Interpreting coronary artery disease GWAS results: A functional genomics approach assessing biological significance

Katherine Hartmann[1*#a], Michał Seweryn[2] , Wolfgang Sadee[1]

1 Department of Cancer Biology and Genetics, Center for Pharmacogenomics, College of Medicine, The Ohio State University, Columbus, OH, USA

2 Biobank Lab, Department of Molecular Biophysics, University of Lodz, Poland

#a Current Address: Department of Internal Medicine, Santa Clara Valley Medical Center, San Jose, CA, USA

* Corresponding Author

E-mail: hartmann.kate@gmail.com (KH)

Abstract

Genome-wide association studies (GWAS) have implicated 58 loci in coronary artery disease (CAD). However, the biological basis for these associations, the relevant genes, and causative variants often remain uncertain. Since the vast majority of GWAS loci reside outside coding regions, most exert regulatory functions. Here we explore the complexity of each of these loci, using tissue specific RNA sequencing data from GTEx to identify genes that exhibit altered expression patterns in the context of GWAS-significant loci, expanding the list of candidate genes from the 75 currently annotated by GWAS to 245, with almost half of these transcripts being non-coding. Tissue specific allelic expression imbalance data, also from GTEx, allows us to uncover GWAS variants that mark functional variation in a locus, *e.g.*, rs7528419 residing in the *SORT1* locus, in liver specifically, and rs72689147 in the *GUYC1A1* locus, across a variety of tissues. We consider the GWAS variant rs1412444 in the LIPA locus in more detail as an example, probing tissue and transcript specific effects of genetic variation in the region. By evaluating linkage disequilibrium (LD) between tissue specific eQTLs, we reveal evidence for multiple functional variants within loci. We identify 3 variants (rs1412444, rs1051338, rs2250781) that when considered together, each improve the ability to account for LIPA gene expression, suggesting multiple interacting factors. These results refine the assignment of 58 GWAS loci to likely causative variants in a handful of cases and for the remainder help to re-prioritize associated genes and RNA isoforms, suggesting that ncRNAs maybe a relevant transcript in almost half of CAD GWAS results. Our findings support a multi-factorial system where a single variant can influence multiple genes and each genes is regulated by multiple variants.

Introduction

Genome-wide association studies (GWAS) have identified dozens of genetic variants (SNPs) associated with cardiovascular disease risk and related clinical phenotypes (*e.g.*, blood pressure, lipid levels) (1–3). However, these findings do not necessarily translate to understanding of heritability, likely because we do not fully understand the link between significant loci, causative genetic variants and complex phenotypes (4). Moreover, the functional variant and even the relevant gene close to a significant locus in many cases remain uncertain. The majority of statistically significant SNPs reside in non-coding regions with poorly defined biological functions and a complex architecture of multiple genes and transcripts (5). Gene assignment is largely based on proximity, usually with little consideration for non-coding transcripts in the locus or the possibility of chromatin looping that places distant regions in close proximity (6), with regulatory domains often interacting with multiple genomic target regions (9). Additionally, localization to non-coding regions means the mechanisms remain unknown as the function is not immediately obvious, while implicating epigenetics and other regulatory processes (5,7). This uncertainty limits the utility of GWAS findings. To interpret and refine GWAS results for coronary artery disease (CAD), we use RNA expression, in addition to physical position, to prioritize the variants and gene(s) most likely to be relevant.

Although largely thought of in a single SNP – single protein-coding gene paradigm, GWAS variants mark regions with various degrees of complexity often including several protein-coding and non-coding RNAs (ncRNAs). SNPs located within RNA exons may not only alter the protein sequence but also influence RNA structure and function in a transcript specific manner (8). Some of these GWAS loci consist of gene clusters that are coordinately regulated (9), and almost all include multiple RNA isoforms expressed from a given gene, including splice isoforms. Within such multi-gene regions, a single variant may affect more than one gene, both protein-coding and non-coding, via chromatin looping between multiple sites or by regulating DNA accessibility for the entire region (9,10). Therefore, a critical question for interpreting

GWAS associations is which gene(s), and what specific transcript(s), are affected within each significant locus.

The potential for multiple variants to affect a single gene is also critical to the interpretation of GWAS. Such interactions between variants, either linear or dynamic (epistasis) and dictated by linkage disequilibrium (LD), may remain hidden in GWAS because of the restrictive nature of multiple hypotheses corrections; however targeted analysis of loci reveals multiple interacting variants modulating gene expression (9,11,12). Failure to identify all main functional variants in a gene locus and their interactions results in false estimates of the genetic influence of a locus, and further impedes discovery of dynamic interactions that are sensitive to partial or confounded estimates (13–18).

