


# Gene expression in African Americans, Puerto Ricans and Mexican Americans reveals ancestry-specific patterns of genetic architecture

Received: 18 August 2021

Accepted: 21 March 2023

Published online: 25 May 2023

 Check for updates

Linda Kachuri <sup>1,2,19</sup>, Angel C. Y. Mak <sup>3,19</sup>, Donglei Hu <sup>3</sup>, Celeste Eng<sup>3</sup>, Scott Huntsman<sup>3</sup>, Jennifer R. Elhawary<sup>3</sup>, Namrata Gupta<sup>4</sup>, Stacey Gabriel<sup>4</sup>, Shujie Xiao<sup>5</sup>, Kevin L. Keys <sup>3,6</sup>, Akinyemi Oni-Orisan <sup>7,8,9</sup>, José R. Rodríguez-Santana<sup>10</sup>, Michael A. LeNoir<sup>11</sup>, Luisa N. Borrell <sup>12</sup>, Noah A. Zaitlen<sup>13,14</sup>, L. Keoki Williams<sup>5,15</sup>, Christopher R. Gignoux <sup>16,17,20</sup> , Esteban González Burchard <sup>3,8,20</sup>  & Elad Ziv <sup>3,9,18,20</sup> 

We explored ancestry-related differences in the genetic architecture of whole-blood gene expression using whole-genome and RNA sequencing data from 2,733 African Americans, Puerto Ricans and Mexican Americans. We found that heritability of gene expression significantly increased with greater proportions of African genetic ancestry and decreased with higher proportions of Indigenous American ancestry, reflecting the relationship between heterozygosity and genetic variance. Among heritable protein-coding genes, the prevalence of ancestry-specific expression quantitative trait loci (anc-eQTLs) was 30% in African ancestry and 8% for Indigenous American ancestry segments. Most anc-eQTLs (89%) were driven by population differences in allele frequency. Transcriptome-wide association analyses of multi-ancestry summary statistics for 28 traits identified 79% more gene–trait associations using transcriptome prediction models trained in our admixed population than models trained using data from the Genotype-Tissue Expression project. Our study highlights the importance of measuring gene expression across large and ancestrally diverse populations for enabling new discoveries and reducing disparities.

Gene expression has been studied extensively as a trait affected by genetic variation in humans<sup>1</sup>. Expression quantitative trait loci (eQTLs) have been identified in most genes<sup>2–4</sup> and extensive analyses across multiple tissues have demonstrated both tissue-specific and shared eQTLs<sup>2</sup>. Genome-wide association studies (GWASs) tend to identify loci that are enriched for eQTLs<sup>5</sup>. Colocalization of eQTLs with GWASs is an important element of identifying causal genes and investigating the biology underlying genetic susceptibility to disease<sup>6</sup>. Transcriptome-wide association

studies (TWASs) have also been developed to leverage eQTL data by imputing transcriptomic profiles in external datasets or GWAS summary statistics, and have identified additional trait-associated genes<sup>7,8</sup>.

Despite the discovery of thousands of loci that influence hundreds of complex human traits, the underrepresentation of admixed and non-European ancestry individuals in GWAS<sup>9,10</sup> and multi-omic studies remains an obstacle for applying these approaches to diverse populations. Gene expression prediction models trained in predominantly

A full list of affiliations appears at the end of the paper. ✉ e-mail: [Chris.Gignoux@cuanschutz.edu](mailto:Chris.Gignoux@cuanschutz.edu); [Esteban.Burchard@ucsf.edu](mailto:Esteban.Burchard@ucsf.edu); [Elad.Ziv@ucsf.edu](mailto:Elad.Ziv@ucsf.edu)

European ancestry reference datasets, such as the Genotype-Tissue Expression (GTEx) project<sup>2</sup>, have substantially lower accuracy to predict gene expression levels when applied to populations of non-European ancestry<sup>3,11,12</sup>. The importance of aligning ancestry between training and testing populations is also reflected by the poor cross-population performance of other multi-single nucleotide polymorphism (SNP) prediction models, such as polygenic risk scores<sup>13–15</sup>.

To address this gap, we leveraged whole-genome sequencing (WGS) and RNA sequencing (RNA-seq) data from 2,733 participants from the Genes-Environments and Admixture in Latino Asthmatics (GALA II) study and the Study of African Americans, Asthma, Genes, and Environments (SAGE) to characterize the genetic architecture of whole-blood eQTLs. The diversity within the GALA II/SAGE population enabled us to evaluate how genetic ancestry relates to the heritability of gene expression, and to systematically quantify the prevalence of ancestry-specific eQTLs. Lastly, we developed a powerful set of TWAS models from these datasets to facilitate genetic association analyses in multi-ancestry populations.

## Results

We analyzed data from a total of 2,733 participants from the GALA II<sup>16</sup> and SAGE<sup>17</sup> asthma case-control studies who self-identified as African American (AA;  $n = 757$ ), Puerto Rican (PR;  $n = 893$ ), Mexican American (MX;  $n = 784$ ) or other Latino American (LA;  $n = 299$ ) (Table 1 and Supplementary Table 1). The median age of the participants varied from 13.2 (PR) to 16.0 years (AA). Genome-wide, or global, genetic ancestry proportions were estimated for all participants (Fig. 1). The median global African ancestry was highest in AA (82.6%), followed by PR (19.7%), and lowest in MX (3.5%).

### Heritability of gene expression in admixed populations

We compared the heritability ( $h^2$ ) and genetic variance ( $V_G$ ) of whole-blood gene expression attributed to common genetic variation (minor allele frequency (MAF)  $\geq 0.01$ ) within the *cis*-region across self-identified race/ethnicity groups and subpopulations defined based on genetic ancestry (see Methods). Across 17,657 genes, *cis*-heritability (Fig. 2a) was significantly higher in AA (median  $h^2 = 0.097$ ) compared with PR ( $h^2 = 0.072$ ;  $P_{\text{Wilcoxon}} = 2.2 \times 10^{-50}$ ) and MX participants ( $h^2 = 0.059$ ;  $P = 3.3 \times 10^{-134}$ ) and in PR compared with MX participants ( $P_{\text{Wilcoxon}} = 2.2 \times 10^{-25}$ ) (Supplementary Table 2). Genetic variance (Fig. 2b) of whole-blood transcript levels in AA participants (median  $V_G = 0.022$ ) was higher than in PR participants ( $V_G = 0.018$ ;  $P_{\text{Wilcoxon}} = 4.0 \times 10^{-19}$ ) and in MX participants ( $V_G = 0.013$ ;  $P_{\text{Wilcoxon}} = 5.6 \times 10^{-135}$ ). The results remained unchanged when the sample size was fixed to  $n = 600$  in all populations (Extended Data Fig. 1).

Next, we compared the distribution of  $h^2$  (Fig. 2c) and  $V_G$  (Fig. 2d) between participants grouped based on proportions of global genetic ancestry (Supplementary Table 3 and Supplementary Fig. 1). Among participants with >50% African ancestry (AFR<sub>high</sub>;  $n = 721$ ), *cis*-heritability ( $h^2 = 0.098$ ) and genetic variance ( $V_G = 0.022$ ) were higher than in  $n = 1,011$  participants with <10% global African ancestry (AFR<sub>low</sub>;  $h^2 = 0.060$  ( $P_{\text{Wilcoxon}} = 9.6 \times 10^{-126}$ ) and  $V_G = 0.013$  ( $P_{\text{Wilcoxon}} = 7.6 \times 10^{-106}$ )). Among individuals with >50% Indigenous American (IAM) ancestry (IAM<sub>high</sub>;  $n = 610$ ), *cis*-heritability ( $h^2 = 0.059$ ) and genetic variance ( $V_G = 0.012$ ) were lower than in participants with <10% IAM ancestry (IAM<sub>low</sub>;  $h^2 = 0.084$  ( $P = 3.1 \times 10^{-103}$ ) and  $V_G = 0.020$  ( $P_{\text{Wilcoxon}} = 3.1 \times 10^{-158}$ )). The results remained consistent when partitioning  $h^2$  and  $V_G$  by coarse MAF bins, with larger differences in  $h^2$  and  $V_G$  among  $0.01 \leq \text{MAF} \leq 0.10$  variants (Extended Data Fig. 2).

We also investigated the impact of ancestry at the locus level, defined as the number of alleles (zero, one or two) derived from each ancestral population at the transcription start site. Heritability was higher in individuals with homozygous local African ancestry (AFR/AFR) compared with AFR/European (EUR) ancestry ( $h^2 = 0.096$  versus  $h^2 = 0.084$ , respectively;  $P_{\text{Wilcoxon}} = 1.4 \times 10^{-14}$ ) and lower in participants

**Table 1 | Study participants**

	Self-identified race/ethnicity				Pooled
	AA	PR	MX	LA	
Sex					
Number (%) female	405 (53.5)	451 (50.5)	427 (54.5)	158 (52.8)	1,441 (52.7)
Asthma status					
Number (%) of cases	433 (57.2)	549 (61.5)	351 (44.8)	156 (52.2)	1,489 (54.5)
Recruitment center (number (%))					
San Francisco Bay Area	757 (100)	0 (0)	348 (44.4)	109 (36.5)	1,214 (44.4)
Chicago	0 (0)	31 (3.5)	247 (31.5)	52 (17.4)	330 (12.1)
Puerto Rico	0 (0)	837 (93.7)	0 (0)	8 (2.7)	845 (30.9)
New York City	0 (0)	22 (2.5)	36 (4.6)	86 (28.8)	144 (5.3)
Houston	0 (0)	3 (0.3)	153 (19.5)	44 (14.7)	200 (7.3)
Median (IQR) age (years)	16.0 (6.6)	13.2 (4.8)	13.8 (6.5)	13.7 (5.7)	14.0 (6.3)
Median (IQR) genetic ancestry (%)					
African	82.6 (9.4)	19.7 (13.3)	3.5 (2.7)	8.3 (14.8)	17.5 (61.8)
Indigenous American	0.3 (0.9)	9.9 (3.6)	55.3 (23.2)	42.3 (43.2)	10.7 (45.2)
European	16.5 (9.5)	69.5 (13.6)	40.3 (21.9)	45.9 (20.8)	44.2 (43.8)
Total	757	893	784	299	2,733

Demographic characteristics of 2,733 participants from GALA II and SAGE included in the present analysis. IQR, interquartile range.

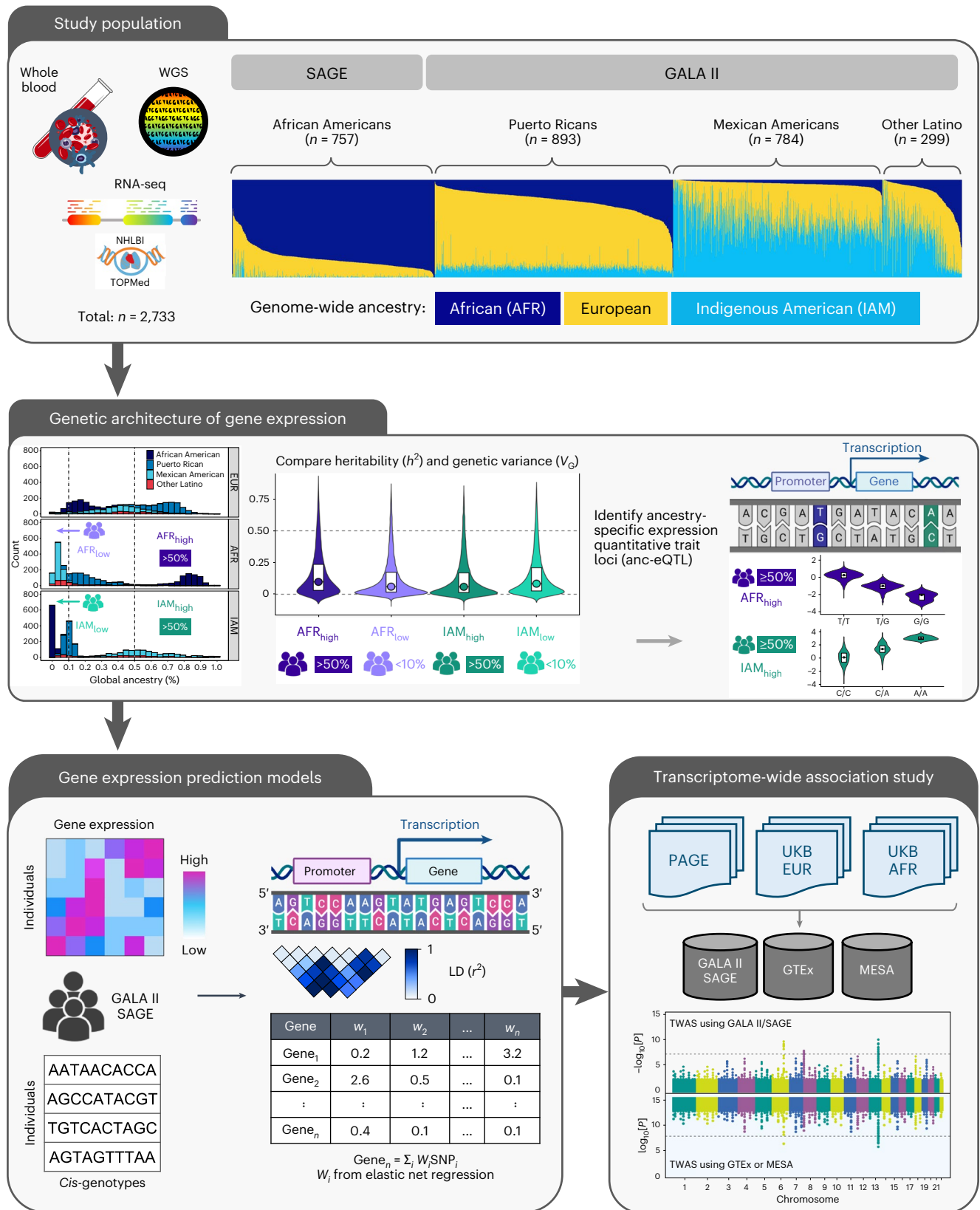
with homozygous Indigenous American ancestry (IAM/IAM) compared with IAM/EUR ancestry ( $h^2 = 0.055$  versus 0.064, respectively;  $P = 1.6 \times 10^{-7}$ ) (Fig. 2e). Differences in  $V_G$  by local ancestry were statistically significant for AFR/AFR versus AFR/EUR ( $P_{\text{Wilcoxon}} = 2.0 \times 10^{-7}$ ) and IAM/IAM versus IAM/EUR ( $P_{\text{Wilcoxon}} = 1.6 \times 10^{-8}$ ) (Fig. 2f). The results were also consistent for  $V_G$  comparisons within self-identified race/ethnicity groups (Supplementary Table 4).

We calculated heritability using linkage disequilibrium adjusted kinships (LDAKs). This method assumes that SNP-specific variance is inversely proportional not only to the MAF, but also to linkage disequilibrium tagging<sup>18</sup>. Estimates obtained using the LDAK-Thin model and genome-wide complex trait analysis (GCTA) were nearly identical across populations based on self-identified race/ethnicity ( $h^2 = 0.094$  for AA, 0.071 for PR and 0.059 for MX) and genetic ancestry ( $h^2 = 0.104$  for AFR<sub>high</sub>, 0.066 for AFR<sub>low</sub>, 0.062 for IAM<sub>high</sub> and 0.093 for IAM<sub>low</sub>; Supplementary Table 5).

Lastly, we tabulated the number of heritable genes for which global and/or local ancestry was significantly associated (false discovery rate (FDR) < 0.05) with transcript levels (Supplementary Fig. 2 and Supplementary Table 6). Global AFR ancestry was associated with the expression of 326 (2.4%) and 589 (4.5%) heritable genes in AA and PR, respectively. Associations with local, but not global, AFR ancestry were more common (8.9% in AA and 10.9% in PR). Local IAM ancestry was associated with the expression of 9.8% of genes in MX, compared with 2.8% for global IAM ancestry.

