Localizing components of shared transethnic genetic architecture of complex traits from GWAS summary data

Huwenbo Shi,[1-3*†] Kathryn S. Burch,[1*†] 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,†]

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

* These authors contributed equally to this work

† Correspondence: H.S. (hshi@hsph.harvard.edu), K.S.B. (kathrynburch@ucla.edu), or B.P. (pasaniuc@ucla.edu)

Abstract

Methods for transethnic genetic analyses rely on assumptions on the transferability of certain components of genetic architecture across ancestry groups (e.g., same causal variants with similar effect sizes in two populations). However, the extent to which these assumptions hold for different complex traits has not been systematically assessed. Here we propose an empirical Bayes framework that models GWAS summary data in two ancestry groups to estimate (1) the prior expected proportions of population-specific/shared causal SNPs and (2) per-SNP posterior probabilities of being causal in one or both populations. Through extensive simulations, we show that our approach yields nearly unbiased estimates of the prior expected proportions of population-specific/shared causal SNPs. We define a statistic to test for enrichment of causal SNPs in a region – the ratio between the posterior and prior expected numbers of causal SNPs in the region – and show that it is conservative at different levels of polygenicity and GWAS power. We analyze publicly available GWAS data for 9 complex traits and diseases in individuals of East Asian and European ancestry and find that most high-posterior SNPs are shared and have highly correlated GWAS marginal effects. In ancestry-specific GWAS risk regions, we observe a 2.8x enrichment of SNPs with high posterior probability of being causal in both populations, suggesting that GWAS associations that do not replicate across populations can be explained in part by differences in sample size, causal effect sizes, and/or LD patterns. Finally, we report enrichments of population-specific/shared causal SNPs in 53 tissue-specific functional genomic annotations.

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, for a given complex trait, the underlying genetic architecture – defined here in terms of 'polygenicity' (the number of variants with nonzero effects) and the coupling of causal effect sizes with minor allele frequency (MAF), linkage disequilibrium (LD), and other genomic features[6–11] – varies among different ancestral populations. While the vast majority of genome-wide association studies (GWAS) to date have been performed in individuals of European descent[12–15], the number of studies performed in individuals of non-European ancestry is steadily increasing[16–22], creating opportunities for well-powered transethnic genetic studies[16,17,19,21,23–28].

Risk regions identified through GWAS tend to replicate across populations[12,16,17,28–30], suggesting that a typical complex trait likely has a genetic component that is shared among populations. Indeed, for certain post-GWAS analyses such as disease mapping[18,26,31] and statistical fine-mapping[23,32–34], 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 (PRS) across populations[5,35–42]; critically, if applied in a clinical setting, existing non-population-matched PRS may exacerbate health disparities among ethnic groups[43]. In addition to the population-specificity of existing PRS, estimates of transethnic genetic correlations less than one[25,44–47] imply that a sizeable portion of causal variants are population-specific and/or that the effects at shared causal variants vary among populations. As an example, consider a single genetic variant that is significantly associated with a trait in two populations: if the biologically causal variants driving the association are specific to each population or have different causal effects, PRS weights inferred from one population will produce inaccurate risk scores in the second population[41,43]. Conversely, two distinct associations in two populations can be driven by a single variant (known as colocalization) that may or may not have the same effect in both populations (Figure 1). Thus, for a single trait in two populations, knowing the number of shared causal variants as well as the genetic correlation at the shared causal variants could improve transethnic analyses (e.g., transferability of PRS across populations[14,35,36,39,40]).

In this work, we introduce PESCA (Population-spEcific/Shared Causal vAriants), an empirical Bayes approach that requires only GWAS summary association statistics and ancestry-matched estimates of LD to infer (1) the prior expected proportions of population-specific and shared causal variants across the entire genome and (2) the per-SNP posterior probabilities of having nonzero causal effects in one or both populations. To leverage the polygenic architecture of complex traits[48–50], we apply our approach genome-wide on a set of LD-pruned SNPs ($\hat{r}_{ij}^2 < 0.95$ for all $i \neq j$), irrespective of their GWAS significance. Through extensive simulations, we show that our method yields accurate estimates of the prior

expected proportions of population-specific/shared causal SNPs and that the per-SNP posterior probabilities of being causal in a single or both populations are well-calibrated. We define a statistic to test for enrichment of population-specific/shared causal SNPs in a region – the ratio between the posterior and prior expected numbers of causal SNPs in the region – and show that the test statistic is conservative at different levels of polygenicity and GWAS power. We note that our definition of enrichment is related to, but conceptually distinct from, definitions of SNP-heritability enrichment used in previous studies[7,51]. 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, whereas 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.

We apply our approach to publicly available GWAS summary statistics for 9 complex traits and diseases in individuals of East Asian (EAS) and European (EUR) ancestry (average N_{EAS} = 94,621, N_{EUR} = 103,507, MAF > 5%) (Table 1), using the 1000 Genomes Phase 3 reference panel[52] to estimate ancestry-matched LD. On average across the 9 traits, we estimate that approximately 80% (S.D. 15%) of the prior expected number of causal SNPs in EAS individuals and 84% (S.D. 8%) of those in EUR individuals are shared by the other population. Consistent with previous studies based on SNP-heritability[49,50], we find that high-posterior SNPs are distributed relatively evenly across the genome. We observe that population-specific GWAS risk regions have, on average across the 9 traits, a 2.8x enrichment of shared high-posterior SNPs relative to the genome-wide background, suggesting that many population-specific GWAS-significant associations can be explained by differences in LD patterns, GWAS power (including different causal effect sizes), and/or gene-environment interactions. At SNPs with posterior probability > 0.8 of being causal in both populations, we observe a strong correlation between EAS and EUR marginal effects, concordant with previously reported estimates of transethnic genetic correlation (average $\hat{\rho}_g = 0.79 \pm 0.07$ s.e.m. across the 9 traits)[45]. Finally, we show that regions flanking genes that are specifically expressed in trait-relevant tissues[53] harbor a disproportionate number of SNPs with high posterior probability of being causal (in both or a single population) relative to the genome-wide background. Our results suggest that, for a typical complex trait, the genetic architecture at common SNPs is highly similar between populations and can be leveraged to improve trans-ethnic analyses.

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[54,55]

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

which facilitates optimization and interpretation (Supplementary Note). 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, $\mathbf{Z}_1$ and $\mathbf{Z}_2$ (both are $p \times 1$ vectors), conditional on the causal status vectors for each population, $\boldsymbol{c}_1 = (c_{11}, \cdots, c_{p1})^T$ and $\boldsymbol{c}_2 = (c_{12} \cdots c_{p2})^T$. Given $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$, $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ are independent. Thus, for population j,

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

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 $\sigma_j^2 = \frac{n_j h_{gj}^2}{|\boldsymbol{c}_j|}$, 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 (Supplementary Note).

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:

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

where $\boldsymbol{f} = (f_{00}, f_{01}, f_{10}, f_{11})$ is the vector of parameters of the MVB distribution. In practice, we partition the genome into approximately independent regions[56] and model the distribution of Z-scores at all regions as the product of the distribution of Z-scores in each region (Supplementary Note).

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 $Q(\boldsymbol{f} | \boldsymbol{f}^{(t)})$ using an efficient Gibbs sampler; in the maximization step, we maximize $Q(\boldsymbol{f} | \boldsymbol{f}^{(t)})$ using analytical formulae (Supplementary Note). 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 (Supplementary Note).

Evaluating per-SNP posterior probabilities to be population-specific or shared

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 (Supplementary Note).

