Improved multi-ancestry fine-mapping identifies *cis*-regulatory variants underlying molecular traits and disease risk

Zeyun Lu[1], Xinran Wang[1], Matthew Carr[1], Artem Kim[1], Steven Gazal[1,2,3], Pejman Mohammadi[4,5,6], Lang Wu[7], Alexander Gusev[8], James Pirruccello[9], Linda Kachuri[10,11], Nicholas Mancuso[1,2,3,12]

1. Center for Genetic Epidemiology, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA
2. Department of Population and Public Health Sciences, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA
3. Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA
4. Center for Immunity and Immunotherapies, Seattle Children's Research Institute, Seattle, WA, USA
5. Department of Pediatrics, University of Washington School of Medicine, Seattle, WA, USA
6. Department of Genome Sciences, University of Washington, Seattle, WA, USA
7. Cancer Epidemiology Division, Population Sciences in the Pacific Program, University of Hawai'i Cancer Center, University of Hawai'i at Mānoa, Honolulu, HI, USA
8. Harvard Medical School and Dana-Farber Cancer Institute, Boston, MA, USA
9. Division of Cardiology, University of California San Francisco, San Francisco, CA, USA
10. Department of Epidemiology and Population Health, Stanford University School of Medicine, Stanford, CA, USA
11. Stanford Cancer Institute, Stanford University School of Medicine, Stanford, CA, USA
12. Corresponding Author

Contacts:

1. Zeyun Lu (zeyunlu@usc.edu)
2. Nicholas Mancuso (Nicholas.Mancuso@med.usc.edu)

Abstract

Multi-ancestry statistical fine-mapping of *cis*-molecular quantitative trait loci (*cis*-molQTL) aims to improve the precision of distinguishing causal *cis*-molQTLs from tagging variants. However, existing approaches fail to reflect shared genetic architectures. To solve this limitation, we present the Sum of Shared Single Effects (SuShiE) model, which leverages LD heterogeneity to improve fine-mapping precision, infer cross-ancestry effect size correlations, and estimate ancestry-specific expression prediction weights. We apply SuShiE to mRNA expression measured in PBMCs (n=956) and LCLs (n=814) together with plasma protein levels (n=854) from individuals of diverse ancestries in the TOPMed MESA and GENOA studies. We find SuShiE fine-maps *cis*-molQTLs for 16% more genes

compared with baselines while prioritizing fewer variants with greater functional enrichment. SuShiE infers highly consistent *cis*-molQTL architectures across ancestries on average; however, we also find evidence of heterogeneity at genes with predicted loss-of-function intolerance, suggesting that environmental interactions may partially explain differences in *cis*-molQTL effect sizes across ancestries. Lastly, we leverage estimated *cis*-molQTL effect-sizes to perform individual-level TWAS and PWAS on six white blood cell-related traits in AOU Biobank individuals (n=86k), and identify 44 more genes compared with baselines, further highlighting its benefits in identifying genes relevant for complex disease risk. Overall, SuShiE provides new insights into the *cis*-genetic architecture of molecular traits.

Introduction

Characterizing the functional consequences of genetic variation remains a crucial task in deciphering the mechanisms underlying complex disease risk[1,2]. To this end, *cis*-molecular quantitative trait loci (*cis*-molQTL) mapping seeks to identify genetic variants associated with genomically proximal molecular features measured across diverse cellular, tissue, and environmental contexts[3–14]. However, due to linkage disequilibrium (LD), it is challenging to distinguish causal *cis*-molQTLs from tagging variants within a genomic region[3,5]. Statistical fine-mapping aims to resolve precisely this issue[15–19], yet pervasive LD signals limit the resolution of these approaches. Previous efforts have demonstrated that leveraging the heterogeneity of LD and minor allele frequency (MAF) across diverse ancestries improves the precision of statistical fine-mapping and therefore enhances our biological understanding of complex diseases[20–25] and molecular traits[26–32].

While existing multi-ancestry fine-mapping frameworks have been proposed for the analysis of complex traits and diseases[30,33–41], they have several limitations in the context of large-scale *cis*-molQTL data. First, many approaches do not model the correlation of causal variant effect sizes across ancestries or assume that they are a-priori independent across ancestries, which fails to reflect shared or similar genetic architectures[33,35,37,38]. Second, existing multi-ancestry approaches scale poorly, which precludes their application to thousands of molecular traits

commonly measured in *cis*-molQTL studies[33,35,40]. Third, current fine-mapping approaches lack ancestry-specific effect size estimates[33,35,37], which neglects their potential use in post-Genome-wide Association Studies (GWASs) frameworks (e.g., Transcriptome- and Proteome-wide Association Studies (TWASs/PWASs)[42–47]. Last, while recent approaches address some of these limitations, existing software implementations are capable of analyzing only two ancestries, which excludes datasets consisting of ever-increasing diverse ancestries[39].

Here, we describe the Sum of Shared Single Effects (SuShiE) approach to fine-map genetic variants shared across diverse ancestries for thousands of molecular traits. SuShiE integrates genotypic and molecular data from multiple ancestries to identify *cis*-molQTLs while modeling and learning the covariance structures of shared/non-shared signals. SuShiE leverages four key insights. First, SuShiE improves fine-mapping precision of the shared *cis*-molQTLs by leveraging LD across different ancestries. Second, it estimates ancestry-specific effect sizes at shared *cis*-molQTLs. Third, it infers the prior effect size correlation across ancestries to shed light on genetic similarities and differences. Lastly, SuShiE is implemented using a scalable variational inference algorithm that runs seamlessly on CPUs, GPUs, or tensor-processing units (TPUs).

Through extensive simulations, we show that SuShiE outputs higher posterior inclusion probabilities (PIPs) at causal *cis*-molQTLs, outputs smaller credible set sizes, and exhibits better calibration compared with current approaches[15,38]. Using bulk mRNA expression levels measured in peripheral blood mononuclear cells (PBMCs) and lymphoblastoid cell lines (LCLs) together with protein abundance measured in plasma, we fine-map 36,911 molecular phenotypes across American European, African, and Hispanic ancestries from TOPMed-MESA[48,49] (n_{mRNA}=956 and $n_{protein}$=814) and GENOA[26] (n_{mRNA}=854). SuShiE fine-maps significantly more *cis*-molQTLs with smaller credible sets and greater enrichment in relevant functional annotations compared with existing methods. In addition, SuShiE infers shared genetic architecture of *cis*-molQTL in significantly heritable genes and shows the heterogeneity across ancestries of signals associated with multiple measures of loss-of-function (LOF) intolerance. Last, we integrate ancestry-specific *cis*-molQTL effects inferred by SuShiE with six white blood cell-related traits to perform individual-level TWAS and PWAS in the All of Us Biobank (average n=86,345)[50] and observe that SuShiE-

based prediction models identified 44 additional associated genes compared with the baseline approach. Overall, our approach sheds light on understanding the genetic *cis*-architecture of molecular data across multiple ancestries.

Results

SuShiE overview

Here, we briefly introduce the SuShiE model (for a detailed description, see **Methods** and **Supplementary Note**). SuShiE assumes *cis*-molQTLs are present in *all* ancestries (defined as shared *cis*-molQTLs) while allowing for effect sizes at causal *cis*-molQTLs to covary across ancestries a-priori, in contrast to previous multi-ancestry approaches[15,33,35,37,38]. These assumptions provide enough flexibility to model a variety of *cis*-genetic architectures across ancestries, including cases when effects are present only in a subset of ancestries. For instance, when effects are observed only in a subset of ancestries, prior variances can be shrunk towards zero to effectively allow for *ancestry-specific* causal *cis*-molQTLs.

Focusing on the i^{th} out of k ancestries, SuShiE models the normalized levels of a molecular trait $\boldsymbol{g}_i$ measured in n_i individuals as a linear combination of p genotyped variants $\boldsymbol{X}_i$ as

$$\boldsymbol{g}_i = \boldsymbol{X}_i \left(\sum_{l=1}^{L} \boldsymbol{\gamma}_l \cdot b_{i,l} \right) + \boldsymbol{\epsilon}_i ,$$

where L is the number of shared effects, $\boldsymbol{\gamma}_l$ is a $p \times 1$ binary vector selecting the l^{th} causal *cis*-molQTL shared across ancestries, $b_{i,l}$ is the l^{th} effect size in the i^{th} ancestry, and environmental noise distributed as $\boldsymbol{\epsilon}_i \sim N(\boldsymbol{0}, \sigma^2{}_{i,e}\, \boldsymbol{I}_{n_i})$ (**Fig. 1A**). Following previous work[15,51,52], we place a $\text{Multi}(1, \boldsymbol{\pi})$ prior over $\boldsymbol{\gamma}_l$ where $\boldsymbol{\pi}$ is a $p \times 1$ vector representing prior probability for each SNP to be shared *cis*-molQTLs; however, unlike existing approaches[33,35,37,38], we organize ancestry-specific effect sizes under a multivariate normal prior $[b_{1,l}, \ldots, b_{k,l}] \sim MVN(\boldsymbol{0}, \boldsymbol{C}_l)$ where $\boldsymbol{C}_l$ is the l^{th} $k \times k$ prior effect size covariance matrix. To perform scalable inference, we use a

variational Bayesian approach and compute, for each of the L shared effects, the posterior probability of a shared causal *cis*-molQTL ($\boldsymbol{\alpha}_l$), the ancestry-specific posterior effect sizes, and covariances, in addition to prior effect-size correlations (**Fig. 1B**) inferred through a procedure analogous to Empirical Bayes. Through learning prior effect-size correlations, SuShiE can quantify genes' heterogeneity in *cis*-molQTL effects across ancestries.

SuShiE constructs a 90% credible set for each of the L effects along with a posterior inclusion probability (PIP) for each SNP to be putative causal *cis*-molQTL (see **Methods**). SuShiE is implemented in an open-source command-line Python software with JAX (see **Methods** and **Code Availability**) using *Just-In-Time* compilation to achieve high-speed inference that runs seamlessly on CPUs, GPUs, or TPUs at https://github.com/mancusolab/sushie.

SuShiE outperforms other methods in realistic simulations

First, to recapitulate the benefits of multi-ancestry study design[33,35,37–41], we performed simulations varying the number of contributing ancestries under a fixed total sample size (see **Methods**). As the number of ancestries increased, SuShiE produced higher PIPs at causal *cis*-molQTLs, smaller credible set sizes, and better calibration (**Fig. S1**), reaffirming that increasing genetic diversity refines fine-mapping results compared with expanding the sample size of a single ancestry. Next, we evaluated the performance of SuShiE in simulations by varying different parameters and compared against three baselines: SuShiE-Indep (i.e., SuShiE assuming no a-priori correlation of effect sizes across ancestries), meta-SuSiE (i.e., a meta-analysis on single-ancestry SuSiE), and SuSiE (i.e., SuSiE performed over data aggregated across ancestries; see **Methods**). For all simulations, SuShiE output higher PIPs at causal *cis*-molQTLs (~0.06 on average; all P<3.1e-4; **Fig. 2A**, **S2**), smaller credible set sizes (~0.73 on average; 2 out of 3 comparisons P<0.05; **Fig. 2B**, **S3**), and better calibration (~0.08 on average; all P<1.51e-7; **Fig. 2C**, **S4**). SuShiE similarly outperformed competing methods under simulations with differential power (**Fig. S5**) and genetic architectures across ancestries (**Fig. S6**). Next, we evaluated the ability of SuShiE to infer prior effect size correlations from data (see **Methods**). SuShiE accurately estimated primary effect size correlations (**Fig. 2D**) with

higher-order effects having diminishing accuracies. This result was likely due to decreasing statistical power, as evidenced by simulations under increased sample sizes (**Fig. 2D, S7**).

Next, we assessed the robustness of SuShiE when there exist genetic variants causal for only a subset of ancestries in addition to shared causal *cis*-molQTLs (see **Methods**). As the number of ancestry-specific *cis*-molQTLs increased, the performance of all approaches decreased compared with previous simulations. However, SuShiE continued producing higher PIPs at shared causal *cis*-molQTLs (**Fig. S8A**), smaller credible set sizes (**Fig. S8B**), and better calibrated credible sets (**Fig. S8C**), demonstrating SuShiE's robustness when ancestry-specific *cis*-molQTLs are present. We also evaluated performance in simulations where the number of causal effects (i.e., L) differs from the number specified at inference and observed that SuShiE similarly outperformed alternative approaches (**Fig. S9**).

Last, we evaluated the use of SuShiE-derived ancestry-specific effect sizes in *cis*-molQTL data as a means to predict the genetic component of gene expression for downstream TWAS[42–44]. Briefly, we performed simulations under a model in which gene expression mediates disease risk and compared SuShiE predictions with commonly used approaches for prediction-based TWAS (e.g., LASSO[53], Elastic Net[54], and gBLUP[55]) to identify susceptibility genes (see **Methods**). SuShiE-derived prediction models more accurately recapitulated gene expression levels compared with existing approaches and exhibited higher statistical power for TWAS with various study sample sizes and proportion of trait heritability mediated by gene expression (**Fig. 2E-F, S10**).

Overall, SuShiE outperforms existing approaches in realistic parameter settings, remains robust under model misspecifications, and improves statistical power in post-GWAS analyses.

SuShiE identifies more functionally relevant *cis*-molQTL signals

Having verified that SuShiE outperforms other methods under realistic simulations, we next sought to perform fine-mapping on 36, 911 molecular phenotypes from diverse ancestries. Specifically, from the Trans-Omics for Precision Medicine program Multi-Ethnic Study of Atherosclerosis[48,49] (TOPMed-MESA), we analyzed mRNA

147 expression data of 21,747 genes measured in PBMCs (visit-1; n=956) and protein expression data of 1,274 genes
148 measured in plasma (visit-1; n=854) for American European, African, and Hispanic ancestries (EUR, AFR, and HIS),
149 together with mRNA expression data of 13,890 genes measured in LCLs (n=814) for EUR and AFR from the Genetic
150 Epidemiology Network of Arteriopathy study[26] (GENOA; see **Methods**; **Table S1**).
151 Focusing on 1Mb windows for each gene (i.e., *cis*-region), SuShiE fine-mapped *cis*-molQTLs for 21,088 phenotypes
152 (e/pGenes), representing an average increase of 3,378 (16%) compared with existing methods (i.e., SuShiE-Indep,
153 Meta-SuSiE, and SuSiE; all P<2.94e-110; see **Methods**). For example, SuShiE fine-mapped 21% more e/pGenes
154 compared to single-ancestry SuSiE followed by meta-analysis (i.e., Meta-SuSiE; P=7.01e-238), again highlighting
155 the benefit of multi-ancestry study design. SuShiE-based credible sets maintained higher average PIPs (~0.07 on
156 average) and higher frequency of *cis*-molQTLs with PIPs > 0.9 (~0.02 on average), as well as smaller credible sets
157 in most cases (~6.24 on average; **Table S2**). We found the performance advantage slightly diminished in TOPMed-
158 MESA protein and GENOA mRNA datasets, likely due to lower statistical power. Using the number of credible sets
159 identified after purity pruning (see **Methods**), SuShiE estimated most (90.4%) molecular phenotypes to exhibit 1-
160 3 *cis*-molQTL signals (**Fig. 3A**) with PIPs localizing near the transcription start site (TSS; **Fig. 3B**), consistent with
161 previous studies[3,4,26,56,57].
162 To characterize the regulatory function of identified *cis*-molQTL signals, we performed enrichment analysis using
163 PIPs with 89 genomic functional annotations (see **Methods**). We observed that PIPs inferred by SuShiE were
164 enriched in 83/89 annotations across all three datasets, with the highest enrichment occurring in promoter
165 regions (**Table S3**). For example, PIPs were enriched in 4/5 candidate cis-regulatory elements (cCREs) from
166 ENCODE Registry v3[58] (**Fig. 3C**) and in all 10 cell-type/tissue-specific cCREs using single-nucleus(sn) or single-cell(sc)
167 ATAC-Seq[59,60] (**Fig. S11**). Importantly, PIPs inferred by SuShiE were more enriched across functional annotations
168 compared with those computed from existing fine-mapping methods (all P<8.13e-3; **Table S4**), highlighting
169 SuShiE's ability to better prioritize functionally relevant *cis*-molQTLs. Next, to explore how potential regulatory
170 function may differ among *cis*-molQTLs contributing to the same gene, we repeated the above analyses using per-

effect posterior probabilities (α_l), rather than overall inclusion probabilities (i.e., PIPs). First, the initial three shared effects were similarly localized near the TSS (**Fig. S12**) and were more enriched in promoter regions compared to the PIP-based analyses (**Fig. S13; Table S5**), echoing the previous finding that most genes are regulated by 1-3 *cis*-molQTLs[3,4,26,57]. Second, we found *cis*-molQTLs with weaker effects were further away from the TSS on average (**Fig. S14**), likely due to statistical power. For example, we observed the expected distance to TSS for the initial three shared effects was 84.7kb compared with 144.5 kb for the remaining shared effects (i.e., from L=6 to L=10; P=8.39e-113).