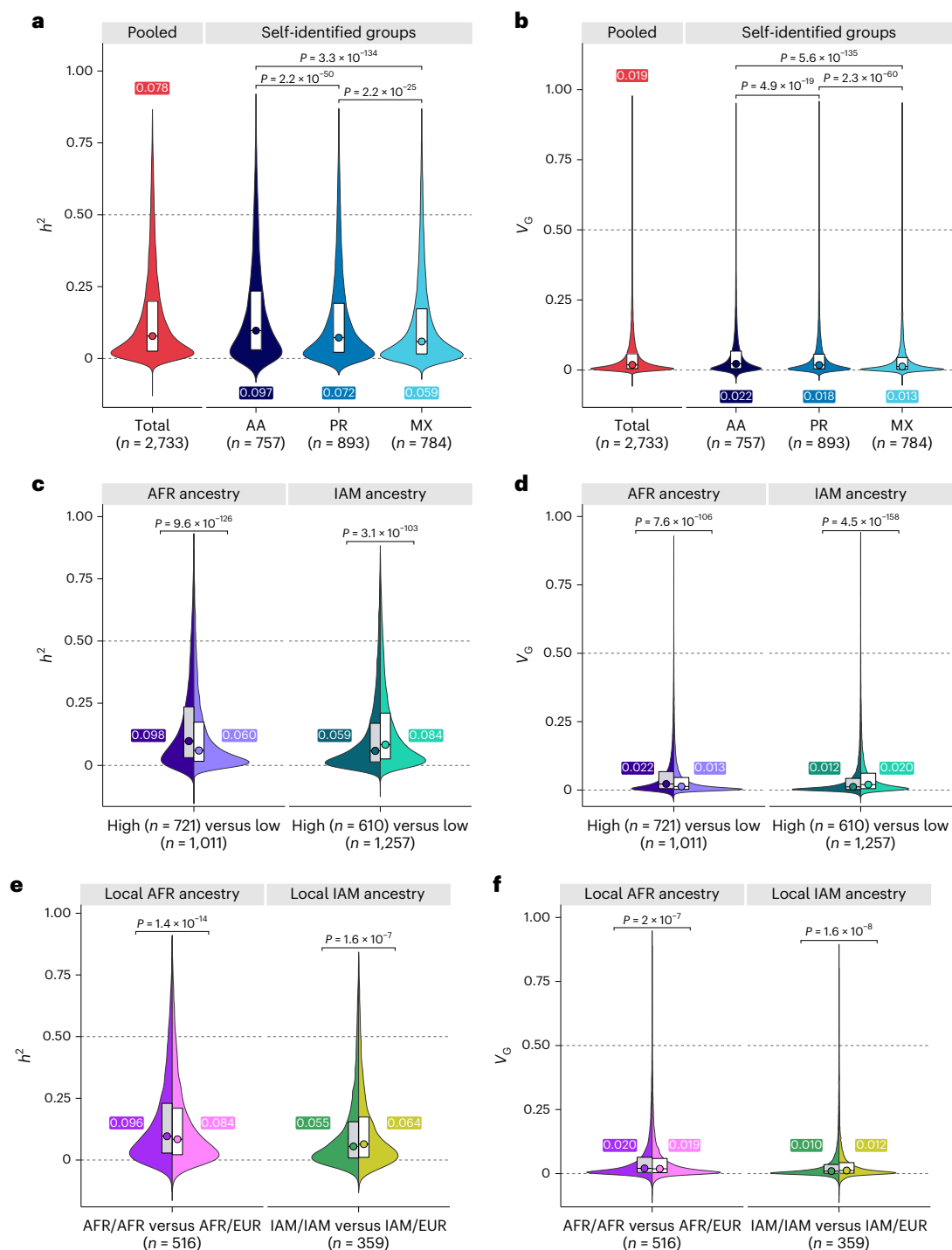
### Assessment of ancestry-specific eQTLs

To understand the control of gene expression at a more granular level, we performed *cis*-eQTL analysis. A total of 19,567 genes with at least one



**Fig. 1 | Study overview.** This study included TOPMed WGS and whole transcriptome data generated from whole-blood samples of SAGE AA and GALA II Latino individuals ( $n = 2,733$ ). We compared elements of the genetic architecture of gene expression, such as *cis*-heritability and genetic variance, across participant groups defined based on self-identified race/ethnicity and genetic ancestry.

We performed eQTL mapping and identified eQTLs that were specific to AFR or IAM ancestry. Finally, we developed genetic prediction models of whole-blood transcriptomes and performed comparative TWASs using GWAS summary statistics generated from the PAGE study and the UKB. Figure created with [BioRender.com](https://www.biorender.com).



**Fig. 2 | Cross-population comparison of *cis*-heritability ( $h^2$ ) and genetic variance ( $V_G$ ) of whole-blood transcript levels. **a,b**, Violin plots showing the distribution of  $h^2$  (**a**) and  $V_G$  (**b**) for all available genes in each population. The box plots extend from the 25th to 75th percentiles and median values are annotated. The analyses were stratified by self-identified race/ethnicity and compared AA, PR and MX participants. **c,d**, Split violin plots showing the distribution of  $h^2$  (**c**) and  $V_G$  (**d**) in individuals with >50% global genetic ancestry (high; darker shading)**

versus those with <10% of the same ancestry (low; lighter shading). Analyses were conducted separately for AFR and IAM ancestry. **e,f**, The local ancestry at the transcription start site of each gene was used to compare  $h^2$  (**e**) and  $V_G$  (**f**) in participants with 100% (AFR/AFR or IAM/IAM; darker shading) versus 50% (AFR/EUR or IAM/EUR; lighter shading) local ancestry. Statistical significance in all panels was determined by Wilcoxon test and all  $P$  values are two sided.

*cis*-eQTL (eGenes) were found in the pooled sample. The largest number of eGenes was detected in AA ( $n=17,336$ ), followed by PR ( $n=16,975$ ) and MX participants ( $n=15,938$ ) (Supplementary Table 7 and Supplementary Fig. 3). The number of eGenes was similar in AFR<sub>high</sub> ( $n=17,123$ ) and AFR<sub>low</sub> ( $n=17,146$ ) groups. When the sample size was fixed to  $n=600$  for

all populations, the number of eGenes observed in AFR<sub>high</sub> ( $n=16,100$ ) was higher than in AFR<sub>low</sub> ( $n=14,344$ ). The numbers of eGenes detected in the IAM<sub>low</sub> ( $n=14,866$ ) and IAM<sub>high</sub> groups ( $n=14,419$ ) were similar. The number of linkage disequilibrium-independent ( $r^2 < 0.10$ ) *cis*-eQTLs per gene was significantly higher in the AFR<sub>high</sub> group than the AFR<sub>low</sub>

group ( $P_{\text{Wilcoxon}} = 2.7 \times 10^{-246}$ ) and lower in the IAM<sub>high</sub> group compared with the IAM<sub>low</sub> group ( $P_{\text{Wilcoxon}} = 2.8 \times 10^{-33}$ ) (Extended Data Fig. 3).

To characterize ancestry-related differences in the regulation of gene expression, we developed a framework for identifying ancestry-specific eQTLs (anc-eQTLs) (see Methods, Fig. 3 and Supplementary Tables 8 and 9). For heritable protein-coding genes, we first compared the overlap in 95% credible sets of *cis*-eQTLs identified in participants with >50% global ancestry (AFR<sub>high</sub> and IAM<sub>high</sub>) and those with <10% of the same global ancestry (AFR<sub>low</sub> and IAM<sub>low</sub>). For genes with nonoverlapping 95% credible sets, we distinguished between population differences in MAF (tier 1) and linkage disequilibrium (tier 2). For genes with overlapping 95% credible sets, eQTLs were further examined for evidence of effect size heterogeneity between ancestry groups (tier 3).

Tier 1 anc-eQTLs were only common (MAF  $\geq$  0.01) in individuals with >50% AFR or IAM ancestry and were thus considered to be the most ancestry specific. Over 28% ( $n = 2,695$ ) of genes contained at least one tier 1 AFR<sub>high</sub> anc-eQTL, while 7% ( $n = 562$ ) of genes contained a tier 1 IAM<sub>high</sub> anc-eQTL (Supplementary Table 9). A representative example of a tier 1 AFR<sub>high</sub> anc-eQTL is rs3211938 (*CD36*), which has an MAF of 0.077 in the AFR<sub>high</sub> group and an MAF of 0.002 in the AFR<sub>low</sub> group (Fig. 4a). This variant has been linked to high-density lipoprotein (HDL) cholesterol levels in several multi-ancestry GWASs that included African Americans<sup>19–21</sup>. Tier 2 anc-eQTLs, with ancestry-specific linkage disequilibrium patterning, had an MAF of 0.01 in both high (>50%) and low (<10%) global ancestry groups and were further fine-mapped using PESCA<sup>22</sup>. There were 109 genes (1.1%) that contained eQTLs with a posterior probability (PP) of being specific to AFR<sub>high</sub> of >0.80 and 33 genes (0.4%) matching the same criteria for IAM<sub>high</sub> (Supplementary Table 9). For instance, two lead eQTLs in low linkage disequilibrium were detected for *TRAPP6A* in AFR<sub>high</sub> (rs12460041) and AFR<sub>low</sub> (rs7247764) populations (Fig. 4b). These variants belonged to nonoverlapping credible sets and PESCA estimated that rs12460041 was specific to AFR<sub>high</sub> with a PP of 0.87 (Fig. 4c). Over 50% of heritable protein-coding genes ( $n = 5,058$  for AFR and 5,355 for IAM) had overlapping 95% credible sets of eQTLs between high and low ancestry groups. Among these shared signals, a small proportion of eQTLs exhibited significant effect size heterogeneity (tier 3; 2.0% for AFR<sub>high</sub> and 1.0% for IAM<sub>high</sub>). For example, the *KCNK17* eQTLs rs34247110 and rs3734618 were detected in the AFR<sub>high</sub> and AFR<sub>low</sub> groups, but with significantly different effect sizes (Cochran's  $Q$   $P$  value =  $1.8 \times 10^{-10}$ ) in each population (Fig. 4d). Consistent with tier 3 eQTLs being observed in multiple ancestries, rs34247110 has been associated with type 2 diabetes in Japanese and multi-ancestry (European, African American, Hispanic and Asian) populations<sup>23,24</sup>.

The prevalence of any tier 1, 2 or 3 anc-eQTL was 30% ( $n = 2,961$ ) for AFR ancestry and 8% ( $n = 679$ ) for IAM ancestry. Overall, 3,333 genes had anc-eQTLs for either ancestry. The remaining genes ( $n = 6,648$  for AFR and 7,836 for IAM) did not contain eQTLs with ancestry-related differences in MAF, linkage disequilibrium or effect size as outlined above. Increasing the global ancestry cut-off to >70% did not have an appreciable impact on anc-eQTLs in the AFR<sub>high</sub> group (28.1% overall and 27.3% for tier 1), but decreased the number of anc-eQTLs in the IAM<sub>high</sub> group (3.3% overall and 3.3% for tier 1), probably due to a greater reduction in sample size in this group ( $n = 212$  versus  $n = 610$ , respectively; Supplementary Table 10). Considering all protein-coding genes without filtering based on heritability ( $n = 13,535$ ), the prevalence of anc-eQTLs was 22% for the AFR<sub>high</sub> group, 5% for the IAM<sub>high</sub> group and 25% overall. The observation that anc-eQTLs were more common in participants with >50% AFR ancestry aligns with the higher  $h^2$  and  $V_G$  values in this population and a greater number of linkage disequilibrium-independent *cis*-eQTLs (Extended Data Fig. 3). Among genes with tier 1 and 2 anc-eQTLs, 83% had higher  $h^2$  estimates in the AFR<sub>high</sub> group than in the AFR<sub>low</sub> group, while this was observed for 57% of genes without any ancestry-specific eQTLs (Extended Data Fig. 4).

We detected 70 unique anc-eQTLs associated with 84 phenotypes from the NHGRI-EBI GWAS catalog<sup>25</sup>, most of which were tier 3 anc-eQTLs (59%) that mapped to blood cell traits, lipids and plasma protein levels (Supplementary Table 11). Colocalization with GWAS results from the multi-ancestry Population Architecture Using Genomics and Epidemiology (PAGE) study<sup>21</sup> identified 78 eQTL–trait pairs with strong evidence of a shared genetic signal (PP<sub>4</sub> > 0.80), 16 of which were anc-eQTLs (Supplementary Table 12). One illustrative example is rs7200153, an AFR<sub>high</sub> tier 1 anc-eQTL for the haptoglobin (*HP*) gene, which colocalized with total cholesterol (PP<sub>4</sub> = 0.997; Extended Data Fig. 5). The 95% credible set included rs7200153 (PP<sub>SNP</sub> = 0.519) and rs5471 (PP<sub>SNP</sub> = 0.481; linkage disequilibrium  $r^2 = 0.75$ ), with rs5471 probably being the true causal variant given its proximity to the *HP* promoter, stronger effect of *HP* expression and decreased transcriptional activity of rs5471-C in West African populations<sup>26–28</sup>, which is supported by previous literature<sup>20,29,30</sup>.

We also performed *trans*-eQTL analyses that largely recapitulated our *cis*-eQTL results, with a larger number of *trans*-eGenes and independent (linkage disequilibrium  $r^2 < 0.10$ ) *trans*-eQTLs in populations with greater levels of AFR ancestry and lower levels of IAM ancestry (see Methods and Supplementary Table 13).

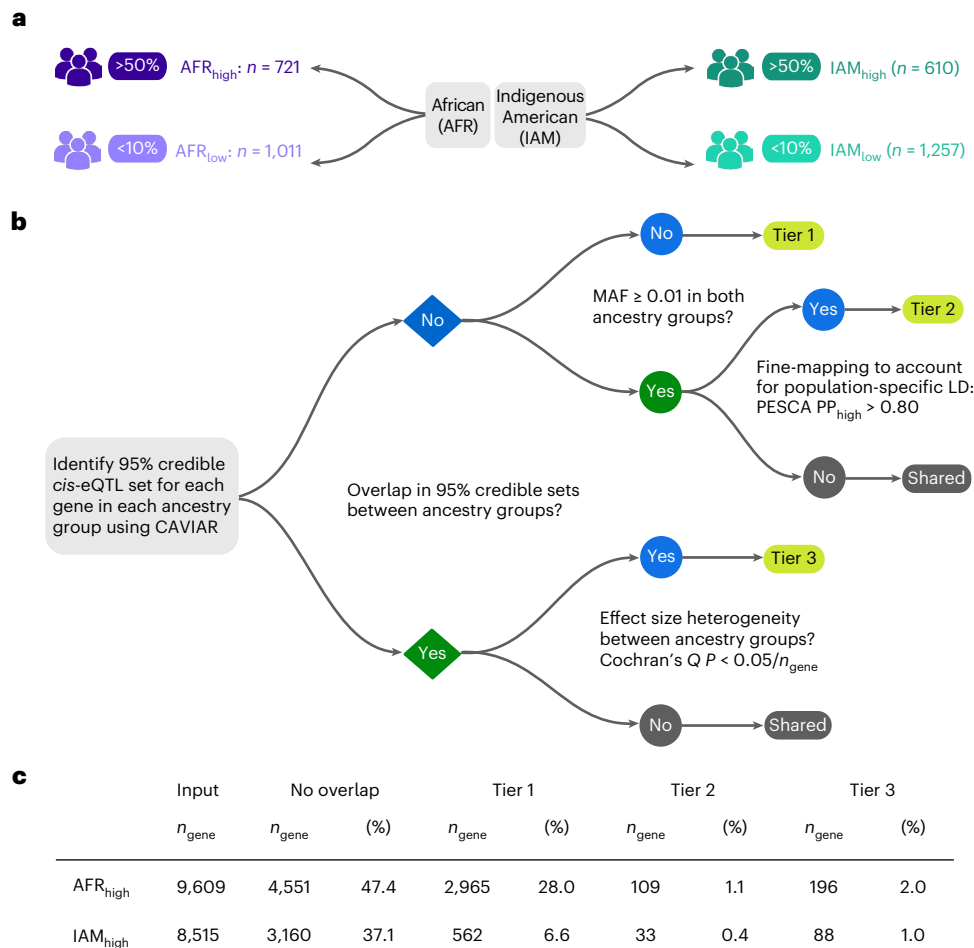
### Performance of TWAS models trained in admixed populations

Following the PrediXcan approach<sup>7</sup>, we developed gene expression imputation models from the pooled GALA II and SAGE population ( $n = 2,733$ ) for 11,830 heritable genes with a mean cross-validation of 0.157 (Supplementary Table 13 and Supplementary Fig. 4). We also generated population-specific models for African Americans (10,090 genes; cross-validation  $r^2 = 0.180$ ), Puerto Ricans (9,611 genes; cross-validation  $r^2 = 0.163$ ) and Mexican Americans (9,084 genes; cross-validation  $r^2 = 0.167$ ). Adjusting for local ancestry did not improve predictive performance (Supplementary Table 14).

We validated GALA II/SAGE TWAS models and compared with GTEx v8 in the Study of Asthma Phenotypes and Pharmacogenomic Interactions by Race-Ethnicity (SAPPHIRE)<sup>31</sup>—a study of 598 African American adults (Supplementary Fig. 5). The prediction accuracy was proportional to the degree of alignment in ancestry between the training and testing study samples. Across 5,254 genes with models available in all studies, the median Pearson's correlation between genetically predicted and observed transcript levels was highest for pooled ( $r = 0.086$ ) and AA ( $r = 0.083$ ) models and lowest for GTEx ( $r = 0.049$ ).

To evaluate the performance of the GALA II/SAGE TWAS models for gene discovery in admixed populations, we applied them to GWAS summary statistics for 28 traits from the PAGE study<sup>21</sup> and compared them with TWASs using GTEx v8 (refs. 2,7) and the Multi-Ethnic Study of Atherosclerosis (MESA)<sup>3</sup>. GTEx v8 whole-blood models are based on 670 participants of predominantly European ancestry (85%)<sup>2</sup>, while MESA models impute monocyte gene expression<sup>3</sup> from African American and Hispanic/Latino individuals (MESA<sub>AFHI</sub>;  $n = 585$ ). The number of genes with available TWAS models was 39–82% higher in GALA II/SAGE compared with GTEx ( $n = 7,249$ ) and MESA<sub>AFHI</sub> ( $n = 5,555$ ). Restricting to 3,143 genes shared across all three studies, the cross-validation  $r^2$  was significantly higher in GALA II/SAGE compared with GTEx ( $P_{\text{Wilcoxon}} = 4.6 \times 10^{-159}$ ) and MESA<sub>AFHI</sub> ( $P_{\text{Wilcoxon}} = 1.1 \times 10^{-64}$ ) (Fig. 5a). TWAS models generated in GALA II/SAGE AA ( $n = 757$ ) attained a higher cross-validation  $r^2$  than GTEx ( $P_{\text{Wilcoxon}} = 2.2 \times 10^{-103}$ ), which had a comparable training sample size (Fig. 5b).