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 $\mathrm{E}[q_{11}|\boldsymbol{Z}_1, \boldsymbol{Z}_2; \boldsymbol{f}^*] = \sum_{i=1}^{p} \mathrm{E}\left[1_{\{\boldsymbol{C}_i=11\}}\middle|\boldsymbol{Z}_1, \boldsymbol{Z}_2; \boldsymbol{f}^*\right] = \sum_{i=1}^{p} \Pr\,(\boldsymbol{C}_i = 11|\boldsymbol{Z}_1, \boldsymbol{Z}_2; \boldsymbol{f}^*)$. 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[56] to define LD blocks within the transethnic LD matrix. Applying this procedure using the 1000 Genomes Phase 3 reference panel[52] to create the transethnic LD

150 matrix produces 1,368 LD blocks (average length of 2-Mb) that are approximately independent in
151 individuals of East Asian and European ancestry.
152

153 **Enrichment of population-specific/shared causal SNPs in functional annotations**

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

158 $$\hat{\alpha}_{k,\boldsymbol{b}} = \frac{E[q_{k,\boldsymbol{b}}|\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{f}^*]}{E[q_{k,\boldsymbol{b}}|\boldsymbol{f}^*]} = \frac{\sum_{i \in \psi(k)} \Pr(\boldsymbol{C}_i = \boldsymbol{b}|\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{f}^*)}{p_k \Pr(\boldsymbol{C}_i = \boldsymbol{b})}$$

159 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 ($\boldsymbol{b} = 11$)
160 causal variants, $\psi(k)$ is the set of SNPs in functional annotation k, and p_k is the number of SNPs in
161 functional annotation k. The numerator, $E[q_{k,\boldsymbol{b}}|\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{f}^*]$, and denominator, $E[q_{k,\boldsymbol{b}}|\boldsymbol{f}^*]$, represent the
162 posterior (conditioned on Z-scores) and prior expected numbers of causal SNPs in functional annotation k,
163 respectively. We estimate the standard error of $\hat{\alpha}_{k,\boldsymbol{b}}$ using block jackknife over 1,368 non-overlapping
164 approximately LD-independent blocks across the entire genome. The resulting enrichment test statistics,
165 $\frac{\hat{\alpha}_{k,\boldsymbol{b}}-1}{\mathrm{SE}(\hat{\alpha}_{k,\boldsymbol{b}})}$, approximately follow a t-distribution with degrees of freedom equal to the number of blocks minus
166 one[57]. Since we are interested in identifying categories of SNPs that harbor more population-specific/shared
167 causal SNPs than expected (i.e. enrichment > 1), we report *P*-values from a one-tailed t-test where the null
168 hypothesis is enrichment ≤ 1.

169 We note that our definition of enrichment of causal SNPs is related to, but conceptually different
170 from, enrichment of SNP-heritability[7,51,58]. A positive enrichment of causal SNPs in a functional category
171 indicates that, compared to the genome-wide background, there are more causal SNPs in that category than
172 expected; a positive enrichment of SNP-heritability in a category indicates that the average per-SNP effect
173 size in the category is larger than the genome-wide average per-SNP effect size.
174

175 **Simulation framework**

176 We used real chromosome 22 genotypes of 10,000 individuals of East Asian ancestry from
177 CONVERGE[59,60] and 50,000 individuals of white British ancestry from the UK Biobank[61,62] to simulate
178 causal effects and phenotypes. First, we used PLINK[63] (v1.9) to perform LD pruning at a threshold of 0.95
179 in the 1000 Genomes Phase 3 reference panel[52]; this process drops SNPs such that there is no pair of SNPs
180 for which $r_{ij}^2 > 0.95$ ($i \neq j$) in the reference panel. 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 $\boldsymbol{y_1} = \boldsymbol{X_1}\boldsymbol{\beta_1} + \boldsymbol{\epsilon_1}$ (population 1) and $\boldsymbol{y_2} = \boldsymbol{X_2}\boldsymbol{\beta_2} + \boldsymbol{\epsilon_1}$ (population 2). We assume the phenotypes are standardized within each population such that $\mathrm{E}[\boldsymbol{y_1}] = \boldsymbol{0}$, $\mathrm{Var}[\boldsymbol{y_1}] = \mathbf{I}$ and $\mathrm{E}[\boldsymbol{y_2}] = \boldsymbol{0}$, $\mathrm{Var}[\boldsymbol{y_2}] = \mathbf{I}$. 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

$$\boldsymbol{\beta}_{1\boldsymbol{c}_1}|\boldsymbol{c}_1 \sim N\left(\boldsymbol{0}, \frac{h_{g1}^2}{|\boldsymbol{c}_1|}\mathbf{I}_{\boldsymbol{c}_1}\right), \qquad \boldsymbol{\beta}_{2\boldsymbol{c}_2}|\boldsymbol{c}_2 \sim N\left(\boldsymbol{0}, \frac{h_{g2}^2}{|\boldsymbol{c}_2|}\mathbf{I}_{\boldsymbol{c}_2}\right)$$

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_{g1}^2 and h_{g2}^2 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 $\epsilon_{1n} \sim N(0, 1 - h_{g1}^2)$ and $\epsilon_{2n} \sim N(0, 1 - h_{g2}^2)$.

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

Application to 9 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[64], which induces downward-bias in the estimated SNP-heritability; we circumvent 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 and perform LD pruning with PLINK[63] (v.19) such that $\hat{r}_{ij}^2 < 0.95$ for all SNPs $i \neq j$ in both ancestry-matched 1000 Genomes[65] 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[58], each of which represents a set of genes that are "specifically expressed" in a GTEx[66] 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).

212 # Results

213 ## Performance of PESCA in simulations

214 We assessed the performance of PESCA in simulations starting from real genotypes of individuals with
215 East Asian[59,60] (EAS) or European[61,62] (EUR) ancestry (N_{EAS} = 10K, N_{EUR} = 50K, M = 8,599 SNPs)
216 (Methods). First, we find that when in-sample LD from the GWAS is available, PESCA yields
217 approximately unbiased estimates of the prior expected numbers of population-specific/shared causal SNPs
218 (Figure 2, top panel). For example, in simulations where we randomly selected 50 EAS-specific, 50 EUR-
219 specific, and 50 shared causal SNPs, we obtained estimates (and corresponding standard errors) of 37.8
220 (4.5) EAS-specific, 40.3 (4.9) EUR-specific, and 64.9 (6.3) shared causal SNPs, respectively. When
221 external reference LD is used (in this case, from 1000 Genomes[52]), PESCA yields a slight upward bias
222 (Figure 2, bottom panel); on the same simulated data, we obtained estimates of 48.0 (5.9) EAS-specific,
223 53.7 (7.44) EUR-specific, and 78.8 (7.6) shared causal SNPs. We observe a slight decrease in accuracy as
224 the product of SNP-heritability and sample size ($N \times h_g^2$) decreases (Figure S1-5). This is expected as the
225 likelihood of the GWAS summary statistics is a function of $N \times h_g^2$ (Methods) – as $N \times h_g^2$ decreases,
226 GWAS summary statistics provide less information on the causal status of each SNP. In general, we
227 recommend applying PESCA to GWAS summary statistics for which $N \times h_g^2 > 2000$.

