


RESEARCH

Open Access



Impact of admixture and ancestry on eQTL analysis and GWAS colocalization in GTEx

Nicole R. Gay¹, Michael Gloude-mans², Margaret L. Antonio², Nathan S. Abell¹, Brunilda Balliu³, YoSon Park^{4,5}, Alicia R. Martin^{6,7}, Shaila Musharoff¹, Abhiram S. Rao⁸, François Aguet⁹, Alvaro N. Barbeira¹⁰, Rodrigo Bonazzola¹⁰, Farhad Hormozdiani^{9,11}, GTEx Consortium, Kristin G. Ardlie⁹, Christopher D. Brown⁴, Hae Kyung Im¹⁰, Tuuli Lappalainen^{12,13}, Xiaoquan Wen¹⁴ and Stephen B. Montgomery^{1,15*} 

* Correspondence: smontgom@stanford.edu

¹Department of Genetics, Stanford University, Stanford, CA, USA

¹⁵Department of Pathology, Stanford University, Stanford, CA, USA

Full list of author information is available at the end of the article

Abstract

Background: Population structure among study subjects may confound genetic association studies, and lack of proper correction can lead to spurious findings. The Genotype-Tissue Expression (GTEx) project largely contains individuals of European ancestry, but the v8 release also includes up to 15% of individuals of non-European ancestry. Assessing ancestry-based adjustments in GTEx improves portability of this research across populations and further characterizes the impact of population structure on GWAS colocalization.

Results: Here, we identify a subset of 117 individuals in GTEx (v8) with a high degree of population admixture and estimate genome-wide local ancestry. We perform genome-wide *cis*-eQTL mapping using admixed samples in seven tissues, adjusted by either global or local ancestry. Consistent with previous work, we observe improved power with local ancestry adjustment. At loci where the two adjustments produce different lead variants, we observe 31 loci (0.02%) where a significant colocalization is called only with one eQTL ancestry adjustment method. Notably, both adjustments produce similar numbers of significant colocalizations within each of two different colocalization methods, COLOC and FINEMAP. Finally, we identify a small subset of eQTL-associated variants highly correlated with local ancestry, providing a resource to enhance functional follow-up.

Conclusions: We provide a local ancestry map for admixed individuals in the GTEx v8 release and describe the impact of ancestry and admixture on gene expression, eQTLs, and GWAS colocalization. While the majority of the results are concordant between local and global ancestry-based adjustments, we identify distinct advantages and disadvantages to each approach.

Keywords: Local ancestry, Population structure, Admixture, eQTL, Colocalization, GTEx, Gene expression



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Introduction

Thousands of genome-wide association studies (GWAS) have been published to date. Subsequently, large-scale expression quantitative trait loci (eQTL) datasets are studied to provide insights for genetic variants associated with complex traits. While the majority of such studies focus on single-ancestry populations or relatively homogeneous populations, the latest Genotype-Tissue Expression (GTEx) project (v8) includes up to 17% of individuals with non-European or admixed ancestry [1]. Genetic studies with individuals of admixed ancestries may suffer from additional challenges due to complex population substructure [2, 3]. Such substructure can confound genetic associations, and insufficient control may increase spurious findings [4, 5].

Global ancestry (GA), or the proportions of different ancestral populations represented across the entire genome, is routinely used to adjust for population structure in genetic association studies [6]. This approach has the advantage of averaging genomic background effects and was used in eQTL mapping for the main GTEx releases [1, 7]. The potential disadvantage of correcting only for GA is that it does not precisely account for ancestry at any specific locus. This can be problematic when genes are differentially expressed in ancestral populations of admixed individuals. In contrast, local ancestry (LA), or the number of alleles derived from distinct ancestral populations at a given locus, may be more appropriate for population structure adjustment in admixed populations but typically suffers from much longer compute time and can be prone to errors in estimation at a variant level [5, 8–12].

LA adjustment in genetic association studies has been shown to reduce type I error rate (false positives) [13–15] and sufficiently control for population stratification [13, 15]. However, the power of adjusting for LA is highly dependent on the underlying genetic architecture of the admixed population [8, 12, 15–17]; some have recommended using LA adjustment as a method for follow-up of candidate loci as opposed to a discovery tool for GWAS [8, 14, 18]. Fewer studies have investigated the effect of LA adjustment on eQTL mapping, demonstrating modest improvements in discovery power [5, 10]. Recently, Zhong et al. have demonstrated that the use of LA adjustment, compared to GA adjustment, can improve eQTL mapping while controlling for type I error rate and increasing statistical power [10]. However, the implications of these differences for GWAS colocalization were not assessed.

In this study, we describe the degree of admixture in the GTEx v8 cohort and estimate LA for a subset of 117 individuals with at least 10% admixture from European, African, and East Asian ancestral populations. LA explains at least 7% of the variance in residual expression for 1% of expressed genes ($M = 1159$). We perform *cis*-eQTL mapping in seven tissues and assess the differences between LA adjustment and GA adjustment in the context of this admixed sub-cohort. For the subset of loci where the two ancestry adjustment methods yield different results, we perform GWAS/eQTL colocalization analyses with 142 previously published GWAS, representing a range of traits, consortia, and cohort ancestry. We characterize 31 loci where a significant colocalization is reported only with one eQTL ancestry adjustment method. Finally, we identify a small subset of GTEx eVariants whose genotypes are highly correlated with LA, providing a resource to enhance functional follow-up of these loci.

Results

GTEx includes African and Asian population admixture

The GTEx v8 release includes whole genome sequencing and gene expression data for 838 individuals, including 103 African American and 12 Asian American individuals (self-reported ancestry). Genome-wide genotype-based principal components (gPCs) reflect GA and have been used to adjust for population structure in both GWAS [6, 9, 13] and eQTL studies [7]. Therefore, to understand the degree of population admixture represented in GTEx, we compared the first two gPCs with self-reported ancestry (Fig. 1a). Figure 1a demonstrates that gPC1 and gPC2 reflect African and Asian ancestry, respectively; the majority of European Americans (698 out of 715 individuals)

