Gene expression in African Americans and Latinos reveals ancestry-specific patterns of genetic architecture

Angel C.Y. Mak[1†*], Linda Kachuri[2†], Donglei Hu[1], Celeste Eng[1], Scott Huntsman[1], Jennifer R. Elhawary[1], Namrata Gupta[3], Stacey Gabriel[3], Shujie Xiao[4], Hongsheng Gui[4], Kevin L. Keys[1,5§], Akinyemi Oni-Orisan[6,7,8], Sam S. Oh[1], Max A. Seibold[9,10], José R. Rodríguez-Santana[11], Michael LeNoir[12], L. Keoki Williams[4,13], Luisa N. Borrell[14], Christopher R. Gignoux[15], Noah A. Zaitlen[16,17], Esteban G. Burchard[1,7‡], Elad Ziv[1,8,18‡]

1. Department of Medicine, University of California San Francisco, San Francisco, CA, USA
2. Department of Epidemiology & Biostatistics, University of California San Francisco, San Francisco, CA, USA
3. Broad Institute of MIT and Harvard, Cambridge, MA, USA
4. Center for Individualized and Genomic Medicine Research, Henry Ford Health System, Detroit, MI, USA
5. Berkeley Institute for Data Science, University of California, Berkeley, CA, USA
6. Department of Clinical Pharmacy, University of California San Francisco, San Francisco, CA, USA
7. Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, CA, USA
8. Institute for Human Genetics, University of California San Francisco, San Francisco, CA, USA
9. Department of Pediatrics, National Jewish Health, Denver, CO, USA
10. Division of Pulmonary Sciences and Critical Care Medicine, University of Colorado School of Medicine, Aurora, CO, USA
11. Centro de Neumología Pediátrica, San Juan, Puerto Rico
12. Bay Area Pediatrics, Oakland, CA, USA
13. Department of Internal Medicine, Henry Ford Health System, Detroit, MI, USA
14. Department of Epidemiology and Biostatistics, Graduate School of Public Health and Health Policy, City University of New York, New York, NY, USA
15. Colorado Center for Personalized Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, USA
16. Department of Neurology, University of California Los Angeles, Los Angeles, CA, USA
17. Department of Computational Medicine, University of California Los Angeles, Los Angeles, CA, USA
18. Division of General Internal Medicine, Helen Diller Family Comprehensive Cancer Center, University of California San Francisco, San Francisco, CA, USA

[§] Current address: Department of Bioinformatics and Computational Biology, Ambys Medicines, San Francisco, CA, USA

[†] These authors contributed equally to this work

36 ‡ These authors jointly supervised this work

37 Corresponding authors:

38 Angel C. Y. Mak
39 Department of Medicine
40 University of California San Francisco
41 Box 2911, 1550 4th Street,
42 San Francisco, CA 94143

43 Tel: 415-514-9931
44 Email: ChiYan.Mak@ucsf.edu

45 Esteban G. Burchard
46 Department of Medicine
47 University of California San Francisco
48 Box 2911, 1550 4th Street,
49 San Francisco, CA 94143

50 Tel: 415-514-9677
51 Email: Esteban.Burchard@ucsf.edu

52

ABSTRACT

We explored the role of genetic ancestry in shaping the genetic architecture of whole blood gene expression using whole genome and RNA sequencing data from 2,733 African American and Hispanic/Latino children. We find that heritability of gene expression significantly increases with greater proportion of genome-wide African ancestry and decreases with higher levels of Indigenous American ancestry. Fine-mapping of expression quantitative trait loci (eQTLs) in individuals with predominantly African or Indigenous American ancestry revealed ancestry-specific eQTLs in over 30% of heritable genes. We leveraged our data to train genetically derived transcriptome prediction models, which identified significantly more associated genes when applied to 28 traits from a multi-ancestry population. Our findings underscore the importance of increasing representation from ancestrally diverse populations in genomic studies to enable new discoveries and ensure their equitable translation.

66 # INTRODUCTION

67 Gene expression has been extensively studied as a trait affected by genetic variation in humans[1].
68 Expression quantitative trait loci (eQTLs) have been identified in most genes[2–4] and extensive
69 analyses across multiple tissues have demonstrated both tissue-specific and shared eQTLs[2].
70 Genome-wide association studies (GWAS) tend to identify loci that are enriched for eQTLs[5].
71 Colocalization of eQTLs with GWAS has become an important element of identifying causal genes
72 and investigating the biology underlying genetic susceptibility to disease[6]. More recently,
73 transcriptome-wide association studies (TWAS) have been developed to systematically leverage
74 eQTL data by imputing transcriptomic profiles in external datasets, which has led to the discovery
75 of trait-associated genes that were often missed by GWAS[7,8].

76 GWAS have identified thousands of loci for hundreds of diseases and disease-related phenotypes
77 in human populations[9]. However, non-European ancestry populations are significantly under-
78 represented in GWAS[10,11] and in studies of gene expression and eQTLs. We and others have shown
79 that gene expression prediction models trained in predominantly European ancestry reference
80 datasets, such as the Genotype-Tissue Expression (GTEx) project[2], have substantially lower
81 accuracy to predict gene expression levels when applied to populations of non-European
82 ancestry[3,12,13]. The importance of having ancestry-matched training datasets for prediction
83 accuracy is also reflected by the limited cross-population portability of other multi-SNP prediction
84 models, such as polygenic risk scores (PRS)[14,15]. Therefore, the limited diversity in genetic
85 association studies and reference datasets is a major obstacle for applying existing integrative
86 genomic studies to non-European populations.

87 To address this gap, we leveraged whole genome and RNA sequencing data from 2,733 African
88 American and Latino children from the Genes-environments and Admixture in Latino Americans
89 (GALA II) study and the Study of African Americans, Asthma, Genes, and Environments (SAGE) to
90 characterize the genetic architecture of whole blood eQTLs. The diversity within the GALA II/SAGE
91 population enabled us to evaluate how genetic ancestry relates to the heritability of gene
92 expression, and systematically quantify the prevalence of ancestry-specific eQTLs. Lastly, we

93 developed a powerful set of TWAS models to facilitate genetic association analyses in multi-
94 ancestry populations.

95 RESULTS

96 *Demographic characteristics of GALA II and SAGE participants*

97 We analyzed data from a total of 2,733 participants from the GALA II and SAGE asthma case-
98 control studies, including 757 self-identified African Americans (AA), 893 Puerto Ricans (PR), 784
99 Mexican Americans (MX), and 299 other Latinos (LA) who did not self-identify as Mexican
100 American or Puerto Rican (Table 1, Table S1). Participants identified all four grandparents as
101 Latino for GALA II or African American for SAGE. The median age of the participants varied from
102 13.2 (PR) to 16.0 (AA) years old. About 50% of the participants were female and 45% (MX) to 62%
103 (PR) had physician-diagnosed asthma. For each participant we estimated genome-wide genetic
104 ancestry (global ancestry) proportions, visualized in Figure 1. Median global African ancestry was
105 highest in AA (82.6%), followed by PR (19.7%), and lowest in MX (3.5%).

106 *Variability of gene expression explained by common genetic variation increases with*
107 *African ancestry*

108 We compared the heritability (h^2) and genetic variance (V_G) of whole blood gene expression
109 across self-identified race/ethnicity groups (AA, PR, MX) and populations defined based on
110 genetic ancestry. V_G corresponds to the observed variance in transcriptome levels attributed to
111 common genetic variation (minor allele frequency [MAF] ≥0.01) within the *cis*-region (see
112 Methods). Heritability (h^2) was calculated as the ratio of V_G to total phenotypic variance. There
113 was an overarching trend whereby increasing proportion of African ancestry was positively
114 associated with h^2 and V_G (Figure 2a-b, Table S2). *Cis*-heritability was significantly higher in AA
115 (h^2=0.158) compared to PR (h^2=0.135; Wilcoxon rank sum test: $p=2.2\times10^{-50}$) and MX (h^2=0.123;
116 $p=3.3\times10^{-134}$), as well as PR compared to MX ($p=2.2\times10^{-25}$).

117 In examining genetic ancestry, we considered both global ancestry and local ancestry. Global
118 ancestry is defined as the genome-wide average ancestry for each individual and local ancestry is

119 defined as the number of alleles (0, 1 or 2) derived from distinct ancestral populations at a given
120 locus. The same pattern was observed in the distribution of h^2 and V_G when we assigned
121 participants to high global African ancestry (AFR_{high}, AFR > 50%) and low global African ancestry
122 (AFR_{low}, AFR < 10%) groups (Table S1-2, Figure S2c-d). *Cis*-heritability and V_G were significantly
123 higher in AFR_{high} compared to AFR_{low} groups (Wilcoxon rank sum test: $p=9.6\times10^{-126}$ for h^2; V_G:
124 $p=7.6\times10^{-106}$ for V_G). The opposite was observed for Indigenous American (IAM) ancestry, with
125 higher h^2 and V_G in IAM_{low} (<10% IAM ancestry) compared to IAM_{high} (≥50% IAM ancestry) groups
126 ($p=3.1\text{x}10^{-103}$ for h2; $p=4.5\text{x}10^{-159}$ for V_G).

127 Next, we refined the observed relationship with heritability by examining ancestry at the locus
128 level. For each gene, individuals with homozygous local African ancestry (AFR/AFR) were
129 compared to those with heterozygous local African and European ancestry (AFR/EUR) at the
130 transcription start site. Consistent with global ancestry, h^2 and V_G were significantly higher in
131 AFR/AFR homozygotes compared to AFR/EUR and lower in IAM/IAM compared to IAM/EUR
132 heterozygotes (Figure 2e-h). Within the pooled GALA/SAGE II population, V_G was significantly
133 higher in AFR/AFR compared to AFR/EUR group (Wilcoxon rank sum test: $p=2.00\times10^{-7}$; Table S3).
134 Conversely, the V_G in IAM/IAM was significantly lower compared to EUR/EUR ($p=1.76\times10^{-7}$ for
135 Pooled;) or IAM/EUR ($p=1.62\times10^{-8}$ for Pooled; Table S3). All associations remained statistically
136 significant in analyses stratified by self-reported race/ethnicity (Table S3; $p=5.70\text{x}10^{-5}$ for AFR/AFR
137 vs. AFR/EUR in AA; $p=2.98\text{x}10^{-4}$ for IAM/IAM vs. EUR/EUR in MX; $p=7.67\text{x}10^{-11}$ for IAM/IAM vs.
138 IAM/EUR in MX).

