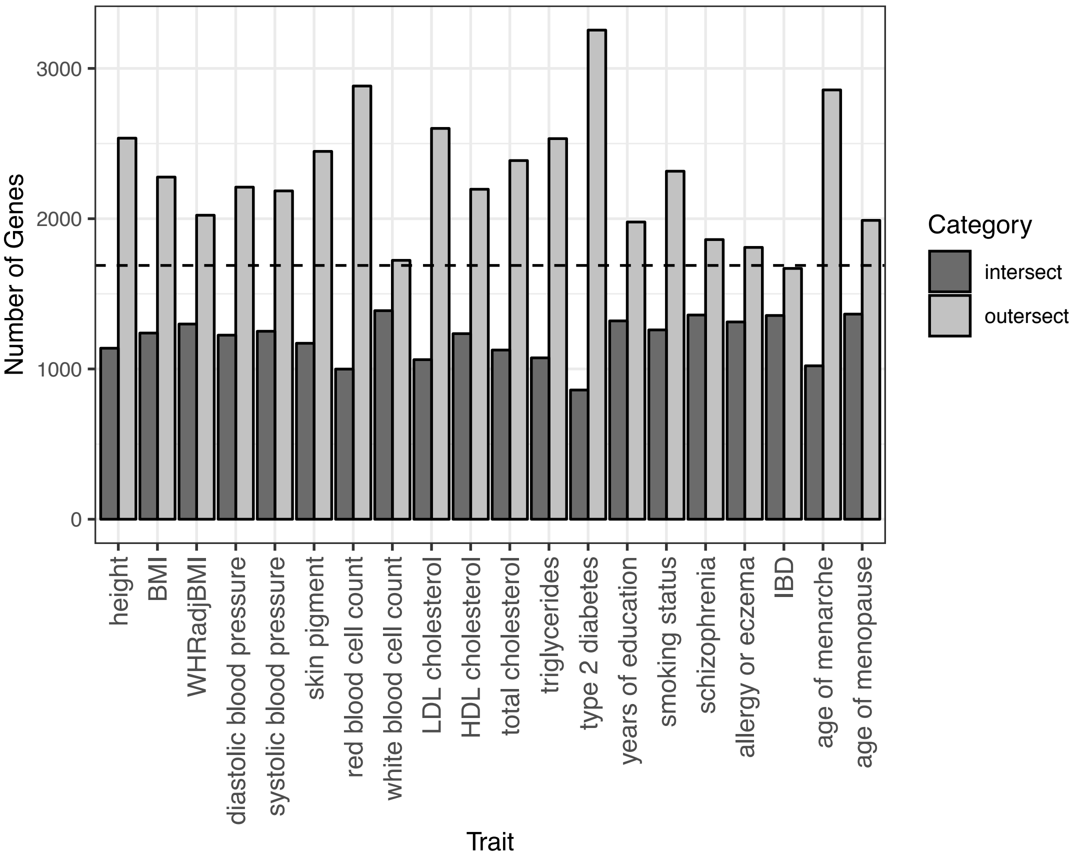
Supplemental Figure 4. Results from the Type 1 error analysis. Panel a) shows results from randomly prioritizing 10% of SNPs and panel b) shows results from randomly prioritizing 10% of genes. Histogram displays the z-score distribution for τ in both sets of simulations (each consisting of 1,000 null simulations, each tested on 10 real GWAS, for a total of 10,000 data points).



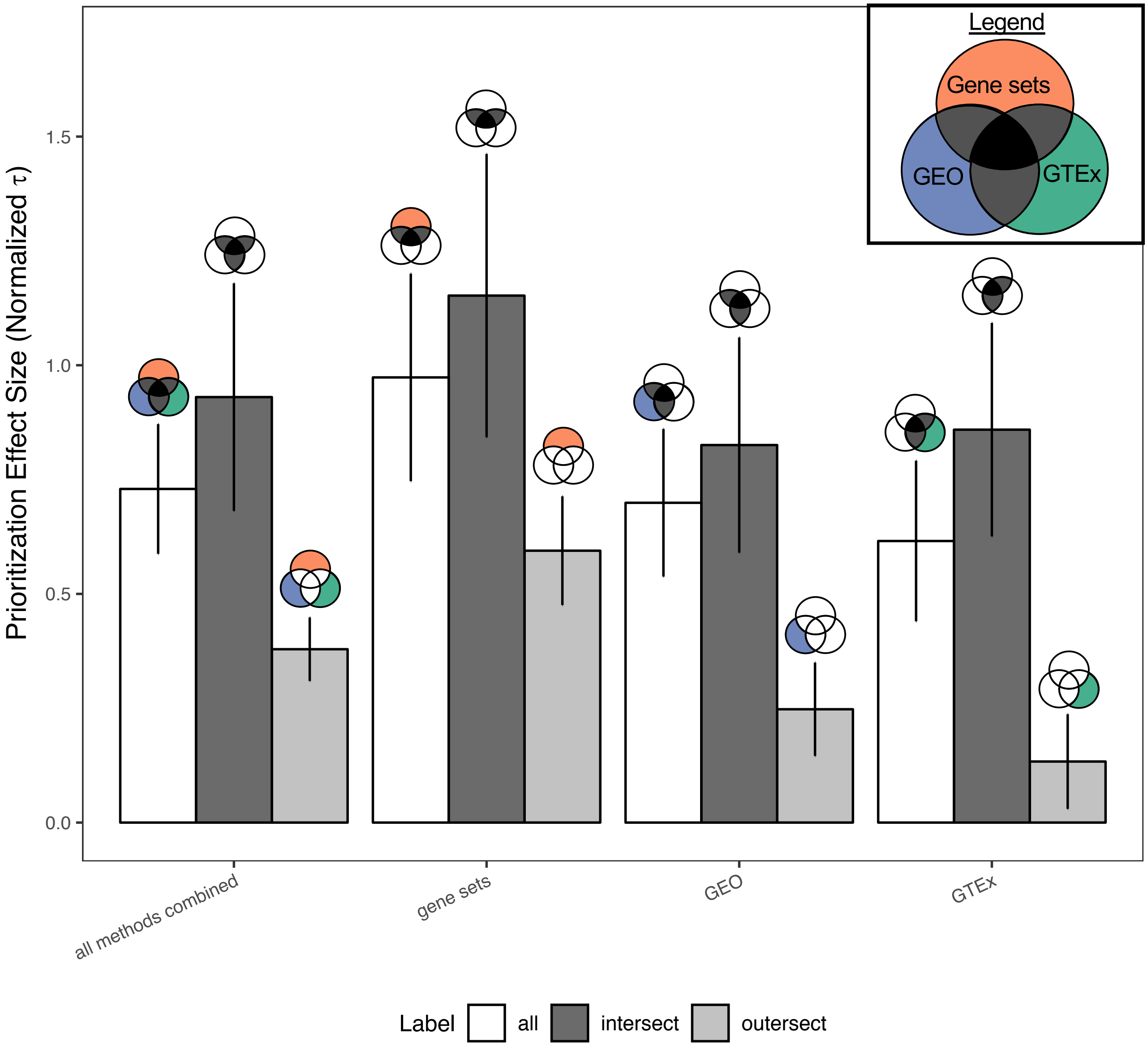
Supplemental Figure 5. Effect sizes (normalized τ) for separate LD score regression models comparing 1) DEPICT-gene-sets, 2) DEPICT-GEO, 3) DEPICT-GTEx, 4) the “intersect” (genes prioritized by at least two of the three preceding methods) and 5) the “outersect” (genes prioritized by only one method). (“Separate LD score regression models” indicates that these annotations are tested individually rather than jointly). Error bars represent 95% confidence intervals. Note that y-axis scales are different for each trait. a) Results from each trait. b) Results meta-analyzed across 16 traits.



