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

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 9 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 2.8x enrichment of shared high-posterior SNPs, suggesting that population-specific GWAS risk regions harbor shared causal SNPs that are undetected in the other GWAS 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 (GWAS) 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 GWAS tend to replicate across populations[17,21,22,33–35], indicating that complex traits have shared genetic components 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 (PRS) across populations[5,41–48]; critically, if applied in a clinical setting, existing PRS may exacerbate health disparities among ethnic groups[49]. The population-specificity of existing PRS 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. For example, due to 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 PRS 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 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]). 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 (Methods) produces well-calibrated per-SNP posterior probabilities and enrichment test statistics. We 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] (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.

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) (Table 1), restricting to common SNPs (MAF > 5%) and using 1000 Genomes[59] to estimate ancestry-matched LD. On average across the 9 traits, we estimate that approximately 80% (S.D. 15%) of common SNPs that are causal in EAS and 84% (S.D. 8%) of those in EUR are shared by the other population. Consistent with previous studies based on SNP-heritability[55,60], 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 2.8x 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 effects 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 $\hat{\rho}_g = 0.79 \pm 0.07$ s.e.m. across the 9 traits)[51]. Finally, we show that regions flanking genes that are specifically expressed in trait-relevant tissues[61] 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 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[62,63]

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

in order to facilitate optimization and interpretation (Supplemental 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, $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ (both are $p \times 1$ vectors), conditional on the causal status vectors for each population, $\boldsymbol{c}_1 = (c_{11}, \cdots, c_{p1})^T$ and $\boldsymbol{c}_2 = (c_{12} \cdots c_{p2})^T$. 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,

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

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 (Supplemental 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:

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

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[64] and model the distribution of Z-scores at all regions as the product of the distribution of Z-scores in each region (Supplemental 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 (Supplemental 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 (Supplemental Note).

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 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[64] to define LD blocks within the transethnic LD matrix. Applying this procedure using the 1000 Genomes Phase 3 reference panel[59] 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

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

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, $E[q_{k,\boldsymbol{b}}|\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{f}^*]$, and denominator, $E[q_{k,\boldsymbol{b}}|\boldsymbol{f}^*]$, 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 $\hat{\alpha}_{k,\boldsymbol{b}}$ using block jackknife over 1,368 non-overlapping approximately LD-independent blocks across the entire genome. The resulting enrichment test statistics, $\frac{\hat{\alpha}_{k,\boldsymbol{b}}-1}{\mathrm{SE}(\hat{\alpha}_{k,\boldsymbol{b}})}$, approximately follow a t-distribution with degrees of freedom equal to the number of blocks minus one[65]. 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[13,16,66]. 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[67,68] and 50,000 individuals of white British ancestry from the UK Biobank[69,70] to simulate causal effects and phenotypes. First, we used PLINK[71] (v1.9) to remove redundant SNPs in the 1000 Genomes Phase 3 reference panel[59] such that there is no pair of SNPs with $r_{ij}^2 > 0.95$ $(i \neq j)$ 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[72], 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[73]. We use PLINK[71] (v.19) to remove redundant SNPs such that

210 $\hat{r}_{ij}^2 < 0.95$ for all SNPs $i \neq j$ in both ancestry-matched 1000 Genomes[73] reference panels. The resulting
211 numbers of SNPs that were analyzed for each trait are listed in Table 1.
212 For each trait, we test for enrichment of population-specific/shared causal SNPs in 53 publicly
213 available tissue-specific gene annotations[66], each of which represents a set of genes that are "specifically
214 expressed" in a GTEx[74] tissue (referred to as "SEG annotations"). We set the threshold for statistical
215 significance to *P*-value < 0.05/53 (Bonferroni correction for the number of tests performed per trait).
216

