Genes with differential expression across ancestries are enriched in ancestry-specific disease effects likely due to gene-by-environment interactions

Authors

Juehan Wang, Zixuan Zhang, Zeyun Lu, Nicholas Mancuso, Steven Gazal

Correspondence

juehanwa@usc.edu (J.W.), gazal@usc.edu (S.G.)

We identified genes differentially expressed across two ancestry groups in seven immune cell types. These genes display cell-type specificity as well as enrichment in environmental interactions and variants with ancestry-specific disease effect sizes. These results suggest the impact of cell-type-specific, gene-by-environment interactions shared between regulatory and disease architectures.

Wang et al., 2024, The American Journal of Human Genetics *111*, 2117–2128 October 3, 2024

https://doi.org/10.1016/j.ajhg.2024.07.021

CellPress

Genes with differential expression across ancestries are enriched in ancestry-specific disease effects likely due to gene-by-environment interactions

Juehan Wang,[1,2,*] Zixuan Zhang,[1,2] Zeyun Lu,[1,2] Nicholas Mancuso,[1,2,3] and Steven Gazal[1,2,3,*]

Summary

Multi-ancestry genome-wide association studies (GWASs) have highlighted the existence of variants with ancestry-specific effect sizes. Understanding where and why these ancestry-specific effects occur is fundamental to understanding the genetic basis of human diseases and complex traits. Here, we characterized genes differentially expressed across ancestries (ancDE genes) at the cell-type level by leveraging single-cell RNA-sequencing data in peripheral blood mononuclear cells for 21 individuals with East Asian (EAS) ancestry and 23 individuals with European (EUR) ancestry (172,385 cells); then, we tested whether variants surrounding those genes were enriched in disease variants with ancestry-specific effect sizes by leveraging ancestry-matched GWASs of 31 diseases and complex traits (average $n \sim 90{,}000$ and $\sim 267{,}000$ in EAS and EUR, respectively). We observed that ancDE genes tended to be cell-type specific and enriched in genes interacting with the environment and in variants with ancestry-specific disease effect sizes, which suggests cell-type-specific, gene-by-environment interactions shared between regulatory and disease architectures. Finally, we illustrated how different environments might have led to ancestry-specific myeloid cell leukemia 1 (*MCL1*) expression in B cells and ancestry-specific allele effect sizes in lymphocyte count GWASs for variants surrounding *MCL1*. Our results imply that large single-cell and GWAS datasets from diverse ancestries are required to improve our understanding of human diseases.

Introduction

Multi-ancestry genome-wide association studies (GWASs) have highlighted that, despite the strong correlation of causal effect sizes across ancestries,[1–9] a non-negligible fraction of causal variants have ancestry-specific effect sizes likely due to gene-by-environment (GxE) interactions.[7–9] Knowing where and why ancestry-specific effects of disease risk variants occur is fundamental for understanding the genetic basis of human diseases and for improving the portability of polygenic risk scores across ancestries.[6]

Differences in gene regulation across ancestries have been observed at different levels (e.g., gene expression,[10–18] expression quantitative trait loci [eQTL] effect sizes,[19–21] methylation,[22–24] and enhancer activity[25]) and could inform which variants have ancestry-specific disease effect sizes. Indeed, gene regulation differences can also be due to GxE (e.g., ancestry-specific eQTL effect sizes), and variants with ancestry-specific disease effect sizes tend to be enriched in regulatory regions and around genes differentially expressed in tissues interacting with the environment.[7] However, investigating whether ancestry-specific regulatory and disease architectures are related (because of shared GxE effects) has been challenging for multiple reasons. First, there is a limited availability of ancestry-matched GWASs and functional datasets from non-European descent. Second, the regulatory differences between ancestries can be explained by multiple factors that are challenging to dissociate. Specifically, ancestry-specific levels of gene expression could be attributed to differences in allele frequencies of the gene's eQTL due to genetic drift or selection (G), different transcriptomic answers to the ancestry group's environment (E), or different effect sizes of the eQTL due to different environments (GxE), or they could be false positives based on batch effects related to how multi-ancestry data have been collected.[12] Finally, although gene regulation is cell-type specific,[26,27] the cell types that are the most subject to ancestry-specific gene regulation and disease effect sizes are unknown.

Here, we aimed to characterize genes differentially expressed across ancestries (ancDE genes) at the cell-type level and to test whether ancDE genes are enriched in disease variants with ancestry-specific effect sizes. We leveraged single-cell RNA-sequencing (scRNA-seq) data in peripheral blood mononuclear cells (PBMCs) for 21 individuals with East Asian (EAS) ancestry and 23 with European (EUR) ancestry[28] (172,385 cells analyzed across seven main cell types) and ancestry-matched GWASs of 31 diseases and complex traits[7] (average $n \sim 90{,}000$ and $\sim 267{,}000$ in EAS and EUR, respectively). We observed that ancDE genes tended to be cell-type specific and enriched in genes interacting with the environment and in variants with ancestry-specific disease effect sizes, which suggests the impact of shared cell-type-specific GxE interactions between regulatory and disease architectures. Our

[1]Department of Population and Public Health Sciences, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA; [2]Center for Genetic Epidemiology, 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, USA
*Correspondence: juehanwa@usc.edu (J.W.), gazal@usc.edu (S.G.)
https://doi.org/10.1016/j.ajhg.2024.07.021.

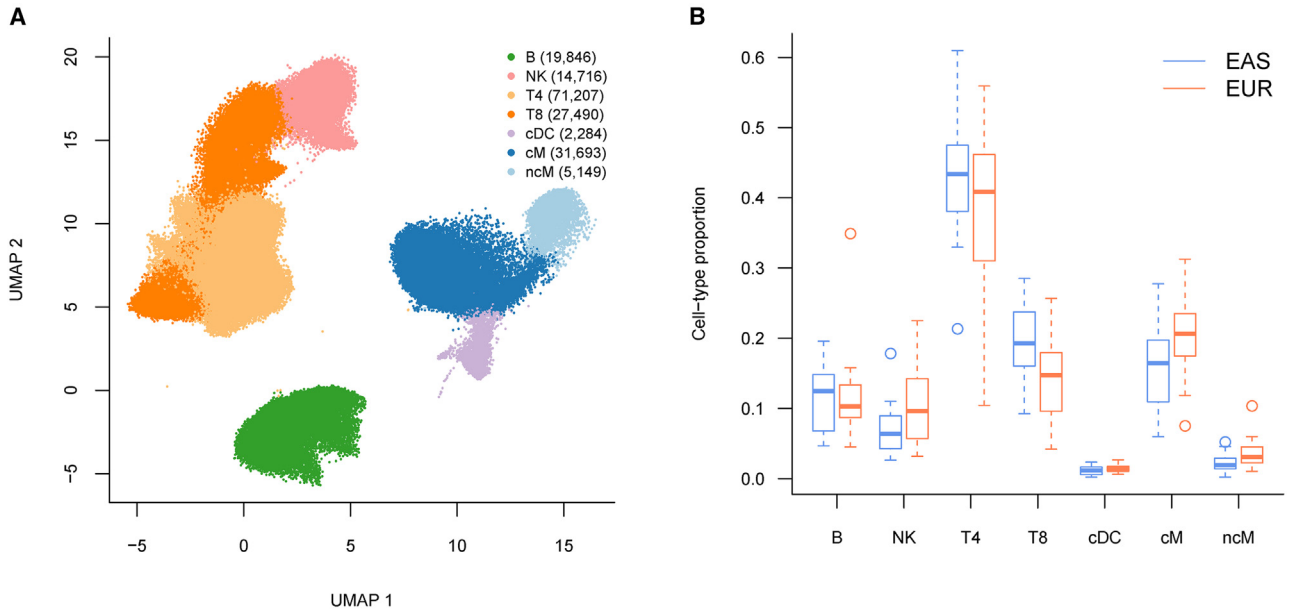

Figure 1. An immune multi-ancestry single-cell dataset
(A) We report the Uniform Manifold Approximation and Projection (UMAP) coordinates and assignment of 172,385 cells to seven immune cell types[28]: B cells (B), natural killer (NK) cells, $CD4^+$ and $CD8^+$ T cells (T4 and T8, respectively), conventional dendritic cells (cDCs), and classical and non-classical monocytes (cMs and ncMs, respectively). The number of cells in each cell type is reported in the legend.
(B) We report cell-type proportions across 21 EAS and 23 EUR individuals; we did not observe significant differences ($p < 0.05/7$) in cell-type proportions across ancestries (min $p = 0.03$ in cM; Table S1). The median value of each proportion is displayed as a band inside each box. Boxes denote values in the second and third quartiles. The length of each whisker is 1.5 times the interquartile range (defined as the height of each box). All values lying outside the whiskers are considered to be outliers.

results imply that large single-cell and GWAS datasets from diverse ancestries are required to improve our understanding of human diseases.

Material and methods

scRNA-seq data in PBMCs

We used processed scRNA-seq data in PBMCs from Perez et al.[28] for 256 individuals of EAS and EUR ancestries. To obtain homogeneous samples, we removed 158 systemic lupus erythematosus cases, 50 controls from the ImmVar study (only European individuals), two outliers in a principal component analysis (PCA) (Figure S1), and two males in the remaining dataset, thus obtaining a dataset of 21 EAS female controls and 23 EUR female controls generated in similar batches. These controls were from the Rheumatology Clinic at University of California, San Francisco, and their cells were reported to come from two different batches (two EAS and four EUR in the first batch, 20 EAS and 21 EUR in the second batch; one EAS and two EUR had cells from the two batches), thus minimizing batch effects between the two groups.

