Cell Genomics Transparent Peer Review Record

Multi-ancestry genetic analysis of gene regulation in coronary arteries prioritizes disease risk loci

Chani J. Hodonsky[1], Adam W. Turner[1], Mohammad Daud Khan[1], Nelson B. Barrientos[1,2], Ruben Methorst[3], Lijiang Ma[4], Nicolas G. Lopez[5], Jose Verdezoto Mosquera[1,6], Gaëlle Auguste[1], Emily Farber[1], Wei Feng Ma[1,7], Doris Wong[1,6], Suna Onengut-Gumuscu[1], Maryam Kavousi[8], Patricia A. Peyser[9], Sander W. van der Laan[3], Nicholas J. Leeper[5], Jason C. Kovacic[10,11,12], Johan L.M. Björkegren[4,13], Clint L. Miller*[1,5,14]

Summary

Initial submission:	Received : 2/27/2023
	Scientific editor: Laura Zahn
First round of review:	Number of reviewers: 2 Revision invited : 4/17/2023 Revision received : 9/7/2023
Second round of review:	Number of reviewers: 2 Accepted : 11/19/2023
Data freely available:	Yes
Code freely available:	Yes

This transparent peer review record is not systematically proofread, type-set, or edited. Special characters, formatting, and equations may fail to render properly. Standard procedural text within the editor's letters has been deleted for the sake of brevity, but all official correspondence specific to the manuscript has been preserved.

Referees' reports, first round of review

Reviewer #1: In this manuscript, Hodonsky et al describes an expression quantitative trait loci (eQTL) and splicing QTL (sQTL) study employing bulk RNA-seq on human coronary artery tissues. Subsequently, QTLs were colocalized with coronary artery disease GWAS and CAD-related risk factors, resulting in a prioritized list of eGenes for further characterization. Given other QTL studies on coronary artery tissues (eg. GTEx, STARNET), the novelty of this study appears to be the use of a genetically diverse set of samples broadly representing the American population, on the presumption that this would lead to the discovery of novel gene targets. This is an interesting manuscript at a time when human population genetics is moving toward requiring more genetic diversity, but the authors should further describe the benefits of such efforts beyond generalizable comments. Generally, some details found in figure legends and methods should be mentioned in-text for readability.

Minor comments:

1) "With regard to local-ancestry-adjusted analyses, 337 eGenes were identified that did not exceed a FDR of 5% in the overall mixQTL analysis, demonstrating the merit of incorporating multiple approaches in a diverse study sample with genetic admixture."

This is couched as an advantage for using multiple approaches. Is there an evaluation for potential errors arising from different methods? Is there benchmarking?

2) "Using mixQTL, 54 eGenes with lead SNPs monomorphic in one or more 1000G superpopulations were identified..."

It would be beneficial to plot a distribution of monomorphic eGenes in the populations for a sense of the value of genetic diversity.

3) "We additionally employed FastPaintor to fine-map associations with epigenomic annotations and relevant GWAS."

Which epigenomics annotations and why?

4) For the "sensitivity analysis in European-ancestry subsample" section, the conclusion appears to be that a larger sample size provides more statistical power, which does not seem to fit the topic sentence. Instead, the authors could randomly downsample the original set to n=87 whilst keeping the proportions of different ancestries constant. This might allow conclusions on the usefulness of having an inclusive study.

5) Is the WGS data re-imputed after liftover to hg38 to capture variants that might be lost?

6) For annotations of eQTL and sQTL with snpEff and Annovar, what was done for predictions that were contrasting to each other?

7) How was the study's summary statistics incorporated with the 1000G GWAS summary statistics in the SMR analysis? Was this through a meta-analysis? If so, please describe in more detail about the parameters and tools used.

Reviewer #2: Non-European populations remain underrepresented in large scale genetic studies (GWAS, molecular QTL), including for highly prevalent diseases like coronary artery disease. Beyond addressing concerns of health equity, increasing diversity of these studies will increase power to detect independent associations and dissect variant contributions at individual loci. To address this gap, Hodonsky and co-authors present an e- and sQTL resource derived from study of coronary arteries from 138 ancestrally diverse American donors. In addition to replicating previous associations, the authors discover novel eQTLs, including at known GWAS loci. This is

enabled, in part, by use of paired approaches to examine both haplotype-specific and ancestry-specific associations. This work represents an expansive resource to complement existing findings reported via STARNET and GTEx. Overall, the manuscript is informative and will be highly useful to researchers. I think there are a few areas you could expand on to enhance overall utility and biological insights that can be gained from this dataset.

1. How representative were ancestries within diagnostic categories? This wasn't clear from the donor characteristics table. I'm empathetic to the challenges of acquiring tissue across ancestries, but it would be a worthwhile clarification to include, if possible.
2. The results presented here rely on bulk RNA profiling. With this in mind, do you expect tissue heterogeneity across diagnostic categories and do you predict that would have an appreciable effect on eQTL/sQTL detection?
3. The splicing results presented are interesting, but I think they could be expanded further.
a. Could you examine the single-cell data included here (or public datasets) to determine whether either of the two genes presented in Figure 5 are uniquely expressed in any specific cell types within the coronary artery?
b. Does the effect size of your sQTLs correlate with predicted variant effects (e.g., protein LOF variants, high scores from other variant effect predictors)?
c. Are sQTLs enriched in any other functional categories beyond open chromatin sites? Other histone modifications, TF motifs esp. for RNA binding proteins, etc.
4. Establishing best practices for multi-ancestry studies is an area of growing interest and appreciation, but there's still not a clear path for how these studies should be performed. I think more information/justification about the methods used in this study (specifically fine-mapping and colocalization) would be helpful for others who are interested in applying similar approaches with samples from diverse ancestries.
5. Minor note: please update your Figure S2 legend text to describe panel b.

Authors’ response to the first round of review

Response to Reviewers
Below we provide a point-by-point response to the Reviewers’ comments. Our responses are in blue text, with changes to the text provided in quotes. To support our responses, we also provide new data within the document, and cite references at the end of the document.

Reviewer #1
In this manuscript, Hodonsky et al describes an expression quantitative trait loci (eQTL) and splicing QTL (sQTL) study employing bulk RNA-seq on human coronary artery tissues. Subsequently, QTLs were colocalized with coronary artery disease GWAS and CAD-related risk factors, resulting in a prioritized list of eGenes for further characterization. Given other QTL studies on coronary artery tissues (eg. GTEx, STARNET), the novelty of this study appears to be the use of a genetically diverse set of samples broadly representing the American population, on the presumption that this would lead to the discovery of novel gene targets. This is an interesting manuscript at a time when human population genetics is moving toward requiring more genetic diversity, but the authors should further describe the benefits of such efforts beyond generalizable comments. Generally, some details found in figure legends and methods should be mentioned in-text for readability.

We thank the reviewer for taking the time to thoroughly read and provide comments on our manuscript. We have revised the results and discussion sections in several places (described in detail below) to be more specific about the benefits of both our study population and methodological approach, including additional citations as necessary. We have further clarified the limitations of publicly available reference datasets, which overrepresent genetically homogeneous populations, inhibiting both discovery and generalization in a genomics field dedicated to promoting health equity. Additionally, we have expanded on the ways in which increasing representation in these resources will improve both statistical power to discover globally low-frequent or rare variants, as well as ancestry-specific associations that are relevant to global populations. In particular, we have added the following text to the discussion section which has not been reproduced in the comments below:

"Nonetheless, our inclusive study design increased statistical power in both our diverse downsampled subset and overall study population compared to a genetically homogeneous European-ancestry-only subset. This is significant given the predominantly European genetic architecture of GTEx and published GWAS—while these resources have been crucial for genomics discovery to date, work highlighting the limitations of genetically restricted samples and technologies developed based on those samples points to the necessity of new, more expansive approaches. This also aligns with current appeals in basic science and public health to promote equitable research benefiting all populations, rather than studies that may extend the health disparity gap"

Please see below for additional responses to individual minor comments, along with relevant citations.