Last, we sought to validate our fine-mapping results by applying SuShiE on molecular phenotypes from three independent datasets: mRNA expression measured in PBMCs of EUR, AFR, and HIS ancestries from TOPMed-MESA[48,49] (visit-5, ten-year after visit-1; n=875), mRNA expression measured in LCLs (n=462) of EUR and Yoruba (YRI) ancestries from GEUVADIS study[61], and protein expression measured in plasma of EUR ancestry (N=3,301; single-ancestry SuSiE) from INTERVAL study[5] (see **Methods**; **Table S1**). First, we confirmed SuShiE identifies 4,361 (21%; all P<2.89e-112) more e/pGenes on average compared with existing methods while obtaining higher average PIPs (~0.07 on average), smaller credible set sizes (~6.54 on average), and more *cis*-molQTLs with PIPs > 0.9 (~0.04 on average) for TOPMed-MESA visit-5 and GEUVADIS (**Table S6**). Second, focusing on 20,502 e/pGenes identified by SuShiE that were also measured in validation datasets, 34% (41%, 32%, and 13% for TOPMed-MESA visit-5, INTERVAL, and GEUVADIS, respectively) *cis*-molQTLs replicated in the validation datasets with an average cosine similarity of 0.70 (0.72, 0.63, and 0.45 for the three mentioned studies; P<2e-200 for all), which increased to 73% and 0.75 respectively after conditioning on significantly heritable genes and the primary signal (see **Methods**). The diminished replication performance of GEUVADIS likely resulted from a combination of significantly reduced sample sizes, admixture differences between African YRI and American Africans in GENOA, and genotyping differences (see **Methods**). Furthermore, SuShiE exhibited similar replication ratios and cosine similarities compared to existing methods, suggesting the higher number of e/pGenes identified by SuShiE were not likely due to false positives (**Table S7**; see **Methods**).

Overall, by jointly modeling multi-ancestry data, SuShiE identifies additional *cis*-regulatory mechanisms for molecular traits.

SuShiE identifies putative eQTL for *URGCP*

Here, we showcase a putative eQTL for *URGCP*, a gene on chromosome 7 that has been implicated in tumor growth and progression[62–66]. SuShiE fine-mapped a single SNP in TOPMed-MESA mRNA (*rs2528382*; GRCh38: 7:43926148; PIP=0.94; **Fig. 4A**), while alternative methods did not produce credible sets for this gene. Importantly, SuShiE replicated *rs2528382* in TOPMed-MESA visit-5 mRNA data. We found *rs2528382* was reported as significant in whole blood eQTL data from the eQTLGen Consortium[4], the Study of African Americans, Asthma, Genes, and Environments (SAGE), and the Genes-Environments and Admixture in Latino Asthmatics (GALA II) study[31], further supporting its role in regulating *URGCP* expression levels. Investigating the functional consequences of *rs2528382* using genomic annotations, we found *rs2528382* represents a non-coding exon variant within the 5' UTR[67], and localizes within a proximal enhancer region (pELS), as evidenced by strong signals of H3K27ac in PBMCs[58] falling within 2kb of the TSS (**Fig. 4B**). Lastly, through snATAC-seq[59] and scATAC-seq[60], we found *rs2528382* localizes within an open chromatin accessibility region measured in different cell types, such as PBMCs, naive T cells, naive B cells, cytotoxic NK cells, and monocytes. Altogether, these results suggest that *rs2528382* regulates *URGCP* expression levels in PBMCs through disruption of regulatory activity.

SuShiE reveals heterogeneity of *cis*-molQTL effect sizes at the loss-of-function intolerant genes

After validating *cis*-molQTLs identified by SuShiE, we next sought to characterize genetic architectures of molecular traits across ancestries. First, we computed *cis*-SNP heritability for all e/pGenes of each ancestry and observed 87% significant heritable genes (in at least one ancestry) across studies (**Fig. S15**), which resulted in highly correlated estimates across ancestries (**Fig. S16**). Next, using SuShiE-derived estimates of *cis*-molQTL correlation across ancestries (see **Methods**), we found highly consistent effect-size correlations on average (0.81,

0.86, and 0.87 for EUR-AFR, EUR-HIS, and AFR-HIS, respectively), which further increased when focusing on genes whose heritabilities are significant in all ancestries (0.94, 0.98 and 0.99, respectively; 9,885 genes; 46.9%; **Figs. S17-S18**). Altogether, these results further affirm previous results[20,21,23,68–74] demonstrating primarily shared genetic architectures for molecular traits across ancestries.

Despite this evidence, we observed a long tail of heterogeneous effect sizes (i.e., SuShiE-estimated effect size correlation <1), suggesting the presence of ancestry-specific *cis*-molQTL effects (**Fig. S19**), which is consistent with previous multi-ancestry *cis*-molQTL studies[27,31,72]. To characterize this apparent heterogeneity across ancestries, we correlated the estimated correlation signals with multiple measures of constraint (pLI[75], LOEUF[76], EDS[77], RVIS[78], and s_{het}[79]) and found highly significant associations (**Table 1**; see **Methods**). Overall, genes with lower effect-size correlations across ancestries exhibited higher intolerance to loss-of-function mutations on average. For example using TOPMed-MESA mRNA dataset, we observed an average *cis*-molQTL effect size correlation of 0.81 (when L=1; SE=0.02) between EUR and AFR individuals at genes that exhibited pLI >0.9, which increased to 0.86 (when L=1; SE=0.01) when focusing on genes with pLI <0.1. Genes with high constraint exhibited lower estimates of *cis*-SNP heritability on average (**Table S8**), which may result in apparent heterogeneity arising from low statistical power. Given this, we re-analyzed putative relationships using estimated covariances, only primary signals (L=1), and bootstrapped standard errors and found broadly consistent results (**Table 1**). In addition, we observed our results were robust to adjusting for Wright's fixation index (F_{st}; **Table 1**; see **Methods**), suggesting heterogeneity/constraint associations are not driven solely by allele frequency differences across ancestries.

To investigate the relationship between *cis*-molQTLs identified by SuShiE and gene constraint, we first observed inverse associations between the number of fine-mapped *cis*-molQTLs per gene and constraint (**Fig. S20**), consistent with several previous studies showing the depletion of *cis*-molQTLs for high constraint genes[56,77,80]. However, we also observed positive associations between expected *cis*-molQTLs' distance to TSS and constraint, affirming previous results that high constraint genes tended to have more complex regulatory regions[56,77] (**Fig. S21**; see **Methods**). In addition, we correlated gene enrichment scores from ENCODE[58] cCREs with constraint

scores. We found that putative causal *cis*-molQTLs for high constraint genes tended to be enriched for distal enhancers (dELS) and depleted for promoter (PLS) and proximal enhancers (pELS) compared with weakly constrained genes, consistent with several previous studies[56,77] (**Fig. S22**). We found these associations remained significant after accounting for F_{st}, suggesting average allele frequency differences across ancestries cannot solely explain the observed heterogeneity.

Overall, SuShiE recapitulates the findings of primarily shared genetic architectures of molecular traits and show that effect size heterogeneity is consistent with gene LOF intolerance.

Posterior *cis*-molQTL effect sizes improve T/PWAS power in white blood cell traits

Lastly, to showcase the downstream benefits of SuShiE, we performed TWAS and PWAS[42–44] on six white blood-cell-related traits in AOU biobank[50] (average n=86,336; **Table S9**). First, we assessed the predictive performance of SuShiE-based weights compared to alternative expression-prediction methods. Specifically, SuShiE obtained better cross-validation estimates (cv-r^2) compared to SuShiE-Indep, Meta-SuSiE, SuSiE, Elastic Net and gBLUP (2 out of 5 comparisons P<0.05) and comparable estimates relative to LASSO (P=0.64; **Fig. S23A**). When focusing on genes with estimated *cis*-molQTL effect size correlation <0.9 across ancestries, we find SuShiE consistently outperformed other approaches (4 out of 6 comparisons P<0.05; **Fig. S23B**), suggesting the benefits in modeling and learning the prior effect size covariances. We observed significantly decreased prediction performance when evaluating cross-ancestry prediction (e.g., predicting mRNA expression of AFR using EUR weights; see **Methods**; P=1.71e-53; **Fig. S24**), consistent with previous works[22,27,36,81] and further motivating ancestry-matched analyses. Given this, we predicted the expression levels of 20,515 genes (mRNA) and 573 proteins using ancestry-matched SuShiE *cis*-molQTL prediction weights from the above analyses and AOU genotypes. Overall, we identified 221 T/PWAS significant associations in white blood count (WBC), eosinophil count (EOS), and monocyte count (MON; **Table S10**; **Fig. S25**). Of these associations, ~90% were identified in WBC due to substantially increased statistical power (i.e., 21,476 more participants on average). We found no significant associations in lymphocyte count (LYM),

neutrophil count (NEU), and basophil count (BAS), likely due to low detected cell counts, similar to previous studies[36,82] that identified fewer associations compared to models based on WBC.

Consistent with our simulation results (**Fig. 2F**), SuShiE demonstrated higher T/PWAS chi-square statistics and identified 44 more T/PWAS associations compared to results driven by SuSiE prediction weights (**Fig. 5A**). In addition, we observed that the SuShiE T/PWAS signals associated with multiple measures of LOF intolerance (**Table S11**), in contrast to previous work demonstrating that high LOF intolerance genes are typically depleted in TWAS models due to weak eQTL signals[56,77,80] (**Fig. 5B**; see **Methods**). We found less support for a relationship between SuSiE-based TWAS signals and LOF intolerance (P=9.21e-10; **Table S11**), further demonstrating SuShiE's advantage. To validate our results, we compared our TWAS statistics with multiple independent white blood cell-related TWASs[31,36,82–84]. Overall, we found SuShiE-based TWAS replicated at rates similar to SuSiE, suggesting that its improved power is unlikely due to false positives and further highlighting its benefit in identifying disease-related genes.

Overall, our work has shown that by jointly modeling the molecular data across different ancestries while allowing effect sizes to differ, SuShiE outputs more accurate *cis*-molQTL prediction weights, thus boosting the downstream statistical power for integrative analyses with GWASs.

Discussion

Here, we present the Sum of Shared Single Effect approach (SuShiE), a novel approach for multi-ancestry SNP fine-mapping of molecular traits using a scalable variational approach. SuShiE assumes the joint *cis*-molQTL effects arise as a linear combination of per-ancestry effect sizes across shared causal variants. Through extensive simulations, SuShiE first improved the fine-mapping precision in disentangling the causal *cis*-molQTLs from tagging SNPs by leveraging LD heterogeneity across diverse ancestries. Second, SuShiE accurately learned prior effect size correlations across ancestries employing a procedure analogous to Empirical Bayes. Third, SuShiE estimated ancestry-specific cis-molQTL prediction weights, boosting findings in the post-GWAS framework (e.g., TWAS and

PWAS), compared to the baselines that did not model effect size covariance across ancestries or ignored ancestry altogether. We applied SuShiE to 36,911 molecular phenotypes of diverse ancestries from three datasets: mRNA and protein expression from TOPMed-MESA and mRNA expression from GENOA. SuShiE fine-mapped 16% more genes on average compared to the existing methods, exhibiting smaller credible set sizes and higher enrichment in relevant functional annotations. SuShiE inferred highly correlated *cis*-molQTL effect sizes across ancestries on average in significantly heritable genes, reflecting primarily shared *cis*-molQTL architectures. In addition, we observed *cis*-molQTL effect size heterogeneity across ancestries associated with multiple constraint measurements, consistent with environmental interactions may partially drive differences in effect sizes across ancestries. Last, we performed TWAS and PWAS on six white blood cell-related traits from AOU biobank using SuShiE-derived ancestry-specific *cis*-molQTL prediction weights and identified 44 more significant genes compared to the existing method. We also observed that SuShiE T/PWAS signals are associated with multiple measures of LOF intolerance, further showing the benefit of multi-ancestry approaches in identifying genes relevant to complex disease risk.

Next, we describe caveats in our real data analysis. First, SuShiE approximates ancestry as a discrete category, allowing us to model *cis*-molQTL effect sizes using a multivariate normal distribution (see **Methods**). While this simplifies modeling and inference tasks, we emphasize that this is a heuristic approach that neglects the complex and shared demographic histories underlying all humans. Indeed, recent work has demonstrated the importance of viewing genetic ancestries as a continuous spectrum rather than discrete categories[85]. Relatedly, previous studies[33,35,37–41] and our simulation results (**Fig. S1**) have shown that increasing the number of ancestries within a multi-ancestry framework improves fine-mapping precision. However, SuShiE and similar frameworks perform inference on variants present after filtering on MAF thresholds (e.g., 1%) within each ancestry. As a result, this requirement can exclude *cis*-molQTLs from analysis due to small sample sizes within an ancestry, suggesting a trade-off in practice between increasing overall sample size versus excluding informative genetic variants. For instance, we obtained mRNA expression data measured in EUR (n=402), AFR (n=175), HIS (n=277), and East Asian

ancestry (EAS; n=96) individuals from TOPMed MESA study visit-1. From two-ancestry fine-mapping (EUR and AFR) to three-ancestry (+HIS), we filtered an additional 29 SNPs per gene on average. However, this number increased to 501 SNPs by including the additional 96 participants of EAS ancestry. As a result, we opted to not include EAS participants in our analysis in order to maximize the genetic variants analyzed. Modeling genetic ancestry continuously can potentially avoid this type of *cis*-molQTL loss, thus improving the fine-mapping precision with a larger sample size.

Second, we note that our data consist of African- (AFR) and Hispanic-American (HIS) individuals, which contain recent admixture events. To account for complex diversity within ancestries, we included genotyping PCs as a covariate in our models. Several works have suggested that admixture can be sufficiently corrected for using global ancestry information (i.e., genotyping PCs) in association testing[73,86–91], especially when causal effect sizes are largely consistent across ancestries[86,87,89] (**Fig. S16-S18**). On the other hand, accounting for local ancestry may increase the associating testing power when causal effects are highly different across ancestries[86,87,92] or aid fine-mapping in post-GWAS analysis[87,89,93], which can be one of the future directions for SuShiE.