We focused on the seven most abundant cell types: B cells (B), natural killer (NK) cells, $CD4^+$ and $CD8^+$ T cells (T4 and T8, respectively), conventional dendritic cells (cDCs), and classical and non-classical monocytes (cMs and ncMs, respectively) (Figure 1A). After removing cells with more than 20% of their reads in 13 mitochondrial (MT) genes and cells with less than 500 reads or more than 10,000 reads, we obtained a total of 172,385 cells across the seven cell types with the number of cells varying from 2,284 to 71,207 (Table S1). Cell-type proportions did not significantly differ across the two ancestries (Figure 1B; Table S1).

For supplementary analyses, we used OneK1K scRNA-seq for 982 Europeans (565 females and 416 males) in PBMCs from Yazar et al.[29] We defined B cells as those labeled "naive B cell," "memory B cell," and "transitional stage B cell"; NK cells as those labeled "natural killer cell," T4 cells as those labeled "central memory CD4-positive, alpha-beta T cell," "naive thymus-derived CD4-positive, alpha-beta T cell," "effector memory CD4-positive, alpha-beta T cell," "CD4-positive, alpha-beta cytotoxic T cell," and "CD4-positive, alpha-beta T cell"; T8 cells as those labeled "effector memory CD8-positive, alpha-beta T cell," "naive thymus-derived CD8-positive, alpha-beta T cell," "central memory CD8-positive, alpha-beta T cell," "CD8-positive, and alpha-beta T cell"; cMs as those labeled "CD14-positive monocyte"; and ncMs as those labeled "CD14-low, CD16-positive monocyte." We applied a similar quality control as described above and obtained a total of 1,175,543 cells for the seven cell types.

Genes differentially expressed across ancestries

For each cell type, we tested whether each gene was differentially expressed between EAS and EUR individuals. We used a Poisson linear mixed-effects model with number of reads as the outcome variable; the donor as a random effect; and ancestry, age, batch, five first principal components of a PCA computed at the cell-type level on the 2,000 most variable genes, log of total number of reads per cell, proportion of genes expressed in a single cell (cellular detection rate), and fraction of reads in MT genes as fixed effect covariates. We restricted our analyses to 19,995 genes,[30] and for each cell type, we restricted our analyses to genes with at least 50 reads across all controls (48% of the genes on average). For main analyses, we defined genes with the top 100 smallest p values for the ancestry covariate as ancDE genes. We performed a similar approach to compute genes differentially expressed between

females and males (sexDE genes) using data from Yazar et al.[29] We note that using a Poisson linear mixed-effects model with principal components computed at the cell-type level should have limited the impact of cell-type heterogeneity on ancDE and sexDE gene results (as reported in Aquino et al.[18] and Oliva et al.[31]). Analyses to estimate power in detecting ancDE genes with this sample size are discussed in Figures S2–S4.

We performed additional analyses in which we recomputed p values by permuting the ancestry label of the individuals (100 permutations; if $p \leq 0.01$, 900 more permutations were performed). Permutations decreased the genomic factor from 1.28 to 1.16 (Figure S5). We observed a large overlap between the top 100 p nominal and top 100 permuted p across cell types: 77% (78% and 83% for top 200 and top 500, respectively), thus demonstrating robust ranking of the ancDE genes.

We performed power analyses to detect ancDE genes with two groups of 21 and 23 individuals by leveraging the Geuvadis dataset[13] (RNA-seq in lymphoblastoid cell lines for $n = 89$ and 373 individuals of African and European ancestry after GTEx recommended quality-control pipeline, respectively). Analyses were performed on standardized gene expression and adjusted on sex and five gene expression principal components (PCs).

Gene ontology enrichment analysis

We performed gene ontology (GO) enrichment analysis of ancDE genes using the R package *goseq*.[32] We restricted analyzed pathways to the biological processes containing from 10 to 1,000 genes. We defined the reference set of genes as the genes with at least 50 reads across all samples within investigated cell types. We computed false discovery rate (FDR) p values using the Benjamini and Hochberg correction[33] implemented in the R p.adjust function.

Cell-type-specific eQTL analyses

To determine whether the cell-type-specific ancDE genes were driven by allele frequency differences, we leveraged independent EAS and EUR cell-type-specific eQTLs from Ishigaki et al.[34] and Yazar et al.[29] For EAS, we leveraged eQTLs in B, T4, T8, NK, and monocytes; we assigned EAS monocytes eQTLs to ancDE genes from cMs or ncMs. Genes with significant eQTL were defined as in Yazar et al.[29] (i.e., by computing eQTL q values within each gene, then computing the FDR-corrected p value of the most significant eQTL within each chromosome). For EUR, we defined B cell eQTL by merging eQTL from cell types labeled as "B IN" and "B Mem"; T4 eQTL by merging eQTLs from cell types labeled as "CD4 ET," "CD4 NC," and "CD4 SOX4"; and T8 eQTL by merging eQTLs from cell types labeled as "CD8 ET," "CD8 NC," and "CD8 S100B." We defined cDC, cM, nCM, and NK eQTL by considering eQTL from cell types labeled as "DC," "Mono C," "Mono NC," and "NK R," respectively. We restricted all analyses to variants with mean allele frequency (MAF) > 5% in EAS or MAF > 5% in EUR.

We compared the EAS and EUR allele frequency of these eQTL using 481 EAS and 489 EUR individuals from 1000 Genomes Project.[35,36] Fixation index (F_{st}) across EAS and EUR individuals were computed using the formula[37]

$$F_{st} = E\left(\frac{(\widehat{p}_{EAS} - \widehat{p}_{EUR})^2 - \left(\frac{1}{2n_{EAS}} + \frac{1}{2n_{EUR}}\right)p_{avg}\left(1 - p_{avg}\right)}{2p_{avg}\left(1 - p_{avg}\right)}\right)$$

where $\widehat{p}_{EAS}$ and $\widehat{p}_{EUR}$ are the allele frequencies estimated in the EAS and EUR individuals, respectively, n_{EAS} and n_{EUR} are the corresponding sample sizes, and p_{avg} is defined as $(p_{EAS} + p_{EUR})/2$.

To test whether the 329 ancDE genes with EUR independent eQTL were still significantly differentially expressed after conditioning to the genotypes of these eQTLs, we extracted those SNPs in the genetic data of Perez et al.[28] (genotypes of eQTL were available for 273 of 329 genes) and replicated our differentially expression analyses for ancDE genes while correcting for the genotypes of the eQTL.

Estimating enrichment of stratified squared multi-ancestry genetic correlation using S-LDXR

S-LDXR[7] is a method to estimate enrichment of stratified squared multi-ancestry genetic correlation across functional categories of SNPs using GWAS summary statistics and ancestry-matched linkage disequilibrium (LD) reference panels. S-LDXR models per-allele effect sizes (accounting for differences in allele frequency differences between ancestries) of SNP j in two ancestries (labeled as β_{1j} and β_{2j}) with variance and covariance,

$$Var\left[\beta_{1j}\right] = \sum_C a_C(j)\tau_{1C},\ Var\left[\beta_{2j}\right] = \sum_C a_C(j)\tau_{2C},\ Cov\left[\beta_{1j}, \beta_{2j}\right] = \sum_C a_C(j)\theta_C$$

where $a_C(j)$ is the value of SNP j for annotation C, τ_{1C} and τ_{2C} are the net contribution of annotation C to the variance of β_{1j} and β_{2j}, respectively, and θ_C is the net contribution of annotation C to the covariance of β_{1j} and β_{2j}.

S-LDXR estimates the stratified squared multi-ancestry genetic correlation, which is defined as

$$r_g^2(C) = \frac{\rho_g^2(C)}{h_{g1}^2(C)h_{g2}^2(C)}$$

where h_{g1}^2 and h_{g2}^2 are heritabilities in each ancestry, and ρ_g is the multi-ancestry genetic covariance of each binary annotation C:

$$\rho_g(C) = \Sigma_{j\epsilon C}(\Sigma_{C'} a_{C'}(j)\theta_{C'})$$

where $a_{C'}(j)$ and $\theta_{C'}$ are annotations and coefficients for all annotations C' included in the analysis, respectively.