217 # Results

218 ## Performance of PESCA in simulations

219 We assessed the performance of PESCA in simulations starting from real genotypes of individuals with
220 East Asian[67,68] (EAS) or European[69,70] (EUR) ancestry (N_{EAS} = 10K, N_{EUR} = 50K, M = 8,599 SNPs)
221 (Methods). First, we find that when in-sample LD from the GWAS is available, PESCA yields
222 approximately unbiased estimates of the numbers of population-specific/shared causal SNPs (Figure 2, top
223 panel). For example, in simulations where we randomly selected 50 EAS-specific, 50 EUR-specific, and
224 50 shared causal SNPs, we obtained estimates (and corresponding standard errors) of 37.8 (4.5) EAS-
225 specific, 40.3 (4.9) EUR-specific, and 64.9 (6.3) shared causal SNPs, respectively. When external reference
226 LD is used (in this case, from 1000 Genomes[73]), PESCA yields a slight upward bias (Figure 2, bottom
227 panel); on the same simulated data, we obtained estimates of 48.0 (5.9) EAS-specific, 53.7 (7.44) EUR-
228 specific, and 78.8 (7.6) shared causal SNPs.
229 We observe a slight decrease in accuracy as the effective sample size, the product of SNP-
230 heritability and sample size ($N \times h_g^2$), decreases (Figures S1-S5). This is expected as the likelihood of the
231 GWAS summary statistics is a function of $N \times h_g^2$ (Methods) – as the expected per-SNP variance at causal
232 SNPs ($N \times h_g^2$ divided by the number of causal SNPs) decreases, GWAS summary statistics provide less
233 information on the causal status of each SNP. Since it is often the case that the sample size of one GWAS
234 is larger than that of the other, we perform simulations in which SNP-heritability is fixed to 0.05 in both
235 populations, the EAS sample size is fixed to $N_{EAS} = 10^4$, and the EUR sample size is varied such that the
236 effective sample size of the EUR GWAS is 1-5x larger than that of the EAS GWAS. We find that the
237 genome-wide estimators are relatively robust with in-sample LD; with external estimates of LD, when
238 effective sample size differs by a factor of 2 or more, the estimator for the number of EUR-specific causal
239 SNPs becomes less biased while the EAS-specific and shared causal estimators become increasingly
240 inflated (Figure S6). In addition, while it seems likely that the effect sizes of shared causal SNPs would be
241 positively correlated across populations, the PESCA model assumes zero cross-population correlation in
242 order to facilitate inference (Methods). We therefore perform simulations under an alternative model in

243 which EAS and EUR effect sizes at shared causal SNPs are positively correlated and find that our estimates
244 of the genome-wide numbers of shared and population-specific causal SNPs become increasingly inflated
245 and deflated, respectively, as the correlation increases from 0 to 1 (Figure S7).

246 Next, we use the estimated genome-wide proportions of population-specific/shared causal SNPs to
247 evaluate per-SNP posterior probabilities of being causal in a single population (EAS only or EUR only) or
248 in both populations (Methods). For each of the three causal configurations of interest (EAS only, EUR only,
249 and shared), we observe an increase in the average correlation between the per-SNP posterior probabilities
250 and the true causal status vector for that configuration as $N \times h_g^2$ increases and as the total number of causal
251 SNPs decreases (i.e. as per-SNP causal effect sizes increase) (Figures S8-S9). As expected, as the simulated
252 proportion of shared causal SNPs increases, the average correlation between the posterior probabilities and
253 true causal status vectors increases for the shared causal configuration and decreases for the population-
254 specific causal configurations (Figures S8-S9). Since we did not have access to individual-level genotypes
255 sampled from an ancestral group with shorter LD blocks (e.g., African-ancestry individuals), we use the
256 EAS and EUR LD scores of each SNP as proxies for the strength of LD in the region housing the SNP to
257 investigate the impact of population-specific LD patterns on the per-SNP posterior probabilities. Among
258 the true causal SNPs (shared or population-specific), the posterior probabilities are relatively invariant to
259 the magnitude of the EAS and EUR LD scores (Figure S10). In other words, under the PESCA framework,
260 power to detect a given true causal SNP does not depend on its LD score in either population. Restricting
261 to a set of "high-posterior SNPs" (defined here as SNPs with posterior probability greater than some
262 threshold t), we investigate whether PESCA systematically misclassifies SNPs based on the magnitude of
263 their LD scores. Again, we observe that the average EAS and EUR LD scores do not vary significantly
264 between the true and false positive classifications (Table S1). We then assessed whether our proposed
265 statistics for testing for enrichment of population-specific/shared causal SNPs in functional annotations
266 (Methods) are well-calibrated under the null hypothesis of no enrichment. Overall, when both population-
267 specific and shared causal SNPs are drawn at random, the enrichment test statistics are conservative at
268 different levels of polygenicity and GWAS power ($N \times h_g^2$), irrespective of whether in-sample LD or
269 external reference LD is used (Figures S11-S16).

