The genetic architecture of gene expression in individuals of African and European ancestry

Kipper Fletez-Brant[1,6,&], Renan Sauteraud[1], Yanyu Liang[2,^], Steven Micheletti[2,^], Priyanka Nandakumar[2,^], Aarathi Sugathan[1, 7,^], Kijoung Song[3,8], Taylor B. Cavazos[1,9], Amal Thomas[1,10], Robert J. Tunney[1], Barry Hicks[2], Jared O'Connell[2,11], Suyash Shringarpure[2], Katelyn Kukar[2], Meghan Moreno[2], Emily DelloRusso[2], Corinna D. Wong[2], Aaron Petrakovitz[2,12], Goutham Atla[4,13], Adrian Cortes[4], Padhraig Gormley[5], Laurence Howe[4], Rajashree Mishra[3], Daniel Seaton[4], the 23andMe Research Team[&], Robert C. Gentleman[1,14], Steven J. Pitts[1,2], Vladimir Vacic[1,&]

[1] Therapeutics Division, 23andMe, Inc., South San Francisco, CA, USA. [2] 23andMe, Inc., Sunnyvale, CA, USA. [3] Genomic Sciences, GSK, Upper Providence, PA, USA. [4] Genomic Sciences, GSK, Stevenage, UK. [5] Genomic Sciences, GSK, Cambridge, MA, USA. [6] Currently at Genentech, Inc., South San Francisco, CA, USA. [7] Currently at CSL, Melbourne, VIC, Australia. [8] Currently at Alumis, South San Francisco, CA, USA. [9] Currently at Exai Bio, Inc., Palo Alto, CA, USA. [10] Currently at Gilead Sciences, Inc., Foster City, CA, USA. [11] Currently at AstraZeneca PLC, Barcelona, Catalonia, Spain. [12] Currently at Genomenon, Inc., Ann Arbor, MI, USA. [13] Currently at Novo Nordisk Research Centre Oxford Ltd, Oxford, UK. [14] Currently at the Center for Computational Biomedicine, Harvard Medical School, Boston, MA, USA. [^] These authors contributed equally. [&] To whom correspondence should be addressed at fletezbk@gene.com, research-inquiries@23andme.com and vvacic@gmail.com.

Abstract

We conducted two large scale studies of the genetics of gene expression in individuals of African ancestry within a cohort of consented 23andMe research participants and in LCL samples from the 1000 Genomes Project African superpopulation. We discovered nearly four times as many eQTLs, compared to tissue-matched eQTL studies in European cohorts. Additionally, we found that the majority of eQTLs were not detectable across populations; those that were, however, were found to be highly concordant. Performing eQTL studies in African ancestry cohorts resulted in more signals per gene and smaller credible sets of causal variants. We showed that comparisons of heritability of gene expression could be confounded by population substructure, but that variation in local genetic ancestry did not majorly impact eQTL discovery. Finally, we showed improvements in variant-to-gene mapping of African-American GWAS signals when using African compared to European ancestry eQTL studies.

Introduction

Underrepresentation of individuals of African ancestry in biomedical research is a serious ethical issue leading to healthcare disparities worldwide, and a major missed scientific opportunity to understand the genetic basis of disease[1]. Initiatives like H3Africa[2], Southern African Human Genome Programme[3] and MalariaGen[4] have conducted genomic studies with research participants from African countries, while the All of US[5] and Million Veteran Program (MVP)[6] have performed large GWAS in African-American cohorts in the United States. While associations can be identified more easily in African compared to non-African populations as a result of increased genetic diversity[7], interpreting GWAS results and mapping disease risk genes requires a functional link between genetics and an intermediate molecular phenotype, such as expression quantitative trait loci (eQTL). Although eQTL consortia such as GTEx[8] included African-American samples and performed follow-ups that focused on analysis of these samples[9], eQTL studies specifically conducted in African[10,11] or African-American cohorts[12,13], as well as QTL studies broadly[14], are fewer. Comparative studies of regulatory variation in diverse human populations in the HapMap3 cohort demonstrated significant sharing of *cis*-regulatory variation across populations, and for shared eQTLs, near-perfect concordance of directionalities and effect sizes[15]. However, differences in allele frequencies across populations cause differential eQTL discovery power[15] and limit transferability of gene expression prediction models[16]. Genetic European Variation in Disease (GEUVADIS)[10], the first large RNA-seq based cross-population study of regulatory variation, highlighted differences in transcript usage and differences in overall gene expression, but concluded that these two were largely mediated by separate genetic variants. eQTL analyses in the Human Genome Diversity Panel[17] cohort reported that 25% of variation in expression was attributable to ancestry, and that 76% of this variation is due to expression rather than splicing. A recent study of eQTLs in African-American and Latino populations linked heritability of gene expression and population heterozygosity, and showed prevalence of population specific eQTLs to be 30% and 8% within ancestral African and indigenous American genomic segments respectively[13]. African Functional Genomics Resource (AFGR) compiled gene expression measured in 1000 Genomes Project African samples and additional Maasai individuals, as well as open chromatin in a subset of 100 individuals, and assembled a comprehensive dataset of expression, splicing and chromatin accessibility QTLs[18].

To address underrepresentation of individuals of African ancestry (AFR) in research, we conducted eQTL studies in two different cohorts. We recruited a cohort of consented 23andMe research participants, which we refer to as the Black Representation in Genomic Research (BRGR) study. We collected saliva and venous blood samples for whole-genome sequencing (WGS) and RNA-seq, respectively, from 737 individuals. Additionally, we sequenced RNA extracted from 659 lymphoblastoid cell lines (LCLs) belonging to the African ancestry

superpopulation in the 1000 Genomes Project[19], comprised of individuals from continental Africa (including GEUVADIS Yoruba individuals[10] and most AFGR samples[18]) and admixed individuals from the African diaspora. Using publicly available WGS of these individuals[20], we analyzed the genetics of gene expression in LCLs (Fig. 1a). As comparisons, we analyzed eQTLs in venous blood and LCLs from two large European cohorts[10,21,22] (Table 1) and subsequently investigated sharing of eQTLs. We used AFR and EUR eQTLs to annotate African-American GWAS signals in 6 studies from MVP[23–25], 11 studies from the Blood Cell Consortium (BCX)[14] and in the 23andMe GWAS of height in African-Americans.

Results

eQTL cohorts capture African and European ancestral backgrounds

We applied 23andMe's Ancestry Composition (AC)[26] algorithm to genetic variants called from WGS data in the four cohorts and confirmed African and European ancestry components (Fig. 1b). Plurality of BRGR participants are from the South Census Region (Suppl. Fig. 1, Suppl. Table 1). Genome-wide AC proportions of BRGR participants are similar to previously published ancestry of individuals of African descent in the U.S.[27], with high representation of Sub-Saharan African ancestry ($\mu^{African}$=0.80±0.11, mean±sd) and moderate representation of European ancestry ($\mu^{European}$=0.17±0.10, Suppl. Table 2). Plurality of local African ancestry is Nigerian ($\mu^{Nigerian}$=0.29±0.09), with the majority of individuals grouping closely to the Igbo reference group according to a graph method based on shared identical-by-descent (IBD) segments[28] (Suppl. Fig. 2a). Additionally, the distribution of ancestry found in the BRGR cohort is representative of all genotyped African-American customers in the 23andMe database (based on a randomization test, Suppl. Table 2). The Parkinson Progression Marker Initiative (PPMI)[22,29] dataset was predominantly European ($\mu^{European}$=0.97±0.09), with the plurality of local European ancestry being British and Irish ($\mu^{British\text{-}Irish}$=0.35±0.37), followed by Iberian ($\mu^{Iberian}$=0.13±0.31). For the 1000 Genomes Project eQTL cohorts, our genome-wide AC proportions recapitulate established population genetics results[19] (Fig. 1b, Suppl. Table 2,3).

Increased genetic diversity improves eQTL discovery power

RNA-seq from each of the four datasets underwent the same QC and gene expression quantification procedures, and yielded comparable numbers of protein-coding genes for eQTL testing (slightly more in PPMI, see Table 1). Approximately 15.6M variants pass QC filters in each African cohort compared to 8.6M variants in European cohorts, in accordance with the known larger number of variants per genome in individuals of African ancestry[19]. Comparing

MAF spectra across ancestry groups reveals enrichment of rare variants in African versus tissue-matched European cohorts (Fig. 2a, “all variants”).

To discover eQTLs, we applied SuSiE[30,31] to individual level data from each study, using default parameters. For each discovered eGene SuSiE returns independent eQTL signals as a collection of credible sets (CSs). Within each CS, we reported the variant with the largest posterior inclusion probability (PIP) as the index eVariant. In BRGR (sample size n=737) we detected 21,130 eQTLs regulating 10,044 eGenes via 20,295 eVariants, compared to 5,661 eQTLs (4,277 eGenes; 5,380 eVariants) in PPMI (n=752; Table 1; Fig. 2b). In LCL_AFR (n=659) we found 16,265 eQTLs (13,584 eGenes; 15,832 eVariants) compared to 2,904 eQTLs (2,527 eGenes; 2,843 eVariants) in GEUVADIS (n=358), however we here note the differential sample size in LCLs. MAF spectra (Fig. 2a, “all eQTLs”) showed that eQTLs called in African cohorts are enriched for rare variants, compared to a lack in European eQTLs. MAF relative increase in AFR compared to EUR is bimodal in both cell types, with a majority of eQTLs (11,662 in BRGR and 10,032 in LCL_AFR) with at least 3.16x ($\log_{10}$ ratio = 0.5) greater MAF (Fig. 2f, “all eQTLs”).

We discovered fewer eVariants per eGene in Europeans, with 77% of PPMI and 88% of GEUVADIS eGenes having only one eVariant. Both African cohorts have comparable trends in numbers of eVariants per eGene (Fig. 2c) and CS sizes are two times smaller in African compared to European cohorts (Fig. 2d).

Majority of eQTLs are population- and tissue-specific, but the effects of shared eQTLs are concordant

We first assessed sharing of eQTLs conservatively defined as exactly matching eGene and eVariant. We found 872 eQTLs shared between BRGR and PPMI, and 341 between LCL_AFR and GEUVADIS. As fraction of constituent datasets, eQTL sharing was infrequent (15.4% of PPMI, 4.1% of BRGR eQTLs; 2.1% of LCL_AFR, 11.7% of GEUVADIS eQTLs; Fig. 2e). However, shared eQTLs exhibited consistent effect sizes (Fig. 2g), with Pearson correlation r^2 = 0.89 (p-value<2.2×10^{-16}) and 0.91 (p-value<2.2×10^{-16}) for venous blood and LCL datasets, respectively. Shared eQTLs had smaller CSs (Suppl. Fig. 3), and 75% of CSs in African cohorts contained 1 or 2 variants, while in PPMI 75% of CSs had up to 7 variants, and over 10 in GEUVADIS. Shared eQTLs had similar MAF spectra, except for rare alleles, which were underrepresented in Europeans (Fig. 2f). Shared venous blood eQTLs either had comparable or greater MAF in AFR, while LCL eQTL MAFs were comparable. Within-population cross-tissue sharing was rare: 1,987 eQTLs were shared within African datasets (9% BRGR; 12% LCL_AFR), while 176 eQTLs were shared within European datasets (3% PPMI; 6% GEUVADIS).

Next, we considered eQTLs shared if eGenes matched and pairs of signals colocalized[32] with posterior probability $H_4 \geq 0.5$. We found that 55.5% of eQTLs identified in PPMI were shared with BRGR, and 14.9% conversely (Table 2). 54% of eQTLs identified in GEUVADIS were shared with LCL_AFR and 9.6% conversely. eQTL effect sizes were consistent, with r^2=0.83 (p-value<2.2×10^{-16}) and 0.79 (p-value<2.2×10^{-16}) for venous blood and LCLs, respectively. Across tissues, African cohorts shared 4,545 eQTLs (22% of BRGR and 27% of LCL_AFR eQTLs, respectively) while European cohorts shared 769 eQTLs (14% of PPMI and 26% of GEUVADIS eQTLs, respectively). CSs of colocalized eQTLs were smaller in African than in European cohorts (Suppl. Fig. 3) and shared eQTLs have comparable MAF. Venous blood specifically shows a striking enrichment in BRGR eQTLs of eVariants more common in BRGR versus PPMI (Fig. 2f, x-axis=+0.5), and also that variants more common in PPMI (compared to BRGR) are detected in BRGR (Fig. 2f, x-axis=-0.5). The converse is not true: variants more prevalent in BRGR are not detected in PPMI, although PPMI is also enriched for variants more common in PPMI.

When comparing MAF spectra of shared (either exact match or coloc) and unshared eQTLs for a given tissue, all shared eQTLs have either comparable or larger MAF in EUR versus AFR (Fig. 2f). eQTLs with meaningfully larger allele frequencies in individuals of African ancestry tend to be unshared, highlighting the value of conducting eQTL studies in diverse cohorts.

We next measured replication, which within the SuSiE framework we defined as matching eGenes and the variant with the largest PIP in one CS (aka eVariant) observed in the other CS (Table 2). 23% of eQTLs found in PPMI replicated in BRGR and 21% vice versa, while for LCL cohorts, 18% of GEUVADIS eQTLs replicated in LCL_AFR and 16% vice versa. Comparing tissues, 16% of BRGR eQTLs were found in LCL_AFR and 21% vice versa, and 15% of PPMI eQTLs were found in GEUVADIS and 18% vice versa.

Finally, we compared detected eGenes, irrespective of eVariants or signal colocalization. 3,822 eGenes were shared between BRGR and PPMI (respectively, 27% and 22% of genes). 2,153 eGenes were shared between LCL_AFR and GEUVADIS (5.8% and 14.7% of genes). 6,843 of eGenes were shared between BRGR and LCL_AFR (48% and 50% of eGenes, respectively) while 1,331 eGenes were shared PPMI and GEUVADIS (7% and 9% of eGenes, respectively). In summary, our results indicated that only a small fraction of discovered eQTLs were shared across populations.

Heritability analyses of gene expression are confounded in cross-ancestry comparisons

We measured h^2 of gene expression in BRGR and PPMI using GCTA-GREML[33] and identified 9,777 eGenes that showed significant heritability in both cohorts. Average heritability in BRGR is significantly higher (h^2=0.31) than in PPMI (h^2=0.20). Out of the 9,777, we identified a subset of 1,016 eGenes with eQTL signals that colocalize across cohorts and measured the relationship of differential genetic variance of index SNPs to differential h^2 (Δh^2) across genes. Despite the assumption of shared causal variants for signals that colocalize across cohorts, we find that Δh^2, the difference in heritability, is positively associated with differential genetic variance (*p-value*=2.9×10^{-43}), accounting for 17% of variation in Δh^2 (Fig. 3b, Suppl. Fig. 4a). We also find that the ratio of unexplained to heritable variation, defined as $q = h^2 / (1-h^2)$ is associated with differential genetic variance of eVariants (Fig. 3c; Suppl. Fig. 4b,c; Suppl. Table 4), indicating that comparisons of heritability of shared casual signals are confounded by differential MAF.