Then S-LDXR estimates the enrichment of squared multi-ancestry genetic correlation, which is defined as

$$\lambda^2(C) = \frac{r_g^2(C)}{r_g^2}$$

where r_g^2 is the squared correlation of the per-allele effect sizes in the two ancestries (i.e., squared multi-ancestry genetic correlation), and $r_g^2(C)$ is the squared multi-ancestry genetic correlation computed within SNPs in C. We note that λ^2 is not affected by allele frequency differences across the two ancestries (because it computes the correlation of per-allele effect sizes and not effect sizes on normalized genotypes) and that S-LDXR estimates of λ^2 were unbiased for annotations with high allele frequency differences in simulations (and conservative for other annotations).[7]

We applied S-LDXR using recommended settings, reference files (i.e., 481 EAS and 489 EUR samples in the 1000 Genomes Project[36]), and a background set of functional annotations (i.e., the baseline-LD-X model, a set of 62 functional SNP-annotations known to impact per-allele effect sizes). We applied S-LDXR on 31 diseases and complex traits[7] (average $n \sim 90{,}000$ and $\sim 267{,}000$ in EAS and EUR individuals, respectively); most of the results were meta-analyzed across 20 approximately independent traits, including 10 approximately independent blood and

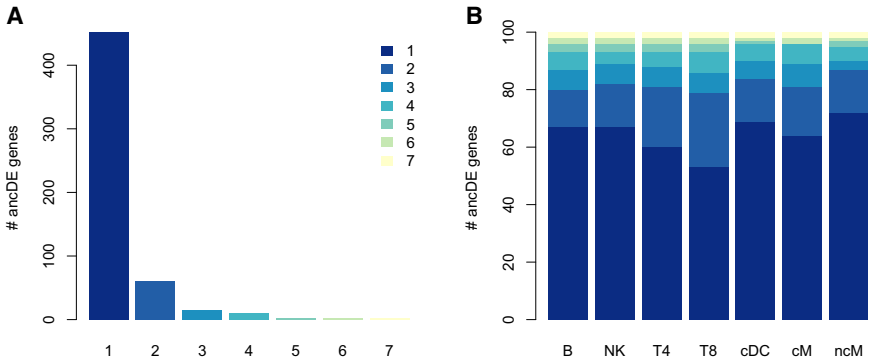

Figure 2. Cell-type specificity of ancDE genes

(A) We report the number of cell-type-specific ancDE genes (top 100 smallest p values for each cell type) shared across all the cell types. We observed that 83% of ancDE genes were differentially expressed in a single cell type.

(B) For each cell type, we report the number of ancDE genes shared across all the cell types. List of ancDE genes is reported in Table S3. Across the cell types, 53%–72% of their ancDE genes were cell-type specific. Similar patterns were observed when defining ancDE genes using the 200 and 500 smallest p values and the 100 smallest permuted p values (Figures S7–S9) and for genes differentially expressed in males and females[29] (Figure S10).

immune-related traits (Table S2). S-LDXR estimated a cross-ancestry genetic correlation of $r_g = 0.88 \pm 0.06$ meta-analyzed across the 20 independent traits (Table S2), consistent with recent estimates.[8,9] Reported p values for heritability enrichments were two sided (i.e., testing if heritability enrichment is different from 1); reported p values for λ^2 and θ were one sided (i.e., testing whether λ^2 and θ are lower than 1 and 0, respectively).

Our analyses included SNP-annotations related to gene sets, which were constructed by adding 100-kb windows on either side of the transcribed region of each gene in the set.[7,38] All analyses included a SNP-annotation for the 19,995 genes, and seven SNP-annotations representing the set of genes expressed in each cell type (i.e., genes with at least 50 reads across all controls).

Extending S-LDXR to estimate enrichment of stratified squared sex genetic correlation

We extended S-LDXR to estimate enrichment of stratified squared sex genetic correlation using GWAS summary statistics computed in males and females of the same ancestry and a corresponding LD reference panel. Here, we applied S-LDXR using recommended settings and the EUR reference file and the baseline-LD model version 2.2 used by S-LDSC.[39] We downloaded the male and female GWASs previously identified with sex genetic correlation significantly <1 (Bernabeu et al.[40]) and defined a set of 17 independent traits with genetic correlation[41] < 0.1. We observed consistent squared sex genetic correlation and multi-ancestry genetic correlation among annotations of the baseline-LD and baseline-LD-X models (Figure S6).

Results

Cell-type specificity of genes differentially expressed between EAS and EUR ancestries

We tested differential gene expression in seven immune cell types within 21 and 23 healthy individuals of EAS and EUR ancestry, respectively. We restricted the main analyses to the top 100 differentially expressed genes with the smallest p values within each cell type, which led to a list of 545 unique ancDE genes (including only 15 genes of the major histocompatibility complex [MHC] region) (Table S3). All further analyses were replicated by using top 200 and 500 genes with the smallest p values within each cell type; we also replicated our analyses by recomputing p values using permutations.

Among the 545 unique ancDE genes, 452 (83%) were differentially expressed in a single cell type, thereby suggesting high cell type specificity of differential expression across ancestries (Figure 2A). Within each cell type, 53%–72% of ancDE genes were specific to this cell type (Figure 2B). Genes differentially expressed in at least two cell types tended to cluster as in hematopoietic lineages, such as lymphoid (NK, B, T4, and T8) and myeloid cell types (cM, ncM, and cDC) (e.g., 28 genes were differentially expressed in both T4 and T8 cell types, and 18 genes were differentially expressed in cMs and ncMs). Among ancDE genes differentially expressed in at least two cell types, a large proportion (86 of 93; 92%) had consistent directions across the cell types (i.e., overexpressed or underexpressed in all cell types) (Table S4).

To validate that the cell-type specificity of ancDE genes was not an artifact of a relatively low number of samples (despite the high number of cells), we performed the following supplementary analyses. First, we replicated this observation when defining ancDE genes using the 200 and 500 smallest p values and the top 100 permuted p values (82%, 77%, and 87% of unique ancDE genes, respectively, were differentially expressed in a single cell type) (Figures S7–S9; Tables S3 and S5). Second, we leveraged a larger scRNA-seq with 416 EUR males and 565 EUR females[29] (1,175,543 cells across seven similar cell types) and identified the top 100 genes that were the most significantly differentially expressed by sex for each cell type (578 unique sexDE genes). We observed that 86% of sexDE genes were differentially expressed in a single cell type, thus confirming that environment differences affected gene expression at the cell-type-specific level (Figure S10; Table S6).

AncDE genes are enriched in genes interacting with the environment

We next sought to investigate whether ancDE genes tended to be driven by environmental differences

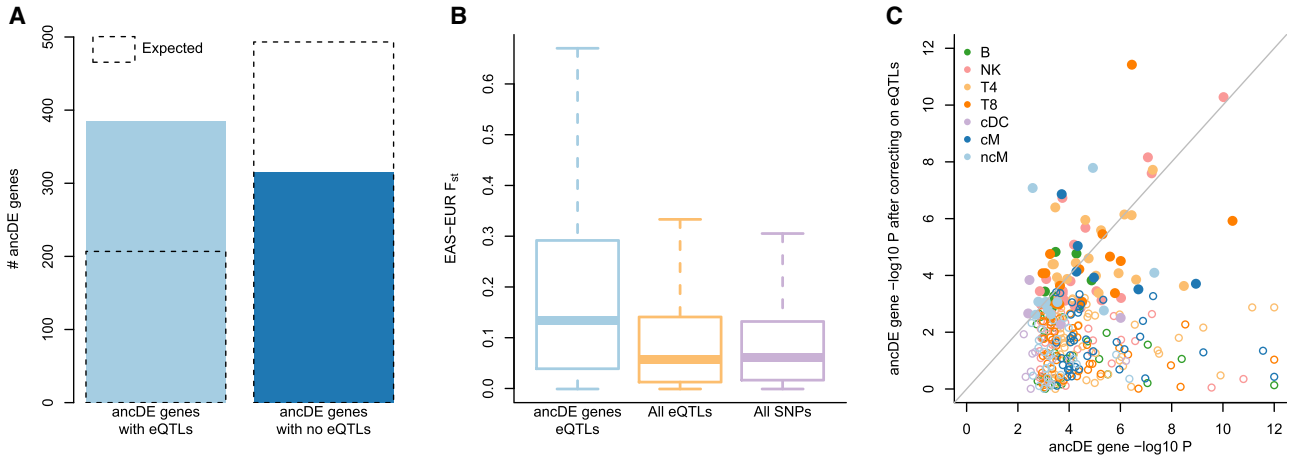

Figure 3. ancDE genes are driven by allele frequency differences of their eQTL
(A) We report the number of cell-type-specific ancDE genes with at least one EUR eQTL and without eQTL in the corresponding cell type. Dotted boxes represent the number of ancDE genes that would have been observed by chance.
(B) We report mean fixation index (F_{st}) across EAS and EUR reference populations[37] for all ancDE gene eQTL, eQTL of all expressed genes, and all SNPs. The median value of each expression is displayed as a band inside each box. Boxes denote values in the second and third quartiles. The length of each whisker is 1.5 times the interquartile range (defined as the height of each box).
(C) Scatterplot of ancDE genes $-\log_{10}(P)$ before and after conditioning on their eQTL. Solid points represent ancDE genes that remain in the top 100 most significantly differentially expressed genes after conditioning on their eQTL.

(i.e., GxE interactions of their eQTL or different gene expression response to environments with no genetic mediation), genetic differences (ancestry differences in allele frequencies of their eQTL), or both.

First, GO enrichment analyses[32] revealed that the 545 ancDE genes were enriched in genes involved in immune response to the environment (FDR-corrected $p = 7.88 \times 10^{-4}$ for leukocyte activation [GO: 0045321] pathway; Table S7). At the cell-type level, the most significant enriched GO categories were in ncMs (FDR-corrected $p = 7.01 \times 10^{-7}$ secretory granule [GO: 0030141] pathway), NK cells (FDR-corrected $p = 1.7 \times 10^{-5}$ for interferon-gamma-mediated signaling [GO: 0060333] pathway), and cMs (FDR-corrected $p = 1.48 \times 10^{-2}$ for neutrophil degranulation [GO: 0043312] pathway) (Table S7). We detected even more significantly enriched pathways involved in immune response when defining ancDE genes by using the 200 and 500 smallest p values and the 100 smallest permuted p values (FDR-corrected $p = 1.01 \times 10^{-5}$, 9.26×10^{-7}, and 7.25×10^{-5} for regulated exocytosis [GO: 0045055], symbiotic process [GO: 0044403], and leukocyte activation [GO: 0045321] pathways, respectively). Similar conclusions were obtained when removing MHC genes from the analysis (Table S7).