139 As a parallel approach for evaluating the role of genetic ancestry in the context of whole blood
140 gene expression, we tabulated the number of heritable genes for which global and/or local
141 ancestry was significantly associated (FDR<0.05) with transcript levels. The proportion of global
142 AFR ancestry was associated with the expression of 326 (2.4%) and 589 (4.5%) of heritable genes
143 in AA and PR, respectively (Table S4). Associations with local, but not global AFR ancestry were
144 more common (8.9% in AA; 10.9% in PR). Few genes were associated with both global and local
145 AFR ancestry (1.5% in AA and 2.5% in PR). Among genes associated with both global and local AFR

146 ancestry in AA, global AFR ancestry explained a median of 1.8% of variation in gene expression,
147 while local AA accounted for 3.8% of variation (Figure S1). Local IAM ancestry was significantly
148 associated with the expression of 9.8% of genes in MX, compared to 2.8% for global. The variation
149 in gene expression accounted for by local IAM ancestry was also higher (3.5%) than global (1.8%)
150 in genes associated with both global and local IAM ancestry.

151 *Identification of ancestry-specific eQTLs*

152 We next sought to understand patterns of eQTLs in the GALA and SAGE datasets and identify
153 potential novel population-specific eQTLs. We identified 19,568 genes with at least one eQTL
154 (eGenes) in the pooled samples and 17,351, 16,980, and 15,975 eGenes in African American,
155 Puerto Rican, and Mexican American samples, respectively (Table S5, Figure S2).

156 To characterize ancestry-related differences in the genetic determinants of gene expression, we
157 built a framework to identify eQTLs with potential ancestry-specific effects on gene expression
158 levels. We then classified them based on their ancestry-specific properties, including allele
159 frequency and effect size (Figure 3a). We termed these eQTLs ancestry-specific eQTLs (anc-
160 eQTLs). To identify anc-eQTLs, we compared the lead eQTL signals of *cis*-heritable genes (Table
161 S6) between high and low ancestry groups (AFR_{high} vs AFR_{low}; IAM_{high} vs IAM_{low}). We identified the
162 lead eQTL signal of each gene in each ancestry group (AFR_{high}, AFR_{low}, IAM_{high} and IAM_{low}) by fine-
163 mapping 95% credible set of causal variants using CAVIAR[16].

164 We divided anc-eQTLs into three tiers of ancestry-specificity in multiple steps, first based on
165 overlap in the 95% credible sets for each gene, followed by analyses accounting for differences in
166 eQTL effect size, allele frequency, and linkage disequilibrium (LD) between ancestry groups (see
167 Methods). Over 30% (n=2,961) and 8% (n=679) of heritable genes had at least one AFR_{high}-eQTL
168 and IAM_{high}-eQTL, respectively. We define the three tiers of ancestry-associated eQTLs below.

169 Tier 1 anc-eQTLs had non-overlapping 95% credible sets because they were only common (MAF
170 ≥ 0.01) in high ancestry groups and were thus considered to be the most ancestry specific. Over
171 28% (n=2,695) of genes contained at least one Tier 1 AFR_{high} lead eQTL, while 7% (n=562) of genes

172 contained a Tier 1 IAM_{high} eQTL (Figure 3a, left branch; Table S7). Tier 2 anc-eQTLs also had non-
173 overlapping 95% credible sets for each gene but were commonly segregating (MAF ≥ 0.01) in both
174 high and low ancestry groups. This class of anc-eQTLs was identified using PESCA[17], which
175 accounted for population-specific LD patterns. We identified 109 genes (1.1%) with eQTLs having
176 a posterior probability (PP) > 80% of being specific to AFR_{high} and 33 genes (0.4%) matching the
177 same criteria for IAM_{high}. Lastly, we found that over 50% of genes (5,058 for AFR and 5,355 for
178 IAM) contained eQTLs that were shared between high and low ancestry groups. Among these
179 shared signals, only a small proportion of eQTLs exhibited effect size heterogeneity, as assessed
180 using Cochran's Q test. We refer to these as Tier 3 anc-eQTLs (2.0% for AFR; 1.0% for IAM) and
181 note this tier has the lowest level of ancestry-specificity.

182 Next, we explored whether the ancestry-specific eQTLs were previously associated with
183 diseases/traits based on the NHGRI-EBI GWAS catalog[18]. We found 70 unique anc-eQTLs in 61
184 genes associated with 84 diseases/traits reported in 65 publications (Table S8, supplementary file
185 1). Representative examples of disease-associated anc-eQTLs are shown in (Figure 3b-d). A Tier 1
186 AFR_{high}-specific eQTL of the *ABAT* gene (rs1641047) was associated with systemic sclerosis (SSc)
187 in a study that compared European and African American SSc patients[19]. It is common in African
188 Americans in SAGE (MAF = 0.04) but rare in Mexican Americans in GALA II (MAF = 2×10^{-3}, Figure
189 3b) and Europeans (MAF = 2×10^{-4}, n=15,666). A Tier 2 IAM_{high}-specific eQTL of *SPSB2* gene
190 (rs11064437) was previously reported as a risk variant for colorectal cancer in East Asian
191 populations[20], likely due to the fact that it is present at higher frequencies in this population (East
192 Asian MAF = 0.26). In our dataset, the lead *SPSB2* eQTL in the IAM_{high} group (rs11064437) was
193 independent from that of the IAM_{low} group (rs6489739), as indicated by low LD (r^2=0.16, Figure
194 3d).A Tier 3 AFR_{high}-specific eQTL of *KCNK17* gene (rs34246110) was associated with type 2
195 diabetes in two independent studies performed in Japanese and multi-ancestry (European,
196 African American, Hispanic and Asian) populations[21,22]. The detection of this association in
197 multiple populations was consistent with the characteristics of Tier 3 anc-eQTLs in that they share
198 lead eQTL signals with other populations but likely exhibit heterogenous effect sizes.

199 *Gene expression prediction models from admixed populations increase power for*
200 *gene discovery*

201 We generated gene expression imputation models from GALA II and SAGE following the PrediXcan
202 approach[7]. We used the pooled population (n=2,733) to generate models with significant
203 prediction (see Methods) for 11,830 heritable genes with mean cross-validation (CV) R^2=0.157
204 (Table S9 and Figure S3). We also generated population-specific models for African Americans
205 (10,090 genes, CV R^2=0.180), Puerto Ricans (9,611 genes, CV R^2=0.163), and Mexican Americans
206 (9,084 genes, CV R^2=0.167). In sensitivity analyses adjusting for local ancestry, we did not observe
207 gains in predictive performance (Table S10). For genes with local ancestry estimates, adjusting for
208 local ancestry did not result in improved model performance (AA: CV R^2=0.177; PR: CV R^2=0.154;
209 MX: CV R^2=0.159).

210 Our TWAS models were successfully validated in the Study of Asthma Phenotypes and
211 Pharmacogenomic Interactions by Race-Ethnicity (SAPPHIRE)[23], an independent adult population
212 of African Americans (Figure S4). As expected, both pooled and AA-specific TWAS models, which
213 contained the highest number of ancestry-matched individuals, had significantly highest gene
214 expression prediction accuracy in SAPPHIRE compared to other models (correlation R^2=0.026 for
215 AA compared to 0.023 for PR [Wilcoxon rank sum test: $p=2.39 \times 10^{-5}$], 0.020 for MX [$p=1.19 \times 10^{-}$
216 ^{17} and 0.015 for GTEx v8 [$p=5.43 \times 10^{-85}$] models).

217 To evaluate the potential of TWAS models built from GALA II and SAGE pooled samples (hereafter
218 referred to as GALA/SAGE models) to improve gene discovery in admixed populations, we applied
219 our models to GWAS summary statistics for 28 traits from the multi-ancestry Population
220 Architecture using Genomics and Epidemiology (PAGE) study[24] and conducted parallel analyses
221 using TWAS models based on GTEx v8[2,7] and the Multi-Ethnic Study of Atherosclerosis (MESA)[3].
222 GTEx v8 whole blood models are based on 670 subjects of predominantly European ancestry
223 (85%) with median age between 50-60 years old[2]. MESA models impute monocyte gene
224 expression[3] based on a sample of African American and Hispanic/Latino individuals ($MESA_{AFHI}$:

225 n=585). As such, MESA more closely resembles the ancestry composition of our population as
226 well as subjects included in the PAGE study.

227 The total number of genes examined was 39% to 82% higher in GALA/SAGE compared to GTEx
228 (n=7,249) and $MESA_{AFHI}$ (n=5,555), with an intersection of 3,146 genes (Figure 4a, left). Restricting
229 to genes shared across all three models, CV R^2 was significantly higher in GALA/SAGE compared
230 to GTEx ($p=5.1\times10^{-159}$) and $MESA_{AFHI}$ ($p=9.9\times10^{-65}$), which is expected based on the substantially
231 larger sample size of GALA/SAGE (Figure 4a). More importantly, we observed the same trend with
232 an intersection of 3,003 genes using GALA/SAGE TWAS models built in African Americans only
233 (n=757), which was a population of comparable sample size (Figure 4b). Association results across
234 28 PAGE traits demonstrate that TWAS using GALA/SAGE (pooled) models identified a larger
235 number of significant gene-trait pairs (n=380, FDR<0.05), followed by $MESA_{AFHI}$ (n=303), and GTEx
236 (n=268), with only 30 genes (35 gene-trait pairs) significant in all three analyses (Figure 4c).

237 GALA/SAGE models yielded a larger number of associated genes than MESA in 80% of analyses
238 (binomial test: p=0.012) and 79% compared to GTEx (p=0.019). Of the 330 genes with FDR<0.05
239 in GALA/SAGE, 143 (43%) were not present in GTEx and 199 (60%) were not present in $MESA_{AFHI}$.
240 For genes that were significant in at least one TWAS, z-scores in GALA/SAGE were highly
241 correlated with GTEx (Figure 4c; r=0.74, $p=3.5\times10^{-64}$) and $MESA_{AFHI}$ (Figure 4d; r=0.55, p=
242 $p=8.5\times10^{-27}$), suggesting that most genes have concordant effects even if they fail to achieve
243 statistical significance in both analyses (Figure 4c). Despite the higher correlation with GTEx TWAS
244 z-scores, we observed a higher proportion of gene-trait pairs significant (FDR<0.05) in GALA/SAGE
245 that were not even nominally associated (p<0.05) in GTEx (33%), compared to 18% in $MESA_{AFHI}$.

