


RESEARCH

Open Access



The genetic and evolutionary basis of gene expression variation in East Africans

Derek E. Kelly^{1,2}, Shweta Ramdas², Rong Ma³, Renata A. Rawlings-Goss², Gregory R. Grant², Alessia Ranciaro², Jibril B. Hirbo^{4,5}, William Beggs², Meredith Yeager⁶, Stephen Chanock⁷, Thomas B. Nyambo⁸, Sabah A. Omar⁹, Dawit Woldemeskel¹⁰, Gurja Belay¹⁰, Hongzhe Li³, Christopher D. Brown^{1,2} and Sarah A. Tishkoff^{2,11*} 

*Correspondence:
tishkoff@pennmedicine.upenn.
edu

² Genetics, University
of Pennsylvania, Philadelphia,
PA, USA
Full list of author information is
available at the end of the article

Abstract

Background: Mapping of quantitative trait loci (QTL) associated with molecular phenotypes is a powerful approach for identifying the genes and molecular mechanisms underlying human traits and diseases, though most studies have focused on individuals of European descent. While important progress has been made to study a greater diversity of human populations, many groups remain unstudied, particularly among indigenous populations within Africa. To better understand the genetics of gene regulation in East Africans, we perform expression and splicing QTL mapping in whole blood from a cohort of 162 diverse Africans from Ethiopia and Tanzania. We assess replication of these QTLs in cohorts of predominantly European ancestry and identify candidate genes under selection in human populations.

Results: We find the gene regulatory architecture of African and non-African populations is broadly shared, though there is a considerable amount of variation at individual loci across populations. Comparing our analyses to an equivalently sized cohort of European Americans, we find that QTL mapping in Africans improves the detection of expression QTLs and fine-mapping of causal variation. Integrating our QTL scans with signatures of natural selection, we find several genes related to immunity and metabolism that are highly differentiated between Africans and non-Africans, as well as a gene associated with pigmentation.

Conclusion: Extending QTL mapping studies beyond European ancestry, particularly to diverse indigenous populations, is vital for a complete understanding of the genetic architecture of human traits and can reveal novel functional variation underlying human traits and disease.

Keyword: Human African genomics, Gene expression, eQTL, Human diversity, Natural selection



© The Author(s) 2023. **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.

Background

Gene regulation is a principal mechanism by which genetic variation contributes to phenotypic variation, making its study essential for understanding human evolution and disease. The genomics era has underscored the importance of noncoding variation in human evolution and disease: ~90% of the genotype–phenotype associations identified by genome-wide association studies (GWAS) cannot be explained by coding variation [1, 2], and similarly, genomic regions harboring evidence of selection in humans are significantly more enriched for variants altering expression than protein coding [3].

While GWAS and scans of selection can identify genomic regions of interest, they often lack the resolution to identify the specific genes underlying traits or targeted by selection. To bridge this gap, studies have aimed to identify genetic variation associated with fine-scale, molecular phenotypes, through quantitative trait locus (QTL) mapping [4]. Combining these molecular QTL maps with GWAS through colocalization, transcriptome-wide association studies, or Mendelian randomization continues to prove a fruitful approach for identifying genes causally linked to traits and potential drug targets. Unfortunately, there is a persistent ancestry bias in human genomics research, with nearly 80% of GWAS participants being of recent European ancestry [5, 6], as well as the majority of participants of molecular trait studies [7], greatly limiting our ability to translate findings from GWAS to diverse populations, as well as discover population-specific variation of interest [8].

While there is an established and active field of study identifying novel GWAS associations and genetic variation contributing to gene expression differences across populations [7, 9–14], most global populations remain understudied, particularly in sub-Saharan Africa. Africa is the birthplace of anatomically modern humans and harbors the greatest levels of human genetic diversity across continents. The majority of genomic studies of sub-Saharan African individuals have focused on populations of primarily West African descent, which fails to capture much of the genetic and phenotypic diversity within sub-Saharan Africa [15]. Moreover, Africa is home to a large array of biomes and terrains, and indigenous Africans continue to practice diverse cultural and subsistence strategies. Together, these environmental pressures have driven genetic adaptations to infectious disease [16], diet [17], and climate [10, 18], sometimes in a population-specific manner. These adaptive variants can have important implications for human health in Africa, and elsewhere [19], and inclusion of African populations is therefore vital for our understanding of human evolutionary history and health.

In this study, we probe the genetic architecture of gene regulation in whole blood from indigenous East Africans by performing expression QTL (eQTL) and splicing QTL (sQTL) mapping in a cohort of 162 individuals, representing nine ethnic groups, from Ethiopia and Tanzania. We measure the degree to which African architecture is shared with that of non-Africans, test whether Africans harbor functional variation absent from existing cohorts, and investigate the demographic and genetic forces that may contribute to variation in gene regulatory architecture. We test whether fine-mapping of QTL signals is improved in Africans relative to an equivalently sized cohort of European Americans, and highlight individual genes with improved fine-mapping in Africans. Finally, we measure the effect of selective forces on shaping gene regulatory architecture and identify candidate genes under selection.

Results

Population structure

The cohort for this study consists of 162 Ethiopian and Tanzanian individuals belonging to nine ethnically and culturally diverse sub-Saharan groups previously unsampled in gene expression studies, including the Cushitic speaking Agaw and Weyto, the Semitic speaking Argoba and Amhara, the Omotic speaking Dizi, the Nilo-Saharan speaking Mursi, the Chabu who speak an unclassified language similar to Nilo-Saharan, and the Khoesan speaking Hadza and Sandawe (Fig. 1A). These populations practice a variety of subsistence strategies, including foraging (Hadza and Chabu currently, Sandawe and Weyto formerly), pastoralism (Mursi), agriculturalism (Agaw, Amhara, and Argoba), and agropastoralism (Dizi), and live in diverse environments with differing pathogen exposures.

To investigate the genetic diversity and structure of these populations, a subset of 162 individuals were genotyped at approximately 4.5 million SNPs on the Illumina Omni5 Beadchip array. These data were further imputed using a reference panel composed of the 1000 Genomes Project (1kGP) dataset [20] and a dataset of whole genome sequences (WGS) from 180 sub-Saharan African individuals [21]. To place their genetic variation in a global context, genotype data from the nine study populations were merged with 1kGP WGS data from 20 individuals each of Yoruban (YRI), Northern and Western European (CEU), and Han Chinese (CHB) ancestry (methods). Principal component analysis (PCA) of this merged dataset recapitulates a primary separation between African and non-African individuals along the first PC, explaining 3.8% of the variance. The second PC, explaining 1.8% of the variance, further separates CEU and CHB individuals, as well as East Africans and the YRI (Fig. 1B). Higher PCs further

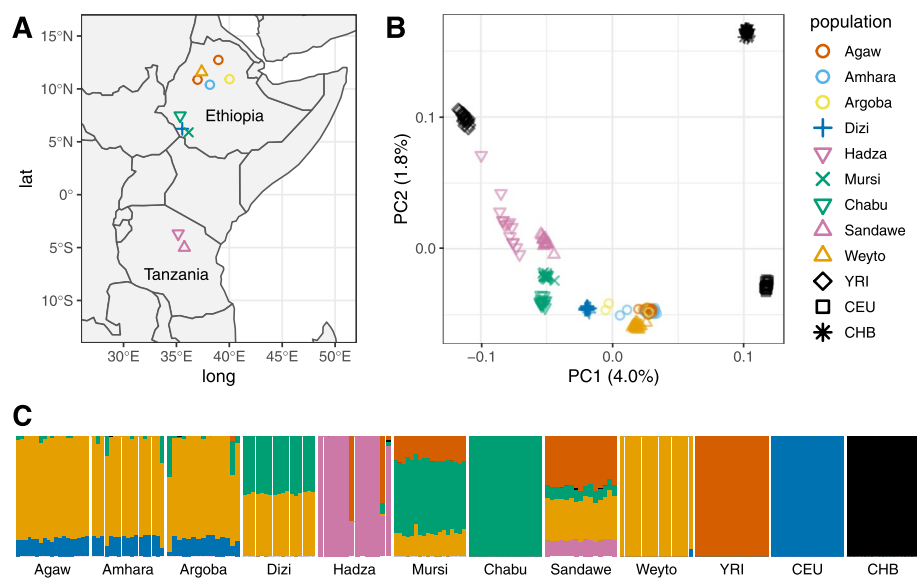


Fig. 1 Global and genetic structure of study populations. **A** Locations of East African populations sampled in this study across Ethiopia and Tanzania. **B** Principal component analysis of genetic data across 162 East Africans, combined with 20 West African Yoruba (YRI), 20 European Americans (CEU), and 20 Han Chinese (CHB) from the 1000 Genomes Project. The percent variance explained by each principal component is indicated in parentheses. **C** ADMIXTURE analysis of East African, YRI, CEU, and CHB populations

separate variation in Africa; PC3 captures variation between the Hadza and YRI, and PC4 between the Hadza and Chabu. Several groups cluster relatively nearer to CEU Europeans along PC1, most notably the Ethiopian Agaw, Amhara, Argoba, and Weyto, which are known to have moderate levels of Eurasian admixture [15, 22, 23]. Inferred ancestry components from *ADMIXTURE* [24] also indicates non-African admixture among these Ethiopian groups, as well as admixture with Bantu-speaking populations of Western African origin [20], represented by the YRI, in the Sandawe, Mursi, and Hadza (Fig. 1C).

Transcriptomic traits in Africans

To assess the contribution of genetic variation to transcriptomic trait variation, we performed genome-wide QTL mapping for expression (eQTL) and splicing (sQTL) transcriptomic traits in *cis* for expressed protein-coding and long-noncoding RNA genes; collectively, we refer to eQTLs and sQTLs as transcriptomic QTLs (tQTLs). We first correct our phenotypes (expression and splicing) for a number of covariates, including age, sex, delivery date, hidden covariates inferred by *PEER* [25], and cell-type fractions inferred by *CIBERSORT* [26]. Cell-type composition of whole blood is known to vary between individuals, and to be a source of confounding in QTL studies [27]. To account for ancestry and relatedness, we generate a genetic relatedness matrix (GRM) and perform tQTL mapping using the linear mixed model tool *GEMMA* [28]. Testing all autosomal SNPs with minor allele frequency (MAF) greater than 0.05 and within 100 kb of the target gene transcription start site (TSS) for eQTLs or within 100 kb of the target intron for sQTLs, we identify 99,685 SNPs associated with the expression of 1330 genes (eGenes) and 74,445 SNPs associated with splicing of 1118 introns (sIntrons) in 776 genes (sGenes) at $FDR < 0.05$ (Methods).