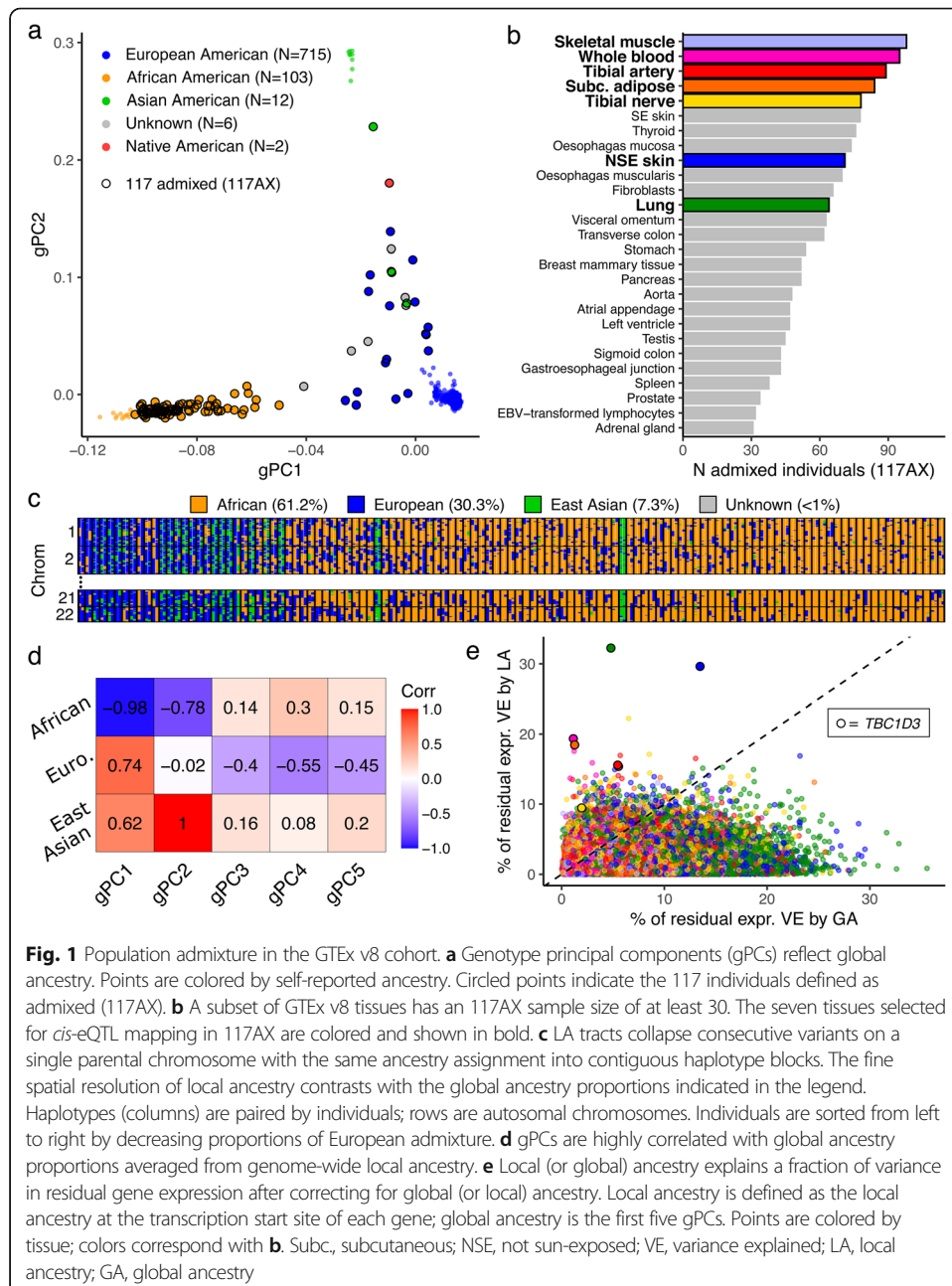


Fig. 1 Population admixture in the GTEx v8 cohort. **a** Genotype principal components (gPCs) reflect global ancestry. Points are colored by self-reported ancestry. Circled points indicate the 117 individuals defined as admixed (117AX). **b** A subset of GTEx v8 tissues has an 117AX sample size of at least 30. The seven tissues selected for *cis*-eQTL mapping in 117AX are colored and shown in bold. **c** LA tracts collapse consecutive variants on a single parental chromosome with the same ancestry assignment into contiguous haplotype blocks. The fine spatial resolution of local ancestry contrasts with the global ancestry proportions indicated in the legend. Haplotypes (columns) are paired by individuals; rows are autosomal chromosomes. Individuals are sorted from left to right by decreasing proportions of European admixture. **d** gPCs are highly correlated with global ancestry proportions averaged from genome-wide local ancestry. **e** Local (or global) ancestry explains a fraction of variance in residual gene expression after correcting for global (or local) ancestry. Local ancestry is defined as the local ancestry at the transcription start site of each gene; global ancestry is the first five gPCs. Points are colored by tissue; colors correspond with **b**. Subc., subcutaneous; NSE, not sun-exposed; VE, variance explained; LA, local ancestry; GA, global ancestry

cluster together near the origin, suggesting that the samples in this cluster are relatively homogeneously European-descendent. These patterns are observed with finer resolution when genotype PCA is performed with combined GTEx and 1000 Genomes data [19] (Additional file 1, Figure S1). A subset of 117 individuals with more than 10% population admixture, referred to as 117AX, was retained for downstream analyses (Fig. 1a; Additional file 2, Table S1).

The 49 tissues used for QTL discovery in the GTEx v8 release have a varying representation of 117AX. Twenty-seven of these tissues have a sample size of at least 30 admixed individuals (Fig. 1b). Sample sizes for all 49 tissues are provided in Figure S2 (Additional file 1). The pituitary and 13 central nervous system tissues have the lowest representation of 117AX relative to total sample sizes per tissue (mean 7%). We selected seven tissues in which to perform *cis*-eQTL calling based on a minimum admixed sample size of 60 [20] and relevance to phenotypes with known population differences (e.g., subcutaneous adipose and body fat distribution [21, 22], $N = 84$; not-sun-exposed (NSE) skin and epidermal gene expression [23], $N = 71$; lung and asthma prevalence [24], $N = 64$; skeletal muscle and lean muscle mass [25], $N = 98$). Whole blood ($N = 95$) and tibial artery ($N = 89$) were also included because they have large 117AX sample sizes.

Using RFMix [26], we performed three-population (European, African, and East Asian) LA estimation on 117AX (see the “Methods” section; Fig. 1c; Additional file 1, Figure S3). We provide these LA calls as a resource for further investigation of GTEx data (Additional file 3, Table S2). For each individual, genome-wide LA was averaged to provide GA estimates. Every sample in 117AX has less than 90% GA from any one ancestral population out of Europe, Africa, and East Asia. We correlated these GA proportions with the first five gPCs, which quantitatively demonstrates the strong relationships between gPC1 and African ancestry ($r = -0.98$) and gPC2 and East Asian ancestry ($r = 1.0$; Fig. 1d).

In order to assess the importance of LA in the context of gene expression, we adapted an existing approach [27] to calculate the proportion of variance explained in 117AX gene expression by LA after accounting for GA and vice versa (see the “Methods” section; Fig. 1e; Additional file 4, Table S3). On average, across genes in our seven tissues of interest, GA explains more variance in gene expression than LA at the transcription start site for each gene (P value $< 2.2 \times 10^{-16}$, two-sided t test). However, LA explains at least 7% of the variance in residual expression for 1% of expressed genes ($M = 1159$). At the extreme, LA explains 32% of the variance in residualized expression of *TBC1 domain family member 3* (*TBC1D3*), a hominoid-specific oncogene [28], in the lung; LA also explains significantly more variance in *TBC1D3* expression than GA in all seven tissues tested (P value = 0.0018, two-sided t test). In a separate study of copy number, *TBC1D3* was among the most variable (median 38.13, variance 93.2 copies among 159 individuals) and population-stratified (mean 29.28, 34.17, and 43.86 copy numbers in European, Asian, and Yoruban samples, respectively) human gene families [29]. Such biological evidence for residual variance in gene expression captured by LA supports the importance of considering LA in the context of eQTL mapping.

Local ancestry adjustment increases power for discovery in *cis*-eQTL mapping

We performed *cis*-eQTL mapping in the admixed population (117AX) to identify associations between variants and gene expression within each of the seven tissues

indicated in Fig. 1b (see the “Methods” section; Additional file 5, Table S4). We implemented linear models to test for an association between each gene-*cis*-variant pair. For each pair, two association tests were performed: the first to adjust for global ancestry (GlobalAA) and the second to adjust for local ancestry (LocalAA). Importantly, LocalAA accounts for the number of European, African, and East Asian alleles for each variant while GlobalAA uses the first five genotype principal components as a proxy for global ancestry, implementing the same ancestry adjustment used in the GTEx eQTL calling pipeline.

A quantile-quantile plot of the nominal P values ($-\log_{10}$) of all association tests in GlobalAA and LocalAA demonstrates that LocalAA has more significant P values (represented in the highest quantiles) relative to GlobalAA for six of the seven tissues, with NSE skin showing more similar P value distributions between the two methods (Fig. 2a). This corroborates previous findings that LA adjustment results in more significant nominal P values than GA adjustment in the context of *cis*-eQTL mapping [10].

We applied a nominal P value cutoff of $1e-6$ to identify significant eQTLs; this threshold closely approximates the threshold required for an eQTL to subsequently pass a false discovery rate cutoff of 5% (Additional file 1, Figure S4). More eGenes are called with LocalAA than GlobalAA in all seven tissues (P value = 0.0078, binomial probability) (Fig. 2b). The majority of the eGenes overlap between the two methods, a

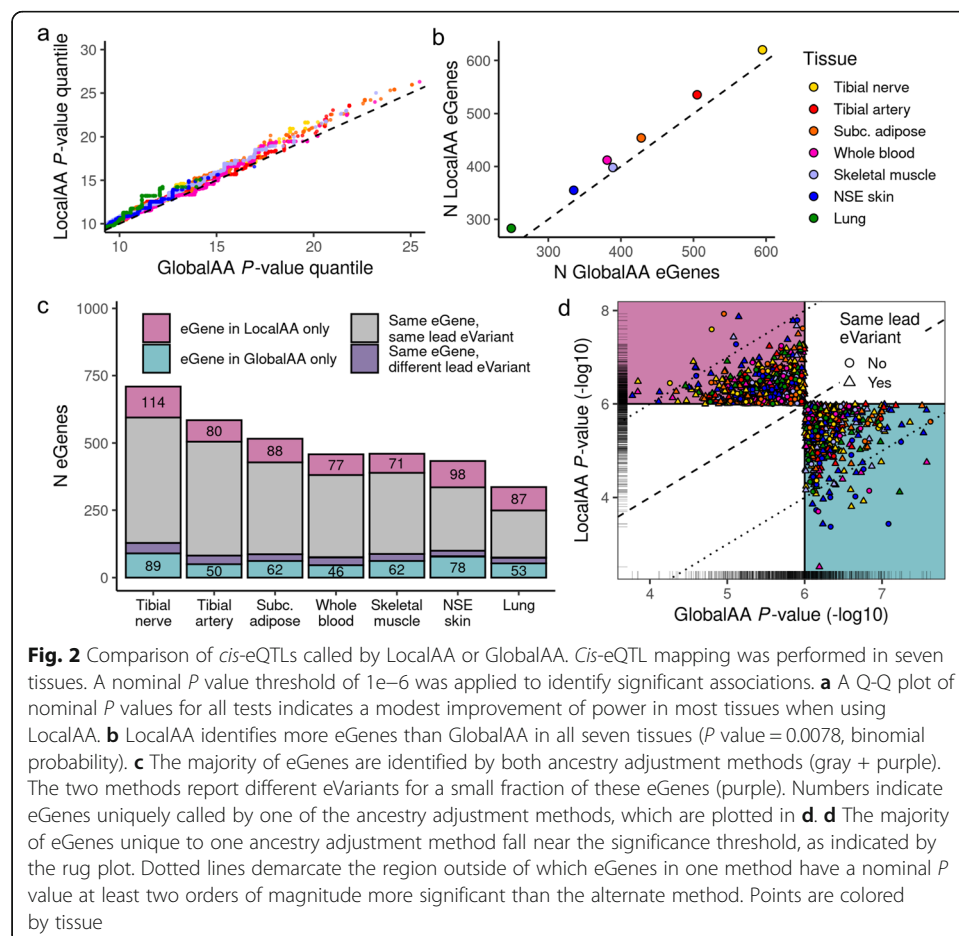


Fig. 2 Comparison of *cis*-eQTLs called by LocalAA or GlobalAA. *Cis*-eQTL mapping was performed in seven tissues. A nominal P value threshold of $1e-6$ was applied to identify significant associations. **a** A Q-Q plot of nominal P values for all tests indicates a modest improvement of power in most tissues when using LocalAA. **b** LocalAA identifies more eGenes than GlobalAA in all seven tissues (P value = 0.0078, binomial probability). **c** The majority of eGenes are identified by both ancestry adjustment methods (gray + purple). The two methods report different eVariants for a small fraction of these eGenes (purple). Numbers indicate eGenes uniquely called by one of the ancestry adjustment methods, which are plotted in **d**. **d** The majority of eGenes unique to one ancestry adjustment method fall near the significance threshold, as indicated by the rug plot. Dotted lines demarcate the region outside of which eGenes in one method have a nominal P value at least two orders of magnitude more significant than the alternate method. Points are colored by tissue

subset of which has different associated lead eVariants between LocalAA and GlobalAA (Fig. 2c). This subset of eGenes provided an opportunity to characterize differences in lead eVariants identified between the two ancestry adjustment methods and was the focus of downstream analyses.