228 Next, we evaluated the accuracy of the per-SNP posterior probabilities of being causal in a single
229 population (EAS only or EUR only) or in both populations (Methods). For the three causal configurations
230 of interest (EAS only, EUR only, and shared), we observe an increase in the average correlation between
231 the per-SNP posterior probabilities of having that causal configuration and the true causal status vectors for
232 that configuration as $N \times h_g^2$ increases and as the total number of causal SNPs decreases (i.e. as per-SNP
233 causal effect sizes increase) (Figure S6-7). As expected, as the simulated proportion of shared causal SNPs
234 increases, the average correlation between the posterior probabilities of being shared and the true causal
235 status vectors for being shared increases; for the population-specific causal configurations, the average
236 correlation between the posterior probabilities and true causal status vectors decreases (Figure S6-7). We
237 then assessed whether our proposed statistics for testing enrichment of population-specific/shared causal
238 SNPs in functional annotations (Methods) are well-calibrated under the null hypothesis of no enrichment
239 (note that our definition of enrichment is related to, but distinct from, the definition of SNP-heritability
240 enrichment used in previous studies[7,51]). Overall, when both population-specific and shared causal SNPs
241 are drawn at random, the enrichment test statistics are conservative at different levels of polygenicity and
242 GWAS power ($N \times h_g^2$), irrespective of whether in-sample LD or external reference LD is used (Figure S8-
243 13).

Finally, we evaluated the computational efficiency of each stage of inference. In the first stage of estimating the prior expected proportions of population-specific/shared causal SNPs, PESCA typically converged within 200 EM iterations (Figure S14-16), with run time increasing with the number of simulated causal SNPs (Figure S17). 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 minutes to obtain estimates in simulations with 20 randomly selected causal variants and 360 minutes 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 (Methods). In the second stage of inference – evaluating posterior probabilities for each SNP – PESCA took an average of 5 minutes in simulations with 20 causal variants and 28 minutes in simulations with 100 causal variants (Figure S17). We note that both stages of the inference procedure can be parallelized to decrease run time.

Expected genome-wide proportions of shared causal SNPs for 9 complex traits

We obtained publicly available GWAS summary statistics for 9 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 prior expected proportions of population-specific/shared common causal SNPs (MAF > 5%) (Methods). To ensure convergence, we applied 750 EM iterations for each trait (Figure S18-20). While the 9 traits display a range in the expected numbers of common causal SNPs in each population (the sum of the expected numbers of population-specific and shared causal SNPs), this is explained in part by differences in GWAS sample size. For example, in both EAS and EUR, we estimate that approximately 0.2% of common SNPs are causal for rheumatoid arthritis (N_{EAS} = 23K, N_{EUR} = 58K) and that 10% are causal for BMI (N_{EAS} = 225K, N_{EUR} = 158K) (Table 1). Our estimated proportion of causal SNPs for BMI in EUR is consistent with a previously reported estimate obtained from individual-level data[67]. While there is heterogeneity in the estimated proportions of shared causal SNPs across the 9 traits, we find that most common causal SNPs are shared between the populations, consistent with findings from previous studies[28]. For example, for BMI, we estimate that approximately 96% of the common causal SNPs for 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 the other population (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 aggregate the per-SNP posterior probabilities of a given causal configuration in each region to obtain the posterior expected number of SNPs with that causal configuration in each region (Methods). For the 9 traits analyzed, high-posterior SNPs for both the population-specific and shared causal configurations are spread nearly

uniformly across the genome (Figure 3, Figure S21-28). For example, mean corpuscular hemoglobin (MCH) harbored, on average, 0.68 (S.D. 0.42) EAS-specific, 0.53 (S.D. 0.40) EUR-specific, and 2.19 (S.D. 1.46) shared high-posterior SNPs per region (Figure 3, Figure S22, S26). 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 (Figure S29-31), recapitulating previous findings based on regional SNP-heritability[49,50].

Distributions of high-posterior SNPs across GWAS risk regions

We investigated whether genomic regions harboring population-specific GWAS-significant associations contain SNPs with high posterior probabilities of being causal in one or both populations. We aggregate per-SNP posterior probabilities within GWAS risk regions that are EAS-specific, EUR-specific, or shared by both populations and find that all GWAS risk regions harbor at least one common SNP with high posterior probability of being causal in both populations (Figure 4, Figure S32–36), recapitulating previous findings on allelic heterogeneity of complex traits[50,68,69]. For example, EAS-specific and EUR-specific GWAS risk regions for mean corpuscular hemoglobin (MCH) harbor, on average, 3.0 (S.D. 1.7) and 3.3 (S.D. 1.5) SNPs with high posterior probability of being causal in both populations, respectively (Figure 4). We do not observe a noticeable difference in the posterior expected numbers of EAS-specific and EUR-specific SNPs in the population-specific GWAS risk regions. These results are consistent with a previous study suggesting that GWAS risk regions that do not replicate across populations can likely be explained in part by heterogeneity in LD patterns and causal effect sizes between the two populations[28]. Regions containing no GWAS-significant associations for either population harbor, on average, approximately two shared high-posterior SNPs (Figure 4, Figure S32–36).

For each of the causal configurations of interest (EAS-specific, EUR-specific, or shared), we examine the relationship between the EAS and EUR GWAS marginal effects at SNPs with posterior probability > 0.8 (Figure 5). Across the 9 traits, the majority of EAS-specific high-posterior SNPs are nominally significant ($p_{GWAS} < 5 \times 10^{-6}$) either in the EAS GWAS only or in both GWASs; none are nominally significant in the EUR GWAS only. 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. The correlations between the EAS and EUR marginal effects for all three sets of high-posterior SNPs are relatively strong (Pearson r^2 of 0.79 [EAS-specific], 0.73 [EUR-specific], and 0.80 [shared]) and are driven by SNPs that are nominally significant in both GWASs (Figure 5).

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 investigated whether there is an enrichment of population-specific and/or shared causal SNPs in the same 53 tissue-specific SEG annotations[53]. For a given causal configuration, enrichment in an annotation is defined as the ratio between the posterior and prior expected numbers of SNPs with that causal configuration in the annotation (Methods). For 8 of the 9 traits, we find significant enrichment of shared high-posterior SNPs in at least one SEG annotation after Bonferroni correction (P-value < 0.05/53 to correct for 53 tests per trait) (Figure 6, Figure S37–41). We do not find enrichment of any high-posterior SNPs in any SEG annotation for major depressive disorder (MDD) (Figure 6, Figure S41), which could be due to low GWAS sample sizes (Table 1).

Next, we investigated whether there is heterogeneity across traits in the enrichments of high-posterior SNPs among different SEG annotations. For HDL and triglycerides, we find enrichment of EAS-specific (but not EUR-specific) high-posterior SNPs in at least one SEG annotation (Figure 6, Figure S39-40); for LDL, total cholesterol, and rheumatoid arthritis, we find enrichment of EUR-specific (but not EAS-specific) high-posterior SNPs in at least one SEG annotation (Figure 6, Figure S39-41). We find significant enrichment of both population-specific and shared high-posterior SNPs in several SEG annotations for the two hematological traits, mean corpuscular hemoglobin (MCH) and mean corpuscular volume (MCV) (Figure 6, Figure S38). For example, for MCV, we observe a 1.3x (S.E. 0.0041, $P = 5.3 \times 10^{-14}$), 1.1x (S.E. 0.0031, $P = 2.0 \times 10^{-4}$), and 1.3x (S.E. 0.0062, $P = 3.6 \times 10^{-6}$) enrichment of shared, EAS-specific and EUR-specific high-posterior SNPs, respectively, in the SEG annotation for whole blood.

