Localizing Components of Shared Transethnic Genetic Architecture of Complex Traits from GWAS Summary Data

Huwenbo Shi,[1,2,3,11,*] Kathryn S. Burch,[1,11,*] Ruth Johnson,[4] Malika K. Freund,[5] Gleb Kichaev,[1] Nicholas Mancuso,[6] Astrid M. Manuel,[7] Natalie Dong,[8] and Bogdan Pasaniuc[1,5,9,10]

Despite strong transethnic genetic correlations reported in the literature for many complex traits, the non-transferability of polygenic risk scores across populations suggests the presence of population-specific components of genetic architecture. We propose an approach that models GWAS summary data for one trait in two populations to estimate genome-wide proportions of population-specific/shared causal SNPs. In simulations across various genetic architectures, we show that our approach yields approximately unbiased estimates with in-sample LD and slight upward-bias with out-of-sample LD. We analyze nine complex traits in individuals of East Asian and European ancestry, restricting to common SNPs (MAF $> 5\%$), and find that most common causal SNPs are shared by both populations. Using the genome-wide estimates as priors in an empirical Bayes framework, we perform fine-mapping and observe that high-posterior SNPs (for both the population-specific and shared causal configurations) have highly correlated effects in East Asians and Europeans. In population-specific GWAS risk regions, we observe a 2.8× enrichment of shared high-posterior SNPs, suggesting that population-specific GWAS risk regions harbor shared causal SNPs that are undetected in the other GWASs due to differences in LD, allele frequencies, and/or sample size. Finally, we report enrichments of shared high-posterior SNPs in 53 tissue-specific functional categories and find evidence that SNP-heritability enrichments are driven largely by many low-effect common SNPs.

Introduction

Genetic and phenotypic variations among humans have been shaped by many factors, including migration histories, geodemographic events, and environmental background.[1–5] As a result, the underlying genetic architecture of a given complex trait—defined here in terms of “polygenicity” (the number of variants with nonzero effects)[6–10] and the coupling of causal effect sizes with minor allele frequency (MAF),[11,12] linkage disequilibrium (LD),[13–15] and other genomic features[16]—varies among ancestral populations. While the vast majority of genome-wide association studies (GWASs) to date have been performed in individuals of European descent,[17–20] growing numbers of studies performed in individuals of non-European ancestry[21–27] have created opportunities for well-powered transethnic genetic studies.[21,22,24,26,28–33]

Risk regions identified through GWASs tend to replicate across populations,[17,21,22,33–35] indicating that complex traits have genetic components that are shared among populations. Indeed, for certain post-GWAS analyses such as disease mapping[23,31,36] and statistical fine-mapping,[28,37–40] under the assumption that two populations share one or more causal variants, population-specific LD patterns can be leveraged to improve performance over approaches that model a single population. On the other hand, several studies have shown that heterogeneity in genetic architectures limits transferability of polygenic risk scores (PRSs) across populations;[5,41–48] critically, if applied in a clinical setting, existing PRSs may exacerbate health disparities among ethnic groups.[49] The population specificity of existing PRSs as well as estimates of transethnic genetic correlations less than one reported in the literature[30,50–53] indicate that (1) LD tagging and allele frequencies of shared causal variants vary across populations, (2) that a sizeable number of causal variants are population specific, and/or (3) that causal effect sizes vary across populations due to, for example, different gene-environment interactions. In a region with population-specific LD, a single genetic variant that is significantly associated with a trait in two populations may actually be tagging distinct population-specific causal variants (Figure 1). Conversely, two distinct associations in two populations may be driven by the same underlying causal variants (i.e., colocalization). Thus, identifying shared and population-specific components of genetic architecture could help improve transethnic analyses (e.g., transferability of PRSs across populations[19,41,42,45,46]) and uncover novel disease etiologies.

In this work, we introduce PESCA (population-specific/shared causal variants), an approach that requires only

[1]Bioinformatics Interdepartmental Program, University of California, Los Angeles, Los Angeles, CA 90095, USA; [2]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA; [3]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA; [4]Department of Computer Science, University of California, Los Angeles, Los Angeles, CA 90095, USA; [5]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA; [6]Center for Genetic Epidemiology, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA; [7]Department of Biological Sciences, Florida International University, Miami, FL 33199, USA; [8]Department of Biomedical Engineering, Boston University, Boston, MA 02215, USA; [9]Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA; [10]Department of Computational Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA

[11]These authors contributed equally to this work

*Correspondence: hshi@hsph.harvard.edu (H.S.), kathrynburch@ucla.edu (K.S.B.)

https://doi.org/10.1016/j.ajhg.2020.04.012.

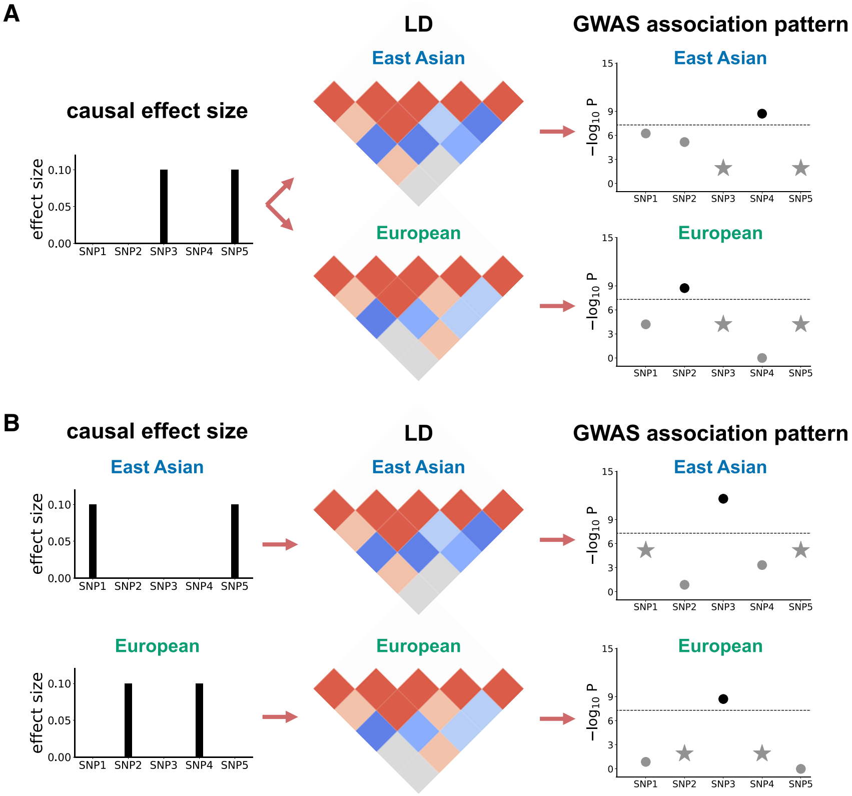


Figure 1. Toy Examples to Illustrate How Population-Specific LD Patterns Affect GWAS Associations

(A) SNPs 3 and 5 are causal in both East Asians and Europeans and have the same causal effect size of 0.1. However, due to different LD patterns in East Asians and Europeans, SNPs 2 and 4 are observed to be GWAS significant, respectively.

(B) Different SNPs are causal in East Asians (SNPs 1 and 5) and Europeans (SNPs 2 and 4). However, due to population-specific LD, SNP 3 is observed to be GWAS significant in both populations. The stars in the rightmost plots represent the SNPs with true nonzero effects; the GWAS-significant SNP is highlighted in a darker color.

GWAS summary association statistics and ancestry-matched estimates of LD to infer genome-wide proportions of population-specific and shared causal variants for a single trait in two populations. These estimates are then used as priors in an empirical Bayes framework to localize and test for enrichment of population-specific/shared causal variants in regions of interest. In this context, a "causal variant" is a variant measured in the given GWAS that either has a nonzero effect on the trait (e.g., a nonsynonymous variant that alters protein folding) or tags a nonzero effect at an unmeasured variant through LD. It is therefore important to note that the set of "causal variants" that PESCA aims to identify is defined with respect to the set of variants included in the GWAS and can contain variants with indirect nonzero effects that are statistical rather than biological in nature (this is analogous to the definition of SNP-heritability, which is also a function of a specific set of SNPs[11,54–56]). We also note that the definition of enrichment used here is related to, but conceptually distinct from, definitions of SNP-heritability enrichment.[13,16] Under our framework, an enrichment of causal SNPs greater than 1 indicates that, compared to the genome-wide background, there are more causal SNPs in that region than expected[57,58] (Material and Methods). In contrast, an enrichment of SNP-heritability greater than 1 indicates that the average per-SNP effect size in the region is larger than the genome-wide average per-SNP effect size.

Through extensive simulations, we show that our method yields approximately unbiased estimates of the proportions of population-specific/shared causal variants if in-sample LD is used and slightly upward-biased estimates if LD is estimated from an external reference panel. We then show that using these estimates as priors to perform fine-mapping (Material and Methods) produces well-calibrated per-SNP posterior probabilities and enrichment test statistics. We apply our approach to publicly available GWAS summary statistics for nine complex traits and diseases in individuals of East Asian (EAS) and European (EUR) ancestry (average $N_{EAS} = 94{,}621$, $N_{EUR} = 103{,}507$) (Table 1), restricting to common SNPs (MAF > 5%) and using 1000 Genomes[59,60] to estimate ancestry-matched LD. On average across the nine traits, we estimate that approximately 80% (SD 15%) of common SNPs that are causal in EAS and 84% (SD 8%) of those in EUR are shared by the other population. Consistent with previous studies based on SNP-heritability,[55,61] we find that high-posterior SNPs are distributed uniformly across the genome. We observe that population-specific GWAS risk regions have, on average across the 9 traits, a 2.8× enrichment of shared high-posterior SNPs relative to the genome-wide background, suggesting that many EAS-specific and EUR-specific GWAS risk regions harbor shared causal SNPs that are undetected in the other population due to differences in LD, allele frequencies, and/or GWAS sample size. The effect sizes of SNPs with posterior probability > 0.8 of being causal (for any causal configuration) are highly correlated between EAS and EUR, concordant with replication slopes between EAS and EUR marginal effects close to 1 that have been reported for several complex diseases[33] and with strong transethnic genetic correlations previously reported for the same traits analyzed in this work (average $\hat{\rho}_g = 0.79 \pm 0.07$ SEM across the 9 traits).[51] Finally, we show that regions flanking genes that are specifically expressed in trait-relevant tissues[62] harbor a disproportionate number of shared high-posterior SNPs. Many of the same tissue-specific gene sets are also enriched with SNP-heritability, implying that SNP-heritability enrichments are

Table 1. Estimated Numbers and Percentages of Population-Specific/Shared Common Causal SNPs for Nine Complex Traits

Trait Name (abbrev.)	Pop.	Ref.	$\hat{h}_g^2$ (SE) %	Sample Size (n)	Total # SNPs (MAF $>$ 5%)	EAS-Specific Causals (SE)	EUR-Specific Causals (SE)	Shared Causals (SE)	$\hat{\rho}_g$ (SE)[51]
Body mass index (BMI)	EAS	22	19.8 (0.6)	224,698	258,130	982 (2); 0.4%	1,033 (2); 0.4%	25,641 (16); 10%	0.80 (0.02)
	EUR	63	20.6 (0.9)	158,284					
Mean corpuscular hemoglobin (MCH)	EAS	21	18.6 (2.2)	108,054	480,684	1,165 (6); 0.2%	728 (3); 0.2%	3,082 (4); 0.6%	0.88 (0.05)
	EUR	64	22.7 (3.2)	172,332					
Mean corpuscular volume (MCV)	EAS	21	21.0 (2.1)	108,256	480,678	1,004 (4); 0.2%	737 (5); 0.2%	3,256 (8); 0.7%	0.89 (0.05)
	EUR	64	23.6 (3.1)	172,433					
High-density lipoprotein (HDL)	EAS	21	20.7 (3.0)	70,657	268,198	3,167 (12); 1%	652 (2); 0.2%	4,789 (9); 2%	0.89 (0.06)
	EUR	65	16.4 (2.2)	89,614					
Low-density lipoprotein (LDL)	EAS	21	9.5 (1.3)	72,866	268,201	969 (5); 0.4%	742 (2); 0.3%	3,129 (6); 1%	0.66 (0.11)
	EUR	65	13.6 (1.9)	85,491					
Total cholesterol (TC)	EAS	21	8.1 (0.8)	128,305	268,197	1,892 (3); 0.7%	1,493 (5); 0.6%	5,058 (12); 2%	0.91 (0.07)
	EUR	65	22.5 (2.1)	89,865					
Triglyceride (TG)	EAS	21	13.5 (3.3)	105,597	268,198	2,245 (3); 0.8%	511 (4); 0.2%	3,432 (7); 1%	0.93 (0.07)
	EUR	65	13.6 (2.2)	86,502					
Major depressive disorder (MDD)	EAS	66	35.6 (3.4)	10,640	389,593	88 (4); 0.02%	3,280 (6); 0.8%	7,830 (6); 2%	0.34 (0.07)
	EUR	67	19.0 (1.8)	18,759					
Rheumatoid arthritis (RA)	EAS	36	28.9 (18.3)	22,515	526,206	3 (0.3); 6e−04%	124 (2); 0.02%	1,080 (6); 0.2%	0.87 (0.10)
	EUR	36	9.5 (1.9)	58,284					

We estimated genome-wide SNP-heritability using LD score regression[54] with the intercept constrained to 1 (i.e., assuming no population stratification). Trans-ethnic genetic correlation estimates ($\hat{\rho}_g$) computed from a similar set of summary statistics were obtained from a previous study.[51] Standard errors of the estimated numbers of population-specific/shared causal SNPs were computed using the last 50 iterations of the EM-MCMC algorithm.

driven by many low-effect SNPs rather than a small number of high-effect SNPs. Our results suggest that common causal SNPs have similar etiological roles in EAS and EUR and that transferability of PRS and other GWAS findings across populations can be improved by explicitly correcting for population-specific LD and allele frequencies.

Material and Methods

Distribution of GWAS Summary Statistics in Two Populations

For a given complex trait, we model the causal statuses of SNP *i* in two populations as a binary vector of size two, $\boldsymbol{C}_i = c_{i1}c_{i2}$, where each bit, $c_{i1} \in \{0,1\}$ and $c_{i2} \in \{0,1\}$, represents the causal status of SNP *i* in populations 1 and 2, respectively. $\boldsymbol{C}_i = 00$ indicates that SNP *i* is not causal in either population; $\boldsymbol{C}_i = 01$ and $\boldsymbol{C}_i = 10$ indicate that SNP *i* is causal only in the first and second population, respectively; and $\boldsymbol{C}_i = 11$ indicates that SNP *i* is causal in both populations. We assume $\boldsymbol{C}_i$ follows a multivariate Bernoulli (MVB) distribution[68,69]

$$\boldsymbol{C}_i \sim \mathrm{MVB}\left(f_{00}, f_{01}, f_{10}, f_{11}\right)$$

in order to facilitate optimization and interpretation (Supplemental Material and Methods). Assuming the causal status vector of a SNP is independent from those of other SNPs ($\boldsymbol{C}_i \perp \boldsymbol{C}_j$ for $i \neq j$), the joint probability of the causal statuses of p SNPs is $\Pr(\boldsymbol{C}_1, \cdots, \boldsymbol{C}_p) = \prod_{i=1}^{p} \Pr(\boldsymbol{C}_i)$.

Given two genome-wide association studies with sample sizes n_1 and n_2 for the first and second populations, respectively, we derive the distribution of Z-scores, $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ (both are $p \times 1$ vectors), conditional on the causal status vectors for each population, $\boldsymbol{c}_1 = \left(c_{11}, \cdots, c_{p1}\right)^T$ and $\boldsymbol{c}_2 = \left(c_{12}, \cdots, c_{p2}\right)^T$. Although it is reasonable to suspect that there are nonzero cross-population correlations of effect sizes at shared causal SNPs, to facilitate inference, we impose the (potentially strong) assumption that $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ are independent given $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$. Thus, for population *j*,

$$\boldsymbol{Z}_j | \boldsymbol{c}_j \sim MVN\left(\boldsymbol{0}, \boldsymbol{V}_j + \sigma_j^2 \boldsymbol{V}_j \mathrm{diag}\left(\boldsymbol{c}_j\right) \boldsymbol{V}_j\right)$$

where $\boldsymbol{V}_j$ is the $p \times p$ LD matrix for population *j*; $\mathrm{diag}(\boldsymbol{c}_j)$ is a diagonal matrix in which the k^{th} diagonal element is 1 if $c_{kj} = 1$ and 0 if $c_{kj} = 0$; and $\sigma_j^2 = \left(n_j h_{gj}^2 / |\boldsymbol{c}_j|\right)$, where h_{gj}^2 and $|\boldsymbol{c}_j|$ are the SNP-heritability of the trait and the number of causal SNPs, respectively, in population *j* (Supplemental Material and Methods).

Finally, we derive the joint probability of $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ by integrating over all possible causal status vectors in the two populations:

$$\Pr(\boldsymbol{Z}_1, \boldsymbol{Z}_2; \boldsymbol{f}) = \sum_{\mathbf{c}_1} \sum_{\mathbf{c}_2} \left[\prod_{i=1}^{p} \Pr(\mathbf{C}_i = c_{i1}c_{i2}) \prod_{j=1}^{2} MVN\left(\boldsymbol{Z}_j; \boldsymbol{0}, \boldsymbol{V}_j + \sigma_j^2 \boldsymbol{V}_j \mathrm{diag}(\mathbf{c}_j) \boldsymbol{V}_j\right) \right] \quad \text{(Equation 1)}$$

where $\boldsymbol{f} = \left(f_{00}, f_{01}, f_{10}, f_{11}\right)$ is the vector of parameters of the MVB distribution. In practice, we partition the genome into approximately independent regions[70] and model the distribution of Z-scores at all regions as the product of the distribution of Z-scores in each region (Supplemental Material and Methods).

Estimating Genome-wide Proportions of Population-Specific/Shared Causal SNPs

We use expectation-maximization (EM) coupled with Markov Chain Monte Carlo (MCMC) to maximize the likelihood function in Equation 1 over the MVB parameters $\boldsymbol{f}$. We initialize $\boldsymbol{f}$ to $\boldsymbol{f} = (0, -3.9, -3.9, 3.9)$ which corresponds to 2% of SNPs being causal in population 1, 2% being causal in population 2, and 2% being shared causals. In the expectation step, we approximate the surrogate function $Q\left(\boldsymbol{f}|\boldsymbol{f}^{(t)}\right)$ using an efficient Gibbs sampler; in the maximization step, we maximize $Q\left(\boldsymbol{f}|\boldsymbol{f}^{(t)}\right)$ using analytical formulae (Supplemental Material and Methods). From the estimated $\boldsymbol{f}$, denoted $\boldsymbol{f}^*$, we recover the proportions of population-specific and shared causal SNPs. For computational efficiency, we apply the EM algorithm to each chromosome in parallel and aggregate the chromosomal estimates to obtain estimates of the genome-wide proportions of population-specific/shared causal SNPs.

Evaluating per-SNP Posterior Probabilities of Being Causal in a Single or Both Populations

We estimate the posterior probability of each SNP to be causal in a single population (population-specific) or both populations (shared), using the estimated genome-wide proportions of population-specific and shared causal variants (obtained from $\boldsymbol{f}^*$) as prior probabilities in an empirical Bayes framework. Specifically, for each SNP i, we evaluate the posterior probabilities $\Pr(\boldsymbol{C}_i = 01|\boldsymbol{Z}_1, \boldsymbol{Z}_2; \boldsymbol{f}^*)$, $\Pr(\boldsymbol{C}_i = 10|\boldsymbol{Z}_1, \boldsymbol{Z}_2; \boldsymbol{f}^*)$, and $\Pr(\boldsymbol{C}_i = 11|\boldsymbol{Z}_1, \boldsymbol{Z}_2; \boldsymbol{f}^*)$. Since evaluating these probabilities requires integrating over the posterior probabilities of all $2^{(2p)}$ possible causal status configurations, we use a Gibbs sampler to efficiently approximate the posterior probabilities (Supplemental Material and Methods).

Estimating the Numbers of Population-Specific/Shared Causal SNPs in a Region

We infer the posterior expected numbers of population-specific/shared causal SNPs in a region (e.g., an LD block or a chromosome) conditional on the Z-scores ($\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$) by summing, across all SNPs in the region, the per-SNP posterior probabilities of being causal in a single or both populations. For example, in a region with p SNPs, the posterior expected number of shared causal SNPs is $\mathrm{E}[q_{11}|\boldsymbol{Z}_1, \boldsymbol{Z}_2; \boldsymbol{f}^*] = \sum_{i=1}^{p} \mathrm{E}[1_{\{\boldsymbol{C}_i=11\}}|\boldsymbol{Z}_1, \boldsymbol{Z}_2; \boldsymbol{f}^*] = \sum_{i=1}^{p} \Pr(\boldsymbol{C}_i = 11|\boldsymbol{Z}_1, \boldsymbol{Z}_2; \boldsymbol{f}^*)$. Since SNPs in a region are highly correlated, invalidating the use of jackknife to estimate standard errors, we refrain from reporting standard errors of the posterior expected regional numbers of population-specific/shared causal SNPs.

Defining LD Blocks that Are Approximately Independent in Two Populations

For computational efficiency, PESCA assumes that, in both populations, a SNP in a given block is independent from all SNPs in all other blocks. This assumption requires defining blocks of SNPs that are approximately LD independent in both populations. To this end, we first compute the "transethnic LD matrix" ($\boldsymbol{V}_{trans}$) from the East Asian- and European-ancestry LD matrices ($\boldsymbol{V}_{EAS}$ and $\boldsymbol{V}_{EUR}$) by setting each element in the transethnic LD matrix to the larger of the East Asian-specific and European-specific pairwise LD; i.e., $\boldsymbol{V}_{trans,ij} = \boldsymbol{V}_{EAS,ij}$ if $|\boldsymbol{V}_{EAS,ij}| > |\boldsymbol{V}_{EUR,ij}|$ and $\boldsymbol{V}_{trans,ij} = \boldsymbol{V}_{EUR,ij}$ if $|\boldsymbol{V}_{EUR,ij}| > |\boldsymbol{V}_{EAS,ij}|$. The resulting matrix $\boldsymbol{V}_{trans}$ is block diagonal due to shared recombination hotspots in both populations; in practice, we apply this procedure to each chromosome separately to obtain 22 chromosome-wide transethnic LD matrices. We then apply LDetect[70] to define LD blocks within the transethnic LD matrix. Applying this procedure using the 1000 Genomes Phase 3 reference panel[59,60] to create the transethnic LD matrix produces 1,368 LD blocks (average length of 2 Mb) that are approximately independent in individuals of East Asian and European ancestry.

Enrichment of Population-Specific/Shared Causal SNPs in Functional Annotations

We define the enrichment of population-specific/shared causal SNPs in a functional annotation as the ratio between the posterior and prior expected numbers of population-specific/shared causal SNPs. Specifically, we estimate the enrichment of population-specific/shared causal SNPs in a functional annotation k relative to the genome-wide background as

$$\widehat{\alpha}_{k,\boldsymbol{b}} = \frac{E\left[q_{k,\boldsymbol{b}}\middle|\boldsymbol{Z}_1,\boldsymbol{Z}_2,\boldsymbol{f}^*\right]}{E\left[q_{k,\boldsymbol{b}}\middle|\boldsymbol{f}^*\right]} = \frac{\sum_{i\in\psi(k)}\Pr(\boldsymbol{C}_i=\boldsymbol{b}|\boldsymbol{Z}_1,\boldsymbol{Z}_2,\boldsymbol{f}^*)}{p_k\Pr(\boldsymbol{C}_i=\boldsymbol{b})}$$

where $\boldsymbol{b} \in \{01,\ 10,\ 11\}$, $q_{k,\boldsymbol{b}}$ is the number of population-specific ($\boldsymbol{b} = 01$ or $\boldsymbol{b} = 10$) or shared ($\boldsymbol{b} = 11$) causal variants, $\psi(k)$ is the set of SNPs in functional annotation k, and p_k is the number of SNPs in functional annotation k. The numerator, $E[q_{k,\boldsymbol{b}}\big|\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{f}^*]$, and denominator, $E[q_{k,\boldsymbol{b}}\big|\boldsymbol{f}^*]$, represent the posterior (conditioned on Z-scores) and prior expected numbers of causal SNPs in functional annotation k, respectively. We estimate the standard error of $\widehat{\alpha}_{k,\boldsymbol{b}}$ using block jackknife over 1,368 non-overlapping approximately LD-independent blocks across the entire genome. The resulting enrichment test statistics, $(\widehat{\alpha}_{k,\boldsymbol{b}} - 1)/\mathrm{SE}(\widehat{\alpha}_{k,\boldsymbol{b}})$, approximately follow a t-distribution with degrees of freedom equal to the number of blocks minus 1.[71] Since we are interested in identifying categories of SNPs that harbor more population-specific/shared causal SNPs than expected (i.e., enrichment > 1), we report p values from a one-tailed t test where the null hypothesis is enrichment ≤ 1.

We note that our definition of enrichment of causal SNPs is related to, but conceptually different from, enrichment of SNP-heritability.[13,16,62] A positive enrichment of causal SNPs in a functional category indicates that, compared to the genome-wide background, there are more causal SNPs in that category than expected; a positive enrichment of SNP-heritability in a category indicates that the average per-SNP effect size in the category is larger than the genome-wide average per-SNP effect size.

Simulation Framework

We used real chromosome 22 genotypes of 10,000 individuals of East Asian ancestry from CONVERGE[66] and 50,000 individuals of white British ancestry from the UK Biobank[72,73] to simulate causal effects and phenotypes. First, we used PLINK[74] (v.1.9) to remove redundant SNPs in the 1000 Genomes Phase 3 reference panel[59,60] such that there are no pairs of SNPs with $r_{ij}^2 > 0.95$ ($i \neq j$). We also removed strand-ambiguous SNPs and SNPs with

MAF $<$ 1% in either reference panel, resulting in a total of $M =$ 8,599 SNPs on chromosome 22 to use in simulations.

Given genotypes at M SNPs for n_1 and n_2 individuals in populations 1 and 2, respectively, we assume the standard linear models $\boldsymbol{y}_1 = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}_1$ (population 1) and $\boldsymbol{y}_2 = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}_2$ (population 2). We assume the phenotypes are standardized within each population such that $\mathrm{E}[\boldsymbol{y}_1] = 0$, $\mathrm{Var}[\boldsymbol{y}_1] = \mathbf{I}$ and $\mathrm{E}[\boldsymbol{y}_2] = 0$, $\mathrm{Var}[\boldsymbol{y}_2] = \mathbf{I}$. Given $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$, the index sets of causal SNPs in each population, the effects at the i^{th} causal SNP in each population, β_{1i} and β_{2i}, are drawn from

$$\beta_{1\boldsymbol{c}_1}\Big|\boldsymbol{c}_1 \sim MVN\left(\mathbf{0}, \frac{h^2_{g1}}{|\boldsymbol{c}_1|}\mathbf{I}_{\boldsymbol{c}_1}\right),\ \beta_{2\boldsymbol{c}_2}|\boldsymbol{c}_2 \sim MVN\left(\mathbf{0}, \frac{h^2_{g2}}{|\boldsymbol{c}_2|}\mathbf{I}_{\boldsymbol{c}_2}\right)$$

where $|\boldsymbol{c}_1| = \sum_{i=1}^{M} c_{i1}$ and $|\boldsymbol{c}_2| = \sum_{i=1}^{M} c_{i2}$ are the total numbers of causal SNPs in each population, h^2_{g1} and h^2_{g2} are the total SNP-heritabilities in each population, and $\mathrm{E}[\beta_{1i}\beta_{1j}] = \mathrm{Cov}[\beta_{1i}, \beta_{1j}] = 0$ and $\mathrm{E}[\beta_{2i}\beta_{2j}] = \mathrm{Cov}[\beta_{2i}, \beta_{2j}] = 0$ for SNPs $i \neq j$. The effects at non-causal SNPs are set to 0. The environmental effects for the n^{th} individual in each population are drawn i.i.d. from $\epsilon_{1n} \sim N\left(0, 1 - h^2_{g1}\right)$ and $\epsilon_{2n} \sim N(0,\ 1 - h^2_{g2})$.

Finally, given the real genotypes and simulated phenotypes for each population, we compute Z-scores for all SNPs in population k as $\boldsymbol{Z}_k = (1/\sqrt{n_k})\boldsymbol{y}_k^T\boldsymbol{X}_k$.

Application to Nine Complex Traits and Diseases

We downloaded publicly available East Asian- and European-ancestry GWAS summary statistics for body mass index (BMI), mean corpuscular hemoglobin (MCH), mean corpuscular volume (MCV), high-density lipoprotein (HDL), low-density lipoprotein (LDL), total cholesterol (TC), triglycerides (TG), major depressive disorder (MDD), and rheumatoid arthritis (RA) from various sources (Table 1). The European-ancestry BMI GWAS is doubly corrected for genomic inflation factor,[63] which induces downward-bias in the estimated SNP-heritability; we correct this bias by re-inflating the Z-scores for this GWAS by a factor of 1.24. For all traits, we restrict to SNPs with MAF $>$ 5% in both populations to reduce noise in the LD matrices estimated from 1000 Genomes.[59,60] We use PLINK[74] (v.19) to remove redundant SNPs such that $\hat{r}^2_{ij} < 0.95$ for all SNPs $i \neq j$ in both ancestry-matched 1000 Genomes[59,60] reference panels. The resulting numbers of SNPs that were analyzed for each trait are listed in Table 1.

For each trait, we test for enrichment of population-specific/shared causal SNPs in 53 publicly available tissue-specific gene annotations,[62] each of which represents a set of genes that are "specifically expressed" in a GTEx[75] tissue (referred to as "SEG annotations"). We set the threshold for statistical significance to p value $<$ 0.05/53 (Bonferroni correction for the number of tests performed per trait).

Results

Performance of PESCA in Simulations

We assessed the performance of PESCA in simulations starting from real genotypes of individuals with East Asian[66] (EAS) or European[72,73] (EUR) ancestry ($\mathrm{N_{EAS}}$ = 10K, $\mathrm{N_{EUR}}$ = 50K, M = 8,599 SNPs) (Material and Methods). First, we find that when in-sample LD from the GWAS is available, PESCA yields approximately unbiased estimates of the numbers of population-specific/shared causal SNPs (Figure 2, top panel). For example, in simulations where we randomly selected 50 EAS-specific, 50 EUR-specific, and 50 shared causal SNPs, we obtained estimates (and corresponding standard errors) of 37.8 (4.5) EAS-specific, 40.3 (4.9) EUR-specific, and 64.9 (6.3) shared causal SNPs, respectively. When external reference LD is used (in this case, from 1000 Genomes[59,60]), PESCA yields a slight upward bias (Figure 2, bottom panel); on the same simulated data, we obtained estimates of 48.0 (5.9) EAS-specific, 53.7 (7.44) EUR-specific, and 78.8 (7.6) shared causal SNPs.

We observe a slight decrease in accuracy as the effective sample size, the product of SNP-heritability and sample size ($N \times h^2_g$), decreases (Figures S1–S5). This is expected as the likelihood of the GWAS summary statistics is a function of $N \times h^2_g$ (Material and Methods)—as the expected per-SNP variance at causal SNPs ($N \times h^2_g$ divided by the number of causal SNPs) decreases, GWAS summary statistics provide less information on the causal status of each SNP. Since it is often the case that the sample size of one GWAS is larger than that of the other, we perform simulations in which SNP-heritability is fixed to 0.05 in both populations, the EAS sample size is fixed to $\mathrm{N_{EAS}} = 10^4$, and the EUR sample size is varied such that the effective sample size of the EUR GWAS is 1–5× larger than that of the EAS GWAS. We find that the genome-wide estimators are relatively robust with in-sample LD; with external estimates of LD, when effective sample size differs by a factor of 2 or more, the estimator for the number of EUR-specific causal SNPs becomes less biased while the EAS-specific and shared causal estimators become increasingly inflated (Figure S6). In addition, while it seems likely that the effect sizes of shared causal SNPs would be positively correlated across populations, the PESCA model assumes zero cross-population correlation in order to facilitate inference (Material and Methods). We therefore perform simulations under an alternative model in which EAS and EUR effect sizes at shared causal SNPs are positively correlated and find that our estimates of the genome-wide numbers of shared and population-specific causal SNPs become increasingly inflated and deflated, respectively, as the correlation increases from 0 to 1 (Figure S7).