eGenes are considered unique to an ancestry adjustment method if the association reaches significance only with that method (nominal P value cutoff of $1e-6$; 1055 total instances across tissues for 988 unique genes). The majority (65%) of eGenes that are unique to one method replicate at a P value within one order of magnitude of the other method (Fig. 2d). However, 44 of these eGenes only replicate in the other method at a P value more than two orders of magnitude less significant (14 and 30 eGenes unique to LocalAA and GlobalAA, respectively). Twenty of these 44 eGenes are in NSE skin; none is in the tibial artery. Interestingly, for 29 out of these 44 eGenes, despite the large difference in the statistical significance, the lead variants between the two adjustment methods are identical.

Different eQTL ancestry adjustments yield minor differences in GWAS colocalization

Colocalization analyses assess the degree to which independent signals of association, including eQTL and GWAS signals, share the same causal variant. We performed colocalization with two different methods: COLOC [30] and FINEMAP [31]. COLOC estimates the posterior probability that a single variant affects both traits (PP4). FINEMAP estimates the posterior probability of single trait causality for all variants in a region; as previously described, these probabilities can be used to derive a colocalization posterior probability (CLPP) for two independent association signals [32] (see the “Methods” section). Importantly, FINEMAP explicitly accounts for linkage disequilibrium (LD) while COLOC does not; this is particularly relevant given the admixed ancestry of the eQTL cohort.

We selected 142 GWAS to perform colocalization with our eQTLs. Previously, 114 of these GWAS were used to perform colocalization with all GTEx v8 eQTLs [33]. These GWAS were originally chosen to include a broad representation of different trait classes and some replication between GWAS from the UK Biobank (UKB) and other consortia. We included an additional 28 multi-ethnic GWAS from the PAGE study to increase the representation of admixed cohorts in our colocalization analyses [34]. More information about each GWAS is available in Table S5 (Additional file 6).

We performed colocalization between our fourteen sets of eQTL summary statistics (one per ancestry adjustment method per seven tissues) and 142 GWAS. Here, we define a locus as a gene and GWAS trait pair in a specific tissue. For a single locus, two colocalization tests are performed with each colocalization method: one test between the GWAS and each set of eQTL summary statistics (LocalAA or GlobalAA). Therefore, there are up to four colocalization scores (COLOC PP4 or FINEMAP CLPP) for a single locus. For colocalization analyses with COLOC, we restricted tested loci to the subset of eGenes with different lead eVariants between LocalAA and GlobalAA at a relaxed nominal P value threshold (Fig. 3a). We subsequently performed colocalization analyses with FINEMAP for the subset of loci with at least one COLOC colocalization (Fig. 3b). We define evidence for colocalization at a locus as $PP4 > 0.5$ or $CLPP > 0.01$ for COLOC and FINEMAP, respectively.

While GWAS colocalization was only tested at loci for which the two eQTL ancestry adjustment methods yielded different lead eVariants, colocalization probabilities are

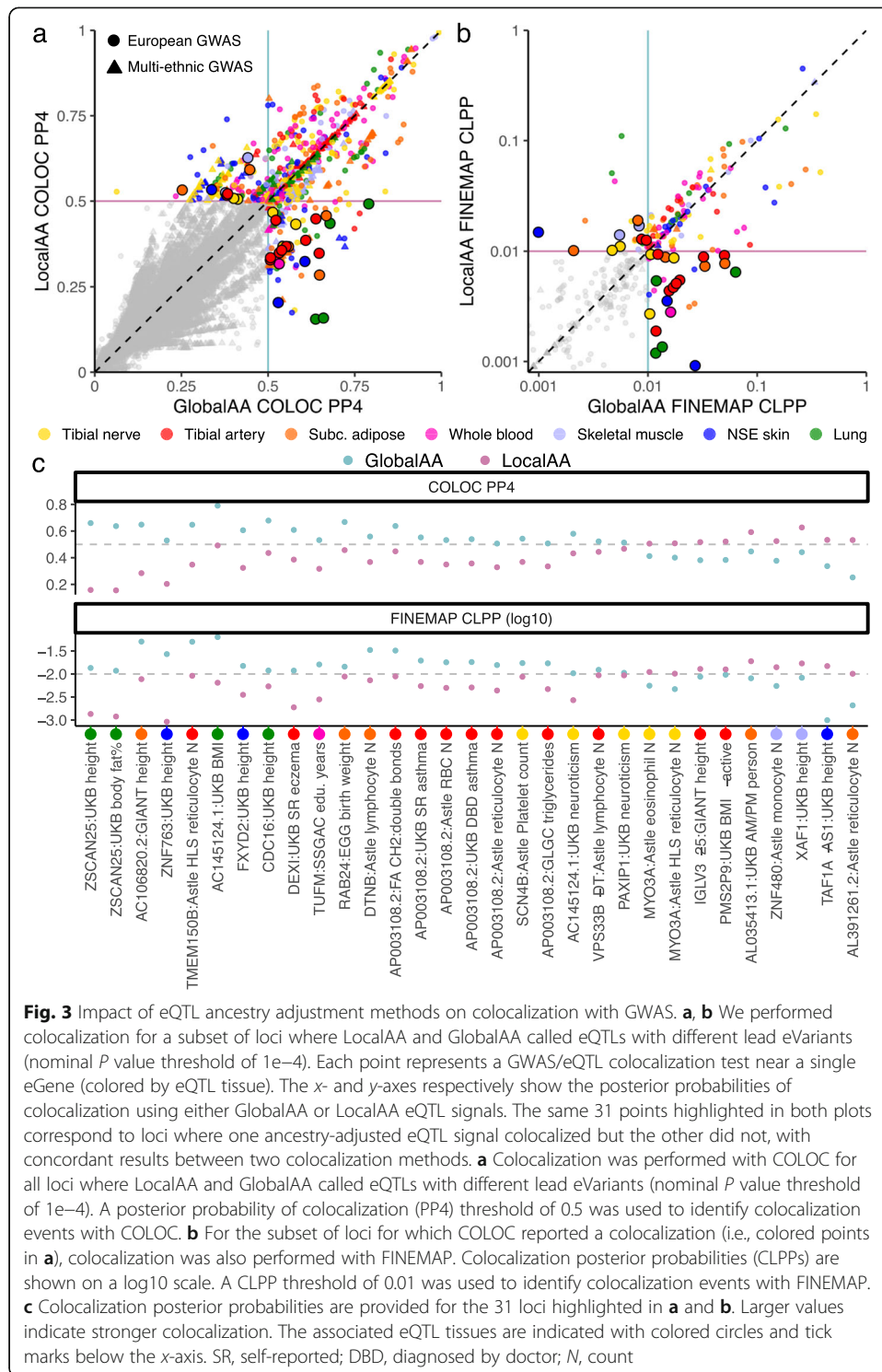


Fig. 3 Impact of eQTL ancestry adjustment methods on colocalization with GWAS. **a, b** We performed colocalization for a subset of loci where LocalAA and GlobalAA called eQTLs with different lead eVariants (nominal P value threshold of $1e-4$). Each point represents a GWAS/eQTL colocalization test near a single eGene (colored by eQTL tissue). The x- and y-axes respectively show the posterior probabilities of colocalization using either GlobalAA or LocalAA eQTL signals. The same 31 points highlighted in both plots correspond to loci where one ancestry-adjusted eQTL signal colocalized but the other did not, with concordant results between two colocalization methods. **a** Colocalization was performed with COLOC for all loci where LocalAA and GlobalAA called eQTLs with different lead eVariants (nominal P value threshold of $1e-4$). A posterior probability of colocalization (PP4) threshold of 0.5 was used to identify colocalization events with COLOC. **b** For the subset of loci for which COLOC reported a colocalization (i.e., colored points in **a**), colocalization was also performed with FINEMAP. Colocalization posterior probabilities (CLPPs) are shown on a log10 scale. A CLPP threshold of 0.01 was used to identify colocalization events with FINEMAP. **c** Colocalization posterior probabilities are provided for the 31 loci highlighted in **a** and **b**. Larger values indicate stronger colocalization. The associated eQTL tissues are indicated with colored circles and tick marks below the x-axis. SR, self-reported; DBD, diagnosed by doctor; N, count

not systematically different between the two methods (P value = 0.791 and P value = 0.324 for COLOC and FINEMAP, respectively; two-sided t test). Furthermore, loci with strong evidence of colocalization (COLOC PP4 > 0.5 or FINEMAP CLPP > 0.01) have similarly high posterior probabilities of colocalization regardless of the correction method, indicating that robust effects are captured by both ancestry adjustments.