270 Finally, we evaluated the computational efficiency of each stage of inference. In the first stage of
271 inference – estimating genome-wide proportions of population-specific/shared causal SNPs – the
272 maximization step of the EM algorithm uses Gibbs sampling to efficiently sample from the posterior of the
273 causal status vectors (Supplemental Note). We set both the number of burn-in iterations and the number of
274 samples to 5,000 for the MCMC within the maximization step and found that the overall EM typically
275 converged within 200 iterations (Figures S17-S19). Run-time per EM-iteration increases with the number
276 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 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 (Supplemental Note). 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 (Supplemental Note). The average run-time was 5 minutes in simulations with 20 causal variants and 28 minutes in simulations with 100 causal 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 9 complex traits

We obtained publicly available GWAS summary statistics for 9 (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 (Methods). To ensure convergence, we applied 750 EM iterations for each trait (Figures S21-S23). Across the 9 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[7,8,55,75,76]. 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 9 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 (Methods). For all 9 traits, high-posterior SNPs for both the population-specific and shared causal configurations are spread nearly uniformly across the genome (Figure 3, Figures S24-S31). 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,

311 and 2.19 (S.D. 1.46) shared high-posterior SNPs per region (Figure 3, Figure S29). Aggregating posterior
312 probabilities by chromosome, we find that the posterior expected numbers of EAS-specific, EUR-specific,
313 and shared causal SNPs per chromosome are highly correlated with chromosome length (Figures S32-S34),
314 recapitulating previous findings based on regional SNP-heritability[55,60].
315

316 **Distributions of high-posterior SNPs across GWAS risk regions**

317 We aggregate per-SNP posterior probabilities within GWAS risk regions that are EAS-specific, EUR-
318 specific, or shared by both populations and find that most GWAS risk regions harbor two or more shared
319 high-posterior SNPs (Figure 4, Figures S35-S39), concordant with previous findings on allelic
320 heterogeneity of complex traits[55,77,78]. On average across the 9 traits, we observe a 2.8x enrichment of shared
321 high-posterior SNPs in population-specific GWAS risk regions relative to the genome-wide background.
322 For example, for mean corpuscular hemoglobin (MCH), the EAS-specific and EUR-specific GWAS risk
323 regions harbor an average of 3.0 (S.D. 1.7) and 3.3 (S.D. 1.5) shared high-posterior SNPs per region,
324 respectively, whereas the average number of shared high-posterior SNPs per region across all regions is 2.0
325 (S.D. 1.3) (Figure 4). While BMI, the blood traits (MCH and MCV), and rheumatoid arthritis have similar
326 numbers of EAS-specific and EUR-specific high-posterior SNPs in their population-specific GWAS risk
327 regions, the lipids traits (HDL, LDL, total cholesterol and triglycerides) have significantly more EAS-
328 specific high-posterior SNPs in all GWAS risk regions (Figure 4, Figures S35-S39).

329 For each causal configuration (EAS-specific, EUR-specific, or shared), we examine the effect sizes
330 of high-posterior SNPs (posterior probability > 0.8) in EAS and EUR (Figure 5). Across the 9 traits, the
331 majority of EAS-specific high-posterior SNPs are nominally significant ($p_{GWAS} < 5 \times 10^{-6}$) either in the
332 EAS GWAS only or in both GWASs. While five EUR-specific high-posterior SNPs are nominally
333 significant in only the EAS GWAS, the majority are nominally significant either in the EUR GWAS only
334 or in both GWASs. We observe strong correlations between the effect sizes in EAS and EUR for all three
335 sets of high-posterior SNPs (Pearson r^2 of 0.79 [EAS-specific], 0.73 [EUR-specific], and 0.80 [shared])
336 that are driven by SNPs that are nominally significant in both GWASs (Figure 5). Taken together, these
337 results suggest that most population-specific GWAS risk regions harbor shared causal variants that are
338 undetected in the other population due to heterogeneity in LD structures, allele frequencies, and/or GWAS
339 sample sizes[33].
340

341 **Enrichment of high-posterior SNPs near genes expressed in trait-relevant tissues**

342 Motivated by recent work that found enrichment of SNP-heritability in regions near genes that are
343 "specifically expressed" in trait-relevant tissues and cell types (referred to as "SEG annotations"), we tested
344 for enrichments of population-specific and shared causal SNPs in the same 53 tissue-specific SEG

annotations[61]. 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 (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 (which are obtained 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 9 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; moreover, high-posterior SNPs (posterior probability > 0.8 for any causal configuration) have highly correlated effect sizes in EAS and EUR, recapitulating results of previous studies[33]. For all traits except MDD, we identify tissue-specific SEG annotations[66] 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 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 Europeans in the presence of the same environmental factors.