SNPs associated with expression (eSNPs) and splicing (sSNPs) show a characteristic enrichment near the transcription start site or intron boundary of their target gene, respectively [29] (Additional file 1: Fig. S3A and B), and are enriched in a variety of functional categories, including transcription start sites, enhancers, and splice sites, and are depleted in repressed chromatin regions. We also find a significant overlap with chromatin QTLs (caQTLs) identified in lymphoblastoid cell lines (LCLs, Additional file 1: Fig. S3C). Further, alleles associated with increased chromatin accessibility are significantly more likely to be associated with increased gene expression ($OR = 2.9$, $p = 8.2 \times 10^{-37}$ Fisher's exact test) and slightly less likely to be associated with increased junction inclusion ($OR = 0.82$, $p = 0.03$ Fisher's exact test), suggesting that regulatory mechanisms altering chromatin accessibility play a greater role in regulation of gene expression than splicing. When we restrict to variants with a greater than 10% probability of being causal (Methods), we find a further enrichment in functional categories, particularly for caQTLs among eQTLs and splice regions among sQTLs, indicating we are capturing true causal variation (Additional file 1: Fig. S3C).

Of the genes tested, 198 have both an eQTL and sQTL in our cohort, suggesting possible shared genetic architecture between these transcriptomic traits. To evaluate whether eQTLs are enriched for sQTLs overall, we first compute the π_1 statistic, which measures the estimated fraction of sQTLs that are true positives in the eQTL scan. A π_1 value of 0.61 suggests that the majority of sQTLs affect expression or are in LD

with variants affecting expression (Additional file 1: Fig. S4), though many of these fail to reach genome-wide significance. To account for the possibility that our findings are related to technical artifacts of RNA-seq mapping across different transcript lengths, we measure π_1 across gene-length deciles. We find that smallest transcripts have the strongest replication overall, but all deciles show appreciable π_1 (min 0.25, max 0.79, Additional file 1: Fig. S5), suggesting our findings are robust to these artifacts. To further evaluate whether the genome-wide significant eQTL and sQTL signals are driven by shared causal variants, we estimated 90% credible sets for each set of QTLs, defined as the minimal set of variants which have at least a 90% probability of containing the causal variant, using the probabilities estimated above (Methods). Overall, we find overlapping credible sets for 114 of the genes with both a significant eQTL and sQTL, which makes up about 9% (114/1,330) of all eGenes in our cohort, comparable to the 12% overlap observed in GTEx [30]. Taken together, this observation suggests that splicing variants likely cause subtle but detectable changes in gene read counts, but that the genetic variants driving genome-wide significant eQTLs and sQTLs are largely independent.

Replication of tQTLs in non-Africans

To validate our tQTLs, and to assess sharing of molecular trait architecture between cohorts of predominantly African vs. predominantly European ancestry, we compared our tQTL results to whole blood eQTL and sQTL summary statistics from the Genotype-Tissue Expression project (GTEx) v8, which is comprised of 85% European Americans [30]. An advantage of using this dataset for replication is availability of both eQTL and sQTL summary statistics for the same RNA-seq samples, though the post-mortem nature of the samples is known to affect gene expression in whole blood [31]. For those QTLs tested in both cohorts, we find that both eQTLs and sQTLs identified in the African cohort show overall high reproducibility in GTEx, with π_1 values for eQTLs and sQTLs of 0.88 and 0.91, respectively (Additional file 1: Fig. S6, Methods). For eQTLs, we also found a high π_1 replication of 0.97 with results from the eQTLGen consortium, a meta-analysis of 37 blood expression datasets [32]. In addition to π_1 , effect sizes between our cohort and GTEx also show overall strong concordance (Pearson's $\rho = 0.73$ for eQTLs and 0.82 for sQTLs, Fig. 2B). To assess whether the observed replication is significantly affected by the different genome versions used between our study and GTEx v8, we also measured π_1 of eQTLs in GTEx v7, finding a π_1 of 0.83 (Additional file 1: Fig. S6).

While tQTLs as a whole show strong replication using π_1 , we also investigate the degree to which individual loci show evidence of shared causal variation. Estimating credible sets for all eGenes and sIntrons in GTEx v8 as described above, we find that 715/1262 (57%) of eGene credible sets and 619/852 (73%) of sIntron credible sets in Africans overlap with credible sets in GTEx v8. While the majority of tQTL credible sets overlap, the many non-overlapping sets suggest many tQTL signals identified in Africans may be driven by independent causal variants. To further evaluate this independence, we remapped tQTLs in Africans, conditioning on sets of independent tQTLs identified in GTEx by forward regression [30]. In cases where there are no genome-wide significant eQTLs or sQTLs in GTEx (169 genes and 541 introns, respectively), we instead condition on the lead eSNP or sSNP in GTEx. Using the

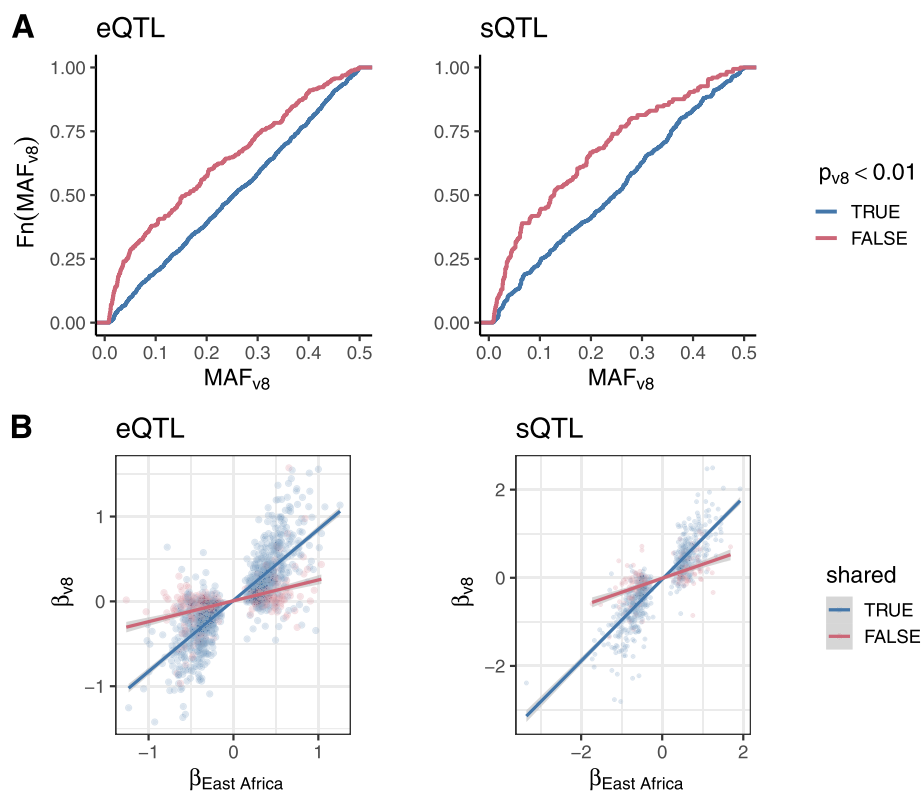


Fig. 2 Replication of tQTLs between East Africans and GTEx v8. **A** Minor allele frequency distribution in GTEx v8 of FDR-significant tQTLs identified in East Africans, colored by whether they have a p -value less than 0.01 in GTEx v8. **B** Comparison effect sizes of tQTLs identified in East Africans. Lines show the best fit regression line between East Africans and GTEx v8 effect sizes, colored by whether the tQTL is shared (i.e., is no longer significant after conditioning) or is independent (remains significant after conditioning)

original FDR significance thresholds for calling eQTLs and sQTLs, we find that 362 (27%) of eGenes and 224 (20%) of sIntrons remain significant after conditioning on GTEx SNPs, including the top variants for 328 eGenes and 199 sIntrons, suggesting widespread independent causal variation in Africa.

The source of replication differences between our cohort and GTEx may be due to several factors, including MAF differences, false positives, differences in LD between cohorts, differences in expression, and/or differences in effect size. Investigating those African tSNPs that fail to replicate in GTEx ($p > 0.01$), we find that non-replicating tSNPs have consistently lower MAF in GTEx when compared with replicating tSNPs (Fig. 2A). Of the 308 lead eSNPs that fail to replicate in GTEx, 60 have a p -value > 0.01 in the larger eQTLGen dataset, close to our specified false discovery rate of 5% ($60/1330 = 4.5\%$), suggesting we are not detecting an excess of false positives. Further restricting to those independent tQTLs identified above, we investigate whether MAF and LD differences can account for our findings. For eight genes, *INPP5K*, *TMEM140*, *ACSM3*, *CNTNAP3*, *PPP1R14C*, *PDZK1TP1*, *GPR56*, and *TRAM2*, the top eSNP in Africans is untested in GTEx and has a $MAF < 0.01$ (the threshold used by GTEx) in 1kGP EUR populations. The top eSNP for these eight genes are also non-significant or absent from the FIVE browser [33]. Similarly, the top sSNPs for introns in four

genes, *ADAM8*, *ICAM2*, *LINC00694*, and *MAPK1*, are absent in GTEx, absent or non-significant in the FIVEx browser, and have a EUR MAF ≤ 0.01 . Overall, however, we find that frequency differences between Africans and EUR do not differ significantly between shared and independent eQTLs ($p=0.49$, one-sided Kolmogorov–Smirnov (KS) test, Methods), while we do find a significant, though slight, enrichment for larger frequency differences among independent sQTLs ($p=5.74 \times 10^{-3}$, one-sided KS test). To investigate the impact of LD variation on tQTL replication, we estimate r^2 between tQTL lead SNPs and SNPs within 100 kb of lead SNPs in 1kGP CEU and YRI populations. We find that correlations between CEU and YRI r^2 values do not differ significantly between shared and independent tQTLs (Additional file 1: Fig. S8, $p=0.25$ for eQTLs, $p=0.43$, one-sided KS test). Finally, comparing effect size estimates between the African cohort and GTEx at top tSNPs, we find a significantly lower correlation of independent tQTLs when compared with shared signals (Fig. 2B, $p < 2.2 \times 10^{-16}$), which may reflect true effect size variation, GxE effects [34–36], or possibly more subtle differences in MAF and local LD between these cohorts [37].