TWASs using GALA II/SAGE models across 28 PAGE traits identified a larger number of significant gene–trait pairs ( $n = 380$ ; FDR < 0.05) than MESA<sub>AFHI</sub> ( $n = 303$ ) and GTEx ( $n = 268$ ), with only 35 significant gene–trait pairs detected in all three analyses (Fig. 5c). GALA II/SAGE models yielded a larger number of associated genes than MESA in 80% of analyses (binomial test;  $P = 0.012$ ) and a larger number than GTEx in 79% of analyses (binomial test;  $P = 0.019$ ). Of the 330 genes with an FDR of <0.05 in GALA II/SAGE, 143 (43%) were not present in GTEx and



**Fig. 3 | Framework for the classification of anc-eQTLs. a**, eQTL mapping analyses were conducted in participant groups derived based on genetic ancestry. Associations identified in participants with >50% global African ancestry (AFR<sub>high</sub>;  $n = 721$ ) were compared with eQTLs detected in participants with <10% African ancestry (AFR<sub>low</sub>;  $n = 1,011$ ), and eQTLs identified in participants with >50% global Indigenous American ancestry (IAM<sub>high</sub>;  $n = 610$ ) were compared with associations detected in participants with <10% Indigenous

American ancestry (IAM<sub>low</sub>;  $n = 1,257$ ). **b**, Decision tree for the classification of anc-eQTLs. Tier 1 is based on MAF differences and represents the most ancestry-specific class. Tier 2 anc-eQTLs were identified using additional fine-mapping with PESCA<sup>22</sup>. Tier 3 anc-eQTLs were detected in both ancestry groups, but with statistically significant effect size heterogeneity based on Cochran's  $Q$  test. **c**, Prevalence of anc-eQTLs in each ancestry group. Figure created with BioRender.com.

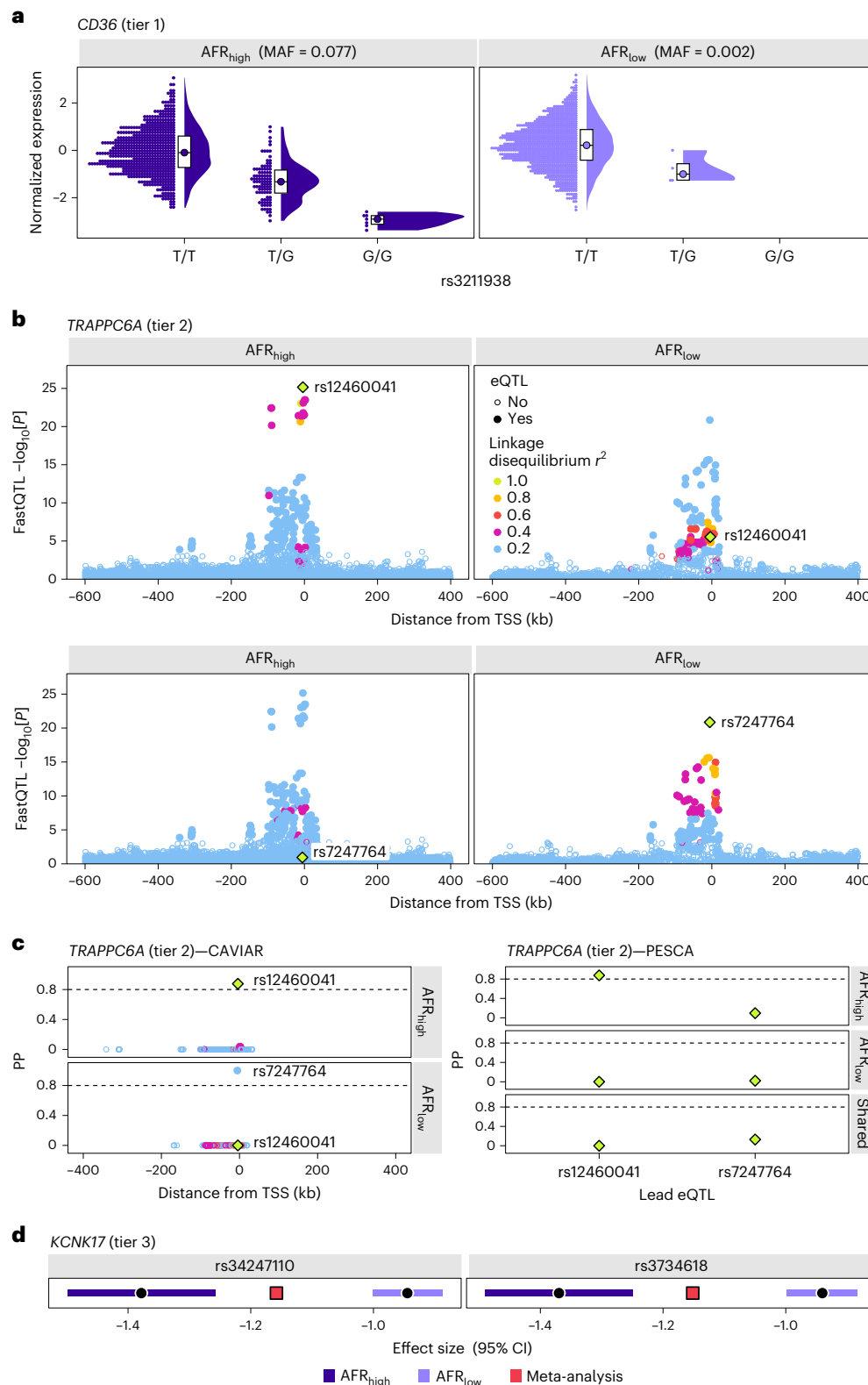
199 (60%) were not present in MESA<sub>AFHI</sub>. For genes that were significant in at least one TWAS,  $z$  scores in GALA II/SAGE were highly correlated with GTEx ( $r = 0.74$ ;  $P = 3.5 \times 10^{-64}$ ; Fig. 5c) and MESA<sub>AFHI</sub> ( $r = 0.55$ ;  $P = 8.5 \times 10^{-27}$ ; Fig. 5d), suggesting that most genes have concordant effects even if they are not significant in both analyses.

HDL cholesterol exhibited one of the largest differences in TWAS associations (Fig. 6a), with over 60% more significant genes identified using GALA II/SAGE models ( $n = 29$ ) than GTEx predictions ( $n = 11$ ). TWAS models for several associated genes, including those with established effects on cholesterol transport and metabolism, such as *CETP*, were not available in GTEx. The top HDL-associated gene, *CD36* ( $z$  score =  $-10.52$ ;  $P_{\text{TWAS}} = 6.9 \times 10^{-26}$ ) had tier 1 AFR<sub>high</sub> anc-eQTLs (rs3211938) that were rare in European ancestry populations (MAF = 0.00013). The difference in MAF may explain why *CD36* was not detected using GTEx ( $z$  score = 0.057;  $P_{\text{TWAS}} = 0.95$ ), even though all 43 variants from the GTEx model were available in PAGE summary statistics.

Although GALA II/SAGE multi-ancestry TWAS models showed robust performance, in some cases population-specific models may be preferred. For instance, benign neutropenia is a well-described phenomenon in persons of African ancestry and is attributed to variation in the Iq23.2 region. Applying GALA II/SAGE AA models

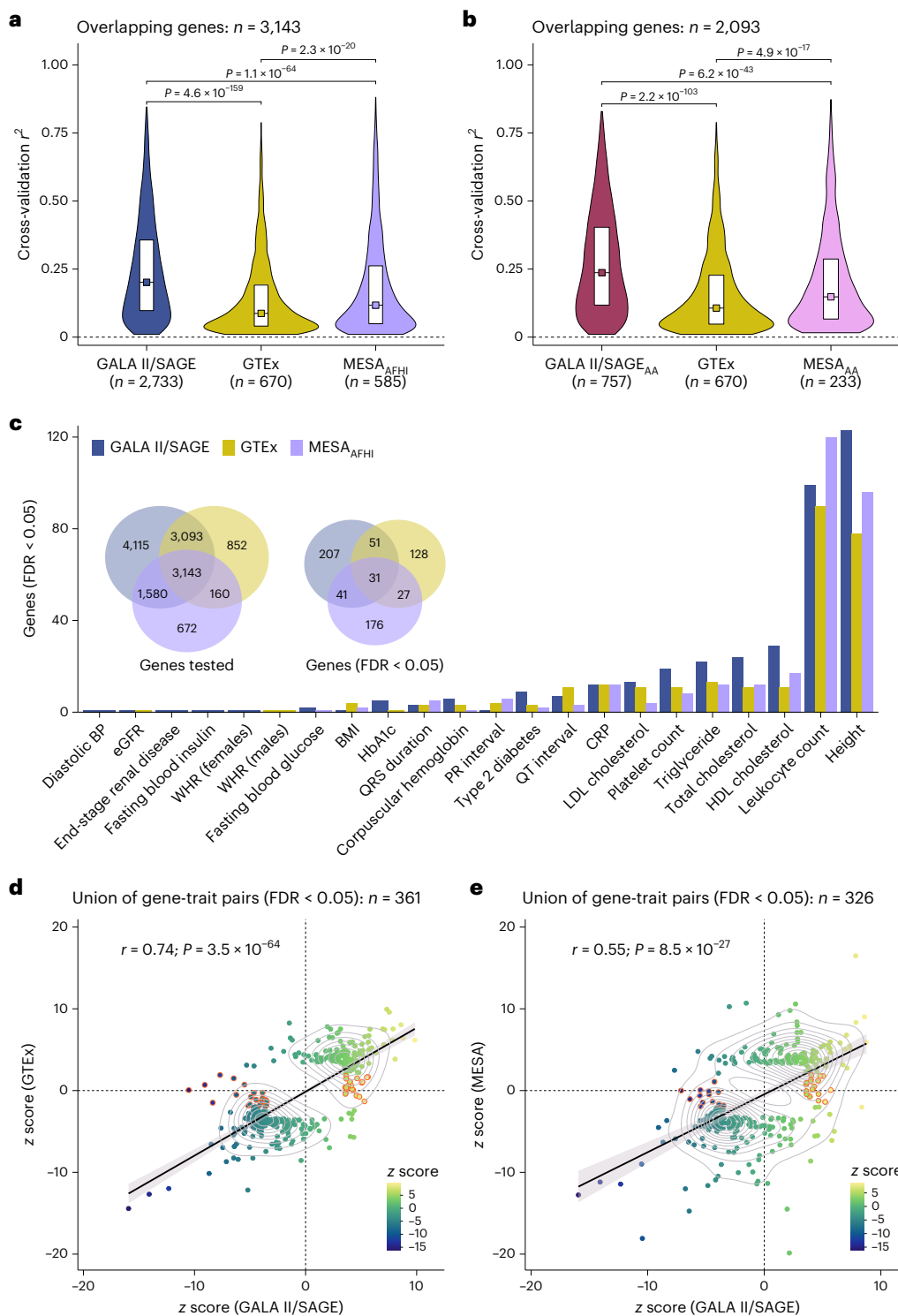
to a meta-analysis of 13,476 individuals of African ancestry<sup>32</sup> identified *ACKR1* ( $P_{\text{TWAS}} = 1.5 \times 10^{-234}$ ), the atypical chemokine receptor gene that is the basis of the Duffy blood group system (Fig. 6b). This causal gene was missed by GTEx and MESA<sub>AFHI</sub>. After conditioning on the Duffy-null rs2814778-CC genotype, no statistically significant TWAS associations remained on chromosome 1 (Supplementary Fig. 6). GALA II/SAGE AA models also detected seven genes outside of Iq23.2 that were not previously reported in GWASs of neutrophil counts: *CREB5* ( $P_{\text{TWAS}} = 1.5 \times 10^{-14}$ ), *DARS* ( $P_{\text{TWAS}} = 2.9 \times 10^{-8}$ ), *CD36* ( $P_{\text{TWAS}} = 1.1 \times 10^{-5}$ ), *PPT2* ( $P_{\text{TWAS}} = 1.3 \times 10^{-5}$ ), *SSH2* ( $P_{\text{TWAS}} = 4.7 \times 10^{-5}$ ), *TOMMS* ( $P_{\text{TWAS}} = 2.9 \times 10^{-4}$ ) and *ARF6* ( $P_{\text{TWAS}} = 3.4 \times 10^{-4}$ ).

Next, we performed TWASs of 22 blood-based biomarkers and quantitative traits using summary statistics from the UK Biobank (UKB). Ancestry-matched TWASs of UKB AFR (median GWAS  $n = 6,190$ ) identified 56 gene-trait associations (FDR < 0.05), whereas GTEx detected only five genes (Extended Data Fig. 6). TWAS  $z$  scores for associated genes were modestly correlated ( $r = 0.37$ ; 95% confidence interval (CI) =  $-0.01$ – $0.66$ ). TWASs in UKB EUR (median GWAS  $n = 400,223$ ) also illustrated the advantage of ancestry-matched analyses, but the difference was less dramatic, with a 15% decrease in the number of genes that reached an FDR of <0.05 using GALA II/SAGE AA models and strong correlation between  $z$  scores ( $r = 0.77$ ; 95% CI =  $0.76$ – $0.78$ ).



**Fig. 4 | Examples of anc-eQTLs. a**, The tier 1 anc-eQTL (*rs3211938*) for *CD36* is present in individuals with >50% African ancestry (AFR<sub>high</sub>) with an MAF of 0.077 and in individuals with <10% African ancestry (AFR<sub>low</sub>) with an MAF of 0.002. *CD36* expression by *rs3211938* genotype is shown for each individual and summarized using split violin plots and box plots, which extend from the 25th to the 75th percentile. **b**, Regional plots of FastQTL association results for *TRAPPC6A* show two independent eQTL signals in AFR<sub>high</sub> and AFR<sub>low</sub> participants, with linkage disequilibrium  $r^2 < 0.2$  in each population. Variants are colored based on linkage disequilibrium  $r^2$  with respect to the index variant, as indicated by the diamond.

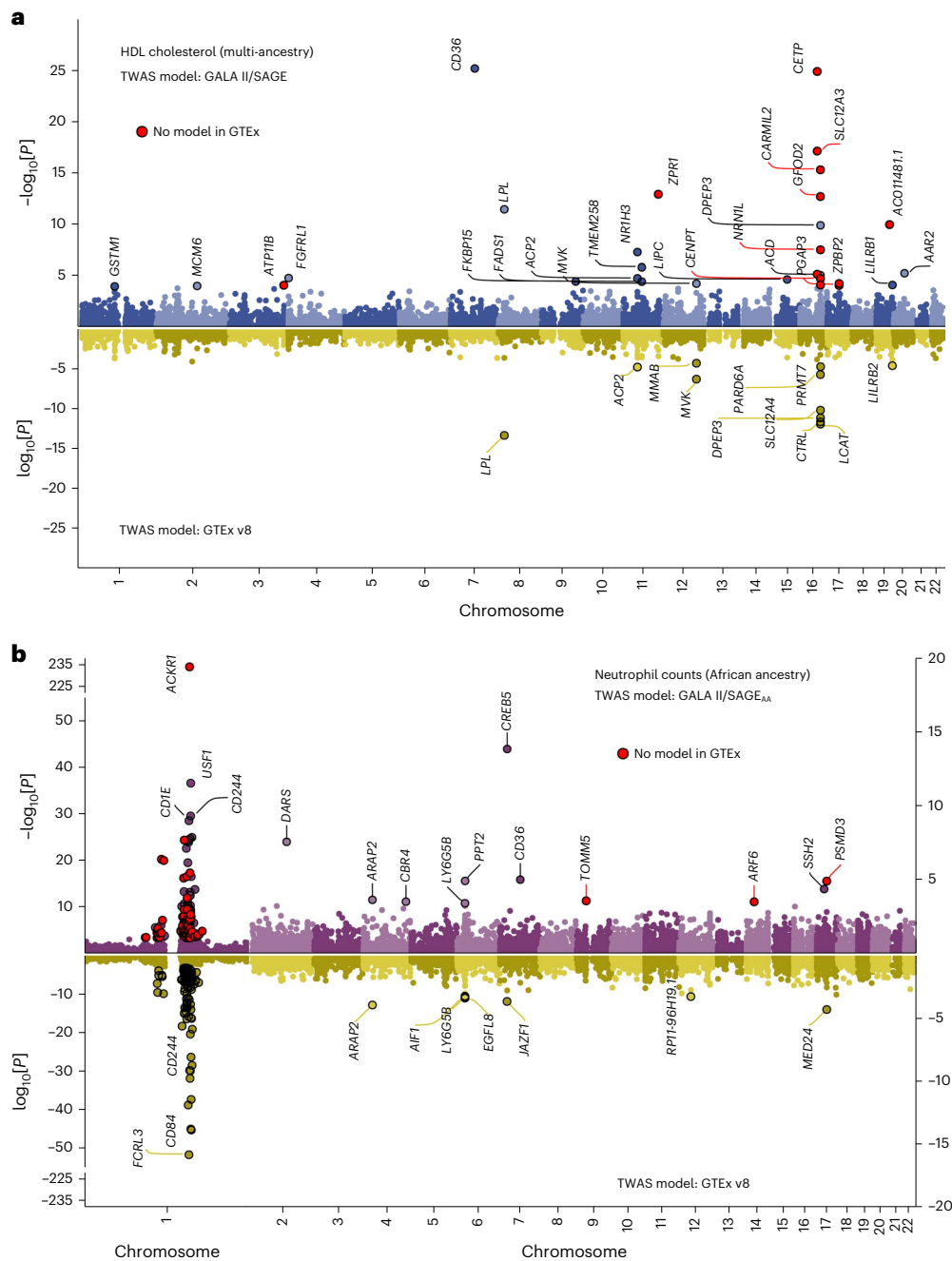
Hollow circles show variants that were not associated with *TRAPPC6A* expression at the gene-specific  $P$  value threshold. kb, kilobases; TSS, transcription start site. **c**, Left, fine-mapping using CAVIAR identified nonoverlapping 95% credible sets in AFR<sub>high</sub> and AFR<sub>low</sub> groups. The lead AFR<sub>high</sub> eQTL, *rs12460041*, had a PP of 0 in AFR<sub>low</sub>. Right, fine-mapping using PESCA to account for linkage disequilibrium differences between populations confirmed that *rs12460041* is a tier 2 anc-eQTL with a PP of >0.80 in AFR<sub>high</sub>. **d**, Tier 3 anc-eQTLs *rs34247110* and *rs3734618* were included in the 95% credible set for *KCNK17* in AFR<sub>high</sub> and AFR<sub>low</sub> groups, but have different eQTL effect sizes in each population, as indicated by the nonoverlapping shaded 95% CIs.