Second, we investigated whether ancDE genes were due to allele frequency differences of their eQTLs by leveraging cell-type-specific eQTLs from 105 EAS[34] and 982 EUR individuals[29]; cDC eQTLs were not available for EAS individuals, and we used EAS monocyte eQTL for both cMs and ncMs. More than half of cell-type-specific ancDE genes (385 of 700; 55%) had at least one independent eQTL in the corresponding cell type, which is 1.9 times more than what would be expected by chance (Figure 3A); these numbers should be considered a lower bound because the statistical power to detect eQTLs is highly associated with the eQTL dataset sample size.[42] Of note, these eQTLs tended to have extremely high fixation index (F_{st}) across EAS and EUR reference populations[36] (mean $F_{st} = 0.19$ across all ancDE gene eQTLs vs. mean $F_{st} = 0.10$ across eQTL for all expressed genes, $p = 6 \times 10^{-37}$ for difference; Figure 3B). When replicating differentially expression analyses for ancDE genes for which genotypes of the eQTL were available (329 of 385), we found that 74% of these genes (243 of 329) did not remain in the top 100 most significantly differentially expressed genes after conditioning on their eQTL (Figure 3C). By multiplying the proportion of ancDE genes with known eQTLs (55%) to the proportion of these genes that were not differentially expressed after conditioning to their eQTL (74%), we estimated that at least 41% of ancDE genes (0.55×0.74) could be driven by allele frequency differences of their eQTL across ancestries. We replicated all our analyses and conclusions by defining ancDE genes using the 200 and 500 genes with the smallest differential gene expression p values within each cell type (Figure S11) and when using eQTL from EAS and EUR individuals separately (Figure S12).

Finally, we investigated whether allele frequency differences of eQTLs across ancestries were due to adaptation to new environments or to genetic drift. We observed that ancDE genes with eQTLs were also enriched in genes involved in immune response (minimum FDR-corrected $p = 4.47 \times 10^{-3}$, 4.13×10^{-3}, 2.56×10^{-4}, and 7.98×10^{-4} for leukocyte activation [GO: 0045321], response to interferon-gamma [GO: 0034341], defense response [GO: 0006952], and fatty acid derivative biosynthetic process [GO: 1901570] pathways when considering the 100, 200, and 500 most-significant ancDE genes and the top 100

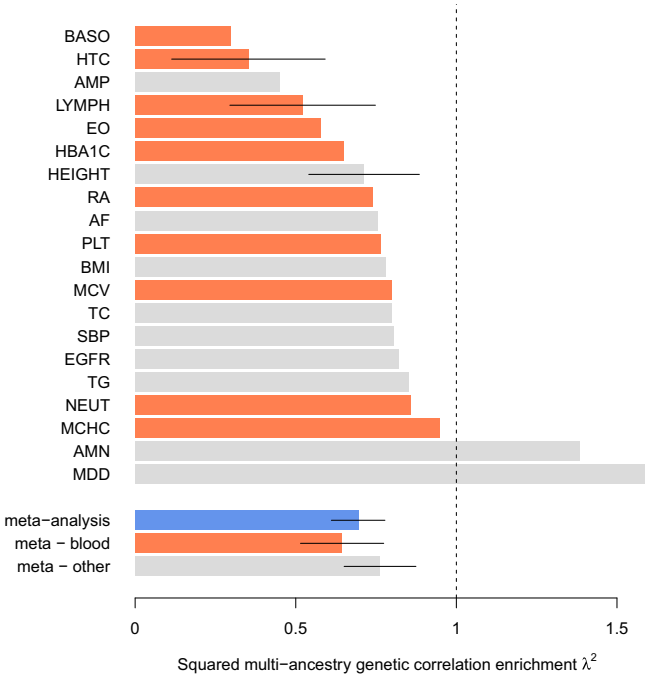

Figure 4. Squared multi-ancestry genetic correlation enrichment for variants surrounding ancDE genes

We report squared multi-ancestry genetic correlation enrichment (λ^2) for each independent trait and meta-analyses results across groups of traits. Orange bars represent blood and immune-related traits, gray bars represent other traits, and the blue bar represents the meta-analysis across all traits. Error bars represent 95% confidence intervals (CIs) and were only plotted for traits with λ^2 value significantly <1 for visualization purposes; CIs for all traits are reported in Figure S14. Numerical results are reported in Tables S8 and S9. We observed $\lambda^2 > 1$ for AMN (1.39, 95% CI = [0.61, 2.17]) and MDD (1.59, 95% CI = [−0.61, 2.17]), two of the most underpowered traits of this study; although multi-ancestry genetic correlation r_g has the biologically plausible [−1,1] range, λ^2 can biologically be >1 if within the annotation is greater than the genome-wide r_g; also, $\lambda^2 > 1$ for these two traits has no biological meaning because of the large 95% CIs.

AF, atrial fibrillation; AMN, age at menarche; AMP, age at menopause; BASO, basophil count; BMI, body mass index; EGFR, estimated glomerular filtration rate; EO, eosinophil count; HBA1C, hemoglobin A1c; HTC, hematocrit; LYMPH, lymphocyte count; MCHC, MCH concentration; MCV, mean corpuscular volume; MDD, major depressive disorder; NEUT, neutrophil count; PLT, platelet count; RA, rheumatoid arthritis; SBP, systolic blood pressure; TC, total cholesterol; TG, triglycerides.

permuted p values, respectively; Table S7), which suggests that allele frequency differences of their eQTL might have been driven by adaptation to new environments rather than genetic drift. Similarly, 27% of cell-type-specific sexDE genes (186 of 700) had at least one independent EUR eQTL (Figure S13), so genes with eQTL can be differentially expressed even without differences in allele frequencies.

All together, these results suggest that ancDE genes are enriched in genes interacting with the environment. While we estimated that at least 41% of ancDE genes could be explained by allele frequency differences of their eQTLs, it is likely that a significant fraction of these differences was driven by adaptation to new environments.

ancDE genes are enriched in ancestry-specific causal effect sizes of complex traits

We created a SNP annotation for the 545 unique ancDE genes (annotation representing 4.8% of investigated common SNPs) and analyzed it using S-LDXR on 31 diseases and complex traits. We then meta-analyzed results across 20 approximately independent traits. SNPs in ancDE genes were significantly enriched in SNP-heritability (h^2) in EAS and EUR GWASs (h^2 enrichment = 2.07 ± 0.18, $p = 3 \times 10^{-9}$ in EAS; h^2 enrichment = 1.71 ± 0.13, $p = 2 \times 10^{-8}$ in EUR), which highlights the impact of ancDE genes on human diseases and complex traits.

SNPs within ancDE genes were extremely depleted of squared multi-ancestry genetic correlation ($\lambda^2 = 0.69 \pm 0.04$, $p = 6 \times 10^{-13}$; Figures 4 and S14; Tables S8 and S9) and more depleted (and most significantly depleted) than any other annotation from the baseline-LD-X model (Table S10). We detected significant depletions ($p < 0.05/31$) for 6 traits, including hematocrit ($\lambda^2 = 0.35 \pm 0.12$, $p = 5 \times 10^{-8}$), lymphocyte count ($\lambda^2 = 0.52 \pm 0.12$, $p = 2 \times 10^{-5}$), and height ($\lambda^2 = 0.71 \pm 0.09$, $p = 5 \times 10^{-4}$) (Table S8). The ancDE gene λ^2 was also significantly lower than the λ^2 estimated on all genes ($\lambda^2 = 0.95 \pm 0.01$, $p = 3 \times 10^{-9}$ for difference with ancDE genes) and on genes expressed in the seven cell types that were not ancDE genes ($\lambda^2 = 0.90 \pm 0.01$, $p = 4 \times 10^{-6}$ for difference with ancDE genes). By performing 100 random sampling of 545 genes, we also validated that ancDE genes were significantly depleted in multi-ancestry genetic correlation ($\lambda^2 = 0.84 \pm 0.01$, $p < 1/100$ for difference with ancDE genes; Figure S15). Finally, the net contribution of the ancDE gene annotation to the covariance of effect sizes (θ, see material and methods) was among the most significant binary annotations ($p = 0.02$, $\theta < 0$ for 15 of 20 traits; Tables S8–S10), meaning that multi-ancestry genetic correlation depletion of this annotation was not fully captured by existing annotations of the baseline-LD-X model.

As expected, the λ^2 was even smaller when meta-analyzed across 10 approximately independent blood and immune-related traits ($\lambda^2 = 0.64 \pm 0.07$, $p = 7 \times 10^{-8}$) and remained significantly depleted in the 10 remaining traits ($\lambda^2 = 0.76 \pm 0.06$, $p = 3 \times 10^{-5}$). Similar conclusions were obtained when defining ancDE genes using the 200 and 500 smallest p values and the permuted p values (Table S9). We reran S-LDXR on two distinct annotations corresponding to ancDE genes with and without eQTL (each annotation represented 2.8% and 2.1% of investigated SNPs, respectively). We observed depletion of squared multi-ancestry genetic correlation for the two annotations ($\lambda^2 = 0.71 \pm 0.07$ and $\lambda^2 = 0.62 \pm 0.02$, respectively) (Table S9). These results suggest that even if genes were differentially expressed due to allele frequency differences of their eQTL, these genes are likely enriched in ancestry-specific causal effect sizes.