246 The high-density lipoprotein (HDL) cholesterol trait exhibited one of the largest differences in
247 TWAS associations, with over 60% more significant genes identified using GALA/SAGE models
248 (n=29) than GTEx predictions (n=11; Figure 4c). Models for several associated genes, including
249 *CETP* which has been consistently implicated in HDL cholesterol transport, were not available in
250 GTEx. The top-ranking HDL cholesterol-associated gene, *CD36* (z-score= -10.52, $p=6.9\times10^{-26}$) had
251 Tier 1 AFR_{high}-specific eQTLs (rs3211938) that were not present at an appreciable frequency in

252 populations with low African ancestry (MAF in European = 1.3 x 10^{-4}). The difference in MAF may
253 explain why *CD36* was not detected using GTEx (z-score=0.057, p=0.95), even though all 43
254 variants used in GTEx prediction model were available in PAGE summary statistics. In addition to
255 HDL cholesterol levels, *CD36* expression was also associated with levels of C-reactive protein (z-
256 score= 5.30, p=1.1×10^{-7}).

257 In addition to multi-ancestry analyses, we also demonstrate the utility of ancestry-specific gene
258 expression prediction models for the discovery of trait-gene association using TWAS. We
259 conducted a TWAS of neutrophil counts in African ancestry subjects using population-specific
260 summary statistics from Chen et al.[25] and ancestry-matched GALA/SAGE models, which identified
261 139 genes. Eleven of the 139 genes identified were outside of the Duffy locus on 1q23 with
262 FDR<0.05 (Figure 5b). Novel findings included *CREB5* (p=1.5×10^{-14}), *DARS* (p=2.9×10^{-8}), *CD36*
263 (p=1.1×10^{-5}), *PPT2* (p=1.3×10^{-5}), *SSH2* (p=4.7×10^{-5}), *TOMM5* (p=2.9×10^{-4}), and *ARF6* (p=3.4×10^{-4}).

264 DISCUSSION

265 Our comprehensive analysis in a uniquely large and diverse multi-racial/ethnic population
266 revealed new insight into the role of genetic ancestry in shaping the genetic architecture of whole
267 blood gene expression that may have broad implications for other complex traits. We found that
268 *cis*-heritability of gene expression increased with higher proportions of African ancestry. We also
269 demonstrated that in populations of mixed ancestry, *cis*-heritability was highest with increasing
270 local African ancestry. The consistency of this pattern in analyses considering global and local
271 ancestry, as well as in pooled and stratified analyses within populations (African Americans or
272 Puerto Ricans), suggests this is unlikely to be attributed to confounding by social or environmental
273 factors. Parallel analyses examining Indigenous American Ancestry revealed an inverse
274 relationship – with genetic variance and *cis*-heritability of gene expression decreasing in
275 individuals with higher levels Indigenous American ancestry compared to European ancestry.

276 These observations are consistent with the overall pattern of heterozygosity of these populations.
277 African populations consistently show the highest heterozygosity since all other populations

278 passed through a bottleneck during their migration out of Africa[26,27]. Indigenous American
279 populations have passed through additional bottlenecks which possibly date to, or prior to, the
280 settling of the Americas[28,29]. The finding that heritability tracks with heterozygosity across human
281 populations may be expected from fundamental assumptions that there are more segregating
282 causal variants in populations with higher heterozygosity[30]. However, to our knowledge, this has
283 not yet been systematically demonstrated for whole blood gene expression, particularly in a
284 sufficiently large racially/ethnically diverse sample size.

285 A second major finding of our work is that over 30% of heritable genes have ancestry-specific
286 eQTLs and these are predominantly comprised of Tier 1 variants that were rare (MAF < 0.01) or
287 not polymorphic in another population. Tier 2 ancestry eQTLs are a less striking, albeit interesting
288 class of variants which were present at a sufficient frequency (MAF>0.01) in both high and low
289 ancestry groups but did not map to the same eQTL signal. Tier 2 eQTLs could arise due to
290 differences in environmental effects on gene expression, multiple causal variants at the same
291 locus that are in different degrees of LD with each other, or *trans*-effects, which are contributed
292 by genetic variation outside of the *cis*-region. Additional research is needed to elucidate the
293 underlying reasons for more subtle ancestry-driven differences in eQTL effects.

294 Given genetic architecture of complex traits is, to variable degree, mirrored or driven by the
295 genetics of gene expression[31], our work has several important implications. Observations
296 regarding heritability suggest genetic prediction of complex traits in populations with at least 50%
297 global African ancestry should be at least as accurate, if not more effective, than in populations
298 with low proportion of African ancestry. However, for most complex traits the accuracy of
299 polygenic prediction models in populations of predominantly African ancestry significantly lags
300 behind other populations[15], particularly those of European ancestry. Although there are caveats
301 associated with using heritability to gauge the future potential of genetic risk prediction models[32],
302 our results imply that current deficits in predictive performance in admixed populations are likely
303 to be a consequence of insufficient sample sizes and systematic underrepresentation. The
304 substantial prevalence of ancestry-specific eQTLs implies that analytic approaches alone will yield

305 limited improvements in the generalizability of genetic prediction models, including European-
306 ancestry derived polygenic risk scores (PRS). For instance, fine-mapping approaches that rely on
307 accounting for differential LD tagging to identify causal variants will recover some deficits in
308 predictive performance, but will not compensate for missing genetic predictors. Our results
309 reinforce the conclusion that developing truly generalizable genetic prediction models requires
310 capturing the full spectrum of genetic variance in non-European populations. As such, access to
311 sufficiently large and racially/ethnically and ancestrally diverse populations is the main rate
312 limiting step.

313 Previous work by Gay *et al.* assessed the amount of variance in gene expression explained by
314 global and/or local ancestry[33]. It was reported that local ancestry explains at least 7% of the
315 variance in residual expression for 1% of expressed genes in an analysis of 117 admixed individuals
316 from GTEx. Using sample sizes ranging from 757 (AA) to 893 (PR) in GALA II/SAGE, we found that
317 local ancestry was a significant predictor of transcript levels for at least 10% of heritable genes,
318 explaining between 2.1% (PR) and 5.1% (AA) of residual variance. Consistent with Gay et al., we
319 observed that local ancestry explains a larger proportion of variance in gene expression corrected
320 for global ancestry. However, contrary to their findings, we found that, on average, local ancestry
321 also explains more variance in residual gene expression than global ancestry.

322 The importance of comprehensively accounting for genetic determinants of trait variation in
323 multi-ethnic populations is illustrated in our TWAS results for 28 traits from the PAGE study. TWAS
324 models trained in the racially/ethnically and ancestrally diverse GALA II/SAGE study identified
325 significantly more trait-associated genes than GTEx and MESA respectively. This is consistent with
326 the findings of Geoffroy et al.[13] using GTEx and MESA models, as well as other observations[12,13]
327 that ancestry-matched models improve power for gene discovery in admixed populations. Over
328 40% of significantly associated TWAS genes detected using GALA/SAGE models were not available
329 in GTEx, which underscores how biologically meaningful associations may be overlooked in
330 studies that exclusively rely on European ancestry-based predictions. The top two HDL
331 cholesterol-associated genes, *CETP* in 16q13 and *CD36* in 7q21, with established effects on lipid

332 metabolism[34–37], were not detected in TWAS using GTEx due differences in eQTLs. The finding for
333 *CD36* is compelling since this gene was associated with multiple phenotypes and contains Tier 1
334 anc-eQTLs that are specific to individuals with high African ancestry, consistent with earlier
335 findings that evolutionary pressures have elevated genetic divergence at this locus[38,39]. *CD36*
336 encodes a transmembrane protein that binds many ligands, including collagen, thrombospondin,
337 and long-chain fatty acids, and serves as a negative regulator of angiogenesis[40]. Beyond lipid
338 metabolism, its most extensively studied function involves mediating the adherence of
339 erythrocytes infected with *Plasmodium falciparum*, the parasite responsible for severe
340 malaria[41,42].

341 An even more striking example of ancestry-specific genetic architecture involves the Duffy
342 antigen receptor gene (*ACKR1*) on 1q23, which is responsible for persistently lower white blood
343 cell and neutrophil counts in African ancestry individuals[43,44]. Common African-derived alleles at
344 this locus confer a selective advantage against *Plasmodium vivax* malaria and are extremely rare
345 in European populations. Expression of *ACKR1* could not be imputed using GTEx or MESA, but this
346 causal gene was captured by the pooled and AA-specific GALA/SAGE TWAS models. Furthermore,
347 we identified 11 genes associated with neutrophil counts outside of the Duffy locus, including
348 associations not reported in the source GWAS by Chen *et al*[25]. We replicated the *PSMD3* signal in
349 17q21[45], which was previously identified in African Americans, as well as *CREB5*, *SSH2*, and *PPT2,*
350 which were discovered in European ancestry populations[46]. We also identified several novel
351 genes that have not previously been linked to hematologic traits: *DARS1* in 2q31.1 modulates
352 reactivity to mosquito antigens[47], while *TOMM5* has been implicated in lipoprotein phospholipase
353 A2 activity[48].

354 Taken together, our TWAS findings suggest that models trained in racially/ethnically and
355 ancestrally diverse populations improve power for identifying trait-associated genes in shared
356 biological pathways. When applied to admixed populations, GALA II/SAGE imputation models
357 benefit from having more similar allele frequency profiles to the target datasets, such as PAGE,
358 as well as more accurate modeling of LD. However, we note that adjustment for local ancestry

359 did not improve the predictive performance of our TWAS models. Previous analyses of GTEx data
360 indicated that local ancestry adjustment increases power for discovery in cis-eQTL mapping[33,49],
361 however this has not been assessed explicitly within the TWAS framework. It is possible that the
362 slight decrease in CV R^2 occurs due to overadjustment, since local ancestry may serve as a proxy
363 for information already captured by genetic variants in the *cis*-region.

364 Interestingly, the correlation of TWAS summary statistics was higher with GTEx than $MESA_{AFHI}$,
365 which may partly reflect the lack of neutrophils present in monocyte gene expression in $MESA_{AFHI}$
366 compared to whole blood in GALA/SAGE and GTEx. However, the scarcity of discordant TWAS
367 findings (null in one analysis or significant associations with opposite directions of effect) or
368 missing genes between GALA II/SAGE and MESA, indicate that similarity in ancestry may partly
369 compensate for differences in the cell type analyzed. Furthermore, GALA II/SAGE and GTEx
370 conducted whole-genome sequencing, whereas MESA TWAS models are based on imputed
371 genotype data, although we this effect should be limited as we filter to common variants. While
372 our study population is comprised of pediatric studies, TWAS analyses of medically relevant traits
373 and chronic conditions in PAGE identified more genes than adult-derived prediction models. This
374 implies that the power gained from ancestry-matched prediction and large sample size of the
375 training set may outweigh substantial differences in age between participants in our cohort and
376 participants in PAGE studies that range across adulthood.