Third, we observed significant associations between gene LOF intolerance and several SuShiE-estimated metrics, including effect size heterogeneity across ancestries, the number of *cis*-molQTLs, *cis*-molQTL distance to TSS, and functional enrichments. The relationship remained significant after adjusting for F_{st}, suggesting allele frequency differences across ancestries are not sufficient to fully explain estimated heterogeneity. As a result, we hypothesized that *cis*-molQTL effect size heterogeneity could be in part due to gene-by-environment (GxE) interactions[69,77,94–96]. Highly constrained genes exhibit more complex regulatory landscapes with fewer *cis*-molQTLs (or apparent *cis*-molQTLs due to smaller effect sizes)[56,77]. As a result, these genes may be less resilient to environmental perturbations[77], which may induce effect-size heterogeneity across different ancestries. On the other hand, it is possible that our F_{st} estimates are underpowered to detect subtle allele frequency differences across ancestries. Therefore, these associations may provide indirect evidence for natural selection partially driving *cis*-molQTL effect size heterogeneity across ancestries. To explicitly investigate the role of selection in

335 molecular differences across ancestries, we likely require a more principled modeling procedure based in
336 population genetics together with higher-resolution molecular data measured in diverse ancestries[56,80,97–100]. For
337 instance, recent work has shown the promise of using single-cell data to demonstrate how selection impacts genes
338 expressed differentially across ancestries[101].

339 Fourth, SuShiE assumes causal *cis*-molQTLs are shared across ancestries. Our simulations show that SuShiE
340 remains robust when ancestry-specific *cis*-molQTLs are present (**Fig. S8**). However, in situations where there exist
341 shared *cis*-molQTLs but ancestries have different sample sizes, SuShiE may prioritize shared *cis*-molQTLS along
342 with SNPs tagged in LD of the ancestry with larger sample sizes, evidenced through simulations (**Fig. S5B**). However,
343 through our case study in *URGCP* (**Fig. 4**), we observed relatively higher signals in AFR but not in EUR and HIS,
344 despite AFR having the smallest sample size, suggesting this limitation may be minimal overall.

345 Lastly, in our T/PWAS analysis, we selected six white blood-cell related traits to best match PBMC and LCL contexts.
346 However, alternative cell-types not included in our analyses may better capture relevant contexts. For example,
347 PBMCs and LCLs do not contain neutrophils, basophils, and eosinophils, and LCLs additionally do not include
348 monocytes, which may result in a loss in statistical power. As single-cell RNA-seq datasets become more
349 available[102], one possible direction would be to perform TWAS in fine-grained cellular contexts and backgrounds.
350 In addition, after predicting expression levels using ancestry-matched weights for each individual, we performed
351 individual-level T/PWAS by concatenating the predicted expression levels across ancestries rather than perform
352 ancestry-specific TWAS followed by meta-analysis[103]. The premise of the meta-analysis approach is that
353 researchers obtain ancestry-specific GWAS and then integrate with corresponding eQTL weights. Because the
354 causal genes for complex traits are likely shared across ancestries[20,21,23,36,68–74], a regression framework with
355 individual-level data concatenated across ancestries (the largest sample size) can maximize power.

356 We briefly discuss potential directions for future work. First, recent studies have shown that incorporating
357 functional annotation in the prior distribution can improve the fine-mapping precision[33,79,104]. SuShiE currently
358 employs a uniform distribution for prior causal probability. Including functionally-informed priors is likely to

improve its performance further. Second, SuShiE fine-maps individual-level molecular and genotypic data in a prespecified locus flanking the TSS and TES regions of a gene. In theory, users can apply SuShiE on individual-level complex trait data, however, this likely will require additional analyses (e.g., pre-specifying GWAS significant loci) and care in controlling for genome-wide backgrounds and population structure. In addition, the limited accessibility to the individual-level complex trait data allows the extension of SuSiE-like models to be compatible with summary statistics[38,39,41,51], which typically requires external LD reference panels. As more *cis*-molQTL summary statistics are available to the community[4,102], we foresee a potential demand to implement this compatibility in SuShiE. Last, SuShiE currently cannot model molecular data in their original read-count format, which is usually transformed to a continuous scale (i.e., inverse normal transformation). Extending SuShiE to a GLM-like model naturally would encompass this scenario and present an exciting direction for SuShiE.

Overall, SuShiE, together with its application on large-scale molecular data of diverse ancestries, identifies more *cis*-regulatory mechanisms and reveals its genetic architecture. We anticipate considerable demand for our approach in the genetics field characterized by forthcoming multi-ancestry and multi-omics research.

Online Methods

Sum of Shared Single Effects Model

Here, we describe the statistical model underlying SuShiE (see **Supplementary Note** for a detailed description).

SuShiE assumes *cis*-molQTLs are present in *all* ancestries, defined as shared *cis*-molQTLs while allowing for effect sizes at causal *cis*-molQTLs to covary across ancestries a-priori. For the i^{th} of total k ancestries, SuShiE models the centered and standardized levels of a molecular trait $\boldsymbol{g}_i$ measured in n_i individuals as a linear combination of p genotyped variants $\boldsymbol{X}_i$ as

$$\boldsymbol{g}_i = \boldsymbol{X}_i\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i$$

where $\boldsymbol{\beta}_i$ is a $p \times 1$ vector of ancestry-specific *cis*-molQTL effects, and $\boldsymbol{\epsilon}_i \sim N(0, \sigma_{i,e}^2 \boldsymbol{I}_{n_i})$ is environmental noise. In addition, we model $\boldsymbol{\beta}_i = \sum_{l=1}^{L} \boldsymbol{\beta}_{i,l}$ as the sum of L effects $\boldsymbol{\beta}_{i,l} = \boldsymbol{\gamma}_l \cdot b_{i,l}$ where $\boldsymbol{\gamma}_l$ is a $p \times 1$ binary vector indicating which variant is the shared *cis*-molQTL for the l^{th} effect while allowing ancestry-specific effect sizes $b_{i,l}$. Furthermore, we model $\boldsymbol{\gamma}_l \sim \text{Multi}(1, \boldsymbol{\pi})$ where $\boldsymbol{\pi}$ is a $p \times 1$ vector representing prior probability for each SNP to be a *cis*-molQTL, and model $b_l = \left[b_{1,l} \ \cdots \ b_{i,l} \ \cdots \ b_{k,l} \right] \sim N(\mathbf{0}, \boldsymbol{C}_l)$ where

$$\boldsymbol{C}_{i,i',l} = \begin{cases} \sigma_{i,b,l}^2 & \text{if } i = i' \\ \rho_{i,i',l} \sigma_{i,b,l} \sigma_{i',b,l} & \text{otherwise'} \end{cases}$$

$\boldsymbol{C}_l$ is the l^{th} prior $k \times k$ effect size covariance matrix with $\sigma_{i,b,l}^2$ as variance, and ρ_l as correlation.

Variational inference of model parameters

To infer the *cis*-molQTL effects, we seek to estimate the posterior distribution of $\Pr(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k | \mathbf{Data}) = \Pr(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k \mid \boldsymbol{g}_1, \ldots, \boldsymbol{g}_k, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_k, \boldsymbol{C}, \boldsymbol{\pi}, \sigma_{1,e}^2, \ldots, \sigma_{k,e}^2)$ where $\boldsymbol{C} = \{\boldsymbol{C}_1, \ldots, \boldsymbol{C}_L\}$. We regard $\boldsymbol{\beta}_*$ as latent variables, $\boldsymbol{g}_*$, and $\boldsymbol{X}_*$, as observed data, and $\boldsymbol{C}$, $\boldsymbol{\pi}$, and $\sigma_{*,e}^2$ are the hyperparameters. However, inferring the exact distributions of latent variables is computationally intractable due to non-conjugacy with the prior distribution. Therefore, we seek a surrogate distribution $Q(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k)$, which minimizes the Kullback–Leibler (KL) divergence with $\Pr(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k | \mathbf{Data})$. Specifically, we have:

$$D_{KL}\left(Q(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k) \,||\, \Pr(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k | \mathbf{Data})\right)$$

$$= \log \Pr(\boldsymbol{g}_1, \ldots, \boldsymbol{g}_k | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_k, \boldsymbol{C}, \boldsymbol{\pi}, \sigma_{1,e}^2, \ldots, \sigma_{k,e}^2)$$

$$- E[\log \Pr(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k, \boldsymbol{g}_1, \ldots, \boldsymbol{g}_k \mid \boldsymbol{C}, \boldsymbol{\pi}, \sigma_{1,e}^2, \ldots, \sigma_{k,e}^2) - \log Q(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)]$$

where the first term is the log evidence, and the expectation term is the evidence lower bound (ELBO). Since the log evidence is constant with respect to model variables, minimizing the KL divergence is equivalent to maximizing the ELBO. Furthermore, to limit the universe of possible forms that the surrogate distribution $Q(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k)$ may take, we impose an additional mean-field assumption[105]. Namely, SuShiE assumes that each of the L shared effects β_l are mutually independent under Q:

$$Q(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k) = \prod_{l=1}^{L} Q(\boldsymbol{\beta}_{1,l}, \ldots, \boldsymbol{\beta}_{k,l}) = \prod_{l=1}^{L} Q(\boldsymbol{b}_l|\boldsymbol{\gamma}_l) Q(\boldsymbol{\gamma}_l).$$

Therefore, to approximate the posterior distributions $Q(\cdot)$ for latent variables $\boldsymbol{b}_{l,j}$ (a $k \times 1$ vector) and $\boldsymbol{\gamma}_{l,j}$ (a scaler) at SNP $j \in [1, p]$ of l^{th} shared effect, we need to compute the expectation of complete data log-likelihood $L(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k, \boldsymbol{g}_1, \ldots, \boldsymbol{g}_k \mid \boldsymbol{C}, \boldsymbol{\pi}, \sigma_{1,e}^2, \ldots, \sigma_{k,e}^2)$ (i.e., the joint distribution) while holding other variables constant. Through the principles of coordinate-ascent variational inference (CAVI)[105], we can identify each $Q(\cdot)$ surrogate as,

$$Q\left(\boldsymbol{b}_{l,j} \mid \boldsymbol{\gamma}_{l,j} = 1\right) = N\left(\boldsymbol{b}_{l,j} \mid \boldsymbol{\mu}_{l,j}, \boldsymbol{\Sigma}_{l,j}\right)$$

$$Q\left(\boldsymbol{\gamma}_{l,j} = 1\right) \propto \text{softmax}(\log \boldsymbol{\pi}_j - \log N\left(\boldsymbol{\mu}_{l,j} \mid \boldsymbol{0}, \boldsymbol{\Sigma}_{l,j}\right))$$

$$Q(\boldsymbol{\gamma}_l) = \text{Multi}\left(\boldsymbol{\gamma}_l \mid 1, \boldsymbol{\alpha}_l\right)$$

where $\boldsymbol{\mu}_{l,j} \in \mathbb{R}^{k \times 1}$ and $\boldsymbol{\Sigma}_{l,j} \in \mathbb{R}^{k \times k}$ are the corresponding posterior mean and covariance, and $\boldsymbol{\alpha}_l \in \mathbb{R}^{p \times 1}$ is each SNP's posterior probability to explain the l^{th} effect. We provide the complete mathematical derivations, inference algorithms, and detailed definitions in the **Supplementary Note**.

Computing posterior inclusion probability and η-credible sets

We define the posterior inclusion probability (PIP) for SNP j with $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_L$ as $\text{PIP}_j \coloneqq 1 - \prod_{l=1}^{L}(1 - \boldsymbol{\alpha}_{l,j})$. To compute an η-credible set for each L, where η represents the desired probability that the set contains *cis*-molQTLs, we decreasingly sort $\boldsymbol{\alpha}_l$ and take a greedy approach to include SNPs until their cumulative sum exceeds η.

In the case that the inferred number of effects L surpasses the actual number of *cis*-molQTLs, the unnecessary credible sets will contain most SNPs with low posterior probability close to $\boldsymbol{\alpha}_{l,j} = 1/p$, where p is the number of SNPs. To refine the final inference results, we remove the credible sets whose lowest absolute pairwise correlation, which is defined as "purity"[15] and weighted by sample size across all ancestries, among SNPs is less than 0.5. In practice, following the previous work[15], we empirically specify L as 10.

Inferring cross-ancestry effect size correlations

SuShiE features the capability to estimate the correlation of *cis*-molQTL effect sizes across multiple ancestries. For some gene t, SuShiE by default outputs L estimates of the effect size correlation $\hat{\rho}_{t,1}, \ldots, \hat{\rho}_{t,L}$ for each credible set. If we apply SuShiE to T genes in total, we empirically recommend computing effect size correlation across ancestries with $\hat{\rho} = \frac{1}{T}\sum_{t=1}^{T} \hat{\rho}_{t,1}$.

Simulating genotypes and quantitative molecular traits

To evaluate SuShiE's performance in simulations, we first simulated genotypes and quantitative molecular traits to mimic the real-world scenarios using our previous simulation frameworks[36,106,107]. To simulate genotype data, we used LD estimates from individuals of European (EUR; n=489), African (AFR; n=639), and East Asian (EAS; n=481) ancestries from the 1000 Genomes Project (1000G) phase three data[108]. We limited LD to biallelic HapMap SNPs[109], discarded those with missingness (>1%), MAF (<1%), and violated Hardy-Weinberg equilibrium (HWE mid-adjusted P< 1e-6). We obtained chromosome, transcription start site (TSS), and transcription end site (TES) information for 19,279 protein-coding autosomal genes using GENCODE release 26 (GRCh37)[110]. We extended each gene 500,000 base pairs (bp) upstream of TSS and 500,000 bp downstream of TES, and randomly selected 500 genes that have at least 500 common SNPs across EUR, AFR, and EAS genotypes.

We first focused on simulations using EUR and AFR ($k = 2$). At each gene, we simulated centered and standardized genotype matrix $\boldsymbol{X}_i \in \mathbb{R}^{n_i \times p}$ for i^{th} ancestry using a multivariate normal distribution $N(0, \boldsymbol{V}_i)$ where $n_i \in \{200, 400, 600, 800\}$ is the *cis*-molQTL study sample size, p is the number of common SNPs across ancestries in the locus, and $\boldsymbol{V}_i \in \mathbb{R}^{p \times p}$ is the ancestry-specific LD matrix estimated from 1000G genotypes[108]. Next, we uniformly chose $m \in \{1, 2, 3\}$ out of p common SNPs as *cis*-molQTLs and simulated their ancestry-specific effect sizes $\widetilde{\boldsymbol{\beta}}_1, \widetilde{\boldsymbol{\beta}}_2 \in \mathbb{R}^{m \times 1}$ under a bivariate normal distribution as

$$(\widetilde{\boldsymbol{\beta}}_1, \widetilde{\boldsymbol{\beta}}_2) \sim N\left(\mathbf{0}, \begin{bmatrix} h^2_{g,1} & \rho \cdot \sqrt{h^2_{g,1} \cdot h^2_{g,2}} \\ \rho \cdot \sqrt{h^2_{g,1} \cdot h^2_{g,2}} & h^2_{g,2} \end{bmatrix} / m \right) \otimes I_m\Bigg),$$

where $h^2_{g,i} \in \{0.01, 0.05, 0.1, 0.2\}$ is the proportion of variance in gene expression explained by *cis*-molQTLs (i.e., *cis*-SNP heritability of the molecular trait) and $\rho \in \{0.01,\ 0.4, 0.8, 0.99\}$ is the effect size correlation. Then, we constructed effect-size vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, where $\widetilde{\boldsymbol{\beta}}_1$ and $\widetilde{\boldsymbol{\beta}}_2$ are the m non-zero entries at the same index representing shared *cis*-molQTLs and the rest $p - m$ entries are zero representing the null SNPs. Next, we computed the quantitative molecular traits $\boldsymbol{g}_i$ using $\boldsymbol{X}_i\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i$ where $\boldsymbol{\epsilon}_i \sim N\left(\mathbf{0}, s^2_{g,i}\left(\frac{1}{h^2_{g,i}} - 1\right)\boldsymbol{I}_{n_i}\right)$ is the random environmental noise and $s^2_{g,i} = \boldsymbol{\beta}_i^T \boldsymbol{V}_i \boldsymbol{\beta}_i$ is the genetic variance after accounting for LD. To reflect cases where heterogeneity exists in the genetic architecture of molecular traits across ancestries[31,72], we allowed *cis*-SNP heritability to be ancestry-specific with $h^2_{g,1} = 0.05$ and $h^2_{g,2} \in \{0.01, 0.05, 0.1, 0.2\}$; we also evaluated the performance under different statistical power where $n_1 = 400$ and $n_2 \in \{200, 400, 600, 800\}$. To determine whether incorporating additional ancestry improves SuShiE's performance, we simulated the genotypic and phenotypic data for EAS with the same total sample sizes and genetic architecture.