Supplemental Figure 6. For each trait, proportion of SNPs prioritized by DEPICT-gene-sets, DEPICT-GEO, DEPICT-GTEx, the intersect, and the outersect.

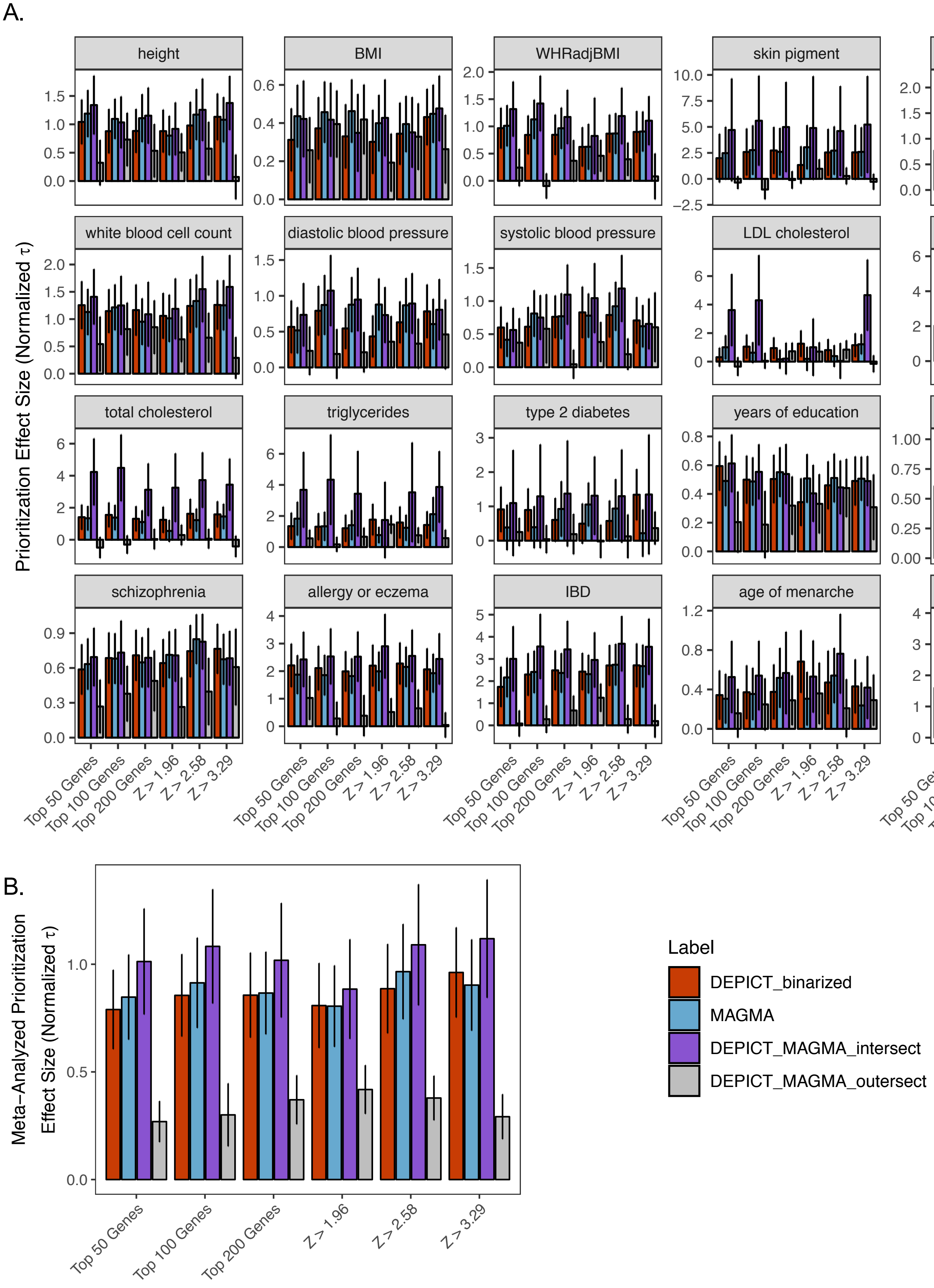


Supplemental Figure 7. For each trait, number of genes in the “intersect” and “outersect” conditions for DEPICT-gene-sets, DEPICT-GEO, and DEPICT-GTEx. Dashed line is at 10% of the genome, which represents the number of genes included in each individual DEPICT condition.