Of 174,388 loci tested for colocalization, 793 loci (< 0.5%) have at least one colocalization reported by *either* COLOC or FINEMAP. Only 159 of these loci have at least one concordant colocalization reported by *both* COLOC and FINEMAP (i.e., both methods report a colocalization for LocalAA or GlobalAA or both). For a subset of 31 loci, one ancestry-adjusted eQTL signal colocalized but the other did not, with concordant results between the two colocalization methods. Twenty-two and 9 loci demonstrate stronger colocalization with GlobalAA and LocalAA, respectively (highlighted points, Fig. 3a, b; Fig. 3c; Additional file 1, Figure S5). Interestingly, all 31 loci correspond with GWAS in primarily European cohorts, regardless of whether colocalization is stronger with GlobalAA or LocalAA.

Six of the loci with stronger GlobalAA colocalizations are associated with the same eGene, *AP003108.2* in the tibial artery. The six colocalized GWAS are associated with three types of traits: asthma (UKB self-reported asthma; UKB diagnosed-by-doctor asthma); red blood cell counts (Astle et al. red blood cell count; Astle et al. reticulocyte count); and fatty acids (GLGC triglycerides; MAGNETIC CH₂:double bond ratio in circulating fatty acids). Despite this replicated colocalization, neither the unannotated gene *AP003108.2* nor the GlobalAA lead eVariant, rs492751, has reported associations in the GWAS Catalog [35]. We further observed that rs492751 has highly variable allele frequencies between 1000 Genomes superpopulations (alternative allele frequencies of 0.02, 0, and 0.76 in European, East Asian, and African populations, respectively). This suggests that these stronger colocalizations with the GlobalAA tibial artery *AP003108.2* eQTL signal may in fact be driven by spurious associations confounded by local ancestry. Notably, a stronger colocalization with one eQTL ancestry adjustment is not synonymous with a more accurate eQTL signal; confounded associations can yield false discoveries.

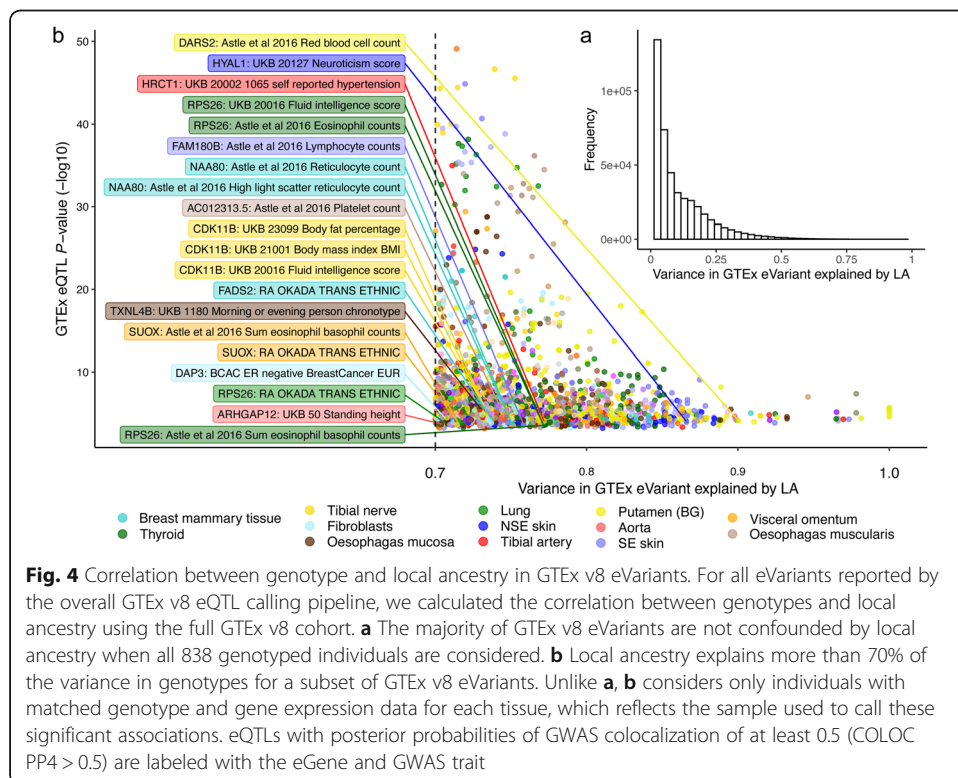
Two loci with stronger LocalAA colocalizations correspond with *MYO3A* in the tibial nerve. The associated traits are eosinophil counts and high light scatter reticulocyte counts (Astle et al.). *MYO3A* associations with interleukin-6, cortisol secretion, and BMI-adjusted waist circumference have previously been reported [35]; in other studies, eosinophil counts and characteristics of red blood cells have been correlated with obesity or BMI [36, 37], and obesity is associated with an inflammatory response [38, 39]. Therefore, a true colocalization between the tibial nerve *MYO3A* eQTL and traits related to properties of immature red blood cells and white blood cells is plausible. This locus provides an example of where LocalAA may outperform GlobalAA in terms of capturing true eQTL signals. However, we acknowledge that the differences in colocalization probabilities are smaller when LocalAA has a stronger colocalization compared to when GlobalAA has a stronger colocalization. In general, LocalAA may reduce false associations more often than it discovers true associations not also identified with GlobalAA. Overall, we observe that neither LocalAA nor GlobalAA performs significantly better in the context of colocalization, regardless of GWAS ancestry or colocalization method.

A subset of GTEx v8 eVariants is highly correlated with local ancestry

One justification for performing LocalAA as opposed to GlobalAA is the unique ability to avoid confounding by local population structure [15]. We examined all significant associations reported by the overall GTEx v8 eQTL calling pipeline for evidence of confounding with LA. Note that this analysis is expanded to include the full GTEx v8

cohort, not just the admixed sub-cohort involved in preceding analyses. For each GTEx eVariant in the set of all significant associations across 49 tissues, we found the variance in genotype explained by LA (the number of African and East Asian alleles at the locus) across all 838 genotyped individuals (see the “Methods” section). The vast majority of GTEx eVariants are not strongly correlated with LA when the entire genotyped population of 838 individuals is considered (Fig. 4a).

However, transcriptome sample sizes within each GTEx v8 eQTL tissue are often less than the full sample size (mean 310; standard deviation 171). Therefore, the degree of confounding between a variant’s genotype and LA in the context of eQTL mapping can vary between tissues. To this point, Fig. 4b provides the variance in genotype explained by LA for GTEx eVariants when only subjects with matched genotype and expression data are included in the regression. Unlike Fig. 4a, an eVariant has as many data points as tissues in which it is reported in a significant association. Twenty GTEx v8 eVariants whose corresponding eGenes have a colocalization probability of greater than 0.5, as reported by Barbeira et al., are also annotated [33]. Notably, 19 unique eVariants have proportions of variance explained by LA greater than 0.9 (Additional file 7, Table S6). These variants have large differences in reference allele frequencies between 1000 Genomes populations. For example, one such variant, chr1_1170732_A_G_b38, has reference allele frequencies of 0.993, 0.996, and 0.124 in European, East Asian, and African populations, respectively. A comprehensive list of the 2556 GTEx v8 significant associations where LA explains more than 70% of the variance in the eVariant genotype is provided in Table S7 (Additional file 8). We expect that functional follow-ups of eQTL/GWAS colocalizations will benefit from cross-referencing with these data.



Discussion

In this study, we describe population admixture in the GTEx v8 release and assess the impact of ancestry adjustment on eQTLs discovered in an admixed sub-cohort (117AX).

GTEx expands representation from non-European populations, including up to 17% of non-European or admixed individuals. For eQTL mapping, the selection of tissues was limited to those with adequate 117AX sample sizes (> 60). We recognize that these relatively small sample sizes will remain an important limitation of multi-population analyses in the GTEx study. Future comparable multi-tissue studies will benefit from increased representation of diverse populations.

The observed trend that GA explains more variance in residual gene expression than LA, on average, agrees with the previous finding that GA explains significantly more heritability of gene expression than LA [40]. However, LA can explain a large proportion of variance in GA-corrected gene expression for a subset of genes. Interestingly, a gene whose expression is largely explained by LA, *TBC1D3*, is a highly expanded gene whose copy number is stratified by ancestral population [29, 41]. Given that copy number expansion is a local phenomenon that has limited effects on global gene expression, population differences in gene copy numbers creates a scenario in which we would expect LA to explain more variance in gene expression than GA. This biological explanation for the differences in *TBC1D3* expression explained by ancestry highlights a specific benefit of considering LA during eQTL mapping.

We decided to include only admixed samples in eQTL mapping on the basis that we would not expect LocalAA to perform any better than GlobalAA in homogeneously European individuals, where the LA covariates are expected to be constant across the majority of the genome. For this same reason, we also excluded homogeneously African ($N=14$) and East Asian ($N=9$) samples from eQTL calling. However, this does not preclude the use of LocalAA as an ancestry adjustment approach in a cohort with individuals of both homogeneous and heterogeneous ancestry. To this point, Zhong et al. reached similar conclusions when comparing LA and GA adjustments in either a strictly African American population or a cohort of mostly European-ancestry individuals with less than 25% African Americans [10].