Next, we use the estimated genome-wide proportions of population-specific/shared causal SNPs to evaluate per-SNP posterior probabilities of being causal in a single population (EAS only or EUR only) or in both populations (Material and Methods). For each of the three causal configurations of interest (EAS only, EUR only, and shared), we observe an increase in the average correlation between the per-SNP posterior probabilities and the true causal status vector for that configuration as $N \times h^2_g$ increases and as the total number of causal SNPs decreases (i.e., as per-SNP causal effect sizes increase) (Figures S8 and S9). As expected, as the simulated proportion of shared causal SNPs increases, the average correlation between the

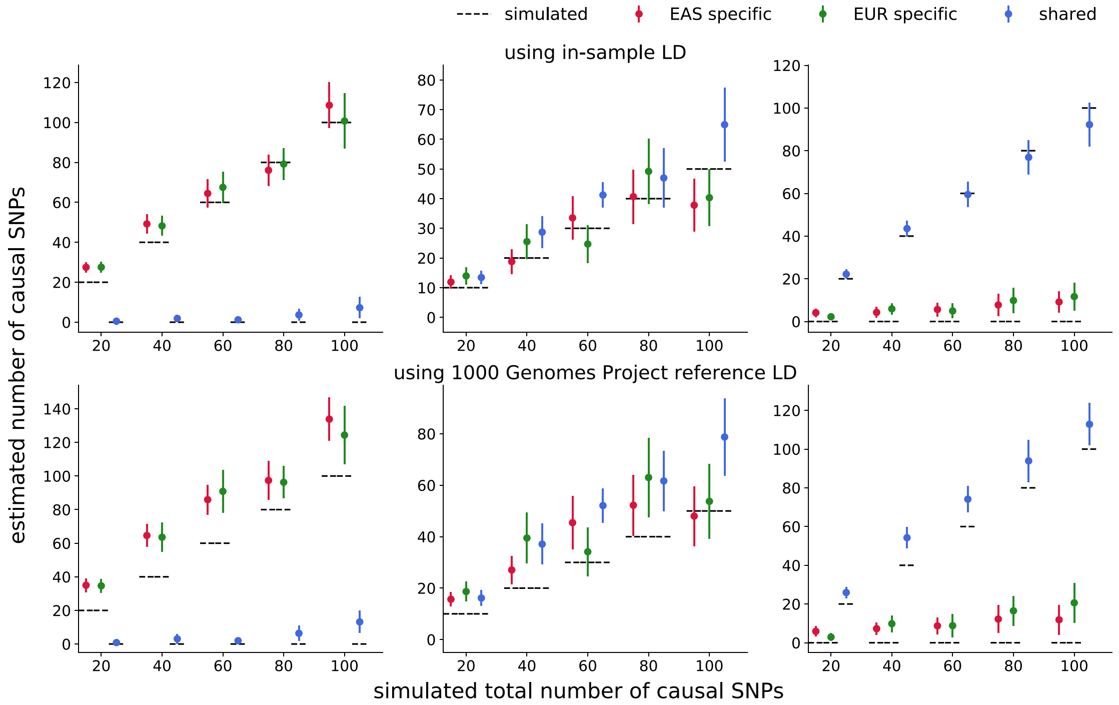


Figure 2. Genome-wide Estimates of the Numbers of Population-Specific/Shared Causal SNPs in Simulations
The estimates are approximately unbiased when in-sample LD is used (top) and upward-biased when external reference LD is used (bottom). For both populations, we simulate such that the product of SNP-heritability and GWAS sample size is 500. Mean and standard errors were obtained from 25 independent simulations. Error bars represent ± 1.96 of the standard error.

posterior probabilities and true causal status vectors increases for the shared causal configuration and decreases for the population-specific causal configurations (Figures S8 and S9). Since we did not have access to individual-level genotypes sampled from an ancestral group with shorter LD blocks (e.g., African-ancestry individuals), we use the EAS and EUR LD scores of each SNP as proxies for the strength of LD in the region housing the SNP to investigate the impact of population-specific LD patterns on the per-SNP posterior probabilities. Among the true causal SNPs (shared or population-specific), the posterior probabilities are relatively invariant to the magnitude of the EAS and EUR LD scores (Figure S10). In other words, under the PESCA framework, power to detect a given true causal SNP does not depend on its LD score in either population. Restricting to a set of "high-posterior SNPs" (defined here as SNPs with posterior probability greater than some threshold *t*), we investigate whether PESCA systematically misclassifies SNPs based on the magnitude of their LD scores. Again, we observe that the average EAS and EUR LD scores do not vary significantly between the true and false positive classifications (Table S1). We then assessed whether our proposed statistics for testing for enrichment of population-specific/shared causal SNPs in functional annotations (Material and Methods) are well calibrated under the null hypothesis of no enrichment. Overall, when both population-specific and shared causal SNPs are drawn at random, the enrichment test statistics are conservative at different levels of polygenicity and GWAS power ($N \times h_g^2$), irrespective of whether in-sample LD or external reference LD is used (Figures S11 and S16).

Finally, we evaluated the computational efficiency of each stage of inference. In the first stage of inference—estimating genome-wide proportions of population-specific/shared causal SNPs—the maximization step of the EM algorithm uses Gibbs sampling to efficiently sample from the posterior of the causal status vectors (Supplemental Material and Methods). We set both the number of burn-in iterations and the number of samples to 5,000 for the MCMC within the maximization step and found that the overall EM typically converged within 200 iterations (Figures S17–S19). Run-time per EM-iteration increases with the number of causal SNPs (Figure S20); for example, in simulations with a total of 8,589 SNPs, when the maximum number of EM iterations was set to 200, PESCA took an average of 90 min to obtain estimates in simulations with 20 randomly selected causal variants and 360 min in simulations with 100 randomly selected causal SNPs. This is expected because the likelihood function being maximized is proportional to the Bayes factor of only the causal SNPs (Supplemental Material and Methods). In the second stage of inference—evaluating posterior probabilities for each SNP—we set both the number of burn-in iterations and the number of samples to 5,000 for the MCMC and, to ensure stable estimates of the posterior probability, we report the average posterior probability from 20 iterations of the Gibbs sampling procedure. The average run-time was 5 min in simulations with 20 causal variants and 28 min in simulations with 100 causal

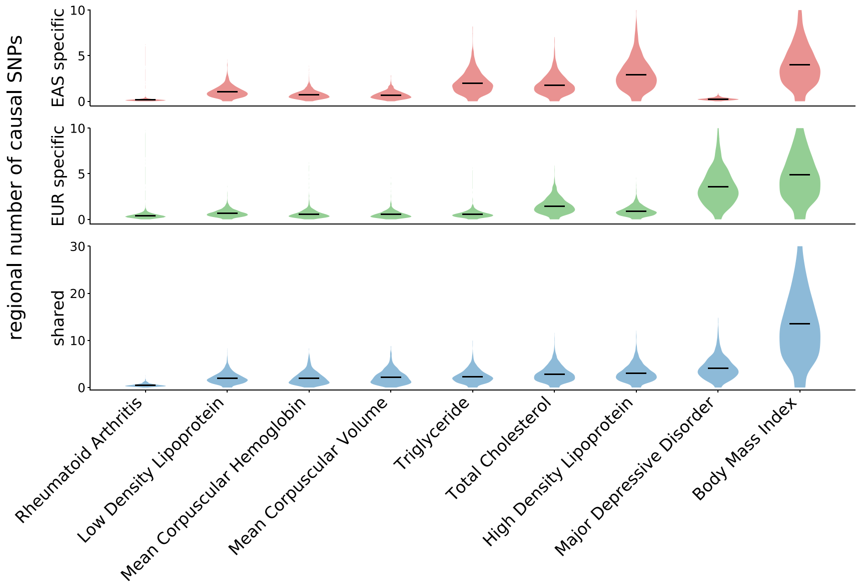


Figure 3. Distributions of the Numbers of Population-Specific/Shared Causal SNPs across 1,368 Regions that are Approximately Independent in Both EAS and EUR
Each violin plot represents the distribution of the posterior expected number of population-specific (red/green) or shared (blue) causal SNPs per region; details on how the regions were defined can be found in the Material and Methods. For a single region, the posterior expected number of SNPs in a given causal configuration is estimated by summing, across all SNPs in the region, the per-SNP posterior probabilities of having that causal configuration. The dark lines mark the means of the distributions. The traits are sorted on the x-axis by the average number of shared high-posterior SNPs per region.

variants (Figure S20). We note that both stages of inference can be parallelized to decrease run time.

Expected Genome-wide Proportions of Shared Causal SNPs for Nine Complex Traits

We obtained publicly available GWAS summary statistics for nine (non-independent) complex traits and diseases in individuals of EAS and EUR ancestry (average $N_{EAS} = 94{,}621$, $N_{EUR} = 103{,}507$) (Table 1) and applied PESCA to estimate the genome-wide proportions of population-specific/shared common causal SNPs (Material and Methods). To ensure convergence, we applied 750 EM iterations for each trait (Figures S21–S23). Across the nine traits, the estimated proportions of common causal SNPs in each population (the sum of the numbers of population-specific and shared causal SNPs) are consistent with previously reported estimates of polygenicity in single populations.[7,8,55,76,77] For example, we estimate that approximately 10% of common SNPs have nonzero effects on BMI in both EAS and EUR and that 2%–3% have nonzero effects on the lipids traits (Table 1). The low estimates for major depressive disorder and rheumatoid arthritis may be explained in part by their small GWAS sample sizes. While there is heterogeneity in the estimated proportions of shared causal SNPs across the nine traits, we find that most common causal SNPs are shared between the populations, consistent with findings from previous studies.[33] For example, for BMI, we estimate that approximately 96% of common causal SNPs in each population are also causal in the other; for total cholesterol (TC), we estimate that 73% of common causal SNPs in EAS and 77% of those in EUR are shared by both populations (Table 1).

High-Posterior SNPs Are Distributed Nearly Uniformly across the Genome

We define 1,368 regions that are approximately LD independent in both populations and estimate the posterior expected numbers of population-specific/shared causal SNPs in each region (Material and Methods). For all nine traits, high-posterior SNPs for both the population-specific and shared causal configurations are spread nearly uniformly across the genome (Figures 3 and S24–S31). For example, mean corpuscular hemoglobin (MCH) harbored, on average, 0.68 (SD 0.42) EAS-specific, 0.53 (SD 0.40) EUR-specific, and 2.19 (SD 1.46) shared high-posterior SNPs per region (Figures 3 and S29). Aggregating posterior probabilities by chromosome, we find that the posterior expected numbers of EAS-specific, EUR-specific, and shared causal SNPs per chromosome are highly correlated with chromosome length (Figures S32–S34), recapitulating previous findings based on regional SNP-heritability.[55,61]

Distributions of High-Posterior SNPs across GWAS Risk Regions

We aggregate per-SNP posterior probabilities within GWAS risk regions that are EAS-specific, EUR-specific, or shared by both populations and find that most GWAS risk regions harbor two or more shared high-posterior SNPs (Figures 4 and S35–S39), concordant with previous findings on allelic heterogeneity of complex traits.[55,78,79] On average across the 9 traits, we observe a 2.8× enrichment of shared high-posterior SNPs in population-specific GWAS risk regions relative to the genome-wide background. For example, for MCH, the EAS-specific and EUR-specific GWAS risk regions harbor an average of 3.0 (SD 1.7) and 3.3 (SD 1.5) shared high-posterior SNPs per region, respectively, whereas the average number of shared high-posterior SNPs per region across all regions is 2.0 (SD 1.3) (Figure 4). While BMI, the blood traits (MCH and MCV), and rheumatoid arthritis have similar numbers of EAS-specific and EUR-specific high-posterior SNPs in their population-specific GWAS risk regions, the lipids traits (HDL, LDL, total cholesterol, and triglycerides) have significantly

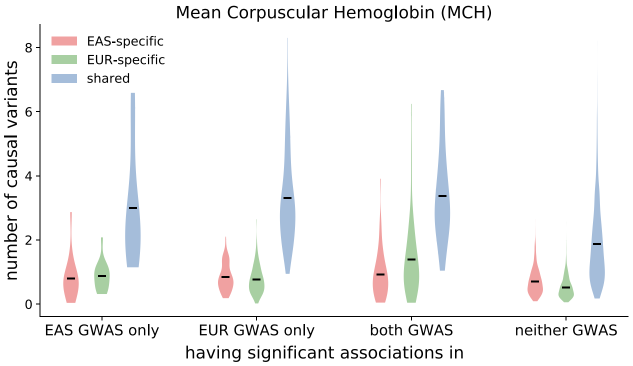


Figure 4. Distributions of the Numbers of Population-Specific/Shared Causal Variants at GWAS Risk Regions for Mean Corpuscular Hemoglobin (MCH)
Each violin plot represents the distribution of the posterior expected number of population-specific (red/green) or shared (blue) causal SNPs at regions with significant associations ($p_{GWAS} < 5\times10^{-8}$) in EAS GWAS only, EUR GWAS only, both EAS and EUR, and neither GWAS. The dark lines mark the means of the distributions.

more EAS-specific high-posterior SNPs in all GWAS risk regions (Figures 4 and S35–S39).

For each causal configuration (EAS-specific, EUR-specific, or shared), we examine the effect sizes of high-posterior SNPs (posterior probability > 0.8) in EAS and EUR (Figure 5). Across the 9 traits, the majority of EAS-specific high-posterior SNPs are nominally significant ($p_{GWAS} < 5\times10^{-6}$) either in the EAS GWAS only or in both GWASs. While five EUR-specific high-posterior SNPs are nominally significant in only the EAS GWAS, the majority are nominally significant either in the EUR GWAS only or in both GWASs. We observe strong correlations between the effect sizes in EAS and EUR for all three sets of high-posterior SNPs (Pearson r^2 of 0.79 [EAS-specific], 0.73 [EUR-specific], and 0.80 [shared]) that are driven by SNPs that are nominally significant in both GWASs (Figure 5). Taken together, these results suggest that most population-specific GWAS risk regions harbor shared causal variants that are undetected in the other population due to heterogeneity in LD structures, allele frequencies, and/or GWAS sample sizes.

Enrichment of High-Posterior SNPs near Genes Expressed in Trait-Relevant Tissues

Motivated by recent work that found enrichment of SNP-heritability in regions near genes that are "specifically expressed" in trait-relevant tissues and cell types (referred to as "SEG annotations"), we tested for enrichments of population-specific and shared causal SNPs in the same 53 tissue-specific SEG annotations.[62] For a given causal configuration, the enrichment of causal SNPs in an annotation is defined as the ratio between the posterior and prior expected numbers of causal SNPs in the annotation (Material and Methods). For 8 of the 9 traits, we find significant enrichment of shared high-posterior SNPs in at least one SEG annotation (p value < 0.05/53 to correct for 53 tests per trait) (Figures S40–S44). All SEG annotations with significant enrichments of population-specific high-posterior SNPs are also enriched with shared high-posterior SNPs for the same trait, providing additional evidence that many signatures of population-specific genetic architecture are induced by population-specific LD and allele frequencies rather than distinct genetic etiologies. We do not find enrichment of any high-posterior SNPs in any SEG annotation for major depressive disorder (MDD) (Figure S44), which could be due to low GWAS sample sizes (Table 1). Finally, for each SEG annotation, we obtain a meta-analyzed transethnic SNP-heritability enrichment by computing the inverse-variance weighted average of the EAS and EUR SNP-heritability enrichments (estimated separately using stratified LD score regression[13,16]). We observe a strong correlation between the meta-analyzed SNP-heritability enrichments and the enrichments of shared high-posterior SNPs (Figure 6), suggesting that SNP-heritability enrichments are largely driven by many low-effect SNPs rather than a small number of high-effect SNPs.

Discussion

We have presented PESCA, a method for estimating the genome-wide proportions of SNPs with nonzero effects in a single population (population-specific) or in two populations (shared) from GWAS summary statistics and estimates of LD. We applied PESCA to EAS and EUR GWAS summary statistics for nine complex traits and find that, while the lipids traits have significantly more EAS-specific common causal SNPs compared to the remaining traits, the majority of common causal SNPs are shared by both populations. Regions that harbor statistically significant GWAS associations for one population are enriched with SNPs with high-posterior probability of being causal in both populations. Morever, high-posterior SNPs (posterior probability > 0.8 for any causal configuration) have highly correlated effect sizes in EAS and EUR, recapitulating findings of previous studies.[33] For all traits except MDD, we identify tissue-specific SEG annotations[62] enriched with shared high-posterior SNPs and observe that all SEG annotations enriched with population-specific high-posterior SNPs are a subset of those enriched with shared high-posterior SNPs. Taken together, our results indicate that most population-specific GWAS risk regions contain shared common causal SNPs that are undetected in the second population due to differences in LD or allele frequencies. This suggests that localizing shared components of genetic architecture and explicitly correcting for population-specific LD and allele frequencies may help improve transferability of results from well-powered European-ancestry studies to other understudied populations. Based on the simulation results in Figure S1 (in which 100% of causal SNPs are shared) and our estimates of SNP-heritability for the traits in Table 1, we recommend applying PESCA to

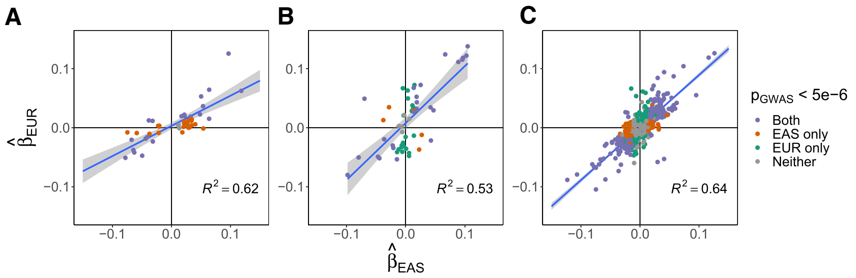


Figure 5. Marginal Regression Coefficients of High-Posterior SNPs for Nine Complex Traits
Each plot corresponds to one of the three causal configurations of interest: EAS-specific (A), EUR-specific (B), and shared (C). Each point represents a SNP with posterior probability > 0.8 for a single trait. The x-axis and y-axis mark the estimated marginal effect sizes in EAS and EUR, respectively. The colors indicate whether the SNP is nominally significant ($p_{GWAS} < 5\times10^{-6}$) in both GWASs (purple), the EAS GWAS only (orange), the EUR GWAS only (green), or in neither GWAS (gray). The gray band marks the 95% confidence interval of the regression line.

summary statistics for which the effective per-SNP sample size, $N\times h_g^2$ divided by the number of causal SNPs, is at least 3 for both GWASs. For a typical quantitative trait (e.g., Table 1), this corresponds to a total effective sample size of approximately $N\times h_g^2 > 10,000$.

We conclude by discussing the caveats and limitations of our analyses. First, the estimated proportions of causal SNPs must be interpreted with caution as they can be influenced by gene-environment interactions. For example, if a SNP has a nonzero effect on a trait only in the presence of environmental factors that are specific to EAS-ancestry individuals, PESCA will interpret that SNP as an EAS-specific causal SNP even though it would have a nonzero effect in EUR-ancestry individuals in the presence of the same environmental factors.

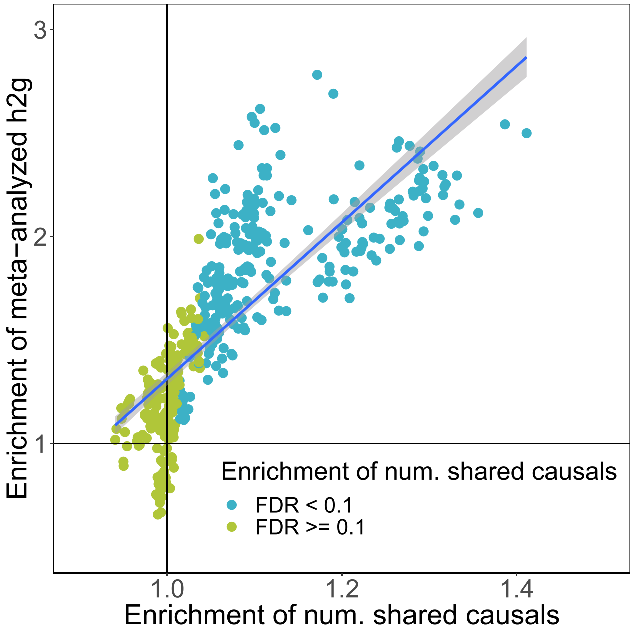


Figure 6. Enrichments of Shared High-Posterior SNPs in 53 Tissue-Specific Functional Categories are Highly Correlated with SNP-Heritability Enrichments
Each point is a trait-tissue pair; each tissue-specific functional category (SEG annotation) is a set of genes that are "specifically expressed" in one of 53 GTEx tissues. The x-axis is the estimated enrichment of shared high-posterior SNPs in the SEG annotation from PESCA. The y-axis is the meta-analyzed transethnic SNP-heritability explained by the SEG annotation, defined as the inverse-variance weighted average of the EAS and EUR SNP-heritability enrichments (estimated separately using stratified LD score regression). The points are colored by whether the trait has a statistically significant enrichment of shared high-posterior SNPs in the corresponding SEG annotation (FDR < 0.1). The gray band marks the 95% confidence interval of the regression line. Enrichment estimates and standard errors for each trait-tissue pair can be found in Figures S40–S44.

Second, we chose to analyze a set of traits for which EAS and EUR GWAS summary statistics were publicly available. Since most publicly available summary statistics of large-scale GWAS are meta-analyses of smaller studies, in-sample LD is often unavailable. While PESCA with in-sample LD is relatively robust to differential GWAS power, with external LD, performance decreases when the GWAS effective sample sizes differ by more than a factor of 2×. We note, however, that for the real traits analyzed in this work, effective sample size differs by a maximum factor of 2× (mean corpuscular hemoglobin; Table 1). Additionally, PESCA currently cannot be applied to admixed populations if in-sample LD is unavailable. An extension of PESCA to properly account for external/noisy estimates of LD would thus increase its utility; we defer a thorough investigation of this to future work. In parallel, in light of ongoing efforts at several institutions to establish biobanks,[72,73,80–82] we believe that well-powered GWASs (with in-sample LD) will become increasingly available for diverse and admixed populations. Another challenge is that many publicly available summary statistics were computed from fixed-effect meta-analyses or linear mixed models. Since the PESCA model is defined with respect to GWAS marginal effects estimated by ordinary least-squares (OLS) regression, it is unclear whether PESCA is sensitive to non-OLS association statistics, which have different statistical properties. We defer a thorough investigation of this to future work.

Third, we restricted our analyses to SNPs with MAF $>$ 5% in both populations to reduce noise in the LD matrices estimated from external reference panels. Consequently, the estimates we report in this work do not capture effects of low frequency or rare variants that are not well-tagged by common SNPs. Furthermore, since most common variants are shared across continental populations and rarer variants tend to localize among closely related populations,[60] our study design undersamples population-specific causal variants. We note, however, that lower MAF thresholds can be used if in-sample LD is available. We also note that for the purpose of improving transferability of polygenic risk scores (PRSs) across populations, prediction accuracy depends largely on the accuracy of the PRS weights at common SNPs (the average per-SNP contribution to total SNP-heritability is larger for common SNPs than for low frequency or rare variants[11]).

Finally, PESCA can be sensitive to model misspecification. For computational efficiency, PESCA relies on having regions that are approximately LD independent in both populations; if there is LD leakage between regions, the estimated proportions of causal SNPs will be biased. We therefore recommend defining LD blocks for each pair of populations one analyzes. Similarly, to facilitate inference, PESCA does not explicitly model cross-population correlations of effect sizes at shared causal variants. We conjecture that modeling these correlations can further improve performance.

Supplemental Data

Supplemental Data can be found online at https://doi.org/10.1016/j.ajhg.2020.04.012.

Acknowledgments

We are grateful to Kangcheng Hou, Alkes L. Price, and Steven Gazal for helpful discussions that greatly improved the quality of this manuscript. We also thank Na Cai, Sriram Sankararaman, Jonathan Flint, and the UK Biobank (application #33297) for providing resources that made this work possible. This work was funded in part by the National Institutes of Health (NIH) under awards R01HG009120, R01HG006399, R01MH115676, U01CA194393, T32NS048004, T32MH073526, and T32HG002536.

Declaration of Interests

The authors declare no competing interests.

Received: December 13, 2019
Accepted: April 20, 2020
Published: May 21, 2020

Web Resources

Biobank Japan summary statistics, http://jenger.riken.jp/en/result
CONVERGE genotype data, https://www.ebi.ac.uk/eva/?eva-study=PRJNA289433
GIANT summary statistics, http://portals.broadinstitute.org/collaboration/giant
GWAS summary statistics for hematological traits, http://www.bloodcellgenetics.org
LD score regression, https://github.com/bulik/ldsc
PESCA, https://github.com/huwenboshi/pesca
PLINK 1.9, https://www.cog-genomics.org/plink/1.9/
Popcorn, https://github.com/brielin/Popcorn
SEG annotations, https://data.broadinstitute.org/alkesgroup/LDSCORE/LDSC_SEG_ldscores

References

1. Campbell, M.C., and Tishkoff, S.A. (2010). The evolution of human genetic and phenotypic variation in Africa. Curr. Biol. *20*, R166–R173.
2. Cavalli-Sforza, L.L., Menozzi, P., and Piazza, A. (1993). Demic expansions and human evolution. Science *259*, 639–646.
3. Pritchard, J.K., Pickrell, J.K., and Coop, G. (2010). The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr. Biol. *20*, R208–R215.
4. Laland, K.N., Odling-Smee, J., and Myles, S. (2010). How culture shaped the human genome: bringing genetics and the human sciences together. Nat. Rev. Genet. *11*, 137–148.
5. Martin, A.R., Gignoux, C.R., Walters, R.K., Wojcik, G.L., Neale, B.M., Gravel, S., Daly, M.J., Bustamante, C.D., and Kenny, E.E. (2017). Human demographic history impacts genetic risk prediction across diverse populations. Am. J. Hum. Genet. *100*, 635–649.
6. Timpson, N.J., Greenwood, C.M.T., Soranzo, N., Lawson, D.J., and Richards, J.B. (2018). Genetic architecture: the shape of the genetic contribution to human traits and disease. Nat. Rev. Genet. *19*, 110–124.
7. O'Connor, L.J., Schoech, A.P., Hormozdiari, F., Gazal, S., Patterson, N., and Price, A.L. (2019). Extreme Polygenicity of Complex Traits Is Explained by Negative Selection. Am. J. Hum. Genet. *105*, 456–476.
8. Zeng, J., de Vlaming, R., Wu, Y., Robinson, M.R., Lloyd-Jones, L.R., Yengo, L., Yap, C.X., Xue, A., Sidorenko, J., McRae, A.F., et al. (2018). Signatures of negative selection in the genetic architecture of human complex traits. Nat. Genet. *50*, 746–753.
9. Zhang, Y., Qi, G., Park, J.-H., and Chatterjee, N. (2018). Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. Nat. Genet. *50*, 1318–1326.
10. Zhu, X., and Stephens, M. (2018). Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes. Nat. Commun. *9*, 4361.
11. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. *42*, 565–569.
12. Schoech, A.P., Jordan, D.M., Loh, P.-R., Gazal, S., O'Connor, L.J., Balick, D.J., Palamara, P.F., Finucane, H.K., Sunyaev, S.R., and Price, A.L. (2019). Quantification of frequency-dependent genetic architectures in 25 UK Biobank traits reveals action of negative selection. Nat. Commun. *10*, 790.
13. 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.
14. Speed, D., Hemani, G., Johnson, M.R., and Balding, D.J. (2012). Improved heritability estimation from genome-wide SNPs. Am. J. Hum. Genet. *91*, 1011–1021.
15. Speed, D., Cai, N., Johnson, M.R., Nejentsev, S., Balding, D.J.; and UCLEB Consortium (2017). Reevaluation of SNP heritability in complex human traits. Nat. Genet. *49*, 986–992.
16. 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.

17. 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.
18. Popejoy, A.B., and Fullerton, S.M. (2016). Genomics is failing on diversity. Nature *538*, 161–164.
19. Rosenberg, N.A., Huang, L., Jewett, E.M., Szpiech, Z.A., Jankovic, I., and Boehnke, M. (2010). Genome-wide association studies in diverse populations. Nat. Rev. Genet. *11*, 356–366.
20. Visscher, P.M., Brown, M.A., McCarthy, M.I., and Yang, J. (2012). Five years of GWAS discovery. Am. J. Hum. Genet. *90*, 7–24.
21. Kanai, M., Akiyama, M., Takahashi, A., Matoba, N., Momozawa, Y., Ikeda, M., Iwata, N., Ikegawa, S., Hirata, M., Matsuda, K., et al. (2018). Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. Nat. Genet. *50*, 390–400.
22. Akiyama, M., Okada, Y., Kanai, M., Takahashi, A., Momozawa, Y., Ikeda, M., Iwata, N., Ikegawa, S., Hirata, M., Matsuda, K., et al. (2017). Genome-wide association study identifies 112 new loci for body mass index in the Japanese population. Nat. Genet. *49*, 1458–1467.
23. Li, Z., Chen, J., Yu, H., He, L., Xu, Y., Zhang, D., Yi, Q., Li, C., Li, X., Shen, J., et al. (2017). Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia. Nat. Genet. *49*, 1576–1583.
24. 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.
25. Ng, M.C.Y., Shriner, D., Chen, B.H., Li, J., Chen, W.-M., Guo, X., Liu, J., Bielinski, S.J., Yanek, L.R., Nalls, M.A., et al.; FIND Consortium; eMERGE Consortium; DIAGRAM Consortium; MuTHER Consortium; and MEta-analysis of type 2 DIabetes in African Americans Consortium (2014). Meta-analysis of genome-wide association studies in African Americans provides insights into the genetic architecture of type 2 diabetes. PLoS Genet. *10*, e1004517.
26. Franceschini, N., Fox, E., Zhang, Z., Edwards, T.L., Nalls, M.A., Sung, Y.J., Tayo, B.O., Sun, Y.V., Gottesman, O., Adeyemo, A., et al.; Asian Genetic Epidemiology Network Consortium (2013). Genome-wide association analysis of blood-pressure traits in African-ancestry individuals reveals common associated genes in African and non-African populations. Am. J. Hum. Genet. *93*, 545–554.
27. Schick, U.M., Jain, D., Hodonsky, C.J., Morrison, J.V., Davis, J.P., Brown, L., Sofer, T., Conomos, M.P., Schurmann, C., McHugh, C.P., et al. (2016). Genome-wide association study of platelet count identifies ancestry-specific loci in Hispanic/Latino Americans. Am. J. Hum. Genet. *98*, 229–242.
28. Kichaev, G., and Pasaniuc, B. (2015). Leveraging Functional-Annotation Data in Trans-ethnic Fine-Mapping Studies. Am. J. Hum. Genet. *97*, 260–271.
29. Mancuso, N., Rohland, N., Rand, K.A., Tandon, A., Allen, A., Quinque, D., Mallick, S., Li, H., Stram, A., Sheng, X., et al.; PRACTICAL consortium (2016). The contribution of rare variation to prostate cancer heritability. Nat. Genet. *48*, 30–35.
30. Brown, B.C., Ye, C.J., Price, A.L., Zaitlen, N.; and Asian Genetic Epidemiology Network Type 2 Diabetes Consortium (2016). Transethnic genetic-correlation estimates from summary statistics. Am. J. Hum. Genet. *99*, 76–88.
31. Morris, A.P. (2011). Transethnic meta-analysis of genomewide association studies. Genet. Epidemiol. *35*, 809–822.
32. Lam, M., Chen, C.-Y., Li, Z., Martin, A.R., Bryois, J., Ma, X., Gaspar, H., Ikeda, M., Benyamin, B., Brown, B.C., et al.; Schizophrenia Working Group of the Psychiatric Genomics Consortium; Indonesia Schizophrenia Consortium; and Genetic REsearch on schizophreniA neTwork-China and the Netherlands (GREAT-CN) (2019). Comparative genetic architectures of schizophrenia in East Asian and European populations. Nat. Genet. *51*, 1670–1678.
33. Marigorta, U.M., and Navarro, A. (2013). High trans-ethnic replicability of GWAS results implies common causal variants. PLoS Genet. *9*, e1003566.
34. Kraft, P., Zeggini, E., and Ioannidis, J.P.A. (2009). Replication in genome-wide association studies. Stat. Sci. *24*, 561–573.
35. Li, Y.R., and Keating, B.J. (2014). Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations. Genome Med. *6*, 91.
36. Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., Kochi, Y., Ohmura, K., Suzuki, A., Yoshida, S., et al.; RACI consortium; and GARNET consortium (2014). Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature *506*, 376–381.
37. Wu, Y., Waite, L.L., Jackson, A.U., Sheu, W.H.-H., Buyske, S., Absher, D., Arnett, D.K., Boerwinkle, E., Bonnycastle, L.L., Carty, C.L., et al. (2013). Trans-ethnic fine-mapping of lipid loci identifies population-specific signals and allelic heterogeneity that increases the trait variance explained. PLoS Genet. *9*, e1003379.
38. Asimit, J.L., Rainbow, D.B., Fortune, M.D., Grinberg, N.F., Wicker, L.S., and Wallace, C. (2019). Stochastic search and joint fine-mapping increases accuracy and identifies previously unreported associations in immune-mediated diseases. Nat. Commun. *10*, 3216.
39. Zaitlen, N., Paşaniuc, B., Gur, T., Ziv, E., and Halperin, E. (2010). Leveraging genetic variability across populations for the identification of causal variants. Am. J. Hum. Genet. *86*, 23–33.
40. Wen, X., Luca, F., and Pique-Regi, R. (2015). Cross-population joint analysis of eQTLs: fine mapping and functional annotation. PLoS Genet. *11*, e1005176.
41. Vilhjálmsson, B.J., Yang, J., Finucane, H.K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P.-R., Bhatia, G., Do, R., et al.; Schizophrenia Working Group of the Psychiatric Genomics Consortium, Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) study (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. Am. J. Hum. Genet. *97*, 576–592.
42. Márquez-Luna, C., Loh, P.-R., Price, A.L.; South Asian Type 2 Diabetes (SAT2D) Consortium; and SIGMA Type 2 Diabetes Consortium (2017). Multiethnic polygenic risk scores improve risk prediction in diverse populations. Genet. Epidemiol. *41*, 811–823.
43. Lewis, C.M., and Vassos, E. (2017). Prospects for using risk scores in polygenic medicine. Genome Med. *9*, 96.