In addition, we simulated two cases under model misspecification. We first evaluated SuShiE's performance when ancestry-specific *cis*-molQTLs exist, we simulated $m_{i,AS} \in \{1, 2, 3\}$ *cis*-molQTLs for both ancestries in addition to shared *cis*-molQTLs $m = 2$ while fixing $h^2_{g,i} = 0.05$ for ancestry i. Second, to reflect cases where the number of shared *cis*-molQTL (m) is different from inferred L by fixing $m = 2$ and varying the inferred $L \in \{2, 5, 10\}$.

Default parameters and performance metrics

We performed SNP fine-mapping using SuShiE on simulated genotypes and molecular data across EUR and AFR individuals. In terms of variational inference parameters, we specified $L \in \{1, 2, 3\}$ to match the actual number of simulated effects and initialized *cis*-molQTL effects $\widehat{\boldsymbol{b}}_{l,j}$ as $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$, their covariance matrix $\hat{C}_l$ as $\begin{bmatrix} 0.001 & 0.1 \\ 0.1 & 0.001 \end{bmatrix}$, the

prior estimates of environmental noises $\widehat{\sigma}^2_{i,e}$ as 0.001, the prior probability for SNPs to be *cis*-molQTLs as $1/p$ where p is the number of common SNPs.

To evaluate the gain in parametrizing the effect size correlation across ancestries, we compared our method SuShiE to "SuShiE-Indep" which assumes the *cis*-molQTL effect sizes are independent across ancestries; that is, we fixed the effect size correlation prior $\rho = 0$, and did not learn it through the Empirical-Bayes-like procedure. To demonstrate that SuShiE's improvement does not result from the accumulation of samples across ancestries, we compared SuShiE's performance to two "baseline" methods: first, we performed single-ancestry SuSiE and then meta-analyzed the resulting PIPs by $\mathrm{PIP}_{\mathrm{meta}} = 1 - (1 - \mathrm{PIP}_{\mathrm{EUR}}) \cdot (1 - \mathrm{PIP}_{\mathrm{AFR}})$; we refer to this method as "meta-SuSiE". Second, we row-stacked the genotype matrices and molecular trait vectors across ancestries and then performed single-ancestry SuSiE as "SuSiE." Overall, we performed four methods (SuShiE, SuShiE-Indep, meta-SuSiE, and SuSiE) on 500 genes' simulated genotypes and molecular traits to output corresponding PIPs, credible sets, and ancestry-specific effect size estimates. We varied four parameters: per-ancestry *cis*-molQTL study sample size (n_i), the number of *cis*-molQTLs (m), the *cis*-SNP heritability of molecular traits (h^2_g) for each ancestry, and the effect size correlation (ρ). To reflect a practical study design, the default parameters were fixed at $n_i = 400$, $m = 2$, $h^2_g = 0.05$, and $\rho = 0.8$ unless stated otherwise. Furthermore, we evaluated the fine-mapping performance with three metrics across 500 simulated genes: PIPs at causal *cis*-molQTLs, credible set size, and frequency that causal *cis*-molQTLs are contained in 90% credible sets (calibration). We computed the metrics of meta-SuSiE based on the union of the credible sets across two single-ancestry SuSiE. As different methods may or may not prune credible sets at the same simulated gene, to show a fair comparison, we computed the credible set size metric only using the credible set that none of the four methods pruned out. To compare metrics across methods, we ran linear regression adjusted for relevant simulation parameters and reported one-sided Wald test P values.

Simulating GWAS and TWAS

Transcriptome-wide Association Studies (TWASs) leverage GWAS summary statistics, eQTL prediction weights, and LD reference to identify genes whose predicted expression levels are associated with complex traits[42–44]. A more accurate eQTL prediction weight will increase the power of the TWAS framework. Therefore, we compared the prediction weights inferred by SuShiE to other methods: SuShiE-Indep, Meta-SuSiE, SuSiE, least absolute shrinkage and selection operator (LASSO)[53], elastic net regularization (Elastic Net)[54], and genomic best linear unbiased prediction (gBLUP)[55]. We simulated the expression and genotype data for the training and testing set separately, with the same method mentioned in the previous sections. For the training set, we varied the per-ancestry sample size $n_t \in \{200, 400, 600, 800\}$ and set the out-of-sample testing set sample size $n_v = 200$. Then, we predicted the expressions using ancestry-matched fitted weights on testing genotype data, and computed the coefficients of determination (r^2) between the predicted and simulated expression. For Meta-SuSiE, we trained the prediction weights for each ancestry using per-ancestry sample size. For SuSiE, LASSO, Elastic Net, and gBLUP, we trained the prediction weights after concatenating data across ancestries to guarantee that the total sample sizes were the same as SuShiE as fair comparisons.

To showcase that SuShiE's prediction weights introduce more power in TWAS, we simulated GWAS summary statistics and computed TWAS statistics using different prediction weights. First, because individuals in GWASs are usually different from ones in the eQTL studies, we re-simulated the genotype matrix $\boldsymbol{X}_{\mathrm{GWAS},i} \in \mathbb{R}^{n_{\mathrm{GWAS},i} \times p}$ where $n_{\mathrm{GWAS},i}$ is the GWAS sample size for ancestry i using the same generating approach above. Then, we used the eQTL effect size vectors $\boldsymbol{\beta}_i$ generated in the previous section to simulate a complex trait $\boldsymbol{y}_i \in \mathbb{R}^{n_{\mathrm{GWAS},i} \times 1}$ as a linear combination of expression levels $\boldsymbol{g}_i \in \mathbb{R}^{n_{\mathrm{GWAS},i} \times 1}$ as

$$\boldsymbol{y}_i = \boldsymbol{g}_i \delta + \boldsymbol{e}_i = \boldsymbol{X}_{\mathrm{GWAS},i} \boldsymbol{\beta}_i \delta + \boldsymbol{e}_i,$$

where $\delta \sim N(0,1)$ is the gene expression effect on the complex trait, $\boldsymbol{e}_i \sim N\left(0, s_i^2 \left(\frac{1}{h_{GE}^2} - 1\right) I_{n_{\mathrm{GWAS},i}}\right)$ is the random noises for the complex traits, $s_i^2 = \boldsymbol{\beta}_i^T \boldsymbol{V}_i \boldsymbol{\beta}_i \delta^2$, $\boldsymbol{V}_i$ is the LD matrix generated from 1000G[108], and $h_{GE}^2 \in$

$\{6 \times 10^{-5}, 1.5 \times 10^{-4}, 3 \times 10^{-4}, 6 \times 10^{-4}\}$ is the proportion of variation of the complex trait explained by the expression of a single gene[111]. Then, we regressed the complex trait $\boldsymbol{y}_i$ on each SNP in $\boldsymbol{X}_{\mathrm{GWAS},i}$ marginally to compute the GWAS summary statistics $\boldsymbol{z}_{\mathrm{GWAS},i} \in \mathbb{R}^{p \times 1}$. Last, we computed TWAS summary statistics with $\boldsymbol{z}_{\mathrm{TWAS},*,i} = \frac{\boldsymbol{w}_{*,i}^T \boldsymbol{z}_{\mathrm{GWAS},i}}{\boldsymbol{w}_{*,i}^T \boldsymbol{V}_i \boldsymbol{w}_{*,i}}$ along with its P value where $\boldsymbol{w}_{*,i}$ is the prediction weights fitted by different methods. We define the TWAS power as the frequency of the Bonferroni-corrected P value is less than 0.05.

Overview of real-data analyses

We applied SuShiE and other methods (e.g., SuShiE-Indep, Meta-SuSiE, and SuSiE) to three datasets: mRNA (visit-1) measured in peripheral blood mononuclear cells (PBMCs) and protein expression measured in plasma of three EUR, AFR, and HIS ancestries from Trans-Omics for Precision Medicine program Multi-Ethnic Study of Atherosclerosis (TOPMed MESA)[48,49] and mRNA expression measured in lymphoblastoid cell lines (LCLs) of EUR and AFR ancestries from the Genetic Epidemiology Network of Arteriopathy (GENOA) study[26]. We excluded the mRNA expression levels data measured in T cells and monocytes from TOPMed MESA study due to relatively smaller sample sizes. We explain the detailed quality control (QC) procedure in the sections below. We conducted pairwise comparisons of methods on four basic summary statistics, focusing on the genes for which both methods output credible sets; the summary statistics included the number of genes identified with *cis*-molQTLs (e/pGenes), the average PIPs of the SNPs in the credible sets, the average single-effect-specific credible set sizes, and the frequency of having genes whose credible sets contained SNPs with PIPs greater than 0.9. We defined the number of *cis*-molQTLs as the number of credible sets output after pruning for purity (see previous section for the definition). Next, we performed enrichment analyses using 89 functional annotations and a case study focusing on a gene that was only identified by SuShiE, and missed by other methods. Last, using SuShiE-derived ancestry-specific *cis*-molQTL effect sizes, we performed individual-level TWAS and PWAS with All Of Us (AOU) biobank[50] individuals and compared to the results derived from SuSiE.

To validate SuShiE's results on the three main datasets mentioned above, we applied SuShiE and other methods to three separate datasets: mRNA expression (visit-5) measured in PBMC of EUR, AFR, and HIS ancestries from TOPMed MESA, protein expression measured in plasma of EUR ancestry from INTERVAL study[5], and the mRNA expression measured in LCL of EUR and Yoruba in Ibadan (YRI) ancestries from the GEUVADIS study[61]. We computed two statistics to evaluate validation performance: first, focusing on the *cis*-molQTLs of e/pGenes identified by SuShiE, the percentage for which SuShiE identified *cis*-molQTLs in the validation datasets. Second, focusing on the credible sets for which we identified the same *cis*-molQTLs in both main and validation studies, we computed the cosine similarity of posterior probabilities ($\boldsymbol{\alpha}_l$) to see whether they prioritized the same SNPs. For SNPs that are not in the overlap between main and validation studies, we manually assigned them a value of 0 for cosine similarity calculation. For each credible set, we randomly shuffled the $\boldsymbol{\alpha}_l$ in validation studies 500 times to construct the null distribution of the cosine similarity and compute its z score. We computed the average cosine similarity and z scores across all credible sets as an aggregation estimate and its corresponding significance. For all the fine-mapping analysis, we used the SNPs that are shared across ancestries on the genomic window of each gene that is 500,000 bp upstream and downstream of each gene's TSS and TES (one million bp in total), respectively, based on the GENCODE v34[110,112]. In addition, we only included genes that are located on the autosomes, do not overlap with the major histocompatibility complex (MHC) region, have more than 100 SNPs on the genomic window present in all ancestries, and whose ENSEMBL gene IDs match the records in GENCODE v34[110,112]. We adjusted for covariates by regressing them from both mRNA/protein levels and each SNP. In addition, we computed the *cis*-SNP heritability using the limix python package (see **Code Availability**) for each analyzed molecule within each ancestry. We used PLINK2.0, vcftools, and bcftools for genotype manipulation[113–116].

Genotype data in the TOPMed MESA study

We obtained the whole-genome sequencing (WGS) data (freeze 9; GRCh 38) of 5,379 individuals from the TOPMed MESA[48,49]. Specifically, we removed the SNPs with the following criteria: both duplicate genotype

discordance and mendelian genotype discordance are greater than 2%, genotype missing rate at depth 10 is greater than 2%, Milk-SVM score for variant quality is less than -0.5, variants that overlap with centromeric regions, HWE p-value is less than 1e-6, and MAF is less than 1%, resulting in a total of 125,089,612 SNPs. In addition, we computed the genotype principal components (PCs) with SNPs that are pruned for LD using PLINK2.0 (--indep-pairwise 200 1 0.3)[113,114,117]. Last, we retained individuals who are self-identified as EUR, AFR, or HIS ancestries and have measurements in mRNA (both visits 1 and 5) and protein datasets, resulting in a total of 1,292 individuals.

mRNA expression data in the TOPMed MESA study

We obtained RNA-seq data in gene-level read counts and reads per kilobase of transcript per million mapped reads (RPKM) of 57,615 genes for 2,137 samples (both visits-1 and visit-5) measured in PBMC using RNA-SeQC v2.0.0 from the TOPMed MESA study. The data was pre-processed based on the TOPMed RNA-seq harmonization pipeline (see **Code Availability**). We first calculated the gene expression PCs on all samples' read counts using the PCA function of the scikit-learn package[118], and normalized it across all samples within each PC. Then, focusing on the samples measured in visit-1, we followed the GTEx[3] eQTL analysis preparation script to select gene whose transcript per million (TPM) is >0.1 and raw read counts >6 reads in at least 20% of samples (see **Code Availability**). For individuals with replicate samples, we only kept one sample with the greatest sum of reads across all genes; we also removed individuals with whom we did not have self-identified ancestry information, resulting in 402 EUR, 175 AFR, and 277 HIS individuals. Then, within each ancestry, we normalized expression levels between samples using edger_cpm function in the pyqtl package, (see **Code Availability**) with normalized_lib_sizes=True, which is a Python implementation of edgeR[119] ; we next performed inverse-rank normalization using the inverse_normal_transform function. Last, focusing on 21,747 genes filtered based on inclusion criteria and using SNPs whose MAF >1% and HWE mid-adjusted P>1e-6 within each ancestry, we ran SuShiE and other methods using SNPs on the genomic windows of each gene, adjusting for 15 gene expression PCs, 10 genotype PCs, age, sex, and the assay lab. We did not include individuals who self-identified as East Asian in TOPMed MESA study due

578 to the small sample size (n=96). We removed SNPs based on MAF <1%, and including EAS participants would
579 exclude 501 more SNPs on average per gene from downstream analyses.