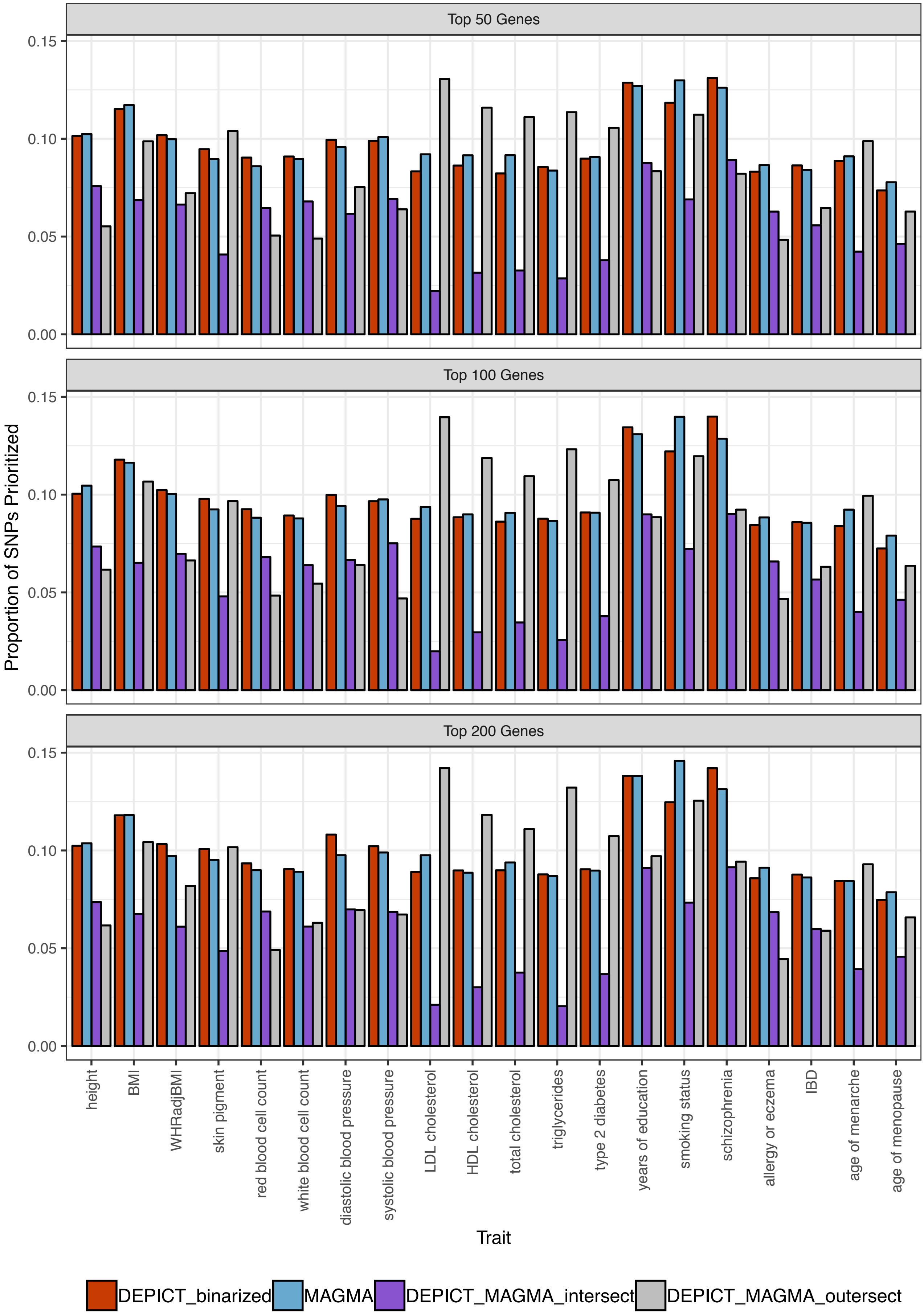


Supplemental Figure 8. Effect sizes (normalized τ) for separate LD score regression models in which each of DEPICT-gene-sets, DEPICT-GEO, and DEPICT-GTEx are split into individual intersects and outersects. For example, the DEPICT-gene-sets intersect consists of genes prioritized by DEPICT-gene-sets and at least one of the other two methods. The DEPICT-gene-sets outersect consists of genes prioritized by DEPICT-gene-sets only.

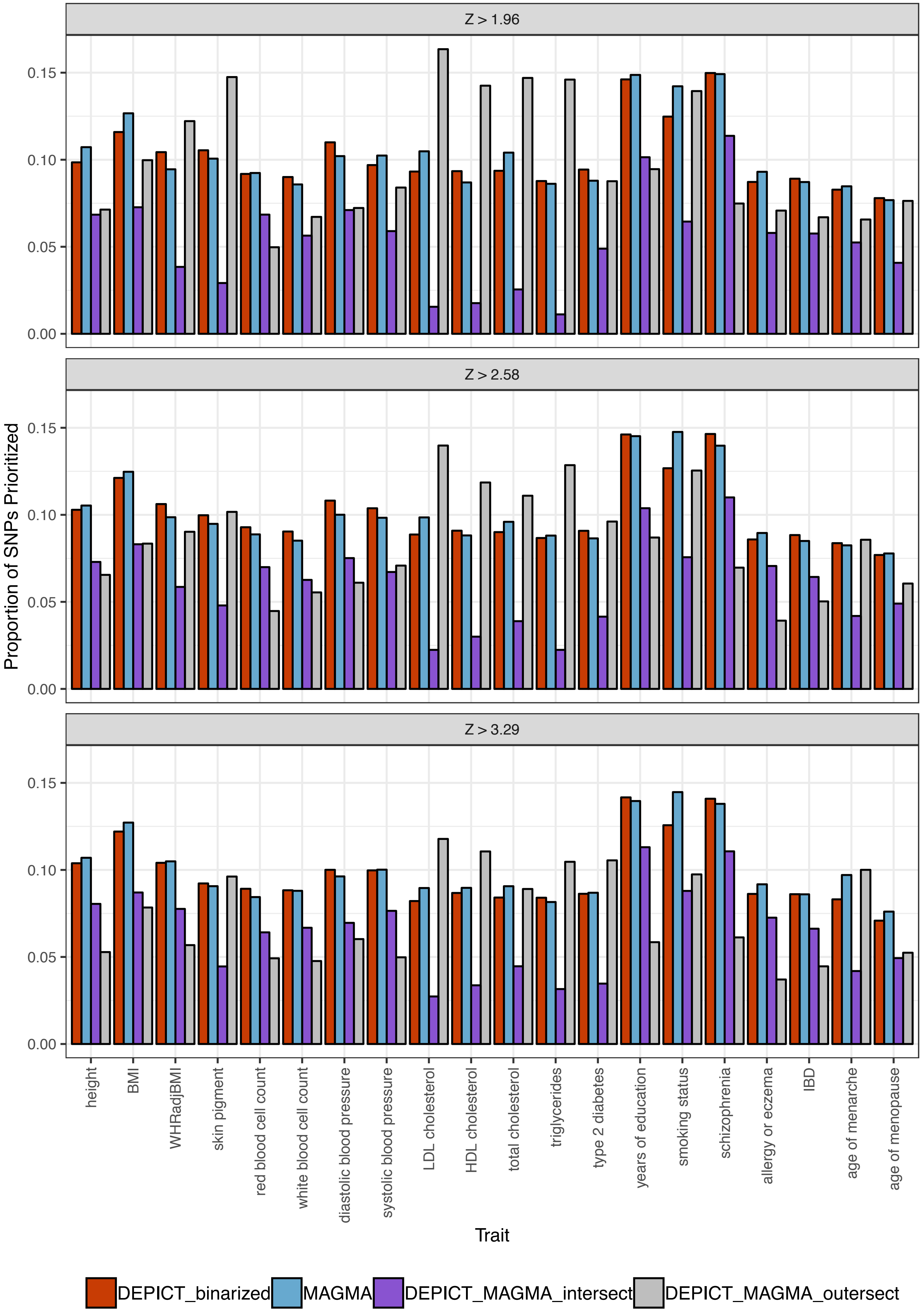
The “all methods combined” column represents the union of all genes prioritized by any method (and the black and grey bars for that column represent the original intersect and outersect data). Venn diagrams show which sets of genes are included for each analysis. Error bars represent 95% confidence intervals. Results meta-analyzed across 16 traits.