Finally, we investigate whether expression differences may affect replication between cohorts. Of the 1330 eGenes identified in Africans, the expression of 98 in GTEx v8 whole blood is too low to be tested for eQTLs. These 98 genes are significantly enriched in two KEGG pathways, “Hypertrophic cardiomyopathy” (FDR=0.032) and “Dilated cardiomyopathy” (FDR=0.038). Investigating what may be driving broader expression differences for testable genes, we identify those genes measured in Africans that fail to reach expression thresholds for testing in GTEx whole blood and vice versa. Altogether, 951 out of 12,377 genes measured in both cohorts and tested for eQTLs in Africans were not tested in GTEx. These genes are enriched for a number of biological processes related to sensory perception, including perception of smell (FDR= 2.85×10^{-6}), sound (FDR= 1.60×10^{-5}), mechanical stimulus (FDR= 5.60×10^{-5}), and chemical stimulus (FDR= 5.22×10^{-4}). Similarly, 6728 out of 18,168 tested for eQTLs in GTEx were not tested in Africans and are enriched for several biological processes related to immunity, including “complement activation, classical pathway” (FDR= 1.78×10^{-22}), “humoral immune response mediated by circulating immunoglobulin” (FDR= 7.32×10^{-18}), and “B cell mediated immunity” (FDR= 2.02×10^{-2}). This observation suggests that disease status, sample collection, and response to environmental factors, in addition to genetics, may account in part for incongruent findings between eQTL cohorts.

Fine-mapping

In addition to assessing the replication of transcriptional QTLs in the larger GTEx v8 dataset, we are interested in the relative power to detect and fine-map tQTLs between cohorts of predominantly African versus European ancestry. To account for sample size differences between our cohort and GTEx, we performed eQTL mapping in a size-matched sample of 162 European American (EA) individuals from GTEx v8 using *FastQTL* [38], with sex, sequencing platform, PCR batch, the top 15 *PEER* factors, and top 5 genotype PCs as covariates. The number of *PEER* factors and genotype PCs was chosen based on prior GTEx analyses [30]. Testing all SNPs with MAF > 0.05 within 100 kb of the target TSS, we identify 1029 eGenes in the 162 EA individuals at FDR < 0.05, compared with 1330 identified in Africans, of which 326 eGenes are

FDR-significant in both cohorts. Despite only 326 eGenes being shared, we find consistently high replication in an independent whole blood meta-analysis [32]; eQTLs that are FDR-significant in both cohorts reach a π_1 of 0.999, while eQTLs discovered only in Africans reach a π_1 of 0.958 and eQTLs discovered only in EAs reach a π_1 of 0.989. This observation suggests that the greater number of eGenes discovered in Africans is not driven by an increase in false positives and that, at similar sample sizes, there is greater power to detect eQTLs in samples from African individuals when compared with samples from individuals of European ancestry.

We next investigate the relative ability to fine-map eQTLs between our African cohort and the 162 EA individuals from GTEx v8. Considering eGenes that are FDR-significant in either cohort (Methods), we perform fine-mapping in both our African cohort and the 162 EAs using the approach described above. Overall, most genes do not fine-map well at this modest sample size, with 57% of genes having a credible set larger than 50 in both cohorts (Fig. 3A). Excluding these genes, we find that Africans have a smaller credible set in 63% of cases (437/697, $p = 2.06 \times 10^{-11}$ binomial test), with a median credible set size of 25 in Africans vs 58 in EAs, and 23 genes fine-mapped to a single variant in Africans vs. 13 in EAs, demonstrating that using ethnically diverse populations facilitates fine-mapping, as has been shown previously [39]. One possible explanation of the smaller credible sets in Africans is that Africans simply have fewer SNPs tested per gene; however, we find the opposite, with 94% of genes have fewer tested SNPs in EAs.

We further compare our credible sets in African eQTLs to credible sets estimated in the full GTEx dataset. As expected, the majority of eGenes have smaller credible sets in GTEx due to the considerably larger sample size (670 vs 162), though we do identify several examples of greatly reduced credible sets in the African cohort. For 18 eGenes and 32 sGenes, we are able to fine-map the QTL signals to a single variant in Africans and find that these variants overlap a lead GWAS association for 10 eGenes and 3 sGenes (supplement). We highlight rs883871 (Fig. 3B), an eQTL for both *THRA* and *NR1B1*, which is FDR-significant in GTEx whole blood but is not the lead eSNP. SNP rs883871 is a strong chromatin QTL in lymphoblastoid cell lines (LCLs) [44], overlaps the binding sites of numerous transcription factors (TFs) in the LCL GM12787 [43], is predicted to disrupt a consensus motif for the ETS family of TFs, which share a core “CCGGAA” motif, and is the lead SNP for a Multiple Sclerosis GWAS association [45]; variants in *ETS1* itself have been previously associated with multiple sclerosis [46].

Signatures of selection

Gene regulation is known or suspected to underlie many adaptive traits in humans, including diet [17, 47], immunity [48], and skin pigmentation [10], and transcriptomic traits show evidence of both purifying and positive selection [35, 36, 49]. Consistent with previous tQTL studies, we find decreasing effect size with increasing MAF among eQTLs and sQTLs, indicative of negative selection against variants of large effects (Additional file 1: Fig. S9). To identify QTLs with evidence of positive selection, we measure genome-wide F_{ST} between our broader African dataset and the 1kGP European (EUR) individuals, with the expectation that selection for expression-altering alleles will lead to increased differentiation at these loci. To assess whether tQTLs are enriched for evidence of positive selection, we identify the highest F_{ST} value for all SNPs in high LD

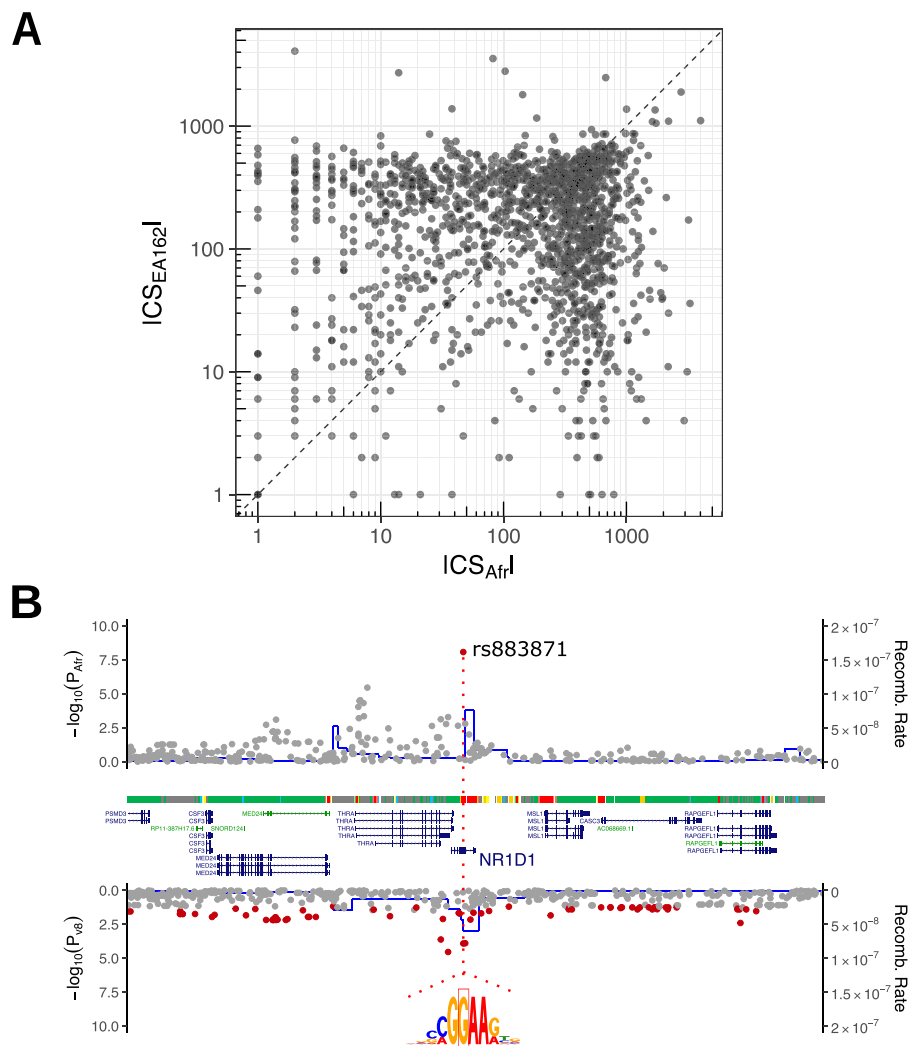


Fig. 3 Fine-mapping in East Africans vs. GTEx v8. **A** Credible set (CS) sizes for eGenes identified in 162 East Africans (Afr) or a subset of 162 European Americans from GTEx v8 (EA162). **B** Locus plot of *NR1D1* eQTLs identified in 162 East Africans (Afr) or the full GTEx v8 cohort (v8). *P*-values are overlaid with African (YRI) and European American (CEU) recombination rates, GENCODE v19 [40] gene models from the UCSC genome browser [41] (<http://genome.ucsc.edu>) and inferred ChromHMM [42] states for GM12878 [43]. The top SNP in Africans, rs883871, disrupts a nucleotide for the core motif of ETS family transcription factors (motif of *ETS1* shown)

($r^2 > 0.8$) with the top eQTL or sQTL and compare these values with null SNPs matched on MAF and the number of SNPs in LD (Methods). Overall, we do not find an enrichment of high F_{ST} among eQTLs or sQTLs, either when combining all populations or testing populations individually, suggesting that selection has not driven significant frequency differentiation at the majority of tQTLs (Additional file 1: Fig. S10 and S11).

We next investigate evidence of selection at individual loci. To account for the fact that the top eSNP may not be the true causal SNP, we score an individual gene's evidence of selection by taking a weighted sum of each SNP's F_{ST} value multiplied by the probability of that SNP being causal. We also perform repeated permutations between F_{ST} and causal probabilities at a locus to generate a locus-specific background expectation (Methods). Considering as candidates loci with a score within the 99th percentile

threshold of all SNP F_{ST} values, and greater than 99% of background values, we identify 23 eGenes and 20 sGenes with evidence of selection (supplement). The most differentiated eGene is *TTC26* (weighted $F_{ST}=0.59$); a mutation in this gene has been associated with abnormal cilia in model organisms and biliary ciliopathy in human liver [50]. We also identified a strong signature of selection at *ARPC1B* (weighted $F_{ST}=0.59$), deficiency of which can result in severe immunodeficiency [51]. Other highly differentiated loci include Platelet Factor 4 Variant 1 (*PF4V1*, $F_{ST}=0.50$), *IL8* ($F_{ST}=0.49$), a major inducer of immune cell chemotaxis and activation [52], and *CCR1* ($F_{ST}=0.43$), a chemokine receptor. Among the most differentiated sGenes, we find several related to immunity and metabolism, including *NADSYN1* (weighted $F_{ST}=0.50$), a gene associated with vitamin D concentration [53], *BTN3A3* (weighted $F_{ST}=0.50$), a butyrophilin gene implicated in activation of T cells [54], and *GANC* (weighted $F_{ST}=0.43$), a member of the glycosyl hydrolase family 31, which play a key role in glycogen metabolism [55].