44. Curtis, D. (2018). Polygenic risk score for schizophrenia is more strongly associated with ancestry than with schizophrenia. Psychiatr. Genet. *28*, 85–89.
45. Chen, C.-Y., Han, J., Hunter, D.J., Kraft, P., and Price, A.L. (2015). Explicit Modeling of Ancestry Improves Polygenic Risk Scores and BLUP Prediction. Genet. Epidemiol. *39*, 427–438.
46. Wojcik, G.L., Graff, M., Nishimura, K.K., Tao, R., Haessler, J., Gignoux, C.R., Highland, H.M., Patel, Y.M., Sorokin, E.P., Avery, C.L., et al. (2019). Genetic analyses of diverse populations improves discovery for complex traits. Nature *570*, 514–518.
47. Sirugo, G., Williams, S.M., and Tishkoff, S.A. (2019). The Missing Diversity in Human Genetic Studies. Cell *177*, 26–31.
48. Gurdasani, D., Barroso, I., Zeggini, E., and Sandhu, M.S. (2019). Genomics of disease risk in globally diverse populations. Nat. Rev. Genet. *20*, 520–535.
49. Martin, A.R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B.M., and Daly, M.J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. Nat. Genet. *51*, 584–591.
50. Ikeda, M., Takahashi, A., Kamatani, Y., Momozawa, Y., Saito, T., Kondo, K., Shimasaki, A., Kawase, K., Sakusabe, T., Iwayama, Y., et al. (2019). Genome-Wide Association Study Detected Novel Susceptibility Genes for Schizophrenia and Shared Trans-Populations/Diseases Genetic Effect. Schizophr. Bull. *45*, 824–834.
51. Shi, H., Gazal, S., Kanai, M., Koch, E.M., Schoech, A.P., Kim, S.S., Luo, Y., Amariuta, T., Okada, Y., Raychaudhuri, S., et al. (2019). Population-specific causal disease effect sizes in functionally important regions impacted by selection. bioRxiv. https://doi.org/10.1101/803452.
52. Galinsky, K.J., Reshef, Y.A., Finucane, H.K., Loh, P.-R., Zaitlen, N., Patterson, N.J., Brown, B.C., and Price, A.L. (2019). Estimating cross-population genetic correlations of causal effect sizes. Genet. Epidemiol. *43*, 180–188.
53. Guo, J., Bakshi, A., Wang, Y., Jiang, L., Yengo, L., Goddard, M.E., Visscher, P.M., and Yang, J. (2019). Quantifying genetic heterogeneity between continental populations for human height and body mass index. bioRxiv. https://doi.org/10.1101/839373.
54. 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.
55. Shi, H., Kichaev, G., and Pasaniuc, B. (2016). Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. Am. J. Hum. Genet. *99*, 139–153.
56. Hou, K., Burch, K.S., Majumdar, A., Shi, H., Mancuso, N., Wu, Y., Sankararaman, S., and Pasaniuc, B. (2019). Accurate estimation of SNP-heritability from biobank-scale data irrespective of genetic architecture. Nat. Genet. *51*, 1244–1251.
57. Farh, K.K.-H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W.J., Beik, S., Shoresh, N., Whitton, H., Ryan, R.J.H., Shishkin, A.A., et al. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature *518*, 337–343.
58. Huang, H., Fang, M., Jostins, L., Umićević Mirkov, M., Boucher, G., Anderson, C.A., Andersen, V., Cleynen, I., Cortes, A., Crins, F., et al.; International Inflammatory Bowel Disease Genetics Consortium (2017). Fine-mapping inflammatory bowel disease loci to single-variant resolution. Nature *547*, 173–178.
59. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., McVean, G.A.; and 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.
60. 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.
61. Loh, P.-R., Bhatia, G., Gusev, A., Finucane, H.K., Bulik-Sullivan, B.K., Pollack, S.J., de Candia, T.R., Lee, S.H., Wray, N.R., Kendler, K.S., et al.; Schizophrenia Working Group of Psychiatric Genomics Consortium (2015). Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. Nat. Genet. *47*, 1385–1392.
62. 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.
63. 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.
64. Astle, W.J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A.L., Mead, D., Bouman, H., Riveros-Mckay, F., Kostadima, M.A., et al. (2016). The allelic landscape of human blood cell trait variation and links to common complex disease. Cell *167*, 1415–1429.e19.
65. 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.
66. Cai, N., Bigdeli, T.B., Kretzschmar, W., Li, Y., Liang, J., Song, L., Hu, J., Li, Q., Jin, W., Hu, Z., et al.; CONVERGE consortium (2015). Sparse whole-genome sequencing identifies two loci for major depressive disorder. Nature *523*, 588–591.
67. Wray, N.R., Ripke, S., Mattheisen, M., Trzaskowski, M., Byrne, E.M., Abdellaoui, A., Adams, M.J., Agerbo, E., Air, T.M., Andlaur, T.M.F., et al. (2018). Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. Nat. Genet. *50*, 668–681.
68. Dai, B., Ding, S., and Wahba, G. (2013). Multivariate bernoulli distribution. Bernoulli *19*, 1465–1483.
69. Shi, H., Pasaniuc, B., and Lange, K.L. (2015). A multivariate Bernoulli model to predict DNaseI hypersensitivity status from haplotype data. Bioinformatics *31*, 3514–3521.
70. Berisa, T., and Pickrell, J.K. (2016). Approximately independent linkage disequilibrium blocks in human populations. Bioinformatics *32*, 283–285.

71. Miller, R.G. (1968). Jackknifing variances. Ann. Math. Stat. *39*, 567–582.

72. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. *12*, e1001779.

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

74. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.

75. Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al.; GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. Nat. Genet. *45*, 580–585.

76. Johnson, R., Shi, H., Pasaniuc, B., and Sankararaman, S. (2018). A unifying framework for joint trait analysis under a non-infinitesimal model. Bioinformatics *34*, i195–i201.

77. Holland, D., Frei, O., Desikan, R., Fan, C.-C., Shadrin, A.A., Smeland, O.B., Sundar, V.S., Thompson, P., Andreassen, O.A., and Dale, A.M. (2019). Beyond SNP Heritability: Polygenicity and Discoverability of Phenotypes Estimated with a Univariate Gaussian Mixture Model. bioRxiv. https://doi.org/10.1101/133132.

78. Hormozdiari, F., Zhu, A., Kichaev, G., Ju, C.J.-T., Segrè, A.V., Joo, J.W.J., Won, H., Sankararaman, S., Pasaniuc, B., Shifman, S., and Eskin, E. (2017). Widespread allelic heterogeneity in complex traits. Am. J. Hum. Genet. *100*, 789–802.

79. Gusev, A., Bhatia, G., Zaitlen, N., Vilhjalmsson, B.J., Diogo, D., Stahl, E.A., Gregersen, P.K., Worthington, J., Klareskog, L., Raychaudhuri, S., et al. (2013). Quantifying missing heritability at known GWAS loci. PLoS Genet. *9*, e1003993.

80. Nagai, A., Hirata, M., Kamatani, Y., Muto, K., Matsuda, K., Kiyohara, Y., Ninomiya, T., Tamakoshi, A., Yamagata, Z., Mushiroda, T., et al.; BioBank Japan Cooperative Hospital Group (2017). Overview of the BioBank Japan Project: Study design and profile. J. Epidemiol. *27* (3S), S2–S8.

81. Gaziano, J.M., Concato, J., Brophy, M., Fiore, L., Pyarajan, S., Breeling, J., Whitbourne, S., Deen, J., Shannon, C., Humphries, D., et al. (2016). Million Veteran Program: A mega-biobank to study genetic influences on health and disease. J. Clin. Epidemiol. *70*, 214–223.

82. Leitsalu, L., Haller, T., Esko, T., Tammesoo, M.-L., Alavere, H., Snieder, H., Perola, M., Ng, P.C., Mägi, R., Milani, L., et al. (2015). Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. Int. J. Epidemiol. *44*, 1137–1147.

The American Journal of Human Genetics, Volume *106*

Supplemental Data

Localizing Components of Shared Transethnic Genetic Architecture of Complex Traits from GWAS Summary Data

Huwenbo Shi, Kathryn S. Burch, Ruth Johnson, Malika K. Freund, Gleb Kichaev, Nicholas Mancuso, Astrid M. Manuel, Natalie Dong, and Bogdan Pasaniuc

Contents

1 Supplemental Figures

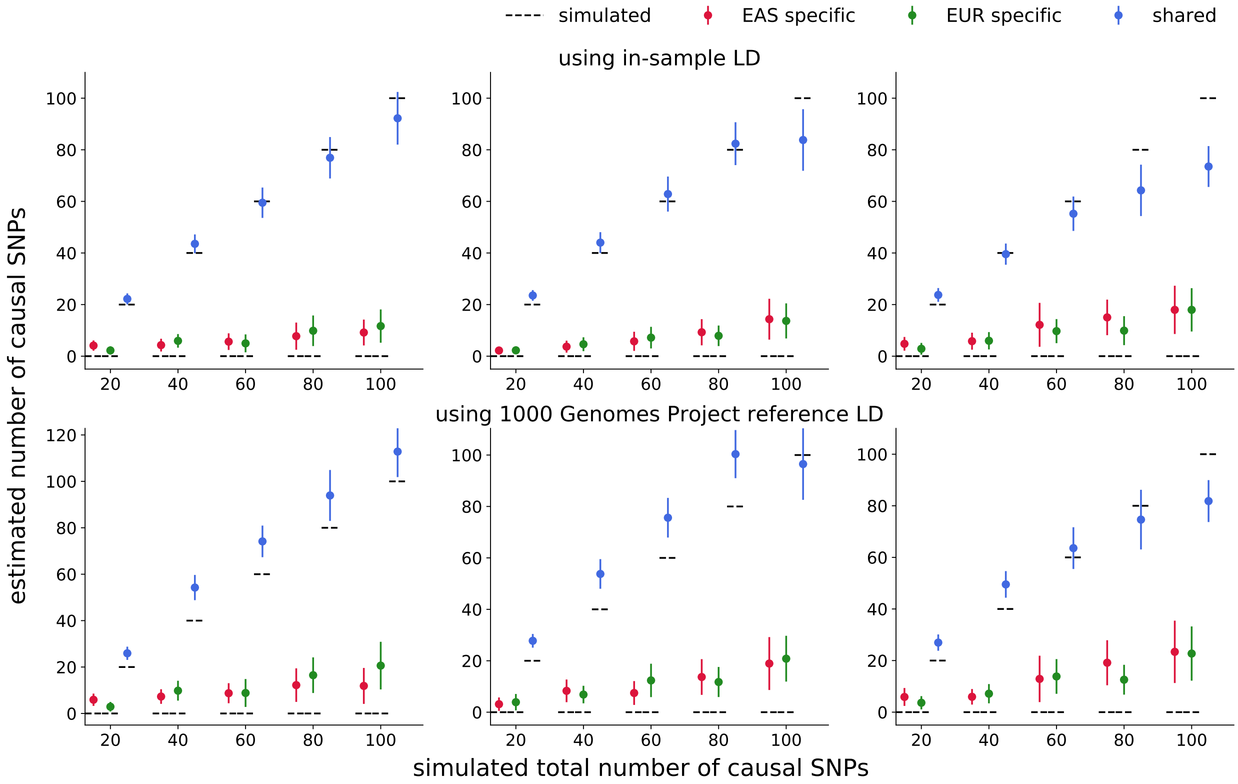


Figure S1: **PESCA estimators for the genome-wide numbers of population-specific/shared causal SNPs when 100% of causal variants are shared.** We simulated 20 to 100 causal variants per population (x-axis), all of which were shared by both populations. We set the product of SNP-heritability and sample size of the GWAS to 500 (left column), 375 (middle column), and 250 (right column), which correspond to per-SNP effective sample sizes (N x per-snp variance) that decrease from 25 to 5 (left), 18.75 to 3.75 (middle), and 12.5 to 2.5 (right). Each dot represents the mean across 25 simulations and error bars represent ±1.96 s.e.m.

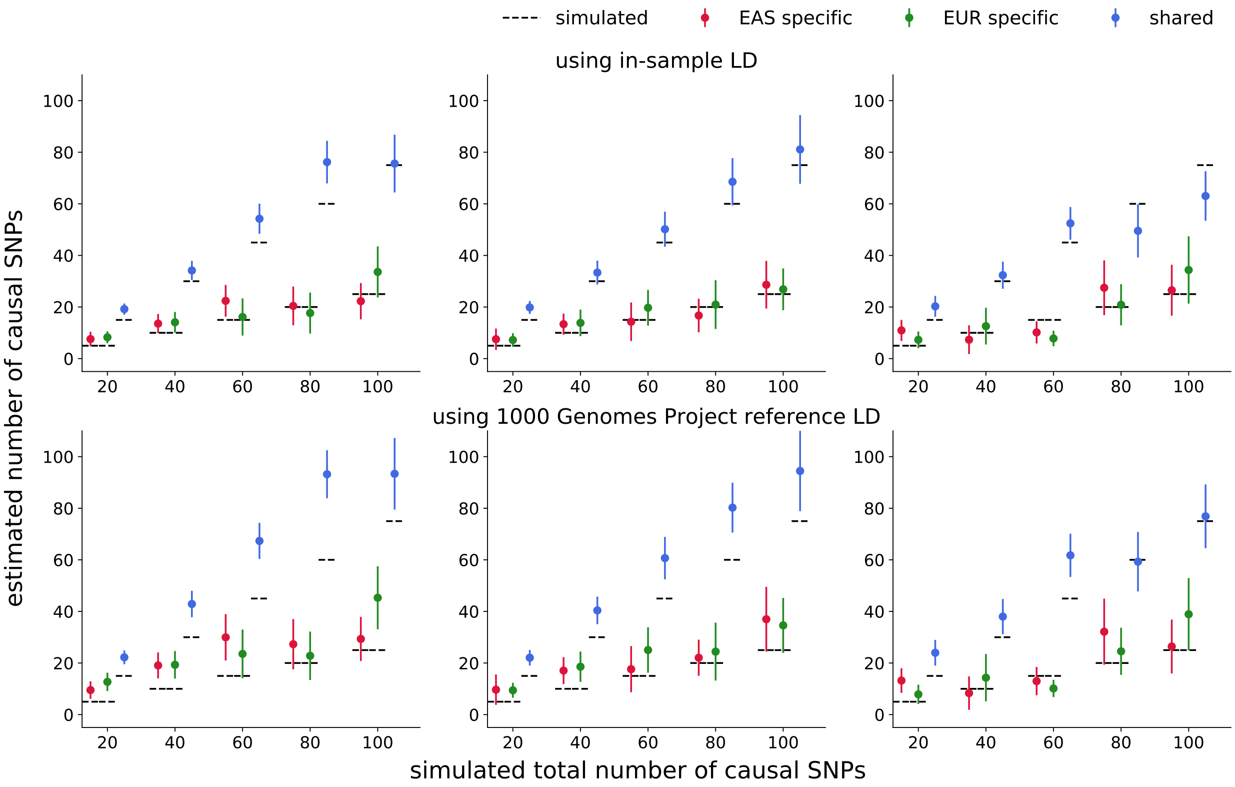


Figure S2: **PESCA estimators for the genome-wide numbers of population-specific/shared causal SNPs when 75% of causal variants are shared.** We simulated 20 to 100 causal variants per population (x-axis), 75% of which were shared; the remaining 25% were population-specific. We set the product of SNP-heritability and sample size of the GWAS to 500 (left column), 375 (middle column), and 250 (right column). Each dot represents the mean across 25 simulations and error bars represent ±1.96 s.e.m.

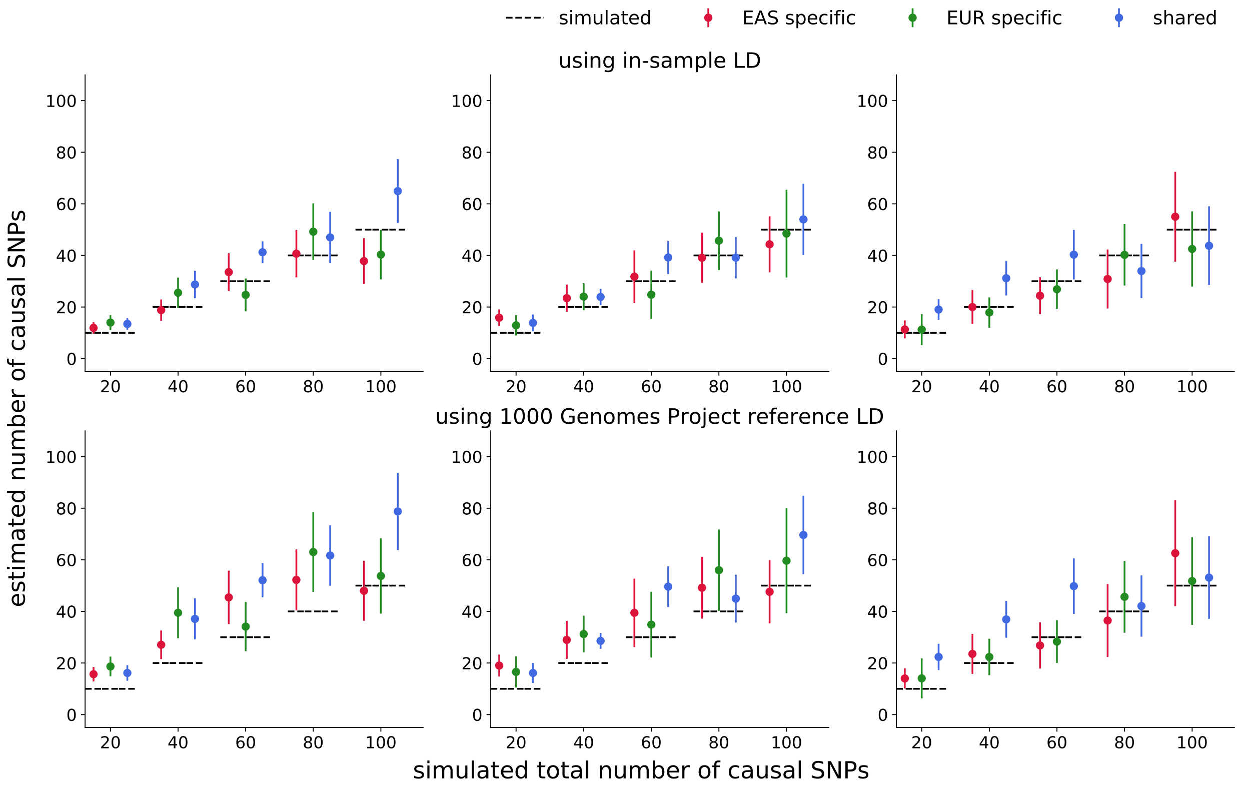


Figure S3: **PESCA estimators for the genome-wide numbers of population-specific/shared causal SNPs when 50% of causal variants are shared.** We simulated 20 to 100 causal variants per population (x-axis), 50% of which were shared; the remaining 50% were population-specific. We set the product of SNP-heritability and sample size of the GWAS to 500 (left column), 375 (middle column), and 250 (right column). Each dot represents the mean across 25 simulations and error bars represent ±1.96 s.e.m.

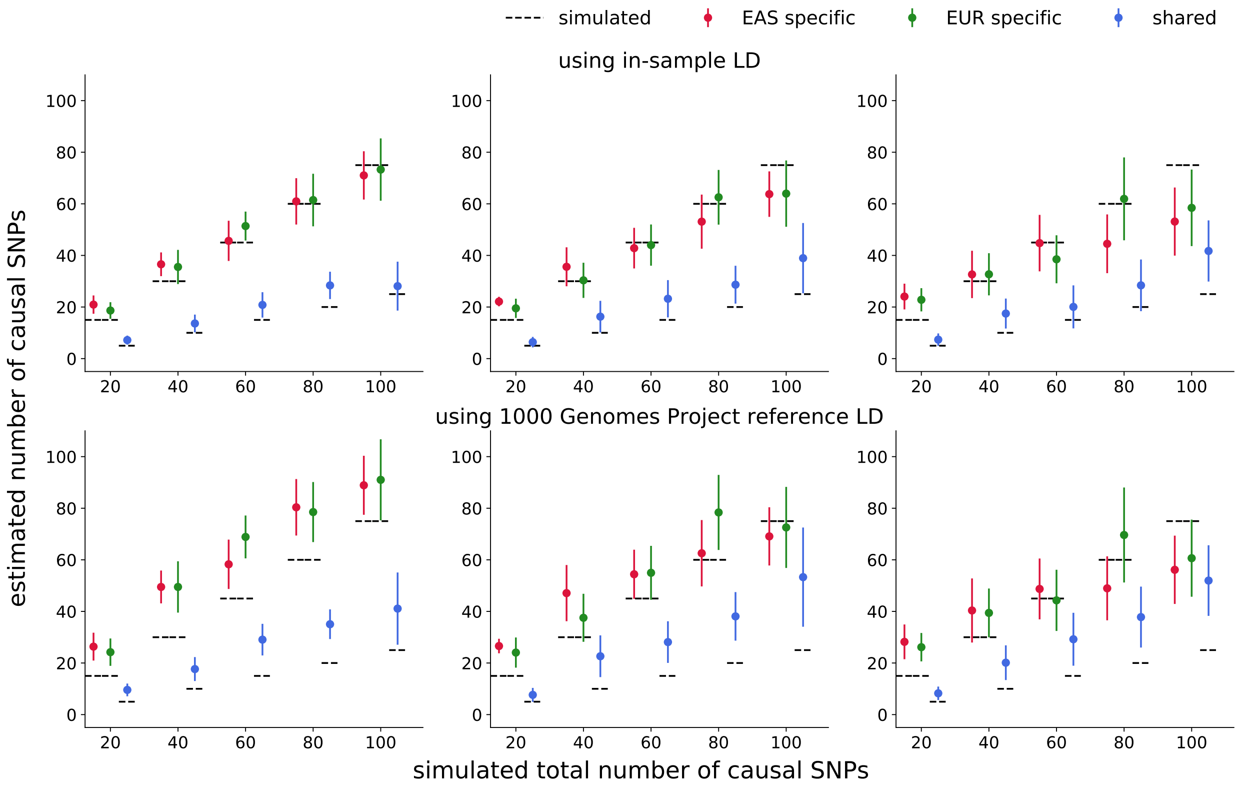


Figure S4: **PESCA estimators for the genome-wide numbers of population-specific/shared causal SNPs when 25% of causal variants are shared.** We simulated 20 to 100 causal variants per population (x-axis), 25% of which were shared; the remaining 75% were population-specific. We set the product of SNP-heritability and sample size of the GWAS to 500 (left column), 375 (middle column), and 250 (right column). Each dot represents the mean across 25 simulations and error bars represent $\pm$1.96 s.e.m.

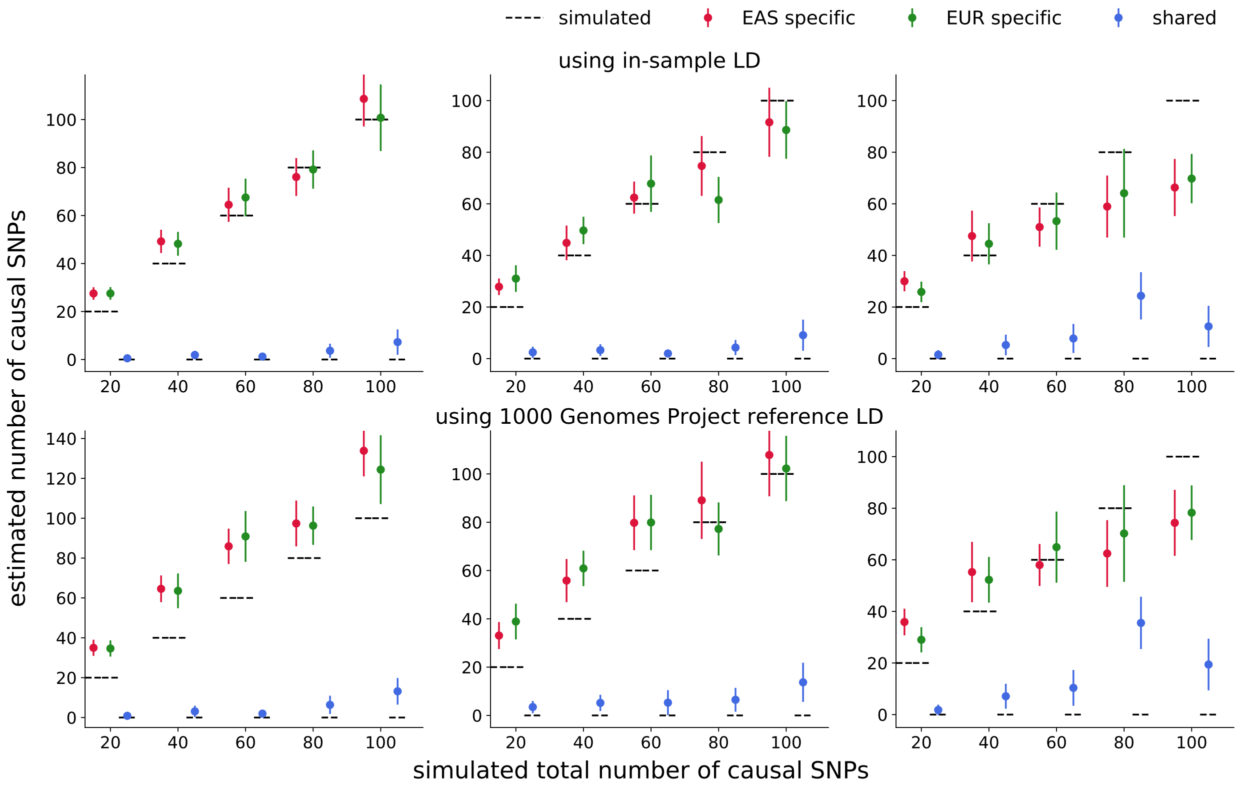


Figure S5: **PESCA estimators for the genome-wide numbers of population-specific/shared causal SNPs when 0% of causal variants are shared.** We simulated 20 to 100 causal variants per population (x-axis), all of which were population-specific. We set the product of SNP-heritability and sample size of the GWAS to 500 (left column), 375 (middle column), and 250 (right column). Each dot represents the mean across 25 simulations and error bars represent ± 1.96 s.e.m.

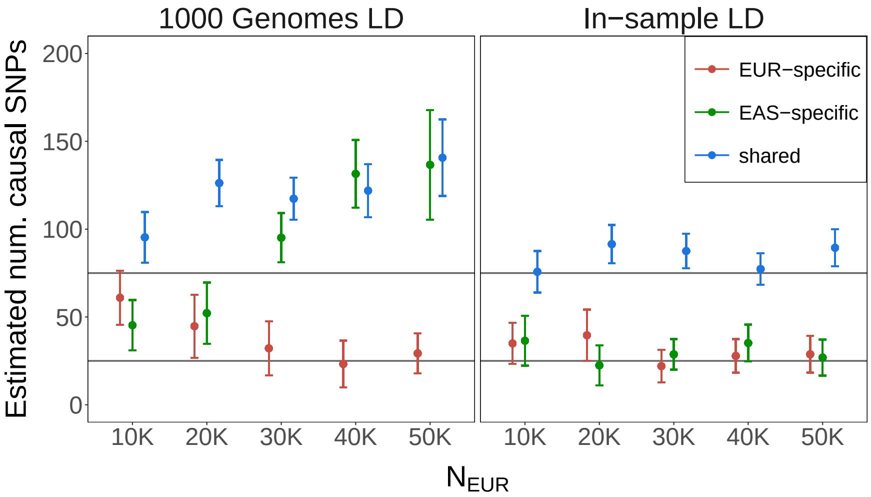


Figure S6: **Effect of differential effective sample size on PESCA estimates of the genome-wide numbers of population-specific/shared causal SNPs.** Total SNP-heritability was fixed to $h_g^2 = 0.05$ for both populations. $N_{EAS} = 10^4$ in all simulations; N_{EUR} was varied from 1×10^4 to 5×10^4 (x-axis). Horizontal lines mark the number of shared (75), EAS-specific (25), and EUR-specific (25) causal SNPs. Each dot represents the mean across 25 simulations and error bars represent $\pm$1.96 s.e.m. The colors correspond to the estimators for the numbers of population-specific (red and green) and shared (blue) causal variants.

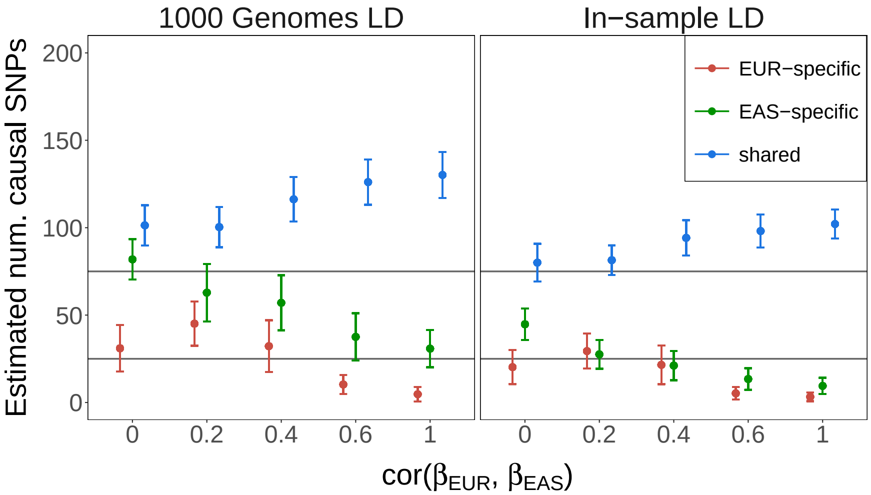


Figure S7: **Effect of cross-population correlation of causal effects on PESCA estimates of the genome-wide numbers of population-specific/shared causal SNPs.** Total SNP-heritability was fixed to $h_g^2 = 0.05$ for both populations. $N_{EAS} = 1 \times 10^4$ and $N_{EUR} = 2 \times 10^4$ in all simulations. Horizontal lines mark the number of shared (75), EAS-specific (25), and EUR-specific (25) causal SNPs. The correlation of effect sizes at causal SNPs was varied from 0 to 1 (x-axis). Each dot represents the mean across 25 simulations and error bars represent $\pm$1.96 s.e.m. The colors correspond to the estimators for the numbers of population-specific (red and green) and shared (blue) causal variants.

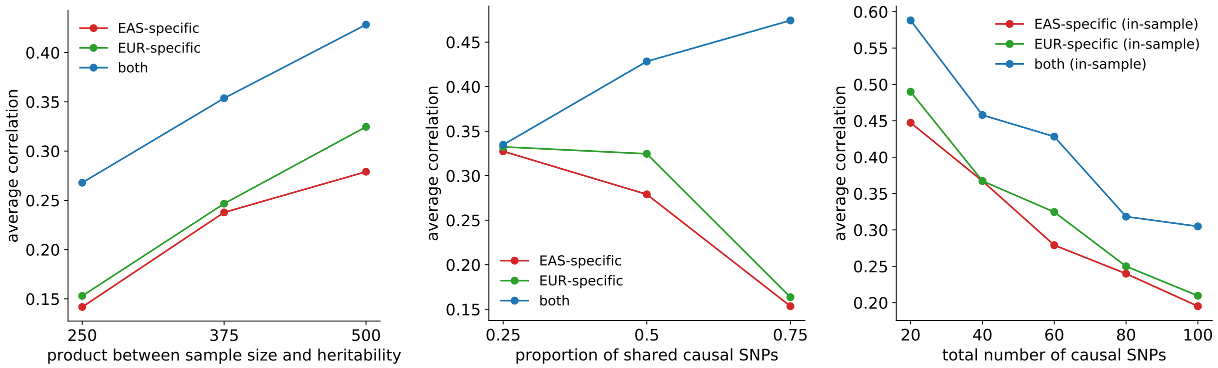


Figure S8: **Accuracy of PESCA posterior probabilities in simulations using in-sample LD.** Each point represents the average correlation (across 25 simulation replicates) between the vector of per-SNP posterior probabilities of causality and the vector of simulated causal statuses for one of the possible causal configurations (EAS-specific, EUR-specific, or both) as a function of $N \times h_g^2$ (left), the total number of causal SNPs in both populations (middle), and the proportion of shared causal SNPs (right). The correlations are calculated from a set of SNPs with MAF $>$ 5% in both populations that satisfy $r_{ij}^2 < 0.95$ for all pairs of SNPs ($i \neq j$) in both populations (Methods).

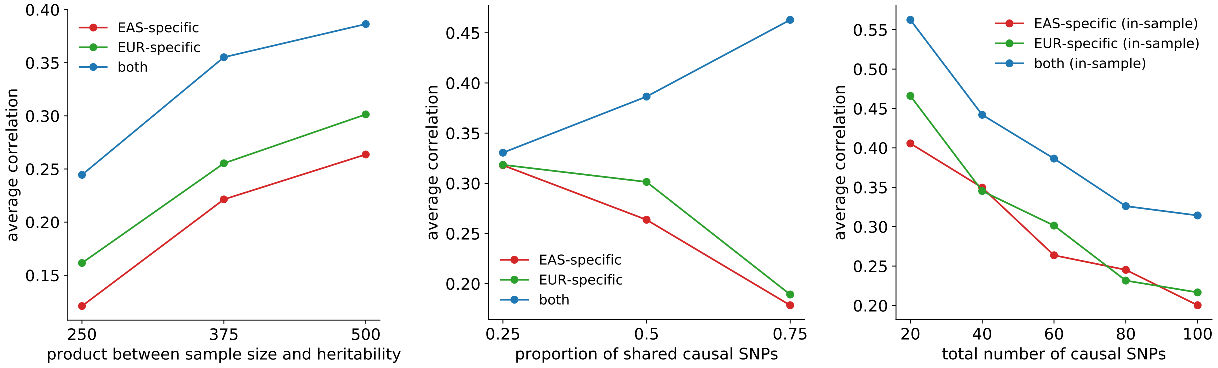


Figure S9: **Accuracy of PESCA posterior probabilities in simulations using external reference panel LD (1000 Genomes).** Each point represents the average correlation (across 25 simulation replicates) between the vector of per-SNP posterior probabilities of causality and the vector of simulated causal statuses for one of the possible causal configurations (EAS-specific, EUR-specific, or both) as a function of $N \times h_g^2$ (left), the total number of causal SNPs in both populations (middle), and the proportion of shared causal SNPs (right). The correlations are calculated from a set of SNPs with MAF $>$ 5% in both populations that satisfy $r_{ij}^2 < 0.95$ for all pairs of SNPs ($i \neq j$) in both populations (Methods).