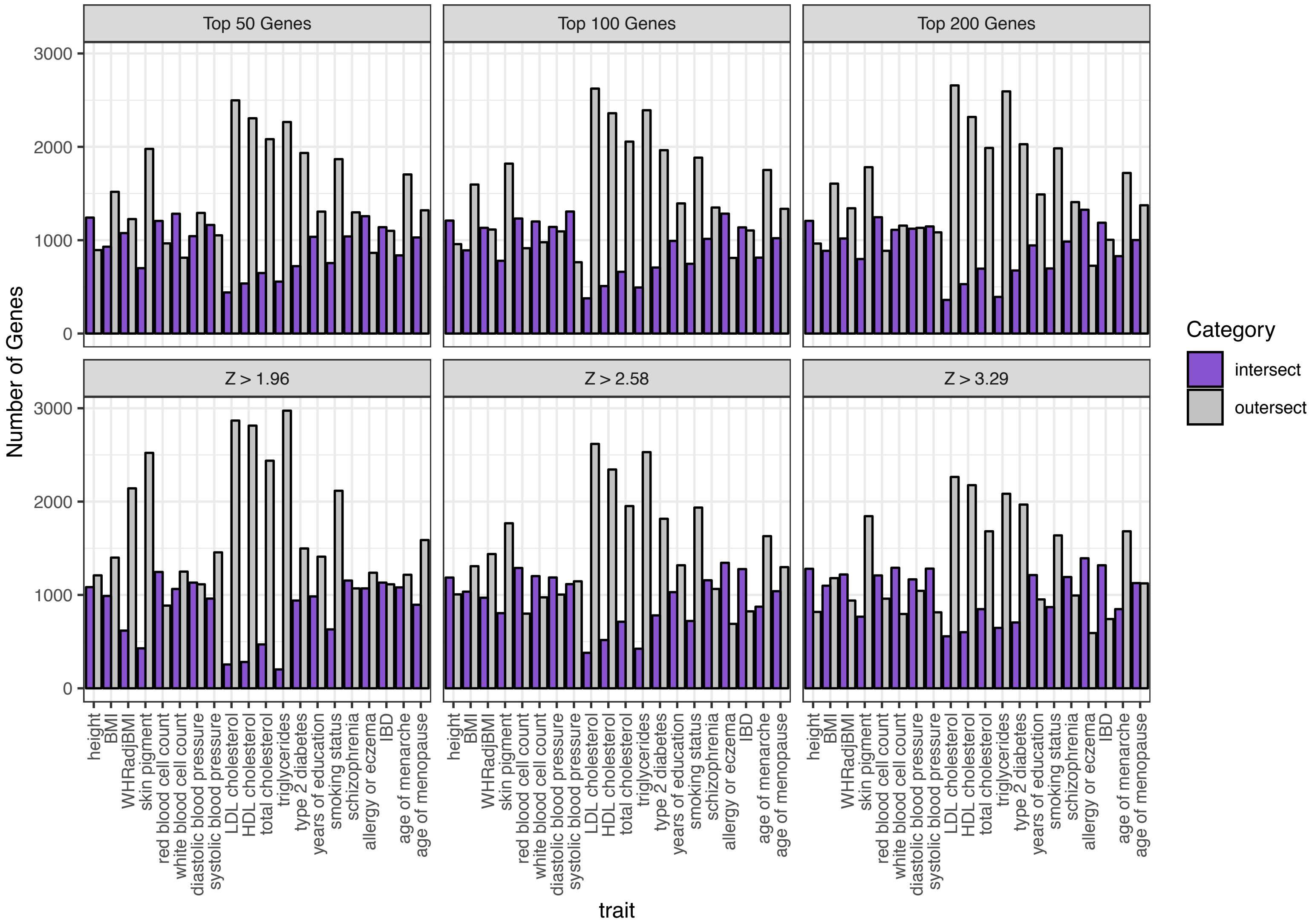
Supplemental Figure 9. Effect sizes (normalized τ) for separate LD score regression models comparing 1) DEPICT with binarized gene sets, 2) MAGMA, 3) the "intersect" (genes prioritized by both DEPICT and MAGMA) and 5) the "outersect" (genes prioritized by only DEPICT or only MAGMA). ("Separate LD score regression models" indicates that these annotations are tested individually rather than jointly). Error bars represent 95% confidence intervals. Note that y-axis scales are different for each trait. a) Results from each trait. b) Results meta-analyzed across 16 traits.