**Fig. 5 | Comparison of transcriptome imputation model performance and TWAS results.** TWAS models were developed using data from GALA II and SAGE participants following the PredictDB pipeline and compared with TWAS models generated in GTEx v8 and MESA using the same approach. **a,b**, Violin plots showing the distribution of internal cross-validation  $r^2$  values for genes with TWAS models available in each study. The box plots extend from the 25th to the 75th percentiles of the  $r^2$  distributions, with median  $r^2$  values indicated by squares. Predictive performance was compared between TWAS models trained in the pooled GALA II/SAGE population and MESA models trained in African American and Hispanic participants (MESA<sub>AFHI</sub>) (**a**) and between TWAS models trained in African Americans (**b**). Statistical significance was determined by

Wilcoxon test and the  $P$  values for differences in  $r^2$  are two sided. **c**, Summary of TWAS associations with  $FDR < 0.05$  across 28 traits in PAGE. BMI, body mass index; BP, blood pressure; CRP, C-reactive protein; eGFR, estimated glomerular filtration rate; HbA1c, hemoglobin A1c; LDL, low-density lipoprotein; WHR, waist-to-hip ratio. **d,e**, Correlation between TWAS z scores based on GALA II/SAGE pooled models and z scores using GTEx (**d**) or MESA<sub>AFHI</sub> models (**e**) for the union of genes with  $FDR < 0.05$  detected with either model. Associations between TWAS z scores from each model are visualized by linear regression lines with shaded 95% CIs. Genes highlighted in orange had  $FDR < 0.05$  using GALA II/SAGE models but did not reach nominal significance (TWAS  $P$  value  $> 0.05$ ) using GTEx or MESA models.





**Fig. 6 | Comparison of TWAS results for selected traits. a,** A TWAS of HDL was performed by applying GALA II/SAGE pooled models and GTEx v8 models to GWAS summary statistics from the multi-ancestry PAGE study ( $n = 33,063$ ). **b,** A TWAS of neutrophil counts was performed by applying GALA II/SAGE models trained in African Americans and GTEx v8 to summary statistics from a GWAS

meta-analysis of individuals of African ancestry ( $n = 13,476$ ) by Chen et al.<sup>32</sup> All genes with  $FDR < 0.05$  are labeled, except for chromosome 1 in **b** due to the large number of statistically significant associations. Significantly associated genes for which expression levels could not be predicted using GTEx v8 elastic net models are highlighted in red.

Concordance between significant associations across all traits was 28%, ranging from 32.7% for height to 7.6% for hemoglobin.

### Discussion

Our comprehensive analysis in a diverse population elucidated the role of genetic ancestry in shaping the genetic architecture of whole-blood gene expression that may be applicable to other complex traits. We found that *cis*-heritability and genetic variance of gene expression increased with a higher proportion of global African ancestry, and that in admixed populations heritability was also highest in individuals with predominantly local African ancestry. We also found that *cis*- $h^2$

and  $V_G$  were lower in individuals with higher levels of Indigenous American compared with European ancestry. The consistent effects of locus-specific ancestry on  $h^2$  and  $V_G$  within each population suggest that confounding by social or environmental factors is unlikely to explain these results. The relationship between ancestry and heritability that we demonstrated for whole-blood gene expression has not been previously shown in a sufficiently large and diverse population with WGS data.

Our findings are consistent with the pattern of heterozygosity in African and Indigenous American populations. Sub-Saharan African populations have the highest heterozygosity since the ancestors of

all other populations passed through a bottleneck during their migration out of Africa<sup>33,34</sup>. Indigenous American populations have passed through additional bottlenecks<sup>35,36</sup>, leading to lower heterozygosity<sup>37</sup>. Therefore, greater genetic variance of gene expression in African ancestry populations may be due to more segregating functional variants in the *cis*-region<sup>38</sup>. This is also supported by the higher number of linkage disequilibrium-independent *cis*-eQTLs, overall and per gene, in AFR<sub>high</sub> compared with AFR<sub>low</sub> groups.

Our second major finding was that over 30% of heritable protein-coding genes have ancestry-specific eQTLs, most of which are tier 1 variants that are rare (MAF < 0.01) or nonpolymorphic in another population. The prevalence of these anc-eQTLs remained stable when the global ancestry cut-off was increased from 50 to 70%. These findings align with a recent plasma proteome analysis of the Atherosclerosis Risk in Communities study, which found that nearly 33% of protein QTLs identified in a large sample of African Americans ( $n = 1,871$ ) were nonexistent or rare in the 1000 Genomes EUR population<sup>39</sup>. Tier 2 anc-eQTLs were defined as variants present at an MAF of  $\geq 0.01$  in both ancestry groups, but which do not belong to the same gene-specific credible set. This eQTL class was far less common than tier 1 and could arise due to differences in environmental effects on gene expression, gene-by-gene and gene-by-environment interactions or multiple causal variants at the same locus. Among eQTL signals that were shared between ancestry groups, effect size heterogeneity was rare and tier 3 anc-eQTLs were effectively eliminated when AFR<sub>high</sub> and IAM<sub>high</sub> were defined using 70% as the global ancestry cut-off. However, comparisons of marginal effect sizes are confounded by differences in sampling error, particularly when there is an imbalance in sample size between populations. Therefore, we may have underestimated ancestry-related heterogeneity in eQTL effects.

Our third major finding was that TWAS models trained in the ancestrally diverse GALA II/SAGE population identified significantly more trait-associated genes than models trained in GTEx and MESA when applied to GWAS results from the multi-ancestry PAGE study. GALA II/SAGE TWAS models benefit from having more similar allele frequency profiles and more accurate modeling of linkage disequilibrium, which is consistent with previous observations<sup>11,12,40</sup> that ancestry-concordant models improve the power for gene discovery in TWASs. Furthermore, over 40% of significantly associated TWAS genes detected using GALA II/SAGE models were not available in GTEx. Biologically informative associations may be missed by relying exclusively on European ancestry-based TWAS models, such as the top two HDL cholesterol-associated genes (*CETP* in 16q13 and *CD36* in 7q21) with established effects on lipid metabolism<sup>20,41–43</sup>. *CD36* expression was associated with multiple phenotypes and contains tier 1 anc-eQTLs found in individuals with >50% African ancestry, consistent with findings of evolutionary pressures at this locus<sup>44,45</sup>. *CD36* encodes a transmembrane protein that binds many ligands, including collagen, thrombospondin and long-chain fatty acids<sup>46</sup>. *CD36* also mediates the adherence of erythrocytes infected with *Plasmodium falciparum*, which causes severe malaria<sup>47,48</sup>.

One of the striking examples of ancestry-specific genetic architecture in our TWASs is the Duffy antigen receptor gene (*ACKR1*) on 1q23.2, where rs2814778 is responsible for constitutively lower white blood cell and neutrophil counts in populations of African ancestry<sup>49,50</sup>. The Duffy-null rs2814778-CC genotype confers resistance to *Plasmodium vivax* malaria and is present at 60–80% frequency in African populations, but is extremely rare in European and Asian populations. The expression of *ACKR1* could not be imputed using GTEx or MESA, but this gene was captured by the pooled and African American GALA II/SAGE models and accounted for the TWAS signal for neutrophil counts in chromosome 1. We also identified 11 genes outside of the Duffy locus, including *DARSI* (which modulates reactivity to mosquito antigens<sup>51</sup>) and *TOMMS* (which has been implicated in lipoprotein phospholipase A2 activity<sup>52</sup>).

Analyses in UKB illustrated that while ancestry-matched training and testing populations are clearly optimal, there is some evidence that transcriptome prediction models developed in African Americans may have better cross-population portability than models trained in predominantly European ancestry cohorts. Similar results were observed for proteome-wide models in the Atherosclerosis Risk in Communities study<sup>37</sup>, where predicted  $r^2$  standardized by *cis*- $h^2$  was higher for AA models applied to EU than for EU models in AA. Greater genetic diversity of African ancestry populations probably captures a more comprehensive set of genetic predictors, only a fraction of which may be present in populations that underwent additional bottlenecks. These findings highlight the value of genetic prediction models trained in admixed, and in particular African ancestry, populations as a resource for identifying new trait-associated genes.

PAGE TWAS z scores were highly correlated across transcriptome models, although the correlation with GALA II/SAGE estimates was higher for GTEx whole blood than MESA<sub>AFHI</sub>, which may partly reflect differences between whole-blood and monocyte transcriptomes in MESA<sub>AFHI</sub><sup>2</sup>. In addition, GALA II/SAGE and GTEx conducted whole-genome sequencing, whereas MESA models are based on imputed genotype data.

Since the genetic architecture of complex traits may mirror the genetics of gene expression<sup>53</sup>, higher heritability in individuals with at least 50% global African ancestry implies that genetic prediction of complex traits in this population should be at least as accurate, if not more so, compared with in European populations. However, polygenic prediction of complex traits in almost all populations, especially African ancestry, lags substantially behind European ancestry<sup>14</sup> due to insufficient sample size and underrepresentation in discovery studies. This is not observed in simulation studies or well-powered analyses of diverse cohorts<sup>38,54,55</sup>. The substantial prevalence of ancestry-specific eQTLs driven by allele frequency differences also implies that analytic approaches alone will yield limited improvements in the cross-population portability of genetic prediction models. For instance, fine-mapping methods that account for differential linkage disequilibrium tagging to identify causal variants will recover some deficits in prediction performance but will not compensate for unobserved risk variants. Our results reinforce the conclusion that developing truly generalizable genetic prediction models requires capture of the full spectrum of genetic variation across human populations. As such, studies that engage with and recruit diverse populations across the globe are more likely to identify novel associations of both public health and biological relevance.

Several limitations of our work should be acknowledged. We only examined whole-blood transcriptomes and analyses of ancestry-specific effects should be conducted for other tissues. However, whole blood is one of the most clinically informative and commonly collected samples, and with the high degree of correlation between gene expression in whole blood and other tissues<sup>56</sup>, our observations regarding the eQTL genetic architecture are probably generalizable. Another limitation of our approach is that we assumed that each gene had one causal eQTL locus in our ancestry comparisons, which may underestimate the number of ancestry-specific eQTLs for genes with multiple independent signals. A caveat of the tier 2 anc-eQTL classification is that PESCA relies on having linkage disequilibrium regions that are approximately independent in both populations to estimate the proportion of causal variants. This is a challenge in admixed populations with longer-range linkage disequilibrium and may lead to biased estimates. Lastly, we compared our TWAS models with elastic net TWAS models from GTEx because they were developed using the same analytic pipeline, although TWAS MASHR models predict a larger number of genes using fine-mapped eQTLs<sup>57</sup>.

Consistent with Gay et al.<sup>58</sup>, we observed that local ancestry explains a larger proportion of variance in gene expression corrected for global ancestry. However, adjustment for local ancestry as a

covariate did not improve the predictive performance of our TWAS models. This may be due to overadjustment, as local ancestry may serve as a proxy for information already captured by population-specific genetic variants, or because of how local ancestry was modeled in our analyses.

Despite these limitations, our study leveraged a uniquely large and diverse sample of 2,733 African American, Puerto Rican and Mexican American participants to explore the interplay between genetic ancestry and the regulation of gene expression. Our cross-ancestry analysis of eQTLs serves as a resource for investigators performing colocalization analyses in genetic association studies of populations of African and Indigenous American ancestry. In addition, we provide genetic prediction models of whole-blood transcriptomes that cover a greater number of genes and facilitate more powerful TWASs when applied to studies of admixed individuals and multi-ancestry GWAS meta-analyses. In summary, our study highlights the need for larger genomic studies in globally representative populations for characterizing the genetic basis of complex traits and ensuring equitable translation of precision medicine efforts.

## Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41588-023-01377-z>.

## References

- Majewski, J. & Pastinen, T. The study of eQTL variations by RNA-seq: from SNPs to phenotypes. *Trends Genet.* **27**, 72–79 (2011).
- The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
- Mogil, L. S. et al. Genetic architecture of gene expression traits across diverse populations. *PLoS Genet.* **14**, e1007586 (2018).
- Wen, X., Luca, F. & Pique-Regi, R. Cross-population joint analysis of eQTLs: fine mapping and functional annotation. *PLoS Genet.* **11**, e1005176 (2015).
- Kim-Hellmuth, S. et al. Cell type-specific genetic regulation of gene expression across human tissues. *Science* **369**, eaaz8528 (2020).
- Porcu, E. et al. Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. *Nat. Commun.* **10**, 3300 (2019).
- Gamazon, E. R. et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).
- Gusev, A. et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).
- Sirugo, G., Williams, S. M. & Tishkoff, S. A. The missing diversity in human genetic studies. *Cell* **177**, 26–31 (2019).
- Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538**, 161–164 (2016).
- Keys, K. L. et al. On the cross-population generalizability of gene expression prediction models. *PLoS Genet.* **16**, e1008927 (2020).
- Geoffroy, E., Gregga, I. & Wheeler, H. E. Population-matched transcriptome prediction increases TWAS discovery and replication rate. *iScience* **23**, 101850 (2020).
- Martin, A. R. et al. Human demographic history impacts genetic risk prediction across diverse populations. *Am. J. Hum. Genet.* **100**, 635–649 (2017).
- Fatumo, S. et al. A roadmap to increase diversity in genomic studies. *Nat. Med.* **28**, 243–250 (2022).
- Patel, R. A. et al. Genetic interactions drive heterogeneity in causal variant effect sizes for gene expression and complex traits. *Am. J. Hum. Genet.* **109**, 1286–1297 (2022).
- Oh, S. S. et al. Effect of secondhand smoke on asthma control among Black and Latino children. *J. Allergy Clin. Immunol.* **129**, 1478–1483.e7 (2012).
- White, M. J. et al. Novel genetic risk factors for asthma in African American children: precision medicine and the SAGE II study. *Immunogenetics* **68**, 391–400 (2016).
- Speed, D., Holmes, J. & Balding, D. J. Evaluating and improving heritability models using summary statistics. *Nat. Genet.* **52**, 458–462 (2020).
- Hoffmann, T. J. et al. A large electronic-health-record-based genome-wide study of serum lipids. *Nat. Genet.* **50**, 401–413 (2018).
- Klarin, D. et al. Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. *Nat. Genet.* **50**, 1514–1523 (2018).
- Wojcik, G. L. et al. Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**, 514–518 (2019).
- 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).
- Suzuki, K. et al. Identification of 28 new susceptibility loci for type 2 diabetes in the Japanese population. *Nat. Genet.* **51**, 379–386 (2019).
- Vujkovic, M. et al. Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis. *Nat. Genet.* **52**, 680–691 (2020).
- Buniello, A. et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
- Grant, D. J. & Maeda, N. A base substitution in the promoter associated with the human haptoglobin 2-1 modified phenotype decreases transcriptional activity and responsiveness to interleukin-6 in human hepatoma cells. *Am. J. Hum. Genet.* **52**, 974–980 (1993).
- Teye, K. et al. A-61C and C-101G Hp gene promoter polymorphisms are, respectively, associated with ahaptoglobinaemia and hypohaptoglobinaemia in Ghana. *Clin. Genet.* **64**, 439–443 (2003).
- Soejima, M., Teye, K. & Koda, Y. The haptoglobin promoter polymorphism rs5471 is the most definitive genetic determinant of serum haptoglobin level in a Ghanaian population. *Clin. Chim. Acta* **483**, 303–307 (2018).
- Boettger, L. M. et al. Recurring exon deletions in the HP (haptoglobin) gene contribute to lower blood cholesterol levels. *Nat. Genet.* **48**, 359–366 (2016).
- Zheng, N. S. et al. A common deletion in the haptoglobin gene associated with blood cholesterol levels among Chinese women. *J. Hum. Genet.* **62**, 911–914 (2017).
- Levin, A. M. et al. Nocturnal asthma and the importance of race/ethnicity and genetic ancestry. *Am. J. Resp. Crit. Care Med.* **190**, 266–273 (2014).
- 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).
- Rosenberg, N. A. et al. Genetic structure of human populations. *Science* **298**, 2381–2385 (2002).
- Henn, B. M., Cavalli-Sforza, L. L. & Feldman, M. W. The great human expansion. *Proc. Natl Acad. Sci. USA* **109**, 17758–17764 (2012).