To support that genes with varying levels of expression in different environments are enriched in context-specific causal effect sizes, we first leveraged genes differentially

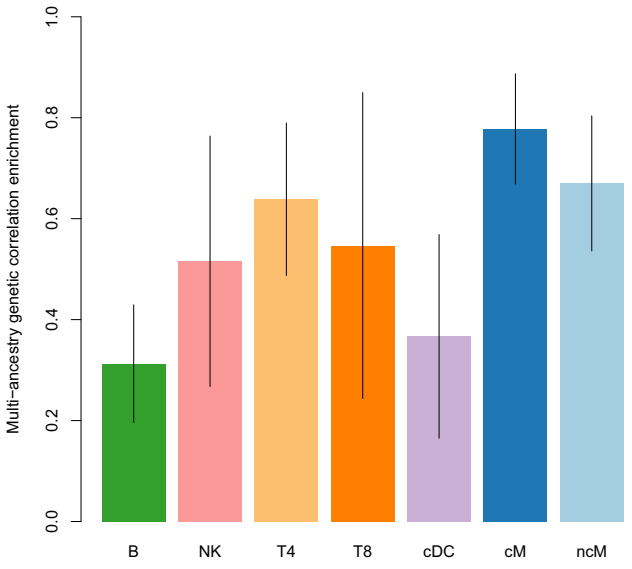

Figure 5. Squared multi-ancestry genetic correlation enrichment for variants surrounding cell-type-specific ancDE genes
We report squared multi-ancestry genetic correlation enrichment (λ^2) for cell-type-specific ancDE gene annotations meta-analyzed across 20 independent traits. Error bars represent 95% CIs. Numerical results are reported in Table S9.

expressed between two other ancestry groups: African and EUR individuals in five PBMC types from Randolph et al.[17] (90 individuals) and Aquino et al.[18] (160 individuals). We observed significant depletion of squared ancestry genetic correlation within our EAS and EUR GWASs ($\lambda^2 = 0.63 \pm 0.05$, $p < 2 \times 10^{-12}$, and $\lambda^2 = 0.82 \pm 0.04$, $p < 9 \times 10^{-6}$, respectively), thereby suggesting that GWAS discordant effects between ancestries around ancDE genes are more likely due to interaction with the environment than different allele frequency differences of their eQTL. Then, we extended S-LDXR to stratify squared genetic correlation between male and female GWASs and applied it to sexDE gene annotations on 17 independent traits previously identified with sex genetic correlation significantly less than 1 (Bernabeu et al.[40]) (see material and methods). We observed significant depletion of squared sex genetic correlation within our male and female GWASs ($\lambda^2 = 0.91 \pm 0.02$, $p < 2 \times 10^{-7}$) and similar depletion when stratifying sexDE genes with and without EUR eQTL ($\lambda^2 = 0.89 \pm 0.03$ and $\lambda^2 = 0.90 \pm 0.02$, respectively) (Table S11).

Finally, to refine the ancDE gene λ^2 signal, we applied S-LDXR analyses by creating SNP-annotations for each of the seven main cell types (each annotation represents 0.8%–1.0% of investigated SNPs). The seven SNP-annotations were all depleted of squared multi-ancestry genetic correlation ($\lambda^2 < 0.78$; Figure 5; Table S9), and all cell types except T8 were significantly depleted. Although none of the λ^2 values significantly differed from each other, we observed the smallest and most significant depletion for the ancDE genes in B cells ($\lambda^2 = 0.35 \pm 0.06$, $p = 1 \times 10^{-24}$) and cDC cells ($\lambda^2 = 0.36 \pm 0.10$, $p = 6 \times 10^{-10}$). The λ^2 values were smaller for the B and cDC annotations when meta-analyzed across 10 approximately independent blood and immune-related traits ($\lambda^2 = 0.30 \pm 0.07$, $p = 4 \times 10^{-21}$ and $\lambda^2 = 0.21 \pm 0.12$, $p = 1 \times 10^{-10}$, respectively); this trend was not observed across the 10 remaining traits (Figure S16; Table S9). Similar trends were obtained when defining ancDE genes using the 200 and 500 smallest p values and the top 100 permuted p values (Figures S17–S19; Table S9).

Altogether, these results demonstrate discordant causal effect sizes between EAS and EUR GWASs for variants surrounding ancDE genes, likely due to GxE interactions. The magnitude of λ^2 was similar for genes with and without eQTL, which suggests that even if a gene is differentially expressed because of different allele frequencies of its eQTL, this gene is likely to also be enriched in GxE effects (because differences in allele frequencies might have been driven by adaptation). Finally, for blood- and immune-related traits, we observed stronger discordant effect sizes for SNPs within ancDE genes in B cells and cDCs, two cell types that initiate and shape the adaptive immune response to new environments.

Illustrating ancDE genes with strong GWAS discordant effect sizes

Here, we illustrate how different environments may have led to differential expression of the ancDE gene *MCL1* in B cells and different allele effect sizes around *MCL1* in lymphocyte count (LYMPH) EAS and EUR GWASs[43] (Figure 6). *MCL1* is essential to B cell development[44–46] and was differentially expressed between EAS and EUR individuals only in B cells (nominal $p = 2 \times 10^{-5}$, permuted $p < 0.001$; Figure 6A). No significant *MCL1* eQTL in B cells were reported in EAS[34] or EUR individuals.[29]

We observed significant associations in the LYMPH EUR GWAS (minimum $p = 3 \times 10^{-42}$ for rs11204702) for *MCL1* but not in the LYMPH EAS GWAS (p at rs11204702 in EAS = 0.63) (Figure 6B). We observed significant different marginal allele per-effect sizes for the EUR most significant loci (per-allele effect size for rs11204702 A allele = 0.002 ± 0.005 and 0.026 ± 0.002 in EAS and EUR, respectively; $p = 6 \times 10^{-6}$ for difference; Figure 6C), similar allele frequencies for the most associated variants in EUR individuals (Figure 6D), and similar LD patterns with the lead SNP in both ancestries (Figure S20). Hence, these discordant effects were not driven by power issues due to different GWAS sample sizes ($n \sim 89{,}000$ in EAS[47] vs. $n \sim 525{,}000$ in EUR[48]) or different allele frequencies and LD structure across the ancestries, respectively.

We also found no significant association in LYMPH GWASs performed in individuals of African and Latino ancestry and significant different marginal allele per-effect sizes for rs11204702 ($p = 0.001$ and 0.01 for EUR-African and EUR-Latino differences, respectively; Figure S21), which suggests a EUR-specific effect for rs11204702. *MCL1* expression in B cells was significantly lower in EUR than African individuals in Aquino et al.[18] ($p = 3 \times 10^{-4}$; this association was not observed in Randolph et al.[17]),