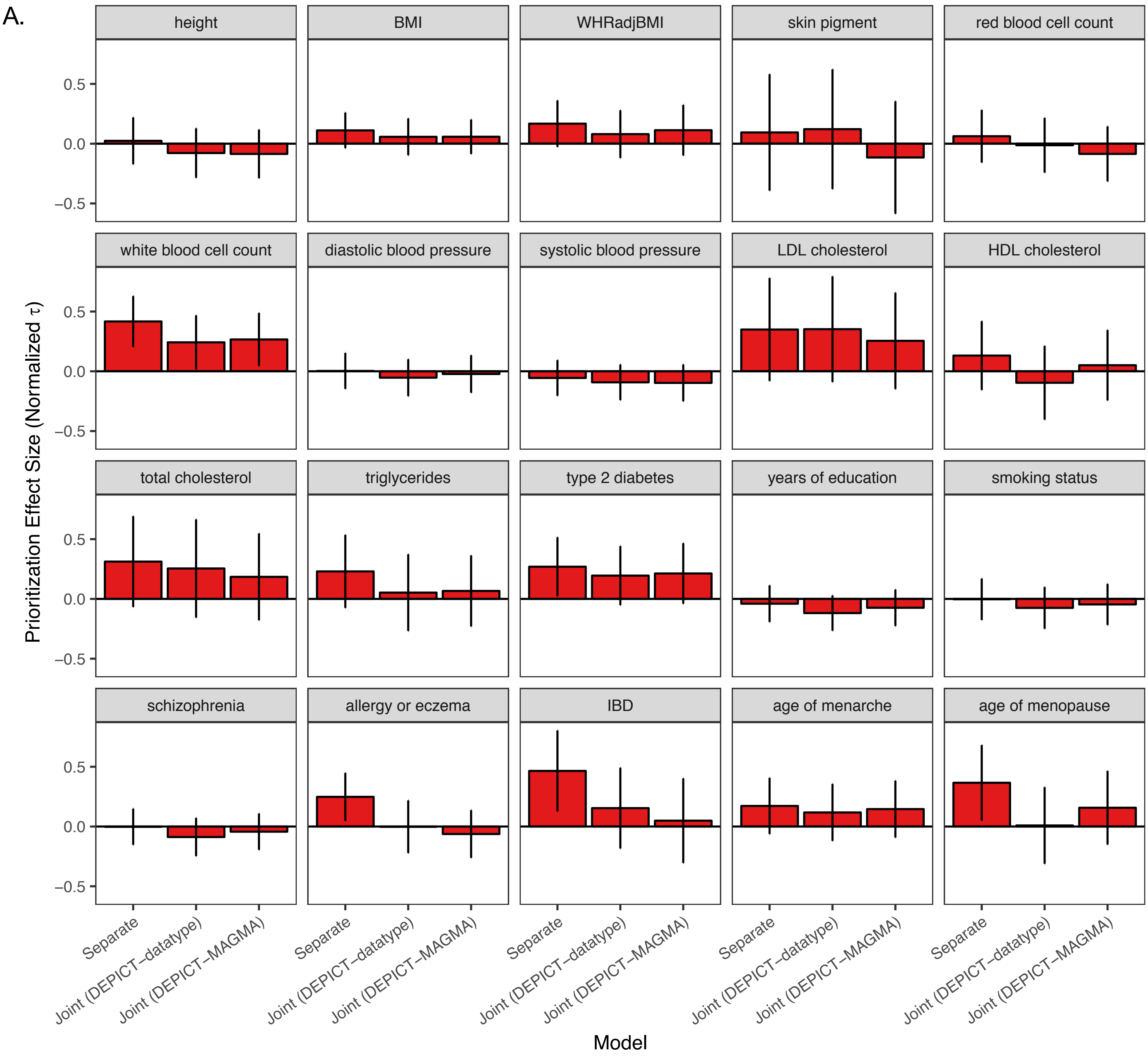
Supplemental Figure 10. For each trait, proportion of SNPs prioritized by 1) DEPICT with binarized gene sets, 2) MAGMA, 3) the "intersect", and 4) the "outersect." Figure displays results for the three rank-based binarizations.

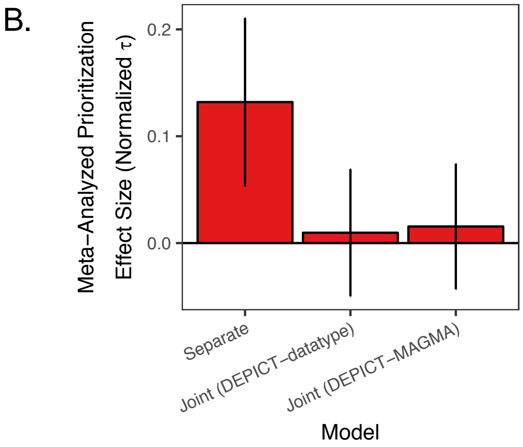


Supplemental Figure 11. For each trait, proportion of SNPs prioritized by 1) DEPICT with binarized gene sets, 2) MAGMA, 3) the "intersect", and 4) the "outersect." Figure displays results for the three z-score-based binarizations.