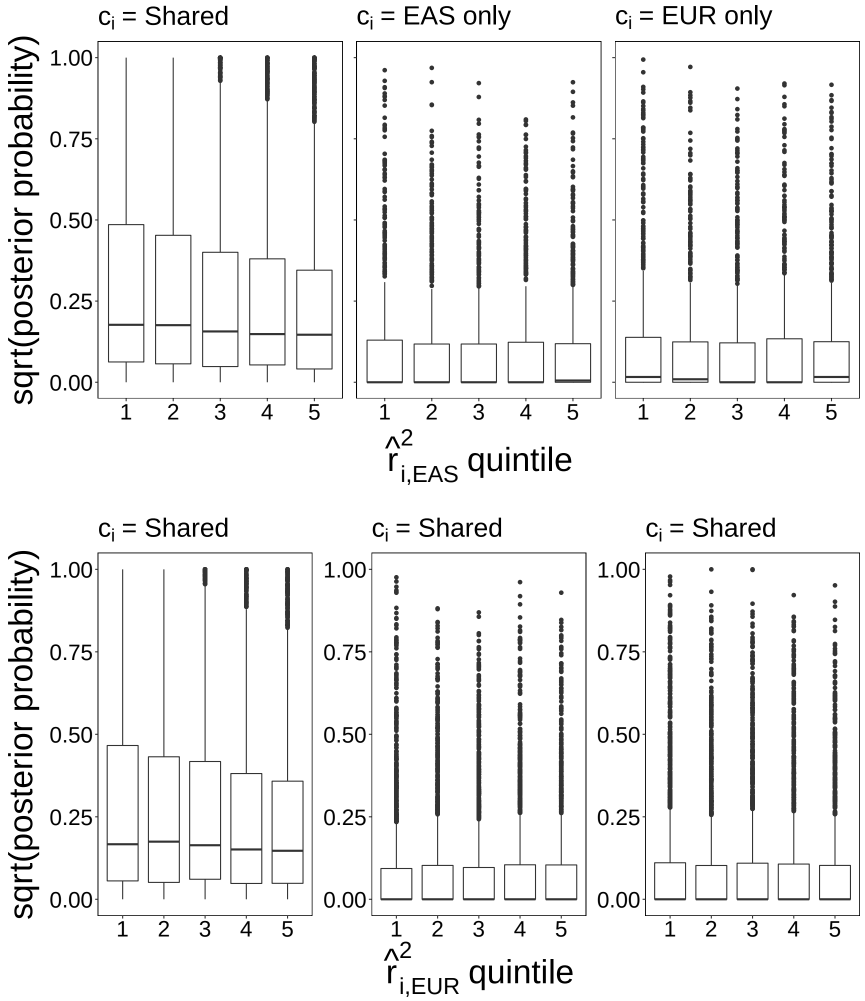


Figure S10: **Posterior probabilities of true causal SNPs with respect to LD score quintiles**. Total SNP-heritability was fixed to $h_g^2 = 0.05$ for both populations and $N_{EAS} = N_{EUR} = 10^4$. In each of the 200 simulation replicates, 75 shared (left panel), 25 EAS-specific (middle panel), and 25 EUR-specific (right panel) causal SNPs were drawn at random from 8,599 SNPs on chromosome 22. Each point in each boxplot represents a single true causal SNP. Each boxplot shows the distribution of per-SNP posterior probabilities for the corresponding correct causal configuration with respect to LD score quintiles in EAS (top) or EUR (bottom). We plot the square root of the posteriors to facilitate visualization.

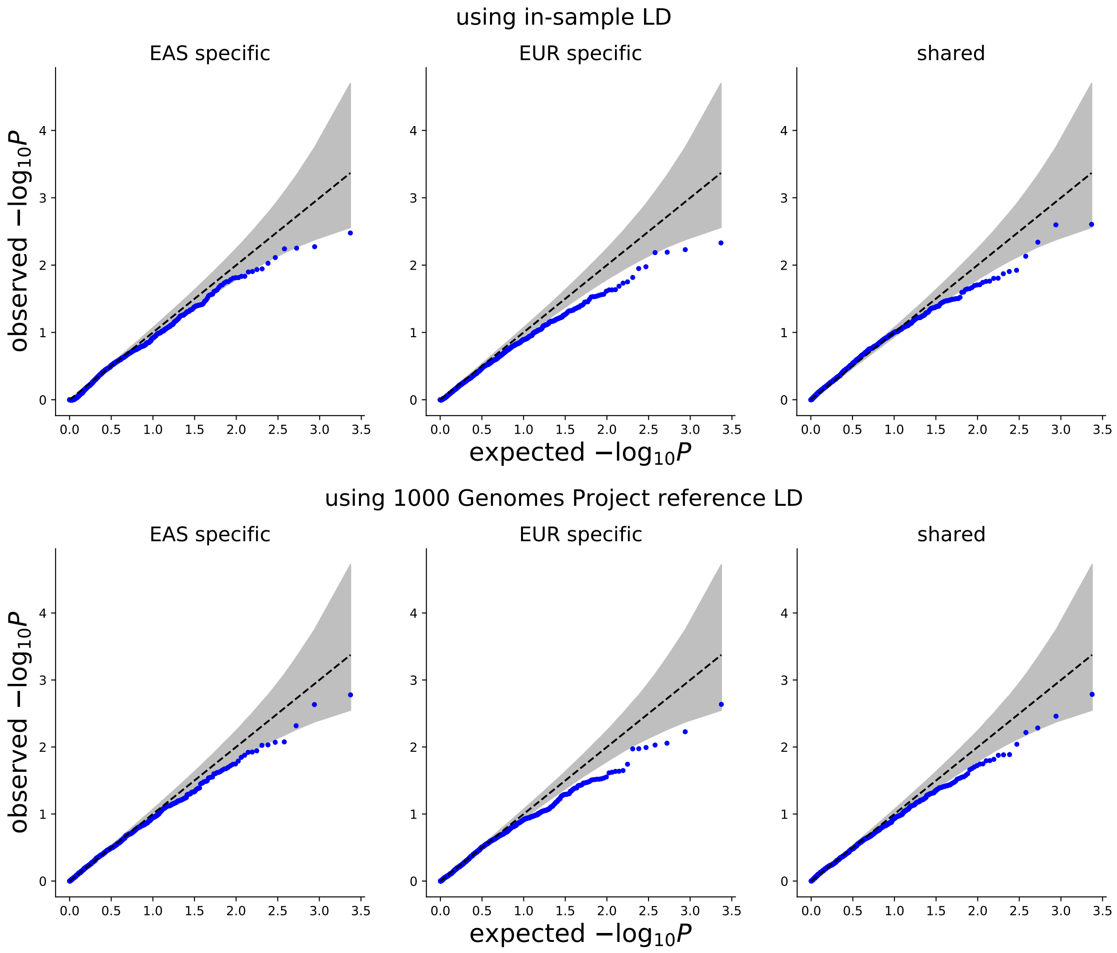


Figure S11: **Q-Q plot of p-values of the enrichments of population-specific/shared causal variants in SEG annotations[1] obtained using in-sample LD (top row) or ancestry-matched 1000 Genomes LD (bottom row).** We computed p-values from the enrichment test statistics of SEG annotations in 53 GTEx tissues from 25 null simulations, where we drew 25 EAS-specific, 25 EUR-specific, and 75 shared causal variants at random. In all simulations, we set $N \times h_g^2 = 500$ in both populations. Columns correspond to enrichment test statistics for the number of EAS-specific (left), EUR-specific (middle), or shared (right) causal variants.

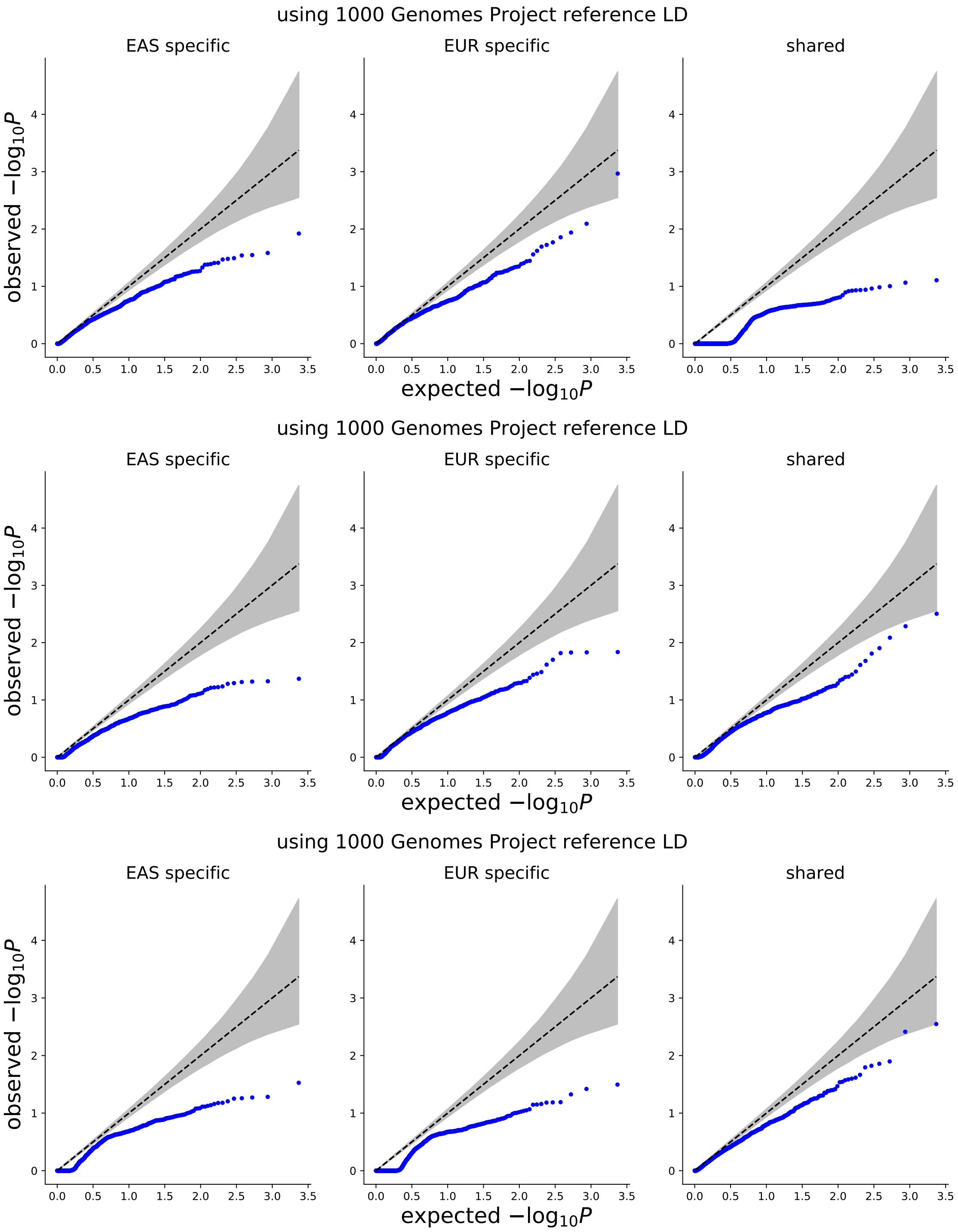


Figure S12: **Q-Q plot of p-values of the enrichments of population-specific/shared causal variants in SEG annotations[1] (20 causal variants per population).** We computed p-values for SEG annotations across 53 GTEx tissues from 25 null simulations, where we drew 20 causal variants at random for each population. In all simulations, we set $N \times h_g^2 = 500$ in both populations. The top, middle, and bottom rows represent results from simulations where 0% (top), 50% (middle), and 100% (bottom) of the causal SNPs were shared. All results were obtained using 1000 Genomes Project reference LD.

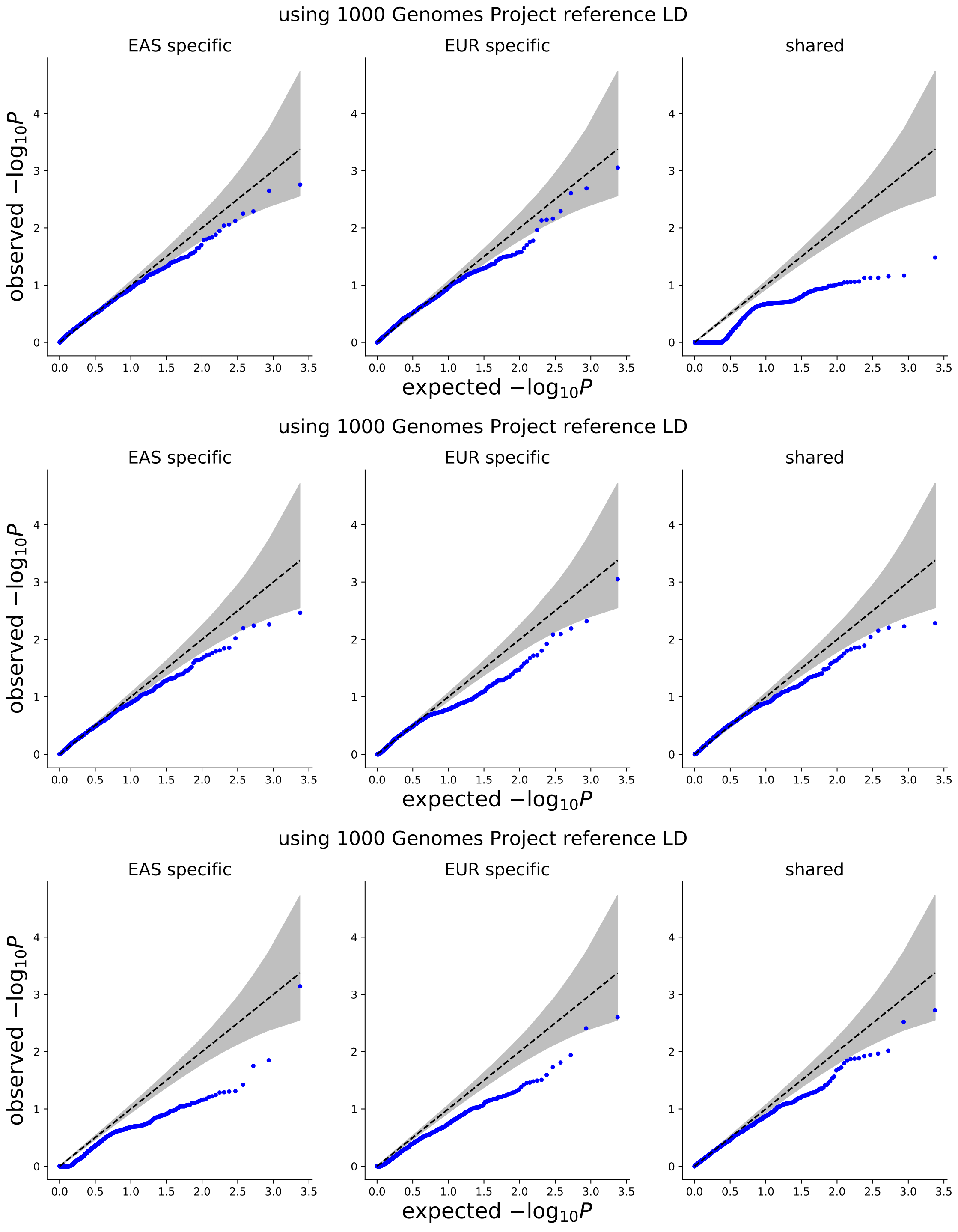


Figure S13: **Q-Q plot of p-values of the enrichments of population-specific/shared causal variants in SEG annotations[1] (60 causal variants per population).** We computed p-values for SEG annotations across 53 GTEx tissues from 25 null simulations, where we drew 60 causal variants at random for each population. In all simulations, we set $N \times h_g^2 = 500$ in both populations. The top, middle, and bottom rows represent results from simulations where 0% (top), 50% (middle), and 100% (bottom) of the causal SNPs were shared. All results were obtained using 1000 Genomes Project reference LD.

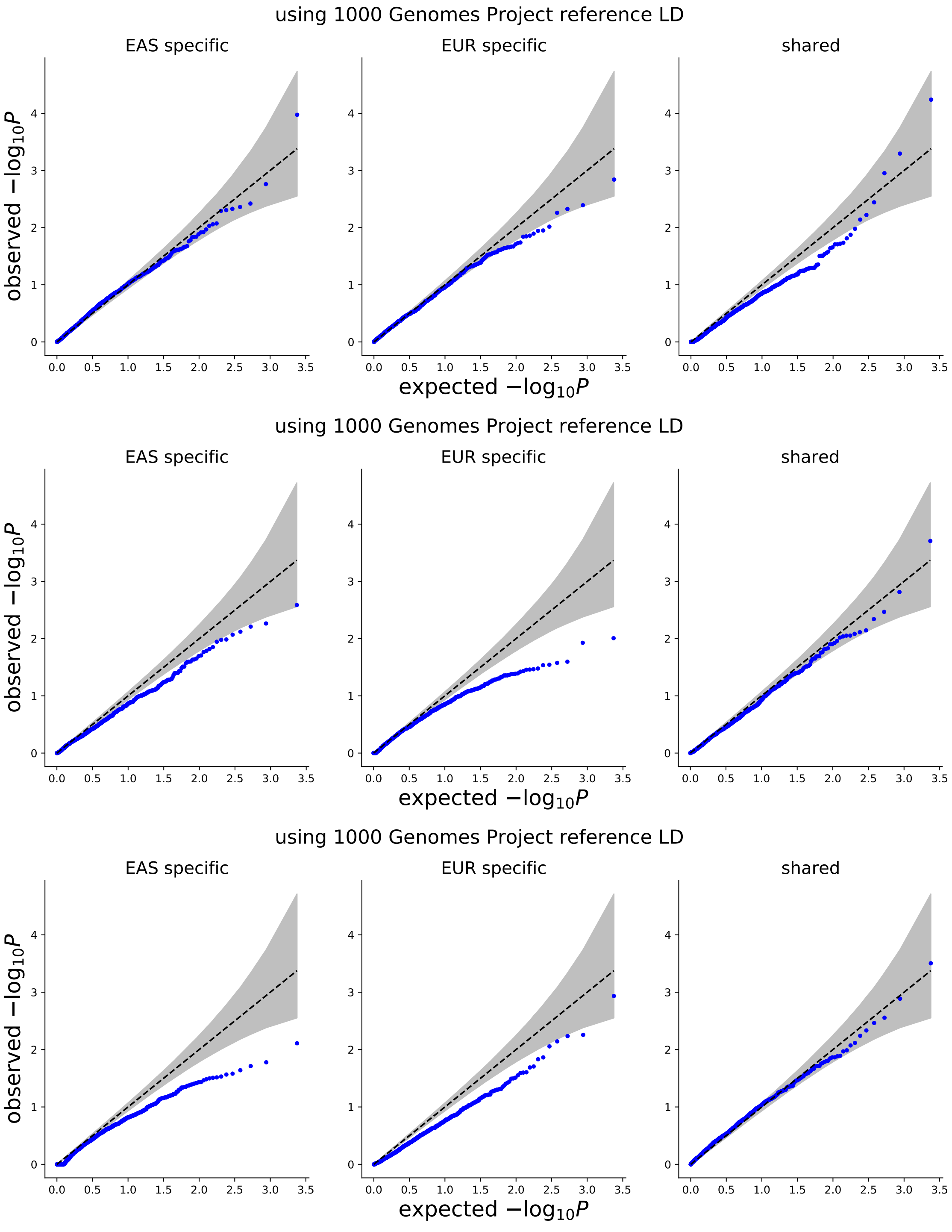


Figure S14: **Q-Q plot of p-values of the enrichments of population-specific/shared causal variants in SEG annotations[1] (100 causal variants per population).** We computed p-values for SEG annotations across 53 GTEx tissues from 25 null simulations, where we drew 100 causal variants at random for each population. In all simulations, we set $N \times h_g^2 = 500$ in both populations. The top, middle, and bottom rows represent results from simulations where 0% (top), 50% (middle), and 100% (bottom) of the causal SNPs were shared. All results were obtained using 1000 Genomes Project reference LD.

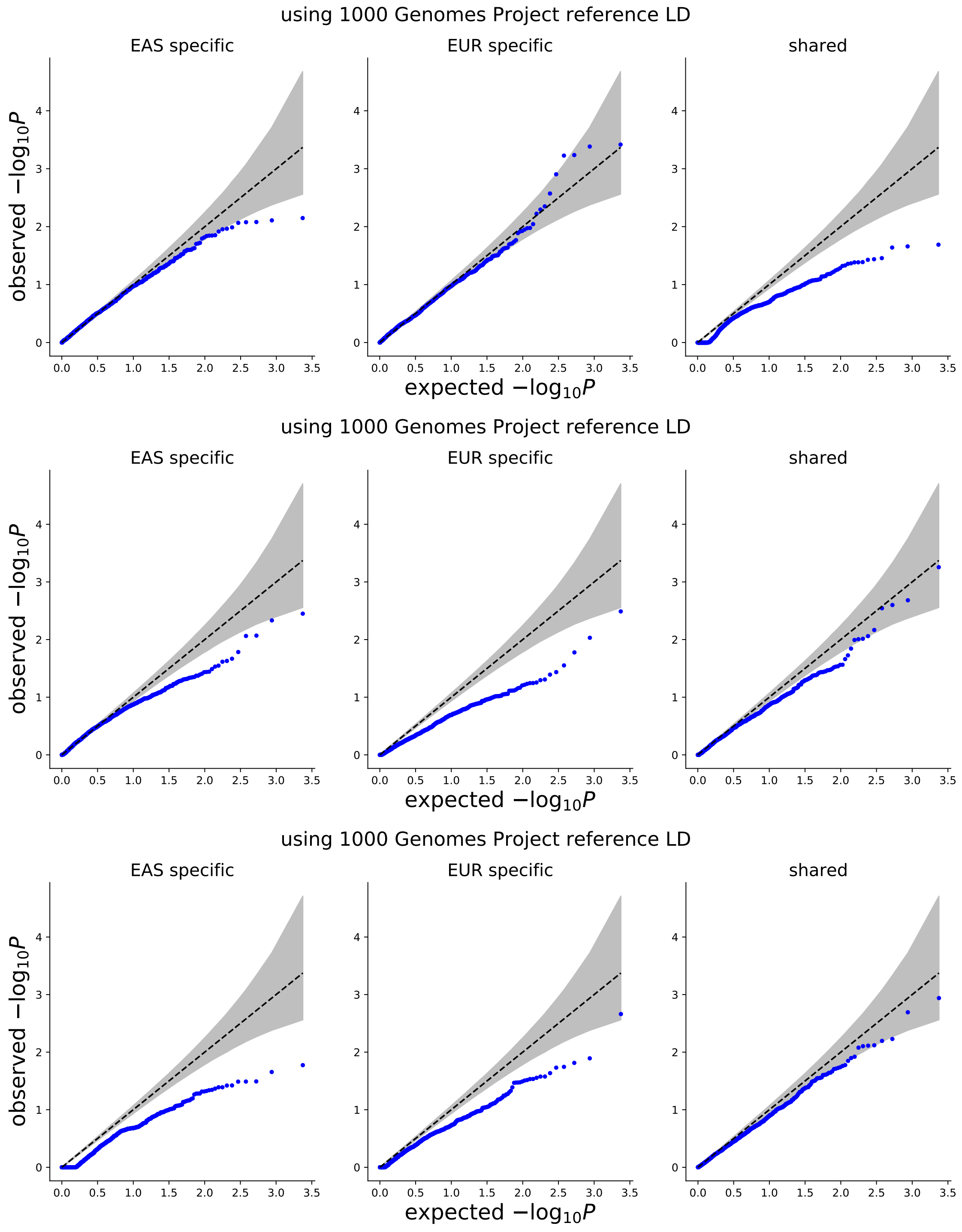


Figure S15: **Q-Q plot of p-values of the enrichments of population-specific/shared causal variants in SEG annotations**[1] **($N \times h_g^2 = 375$, 60 causal variants per population).** We computed p-values for SEG annotations across 53 GTEx tissues from 25 null simulations, where we drew 60 causal variants at random for each population. In all simulations, we set $N \times h_g^2 = 375$ in both populations. The top, middle, and bottom rows represent results from simulations where 0% (left), 50% (middle), and 100% (right) of the causal SNPs were shared. All results were obtained using 1000 Genomes Project reference LD.

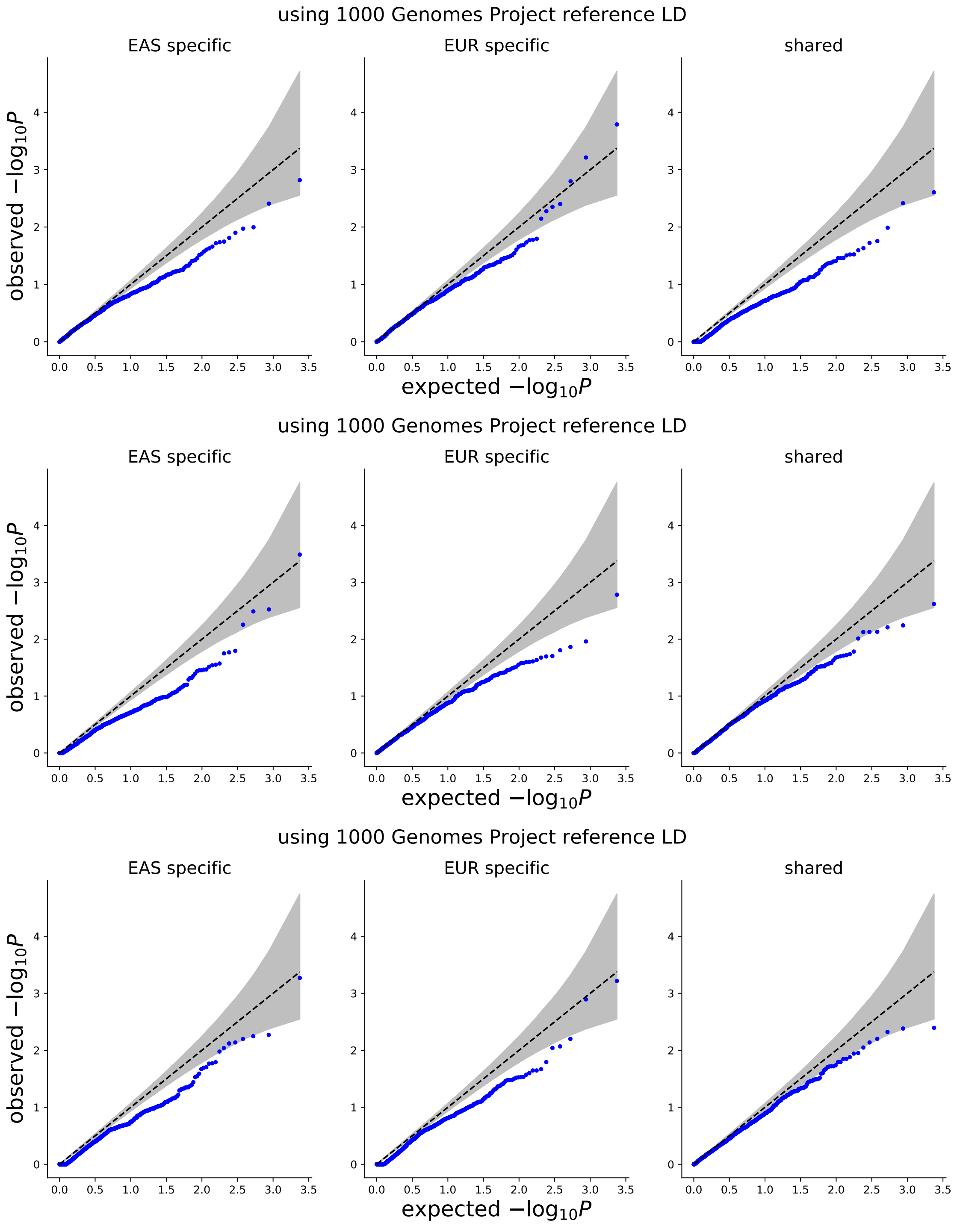


Figure S16: **Q-Q plot of p-values of the enrichments of population-specific/shared causal variants in SEG annotations**[1] (**$N \times h_g^2 = 250$, 60 causal variants per population)** We computed p-values for SEG annotations across 53 GTEx tissues from 25 null simulations, where we drew 60 causal variants at random for each population. In all simulations, we set $N \times h_g^2 = 250$ in both populations. The top, middle, and bottom rows represent results from simulations where 0% (left), 50% (middle), and 100% (right) of the causal SNPs were shared. All results were obtained using 1000 Genomes Project reference LD.

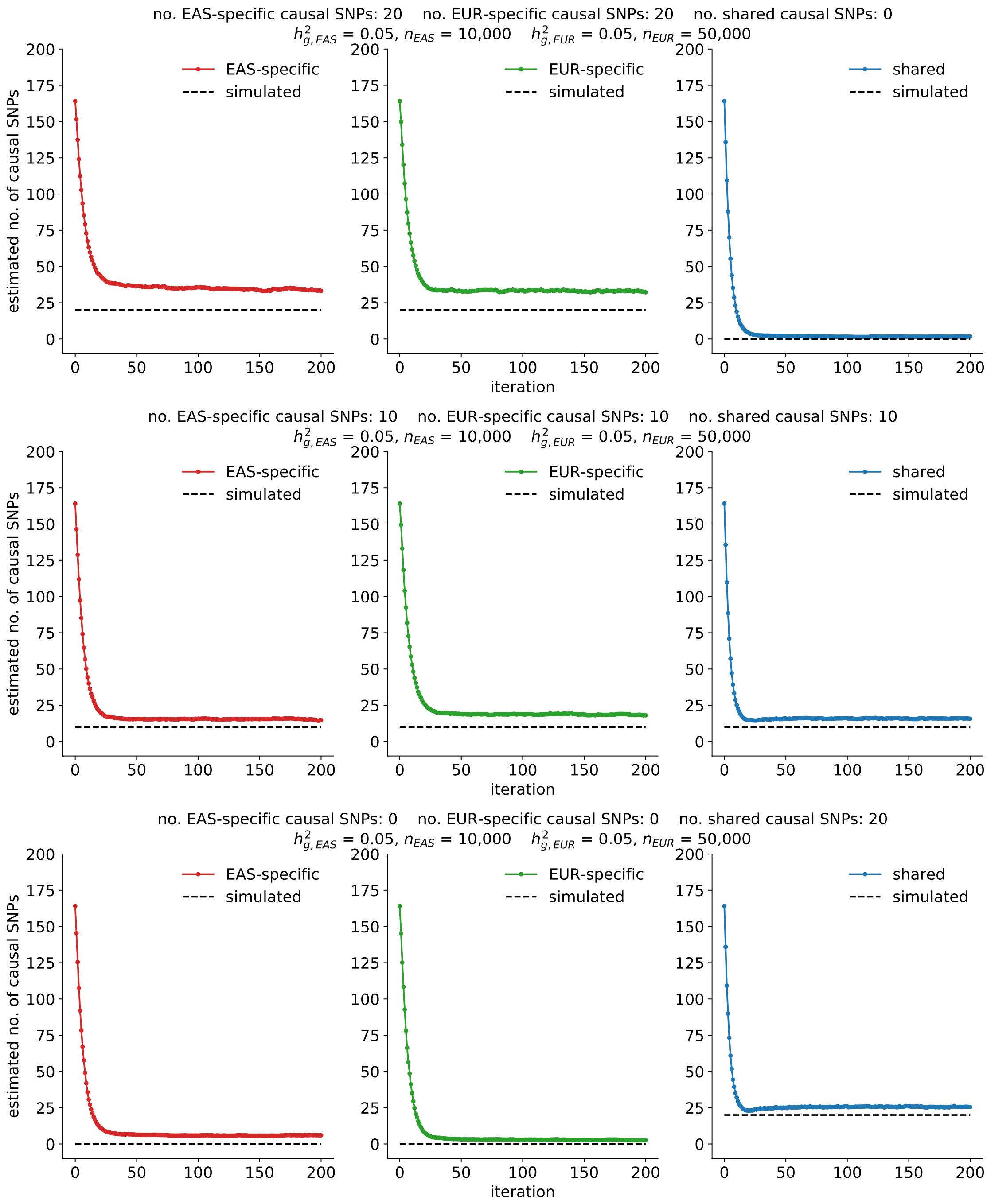


Figure S17: **Estimated numbers of population-specific/shared causal SNPs across iterations of the EM algorithm (20 causal SNPs per population).** We randomly selected 20 causal SNPs on chr22 (out of 8,599) in both populations where either 0% (top), 50% (middle) or 100% (bottom) were shared causal SNPs. $N \times h_g^2 = 500$ for both populations. Each curve represents the average estimate across 25 simulations.

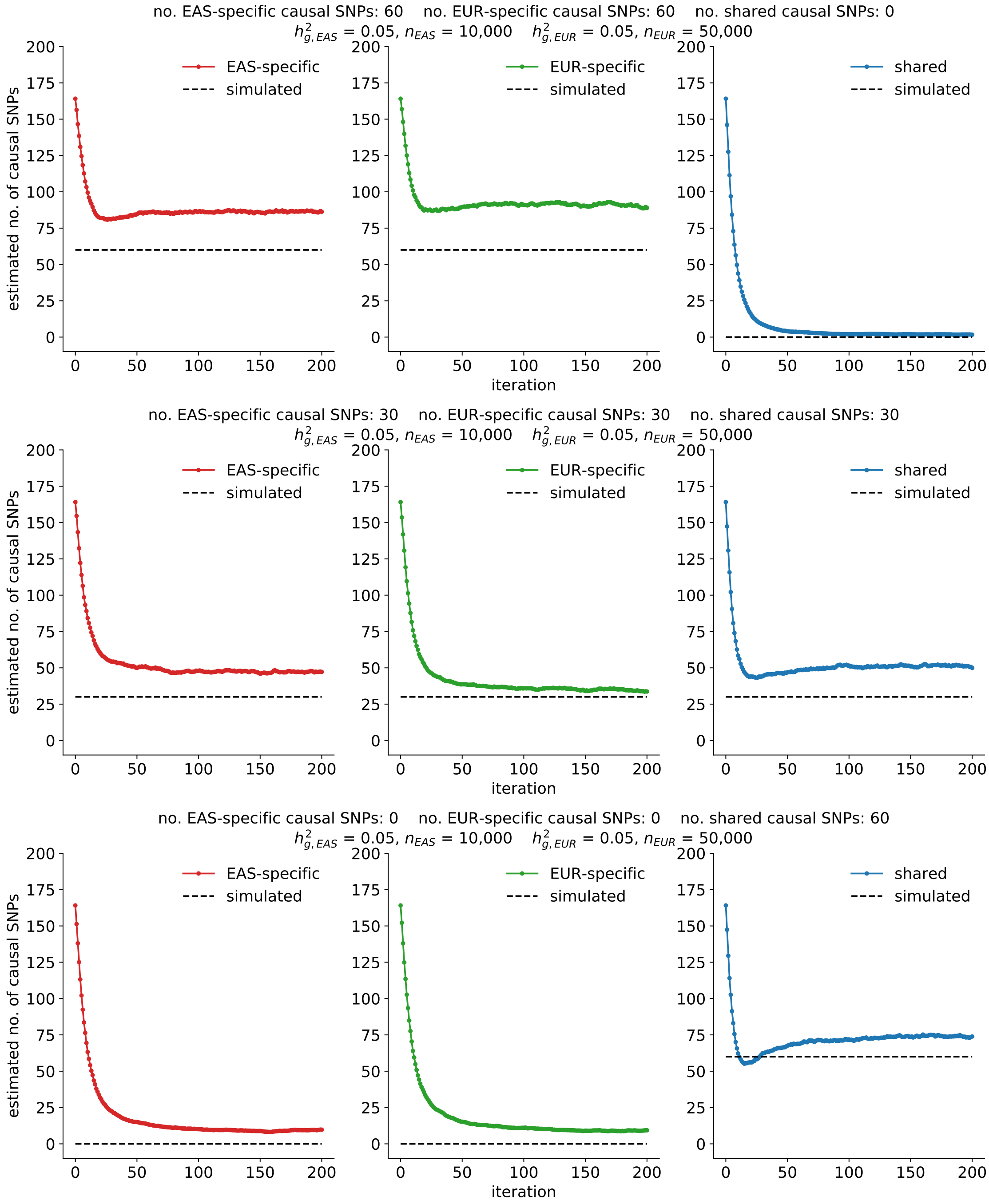


Figure S18: **Estimated number of population-specific and shared causal variants across iterations of the EM algorithm (60 causal SNPs per population).** We randomly selected 60 causal SNPs (out of 8,599) in both populations where either 0% (top), 50% (middle) or 100% (bottom) were shared causal SNPs. $N \times h_g^2 = 500$ for both populations. Each curve represents the average across 25 simulations.