Detailed analysis of RNA expression to evaluate GWAS results is increasingly employed to evaluate co-localization of GWAS and eQTL signals (19–21). However, most methods rely on the a *priori* assumptions that variants are independent of each other (*e.g.,* eCAVIAR), while COLOC assumes that there is only one functional variant per GWAS locus. These assumptions do not allow for a multifactorial system, where a single variant can influence multiple genes and each gene can be regulated by multiple variants. Accordingly, we search for overlap between variants marking GWAS associations and those marking eQTLs rather than using existing methods to co-localize signals. Although this approach limits our power to detect overlap as it requires a single variant appear as a marker in both GWAS and eQTL analysis, we posit it facilitates functional exploration of a multi-factorial system.

A recent CAD GWAS used 1000 genomes to impute insertions/deletions, rare variants and common variants that were not directly genotyped as part of a large-scale meta-analysis of 185 thousand cases and controls (1). While confirming 47 of the 48 previously identified loci, this study identified an additional 10 at genome-wide significance, bringing the total count of CAD associated loci to 58. Each of these loci are based on robust statistical associations for one or more SNPs in the locus. Furthermore, each locus has been assigned one or more

genes based largely on proximity as part of the GWAS annotation. We consider each of these 58 loci in detail, using QTL and position to re-prioritize candidate genes and focusing on a subset of loci, to begin resolving inherent complexities of genomic architecture.

Materials and Methods

Data

1000 Genomes

Genotypes for calculating LD between SNPs of interest were downloaded from: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/GRCh38_positions/. Individuals of the 'EUR' superpopulation were selected for LD calculations.

CATHeritization GENetics (CATHGEN)

Expression, genotypes, and clinical phenotypes were acquired via dbGaP Project #5358 (dbGaP accession phs0000703). Expression levels had been determined using Illumina HumanHT-12-v3 in RNA from whole blood. We considered variables recorded in pht003672: age (phv00197199), gender (phv00197207), hypercholesterolemia (phv00197204), smoking (phv00197208), number of diseased vessels (phv00197295), and history of myocardial infarction (MI) (phv00197212). We restricted analysis to Caucasians (race (phv00197206)) for sample size considerations (862 Caucasians; 259 African Americans). The approach developed here can be extended to other ethnic groups as these datasets become available. Data access was approved by the Ohio State University IRB (Protocol #2013H0096).

Genotype and Tissue Expression Project (GTEx)

Tissue-specific RNAseq data was acquired via dbGaP Project #5358 (dbGaP accession phs000424). For details see Lonsdale et al. and http://www.gtexportal.org/home/documentationPage (22). MI was defined as recorded history of

heart disease (MHHRTDIS) or heart attack (MHHRTATT). Data access was approved by the Ohio State University IRB (Protocol #2013H0096).

Gene Information

Transcripts, coding status, GO Ids, number of publications indexed in PubMed, gene/transcript expression, GWAS hits, GTEx eQTLs (expression quantitative trait loci) and sQTLs (splicing quantitative trait loci) including tissue specific expression, and allelic ratios in DNAse hypersensitivity sites were obtained for each gene using the package 'mglR' implemented in R (https://cran.r-project.org/web/packages/mglR/index.html). Protein-coding transcripts were defined as those annotated by BiomaRt as "IGC gene", "IGD gene", "IG gene", "IGJ gene", "IGLV gene", "IGM gene", "IGV gene", "IGZ gene", "nonsense_mediated_decay", "nontranslating CDS", "non stop decay", "polymorphic pseudogene", "TRC gene", "TRD gene", "TRJ gene", "protein_coding", "TEC". The remaining designations were considered non-coding and include "disrupted domain", "IGC pseudogene", "IGJ pseudogene", "IG pseudogene", "IGV pseudogene", "processed_pseudogene", "transcribed_processed_pseudogene", "transcribed unitary pseudogene", "transcribed_unprocessed_pseudogene", "translated processed pseudogene", "TRJ pseudogene", "unprocessed_pseudogene", "unitary_pseudogene", "3prime overlapping ncrna", "ambiguous orf", "antisense", "antisense RNA", "lincRNA", "ncrna host", "processed_transcript", "sense intronic", "sense overlapping", "lncRNA", "retained_intron", "miRNA", "miRNA_pseudogene", "miscRNA", "miscRNA_pseudogene", "Mt rRNA", "Mt tRNA", "rRNA", "scRNA", "snlRNA", "snoRNA", "snRNA", "tRNA", "tRNA_pseudogene", and "rRNA_pseudogene". A gene was considered non-coding only if all transcripts were non-coding.

Linkage Disequilibrium (LD)

R^2 was calculated for 1000 genomes 'EUR' super population using the 'ld' function from the package 'snpStats' implemented in R and using LDlink.