377 In evaluating the contributions of our work, a few limitations should be acknowledged. Although
378 our analyses were limited to a single tissue, we note that whole blood is one of the most clinically-
379 informative and commonly-collected samples, and for over 60% of genes whole blood
380 transcriptomes can significantly capture expression levels in other tissues[50]. Thus, our
381 observations regarding the genetic architecture of whole blood eQTLs in non-European ancestry
382 populations are likely generalizable to other tissues. Our approach for ancestry-specific eQTLs
383 may result in an underestimation of the number of these loci. We assumed that each gene had
384 one causal eQTL locus and focused all comparisons on the corresponding 95% credible set. This
385 assumption is likely violated for genes with multiple independent eQTLs, which would limit our

386 ability to assess the ancestry-specificity of all signals. In general, we believe this is a conservative
387 assumption that would lead us to potentially miss some ancestry-specific eQTLs. Detection of Tier
388 2 anc-eQTLs by the PESCA method relies on having regions that are approximately LD
389 independent in both populations to estimate the proportion of causal variants. This estimate may
390 be biased if there is residual LD between regions, which is particularly a challenge in admixed
391 populations with longer-range LD. Lastly, our comparison of TWAS models may be slightly biased
392 against GTEx since we did not apply MASHR models, which predict a larger number of genes using
393 fine-mapped eQTLs[51]. We chose to compare with elastic net GTEx models instead because
394 GALA/SAGE TWAS models were developed using the same analytic pipeline.

395 Despite these limitations, our study leveraged a uniquely large and diverse sample of 2,733
396 African American and Latino youths to advance our understanding of the interplay between
397 genetic ancestry and genetic determinants of gene expression. Our systematic approach for
398 evaluating the ancestry-specificity of whole blood eQTLs revealed that most of this signal is due
399 to Tier 1 anc-eQTLs, which are only polymorphic in predominantly AFR and IAM ancestry
400 populations. These ancestry-specific eQTLs are present in a large fraction of genes, which
401 highlights the need for larger genomic studies in globally representative populations to detect
402 functional genetic variants and evaluate their role in disease susceptibility. We also developed
403 genetic prediction models of whole blood transcriptomes and demonstrated that they covered
404 more genes and, in TWAS based on multi-ethnic populations, performed better than similar
405 models built in European populations. In summary, the results of our study have implications not
406 only for the genetic prediction of gene expression, but for other complex traits, and ultimately
407 for the equitable translation of precision medicine efforts to all populations worldwide.

408 # METHODS

409 ## *Study population*

410 This study examined African American, Puerto Rican and Mexican American children between 8-
411 21 years of age with or without physician-diagnosed asthma from the Genes-environments and
412 Admixture in Latino Americans II (GALA II) study and the Study of African Americans, Asthma,
413 Genes & Environments (SAGE). The inclusion and exclusion criteria are previously described in
414 detail[52,53]. Briefly, participants were eligible if they were 8-21 years of age and identified all four
415 grandparents as Latino for GALA II or African American for SAGE. Study exclusion criteria included
416 the following: 1) any smoking within one year of the recruitment date; 2) 10 or more pack-years
417 of smoking; 3) pregnancy in the third trimester; 4) history of lung diseases other than asthma (for
418 cases) or chronic illness (for cases and controls).

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

422 ## *Whole genome sequencing data and processing*

423 Genomic DNA samples extracted from whole blood were sequenced as part of the Trans-Omics
424 for Precision Medicine (TOPMed) whole genome sequencing (WGS) program[54] and the Centers
425 for Common Disease Genomes of the Genome Sequencing Program. WGS was performed at the
426 New York Genome Center and Northwest Genomics Center on a HiSeq X system (Illumina, San
427 Diego, CA) using a paired-end read length of 150 base pairs (bp), with a minimum of 30x mean
428 genome coverage. DNA sample handling, quality control, library construction, clustering, and
429 sequencing, read processing and sequence data quality control are previously described in
430 detail[54]. All samples were jointly genotyped by the TOPMed Informatics Research Center. Variant
431 calls were obtained from TOPMed data freeze 8 VCF files generated based on the GRCh38
432 assembly. Variants with a minimum read depth of 10 (DP10) were used for analysis unless
433 otherwise stated.

434 *RNA sequencing data generation and processing*

435 Total RNA was isolated from PAXgene tube using MagMax™ for Stabilized Blood Tubes RNA
436 Isolation Kit (Applied Biosystem, P/N 4452306). Globin depletion was performed using
437 GLOBINcleasr™ Human (Thermo Fisher Scientific, cat. no. AM1980). RNA integrity and yield were
438 assessed using an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA).

439 Total RNA was quantified using the Quant-iT™ RiboGreen® RNA Assay Kit and normalized to
440 5ng/ul. An aliquot of 300ng for each sample was transferred into library preparation which was
441 an automated variant of the Illumina TruSeq™ Stranded mRNA Sample Preparation Kit. This
442 method preserves strand orientation of the RNA transcript. It uses oligo dT beads to select mRNA
443 from the total RNA sample. It is followed by heat fragmentation and cDNA synthesis from the RNA
444 template. The resultant cDNA then goes through library preparation (end repair, base 'A' addition,
445 adapter ligation, and enrichment) using Broad-designed indexed adapters substituted in for
446 multiplexing. After enrichment the libraries were quantified with qPCR using the KAPA Library
447 Quantification Kit for Illumina Sequencing Platforms and then pooled equimolarly. The entire
448 process is in 96-well format and all pipetting is done by either Agilent Bravo or Hamilton Starlet.

449 Pooled libraries were normalized to 2nM and denatured using 0.1 N NaOH prior to sequencing.
450 Flowcell cluster amplification and sequencing were performed according to the manufacturer's
451 protocols using the HiSeq 4000. Each run was a 101bp paired-end with an eight-base index
452 barcode read. Each sample was targeted to 50M reads. Data was analyzed using the Broad Picard
453 Pipeline which includes de-multiplexing and data aggregation.

454 RNA-seq reads were further processed using the TOPMed RNA-seq pipeline for Year 3 and Phase
455 5 RNA-seq data (supplementary file 2 obtained from https://topmed.nhlbi.nih.gov/sites/default/
456 files/TOPMed_RNAseq_pipeline_COREyr3.pdf) . Count-level data were generated using GRCh38
457 human reference genome and GENCODE 30 for transcript annotation.

458 Count-level quality control (QC) and normalization were performed following the Genotype-
459 Tissue Expression (GTEx) project v8 protocol (https://github.com/hakyimlab/PredictDB_Pipeline_

460 GTEx_v7). Sample-level QC included removal of RNA samples with RIN < 6, genetically related
461 samples (equal or more related than third degree relative), and sex-discordant samples based on
462 reported sex and their *XIST* and *RPS4Y1* gene expression profiles. Count distribution outliers were
463 detected as follows: (i) Raw counts were normalized using the trimmed mean of M values (TMM)
464 method in edgeR[55] as described in GTEx v8 protocol. (ii) The log2 transformed normalized counts
465 at the 25th percentile of every sample were identified ($count_{q25}$). (iii) The 25th percentile (Q25)
466 of $count_{q25}$ was calculated. (iv) Samples were removed if their $count_{q25}$ was lower than -4 as
467 defined by visual inspection (Figure S6).

468 *Estimation of hidden confounding factors*

469 To account for hidden confounding factors such as batch effects, technical and biological variation
470 in the sample preparation, and sequencing and/or data processing procedures, latent factors
471 were estimated using the Probabilistic Estimation of Expression Residuals (PEER) method[56].
472 Optimization was performed according to approach adopted by GTEx with the goal to maximized
473 eQTL discovery[57]. A total of 50 (for AA, PR, MX, pooled samples) and 60 (for AFR_{high}, AFR_{low},
474 IAM_{high}, IAM_{low}) PEER factors were selected for downstream analyses (Figure S7).

475 *Estimation of global and local genetic ancestry*

476 Genetic principal components (PCs), global and local ancestry, and kinship estimation on genetic
477 relatedness were computed using biallelic single nucleotide polymorphisms (SNPs) with a PASS
478 flag from TOPMed freeze 8 DP10 data as described previously[58,59]. Briefly, genotype data from
479 European (CEU), African (YRI) and Indigenous American (IAM) ancestral populations were used as
480 the reference panels for global and local ancestry estimation assuming three ancestral
481 populations. Global ancestry was estimated by ADMIXTURE[60] in supervised while local ancestry
482 was estimated by RFMIX version 2 with default settings[61]. Throughout this study, local ancestry
483 of a gene was defined as the number of ancestral alleles (0, 1, or 2) at the transcription start site.

484 Comparative analyses were performed based on two different sample grouping strategies, by
485 self-identified race/ethnicity or by global ancestry. Self-identified race/ethnicity included four

486 groups – African Americans (AA), Puerto Ricans (PR), Mexican Americans (MX), and the pooling
487 of AA, PR, MX and other Latinos (pooled). For groups defined by global ancestry, samples were
488 grouped into high (> 50%, AFR_{high} or IAM_{high}) or low (< 10%, AFR_{low} or IAM_{low}) global African or
489 Indigenous American ancestry (Figure S5). The sample size for each group is shown in Table S1.

490 *Cis-heritability of gene expression*

491 The genetic region of *cis*-gene regulation was defined by 1MB region flanking each side of the
492 transcription start site (*cis*-region). *Cis*-heritability (h^2) of gene expression was estimated using
493 unconstrained GREML[62] analysis (--reml-no-constrain), and estimation was restricted to common
494 autosomal variants (MAF ≥ 0.01). Inverse-normalized gene expression was regressed on PEER
495 factors, and the residuals were used as the phenotype for GREML analysis. Sex and asthma case-
496 control status was used as categorical covariates, while age at blood draw and the first 5 genetic
497 PCs were used as quantitative covariates. *Cis*-heritability was estimated separately for each self-
498 identified race/ethnicity group (AA, PR, MX and pooled) and groupings based on global (AFR_{high},
499 AFR_{low}, IAM_{high} and IAM_{low}) and local ancestry (described below). Differences in the distribution
500 of h^2 and genetic variance (V_G) between groups were tested using two-sided Wilcoxon tests.
501 These analyses included all genes and were not restricted to *cis*-heritable genes. For analyses
502 based on local ancestry, we also tested whether the proportion of genes with higher V_G in the
503 AFR/AFR group was higher than that in AFR/EUR group using a two-sided binomial test. Parallel
504 analyses were also conducted for Indigenous American ancestry (IAM/IAM vs. EUR/EUR and
505 IAM/IAM vs. IAM/EUR).