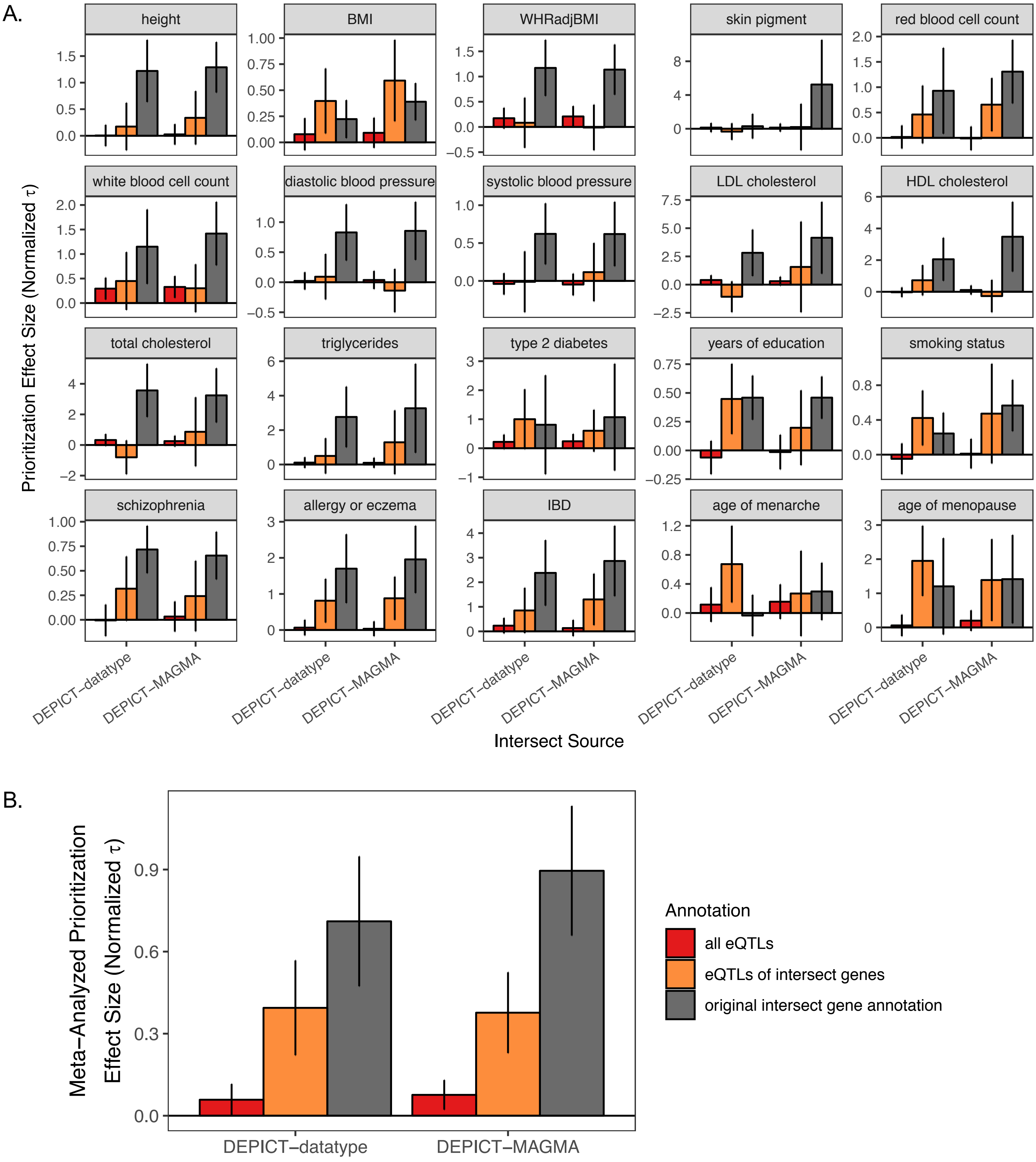


Supplemental Figure 12. For each trait, number of genes in the “intersect” and “outersect” conditions for binarized DEPICT and MAGMA. Each version of the binarized gene sets is shown separately.