Given our genetically and culturally diverse cohort, we are also interested in tQTLs with evidence of population-specific differentiation and selection. For each of the nine populations in the African dataset, we calculate a modified version of the d -statistic [56], a summation of normalized, pairwise F_{ST} , which tests for variants that are highly differentiated in a focal population versus other populations (Methods). As above, we weight these d -statistics by the probability of a SNP being causal to derive a “ d -score” for each gene or intron. Genes with high d -scores in populations with evidence of non-African admixture (i.e., Agaw, Amhara, Argoba, and Weyto) are more genetically similar to EUR samples from the 1kGP, based on F_{ST} . Conversely, populations with evidence of West African admixture (i.e., the Hadza, Mursi, and Sandawe) are more genetically similar to YRI samples at high d -score genes, suggesting that in many cases the genetic differentiation at these loci is driven by population-specific patterns of admixture. We therefore calculate the population branch statistic (PBS) [57] between individual populations in our study and 1kGP CEU and YRI populations. Considering genes with a weighted d and PBS score in the top 99.5th percentile as significant, we identify 22 eGenes and 22 sGenes with significant evidence of population-specific selection (Fig. 4A, B).

Among the top eGenes with evidence of population-specific selection is *TMEM216* among the Nilo-Saharan speaking Mursi pastoralists (Fig. 4A). This gene is located near a skin pigmentation GWAS locus discovered in a cohort with the same sub-Saharan African populations [10]. This association signal overlaps the UV-repair gene *DDI1*, as well as several other genes expressed in melanocytes. Colocalization analyses show strong overlap between the African *TMEM216* eQTL and pigmentation GWAS signals (PP4 = 0.95, Fig. 5, Methods), suggesting possible shared causal variation between *TMEM216* expression and pigmentation variation. LD patterns around *TMEM216* shows evidence of three independent eQTLs segregating for this gene, tagged by rs7948623, rs11230664, and rs3741265. Two of these SNPs, rs7948623, rs11230664, are also genome-wide significant GWAS SNPs for pigmentation variation in Africans, while the third, rs3741265, is marginally significant ($p < 10^{-5}$, Fig. 5). All three SNPs show strong population-specific differentiation in Ethiopian Nilo-Saharan groups, who have amongst the highest levels of skin melanin of any global population (Additional file 1: Fig. S12). Previous analyses of these populations have shown

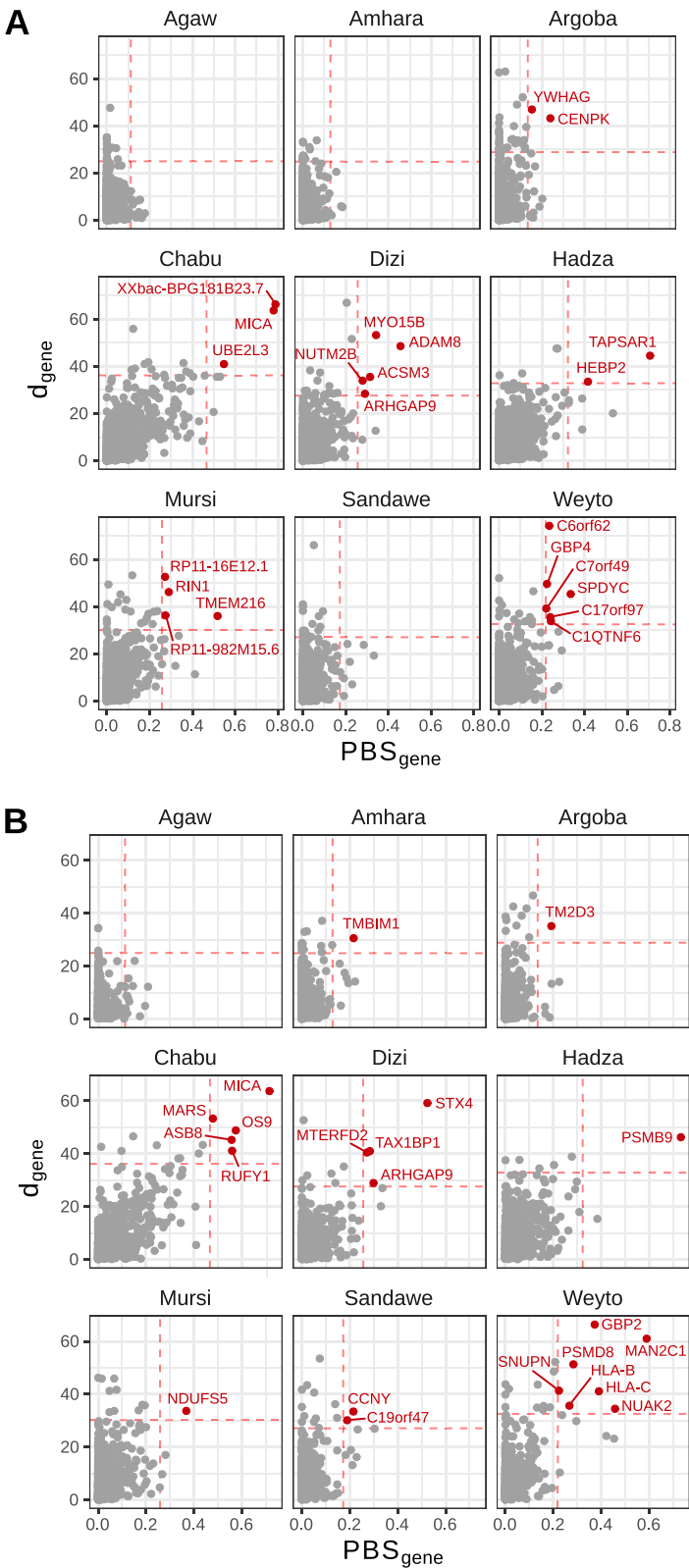


Fig. 4 Population-specific selection in East Africa. Gene scores for the d -statistics plotted against the population branch statistics (PBS) for each population. PBS is calculated for each focal population versus the CEU and YRI populations from the 1000 Genomes Project. Genes with a score above the 99.5th percentile of genome-wide statistics for d and PBS are highlighted in red for eGenes (**A**) and sGenes (**B**)

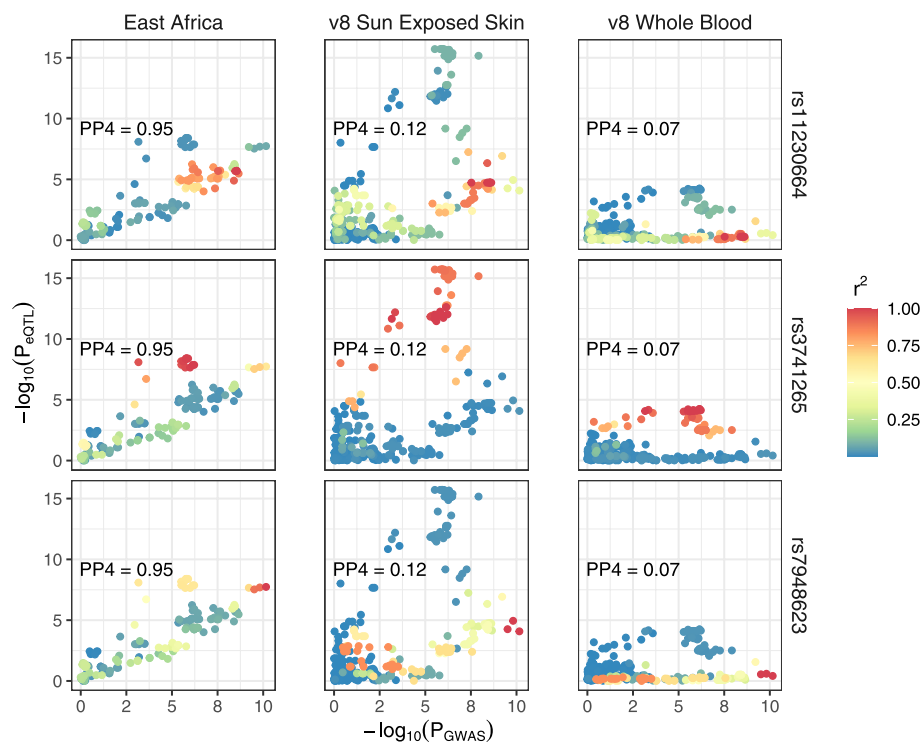


Fig. 5 Colocalization of whole blood eQTLs and pigmentation GWAS. Comparison of pigmentation GWAS p -values from Crawford et al. [10] against eQTL p -values from our study (East Africa), GTEx v8 Whole Blood, or GTEx v8 Sun-exposed skin (lower leg), in the style of LocusCompare [58]. Variants are colored by their degree of LD with three top pigmentation GWAS variants, rs7948623, rs11230664, and rs2512809. Colocalization probabilities from *coloc* [59] (PP4) are indicated for each eQTL group

evidence of a selective sweep near this pigmentation GWAS locus, including high PBS and d values among GWAS variants (Additional file 1: Fig. S14) and extreme negative Tajima's D values overlapping the *TMEM138/TMEM216* locus [10].

The top GWAS variant, rs7948623, overlaps an active enhancer in keratinocytes and melanocytes and has been demonstrated to alter enhancer activity in melanocytes via luciferase reporter assays [10]. SNPS rs7948623 is a significant eQTL for *TMEM216* in our study, as well as an eQTL mapping study of stimulated monocytes from European and African ancestry individuals [35], but is not significant in GTEx whole blood, though it has been identified in ovary, nerve, and exposed skin. In addition, rs7948623 is a significant sQTL for *TMEM216* in multiple GTEx tissues, including exposed skin (Fig. 5). A second group of *TMEM216* eQTL and pigmentation GWAS variants are tagged by rs11230664 and include the indel rs148172827, which overlaps an active melanocyte enhancer and shows significant correlation with *TMEM216* expression in GTEx exposed skin (Fig. 5). We do not identify significant sQTLs in Africans for *TMEM216*; however, the top sSNP for *TMEM216* in GTEx exposed skin, rs3741265 ($p = 1.43 \times 10^{-322}$), is in high LD with the top *TMEM16* eQTL in Africans, rs7934229 ($r^2 = 0.99$). Both of these SNPs are moderately associated with skin pigmentation in Africans ($p < 5 \times 10^{-6}$) but do not reach genome-wide significance (Additional file 1: Fig. S14).

Discussion

This study extends our understanding of the genetic basis of human gene regulation, with the inclusion of whole blood samples for 162 ethnically diverse sub-Saharan Africans from Ethiopia and Tanzania. We find that variation underlying expression and splicing is broadly shared between African and European cohorts, though there is considerable independent variation at individual loci in Africans, often driven by variation in frequency and effect sizes of tQTLs. When matched for sample size, Africans show improved fine-mapping of molecular traits, facilitating the identification of causal variants and candidate genes underlying GWAS traits. This diverse cohort also allows for inference of tQTLs with evidence of local adaptation, identifying *TMEM216* as a target of selection in Nilo-Saharan speakers and a candidate gene that may play a role in skin pigmentation.