Standard best practices in eQTL calling adequately control for ancestry as confounder

To assess unmodeled effects of ancestry on our findings in eQTL sharing and heritability, we compared the contribution of global ancestry using the standard approach of genetic PC covariates to modeling local ancestry using Tractor[34], and found that 80% of eGenes were identified in both models (Suppl. Fig. 5a). Another 15% of eGenes were significant in both models but with different eVariants. As Hou *et al.*[35], we found that Tractor was well-powered to identify effect size heterogeneity by ancestry and underpowered otherwise (Suppl. Fig. 6b), and for the majority of eQTLs we did not observe effect heterogeneity. We concluded that individual eQTL datasets were not impacted by ancestry-induced biases.

Genetic determinants of Duffy-null associated neutrophil count connect gene expression to cellular and physiological phenotypes

Reasoning that blood cell phenotypes with greater prevalence in Africans compared to Europeans should have detectable correlates in BRGR eQTLs, we investigated eQTL signals for Duffy-null associated neutrophil count (DnANC), a phenotype characterized by lower neutrophil and leukocyte counts without increased infection risk[36,37]. DnANC is observed in individuals of African, Middle Eastern, and West Indian descent, with estimated prevalence as high as 25-50% in AFR[38]. DnANC is driven by the Duffy-null genotype CC of rs2814778 (ClinVar:18395) in *ACKR1*[37,39]. Individuals with Duffy-null allele have increased protection against *Plasmodium vivax* malaria infection and increased susceptibility to HIV-1 trans-infection[40]. *ACKR1* venous blood expression correlates with neutrophil counts because of its effects on neutrophil

homeostasis; however *ACKR1* is a chemokine scavenger receptor expressed in erythrocytes and endothelial cells, not neutrophils.

We found multiple eQTLs for *ACKR1* in the BRGR and PPMI cohorts (Fig. 4): rs2814778 is an eQTL in the BRGR study (*β*±se=*-0.96±0.02,* p-value<2.2×10^{-308}), is more common in Africans (MAF_{AFR}=0.16, MAF_{EUR}=0.002, based on gnomAD[41] v4) and is located in the promoter of *ACKR1*. In BCX, rs2814778 has the strongest association of any blood-cell trait in African-ancestry individuals, and is associated with counts of neutrophils (β=0.86±0.02, p-value=3.99×10^{-432}), white blood cells (β=0.70±0.02, p-value=1.02×10^{-330}) and monocytes (β=0.32±0.02, p-value=1.92×10^{-63}). We also identified rs863005 as a secondary *ACKR1* eQTL in the BRGR cohort. We detected rs12075 as an eQTL for *ACKR1* in PPMI. This variant is more common in Europeans (MAF_{AFR}=0.07, MAF_{EUR}=0.42) and is a QTL for monocyte (β=0.027±0.002, p-value=3.24×10^{-41}) and basophil counts (β=0.028±0.002, p-value=6.63×10^{-39}) in Europeans in BCX. rs12075 is a coding variant for *ACKR1*[42] and has been reported as a regulator of the monocyte chemokine *MCP-1*[43]. This European eQTL colocalizes (H_4=0.6) with the rs863005 eQTL from BRGR. We didn't detect any eQTLs for *ACKR1* in the LCL datasets, as this gene is not expressed in B cells[44,45].

African ancestry eQTLs improve annotation of African-American GWAS

To assess the utility of African ancestry eQTL datasets for annotating African-American GWAS, we mapped associations in 22 GWAS performed in African-American cohorts: 8 cardio-metabolic traits from the Million Veteran Program (MVP), 13 blood measurement traits from the Blood Cell Consortium (BCX) and 23andMe's GWAS of height (Suppl. Table 5). Across the board, using AFR eQTLs lead to higher fractions of African-American GWAS signals annotated with at least one gene mapping hypothesis. In MVP, BCX and height studies BRGR eQTLs mapped 1.3-3x more signals to genes compared to PPMI. In BCX and height studies, LCL_AFR eQTLs mapped 1.75-2.77x more signals than GEUVADIS. GEUVADIS eQTLs didn't contribute to gene mapping in MVP while LCL_AFR eQTLs mapped 1.3% of association signals to genes (Table 3). BRGR eQTLs map at least one GWAS hit per phenotype in 71.4% of BCX phenotypes and 80% of MVP phenotypes, compared to 42.9% and 40% for PPMI respectively. Similarly, LCL_AFR eQTLs annotate at least one GWAS hit per phenotype in 50% of BCX and 60% of MVP phenotypes compared 35.7% and 0% for GEUVADIS. All four eQTL datasets annotated at least one GWAS hit in 23andMe height.

eQTLs discovered in BRGR and not shared with PPMI contributed 50 gene-phenotype hypotheses relating 38 GWAS signals in 15 phenotypes to 40 genes, while only 9 geno-phenotype pairs (9 GWAS signals in 3 phenotypes and 9 genes) were discovered with eQTLs unique to PPMI. eQTLs discovered in LCL_AFR and not shared with GEUVADIS

contributed 34 gene-phenotype hypotheses (31 GWAS signals in 8 phenotypes, 32 genes), while only 4 were discovered with eQTLs unique to GEUVADIS (4 GWAS signals in 2 phenotypes, 4 genes).

We next explored the novelty of our gene-phenotype hypotheses in OpenTargets Genetics v22.10, regardless of the specific eQTL or discovery tissue or cohort. Of the gene-phenotype hypotheses contributed by eQTLs only discovered in BRGR, 20 out of 50 hypotheses have not been previously reported. Notably this includes thalassemia variant rs33930165 (Clinvar ID: 15126) and the hemoglobin gene *HBD* and BCX phenotypes MCHC, RBC, RDW (all H_4=1.0; see Suppl. Table 6). Of those that have a close matching phenotype in OpenTargets — e.g. *TRIP10* and BCX phenotype mean platelet volume (MVP)[46]; max $H_{4,\text{OpenTargets}}$=0.37 and $H_{4,\text{ BRGR}}$=0.99 — the average relative increase of H_4 under African eQTLs was 6.1x greater than H_4 reported previously. For LCL_AFR, 17 out of 34 hypotheses were not reported in OpenTargets, including *ITM2* and height[47] (H_4=0.97; see Suppl. Table 6). Of those reported previously, the average relative increase in $P(H_4)$ was 6.8x greater using LCL_AFR eQTLs in comparison to reported H_4; with *EDC3* and height as example, we find H_4 2.3x greater in LCL_AFR compared to OpenTargets' reporting of height[48].

Out of gene-phenotype pairs derived from eQTLs unique to African cohorts, only 2 were seen in both BRGR and LCL_AFR: *DRICH1* and eosinophil count in BCX (rs5759953 is eVariant in both cohorts) and *LLGL1* and height (rs112521610 eVariant in both). While colocalizations between *DRICH1* and lymphocyte, white blood cell and neutrophil count GWAS were observed in BCX[49] using European-based eQTLs, colocalization with eosinophil count has not previously been reported. *LLGL1* eQTL signals have previously been observed to colocalize with GWAS of height, standing height and height at 10 years of age in Europeans[48–50].

Discussion

The increased genetic diversity of individuals with African ancestry improves discovery power for regulatory mechanisms compared to Europeans. We performed eQTL studies in venous blood and LCLs in African and European cohorts and discovered that at the level of variants that passed QC, African ancestry cohorts have approximately 2x as many variants by tissue compared to European. We found enrichment of eQTLs in African compared to European cohorts: 3.7x more eQTLs in venous blood, and 5.6x more eQTLs in LCLs. While the discrepancy in number of variants or eQTLs in LCLs can be partially attributed to a nearly 2x greater sample size of the LCL_AFR compared to GEUVADIS, for venous blood PPMI sample size is slightly greater than BRGR.

Variant-for-variant, power increases with the increase in MAF from one population to another[51,52]. Here, 55.1% of eQTLs in BRGR and 61.7% in LCL_AFR are at least 3.16x more frequent as compared to PPMI or GEUVADIS. To facilitate comparisons between African and European eQTLs, we used a common human genome reference (GRCh38) and did not include the 297Mbp of genomic contigs identified in a large African WGS cohort[53] that are absent from GRCh38. Using this updated reference may yield more ancestry-specific eQTLs than were discovered in the present study.

We avoided ancestry-based confounding in eQTL calling by using genetic and expression PCs, and showed that results were not affected by local ancestry. Interestingly, it was recently shown that DNA methylation QTLs are sensitive to local ancestry[54], indicating that further work is needed to fully disentangle the contribution of local ancestry to different modalities of molecular phenotypes in *cis*.

For variants with comparable allele frequencies in African and European cohorts, we quantified sharing of eQTLs using four measurements: exact (eGene, eVariant) pair, colocalization, replication and sharing at the level of eGenes. Within a tissue and across ancestry groups, we find QTLs shared under any definition to be low when considered as a fraction of total. Low eQTL sharing across tissues has previously been observed to hold generally in cross-tissue comparisons[55], which we also observe in comparisons across tissues within an ancestry group.

Kachuri *et al.*[13] stratified eQTLs into multiple tiers of sharing according to a decision tree, which does not readily allow direct comparison to other work. Nonetheless, of the genes they identified as heritable, 47.4% had no overlapping CS in their AFR_{high} cohort as compared to their AFR_{low} cohort. As only 9,609 genes met their definition of heritable, their conclusions broadly agree with ours. Recently, the AFGR[18] cohort released a preprint studying the same HapMap samples comprising the LCL_AFR cohort, and report extensive sharing of eQTL effects with European LCL eQTLs. However, they measure sharing with mashr[56], which specifically assesses whether a variant's effect size is 0 or not in one or more groups. While interesting, this question is per-variant, and is only comparable to our exact match analysis, whereas our use of colocalization is targeted at identifying shared causal signals, and our replication analysis considers credible sets, both of which avoid confounding by unmodeled linkage disequilibrium. Despite the low eQTL sharing, shared eQTLs have similar effect sizes, agreeing with Stranger *et al.*[15] that 34% of genetic effects are shared between ancestry groups with similar effect sizes. Beyond eQTLs, Kanai *et al*.[57] studied the replication of fine-mapped GWAS variants (PIP > 0.9; "hits") across three biobanks of different ancestry groups. Although their definitions of replication are both different from ours and multi-tiered, they find that 55% of hits meet their definition of fine-mapping-based replication, in broad agreement with our

findings. Of the variants found not to replicate, they note that approximately 42% can be attributed to simply lower power in the other cohort(s), and another 42% can be attributed to differential allele frequencies. We speculate that the low replication rates observed in the present work may also be attributable to either lack of power, or differential allele frequency, especially given the pronounced differences in allele spectra we observe between ancestry groups (Fig. 2a, 2f).

Kachuri *et al.*[13] found heritability of gene expression to differ by ancestry group, specifically their AFR_{high} group having more heritability on average as compared to their AFR_{low} group. We observed similar heritability differences when comparing BRGR and PPMI, and further showed that this comparison was confounded by differential genetic variance, and caution against overinterpretation of these findings in either the previous or current work.

In terms of gene mapping, we observed a meaningful increase in gene hypotheses when matching ancestries of GWAS and eQTL cohorts, and discovered colocalizations not previously reported. We also observed increased confidence in a number of previously reported gene hypotheses. While colocalization between GWAS and eQTL signals doesn't necessarily imply that changes in gene expression levels of the eGene mediate genetic effects on disease in every instance[58], the fact that only about 10% of index SNPs in the GWAS Catalog are located in the coding regions[55] does mean that the use of eQTLs will remain a major strategy for nominating gene mapping hypotheses in a vast majority of GWAS loci. GWAS annotation analyses indicate the existence of novel gene-phenotype pairs discoverable only in African cohorts. We anticipate the release of the BRGR and LCL_AFR datasets will enable further research on these and other important questions pertaining to genetic regulation of gene expression in individuals of African descent.

Online Methods

Black Representation in Genomic Research (BRGR). We recruited a cohort of 23andMe consented research participants who self-identified as being of African descent, were predicted to have ≥ 50% African ancestry by the 23andMe Ancestry Composition algorithm[26], had no known blood related cancers or illness, and resided in the continental United States (see Suppl. Fig. 1 for details). WGS was performed on biobanked DNA from saliva samples of BRGR participants, and a venous blood sample was collected for RNA-seq. RNA extraction of 787 venous blood samples that were collected from 23andMe research participants using PAXgene blood RNA tubes was performed at the New York Genome Center (New York, NY). 2 samples were dropped due to contamination and an additional 14 were removed due to sample swaps. DNA extracted from blood cells in saliva samples of 23andMe customers was whole-genome sequenced at the Broad Institute (Cambridge, MA) with aligned CRAMs produced using their

standard pipeline[59]. The average passing aligned read depth was 22.5x in the BRGR cohort. Randomization of BRGR WGS samples was performed to prevent batch effects using the blockTools R library[60]. 976 subjects had usable WGS samples. A total of 737 individuals from this cohort had saliva and venous blood samples of sufficiently high quality for downstream WGS and RNA-seq, respectively (see Suppl. Table 1 for details).

Ethical approval. All biological sample collection was performed in accordance with the terms of informed consents and under an IRB approved protocol. Participants provided informed consent and participated in the research under a research protocol reviewed and approved by an external AAHRPP-accredited IRB, Ethical and Independent Review Services (www.eandireview.com). Participants consented to sharing of genetic and transcriptomic data via the NIH Database of Genotypes and Phenotypes (dbGaP).

LCL_AFR. Lymphoblastoid cell lines (LCLs) from 660 individuals from the 1000 Genomes Project African superpopulation (all except HG02756, which is no longer available) were thawed and clones expanded at the Coriell Institute for Medical Research (Camden, NJ). This included individuals from the Yoruban, Esan, Gambian, Mende and Luhya continental African populations (including all YRI samples from the GEUVADIS project), as well as admixed African-American individuals from the U.S. Southwest and Afro-Caribbean individuals from Barbados (see Suppl. Table 3 for details). RNA from LCLs was extracted at the Coriell Institute. The final sample count was 659.

RNA library preparation and sequencing. BRGR and LCL_AFR samples that met the following QC criteria: (1) minimum of 2μg total DNase-treated RNA, (2) absorbance values of OD260/280 ≥ 1.9 and (3) BioAnalyzer RIN value ≥ 8, were fragmented to 350bp average fragment length and prepared for sequencing using mRNA TruSeq Stranded kits (Illumina, San Diego, CA). Library preparation and RNA-sequencing of paired end 2x100bp reads was performed on Illumina NovaSeq sequencers at the New York Genome Center (New York, NY) to an average coverage of 60M reads.