Second, we chose to analyze a set of traits that were present in both the UK Biobank and Biobank Japan and for which 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 2x. We note, however, that for the real traits analyzed in this work, effective sample size differs by a maximum factor of 2x (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[69,70,79–81], 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[73], 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 (PRS) 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.

Acknowledgements

We are grateful to 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, R01MH115676, U01CA194393, T32NS048004, T32MH073526, and T32HG002536.

Declaration of Interests

The authors declare no competing interests.

Web Resources

GIANT consortium GWAS summary statistics: http://portals.broadinstitute.org/collaboration/giant

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/pesca

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. 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

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

9. Zhang, Y., Qi, G., Park, J.-H. & Chatterjee, N. Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. *Nat. Genet.* **50**, 1318–1326 (2018).

10. Zhu, X. & Stephens, M. Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes. *Nat. Commun.* **9**, 4361 (2018).

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

12. Schoech, A. P. *et al.* Quantification of frequency-dependent genetic architectures in 25 UK Biobank traits reveals action of negative selection. *Nat. Commun.* **10**, 790 (2019).

13. Gazal, S. *et al.* Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* **49**, 1421 (2017).

14. 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).

15. Speed, D. *et al.* Reevaluation of SNP heritability in complex human traits. *Nat. Genet.* **49**, 986 (2017).

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

17. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).

18. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538**, 161 (2016).

19. Rosenberg, N. A. *et al.* Genome-wide association studies in diverse populations. *Nat. Rev. Genet.* **11**, 356 (2010).

20. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five Years of GWAS Discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012).

21. 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).

22. 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).

23. Li, Z. *et al.* Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia. *Nat. Genet.* **49**, 1576 (2017).

24. 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).

25. 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).

26. 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).

27. 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).

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

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

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

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

32. 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

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

34. 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).

35. 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).

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

37. 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).

38. 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).

39. 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).

40. Wen, X., Luca, F. & Pique-Regi, R. Cross-population joint analysis of eQTLs: fine mapping and functional annotation. *PLoS Genet.* **11**, e1005176 (2015).

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

42. 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).

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

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

45. 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).

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

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

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

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

50. 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).

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

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

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

54. 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).

55. 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).

56. Hou, K. *et al.* Accurate estimation of SNP-heritability from biobank-scale data irrespective of genetic architecture. *Nat. Genet.* **51**, 1244–1251 (2019).

57. Farh, K. K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015).

58. Huang, H. *et al.* Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature* **547**, 173–178 (2017).

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

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

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

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

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

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

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

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

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

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

69. 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).

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

71. 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).

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

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

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

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

76. Holland, D. *et al.* Beyond SNP Heritability: Polygenicity and Discoverability of Phenotypes Estimated with a Univariate Gaussian Mixture Model. *bioRxiv* 133132 (2019). doi:10.1101/133132

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

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

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

80. 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).

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

82. 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).

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

84. 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).

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.)[51]
Body Mass Index (BMI)	EAS	22	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	72	20.6 (0.91)	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	82	22.7 (3.2)	172,332					
Mean Corpuscular Volume (MCV)	EAS	21	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	82	23.6 (3.1)	172,433					
High Density Lipoprotein (HDL)	EAS	21	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	83	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	83	13.6 (1.93)	85,491					
Total Cholesterol (TC)	EAS	21	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	83	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	83	13.6 (2.2)	86,502					
Major Depressive Disorder (MDD)	EAS	67	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	84	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					

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[54] 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[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.