580 Protein expression data in the TOPMed MESA study

581 We obtained the protein expression levels of 1,317 target proteins for 1,966 samples (both visits-1 and -5) from
582 the TOPMed MESA study using SOMAscan, an aptamer-based technology. First, we computed the protein
583 expression PCs on all samples using the PCA function of the scikit-learn package[118], and normalized it across all
584 samples for each PC. Then, focusing on the samples measured in visit-1, we removed individuals with whom we
585 did not have self-identified ancestry information, resulting in 398 EUR, 297 AFR, and 261 HIS individuals. Within
586 each ancestry, we inverse-rank normalized the protein expression data using the inverse_normal_transform
587 function in the pyqtl package (see **Code Availability**). As some proteins may be targeted by multiple aptamers,
588 which correspond to different isoforms of proteins[120], we regarded each isoform as a unique protein. As a result,
589 we obtained 1,274 proteins based on gene inclusion criteria and performed fine-mapping using SuShiE and other
590 methods on the genomic windows adjusted for 15 protein expression PCs, 10 genotype PCs, sex, and age, using
591 SNPs whose MAF > 1% and HWE mid-adjusted P>1e-6 within each ancestry.

592 Genotype and mRNA expression data in the GENOA study

593 From the GENOA study[26], we obtained paired genotype and LCL mRNA expression data of 373 EUR and 441 AFR
594 individuals, together with corresponding covariates, processed by previous works[26,36]. Briefly, we restricted
595 TOPMed-imputed[121] genotype data on biallelic SNPs with imputation score $r^2 > 0.6$, MAF >1%, and HWE mid-
596 adjusted P>1e-6 within each ancestry. Focusing on 14,797 genes based on gene inclusion criteria, we performed
597 fine-mapping on the genomic window, adjusted for 30 gene expression PCs, five genotype PCs, age, sex, and
598 genotyping platform.

Genotype and molecular data in three validation datasets

To validate SuShiE's results of PBMC mRNA expression (visit-1) in TOPMed MESA[48,49], we used the mRNA expression data measured in PBMC of the same study but collected from visit-5, a 10-year-later follow-up visit. We performed the identical pipeline mentioned in the previous section, resulting in 21,695 genes (21,240 overlapped with visit-1) from 422 EUR, 168 AFR, and 285 HIS individuals.

To validate the plasma protein expression results in TOPMed MESA, we obtained the inverse-rank normalized protein expression levels of 3,301 EUR individuals measured in plasma from the INTERVAL study[5]. The genotype data was pre-processed, imputed, and annotated with dbSNP v153 by previous studies[5,122,123]. We obtained 3,187 ENSEMBLE-UniProt-SOMAmer ID triplets (1,313 overlapped with the TOPMed MESA) based on gene selection criteria and performed singe-ancestry SuSiE fine-mapping on the genomic window, adjusted for sex, age, duration between blood draw and process, 3 genotype PCs, and subcohort, and 5 expression PCs, using SNPs whose MAF >1% and HWE mid-adjusted P>1e-6.

To validate the mRNA expression data measured in LCLs from the GENOA study, we obtained paired genotype and gene expression data measured in LCLs in gene-level RPKM of 23,722 genes for 373 EUR and 89 YRI individuals from the GEUVADIS study[61]. First, we computed the expression PCs on all the individuals using the PCA function of the scikit-learn package[118]. Then, we kept high-expressed genes whose TPM >0.1 in at least 20% of all the individuals[3] and filtered based on gene selection criteria, resulting in a total of 19,882 genes (10,439 overlapped with GENOA). Last, using SNPs whose MAF >1% and HWE mid-adjusted P>1e-6 within each ancestry, we performed SuShiE fine-mapping on the genomic window, adjusted for sex, five expression PCs, and five genotype PCs, which is calculated on the LD-pruned pipeline defined in the previous section.

Functional enrichment analyses and case study

We ran functional enrichment analysis only on the genes identified with *cis*-molQTLs (i.e., SuShiE outputs credible sets; e/pGenes). To visualize the relationship between the PIPs inferred by SuShiE and their distance to the TSS,

we grouped fine-mapped SNPs into 2,000 bins that are 500 bp long to cover the one-million-bp window around the TSS for each gene and computed the average PIPs within each bin. To visualize the relationship between single effects' posterior probabilities and their distance to the TSS, we performed the same procedure focusing on the shared effects that had credible set output (i.e., passed the purity threshold; see previous method section).

We performed enrichment analysis using 89 functional annotations. First, we downloaded 5 candidate cis-regulatory elements (cCREs) from ENCODE Registry v3[58]. Then, we obtained 9 cell-type specific cCREs measured in PBMC using snATAC-Seq[59] and one cCRE measured in frozen PBMC using scATAC-seq[60]. Last, we obtained the 74 categorical functional annotations from LDSC baseline annotations v2.2[124,125], and remapped to GRCh38 using LiftOver (see **Code Availability**). To compute the functional enrichment scores, we employed an approach that is similar to TORUS[126]. Briefly, for each functional annotation and each gene, we performed the logistic regression $g(\boldsymbol{P}) = \boldsymbol{a}\omega$ where $g(\cdot)$ is the logit link function, $\boldsymbol{P}$ is the vector for the PIPs of all the SNPs, $\boldsymbol{a}$ is the binary vector indicating whether the SNPs fall into the annotation, and ω is the desired log-enrichment scores. After removing the genes on which logistic regression does not converge, we meta-analyzed the log-enrichment scores across genes by $\omega_{\text{meta}} = \frac{\sum \phi_i \omega_i}{\sum \phi_i}$ and $z_{\omega_{\text{meta}}} = \frac{\sum \phi_i \omega_i}{\sqrt{\sum \phi_i}}$ where ϕ_i is the inverse of the squared standard error for gene i. When comparing enrichment results across methods, we focused on e/pGenes fine-mapped by both methods. We computed the comparison z score as $\frac{\omega_{\text{meta},j} - \omega_{\text{meta},j'}}{\sqrt{se^2{}_{\omega_{\text{meta},j}} + se^2{}_{\omega_{\text{meta},j'}}}}$ for method j and j'. For the enrichment analyses focusing on individual shared effect using $\boldsymbol{\alpha}_l$, rather than PIPs, we limited analyses to those single effects that had corresponding credible sets (i.e., were not pruned).

To perform a case study, we selected *URGCP,* which was fine-mapped by SuShiE, but missed by other methods. To annotate the genomic region around *URGCP*, we downloaded the ChiP-Seq H3K27ac data of ENCODE[58] from WashU Epigenome Browser[127] (see **Code Availability**) and proximal enhancer (pELS) cCREs from ENCODE Registry v3, PBMC annotation using scATAC-seq in Satpathy et al.[60], naive T cells, naive B cells, cytotoxic natural killer (cNK) cells, and monocytes annotations using snATAC-seq in Chiou et al.[59]

Prior *cis*-molQTL correlation analyses

To shed light on the relationship between heterogeneity of effect-sizes across ancestries and genes' constraint, using all the credible sets output by SuShiE, we tested for association between SuShiE-inferred effect size correlations across ancestries ($\boldsymbol{\rho}_l$) and five measures of constraint ($\boldsymbol{s}$) using all the fine-mapped e/pGenes: probability of being Loss-of-Function Intolerant (pLI)[75], loss-of-function observed/expected upper bound fraction (LOEUF)[76], enhancer-domain score (EDS)[77], the Residual Variation Intolerance Score (RVIS)[78], and s_{het}[79]. We downloaded pLI and LOEUF from gnomAD browser v4.0 (see **Code Availability**), we downloaded EDS, RVIS, and s_{het} from their original papers. Our base model is according to:

$$E(\boldsymbol{s}) = \boldsymbol{\upsilon}_0 + \boldsymbol{\rho}_l \upsilon_1 + \boldsymbol{L}\upsilon_2 + \boldsymbol{d}\upsilon_3 + \boldsymbol{r}\upsilon_4$$

where $\boldsymbol{\upsilon}_0$ is the intercept term, $\boldsymbol{L}$ is the ordered and categorical single effect index representing the order of variance explained, $\boldsymbol{d}$ is the corresponding ancestry pair indicator (e.g., the correlation of EUR-AFR, EUR-HIS, or HIS-AFR), $\boldsymbol{r}$ is the study indicator (e.g., TOPMed MESA mRNA, TOPMed MESA proteins, or GENOA mRNA) , υ_is are the corresponding coefficients. We test the significance of $\upsilon_1$in a linear regression framework. A negative value of $\upsilon_1$for pLI, EDS, and s_{het} is taken to indicate stronger associations between *cis*-molQTL effect size heterogeneity across ancestries and gene constraint, while a lower value of LOEUF and RVIS is suggestive of stronger associations. In addition, to show robustness, we re-tested these associations using estimated covariance by replacing $\boldsymbol{\rho}_l$ by $\boldsymbol{\sigma}_b{}^2$. We also only focused on correlations estimated only from the primary effect (i.e., L=1); in this case, we removed $\boldsymbol{L}$ from the base model. We also re-computed the standard error using bootstrap. Specifically, for each study, each ancestry pair, and each L, we sampled the genes with replacement and computed the υ_1. We repeated 100 times to construct the null distributions for υ_1 and used its standard deviation as a new standard error. In addition, to adjust for allele frequency differences across ancestries, we added Wright's fixation index (F_{st}) as an additional term. To compute $F_{s,}$, we only used the fine-mapped SNPs to compute the F_{st} using PLINK2[113,114] with the "Hudson" method[128,129] for each gene. To investigate the relationship between expected *cis*-

668 molQTLs's distance to TSS and genes' constraint, we computed the expected distance to TSS for each gene
669 according to $\frac{\sum \mathrm{PIP}_i * D_i}{\sum \mathrm{PIP}_i}$ where D_i is the distance (absolute value) to the TSS for SNP i.

670 ## TWAS and PWAS analyses in All Of Us biobank

671 We performed individual-level Transcriptome- and Proteome-wide Association Studies (TWASs and PWASs)[42–44,47]
672 on 6 white blood cell-related traits: basophil count (BAS), eosinophil count (EOS), lymphocyte count (LYM),
673 monocyte count (MON), neutrophil count (NEU), and white blood cell count (WBC; **Table S9**) measured in AOU
674 biobank[50]. We excluded individuals who had acute abdomen, acute appendicitis, acute cholangitis, acute
675 cholecystitis, acute pancreatitis, anemia due to and following chemotherapy, bone marrow transplant present,
676 chemotherapy-induced nausea and vomiting, cirrhosis of liver, clostridium difficile colitis, complication of
677 chemotherapy, congenital anemia, congenital hemolytic anemia, convalescence after chemotherapy, dermatosis
678 resulting from cytotoxic therapy, diverticulitis of intestine, end-stage renal disease, fatigue due to chemotherapy,
679 hereditary hemolytic anemia, human immunodeficiency virus infection, leukemia, mucositis following
680 chemotherapy, myelodysplastic syndrome (clinical), neutropenia due to and following chemotherapy,
681 pancytopenia due to antineoplastic chemotherapy, peripheral neuropathy due to and following antineoplastic
682 therapy, post-splenectomy disorder and post-splenectomy thrombocytosis. For WBC, we only included
683 measurements <200e9/L. For all the traits, we also excluded measurements that were 3 standard deviations away
684 from the mean, resulting in a total of 86,345 individuals on average. We identified individual ancestry information
685 based on AOU precomputed information (i.e., "eur", "afr", and "amr" labels), resulting in 53,268 EUR, 16,748 AFR,
686 and 16,329 HIS individuals on average.
687 From our previous analysis, we obtained the eQTL prediction weights of EUR, AFR, and HIS in the TOPMed MESA
688 mRNA dataset, the pQTL prediction weights of EUR, AFR, and HIS in the TOPMed MESA protein dataset, and the
689 eQTL prediction weights of EUR and AFR in the GENOA mRNA dataset. We evaluated the prediction accuracy for
690 SuShiE SuShiE-Indep, Meta-SuSiE, SuSiE, LASSO[53], Elastic Net[54], and gBLUP[55] with five-fold cross-validation. For

691 Meta-SuSiE, we trained the prediction weights for each ancestry. For SuSiE, LASSO, Elastic Net, and gBLUP, we
692 trained the prediction weights after concatenating genotype and phenotype data across ancestries to ensure the
693 equal sample sizes as SuShiE (i.e., the same prediction weights for all ancestries). We computed cross validation
694 r^2 (cv-r^2) between the measured expression levels and predicted expression levels concatenated across each fold
695 and each ancestry. We also used the SuShiE-based ancestry-specific prediction weights to evaluate the prediction
696 performance using cross-ancestry weights. Specifically, we predicted the expression levels of EUR individuals using
697 AFR weights (of AFR individuals using HIS weights and of HIS individuals using EUR weights).
698 To perform T/PWAS, we first predicted expression levels (either mRNA or proteins) for EUR, AFR, and HIS
699 individuals in AOU using each ancestry-matched e/pQTL prediction weights with the score function in PLINK2[113,114].
700 Then, we standardized the expression vector (centered by mean and scaled by standard deviation) within each
701 ancestry and then concatenated them into a single vector across ancestries. Then, we regressed out sex, age,
702 squared age, and ten genotype PCs from the trait measurements. Last, we regressed the inverse-rank normalized
703 residuals on the predicted expression levels to compute the TWAS or PWAS statistics. We re-performed the
704 procedure using SuSiE-derived e/pQTL prediction weights as comparisons. We applied the Bonferroni correction
705 to adjust the reported P-values with n=23,000. To validate our TWAS results, we compared them to five
706 independent TWAS studies: Lu and Gopalan et al.[36], Kachuri et al.[31], Tapia et al.[82], Rowland et al.[84], and Wen et
707 al.[83] We released our *cis*-molQTL prediction weights to the public, which can be found at the Data Availability
708 section. To test the association between T/PWAS chi-square statistics and genes' constraint scores: pLI[75], LOEUF[76],
709 EDS[77], RVIS[78], and s_{het}[79], we used linear regression adjusted for phenotype and study and reported one-sided P
710 values. To compare significance of these associations between SuShiE and SuSiE, we computed the z score as
711 $\frac{\sum_{i=1}^{I} \mathrm{X}^2_{i,\mathrm{SuShiE}} - \sum_{i=1}^{I} \mathrm{X}^2_{i,\mathrm{SuSiE}}}{\sqrt{2*(2I)}}$ where $\mathrm{X}^2_{i,*}$ is the chi-square statistics for constraint score i. We classified genes into three
712 groups: Low, Middle, and High based on different scores, respectively. For pLI, we labeled genes with pLI >0.9 as
713 High, <0.1 as Low, and otherwise middle. For other scores, we labeled genes whose value is greater than 90%
714 quantile as High, smaller than 10% quantile as Low, and otherwise middle.

High-speed inference of SuShiE using JAX

We implemented SuShiE in an open-sourced command-line Python software *sushie*, which can read individual-level genotype data in three formats: PLINK1.9[113,114], bgen1.3[130], and vcf[116], together with phenotypic and covariates data in tab-separated-values format. We leveraged *Just In Time* compilation in JAX (see **Code Availability**) to facilitate high-speed inference on CPUs, GPUs, or TPUs. This technique allows users to process, in a scalable fashion, thousands of molecular phenotypes with the backgrounds of diverse ancestries specified by the user. Not only can *sushie* perform our method, but it can also perform single-ancestry SuSiE[15], effect size correlation estimation, *cis*-SNP heritability estimation, cross-validation for the *cis*-molQTL prediction weights, and contain the script to convert the *cis*-molQTL prediction results to FUSION format[42], thus can be used in TWAS framework. We also implemented basic QC on the input data. Users can also customize the *sushie* inference function according to their preferences. We have compiled comprehensive documentation about the software at https://mancusolab.github.io/sushie/.