Parkinson Progression Marker Initiative (PPMI). The PPMI[22,29] cohort contains WGS and a series of functional measurements performed during the course of progression of Parkinson disease. We downloaded RNA-seq data and VCF files (Tier 2 Data) from the Image and Data Archive run by the Laboratory of Neuro Imaging (LONI) at the USC Mark and Mary Stevens Neuroimaging and Informatics Institute. For each sample, we took the RNA-seq from the earliest time point. Average coverage was about 100M reads per sample. The 1,379 initial samples as downloaded from USC were predominantly European ($\mu^{European}$=0.97±0.07), with the plurality of local European ancestry being Ashkenazi Jewish ($\mu^{Ashkenazi}$=0.36±0.48), followed by British and Irish ($\mu^{British\text{-}Irish}$=0.22±0.03). 1,238 samples were of broadly European ancestry. Relatedness was highest in individuals of Ashkenazi descent (defined as $\mu^{Ashkenazi} \geq 0.25$) with

49.2% of pairwise comparisons having at least 0.05% of their genome that is IBD (Suppl. Fig. 2, Suppl. Table 7). In order to remove a potential bias in relatedness and heritability calculations due to including individuals from this founder population, and to standardize sample numbers in BRGR and PPMI cohorts so we remove an obvious confounder in eQTL statistical discovery power, we have excluded the 481 Ashkenazi Jewish individuals from our PPMI cohort. After removing 5 additional samples with incomplete gene expression data, the final sample count was 752.

GEUVADIS. We downloaded RNA-seq data for the 358 European ancestry samples from the Geuvadis Consortium from https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-GEUV-1. Average coverage was about 20M reads per sample. LCL_AFR and GEUVADIS samples were part of the 1000 Genomes Project, and we downloaded high-coverage WGS data generated by the New York Genome Center[20] from https://www.internationalgenome.org/data-portal/data-collection/30x-grch38.

Variant calling. Variant calls were made using DeepVariant[61] with joint calling performed by GLnexus[62,63]. The 1000 Genomes Project WGS samples (LCL_AFR and GEUVADIS cohorts) were processed with DeepVariant-0.8.0 and GLNexus-1.2.3 while for BRGR we used DeepVariant-1.1.0 and GLNexus-1.2.7. For PPMI, variant call files were downloaded through the LONI portal; variant calling in this cohort has been previously described[64]. Multi-allelic sites were split into individual alleles using `bcftools norm -m -any`. Genotypes with GQ<20 were set to missing, then variants were excluded where any of: >20% of genotypes missing, had no genotype with an alternative allele present or HWE exact test p-value<10^{-50}.

Ancestry inference. To identify the ancestral origins of chromosomal segments across cohorts, we performed local ancestry inference using 23andMe's Ancestry Composition[26]. Ancestry Composition uses support vector machine classifiers to assign one of 45 fine-scale ancestry populations to locally phased 300-SNP windows based on 541,948 SNPs. These preliminary assignments are next processed with an autoregressive pair hidden Markov model that smooths and corrects any phasing errors. The resulting posterior probabilities are recalibrated with an isotonic regression model. Local ancestry segments within each individual are finally summarized to produce a genome-wide proportion. For the purpose of this study, any population assignment with a mean less than 5% in either PPMI, African LCL, or BRGR was binned into a trace ancestry category. Finally, we performed independent randomization tests on 18 relevant ancestry populations to determine if the BRGR cohort ancestral representation is significantly different from that of a larger subset of individuals of African American participants. In each case, we performed a randomization test by randomly sampling (with replacement) 1000 23andMe research participants who identified as African-American based on survey answers and have >50% African ancestry (n=203,916) across 1000 iterations. For each iteration, we calculated the difference in mean genome-wide ancestry proportions

between BRGR and the initial random subset ($\Delta\mu_{Ancestry^1}$), then created two additional cohorts of 1000 individuals each by randomly sampling individuals from both starting cohorts and determined the difference in the means of these second cohorts ($\Delta\mu_{Ancestry^2}$). We determined p-values as the number of times that $\Delta\mu_{Ancestry^2}$ was greater than or equal to $\Delta\mu_{Ancestry^1}$ out of the 1000 iterations.

Identity-by-descent (IBD). To determine relatedness between individuals we calculated the amount of DNA that is identical-by-descent between all pairs of individuals using phase-aware templated positional Burrows-Wheeler transform IBD detection (TPBWT-IBD)[65]. TPBWT-IBD was performed on all pairwise combinations of individuals across the PPMI, African LCL, and BRGR cohorts and a subset of additional 1000 Genomes Project[19] and Human Genome Diversity Cell Line Panel (HGDP)[66] populations. To maximize the accuracy of IBD detection, we used default parameters on an optimized set of 541,948 SNPs and retained IBD segments ≥ 5 centimorgans (cM). Finally, to visualize fine-scale ancestry, we arranged individuals in a graph based on the total amount of IBD they share using the ForceAtlas2[28] layout. Force Atlas is an algorithm that situates individuals (or nodes) in a graph using a physical magnetic model. In this case, individuals with more IBD sharing will be attracted to one another and individuals with less IBD sharing are repelled. ForceAtlas2 runs until balance between repulsion and attraction is achieved, essentially illustrating fine-structure of individuals using the total IBD shared.

RNA-seq mapping. RNA-seq reads were aligned to the human reference genome GRCh38 using STAR[67] with 2 passes. Quality control for technical factors was done with FastQC[68] and MultiQC[69], and sample swaps were checked for using verifyBamID[70]. Per sample strand orientation was verified using the `infer_experiment.py` module in RSeQC[71]. Gene-level expression was quantified by HTSeq[72] (`--stranded=reverse`) using GENCODE[73,74] v28 gene models. RNA-seq reads were aligned to the whole transcriptome, but subsequently all analyses here and throughout the manuscript were limited to protein-coding genes.

RNA-seq normalization. RNA-seq datasets were filtered for genes where ≥20% samples had a CPM (count per million) ≥0.1. Genes passing this threshold were further normalized first by scaling by library size using edgeR[75,76], then converting to $\log_2$ scale. Finally, for all genes with more than 95% of samples exhibiting normalized expression value less than 4 standard deviations away from the expression mean, samples were right-truncated to the 95%-ile of normalized expression values; genes with 5% or more samples exhibiting normalized expression greater than 4 standard deviations from the mean were discarded from downstream analysis.

eQTL discovery: eQTL discovery was done using the susieR package[30] with default parameters: maximum 10 independent signals per gene, sum of credible set-level posterior inclusion probabilities (“coverage”) of 0.95 and minimum within-credible-set correlation

("purity") of 0.5. The analysis was restricted to protein-coding genes and variants within a ±1Mbp window centered on the tested gene's transcription start site (TSS). We tested SNVs and indels ≤ 500bp that had in-sample MAF≥1% and missingness < 5%. Covariates used across all datasets for eQTL calling included age, sex and 10 genetic PCs. Expression PCs were used to adjust for hidden covariates in RNA-seq data. The number of PCs was selected per dataset using the elbow method[77], leading to 18 expression PCs for BRGR, 31 for LCL_AFR, 19 for PPMI and 23 for GEUVADIS. BRGR also included unique read fraction, RNA integrity number[78] and % African ancestry as inferred by 23andMe Ancestry Classifier model[26]. GTEx also included sequencing platform and cohort (post-mortem, organ donor or surgical) as covariates.

African-American GWAS. To assess the utility of the eQTL datasets for annotating GWAS hits, we downloaded African-American GWAS summary statistics from the VA Million Veteran Program (specifically blood lipids[23], VTE[24] and T2D[25]), and GWAS of blood traits from the Blood Cell Consortium[14]. Summary of all included GWAS are shown in Suppl. Table 5. As these GWAS had moderate sample sizes and numbers of genome-wide significant hits, we did not fully condition association signals in these studies.

In addition, we ran a GWAS of height in 23andMe's African-American cohort, using our standard GWAS pipeline as described previously[79]. In short, we compute association test results for the genotyped and the imputed SNPs. For tests using imputed data, we use the imputed dosages rather than best-guess genotypes. As standard, we include covariates for age, gender, the first 6 PC of genetic ancestry to account for residual population structure, and indicators for genotype platforms to account for genotype batch effects. The association test p-value we report is computed using a likelihood ratio test. Association tests are performed by linear regression. Results for the chrX are computed similarly, with male genotypes coded as if they were homozygous diploid for the observed allele. Height GWAS signals were fully conditioned using the step-down conditional process: for each association genome wide, we re-ran the association test with the top variant from the preceding step in the model as an additional covariate at each iteration. The process is repeated up to 20 times or until no association is detected at p-value≤10^{-5}. All conditionally independent variants identified were then introduced in a joint model. At each iteration one of the variants is left out to compute conditional leave each out (CLEO) statistics to be used in downstream analysis.

Variant-to-gene mapping. African-American GWAS hits were linked to eGenes via colocalization analysis. Approximate Bayes factors[80] were derived directly from MVP GWAS marginal summary statistics as the GWAS were likely underpowered to confidently detect secondary signals. For 23andMe African American height, each marginal association was further analyzed for conditionally independent signals using conditional leave each out (CLEO) analysis, and ABFs were computed using these conditionally-resolved statistics. ABFs

representing independent signals were colocalized against SuSiE-derived eGenes using the coloc R package function coloc.susie_bf, with default parameters. Gene-trait pairs with a posterior probability of colocalization $H_4 \geq \{0.5, 0.8\}$ are reported as colocalizing.

Local ancestry-based eQTL calling. Local ancestry was incorporated into cis-eQTL calling using the Tractor[34] model. We used Tractor to estimate European and African ancestry-specific effects and p-values by including alternate allele counts for each ancestry into the model. This was compared to a "standard" generalized linear regression model which measured the alternate allele effect, regardless of haplotype ancestry. In both models, we tested all variants within a 1Mb cis-window of each gene with MAF ≥ 5% and MAC ≥ 10 for both ancestry tracts. We adjusted for covariates including age, sex, 10 genetic PCs, 35 PEER factors, and sequencing factors in both models and for the Tractor model we also adjusted for the number of African ancestry haplotypes per locus.

Heritability analysis. GCTA GREML[33] was used to estimate cis-heritability for significant eGenes in BRGR and PPMI, separately, for autosomal variants with MAF ≥ 0.01. The phenotype used for GREML analysis was the normalized expression residualized on age, sex, 10 PCs of genetic ancestry, 35 PEER factors and sequencing covariates. We identified 9,777 eGenes that showed significant (p-value < 0.05) heritability in both cohorts. To avoid inflating estimates of h^2 through relatedness, we filtered each cohort to the subset of individuals for which pairwise IBD < 0.025. To identify a high-confidence gene set for comparing heritability between BRGR and PPMI, we first derived fine-mapped[30,81] credible sets for eQTLs in each cohort.

We focused our comparison on the 183 genes which had significant heritability estimates in both cohorts, and for which all 95% credible sets colocalized[29] across the two cohorts with $P(H_{12}) > 0.5$. For each gene *g* in each cohort, we approximately quantified the genetic variation (V_g) that contributes to gene expression as the sum of genotype variance across the index variants in BRGR of *k* credible sets as

$$v_g := \sum_k 2 f_k \left(1 - f_k\right)$$

where f_k is the cohort-specific in-sample MAF for gene *g*. Here we assume that the genetic effect sizes are the same across cohorts and independent on variants so that, on average across the genes, we have $V_g^{BRGR} / V_g^{PPMI} \approx v_g^{BRGR} / v_g^{PPMI}$. We use the BRGR index variants under the assumption that these colocalizing signals are shared causal eQTLs, and note that using PPMI index variants does not change our conclusions (Fig. 3). We compared the difference in heritability between BRGR and PPMI ($\Delta h^2 = h^2_{BRGR} - h^2_{PPMI}$) and the difference in v_g ($\Delta v_g = v_{g,\,BRGR} - v_{g,\,PPMI}$).

For each cohort, we define the ratio of genetic to environmental variance explained for a given gene *g* as

$$q_g = \frac{h^2}{1-h^2} = \frac{V_g}{V_e}.$$

With terms $k_1 = V_e^{BRGR} / V_e^{PPMI}$ (ratio of environmental variances) and $k_2 = V_g^{BRGR} / V_g^{PPMI}$ (ratio of genetic variances), for gene *g* we can relate BRGR and PPMI as

$$q_g^{BRGR} = \frac{V_g^{BRGR}}{V_e^{BRGR}} = \frac{k_2}{k_1} q_g^{PPMI}.$$

That is, the ratio of unexplained to explained variance for a gene in PPMI is proportional to the same quantity in BRGR. Under the assumption of approximately equal environmental variation, $k_1 \approx 1$, and

$$\frac{q_g^{BRGR}}{q_g^{PPMI}} \approx k_2 = \frac{V_g^{BRGR}}{V_g^{PPMI}} \approx \frac{v_g^{BRGR}}{v_g^{PPMI}}$$

across genes broadly. This latter hypothesis can be tested by defining similar difference terms $\Delta\log(q)$ and $\Delta\log(v_g)$, which are expected to be correlated. Relatedly, under this line of reasoning, $\log(q_{BRGR}) = \Delta\log(v_g) + \log(q_{PPMI})$.

OpenTargets Genetics access. OpenTargets Genetics v22.10 was accessed via API with queries against colocalization tables. We consider all gene-phenotype pairs reported in OpenTargets, regardless of eQTL or GWAS discovery cohorts, or tissues or cell types of discovery in the case of eQTLs. For a given eGene-phenotype pair discovered through one of the AFR cohorts, we search for similar OpenTargets Genetics phenotypes (i.e. MPV in BCX and 'Mean platelet volume' in UKB), and report the result from phenotype with the largest colocalization H_4 with the eGene.

Data availability

Consistent with the research consent provided, we made these datasets publicly available to all qualified researchers. RNA-seq and WGS data for Black Representation in Genomic Research is available as dbGaP study phs002969.v1.p1, and RNA-seq for the 1000 Genomes Project African superpopulation LCLs as SRA project PRJNA1108327. Upon publication the full summary statistics for the 23andMe African-American height GWAS and the BRGR eQTL study will be made available through 23andMe to qualified researchers under an agreement with 23andMe that protects the privacy of the 23andMe participants. Please visit https://research.23andme.com/collaborate/#dataset-access for more information and to apply to access the data. Upon publication genome-wide association summary statistics for the African LCL eQTL study will be freely available for download via https://www.internationalgenome.org.

Acknowledgements

We thank the research participants and employees of 23andMe for making this work possible. We in particular thank the research participants of the Black Representation in Genomic Research (BRGR) study. BRGR participants provided informed consent and participated in the research under a protocol approved by the external AAHRPP-accredited IRB, Ethical & Independent Review Services (E&I Review).