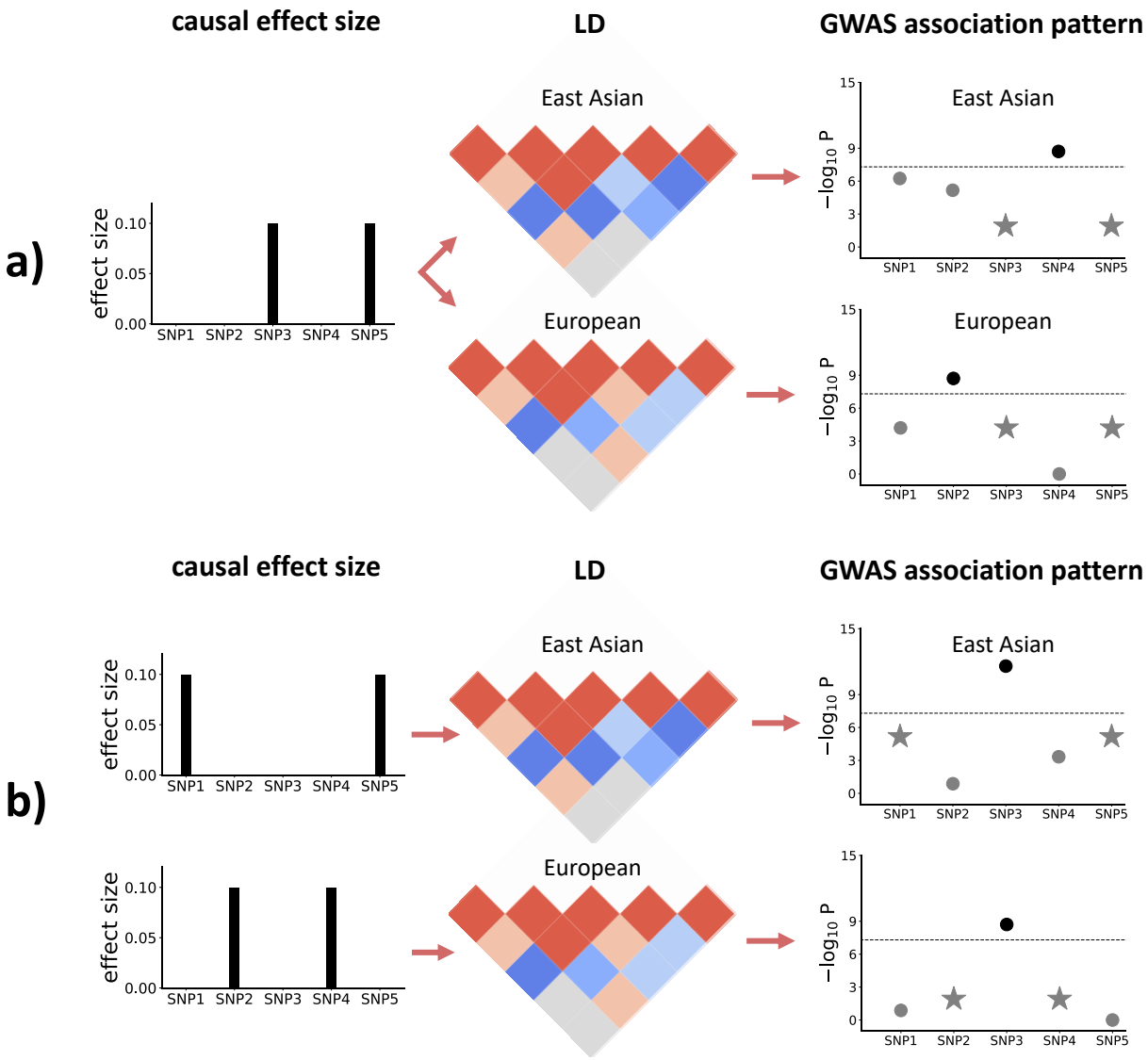

Figure 1**: Toy examples to illustrate how population-specific LD patterns affect GWAS associations.** a) SNPs 3 and 5 are causal in both East Asians and Europeans and have the same population-specific causal effect size of 0.1. However, due to different LD patterns in East Asians and Europeans, SNPs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