Minor comments:
1) "With regard to local-ancestry-adjusted analyses, 337 eGenes were identified that did not exceed a FDR of 5% in the overall mixQTL analysis, demonstrating the merit of incorporating multiple approaches in a diverse study sample with genetic admixture."
This is couched as an advantage for using multiple approaches. Is there an evaluation for potential errors arising from different methods? Is there benchmarking?

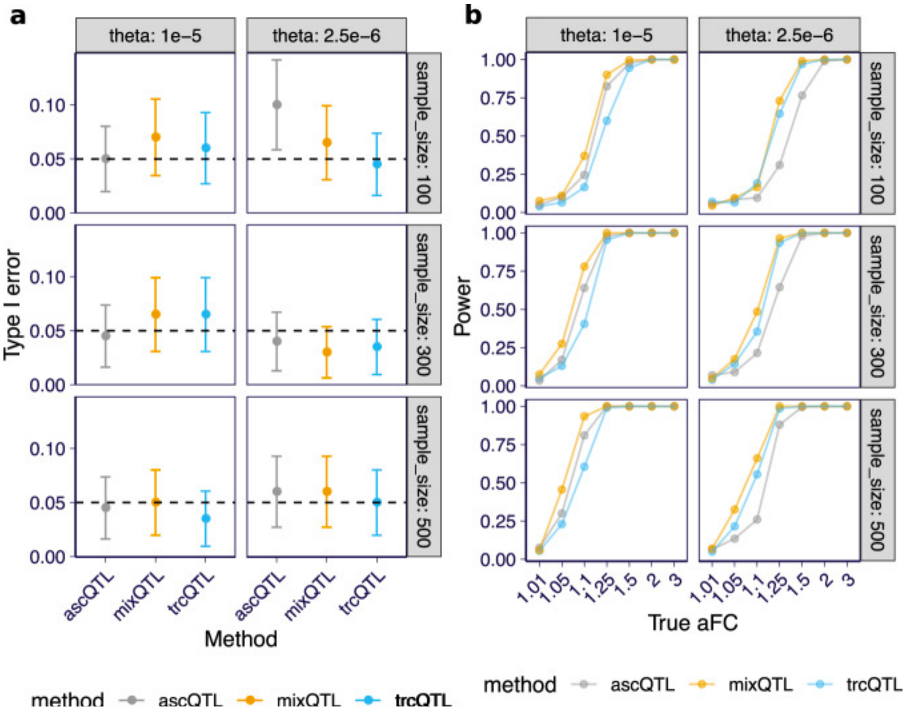
With the expansion of multi-ancestry studies of gene regulation, there is an increasing need to develop more appropriate methods than those applied to homogeneous populations. The approach we applied here is beneficial compared to using a single method for two reasons: 1) there is not a clear 'gold standard' approach for using any particular method in a genetically diverse sample of restricted size, so the number of "true" coronary eQTLs at any disease stage is unknown, and a stringent significance threshold can improve ability to detect likely true-positives for which one approach is comparatively better powered; 2) the tools and resources required for running mixQTL (phasing RNAseq data in particular) may not be broadly available so comparing multiple options is ideal to provide potential expectations for researchers looking to maximize available resources. We applied a 5% minor allele frequency (MAF) threshold across the total study population to minimize false-positive findings due to population stratification, while acknowledging future studies could identify low frequency variants (1-5% MAF) in one or more ancestries.

We have added information about the LA analysis results (the 337 genes which were either not evaluated or not reported as significant in mixQTL) in the Results sections (under headings "Local-ancestry-adjusted and ancestry-specific eQTLs" and "Colocalization of eQTLs"). We have also incorporated LA colocalization results into the shared coloc figure (now Figures S4a) with the downsampled sensitivity analyses referenced below in a response to comment (4), and added Table S13 to the supplement. Specifically, the following two paragraphs:
"Among the genes with no eQTLs exceeding genome-wide-significance in the mixQTL were several interesting genes with sub-significant associations, including VPS37B (Vacuolar Protein Sorting-Associated Protein 37B). VPS37B is involved in endosomal protein binding activity, and the genomic region has been associated with CAD-relevant traits including adiponectin levels, BMI, and cholesterol traits. MixQTL and LA methods resulted in the same lead variant, rs897392, in the intron of neighboring gene HIP1R, which exhibits modest to strong LD with reported arterial eQTLs for VPS37B (Table S11) as well as different expression between CAD cases and controls in multiple tissues in the STARNET study population. rs897392 had a mixQTL adjusted p-value of 0.14, meaning VPS37B would not be considered an eGene using this method alone, despite evidence favoring genetic regulation of this gene in cardiac tissues."

And

"Among LA eGenes, colocalization was limited to 27 associations, but four of these were not eGenes in mixQTL and a further three were mixQTL eGenes that did not colocalize to any GWAS trait (Table S14, Figure S4). Of particular interest is ANAPC13, a component of an anaphase-associated E3 ubiquitin ligase for which the LA eQTL (led by rs9809619, pLA=2.3E-6) colocalized to the MVP CAD association signal, but did not meet the threshold for colocalization (PPH4 ≥0.8) for any trait in mixQTL. Rs9809619 is in close proximity to and exhibits near-perfect LD globally with the lead mixQTL variant, rs4367113 (pmixQTL=1.6E-5, Table S11), across 1000 Genomes populations. LA adjustment resulted in lower p-values for ANAPC13-associated variants compared to mixQTL (Figure S4), showing the benefit of complementary approaches for a locus with similar associations but different significance between methods."

With regard to potential errors, we refer to Liang, et al[1], for assessments of false-positives in simulations utilizing mixQTL. Using mixQTL, a 29% power increase was demonstrated in GTEx compared to standard regression-based methods, and Type I error is well controlled, as shown in figure 2 from their manuscript, reproduced below. Compared to their simulations, our study population is most similar to the top row of panels and we could be expected to identify approximately 80% of true associations with an allelic fold change ≥1.5 with a Type I error falling between ~3.5 and 9% using mixQTL, well within acceptable ranges in the field.



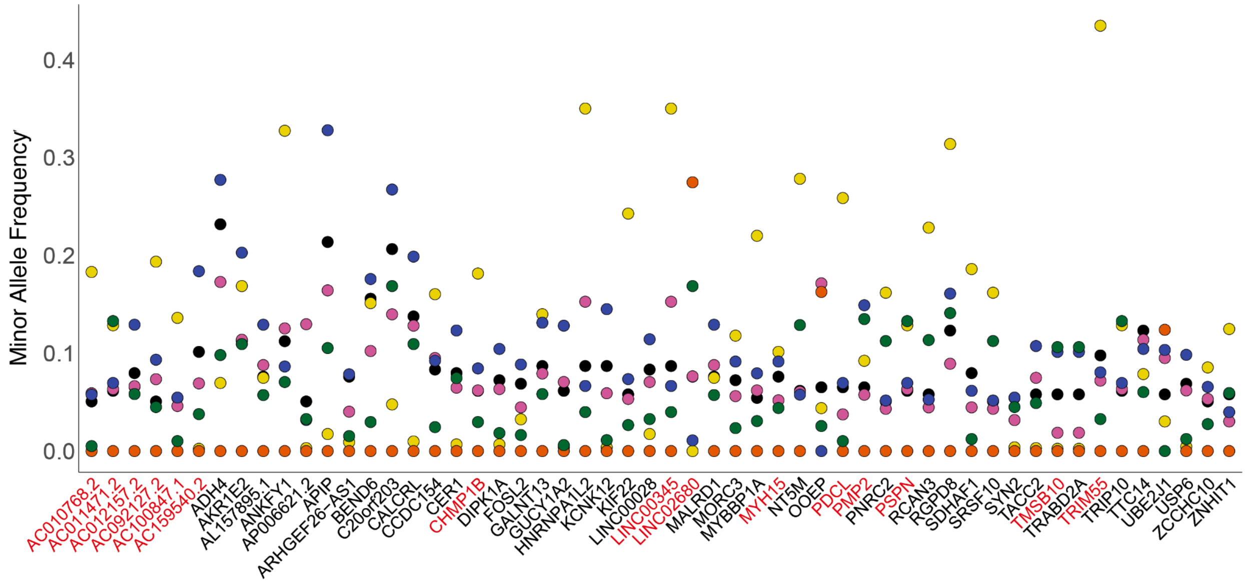
Text of Liang, et al, figure legend: "Fig. 2. QTL mapping performance for mixQTL and approaches based on either total reads (trcQTL) or allele-specific reads (ascQTL) on simulated data. Each panel presents the results for two relative abundances of the gene, θ, and three sample sizes. a Type I error (y-axis) at a 5% significance level across methods (x-axis) are shown. The dashed line represents the desired error rate under the null hypothesis. The error bar indicates the 95% confidence interval of the estimated error rate from 200 replicates. b Power (y-axis) at a 5% significance level across methods under a range of true aFC values (x-axis) are shown. Power is defined as the fraction of eQTLs passing the significance threshold."