The following members of the 23andMe Research Team contributed to this study: Stella Aslibekyan, Adam Auton, Elizabeth Babalola, Robert K. Bell, Jessica Bielenberg, Ninad S. Chaudhary, Zayn Cochinwala, Sayantan Das, Emily DelloRusso, Payam Dibaeinia, Sarah L. Elson, Nicholas Eriksson, Chris Eijsbouts, Teresa Filshtein, Pierre Fontanillas, Davide Foletti, Will Freyman, Zach Fuller, Julie M. Granka, Chris German, Éadaoin Harney, Alejandro Hernandez, Barry Hicks, David A. Hinds, M. Reza Jabal-Ameli, Ethan M. Jewett, Yunxuan Jiang, Sotiris Karagounis, Lucy Kaufmann, Matt Kmiecik, Katelyn Kukar, Alan Kwong, Keng-Han Lin, Yanyu Liang, Bianca A. Llamas, Aly Khan, Steven J. Micheletti, Matthew H. McIntyre, Meghan E. Moreno, Priyanka Nandakumar, Dominique T. Nguyen, Jared O'Connell, Steven J. Pitts, G. David Poznik, Alexandra Reynoso, Shubham Saini, Morgan Schumacher, Leah Selcer, Anjali J. Shastri, Jingchunzi Shi, Suyash Shringarpure, Keaton Stagaman, Teague Sterling, Qiaojuan Jane Su, Joyce Y. Tung, Susana A. Tat, Vinh Tran, Xin Wang, Wei Wang, Catherine H. Weldon, Amy L. Williams, Peter Wilton.

Some of the data used in the preparation of this manuscript was obtained in 2019 from the Parkinson's Progression Markers Initiative (PPMI) database (www.ppmi-info.org/access-data-specimens/download-data), RRID:SCR_006431. For up-to-date information on the study, visit www.ppmi-info.org. Our analysis used data from PPMI made available after either a pre-defined embargo or investigator submission of an associated cloud transfer request to the PPMI Data Access Committee. Protocol information for the Parkinson's Progression Markers Initiative (PPMI) Clinical - Establishing a Deeply Phenotyped PD Cohort AM 3.2. can be found on protocols.io or by following this link: https://dx.doi.org/10.17504/protocols.io.n92ldmw6ol5b/v2. PPMI – a public-private partnership – is funded by the Michael J. Fox Foundation for Parkinson's Research, and funding partners; including 4D Pharma, Abbvie, AcureX, Allergan, Amathus Therapeutics, Aligning Science Across Parkinson's, AskBio, Avid Radiopharmaceuticals, BIAL, BioArctic, Biogen, Biohaven, BioLegend, BlueRock Therapeutics, Bristol-Myers Squibb, Calico Labs, Capsida Biotherapeutics, Celgene, Cerevel Therapeutics, Coave Therapeutics, DaCapo Brainscience, Denali, Edmond J. Safra Foundation, Eli Lilly, Gain Therapeutics, GE HealthCare, Genentech, GSK, Golub Capital, Handl Therapeutics, Insitro, Jazz Pharmaceuticals, Johnson & Johnson Innovative Medicine, Lundbeck, Merck, Meso Scale Discovery, Mission Therapeutics, Neurocrine Biosciences,

Neuron23, Neuropore, Pfizer, Piramal, Prevail Therapeutics, Roche, Sanofi, Servier, Sun Pharma Advanced Research Company, Takeda, Teva, UCB, Vanqua Bio, Verily, Voyager Therapeutics, the Weston Family Foundation and Yumanity Therapeutics.

Author contributions

All authors contributed substantially to this manuscript.

Competing interests

KFB, RS, YL, SM, PN, AS, AT, RJT, BH, JOC, SS, KK, MM, EB, CW, AP, RG, SJP, VV are current or former employees of 23andMe, Inc. and may own shares of stock or stock options of 23andMe. KFB is an employee of Genentech, Inc. AS is an employee of CSL. KS, GA, AC, PG, LH, RM, DS are employees of GSK plc. TC is an employee of Exai Bio. AT is an employee of Gilead Sciences, Inc. JOC is an employee of AstraZeneca plc. AP is an employee of Genomenon, Inc.

Figure Legends

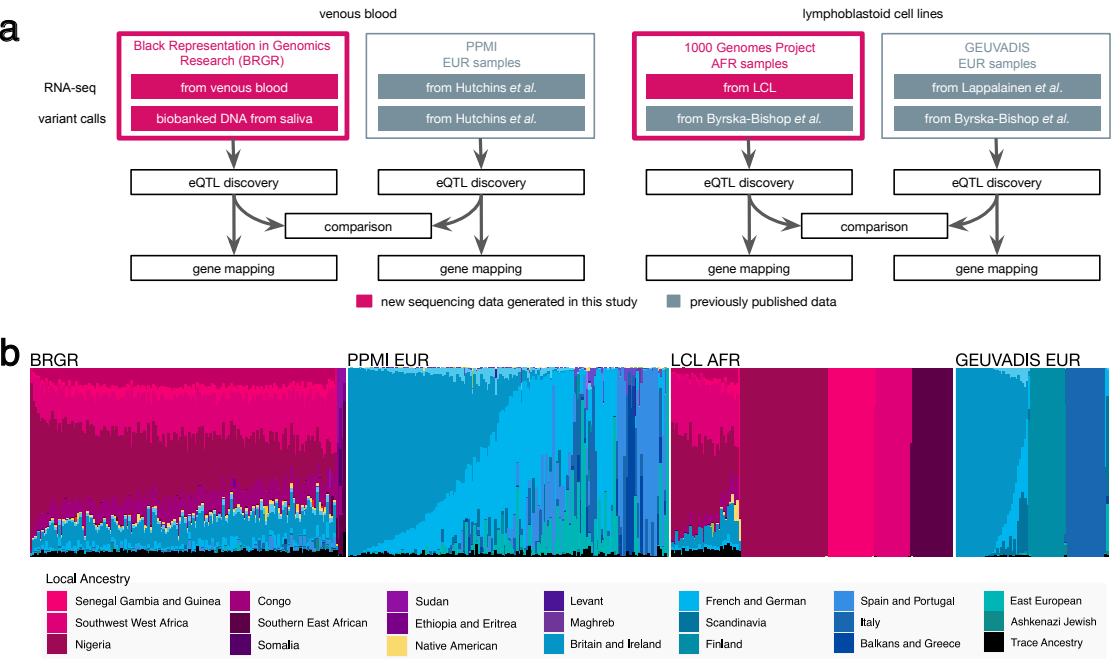
Figure 1 | Schematic diagram of the study design and Ancestry Composition estimates for each eQTL cohort. (a) Schematic diagram of the four eQTL study arms. Fuchsia-colored boxes indicate new RNA-seq or WGS data generated for this study. **(b)** Results of the Ancestry Composition algorithm applied to the four eQTL cohorts. Each vertical line represents genome-wide proportions for a single individual.

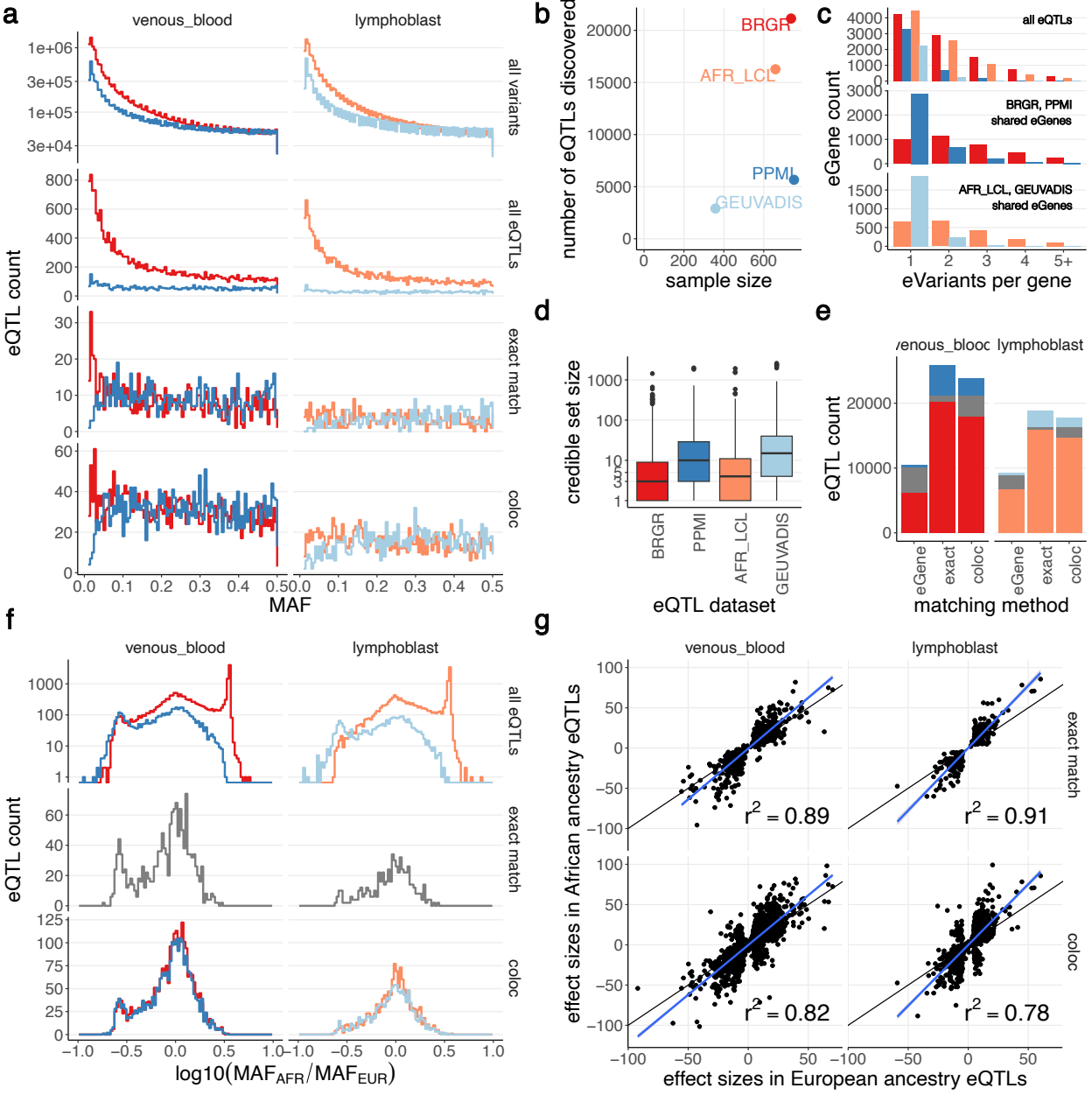
Figure 2 | Statistics of eQTL calls and eQTL sharing across cohorts. (a) Minor allele frequency histograms of all variants and all detected eQTLs, as well as for eQTLs that were also observed in the matching tissue in the complementary population. **(b)** Scatter plot of number of eQTLs discovered against cohort size shows higher numbers of eQTLs in African ancestry groups even when normalized by study size. **(c)** Distributions of numbers of distinct eVariants per eGene show a higher number of eVariants in AFR cohorts. **(d)** Distributions of credible set sizes have smaller medians in European ancestry datasets. **(e)** Sharing of eGenes and eQTLs between African and European studies in the same tissue or cell type is broadly low, irrespective of the method of comparison. **(f)** Spectra of $\log10(MAF_{AFR} / MAF_{EUR})$ for eQTLs from African (red) and European (blue) ancestry cohorts, in venous blood and LCLs respectively. **(g)** Shared eQTLs are largely concordant in their direction of effect (Pearson r^2 = 0.89-0.91 for exactly the same eVariants, r^2 = 0.79-0.83 for colocalizing eVariants), with on

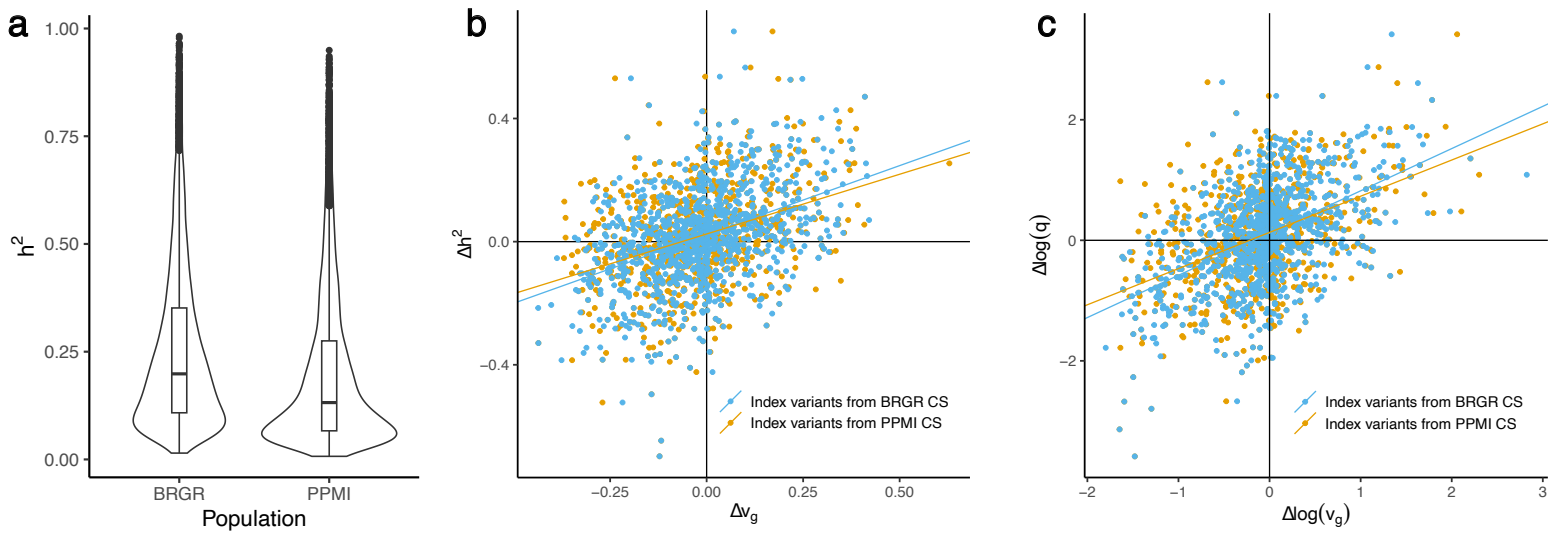
average slightly elevated magnitude of effect in African cohorts (blue line is best linear fit, black line x=y for comparison).