After performing *cis*-eQTL mapping in seven tissues, we observe that LocalAA has a modest improvement in power, consistent with previous observations [10, 42]. We also observe that most eQTLs agree between LocalAA and GlobalAA; the majority of eGenes that are called uniquely by one ancestry adjustment method are at the threshold of significance. Both of these observations are consistent with previous findings by Zhong et al. [10]. Further, eGenes called uniquely by GlobalAA are not confounded by LA. Neither do differences in variance in gene expression explained by LA or GA explain these eGenes uniquely called by one method. This, combined with the fact that both methods indicate the same lead eVariant more often than not, even when the association only reaches significance with one method, suggests that eGenes uniquely called by GlobalAA may not in fact be driven by confounding with LA. Instead, LocalAA and GlobalAA may have relatively more power for eQTL discovery in different contexts.

To our knowledge, the effects of LA adjustment in eQTL mapping on GWAS colocalization have not previously been explored. In general, stronger colocalization events

are captured by both ancestry adjustment methods. For 31 loci, only one of the two ancestry-adjusted eQTLs colocalizes with the GWAS, reported by both COLOC and FINEMAP. Interestingly, all 31 loci correspond with GWAS with European cohorts; no loci from the multi-ethnic GWAS robustly colocalize more strongly with either LocalAA or GlobalAA eQTLs. Six loci with stronger GlobalAA colocalization correspond with *AP003108.2* in the tibial artery; the GlobalAA lead eVariant has large differences in superpopulation allele frequencies, suggesting that confounding with local population structure is driving a spurious association signal. We also describe stronger colocalizations with LocalAA *MYO3A* eQTL signals in the tibial nerve that are supported by previously reported phenotypic associations. However, we find that neither LocalAA nor GlobalAA in eQTL mapping of seven different tissues yield systematically stronger colocalizations across 142 GWAS. Limitations of our colocalization analyses include our use of the assumption of one causal variant per trait and our lack of an attempt to colocalize secondary signals [43, 44].

Population-stratified eQTL calling is another potential approach in heterogeneous cohorts. To our knowledge, population-stratified eQTL calling has not yet been performed in the GTEx v8 cohort. However, the consortium did characterize population-biased *cis*-eQTLs (pb-eQTLs), where a variant's molecular effect on gene expression differs between individuals of European and African ancestry [1]. Only 178 pb-eQTLs for 141 unique eGenes (FDR \leq 25%) were identified across 31 tissues, which indicates that pb-eQTLs are hard to find and generally have small effects. Relatedly, Mogil et al. performed population-stratified eQTL calling independently in African American, Hispanic American, and European American samples in MESA; among several replication cohorts, the highest replication rate for all three discovery populations was in the Framingham Heart Study, a European cohort, simply because the sample size was much larger than the other population-matched replication cohorts [45]. This result, combined with the paucity of eQTLs with robust differences in effect sizes between populations, suggests that population-stratified eQTL calling at current sample sizes is limited in its ability to discover eQTLs not found in a pooled analysis.

One limitation of local ancestry inference is its dependence on the availability of appropriate reference panels. Access to genetic data for some populations remains limited, which makes it challenging to estimate local ancestry from those groups [26, 46]. Even with access to sufficient numbers of reference panels, there is a limit to the resolution that can be achieved with local ancestry inference given that local ancestry becomes more difficult to estimate as the genetic similarity between reference populations increases [11]. Addressing these challenges in future, larger functional genomics studies stands to improve our understanding of genetic risk across populations [47, 48] and resolution for the identification of causal variants [49].

Finally, the additional step of LA inference and the incorporation of LA into models for eQTL calling or GWAS makes LocalAA much more computationally intensive than GlobalAA. Therefore, a significant improvement of power for discovery or fine-mapping would be required to motivate the widespread implementation of LocalAA in large genetic association studies. Several groups recommend that GlobalAA is sufficient to control for type I error during screening for genetic associations, but LocalAA at loci of interest may improve fine-mapping or provide better effect estimates [5, 8, 9, 18]. Thus, a candidate approach may be taken to adjust for LA only at a subset of loci

where LA is expected to improve fine-mapping, which would reduce computational cost and maximize the potential benefit of LA adjustment.

A practical example of this is performing eQTL mapping with GlobalAA and subsequently assessing residual variance explained by LA for discovered eQTLs. To assess this, we post hoc analyzed GTEx release eVariants to discover 2556 associations that have a large amount of variance explained by local ancestry (> 70%). It remains a challenge to select a threshold for simply excluding QTLs based on the degree of variance explained by local ancestry. We provide this list to enhance the future analysis of eQTL/GWAS associations.

Conclusions

Despite claims of the importance of accounting for LA when performing genetic association studies in admixed populations [15, 16], the impact of LocalAA in the context of eQTL mapping and GWAS interpretation has been relatively underexplored. We performed genome-wide LA inference in an admixed sub-cohort of GTEx v8 and provide these LA calls as a resource to further investigate GTEx data. We then performed *cis*-eQTL mapping in this admixed sub-cohort to compare GlobalAA and LocalAA ancestry adjustment methods. We observe a modest improvement in power with LocalAA relative to GlobalAA. While both methods yield the same lead eVariant for the majority of eGenes, small subsets of eGenes have different lead eVariants between methods or pass the eQTL significance threshold in only one of the methods. We do not see large-scale or systematic differences in colocalization probabilities when we perform colocalization between GWAS and eQTLs where the two ancestry adjustments yield different lead eVariants. Finally, we provide a resource of GTEx v8 eVariants that are potentially confounded by LA. Together, these results describe the population structure of admixed individuals in the GTEx v8 release and demonstrate limited confounding based on local ancestry.

Methods

Genotype data

We used GTEx v8 release genotype data [1]. Briefly, whole genome sequencing (WGS) was performed for 899 samples from 869 unique GTEx donors, to a median depth of 32×. Alignment to the human reference genome build GRCh38 was performed with BWA-MEM [50]. Variants were called with GATK HaplotypeCaller v3.5, and multi-allelic sites were split into biallelic sites using Hail v0.1 [51]. After performing quality control, the final analysis freeze set contained variant calls from 838 donors. SHAPEIT v2 [52] was used to impute missing calls and phase the sample- and variant-QCed variant call file (VCF).

Genotype principal component analysis

We used GTEx v8 release genotype principal components (gPCs) [1]. gPCs were computed based on the sample- and variant-QCed WGS VCF using EIGENSTRAT [6]. PCA was performed on a set of LD-independent variants with a call rate $\geq 99\%$ and $MAF \geq 0.05$. LD pruning was performed using PLINK 1.9 [53].

Gene expression data

We used GTEx v8 release normalized gene expression data; detailed method descriptions can be found in the main GTEx publication [1]. RNA sequencing (RNA-seq) was

performed at the Broad Institute using the Illumina TruSeq™ RNA sample preparation protocol, which was based on polyA+ selection of mRNA and was not strand-specific. RNA-seq data were aligned to the human reference genome GRCh38/hg38 with STAR v2.5.3a [54]. Gene-level expression quantification was performed using RNA-SeQC [55] with a gene annotation available on the GTEx Portal (gencode.v26.GRCh38.genes.gtf). Quantified gene expression (TPM and raw counts) for each tissue was filtered and normalized according to the GTEx eQTL discovery pipeline [56]. For each of the seven tissues in which we chose to perform eQTL mapping, we subsetted normalized gene expression to include only 117AX samples.

Local ancestry inference

LiftOver [57] was used to convert the phased GTEx v8 whole genome sequencing variant call file (VCF) (dbGaP accession number phs000424.v8) from reference genome Human Build 38 (hg38) to Human Build 37 (hg19) for compatibility with 1000 Genomes and the hg19 HapMap genetic map. The resulting GTEx VCF was filtered to include self-reported African Americans and Asian Americans (103 and 12 individuals, respectively) as well as 25 admixed individuals as identified by the genotype PCA (Fig. 1a), resulting in 140 individuals. 1000 Genomes Phase 3 phased VCFs [58] were filtered to include biallelic variants and only individuals in the following populations: Han Chinese in Beijing, China (CHB); Japanese in Tokyo, Japan (JPT); Utah residents (CEPH) with Northern and Western European ancestry (CEU); Yoruba in Ibadan, Nigeria (YRI); Gambian in Western Divisions in the Gambia (GWD); Mende in Sierra Leone (MSL); and Esan in Nigeria (ESN) [19]. The intersection of autosomal variants in the resulting GTEx and 1000 Genomes VCFs ($N \sim 28$ M) was identified for LA inference. For compatibility with RFMix v1.5.4, variant positions were converted from base pairs to centimorgans [59] using the HapMap hg19 genetic map [60].

RFMix v1.5.4 [61] was run in PopPhased mode with the additional --forward-backward option [26]. All other parameters were set to the default values. The 1000 Genomes populations were used as reference panels for European (EUR), East Asian (ASN), and African (AFR) populations as follows: EUR (CEU, $N = 99$), ASN (CHB, JPT, $N = 207$), and AFR (YRI, GWD, MSL, ESN, $N = 405$) [19]. This generated posterior probabilities for the assignment of each phased allele to each of the three reference populations (EUR, AFR, ASN). An allele was assigned to a reference population only if the posterior probability was at least 0.9; otherwise, the local ancestry was indicated as “unknown.” For each individual, consecutive phased alleles with the same LA assignment were collapsed into BED files of haplotype blocks with the same LA (Additional file 3, Table S2). These BED files were then used to calculate global ancestry fractions per individual. Scripts used to collapse LA into BED files and calculate global ancestry fractions are available [62].