With regard to false positives in the local ancestry adjusted analyses, previous work has suggested that adjustment for local ancestry reduces Type I error in studies with genetically heterogeneous populations.[2,3] Potential population stratification leading to associations with rare, ancestry-specific variants should be further mitigated by the inclusion of only sample-wide ≥5% MAF variants in the analyses. This is supported in part by the fact that we identify fewer ancestry-specific (or rather single-ancestry-monomorphic) associations using LA compared to mixQTL (17 vs 54 eGenes). While this specific approach has not been applied to eQTL analyses to date, there is evidence supporting our approach in a recently published multi-ancestry analysis in GTEx [4] as well as GBMI efforts [5]. These consortia recommend local ancestry inference on a broader scale to improve global and ancestry-specific association detection in globally representative populations based on findings from the aforementioned references by Zhong, et al and Qin, et al.

2) "Using mixQTL, 54 eGenes with lead SNPs monomorphic in one or more 1000G superpopulations were identified…"

It would be beneficial to plot a distribution of monomorphic eGenes in the populations for a sense of the value of genetic diversity.

Thank you for this suggestion. We have generated the following figure to demonstrate the differences in allele frequency by 1000G ancestry as well as the overall AF in our study population for these 54 lead variants (highlighting discovery eGenes in red font), added it to Figure S2d, and referenced it in the text.



3) "We additionally employed FastPaintor to fine-map associations with epigenomic annotations and relevant GWAS."

Which epigenomics annotations and why?

Thank you for pointing out this omission from the results. Published eQTL studies in tissues affected by common complex diseases support using epigenetic marks representing chromatin accessibility as well as enhancer activity or repression for refining eQTL associations and prioritizing candidate causal variants.6–10 Coronary artery epigenetic data are limited, so we utilized available options representing (as much as possible) gene repression, activation, transcription, and chromatin-contact-based enhancer activity. We have updated both the results and the methods to reflect the specific annotations and clarified our reasoning as follows:

Methods

“We evaluated four annotations based on gene repression, activation, transcription, and chromatin-contact-based enhancer activity: CTCF only, H3K4me3 only, and H3K4me3 / H3K27Ac combined. To do this, we included ENCODE binary SNP annotations for chromatin accessibility (CTCF, H3K27ac, and H3K4me3 in coronary artery tissue from one 53 yo female). To apply enriched enhancer regulatory activity in coronary artery tissue, we also used the activity-by-contact model (ABC; https://github.com/broadinstitute/ABC-Enhancer-Gene-Prediction). ABC scores were obtained from previous H3K27ac HiChIP and ATAC data in HCASMC and coronary artery from our previous work are publicly available (see “Data and Code Availability”).”

Results

“Including prior functional annotations in relevant tissues can refine association signals and prioritize variants and candidate cis-regulatory mechanisms. Therefore, we employed FastPaintor to fine-map associations with epigenomic annotations (ENCODE coronary artery H3Kme3 and H3K27Ac marks and activity-by-contact scores for human coronary artery SMCs) as well as BP and CAD GWAS, which both exhibited strong evidence of colocalization.”

4) For the "sensitivity analysis in European-ancestry subsample" section, the conclusion appears to be that a larger sample size provides more statistical power, which does not seem to fit the topic sentence. Instead, the authors could randomly downsample the original set to n=87 whilst keeping the proportions of different ancestries constant. This might allow conclusions on the usefulness of having an inclusive study.

Thank you for this useful suggestion, we agree that directly comparing similarly sized samples would better portray both the ostensible benefits of inclusive study samples, as well as the increased discovery with increased sample size across ancestries. To address this, we followed the reviewer’s suggestion and downsampled randomly within each ancestry group to approximately 63% of the total study to achieve a diverse sample of 80 individuals (this was correctly noted in the supplement and methods but unfortunately reported as 87 [a typo] in the manuscript, which we have changed), then re-ran mixQTL and performed colocalization and generalization analyses within this subset to directly compare to the European-ancestry-only subset.

Our findings indicate that an inclusive study design outweighs the potential for limiting ancestry-specific low-frequency variants when compared to exclusively subsetting to a highly homogeneous population such as European-ancestry individuals. In comparison to 1,311 eGenes (210 discovery) in the European-restricted subset, we identified 1,469 eGenes (395 discovery) in a proportionally representative subset of the total study. Among combined-study-population eGenes, 1,043 were also eGenes in the diverse subset (median minor AF of lead eQTL 30%, 295 not in the Euro dataset) compared to 1,033 in the European-only subset (median minor AF of lead eQTL 29%, 285 not in the diverse subset). Generalization was improved for the diverse subset (1,074 generalized eGenes compared to 983 in the Euro-only sample), despite the genetic ancestry of published arterial eQTL datasets being overwhelmingly of European origin. We expand on these findings in the results in comparison to the European-ancestry-only findings, and have added a description of this subset derivation in the corresponding Methods section.

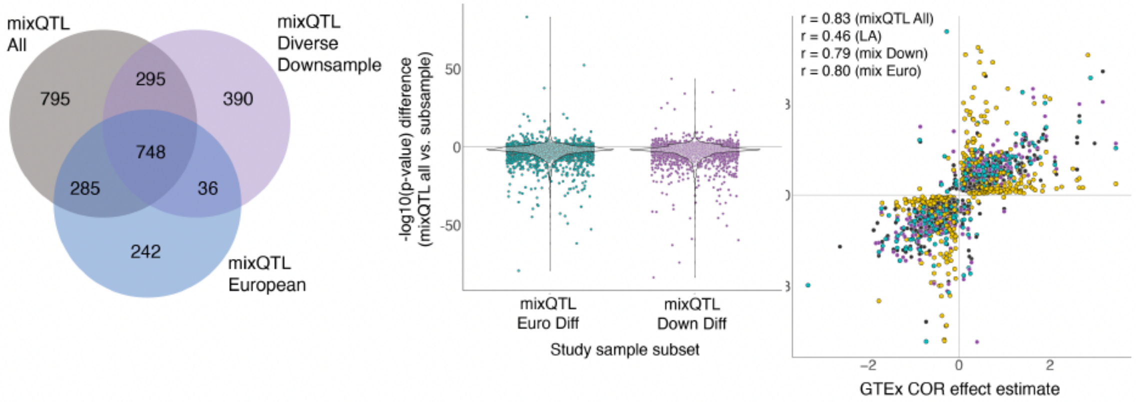
Results

“We also assessed whether our inclusive study design affected discovery, colocalization, and fine-mapping by restricting our sample to European-ancestry individuals (n=80), as well as a random subset of 80 members approximating the representation of the genetic ancestry of the total sample. In the European-ancestry-only subset, we identified 1,311 eGenes (16% discovery eGenes; 79% eGenes in the combined analysis), compared to 1,469 eGenes in the genetically diverse subset (27% discovery, 71% present in combined analysis, Figure S8, Tables S18A,B). With regard to generalization of published arterial eQTLs, 983 and 1,074 eGenes in the European-only and representative subsets respectively also had eQTLs in a GTEx or STARNET arterial tissue, and directional consistency with GTEx coronary was over 90% (Figure S8, Tables S19A,B). Compared to the overall sample, less than half the number of eGenes from either subset colocalized to relevant GWAS traits–fewer than would be expected if colocalization were linearly correlated with sample size (Figure S4, Tables S20A,B). The reduction in associations across all analyses in the European-only subset compared to the genetically diverse subset reinforces the benefits of methodological approaches designed to maximize study sample size and diverse genetic ancestry representation.”