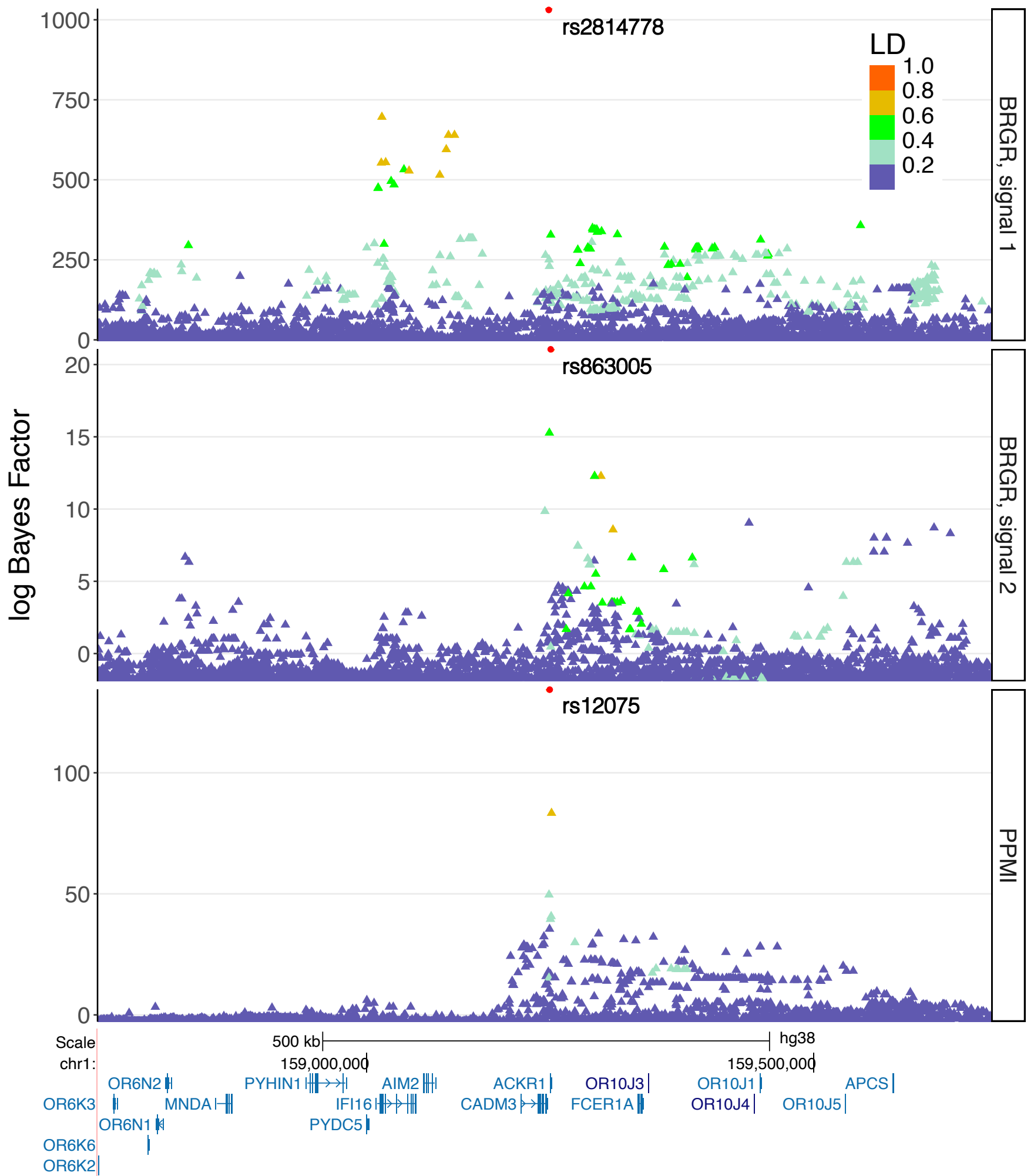
Figure 3 | Cis-heritability of gene expression in venous blood and difference in genetic diversity explaining population differences. (a) Comparison of h^2 across populations (BRGR in blue and PPMI in yellow) showing a significant difference in h^2 (p-value < 2.2×10^{-16}). **(b), (c)** Among the 183 genes whose credible sets are colocalized between BRGR and PPMI, we define the causal eQTLs as the index variants from BRGR (blue) or PPMI (yellow) credible sets and calculate the genetic diversity v_g of the causal eQTLs in BRGR and PPMI respectively. To examine the relation between genetic diversity and heritability, we introduce a transformed heritability, $q = h^2 / (1 - h^2)$. **(b)** For each of the 183 genes, differences in genetic diversity $\Delta v_g = v_{g,\,BRGR} - v_{g,\,PPMI}$ are shown on x-axis and differences in heritability $\Delta h^2 = h^2_{BRGR} - h^2_{PPMI}$ are shown on y-axis. The line is drawn from the linear fit $\Delta h^2 \sim 1 + \Delta v_g$. **(c)** For each of the 183 genes, differences in logarithm of genetic diversity between BRGR and PPMI ($\Delta\log(v_g) = \log(v_{g,\,BRGR}) - \log(v_{g,\,PPMI})$) are shown on x-axis and differences in logarithm of transformed heritability ($\Delta\log(q) = \log(q_{BRGR}) - \log(q_{PPMI})$) are shown on y-axis. The line is drawn from the linear fit $\Delta\log(q) \sim 1 + \Delta\log(v_g)$.

Figure 4 | *ACKR1* eQTLs in BRGR and PPMI, colored by in-sample LD to highest logBF variant. Shown are SuSiE log Bayes factors (logBFs) for each SNP in a given eQTL signal. logBFs reflect the evidence that a SNP is the causal SNP, and are proportional to posterior inclusion probabilities. BGRG signal #1 (PIP = 1.0), with the Duffy null allele variant rs2814778 as the eVariant. BRGR signal #2 (PIP = 0.996), with rs863005 as the eVariant. Compared to signal #2, #1 exhibits a longer segment of variants in linkage disequilibrium with each other, which may reflect the selective pressure specifically on the haplotype carrying the CC allele of rs2814778. PPMI signal for rs12075 (PIP = 0.999). The signal is localized to the eVariant, and is in linkage disequilibrium with few nearby variants.


a
venous blood
lymphoblastoid cell lines
Black Representation in Genomics Research (BRGR)
PPMI EUR samples
1000 Genomes Project AFR samples
GEUVADIS EUR samples
RNA-seq
variant calls
from venous blood
biobanked DNA from saliva
from Hutchins et al.
from Hutchins et al.
from LCL
from Byrska-Bishop et al.
from Lappalainen et al.
from Byrska-Bishop et al.
eQTL discovery
eQTL discovery
eQTL discovery
eQTL discovery
comparison
comparison
gene mapping
gene mapping
gene mapping
gene mapping
new sequencing data generated in this study
previously published data
b
BRGR
PPMI EUR
LCL AFR
GEUVADIS EUR
Local Ancestry
Senegal Gambia and Guinea
Southwest West Africa
Nigeria
Congo
Southern East African
Somalia
Sudan
Ethiopia and Eritrea
Native American
Levant
Maghreb
Britain and Ireland
French and German
Scandinavia
Finland
Spain and Portugal
Italy
Balkans and Greece
East European
Ashkenazi Jewish
Trace Ancestry


a
venous_blood
lymphoblast
all variants
all eQTLs
exact match
coloc
eQTL count
MAF
b
number of eQTLs discovered
sample size
BRGR
AFR_LCL
PPMI
GEUVADIS
c
all eQTLs
BRGR, PPMI
shared eGenes
AFR_LCL, GEUVADIS
shared eGenes
eGene count
eVariants per gene
d
credible set size
eQTL dataset
e
eQTL count
matching method
eGene
exact
coloc
f
log10(MAF_AFR/MAF_EUR)
g
effect sizes in African ancestry eQTLs
effect sizes in European ancestry eQTLs
r² = 0.89
r² = 0.91
r² = 0.82
r² = 0.78


a
1.00
0.75
0.50
0.25
0.00
h2
BRGR
PPMI
Population
b
0.4
0.0
−0.4
Δh2
−0.25
0.00
0.25
0.50
Δvg
Index variants from BRGR CS
Index variants from PPMI CS
c
2
0
−2
Δlog(q)
−2
−1
0
1
2
3
Δlog(vg)
Index variants from BRGR CS
Index variants from PPMI CS


log Bayes Factor
rs2814778
rs863005
rs12075
LD
1.0
0.8
0.6
0.4
0.2
BRGR, signal 1
BRGR, signal 2
PPMI
1000
750
500
250
0
20
15
10
5
0
100
50
0
Scale
500 kb
hg38
chr1:
159,000,000
159,500,000
OR6N2
PYHIN1
AIM2
ACKR1
OR10J3
OR10J1
APCS
OR6K3
MNDA
IFI16
CADM3
FCER1A
OR10J4
OR10J5
OR6N1
PYDC5
OR6K6
OR6K2

Tables

Table 1 | Summary of the four cohorts and eQTLs discoveries. eQTLs are defined as (eVariant, eGene) pairs, eVariants as variants associated with a change in gene expression of one or more protein-coding eGenes, and eGenes as protein-coding genes having one or more eQTLs.

Cell line / tissue	**Venous blood**		**LCL**	
Dataset	**BRGR**	**PPMI**	**LCL_AFR**	**GEUVADIS**
Population	**AFR**	**EUR**	**AFR**	**EUR**
Sample size	737	752	659	358
Total variants in cohort	15,598,612	8,672,989	15,585,478	8,895,186
Genes tested	14,162	17,233	13,584	14,624
eGenes discovered	10,044 (70.9%)	4,277 (24.8%)	8,843 (65.1%)	2,527 (17.3%)
eQTLs discovered	21,130	5,661	16,265	2,904
eVariants discovered	20,295	5,380	15,832	2,843

Table 2 | Summary of eQTLs and eGenes overlap across populations and tissues.
Replicating here means that (eGene, eVariant) pair from the first dataset appeared in one of the credible sets for the same eGene in the second dataset. We report both directions of replication (1^{st} replicating in 2^{nd}; 2^{nd} replicating in 1^{st}) separately. Where applicable, numbers in square brackets are values of Jaccard similarity coefficient defined as the size of the intersection divided by the size of the union. For replication, numbers in square brackets are % of hits replicating. Numbers and % in parentheses indicate eQTLs with a common directionality of effect. For exact eVariant match and replication, direction was based on the sign of coefficient of the eVariant; for coloc, direction was based on the sign of the z-score correlation coefficient.

	BRGR and PPMI (both venous blood)	LCL_AFR and GEUVADIS (both LCL)	BRGR and LCL_AFR (both AFR)	PPMI and GEUVADIS (both EUR)
Exact eVariant match	872 [0.03] (866; 99%)	341 [0.02] (341; 100%)	1,987 [0.06] (1,938; 98%)	176 [0.02] (168; 95%)
Colocalizing eQTLs ($H_4 \geq 0.8$)	2,772 [0.11] (2,534; 91%)	1,330 [0.07] (1,199; 90%)	4,161 [0.10] (3,858; 93%)	656 [0.08] (558; 85%)
Colocalizing eQTLs ($H_4 \geq 0.5$)	3,145 [0.13] (2,820, 90%)	1,569 [0.09] (1,398, 89%)	4,545 [0.14] (4,159; 92%)	769 [0.10] (630; 82%)
1^{st} replicating in 2^{nd}	4,508 [21%] (4,408; 98%)	2,621 [16%] (2,598; 99%)	3,406 [16%] (3,279; 96%)	836 [15%] (763; 91%)
2^{nd} replicating in 1^{st}	1,295 [23%] (1,283; 99%)	511 [18%] (511; 100%)	3,405 [21%] (3,278; 96%)	536 [18%] (492; 92%)
Shared eGenes	3,822 [0.37]	2,153 [0.23]	6,843 [0.57]	1,331 [0.24]

Table 3 | Summary of variant-to-gene mapping of African-American GWAS signals. MVP - the VA Million Veteran Program; BCX - Blood Cell Consortium. Detailed per-phenotype GWAS breakdown is provided in Suppl. Table 5.

GWAS dataset	GWAS signals	H_4 cutoff	Number of GWAS-to-eQTL signal colocalizations				
			BRGR AFR	PPMI EUR	LCL_AFR AFR	GEUVADIS EUR	r^2 with coding
MVP	312	0.8	6	2	4	0	15
		0.5	9	2	4	1	16
BCX	461	0.8	32	12	14	8	5
		0.5	41	16	19	9	5
Height	810	0.8	32	22	26	8	21
		0.5	52	35	46	17	37

References

1. Fatumo, S. *et al.* A roadmap to increase diversity in genomic studies. *Nat. Med.* **28**, 243–250 (2022).
2. The H3Africa Consortium *et al.* Enabling the genomic revolution in Africa. *Science* **344**, 1346–1348 (2014).
3. Choudhury, A. *et al.* Whole-genome sequencing for an enhanced understanding of genetic variation among South Africans. *Nat. Commun.* **8**, 2062 (2017).
4. The Malaria Genomic Epidemiology Network. A global network for investigating the genomic epidemiology of malaria. *Nature* **456**, 732–737 (2008).
5. The All of Us Research Program Genomics Investigators *et al.* Genomic data in the All of Us Research Program. *Nature* (2024) doi:10.1038/s41586-023-06957-x.
6. Verma, A. *et al.* Diversity and scale: Genetic architecture of 2068 traits in the VA Million Veteran Program. *Science* **385**, eadj1182 (2024).
7. McClellan, J. M., Lehner, T. & King, M.-C. Gene Discovery for Complex Traits: Lessons from Africa. *Cell* **171**, 261–264 (2017).
8. The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* 1318–1330 (2020).
9. Gay, N. R. *et al.* Impact of admixture and ancestry on eQTL analysis and GWAS colocalization in GTEx. *Genome Biol.* **21**, 233 (2020).
10. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
11. Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768–772 (2010).
12. Zhong, Y. *et al.* Discovery of novel hepatocyte eQTLs in African Americans. *PLOS Genet.* **16**, e1008662 (2020).

13. Kachuri, L. *et al.* Gene expression in African Americans, Puerto Ricans and Mexican Americans reveals ancestry-specific patterns of genetic architecture. *Nat. Genet.* **55**, 952–963 (2023).
14. 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).
15. Stranger, B. E. *et al.* Patterns of Cis Regulatory Variation in Diverse Human Populations. *PLoS Genet.* **8**, e1002639 (2012).
16. Mogil, L. S. *et al.* Genetic architecture of gene expression traits across diverse populations. *PLOS Genet.* **14**, e1007586 (2018).
17. Martin, A. R. *et al.* Transcriptome Sequencing from Diverse Human Populations Reveals Differentiated Regulatory Architecture. *PLoS Genet.* **10**, e1004549 (2014).
18. DeGorter, M. K. *et al. Transcriptomics and Chromatin Accessibility in Multiple African Population Samples*. http://biorxiv.org/lookup/doi/10.1101/2023.11.04.564839 (2023) doi:10.1101/2023.11.04.564839.
19. The 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
20. Byrska-Bishop, M. *et al.* High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**, 3426-3440.e19 (2022).
21. Kern, F. *et al.* Deep sequencing of sncRNAs reveals hallmarks and regulatory modules of the transcriptome during Parkinson's disease progression. *Nat. Aging* **1**, 309–322 (2021).
22. Craig, D. W. *et al.* RNA sequencing of whole blood reveals early alterations in immune cells and gene expression in Parkinson's disease. *Nat. Aging* **1**, 734–747 (2021).
23. Global Lipids Genetics Consortium *et al.* Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. *Nat. Genet.* **50**, 1514–1523 (2018).
24. INVENT Consortium *et al.* Genome-wide association analysis of venous thromboembolism

identifies new risk loci and genetic overlap with arterial vascular disease. *Nat. Genet.* **51**, 1574–1579 (2019).