Discussion

We have presented PESCA, a method for identifying SNPs with high posterior probability of being causal in a single population (population-specific) or in two populations (shared) from GWAS summary statistics and estimates of LD. We show that, using either in-sample or external reference LD, PESCA yields accurate and robust estimates of the prior expected proportions of population-specific/shared causal SNPs and well-calibrated enrichment test statistics. We applied PESCA to East Asian- and European-ancestry GWAS summary statistics for 9 complex traits and find that, while there is a non-negligible number of population-specific causal SNPs, most common causal SNPs in East Asian and European populations are shared with the other population. We observe that regions that harbor GWAS risk variants for one population are enriched with SNPs with high-posterior probability of being causal in both populations, and that population-specific high-posterior SNPs are distributed relatively evenly across all regions (whether or not they contain

343 a GWAS-significant association). For all traits except MDD, we find enrichments of population-specific
344 and/or shared high-posterior SNPs in tissue-specific SEG annotations[58] relative to the genome-wide
345 background. Overall, we conclude that many population-specific GWAS risk regions are likely be
346 explained by differences in LD patterns and GWAS power, including differences in causal effect sizes due
347 to different SNP-heritabilities and/or gene-environment interactions among populations. The vast majority
348 of common high-posterior SNPs are shared between populations, motivating the development of new
349 methodologies that can leverage this shared architecture to improve transferability of results from well-
350 powered European-ancestry studies to other understudied populations.

351 We conclude by discussing the caveats and limitations of our analyses as well as possible future
352 directions. First, the estimated proportions of population-specific causal SNPs must be interpreted with
353 caution as they can be influenced by gene-environment interactions. For example, if a SNP has a nonzero
354 effect on a trait only in the presence of environmental factors that are specific to individuals of East Asian
355 ancestry, PESCA will interpret that SNP as an EAS-specific causal SNP even though it can have a nonzero
356 effect in Europeans in the presence of the same environmental factors. Second, we restricted our analyses
357 to an LD-pruned set of SNPs with MAF > 5% in both populations to reduce noise in the LD matrices
358 estimated from external reference panels. Thus, the estimates we report in this work do not capture nonzero
359 effects at low frequency or rare variants that are not well-tagged by common SNPs. Furthermore, since
360 most common variants are shared among continental populations and rarer variants tend to be localized to
361 more closely related populations[65], this design undersamples population-specific causal variants. However,
362 lower MAF thresholds can be used if in-sample LD is available; in light of ongoing efforts at several
363 institutions to establish biobanks[61,62,70–72], we believe that well-powered GWASs with in-sample LD will
364 become increasingly available for more diverse populations. We also note that for the purpose of improving
365 transferability of polygenic risk scores (PRS) across populations, prediction accuracy depends largely on
366 the accuracy of the PRS weights at common SNPs (the average per-SNP contribution to total SNP-
367 heritability is larger for common SNPs than for low frequency or rare variants[48]). Third, for computational
368 efficiency, PESCA relies on having regions that are approximately LD-independent in both populations; if
369 there is LD leakage between regions, the estimated proportions of causal SNPs will be biased. We therefore
370 recommend defining LD blocks for each pair of populations one analyzes. Finally, PESCA does not
371 explicitly model cross-population correlations of effect sizes at shared causal variants; we conjecture that
372 modeling these correlations can further improve performance.
373

374 ## Acknowledgements

375 We are grateful to Alkes L. Price and Steven Gazal for helpful discussions that greatly improved the quality
376 of this manuscript. We also thank Sriram Sankararaman, Jonathan Flint, and the UK Biobank (application

#33297) for providing access to datasets that made this work possible. This work was funded in part by the National Institutes of Health (NIH) under awards R01HG009120, R01MH115676, U01CA194393, T32NS048004, T32MH073526, and T32HG002536.

Declaration of Interests

The authors declare no competing interests.

Web Resources

GIANT consortium GWAS summary statistics:
http://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files
Biobank Japan GWAS summary statistics:
http://jenger.riken.jp/en/result
GWAS summary statistics for hematological traits:
http://www.bloodcellgenetics.org
LD score regression:
https://github.com/bulik/ldsc
PLINK 1.9:
https://www.cog-genomics.org/plink/1.9/
Popcorn:
https://github.com/brielin/Popcorn
PESCA:
https://github.com/huwenboshi/posc
Specifically expressed genes:
https://data.broadinstitute.org/alkesgroup/LDSCORE/LDSC_SEG_ldscores

References

1. Campbell, M. C. & Tishkoff, S. A. The evolution of human genetic and phenotypic variation in Africa. *Curr. Biol.* **20**, R166--R173 (2010).
2. Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. Demic expansions and human evolution. *Science (80-.).* **259**, 639–646 (1993).
3. Pritchard, J. K., Pickrell, J. K. & Coop, G. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr. Biol.* **20**, R208--R215 (2010).
4. Laland, K. N., Odling-Smee, J. & Myles, S. How culture shaped the human genome: bringing genetics and the human sciences together. *Nat. Rev. Genet.* **11**, 137 (2010).
5. Martin, A. R. *et al.* Human demographic history impacts genetic risk prediction across diverse populations. *Am. J. Hum. Genet.* **100**, 635–649 (2017).
6. Timpson, N. J., Greenwood, C. M. T., Soranzo, N., Lawson, D. J. & Richards, J. B. Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nat. Rev. Genet.* **19**, 110 (2017).
7. Gazal, S. *et al.* Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* **49**, 1421 (2017).
8. Gazal, S. *et al.* Functional architecture of low-frequency variants highlights strength of negative selection across coding and non-coding annotations. *Nat. Genet.* **50**, 1600–1607 (2018).
9. O'Connor, L. J. *et al.* Extreme Polygenicity of Complex Traits Is Explained by Negative Selection. *Am. J. Hum. Genet.* (2019). doi:https://doi.org/10.1016/j.ajhg.2019.07.003
10. Speed, D., Hemani, G., Johnson, M. R. & Balding, D. J. Improved Heritability Estimation from Genome-wide SNPs. *Am. J. Hum. Genet.* **91**, 1011–1021 (2012).
11. Speed, D. *et al.* Reevaluation of SNP heritability in complex human traits. *Nat. Genet.* **49**, 986 (2017).
12. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
13. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538**, 161 (2016).
14. Rosenberg, N. A. *et al.* Genome-wide association studies in diverse populations. *Nat. Rev. Genet.* **11**, 356 (2010).
15. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five Years of GWAS Discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012).
16. Kanai, M. *et al.* Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nat. Genet.* **50**, 390 (2018).
17. Akiyama, M. *et al.* Genome-wide association study identifies 112 new loci for body mass index in the Japanese population. *Nat. Genet.* **49**, 1458 (2017).
18. Li, Z. *et al.* Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia. *Nat. Genet.* **49**, 1576 (2017).
19. Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat Genet* **47**, 979–986 (2015).
20. Ng, M. C. Y. *et al.* Meta-analysis of genome-wide association studies in African Americans provides insights into the genetic architecture of type 2 diabetes. *PLoS Genet.* **10**, e1004517 (2014).
21. Franceschini, N. *et al.* 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 (2013).
22. Schick, U. M. *et al.* Genome-wide association study of platelet count identifies ancestry-specific loci in Hispanic/Latino Americans. *Am. J. Hum. Genet.* **98**, 229–242 (2016).

23. Kichaev, G. & Pasaniuc, B. Leveraging Functional-Annotation Data in Trans-ethnic Fine-Mapping Studies. *Am. J. Hum. Genet.* **97**, 260–271 (2015).

24. Mancuso, N. *et al.* The contribution of rare variation to prostate cancer heritability. *Nat. Genet.* **48**, 30 (2015).

25. Brown, B. C. *et al.* Transethnic genetic-correlation estimates from summary statistics. *Am. J. Hum. Genet.* **99**, 76–88 (2016).

26. Morris, A. P. Transethnic meta-analysis of genomewide association studies. *Genet. Epidemiol.* **35**, 809–822 (2011).