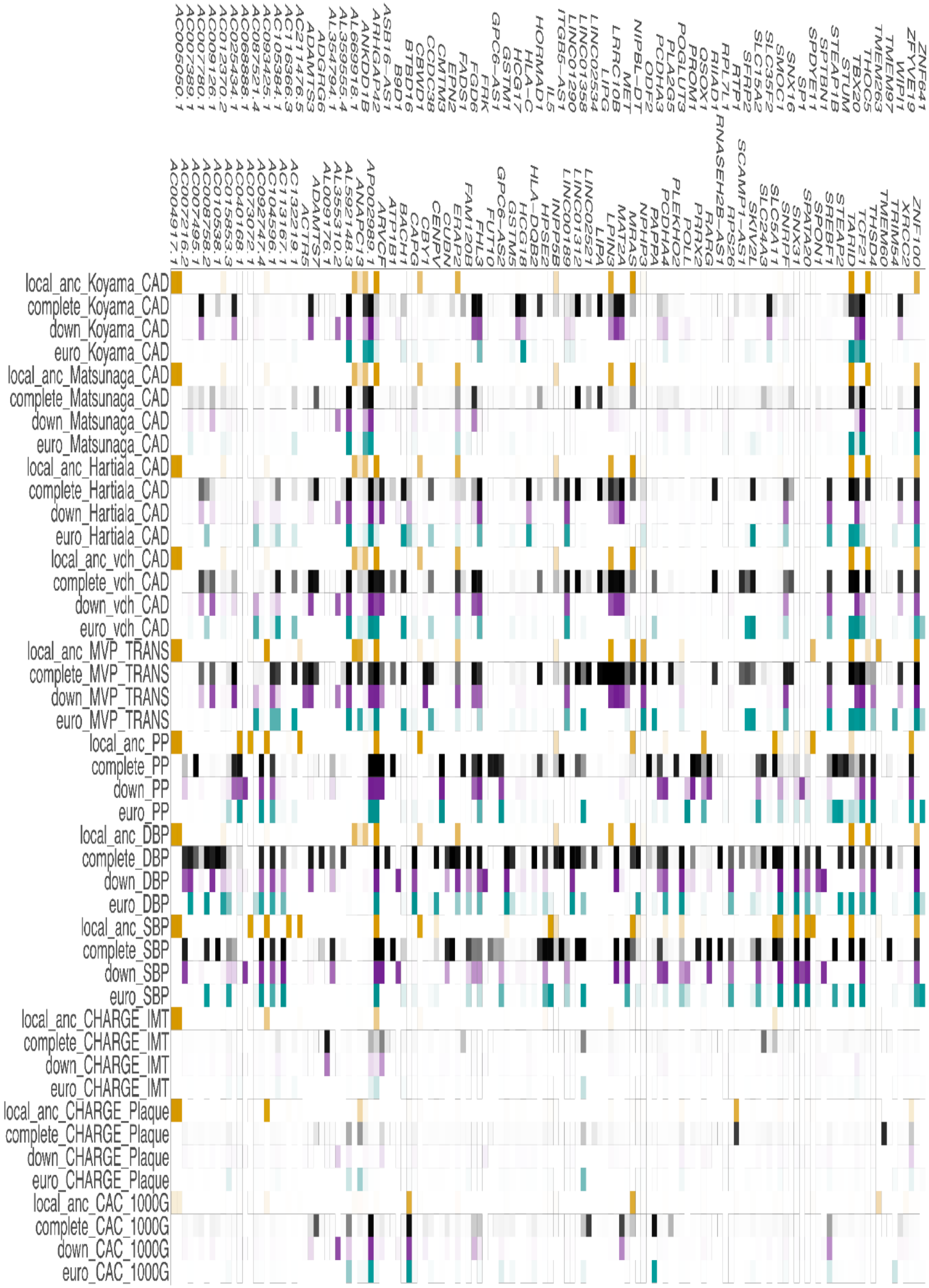
Methods

We have updated the Methods section to incorporate the above information as follows, in a new subsection entitled "Sensitivity analyses in European-ancestry-specific and ancestrally proportional sample subsets":

"Most statistical fine-mapping methods and publicly available genomics references continue to over-represent ancestrally homogeneous European-ancestry populations. To evaluate whether discovery or replication could be improved in a genetically homogeneous study sample despite a meaningful decrease in sample size, we performed mixQTL and downstream analyses in subsets of our sample restricted either to individuals with 100% European ancestry (n=80) or an 80-person subset of the total study population randomly selected within each majority-assigned-ancestry group. Regression, generalization, and colocalization were performed within both subsets as described above for the combined sample."
With regard to figures and tables, we have added summary information about the diverse subset findings to Tables 1 and S2; added Supplemental Tables S17B, S18B, and S19B for lead variants, generalization, and colocalization of the diverse subset respectively; and added three Supplemental Figures (new Figures S4a and S8a-c, copied below) comparing the findings from the separate 80-person subsets to each other as well as to the overall study population results (coloc plot includes local ancestry results as referenced above in response to comment 2).



Cell Genomics Transparent Peer Review Record



Finally, we also uploaded summary statistics from this new subset analysis to Zenodo in combination with the previously reported summary statistics.
(new version DOI: 10.5281/zenodo.7992146)

5) Is the WGS data re-imputed after liftover to hg38 to capture variants that might be lost?

Re-imputation of low-pass WGS was necessary in our case because the original VCFs received from Gencove were not phased, which is required for all analyses. No additional variants were captured, rather the re-imputation was solely a function of our phasing & imputation pipeline. We have clarified this in the methods as follows:
"Because phasing was necessary for downstream analyses, samples were phased and subsequently re-imputed to 1000G phase 3 b37 reference panel using Beagle with impute=true and gp=true options. No additional variants had been imputed at the conclusion of this process."

6) For annotations of eQTL and sQTL with snpEff and Annovar, what was done for predictions that were contrasting to each other?

In order to maximize interpretability, we performed a comparison of our Annovar and SnpEff annotations and have provided the results here, but given restricted annotation categorizations presenting solely SnpEff in the revised manuscript. Based on the results of our fine-mapping and the understanding that lead eQTLs are often proxy variants rather than the causal SNP, we have additionally run SnpEff with the relevant Gencode v32 reference annotation on all variants included in any Paintor credible set. These results are presented as four additional columns in Tables S15A/B, representing the assigned primary and secondary annotations by SnpEff, the name of the transcript affected, and the sequence change for each respective gene-variant pair. We have similarly presented the SnpEff annotations for all available lead sQTLs in a new supplementary table (now Table S22).

7) How was the study's summary statistics incorporated with the 1000G GWAS summary statistics in the SMR analysis? Was this through a meta-analysis? If so, please describe in more detail about the parameters and tools used.

We have attempted to clarify this analysis in both the methods and results section to provide additional context. The functionality of SMR does not support comparing more than two studies, therefore we compared each relevant GWAS with our eQTL associations in individual analyses and combined the results for presentation in the supplement. The results have been revised as follows:

"We further assessed the possibility that GWAS associations overlapping our eQTL associations were putatively mediated by genetically regulated gene expression using summarized Mendelian randomization (SMR). Given LD-reliant methodologic restrictions for both coloc and SMR, as expected we identified fewer associations but notable overlap between the two methods (25 overlapping signals and 18 unique to SMR, Table S13)."