We find that the majority of tQTLs replicate between Africans and GTEx v8, with π_1 values near 0.9 among both eQTLs and sQTLs, on par with the 0.919 value estimated between African Americans in the GENOA cohort [60] and EUR populations from the Geuvadis project [13]. We also observe strong effect size correlation between tQTLs in our study and GTEx v8. Investigating individual loci, however, we find that many genome-wide signals are driven by distinct causal variation; 43% of eQTL and 27% of sQTL credible sets in Africans do not overlap those in GTEx v8, and 27% of eGenes and 20% of sIntrons have QTL signals that remain significant after conditioning on all tQTLs in GTEx.

Investigating what may account for QTL differences between Africans and non-Africans, we find that genes relating to sensory perception and immunity show differential expression between our African cohort and the GTEx cohorts, pathways known to vary across populations and environments [11, 61], though the explained variability is generally low. Additionally, the post-mortem nature of GTEx samples may contribute to expression differences. An analysis of the effects of death on gene expression in GTEx found that immune genes in whole blood are significantly dysregulated following death; however, this change was characterized by an overall deactivation of immune genes, along with an overall increase in NK cells and CD8 T cells and a reduction in neutrophils [31]. In addition to expression differences, we find an enrichment for low-frequency variants in GTEx among non-replicating tQTLs. However, the majority of tQTLs that are conditionally independent show similar frequency differences with shared tQTLs, suggesting that frequency variation alone cannot account for independent tQTLs. This issue of trans-ethnic GWAS replication is an ongoing area of research [62, 63], and non-replication may occur for many reasons including frequency variation, differences in power, LD, or true differences in effect size, including $G \times E$ effects. While we do not find a significant difference in local LD structure between shared and independent QTL signals, we do find significant differences in estimated effect sizes. Using a Bayesian approach to account for frequency and LD variation, Brown et al. also found eQTL effect size differences between EUR and YRI individuals from Geuvadis [13], which become more pronounced as genetic effects become weaker [64]. However, for strong, genome-wide significant effects, Zanetti and Weale demonstrated using simulations that most trans-ethnic differences in GWAS effect sizes can largely be accounted for by a combination of frequency and LD variation, though they could not rule out effect size differences

[37]. More recently, Patel et al. leveraged local ancestry information to infer differences in causal effect sizes between variants on European and African ancestry-derived haplotypes, finding a significant effect of haplotype background on variant effect size [65].

Beyond replication, we demonstrate that at comparable sample sizes, African cohorts have improved sensitivity to detect tQTLs and improved ability to fine-map causal variants, compared with cohorts of European ancestry. It is well established that non-African populations have more extensive LD relative to Africans [66, 67], resulting from the out-of-Africa bottleneck [68, 69], and that multi-population analyses can improve causal variant detection [39], which likely account for the observed improvement in fine-mapping in African populations. As to the increased sensitivity to detect tQTLs in Africans, one hypothesis is a higher false-positive rate in the African cohort. However, we find comparable replication of African-specific tQTLs in a large, independent meta-analysis [32], suggesting that false positives do not account for the observed improvement. Moreover, Quach et al. found a similar pattern of improved sensitivity to detect eQTLs in individuals of self-reported African ancestry in an analysis of stimulated and unstimulated monocytes from 200 Belgians, 100 of European and 100 of African ancestry [35]. Among African Belgians, they found 13% more eQTLs in unstimulated monocytes, and 10% more eQTLs across all conditions. While several other studies have mapped eQTLs across multiple ancestry groups [12, 13, 36, 70], variation in sample size precludes direct comparison of sensitivities across ethnicities.

In addition to the inclusion in our study of ancestral groups not represented in existing reference cohorts (e.g., the 1kGP), which enables the detection of novel regulatory variation, these populations live in diverse climates and have distinct cultural and subsistence practices, which may have driven unique local adaptations. Using an outlier approach based on the F_{ST} based d and PBS statistics [56, 57], we identify population-specific differentiation of tQTLs among East African populations. One notable example is the eQTL *TMEM216* among the Mursi, which is near a recently identified pigmentation locus specific to sub-Saharan Africans [10]. *TMEM216*, and the nearby *TMEM138* gene, form an evolutionarily conserved *cis*-regulatory module vital for ciliogenesis and have been identified as causal genes underlying Joubert and Merkel syndromes [71, 72]. *TMEM216* has not been previously associated with pigmentation variation, though activation and suppression of primary cilia have been shown to inhibit and activate melanogenesis, respectively, in a human skin model [73]. Consistent with this, we find that the expression decreasing allele is associated with increased melanin levels for rs7948623, rs11230664, and rs3741265 and is most common in the Mursi, a population with darkly pigmented skin (Additional file 1: Fig. S12) [10]. In addition, recurrent somatic mutations driving alternative splicing of *TMEM216* are significantly associated with melanoma in The Cancer Genome Atlas (TCGA), suggesting possible tumor suppressor function for this gene [74]. While the strong colocalization between the *TMEM216* eQTL and pigmentation GWAS signals suggests *TMEM216* as a possible pigmentation gene, there are several haplotypes segregating in this region, some of which carry tQTLs for other genes in GTEx (Additional file 1: Fig. S16 and S17). In addition, several nearby genes show melanocyte-specific expression or have been previously associated with pigmentation in other organisms, complicating identification of the gene or genes that are causally associated with pigmentation variation [10, 75].

There are several limitations to our study, foremost being our modest sample size of 162 individuals, with current eQTL datasets reaching sample sizes an order of magnitude larger [60]. Many of the populations participating in this study live at considerable distances from medical or scientific facilities, and all necessary tools and supplies must be transported to field sites, greatly limiting the capacity for sample collection. Additionally, we are limited to studying blood tissues among these populations. Generation of induced pluripotent stem cells (iPSC) may allow for the study of gene regulation across developing tissues or differentiated cells within diverse populations [76, 77], but such approaches remain technically difficult. This study is also restricted to steady-state gene expression, which may miss cell-type- or dynamic, environment-specific genetic effects, which cannot be captured in bulk and/or steady-state tissues [34–36, 78–80]. Despite these limitations, this study makes important contributions to our understanding of gene expression variation and the molecular basis of human adaptation in sub-Saharan Africa.

Conclusion

We have presented a comprehensive analysis of transcriptomic variation in a cohort of previously unstudied indigenous sub-Saharan Africans. We identify extensive novel regulatory variation in Africans and show that the study of African populations improves the detection of transcriptomic QTLs and fine-mapping of causal variation. Studying diverse populations within Africa also allows for the detection of genes targeted by population-specific selection, including evidence of selection on *TMEM216* expression in the Mursi and strong colocalization between *TMEM216* eQTLs and a pigmentation GWAS locus.

Methods

Sample collection

Phenotypic, genealogical, and biological data were collected from individuals belonging to nine populations in Ethiopia and Tanzania. Prior to sample collection, IRB approval for this project was obtained from the University of Pennsylvania. Written informed consent was obtained from all participants and research/ethics approval and permits were obtained from the following institutions prior to sample collection: the University of Addis Ababa and the Federal Democratic Republic of Ethiopia Ministry of Science and Technology National Health Research Ethics Review Committee; COSTECH, NIMR, and Muhimbili University of Health and Allied Sciences in Dar es Salaam, Tanzania. To obtain DNA and RNA data, whole blood was collected using vacutainers and RNA was stabilized in the field using LeukoLOCK Total RNA Isolation System (Ambion life Technologies). The Poly(A)Purist Kit (Ambion Life Technologies, CA) was used for mRNA selection, and Ampure XP magnetic beads (Beckman Coulter, CA) were used for size selection after amplification.

Genotyping and imputation

A subset 162 individuals were genotyped as part of the 5 M dataset using the whole genome Illumina Omni5 Beadchip array, which includes approximately 4.5 million

SNPs. The full 5 M dataset was phased using Beagle 4.0 [81] and the 1kGP reference panel [20]. These data were further imputed using minimac3 [82] and a reference panel consisting of the 1kGP and 180 WGS from the Tishkoff lab [21]. The 180 WGS data include 15 individuals from each of the following populations used in our study: Amhara, Dizi, Hadza, Mursi, Chabu, and Sandawe.

PCA and ADMIXTURE

To identify related individuals, relatedness was inferred in the imputed 5 M dataset using the KING extension of plink 2.0 [83]. To place the genetic variation in this study within a global context, the 5 M imputed dataset was merged with the 1KGP. Individuals from the 162 in this study with inferred relatedness more distant than third degree were then extracted from the merged dataset (145 total), along with 20 individuals each from the YRI, CEU, and CHB populations, restricting to unambiguous SNPs (i.e., excluding A/T and C/G) with MAF > 0.01 and with imputation accuracy (r^2) greater than 0.99 reported from minimac3. SNPs were LD-pruned using plink v1.90 [84] and parameters “-indep-pairwise 50 10 0.1.” PCA was performed on this dataset using smartpca from EIGENSOFT v6.1.4 [85], with “numoutlieriter” set to 0. ADMIXTURE [86] was run on the same dataset using parameters “-cv -j8 -B100 -s7.”

mRNA sequencing and molecular trait quantification

Samples were sequenced on an Illumina HiSeq to a median depth of 56,122,076 reads (11,727,716 min., 228,660,534 max.). Prior to mapping, all reads aligned to rRNA genes with BLAST [87] were removed. Remaining reads were mapped to the hg19 genome with STAR v2.5.3a [88] and the GTEx GENCODE v19 gene annotations [40] using two-pass mapping. Expression was quantified at the gene level using feature-Counts v1.5.3 [89] as fragments per gene, as well as using RSEM v1.2.31 [90] as transcripts per million (TPM). Splicing was quantified using leafcutter [91] as fraction of intron exclusion reads per cluster (JPC).

Cell-type inference

Cell-type fractions for each individual were inferred using CIBERSORT [26]. The LM22 signature gene file from Abbas et al. [92] was used to infer frequencies of 22 immune cell types for a mixture file of TPM values for all 171 individuals with RNA-seq data. Quantile normalization was disabled, and 1000 permutations were used.