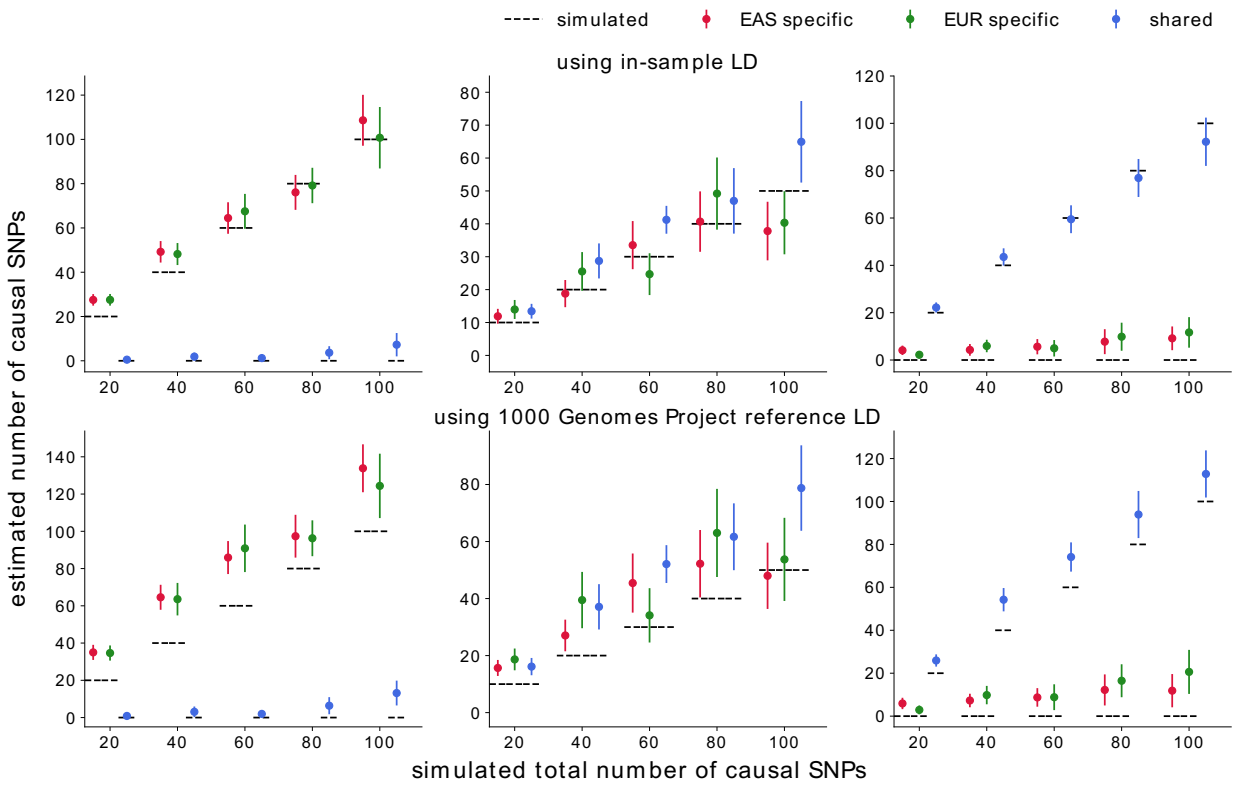

Figure 2: **Genome-wide estimates of the numbers of population-specific/shared causal SNPs in simulations.** The estimates are approximately unbiased when in-sample LD is used (top panel) and upward-biased estimates when external reference LD is used (bottom panel). For both populations, we simulate 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.