Additionally we have revised the methods to add information about the relevant thresholds and parameters used:

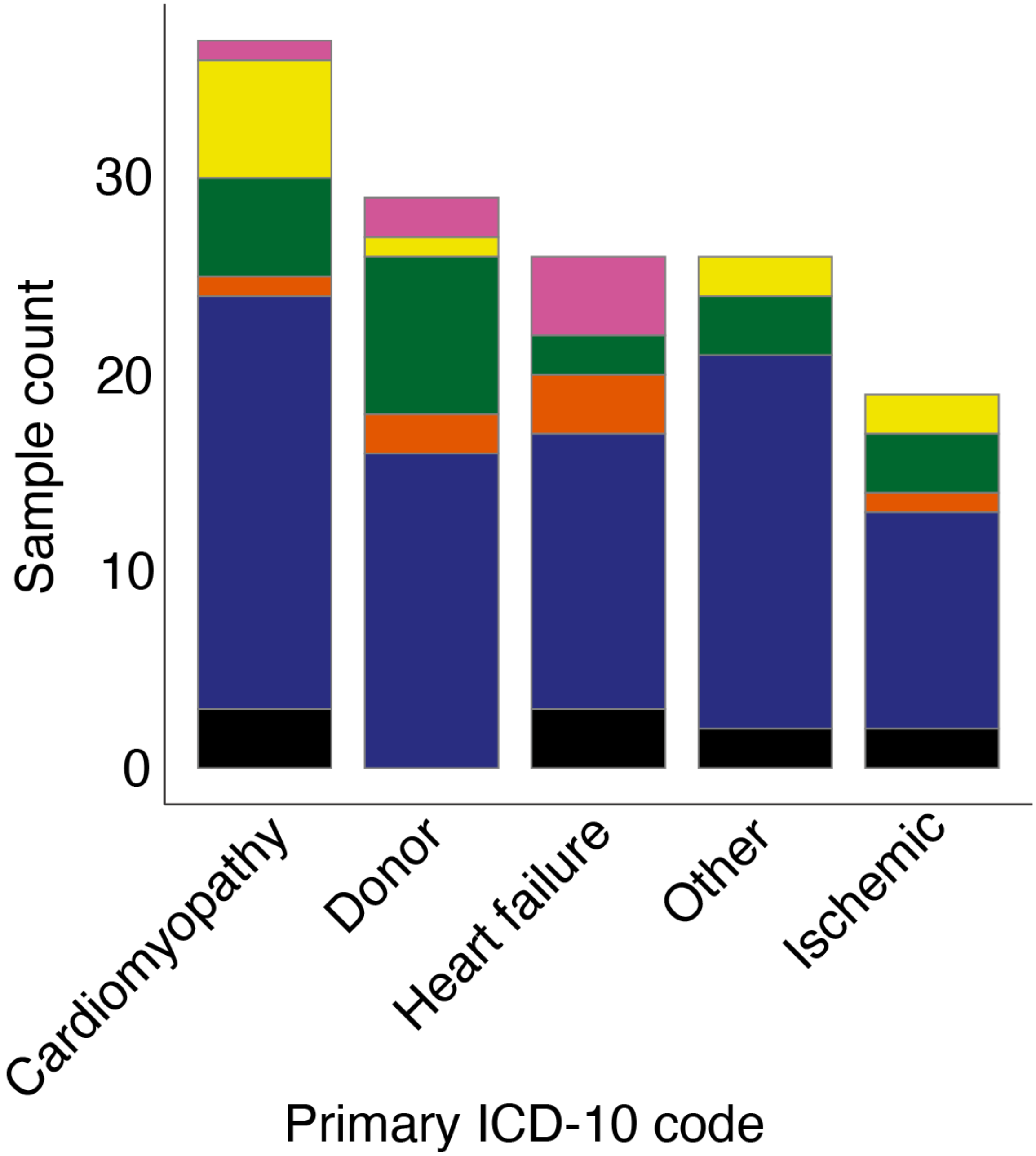
"Due to the complexity and low informativeness likely in LD generated from our genetically diverse but modestly sized sample, we generated bed files in Plink using 1000G EUR population (excluding Finnish samples due to genetic distinction of that population, which was not represented in our samples) as our LD reference. We then performed SMR (https://yanglab.westlake.edu.cn/software/smr) using mixQTL results and GWAS summary statistics from CAD and BP traits included in colocalization analyses. We were not able to evaluate Japanese CAD GWAS due to the large number of variants with allele frequency differences >30% between the GWAS study populations and ours. Options included a minor allele frequency threshold of 0.01, a 1Mb window surrounding the most significant eQTL per locus, and a maximum mixQTL p-value of 5E-06 required for variant inclusion."

Reviewer #2
Non-European populations remain underrepresented in large scale genetic studies (GWAS, molecular QTL), including for highly prevalent diseases like coronary artery disease. Beyond addressing concerns of health equity, increasing diversity of these studies will increase power to detect independent associations and dissect variant contributions at individual loci. To address this gap, Hodonsky and co-authors present an e- and sQTL resource derived from study of coronary arteries from 138 ancestrally diverse American donors. In addition to replicating previous associations, the authors discover novel eQTLs, including at known GWAS loci. This is enabled, in part, by use of paired approaches to examine both haplotype-specific and ancestry-specific associations. This work represents an expansive resource to complement existing findings reported via STARNET and GTEx. Overall, the manuscript is informative and will be highly useful to researchers. I think there are a few areas you could expand on to enhance overall utility and biological insights that can be gained from this dataset.

1. How representative were ancestries within diagnostic categories? This wasn't clear from the donor characteristics table. I'm empathetic to the challenges of acquiring tissue across ancestries, but it would be a worthwhile clarification to include, if possible.

We agree with the reviewer that this is indeed pertinent for interpreting potential stratification of results by ancestry/diagnosis information given healthy equity issues such as differences in access to care, different ages at diagnosis, and differential recommendations for treatment course by race and ethnic group, which often correlate to genetic ancestry.
We have revised the relevant sentence in the results as follows to clarify this information:
"Majority inferred ancestry groups were represented across diagnoses, but only European and South Asian genetic ancestries were represented in all primary diagnostic categories (Figure 1E, Table S1)."

We have also updated Figure 1E as shown below to visually portray the distribution of majority assigned ancestry by diagnostic category, which is more informative than coloring according to explant/donor status. While differences do exist, most ancestries are represented in most categories (specifically rejected donor hearts, which are most likely to represent "healthy" or



"normal" tissue samples compared to transplant recipients), suggesting our sample is not particularly biased toward ancestry-specific representation in one diagnostic category.

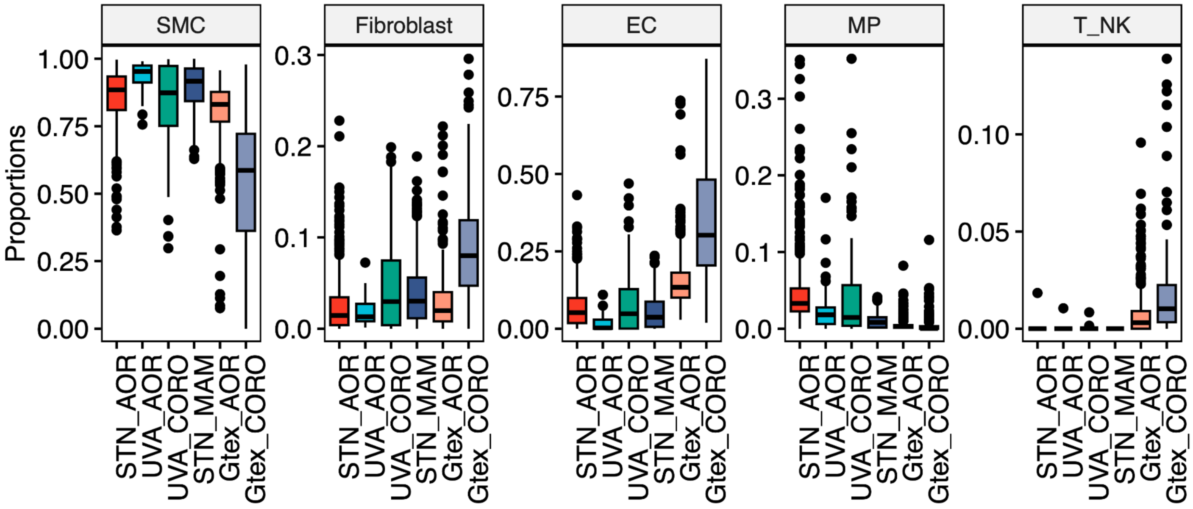
2. The results presented here rely on bulk RNA profiling. With this in mind, do you expect tissue heterogeneity across diagnostic categories and do you predict that would have an appreciable effect on eQTL/sQTL detection?