Figures

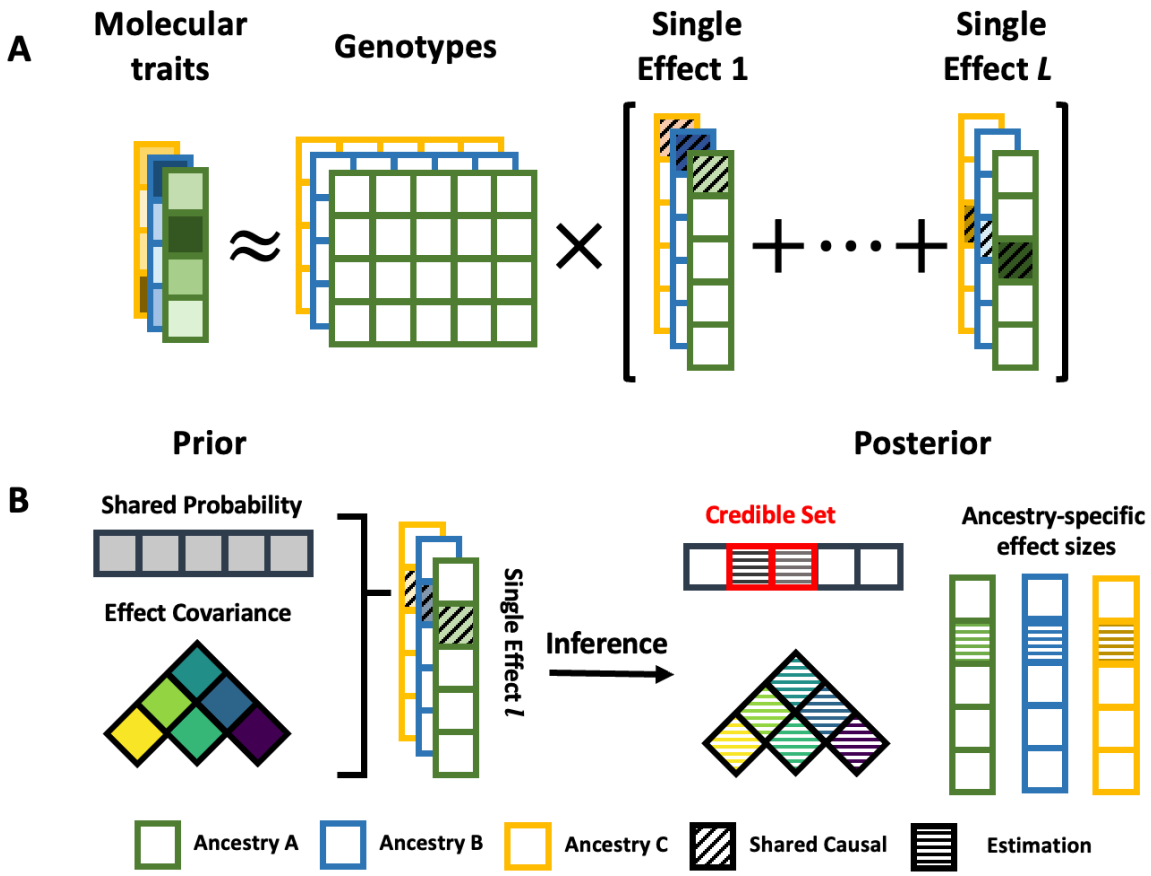

Fig. 1: SuShiE infers ancestry-specific effect sizes, PIPs, and credible sets by leveraging shared genetic architectures and LD heterogeneity.

A) SuShiE takes individual-level phenotypic and genotypic data as input and assumes the shared *cis*-molQTL effects as a linear combination of single effects.

B) For each single shared effect, SuShiE models the *cis*-molQTL effect size follows a multivariate normal prior distribution with a covariance matrix, and the probability for each SNP to be moQTL follows a uniform prior distribution; through the inference, SuShiE outputs a credible set that includes putative causal *cis*-molQTLs, learns the effect-size covariance prior, and estimates the ancestry-specific effect sizes.

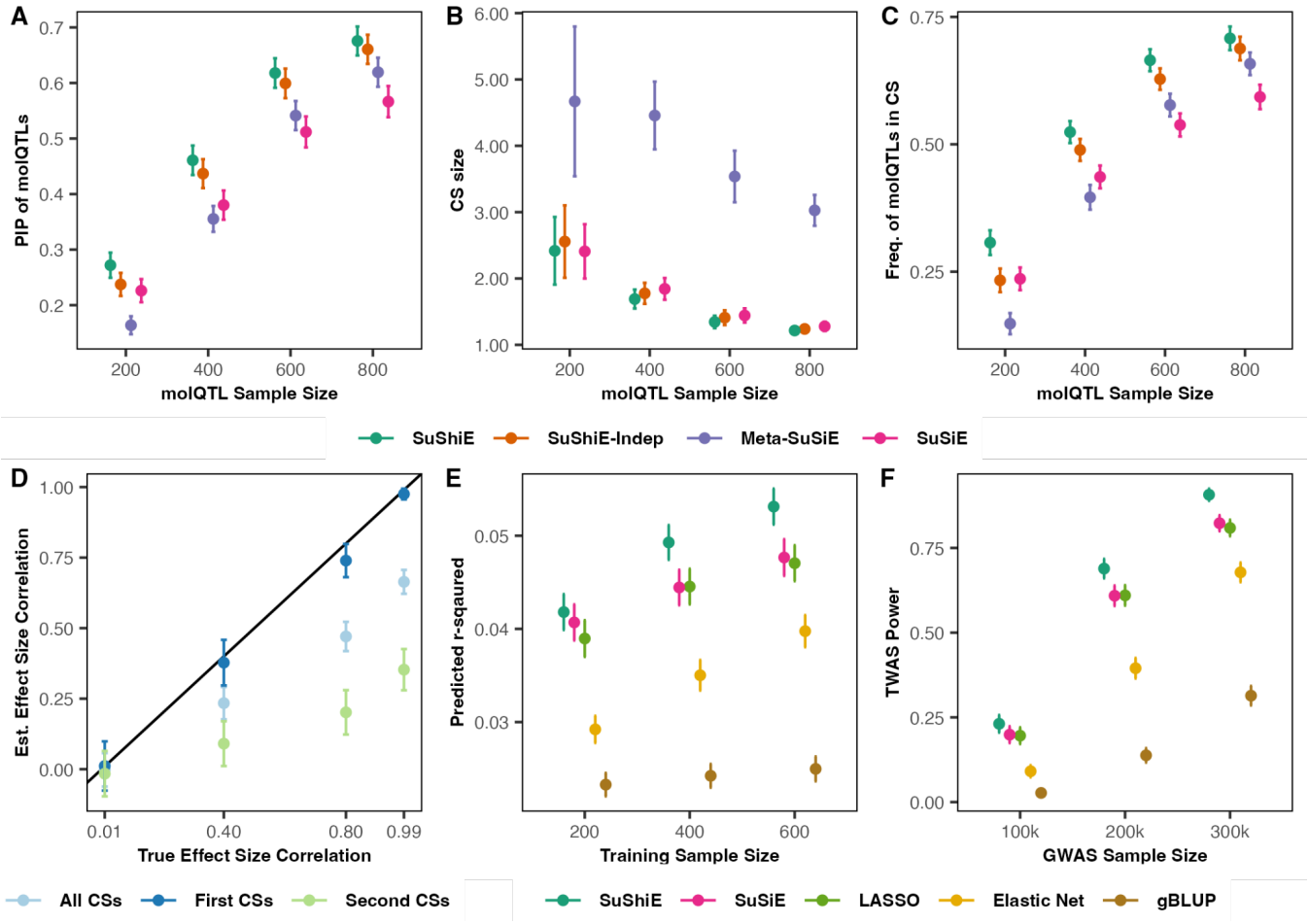

Fig. 2: SuShiE outperforms other methods, estimates accurate effect-size correlation, and boosts higher power of TWAS in realistic simulations

A-C) SuShiE outputs higher posterior inclusion probabilities (PIPs; A), smaller credible set sizes (B), and higher frequency of *cis*-molQTLs in the credible sets (calibration; C) compared to SuShiE-Indep (2.60e-4, 1.5e-1, and 1.30e-11), Meta-SuSiE (P=9.67e-43, 9.35e-231, and 1.17e-76), and SuSiE (P=6.98e-63, 6.65e-2, and 1.58e-104).
D) SuShiE accurately estimates the true effect-size correlation across ancestries using the primary effect (First credible sets; CSs) while exhibiting an underestimation using the secondary effects (Second CSs) or combined (All CSs) because the variance explained by the secondary effect decreases, thus requiring higher statistical power. The error bar is a 95% confidence interval.
E) SuShiE outputs higher ancestry-specific prediction accuracy compared against SuSiE, LASSO, Elastic Net, and gBLUP (all P<9.57e-8) with the fixed sample size. The plots are aggregation across two ancestries.
F) SuShiE induces higher TWAS power compared to SuSiE, LASSO, Elastic Net, and gBLUP (all P<4.34e-14) with the fixed sample size. The plots are aggregation across two ancestries.

By default, the simulation assumes that there are 2 causal *cis*-molQTLs, the per-ancestry training sample size is 400, and the testing sample size is 200, *cis*-SNP heritability is 0.05, the effect size correlation is 0.8 across ancestries, and the proportion of *cis*-SNP heritability of complex trait explained by gene expression is 1.5e-14. The error bar is a 95% confidence interval.

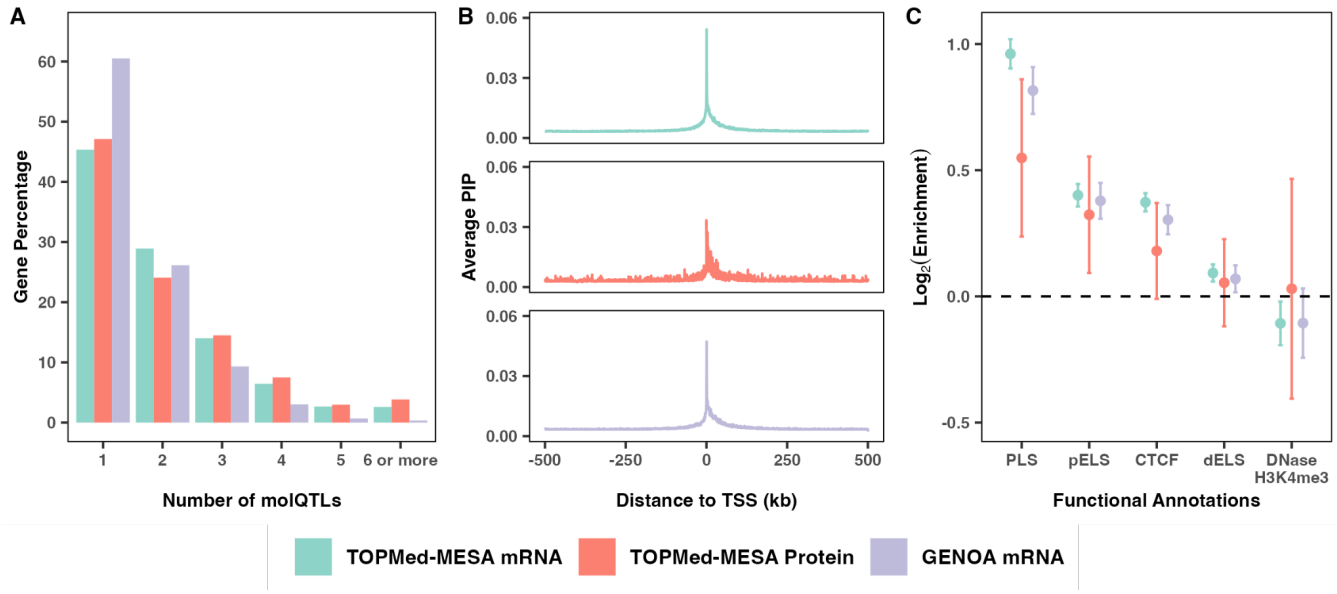

Fig. 3: SuShiE reveals cis-regulatory mechanisms for mRNA and protein expression

A) SuShiE identified *cis*-molQTLs for 14,590, 573, and 5,925 genes whose 88%, 86%, and 96% contain 1-3 *cis*-molQTLs for the TOPMed-MESA mRNA, TOPMed-MESA protein, and GENOA mRNA dataset, respectively.
B) Posterior inclusion probabilities (PIPs) of *cis*-molQTLs inferred by SuShiE are mainly enriched around the TSS region of genes. We grouped SNPs into 500-bp-long bins and computed their PIP average. There are 2,000 bins to cover a one-million-bp-long genomic window around the genes' TSS.
C) Across all three studies, *cis*-molQTLs identified by SuShiE are enriched in four out of five candidate *cis*-regulatory elements (cCREs) from ENCODE[58], with the promoter (PLS) as the most enriched category. Specifically, the mRNA expression from TOPMed-MESA and GENOA showed enrichment in the promoter, proximal enhancer (pELS), CTCF, and distal enhancer (dELS) but depletion in DNase-H3K4me3. Protein expression from TOPMed-MESA showed enrichment in PLS and pELS but non-significant enrichment in CTCF and dELS because of the low number of genes identified with pQTLs (n=573). The error bar is a 95% confidence interval.

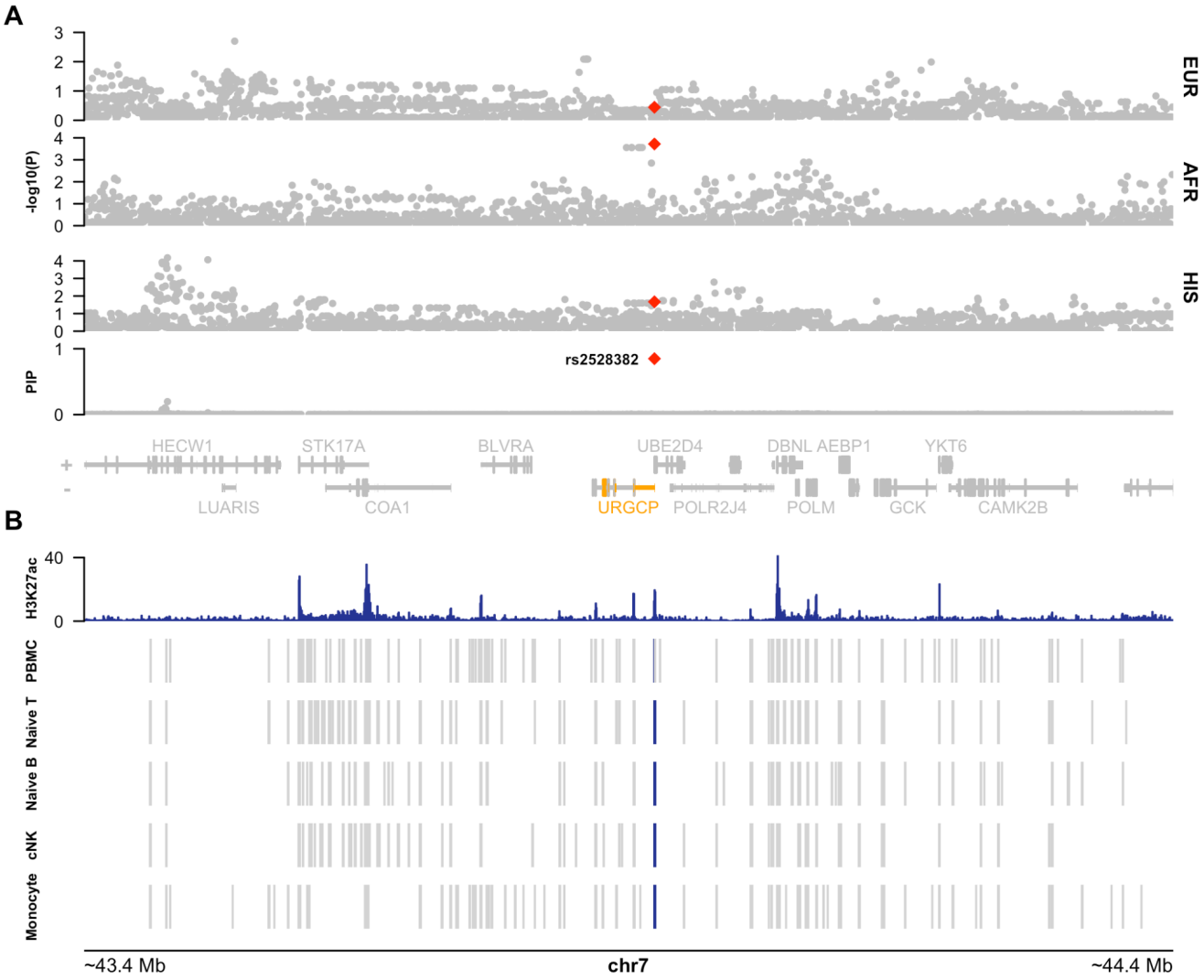

Fig. 4: SuShiE identifies eQTL *rs2528382* for *URGCP* with functional support

A) Manhattan plot of *cis*-eQTL scans of *URGCP* (denoted in orange) for each ancestry (above) with SuShiE fine-mapping results (below). SuShiE was the only method to output credible sets for *URGCP* and prioritized a single SNP (*rs2528382*; denoted in red).