25. 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).
26. Durand, E. Y., Do, C. B., Mountain, J. L. & Macpherson, J. M. *Ancestry Composition: A Novel, Efficient Pipeline for Ancestry Deconvolution*. http://biorxiv.org/lookup/doi/10.1101/010512 (2014) doi:10.1101/010512.
27. Micheletti, S. J. *et al.* Genetic Consequences of the Transatlantic Slave Trade in the Americas. *Am. J. Hum. Genet.* **107**, 265–277 (2020).
28. Jacomy, M., Venturini, T., Heymann, S. & Bastian, M. ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLoS ONE* **9**, e98679 (2014).
29. Hutchins, E. *et al. Quality Control Metrics for Whole Blood Transcriptome Analysis in the Parkinson's Progression Markers Initiative (PPMI)*. http://medrxiv.org/lookup/doi/10.1101/2021.01.05.21249278 (2021) doi:10.1101/2021.01.05.21249278.
30. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A Simple New Approach to Variable Selection in Regression, with Application to Genetic Fine Mapping. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **82**, 1273–1300 (2020).
31. Wallace, C. A more accurate method for colocalisation analysis allowing for multiple causal variants. *PLOS Genet.* **17**, e1009440 (2021).
32. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genet* **10**, e1004383 (2014).
33. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A Tool for Genome-wide

Complex Trait Analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).

34. Atkinson, E. G. *et al.* Tractor uses local ancestry to enable the inclusion of admixed individuals in GWAS and to boost power. *Nat. Genet.* **53**, 195–204 (2021).
35. Hou, K., Bhattacharya, A., Mester, R., Burch, K. S. & Pasaniuc, B. On powerful GWAS in admixed populations. *Nat. Genet.* **53**, 1631–1633 (2021).
36. Rappoport, N., Simon, A. J., Amariglio, N. & Rechavi, G. The Duffy antigen receptor for chemokines, *ACKR 1* ,– 'Jeanne DARC ' of benign neutropenia. *Br. J. Haematol.* **184**, 497–507 (2019).
37. Merz, L. E. *et al.* Absolute neutrophil count by Duffy status among healthy Black and African American adults. *Blood Adv.* **7**, 317–320 (2023).
38. Atallah-Yunes, S. A., Ready, A. & Newburger, P. E. Benign ethnic neutropenia. *Blood Rev.* **37**, 100586 (2019).
39. Legge, S. E. *et al.* The Duffy-null genotype and risk of infection. *Hum. Mol. Genet.* **29**, 3341–3349 (2020).
40. He, W. *et al.* Duffy Antigen Receptor for Chemokines Mediates trans-Infection of HIV-1 from Red Blood Cells to Target Cells and Affects HIV-AIDS Susceptibility. *Cell Host Microbe* **4**, 52–62 (2008).
41. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
42. Srivastava, K. *et al.* ACKR1 Alleles at 5.6 kb in a Well-Characterized Renewable US Food and Drug Administration (FDA) Reference Panel for Standardization of Blood Group Genotyping. *J. Mol. Diagn.* **22**, 1272–1279 (2020).
43. Schnabel, R. B. *et al.* Duffy antigen receptor for chemokines (Darc) polymorphism regulates circulating concentrations of monocyte chemoattractant protein-1 and other inflammatory mediators. *Blood* **115**, 5289–5299 (2010).

44. Guo, X. *et al.* Endothelial ACKR1 is induced by neutrophil contact and down-regulated by secretion in extracellular vesicles. *Front. Immunol.* **14**, 1181016 (2023).
45. Peiper, S. C. *et al.* The Duffy antigen/receptor for chemokines (DARC) is expressed in endothelial cells of Duffy negative individuals who lack the erythrocyte receptor. *J. Exp. Med.* **181**, 1311–1317 (1995).
46. Vuckovic, D. *et al.* The Polygenic and Monogenic Basis of Blood Traits and Diseases. *Cell* **182**, 1214-1231.e11 (2020).
47. Tukiainen, T. *et al.* Chromosome X-Wide Association Study Identifies Loci for Fasting Insulin and Height and Evidence for Incomplete Dosage Compensation. *PLoS Genet.* **10**, e1004127 (2014).
48. Neale, B. M. UK Biobank GWAS Round 2. http://www.nealelab.is/uk-biobank.
49. Ochoa, D. *et al.* The next-generation Open Targets Platform: reimagined, redesigned, rebuilt. *Nucleic Acids Res.* **51**, D1353–D1359 (2023).
50. Barton, A. R., Sherman, M. A., Mukamel, R. E. & Loh, P.-R. Whole-exome imputation within UK Biobank powers rare coding variant association and fine-mapping analyses. *Nat. Genet.* **53**, 1260–1269 (2021).
51. Skol, A. D., Scott, L. J., Abecasis, G. R. & Boehnke, M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat. Genet.* **38**, 209–213 (2006).
52. Sham, P. C. & Purcell, S. M. Statistical power and significance testing in large-scale genetic studies. *Nat. Rev. Genet.* **15**, 335–346 (2014).
53. Sherman, R. M. *et al.* Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat. Genet.* **51**, 30–35 (2019).
54. Li, B. *et al.* Incorporating local ancestry improves identification of ancestry-associated methylation signatures and meQTLs in African Americans. *Commun. Biol.* **5**, 401 (2022).

55. Arvanitis, M., Tayeb, K., Strober, B. J. & Battle, A. Redefining tissue specificity of genetic regulation of gene expression in the presence of allelic heterogeneity. *Am. J. Hum. Genet.* **109**, 223–239 (2022).

56. Urbut, S. M., Wang, G., Carbonetto, P. & Stephens, M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat. Genet.* **51**, 187–195 (2019).

57. Kanai, M. *et al.* Insights from complex trait fine-mapping across diverse populations. Preprint at https://doi.org/10.1101/2021.09.03.21262975 (2021).

58. Yao, D. W., O'Connor, L. J., Price, A. L. & Gusev, A. Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nat. Genet.* **52**, 626–633 (2020).

59. Regier, A. A. *et al.* Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. *Nat. Commun.* **9**, 4038 (2018).

60. Moore, Ryan T. & Keith Schnakenberg. blockTools: Blocking, Assignment, and Diagnosing Interference in Randomized Experiments. (2023).

61. Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).

62. Lin, M. F. *et al. GLnexus: Joint Variant Calling for Large Cohort Sequencing*. http://biorxiv.org/lookup/doi/10.1101/343970 (2018) doi:10.1101/343970.

63. Yun, T. *et al.* Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *Bioinformatics* **36**, 5582–5589 (2021).

64. Iwaki, H. *et al.* Accelerating Medicines Partnership: Parkinson's Disease. Genetic Resource. *Mov. Disord.* **36**, 1795–1804 (2021).

65. Freyman, W. A. *et al.* Fast and Robust Identity-by-Descent Inference with the Templated Positional Burrows–Wheeler Transform. *Mol. Biol. Evol.* **38**, 2131–2151 (2021).

66. Cann, H. M. *et al.* A Human Genome Diversity Cell Line Panel. *Science* **296**, 261–262

(2002).

67. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
68. Andrews, Simon. FastQC: A quality control tool for high throughput sequence data. (2019).
69. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
70. Jun, G. *et al.* Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data. *Am. J. Hum. Genet.* **91**, 839–848 (2012).
71. Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184–2185 (2012).
72. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
73. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).
74. Frankish, A. *et al.* GENCODE 2021. *Nucleic Acids Res.* **49**, D916–D923 (2021).
75. 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).
76. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* **40**, 4288–4297 (2012).
77. Zhou, H. J., Li, L., Li, Y., Li, W. & Li, J. J. PCA outperforms popular hidden variable inference methods for molecular QTL mapping. *Genome Biol.* **23**, 210 (2022).
78. Schroeder, A. *et al.* The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Mol. Biol.* **7**, 3 (2006).
79. Wang, W. *et al.* Prospective analysis of incident disease among individuals of diverse

ancestries using genetic and conventional risk factors. Preprint at https://doi.org/10.1101/2023.10.23.23297414 (2023).

80. Wakefield, J. Bayes factors for genome-wide association studies: comparison with P-values. *Genet. Epidemiol.* **33**, 79–86 (2009).

81. Zou, Y., Carbonetto, P., Wang, G. & Stephens, M. Fine-mapping from summary data with the "Sum of Single Effects" model. *PLOS Genet.* **18**, e1010299 (2022).

The genetic architecture of gene expression in individuals of African and European ancestry

Supplementary Figures

Supplementary Fig. 1 | U.S. states / countries of birth for the Black Representation in Genomics Research cohort. Geographic representation of the BRGR cohort based on self-reported **(a)** birth locations (number of participants) and **(b)** birthplace of each grandparent (number of participants' grandparents).

Supplementary Fig. 2 | Summary of identical-by-descent (IBD) sharing in the BRGR and PPMI cohorts. **(a)** ForceAtlas2 network plot of individuals of African, European, and American descent using the total amount of DNA that is identical-by-descent between individuals. PPMI AJ individuals are those from the PPMI dataset that have $\geq$ 25% Ashkenazi Jewish ancestry, whereas PPMI NOAJ constitutes individuals with < 25 % Ashkenazi ancestry. **(b)** Difference in the proportion of genome that is identical by descent between BRGR, PPMI, and PPMI subset by individuals of Ashkenazi descent (PPMI AJ; $\geq$ 25% Ashkenazi ancestry) and individuals with < 25% Ashkenazi ancestry (PPMI NO AJ).

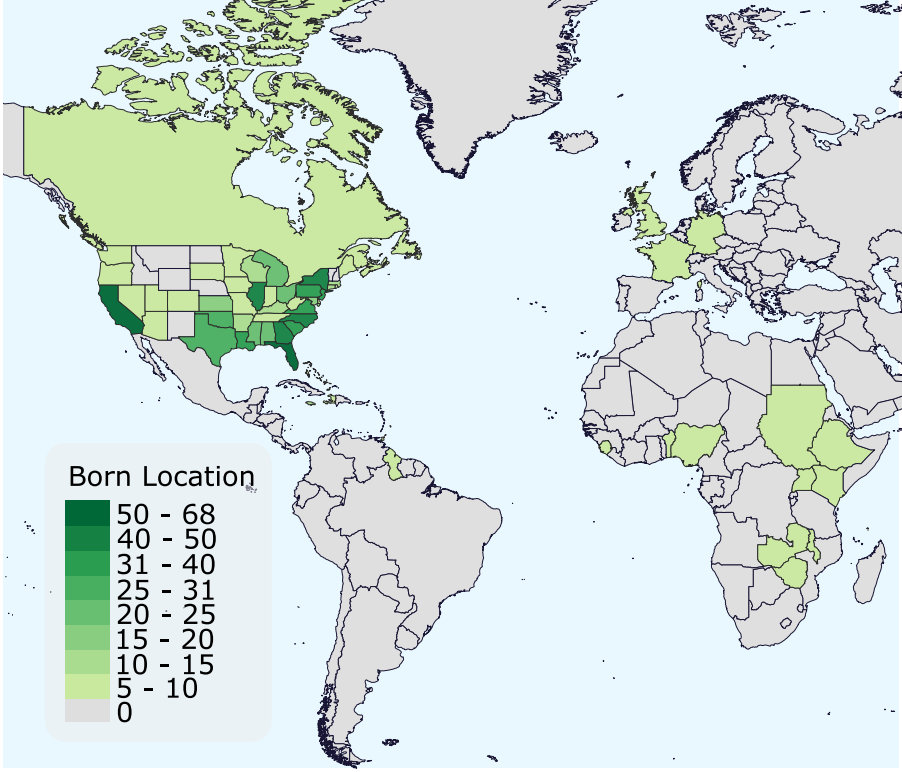
Supplementary Fig. 3 | Distributions of credible set sizes in the four eQTL datasets and intersection of datasets.

Supplementary Fig. 4 | Explaining population differences in heritability using the differences in genetic diversity. Among the 183 genes whose credible sets are colocalized between BRGR and PPMI, we define the genetic diversity in three ways. Firstly, we assume that the causal eQTLs come from index variants of BRGR credible sets. Then the genetic diversity is calculated from the cohort-specific MAF of BRGR index variants (colored in blue). Similarly, we can define the genetic diversity based on the cohort-specific MAF of PPMI index variants (colored in yellow). Lastly, without determining the causal variants, we define the genetic diversity based on the index variants of the credible sets identified in the corresponding cohort (colored in black). **(a)** For each of the 183 genes, differences in genetic diversity $\Delta v_g = v_{g,\,BRGR} - v_{g,\,PPMI}$ are shown on x-axis and differences in heritability $\Delta h^2 = h^2_{BRGR} - h^2_{PPMI}$ are shown on y-axis. The line is drawn from the linear fit $\Delta h^2 \sim 1 + \Delta v_g$. **(b)** For each of the 183 genes, differences in logarithm of genetic diversity between BRGR and PPMI ($\Delta\log(v_g) = \log(v_{g,\,BRGR}) - \log(v_{g,\,PPMI})$) are shown on x-axis and differences in logarithm of transformed heritability ($\Delta\log(q) = \log(q_{BRGR}) - \log(q_{PPMI})$) are shown on y-axis. The line is drawn from the linear fit $\Delta\log(q) \sim 1 + \Delta\log(v_g)$. **(c)** For each of the 183 genes, the residuals of $\log(q_{BRGR})$ and $\Delta\log(v_g)$ after regressing out $\log(q_{PPMI})$ (intercept is also regressed out) are shown on y-axis and x-axis. The line is drawn from the linear fit residual($\log(q_{BRGR})$) ~ residual($\Delta\log(v_g)$) whose slope is equal to the coefficient of $\Delta\log(v_g)$ in the linear fit $\log(q_{BRGR}) \sim 1 + \log(q_{PPMI}) + \Delta\log(v_g)$.