This is an excellent point, and we acknowledge that sample composition heterogeneity by disease status as well as diagnostic categories is likely in our study sample, in particular given the difficulties inherent in isolating RNA from samples with highly calcified lesions/advanced disease. Incorporating cell-type-proportion estimates is expected to improve precision of QTL associations–to this end, we are contributing these results to a STARNET manuscript incorporating cell-type-proportion estimates for multiple human arterial tissue types from our coronary arteries as well as STARNET arterial tissues using deconvolution via CibersortX with a meta-analyzed arterial reference generated in our lab11. Based on preliminary findings from these analyses (Ma, et al, in preparation, figure below), we see an increase in eGene discovery using standard regression methods as well as discernable differences between samples collected in living vs deceased study participants when calculating proportions for five major cell types (EC, SMC, Macrophage, fibroblast, and T/NK cells; see unpublished preliminary figure below, from left to right in each box plot is STARNET aortic root, UVA aortic root [insufficient sample size to include in our eQTL study], UVA coronary artery, STARNET mammary artery, and GTEx coronary artery).

We have also added the following text to the limitations paragraph in the discussion:

"Additionally, we acknowledge that both sample quality and disease status may affect interindividual cell-type proportions and therefore eQTL detection. Adjusting for estimated cell type proportions using single-cell reference based deconvolution may improve discovery across tissues and complement cell-type-specific QTL studies."

Deconvolution results using meta data non-lesion samples-5 cell types (GTExV8)



Unfortunately, while adjusting for cell-type proportions generated from single-cell data may prove informative for identifying eQTLs when phenotype-representative references were available, this would not be expected to be equally applicable to sQTL identification unless splicing differences were also affected by total gene expression across cell types. (We discuss this further in the response to the following comment.) As it stands, there is no splicing/isoform-specific expression reference dataset to apply for deconvolution in this way. The availability of such a resource would certainly be a boon to the atherosclerosis research community moving forward.

3. The splicing results presented are interesting, but I think they could be expanded further.

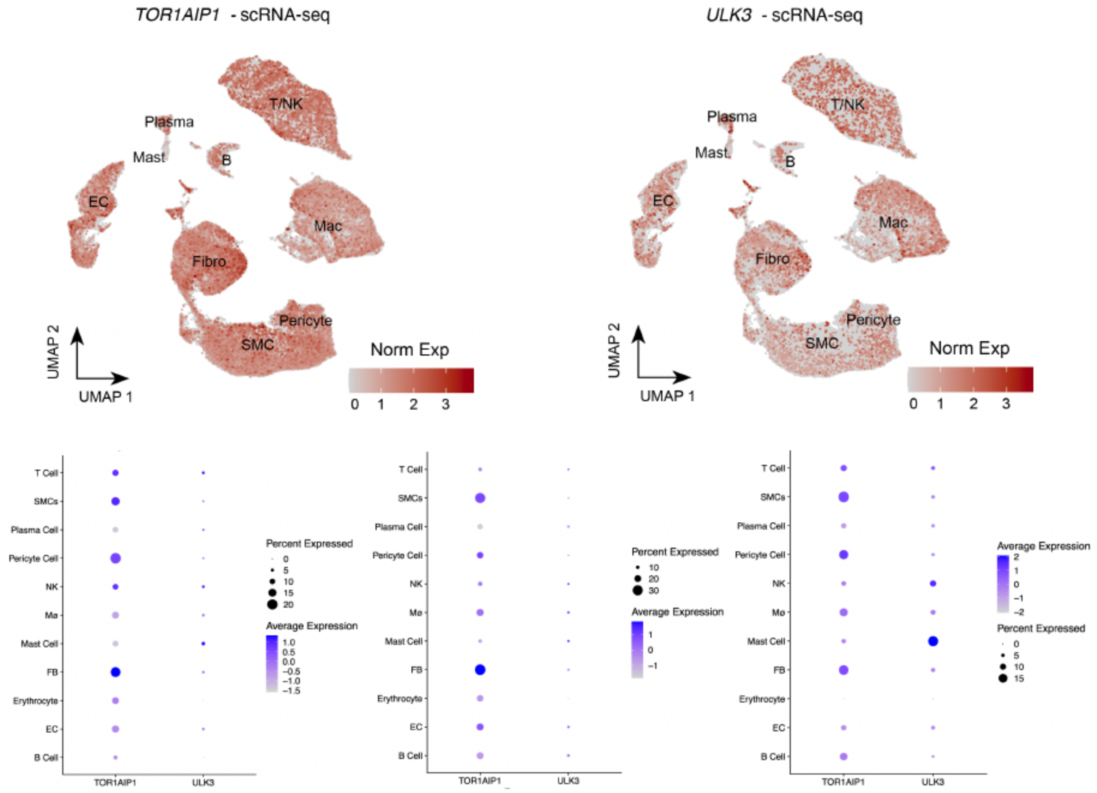
a. Could you examine the single-cell data included here (or public datasets) to determine whether either of the two genes presented in Figure 5 are uniquely expressed in any specific cell types within the coronary artery?

Since isoform differences are not captured in short-read data (nor well represented in published arterial datasets), the interpretation of our findings is limited to overall gene expression. While isoform-specific changes cannot be identified in our study, there is utility for future functional characterization in ascertaining whether genetically regulated isoform switching is more likely to be impactful in specific cell types. We have added the following text to the Results section entitled "Generalization and colocalization of lead sQTLs" to expand on this idea:

"Tissue-specific differences in isoform proportion cannot necessarily be detected with bulk sequencing, implicating potentially distinct regulatory mechanisms for splicing compared to overall expression. Therefore, we first directly compared the gene sets represented by eQTLs and sQTLs. Only modest overlap was observed (297 eGenes were also sGenes, 23 of which shared the same lead variant, Table S23), suggesting that splicing analyses likely represent unique disease-relevant pathways compared to overall expression. Among 71 sGenes with lead sQTLs reported to have splicing functions in SnpEff, 59 have no eQTLs, with 5 having no expression variants exceeding even nominal significance (Table S23). This difference is further exemplified by differences in cell-type-specificity in our scRNAseq reference dataset: while nearly all cell-type-specific (score >0.7 as described in methods) reference genes were eGenes (408 of 411 total), only 37 sGenes (of 83 total) exhibited cell-type specificity."

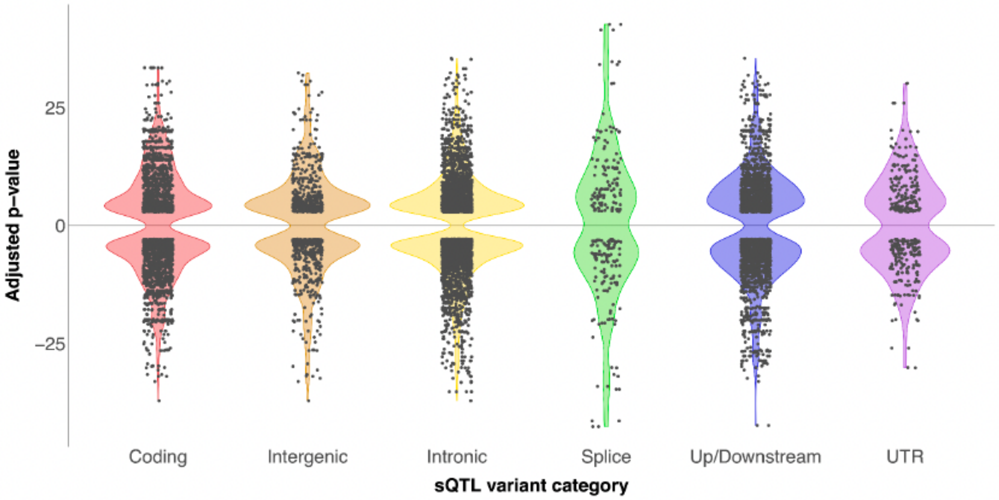
To visualize the distribution of these genes despite lacking significant cell type specificity, we generated UMAP plots for TOR1AIP1 and ULK3 as well as study-specific scRNA expression from human studies available in Plaqview.12–14 Both genes are broadly expressed, with modest evidence of higher expression of TOR1AIP1 in mural cells and pericytes.