35. Reich, D. et al. Reconstructing Native American population history. *Nature* **488**, 370–374 (2012).
36. Wall, J. D. et al. Genetic variation in Native Americans, inferred from Latino SNP and resequencing data. *Mol. Biol. Evol.* **28**, 2231–2237 (2011).
37. DeGiorgio, M., Jakobsson, M. & Rosenberg, N. A. Out of Africa: modern human origins special feature: explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa. *Proc. Natl Acad. Sci. USA* **106**, 16057–16062 (2009).
38. Lin, M., Park, D. S., Zaitlen, N. A., Henn, B. M. & Gignoux, C. R. Admixed populations improve power for variant discovery and portability in genome-wide association studies. *Front. Genet.* **12**, 673167 (2021).
39. Zhang, J. et al. Plasma proteome analyses in individuals of European and African ancestry identify *cis*-pQTLs and models for proteome-wide association studies. *Nat. Genet.* **54**, 593–602 (2022).
40. Wen, J. et al. Transcriptome-wide association study of blood cell traits in African ancestry and Hispanic/Latino populations. *Genes (Basel)* **12**, 1049 (2021).
41. Barter, P. J. et al. Cholesteryl ester transfer protein: a novel target for raising HDL and inhibiting atherosclerosis. *Arterioscler. Thromb. Vasc. Biol.* **23**, 160–167 (2003).
42. Armitage, J., Holmes, M. V. & Preiss, D. Cholesteryl ester transfer protein inhibition for preventing cardiovascular events: JACC review topic of the week. *J. Am. Coll. Cardiol.* **73**, 477–487 (2019).
43. Dewey, F. E. et al. Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science* **354**, aaf6814 (2016).
44. Fry, A. E. et al. Positive selection of a CD36 nonsense variant in sub-Saharan Africa, but no association with severe malaria phenotypes. *Hum. Mol. Genet.* **18**, 2683–2692 (2009).
45. Bhatia, G. et al. Genome-wide comparison of African-ancestry populations from CARE and other cohorts reveals signals of natural selection. *Am. J. Hum. Genet.* **89**, 368–381 (2011).
46. Silverstein, R. L. & Febbraio, M. CD36, a scavenger receptor involved in immunity, metabolism, angiogenesis, and behavior. *Sci. Signal.* **2**, re3 (2009).
47. Oquendo, P., Hundt, E., Lawler, J. & Seed, B. CD36 directly mediates cytoadherence of *Plasmodium falciparum* parasitized erythrocytes. *Cell* **58**, 95–101 (1989).
48. Hsieh, F.-L. et al. The structural basis for CD36 binding by the malaria parasite. *Nat. Commun.* **7**, 12837 (2016).
49. Nalls, M. A. et al. Admixture mapping of white cell count: genetic locus responsible for lower white blood cell count in the Health ABC and Jackson Heart studies. *Am. J. Hum. Genet.* **82**, 81–87 (2008).
50. Reich, D. et al. Reduced neutrophil count in people of African descent is due to a regulatory variant in the Duffy antigen receptor for chemokines gene. *PLoS Genet.* **5**, e1000360 (2009).
51. Jones, A. V. et al. GWAS of self-reported mosquito bite size, itch intensity and attractiveness to mosquitoes implicates immune-related predisposition loci. *Hum. Mol. Genet.* **26**, 1391–1406 (2017).
52. Yeo, A. et al. Pharmacogenetic meta-analysis of baseline risk factors, pharmacodynamic, efficacy and tolerability endpoints from two large global cardiovascular outcomes trials for darapladib. *PLoS ONE* **12**, e0182115 (2017).
53. Cookson, W., Liang, L., Abecasis, G., Moffatt, M. & Lathrop, M. Mapping complex disease traits with global gene expression. *Nat. Rev. Genet.* **10**, 184–194 (2009).
54. Holland, D. et al. The genetic architecture of human complex phenotypes is modulated by linkage disequilibrium and heterozygosity. *Genetics* **217**, iyaa046 (2021).
55. Luo, Y. et al. Estimating heritability and its enrichment in tissue-specific gene sets in admixed populations. *Hum. Mol. Genet.* **30**, 1521–1534 (2021).
56. Basu, M., Wang, K., Ruppin, E. & Hannenhalli, S. Predicting tissue-specific gene expression from whole blood transcriptome. *Sci. Adv.* **7**, eabd6991 (2021).
57. Barbeira, A. N. et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat. Commun.* **9**, 1825 (2018).
58. Gay, N. R. et al. Impact of admixture and ancestry on eQTL analysis and GWAS colocalization in GTEx. *Genome Biol.* **21**, 233 (2020).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023

<sup>1</sup>Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA, USA. <sup>2</sup>Department of Epidemiology and Population Health, Stanford University, Stanford, CA, USA. <sup>3</sup>Department of Medicine, University of California, San Francisco, San Francisco, CA, USA. <sup>4</sup>Broad Institute of MIT and Harvard, Cambridge, MA, USA. <sup>5</sup>Center for Individualized and Genomic Medicine Research, Henry Ford Health System, Detroit, MI, USA. <sup>6</sup>Berkeley Institute for Data Science, University of California, Berkeley, Berkeley, CA, USA. <sup>7</sup>Department of Clinical Pharmacy, University of California, San Francisco, San Francisco, CA, USA. <sup>8</sup>Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, CA, USA. <sup>9</sup>Institute for Human Genetics, University of California, San Francisco, San Francisco, CA, USA. <sup>10</sup>Centro de Neumología Pediátrica, San Juan, Puerto Rico. <sup>11</sup>Bay Area Pediatrics, Oakland, CA, USA. <sup>12</sup>Department of Epidemiology and Biostatistics, Graduate School of Public Health and Health Policy, City University of New York, New York, NY, USA. <sup>13</sup>Department of Neurology, University of California, Los Angeles, Los Angeles, CA, USA. <sup>14</sup>Department of Computational Medicine, University of California, Los Angeles, Los Angeles, CA, USA. <sup>15</sup>Department of Internal Medicine, Henry Ford Health System, Detroit, MI, USA. <sup>16</sup>Colorado Center for Personalized Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, USA. <sup>17</sup>Department of Biomedical Informatics, University of Colorado Anschutz Medical Campus, Aurora, CO, USA. <sup>18</sup>Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, San Francisco, CA, USA. <sup>19</sup>These authors contributed equally: Linda Kachuri, Angel C. Y. Mak. <sup>20</sup>These authors jointly supervised this work: Christopher R. Gignoux, Esteban González Burchard, Elad Ziv. ✉e-mail: [Chris.Gignoux@ucsf.edu](mailto:Chris.Gignoux@ucsf.edu); [Esteban.Burchard@ucsf.edu](mailto:Esteban.Burchard@ucsf.edu); [Elad.Ziv@ucsf.edu](mailto:Elad.Ziv@ucsf.edu)

## Methods

### Study population

This study examined African American, Puerto Rican and Mexican American children between 8 and 21 years of age with or without physician-diagnosed asthma from the GALA II study and SAGE. The inclusion and exclusion criteria have previously been described in detail<sup>16,17</sup>. Briefly, participants were eligible if they were 8–21 years of age and identified all four grandparents as Latino for GALA II or African American for SAGE. Study exclusion criteria included the following: (1) any smoking within 1 year of the recruitment date; (2) ten or more pack-years of smoking; (3) pregnancy in the third trimester; and (4) a history of lung diseases other than asthma (for cases) or chronic illness (for cases and controls). Other Latino participants did not self-identify as Mexican American or Puerto Rican.

The local Institutional Review Board from the University of California, San Francisco Human Research Protection Program approved the studies (Institutional Review Board numbers 10-02877 (SAGE) and 10-00889 (GALA II)). All participants and their legal guardians provided written informed consent.

### WGS data and processing

Genomic DNA samples extracted from whole blood were sequenced as part of the Trans-Omics for Precision Medicine (TOPMed) WGS program<sup>59</sup> and the Centers for Common Disease Genomics Genome Sequencing Program (GSP). WGS was performed at the New York Genome Center (NYGC) and Northwest Genomics Center (NWGC) on a HiSeq X system (Illumina) using a paired-end read length of 150 base pairs, with a minimum of 30× mean genome coverage. DNA sample handling, quality control, library construction, clustering, and sequencing, read processing and sequence data quality control have previously been described in detail<sup>59</sup>. All samples were jointly genotyped at the TOPMed Informatics Research Center. Variant calls were obtained from TOPMed data Freeze 8 VCF files generated based on the GRCh38 assembly. Variants with a minimum read depth of 10 (DP10) were used for analysis unless otherwise stated.

### RNA-seq data generation and processing

Total RNA was isolated from a PAXgene tube using a MagMAX for Stabilized Blood Tubes RNA Isolation Kit (4452306; Applied Biosystems). Globin depletion was performed using GLOBINclear Human (AM1980; Thermo Fisher Scientific). RNA integrity and yield were assessed using an Agilent 2100 Bioanalyzer (Agilent Technologies).

Total RNA was quantified using the Quant-iT RiboGreen RNA Assay Kit and normalized to 5 ng  $\mu\text{l}^{-1}$ . An aliquot of 300 ng for each sample was transferred into library preparation, which was an automated variant of the Illumina TruSeq Stranded mRNA Sample Preparation Kit. This method preserves strand orientation of the RNA transcript. It uses oligo dT beads to select messenger RNA from the total RNA sample. It is followed by heat fragmentation and complementary DNA synthesis from the RNA template. The resultant complementary DNA then goes through library preparation (end repair, base A addition, adapter ligation and enrichment) using Broad-designed indexed adapters substituted in for multiplexing. After enrichment, the libraries were quantified with quantitative PCR using the KAPA Library Quantification Kit for Illumina Sequencing Platforms and then pooled equimolarly. The entire process was in 96-well format and all pipetting was done using either an Agilent Bravo or a Hamilton Starlet instrument.

Pooled libraries were normalized to 2 nM and denatured using 0.1 N NaOH before sequencing. Flow cell cluster amplification and sequencing were performed according to the manufacturer's protocols using the HiSeq 4000. Each run was a 101-base pair paired-end read with an eight-base index barcode. Each sample was targeted to 50 million reads. Data were analyzed using the Broad Picard Pipeline, which includes demultiplexing and data aggregation.

RNA-seq reads were further processed using the TOPMed RNA-seq pipeline for year 3 and phase 5 RNA-seq data. Count-level data were generated using the GRCh38 human reference genome and GENCODE 30 for transcript annotation. Count-level quality control and normalization were performed following the GTEx project v8 protocol (<https://gtexportal.org/home/methods>). Sample-level quality control included the removal of RNA samples with an RNA integrity number of <6, genetically related samples (equally or more related than third-degree relative) and sex-discordant samples based on reported sex and *XIST* and *RPS4Y1* gene expression profiles. Count distribution outliers were detected as follows: (1) raw counts were normalized using the trimmed mean of *M* values method in edgeR<sup>60</sup>, as described in GTEx v8 protocol; (2) the  $\log_2$ -transformed normalized counts at the 25th percentile of every sample were identified ( $\text{count}_{q_{25}}$ ); (3) the 25th percentile (Q25) of  $\text{count}_{q_{25}}$  was calculated; and (4) samples were removed if their  $\text{count}_{q_{25}}$  was lower than  $-4$ , as defined by visual inspection.

To account for hidden confounding factors, such as batch effects, technical and biological variation in the sample preparation and sequencing and/or data processing procedures, latent factors were estimated using the probabilistic estimation of expression residuals (PEER) method<sup>61</sup>. Optimization was performed according to an approach adopted by GTEx with the goal to maximize eQTL discovery<sup>62</sup>. A total of 50 (for AA, PR, MX and pooled samples) and 60 (for  $\text{AFR}_{\text{high}}$ ,  $\text{AFR}_{\text{low}}$ ,  $\text{IAM}_{\text{high}}$  and  $\text{IAM}_{\text{low}}$ ) PEER factors were selected for downstream analyses (Supplementary Fig. 6).

### Estimation of global and local genetic ancestry

Genetic principal components, global and local ancestry and kinship estimation on genetic relatedness were computed using biallelic SNPs with a PASS flag from TOPMed Freeze 8 DP10 data, as described previously<sup>63,64</sup>. Briefly, genotype data from European, African and Indigenous American ancestral populations were used as the reference panels for global and local ancestry estimation, assuming three ancestral populations.

Reference genotypes for European (HapMap CEU) and African (HapMap YRI) ancestries were obtained from the Axiom Genotype Data Set (<https://www.thermofisher.com/us/en/home/life-science/microarray-analysis/microarray-data-analysis/microarray-analysis-sample-data/axiom-genotype-data-set>). The CEU populations were recruited from Utah residents with Northern and Western European ancestry from the CEPH collection. The YRI populations were recruited from Yoruba in Ibadan, Nigeria. The Axiom Genome-Wide LAT 1 array was used to generate the Indigenous American ancestry reference genotypes from 71 Indigenous Americans (14 Zapotec, two Mixe and 11 Mixtec from Oaxaca and 44 Nahua from Central Mexico)<sup>65,66</sup>. ADMIXTURE was used with the reference genotypes in a supervised analysis assuming three ancestral populations. Global ancestry was estimated by ADMIXTURE<sup>67</sup> in supervised analysis mode, whereas local ancestry was estimated by RFMIX version 2 with default settings<sup>68</sup>. Throughout this study, the local ancestry of a gene was defined as the number of ancestral alleles (zero, one or two) at the transcription start site.

Comparative analyses were performed based on two different sample grouping strategies, by self-identified race/ethnicity or by global ancestry. Self-identified race/ethnicity included four groups: AA, PR, MX and the pooling of AA, PR, MX and other Latino Americans (pooled). For groups defined by global ancestry, samples were grouped into high (>50%;  $\text{AFR}_{\text{high}}$  and  $\text{IAM}_{\text{high}}$ ) or low (<10%;  $\text{AFR}_{\text{low}}$  and  $\text{IAM}_{\text{low}}$ ) global African or Indigenous American ancestry. The sample size for each group is shown in Supplementary Table 1.

### Cis-heritability of gene expression

The genetic region of *cis*-gene regulation was defined by a 1-megabase (Mb) region flanking each side of the transcription start site (*cis*-region). *Cis*-heritability ( $h^2$ ) of gene expression was estimated

using unconstrained genome-based restricted maximum likelihood (GREML)<sup>69</sup> analysis (`--reml-no-constrain`) and estimation was restricted to common autosomal variants ( $MAF \geq 0.01$ ). Inverse-normalized gene expression was regressed on PEER factors and the residuals were used as the phenotype for GREML analysis. Sex and asthma case–control status were used as categorical covariates, while age at blood draw and the first five genetic principal components were used as quantitative covariates. *Cis*-heritability was estimated separately for each self-identified race/ethnicity group (AA, PR, MX and pooled) and groupings based on global ( $AFR_{high}$ ,  $AFR_{low}$ ,  $IAM_{high}$  and  $IAM_{low}$ ) and local ancestry (described below). Differences in the distribution of  $h^2$  and genetic variance ( $V_G$ ) between groups were tested using two-sided Wilcoxon tests. Parallel analyses were also conducted for Indigenous American ancestry (IAM/IAM versus EUR/EUR and IAM/IAM versus IAM/EUR).

The following sensitivity analyses were conducted using GCTA: (1) using the same sample size in each self-identified group ( $n = 600$ ); and (2) partitioning heritability and genetic variance by two MAF bins (0.01–0.10 and 0.1–0.5). We also estimated heritability using the LDK-Thin model<sup>70</sup>, following the recommended genetic relatedness matrix processing. Thinning of duplicate SNPs was performed using the arguments `--window-prune .98 --window-kb 100`. The direct method was applied to calculate kinship using the thinned data and, lastly, generalized restricted maximum likelihood (REML) was used to estimate heritability.

### Association of global and local ancestry with gene expression

Methods from Gay et al.<sup>58</sup> were modified to identify genes associated with global and local ancestry. In step 1, inversed normalized gene expression was regressed on age, sex and asthma status (model 0). In step 2, the residuals from model 0 were regressed on global ancestry (model 1). In step 3, the residuals from model 1 were regressed on local ancestry (model 2) to identify genes that are associated with local ancestry. An FDR of 0.05 was applied to steps 2 and 3 separately to identify genes that were significantly associated with global and/or local ancestry. Steps 1–3 were run separately for African and Indigenous American ancestry. For heritable genes that were associated with global and/or local ancestry, a joint model of regressing global and local ancestry from residuals from model 0 was also examined to assess the percentage of variance of gene expression explained by global and/or local ancestry.