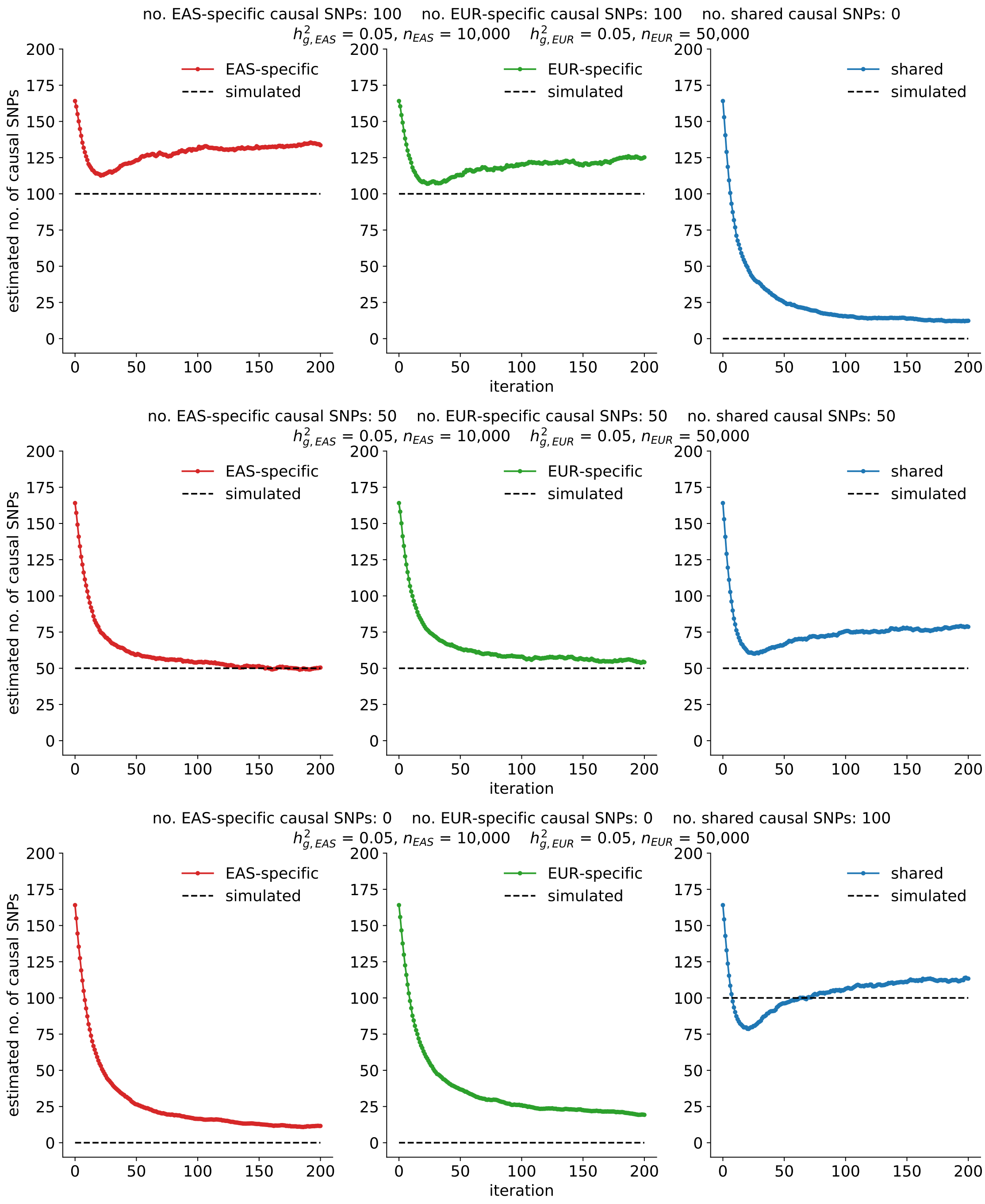


Figure S19: **Estimated number of population-specific and shared causal variants across iterations of the EM algorithm (100 causal SNPs per population).** We randomly selected 100 causal SNPs (out of 8,599) in both populations where either 0% (top), 50% (middle) or 100% (bottom) were shared causal SNPs. $N \times h_g^2 = 500$ for both populations. Each curve represents the average across 25 simulations.

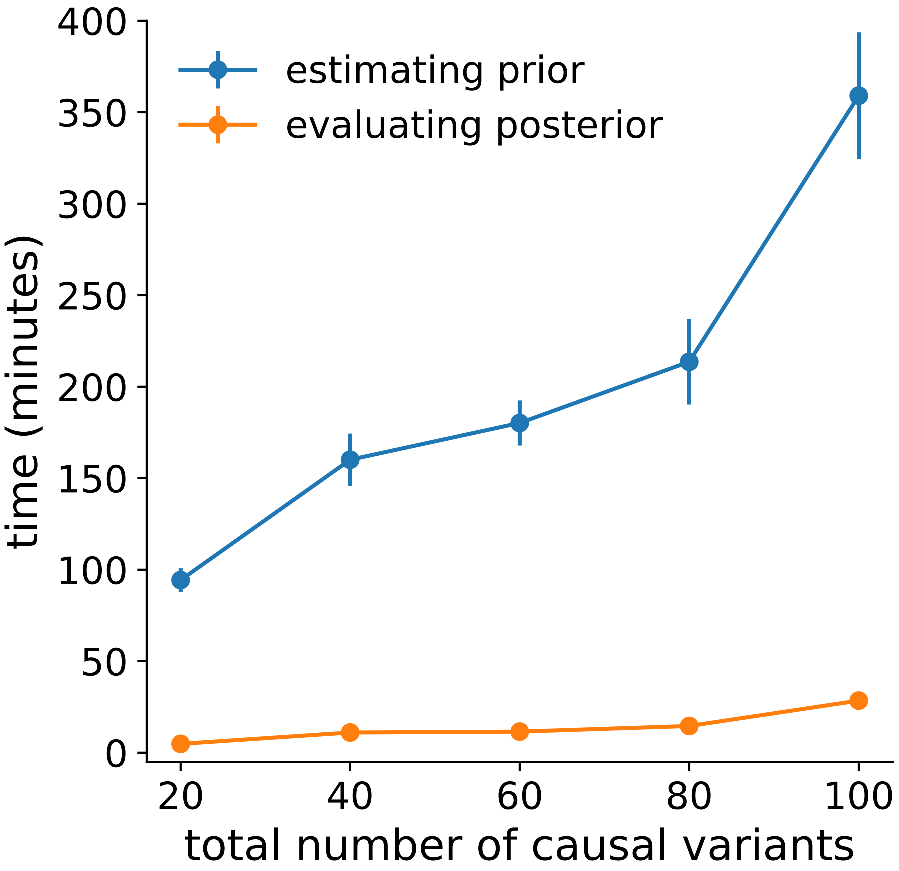


Figure S20: **Average run-times for estimating the prior (MVB parameters) and evaluating the per-SNP posterior probabilities of being causal in one or both populations.** Each dot represents the average run-time across 25 simulations; the total number of causal variants per population is specified on the x-axis. $N \times h_g^2 = 500$ for both populations. Error bars represent $\pm$1.96 s.e.m.

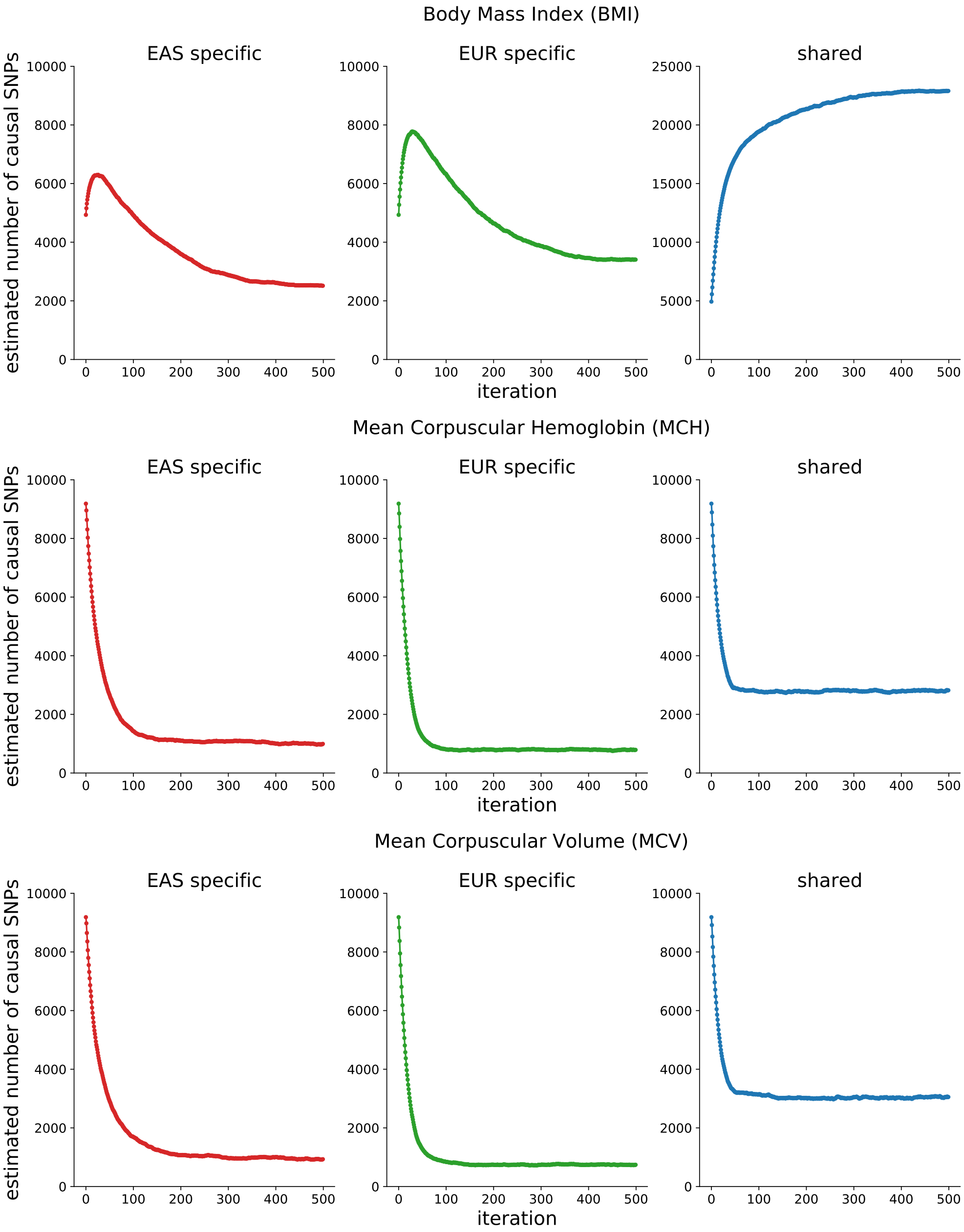


Figure S21: **Estimated numbers of population-specific/shared causal variants across EM iterations for BMI, MCH, and MCV.**

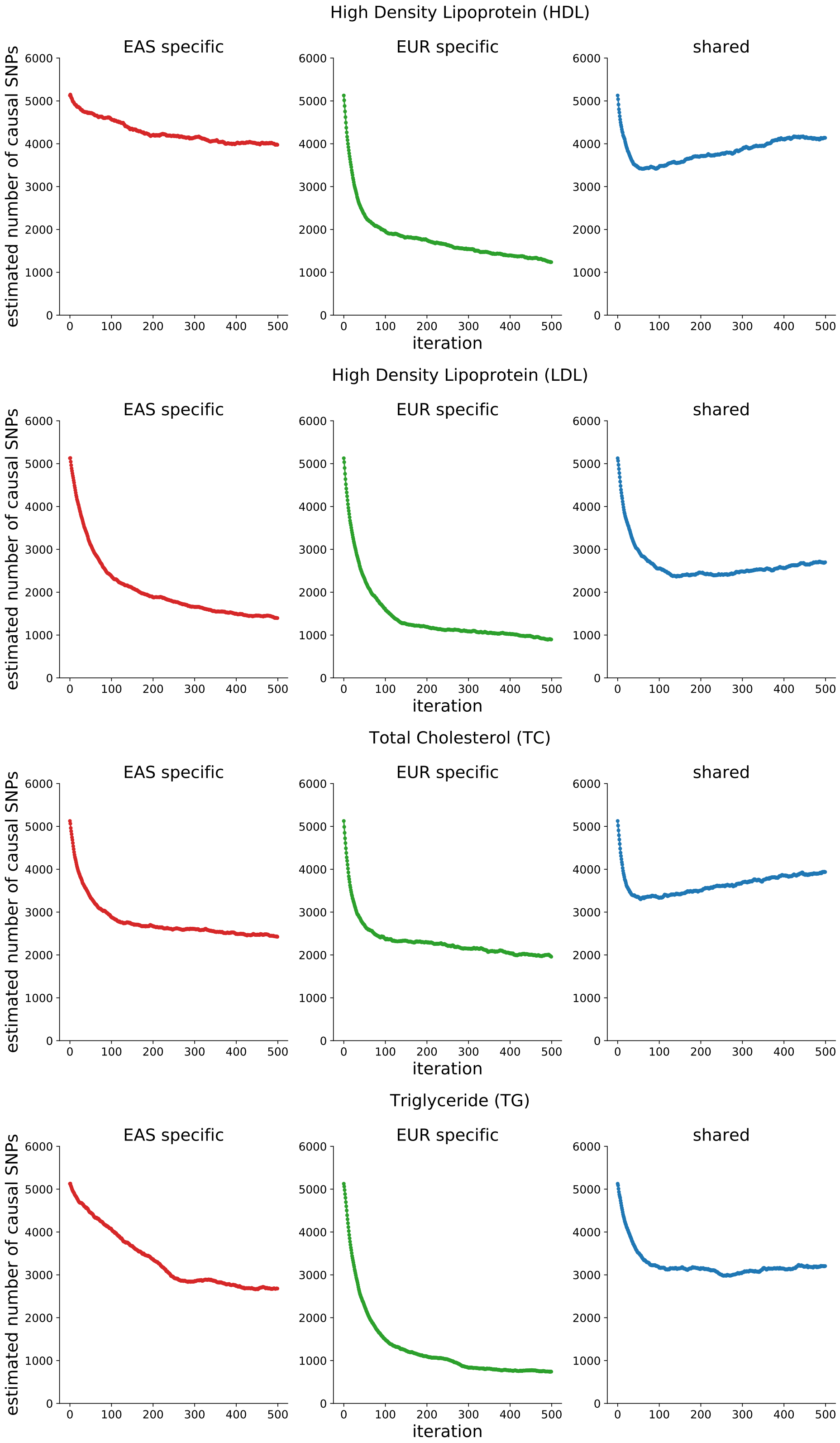


Figure S22: **Estimated numbers of population-specific/shared causal variants across EM iterations for HDL, LDL, TC, and TG.**

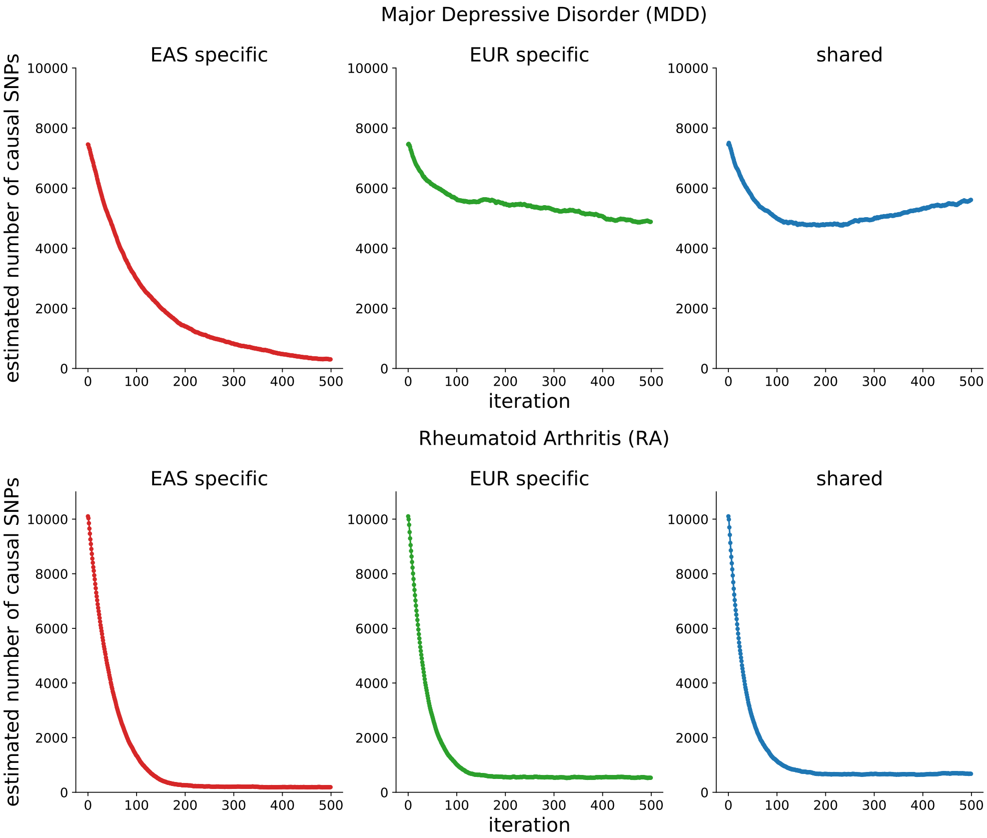


Figure S23: **Estimated numbers of population-specific/shared causal variants across EM iterations for MDD and RA.**

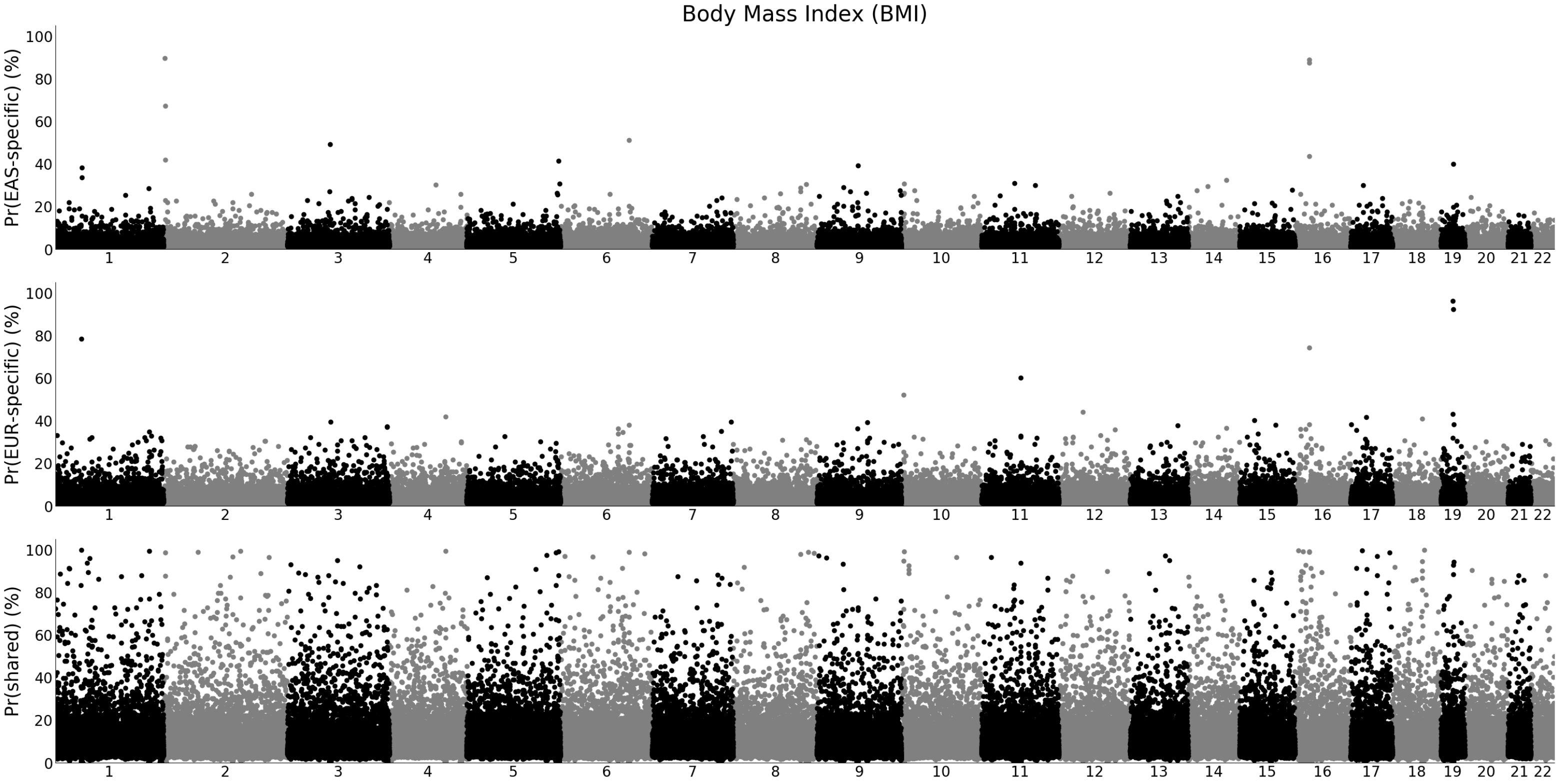
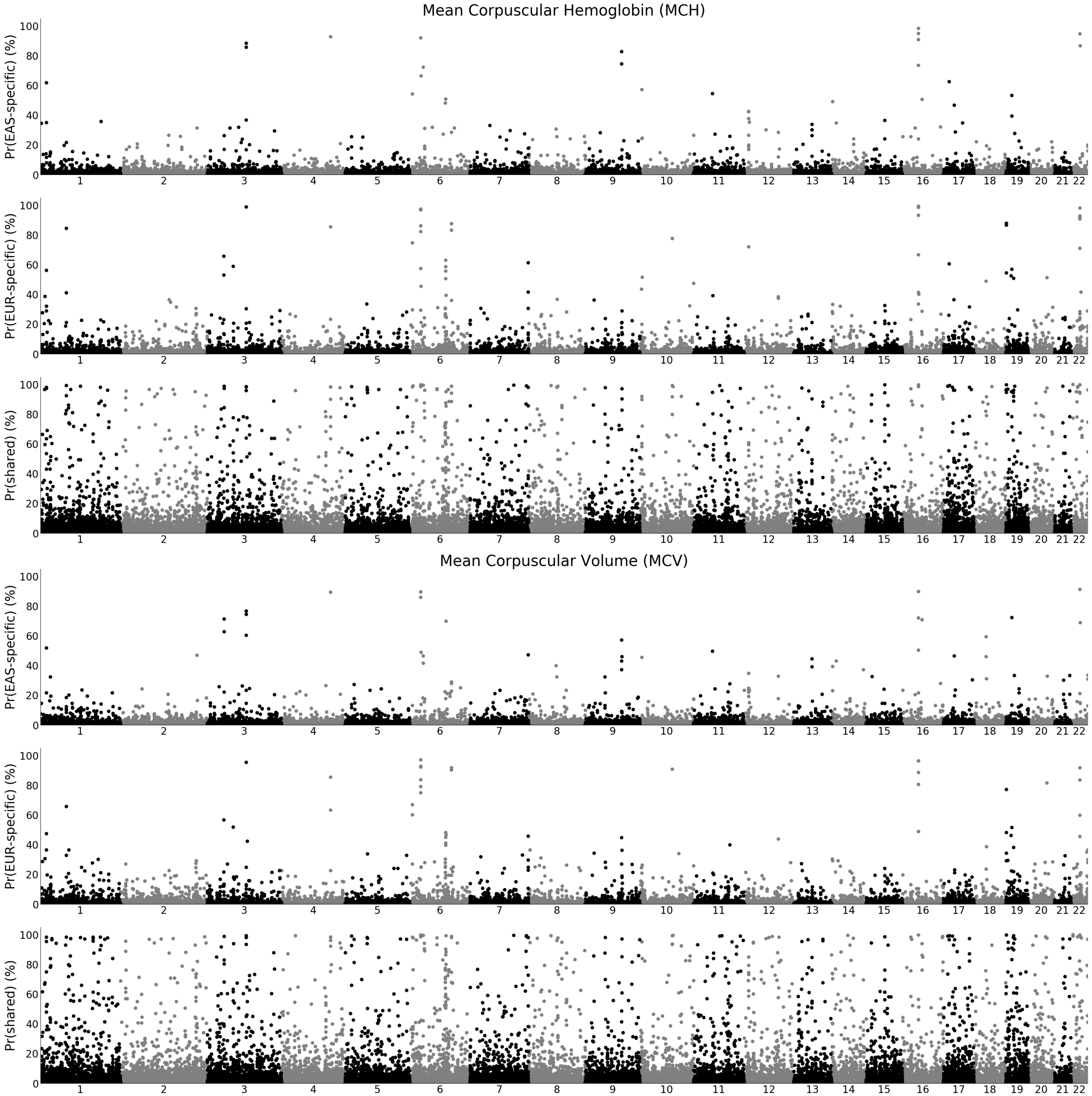


Figure S24: **Manhattan-style plots for posterior probability of each SNP to population-specific or shared for BMI.**


Mean Corpuscular Hemoglobin (MCH)

Mean Corpuscular Volume (MCV)

Figure S25: **Manhattan-style plots for posterior probability of each SNP to population-specific or shared for MCH and MCV.**

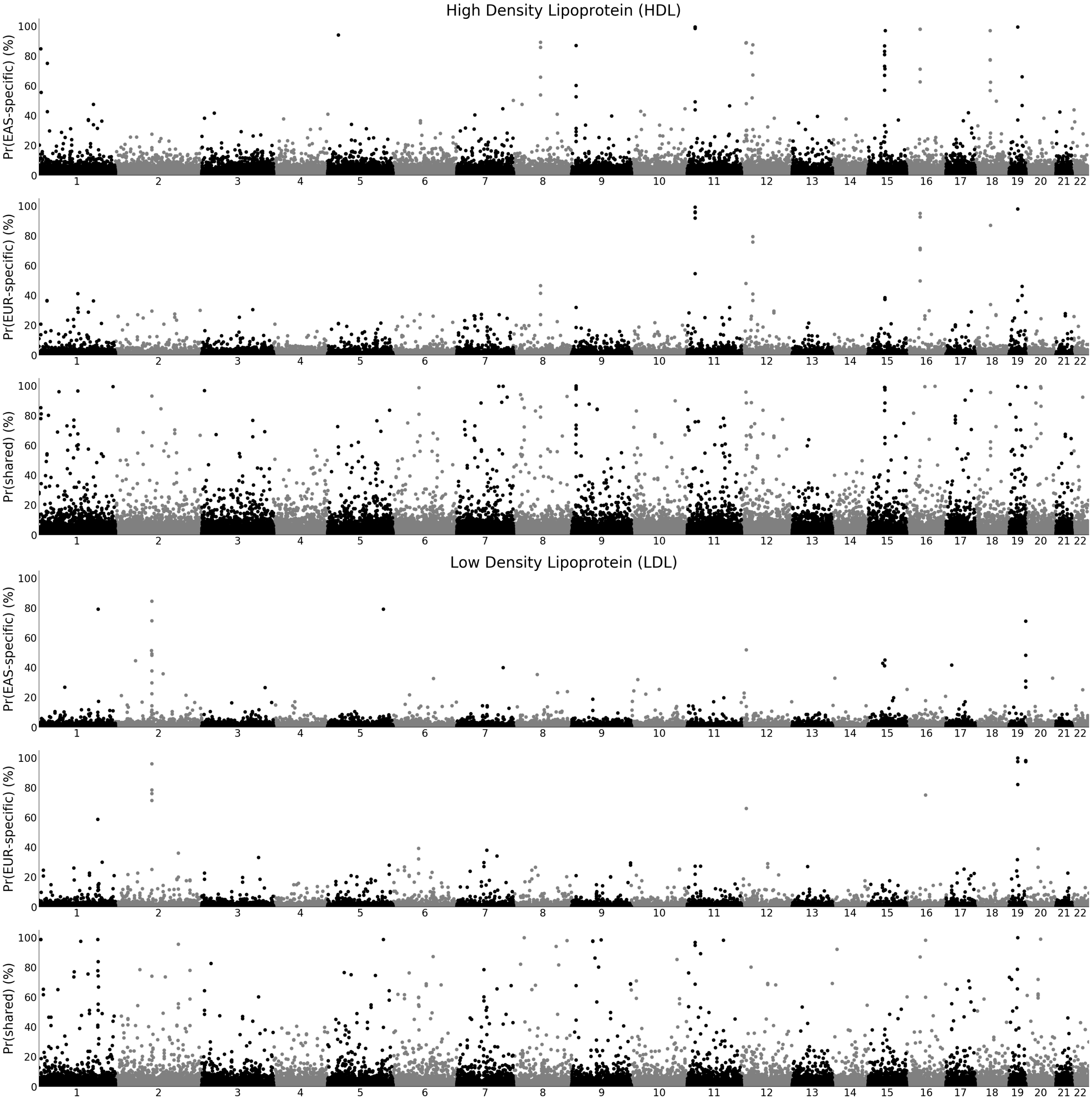


Figure S26: **Manhattan-style plots for posterior probability of each SNP to population-specific or shared for HDL and LDL.**

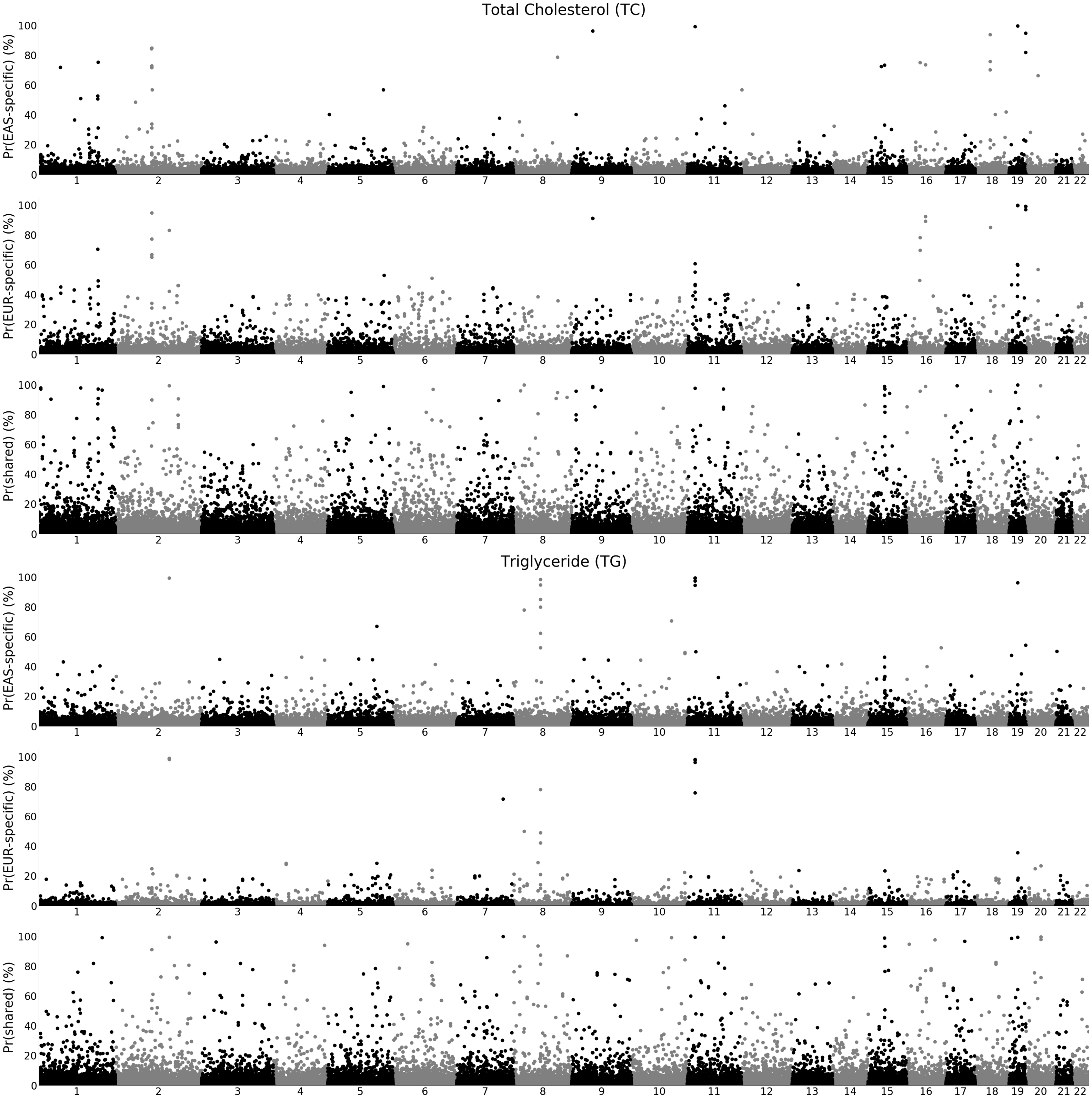


Figure S27: **Manhattan-style plots for posterior probability of each SNP to population-specific or shared for TC and TG.**