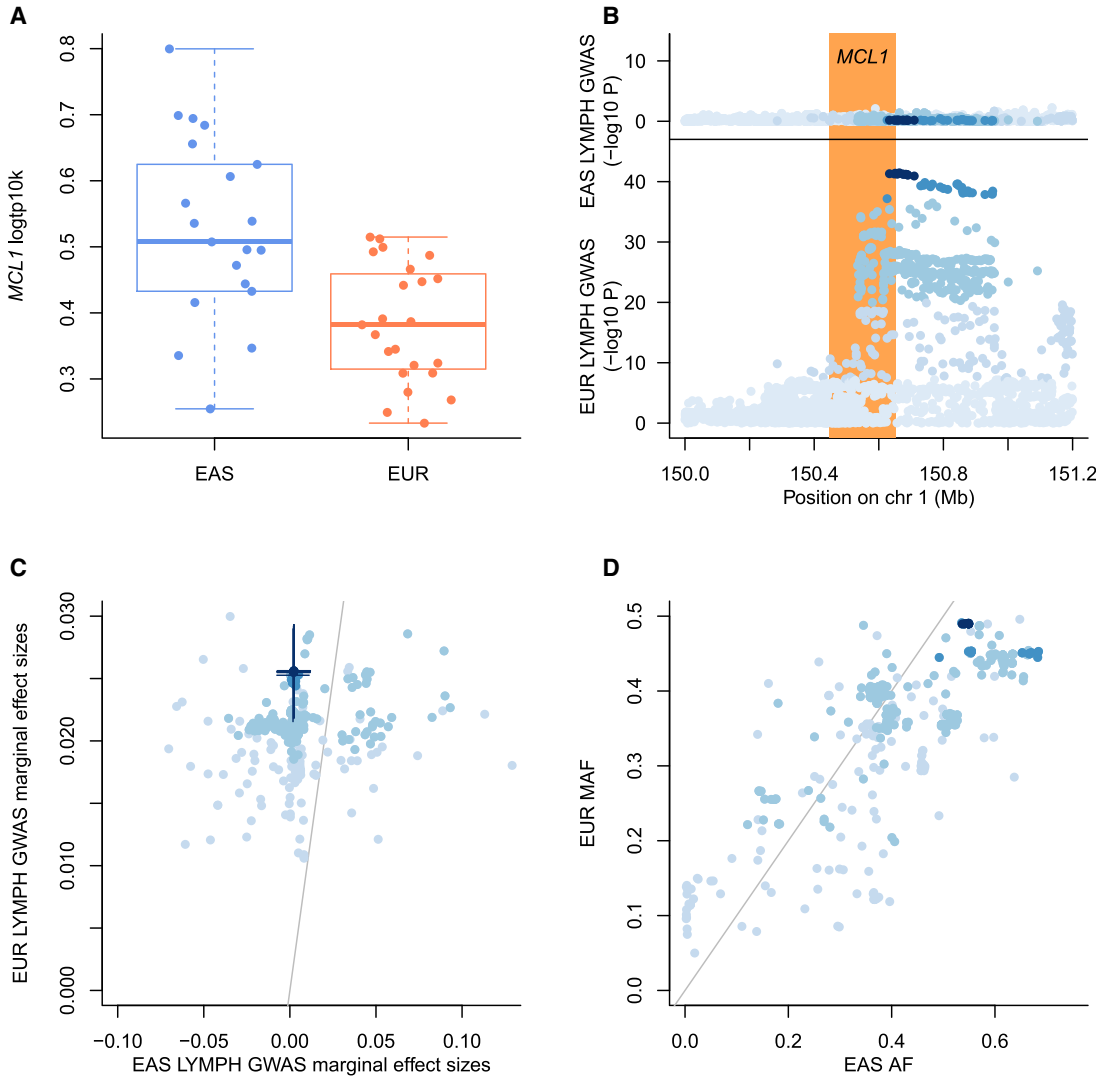

Figure 6. Discordant results between EAS and EUR lymphocyte count GWAS around the B cell ancDE gene *MCL1*
(A) We report logtp10k for *MCL1* pseudo-bulk gene expression in B cells between EAS and EUR individuals. The median value of each expression is displayed as a band inside each box. Boxes denote values in the second and third quartiles. The length of each whisker is 1.5 times the interquartile range (defined as the height of each box). All dots represent observed values.
(B) We report lymphocyte count (LYMPH) $-\log_{10}$ GWAS p values computed for individuals of EAS and EUR ancestry. The orange region represents 100-kb windows on either side of the *MCL1* gene body.
(C) We report LYMPH marginal effect sizes computed for individuals of EAS and EUR ancestry. Marginal effects are plotted using the EUR risk allele as the reference. Blue lines represent 95% CIs for the most associated SNPs in the EUR GWAS.
(D) We report EAS allele frequency (AF) and EUR MAF in 1000 Genomes Project. Color intensity in (B–D) represents GWAS $-\log_{10}(p)$ in EUR GWAS; only SNPs with $p < 5 \times 10^{-8}$ in the EUR GWAS were plotted in (C and D).

which is consistent with our observations for EAS and EUR individuals. All together, these results illustrate that different environments can lead to both ancestry-specific gene expression and ancestry-specific GWAS allele effect sizes for this gene.

Discussion

Although causal effect sizes of human diseases and complex traits tend to be highly correlated across ancestries ($r_g = 0.88 \pm 0.06$ for the 20 independent traits analyzed in this study, consistent with recent estimates[8,9]), understanding where and why ancestry-specific effects of disease risk variants occur is fundamental for understanding the genetic basis of human diseases and for improving the portability of polygenic risk scores across ancestries.[6] Here, we characterized ancestry-specific gene regulation architectures at the cell-type level and investigated their overlap with ancestry-specific disease architectures. We analyzed scRNA-seq data for PBMCs from 44 individuals of EAS or EUR ancestry and observed that ancDE genes tended to be differentially expressed in a single cell type and were enriched in genes involved in immune response to the environment. At least one-third of ancDE genes could be due to allele frequency differences of their eQTLs

(although a large fraction of these eQTLs likely have allele frequency differences due to adaptation to a new environment). Then, by leveraging ancestry-matched GWAS of 31 diseases and complex traits, we determined that squared multi-ancestry genetic correlation enrichment was $\lambda^2 = 0.69 \pm 0.04$ for SNPs surrounding ancDE genes, representing the lowest correlation reported by S-LDXR; numbers were similar when stratifying genes with and without eQTL, which suggests that even if genes were differentially expressed due to allele frequency differences of their eQTL, they are likely enriched in ancestry-specific effect sizes. These depletions were driven by ancDE genes from B cells ($\lambda^2 = 0.35 \pm 0.06$) and cDCs ($\lambda^2 = 0.36 \pm 0.10$). Finally, we illustrated how GxE interactions may have led to differential expression of the ancDE gene *MCL1* in B cells, different *MCL1* eQTL effect sizes in blood, and different allele effect sizes around *MCL1* in LYMPH EAS and EUR GWASs.

To validate that cell-type specificity of ancDE genes were not driven by low single-cell sample size and that our S-LDXR results were not driven by different allele frequency and LD structure across ancestries, we also extended our approach to sex-specific regulatory and complex trait architectures and observed similar patterns. Specifically, by detecting sexDE genes in a larger single-cell dataset[29] (1,175,543 cells from 982 donors), we showed that sexDE genes were also cell-type specific (Figure S7; Table S6) and that nearly one-quarter had at least one independent eQTL (Figure S13), so genes with eQTLs can be differentially expressed even without differences in allele frequencies. Then, by extending S-LDXR to sex-specific effect sizes in sex-specific GWASs (not subject to allele frequency differences across investigated ancestries), we observed a significant depletion of squared sex genetic correlation within 17 independent male and female GWASs[40] ($\lambda^2 = 0.91 \pm 0.02$, $p < 2 \times 10^{-7}$) (Table S11), which confirms the impact of GxE interactions on GWAS effect sizes. In supplementary analyses, we also assessed discordant effects of sex-stratified GWAS within functional annotations from the baseline-LD model[39] and showed similar enrichment of squared multi-ancestry and sex genetic correlations across annotations (Figure S6).

Our findings have several implications for downstream analyses. First, they provide a partial source of explanation for the non-transportability of polygenic risk scores across ancestries.[6] Although modeling the environment in risk prediction is challenging, accounting for genes interacting with the environment (or ancDE genes) in relevant cell types could help downweigh variant effects when computing polygenic risk scores. Second, our results highlight the benefits of generating single-cell datasets for individuals of non-European ancestry because cell-type specificity is crucial to identify ancestry-specific and disease regulatory mechanisms.[38,49–54] Also multi-ancestry functional data may increase power in multi-ancestry GWAS meta-analysis,[7,55] fine-mapping,[7,56] and transcriptome-wide association studies[57] when disease effect sizes and/or gene regulation is ancestry specific. Third, our results highlight the disease relevance of ancDE genes that are driven by allele frequency differences of their eQTL across ancestries, thereby suggesting the impact of selection (rather than genetic drift) on variants regulating genes in the immune system. Characterizing ancDE genes across multiple ancestries should shed light on human adaptation to new environments. Finally, although our results broadly highlight the impact of GxE on the multi-ancestry genetic architecture of human diseases, they also suggest the impact of GxE within a given ancestry and that accounting for environment heterogeneity within a sample can shed light on disease genetic architectures. We propose a framework leveraging single-cell and GWAS datasets that could be extended to analyze the impact of any environment interactions into complex traits, as performed here by extending S-LDXR to analyze the impact of sex on the genetic architectures of complex traits within Europeans.

Our work has several limitations. First, although our dataset was (to our knowledge) the largest multi-ancestry scRNA-seq dataset publicly available, it included only 44 individuals. However, we observed (1) reasonable power to detect the most differentially expressed ancDE genes (Figures S2–S4), (2) that our ancDE genes were significantly enriched in genes interacting with the environment (Table S7), (3) that ancDE genes with eQTL are significantly enriched in eQTL with high fixation index (Figure 3B), (4) extremely significant S-LDXR results obtained for the top 100 ancDE genes in ncMs (one of the lowest abundant cell types with 5,149 cells), thus suggesting that the gene ranking of our analyses is robust, and (5) similar conclusions for ancDE cell-type specificity and enrichment in GWAS discordant effect sizes when detecting sexDE genes in a larger dataset. Altogether, these results suggest that our conclusions on ancDE genes are robust for the low sample size of the dataset that has been used to detect them, although we caution that our list of ancDE genes (Table S3) is imperfect. Low sample size also prevented us from performing eQTL analyses and directly quantifying ancestry-specific eQTL effect sizes at the cell-type level. Despite available cell-type-specific eQTLs from Biobank Japan and OneK1K, these eQTLs were obtained using different technologies (RNA-seq vs. scRNA-seq, respectively) on different sample sizes (105 vs. 982, respectively), which prevents any cross-ancestry eQTL comparison (see also Figure S12). Second, our analyses were restricted to datasets of only two ancestries, which are the only ones with both large functional and GWAS datasets available. Ongoing efforts to generate both functional and GWAS datasets in diverse ancestries would help in replicating our results. Third, our analyses were restricted to gene expression and did not investigate the impact of ancestry-specific regulatory elements (such as enhancers), whereas chromatin signals have been found more ancestry-specific than gene expression.[25] We anticipate that generating diverse

functional datasets (such as single-cell ATAC-seq) in diverse ancestries will help in investigating ancestry-specific regulation at a finer scale. Fourth, our analyses were restricted to seven main PBMC types, which limits the characterization of rarer cell types as well as that of the cellular composition of each main cell type. Fifth, our approach assumes the need for distinct ancestry groups and is not applicable to admixed individuals. Admixture can be leveraged to estimate the correlation of causal effect sizes of an ancestry within two different populations (i.e., European ancestry within European Americans and admixed African Americans, as in Patel et al.[8]) or to estimate correlation of causal effect sizes of multiple ancestries within an admixed population (i.e., African and European ancestries within admixed African Americans, as in Patel et al.[8,9]). Thus, we anticipate that methodological development leveraging admixed individuals to partition ancestry-specific effects across functional annotations will improve our understanding of ancestry-specific causal effect sizes. Finally, because genetic ancestry is multidimensional and continuous,[58–60] assuming distinct ancestry groups also fails to adequately capture human genetic diversity and gene expression variability within an ancestry group. Specifically, by comparing pseudo-bulk expression of ancDE genes across EAS and EUR individuals, we observed gene expression variability across the two ancestry groups (i.e., different variance in EAS and EUR), which suggests that differential expression also occurs within ancestry groups as defined in our analyses (Figure S22). Despite these limitations, our results convincingly demonstrate that ancestry-specific effect sizes are enriched in genes with ancestry-specific regulation and demonstrate the need to generate large single-cell and GWAS datasets in diverse ancestries to improve our understanding of human diseases.