Of the 140 GTEx v8 individuals whose LA was inferred, 117 individuals with less than 90% global ancestry in a single population (among EUR, AFR, and ASN) were defined as admixed and retained for downstream analyses. This cohort is referred to as 117AX in this paper. VCFtools [63] was used to filter the hg19 GTEx VCF down to variants with a minor allele count (MAC) of at least 10 in 117AX. For the remaining 8,088,666 variants, the LA BED files (Additional file 3, Table S2) were used to count

the number of EUR, AFR, ASN, and unknown alleles at each SNP within 117AX. These allele counts were used as LA covariates in eQTL mapping with LocalAA.

Variance in gene expression explained by ancestry

We adapted an existing approach [27] to quantify variance in gene expression explained independently by LA or GA. For each expressed gene in each tissue, we performed two-step regressions to quantify variance explained by LA (or GA) in the gene expression residualized by GA (or LA). First, we regressed out the effects of one type of ancestry (LA or GA) on the gene expression using the following multiple linear regression, where γ_i is the effect of ancestry covariate a_i on gene expression g , and e_g is the residual:

$$g = \sum_{i=1}^m \gamma_i a_i + e_g$$

m is five for GA (five genotype PCs) and two for LA (numbers of alleles assigned to African or East Asian ancestry at the gene's transcription start site). Then, we quantified variance in e_g explained by the other type of ancestry (a^* , LA or GA covariates) by taking the coefficient of determination from the following linear regression:

$$e_g = \sum_{i=1}^m \gamma_i a_i^* + \epsilon$$

This process was performed for both LA and GA. All regressions were performed with the `lm()` function in R.

cis-eQTL mapping with LocalAA and GlobalAA

Genome-wide *cis*-eQTL mapping in 117AX was performed in seven GTEx v8 tissues: subcutaneous adipose (subc. adipose), tibial artery, lung, skeletal muscle, tibial nerve, whole blood, and not-sun-exposed suprapubic skin (NSE skin). All methods in this section were performed independently for each tissue. Normalized gene expression files filtered to include only 117AX samples were used to calculate 15 hidden confounders with PEER [64] according to the GTEx eQTL discovery pipeline [56]. Additional sample-level covariates, including gPCs, WGS sequencing platform (HiSeq 2000 or HiSeq X), WGS library construction protocol (PCR-based or PCR-free), and donor sex, were extracted from GTEx v8 release covariate files.

We assumed an additive genetic effect on gene expression and fit the following linear model for each gene-variant pair (gene g , variant ν):

$$G = \beta V + \sum_{i=1}^k \alpha_i c_i + \sum_{i=1}^m \gamma_i a_i + e$$

where G is the expression of gene g across 117AX samples in the given tissue; V is the number of alternate alleles at variant ν , coded as 0, 1, or 2; β is the effect of the alternate allele of variant ν on gene g expression; α_i is the effect of the technical or biological covariate c_i on gene g expression, including donor sex, sequencing platform, library construction protocol, and fifteen hidden confounders; γ_i is the effect of ancestry covariate a_i on gene g expression; and e is the residual. Any of the 8,088,666 filtered variants within a megabase of the transcription start site of a gene were tested for an association

with that gene's expression. The significance of an association was taken to be the two-sided P value corresponding to the t -statistic of the β coefficient estimate. All regressions were performed with the `lm()` function in R.

For each gene-variant pair, two iterations of this regression were performed: one to adjust for global ancestry (GlobalAA), in which case each a_i is one of the first five genotype principal components (gPCs), and one to correct for local ancestry (LocalAA), in which case there are two ancestry covariates, coded as the number of alleles at variant v assigned to African and East Asian populations, respectively. gPCs were not included as covariates in the LocalAA model. For LocalAA, samples with any number of alleles with unknown ancestry for the given variant were excluded; the covariate matrix was necessarily reconstructed for each variant tested. This is unlike GlobalAA, where the GA covariates are also sample-level covariates and can be reused for every association test.

After eQTL mapping was completed, the most significant, i.e., lead, eVariant (or eVariants, in the case of tied P values) was identified for each gene, independently for the two ancestry adjustment methods. A nominal P value cutoff of $1e-6$ was applied to identify significant associations. This threshold approximates a 5% FDR (Additional file 1, Figure S4). LD (R^2) was calculated between single pairs of GlobalAA and LocalAA lead eVariants for each eGene using PLINK [53]; an eGene was defined as having different lead eVariants between the two ancestry adjustment methods if (1) there was no intersection between the two sets of lead eVariants and (2) the LD between the tested pair of GlobalAA and LocalAA lead eVariants was less than 1.0.

Variance in GTEx eVariant genotype explained by local ancestry

In order to identify potential confounding by LA in GTEx v8 eQTLs, we first needed LA calls for all 838 individuals with both WGS and RNA-seq data [1]. The remaining 698 individuals for which we did not perform LA inference have self-reported European ancestry and cluster tightly together in gPC space (Fig. 1a). Therefore, we approximated LA in these 698 individuals to two European alleles at all tested loci. Then, LA covariates for this analysis were the union of computationally inferred LA in 140 admixed or non-European individuals and approximated LA in the remaining 698 homogeneously European individuals.

We calculated the variance explained by LA in the genotype of each eVariant implicated in reported GTEx v8 eQTLs. The following linear model was fit for each eVariant:

$$V = \alpha \times AFR + \beta \times ASN + e$$

where V is the genotype vector (number of minor alleles), and AFR and ASN are the two LA covariate vectors, representing the number of alleles assigned to African and East Asian populations, respectively. The resulting coefficient of determination of each regression was recorded. We did this in two settings: (1) for the set of unique eVariants across all GTEx v8 eQTLs, where genotypes and LA for all 838 individuals were included in the regression (Fig. 4a), and (2) for all eVariants within each tissue, with samples subset to those with matched gene expression in the given tissue (Fig. 4b). (1) provides a global picture of the degree of correlation between eVariant genotypes and LA while (2) reflects the actual samples used to call eQTLs in each tissue. For (2), we also intersected GTEx v8 eQTLs with GTEx v8 GWAS colocalization results (see

below) to identify loci with high posterior probabilities of colocalization between eQTLs and GWAS ($PP4 > 0.5$) associated with eVariants whose genotypes are highly correlated with LA ($R^2 > 0.7$).

Imputation of GWAS summary statistics

Harmonization and imputation of 114 previously published GWAS are described in detail by [33] and [1]. Briefly, summary statistics were harmonized and lifted over to hg38; an in-house implementation of best linear unbiased prediction (BLUP) [65, 66] was used to impute z -scores for those variants reported in GTEx without matching data in the GWAS summary statistics.

Colocalization between eQTL and GWAS signals

We performed colocalization between 142 GWAS and 14 sets of 117AX eQTL summary statistics (one set for each ancestry adjustment in each of seven tissues). Colocalization tests were restricted to the subset of genes where the two eQTL ancestry adjustments yielded different lead variants with nominal P values less than $1e-4$. COLOC was used to test for colocalization at all of these loci [30]; an implementation of FINEMAP was used to test for colocalization at the subset of loci for which COLOC reported a colocalization ($PP4 > 0.5$) [31]. Inputs were prepared similarly for COLOC and FINEMAP analyses. Each GWAS was scanned for putative association signals, defined as variants with a nominal P value less than $1e-5$. If multiple variants within a 1-MB window had a P value less than this threshold, the variant with the smallest P value was selected as the seed variant. For each GWAS seed variant, if there was an eQTL with a P value of less than $1e-4$ within 1 MB, the intersection of GWAS and eQTL variants within 1 MB of the GWAS seed variant was tested for colocalization. The same GWAS seed variant was used to perform colocalization with GlobalAA and LocalAA eQTL signals at each locus. Colocalization method-specific parameters are detailed below.

COLOC

For colocalization analyses between 142 GWAS and 117AX eQTL summary statistics, the same GTEx VCF used for eQTL mapping in 117AX was used to calculate eQTL effect allele frequencies; GWAS effect allele frequencies were extracted from the GWAS summary statistics. The `coloc.abf()` function in the “coloc” R package was used to run COLOC. For binary GWAS traits, case proportion and “cc” trait type parameters were used. For continuous GWAS traits, sample size and “quant” trait type parameters were used. These GWAS characteristics are provided in Table S5 (Additional file 6).

Figure 4b references colocalizations identified by an independent analysis of the 114 imputed GWAS and eQTLs reported in the GTEx v8 release [33]. Briefly, COLOC was used to perform colocalization with variants in the *cis*-window of each gene with at least one eVariant (*cis*-eQTL per-tissue q value < 0.05). For binary GWAS traits, case proportion and “cc” trait type parameters were used. For continuous GWAS traits, sample size and “quant” trait type parameters were used. In both cases, imputed or calculated z -scores were used as effect coefficients in Bayes factor calculations. Enloc enrichment estimates [67] were used to define data-based priors for COLOC in a consistent manner with other GTEx companion papers [33].

FINEMAP

An implementation of FINEMAP was used to test for colocalization at the subset of loci for which COLOC reported a colocalization ($PP4 > 0.5$) between a GWAS and an 117AX eQTL. After the intersection of GWAS and eQTL variants within 1 MB of the GWAS seed variant was identified for a locus, FINEMAP v1.1 was run independently for the GWAS and eQTL association signals using parameters `--n-causal-max 1 --n-iterations 1000000 --n-convergence 1000`. The 1000 Genomes Phase 3 VCF was used for LD calculations [19]. As previously described [32], the marginal posterior inclusion probabilities (PIPs) for each of K variants were then multiplied to calculate a colocalization posterior probability (CLPP):

$$CLPP = 1 - \left[\prod_{i=1}^K 1 - (PIP_{GWAS,i} \times PIP_{eQTL,i}) \right]$$

$PIP_{GWAS,i}$ is the PIP for the i th variant in the vector of K variants tested for the causality of the GWAS signal; $PIP_{eQTL,i}$ is the PIP for the i th variant in the vector of K variants tested for the causality of the eQTL signal. The i th variant in the list of tested GWAS variants is the same as the i th variant in the list of tested eQTL variants for all i .

Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s13059-020-02113-0>.

Additional file 1: Figure S1. Supplementary figures (Figures S1-S5).

Additional file 2: Table S1. GTEx IDs for individuals in 117AX cohort.

Additional file 3: Table S2. Local ancestry calls for 140 GTEx individuals.

Additional file 4: Table S3. Variance in residual 117AX gene expression explained by ancestry.

Additional file 5: Table S4. 117AX GlobalAA and LocalAA lead eVariants for all tested genes.

Additional file 6: Table S5. Characteristics of GWAS used in colocalization analyses.

Additional file 7: Table S6. GTEx v8 eVariants highly correlated with LA ($R^2 > 0.9$).

Additional file 8: Table S7. GTEx v8 eVariants correlated with LA ($R^2 > 0.7$).

Additional file 9. GTEx Consortium author list.

Additional file 10. Review History.

Acknowledgements

We thank Daniel Nachun and Jonathan Pritchard for the helpful conversations. We thank Joe Pickrell for his publicly available code. We thank Carlos Bustamante, Brian K. Maples, and Christopher DeBoever for the development and maintenance of RFMix. The full GTEx Consortium author list is provided in Additional file 9.

Review history

The review history is available as Additional file 10.

Peer review information

Tim Sands was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions

N.R.G. and S.B.M. designed the study. N.R.G. performed all analyses involving 117AX or local ancestry, wrote the first draft of the manuscript, and prepared all figures, tables, and supplementary information. M.G. developed the wrapper pipeline that N.R.G. used to perform the colocalization. M.G. and A.S.R. provided feedback about the colocalization analyses. N.S.A., B.B., and S.M. provided advice about the statistical analyses. A.R.M., M.L.A., and S.M. contributed to discussions about ancestry inference algorithms and software. F.A. generated the genotype principal components used as a proxy for global ancestry. F.A. and K.G.A. generated the expression read counts that were then normalized by F.A. and used for *cis*-eQTL mapping. F.A. and K.G.A. generated the GTEx v8 release *cis*-eQTL call sets. A.N.B., R.B., and H.K.I. generated the harmonized and imputed summary statistics for 114 of the GWAS used for colocalization. Y.P., A.N.B., F.H., C.D.B., X.W., and H.K.I. performed the colocalization between GTEx v8 eQTLs and the imputed GWAS summary statistics.

Y.P., M.L.A, B.B., and M.G. helped revise the manuscript. T.L. provided critical feedback. S.B.M. provided guidance for the analyses and helped write and edit the manuscript. All authors read and approved the final manuscript.

Authors' information

Twitter handles: @genetisaur (Alicia R. Martin); @wenxqwen (Xiaoquan Wen); @sbmontgom (Stephen B. Montgomery).

Funding

This work was funded by NIH grants R01 HG008150, U01 HG009080, R01 HL142015, and U01 HG009431. N.R.G. was funded by the Stanford Genome Training Program NIH Training Grant (5T32HG000044-22) and the 2018 NSF Graduate Research Fellowship Program. Y.P. is funded by the NHGRI award R01HG010067.

Availability of data and materials

GTEX v8 release gene expression data and *cis*-eQTL call sets are available through the GTEX Portal [68]. GTEX v8 genotype data are available through the dbGaP website under dbGaP accession phs000424.v8.p2 [69]. LocalAA and GlobalAA eQTL summary statistics and colocalization posterior probabilities are available through Zenodo [70]. The 114 GWAS summary statistics imputed and harmonized by the GWAS GTEX Subgroup are available through Zenodo [71]. The source code is published on Zenodo under a Creative Commons Attribution 4.0 International License and available on GitHub [72]; a publication version of the code has been deposited on Zenodo [73].

Ethics approval and consent to participate

While deceased individuals do not require consent for research, GTEX consent is described in detail in [74].

Consent for publication

Not applicable.

Competing interests

F.A. is an inventor on a patent application related to TensorQTL; H.K.I. has received speaker honoraria from GSK and AbbVie; T.L. is a scientific advisory board member of Variant Bio with equity and Goldfinch Bio; S.B.M. is on the scientific advisory board of Myome.

Author details

¹Department of Genetics, Stanford University, Stanford, CA, USA. ²Biomedical Informatics, Stanford University, Stanford, CA, USA. ³Department of Biomathematics, University of California, Los Angeles, Los Angeles, CA, USA. ⁴Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. ⁵Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. ⁶Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA. ⁷Stanley Center for Psychiatric Research, Broad Institute, Cambridge, MA, USA. ⁸Department of Bioengineering, Stanford University, Stanford, CA, USA. ⁹The Broad Institute of MIT and Harvard, Cambridge, MA, USA. ¹⁰Section of Genetic Medicine, Department of Medicine, The University of Chicago, Chicago, IL, USA. ¹¹Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA. ¹²New York Genome Center, New York, NY, USA. ¹³Department of Systems Biology, Columbia University, New York, NY, USA. ¹⁴Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA. ¹⁵Department of Pathology, Stanford University, Stanford, CA, USA.

Received: 8 November 2019 Accepted: 19 July 2020

Published online: 11 September 2020

References

1. Aguet F, Barbeira AN, Bonazzola R, Brown A, Castel SE, Jo B, et al. The GTEX Consortium atlas of genetic regulatory effects across human tissues. *bioRxiv*. 2019;787903. <https://doi.org/10.1101/787903>.
2. Hellwege J, Keaton J, Giri A, Gao X, Velez Edwards DR, Edwards TL. Population stratification in genetic association studies. *Curr Protoc Hum Genet*. 2017;95:1.22.1–1.22.23.
3. Thornton TA, Bermejo JL. Local and global ancestry inference and applications to genetic association analysis for admixed populations. *Genet Epidemiol*. 2014;38(S1):S5–12.
4. Chakraborty R, Weiss KM. Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proc Natl Acad Sci U S A*. 1988;85(23):9119–23.
5. Martin ER, Tunc I, Liu Z, Slifer SH, Beecham AH, Beecham GW. Properties of global- and local-ancestry adjustments in genetic association tests in admixed populations. *Genet Epidemiol*. 2018;42(2):214–29.
6. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 2006;38(8):904–9.
7. GTEX Consortium. Genetic effects on gene expression across human tissues. *Nature*. 2017;550(7675):204–13.
8. Liu J, Lewinger JP, Gilliland FD, Gauderman WJ, Conti DV. Confounding and heterogeneity in genetic association studies with admixed populations. *Am J Epidemiol*. 2013;177(4):351–60.
9. Zhang J, Stram DO. The role of local ancestry adjustment in association studies using admixed populations. *Genet Epidemiol*. 2014;38(6):502–15.
10. Zhong Y, Perera MA, Gamazon ER. 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*. 2019;104(6):1097–115.
11. Paşaniuc B, Sankararaman S, Kimmel G, Halperin E. Inference of locus-specific ancestry in closely related populations. *Bioinformatics*. 2009;25(12):213–21.
12. Duan Q, Xu Z, Raffield LM, Chang S, Wu D, Lange EM, et al. A robust and powerful two-step testing procedure for local ancestry adjusted allelic association analysis in admixed populations. *Genet Epidemiol*. 2018;42(3):288–302.