27. Lam, M. *et al.* Comparative genetic architectures of schizophrenia in East Asian and European populations. *Nat. Genet.* (2019). doi:10.1038/s41588-019-0512-x

28. Marigorta, U. M. & Navarro, A. High trans-ethnic replicability of GWAS results implies common causal variants. *PLoS Genet.* **9**, e1003566 (2013).

29. Kraft, P., Zeggini, E. & Ioannidis, J. P. A. Replication in genome-wide association studies. *Stat. Sci. A Rev. J. Inst. Math. Stat.* **24**, 561 (2009).

30. Li, Y. R. & Keating, B. J. Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations. *Genome Med.* **6**, 91 (2014).

31. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381 (2014).

32. Wu, Y. *et al.* Trans-Ethnic Fine-Mapping of Lipid Loci Identifies Population-Specific Signals and Allelic Heterogeneity That Increases the Trait Variance Explained. *PLOS Genet.* **9**, e1003379 (2013).

33. Asimit, J. L. *et al.* Stochastic search and joint fine-mapping increases accuracy and identifies previously unreported associations in immune-mediated diseases. *Nat. Commun.* **10**, 3216 (2019).

34. Zaitlen, N., Paşaniuc, B., Gur, T., Ziv, E. & Halperin, E. Leveraging genetic variability across populations for the identification of causal variants. *Am. J. Hum. Genet.* **86**, 23–33 (2010).

35. Vilhjálmsson, B. J. *et al.* Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).

36. Márquez-Luna, C., Loh, P.-R. & Price, A. L. Multiethnic polygenic risk scores improve risk prediction in diverse populations. *Genet. Epidemiol.* **41**, 811–823 (2017).

37. Lewis, C. M. & Vassos, E. Prospects for using risk scores in polygenic medicine. *Genome Med.* **9**, 96 (2017).

38. Curtis, D. Polygenic risk score for schizophrenia is more strongly associated with ancestry than with schizophrenia. *Psychiatr. Genet.* **28**, (2018).

39. Chen, C.-Y., Han, J., Hunter, D. J., Kraft, P. & Price, A. L. Explicit Modeling of Ancestry Improves Polygenic Risk Scores and BLUP Prediction. *Genet. Epidemiol.* **39**, 427–438 (2015).

40. Wojcik, G. L. *et al.* Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**, 514–518 (2019).

41. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The Missing Diversity in Human Genetic Studies. *Cell* **177**, 26–31 (2019).

42. Gurdasani, D., Barroso, I., Zeggini, E. & Sandhu, M. S. Genomics of disease risk in globally diverse populations. *Nat. Rev. Genet.* **20**, 520–535 (2019).

43. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).

44. Ikeda, M. *et al.* Genome-Wide Association Study Detected Novel Susceptibility Genes for Schizophrenia and Shared Trans-Populations/Diseases Genetic Effect. *Schizophr. Bull.* **45**, 824–834 (2019).

45. Shi, H. *et al.* Population-specific causal disease effect sizes in functionally important regions impacted by selection. *bioRxiv* 803452 (2019). doi:10.1101/803452

46. Galinsky, K. J. *et al.* Estimating cross-population genetic correlations of causal effect sizes. *Genet. Epidemiol.* **43**, 180–188 (2019).

47. Guo, J. *et al.* Quantifying genetic heterogeneity between continental populations for human height and body mass index. *bioRxiv* 839373 (2019). doi:10.1101/839373

48. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).

49. Loh, P.-R. *et al.* Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat. Genet.* **47**, 1385 (2015).

50. Shi, H., Kichaev, G. & Pasaniuc, B. Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. *Am. J. Hum. Genet.* **99**, 139–153 (2016).

51. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**, 1228–1235 (2015).

52. Consortium, 1000 Genomes Project & others. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).

53. Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).

54. Dai, B., Ding, S., Wahba, G. & others. Multivariate bernoulli distribution. *Bernoulli* **19**, 1465–1483 (2013).

55. Shi, H., Pasaniuc, B. & Lange, K. L. A multivariate Bernoulli model to predict DNaseI hypersensitivity status from haplotype data. *Bioinformatics* **31**, 3514–3521 (2015).

56. Berisa, T. & Pickrell, J. K. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics* **32**, 283–285 (2016).

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

58. Finucane, H. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *bioRxiv* 103069 (2017).

59. Cai, N. *et al.* Sparse whole-genome sequencing identifies two loci for major depressive disorder. *Nature* **523**, 588 (2015).

60. Cai, N. *et al.* 11,670 whole-genome sequences representative of the Han Chinese population from the CONVERGE project. *Sci. data* **4**, 170011 (2017).

61. Sudlow, C. *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 Med.* **12**, e1001779 (2015).

62. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

63. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

64. Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).

65. Consortium, T. 1000 G. P. *et al.* A global reference for human genetic variation. *Nature* **526**, 68 (2015).

66. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580 (2013).

67. Zeng, J. *et al.* Signatures of negative selection in the genetic architecture of human complex traits. *Nat. Genet.* **50**, 746–753 (2018).

68. Hormozdiari, F. *et al.* Widespread allelic heterogeneity in complex traits. *Am. J. Hum. Genet.* **100**, 789–802 (2017).

69. Gusev, A. *et al.* Quantifying Missing Heritability at Known GWAS Loci. *PLOS Genet.* **9**, e1003993 (2013).

70. Nagai, A. *et al.* Overview of the BioBank Japan Project: Study design and profile. *J. Epidemiol.* **27**, S2–S8 (2017).

71. Gaziano, J. M. *et al.* Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol.* **70**, 214–223 (2016).

72. Leitsalu, L. *et al.* Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *Int. J. Epidemiol.* **44**, 1137–1147 (2015).

73. Astle, W. J. *et al.* The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell* **167**, 1415–1429 (2016).

74. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707 (2010).

75. Wray, N. R., Sullivan, P. F. & others. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *bioRxiv* 167577 (2017).

76. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **advance on**, 291–295 (2015).

Trait name (abbrev.)	Pop.	Ref.	$\widehat{h}_g^2$ (S.E.) %	Sample size (N)	Total # SNPs (MAF > 5%)	EAS-specific causals (S.E.)	EUR-specific causals (S.E.)	Shared causals (S.E.)	$\widehat{\rho}_g$ (S.E.)[45]
Body Mass Index (BMI)	EAS	17	19.8 (0.64)	224,698	258,130	982 (2) 0.4%	1,033 (2) 0.4%	25,641 (16) 10%	0.80 (0.02)
	EUR	64	20.6 (0.91)	158,284					
Mean Corpuscular Hemoglobin (MCH)	EAS	16	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	73	22.7 (3.2)	172,332					
Mean Corpuscular Volume (MCV)	EAS	16	21.0 (2.13)	108,256	480,678	1004 (4) 0.2%	737 (5) 0.2%	3,256 (8) 0.7%	0.89 (0.05)
	EUR	73	23.6 (3.1)	172,433					
High Density Lipoprotein (HDL)	EAS	16	20.7 (3.03)	70,657	268,198	3,167 (12) 1%	652 (2) 0.2%	4,789 (9) 2%	0.89 (0.06)
	EUR	74	16.4 (2.2)	89,614					
Low Density Lipoprotein (LDL)	EAS	16	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	74	13.6 (1.93)	85,491					
Total Cholesterol (TC)	EAS	16	8.1 (0.84)	128,305	268,197	1,892 (3) 0.7%	1,493 (5) 0.6%	5,058 (12) 2%	0.91 (0.07)
	EUR	74	22.5 (2.1)	89,865					
Triglyceride (TG)	EAS	16	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	74	13.6 (2.2)	86,502					
Major Depressive Disorder (MDD)	EAS	59	35.6 (3.4)	10,640	389,593	88 (4) 0.02%	3,280 (6) 0.84%	7,830 (6) 2%	0.34 (0.07)
	EUR	75	19.0 (1.8)	18,759					
Rheumatoid Arthritis (RA)	EAS	31	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	31	9.5 (1.9)	58,284					

Table 1: **Estimated numbers and percentages of population-specific/shared common causal SNPs for 9 complex traits.** We estimated genome-wide SNP-heritability using LD score regression[76] with the intercept constrained to 1 (i.e. assuming no population stratification). Trans-ethnic genetic correlation estimates ($\widehat{\rho}_g$) computed from a similar set of summary statistics were obtained from a previous study[45]. Standard errors of the estimated numbers of population-specific/shared causal SNPs were computed using the last 50 iterations of the EM-MCMC algorithm.