Data and code availability

The single-cell dataset used in this study, differential gene expression results, S-LDXR files, summary statistics in EAS and EUR, and code to replicate our analyses are available at https://zenodo.org/records/11455096.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2024.07.021.

Acknowledgments

We thank A. de Smith, A.L. Price, H. Shi, N. Zaitlen, C.J. Ye, R. Perez, and G. Gordon for helpful discussion. S.G. is funded by NIH grant R35 GM147789.

Declaration of interests

The authors declare no competing interests.

Received: December 21, 2023
Accepted: July 30, 2024
Published: August 26, 2024

Web resources

1000 Genomes data, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502
dbGaP, genotypes from Perez et al., https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs002812.v1.p1
eQTL from Biobank Japan, http://jenger.riken.jp/en/result
Geuvadis data, https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-GEUV-1
GTEx pipeline, https://www.gtexportal.org/home/methods
OneK1K data, https://cellxgene.cziscience.com/collections/dde06e0f-ab3b-46be-96a2-a8082383c4a1
S-LDXR, https://huwenboshi.github.io/s-ldxr/

References

1. de Candia, T.R., Lee, S.H., Yang, J., Browning, B.L., Gejman, P.V., Levinson, D.F., Mowry, B.J., Hewitt, J.K., Goddard, M.E., O'Donovan, M.C., et al. (2013). Additive genetic variation in schizophrenia risk is shared by populations of African and European descent. Am. J. Hum. Genet. *93*, 463–470.
2. Brown, B.C., Asian Genetic Epidemiology Network Type 2 Diabetes Consortium, Ye, C.J., Price, A.L., and Zaitlen, N. (2016). Transethnic Genetic-Correlation Estimates from Summary Statistics. Am. J. Hum. Genet. *99*, 76–88.
3. Mancuso, N., Rohland, N., Rand, K.A., Tandon, A., Allen, A., Quinque, D., Mallick, S., Li, H., Stram, A., Sheng, X., et al. (2016). The contribution of rare variation to prostate cancer heritability. Nat. Genet. *48*, 30–35.
4. Ikeda, M., Takahashi, A., Kamatani, Y., Momozawa, Y., Saito, T., Kondo, K., Shimasaki, A., Kawase, K., Sakusabe, T., Iwayama, Y., et al. (2019). Genome-Wide Association Study Detected Novel Susceptibility Genes for Schizophrenia and Shared Trans-Populations/Diseases Genetic Effect. Schizophr. Bull. *45*, 824–834.
5. Galinsky, K.J., Reshef, Y.A., Finucane, H.K., Loh, P.-R., Zaitlen, N., Patterson, N.J., Brown, B.C., and Price, A.L. (2019). Estimating cross-population genetic correlations of causal effect sizes. Genet. Epidemiol. *43*, 180–188.
6. Martin, A.R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B.M., and Daly, M.J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. Nat. Genet. *51*, 584–591.
7. Shi, H., Gazal, S., Kanai, M., Koch, E.M., Schoech, A.P., Siewert, K.M., Kim, S.S., Luo, Y., Amariuta, T., Huang, H., et al. (2021). Population-specific causal disease effect sizes in functionally important regions impacted by selection. Nat. Commun. *12*, 1098.
8. Patel, R.A., Musharoff, S.A., Spence, J.P., Pimentel, H., Tcheandjieu, C., Mostafavi, H., Sinnott-Armstrong, N., Clarke, S.L., Smith, C.J.; and VA Million Veteran Program (2022). Genetic interactions drive heterogeneity in causal variant effect sizes for gene expression and complex traits. Am. J. Hum. Genet. *109*, 1286–1297.
9. Hou, K., Ding, Y., Xu, Z., Wu, Y., Bhattacharya, A., Mester, R., Belbin, G.M., Buyske, S., Conti, D.V., Darst, B.F., et al. (2023). Causal effects on complex traits are similar for common

variants across segments of different continental ancestries within admixed individuals. Nat. Genet. *55*, 549–558.
10. Nédélec, Y., Sanz, J., Baharian, G., Szpiech, Z.A., Pacis, A., Dumaine, A., Grenier, J.-C., Freiman, A., Sams, A.J., Hebert, S., et al. (2016). Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Pathogens. Cell *167*, 657–669.e21.
11. Quach, H., Rotival, M., Pothlichet, J., Loh, Y.-H.E., Dannemann, M., Zidane, N., Laval, G., Patin, E., Harmant, C., Lopez, M., et al. (2016). Genetic Adaptation and Neandertal Admixture Shaped the Immune System of Human Populations. Cell *167*, 643–656.e17.
12. Idaghdour, Y., Czika, W., Shianna, K.V., Lee, S.H., Visscher, P.M., Martin, H.C., Miclaus, K., Jadallah, S.J., Goldstein, D.B., Wolfinger, R.D., and Gibson, G. (2010). Geographical genomics of human leukocyte gene expression variation in southern Morocco. Nat. Genet. *42*, 62–67.
13. Lappalainen, T., Sammeth, M., Friedländer, M.R., t Hoen, P.A.C., Monlong, J., Rivas, M.A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506–511.
14. Martin, A.R., Costa, H.A., Lappalainen, T., Henn, B.M., Kidd, J.M., Yee, M.-C., Grubert, F., Cann, H.M., Snyder, M., Montgomery, S.B., and Bustamante, C.D. (2014). Transcriptome sequencing from diverse human populations reveals differentiated regulatory architecture. PLoS Genet. *10*, e1004549.
15. Hughes, D.A., Kircher, M., He, Z., Guo, S., Fairbrother, G.L., Moreno, C.S., Khaitovich, P., and Stoneking, M. (2015). Evaluating intra- and inter-individual variation in the human placental transcriptome. Genome Biol. *16*, 54.
16. Melé, M., Ferreira, P.G., Reverter, F., DeLuca, D.S., Monlong, J., Sammeth, M., Young, T.R., Goldmann, J.M., Pervouchine, D.D., Sullivan, T.J., et al. (2015). Human genomics. The human transcriptome across tissues and individuals. Science *348*, 660–665.
17. Randolph, H.E., Fiege, J.K., Thielen, B.K., Mickelson, C.K., Shiratori, M., Barroso-Batista, J., Langlois, R.A., and Barreiro, L.B. (2021). Genetic ancestry effects on the response to viral infection are pervasive but cell type specific. Science *374*, 1127–1133.
18. Aquino, Y., Bisiaux, A., Li, Z., O'Neill, M., Mendoza-Revilla, J., Merkling, S.H., Kerner, G., Hasan, M., Libri, V., Bondet, V., et al. (2023). Dissecting human population variation in single-cell responses to SARS-CoV-2. Nature *621*, 120–128.
19. Stranger, B.E., Montgomery, S.B., Dimas, A.S., Parts, L., Stegle, O., Ingle, C.E., Sekowska, M., Smith, G.D., Evans, D., Gutierrez-Arcelus, M., et al. (2012). Patterns of cis regulatory variation in diverse human populations. PLoS Genet. *8*, e1002639.
20. Mogil, L.S., Andaleon, A., Badalamenti, A., Dickinson, S.P., Guo, X., Rotter, J.I., Johnson, W.C., Im, H.K., Liu, Y., and Wheeler, H.E. (2018). Genetic architecture of gene expression traits across diverse populations. PLoS Genet. *14*, e1007586.
21. Shang, L., Smith, J.A., Zhao, W., Kho, M., Turner, S.T., Mosley, T.H., Kardia, S.L.R., and Zhou, X. (2020). Genetic Architecture of Gene Expression in European and African Americans: An eQTL Mapping Study in GENOA. Am. J. Hum. Genet. *106*, 496–512.
22. Fagny, M., Patin, E., MacIsaac, J.L., Rotival, M., Flutre, T., Jones, M.J., Siddle, K.J., Quach, H., Harmant, C., McEwen, L.M., et al. (2015). The epigenomic landscape of African rainforest hunter-gatherers and farmers. Nat. Commun. *6*, 10047.
23. Carja, O., MacIsaac, J.L., Mah, S.M., Henn, B.M., Kobor, M.S., Feldman, M.W., and Fraser, H.B. (2017). Worldwide patterns of human epigenetic variation. Nat. Ecol. Evol. *1*, 1577–1583.
24. Hatton, A.A., Cheng, F.-F., Lin, T., Shen, R.-J., Chen, J., Zheng, Z., Qu, J., Lyu, F., Harris, S.E., Cox, S.R., et al. (2024). Genetic control of DNA methylation is largely shared across European and East Asian populations. Nat. Commun. *15*, 2713–2812.
25. Kasowski, M., Kyriazopoulou-Panagiotopoulou, S., Grubert, F., Zaugg, J.B., Kundaje, A., Liu, Y., Boyle, A.P., Zhang, Q.C., Zakharia, F., Spacek, D.V., et al. (2013). Extensive variation in chromatin states across humans. Science *342*, 750–752.
26. Shlyueva, D., Stampfel, G., and Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. Nat. Rev. Genet. *15*, 272–286.
27. Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.
28. Perez, R.K., Gordon, M.G., Subramaniam, M., Kim, M.C., Hartoularos, G.C., Targ, S., Sun, Y., Ogorodnikov, A., Bueno, R., Lu, A., et al. (2022). Single-cell RNA-seq reveals cell type-specific molecular and genetic associations to lupus. Science *376*, eabf1970.
29. Yazar, S., Alquicira-Hernandez, J., Wing, K., Senabouth, A., Gordon, M.G., Andersen, S., Lu, Q., Rowson, A., Taylor, T.R.P., Clarke, L., et al. (2022). Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. Science *376*, eabf3041.
30. Gazal, S., Weissbrod, O., Hormozdiari, F., Dey, K.K., Nasser, J., Jagadeesh, K.A., Weiner, D.J., Shi, H., Fulco, C.P., O'Connor, L.J., et al. (2022). Combining SNP-to-gene linking strategies to identify disease genes and assess disease omnigenicity. Nat. Genet. *54*, 827–836.
31. Oliva, M., Muñoz-Aguirre, M., Kim-Hellmuth, S., Wucher, V., Gewirtz, A.D.H., Cotter, D.J., Parsana, P., Kasela, S., Balliu, B., Viñuela, A., et al. (2020). The impact of sex on gene expression across human tissues. Science *369*.
32. Young, M.D., Wakefield, M.J., Smyth, G.K., and Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol. *11*, R14.
33. Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Series B Stat. Methodol. *57*, 289–300.
34. Ishigaki, K., Kochi, Y., Suzuki, A., Tsuchida, Y., Tsuchiya, H., Sumitomo, S., Yamaguchi, K., Nagafuchi, Y., Nakachi, S., Kato, R., et al. (2017). Polygenic burdens on cell-specific pathways underlie the risk of rheumatoid arthritis. Nat. Genet. *49*, 1120–1125.
35. Gazal, S., Sahbatou, M., Babron, M.-C., Génin, E., and Leutenegger, A.-L. (2015). High level of inbreeding in final phase of 1000 Genomes Project. Sci. Rep. *5*, 17453.
36. (2015). The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature *526*, 68–74.
37. Bhatia, G., Patterson, N., Sankararaman, S., and Price, A.L. (2013). Estimating and interpreting FST: the impact of rare variants. Genome Res. *23*, 1514–1521.
38. Finucane, H.K., Reshef, Y.A., Anttila, V., Slowikowski, K., Gusev, A., Byrnes, A., Gazal, S., Loh, P.-R., Lareau, C., Shoresh, N., et al. (2018). Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. Nat. Genet. *50*, 621–629.