### Identification of eGenes, *cis*-eQTLs and ancestry-specific *cis*-eQTLs

FastQTL<sup>71</sup> was used to process raw gene counts and identify eQTLs, according to the GTEx v8 pipeline (<https://github.com/broadinstitute/gtex-pipeline>). Age, sex, asthma status, the first five genetic ancestry principal components and PEER factors were used as covariates for FastQTL analysis. To account for multiple testing across all tested genes, Benjamini–Hochberg correction was applied to the beta-approximated  $P$  values from the permutation step of FastQTL. For each gene with a significant beta-approximated  $P$  value at an FDR of  $<0.05$ , a nominal  $P$  value threshold was estimated using the beta-approximated  $P$  value. *Cis*-eQTLs were defined as genetic variants with nominal  $P$  values less than the nominal  $P$  value threshold of the corresponding gene. eGenes were defined as genes with at least one eQTL. To summarize the number of independent *cis*-eQTLs in each ancestry group, linkage disequilibrium clumping was performed using PLINK (`--clump-kb 1000 --clump-r2 0.1`) using gene-specific  $P$  value thresholds.

*Trans*-eQTLs were identified using the same protocol as in GTEx v8 (ref. 2). *Trans*-eQTLs were defined as eQTLs that were not located on the same chromosome as the gene. Only protein-coding and long intergenic noncoding RNA genes and SNPs on autosomes were included in the analyses. Briefly, linear regression on the expression of the gene was performed in PLINK2 (version v2.00a3LM; released 28 March 2020)

using SNPs with  $MAF \geq 0.05$  and the same covariates as in *cis*-eQTL discovery. Gene and variant mappability data (GRCh38 and GENCODE v26) were downloaded from Saha and Battle<sup>72</sup> for the following filtering steps: (1) keep gene–variant pairs that pass a  $P$  value threshold of  $1 \times 10^{-5}$ ; (2) keep genes with a mappability of  $\geq 0.8$ ; (3) remove SNPs with a mappability of  $<1$ ; and (4) remove a *trans*-eQTL candidate if genes within 1 Mb of the SNP candidate cross-map with the *trans*-eGene candidate. The Benjamini–Hochberg procedure was applied to control for the FDR at the 0.05 level using the smallest  $P$  value (multiplied by  $10^{-6}$ ) from each gene. An additional filtering step was applied for the  $AFR_{high}$  and  $IAM_{high}$  groups. For  $AFR_{high}$ , all *trans*-eQTLs detected in  $AFR_{low}$  group were removed and the resulting *trans*-eQTLs were referred to as filtered  $AFR_{high}$  *trans*-eQTLs. Similarly, for the  $IAM_{high}$  group, all *trans*-eQTLs detected in the  $IAM_{low}$  group were removed and the resulting *trans*-eQTLs were referred to as filtered  $IAM_{high}$  *trans*-eQTLs. Filtered  $AFR_{high}$  *trans*-eQTLs were checked for the presence of filtered  $IAM_{high}$  *trans*-eQTLs and vice versa. Linkage disequilibrium clumping was performed using PLINK (v1.90b6.26 `--clump-kb 1000 --clump-r2 0.1 --clump-p1 0.00000005 --clump-p2 1`) to group *trans*-eQTLs into independent signals.

Mapping of anc-eQTLs was performed in participants stratified by high and low global African and Indigenous American ancestry. We developed a framework to identify anc-eQTLs by focusing on the lead eQTL signal for each gene and comparing fine-mapped 95% credible sets between high ( $>50\%$ ) and low ( $<10\%$ ) global ancestry groups ( $AFR_{high}$  versus  $AFR_{low}$  and  $IAM_{high}$  versus  $IAM_{low}$ ). Sensitivity analyses were conducted using  $>70\%$  as the cut-off for the  $AFR_{high}$  and  $IAM_{high}$  groups. Anc-eQTLs were classified into three tiers as described below, based on population differences in allele frequency, linkage disequilibrium and effect size (Fig. 3a). For every protein-coding and heritable eGene (GCTA  $h^2$  likelihood ratio test  $P$  value  $< 0.05$ ), the lead eQTL signal was identified using CAVIAR<sup>73</sup>, assuming one causal locus ( $c = 1$ ). The 95% credible sets of eQTLs in the high and low global ancestry groups were compared to determine whether there was any overlap. Variants from nonoverlapping 95% credible sets were further classified as tier 1 anc-eQTLs based on allele frequency differences, or tier 2 after additional fine-mapping using PESCA<sup>22</sup>. For genes with overlapping 95% credible sets, tier 3 anc-eQTLs were detected based on effect size heterogeneity.

eQTLs identified in the  $AFR_{high}$  or  $IAM_{high}$  groups that were common ( $MAF \geq 0.01$ ) in the high groups but rare ( $MAF < 0.01$ ) or monomorphic in the  $AFR_{low}$  or  $IAM_{low}$  groups were classified as tier 1. If the eQTLs were detected at an  $MAF$  of  $\geq 0.01$  in both the high and low ancestry groups, they were further fine-mapped using PESCA<sup>22</sup>, which tests for differential effect sizes while accounting for linkage disequilibrium between eQTLs. Preprocessing for the PESCA analyses involved linkage disequilibrium pruning at  $r^2 > 0.95$ . All eQTL pairs with  $r^2 > 0.95$  were identified in both the high and low groups and only those pairs common to both groups were removed. For each eQTL, PESCA estimated three posterior probabilities: specific to the  $AFR_{high}$  or  $IAM_{high}$  group ( $PP_{high}$ ); specific to the  $AFR_{low}$  or  $IAM_{low}$  group ( $PP_{low}$ ); or shared between the two groups ( $PP_{shared}$ ). Tier 2 anc-eQTLs were selected based on the following criteria: (1) all variants in the credible set have ( $PP_{high} > PP_{low}$ ) and ( $PP_{high} > PP_{shared}$ ); and (2)  $PP_{high} > 0.8$ . The tier 3 class was based on evidence of significant heterogeneity in eQTL effect size, defined as a Cochran's  $Q$   $P$  value of  $<0.05/n_{gene}$ , where  $n_{gene}$  was the number of genes tested. Since we assume that the 95% credible set corresponds to a single lead eQTL signal, all eQTLs in the credible set were required to have a significant heterogeneous effect size to be classified as tier 3 anc-eQTLs.

To systematically assess the overlap in eQTL signals identified in our study and trait-associated loci, we colocalized eQTL summary statistics with GWAS results from PAGE. Colocalization was performed using COLOC<sup>74</sup> within a linkage disequilibrium window of 2 Mb centered on the eQTL with the lowest GWAS  $P$  value. For each eQTL–trait

pair, a posterior probably of a shared causal signal ( $PP_4$ ) of  $>0.80$  was interpreted as strong evidence of colocalization.

### Development and application of multi-ancestry TWAS models

Gene prediction models for *cis*-gene expression were generated using common variants and elastic net modeling implemented in the PredictDBv7 pipeline ([https://github.com/hakyimlab/PredictDB\\_Pipeline\\_GTEEx\\_v7](https://github.com/hakyimlab/PredictDB_Pipeline_GTEEx_v7)). Models were filtered by nested cross-validation prediction performance and heritability  $P$  value ( $\rho_{\text{avg}} > 0.1$ ,  $z_{\text{score\_pval}} < 0.05$  and GCTA  $h^2$   $P$  value  $< 0.05$ ). Sensitivity analyses were performed by generating gene prediction models that included the number of ancestral alleles as covariates to account for local ancestry in the *cis*-region. In the AA group, one covariate indicating the count of African ancestral alleles was used, whereas in the PR, MX and pooled groups, two additional covariates indicating the numbers of European and Indigenous American ancestral alleles were used.

Out-of-sample validation of the gene expression prediction models was done using 598 individuals from the African American asthma cohort SAPPHERE<sup>31</sup>. Predicted gene expression from SAPPHERE genotypes was generated using the predict function from MetaXcan. Genotypes of SAPPHERE samples were generated by whole-genome sequencing through the TOPMed program and were processed in the same way as for GALA II and SAGE. RNA-seq data from SAPPHERE were generated as previously described<sup>75</sup> and were normalized using the trimmed mean of  $M$  values method in edgeR. Predicted and normalized gene expression data were compared to generate correlation  $r^2$  values.

To assess the performance of the resulting GALA II/SAGE models, we conducted TWASs of 28 traits using GWAS summary statistics from the PAGE Consortium study by Wojcik et al.<sup>21</sup>. Analyses were performed using S-PrediXcan with whole-blood gene prediction models from GALA II and SAGE (GALA II/SAGE models) and GTEEx v8, as well as monocyte gene expression models from MESA<sup>3</sup>. In the UKB, we conducted TWASs of 22 blood-based biomarkers and quantitative traits using GALA II/SAGE models generated in African Americans (GALA II/SAGE AA) and GTEEx v8 whole blood. Each set of TWAS models was applied to publicly available GWAS summary statistics (Pan-UKB team; <https://pan.ukbb.broadinstitute.org>) from participants of predominantly European ancestry (UKB EUR) and African ancestry (UKB AFR). Ancestry assignment in UKB was based on a random forest classifier trained on the merged 1000 Genomes and Human Genome Diversity Project reference populations. The classifier was applied to UKB participants projected into the 1000 Genomes and Human Genome Diversity Project principal components.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

### Data availability

WGS and RNA-seq data for the GALA II, SAGE and SAPPHERE studies, generated as part of the NHLBI TOPMed program, are available from the Database of Genotypes and Phenotypes under accession numbers [phs000920](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE1000920) (GALA II), [phs000921](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE1000921) (SAGE) and [phs001467](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE1001467) (SAPPHERE). Comprehensive phenotypic data for GALA II study participants are available through the Database of Genotypes and Phenotypes ([phs001180](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE1001180)). Summary statistics for *cis*- and *trans*-eQTLs, a catalog of ancestry-specific eQTLs, TWAS models developed using data from GALA II and SAGE participants, and normalized individual-level gene expression data have been posted in a public repository at <https://doi.org/10.5281/zenodo.7735723>.

### Code availability

A description of the RNA-seq harmonization pipeline is available at <https://doi.org/10.5281/zenodo.7735723>. Scripts for performing heritability analyses, eQTL mapping and TWAS model development

are available at [https://github.com/angelcymak/gala2\\_sage\\_eQTL](https://github.com/angelcymak/gala2_sage_eQTL). Heritability analyses were performed using GCTA (<https://yanglab.westlake.edu.cn/software/gcta/#GREMLanalysis>). FastQTL analysis was conducted following the GTEx v8 pipeline (<https://github.com/broadinstitute/gtex-pipeline>). TWAS models were generated following the PredictDB v7 pipeline ([https://github.com/hakyimlab/PredictDB\\_Pipeline\\_GTEEx\\_v7](https://github.com/hakyimlab/PredictDB_Pipeline_GTEEx_v7)).

### References

- Taliun, D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
- Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
- Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500–507 (2012).
- Aguet, F. et al. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
- Mak, A. C. Y. et al. Lung function in African American children with asthma is associated with novel regulatory variants of the KIT ligand KITLG/SCF and gene-by-air-pollution interaction. *Genetics* **215**, 869–886 (2020).
- Lee, E. Y. et al. Whole-genome sequencing identifies novel functional loci associated with lung function in Puerto Rican youth. *Am. J. Respir. Crit. Care Med.* **202**, 962–972 (2020).
- Kumar, R. et al. Factors associated with degree of atopy in Latino children in a nationwide pediatric sample: the Genes-Environments and Admixture in Latino Asthmatics (GALA II) study. *J. Allergy Clin. Immunol.* **132**, 896–905.e1 (2013).
- Spear, M. L. et al. A genome-wide association and admixture mapping study of bronchodilator drug response in African Americans with asthma. *Pharmacogenomics J.* **19**, 249–259 (2019).
- Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
- Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* **93**, 278–288 (2013).
- Yang, J. et al. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
- Zhang, Q., Privé, F., Vilhjálmsson, B. & Speed, D. Improved genetic prediction of complex traits from individual-level data or summary statistics. *Nat. Commun.* **12**, 4192 (2021).
- Ongen, H., Buil, A., Brown, A. A., Dermizakis, E. T. & Delaneau, O. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**, 1479–1485 (2016).
- Saha, A. & Battle, A. False positives in trans-eQTL and co-expression analyses arising from RNA-sequencing alignment errors. *F1000Res* **7**, 1860 (2019).
- Hormozdiani, F., Kostem, E., Kang, E. Y., Pasaniuc, B. & Eskin, E. Identifying causal variants at loci with multiple signals of association. *Genetics* **198**, 497–508 (2014).
- Wallace, C. Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. *PLoS Genet.* **16**, e1008720 (2020).
- Levin, A. M. et al. Integrative approach identifies corticosteroid response variant in diverse populations with asthma. *J. Allergy Clin. Immunol.* **143**, 1791–1802 (2019).

### Acknowledgements

The generation of molecular data for the TOPMed program was supported by the National Heart, Lung, and Blood Institute (NHLBI). RNA-seq for the NHLBI TOPMed Genes-Environments and Admixture

in Latino Asthmatics Study (GALA II; phs000920) and Study of African Americans, Asthma, Genes, and Environments (SAGE; phs000921) was performed at the Broad Institute Genomics Platform (HHSN268201600034). WGS for the same studies was performed at the NYGC (3R01HL117004-02S3) and NWGC (HHSN268201600032). Core support including centralized genomic read mapping and genotype calling, along with variant quality metrics and filtering, was provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002). Core support including phenotype harmonization, data management, sample identity quality control and general program coordination was provided by the TOPMed Data Coordinating Center (R01HL-120393 and U01HL-120393; contract HHSN268201800001). WGS as part of GALA II was performed by the NYGC under a grant from the Centers for Common Disease Genomics of the GSP (UM1 HG008901). The GSP Coordinating Center (U24 HG008956) contributed to cross-program scientific initiatives and provided logistical and general study coordination. The GSP is funded by the National Human Genome Research Institute, NHLBI and National Eye Institute. This work and E.G.B. were supported in part by the Sandler Family Foundation, American Asthma Foundation, Robert Wood Johnson Foundation Amos Medical Faculty Development Program, Harry Wm. and Diana V. Hind Distinguished Professor in Pharmaceutical Sciences II, NHLBI (R01HL117004, R01HL135156, X01HL134589 and U01HL138626), National Institute of Environmental Health Sciences (R01ES015794), National Institute on Minority Health and Health Disparities (R56MD013312 and P60MD006902), Tobacco-Related Disease Research Program (24RT-0025 and 27IR-0030) and National Human Genome Research Institute (U01HG009080). L.K. was supported by funding from the National Cancer Institute (K99CA246076). E.Z. was supported by funding from the National Cancer Institute (K24CA169004 and R01CA227466) and State of California Initiative to Advance Precision Medicine (OPR18111). K.L.K. was additionally supported by a diversity supplement of NHLBI R01HL135156, the University of California, San Francisco Bakar Computational Health Sciences Institute, a Gordon and Betty Moore Foundation grant (GBMF3834) and a grant from the Alfred P. Sloan Foundation (2013-10-27) to the University of California, Berkeley through the Moore-Sloan Data Sciences Environment initiative at the Berkeley Institute for Data Science. C.R.G. was supported by funding from the National Human Genome Research Institute (R01HG010297). C.R.G. and N.A.Z. were supported by funding from

the NHLBI (R01HL151152) and National Human Genome Research Institute (R01HG011345). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed. We thank S. Germer, M. C. Zody, L. Winterkorn and C. Reeves from the NYGC and D. A. Nickerson from the NWGC for overseeing the production of GALA II and SAGE WGS data. We thank H. Shi, K. S. Burch and B. Pasaniuc for technical support with the PESCA method. This content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Figures 1 and 3 and Supplementary Fig. 1 were created with [BioRender.com](https://www.biorender.com).

### Author contributions

A.C.Y.M., L.K., K.L.K., C.R.G., N.A.Z., E.G.B. and E.Z. contributed to the conception or design of the work. L.K., A.C.Y.M., D.H., C.E., S.H., J.R.E., N.G., S.G., S.X., A.O.-O., J.R.R.-S., M.A.L., L.K.W., L.N.B., C.R.G., N.A.Z., E.G.B. and E.Z. contributed to the acquisition, analysis or interpretation of the data. L.K., A.C.Y.M., J.R.E., K.L.K., A.O.-O., L.N.B., C.R.G., N.A.Z., E.G.B. and E.Z. drafted the work or substantively revised it. All authors approved the submission of this manuscript.

### Competing interests

The authors declare no competing interests.