B) Functional annotations at *URGCP* locus show colocalization of active enhancer activity and chromatin accessibility with *rs2528382*. H3K27ac CHIP-seq peaks measured in PBMCs (intensity denoted in blue) and 0/1 accessibility annotations determined from scATAC-seq measured in PBMCs and snATAC-seq measured in naive T cells, naive B cells, cytotoxic NK (cNK) cells, and monocytes. Blue rectangles denote a putative cCRE called from sc/snATAC-seq data that colocalize with *rs2528382* (gray no colocalization).

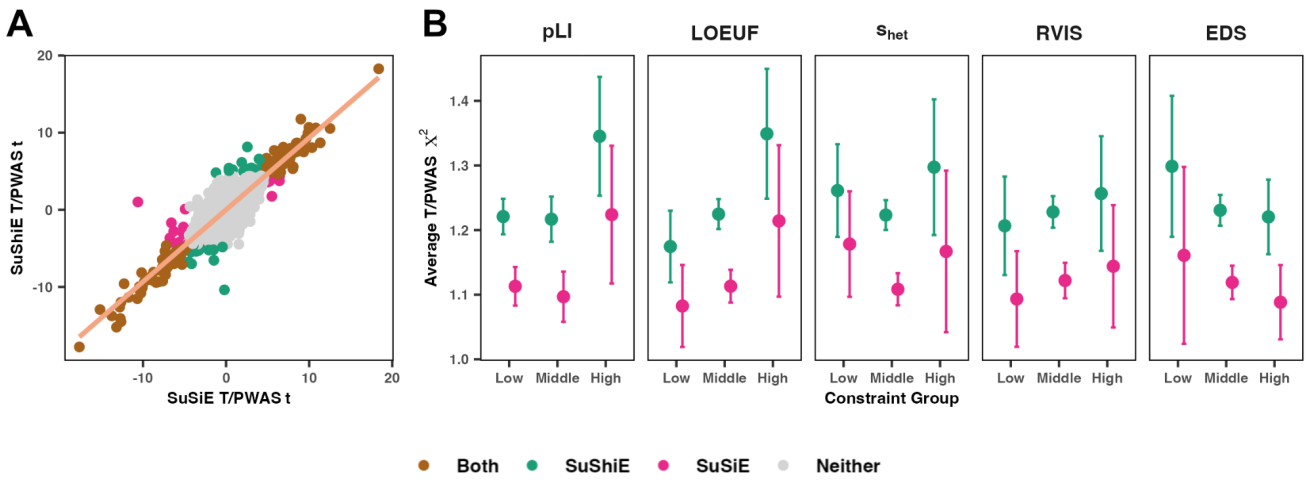

Fig. 5: SuShiE identifies more T/PWAS genes compared with SuSiE

A) Scatter plot of T/PWAS t-statistics between SuShiE (y-axis) and SuSiE (x-axis) across all phenotypes and contributing *cis*-molQTL studies.
B) Average T/PWAS chi-square statistics within low, middle, and high constraint scores (see **Methods**). Error bars represent 95% confidence intervals.

Tables

	pLI	LOEUF	S_{het}	RVIS	EDS
Base Model	-0.022 (4.13e-33)	0.021 (5.92e-20)	-0.007 (4.15e-40)	0.043 (2.04e-14)	-0.002 (1.25e-02)
Bootstrap SE	-0.022 (5.84e-32)	0.021 (4.92e-20)	-0.007 (3.13e-37)	0.043 (1.68e-17)	-0.002 (1.56e-02)
Primary Effect	-0.034 (3.51e-23)	0.027 (7.45e-11)	-0.011 (7.69e-29)	0.055 (4.09e-09)	-0.004 (1.27e-03)
Effect Covariance	-0.339 (7.59e-177)	0.334 (1.77e-109)	-0.089 (1.33e-154)	0.537 (9.93e-49)	-0.053 (3.10e-25)
Adjusted F_{st}	-0.022 (2.00e-32)	0.021 (9.90e-20)	-0.007 (2.22e-39)	0.042 (5.63e-14)	-0.002 (1.08e-02)

Table 1: Across-ancestry *cis*-molQTL effect size correlations are negatively associated with gene constraint scores

The estimates and corresponding P-value in the regression framework testing associations between inferred effect size correlations across ancestries and constraint scores (see **Methods** for the base model). "Bootstrap SE" is to re-estimate standard error using bootstrap. "Primary Effect" is to only use estimates from L=1. "Effect Covariance" is to replace estimated correlation with estimated effect size covariance across ancestries. "Adjusted F_{st}" is to additionally adjusted for F_{st} from the base model. A higher value of pLI, s_{het}, and, EDS is taken to indicate stronger constraint, while a lower value of LOEUF and RVIS is suggestive of more constraint. The reported P-value is one-sided.

Data availability

SuShiE-derived prediction models for TWAS/PWAS, fine-mapping, and other analyzed results across *cis*-molQTL datasets can be found at https://zenodo.org/records/10963034.

Code availability

SuShiE: https://github.com/mancusolab/sushie

The analysis codes for simulation and real-data analysis of this manuscript: https://github.com/mancusolab/sushie-project-codes

TOPMed RNA-seq Harmonization pipeline: https://github.com/broadinstitute/gtex-pipeline/blob/master/TOPMed_RNAseq_pipeline.md

gnomAD v4.0: https://gnomad.broadinstitute.org/news/2023-11-gnomad-v4-0/

GTEx eQTL analysis pipeline: https://www.gtexportal.org/home/methods

pyqtl software: https://github.com/broadinstitute/pyqtl

PLINK: https://www.cog-genomics.org/plink/2.0

BCFTOOLS: https://samtools.github.io/bcftools/bcftools.html

JAX: https://github.com/google/jax

scikit-learn: https://scikit-learn.org/stable/

FUSION: http://gusevlab.org/projects/fusion/

limix: https://github.com/limix/limix

LiftOver: https://genome.ucsc.edu/cgi-bin/hgLiftOver

WashU Epigenome Browser: https://epigenomegateway.wustl.edu/

Acknowledgements

The authors would like to thank members of the Mancuso and Gazal labs for fruitful discussions regarding this manuscript. The authors would also like to specially thank Dr. Michael D. Edge for his thoughtful comments and suggestions. This work was funded in part by National Institutes of Health (NIH) under awards R01HG012133, R01CA258808, R01GM140287, R35GM142783, R01GM140287, U54HG013243, R35GM147789, and K08HL159346.

MESA phenotypes (dbGaP: phs000209.v13.p3): MESA and the MESA SHARe project are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts HHSN268201500003I, N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-001079, UL1-TR000040, UL1-TR-001420, UL1-TR-001881, and DK063491. Funding for SHARe genotyping was provided by NHLBI Contract N02-HL-64278. TOPMed MESA WGS genotype, mRNA, and protein expression data (dbGaP: phs001416.v3.p1): Molecular data for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). WGS genotype data for NHLBI TOPMed: MESA (phs001416.v3.p1) was performed at Broad Genomics (HHSN268201600034I). mRNA expression data for NHLBI TOPMed: MESA (phs001416.v3.p1) was performed at NWGC (HHSN268201600032I). SOMAscan proteomics for NHLBI TOPMed: Multi-Ethnic Study of Atherosclerosis (MESA) (phs001416.v1.p1) was performed at the Broad Institute and Beth Israel Proteomics Platform (HHSN268201600034I). Core support including centralized genomic read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Core support including phenotype harmonization, data management, sample-identity QC, and general program coordination were provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed.

GENOA genotype (dbGaP: phs001238.v2.p1) and gene expression (GEO: GSE138914) data were supported by grants from NIH NHLBI (HL054457, HL054464, HL054481, HL119443, and HL087660). The authors would like to acknowledge Drs. Sharon Kardia and Jennifer Smith in preparing GENOA eQTL data.

The All of Us Research Program is supported by the National Institutes of Health, Office of the Director: Regional Medical Centers: 1 OT2 OD026549; 1 OT2 OD026554; 1 OT2 OD026557; 1 OT2 OD026556; 1 OT2 OD026550; 1 OT2 OD 026552; 1 OT2 OD026553; 1 OT2 OD026548; 1 OT2 OD026551; 1 OT2 OD026555; IAA #: AOD 16037; Federally Qualified Health Centers: HHSN 263201600085U; Data and Research Center: 5 U2C OD023196; Biobank: 1 U24 OD023121; The Participant Center: U24 OD023176; Participant Technology Systems Center: 1 U24 OD023163; Communications and Engagement: 3 OT2 OD023205; 3 OT2 OD023206; and Community Partners: 1 OT2 OD025277; 3 OT2 OD025315; 1 OT2 OD025337; 1 OT2 OD025276. In addition, the All of Us Research Program would not be possible without the partnership of its participants.

Author contributions

Z.L. and N.M. developed the model and study design. Z.L. performed simulations and fine-mapping analyses. Z.L., X.W., J.P., and L.K. performed TWAS and AoU analyses. Z.L., M.C., and N.M. developed the model and inference scheme. Z.L. and A.K. prepared functional genomic annotations and enrichment analyses. Z.L. and N.M. wrote the initial manuscript. All authors edited the final manuscript.

Competing interests

L.W. provided consulting service to Pupil Bio Inc. and reviewed manuscripts for Gastroenterology Report, not related to this study, and received honorarium. No potential conflicts of interest were disclosed by the other authors.

References

1. Claussnitzer, M. *et al.* A brief history of human disease genetics. *Nature* **577**, 179–189 (2020).
2. Cheung, V. G. *et al.* Mapping determinants of human gene expression by regional and genome-wide association. *Nature* **437**, 1365–1369 (2005).
3. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
4. Võsa, U. *et al.* Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat. Genet.* **53**, 1300–1310 (2021).
5. Sun, B. B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).
6. Gate, R. E. *et al.* Genetic determinants of co-accessible chromatin regions in activated T cells across humans. *Nat. Genet.* **50**, 1140–1150 (2018).
7. Battle, A. *et al.* Genomic variation. Impact of regulatory variation from RNA to protein. *Science* **347**, 664–667 (2015).
8. Li, Y. I. *et al.* RNA splicing is a primary link between genetic variation and disease. *Science* **352**, 600–604 (2016).
9. McVicker, G. *et al.* Identification of genetic variants that affect histone modifications in human cells. *Science* **342**, 747–749 (2013).
10. Oliva, M. *et al.* DNA methylation QTL mapping across diverse human tissues provides molecular links between genetic variation and complex traits. *Nat. Genet.* **55**, 112–122 (2023).
11. Wu, L. *et al.* Variation and genetic control of protein abundance in humans. *Nature* **499**, 79–82 (2013).
12. Aguet, F. *et al.* Molecular quantitative trait loci. *Nat. Rev. Methods Primers* **3**, 1–22 (2023).
13. Albert, F. W. & Kruglyak, L. The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.* **16**, 197–212 (2015).
14. Suhre, K., McCarthy, M. I. & Schwenk, J. M. Genetics meets proteomics: perspectives for large population-based studies. *Nat. Rev. Genet.* **22**, 19–37 (2021).
15. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Series B Stat. Methodol.* **82**, 1273–1300 (2020).
16. Hormozdiari, F., Kichaev, G., Yang, W.-Y., Pasaniuc, B. & Eskin, E. Identification of causal genes for complex traits. *Bioinformatics* **31**, i206-13 (2015).
17. Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).
18. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* **19**, 491–504 (2018).
19. Kichaev, G. *et al.* Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet.* **10**, e1004722 (2014).
20. Wojcik, G. L. *et al.* Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**, 514–518 (2019).
21. Chen, M.-H. *et al.* Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations. *Cell* **182**, 1198-1213.e14 (2020).
22. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
23. Sakaue, S. *et al.* A cross-population atlas of genetic associations for 220 human phenotypes. *Nat. Genet.* **53**, 1415–1424 (2021).
24. Conti, D. V. *et al.* Trans-ancestry genome-wide association meta-analysis of prostate cancer identifies new susceptibility loci and informs genetic risk prediction. *Nat. Genet.* **53**, 65–75 (2021).
25. Wang, A. *et al.* Characterizing prostate cancer risk through multi-ancestry genome-wide discovery of 187