Quantile normalization and hidden factor inference

Prior to hidden factor inference and QTL mapping, molecular phenotype matrices were first filtered and quantile-normalized. For eQTL mapping, only lncRNA and protein-coding genes with more than 5 reads in at least 20 individuals and with mean TPM > 0.1 across all populations were considered. For sQTL mapping, introns from lncRNA and protein-coding genes with no more than 5 individuals with 0 reads were included. Furthermore, clusters were required to have at least 20 reads in at least 100 individuals and have 0 reads in fewer than 10 individuals. These filtered phenotype matrices (TPM for eQTL mapping and JPC for sQTL) were then quantile-normalized

using the two-stage procedure implemented by GTEx [30]. Briefly, the distribution of the phenotypes per individual were first quantile-normalized to the mean of the phenotypes across individuals. Next, the distribution of each phenotype was quantile-normalized to the standard normal. Hidden covariates were inferred using *PEER* [25] for these quantile-normalized phenotype matrices.

eQTL and sQTL mapping

Expression and splicing quantitative trait loci were mapped using a linear mixed modelling approach, using the quantile-normalized gene or intron fractions as phenotypes, while correcting for sex, age, cell-type composition, delivery date, latent *PEER* factors, and genetic relatedness. Mapping was performed for SNPs with MAF > 0.05, imputation $r^2 > 0.3$, and within 100 kb of the target phenotype (gene TSS for eQTLs and intron for sQTLs) using *GEMMA* [28] and a genetic relatedness matrix (GRM) generated from all biallelic SNPs across the imputed, 162 individual genotype dataset. tQTL mapping was repeated across a range of *PEER* factors: 0–5, 10, 15, 20, 25, and 30 factors for eQTL mapping, and 0–10 factors for sQTL mapping, and the number of factors maximizing the number of eQTLs or sQTLs discovered were chosen for downstream analysis.

To identify significant QTLs, tested SNPs for each phenotype were first FDR-corrected using Benjamini-Hochberg (BH), yielding single-corrected p -values (P') for each tested SNP-phenotype pair. The minimum P' per phenotype were again FDR-corrected using BH, yielding double-corrected p -values (P'') per phenotype, and phenotypes with $P'' < 0.05$ were considered significant. To identify significant SNPs, a threshold was set equal to the lowest P' for the phenotype with highest significant P'' , and all SNPs with P' lower than this threshold were deemed significant.

Credible sets

For each gene or intron of interest, Approximate Bayes Factors were calculated for each tested SNP using the function “approx.bf.estimates” from the coloc package [59], or the function “approx.bf.p” in cases where effect size or standard error information was not available. The posterior probability of each SNP n being causal (PP_n) was then taken as:

$$PP_n = \frac{ABF_n}{\sum_p ABF_p}$$

Similar to The Wellcome Trust Case Control Consortium et al. [93], where ABF_n is the Approximate Bayes Factor of SNP n , and p indexes all tested SNPs for a given feature of interest. A 90% credible set was then defined as the minimal number of SNPs whose sum of posterior probabilities was > 0.9.

Functional enrichment

All SNPs in the imputed genotype dataset of 162 individuals were annotated for functional consequences using the Variant Effect Predictor (VEP) [94] with parameters “–per_gene –most_severe.” In addition, SNPs were overlapped with 15 state ChromHMM tracks for PBMCs (E062) from the Roadmap Epigenomics Consortium [75], transcription factor binding sites for lymphoblastoid cell lines (LCLs, GM12878) from ENCODE [43], and chromatin QTLs from Tehranchi et al. [44]. To test for enrichment, each

FDR-significant eQTL or sQTL was matched on MAF and distance to nearest TSS or intron boundary, respectively, and the log ratio of tQTL SNPs to matched background SNPs overlapping each functional category was taken as an enrichment score. This was repeated 10,000 times, producing an empirical distribution of enrichment scores for each functional category.

Replication with GTEx v8

All SNPs and intron boundaries were converted to hg38 coordinates using liftOver [95]. For eQTLs, those hg19 SNPs that successfully mapped to locations in hg38 (81,928/82,144) and genes with Ensembl IDs shared between GENCODE v19 and GENCODE v26 (1291/1330) were considered (96,903/99,685 of possible eQTLs). Of these, 77,238 eQTLs were tested in GTEx v8 and could be compared. For sQTLs, SNPs and Ensembl IDs were required to successfully map between versions (49,706/49,794 and 772/776, respectively), and intron boundaries were required to map between GENCODE versions (738/1118). Of these, 55,046 sQTLs were tested in GTEx. The fraction of true positives for successfully mapped tQTLs in GTEx, π_1 , was estimated using the R package *qvalue* [96].

Conditional tQTL mapping

To identify tQTLs in the African cohort that are independent of GTEx v8 tQTLs, we performed eQTL and sQTL scans conditioning on independent GTEx eQTLs and sQTLs identified via step-wise regression [30]. In cases where there are no significant tQTLs in GTEx, we instead use the top variant per feature. To account for these variants, we residualize the quantile-normalized feature matrices used in the original QTL mapping against the genotypes of independent GTEx QTLs. We then perform identical eQTL and sQTL scans and consider genes and introns with variants that pass the original FDR threshold as independent.

LD variation across populations

To compare LD structure between Africans and Europeans at tQTL loci, LD was estimated (using r^2) between lead SNPs for eQTLs and sQTLs and all tested SNPs in the YRI and CEU 1kGP samples, restricting to those variants polymorphic in both, resulting in an r^2 vector per group (YRI and CEU) per locus (eGenes and sIntrons). For each tQTL locus, we estimated the Pearson correlation ρ between the YRI and CEU r^2 vectors, and the distribution of these ρ values was compared for tQTLs shared between East Africans and GTEx and independent tQTLs.

Testing differences in allele frequency and LD between shared and independent tQTLs

To test whether independent tQTLs show greater allele frequency differences between Africans and EUR samples compared with shared tQTLs, we perform a one-sided Kolmogorov–Smirnov (KS) test, with the alternative hypothesis being that the absolute frequency difference for independent tQTLs is right skewed (i.e. has an enrichment of large frequency differences) compared with the distribution of shared tQTLs. Similarly, to test whether independent tQTLs show weaker LD-structure correlation between African

and Europeans compared with shared tQTLs, we perform a one-sided KS test using the ρ values calculated above, with the alternative hypothesis being that independent tQTLs are left skewed (i.e., has an enrichment of low ρ values) compared with shared tQTLs.

eQTL mapping in 162 European Americans from GTEx v8

eQTL mapping was performed on 162 individuals of European ancestry from GTEx v8 using FastQTL [38] with 10,000 permutations for all SNPs with MAF > 0.05 and within 100 kb of the target TSS. Covariates included the top 15 *PEER* factors, top 5 genotype PCs, sex, platform, and PCR batch. Significance was evaluated using the hierarchical Benjamini–Hochberg procedure used for African samples.

Scans of natural selection

To test for genetic differentiation between our African dataset and Europeans, all individuals belonging to the 9 populations in our study were extracted from the full 5 M dataset (664 total) and allele frequencies were combined with frequency information for EUR populations from the 1KGP, restricting to SNPs polymorphic in both datasets. F_{ST} was estimated using the Hudson estimator [97], and SNPs within the top 99th percentile ($F_{ST} > 0.36$) were considered outliers. To test for overall enrichment of F_{ST} outliers among tQTLs, we use an approach similar to that of Quach et al. [35]. The maximum F_{ST} value of SNPs in LD with lead tQTL SNPs ($r^2 > 0.8$) was found, and the fraction of outliers among these maximum F_{ST} values was calculated. To generate a null expectation, each lead tSNP was matched with a random SNP, matching on MAF (bins of 0.05) and number of SNPs in LD (bins of [0], [1, 2], (2,5], (5,10], (10,20], (20,50], and > 50). The maximum F_{ST} of SNPs in LD with these matched SNPs was found, and the fraction of outliers among these matched maximum F_{ST} SNPs calculated. This procedure was repeated 10,000 times, generating a null distribution of expected number of outlier SNPs.

To identify individual eGenes and sGenes with evidence of selection, weighted F_{ST} scores were generated for each eGene and sIntron. For each feature of interest (gene or intron), the posterior probability of each tested SNP was calculated using the approach used to define credible sets, and for each feature, a weighted F_{ST} score was calculated as:

$$\overline{F_{ST}} = \sum_p PP_p F_{ST}^p$$

where PP_p is the posterior probability of SNP p being causal and F_{ST}^p is the F_{ST} of SNP p . Scores higher than the 99th percentile of genome-wide F_{ST} values were considered as candidate genes under selection. To further account for background selection, or other factors that may alter the F_{ST} of the surrounding region, we permuted F_{ST} across variants for each candidate and re-calculated the F_{ST} score. This was repeated 10,000 times to generate a null expectation of weighted F_{ST} scores for each candidate gene, which we compared observed scores against to generate empirical p -values. Those candidate genes with an empirical p -value less than 0.01 were considered significant.

To detect population-specific selection, we use an adapted, polarized version of the d -statistic for each SNP:

$$d_i = \left| \sum_{j \neq i} \mathbf{I}_{p_i \geq p_j} \frac{F_{ST}^{ij} - E[F_{ST}^{ij}]}{sd[F_{ST}^{ij}]} \right|$$

where p_i and p_j are the allele frequencies in populations i and j , respectively, $\mathbf{I}_{p_i \geq p_j}$ is an indicator function that returns 1 if $p_i \geq p_j$ and -1 otherwise, F_{ST}^{ij} is the F_{ST} between focal population i and population j , and $E[F_{ST}^{ij}]$ and $sd[F_{ST}^{ij}]$ are the expected value and standard deviation of F_{ST} between populations i and j across all SNPs. We implement this polarization procedure because SNP frequencies that are at an intermediate frequency in the focal population, but strongly differentiated in others, can show up as strong d_i outliers in the focal population due to the symmetry of F_{ST} . To identify individual eGenes and sGenes with evidence of population-specific selection, we generate weighted d_i scores as described above for F_{ST} .

Due to differential levels of admixture across populations, some d_i outlier loci show genetic similarity with non-African and West African populations, suggesting that these loci are uniquely differentiated in the focal population due to admixture. To eliminate candidates that may be driven by admixture, we also calculate the population branch statistic (PBS_i) [57] between each focal population i and the CEU (a proxy for non-Africans) and the YRI (a proxy for sub-Saharan Africans):

$$PBS_i = \frac{T^{i,YRI} + T^{i,CEU} - T^{YRI,CEU}}{2}$$

where $T^{A,B} = -\log(1 - F_{ST}^{A,B})$ and $F_{ST}^{A,B}$ is F_{ST} calculated between populations A and B . We then go on to create a weighted PBS_i statistic per gene or intron as above. Candidates of selection are then defined as those features with a weighted d_i and PBS_i score above the 99.5th percentile of genome-wide d_i and PBS_i SNP-wise statistics.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-023-02874-4>.