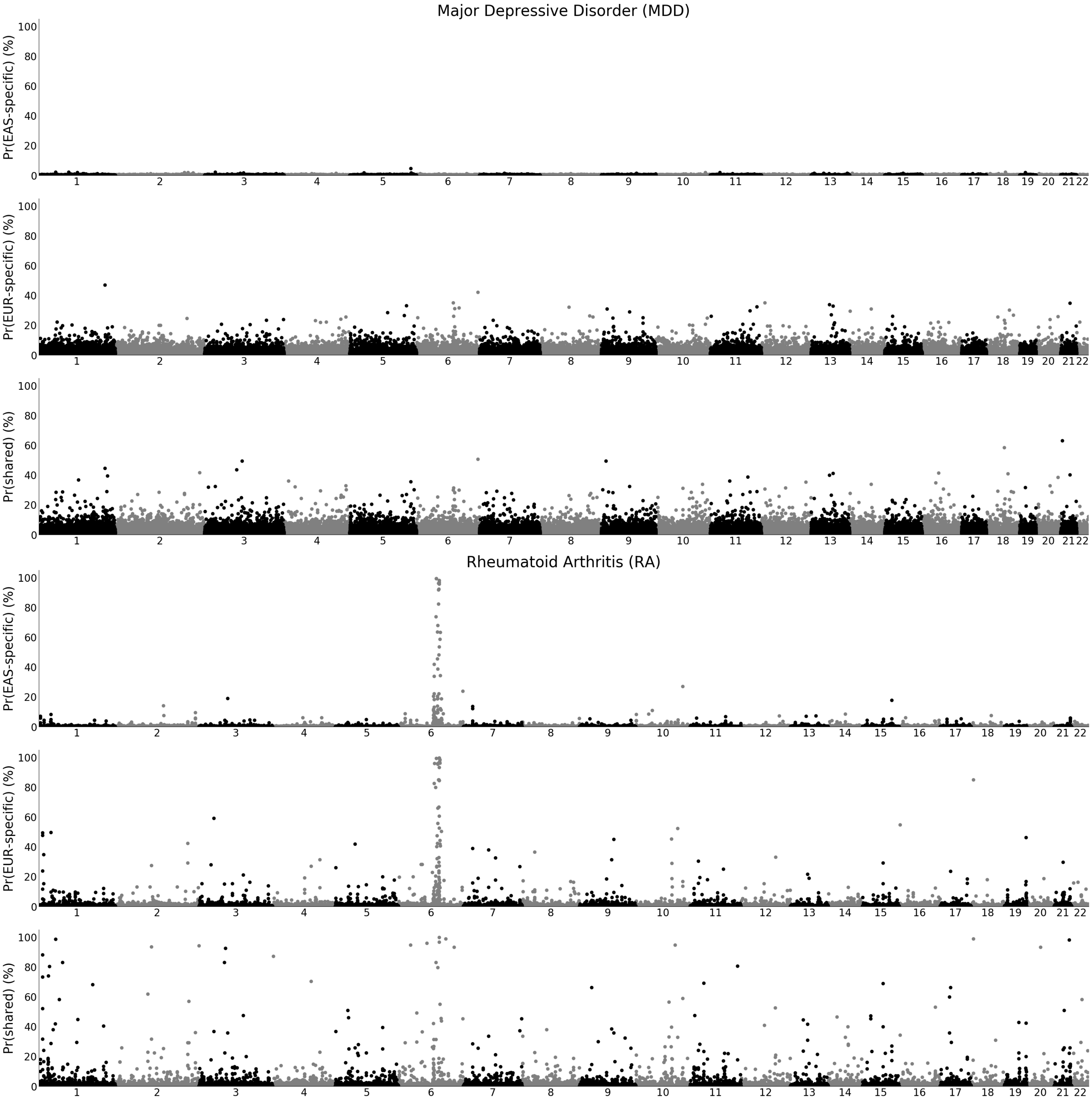
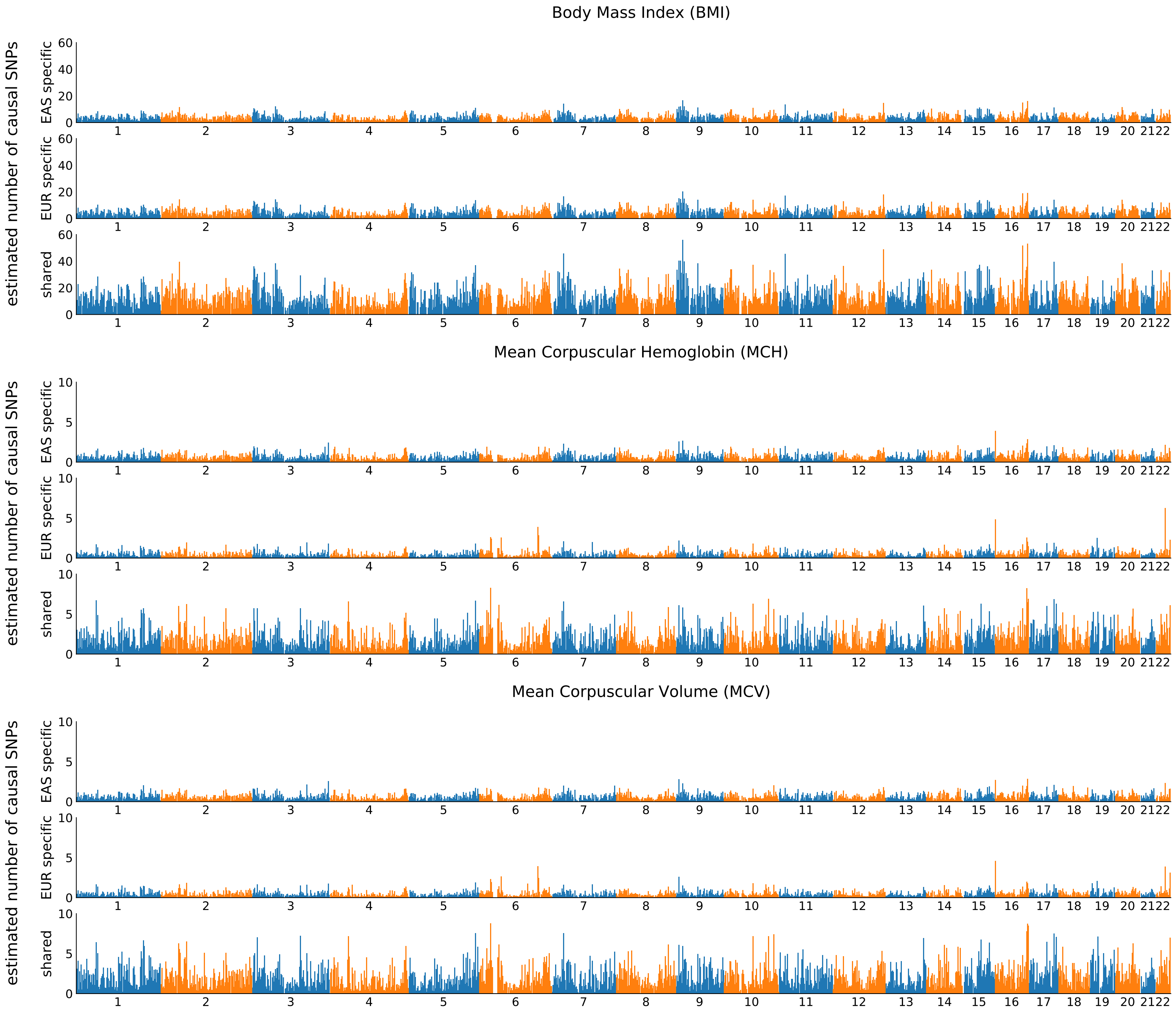


Figure S28: **Manhattan-style plots for posterior probability of each SNP to population-specific or shared for MDD and RA.**

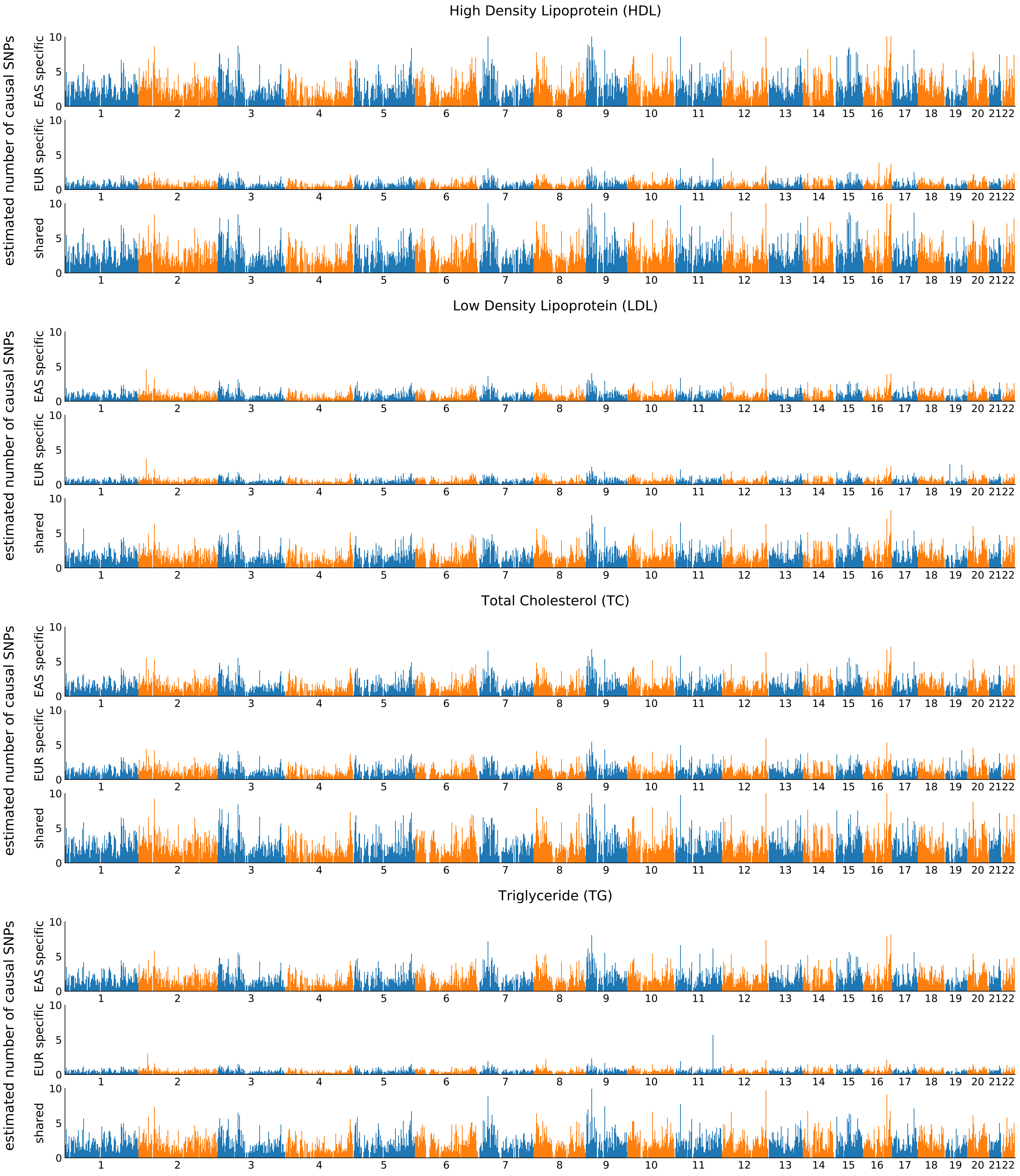


Body Mass Index (BMI)

Mean Corpuscular Hemoglobin (MCH)

Mean Corpuscular Volume (MCV)

Figure S29: **Regional number of causal variants for BMI, MCH, and MCV.**


High Density Lipoprotein (HDL)

Low Density Lipoprotein (LDL)

Total Cholesterol (TC)

Triglyceride (TG)

Figure S30: **Regional number of causal variants for HDL, LDL, TC, and TG.**

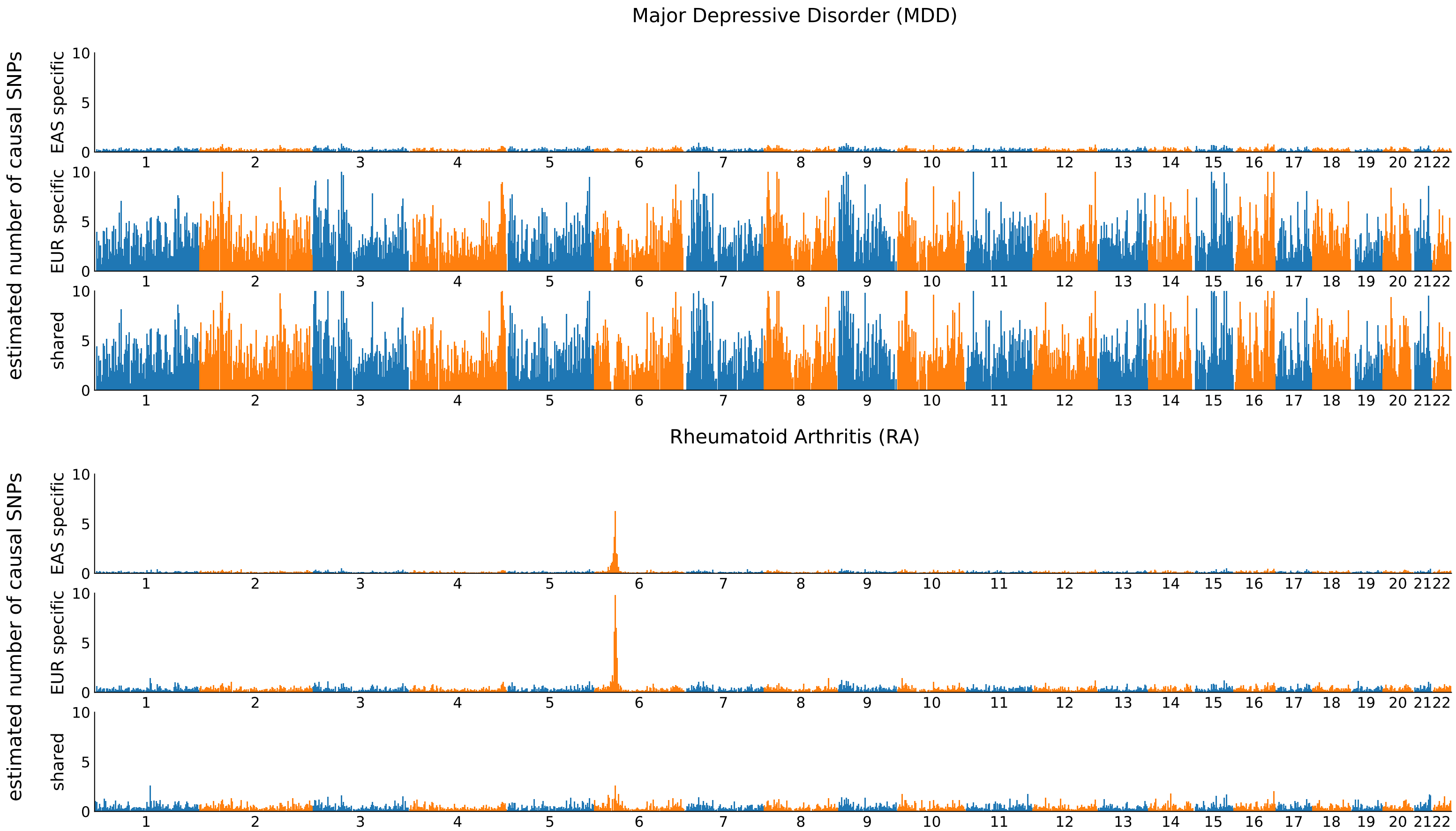


Figure S31: **Regional number of causal variants for MDD and RA.**

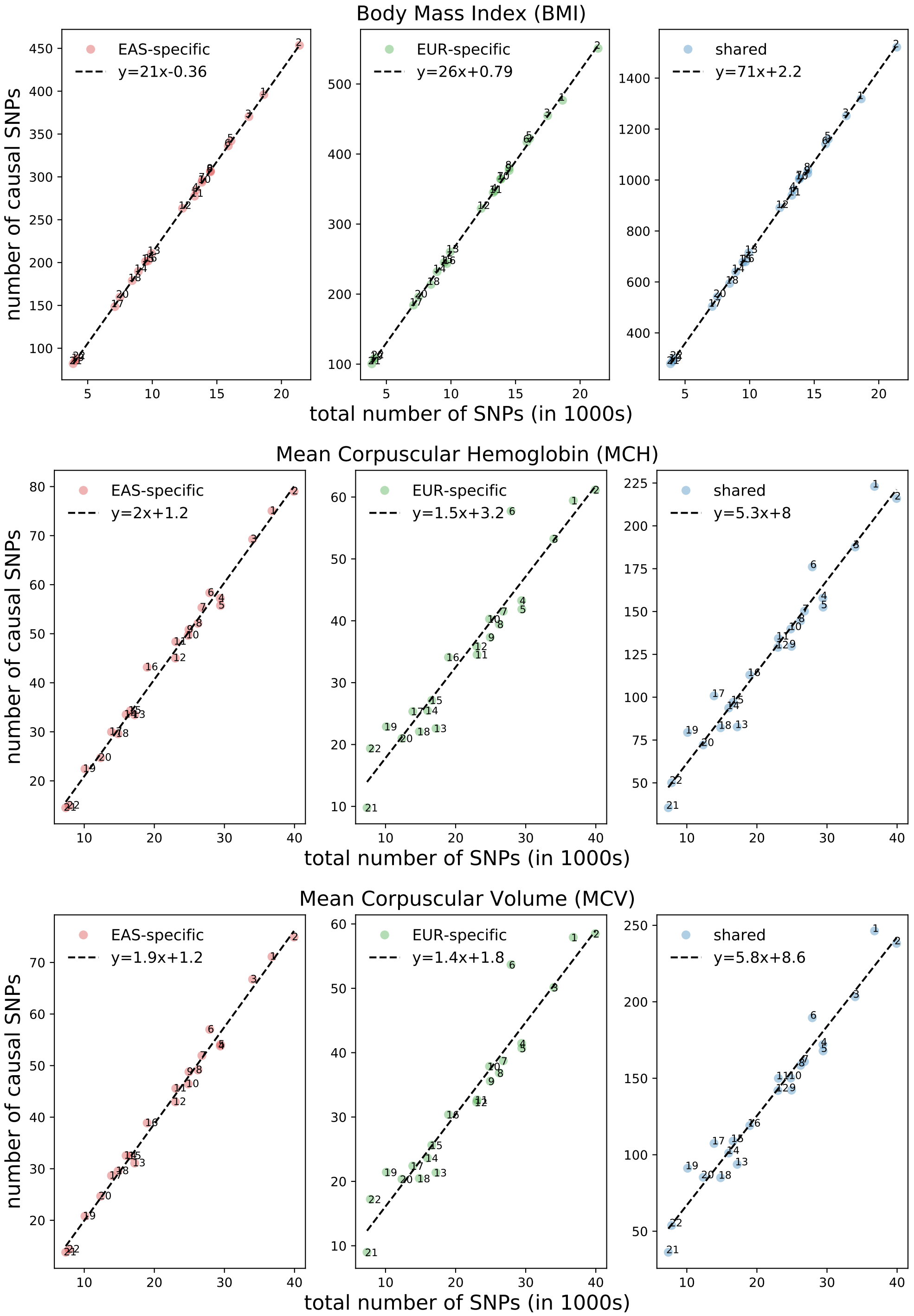


Figure S32: **Chromosomal number of causal variants for BMI, MCH, and MCV.**

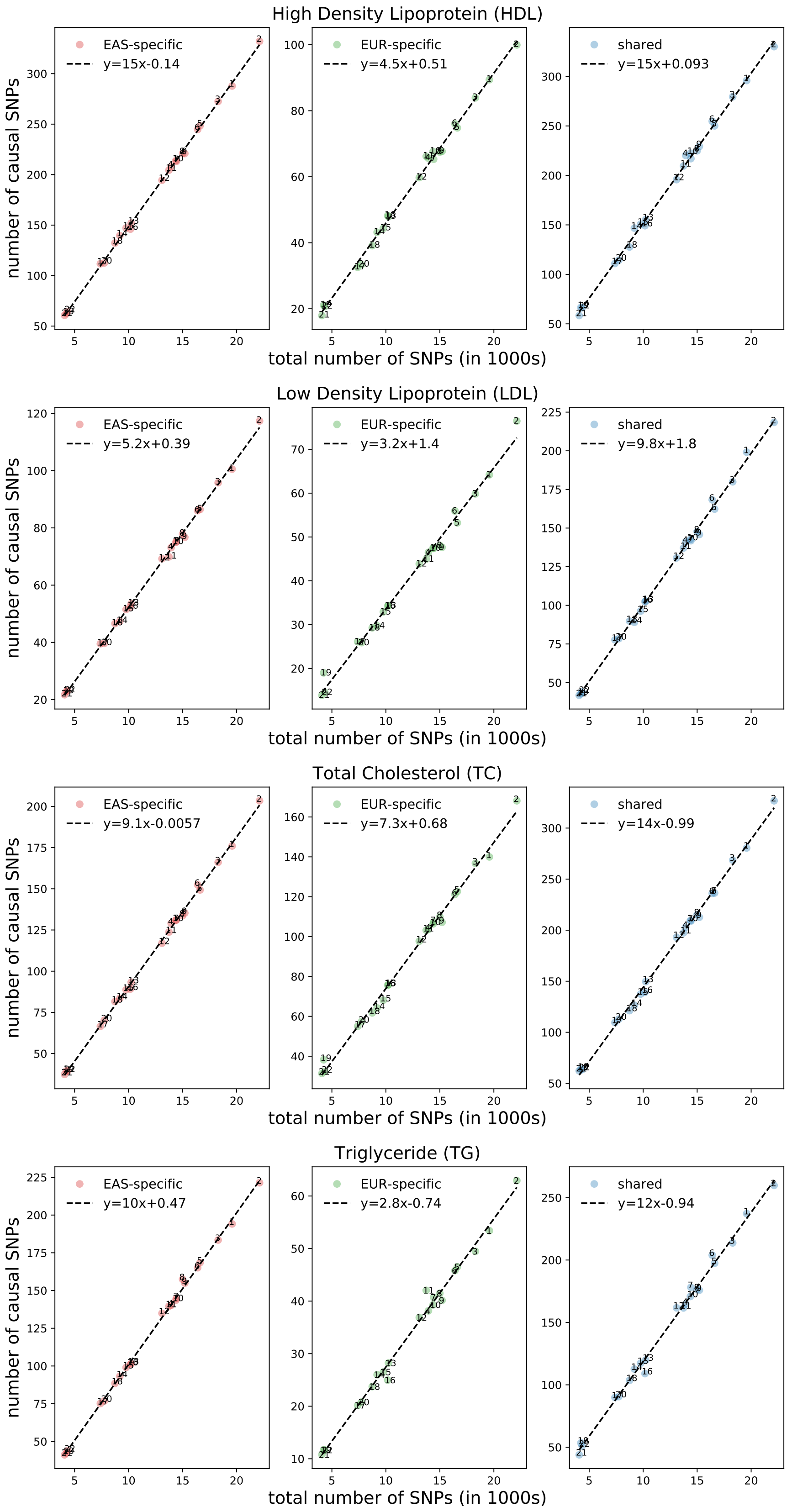


Figure S33: **Chromosomal number of causal variants for HDL, LDL, TC, and TG.**

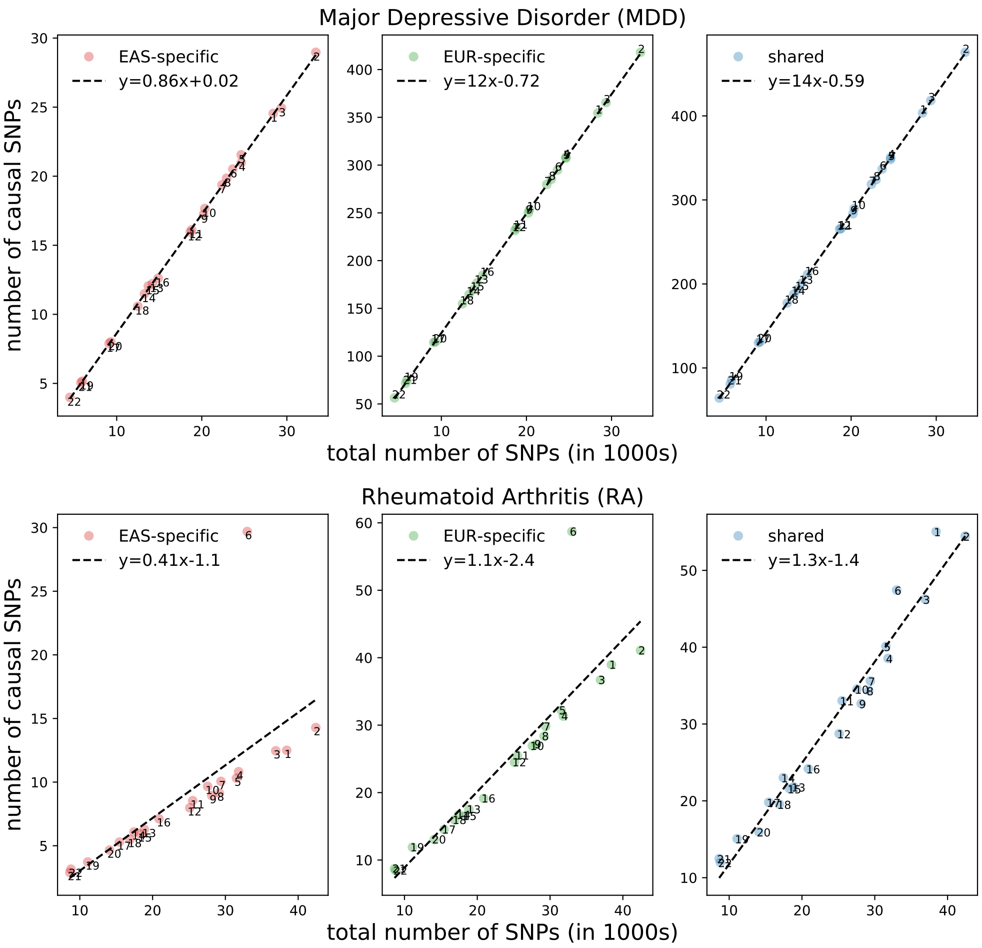


Figure S34: **Chromosomal number of causal variants for MDD and RA.**

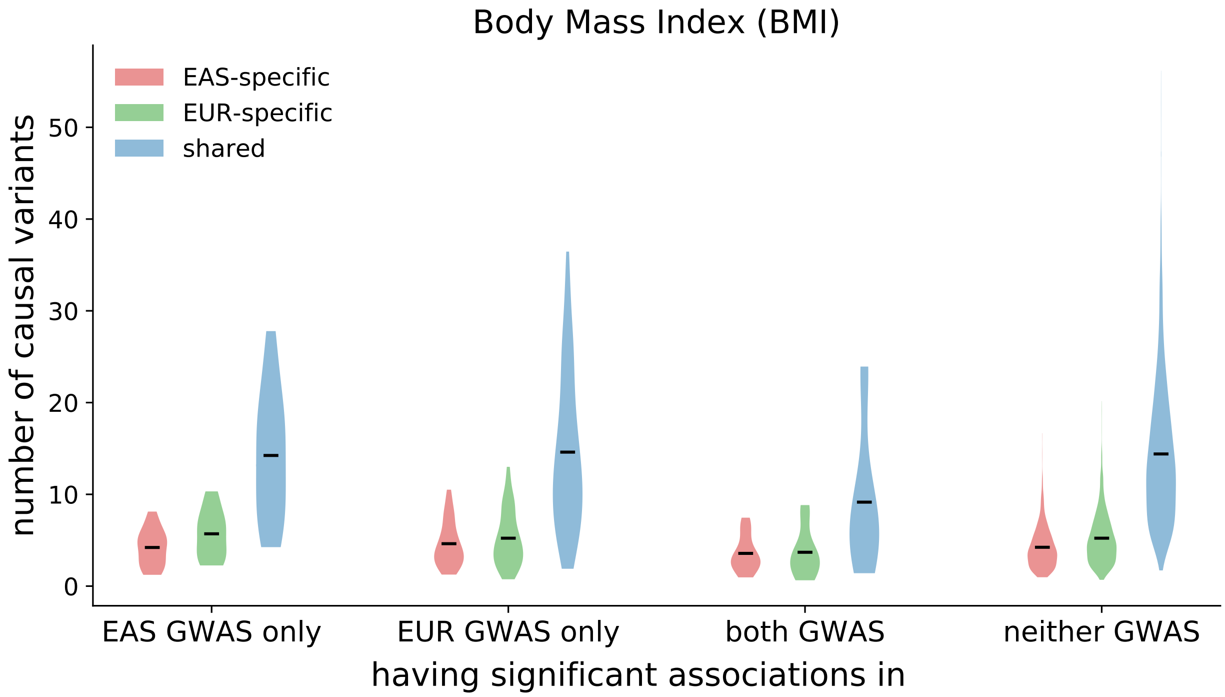


Figure S35: **Distribution of regional number of causal variants at GWAS risk regions.** Each violin plot shows the distribution of population-specific or shared causal variants at regions harboring significant associations ($p < 5 \times 10^{-5}$) in the East Asian GWAS only, in the European GWAS only, in both GWASs, and in neither GWAS. The dark line represents the mean of the distribution.

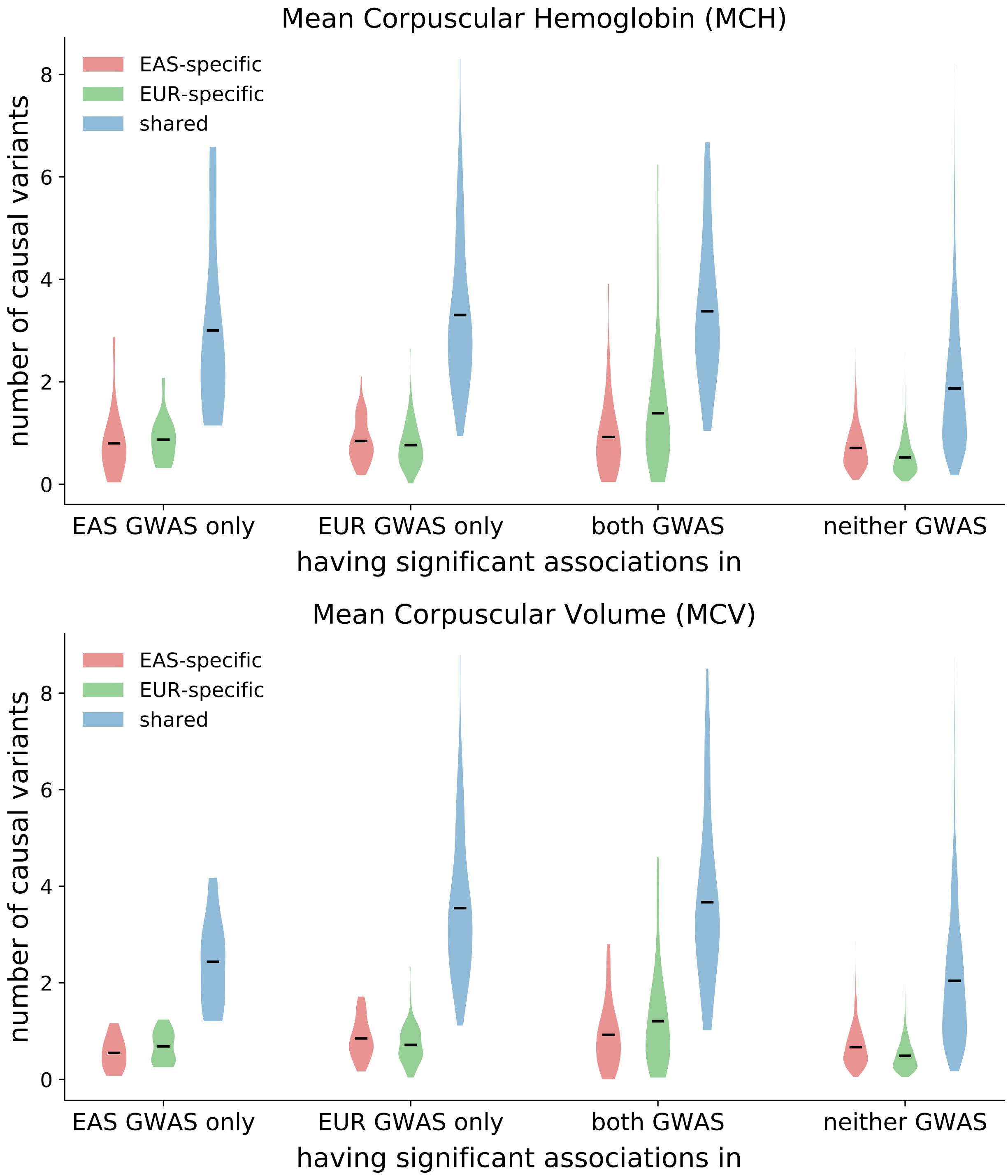
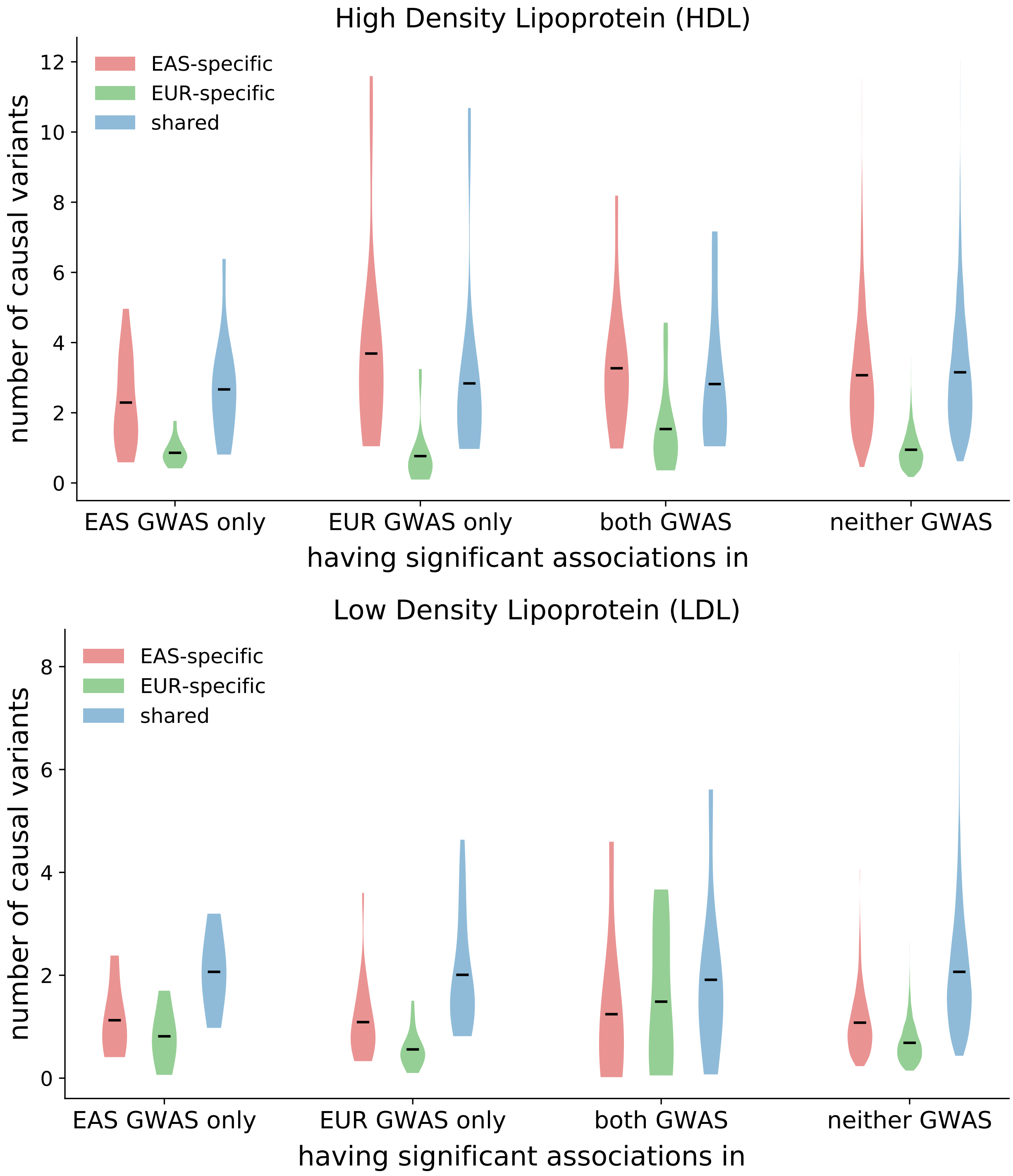


Figure S36: **Distribution of regional number of causal variants at GWAS risk regions.** Each violin plot shows the distribution of population-specific or shared causal variants at regions harboring significant associations ($p < 5 \times 10^{-5}$) in the East Asian GWAS only, in the European GWAS only, in both GWASs, and in neither GWAS. The dark line represents the mean of the distribution.