914 novel risk variants. *Nat. Genet.* (2023) doi:10.1038/s41588-023-01534-4.
915 26. Shang, L. *et al.* Genetic Architecture of Gene Expression in European and African Americans: An eQTL
916 Mapping Study in GENOA. *Am. J. Hum. Genet.* **106**, 496–512 (2020).
917 27. Mogil, L. S. *et al.* Genetic architecture of gene expression traits across diverse populations. *PLoS Genet.* **14**,
918 e1007586 (2018).
919 28. Schubert, R. *et al.* Protein prediction for trait mapping in diverse populations. *PLoS One* **17**, e0264341
920 (2022).
921 29. Tehranchi, A. *et al.* Fine-mapping cis-regulatory variants in diverse human populations. *Elife* **8**, (2019).
922 30. Wen, X., Luca, F. & Pique-Regi, R. Cross-population joint analysis of eQTLs: fine mapping and functional
923 annotation. *PLoS Genet.* **11**, e1005176 (2015).
924 31. Kachuri, L. *et al.* Gene expression in African Americans, Puerto Ricans and Mexican Americans reveals
925 ancestry-specific patterns of genetic architecture. *Nat. Genet.* **55**, 952–963 (2023).
926 32. Kasela, S. *et al.* Interaction molecular QTL mapping discovers cellular and environmental modifiers of
927 genetic regulatory effects. *Am. J. Hum. Genet.* **111**, 133–149 (2024).
928 33. Kichaev, G. & Pasaniuc, B. Leveraging Functional-Annotation Data in Trans-ethnic Fine-Mapping Studies.
929 *Am. J. Hum. Genet.* **97**, 260–271 (2015).
930 34. Asimit, J. L., Hatzikotoulas, K., McCarthy, M., Morris, A. P. & Zeggini, E. Trans-ethnic study design
931 approaches for fine-mapping. *Eur. J. Hum. Genet.* **24**, 1330–1336 (2016).
932 35. LaPierre, N. *et al.* Identifying causal variants by fine mapping across multiple studies. *PLoS Genet.* **17**,
933 e1009733 (2021).
934 36. Lu, Z. *et al.* Multi-ancestry fine-mapping improves precision to identify causal genes in transcriptome-wide
935 association studies. *Am. J. Hum. Genet.* **109**, 1388–1404 (2022).
936 37. Shen, J. *et al.* Fine-mapping and credible set construction using a multi-population Joint Analysis of
937 Marginal summary statistics from Genome-wide Association Studies. *bioRxiv* (2022)
938 doi:10.1101/2022.12.22.521659.
939 38. Yuan, K. *et al.* Fine-mapping across diverse ancestries drives the discovery of putative causal variants
940 underlying human complex traits and diseases. *medRxiv* (2023) doi:10.1101/2023.01.07.23284293.
941 39. Cai, M. *et al.* XMAP: Cross-population fine-mapping by leveraging genetic diversity and accounting for
942 confounding bias. *Nat. Commun.* **14**, 6870 (2023).
943 40. Zhou, F. *et al.* Leveraging information between multiple population groups and traits improves fine-
944 mapping resolution. *Nat. Commun.* **14**, 7279 (2023).
945 41. Gao, B. & Zhou, X. MESuSiE enables scalable and powerful multi-ancestry fine-mapping of causal variants in
946 genome-wide association studies. *Nat. Genet.* (2024) doi:10.1038/s41588-023-01604-7.
947 42. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.*
948 **48**, 245–252 (2016).
949 43. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome
950 data. *Nat. Genet.* **47**, 1091–1098 (2015).
951 44. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene
952 targets. *Nat. Genet.* **48**, 481–487 (2016).
953 45. Gusev, A. *et al.* Transcriptome-wide association study of schizophrenia and chromatin activity yields
954 mechanistic disease insights. *Nat. Genet.* **50**, 538–548 (2018).
955 46. Mancuso, N. *et al.* Large-scale transcriptome-wide association study identifies new prostate cancer risk
956 regions. *Nat. Commun.* **9**, 4079 (2018).
957 47. Zhang, J. *et al.* Plasma proteome analyses in individuals of European and African ancestry identify cis-pQTLs
958 and models for proteome-wide association studies. *Nat. Genet.* **54**, 593–602 (2022).
959 48. Bild, D. E. *et al.* Ethnic differences in coronary calcification: the Multi-Ethnic Study of Atherosclerosis
960 (MESA). *Circulation* **111**, 1313–1320 (2005).
961 49. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**,

290–299 (2021).
50. All of Us Research Program Genomics Investigators. Genomic data in the All of Us Research Program. *Nature* (2024) doi:10.1038/s41586-023-06957-x.
51. Zou, Y., Carbonetto, P., Wang, G. & Stephens, M. Fine-mapping from summary data with the "Sum of Single Effects" model. *PLoS Genet.* **18**, e1010299 (2022).
52. Zou, Y., Carbonetto, P., Xie, D., Wang, G. & Stephens, M. Fast and flexible joint fine-mapping of multiple traits via the Sum of Single Effects model. *bioRxiv* (2023) doi:10.1101/2023.04.14.536893.
53. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Series B Stat. Methodol.* **58**, 267–288 (1996).
54. Zou, H. & Hastie, T. Regularization and Variable Selection Via the Elastic Net. *J. R. Stat. Soc. Series B Stat. Methodol.* **67**, 301–320 (2005).
55. Clark, S. A. & van der Werf, J. Genomic best linear unbiased prediction (gBLUP) for the estimation of genomic breeding values. *Methods Mol. Biol.* **1019**, 321–330 (2013).
56. Mostafavi, H., Spence, J. P., Naqvi, S. & Pritchard, J. K. Systematic differences in discovery of genetic effects on gene expression and complex traits. *Nat. Genet.* **55**, 1866–1875 (2023).
57. Sun, B. B. *et al.* Plasma proteomic associations with genetics and health in the UK Biobank. *Nature* **622**, 329–338 (2023).
58. ENCODE Project Consortium *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
59. Chiou, J. *et al.* Interpreting type 1 diabetes risk with genetics and single-cell epigenomics. *Nature* **594**, 398–402 (2021).
60. Satpathy, A. T. *et al.* Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat. Biotechnol.* **37**, 925–936 (2019).
61. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
62. Tufan, N. L. S. *et al.* Hepatitis Bx antigen stimulates expression of a novel cellular gene, URG4, that promotes hepatocellular growth and survival. *Neoplasia* **4**, 355–368 (2002).
63. Song, J. *et al.* Enhanced cell survival of gastric cancer cells by a novel gene URG4. *Neoplasia* **8**, 995–1002 (2006).
64. Li, W. & Zhou, N. URG4 upregulation is associated with tumor growth and poor survival in epithelial ovarian cancer. *Arch. Gynecol. Obstet.* **286**, 209–215 (2012).
65. Xie, C. *et al.* Upregulator of cell proliferation predicts poor prognosis in hepatocellular carcinoma and contributes to hepatocarcinogenesis by downregulating FOXO3a. *PLoS One* **7**, e40607 (2012).
66. Cai, J. *et al.* URGCP promotes non-small cell lung cancer invasiveness by activating the NF-κB-MMP-9 pathway. *Oncotarget* **6**, 36489–36504 (2015).
67. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **6**, 80–92 (2012).
68. Shi, H., Mancuso, N., Spendlove, S. & Pasaniuc, B. Local Genetic Correlation Gives Insights into the Shared Genetic Architecture of Complex Traits. *Am. J. Hum. Genet.* **101**, 737–751 (2017).
69. Shi, H. *et al.* Population-specific causal disease effect sizes in functionally important regions impacted by selection. *Nat. Commun.* **12**, 1098 (2021).
70. Saitou, M., Dahl, A., Wang, Q. & Liu, X. Allele frequency differences of causal variants have a major impact on low cross-ancestry portability of PRS. *bioRxiv* (2022) doi:10.1101/2022.10.21.22281371.
71. Taylor, D. J. *et al.* Sources of gene expression variation in a globally diverse human cohort. *bioRxiv* (2023) doi:10.1101/2023.11.04.565639.
72. Brown, B. C., Asian Genetic Epidemiology Network Type 2 Diabetes Consortium, Ye, C. J., Price, A. L. & Zaitlen, N. Transethnic Genetic-Correlation Estimates from Summary Statistics. *Am. J. Hum. Genet.* **99**, 76–88 (2016).

73. Hou, K. *et al.* Causal effects on complex traits are similar for common variants across segments of different continental ancestries within admixed individuals. *Nat. Genet.* **55**, 549–558 (2023).
74. Shi, H. *et al.* Localizing Components of Shared Transethnic Genetic Architecture of Complex Traits from GWAS Summary Data. *Am. J. Hum. Genet.* **106**, 805–817 (2020).
75. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
76. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
77. Wang, X. & Goldstein, D. B. Enhancer Domains Predict Gene Pathogenicity and Inform Gene Discovery in Complex Disease. *Am. J. Hum. Genet.* **106**, 215–233 (2020).
78. Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet.* **9**, e1003709 (2013).
79. Zeng, T., Spence, J. P., Mostafavi, H. & Pritchard, J. K. Bayesian estimation of gene constraint from an evolutionary model with gene features. *bioRxiv* (2023) doi:10.1101/2023.05.19.541520.
80. Berg, J. J. *et al.* Reduced signal for polygenic adaptation of height in UK Biobank. *Elife* **8**, (2019).
81. Keys, K. L. *et al.* On the cross-population generalizability of gene expression prediction models. *PLoS Genet.* **16**, e1008927 (2020).
82. Tapia, A. L. *et al.* A large-scale transcriptome-wide association study (TWAS) of 10 blood cell phenotypes reveals complexities of TWAS fine-mapping. *Genet. Epidemiol.* **46**, 3–16 (2022).
83. Wen, J. *et al.* Transcriptome-Wide Association Study of Blood Cell Traits in African Ancestry and Hispanic/Latino Populations. *Genes* **12**, (2021).
84. Rowland, B. *et al.* Transcriptome-wide association study in UK Biobank Europeans identifies associations with blood cell traits. *Hum. Mol. Genet.* **31**, 2333–2347 (2022).
85. Ding, Y. *et al.* Polygenic scoring accuracy varies across the genetic ancestry continuum. *Nature* **618**, 774–781 (2023).
86. Mester, R. *et al.* Impact of cross-ancestry genetic architecture on GWASs in admixed populations. *Am. J. Hum. Genet.* **110**, 927–939 (2023).
87. Hou, K., Bhattacharya, A., Mester, R., Burch, K. S. & Pasaniuc, B. On powerful GWAS in admixed populations. *Nature genetics* vol. 53 1631–1633 (2021).
88. Zhong, Y., Perera, M. A. & Gamazon, E. R. On Using Local Ancestry to Characterize the Genetic Architecture of Human Traits: Genetic Regulation of Gene Expression in Multiethnic or Admixed Populations. *Am. J. Hum. Genet.* **104**, 1097–1115 (2019).
89. Zhang, J. & Stram, D. O. The role of local ancestry adjustment in association studies using admixed populations. *Genet. Epidemiol.* **38**, 502–515 (2014).
90. Pasaniuc, B. *et al.* Enhanced statistical tests for GWAS in admixed populations: assessment using African Americans from CARe and a Breast Cancer Consortium. *PLoS Genet.* **7**, e1001371 (2011).
91. Seldin, M. F., Pasaniuc, B. & Price, A. L. New approaches to disease mapping in admixed populations. *Nat. Rev. Genet.* **12**, 523–528 (2011).
92. Atkinson, E. G. *et al.* Tractor uses local ancestry to enable the inclusion of admixed individuals in GWAS and to boost power. *Nat. Genet.* **53**, 195–204 (2021).
93. Qin, H. *et al.* Interrogating local population structure for fine mapping in genome-wide association studies. *Bioinformatics* **26**, 2961–2968 (2010).
94. Aracena, K. A. *et al.* Epigenetic variation impacts individual differences in the transcriptional response to influenza infection. *Nat. Genet.* **56**, 408–419 (2024).
95. Randolph, H. E. *et al.* Genetic ancestry effects on the response to viral infection are pervasive but cell type specific. *Science* **374**, 1127–1133 (2021).
96. Robinson, M. R. *et al.* Genotype-covariate interaction effects and the heritability of adult body mass index. *Nat. Genet.* **49**, 1174–1181 (2017).
97. Durvasula, A. & Lohmueller, K. E. Negative selection on complex traits limits phenotype prediction accuracy

between populations. *Am. J. Hum. Genet.* **108**, 620–631 (2021).
98. Yair, S. & Coop, G. Population differentiation of polygenic score predictions under stabilizing selection. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **377**, 20200416 (2022).
99. Agarwal, I., Fuller, Z. L., Myers, S. R. & Przeworski, M. Relating pathogenic loss-of-function mutations in humans to their evolutionary fitness costs. *Elife* **12**, (2023).
100. Speidel, L., Forest, M., Shi, S. & Myers, S. R. A method for genome-wide genealogy estimation for thousands of samples. *Nat. Genet.* **51**, 1321–1329 (2019).
101. Wang, J. & Gazal, S. Ancestry-specific regulatory and disease architectures are likely due to cell-type-specific gene-by-environment interactions. *medRxiv* (2023) doi:10.1101/2023.10.20.23297214.
102. Yazar, S. *et al.* Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. *Science* **376**, eabf3041 (2022).
103. Bhattacharya, A. *et al.* Best practices for multi-ancestry, meta-analytic transcriptome-wide association studies: Lessons from the Global Biobank Meta-analysis Initiative. *Cell Genom* **2**, (2022).
104. Selewa, A. *et al.* Single-cell genomics improves the discovery of risk variants and genes of atrial fibrillation. *Nat. Commun.* **14**, 4999 (2023).
105. Blei, D. M., Kucukelbir, A. & McAuliffe, J. D. Variational inference: A review for statisticians. *J. Am. Stat. Assoc.* **112**, 859–877 (2017).
106. Mancuso, N. *et al.* Probabilistic fine-mapping of transcriptome-wide association studies. *Nat. Genet.* **51**, 675–682 (2019).
107. Wang, X., Lu, Z., Bhattacharya, A., Pasaniuc, B. & Mancuso, N. twas_sim, a Python-based tool for simulation and power analysis of transcriptome-wide association analysis. *Bioinformatics* (2023) doi:10.1093/bioinformatics/btad288.
108. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
109. International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
110. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).
111. Yao, D. W., O'Connor, L. J., Price, A. L. & Gusev, A. Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nat. Genet.* **52**, 626–633 (2020).
112. Frankish, A. *et al.* GENCODE 2021. *Nucleic Acids Res.* **49**, D916–D923 (2021).
113. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics* vol. 81 559–575 Preprint at https://doi.org/10.1086/519795 (2007).
114. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* vol. 4 Preprint at https://doi.org/10.1186/s13742-015-0047-8 (2015).
115. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* vol. 25 2078–2079 Preprint at https://doi.org/10.1093/bioinformatics/btp352 (2009).
116. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* vol. 27 2156–2158 Preprint at https://doi.org/10.1093/bioinformatics/btr330 (2011).
117. Price, A. L. *et al.* Long-range LD can confound genome scans in admixed populations. *American journal of human genetics* vol. 83 132–5; author reply 135-9 (2008).
118. Buitinck, L. *et al.* API design for machine learning software: experiences from the scikit-learn project. *arXiv [cs.LG]* (2013).
119. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).
120. Gold, L. *et al.* Aptamer-based multiplexed proteomic technology for biomarker discovery. *PLoS One* **5**, e15004 (2010).
121. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. Preprint at

https://doi.org/10.1101/052308.

122. Di Angelantonio, E. *et al.* Efficiency and safety of varying the frequency of whole blood donation (INTERVAL): a randomised trial of 45 000 donors. *Lancet* **390**, 2360–2371 (2017).
123. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
124. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics* vol. 47 1228–1235 Preprint at https://doi.org/10.1038/ng.3404 (2015).
125. Hujoel, M. L. A., Gazal, S., Hormozdiari, F., van de Geijn, B. & Price, A. L. Disease Heritability Enrichment of Regulatory Elements Is Concentrated in Elements with Ancient Sequence Age and Conserved Function across Species. *Am. J. Hum. Genet.* **104**, 611–624 (2019).
126. Wen, X. MOLECULAR QTL DISCOVERY INCORPORATING GENOMIC ANNOTATIONS USING BAYESIAN FALSE DISCOVERY RATE CONTROL. *Ann. Appl. Stat.* **10**, 1619–1638 (2016).
127. Li, D. *et al.* WashU Epigenome Browser update 2022. *Nucleic Acids Res.* **50**, W774–W781 (2022).
128. Hudson, R. R., Slatkin, M. & Maddison, W. P. Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**, 583–589 (1992).
129. Bhatia, G., Patterson, N., Sankararaman, S. & Price, A. L. Estimating and interpreting FST: the impact of rare variants. *Genome Res.* **23**, 1514–1521 (2013).
130. Band, G. & Marchini, J. BGEN: a binary file format for imputed genotype and haplotype data. *bioRxiv* 308296 (2018) doi:10.1101/308296.