506 *Association of global and local ancestry with gene expression*

507 Methods from Gay *et al* (2020)[33] was modified to identify genes associated with global and local
508 ancestry (Figure S1). In step 1, inversed normalized gene expression was regressed on age, sex
509 and asthma status (model 0). In step 2, the residuals from model 0 were regressed on global
510 ancestry (model 1). In step 3, the residuals from model 1 were regressed on local ancestry (model
511 2) to identify genes that are associated with local ancestry. False discovery rate of 0.05 (R package
512 qvalue) was applied to step 2 and 3 separately to identify genes that were significantly associated

513 with global and/or local ancestry. Step 1 to step 3 were run separately for African and Indigenous
514 American ancestry. For heritable genes that were associated with global and/or local ancestry, a
515 joint model of regressing global and local ancestry from residuals from model 0 was also examined
516 to assess the percentage of variance of gene expression explained by global and/or local ancestry.

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

518 Raw gene counts were processed and eQTLs were identified using FastQTL[63] according to the
519 GTEx v8 pipeline (https://github.com/broadinstitute/gtex-pipeline). Age, sex, asthma status, first
520 5 genetic PCs, and PEER factors were used as covariates for FastQTL analysis. *Cis*-eQTLs were
521 defined as genetic variants that have nominal p values less than the nominal p value threshold
522 estimated from the permutation step of FastQTL. eGenes are defined as genes with at least one
523 eQTL.

524 Ancestry-specific eQTL (anc-eQTL) mapping was performed in participants stratified by high and
525 low global African and Indigenous American ancestry (see "Grouping samples by self-identified
526 race/ethnicity or global ancestry"). We developed a framework to identify anc-eQTLs by focusing
527 on the lead eQTL signal for each gene and compared these between high and low global ancestry
528 groups (AFR_{high} vs AFR_{low}; IAM_{high} vs IAM_{low}) (Figure 3a).

529 Anc-eQTLs were classified into three tiers as described below based on differential effect size and
530 allele frequency, while accounting for linkage disequilibrium (LD) between eQTLs (Figure 3a). For
531 every protein coding and heritable (GCTA h^2 LRT $p < 0.05$) eGene (Table S6), the lead eQTL signal
532 was identified using CAVIAR[16] assuming one causal locus (c=1). The 95% credible set of the high
533 and low global ancestry group were compared to determine if there was any overlap of the 95%
534 credible set (overlapping or non-overlapping lead eQTL signals).

535 For genes with non-overlapping lead eQTL signals (Figure 3a, left branch), eQTLs from the high
536 group were categorized into Tier 1 anc-eQTLs if they were common variants (MAF >= 0.01) in the
537 high group but rare or absent in the low group. If the eQTLs were common variants in both the
538 high and low groups, their specificity was further assessed using PESCA[17] for differential effect

539 size while accounting for LD between eQTLs. LD pruning on eQTLs with r2 > 0.95 was performed
540 prior to PESCA analysis. All eQTL pairs with r2 > 0.95 were identified in both the high and low
541 groups and only those pairs common to both groups were removed. In the PESCA analysis, each
542 eQTL obtained a posterior probability whether it was specific to the high group (PP_{high}), low group
543 (PP_{low}) or shared between the two groups (PP_{shared}). All eQTLs in the credible set of the high group
544 were required to obtain the highest posterior probability in the high group ($PP_{high} > PP_{low}$ and
545 $PP_{high} > PP_{shared}$) and pass the $PP_{high} > 0.8$ cutoff to be considered Tier 2 ancestry-specific eQTLs.

546 For genes with overlapping lead eQTL signals (Figure 3a, right branch), Cochran's Q statistics were
547 used to test for effect size heterogeneity between high and low ancestry groups. An eQTL with
548 significant heterogeneity in effect size was defined as having a p value of Cochran's Q smaller than
549 the bonferroni threshold (0.05/nGene, where nGene was the number of genes tested and was
550 5,058 for AFR and 5,355 for IAM ancestry, see Table S7 row C). Since we assume the 95% credible
551 set corresponds to a single lead eQTL signal, all eQTLs in the credible set of the high group were
552 required to have a significant heterogeneous effect size to be qualified into Tier 3 anc-eQTLs.

553 *Development of gene prediction models and transcriptome-wide association*
554 *analyses*

555 Gene prediction models for *cis*-gene expression were generated using common variants and
556 elastic net modeling implemented in the PredictDB v7 pipeline (https://github.com/hakyimlab/
557 PredictDB_Pipeline_GTEx_v7). Models were filtered by nested cross validation (CV) prediction
558 performance and heritability p value (rho_avg > 0.1, zscore_pval < 0.05 and GCTA h^2 LRT p value
559 < 0.05). Sensitivity analysis was performed by generating gene prediction models by including
560 additional number of ancestral alleles as covariates for local ancestry adjustments. For AA,
561 number of African ancestral allele count was used while European and Indigenous American
562 ancestral alleles were used as covariates.

563 Out-of-sample validation of the gene expression prediction models were done using the African
564 American asthma cohort, Study of Asthma Phenotypes and Pharmacogenomic Interactions by

565 Race-Ethnicity (SAPPHIRE)[23]. Predicted gene expression from SAPPHIRE genotypes was generated
566 using the predict function from MetaXcan. Genotypes of SAPPHIRE samples were generated by
567 whole genome sequencing through the TOPMed program and were processed the same way as
568 GALA II and SAGE. RNA-seq data from SAPPHIRE were generated as previously described[64] and
569 were normalized using TMM in edgeR. Predicted and normalized gene expression data were
570 compared to generate correlation R^2.

571 To assess the performance of the resulting GALA/SAGE models we conducted transcriptome-wide
572 association studies (TWAS) of 28 traits using GWAS summary statistics from the Population
573 Architecture using Genomics and Epidemiology (PAGE) Consortium study by Wojcik et al[24].
574 Analyses were performed using S-PrediXcan with whole blood gene prediction models from GALA
575 II and SAGE (GALA/SAGE models), GTEx v8, and monocyte gene expression models from the Multi-
576 Ethnic Study of Atherosclerosis (MESA) study[3].

577 *Data availability*

578 TOPMed WGS and RNA-seq data from GALA II and SAGE are available on dbGaP under accession
579 number phs000920.v4.p2 and phs000921.v4.p1, respectively. TOPMed WGS data from SAPPHIRE
580 are available under the dbGaP accession number phs001467.v1.p1. Models for performing
581 transcriptome wide association studies developed from GALA II and SAGE are freely available
582 upon request from the corresponding authors. Due to use restrictions and prohibition of public
583 posting of genomic summary results from GALA II and SAGE, please email the corresponding
584 authors to set up data transfer.

585

586 ACKNOWLEDGEMENTS

587 Generation of molecular data for the TOPMed (Trans-Omics in Precision Medicine) program was
588 supported by the National Heart, Lung and Blood Institute (NHLBI). RNA sequencing for "NHLBI
589 TOPMed: Gene-Environment, Admixture and Latino Asthmatics Study" (phs000920, GALA II) and
590 "NHLBI TOPMed: Study of African Americans, Asthma, Genes and Environments" (phs000921,
591 SAGE) was performed at Broad Institute Genomics Platform (HHSN268201600034I). Whole
592 genome sequencing (WGS) for the same studies were performed at the New York Genome Center
593 (NYGC, 3R01HL117004-02S3) and Northwest Genomics Center (NWGC, HHSN268201600032I).
594 Core support including centralized genomic read mapping and genotype calling, along with
595 variant quality metrics and filtering were provided by the TOPMed Informatics Research Center
596 (3R01HL-117626-02S1; contract HHSN268201800002I). Core support including phenotype
597 harmonization, data management, sample-identity QC, and general program coordination were
598 provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract
599 HHSN268201800001I).

600 Whole genome sequencing of part of GALA II was performed by the New York Genome Center
601 under The Centers for Common Disease Genomics of the Genome Sequencing Program (GSP)
602 Grant (UM1 HG008901). The GSP Coordinating Center (U24 HG008956) contributed to cross-
603 program scientific initiatives and provided logistical and general study coordination. GSP is funded
604 by the National Human Genome Research Institute, the National Heart, Lung, and Blood Institute,
605 and the National Eye Institute.

606 This work and EGB were supported in part by the Sandler Family Foundation, the American
607 Asthma Foundation, the RWJF Amos Medical Faculty Development Program, Harry Wm. and
608 Diana V. Hind Distinguished Professor in Pharmaceutical Sciences II, the National Heart, Lung, and
609 Blood Institute (R01HL117004, R01HL135156, X01HL134589, U01HL138626), the National
610 Institute of Health and Environmental Health Sciences (R01ES015794), the National Institute on
611 Minority Health and Health Disparities (R56MD013312, P60MD006902), the Tobacco-Related
612 Disease Research Program (24RT-0025, 27IR-0030), and the National Human Genome Research

613 Institute (U01HG009080). LK was supported by funding from National Cancer Institute
614 (K99CA246076). CRG was supported by funding from the National Human Genome Research
615 Institute (R01HG010297), CRG and NAZ were supported by funding from the National Heart, Lung,
616 and Blood Institute (R01HL151152) and the National Human Genome Research Institute
617 (R01HG011345). We gratefully acknowledge the studies and participants who provided biological
618 samples and data for TOPMed. We thank Soren Germer, Michael C. Zody, Lara Winterkorn and
619 Catherine Reeves from NYGC and Deborah A. Nickerson from NWGC for overseeing the
620 production of GALA II and SAGE whole genome sequencing data. We thank Huwenbo Shi, Kathryn
621 S. Burch and Bodgdan Pasaniuc for their technical support on the PESCA software.

622 The content is solely the responsibility of the authors and does not necessarily represent the
623 official views of the National Institutes of Health.

AUTHOR CONTRIBUTIONS

ACYM, LK, KLK, SSO, MAS, CRG, NZ, EGB, EZ contributed to the conception or design of the work. ACYM, LK, DH, CE, SH, JRE, NG, SG, SX, HG, AOO, JRS, MAL, LKW, LNB, CRG, NZ, EGB, EZ contributed to the acquisition, analysis, or interpretation of data. ACYM, LK, JRE, KLK, AOO, SSO, LNB, CRG, NZ, EGB, EZ have drafted the work or substantively revised it. All authors approved the submission of this manuscript.