39. Gazal, S., Finucane, H.K., Furlotte, N.A., Loh, P.-R., Palamara, P.F., Liu, X., Schoech, A., Bulik-Sullivan, B., Neale, B.M., Gusev, A., and Price, A.L. (2017). Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. Nat. Genet. *49*, 1421–1427.
40. Bernabeu, E., Canela-Xandri, O., Rawlik, K., Talenti, A., Prendergast, J., and Tenesa, A. (2021). Sex differences in genetic architecture in the UK Biobank. Nat. Genet. *53*, 1283–1289.
41. Bulik-Sullivan, B., Finucane, H.K., Anttila, V., Gusev, A., Day, F.R., Loh, P.-R., ReproGen Consortium; Psychiatric Genomics Consortium; and Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3, and Duncan, L., et al. (2015). An atlas of genetic correlations across human diseases and traits. Nat. Genet. *47*, 1236–1241.
42. Aguet, F., Anand, S., Ardlie, K.G., Gabriel, S., Getz, G.A., Graubert, A., Hadley, K., Handsaker, R.E., Huang, K.H., Kashin, S., et al. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318–1330.
43. Chen, M.-H., Raffield, L.M., Mousas, A., Sakaue, S., Huffman, J.E., Moscati, A., Trivedi, B., Jiang, T., Akbari, P., Vuckovic, D., et al. (2020). Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations. Cell *182*, 1198–1213.e14.
44. Opferman, J.T., Letai, A., Beard, C., Korsmeyer, S.J., Sorcinelli, M.D., and Ong, C.C. (2003). Development and maintenance of B and T lymphocytes requires antiapoptotic MCL-1. Nature *426*, 671–676.
45. Vikstrom, I., Carotta, S., Lüthje, K., Peperzak, V., Jost, P.J., Glaser, S., Busslinger, M., Bouillet, P., Strasser, A., Nutt, S.L., et al. (2010). Mcl-1 is essential for germinal center formation and B cell memory. Science *330*.
46. Vikström, I.B., Slomp, A., Carrington, E.M., Moesbergen, L.M., Chang, C., Kelly, G.L., Glaser, S.P., Jansen, J.H.M., Leusen, J.H.W., Strasser, A., et al. (2016). MCL-1 is required throughout B-cell development and its loss sensitizes specific B-cell subsets to inhibition of BCL-2 or BCL-XL. Cell Death Dis. *7*, e2345.
47. Kanai, M., Akiyama, M., Takahashi, A., Matoba, N., Momozawa, Y., Ikeda, M., Iwata, N., Ikegawa, S., Hirata, M., Matsuda, K., et al. (2018). Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. Nat. Genet. *50*, 390–400.
48. Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A.P., and Price, A.L. (2018). Mixed-model association for biobank-scale datasets. Nat. Genet. *50*, 906–908.
49. Finucane, H.K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K., et al. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. *47*, 1228–1235.
50. Jagadeesh, K.A., Dey, K.K., Montoro, D.T., Mohan, R., Gazal, S., Engreitz, J.M., Xavier, R.J., Price, A.L., and Regev, A. (2022). Identifying disease-critical cell types and cellular processes by integrating single-cell RNA-sequencing and human genetics. Nat. Genet. *54*, 1479–1492.
51. Aygün, N., Liang, D., Crouse, W.L., Keele, G.R., Love, M.I., and Stein, J.L. (2023). Inferring cell-type-specific causal gene regulatory networks during human neurogenesis. Genome Biol. *24*, 130.
52. Ma, Y., Deng, C., Zhou, Y., Zhang, Y., Qiu, F., Jiang, D., Zheng, G., Li, J., Shuai, J., Zhang, Y., et al. (2023). Polygenic regression uncovers trait-relevant cellular contexts through pathway activation transformation of single-cell RNA sequencing data. Cell Genomics *0*.
53. Ongen, H., Brown, A.A., Delaneau, O., Panousis, N.I., Nica, A.C., GTEx Consortium, and Dermitzakis, E.T. (2017). Estimating the causal tissues for complex traits and diseases. Nat. Genet. *49*, 1676–1683.
54. Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B.E., Liu, X.S., and Raychaudhuri, S. (2013). Chromatin marks identify critical cell types for fine mapping complex trait variants. Nat. Genet. *45*, 124–130.
55. Morris, A.P. (2011). Transethnic meta-analysis of genomewide association studies. Genet. Epidemiol. *35*, 809–822.
56. Kichaev, G., and Pasaniuc, B. (2015). Leveraging Functional-Annotation Data in Trans-ethnic Fine-Mapping Studies. Am. J. Hum. Genet. *97*, 260–271.
57. Lu, Z., Gopalan, S., Yuan, D., Conti, D.V., Pasaniuc, B., Gusev, A., and Mancuso, N. (2022). Multi-ancestry fine-mapping improves precision to identify causal genes in transcriptome-wide association studies. Am. J. Hum. Genet. *109*, 1388–1404.
58. Kamariza, M., Crawford, L., Jones, D., and Finucane, H. (2021). Misuse of the term "trans-ethnic" in genomics research. Nat. Genet. *53*, 1520–1521.
59. Lewis, A.C.F., Molina, S.J., Appelbaum, P.S., Dauda, B., Di Rienzo, A., Fuentes, A., Fullerton, S.M., Garrison, N.A., Ghosh, N., Hammonds, E.M., et al. (2022). Getting genetic ancestry right for science and society. Science *376*, 250–252.
60. Ding, Y., Hou, K., Xu, Z., Pimplaskar, A., Petter, E., Boulier, K., Privé, F., Vilhjálmsson, B.J., Olde Loohuis, L.M., and Pasaniuc, B. (2023). Polygenic scoring accuracy varies across the genetic ancestry continuum. Nature *618*, 774–781.