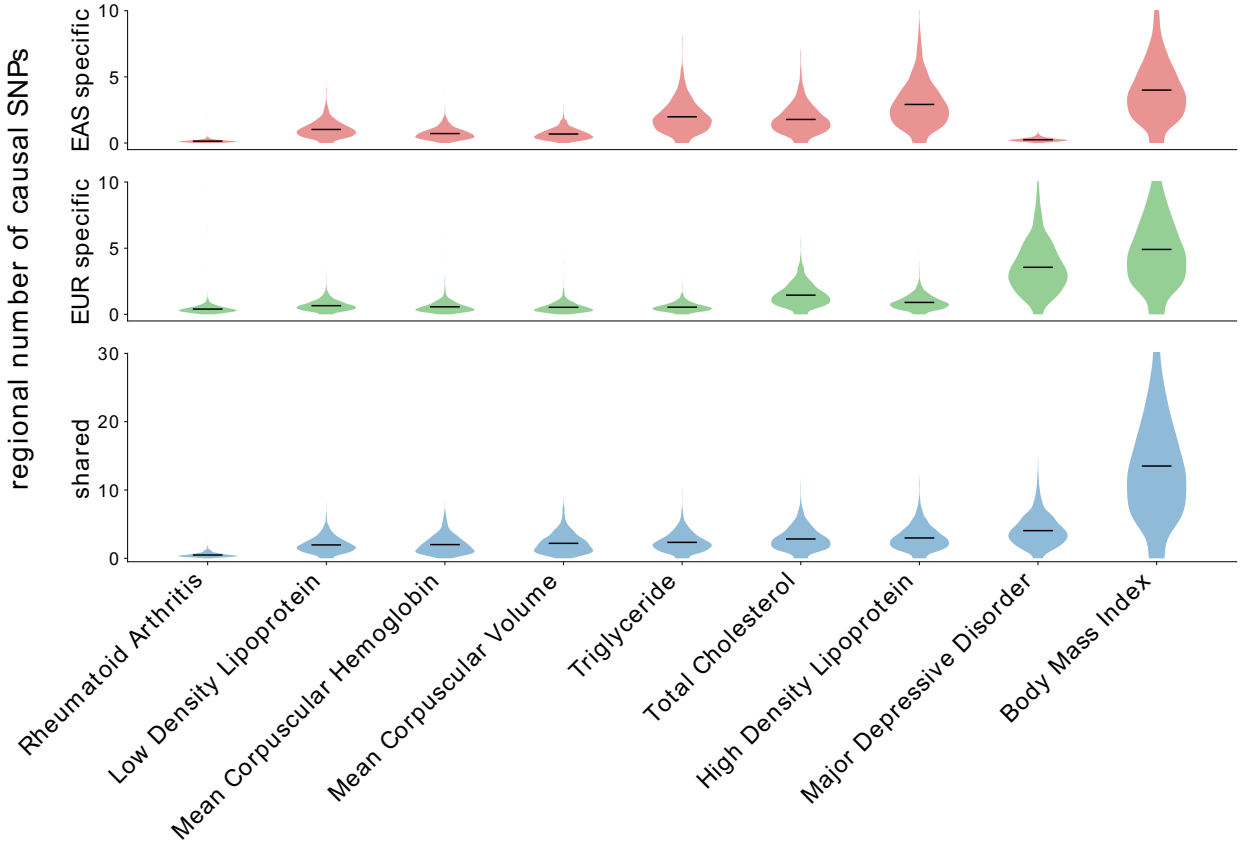

Figure 3: **Distributions of the numbers of population-specific/shared causal SNPs across 1,368 regions that are approximately independent in both EAS and EUR.** Each violin plot represents the distribution of the posterior expected number of population-specific or shared causal SNPs per region; details on how the regions were defined can be found in the 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 (Methods). 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.

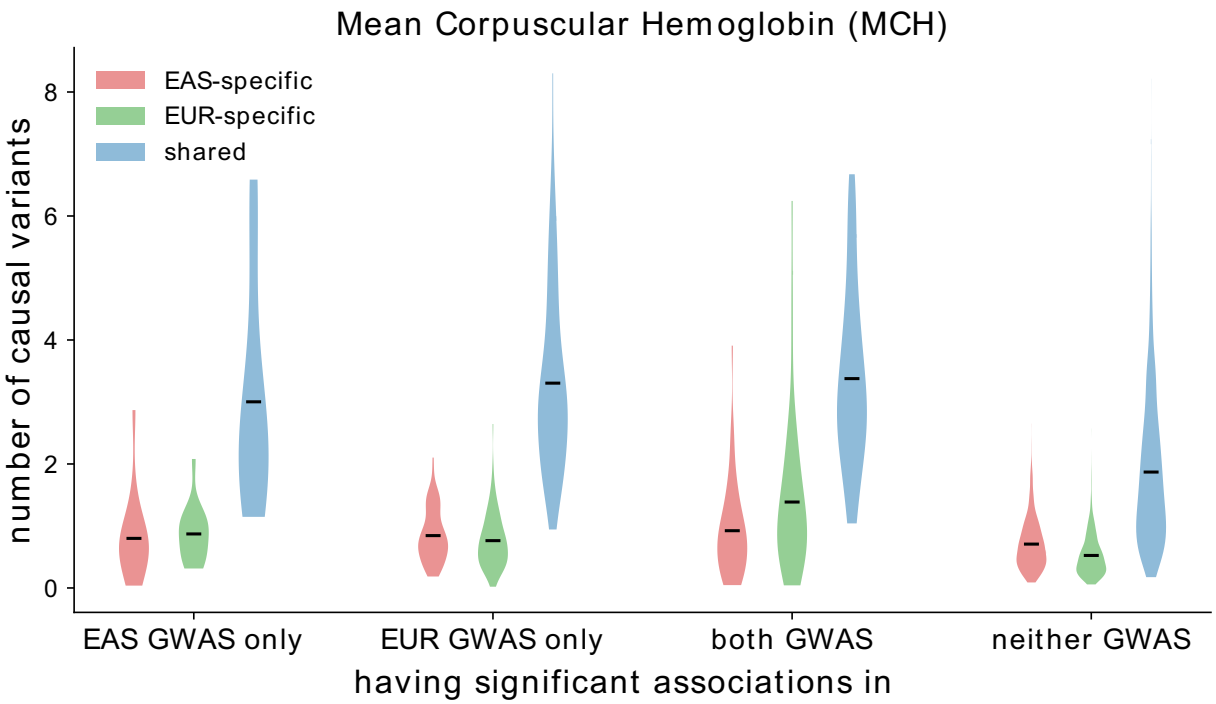

Figure 4**: Distributions of the numbers of population-specific/shared causal variants at GWAS risk regions for mean corpuscular hemoglobin (MCH).** Each violin plot represents the distribution of the posterior expected number of population-specific (red/green) or shared (blue) causal SNPs at regions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