### Additional information

**Extended data** is available for this paper at <https://doi.org/10.1038/s41588-023-01377-z>.

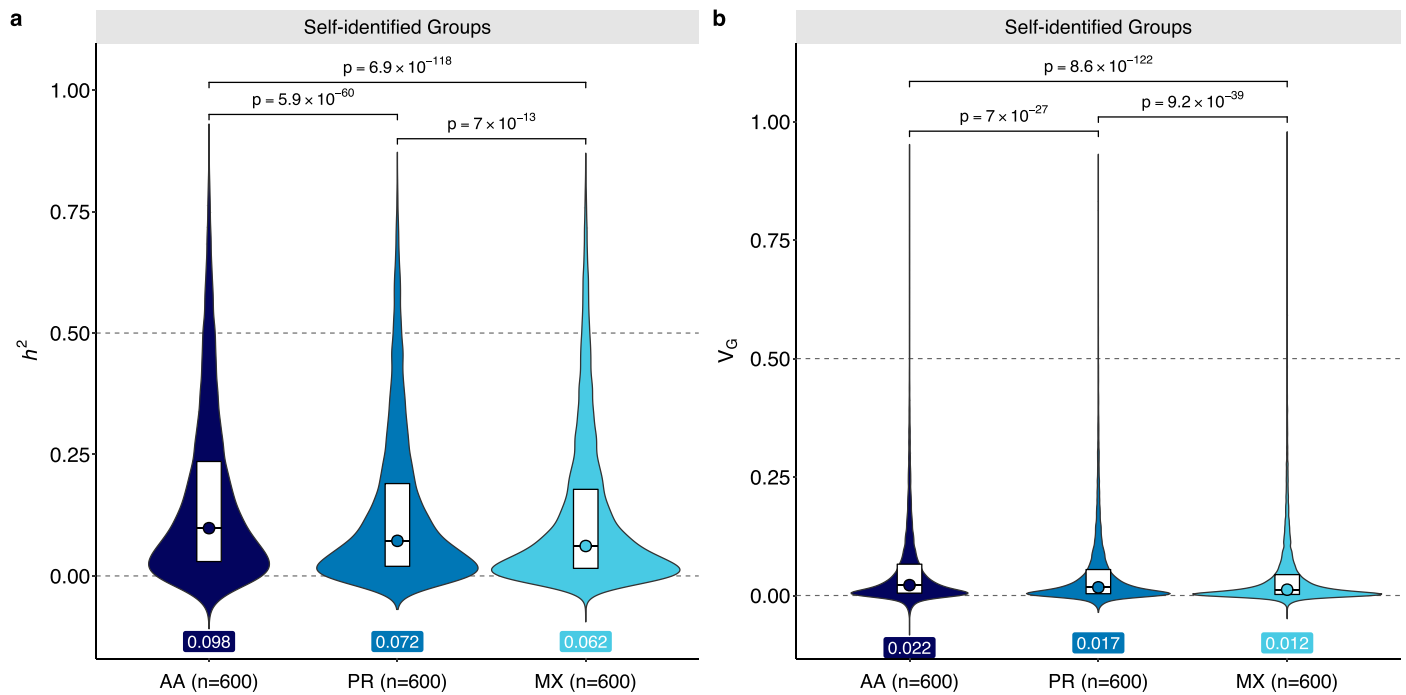
**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41588-023-01377-z>.

**Correspondence and requests for materials** should be addressed to Christopher R. Gignoux, Esteban González Burchard or Elad Ziv.

**Peer review information** *Nature Genetics* thanks Heather Wheeler and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

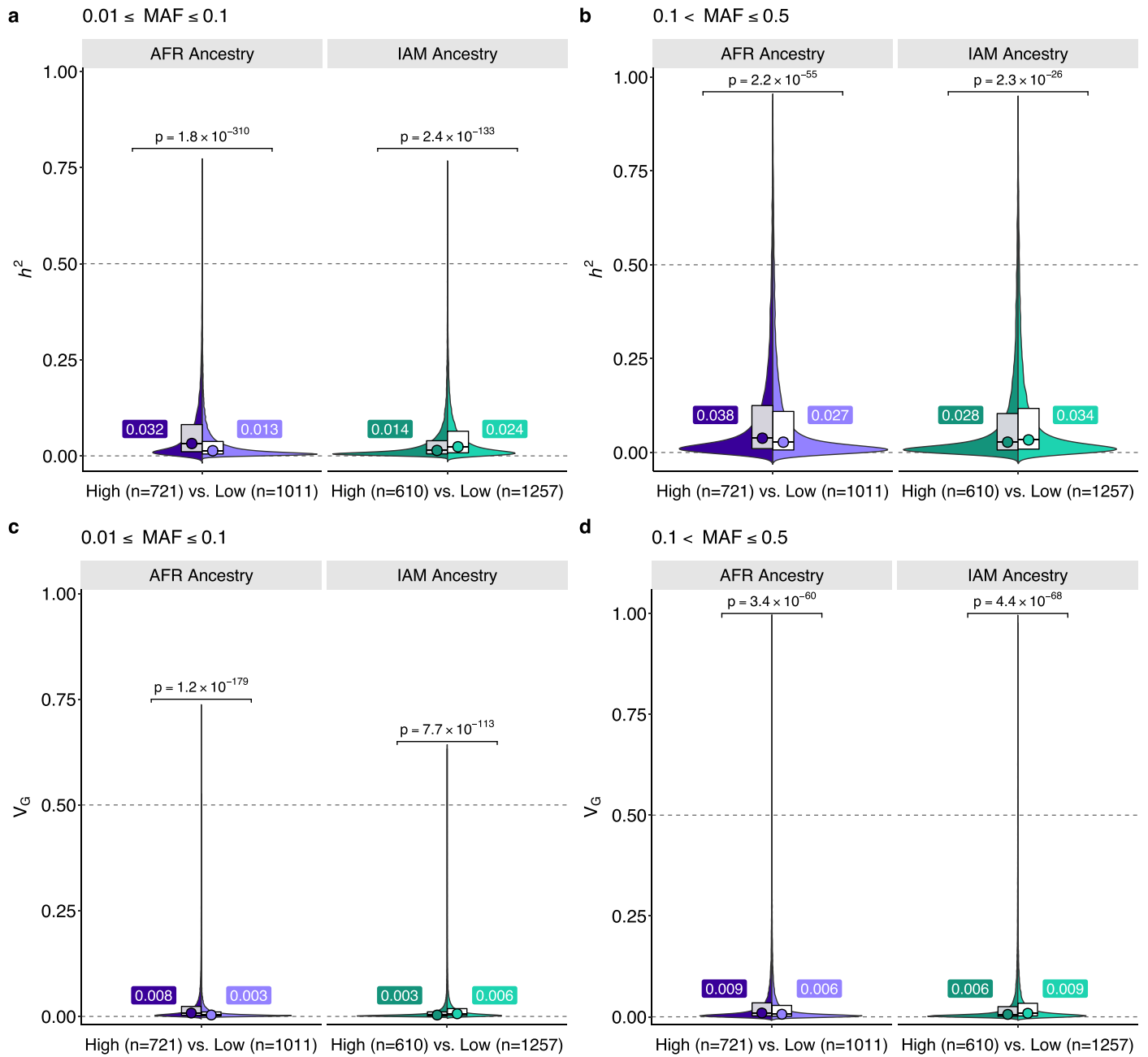
**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).





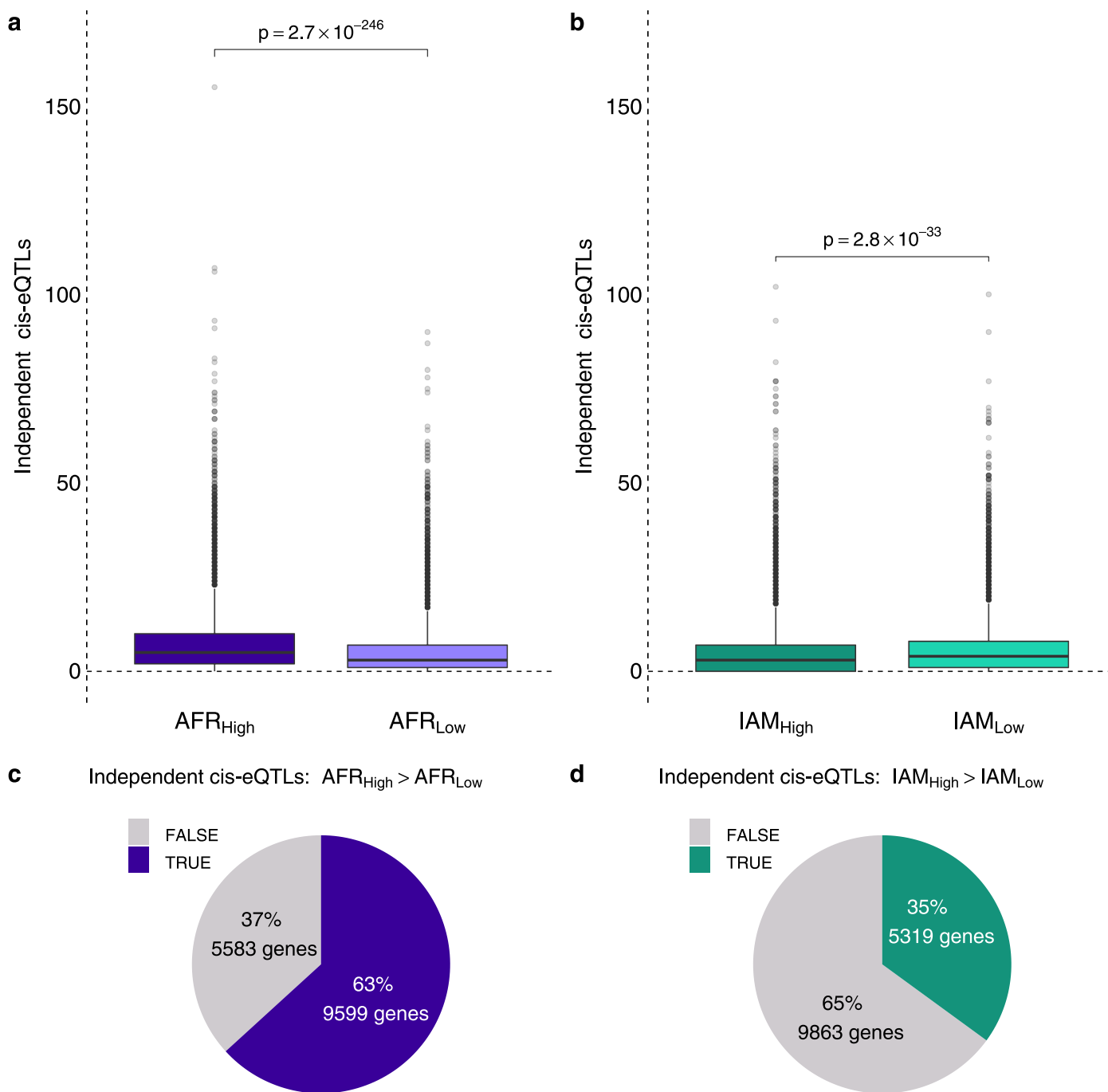
**Extended Data Fig. 1 | Comparison of *cis*-heritability ( $h^2$ ) and genetic variance ( $V_G$ ) of whole blood transcript levels across populations with a fixed sample size.** Analyses of **a**,  $h^2$  and **b**,  $V_G$  were stratified by self-identified race/ethnicity and compared African American (AA), Puerto Rican (PR), Mexican American (MX) participants. Within each self-identified race/ethnicity group, individuals

were down-sampled to  $n = 600$ . Violin plots show the full distribution of  $h^2$  or  $V_G$  values across all genes within each population. Box plots extend from the 25<sup>th</sup> to the 75<sup>th</sup> percentile. Each group-specific median value of  $h^2$  or  $V_G$  is annotated and indicated by the shaded circle. All Wilcoxon  $p$ -values are two-sided.



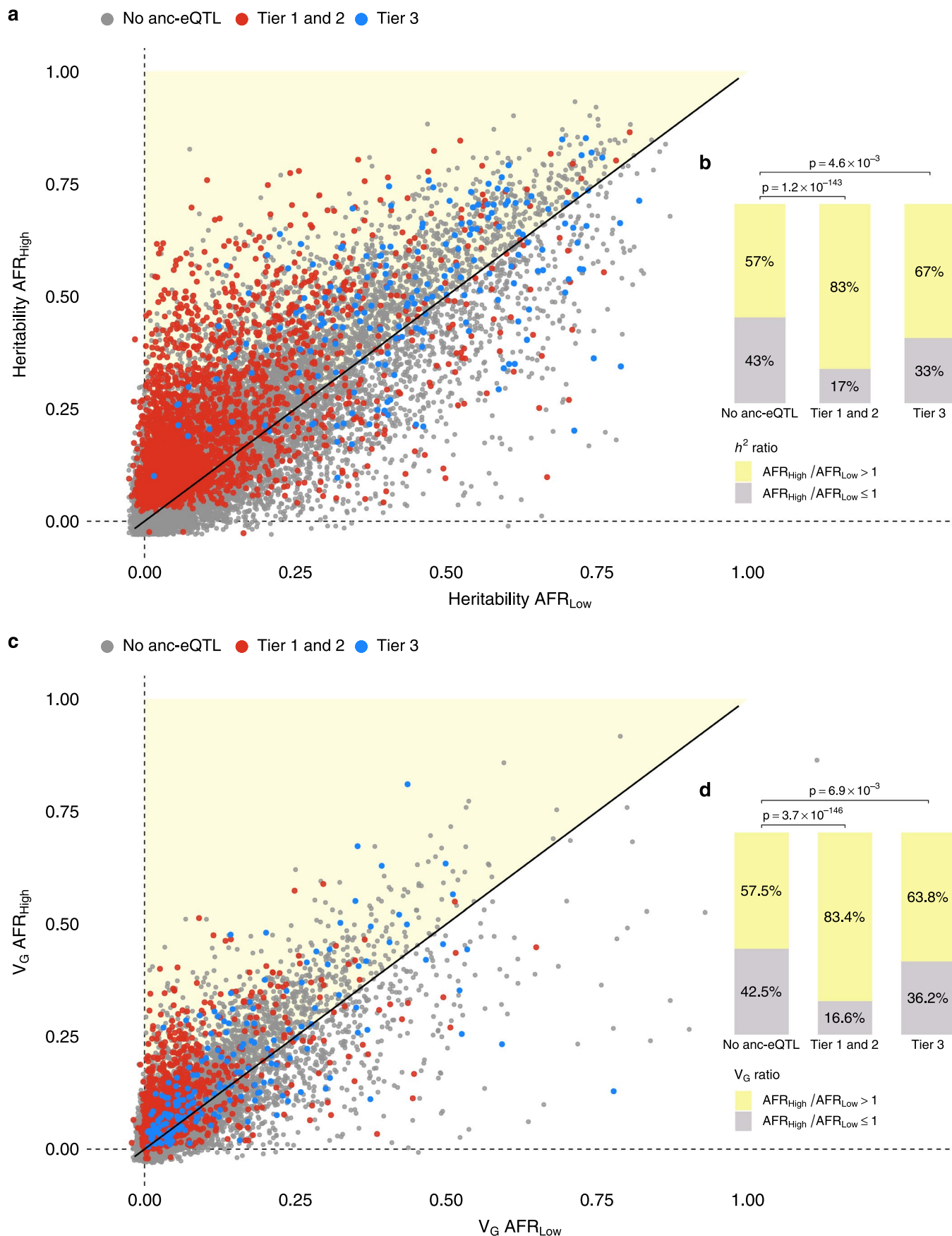
**Extended Data Fig. 2 | Comparison of *cis*-heritability ( $h^2$ ) and genetic variance ( $V_G$ ) of whole blood transcript levels stratified by minor allele frequency (MAF). Analyses were performed separately for low-frequency variants with  $0.01 \leq \text{MAF} \leq 0.1$  and for common variants with  $0.01 \leq \text{MAF} < 0.5$ . Split violin plots show the distribution of **a**, **b**,  $h^2$  and **c**, **d**,  $V_G$  in individuals with >50% global genetic**

ancestry (High) to participants with <10% of the same ancestry (Low). Analyses were conducted separately for African (AFR) and Indigenous American (IAM) ancestry within each MAF bin. Box plots extend from the 25<sup>th</sup> to the 75<sup>th</sup> percentile and the median value for  $h^2$  and  $V_G$  value is annotated. All Wilcoxon p-values are two-sided.