High Density Lipoprotein (HDL)

Low Density Lipoprotein (LDL)

Figure S37: **Distribution of regional number of causal variants at GWAS risk regions.** Each violin plot shows the distribution of population-specific or shared causal variants at regions harboring significant associations ($p < 5 \times 10^{-5}$) in the East Asian GWAS only, in the European GWAS only, in both GWASs, and in neither GWAS. The dark line represents the mean of the distribution.

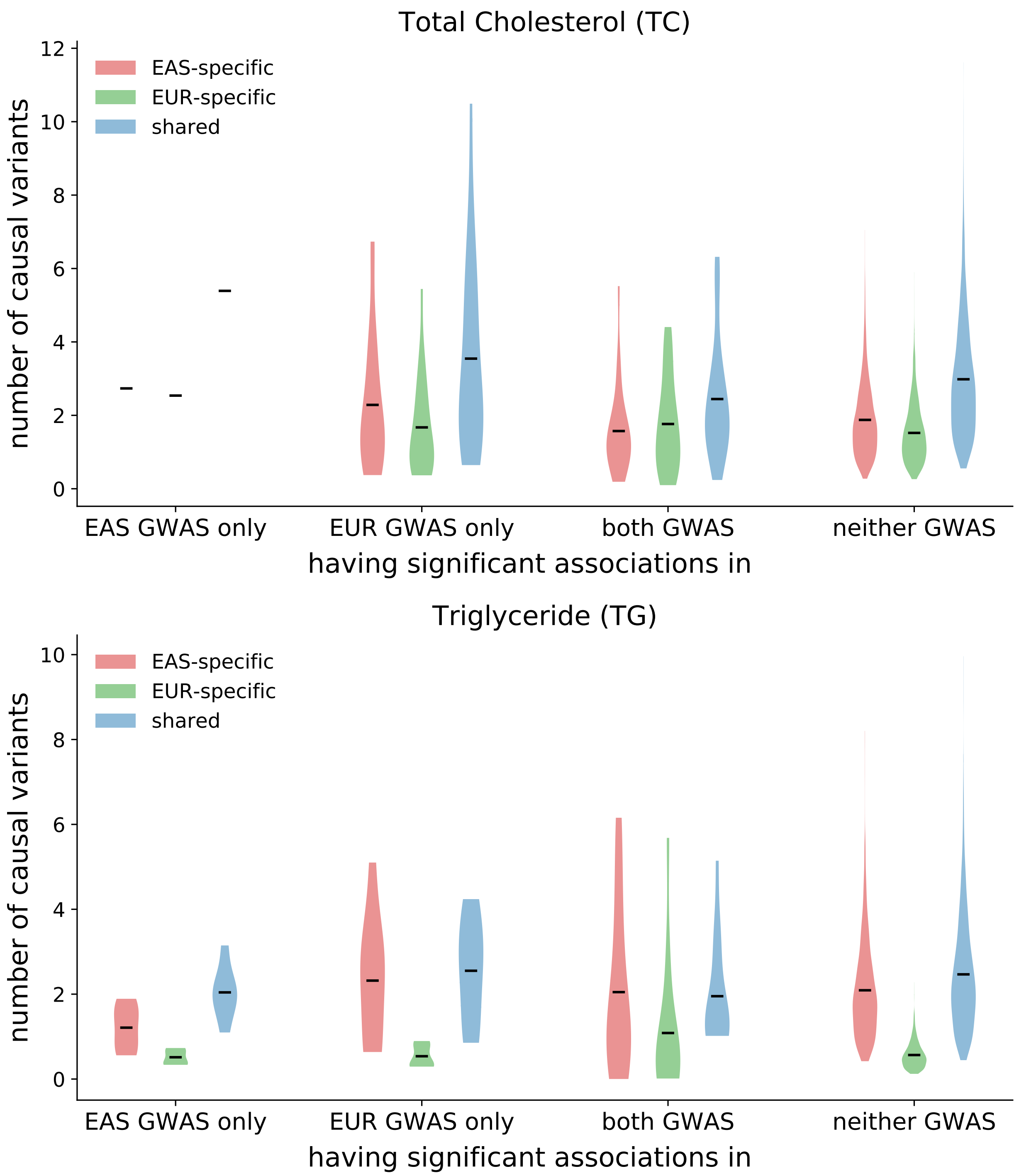


Figure S38: **Distribution of regional number of causal variants at GWAS risk regions.** Each violin plot shows the distribution of population-specific or shared causal variants at regions harboring significant associations ($p < 5 \times 10^{-5}$) in the East Asian GWAS only, in the European GWAS only, in both GWASs, and in neither GWAS. The dark line represents the mean of the distribution.

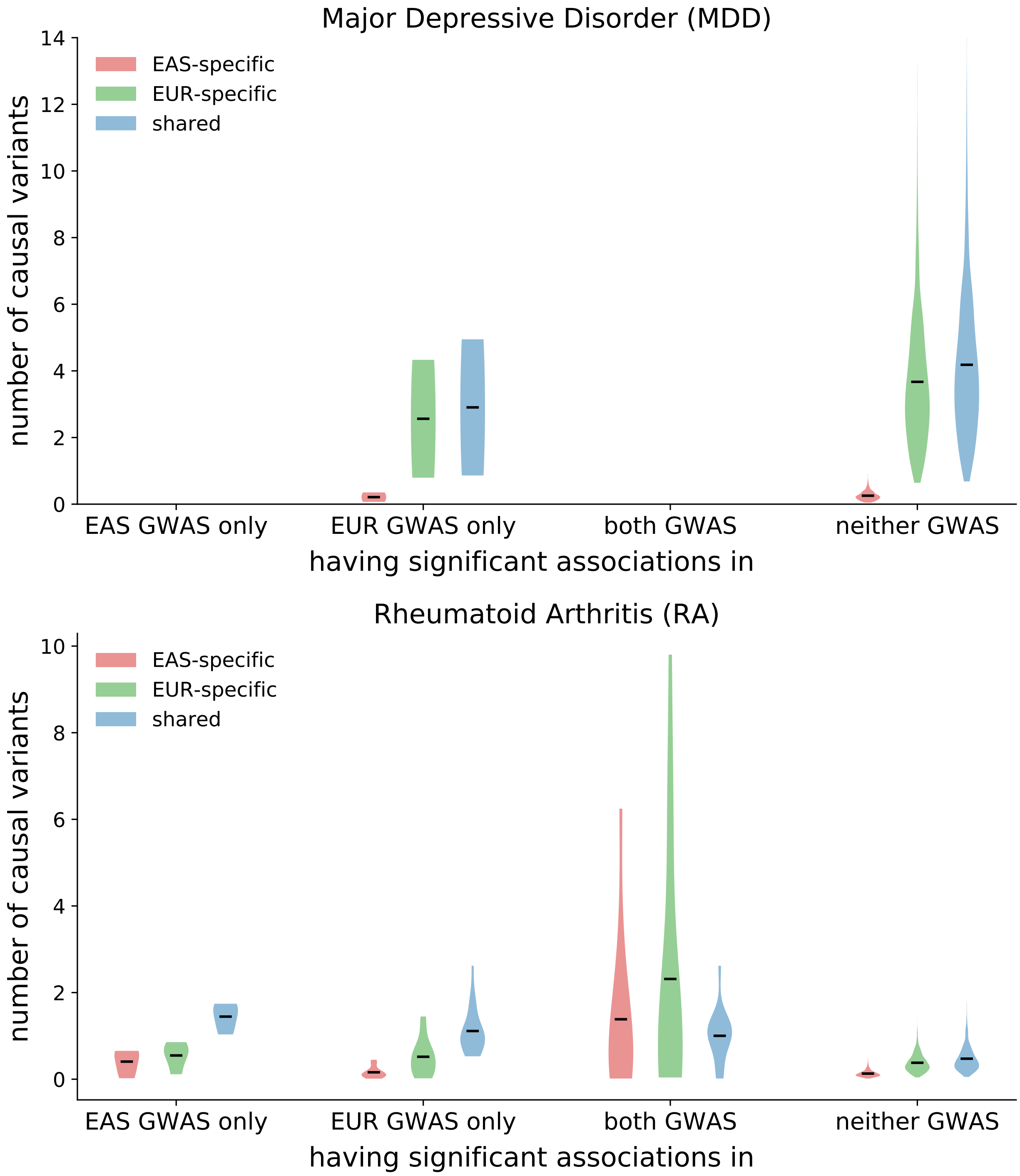


Figure S39: **Distribution of regional number of causal variants at GWAS risk regions.** Each violin plot shows the distribution of population-specific or shared causal variants at regions harboring significant associations ($p < 5 \times 10^{-5}$) in the East Asian GWAS only, in the European GWAS only, in both GWASs, and in neither GWAS. The dark line represents the mean of the distribution.

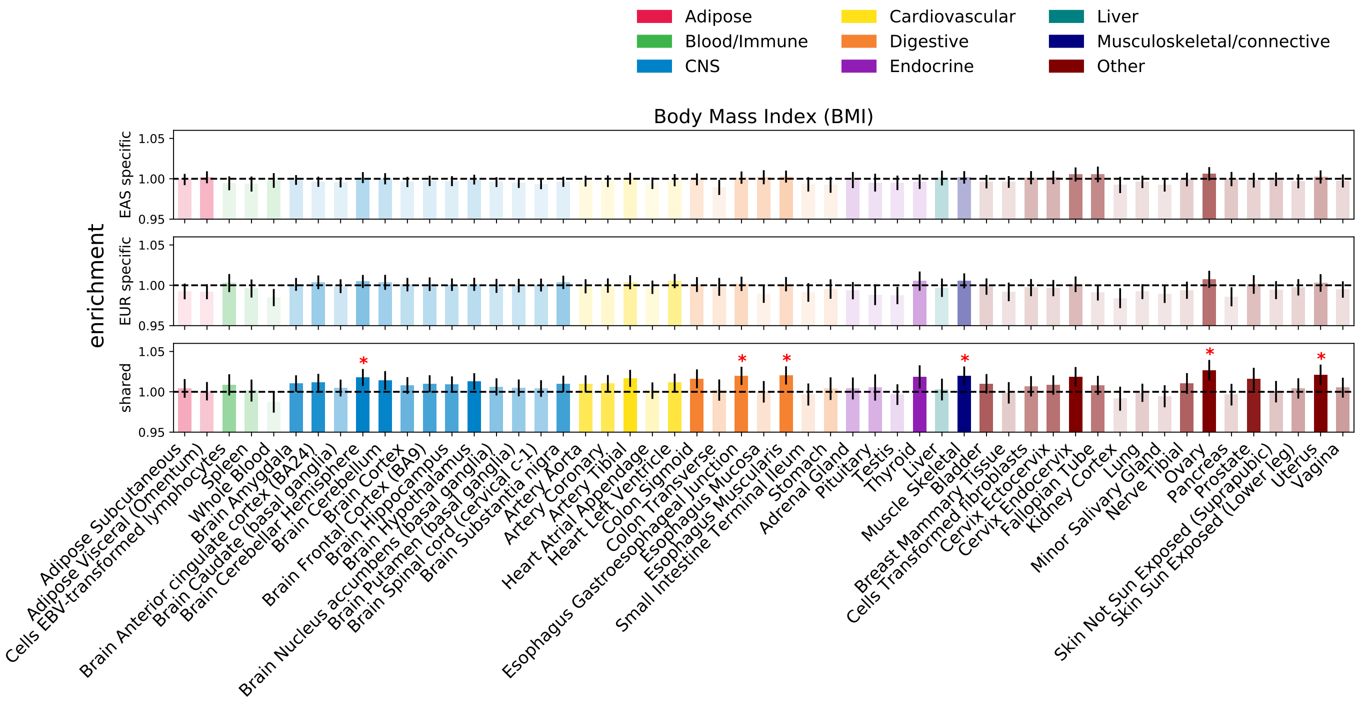


Figure S40: **Enrichment of population-specific and shared causal variants in specifically expressed genes annotation across 53 GTEx tissues.** Error bars represent 1.96 times the standard error on each side. The darker the color, the more significant an enrichment is. We mark enrichment with p-value less than $0.05/53$ with a star.

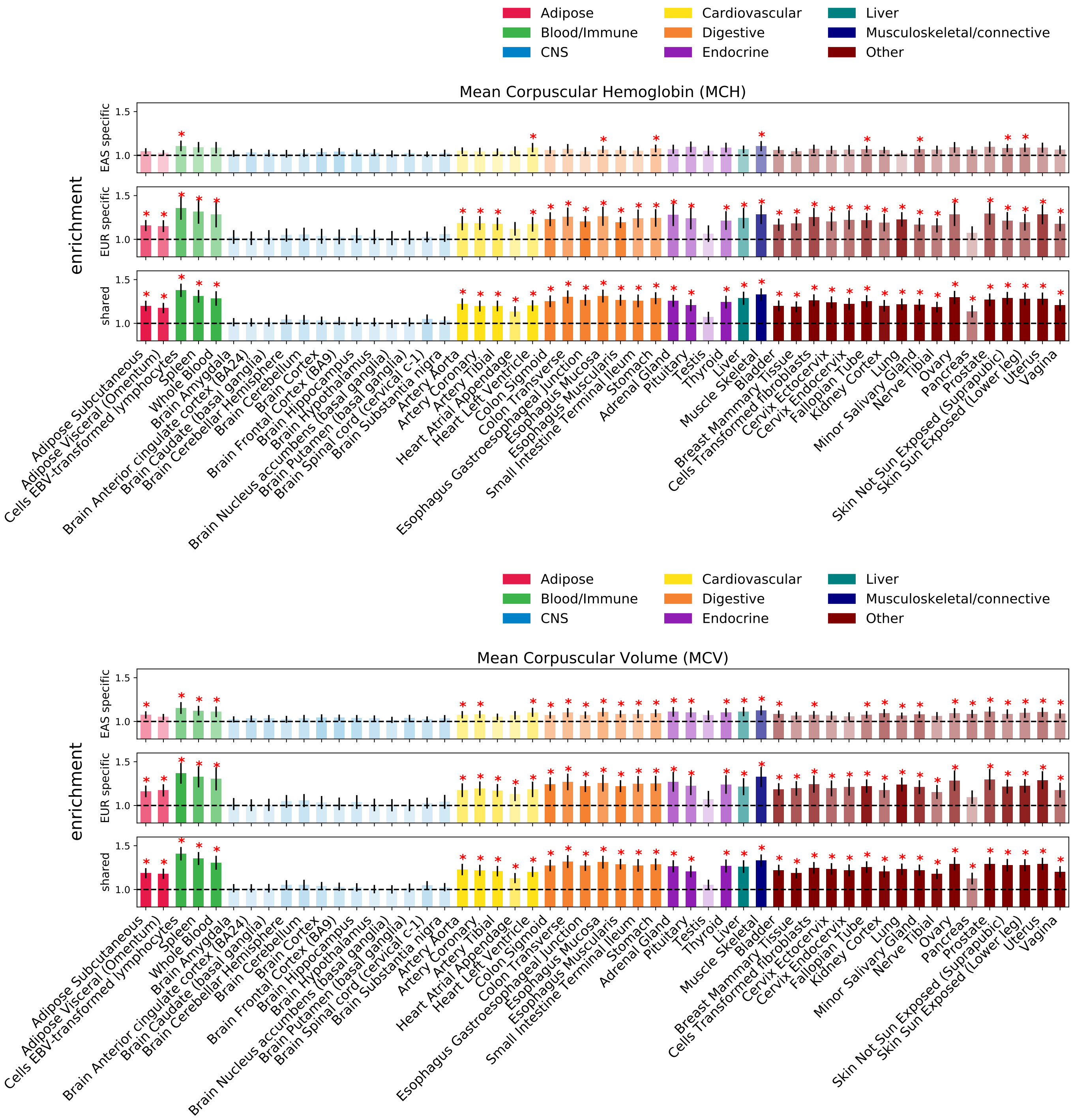


Figure S41: **Enrichment of population-specific and shared causal variants in specifically expressed genes annotation across 53 GTEx tissues.** Error bars represent 1.96 times the standard error on each side. The darker the color, the more significant an enrichment is. We mark enrichment with p-value less than $0.05/53$ with a star.

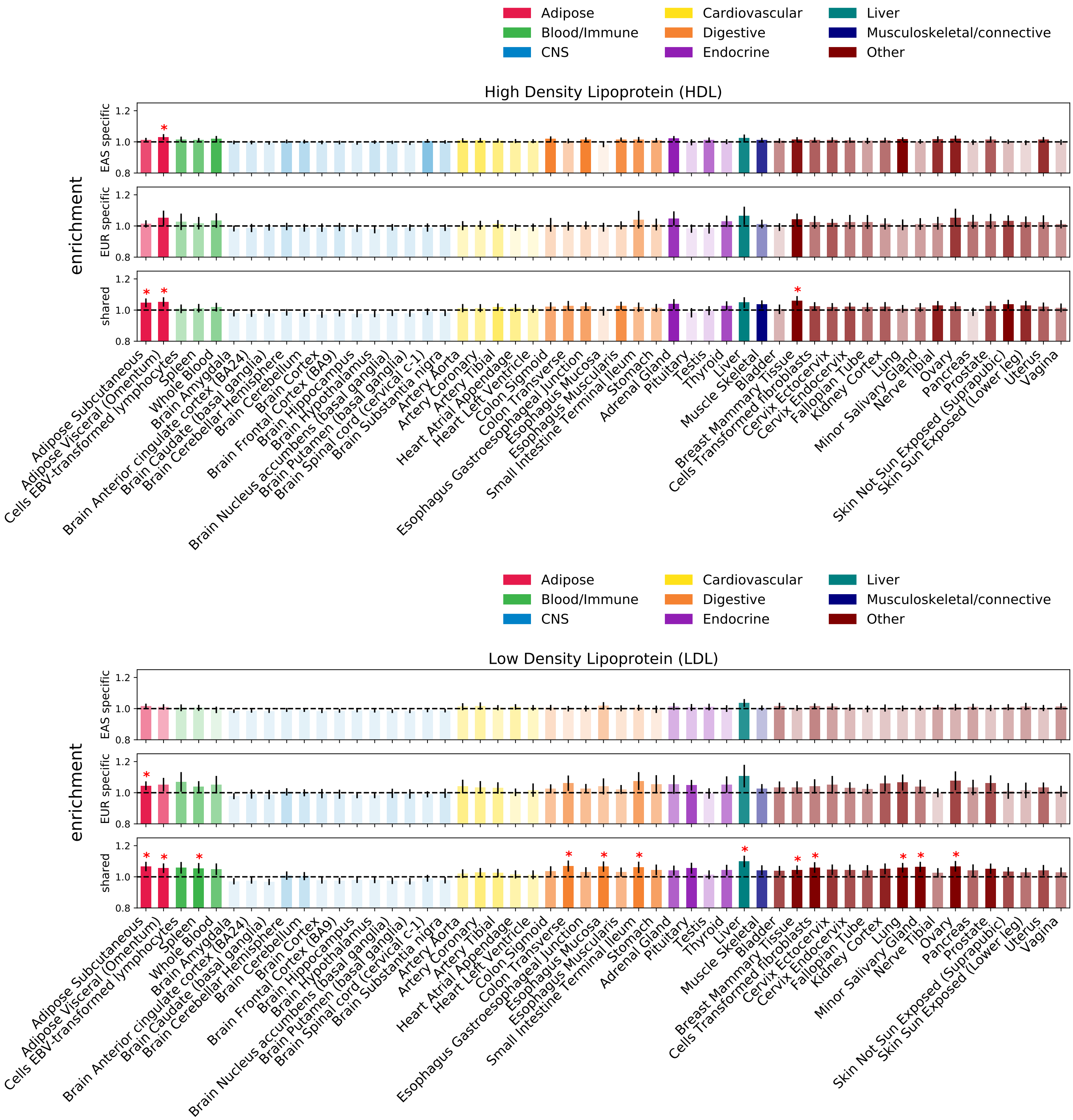


Figure S42: **Enrichment of population-specific and shared causal variants in specifically expressed genes annotation across 53 GTEx tissues.** Error bars represent 1.96 times the standard error on each side. The darker the color, the more significant an enrichment is. We mark enrichment with p-value less than $0.05/53$ with a star.

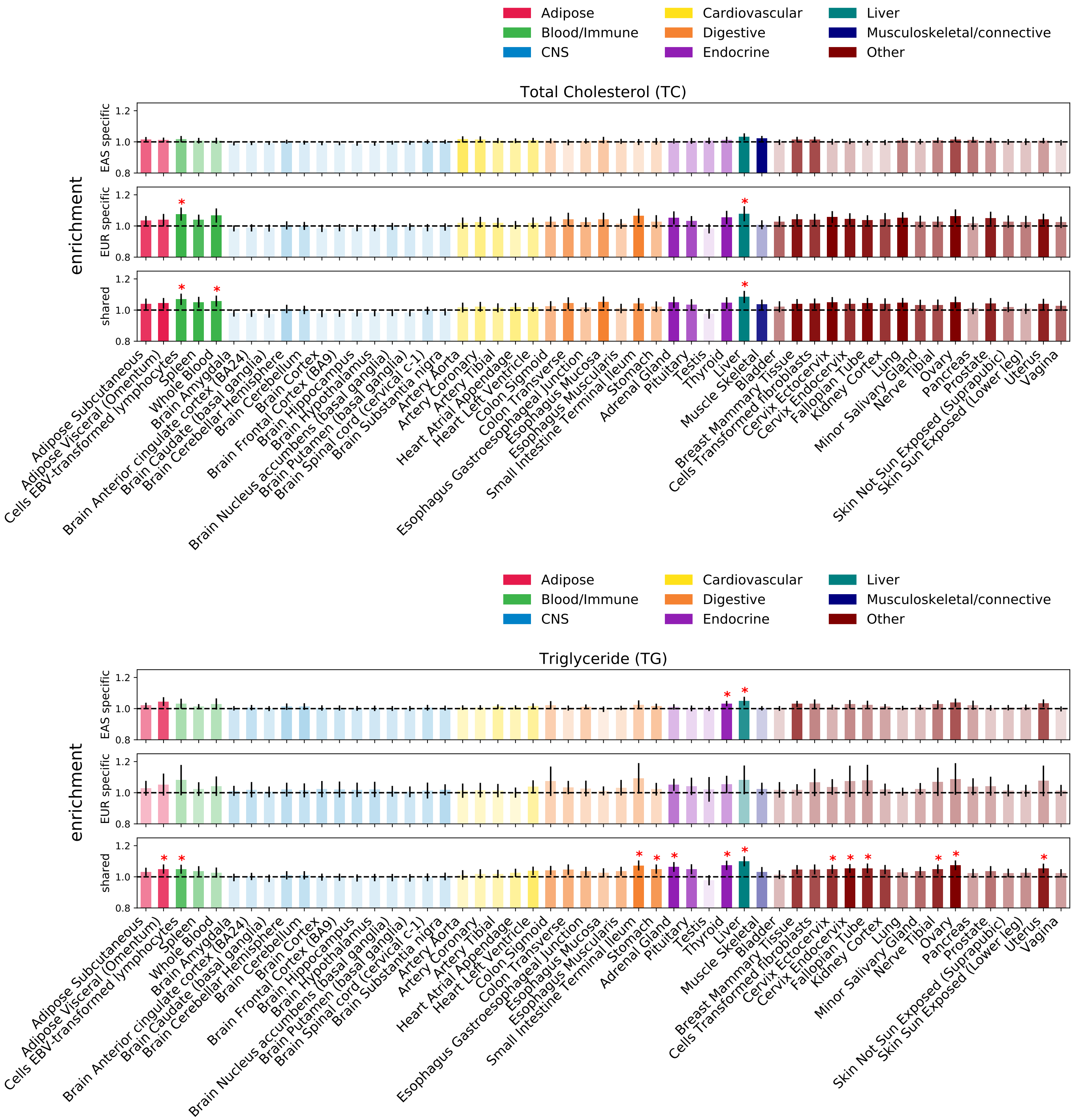


Figure S43: **Enrichment of population-specific and shared causal variants in specifically expressed genes annotation across 53 GTEx tissues.** Error bars represent 1.96 times the standard error on each side. The darker the color, the more significant an enrichment is. We mark enrichment with p-value less than $0.05/53$ with a star.

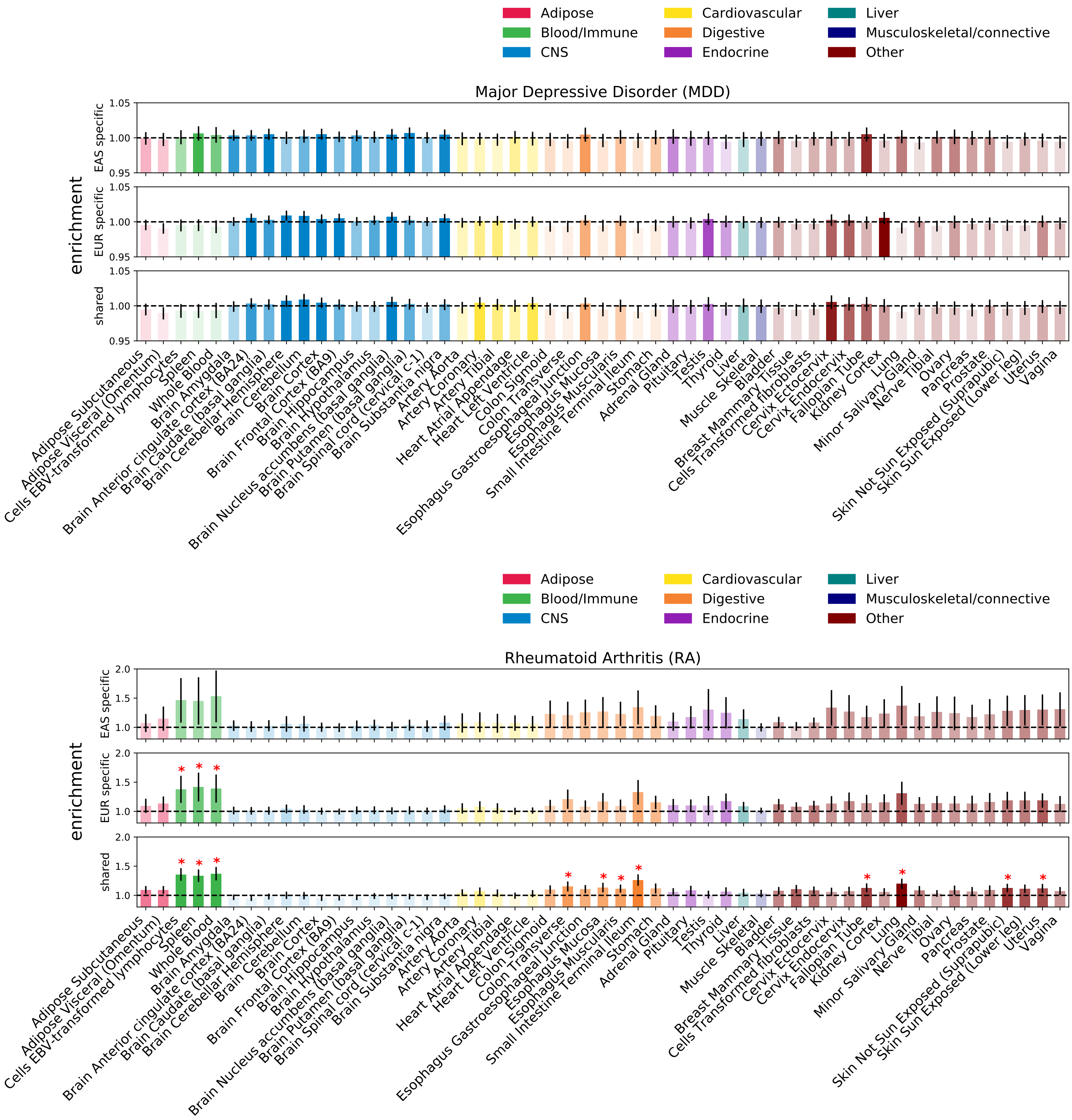


Figure S44: **Enrichment of population-specific and shared causal variants in specifically expressed genes annotation across 53 GTEx tissues.** Error bars represent 1.96 times the standard error on each side. The darker the color, the more significant an enrichment is. We mark enrichment with p-value less than $0.05/53$ with a star.

2 Supplemental Tables

true_cau_status	Posterior > t	mean_l2_EAS	sem_l2_EAS	mean_l2_EUR	sem_l2_EUR	t
shared	shared	6.79	0.08	6.73	0.08	0.25
none	shared	7.37	0.1	7.13	0.09	0.25
EUR_only	shared	6.87	0.21	6.44	0.21	0.25
EAS_only	shared	6.55	0.19	6.72	0.21	0.25
shared	EAS_only	6.57	0.23	6.75	0.25	0.25
none	EAS_only	6.54	0.22	6.58	0.25	0.25
EAS_only	EAS_only	6.49	0.21	6.61	0.24	0.25
shared	EUR_only	6.74	0.2	6.16	0.19	0.25
none	EUR_only	7.02	0.26	6.57	0.22	0.25
EUR_only	EUR_only	7.02	0.29	6.36	0.23	0.25
EAS_only	EUR_only	5.27	0.18	5.46	0.75	0.25
shared	shared	6.52	0.11	6.39	0.1	0.5
EUR_only	shared	6.72	0.34	6.12	0.33	0.5
EAS_only	shared	6.36	0.32	6.54	0.35	0.5
none	shared	7.2	0.2	7.09	0.18	0.5
shared	EAS_only	6.45	0.4	6.49	0.35	0.5
none	EAS_only	7.18	0.73	7.89	0.9	0.5
EAS_only	EAS_only	6.94	0.45	6.82	0.53	0.5
EUR_only	EUR_only	6.3	0.35	5.84	0.3	0.5
shared	EUR_only	7.27	0.39	6.4	0.31	0.5
none	EUR_only	8.1	0.91	7.23	0.72	0.5

Table S1: **Average LD scores of SNPs with posterior probability $> t$ for at least one causal configuration.** For each set of SNPs with posterior probability $> t$ (i.e. SNPs classified as shared, EAS-specific, or EUR-specific with respect to a given threshold), we stratified the SNPs by their true causal statuses and report the mean and S.E.M. of their EAS and EUR LD scores. Column 1 contains the true causal statuses; column 2 contains the causal configurations for which at least two SNPs have posterior probability $> t$.