Additional file 1: Fig S1. Principal Component Analysis of East African and 1000 Genome Project. **Fig S2.** ADMIXTURE analysis across K values 2–12. **Fig S3.** Genomic context of tQTLs. **Fig S4.** π_1 of eQTL p-values of SNP-gene pairs ascertained as sQTLs. **Fig S5.** π_1 between sQTLs and eQTLs across gene length deciles. **Fig S6.** π_1 value of ascertained eQTLs and sQTLs in GTEx. **Fig S7.** Mapping statistics from STAR. **Fig S8.** Frequency and LD differences between African samples and 1000 Genomes EUR populations. **Fig S9.** tQTL effect size vs MAF. **Fig S10.** Fraction of F_{ST} outliers among eQTLs and sQTLs compared with matched background. **Fig S11.** Population-specific F_{ST} outliers. **Fig S12.** Global frequencies of SNPs associated with pigmentation variation and TMEM216 expression and splicing. **Fig S13.** Colocalization of Mursi PBS and d-statistics with TMEM216 eQTLs. **Fig S14.** Colocalization of Mursi PBS and d-statistics with pigmentation GWAS. **Fig S15.** eQTL associations for TMEM216 across populations. **Fig S16.** 'LocusCompare' plots of African pigmentation GWAS and GTEx v8 eQTLs. **Fig S17.** 'LocusCompare' plots of African pigmentation GWAS and GTEx v8 sQTLs.

Additional file 2: Table S1. Genes with weighted East African-EUR eQTL F_{ST} scores greater than 99% of genome-wide SNPs. **Table S2.** Spliced introns and related genes with weighted East African-EUR sQTL F_{ST} scores greater than 99% of genome-wide SNPs. **Table S3.** Genes with weighted East African d-statistic and PBS scores greater than 99.5% of genome-wide SNPs. **Table S4.** Introns and related genes with weighted East African d-statistic and PBS scores greater than 99.5% of genome-wide SNPs.

Additional file 3. Review history.

Acknowledgements

We would like to thank all of the study participants who make this work possible, along with our funding sources. We would also like to thank Dr. Nicholas Lahens and the ITMAT Bioinformatics Group for their assistance in data processing.

Review history

The review history is available as Additional file 3.

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

SAT conceived and supervised the study. TBN, SAO, DWM, GB, WB, JBH, and AR collected and processed samples. MY and SC performed SNP genotyping. CDB, GRG, RAR, RM, and HL assisted in statistical and bioinformatic analysis. SR performed eQTL mapping of European Americans from GTEx. DEK performed all other analyses. DEK and SAT wrote the manuscript with help from other co-authors. All authors read and approved the final manuscript.

Funding

This work was supported by the grant numbers: ADA 1–19-VSN-02, and NIH grants 1R35GM134957, R01DK104339, and R01AR076241 to SAT. Training of DEK was further supported by NIH grant T32AI007532.

Availability of data and materials

Gene expression and covariate data for this study is available through dbGAP Study Accession phs002824.v1.p1 [98]. Genotype data for individuals in this study is available through dbGAP Study Accession phs001396.v1.p1 [99]. Code to replicate the main results of this manuscript is available in a public repository at https://github.com/derkelly/afr_eqtl [100] and archived on Zenodo at [101], and is licensed under the GNU General Public License v3.0.

Declarations**Ethics approval and consent to participate**

Written informed consent was obtained from all participants. IRB approval for this project was obtained from the University of Pennsylvania, and research/ethics approval and permits were obtained from the following institutions prior to sample collection: the University of Addis Ababa and the Federal Democratic Republic of Ethiopia Ministry of Science and Technology National Health Research Ethics Review Committee; COSTECH, NIMR and Muhimbili University of Health and Allied Sciences in Dar es Salaam, Tanzania.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Genomics and Computational Biology, University of Pennsylvania, Philadelphia, PA, USA. ²Genetics, University of Pennsylvania, Philadelphia, PA, USA. ³Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, PA, USA. ⁴Department of Medicine, Division of Genetic Medicine, Vanderbilt University School of Medicine, Nashville, TN, USA. ⁵Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, USA. ⁶Frederick National Laboratory for Cancer Research, Frederick, MD, USA. ⁷Division of Cancer Epidemiology and Genetics, National Institutes of Health, Rockville, MD, USA. ⁸Department of Biochemistry, Kampala International University in Tanzania, Dar Es Salaam, Tanzania. ⁹Center for Biotechnology Research and Development, Kenya Medical Research Institute, Nairobi, Kenya. ¹⁰Microbial Cellular and Molecular Biology Department, Addis Ababa University, Addis Ababa, Ethiopia. ¹¹Department of Biology, University of Pennsylvania, Philadelphia, USA.

Received: 3 March 2022 Accepted: 13 February 2023

Published online: 24 February 2023

References

- Farh KKH, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*. 2015;518(7539):337–43.
- MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res*. 2017 Jan 4;45(Database issue):D896–901.
- Fraser HB. Gene expression drives local adaptation in humans. *Genome Res*. 2013;23(7):1089–96.
- Lappalainen T. Functional genomics bridges the gap between quantitative genetics and molecular biology. *Genome Res*. 2015;25(10):1427–31.
- Popejoy AB, Fullerton SM. Genomics is failing on diversity. *Nature*. 2016;538(7624):161–4.
- Sirugo G, Williams SM, Tishkoff SA. The missing diversity in human genetic studies. *Cell*. 2019;177(1):26–31.
- Kelly DE, Hansen MEB, Tishkoff SA. Global variation in gene expression and the value of diverse sampling. *Curr Opin Syst Biol*. 2017;1(1):102–8.
- Fan S, Hansen MEB, Lo Y, Tishkoff SA. Going global by adapting local: a review of recent human adaptation. *Science*. 2016;354(6308):54–9.
- Minster RL, Hawley NL, Su CT, Sun G, Kershaw EE, Cheng H, et al. A thrifty variant in CREBRF strongly influences body mass index in Samoans. *Nat Genet*. 2016;48(9):1049–54.
- Crawford NG, Kelly DE, Hansen MEB, Beltrame MH, Fan S, Bowman SL, et al. Loci associated with skin pigmentation identified in African populations. *Science*. 2017 Nov 17;358(6365):eaan8433.
- Storey JD, Madeoy J, Strout JL, Wurfel M, Ronald J, Akey JM. Gene-expression variation within and among human populations. *Am J Hum Genet*. 2007;80(3):502–9.

12. Stranger BE, Montgomery SB, Dimas AS, Parts L, Stegle O, Ingle CE, et al. Patterns of cis regulatory variation in diverse human populations. *PLoS Genet.* 2012;8(4):e1002639.
13. Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PAC, Monlong J, Rivas MA, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature.* 2013 Sep;501(7468):506–11.
14. Martin AR, Costa HA, Lappalainen T, Henn BM, Kidd JM, Yee MC, et al. Transcriptome sequencing from diverse human populations reveals differentiated regulatory architecture. *PLoS Genet.* 2014;10(8):e1004549.
15. Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, et al. The genetic structure and history of Africans and African Americans. *Science.* 2009;324(5930):1035–44.
16. Kwiatkowski DP. How malaria has affected the human genome and what human genetics can teach us about malaria. *Am J Hum Genet.* 2005;77(2):171–92.
17. Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, et al. Convergent adaptation of human lactase persistence in Africa and Europe. *Nat Genet.* 2007;39(1):31–40.
18. Scheinfeldt LB, Soi S, Thompson S, Ranciaro A, Woldemeskel D, Beggs W, et al. Genetic adaptation to high altitude in the Ethiopian highlands. *Genome Biol.* 2012;13(1):R1.
19. Yusuf AA, Govender MA, Brandenburg JT, Winkler CA. Kidney disease and APOL1. *Hum Mol Genet.* 2021;30(R1):R129–37.
20. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. *Nature.* 2015;526(7571):68–74.
21. Fan S, Spence JP, Feng Y, Hansen MEB, Terhorst J, Beltrame MH, et al. Whole-genome sequencing reveals a complex African population demographic history and signatures of local adaptation. *Cell.* in press.
22. McQuillan MA, Ranciaro A, Hansen MEB, Fan S, Beggs W, Belay G, et al. Signatures of convergent evolution and natural selection at the alcohol dehydrogenase gene region are correlated with agriculture in ethnically diverse Africans. *Molecular Biology and Evolution.* 2022 Aug 26;msac183.
23. Scheinfeldt LB, Soi S, Lambert C, Ko WY, Coulibaly A, Ranciaro A, et al. Genomic evidence for shared common ancestry of East African hunting-gathering populations and insights into local adaptation. *PNAS.* 2019;116(10):4166–75.
24. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 2009;19(9):1655–64.
25. 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.
26. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods.* 2015;12(5):453–7.
27. Glastonbury CA, Couto Alves A, El-Sayed Moustafa JS, Small KS. Cell-type heterogeneity in adipose tissue is associated with complex traits and reveals disease-relevant cell-specific eQTLs. *Am J Hum Genet.* 2019;104(6):1013–24.
28. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet.* 2012;44(7):821–4.
29. Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.* 2013 Oct 3;gr.155192.113.
30. THE GTex CONSORTIUM. The GTex Consortium atlas of genetic regulatory effects across human tissues. *Science.* 2020;369(6509):1318–30.
31. Ferreira PG, Muñoz-Aguirre M, Reverter F, Sá Godinho CP, Sousa A, Amadoz A, et al. The effects of death and post-mortem cold ischemia on human tissue transcriptomes. *Nat Commun.* 2018;9(1):490.
32. Vösa U, Claringbould A, Westra HJ, Bonder MJ, Deelen P, Zeng B, et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat Genet.* 2021;53(9):1300–10.
33. Kwong A, Boughton AP, Wang M, VandeHaar P, Boehnke M, Abecasis G, et al. FIVEx: an interactive multi-tissue eQTL browser [Internet]. *bioRxiv*; 2021 [cited 2022 Sep 26]. p. 2021.01.22.426874. Available from: <https://www.biorxiv.org/content/https://doi.org/10.1101/2021.01.22.426874v1>
34. Ye CJ, Feng T, Kwon HK, Raj T, Wilson MT, Asinovski N, et al. Intersection of population variation and autoimmunity genetics in human T cell activation. *Science.* 2014;345(6202):1254665.
35. Quach H, Rotival M, Pothlichet J, Loh YHE, Dannemann M, Zidane N, et al. Genetic adaptation and neandertal admixture shaped the immune system of human populations. *Cell.* 2016;167(3):643–656.e17.
36. Nédélec Y, Sanz J, Baharian G, Szpiech ZA, Pacis A, Dumaine A, et al. Genetic ancestry and natural selection drive population differences in immune responses to pathogens. *Cell.* 2016;167(3):657–669.e21.
37. Zanetti D, Weale ME. Transethnic differences in GWAS signals: a simulation study. *Ann Hum Genet.* 2018;82(5):280–6.
38. Ongen H, Buil A, Brown AA, Dermitzakis ET, Delaneau O. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics.* 2016;32(10):1479–85.
39. Zaitlen N, Pasaniuc 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.
40. Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* 2019;47(D1):D766–73.
41. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. *Genome Res.* 2002;12(6):996–1006.
42. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods.* 2012;9(3):215–6.
43. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature.* 2012;489(7414):57–74.
44. Tehranchi A, Hie B, Dacre M, Kaplow I, Pettie K, Combs P, et al. Fine-mapping cis-regulatory variants in diverse human populations. *Morris AP, Wittkopp PJ, editors. eLife.* 2019 Jan 16;8:e39595.