REFERENCES

1. Majewski, J. & Pastinen, T. The study of eQTL variations by RNA-seq: from SNPs to phenotypes. *Trends Genet* **27**, 72–79 (2011).
2. The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
3. Mogil, L. S. *et al.* Genetic architecture of gene expression traits across diverse populations. *PLoS Genet* **14**, e1007586 (2018).
4. Wen, X., Luca, F. & Pique-Regi, R. Cross-population joint analysis of eQTLs: fine mapping and functional annotation. *PLoS Genet* **11**, e1005176 (2015).
5. Kim-Hellmuth, S. *et al.* Cell type–specific genetic regulation of gene expression across human tissues. *Science* **369**, eaaz8528 (2020).
6. Porcu, E. *et al.* Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. *Nat Commun* **10**, 3300 (2019).
7. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics* **47**, 1091–1098 (2015).
8. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet* **48**, 245–252 (2016).
9. Tam, V. *et al.* Benefits and limitations of genome-wide association studies. *Nat Rev Genet* **20**, 467–484 (2019).
10. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The Missing Diversity in Human Genetic Studies. *Cell* **177**, 26–31 (2019).

11. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature News* **538**, 161 (2016).
12. Keys, K. L. *et al.* On the cross-population generalizability of gene expression prediction models. *PLOS Genetics* **16**, e1008927 (2020).
13. Geoffroy, E., Gregga, I. & Wheeler, H. E. Population-Matched Transcriptome Prediction Increases TWAS Discovery and Replication Rate. *iScience* **23**, 101850 (2020).
14. Martin, A. R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *The American Journal of Human Genetics* **100**, 635–649 (2017).
15. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics* **51**, 584–591 (2019).
16. Hormozdiari, 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).
17. Shi, H. *et al.* Localizing Components of Shared Transethnic Genetic Architecture of Complex Traits from GWAS Summary Data. *The American Journal of Human Genetics* **106**, 805–817 (2020).
18. 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).
19. Gorlova, O. Y. *et al.* Gene-level association analysis of systemic sclerosis: A comparison of African-Americans and White populations. *PLOS ONE* **13**, e0189498 (2018).

20. Zeng, C. *et al.* Identification of Susceptibility Loci and Genes for Colorectal Cancer Risk. *Gastroenterology* **150**, 1633–1645 (2016).

21. Suzuki, K. *et al.* Identification of 28 new susceptibility loci for type 2 diabetes in the Japanese population. *Nat Genet* **51**, 379–386 (2019).

22. 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. *Nature Genetics* **52**, 680–691 (2020).

23. Levin, A. M. *et al.* Nocturnal asthma and the importance of race/ethnicity and genetic ancestry. *American journal of respiratory and critical care medicine* **190**, 266–273 (2014).

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

25. 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).

26. Rosenberg, N. A. *et al.* Genetic Structure of Human Populations. *Science* **298**, 2381–2385 (2002).

27. Henn, B. M., Cavalli-Sforza, L. L. & Feldman, M. W. The great human expansion. *Proc Natl Acad Sci U S A* **109**, 17758–17764 (2012).

28. Reich, D. *et al.* Reconstructing Native American population history. *Nature* **488**, 370–374 (2012).

29. Wall, J. D. *et al.* Genetic variation in Native Americans, inferred from Latino SNP and resequencing data. *Mol Biol Evol* **28**, 2231–2237 (2011).

30. 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).

31. 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).

32. Chatterjee, N., Shi, J. & García-Closas, M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat Rev Genet* **17**, 392–406 (2016).

33. Gay, N. R. *et al.* Impact of admixture and ancestry on eQTL analysis and GWAS colocalization in GTEx. *Genome Biology* **21**, 233 (2020).

34. 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).

35. 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).

36. 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).

37. 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).

38. 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).

39. 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).

40. Silverstein, R. L. & Febbraio, M. CD36, a scavenger receptor involved in immunity, metabolism, angiogenesis, and behavior. *Sci Signal* **2**, re3 (2009).

41. Oquendo, P., Hundt, E., Lawler, J. & Seed, B. CD36 directly mediates cytoadherence of Plasmodium falciparum parasitized erythrocytes. *Cell* **58**, 95–101 (1989).

42. Hsieh, F.-L. *et al.* The structural basis for CD36 binding by the malaria parasite. *Nat Commun* **7**, 12837 (2016).

43. 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).

44. 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).

45. Reiner, A. P. *et al.* Genome-Wide Association Study of White Blood Cell Count in 16,388 African Americans: the Continental Origins and Genetic Epidemiology Network (COGENT). *PLOS Genetics* **7**, e1002108 (2011).

46. Astle, W. J. *et al.* The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* **167**, 1415-1429.e19 (2016).

47. 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).

48. 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).

49. Zhong, Y., Perera, M. A. & Gamazon, E. R. On Using Local Ancestry to Characterize the Genetic Architecture of Human Traits: Genetic Regulation of Gene Expression in Multiethnic or Admixed Populations. *Am. J. Hum. Genet.* **104**, 1097–1115 (2019).

50. Basu, M., Wang, K., Ruppin, E. & Hannenhalli, S. Predicting tissue-specific gene expression from whole blood transcriptome. *Science Advances* **7**, eabd6991 (2021).

51. 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).

52. Oh, S. S. *et al.* Effect of secondhand smoke on asthma control among black and Latino children. *The Journal of allergy and clinical immunology* **129**, 1478–83.e7 (2012).

53. 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).

54. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).

55. 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).

56. 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).

57. Aguet, F. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).

58. 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).

59. 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).

60. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome research* **19**, 1655–1664 (2009).

61. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. *The American Journal of Human Genetics* **93**, 278–288 (2013).

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

63. Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T. & Delaneau, O. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**, 1479–1485 (2016).

64. Levin, A. M. *et al.* Integrative approach identifies corticosteroid response variant in diverse populations with asthma. *Journal of Allergy and Clinical Immunology* **143**, 1791–1802 (2019).

Figure 1. Study overview.

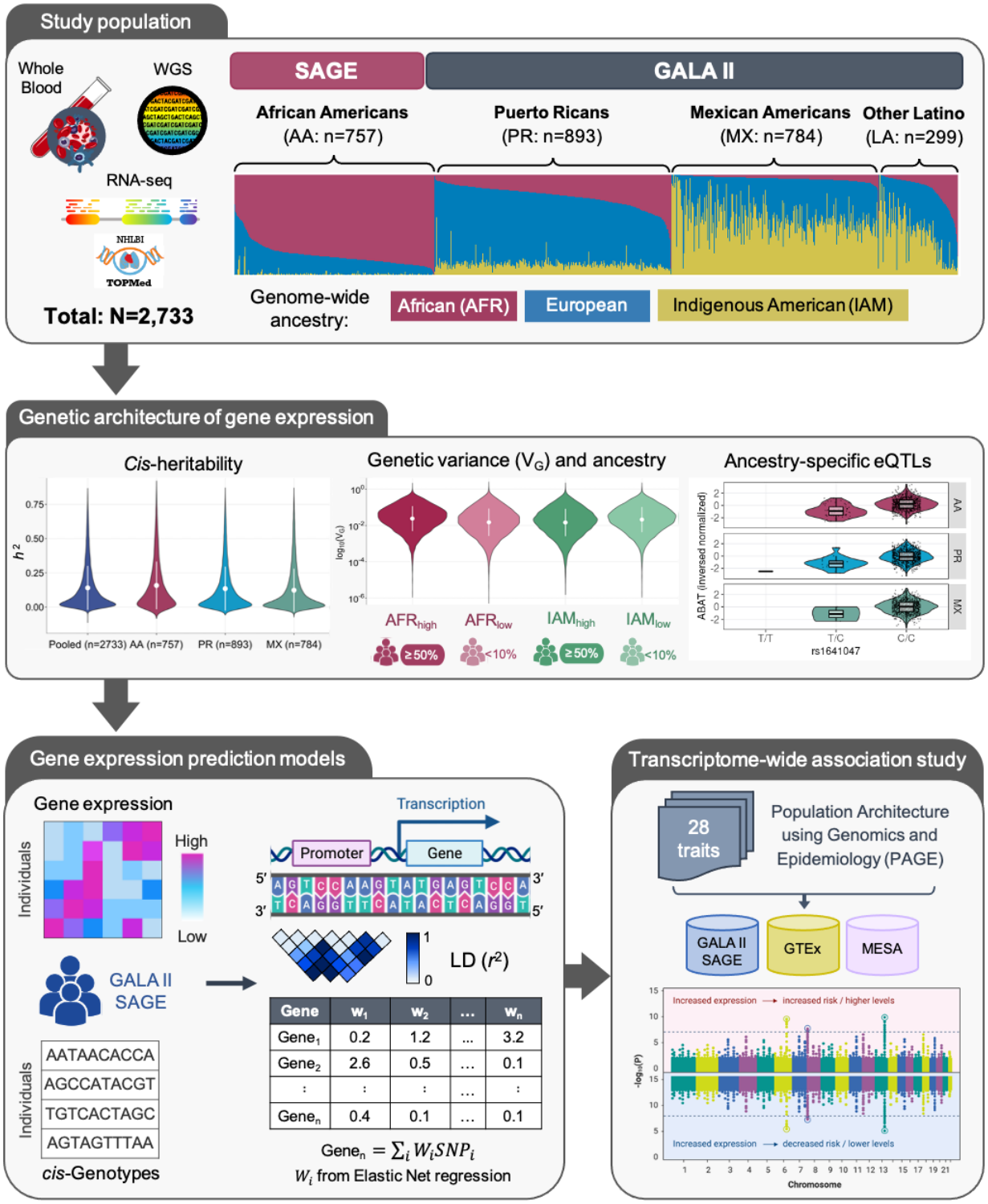

Figure 1. Study overview. This study included TOPMed whole genome sequencing and whole transcriptome data generated from whole blood samples of SAGE African American and GALA II Latino individuals (n=2,733). Global ancestry proportions for all individuals are shown in the top panel. **Middle row:** we studied the genetic architecture of gene expression including *cis*-heritability and genetic component of gene expression variation (see Figure 2 for details), and association of gene expression with global and/or local ancestry (see Figure S1 for details). We also identified eQTLs, eGenes and ancestry-specific eQTLs (see Figure 3 for details). **Bottom:** we generated gene prediction models (left panel) and applied them in TWAS based on GWAS summary statistics generated from the PAGE study. We also compared TWAS results generated from GTEx v8 whole blood and MESA monocyte-derived models.

Figure 2. *Cis*-heritability of gene expression and variance explained by genetic component.