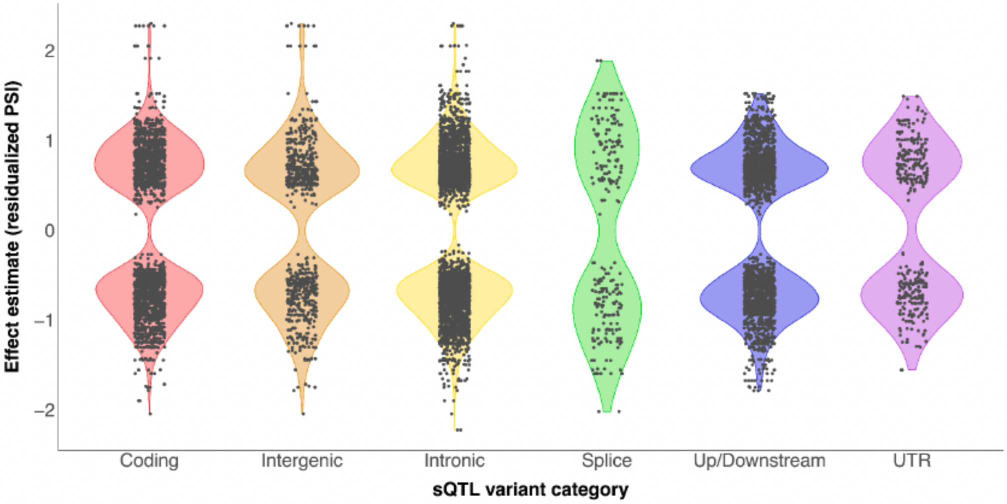
In summary, long-read expression data or single-cell Isoseq15 will be essential for comprehensively addressing these questions, particularly in precious tissues such as coronary artery which do not have large sample sizes or publicly available references, as well as cell types present in lower proportions in these tissues.16 To comprehensively portray our results insofar as we are able, we have added the UMAPs below to the supplement (Figures S9b,c), presented the dot plots for the reviewer's consideration (L to R: Alsaigh, et al; Hu, et al; and Wirka, et al), updated the results section to include the aforementioned data, and extended the discussion paragraph referring to these two genes in the context of limitations of available references as described above.



b. Does the effect size of your sQTLs correlate with predicted variant effects (e.g., protein LOF variants, high scores from other variant effect predictors)?

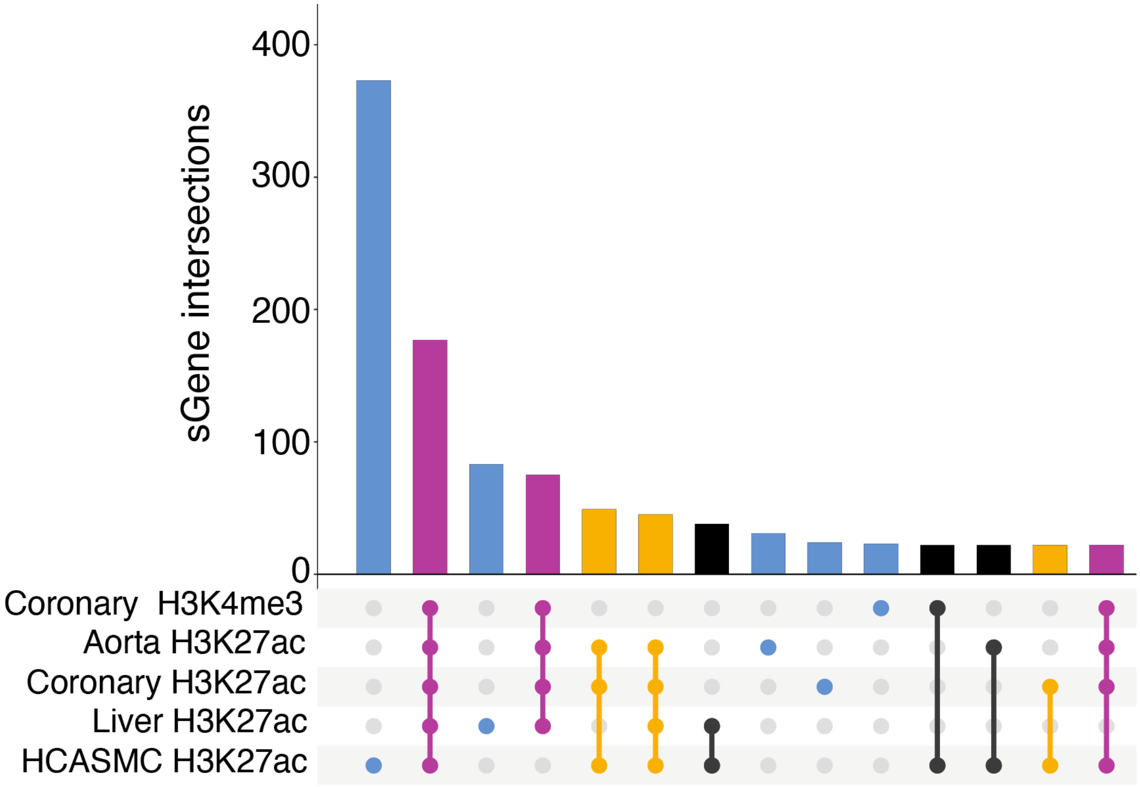
While we did perform limited fine-mapping in the sQTL dataset, it is much more complex to evaluate in Paintor, for example, compared to total gene expression, so we limited functional annotation to the lead sQTLs, which are expected to be proxies for the causal variant and may therefore not proportionally represent the functional categories of the true causal variant or variants at each association signal. However, given that we would indeed expect effect size for PSI to vary according to its causal mechanism, we agree that it is important to clarify this relationship using the available data. While effect size was not correlated with broadly defined variant category (pANOVA=0.62), there was a very strong association with p-value (pANOVA=3.65E-11), suggesting that even low-effect-magnitude sQTLs (which could be simplistically interpreted as small differences in isoform-specific expression) may play a role in physiology. We have generated the plot below and incorporated it into Figure S9a, and provide the effect vs variant annotation plot for your review.





c. Are sQTLs enriched in any other functional categories beyond open chromatin sites? Other histone modifications, TF motifs esp. for RNA binding proteins, etc.

Keeping in mind the caveat of lead sQTLs being potential proxies for causal SNPs, we have evaluated several additional functional categories for these variants in sGenes. First, we evaluated overlap between lead sQTLs and H3K27 acetylation marks from ChIP-seq in HCASMCs as well as coronary, aorta, and liver tissue from one individual in ENCODE, and H3K4me3 in coronary in the same individual. We generated the following UpSet plot to visualize these results, replacing the SnpEff annotation plot in Figure 5b from the original manuscript, which is now Supplemental Figure S8a:



We agree that it would be interesting to evaluate the enrichment for other functional categories such as RNA binding proteins. However these data are currently not available in coronary artery or other human artery tissues or cells, and analysis of available data in non-adherent cells (e.g. K562) would likely not provide reliable information. We also acknowledge that fully addressing this question would require long-read sequencing to capture isoform specificity, which would improve functional characterization for in vitro or in vivo disease modeling. As shown in the UMAP plots, we are not able to separate out cell type specificity for the sGenes, suggesting that single-cell isoform sequencing is required. While this remains an important question, we feel that additional sequencing data are ultimately required and this remains outside the scope of the current work.

4. Establishing best practices for multi-ancestry studies is an area of growing interest and appreciation, but there's still not a clear path for how these studies should be performed. I think more information/justification about the methods used in this study (specifically fine-mapping and colocalization) would be helpful for others who are interested in applying similar approaches with samples from diverse ancestries.