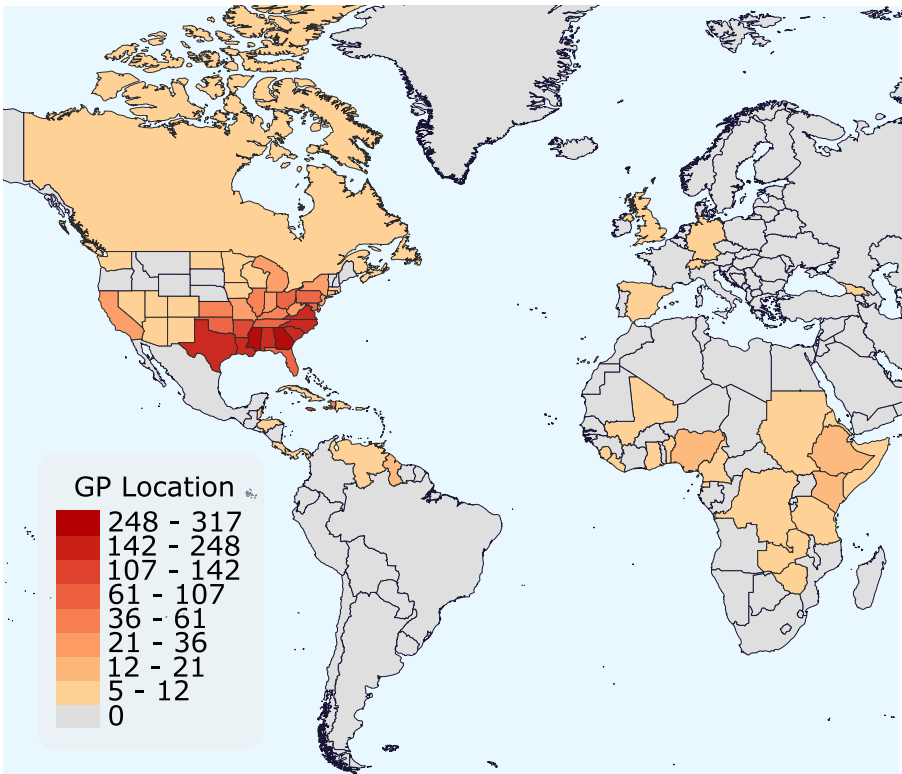
Supplementary Fig. 5 | Local ancestry based cis-eQTL mapping using Tractor for 737 samples with admixed African ancestry from BRGR. Comparisons of genome-wide significant ($p < 5\times10^{-8}$) cis-eQTLs

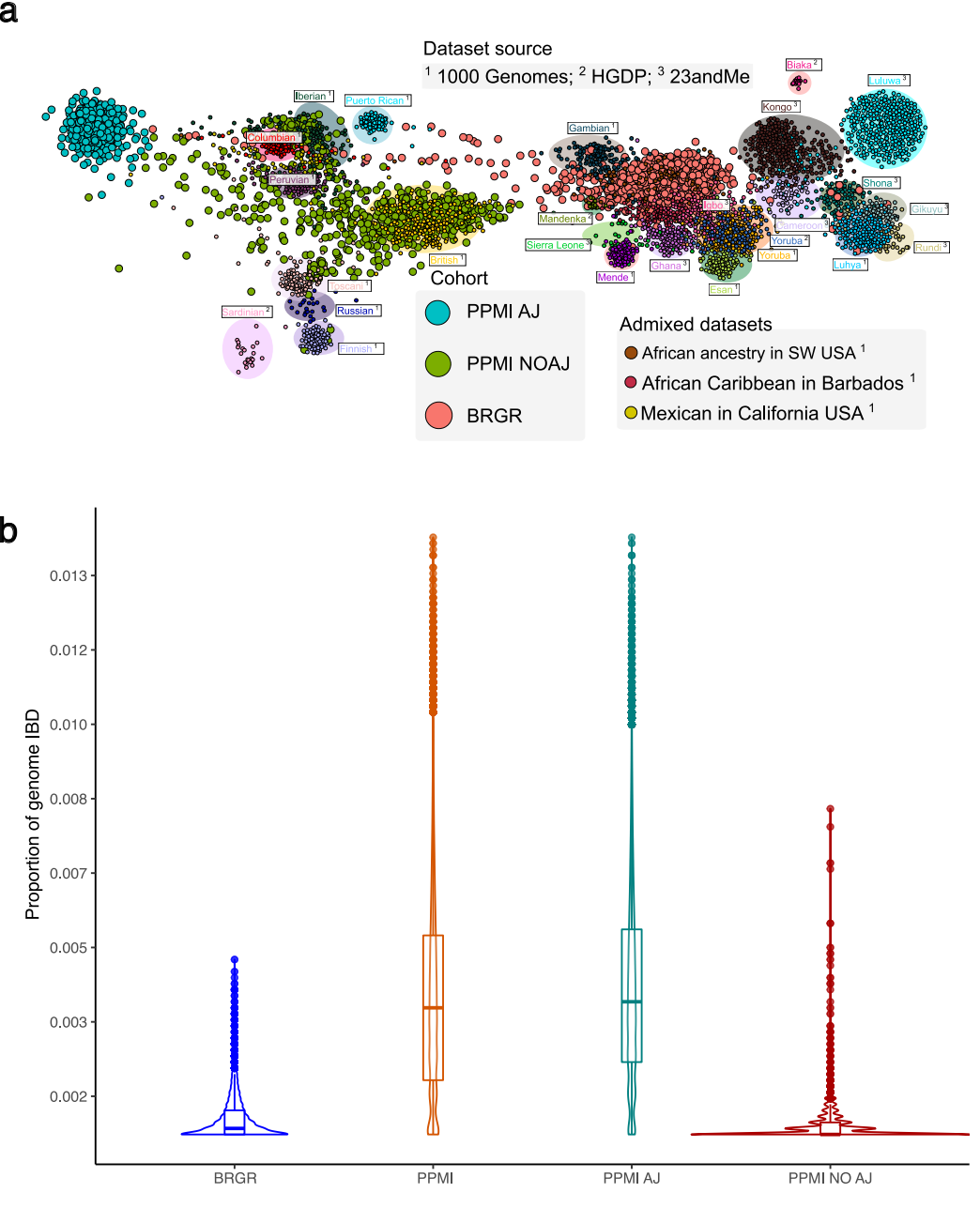
identified using a local-ancestry model, Tractor, or standard generalized linear model. **(a)** Shows the number of eGenes shared by or specific to each model. The majority of eGenes were identified in both models ("Same eGene, same eVar '' or "Same eGene, different eVar"). **(b)** For significant cis-eQTLs across both models, we compared p-values computed for each model. Ancestry-specific effects, estimated by Tractor, were also tested for significant differences. P-values were largely concordant between models for all cis-eQTLs; however, the p-value estimated by Tractor was slightly smaller for a subset of cis-eQTLs with heterogeneous effects between ancestries.

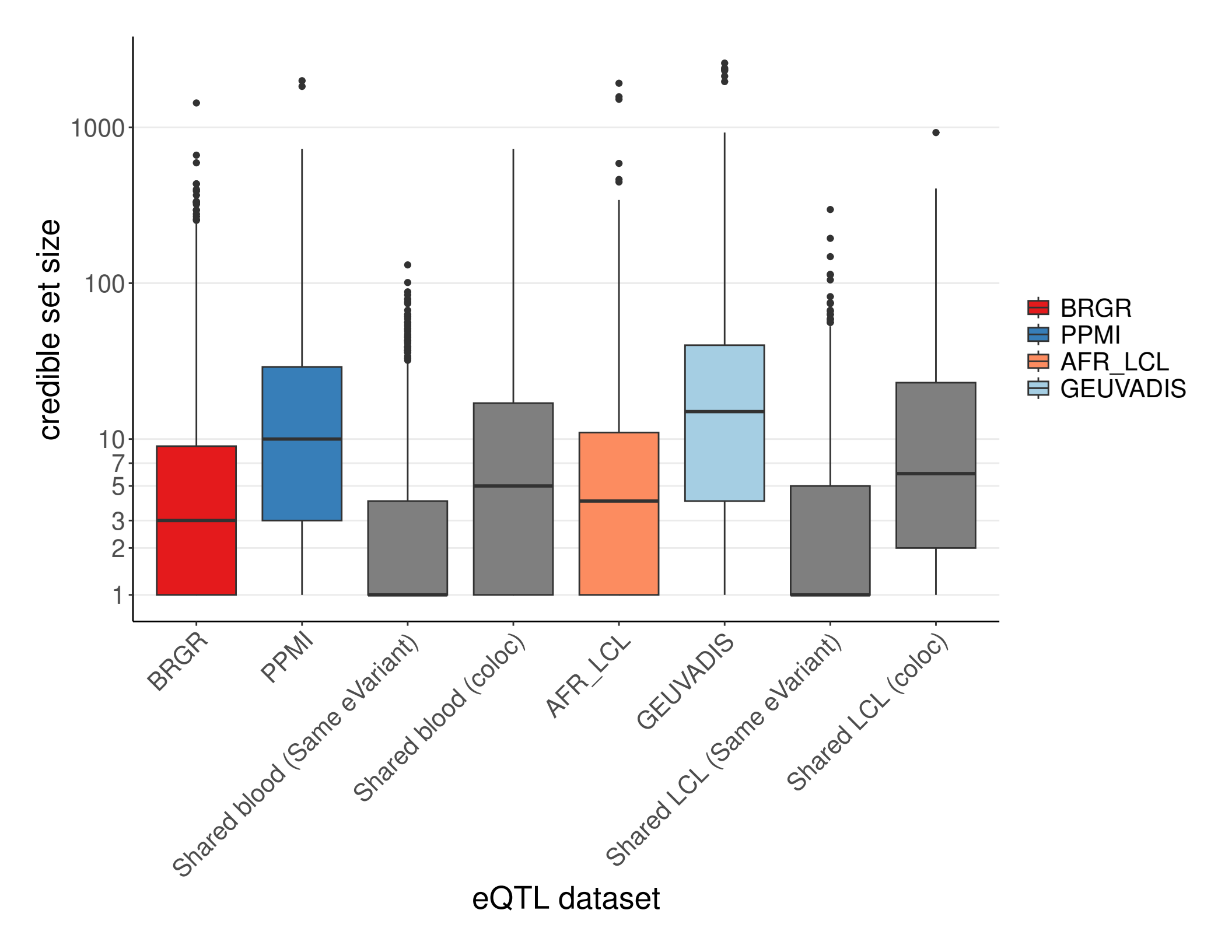
a

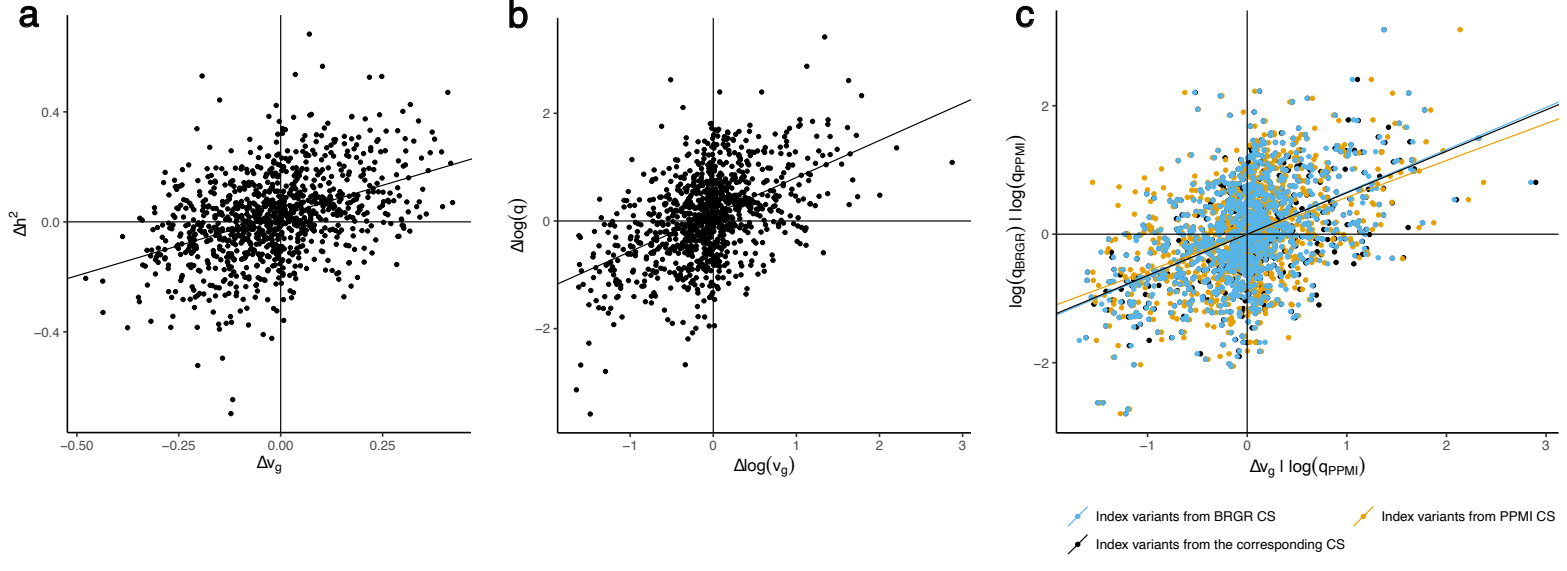

Born Location
50 - 68
40 - 50
31 - 40
25 - 31
20 - 25
15 - 20
10 - 15
5 - 10
0

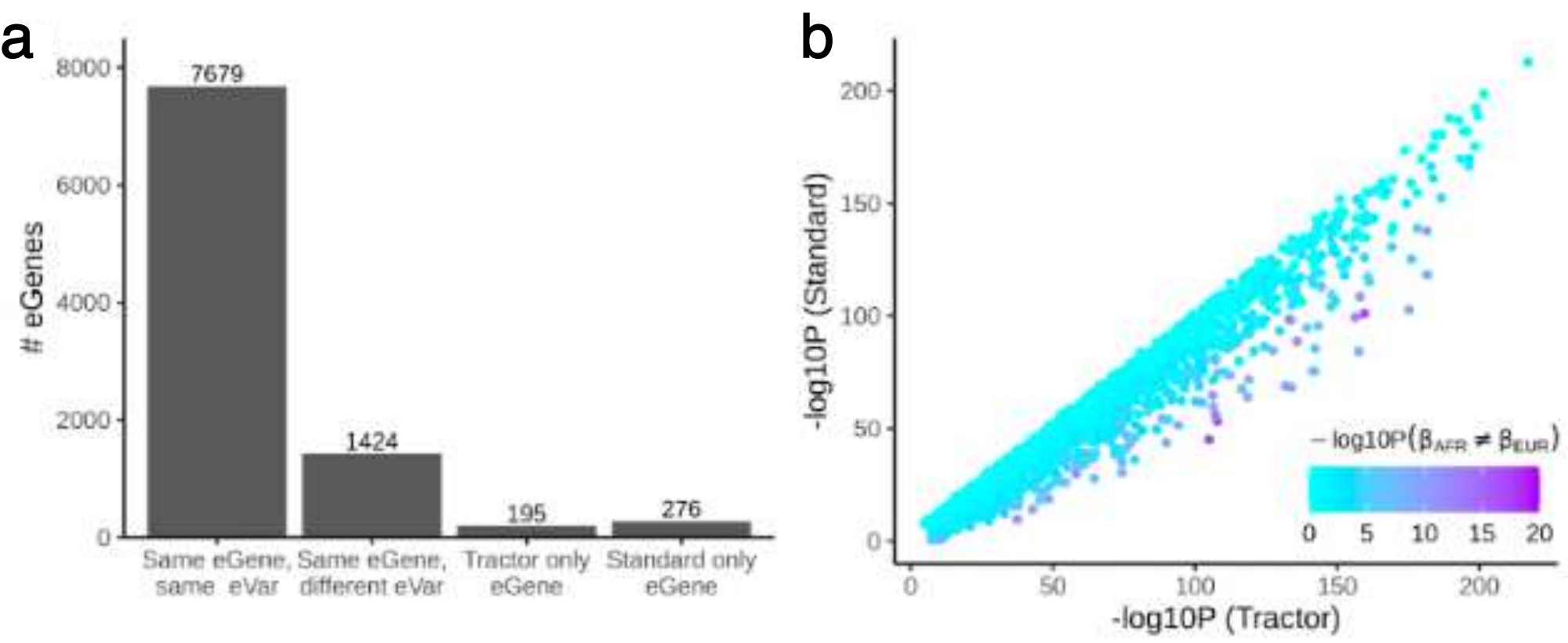
b


GP Location
248 - 317
142 - 248
107 - 142
61 - 107
36 - 61
21 - 36
12 - 21
5 - 12
0