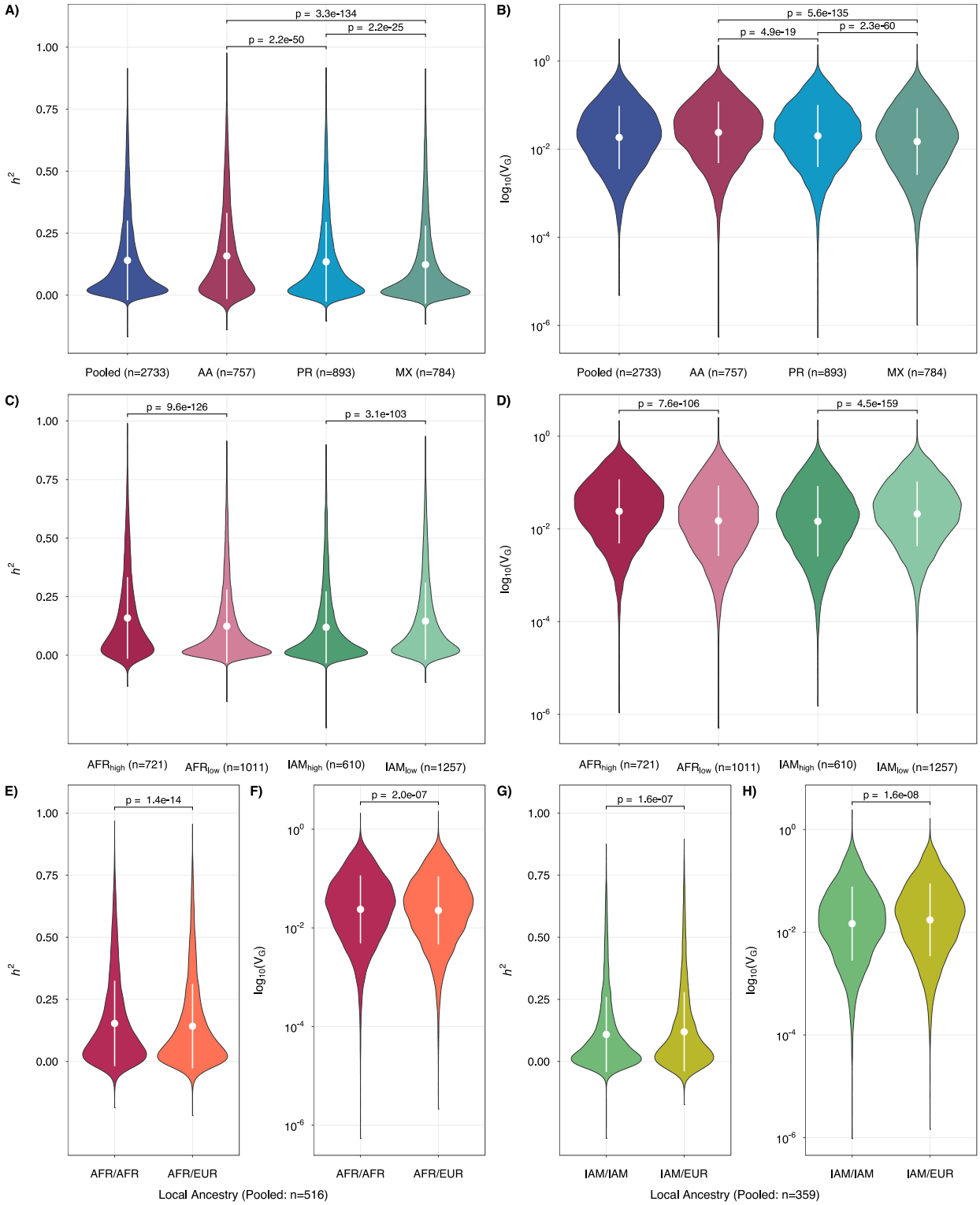

Figure 2. *Cis*-heritability (h^2) of gene expression and genetic component of gene expression variance (V_G). **a, b|** Estimated h^2 and V_G of whole blood gene expression in self-identified race/ethnicity groups . **c, d|** Estimated h^2 and V_G in groups defined by genetic ancestry, comparing subjects with > 50% global ancestry (high) to < 10% of the same ancestry (low). **e, f|** Estimated h^2 and V_G in groups defined by local ancestry at gene's transcriptional start site (TSS), comparing subjects with 100% (AFR/AFR) to 50% (AFR/EUR) local African ancestry. **g, h|** Similar to e and f but for 100% (IAM/IAM) to 50% (IAM/EUR) local Indigenous ancestry at the TSS. Two-sided p-values are based on the Wilcoxon rank sum test.

Figure 3. Identification and examples of ancestry-specific eQTLs.

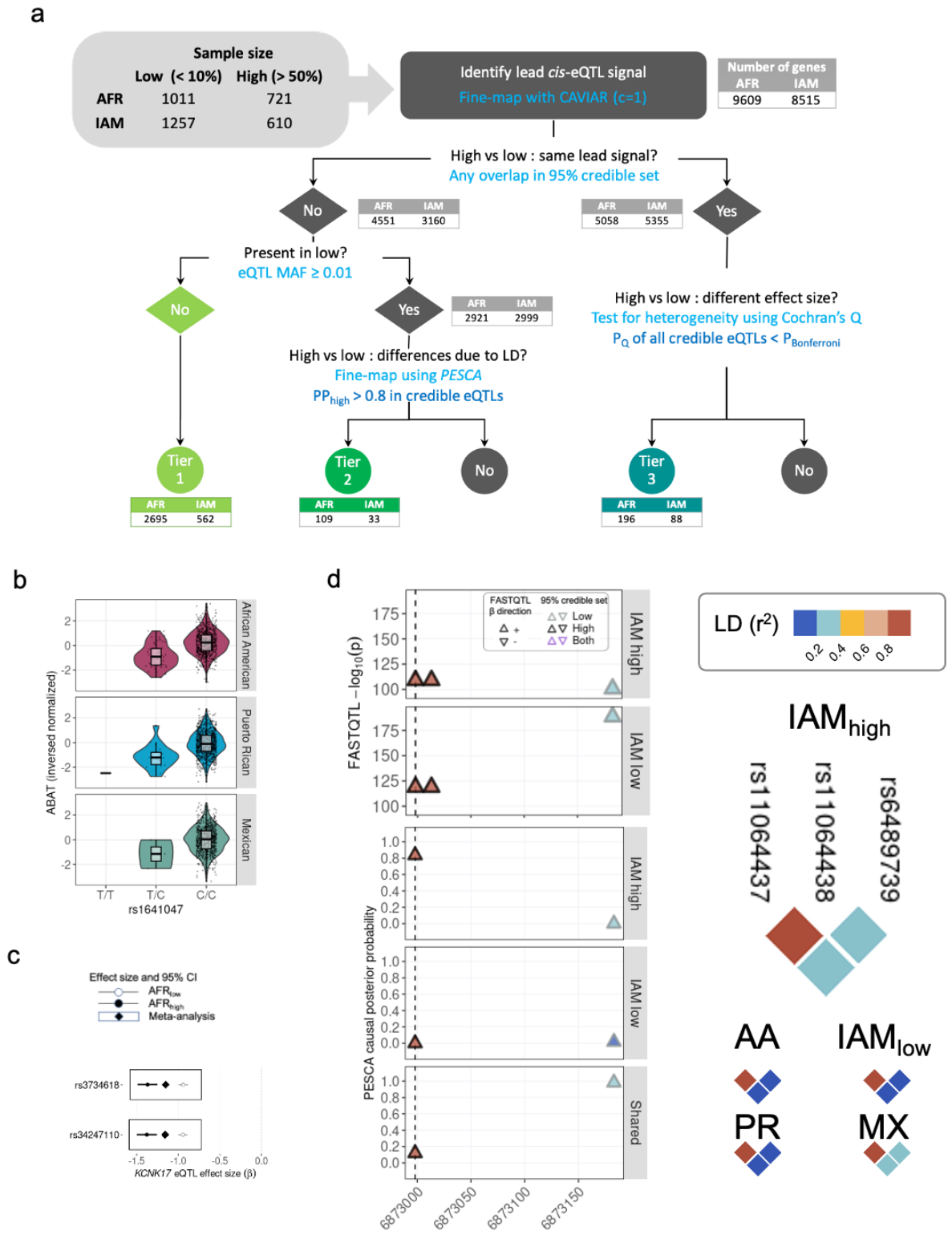

Figure 3. Identification of ancestry-specific eQTLs (anc-eQTL). **a**| Decision tree for the identification of anc-eQTLs. The lead eQTL signal of each heritable gene, represented by the 95% credible set, was identified by CAVIAR assuming 1 causal variant in each global ancestry high and low group separately (AFR_{high}, AFR_{low}, IAM_{high} and IAM_{low}). Sample size and global ancestry cutoff criteria used to classify individuals into each group are shown in the grey box at the top left corner. In genes with non-overlapping eQTL lead signal (left branch), Tier 1 anc-eQTLs were identified based on differences in allele frequency between groups (MAF ≥ 0.01 in high and < 0.01 in low). Tier 2 anc-eQTLs were identified with the consideration of different effect size and LD pattern. To be considered Tier 2 ancestry-specific, all credible eQTLs in the high group were required to show the highest posterior probability to be causal specific to the high group (PP_{high}) and at least one of these eQTLs have to attain $PP_{high} > 0.8$. In genes with overlapping credible sets (right branch), Tier 3 anc-eQTLs were identified based on different effect size assessed by Cochran's Q. All credible eQTLs in the high group were required to have a Cochran's Q p-value smaller than the bonferroni threshold. Number of genes remaining for each ancestry (AFR and IAM) after each filtering step is indicated alongside each branch and in Table S6. The procedures are described in detail in the method section. **b**| An example of a Tier 1 AFR_{high}-specific eQTL (rs1641047) of the *ABAT* gene that was reported in GWAS catalog to be associated with systemic sclerosis. The minor allele, T allele, was associated with systemic sclerosis and was common (MAF ≥ 0.01) in African Americans but rare in Puerto Ricans and Mexican Americans. **c**| An example of Tier 3 AFR_{high}-specific eQTLs (rs34247110 and rs3734618) of the *KCNK17* gene. Both eQTLs were shared between the 95% credible set (CAVIAR) of AFR_{high} and AFR_{low} groups. The eQTL, rs34247110, was reported in GWAS catalog by two different

studies to be associated with type 2 diabetes in European, African American, Hispanic and Asian populations. The detection of rs34247110 from multiple populations was consistent with the nature of overlapping lead eQTL signals for Tier 3 anc-eQTLs. Black and open white points and their whiskers indicate eQTL effect size and their 95% confidence interval (CI) in the AFR_{high} and AFR_{low} group, respectively. The two eQTLs show non-overlapping 95% CI of effect sizes between the AFR_{high} and AFR_{low} groups, supporting their identity as Tier 3 anc-eQTLs, which have significantly different effect size between the two groups (Cochran's Q p < bonferroni threshold). Diamond and rectangle indicate effect size and their 95% confidence interval from meta-analysis using METASOFT. **d**| An example of a Tier 2 IAM_{high}-specific, colorectal cancer-associated eQTL (rs11064437) of the *SPSB2* gene. The top 2 panels on the left show non-overlapping 95% credible set (CAVIAR) between the IAM_{high} (black outline triangle) and IAM_{low} (grey outline triangle). The fill color of the triangle indicates LD of an eQTL with the eQTL marked by the vertical dotted line. The CAVIAR 95% credible eQTLs in IAM_{high} and IAM_{low} were in different LD blocks, as indicated by low LD represented by the fill of the triangle and LD plots on the right. LD plots showed eQTLs in the same order as the top two panels on the left. The bottom three panels on the left show the left eQTL had the highest posterior probability (PESCA) of being IAM-specific while the right eQTL had highest posterior probability of being shared between two groups. In the PESCA analysis, eQTLs were pruned by R^2 of 0.95 to remove eQTLs that have almost perfect LD with each other.