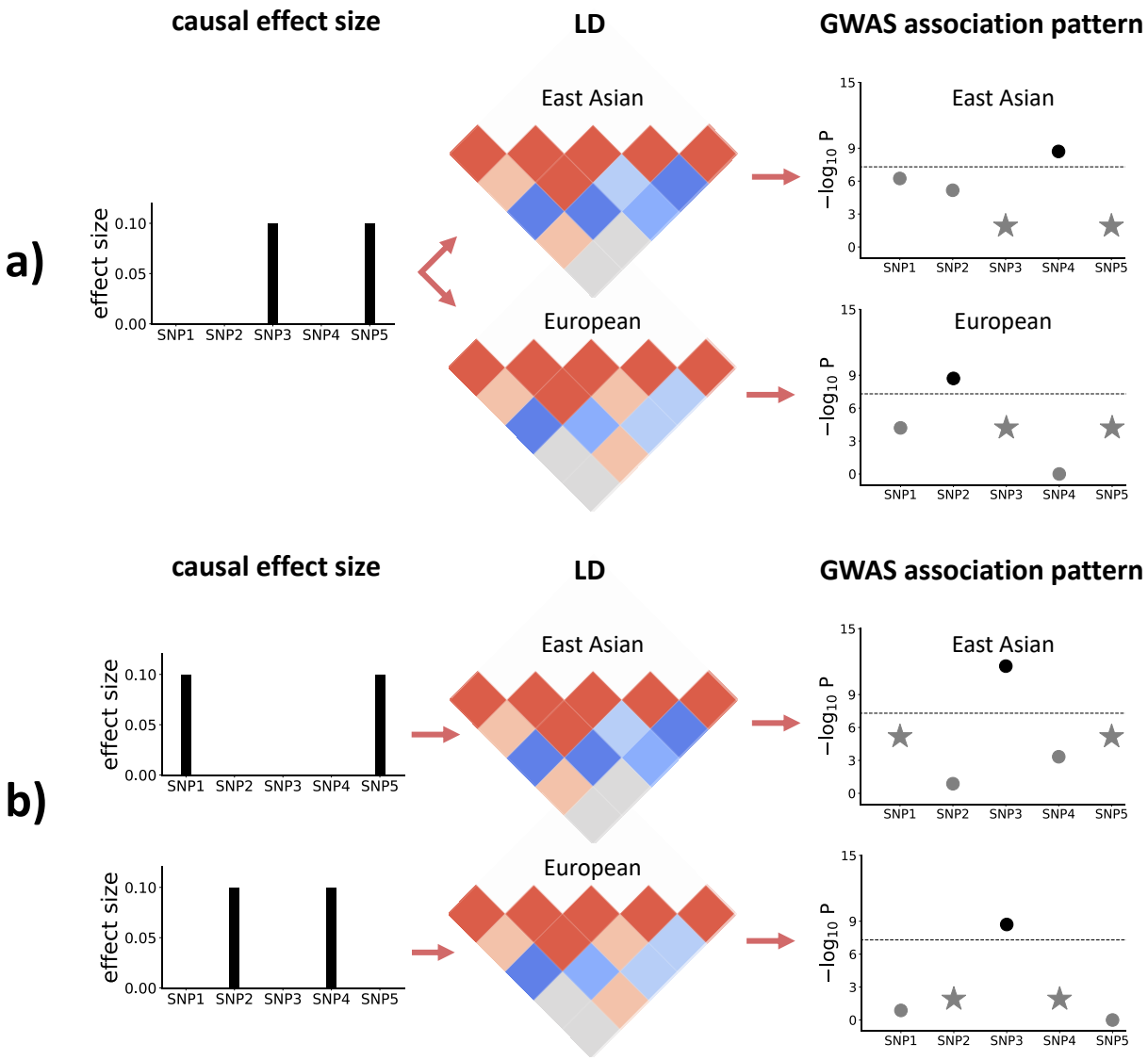

Figure 1: **Toy examples to illustrate how population-specific LD patterns affect GWAS associations.** a) SNPs 3 and 5 are causal in both East Asians and Europeans and have the same population-specific causal effect size of 0.1. However, due to different LD patterns in East Asians and Europeans, SNPs 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 SNPs with true nonzero effects in the; the GWAS-significant SNP is highlighted in a darker color.