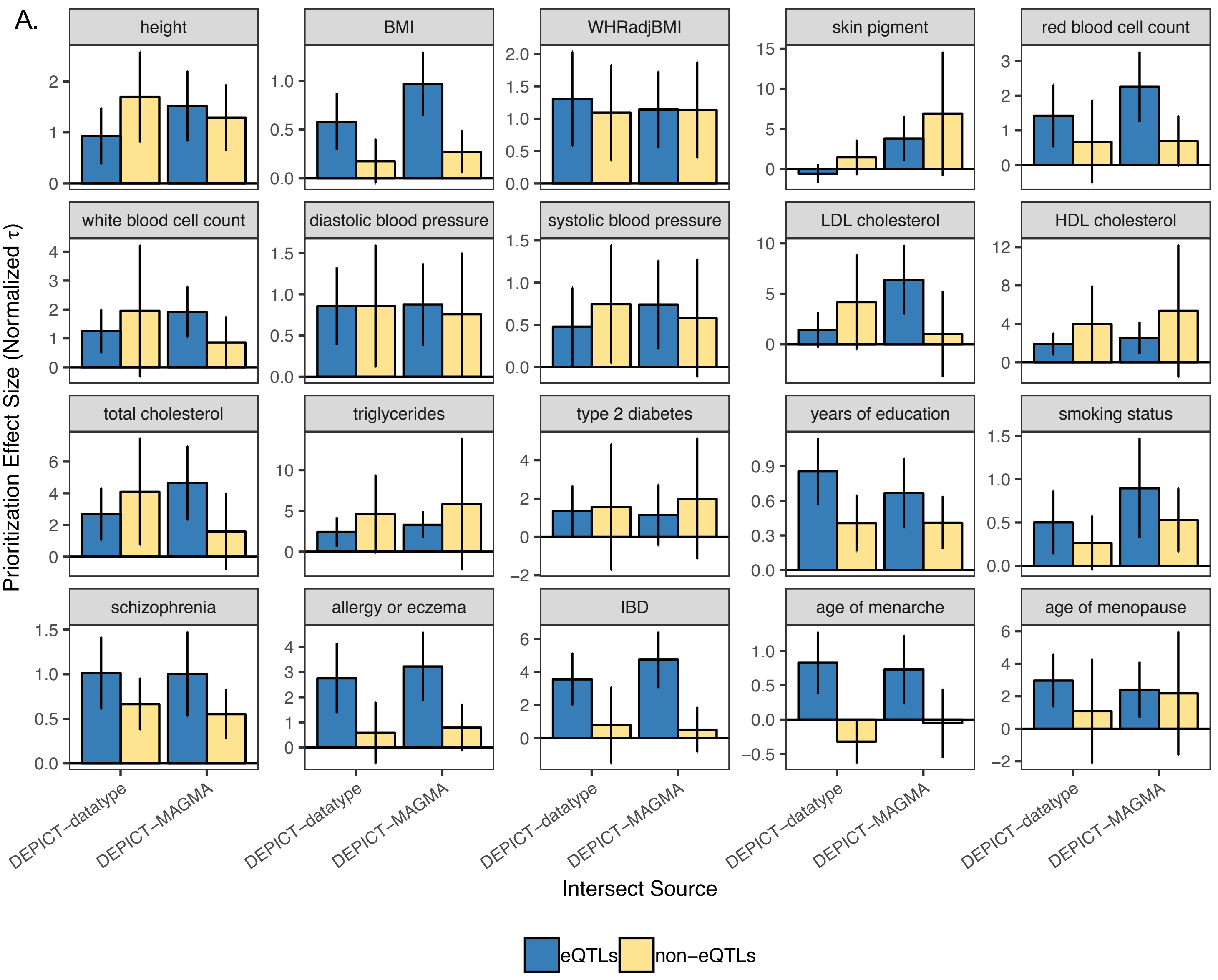


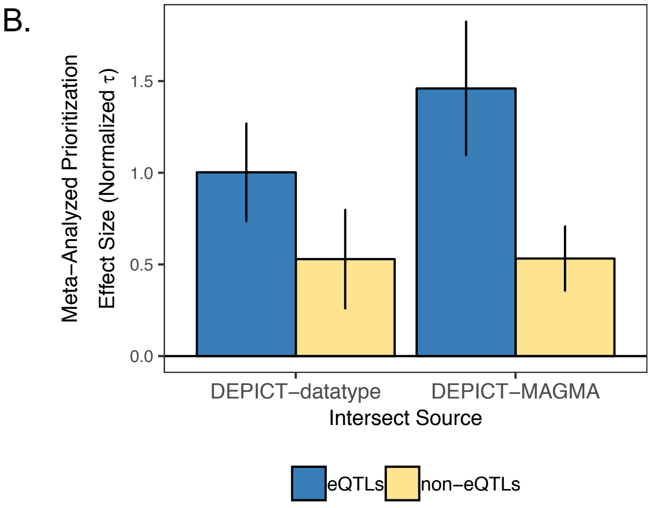


Supplemental Figure 13. Effect sizes (normalized τ) for an annotation consisting of all blood eQTLs regulating at least one gene in our data set. In the "Separate" model, the all-eQTL annotation was modeled with the baseline model only. In the "Joint (DEPICT-datatype)" and "Joint (DEPICT-MAGMA)" models, the all-eQTL annotation was jointly modeled with all eQTLs regulating prioritized genes from either the DEPICT-datatype intersect set or DEPICT-MAGMA intersect set (where the DEPICT-MAGMA intersect set was based on the Z > 3.29 binarization). Error bars represent 95% confidence intervals. a) Results for all traits. b) Results meta-analyzed across 16 traits.