Thank you for your comment. We agree that there are limited standards or best practices for performing these studies. It is a difficult path to forge given that inclusively designed studies often suffer from a lack of publicly available resources and there may be minimal motivation to improve representation within reference datasets when the majority of association studies are performed in genetically homogeneous populations already represented by those resources. We hope this will change with the increasing number of multi-ancestry GWAS publications.
Our intent regarding colocalization and fine-mapping was to maximize our findings in a sample that does have strong representation from European ancestry, while calling for improved methods development and study design to maximize discovery and generalizability of findings to global populations. With the recent accumulation of non-European genetic associations and reference datasets, we anticipate an increased justification for utilizing local ancestry within these expression analyses. As discussed in the response to Reviewer 1 above, benchmarking of these methods is warranted given differences in sample sizes and access to computational resources, which will promote true discoveries and minimize type I errors. The experimental

design and accompanying scripts described in this manuscript provide a foundation for future adoption of more scalable tools, such as FLARE or machine-learning-based methods for fast, accurate local ancestry inference.17,1817,18 We have added the following text to the discussion section to encourage both maximally inclusive study design and increased representation in future implementations of public references:

"Nonetheless, our inclusive study design increased statistical power in both our diverse downsampled subset and overall study population compared to a genetically homogeneous European-ancestry-only subset. This is significant given the predominantly European genetic architecture of GTEx and published GWAS—while these resources have been crucial for genomics discovery to date, work highlighting the limitations of genetically restricted samples and technologies developed based on those samples points to the necessity of new, more expansive approaches. This also aligns with current appeals in basic science and public health to promote equitable research benefiting all populations, rather than studies that may extend the health disparity gap.

We have also added the following text to the discussion:

"With the generation of new eQTL datasets from admixed populations, establishing best practices such as minimizing LD mismatch and using local ancestry estimates are needed to improve data standards, integration, and replication efforts."5,19

5. Minor note: please update your Figure S2 legend text to describe panel b.

Thank you for identifying this omission; we have updated it in the legend for Figure S2.

References
1. Liang, Y., Aguet, F., Barbeira, A.N., Ardlie, K., and Im, H.K. (2021). A scalable unified framework of total and allele-specific counts for cis-QTL, fine-mapping, and prediction. Nat. Commun. 12, 1424. 10.1038/s41467-021-21592-8.
2. Zhong, Y., Perera, M.A., and Gamazon, E.R. (2019). On using local ancestry to characterize the genetic architecture of human traits: genetic regulation of gene expression in multiethnic or admixed populations. Am. J. Hum. Genet. 104, 1097–1115. 10.1016/j.ajhg.2019.04.009.
3. Qin, H., Morris, N., Kang, S.J., Li, M., Tayo, B., Lyon, H., Hirschhorn, J., Cooper, R.S., and Zhu, X. (2010). Interrogating local population structure for fine mapping in genome-wide association studies. Bioinformatics 26, 2961–2968. 10.1093/bioinformatics/btq560.
4. Gay, N.R., Gloudemans, M., Antonio, M.L., Abell, N.S., Balliu, B., Park, Y., Martin, A.R., Musharoff, S., Rao, A.S., Aguet, F., et al. (2020). Impact of admixture and ancestry on eQTL analysis and GWAS colocalization in GTEx. Genome Biol. 21, 233. 10.1186/s13059-020-02113-0.
5. Bhattacharya, A., Hirbo, J.B., Zhou, D., Zhou, W., Zheng, J., Kanai, M., Global Biobank Meta-analysis Initiative, Pasaniuc, B., Gamazon, E.R., and Cox, N.J. (2022). Best practices for multi-ancestry, meta-analytic transcriptome-wide association studies: Lessons from the Global Biobank Meta-analysis Initiative. Cell Genomics 2. 10.1016/j.xgen.2022.100180.
6. Dey, K.K., Gazal, S., van de Geijn, B., Kim, S.S., Nasser, J., Engreitz, J.M., and Price, A.L. (2022). SNP-to-gene linking strategies reveal contributions of enhancer-related and candidate master-regulator genes to autoimmune disease. Cell Genomics 2. 10.1016/j.xgen.2022.100145.
7. Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B.E., Liu, X.S., and Raychaudhuri, S. (2013). Chromatin marks identify critical cell types for fine mapping complex trait variants. Nat. Genet. 45, 124–130. 10.1038/ng.2504.
8. Çalışkan, M., Manduchi, E., Rao, H.S., Segert, J.A., Beltrame, M.H., Trizzino, M., Park, Y., Baker, S.W., Chesi, A., Johnson, M.E., et al. (2019). Genetic and epigenetic fine mapping of complex trait associated loci in the human liver. Am. J. Hum. Genet. 105, 89–107. 10.1016/j.ajhg.2019.05.010.
9. Nora, E.P., Goloborodko, A., Valton, A.-L., Gibcus, J.H., Uebersohn, A., Abdennur, N.,

Dekker, J., Mirny, L.A., and Bruneau, B.G. (2017). Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. Cell 169, 930-944.e22. 10.1016/j.cell.2017.05.004.
10. Wen, X., Luca, F., and Pique-Regi, R. (2015). Cross-population joint analysis of eQTLs: fine mapping and functional annotation. PLoS Genet. 11, e1005176. 10.1371/journal.pgen.1005176.
11. Mosquera, J.V., Wong, D., Auguste, G., Turner, A.W., Hodonsky, C.J., Lino Cardenas, C.L., Theofilatos, K., Bos, M., Kavousi, M., Peyser, P., et al. (2022). Integrative single-cell meta-analysis reveals disease-relevant vascular cell states and markers in human atherosclerosis. BioRxiv. 10.1101/2022.10.24.513520.
12. Alsaigh, T., Evans, D., Frankel, D., and Torkamani, A. (2020). Decoding the transcriptome of atherosclerotic plaque at single-cell resolution. BioRxiv. 10.1101/2020.03.03.968123.
13. Wirka, R.C., Wagh, D., Paik, D.T., Pjanic, M., Nguyen, T., Miller, C.L., Kundu, R., Nagao, M., Coller, J., Koyano, T.K., et al. (2019). Atheroprotective roles of smooth muscle cell phenotypic modulation and the TCF21 disease gene as revealed by single-cell analysis. Nat. Med. 25, 1280–1289. 10.1038/s41591-019-0512-5.
14. Hu, Z., Liu, W., Hua, X., Chen, X., Chang, Y., Hu, Y., Xu, Z., and Song, J. (2021). Single-Cell Transcriptomic Atlas of Different Human Cardiac Arteries Identifies Cell Types Associated With Vascular Physiology. Arterioscler. Thromb. Vasc. Biol. 41, 1408–1427. 10.1161/ATVBAHA.120.315373.
15. Dougherty, M.L., Underwood, J.G., Nelson, B.J., Tseng, E., Munson, K.M., Penn, O., Nowakowski, T.J., Pollen, A.A., and Eichler, E.E. (2018). Transcriptional fates of human-specific segmental duplications in brain. Genome Res. 28, 1566–1576. 10.1101/gr.237610.118.
16. Hardwick, S.A., Hu, W., Joglekar, A., Fan, L., Collier, P.G., Foord, C., Balacco, J., Lanjewar, S., Sampson, M.M., Koopmans, F., et al. (2022). Single-nuclei isoform RNA sequencing unlocks barcoded exon connectivity in frozen brain tissue. Nat. Biotechnol. 40, 1082–1092. 10.1038/s41587-022-01231-3.
17. Browning, S.R., Waples, R.K., and Browning, B.L. (2023). Fast, accurate local ancestry inference with FLARE. Am. J. Hum. Genet. 110, 326–335. 10.1016/j.ajhg.2022.12.010.
18. Pearson, A., and Durbin, R. (2023). Local ancestry inference for complex population histories. BioRxiv. 10.1101/2023.03.06.529121.
19. Zhou, W., Kanai, M., Wu, K.-H.H., Rasheed, H., Tsuo, K., Hirbo, J.B., Wang, Y., Bhattacharya, A., Zhao, H., Namba, S., et al. (2022). Global Biobank Meta-analysis Initiative: Powering genetic discovery across human disease. Cell Genomics 2, 100192. 10.1016/j.xgen.2022.100192.

Referees' reports, second round of review

Reviewer #1: All previous points adequately addressed

Reviewer #2: Overall, the authors have satisfactorily responded to previous comments, which has significantly improved the quality of this manuscript. By including new analyses and added clarification in the text, utility to the broader community has been made much clearer. In particular, added information about how to select tools for analysis in populations with significant genetic admixture is timely and the sensitivity analyses provide helpful context, both in contrast to other similar study designs and for considering the use of diverse ancestries in genetic studies.

Authors' response to the second round of review

none