3 Supplemental Material and Methods

3.1 The multivariate Bernoulli (MVB) distribution

The multivariate Bernoulli (MVB) is a generalization of the Bernoulli for modeling the distribution of a binary vector of arbitrary size[2,3]. Let $\boldsymbol{B} \in \{0,1\}^p$ represent a random binary vector of size p that follows an MVB distribution. The distribution of $\boldsymbol{B}$ can be described by 2^p probabilities, namely $\Pr(\boldsymbol{B} = 0,\cdots,0),\cdots,\Pr(\boldsymbol{B} = 1,\cdots,1)$, one for each of the 2^p possible realizations of $\boldsymbol{B}$[2,3]. Alternatively, one can adopt an index set representation of the binary vector $\boldsymbol{B}$, $\boldsymbol{A} = \{i : \boldsymbol{B}_i = 1\}$, the set of indices of 1's in $\boldsymbol{B}$, and represent the distribution of $\boldsymbol{B}$ as the ratio

$$\Pr(\boldsymbol{B}) = \Pr(\boldsymbol{A}) = \frac{\exp\left(\sum_{\boldsymbol{C}\subseteq\boldsymbol{A}} f_{\boldsymbol{C}}\right)}{\sum_{\boldsymbol{D}} \exp\left(\sum_{\boldsymbol{C}\subseteq\boldsymbol{D}} f_{\boldsymbol{C}}\right)} = \frac{\exp\left(S_{\boldsymbol{A}}\right)}{\sum_{\boldsymbol{D}} \exp\left(S_{\boldsymbol{D}}\right)}, \tag{1}$$

where $f_{\boldsymbol{C}}$ contains the natural parameters of the MVB[2,3] and $S_{\boldsymbol{A}} = \sum_{\boldsymbol{C}\subseteq\boldsymbol{A}} f_{\boldsymbol{C}}$.

We use the convention that the right-most bit in the binary vector is the first bit and the left-most bit is the last bit. For convenience, we use binary string and index set representation of binary vectors interchangeably (e.g., both the binary string 011 and the index set $\{1,2\}$ represent the binary vector $(0,1,1)$).

As a concrete example, consider a binary vector of size 2. The probabilities of each possible realization of a binary vector of size 2 under the MVB are

$$\begin{aligned}
\Pr(00) = \Pr(\phi) &= \frac{\exp(f_{00})}{\exp(f_{00}) + \exp(f_{00}+f_{01}) + \exp(f_{00}+f_{10}) + \exp(f_{00}+f_{01}+f_{10}+f_{11})} \\
\Pr(01) = \Pr(\{1\}) &= \frac{\exp(f_{00}+f_{01})}{\exp(f_{00}) + \exp(f_{00}+f_{01}) + \exp(f_{00}+f_{10}) + \exp(f_{00}+f_{01}+f_{10}+f_{11})} \\
\Pr(10) = \Pr(\{2\}) &= \frac{\exp(f_{00}+f_{10})}{\exp(f_{00}) + \exp(f_{00}+f_{01}) + \exp(f_{00}+f_{10}) + \exp(f_{00}+f_{01}+f_{10}+f_{11})} \\
\Pr(11) = \Pr(\{1,2\}) &= \frac{\exp(f_{00}+f_{01}+f_{10}+f_{11})}{\exp(f_{00}) + \exp(f_{00}+f_{01}) + \exp(f_{00}+f_{10}) + \exp(f_{00}+f_{01}+f_{10}+f_{11})}
\end{aligned} . \tag{2}$$

3.2 MVB prior for a SNP's causal status in two ancestral populations

We use a binary vector of size 2, $\boldsymbol{C}_i = (\boldsymbol{c}_{i1}, \boldsymbol{c}_{i2})$, to model the causal statuses of SNP i in two ancestral populations. In total, there are 4 possible binary vectors of size 2: if $\boldsymbol{C}_i = 00$, the SNP is causal in neither population; if $\boldsymbol{C}_i = 01$, the SNP is causal in population 1 only; if $\boldsymbol{C}_i = 10$, the SNP is causal in population 2 only; and if $\boldsymbol{C}_i = 11$, the SNP is causal in both populations. $\boldsymbol{C}_i$ can be modeled using a multinomial distribution, Mult$(p_{00}, p_{01}, p_{10}, p_{11})$, where p_{00}, p_{01}, p_{10}, and p_{11} represent the probability of each possible binary vector of size 2. Equivalently, one can model

$\boldsymbol{C}_i$ through the MVB as

$$\begin{aligned}
\Pr(\boldsymbol{C}_i = 00) &= \frac{\exp(f_{00})}{\eta} \\
\Pr(\boldsymbol{C}_i = 01) &= \frac{\exp(f_{01} + f_{00})}{\eta} \\
\Pr(\boldsymbol{C}_i = 10) &= \frac{\exp(f_{10} + f_{00})}{\eta} \\
\Pr(\boldsymbol{C}_i = 11) &= \frac{\exp(f_{11} + f_{10} + f_{01} + f_{00})}{\eta},
\end{aligned} \tag{3}$$

where $\eta = \exp(f_{00}) + \exp(f_{01} + f_{00}) + \exp(f_{10} + f_{00}) + \exp(f_{11} + f_{10} + f_{01} + f_{00})$ is the normalization constant and $\boldsymbol{f} = (f_{00}, f_{01}, f_{10}, f_{11})$ are the parameters of the MVB (see Equation (2)).

Since the MVB distribution is invariant with respect to the parameter f_{00}, we enforce $f_{00} = 0$ as a convention[2]. The parameters f_{01} and f_{10} govern the probability of a SNP being causal in a single population and f_{11} governs the dependence of the causal statuses between two populations; $f_{11} = 0$ indicates independence and $f_{11} \neq 0$ indicates dependence[2,3]. Since the MVB parameters are real numbers (i.e. $\boldsymbol{f} \in \mathbb{R}^4$), they can be estimated using unconstrained optimization.

3.3 Joint distribution of GWAS summary statistics in two ancestral populations

We model a phenotype in two ancestral populations using the linear models $\boldsymbol{Y}_1 = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}_1$ and $\boldsymbol{Y}_2 = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}_2$, where $\boldsymbol{Y}_1 \in \mathbb{R}^{n_1}$ and $\boldsymbol{Y}_2 \in \mathbb{R}^{n_2}$ are the phenotype measurements for n_1 individuals in population 1 and n_2 individuals in population 2, respectively; $\boldsymbol{X}_1 \in \mathbb{R}^{n_1 \times p}$ and $\boldsymbol{X}_2 \in \mathbb{R}^{n_2 \times p}$ are column-standardized genotype matrices for p SNPs; $\boldsymbol{\beta}_1 \in \mathbb{R}^p$ and $\boldsymbol{\beta}_2 \in \mathbb{R}^p$ are the standardized causal effect sizes of the p SNPs in the two populations, and $\boldsymbol{\epsilon}_1 \in \mathbb{R}^{n_1}$ and $\boldsymbol{\epsilon} \in \mathbb{R}^{n_2}$ are environmental effects. We further assume that, for population j, the genotype vector of each individual is drawn from a distribution with covariance $\boldsymbol{V}_j$ (the $p \times p$ LD matrix in population j) and that $\boldsymbol{\epsilon}_j \sim N\left(0, \sigma_{ej}^2 \boldsymbol{I}\right)$, where σ_{ej}^2 is the variance of the environmental effects in population j.

In a typical GWAS, one obtains association statistics (Z-scores) of every SNP as

$$\begin{aligned}
\boldsymbol{Z}_1 &= \frac{1}{\sqrt{n_1}} \boldsymbol{X}_1^\intercal \boldsymbol{Y}_1 \\
\boldsymbol{Z}_2 &= \frac{1}{\sqrt{n_2}} \boldsymbol{X}_2^\intercal \boldsymbol{Y}_2
\end{aligned} \tag{4}$$

which have been shown to follow the multivariate normal distributions[4]

$$\begin{aligned} \boldsymbol{Z}_1|\boldsymbol{\beta}_1 &\sim N\left(\sqrt{n_1}\boldsymbol{V}_1\boldsymbol{\beta}_1, \sigma_{e1}^2\boldsymbol{V}_1\right) \\ \boldsymbol{Z}_2|\boldsymbol{\beta}_2 &\sim N\left(\sqrt{n_2}\boldsymbol{V}_1\boldsymbol{\beta}_2, \sigma_{e2}^2\boldsymbol{V}_2\right) \end{aligned} \tag{5}$$

Given the causal status vectors, $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$, of every SNP in each population, one obtains the conditional distributions $\boldsymbol{Z}_1|\boldsymbol{\beta}_1, \boldsymbol{c}_1$ and $\boldsymbol{Z}_2|\boldsymbol{\beta}_2, \boldsymbol{c}_2$ as

$$\begin{aligned} \boldsymbol{Z}_1|\boldsymbol{\beta}_1, \boldsymbol{c}_1 &\sim N\left(\sqrt{n_1}\boldsymbol{V}_1(\boldsymbol{\beta}_1 \circ \boldsymbol{c}_1), \sigma_{e1}^2\boldsymbol{V}_1\right) \\ \boldsymbol{Z}_2|\boldsymbol{\beta}_2, \boldsymbol{c}_2 &\sim N\left(\sqrt{n_2}\boldsymbol{V}_2(\boldsymbol{\beta}_2 \circ \boldsymbol{c}_2), \sigma_{e2}^2\boldsymbol{V}_2\right) \end{aligned} \tag{6}$$

where $\circ$ denotes the Hadamard product[5].

Following Equation (6), one can evaluate the likelihood of $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ given the true causal effect size vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. However, in reality the true causal effect size vectors are not given, and estimating these parameters from data will likely lead to over-fitting. Instead, we impose a normal prior on each causal SNP in $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ to obtain

$$\begin{aligned} \boldsymbol{\beta}_1|\boldsymbol{c}_1 &\sim N\left(\boldsymbol{0}, \frac{h_{g1}^2}{|\boldsymbol{c}_1|}\operatorname{diag}(\boldsymbol{c}_1)\right), \\ \boldsymbol{\beta}_2|\boldsymbol{c}_2 &\sim N\left(\boldsymbol{0}, \frac{h_{g2}^2}{|\boldsymbol{c}_2|}\operatorname{diag}(\boldsymbol{c}_2)\right), \end{aligned} \tag{7}$$

where h_{g1}^2 and h_{g2}^2 are the SNP-heritability of the phenotype in population 1 and 2, respectively, and $|\boldsymbol{c}_1|$ and $|\boldsymbol{c}_2|$ denote the number of 1's (i.e. the number of causal SNPs) in the binary vectors[6,7,8]. With the normal prior on $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, the conditional distributions $\boldsymbol{Z}_1|\boldsymbol{c}_1$ and $\boldsymbol{Z}_2|\boldsymbol{c}_2$ are

$$\begin{aligned} \boldsymbol{Z}_1|\boldsymbol{c}_1 &\sim N\left(\boldsymbol{0}, \boldsymbol{V}_1 + \sigma_1^2\boldsymbol{V}_1\operatorname{diag}(\boldsymbol{c}_1)\boldsymbol{V}_1\right), \\ \boldsymbol{Z}_2|\boldsymbol{c}_2 &\sim N\left(\boldsymbol{0}, \boldsymbol{V}_2 + \sigma_2^2\boldsymbol{V}_2\operatorname{diag}(\boldsymbol{c}_2)\boldsymbol{V}_2\right), \end{aligned} \tag{8}$$

where $\sigma_1^2 = \frac{n_1 h_{g1}^2}{|\boldsymbol{c}_1|}$ and $\sigma_2^2 = \frac{n_2 h_{g2}^2}{|\boldsymbol{c}_2|}$.

Incorporating the MVB prior on the causal status vectors, the joint distribution of $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$, which is parameterized by the MVB parameters, $\boldsymbol{f} = (f_{00}, f_{01}, f_{10}, f_{11})$, is

$$
\begin{aligned}
\Pr(\boldsymbol{Z}_1, \boldsymbol{Z}_2; \boldsymbol{f}) &= \sum_{\boldsymbol{c}_1}\sum_{\boldsymbol{c}_2} \Pr(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{c}_1, \boldsymbol{c}_2; \boldsymbol{f}) = \sum_{\boldsymbol{c}_1}\sum_{\boldsymbol{c}_2} \Pr(\boldsymbol{Z}_1|\boldsymbol{c}_1)\Pr(\boldsymbol{Z}_2|\boldsymbol{c}_2)\Pr(\boldsymbol{c}_1, \boldsymbol{c}_2; \boldsymbol{f}) \\
&= \sum_{\boldsymbol{c}_1}\sum_{\boldsymbol{c}_2} \left[\begin{array}{c} N(\boldsymbol{Z}_1; \boldsymbol{0}, \boldsymbol{V}_1 + \sigma_1^2 \boldsymbol{V}_1 \operatorname{diag}(\boldsymbol{c}_1)\boldsymbol{V}_1) \times \\ N(\boldsymbol{Z}_2; \boldsymbol{0}, \boldsymbol{V}_2 + \sigma_2^2 \boldsymbol{V}_2 \operatorname{diag}(\boldsymbol{c}_2)\boldsymbol{V}_2) \times \prod_{i=1}^{p} \frac{\exp(S_{\boldsymbol{C}_i})}{\sum_{\boldsymbol{B}} \exp(S_{\boldsymbol{B}})} \end{array} \right]
\end{aligned}
\tag{9}
$$

To model the joint distribution of GWAS summary statistics across L LD-independent regions, we take the product of the probability of Z-scores across regions:

$$
\begin{aligned}
&\Pr(\boldsymbol{Z}_{1\{1,\cdots,L\}}, \boldsymbol{Z}_{2\{1,\cdots,L\}\}}; \boldsymbol{f}) = \prod_{l=1}^{L} \Pr(\boldsymbol{Z}_{1l}, \boldsymbol{Z}_{2l}; \boldsymbol{f}) \\
&= \prod_{l=1}^{L} \left\{ \sum_{\boldsymbol{c}_{1l}}\sum_{\boldsymbol{c}_{2l}} \left[\begin{array}{c} N(\boldsymbol{Z}_{1l}; \boldsymbol{0}, \boldsymbol{V}_{1l} + \sigma_{1l}^2 \boldsymbol{V}_{1l} \operatorname{diag}(\boldsymbol{c}_{1l})\boldsymbol{V}_{1l}) \times \\ N(\boldsymbol{Z}_{2l}; \boldsymbol{0}, \boldsymbol{V}_{2l} + \sigma_{2l}^2 \boldsymbol{V}_{2l} \operatorname{diag}(\boldsymbol{c}_{2l})\boldsymbol{V}_{2l}) \times \prod_{i=1}^{p_l} \frac{\exp(S_{\boldsymbol{C}_{li}})}{\sum_{\boldsymbol{B}} \exp(S_{\boldsymbol{B}})} \end{array} \right] \right\}.
\end{aligned}
\tag{10}
$$

3.4 Model fitting using Expectation Maximization

3.4.1 Expectation step

We use expectation-maximization (EM) to estimate the model parameters $\boldsymbol{f}$. First, we derive the complete log-likelihood of the data

$$
\begin{aligned}
&\ell\left(\boldsymbol{f}|\boldsymbol{Z}_{1\{1,\cdots,L\}}, \boldsymbol{Z}_{2\{1,\cdots,L\}}, \boldsymbol{c}_{1\{1,\cdots,L\}}, \boldsymbol{c}_{2\{1,\cdots,L\}}\right) \\
&= \log \left\{ \prod_{l=1}^{L} \left[\begin{array}{c} N(\boldsymbol{Z}_{1l}; \boldsymbol{0}, \boldsymbol{V}_{1l} + \sigma_{1l}^2 \boldsymbol{V}_{1l} \operatorname{diag}(\boldsymbol{c}_{1l})\boldsymbol{V}_{1l}) \times \\ N(\boldsymbol{Z}_{2l}; \boldsymbol{0}, \boldsymbol{V}_{2l} + \sigma_{2l}^2 \boldsymbol{V}_{2l} \operatorname{diag}(\boldsymbol{c}_{2l})\boldsymbol{V}_{2l}) \times \prod_{i=1}^{p_l} \frac{\exp(S_{\boldsymbol{C}_{li}})}{\sum_{\boldsymbol{B}} \exp(S_{\boldsymbol{B}})} \end{array} \right] \right\} \\
&= \sum_{l=1}^{L} \left[\log N(\boldsymbol{Z}_{1l}; \boldsymbol{0}, \boldsymbol{V}_{1l} + \sigma_{1l}^2 \boldsymbol{V}_{1l} \operatorname{diag}(\boldsymbol{c}_{1l})\boldsymbol{V}_{1l}) + \log N(\boldsymbol{Z}_{2l}; \boldsymbol{0}, \boldsymbol{V}_{2l} + \sigma_{2l}^2 \boldsymbol{V}_{2l} \operatorname{diag}(\boldsymbol{c}_{2l})\boldsymbol{V}_{2l}) \right] \\
&\quad + \sum_{l=1}^{L}\sum_{i=1}^{p_l} S_{C_{li}} - \log\left(\sum_{\boldsymbol{B}} \exp(S_{\boldsymbol{B}}) \right) \sum_{l=1}^{L} p_l.
\end{aligned}
\tag{11}
$$

In the expectation step of the EM algorithm, one finds the expectation of the log-likelihood with respect to the causal status vectors $\boldsymbol{c}_{1\{1,\cdots,L\}}$, $\boldsymbol{c}_{2\{1,\cdots,L\}}$, conditioned on the current estimate of the model parameters $\boldsymbol{f}^{(t)}$,

$$\begin{aligned}
Q\left(\boldsymbol{f}|\boldsymbol{f}^{(t)}\right) &= \mathrm{E}\left[\ell\left(\boldsymbol{f}|\boldsymbol{Z}_{1\{1,\cdots,L\}}, \boldsymbol{Z}_{2\{1,\cdots,L\}}, \boldsymbol{c}_{1\{1,\cdots,L\}}, \boldsymbol{c}_{2\{1,\cdots,L\}}\right)\right] \\
&= \sum_{l=1}^{L} \sum_{\boldsymbol{c}_{1l},\boldsymbol{c}_{2l}} \Pr\left(\boldsymbol{c}_{1l}, \boldsymbol{c}_{2l}|\boldsymbol{f}^{(t)}, \boldsymbol{Z}_{1l}, \boldsymbol{Z}_{2l}\right) \left[\begin{array}{c} \log N(\boldsymbol{Z}_{1l}; \mathbf{0}, \boldsymbol{V}_{1l} + \sigma_{1l}^2 \boldsymbol{V}_{1l} \operatorname{diag}(\boldsymbol{c}_{1l}) \boldsymbol{V}_{1l}) \\ + \log N(\boldsymbol{Z}_{2l}; \mathbf{0}, \boldsymbol{V}_{2l} + \sigma_{2l}^2 \boldsymbol{V}_{2l} \operatorname{diag}(\boldsymbol{c}_{2l}) \boldsymbol{V}_{2l}) \end{array}\right] \\
&\quad + \sum_{l=1}^{L} \sum_{\boldsymbol{c}_{1l},\boldsymbol{c}_{2l}} \Pr\left(\boldsymbol{c}_{1l}, \boldsymbol{c}_{2l}|\boldsymbol{f}^{(t)}, \boldsymbol{Z}_{1l}, \boldsymbol{Z}_{2l}\right) \left(\sum_{i=1}^{p_l} S_{\boldsymbol{C}_{li}}\right) - \log\left(\sum_{\boldsymbol{B}} \exp(S_{\boldsymbol{B}})\right) \sum_{l=1}^{L} p_l,
\end{aligned} \tag{12}$$

where $\Pr\left(\boldsymbol{c}_{1l}, \boldsymbol{c}_{2l}|\boldsymbol{f}^{(t)}, \boldsymbol{Z}_{1l}, \boldsymbol{Z}_{2l}\right)$ is

$$\Pr\left(\boldsymbol{c}_{1l}, \boldsymbol{c}_{2l}|\boldsymbol{f}^{(t)}, \boldsymbol{Z}_{1l}, \boldsymbol{Z}_{2l}\right) = \frac{\Pr\left(\boldsymbol{c}_{1l}, \boldsymbol{c}_{2l}, \boldsymbol{Z}_{1l}, \boldsymbol{Z}_{2l}|\boldsymbol{f}^{(t)}\right)}{\sum_{\boldsymbol{b}_{1l},\boldsymbol{b}_{2l}} \Pr\left(\boldsymbol{b}_{1l}, \boldsymbol{b}_{2l}, \boldsymbol{Z}_{1l}, \boldsymbol{Z}_{2l}|\boldsymbol{f}^{(t)}\right)}. \tag{13}$$

3.4.2 Maximization step

The goal of the maximization step is to find

$$\boldsymbol{f}^{(t+1)} = \operatorname{argmax}_{\boldsymbol{f}} Q\left(\boldsymbol{f}|\boldsymbol{f}^{(t)}\right) = \operatorname{argmax}_{\boldsymbol{f}} g(\boldsymbol{f}) \tag{14}$$

where

$$g(\boldsymbol{f}) = \sum_{l=1}^{L} \sum_{\boldsymbol{c}_{1l},\boldsymbol{c}_{2l}} \Pr\left(\boldsymbol{c}_{1l}, \boldsymbol{c}_{2l}|\boldsymbol{f}^{(t)}, \boldsymbol{Z}_{1l}, \boldsymbol{Z}_{2l}\right) \left(\sum_{i=1}^{p_l} S_{\boldsymbol{C}_{li}}\right) - \log\left(\sum_{\boldsymbol{B}} \exp(S_{\boldsymbol{B}})\right) \sum_{l=1}^{L} p_l, \tag{15}$$

removing the irrelevant constant in $Q(\boldsymbol{f}|\boldsymbol{f}^{(t)})$.

Evaluating $g(\boldsymbol{f})$ involves a summation over all possible causal status vectors, which has time complexity on the order of $O(2^{2p_l})$ and is intractable. Instead, we recognize that

$$\begin{aligned}
g(\boldsymbol{f}) &= \sum_{l=1}^{L} \sum_{\boldsymbol{c}_{1l},\boldsymbol{c}_{2l}} \mathrm{E}\left[\sum_{i=1}^{p_l} S_{\boldsymbol{C}_{li}}\right] - \log\left(\sum_{\boldsymbol{B}} \exp(S_{\boldsymbol{B}})\right) \sum_{l=1}^{L} p_l \\
&\approx h(\boldsymbol{f}) = \sum_{l=1}^{L} \left[\frac{1}{J} \sum_{j=1}^{J} \left(\sum_{i=1}^{p_l} S_{\boldsymbol{C}_{li}^{(j)}}\right)\right] - \log\left(\sum_{\boldsymbol{B}} \exp(S_{\boldsymbol{B}})\right) \sum_{l=1}^{L} p_l,
\end{aligned} \tag{16}$$

where $\boldsymbol{C}_{li}^{(j)} = \left(\boldsymbol{c}_{1i}^{(j)}, \boldsymbol{c}_{2i}^{(j)}\right)$ represents the causal status of the i-th SNP at locus l in the two populations, from the causal status vectors, $\boldsymbol{c}_1^{(j)}$, $\boldsymbol{c}_2^{(j)}$, sampled from the posterior distribution $\Pr\left(\boldsymbol{c}_{1l}, \boldsymbol{c}_{2l}|\boldsymbol{Z}_{1l}, \boldsymbol{Z}_{2l}, \boldsymbol{f}^*\right)$. We use Gibbs sampling to efficiently sample causal status vectors from the posterior (see Section 3.5).

It can be shown that the following parameter updates maximizes $h(\boldsymbol{f})$,

$$\begin{aligned}
\boldsymbol{f}_{00}^{(t+1)} &= 0, \\
\boldsymbol{f}_{01}^{(t+1)} &= \log \bar{q}_{01} - \log \bar{q}_{00}, \\
\boldsymbol{f}_{10}^{(t+1)} &= \log \bar{q}_{10} - \log \bar{q}_{00}, \\
\boldsymbol{f}_{11}^{(t+1)} &= \log \bar{q}_{11} - \log \bar{q}_{01} - \log \bar{q}_{10} + \log \bar{q}_{00},
\end{aligned} \tag{17}$$

where $\bar{q}_{00}$, $\bar{q}_{01}$, $\bar{q}_{10}$, and $\bar{q}_{11}$ represent the average count of 01, 10, and 11 causal status at a single SNP in two ancestral populations across MCMC samples from the Gibbs sampler (see Section 3.5).

3.5 Sampling causal status vectors from the posterior distribution

We use Gibbs sampling to sample $\boldsymbol{C} = (\boldsymbol{c}_1, \boldsymbol{c}_2)$ from the posterior distribution,

$$\boldsymbol{C} \sim \Pr\left(\boldsymbol{C} | \boldsymbol{f}, \boldsymbol{Z}_1, \boldsymbol{Z}_2\right) \propto \Pr\left(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{C} | \boldsymbol{f}\right). \tag{18}$$

For notational simplicity, we drop the index l representing different loci. To advance the Markov chain from step j to step $j+1$ in Gibbs sampling, at step j we select SNP k and evaluate the probability of the four possible cross-population causal configurations at that SNP,

$$\begin{aligned}
&\Pr\left(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{C}_k = 00, \boldsymbol{C}_{\neg j}^{(j)} | \boldsymbol{f}\right) \ \Pr\left(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{C}_k = 01, \boldsymbol{C}_{\neg j}^{(j)} | \boldsymbol{f}\right) \\
&\Pr\left(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{C}_k = 10, \boldsymbol{C}_{\neg j}^{(j)} | \boldsymbol{f}\right) \ \Pr\left(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{C}_k = 11, \boldsymbol{C}_{\neg j}^{(j)} | \boldsymbol{f}\right),
\end{aligned} \tag{19}$$

where $\boldsymbol{C}_{\neg j}^{(j)}$ denotes the rest of the causal configurations, excluding that of SNP k in the j-th step. We then sample $\boldsymbol{C}^{(j+1)}$ based on the following probability

$$\Pr\left(\boldsymbol{C}^{(t+1)} = \left(\boldsymbol{C}_k = \boldsymbol{b}', \boldsymbol{C}_{\neg j}^{(j)}\right)\right) = \frac{\Pr\left(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{C}_k = \boldsymbol{b}', \boldsymbol{C}_{\neg j}^{(j)} | \boldsymbol{f}\right)}{\sum_{\boldsymbol{b}} \Pr\left(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{C}_k = \boldsymbol{b}, \boldsymbol{C}_{\neg j}^{(j)} | \boldsymbol{f}\right)}. \tag{20}$$

To evaluate $\Pr(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{c}_1, \boldsymbol{c}_2 | \boldsymbol{f}) = \Pr(\boldsymbol{Z}_1 | \boldsymbol{c}_1) \Pr(\boldsymbol{Z}_2 | \boldsymbol{c}_2) \Pr(\boldsymbol{c}_1, \boldsymbol{c}_2 | \boldsymbol{f})$, we note that previous work has shown that

$$\begin{aligned}
\Pr(\boldsymbol{Z}_1 | \boldsymbol{c}_1) &= N\left(\boldsymbol{Z}_1 | \boldsymbol{0}, \boldsymbol{V}_1 + \sigma_1^2 \boldsymbol{V}_1^2\right) \\
&\propto \frac{N\left(\boldsymbol{Z}_{1\boldsymbol{c}_1} | \boldsymbol{0}, \boldsymbol{V}_{1\boldsymbol{c}_1} + \sigma_1^2 \boldsymbol{V}_{1\boldsymbol{c}_1}^2\right)}{N\left(\boldsymbol{Z}_{1\boldsymbol{c}_1} | \boldsymbol{0}, \boldsymbol{V}_{1\boldsymbol{c}_1}\right)},
\end{aligned} \tag{21}$$

where $BF_1 = \frac{N\left(\boldsymbol{Z}_{1\boldsymbol{c}_1}|\boldsymbol{0},\boldsymbol{V}_{1\boldsymbol{c}_1}+\sigma_1^2\boldsymbol{V}_{1\boldsymbol{c}_1}^2\right)}{N\left(\boldsymbol{Z}_{1\boldsymbol{c}_1}|\boldsymbol{0},\boldsymbol{V}_{1\boldsymbol{c}_1}\right)}$ is the Bayes factor at only the causal SNPs, reducing the time complexity of evaluating the probability from p^3 to p^3_{causal}. Let $\boldsymbol{V}_{1\boldsymbol{c}_1} = \sum_{i=1}^{p_{\text{causal}}} w_i \boldsymbol{u}_i \boldsymbol{u}_i^\intercal$ be the eigenvalue decomposition of $\boldsymbol{V}_{1\boldsymbol{c}_1}$, where w_i and $\boldsymbol{u}_i$ are the eigenvalues and eigenvectors of $\boldsymbol{V}_{1\boldsymbol{c}_1}$, respectively. We further note that BF_1 can be expressed as

$$\begin{aligned} BF_1 &= \frac{\det(\boldsymbol{V}_{1\boldsymbol{c}_1} + \sigma_1^2 \boldsymbol{V}_{1\boldsymbol{c}_1}^2)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\boldsymbol{Z}_{1\boldsymbol{c}_1}^\intercal(\boldsymbol{V}_{1\boldsymbol{c}_1} + \sigma_1^2\boldsymbol{V}_{1\boldsymbol{c}_1}^2)^{-1}\boldsymbol{Z}_{1\boldsymbol{c}_1}\right]}{\det(\boldsymbol{V}_{1\boldsymbol{c}_1})^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{Z}_{1\boldsymbol{c}_1}^\intercal \boldsymbol{V}_{1\boldsymbol{c}_1}^{-1}\boldsymbol{Z}_{1\boldsymbol{c}_1}\right)} \\ &\propto \left(\prod_{i=1}^{p_{\text{causal}}} \frac{1}{1+\sigma_1^2 w_i}\right)^{\frac{1}{2}} \exp\left[\frac{1}{2}\sum_{i=1}^{p_{\text{causal}}} \frac{\sigma_1^2}{1+\sigma_1^2 w_i}\left(\boldsymbol{Z}_{1\boldsymbol{c}_1}^\intercal \boldsymbol{u}_i\right)^2\right], \end{aligned} \tag{22}$$

avoiding numerical instability introduced by small eigenvalues. The Bayes factor for $\boldsymbol{Z}_{2\boldsymbol{c}_2}$ can be obtained using the same approach.

3.6 Posterior probability of each SNP to be ancestry-specific or shared

For each SNP i, we evaluate

$$\Pr(\boldsymbol{C}_i = \boldsymbol{b}|\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{f}^*) \tag{23}$$

for $\boldsymbol{b} \in \{01, 10, 11\}$, the three causal configurations of interest (causal in a single population or both populations), where $\boldsymbol{f}^*$ denotes the estimated MVB parameter. We show below that Equation (23) can be evaluated using the Gibbs sampling procedure outlined in Section 3.5. First, we note that

$$\begin{aligned} \Pr(\boldsymbol{C}_i = \boldsymbol{b}|\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{f}^*) &= \sum_{\boldsymbol{C}_{\neg i}} \Pr(\boldsymbol{C}_i = \boldsymbol{b}, \boldsymbol{C}_{\neg i}|\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{f}^*) \\ &= \sum_{\boldsymbol{C}_{\neg i}} \Pr(\boldsymbol{C}_i = \boldsymbol{b}|\boldsymbol{C}_{\neg i}, \boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{f}^*)\Pr(\boldsymbol{C}_{\neg i}|\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{f}^*) \\ &= \mathrm{E}\left[\Pr(\boldsymbol{C}_i = \boldsymbol{b}|\boldsymbol{C}_{\neg i}, \boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{f}^*)\right] = \mathrm{E}\left[\mathrm{E}[\mathbb{1}_{\{\boldsymbol{C}_i=\boldsymbol{b}\}}|\boldsymbol{C}_{\neg i}, \boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{f}^*]\right] \\ &= \mathrm{E}[\mathbb{1}_{\{\boldsymbol{C}_i=\boldsymbol{b}\}}|\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{f}^*] \approx \frac{\sum_{j=1}^{J} \mathbb{1}_{\{\boldsymbol{C}_i^{(j)}=\boldsymbol{b}\}}}{J}, \end{aligned} \tag{24}$$

where $\boldsymbol{C}^{(j)}$ is the j-th causal status vector sampled from the posterior distribution $\Pr(\boldsymbol{C}|\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{f}^*)$ out of a total of J samples (see Section 3.5). To ensure stable estimates of the posterior probability, we run the Gibbs sampling procedure 20 times and report the average posterior probability.

References

[1] Hilary Finucane, Yakir Reshef, Verneri Anttila, Kamil Slowikowski, Alexander Gusev, Andrea Byrnes, Steven Gazal, Po-Ru Loh, Giulio Genovese, Arpiar Saunders, et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *bioRxiv*, page 103069, 2017.

[2] Bin Dai, Shilin Ding, Grace Wahba, et al. Multivariate bernoulli distribution. *Bernoulli*, 19(4): 1465–1483, 2013.

[3] Huwenbo Shi, Bogdan Pasaniuc, and Kenneth L Lange. A multivariate bernoulli model to predict dnasei hypersensitivity status from haplotype data. *Bioinformatics*, 31(21):3514–3521, 2015.

[4] Huwenbo Shi, Gleb Kichaev, and Bogdan Pasaniuc. Contrasting the genetic architecture of 30 complex traits from summary association data. *The American Journal of Human Genetics*, 99(1):139–153, 2016.

[5] Gleb Kichaev, Wen-Yun Yang, Sara Lindstrom, Farhad Hormozdiari, Eleazar Eskin, Alkes L Price, Peter Kraft, and Bogdan Pasaniuc. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS genetics*, 10(10):e1004722, 2014.

[6] Gleb Kichaev, Megan Roytman, Ruth Johnson, Eleazar Eskin, Sara Lindstroem, Peter Kraft, and Bogdan Pasaniuc. Improved methods for multi-trait fine mapping of pleiotropic risk loci. *Bioinformatics*, 33(2):248–255, 2017.

[7] Christian Benner, Chris CA Spencer, Aki S Havulinna, Veikko Salomaa, Samuli Ripatti, and Matti Pirinen. Finemap: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*, 32(10):1493–1501, 2016.

[8] Farhad Hormozdiari, Emrah Kostem, Eun Yong Kang, Bogdan Pasaniuc, and Eleazar Eskin. Identifying causal variants at loci with multiple signals of association. *Genetics*, 198(2):497–508, 2014.

[9] Tomaz Berisa and Joseph K Pickrell. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics*, 32(2):283, 2016.

[10] Na Cai, Tim B Bigdeli, Warren Kretzschmar, Yihan Li, Jieqin Liang, Li Song, Jingchu Hu, Qibin Li, Wei Jin, Zhenfei Hu, et al. Sparse whole-genome sequencing identifies two loci for major depressive disorder. *Nature*, 523(7562):588, 2015.

[11] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, 2015.

[12] Ludmil B Alexandrov, Serena Nik-Zainal, David C Wedge, Samuel AJR Aparicio, Sam Behjati, Andrew V Biankin, Graham R Bignell, Niccolo Bolli, Ake Borg, Anne-Lise Børresen-Dale, et al. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415, 2013.

[13] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.