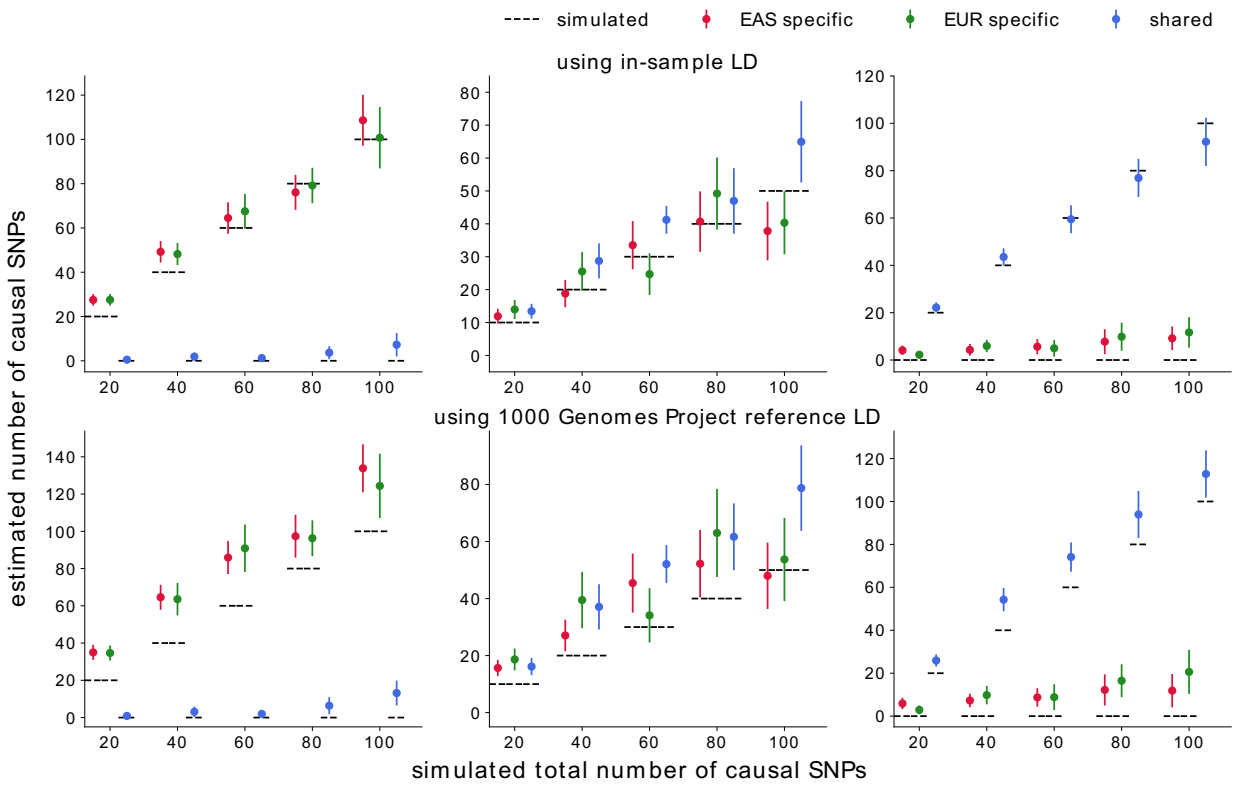

Figure 2**: Performance of PESCA in simulations.** PESCA yields approximately unbiased estimates of the prior expected numbers of population-specific and shared causal SNPs in simulations when in-sample LD is used (top panel), and upward-biased estimates when external reference LD is used (bottom panel). For both populations, we simulate such that the product of the SNP-heritability of the trait and sample size of the GWAS is 500. Mean and standard errors were obtained from 25 independent simulations. Error bars represent ±1.96 of the standard error.

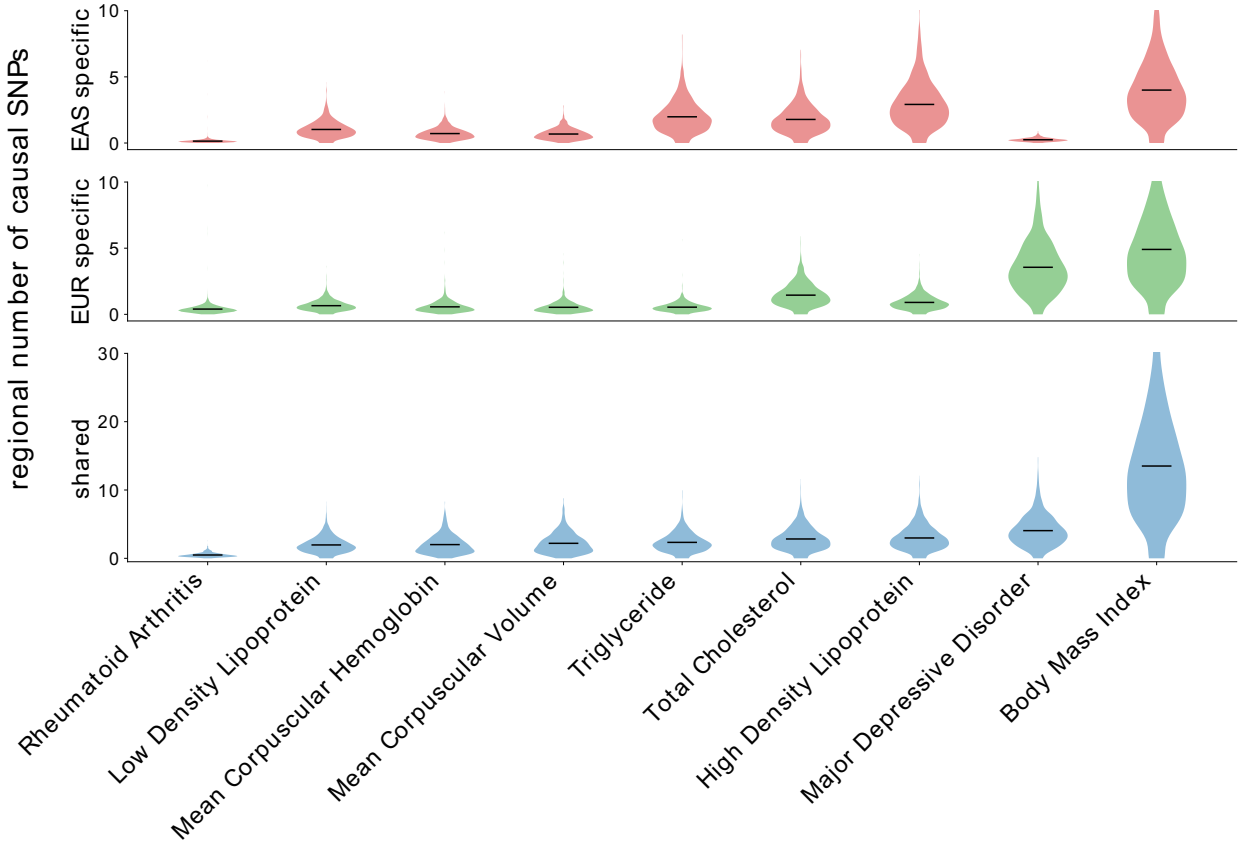

Figure 3: **Distribution of the numbers of population-specific and shared causal SNPs per region.** Each violin plot represents the distribution of the posterior expected number of population-specific or shared causal SNPs per region across the genome. For a single region, the posterior expected number 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 (Methods). The dark lines mark the means of the distributions. The traits are sorted on the x-axis by the average number of shared causal SNPs per region.

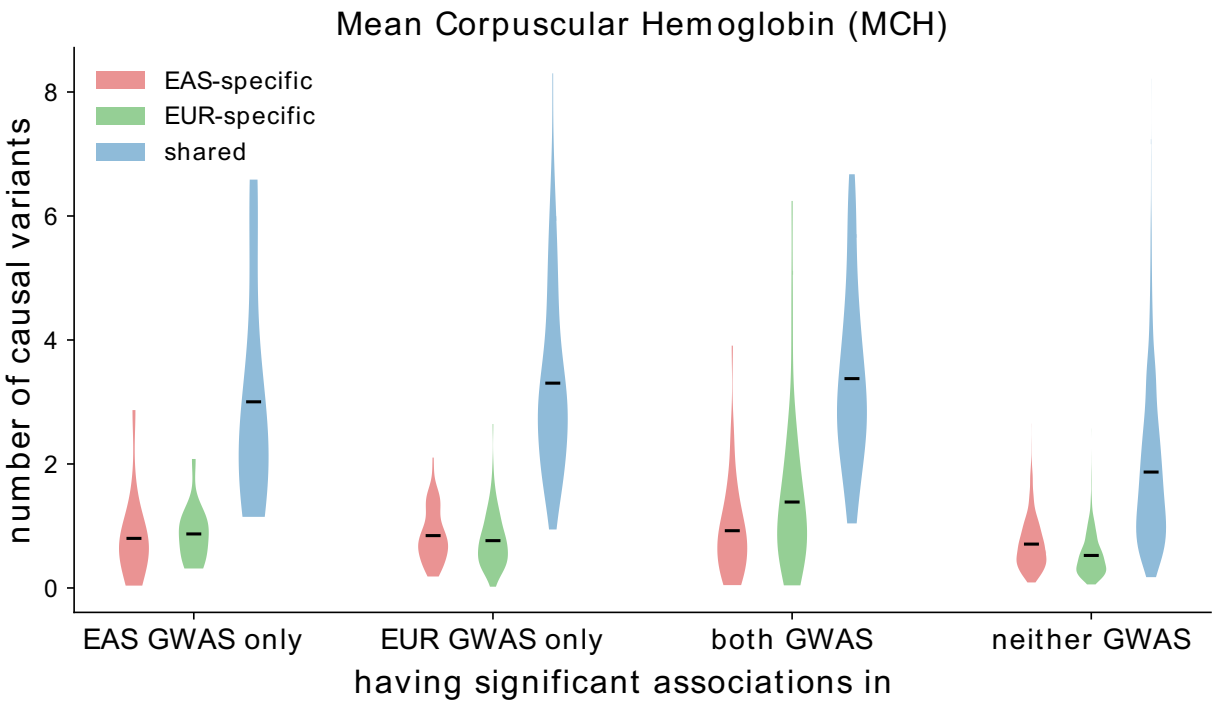

Figure 4**: Distributions of the numbers of population-specific and shared causal variants at GWAS risk regions for mean corpuscular hemoglobin (MCH).** Each violin plot represents the distribution of the posterior expected number of population-specific (red/green) or shared (blue) causal SNPs at regions with significant associations ($p_{GWAS} < 5 \times 10^{-8}$) in EAS GWAS only, EUR GWAS only, both EAS and EUR, and neither GWAS. The dark lines mark the means of the distributions.