13. Kang SJ, Larkin EK, Song Y, Barnholtz-Sloan J, Baechle D, Feng T, et al. Assessing the impact of global versus local ancestry in association studies. *BMC Proc.* 2009;3(Suppl 7):S107.
14. Qin H, Morris N, Kang SJ, Li M, Tayo B, Lyon H, et al. Interrogating local population structure for fine mapping in genome-wide association studies. *Bioinformatics.* 2010;26(23):2961–8.
15. Wang X, Zhu X, Qin H, Cooper RS, Ewens WJ, Li C, et al. Adjustment for local ancestry in genetic association analysis of admixed populations. *Bioinformatics.* 2011;27(5):670–7.
16. Chen M, Yang C, Li C, Hou L, Chen X, Zhao H. Admixture mapping analysis in the context of GWAS with GAW18 data. *BMC Proc.* 2014;8(Suppl 1):S3.
17. Pasaniuc B, Zaitlen N, Lettre G, Chen GK, Tandon A, Kao WHL, et al. Enhanced statistical tests for GWAS in admixed populations: assessment using African Americans from CARE and a breast cancer consortium. *PLoS Genet.* 2011;7(4):e1001371.
18. Pino-Yanes M, Gignoux CR, Galanter JM, Levin AM, Campbell CD, Eng C, et al. Genome-wide association study and admixture mapping reveal new loci associated with total IgE levels in Latinos. *J Allergy Clin Immunol.* 2015;135(6):1502–10.
19. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature.* 2015;526(7571):68–74.
20. Min JL, Taylor JM, Richards JB, Watts T, Pettersson FH, Broxholme J, et al. The use of genome-wide eQTL associations in lymphoblastoid cell lines to identify novel genetic pathways involved in complex traits. *PLoS ONE.* 2011;6(7):e22070.
21. Lovejoy JC, de la Bretonne JA, Klemperer M, Tulley R. Abdominal fat distribution and metabolic risk factors: effects of race. *Metabolism.* 1996;45(9):1119–24.
22. Forouhi NG, Sattar N, McKeigue PM. Relation of C-reactive protein to body fat distribution and features of the metabolic syndrome in Europeans and South Asians. *Int J Obes Relat Metab Disord.* 2001;25(9):1327–31.
23. Yin L, Coelho SG, Ebsen D, Smuda C, Mahns A, Miller SA, et al. Epidermal gene expression and ethnic pigmentation variations among individuals of Asian, European and African ancestry. *Exp Dermatol.* 2014;23(10):731–5.
24. Silva AM, Shen W, Heo M, Gallagher D, Wang Z, Sardinha LB, et al. Ethnicity-related skeletal muscle differences across the lifespan. *Am J Hum Biol.* 2010;22(1):76–82.
25. Stewart KA, Higgins PC, McLaughlin CG, Williams TV, Granger E, Croghan TW. Differences in prevalence, treatment, and outcomes of asthma among a diverse population of children with equal access to care: findings from a study in the military health system. *Arch Pediatr Adolesc Med.* 2010;164(8):720–6.
26. Maples BK, Gravel S, Kenny EE, Bustamante CD. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am J Hum Genet.* 2013;93(2):278–88.
27. Shriner D, Bentley AR, Doumatey AP, Chen G, Zhou J, Adeyemo A, et al. Phenotypic variance explained by local ancestry in admixed African Americans. *Front Genet.* 2015;6:324. <https://doi.org/10.3389/fgene.2015.00324>.
28. Paulding CA, Ruvolo M, Haber DA. The Tre2 (USP6) oncogene is a hominoid-specific gene. *Proc Natl Acad Sci U S A.* 2003;100(5):2507–11.
29. Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, et al. Diversity of human copy number variation and multicopy genes. *Science.* 2010;330(6004):641–6.
30. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* 2014;10(5):e1004383.
31. Benner C, Spencer CCA, Havulinna AS, Salomaa V, Ripatti S, Pirinen M. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics.* 2016;32(10):1493–501.
32. Hormozdiari F, van de Bunt M, Segrè AV, Li X, Joo JWJ, Bilow M, et al. Colocalization of GWAS and eQTL signals detects target genes. *Am J Hum Genet.* 2016;99(6):1245–60.
33. Barbeira AN, Bonazzola R, Gamazon ER, Liang Y, Park Y, Kim-Hellmuth S, et al. Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. *bioRxiv.* 2020;814350. <https://doi.org/10.1101/814350>.
34. Wojcik GL, Graff M, Nishimura KK, Tao R, Haessler J, Gignoux CR, et al. Genetic analyses of diverse populations improves discovery for complex traits. *Nature.* 2019;570(7762):514–8.
35. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* 2019;47(D1):D1005–12.
36. Sunadome H, Matsumoto H, Izuhara Y, Nagasaki T, Kanemitsu Y, Ishiyama Y, et al. Correlation between eosinophil count, its genetic background and body mass index: the Nagahama Study. *Allergol Int.* 2020;69(1):46–52.
37. Altunoglu E, Müderrisoğlu C, Erdenen F, Ülgen E, Ar MC. The impact of obesity and insulin resistance on iron and red blood cell parameters: a single center, cross-sectional study. *Turk J Hematol.* 2014;31(1):61–7.
38. Ferrante AW. Obesity-induced inflammation: a metabolic dialogue in the language of inflammation. *J Intern Med.* 2007;262(4):408–14.
39. Shoelson SE, Herrero L, Naaz A. Obesity, inflammation, and insulin resistance. *Gastroenterology.* 2007;132(6):2169–80.
40. Price AL, Patterson N, Hancks DC, Myers S, Reich D, Cheung VG, et al. Effects of cis and trans genetic ancestry on gene expression in African Americans. *PLoS Genet.* 2008;4(12):e1000294.
41. Hodzic D, Kong C, Wainszelbaum MJ, Charron AJ, Su X, Stahl PD. TBC1D3, a hominoid oncoprotein, is encoded by a cluster of paralogues located on chromosome 17q12. *Genomics.* 2006;88(6):731–6.
42. Yorgov D, Edwards KL, Santorico SA. Use of admixture and association for detection of quantitative trait loci in the Type 2 Diabetes Genetic Exploration by Next-Generation Sequencing in Ethnic Samples (T2D-GENES) study. *BMC Proc.* 2014;8(Suppl 1):S6.
43. Wallace C. Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. *PLoS Genet.* 2020;16(4):e1008720.
44. Dobbyn A, Huckins LM, Boocock J, Sloofman LG, Glicksberg BS, Giambartolomei C, et al. Landscape of conditional eQTL in dorsolateral prefrontal cortex and co-localization with schizophrenia GWAS. *Am J Hum Genet.* 2018;102(6):1169–84.
45. Mogil LS, Andaleon A, Badalamenti A, Dickinson SP, Guo X, Rotter JJ, et al. Genetic architecture of gene expression traits across diverse populations. *PLoS Genet.* 2018;14(8):e1007586.
46. Baran Y, Pasaniuc B, Sankararaman S, Torgerson DG, Gignoux C, Eng C, et al. Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics.* 2012;28(10):1359–67.
47. Amariuta T, Ishigaki K, Sugishita H, Ohta T, Matsuda K, Murakami Y, et al. In silico integration of thousands of epigenetic datasets into 707 cell type regulatory annotations improves the trans-ethnic portability of polygenic risk scores. *bioRxiv.* 2020;959510. <https://doi.org/10.1101/2020.02.21.959510>.

48. Rosenberg NA, Huang L, Jewett EM, Szpiech ZA, Jankovic I, Boehnke M. Genome-wide association studies in diverse populations. *Nat Rev Genet.* 2010;11(5):356–66.
49. Zaitlen N, Paşaniuc B, Gur T, Ziv E, Halperin E. Leveraging genetic variability across populations for the identification of causal variants. *Am J Hum Genet.* 2010;86(1):23–33.
50. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2009;25(14):1754–60.
51. Hail. <https://hail.is/>. Accessed 21 Apr 2020.
52. Delaneau O, Zagury J-F, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods.* 2013;10(1):5–6.
53. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007;81(3):559–75.
54. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* 2013;29(1):15–21.
55. DeLuca DS, Levin JZ, Sivachenko A, Fennell T, Nazaire M-D, Williams C, et al. RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics.* 2012;28(11):1530–2.
56. GTEx Consortium. eQTL discovery pipeline for the GTEx Consortium. GitHub. <https://github.com/broadinstitute/gtex-pipeline/tree/master/qtl>. Accessed 13 Feb 2018.
57. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, et al. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* 2006;34(Database issue):D590–8.
58. 1000 Genomes FTP. <https://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20130502/>. Accessed 19 Jan 2018.
59. Pickrell J. interpolate_maps.py. GitHub. https://github.com/joepickrell/1000-genomes-genetic-maps/blob/master/scripts/interpolate_maps.py. Accessed 19 Jan 2018.
60. International HapMap Project. HapMap genetic map FTP. ftp://ftp.ncbi.nlm.nih.gov/hapmap/recombination/2011-01_phase1_B37. Accessed 19 Jan 2018.
61. Maples BK. RFMix v1.5.4. <https://sites.google.com/site/rfmixlocalancestryinference>. Accessed 12 Jan 2018.
62. Martin AR. Ancestry pipeline. GitHub. https://github.com/armartin/ancestry_pipeline. Accessed 26 Jan 2018.
63. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. *Bioinformatics.* 2011;27(15):2156–8.
64. 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.* 2012;7(3):500–7.
65. Lee D, Bigdeli TB, Riley BP, Fanous AH, Bacanu S-A. DIST: direct imputation of summary statistics for unmeasured SNPs. *Bioinformatics.* 2013;29(22):2925–7.
66. Pasiñuc B, Zaitlen N, Shi H, Bhatia G, Gusev A, Pickrell J, et al. Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics.* 2014;30(20):2906–14.
67. Wen X, Pique-Regi R, Luca F. Integrating molecular QTL data into genome-wide genetic association analysis: probabilistic assessment of enrichment and colocalization. *PLoS Genet.* 2017;13(3):e1006646.
68. GTEx Portal. <https://gtexportal.org>. Accessed 1 July 2020.
69. Common Fund (CF) Genotype-Tissue Expression Project (GTEx). dbGaP. https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000424.v8.p2. Accessed 1 July 2020.
70. Gay NR, Gloudemans M, Antonio ML, Balliu B, Park Y, Martin AR, et al. Extended data: Impact of admixture and ancestry on eQTL analysis and GWAS colocalization in GTEx [Data set]. Zenodo; 2020. <https://doi.org/10.5281/zenodo.3926871>.
71. Barbeira AN, Bonazzola R, Gamazon ER, Liang Y, Park Y, Ardlie K, et al. Publicly available GWAS summary statistics, harmonized and imputed to GTEx v8' variant reference [Data set]. Zenodo; 2020. <https://doi.org/10.5281/zenodo.3629742>.
72. Gay NR. gtex-admixture-la. GitHub. <https://github.com/nicolerg/gtex-admixture-la>. Accessed 1 July 2020.
73. Gay NR. Source code: impact of admixture and ancestry on eQTL analysis and GWAS colocalization in GTEx. Zenodo; 2020. <https://doi.org/10.5281/zenodo.3924788>. Accessed 1 July 2020.
74. Carithers LJ, Ardlie K, Barcus M, Branton PA, Britton A, Buia SA, et al. A novel approach to high-quality postmortem tissue procurement: the GTEx project. *Biopreserv Biobank.* 2015;13(5):311–9.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