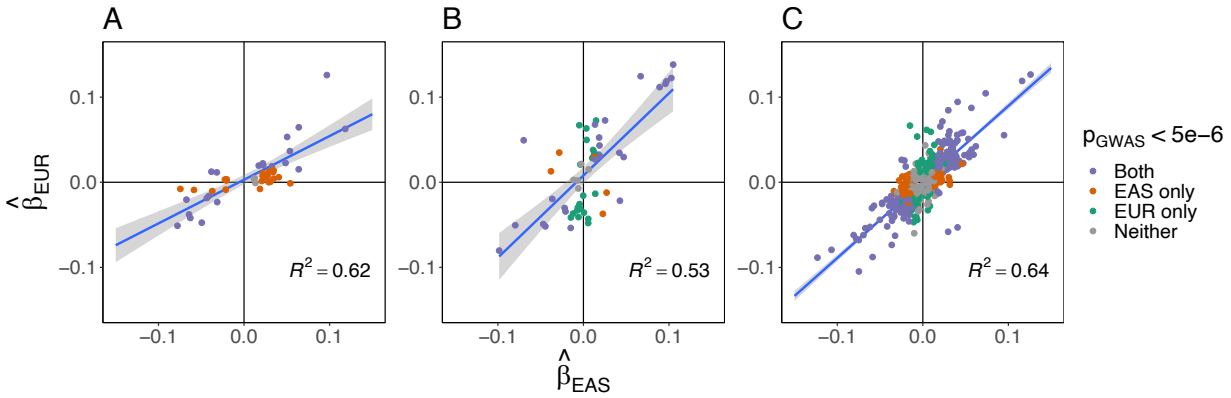

Figure 5: **Marginal regression coefficients of high-posterior SNPs for 9 complex traits.** Each plot corresponds to one of the three causal configurations of interest: EAS-specific (A), EUR-specific (B), and shared (C). Each point represents a SNP with posterior probability > 0.8 for a single trait. The x-axis and y-axis mark the 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.

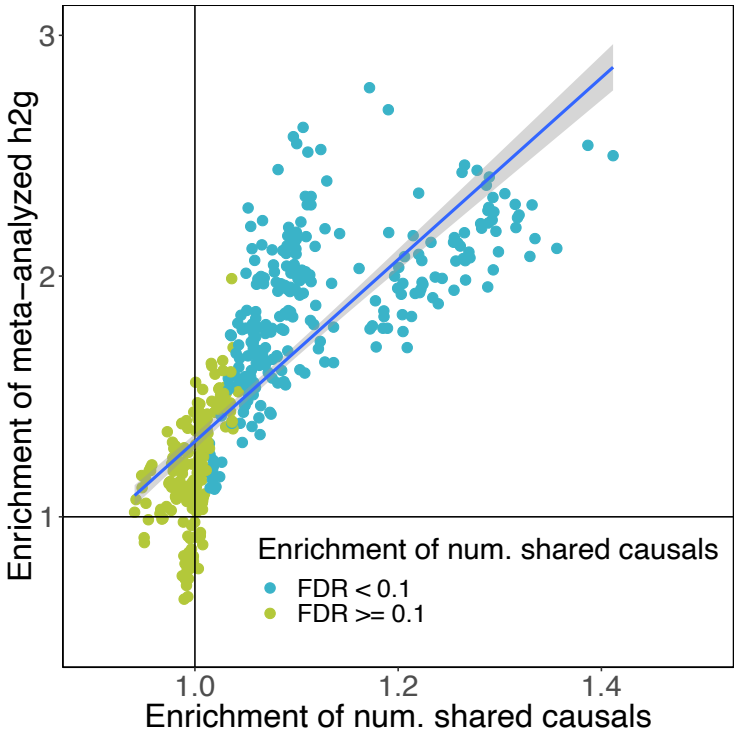

Figure 6**: Enrichments of shared high-posterior SNPs in 53 tissue-specific functional categories are highly correlated with SNP-heritability enrichments.** Each point represents a trait-tissue pair; each tissue-specific functional category represents a set of genes that are "specifically expressed" in one of 53 GTEx tissues (53 SEG annotations). The x-axis is the enrichment of shared high-posterior SNPs in the SEG annotation obtained 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 (obtained 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). Enrichment estimates and standard errors for each trait-tissue pair can be found in Figures S40-S44.