45. International Multiple Sclerosis Genetics Consortium. Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. *Science*. 2019 Sep 27;365(6460):eaav7188.
46. Lill CM, Luessi F, Alcina A, Sokolova EA, Ugidos N, de la Hera B, et al. Genome-wide significant association with seven novel multiple sclerosis risk loci. *J Med Genet*. 2015;52(12):848–55.
47. Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, Järvelä I. Identification of a variant associated with adult-type hypolactasia. *Nat Genet*. 2002;30(2):233–7.
48. Hamblin MT, Di Rienzo A. Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. *Am J Hum Genet*. 2000;66(5):1669–79.
49. Kudaravalli S, Veyrieras JB, Stranger BE, Dermitzakis ET, Pritchard JK. Gene expression levels are a target of recent natural selection in the human genome. *Mol Biol Evol*. 2009;26(3):649–58.
50. Shaheen R, Alsahli S, Ewida N, Alzahrani F, Shamseldin HE, Patel N, et al. Biallelic mutations in Tetratricopeptide Repeat Domain 26 (Intraflagellar Transport 56) cause severe biliary ciliopathy in humans. *Hepatology*. 2020;71(6):2067–79.
51. Volpi S, Cicalese MP, Tuijnenburg P, Tool ATJ, Cuadrado E, Abu-Halaweh M, et al. A combined immunodeficiency with severe infections, inflammation, and allergy caused by ARPC1B deficiency. *J Allergy Clin Immunol*. 2019;143(6):2296–9.
52. Baggiolini M, Dewald B, Moser B. Interleukin-8 and related chemotactic cytokines—CXC and CC chemokines. *Adv Immunol*. 1994;55:97–179.
53. Revez JA, Lin T, Qiao Z, Xue A, Holtz Y, Zhu Z, et al. Genome-wide association study identifies 143 loci associated with 25 hydroxyvitamin D concentration. *Nat Commun*. 2020;11(1):1647.
54. Rhodes DA, Reith W, Trowsdale J. Regulation of immunity by butyrophilins. *Annu Rev Immunol*. 2016;20(34):151–72.
55. Hirschhorn R, Huie ML, Kasper JS. Computer assisted cloning of human neutral α -glucosidase C (GANC): a new paralog in the glycosyl hydrolase gene family 31. *Proc Natl Acad Sci U S A*. 2002;99(21):13642–6.
56. Akey JM, Ruhe AL, Akey DT, Wong AK, Connelly CF, Madeoy J, et al. Tracking footprints of artificial selection in the dog genome. *PNAS*. 2010;107(3):1160–5.
57. Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZXP, Pool JE, et al. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science*. 2010;329(5987):75–8.
58. Liu B, Gloudemans MJ, Rao AS, Ingelsson E, Montgomery SB. Abundant associations with gene expression complicate GWAS follow-up. *Nat Genet*. 2019;51(5):768–9.
59. 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.
60. Shang L, Smith JA, Zhao W, Kho M, Turner ST, Mosley TH, et al. Genetic architecture of gene expression in European and African Americans: an eQTL mapping study in GENOA. *Am J Hum Genet*. 2020;106(4):496–512.
61. Idaghghour Y, Storey JD, Jadallah SJ, Gibson G. A genome-wide gene expression signature of environmental geography in leukocytes of Moroccan Amazighs. *PLoS Genet*. 2008;4(4):e1000052.
62. Marigorta UM, Navarro A. High trans-ethnic replicability of GWAS results implies common causal variants. *PLoS Genet*. 2013;9(6):e1003566.
63. Li YR, Keating BJ. Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations. *Genome Medicine*. 2014;6(10):91.
64. Brown BC, Asian Genetic Epidemiology Network Type 2 Diabetes Consortium, Ye CJ, Price AL, Zaitlen N. Transethnic genetic-correlation estimates from summary statistics. *Am J Hum Genet*. 2016;99(1):76–88.
65. Patel RA, Musharoff SA, Spence JP, Pimentel H, Tcheandjieu C, Mostafavi H, et al. Genetic interactions drive heterogeneity in causal variant effect sizes for gene expression and complex traits. *Am J Hum Genet*. 2022;109(7):1286–97.
66. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, et al. The structure of haplotype blocks in the human genome. *Science*. 2002;296(5576):2225–9.
67. Sawyer SL, Mukherjee N, Pakstis AJ, Feuk L, Kidd JR, Brookes AJ, et al. Linkage disequilibrium patterns vary substantially among populations. *Eur J Hum Genet*. 2005;13(5):677–86.
68. Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, et al. Linkage disequilibrium in the human genome. *Nature*. 2001;411(6834):199–204.
69. Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC, et al. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*. 2008;451(7181):998–1003.
70. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*. 2007;315(5813):848–53.
71. Valente EM, Logan CV, Mougou-Zerelli S, Lee JH, Silhavy JL, Brancati F, et al. Mutations in TMEM216 perturb ciliogenesis and cause Joubert, Meckel and related syndromes. *Nat Genet*. 2010;42(7):619–25.
72. Lee JH, Silhavy JL, Lee JE, Al-Gazali L, Thomas S, Davis EE, et al. Evolutionarily assembled cis-regulatory module at a human ciliopathy locus. *Science*. 2012;335(6071):966–9.
73. Choi H, Shin JH, Kim ES, Park SJ, Bae IH, Jo YK, et al. Primary cilia negatively regulate melanogenesis in melanocytes and pigmentation in a human skin model. *PLoS ONE*. 2016;11(12):e0168025.
74. Guan J, Gupta R, Filipp FV. Cancer systems biology of TCGA SKCM: Efficient detection of genomic drivers in melanoma. *Sci Rep*. 2015;5(1):7857.
75. Kundaje A, Meuleman W, Ernst J, Bilieny M, Yen A, Heravi-Moussavi A, et al. Integrative analysis of 111 reference human epigenomes. *Nature*. 2015;518(7539):317–30.
76. Cuomo ASE, Seaton DD, McCarthy DJ, Martinez I, Bonder MJ, Garcia-Bernardo J, et al. Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. *Nat Commun*. 2020;11(1):810.
77. Ward MC, Banovich NE, Sarkar A, Stephens M, Gilad Y. Dynamic effects of genetic variation on gene expression revealed following hypoxic stress in cardiomyocytes. Stegle O, Wittkopp PJ, editors. *eLife*. 2021 Feb 8;10:e57345.
78. Raj T, Rothamel K, Mostafavi S, Ye C, Lee MN, Replogle JM, et al. Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. *Science*. 2014;344(6183):519–23.

79. Randolph HE, Fiege JK, Thielen BK, Mickelson CK, Shiratori M, Barroso-Batista J, et al. Genetic ancestry effects on the response to viral infection are pervasive but cell type specific. *Science*. 2021;374(6571):1127–33.
80. Neavin D, Nguyen Q, Daniszewski MS, Liang HH, Chiu HS, Wee YK, et al. Single cell eQTL analysis identifies cell type-specific genetic control of gene expression in fibroblasts and reprogrammed induced pluripotent stem cells. *Genome Biol*. 2021;22(1):76.
81. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet*. 2007;81(5):1084–97.
82. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. *Nat Genet*. 2016;48(10):1284–7.
83. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. *Bioinform*. 2010;26(22):2867–73.
84. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*. 2015;4:7.
85. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLOS. Genetics*. 2006;2(12):e190.
86. Alexander DH, Lange K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *Bioinform*. 2011;12(1):246.
87. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215(3):403–10.
88. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinform*. 2013;29(1):15–21.
89. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinform*. 2014;30(7):923–30.
90. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *Bioinform*. 2011;12(1):323.
91. Li YI, Knowles DA, Humphrey J, Barbeira AN, Dickinson SP, Im HK, et al. Annotation-free quantification of RNA splicing using LeafCutter. *Nat Genet*. 2018;50(1):151–8.
92. Abbas AR, Wolslegel K, Seshasayee D, Modrusan Z, Clark HF. Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. *PLoS ONE*. 2009;4(7):e6098.
93. Maller JB, McVean G, Byrnes J, Vukcevic D, Palin K, Su Z, et al. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat Genet*. 2012;44(12):1294–301.
94. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. *Genome Biol*. 2016;17(1):122.
95. liftOver. Bioconductor. [Cited 2021 Nov 24]. Available from: <http://bioconductor.org/packages/liftOver/>
96. Storey JD, Bass AJ, Dabney A, Robinson D, Warnes G. qvalue: Q-value estimation for false discovery rate control. Bioconductor version: Release (3.14); 2021 [Cited 2021 Nov 24]. Available from: <https://bioconductor.org/packages/qvalue/>
97. Bhatia G, Patterson N, Sankararaman S, Price AL. Estimating and interpreting FST: The impact of rare variants. *Genome Res*. 2013;23(9):1514–21.
98. Kelly DE, Ramdas S, Ma R, Rawlings-Goss RA, Grant GR, Ranciaro A, Hirbo JB, Beggs W, Yeager M, Chanock S, Nyambo TB, Omar SA, Woldemeskel D, Belay G, Li H, Brown CD, Tishkoff SA. The genetic and evolutionary basis of gene expression variation in East Africans. Datasets. dbGaP. http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs002824.v1.p1 (2022)
99. Crawford NG, Kelly DE, Hansen MEB, Beltrame MH, Fan S, Bowman SL, Jewett E, Ranciaro A, Thompson S, Lo Y, Pfeifer SP, Jensen JD, Campbell MC, Beggs W, Hormozdiari F, Mpoloka SW, Mokone GG, Nyambo T, Meskel DW, Belay G, Haut J, NISC Comparative Sequencing Program, Rothschild H, Zon L, Zhou Y, Kovacs MA, Xu M, Zhang T, Bishop K, Sinclair J, Rivas C, Elliot E, Choi J, Li SA, Hicks B, Burgess S, Abnet C, Watkins-Chow DE, Oceana E, Song YS, Eskin E, Brown KM, Marks MS, Loftus SK, Pavan WJ, Yeager M, Chanock S, Tishkoff SA. Genetics of Pigmentation in Eastern and Southern African Populations Study. Datasets. dbGaP. https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001396.v1.p1 (2018)
100. Kelly DE. Mapping expression QTLs in East Africans. GitHub. https://github.com/derkelly/afr_eqtl (2022)
101. Kelly DE. Mapping expression QTLs in East Africans. Zenodo. <https://doi.org/10.5281/zenodo.7230625> (2022)

Publisher's Note

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