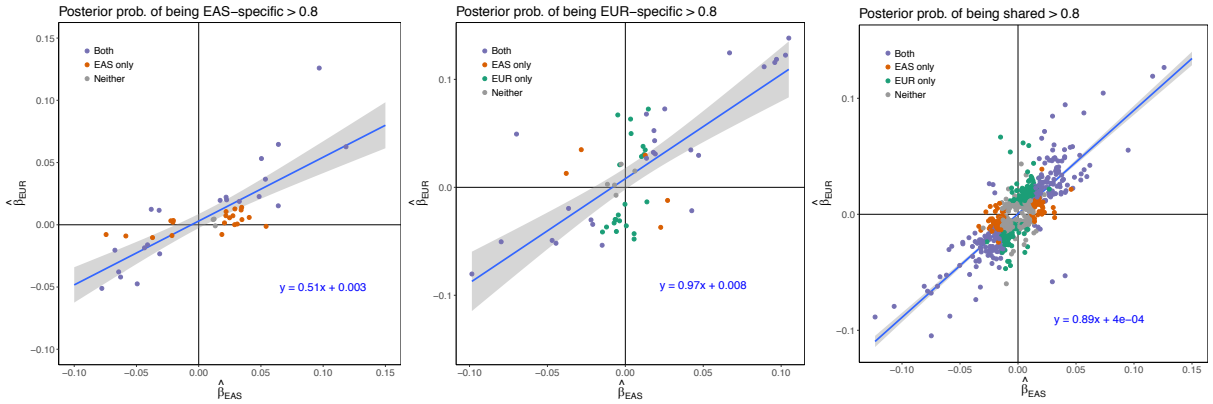

Figure 5: **Marginal regression coefficients of high-posterior SNPs for 9 complex traits.** Each plot corresponds to one of the three causal configurations of interest: EAS-specific (left), EUR-specific (middle), and shared (right). Each point represents a SNP with posterior probability of causality > 0.8 for a single trait. The x-axis and y-axis mark the marginal regression coefficients in the EAS-ancestry GWAS and EUR-ancestry GWAS, respectively. The colors indicate whether the SNP is nominally significant ($p_{GWAS} < 5 \times 10^{-6}$) 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.

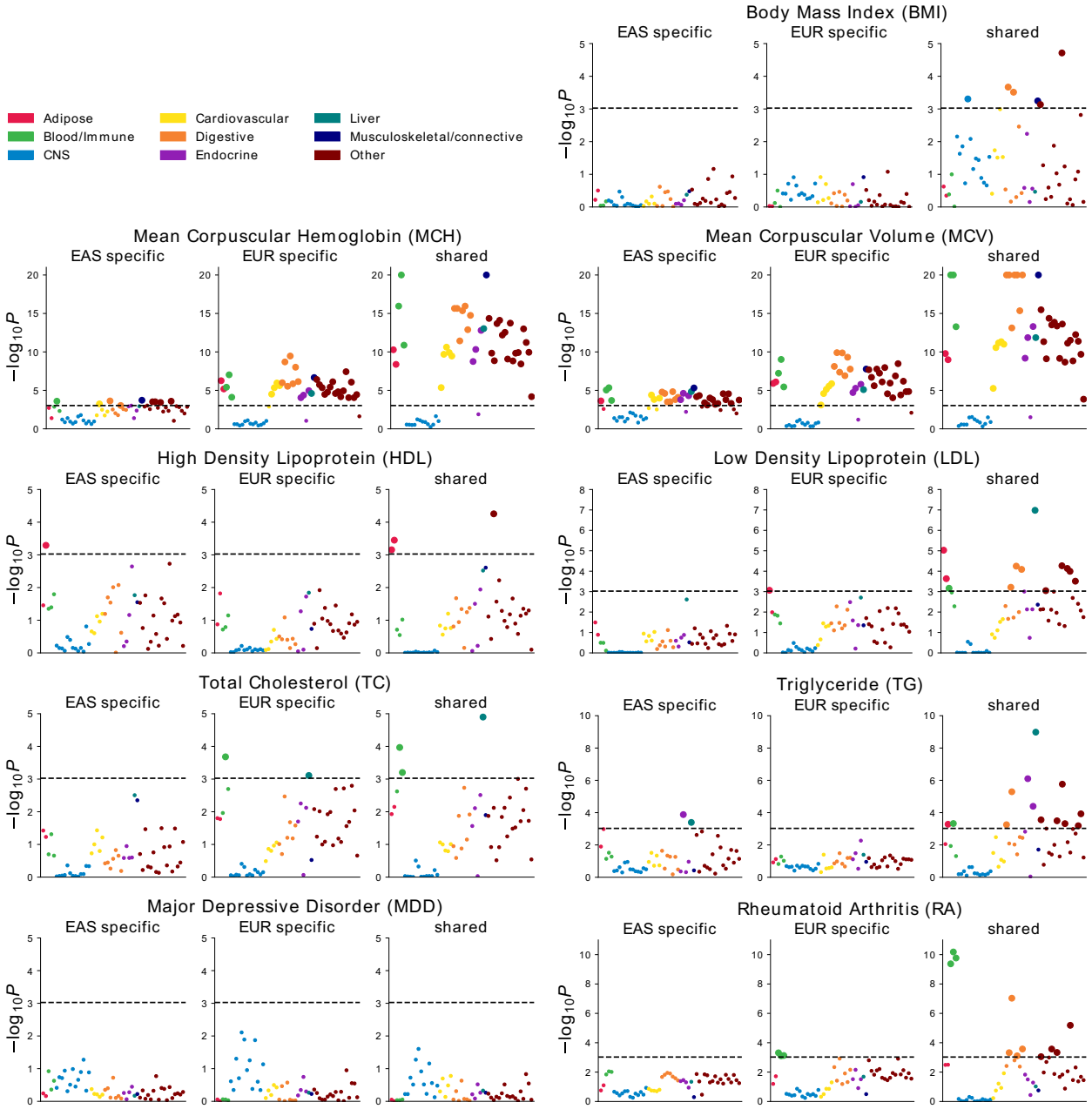

Figure 6**: P-values of enrichment of population-specific/shared causal variants near genes that are "specifically expressed" in one of 53 GTEx tissues.** Each plot contains 53 points; each point represents the p-value from a test of enrichment of EAS-specific, EUR-specific, or shared causal SNPs in one of the 53 SEG annotations (Methods). The dotted line marks the significance threshold applied to all tests (after Bonferroni correction: *P*-value < 0.05/53 which corresponds to $-\log_{10}P > -3.03$). We highlight SEG annotations passing the significance threshold using larger dots. Detailed information, including estimated enrichments and standard errors, can be found in Figures S37-41.