a
Dataset source
[1] 1000 Genomes; [2] HGDP; [3] 23andMe
Iberian [1]
Puerto Rican [1]
Columbian [1]
Peruvian [1]
Gambian [1]
Mandenka [2]
Sierra Leone [3]
Mende [1]
Ghana [3]
Esan [1]
Yoruba [2]
Yoruba [1]
Biaka [2]
Kongo [3]
Luluwa [3]
Shona [3]
Kikuyu [3]
Rundi [3]
Luhya [1]
Tubu [2]
Cameroon [2]
British [1]
Tuscan [1]
Sardinian [2]
Russian [1]
Finnish [1]
Cohort
PPMI AJ
PPMI NOAJ
BRGR
Admixed datasets
African ancestry in SW USA [1]
African Caribbean in Barbados [1]
Mexican in California USA [1]
b
Proportion of genome IBD
0.013
0.012
0.010
0.008
0.007
0.005
0.003
0.002
BRGR
PPMI
PPMI AJ
PPMI NO AJ


credible set size
1000
100
10
7
5
3
2
1
BRGR
PPMI
Shared blood (Same eVariant)
Shared blood (coloc)
AFR_LCL
GEUVADIS
Shared LCL (Same eVariant)
Shared LCL (coloc)
eQTL dataset
BRGR
PPMI
AFR_LCL
GEUVADIS


a
b
c
Δh^2
Δv_g
$\Delta \log(q)$
$\Delta \log(v_g)$
$\log(q_{BRGR}) \mid \log(q_{PPMI})$
$\Delta v_g \mid \log(q_{PPMI})$
Index variants from BRGR CS
Index variants from PPMI CS
Index variants from the corresponding CS


a
8000
6000
4000
2000
0
eGenes
7679
1424
195
276
Same eGene, same eVar
Same eGene, different eVar
Tractor only eGene
Standard only eGene
b
200
150
100
50
0
-log10P (Standard)
0
50
100
150
200
-log10P (Tractor)
− log10P(β_AFR ≠ β_EUR)
0
5
10
15
20

Supplementary Tables

Supplementary Table 1 | Summary of the Black Representation in Genomic Research (BRGR) cohort. The 737 donors that had both RNA-seq and WGS samples were included in the BRGR eQTL study. Location of sample collection was reported in terms of census regions as defined by the U.S. Census Bureau (https://www.census.gov/programs-surveys/economic-census/guidance-geographies/levels.html).

BRGR Descriptor	All donors	eQTL study
Total	1,012	737
Only WGS sample	239	0
Only RNA-seq sample	36	0
Genetically female	698 (69.0%)	511 (69.3%)
Genetically male	314 (31.0%)	226 (30.7%)
Location: Midwest Census Region	145	111
Location: Northeast Census Region	177	128
Location: South Census Region	401	298
Location: West Census Region	195	136
Median age	37	38

Supplementary Table 2 | Mean genome-wide Ancestry Composition (± standard deviation) of represented broad and local populations in each cohort. *23andMe African-American* reference dataset consists of 203,937 research participants in the 23andMe database that self-identify as African-American and have ≥ 50% African ancestry. A Bonferroni corrected p-value < 0.012 indicates that the distribution of ancestry proportions between *23andMe African-American* and BRGR are indistinguishable based on a 1000-iteration randomization test.

Ancestral population	23andMe African-American	BRGR	LCL_AFR	PPMI EUR	GEUVADIS EUR	P-value
Continental						
African	0.807 (± 0.105)	0.802 (± 0.107)	0.96 (± 0.09)	0.001 (± 0.005)	0.001 (± 0.001)	**0.009**
European	0.166 (± 0.097)	0.17 (± 0.101)	0.03 (± 0.07)	0.97 (± 0.07)	0.99 (± 0.002)	**0.004**
Local						
Nigerian	0.296 (± 0.107)	0.29 (± 0.097)	0.41 (± 0.43)	0	0	**0.007**
Sengembian & Guinean	0.061 (± 0.044)	0.06 (± 0.03)	0.18 (± 0.37)	0	0	**0.001**
Coastal West African	0.198 (± 0.084)	0.201 (± 0.076)	0.198 (± 0.36)	0	0	0.019
Congolese Bantu	0.083 (± 0.059)	0.086 (± 0.051)	0.16 (± 0.35)	0	0	0.082
East Bantu	0.012 (± 0.045)	0.014 (± 0.055)	0.15 (± 0.36)	0	0	**0.01**
Somali	0.004 (± 0.062)	0.002 (± 0.038)	0	0	0	0.013
Sudanese	0.004 (± 0.046)	0.007 (± 0.068)	0	0	0	0.014
Hunter-Gatherer	0.002 (± 0.003)	0.002 (± 0.003)	0.002 (± 0.002)	0	0	0
Native American	0.008 (± 0.015)	0.008 (± 0.014)	0.003 (± 0.034)	0.001 (± 0.007)	0	**0.002**
British & Irish	0.073 (± 0.049)	0.074 (± 0.049)	0.023 (± 0.05)	0.22 (± 0.034)	0.39 (± 0.44)	**0.005**
French & German	0.012 (± 0.02)	0.014 (± 0.024)	0.004 (± 0.015)	0.18 (± 0.29)	0.05 (± 0.15)	0.11
Iberian	0.008 (± 0.02)	0.008 (± 0.018)	0.001 (± 0.004)	0.07 (± 0.025)	0	**0.001**
Italian	0.001 (± 0.005)	0.001 (± 0.007)	0.002 (± 0.001)	0.04 (± 0.16)	0.26 (± 0.15)	**0.001**
Finnish	0	0	0	0.001 (± 0.02)	0.24 (± 0.43)	**0**
East European	0	0	0.002 (± 0.001)	0.05 (± 0.16)	0.002 (± 0.04)	**0**
Scandinavian	0.004 (± 0.006)	0.004 (± 0.006)	0.006 (± 0.002)	0.011 (± 0.78)	0.03 (± 0.11)	**0**
Ashkenazi Jewish	0.001 (± 0.011)	0.002 (± 0.015)	0	0.36 (± 0.48)	0.001 (± 0.01)	0.024

Supplementary Table 3 | Summary of the 1000 Genomes Project samples included in the African ancestry superpopulation LCL eQTL cohort, and in the European component of GEUVADIS.

Superpopulation	Code	Population	Females	Males	Both sexes
AFR	ACB	African Caribbeans in Barbados	49	47	96
AFR	ASW	Americans of African Ancestry in SW USA	35	26	61
AFR	ESN	Esan in Nigeria	46	53	99
AFR	GWD	Gambian in Western Divisions in the Gambia	58	53	111
AFR	LWK	Luhya in Webuye, Kenya	55	44	99
AFR	MSL	Mende in Sierra Leone	43	42	85
AFR	YRI	Yoruba in Ibadan, Nigeria	56	52	108
AFR	***Total***		***342 (51.8%)***	***318 (48.2%)***	***659***
EUR	CEU	CEPH	44	45	89
EUR	FIN	Finns	56	36	92
EUR	GBR	British	43	43	86
EUR	TSI	Toscani	44	47	91
EUR	***Total***		***187 (52.2%)***	***171 (47.8%)***	***358***

Supplementary Table 4 | Coefficients of the regression analysis in testing the relation between differences in heritability and differences in genetic diversity. Assuming no differential causal effect and environmental noise, we derived a relation describing how genetic diversity differences result in differences in heritability: $v_{g,\,BRGR} / v_{g,\,PPMI} \approx q_{BRGR} / q_{PPMI}$ or equivalently $\Delta\log(v_g) - \Delta\log(q) \approx 0$. This relation is examined in various regression analysis using the 183 genes whose credible sets are colocalized between BRGR and PPMI. The following linear regression analysis were performed: 1) $\Delta h^2 \sim 1 + \Delta v_g$; 2) $\Delta\log(q) \sim 1 + \Delta\log(v_g)$; 3) $\log(q_{BRGR}) \sim 1 + \log(q_{PPMI}) + \Delta\log(v_g)$. The coefficients, standard errors, and p-values of these regressions are shown.

Model	variable	Coefficient	SE	P-value	Model formula
asis	(Intercept)	0.12573	0.00962	6.25E-28	diff_h2 ~ 1 + diff_vg
asis	diff_vg	0.32275	0.06208	5.38E-07	diff_h2 ~ 1 + diff_vg
BRGR	(Intercept)	0.12086	0.00950	6.42E-27	diff_h2 ~ 1 + diff_vg
BRGR	diff_vg	0.36830	0.06177	1.27E-08	diff_h2 ~ 1 + diff_vg
PPMI	(Intercept)	0.12726	0.00953	9.25E-29	diff_h2 ~ 1 + diff_vg
PPMI	diff_vg	0.35329	0.06400	1.16E-07	diff_h2 ~ 1 + diff_vg
asis	(Intercept)	-0.01985	0.10985	0.857	logq_BRGR ~ 1 + logq_PPMI + diff_logvg
asis	logq_PPMI	0.53315	0.05462	2.47E-18	logq_BRGR ~ 1 + logq_PPMI + diff_logvg
asis	diff_logvg	0.33927	0.07286	6.23E-06	logq_BRGR ~ 1 + logq_PPMI + diff_logvg
BRGR	(Intercept)	0.02254	0.10892	0.836	logq_BRGR ~ 1 + logq_PPMI + diff_logvg
BRGR	logq_PPMI	0.57524	0.05531	3.88E-20	logq_BRGR ~ 1 + logq_PPMI + diff_logvg
BRGR	diff_logvg	0.38416	0.07215	3.00E-07	logq_BRGR ~ 1 + logq_PPMI + diff_logvg
PPMI	(Intercept)	0.01767	0.11045	0.873	logq_BRGR ~ 1 + logq_PPMI + diff_logvg
PPMI	logq_PPMI	0.55065	0.05512	5.63E-19	logq_BRGR ~ 1 + logq_PPMI + diff_logvg
PPMI	diff_logvg	0.36581	0.07634	3.44E-06	logq_BRGR ~ 1 + logq_PPMI + diff_logvg
asis	(Intercept)	0.81573	0.05925	5.70E-30	diff_logq ~ 1 + diff_logvg
asis	diff_logvg	0.44315	0.08495	4.95E-07	diff_logq ~ 1 + diff_logvg
BRGR	(Intercept)	0.76469	0.05773	1.93E-28	diff_logq ~ 1 + diff_logvg
BRGR	diff_logvg	0.54317	0.07942	1.18E-10	diff_logq ~ 1 + diff_logvg
PPMI	(Intercept)	0.82142	0.05810	4.64E-31	diff_logq ~ 1 + diff_logvg
PPMI	diff_logvg	0.50699	0.08675	2.33E-08	diff_logq ~ 1 + diff_logvg

Supplementary Table 5 | Summary of African-American GWAS used and detailed summary of variant-to-gene mapping. MVP - the VA Million Veteran Program; BCX - Blood Cell Consortium.

Source	Trait	dbGaP analysis accession	Cases	Controls	GWAS hits	N coloc ($H_4 \geq .8$)				N coloc ($H_4 \geq .5$)			
						BRGR	PPMI	LCL_AFR	GEUVADIS	BRGR	PPMI	LCL_AFR	GEUVADIS
MVP	blood lipids - HDL	pha004827.1	57,332		63	2	1	3	-	4	1	3	-
MVP	blood lipids - LDL	pha004830.1			53	1	-	-	-	1	-	-	-
MVP	blood lipids - total cholesterol (TC)	pha004833.1			77	1	-	-	-	2	-	-	-
MVP	blood lipids - triglycerides (TG)	pha004836.1			47	-	1	1	-	-	1	1	1
MVP	venous thromboembolism (VTE)	pha004962.1	2,261	49,400	16	1	-	1	-	1	-	-	-
MVP	type 2 diabetes (T2D)	pha004943.1	24,646	31,446	56	1	-	-	-	1	-	-	-
MVP	**total across all traits**				**312**	**6**	**2**	**4**	**0**	**9**	**2**	**4**	**1**
BCX	red blood cell count (RBC count)		Up to 15,171. Varies variant by variant depending on how many studies were included in a meta-analysis.		32	2	-	-	-	2	1	1	-
BCX	hemoglobin concentration (HGB)				22	-	-	1	-	-	-	1	-
BCX	hematocrit (HCT)				17	-	-	-	-	-	-	-	-
BCX	mean corpuscular hemoglobin (MCH)				19	-	-	-	1	-	-	-	1
BCX	mean corpuscular volume (MCV)				21	2	1	1	1	2	2	1	1
BCX	mean corpuscular hemoglobin concentration (MCHC)				39	3	-	1	-	7	1	4	-
BCX	RBC distribution width (RDW)				27	2	-	-	-	2	-	-	-
BCX	total white blood cell count (WBC count)				46	5	-	-	-	6	-	-	-
BCX	neutrophil count (Neutro)				37	4	1	-	-	4	1	-	-
BCX	lymphocyte count (Lympho)				29	1	-	2	-	1	-	2	-
BCX	monocyte count (Mono)				33	3	2	-	-	3	2	1	-
BCX	basophil count (Baso)				35	-	-	-	-	-	-	-	-
BCX	eosinophil count (Eosin)				28	2	2	2	3	2	3	2	3
BCX	platelet count (PLT count)				38	4	2	3	2	7	3	4	2
BCX	mean platelet volume (MPV)				38	4	3	3	1	5	3	3	2
BCX	**total across all traits**				**461**	**32**	**11**	**13**	**8**	**41**	**16**	**19**	**9**
23andMe	**height**	**NA**	**273,215**		**803**	**37**	**28**	**36**	**13**	**69**	**53**	**64**	**24**

Supplementary Table 6 | Variant-to-gene hypothesis derived from eQTLs unique to African ancestry cohorts. MVP - the VA Million Veteran Program; BCX - Blood Cell Consortium.

Supplementary Table 7 | Number of relationships based on the proportion of the genome that is identical-by-descent (IBD) between all pairwise combinations of individuals in each cohort.

	N pairwise relationships			
IBD Range (proportion genome)	**PPMI**	**PPMI (with AJ)**	**BRGR**	**Typical relationship(s) in range**
< 0.001	575,626	118,123	1,015,119	Unrelated
0.001-0.005	4,841	84,903	8,959	4th, 5th and distant cousins
0.005-0.01	92	28,924	27	3rd cousins
0.01-0.025	17	770	11	Half 2nd cousins
0.025-0.05	13	5	5	2nd cousins
0.05-0.1	5	5	5	Great grandparents
0.1-0.2	12	5	5	1st cousins
0.2-0.3	6	12	5	Grandparents, avuncular
0.3-1	32	59	8	Full siblings, parents, twins