Figure 4. Comparison of transcriptome imputation model performance and transcriptome-wide association study (TWAS) results.

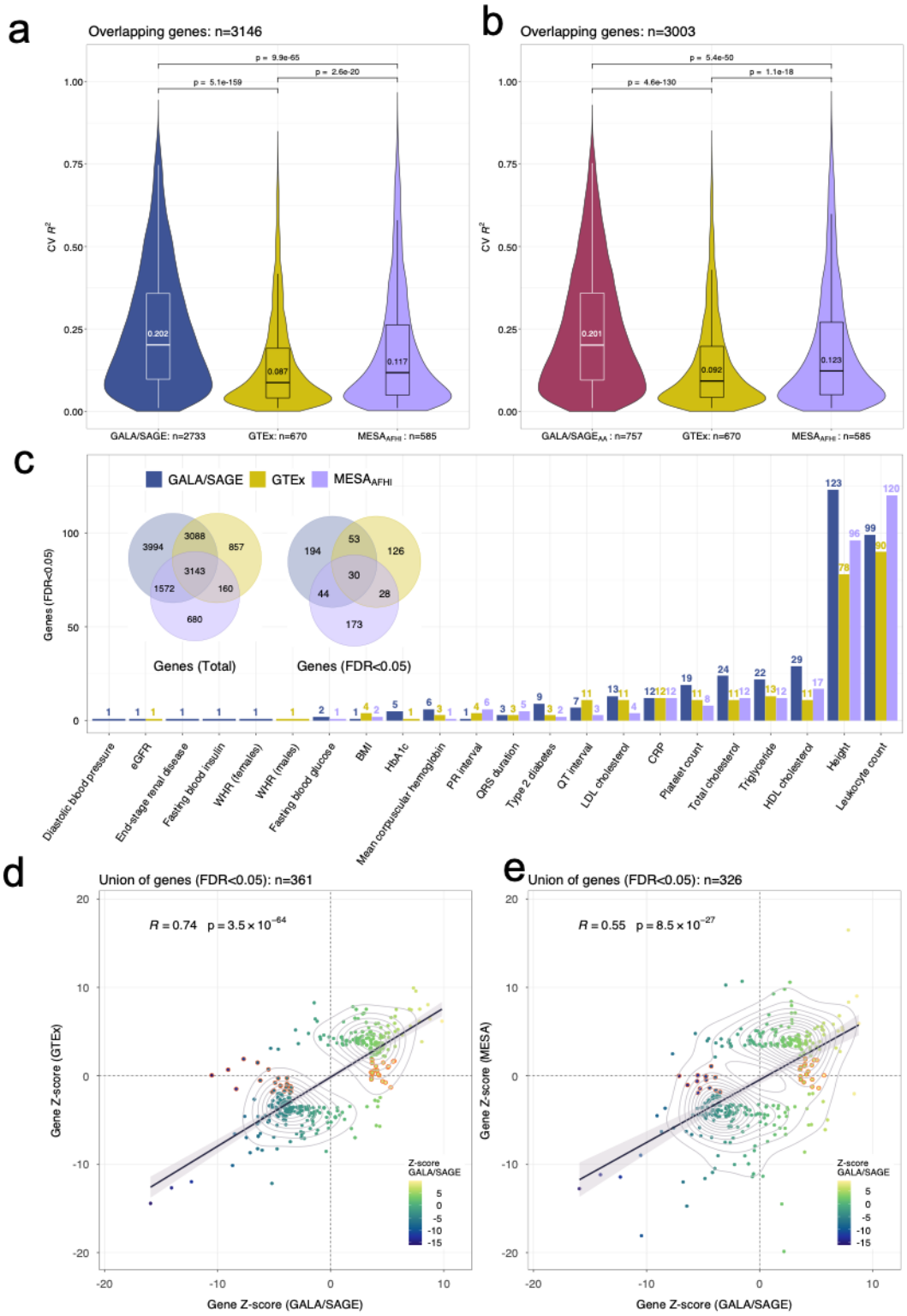

Figure 4. Comparison of transcriptome imputation model performance and transcriptome-wide association study (TWAS) results. **a,b|** Internal cross-validation R2 values for each TWAS model restricting to the subset of overlapping genes. GALA/SAGE models were trained by pooled (a) or AA (b) samples. **c|** Summary of TWAS results across 28 traits from the PAGE GWAS meta-analysis. Venn diagrams depict the number of genes analyzed in each TWAS and the number significant (FDR<0.05) findings across all traits. **d,e|** Correlation in z-scores for the union of genes that were statistically significant in at least one TWAS model. Correlation between GALA/SAGE pooled models with GTEx (d) or MESA (e).

Figure 5. Transcriptome-wide association study (TWAS) results.

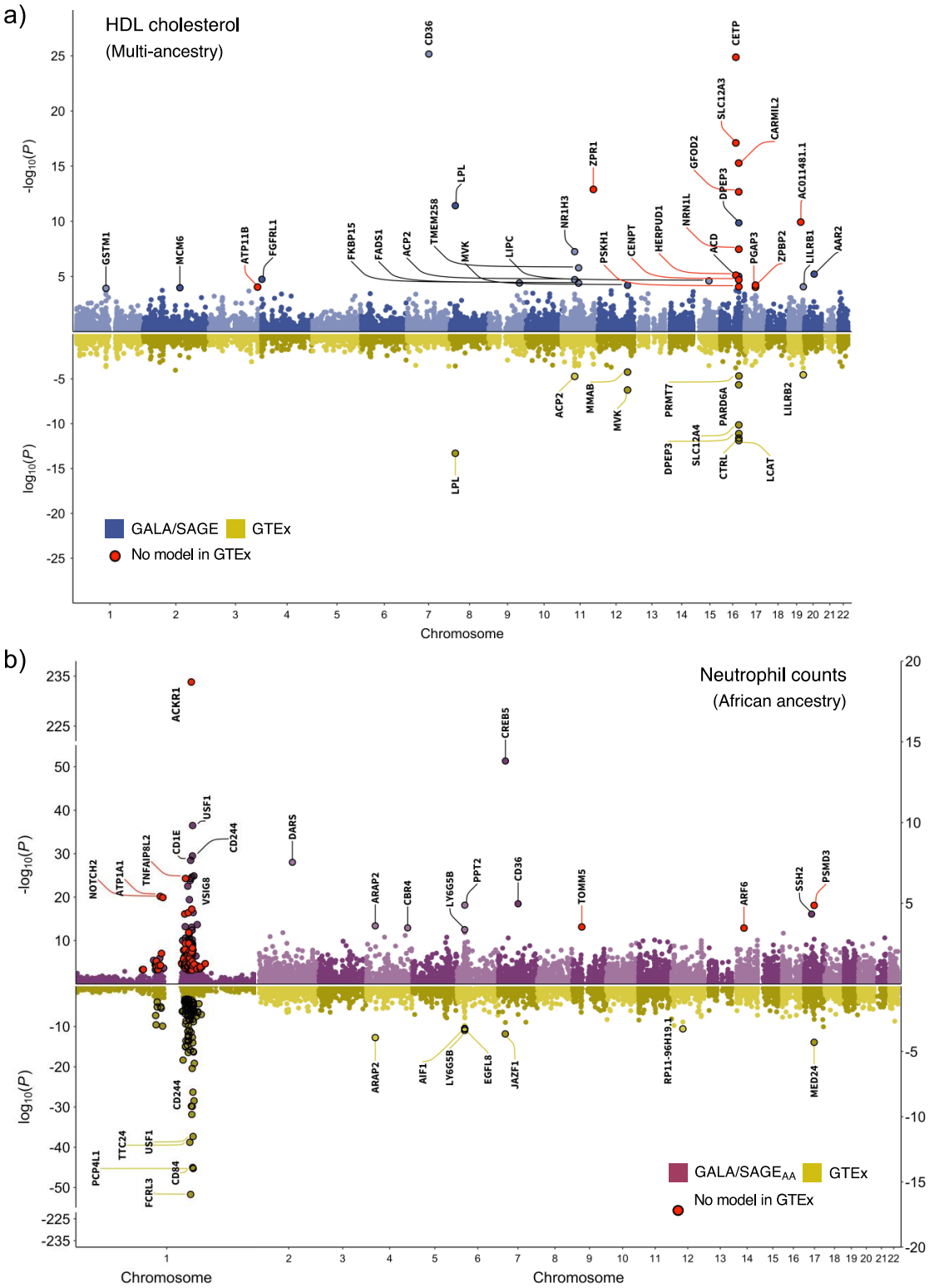

Figure 5. Transcriptome-wide association study (TWAS) results for **a|** HDL cholesterol using summary statistics form the PAGE multi-ancestry GWAS meta-analysis (N=33,063) and **b|** neutrophil counts from a GWAS meta-analysis of African ancestry individuals (N=13,476). Genes that achieved TWAS significance (FDR<0.05) are indicated by circles with a black border and are labeled, except for chromosome 1 for neutrophil counts due to the large number of associations. Significantly associated genes for which expression levels could not be predicted using GTEx v8 elastic net models are indicated in red.

Table 1. Demographic characteristics of GALA II and SAGE participants included in this study.

	African American (N=757)	**Puerto Rican (N=893)**	**Mexican (N=784)**	**Other Latino (N=299)**	**Pooled (N=2,733)**
Age (years)	16.0 [6.6]	13.2 [4.8]	13.8 [6.5]	13.7 [5.7]	14.0 [6.3]
Sex					
Female	405 (53.5%)	451 (50.5%)	427 (54.5%)	158 (52.8%)	1,441 (52.7%)
Asthma status					
Case	433 (57.2%)	549 (61.5%)	351 (44.8%)	156 (52.2%)	1,489 (54.5%)
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]
Recruitment center					
SF 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%)

Summary statistics of quantitative variables are reported as median [interquartile range] while categorial variables were reported as count (percentage).