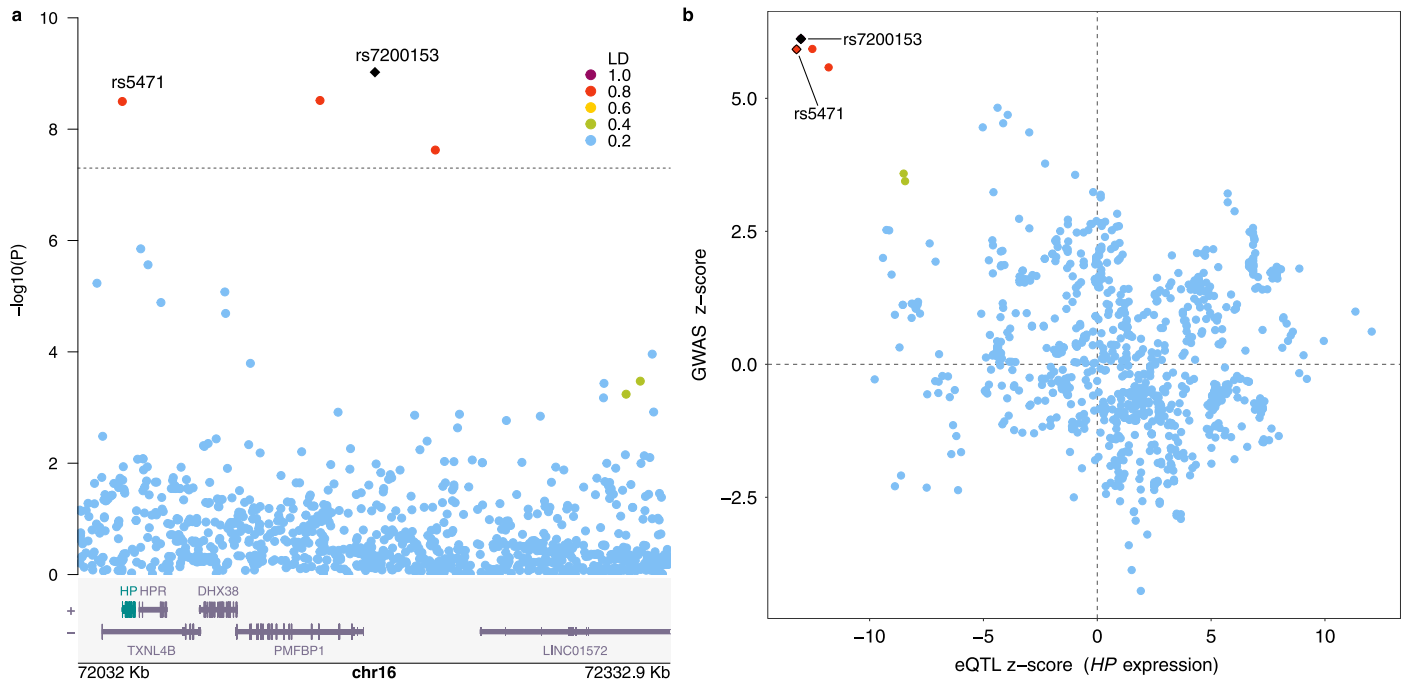
**Extended Data Fig. 3 | Cross-population comparison of independent cis-eQTLs.** Sample size was fixed to  $n = 600$  for eQTL mapping analyses in each ancestry group. Independent cis-eQTLs were identified by performing LD-based clumping ( $r^2 < 0.10$ ) of statistically significant eQTLs identified in each population. Individuals with >50% global genetic ancestry (High) were compared to participants with <10% of the same ancestry (Low). Analyses were conducted

separately for African (AFR) and Indigenous American (IAM) ancestry. The distribution of independent cis-eQTLs per gene between was compared between **a**, AFR<sub>high</sub> and AFR<sub>low</sub> populations and **b**, IAM<sub>high</sub> and IAM<sub>low</sub> groups using a two-sided Wilcoxon test. Pie charts visualize the proportion of genes with a greater number of cis-eQTLs in **c**, AFR<sub>high</sub> compared to AFR<sub>low</sub> and in **d**, IAM<sub>high</sub> compared to IAM<sub>low</sub>.



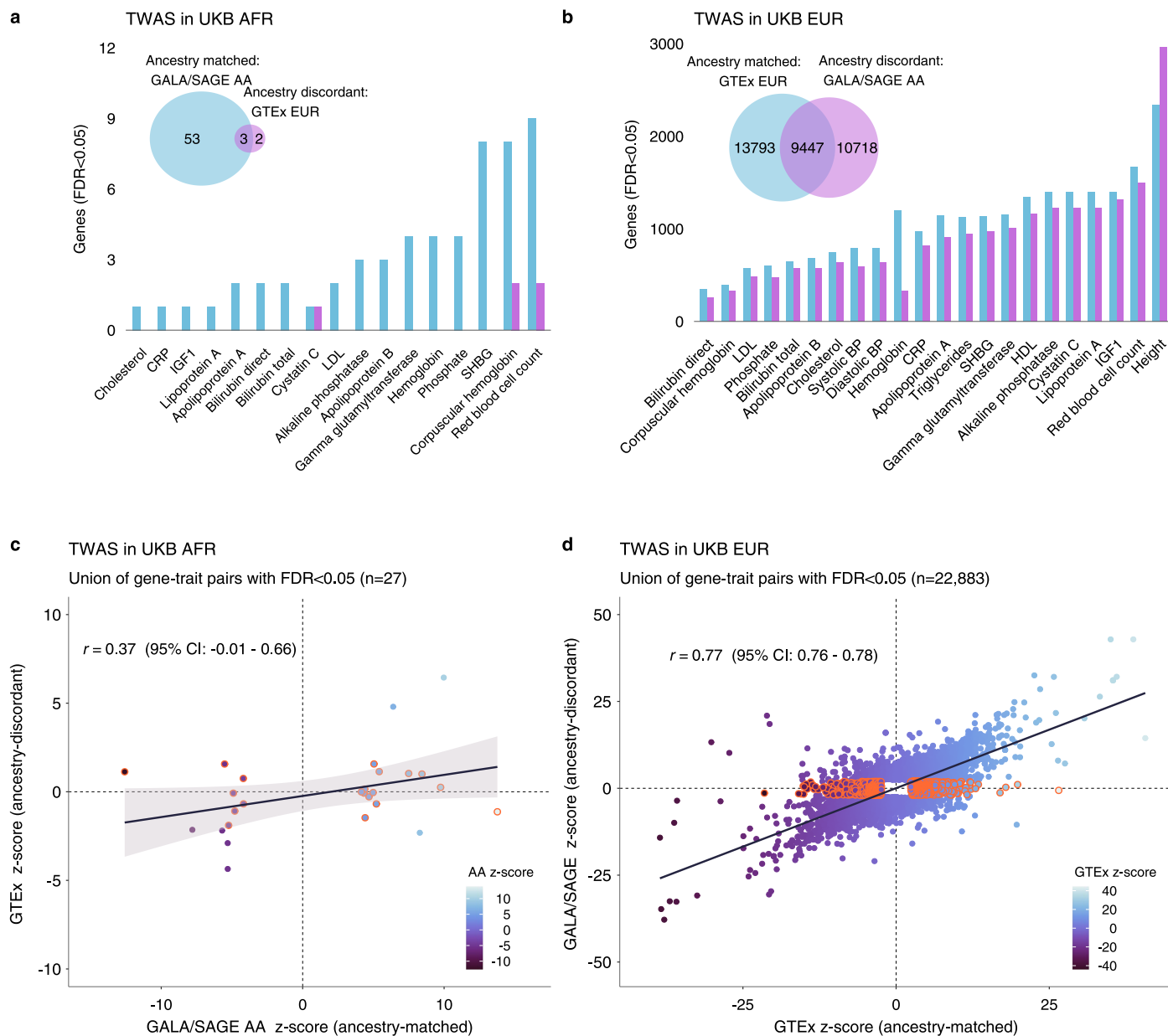
**Extended Data Fig. 4 | Scatter plots comparing gene-specific  $h^2$  and  $V_G$  by levels of African ancestry.** Estimates of **a**,  $h^2$  and **c**,  $V_G$  for each gene in individuals with  $\geq 50\%$  global African ancestry ( $AFR_{High}$ ) are plotted on the y-axis and in participants with  $< 10\%$  AFR ancestry ( $AFR_{Low}$ ) on the x-axis. Genes containing

ancestry-specific eQTLs are highlighted. The proportion of genes falling off the diagonal, with higher **b**,  $h^2$  or **d**,  $V_G$  in  $AFR_{High}$  than  $AFR_{Low}$ , is shown using stacked bar plots and compared using a two-sided binomial test.



**Extended Data Fig. 5 | Colocalization of haptoglobin (*HP*) expression and total cholesterol.** We observed strong evidence of colocalization, with posterior probability (PP) = 0.997, between Tier 1 ancestry-specific eQTLs for *HP* and GWAS associations from PAGE for total cholesterol. **a**, Regional plot showing GWAS results for total cholesterol in PAGE. The labeled eQTLs comprise the 95%

credible set for *HP*: **rs7200153** (PP<sub>SNP</sub> = 0.519) and **rs5471** (PP<sub>SNP</sub> = 0.481). Variants are colored based on LD with respect to **rs7200153**, which had the lowest GWAS p-value in PAGE. **b**, Plot of test statistics for variant effects on gene expression (eQTL z-scores) against effects on total cholesterol observed in PAGE (GWAS z-scores).



**Extended Data Fig. 6 | Summary of cross-ancestry TWAS results in UK Biobank.** TWAS of 22 blood-based biomarkers and quantitative traits was performed using GWAS summary statistics from African and European ancestry subjects in the UK Biobank. We applied GTEX v8 whole blood models and GALA/SAGE models trained in African Americans (AA) to each set of summary statistics. The number of genes with false discovery rate (FDR) < 0.05 detected in ancestry matched and ancestry discordant analyses is summarized for TWAS of

**a**, African (AFR) and **b**, European (EUR) ancestry subjects. Correlation between TWAS z-scores for statistically significant associations detected in **c**, UKB EUR and **d**, UKB AFR are shown for genes that were present in GTEX and GALA/SAGE models. Association between TWAS z-scores from each model is visualized by a linear regression line with shaded 95% confidence intervals. Genes highlighted in orange had FDR < 0.05 using ancestry-matched models but did not reach nominal significance (TWAS p-value > 0.05) using ancestry discordant models.

## Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

Data collection No software was used for data collection.

Data analysis

Open source codes used in this study:

GTEEx FASTQTL docker image. Available at <https://github.com/broadinstitute/gtex-pipeline/tree/master/ctl>  
 GCTA Version 1.93.1beta. Available at <https://cnsgenomics.com/software/gcta>  
 CAVIAR v2.2. Available at <https://github.com/fhormoz/caviar>  
 PESCA version 0.3-beta. Available at <https://github.com/huwenboshi/pesca>  
 METASOFT v2.0.1. Available at <http://genetics.cs.ucla.edu/meta>  
 PredictDB v7 pipeline. Available at [https://github.com/hakyimlab/PredictDB\\_Pipeline\\_GTEEx\\_v7](https://github.com/hakyimlab/PredictDB_Pipeline_GTEEx_v7)  
 MetaXcan version 0.7.4. Available at <https://github.com/hakyimlab/MetaXcan>  
 LDAK version 5.1 Available at <https://dougsspeed.com/ldak/>  
 COLOC R package version 5.1 Available at <https://cran.r-project.org/web/packages/coloc/index.html>  
 Data visualization: ggplot2 (version 3.3.5), colorspace (version 2.0-2), ggrepel (version 0.9.1), ggfittext (version 0.9.1), cowplot (version 1.1.1); parts of Figure 1 and Supplementary Figure 1 were generated using BioRender (<https://biorender.com>)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

## Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

TOPMed WGS and RNA-seq data from GALA II and SAGE are available on dbGaP under accession number phs000920.v4.p2 and phs000921.v4.p1, respectively. TOPMed WGS data from SAPHIRE are available under the dbGaP accession number phs001467.v1.p1. Individual-level normalized gene expression data for from GALA II and SAGE, models for performing transcriptome-wide association studies (TWAS), and eQTL summary statistics based on GALA II and SAGE are freely available from <https://doi.org/10.5281/zenodo.7735723>

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- Life sciences       Behavioural & social sciences       Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	2,733 individuals with available RNA samples with paired whole genome and RNA sequencing data were used.
Data exclusions	<p>GALA II and SAGE recruitment: Participants were excluded if they reported any of the following: (1) 10 or more pack-years of smoking; (2) any smoking within 1 year of recruitment date; (3) history of lung diseases other than asthma (cases) or chronic illness (cases and controls); or (4) pregnancy in the third trimester.</p> <p>Analyses in this manuscript were restricted to participants with whole genome sequencing (WGS) and RNA sequencing data generated by the NHLBI Trans-Omics for Precision Medicine (TOPMed) Program.</p> <p>Samples were excluded if they failed sample quality control standards as described in the GTEx v8 documentation. Sample-level QC included removal of RNA samples with RIN &lt; 6, genetically related samples (equal or more related than third degree relative), and sex-discordant samples based on reported sex and their XIST and RPS4Y1 gene expression profiles. Count distribution outliers were detected as follows: (i) Raw counts were normalized using the trimmed mean of M values (TMM) method in edgeR60 as described in GTEx v8 protocol. (ii) The log2 transformed normalized counts at the 25th percentile of every sample were identified (countq25). (iii) The 25th percentile (Q25) of countq25 was calculated. (iv) Samples were removed if their countq25 was lower than -4 as defined by visual inspection.</p>
Replication	We assessed the out-of-sample performance of our gene expression prediction models using an independent African American study (SAPHIRE) with RNA-seq data and confirmed our models generated prediction with higher correlation to measured gene expression levels than existing models.
Randomization	GALA II and SAGE are observational studies and did not have an intervention or randomization component. To account for hidden confounding factors in RNA-seq data such as batch effects, technical and biological variation in the sample preparation, and sequencing and/or data processing procedures, latent factors were estimated using the Probabilistic Estimation of Expression Residuals (PEER) method. Analyses adjusted for 50 to 60 PEER factors in analyses stratified by self-identified race/ethnicity and genetic ancestry. Inverse-normalized gene expression was regressed on PEER factors, and the residuals were used in downstream analyses. To ensure robust results, additional covariates included age at blood draw, sex, asthma case-control status, and the first 5 genetic ancestry principal components.
Blinding	Genomic DNA samples extracted from whole blood were sequenced as part of the Trans-Omics for Precision Medicine (TOPMed) whole genome sequencing (WGS) program and the Centers for Common Disease Genomes of the Genome Sequencing Program. WGS was performed at the New York Genome Center and Northwest Genomics Center WGS by researchers and technicians who were blinded to race/ethnicity and genetic ancestry of the participants, as well as their asthma case/control status. The researchers who carried out eQTL mapping and heritability analyses for this manuscript had no influence on how WGS and RNA-seq data were collected or processed.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.



## Materials &amp; experimental systems

## Methods

n/a	Involvement
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

n/a	Involvement
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

## Human research participants

Policy information about [studies involving human research participants](#)

## Population characteristics

The Genes-environments and Admixture in Latino Americans II (GALA II) study and the Study of African Americans, Asthma, Genes & Environments (SAGE) include African American, Puerto Rican and Mexican American children between 8-21 years of age with or without physician-diagnosed asthma. Briefly, participants were eligible if they were 8-21 years of age and identified all four grandparents as Latino for GALA II or African American for SAGE.

## Recruitment

Children were enrolled as a part of the ongoing Genes-environments & Admixture in Latino Americans (GALA II) case-control study. From July 2008 through November 2011, children were recruited from five centers (Chicago, Illinois; Bronx, New York; Houston, Texas; San Francisco Bay Area, California; and Puerto Rico) using a combination of community- and clinic-based recruitment. Participants were eligible if they were 8–21 years of age and all four grandparents self-identified as Latino. Asthma cases were defined as participants with a history of physician diagnosed asthma and the presence of two or more symptoms of coughing, wheezing, or shortness of breath in the 2 years preceding enrollment. Healthy controls were recruited from the community and clinics with the same catchment area as cases. Controls were defined as participants with no reported history of asthma, lung disease, or chronic illness over their lifetime, and no reported symptoms of coughing, wheezing or shortness of breath in the last two years. Controls were frequency matched on age (within 1 year), sex, and study center. Participants were excluded if they reported any of the following: (1) 10 or more pack-years of smoking; (2) any smoking within 1 year of recruitment date; (3) history of lung diseases other than asthma (cases) or chronic illness (cases and controls); or (4) pregnancy in the third trimester. All local institutional review boards approved the study, and all participants/parents provided appropriate written assent/consent.

The participant's country of birth, parents' country of birth, and self-reported country of birth of all four grandparents were used to determine country or region of origin. Firstly, for participants born outside the U.S., their origin was determined by their country of birth. For U.S.-born participants, origin was determined by their parents' country of birth (or grandparent's country of origin if their parents were born in the U.S.). For those with missing information for one parent, origin defaulted to the origin of the known parent. Participants were then classified as Puerto Rican, Mexican, South American, Central American, non-Puerto Rican Caribbean, or mixed Latino (for individuals of multiple origins who did not fit in the prior mentioned categories).

The Study of African Americans, Asthma, Genes, and Environments (SAGE II) is the largest ongoing gene-environment interaction study of asthma in African American children in the USA. SAGE II was initiated in 2006 and recruited participants with and without asthma until 2013 through a combination of clinic- and community-based recruitment centers in the San Francisco Bay Area. Institutional review boards of participant centers approved the study, and all participants or, for participants 17 or younger, their parents provided written informed consent. Participants 17 or younger also provided age-appropriate assent. Asthma cases were defined as individuals with a history of physician-diagnosed asthma and asthma controller or rescue medication use within the last 2 years and report of symptoms. Participants were eligible if they were 8–21 years of age and self-identified as African American and had four African American grandparents. Study exclusion criteria included the following: (1) any smoking within 1 year of the recruitment date, (2) 10 or more pack-years of smoking, (3) pregnancy in the third trimester, and (4) history of lung diseases other than asthma (cases) or chronic illness (cases and controls).

## Ethics oversight

The local institutional review board from the University of California San Francisco Human Research Protection Program approved the studies (IRB# 10-02877 for SAGE and 10-00889 for GALA II). All subjects and their legal guardians provided written informed consent.

Note that full information on the approval of the study protocol must also be provided in the manuscript.