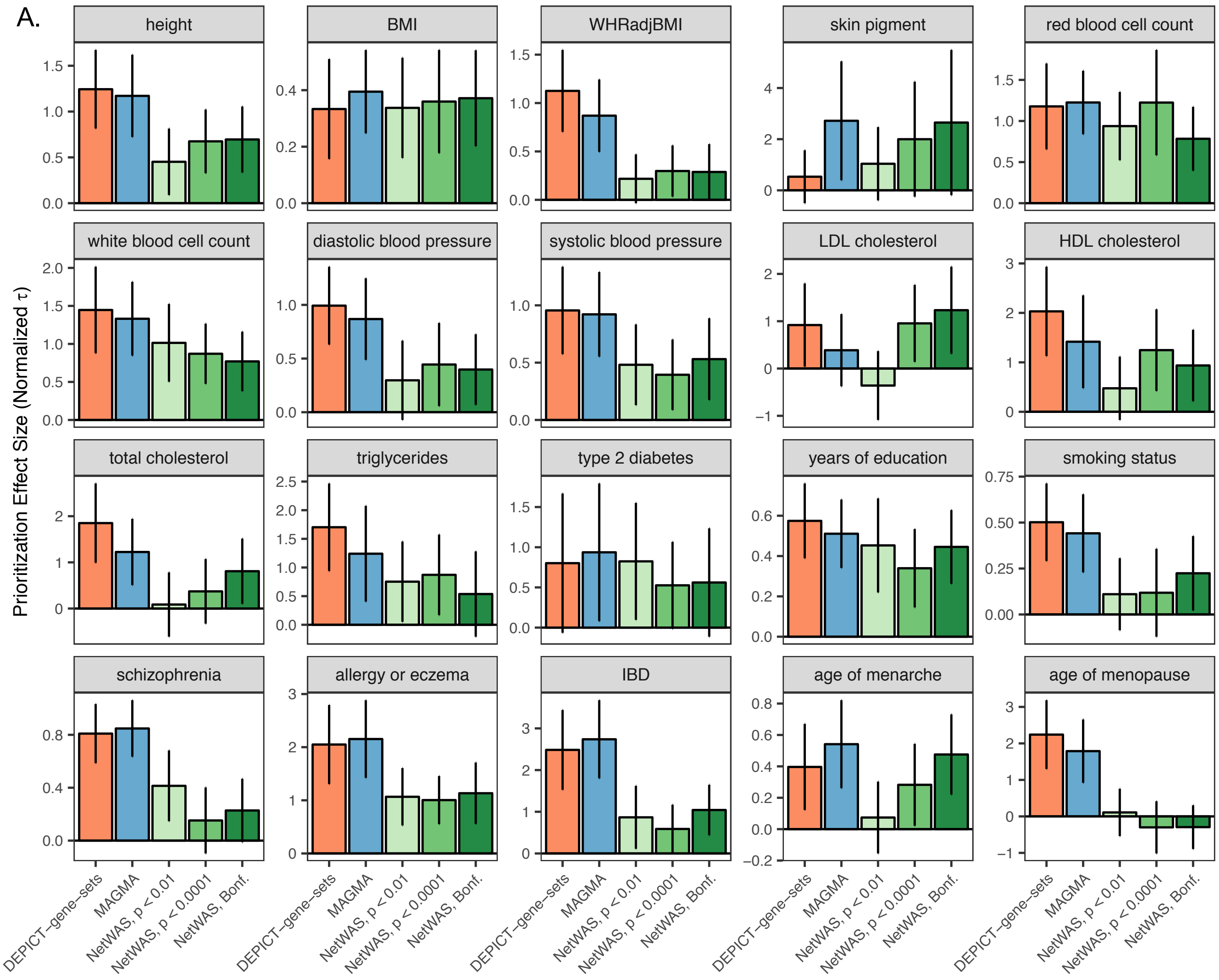


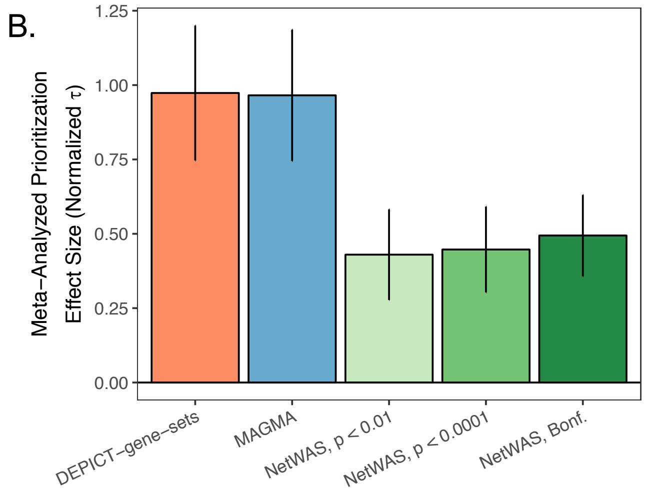
Supplemental Figure 14. Effect sizes (normalized τ) for a joint S-LDSC model including 1) all eQTLs, 2) the original gene intersect prioritization annotation (i.e. all SNPs in the gene body and a 50-kb window), and 3) all eQTLs regulating the prioritized intersect genes. Error bars represent 95% confidence intervals. Results are shown for two intersect sets: DEPICT-datatype and DEPICT-MAGMA (the latter based on the Z > 3.29 binarization). a) Results for all traits. b) Results meta-analyzed across 16 traits.



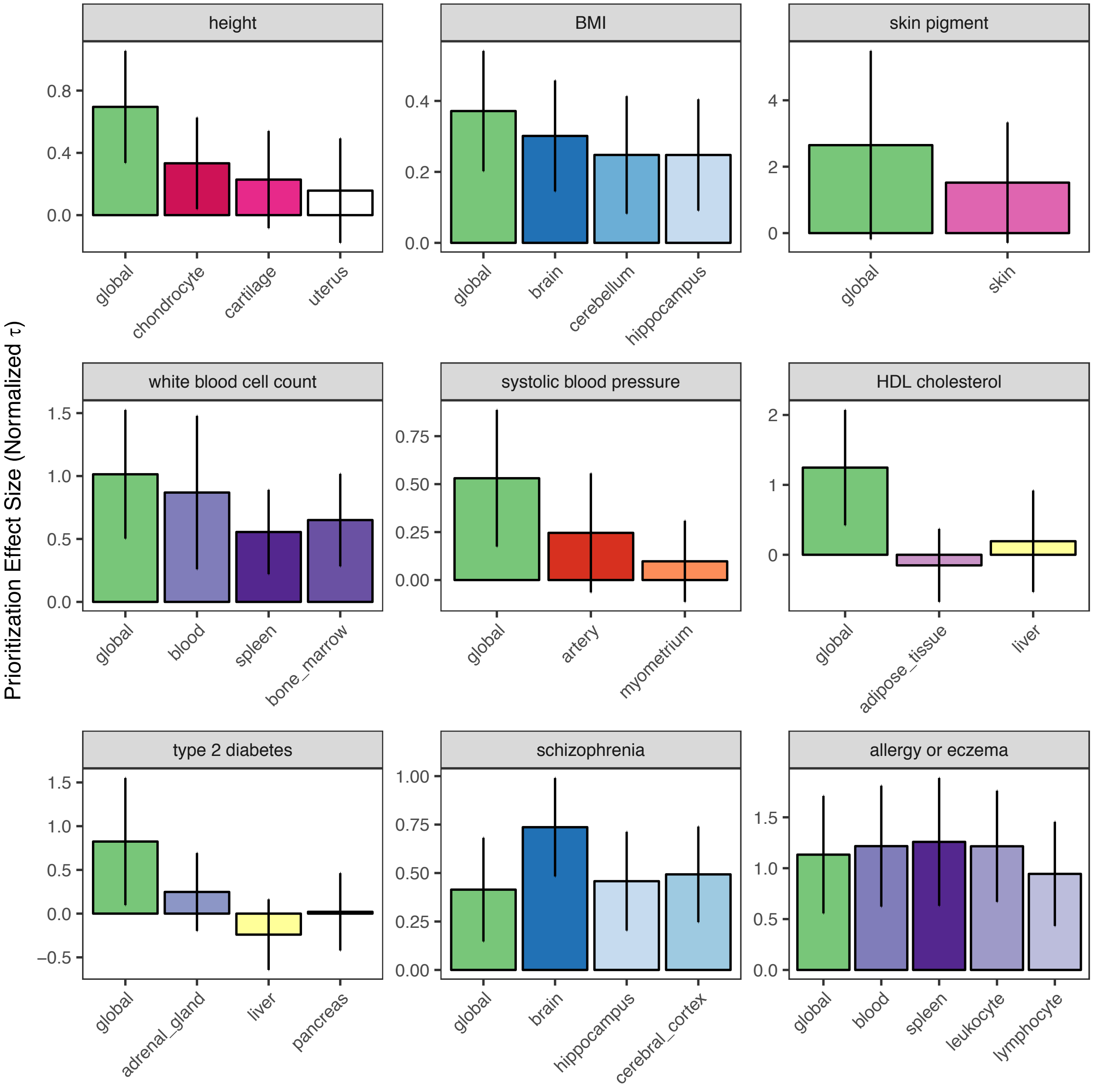


Supplemental Figure 15. Effect sizes (normalized τ) for an analysis in which we divided our intersect annotations into eQTLs and non-eQTLs and modeled them jointly for each trait. Error bars represent 95% confidence intervals. a) Results for all traits. b) Results meta-analyzed across 16 traits.





Supplemental Figure 16. Effect sizes (normalized τ) for separate LD score regression models comparing 1) DEPICT-gene-sets, 2) MAGMA (based on the Z > 2.58 binarization), 3) NetWAS using the global tissue network with three p-value thresholds: p < 0.01, p < 0.0001, and a Bonferroni correction for the number of genes tested. ("Separate LD score regression models" indicates that these annotations are tested individually rather than jointly). Error bars represent 95% confidence intervals. Note that y-axis scales are different for each trait. a) Results from each trait. b) Results meta-analyzed across 16 traits.



Supplemental Figure 17. Effect sizes (normalized τ) for separate LD score regression models comparing NetWAS performance for the global tissue network and several trait-relevant tissues for nine of our GWAS. For each trait, we used the p-value threshold with the best performance from the global network. ("Separate LD score regression models" indicates that these annotations are tested individually rather than jointly). Error bars represent 95% confidence intervals.