Association Testing

Additive logistic models to account for LIPA expression and MI using different combinations of variants were compared using ANOVA with a likelihood ratio test (LRT) implemented in R. Gender and age were included as covariates in models explaining LIPA gene expression, while sex, age, hypercholesterolemia, smoking, and number of diseased vessels were included as covariates in models explaining MI. Differences in LIPA expression between those with or without a history of MI were calculated using the wilcoxin.test function in R. Bonferroni multiple hypothesis correction was implemented.

Allelic Expression Imbalance (AEI)

Allelic RNA expression imbalance (AEI) was assessed using data from GTEx (phe000039.v1.GTEx_v8_ASE.expression-matrixfmt-ase.c1). Candidate variants were subsetted from each individual file, and the deviation of the “REF_RATIO” from the “NULL_RATIO” was plotted for each variant in a given tissue type. Tissue types with 5 or more samples were considered.

Results and discussion

Expanding candidate gene lists using QTL and position

As many functional variants marked by GWAS likely have regulatory functions affecting RNA expression or processing, the same SNPs appearing in GWAS may also mark expression Quantitative Trait Loci (eQTLs) or splicing Quantitative Trait Loci (sQTLs) for their target gene. To assign GWAS hits to target genes, we determine for each of the GWAS SNPs whether it appears as an eQTL or sQTL reported by GTEx, searching all available tissues. Recognizing that often multiple SNPs exist over a genomic region as significant GWAS hits, we consider each one individually in assigning candidate genes and separately assess concordance. We opt

not to use COLOC and other existing tools that search for overlapping signal between GWAS variants and QTLs because they make assumptions about the genetic model that are not in line with the multi-factorial system we test here (23); namely, these methods assume a single causative variant or that each variant acts independently. Instead, although we recognize it limits the overlap we are able to detect and biases our sample to variants that are ideal markers (i.e. frequent), we search for exact matches between GWAS and QTL marker variants. In addition to evaluating associations with gene expression and splicing, we consider the physical position of each GWAS variant as SNPs within the RNA sequence are expected to impact RNA folding, stability, function, etc. Specifically, we consider the corresponding gene for any transcript that physically overlaps the GWAS variant regardless of strand, thus incorporating coding, non-coding, and antisense genes. Using these three approaches (cis-eQTLs, cis-sQTLs, position), we expand the list of potential candidate genes for the 58 GWAS loci from 75 to 245 (Fig 1A, S1 File, comprehensive table is included in S3 File, S1 Fig).

Fig 1. Summary of CAD GWAS loci. (**A**) For each of the 58 loci identified by GWAS, number of candidate genes annotated by GWAS and additional genes added by eQTL, then sQTL, and finally position based reprioritization, if implicating genes other than those annotated previously by GWAS (See S1 Fig for further details about the approach and S3 File for a comprehensive table). Tier 1 (n = 7) denotes those loci where a GWAS annotated gene is supported by QTL-based re-prioritization or position and no other candidate genes are introduced; Tier 2 (n = 50) where QTL-based reprioritization or position introduces new associated genes while supporting all candidates at this locus (Tier2A), only some including the GWAS gene (Tier2B) or new genes except the GWAS genes (Tier2C); and Tier 3 (n = 1) where no eQTLs or sQTLs are identified and no gene physically overlaps the SNP, accordingly annotation by GWAS is not supported and no other genes are implicated. (**B**) For each of the 245 candidate genes displayed along the x-axis (names available in S1 File), the number of

199 transcripts assigned to the gene, the number of antisense transcripts (note: antisense genes are
200 not included among the 245 candidate genes unless their expression is associated with or they
201 physically overlap a GWAS variant), GO terms, Papers indexed in PubMed, *cis*-eQTLs and
202 sQTLs published in v8 of GTEx. Blue bar highlights those genes with only non-coding
203 transcripts.
204
205 In an effort to identify those loci where a target gene(s) is clearly supported by functional
206 markers, we consider the agreement between the gene assignment given by GWAS studies
207 and that derived by eQTL and sQTL analysis as well as by physical position. We group each of
208 the 58 GWAS loci as follows: GWAS annotation is supported by QTL-based re-prioritization or
209 position and no other candidate genes are introduced (Tier 1); QTL-based reprioritization or
210 position introduces new genes, while supporting all (Tier2A), some (Tier2B), or none (Tier2C) of
211 the genes annotated by GWAS so that multiple genes are implicated; no eQTLs or sQTLs are
212 identified and no gene or annotated RNA transcript physically overlaps the SNP, accordingly
213 annotation by GWAS is not supported (but also not negated) and no other genes are implicated
214 (Tier 3), see Fig 1, S2 Fig, S3 File.
215 In some instances, the number of candidate genes implicated changes substantially with
216 the particular GWAS SNP considered (S3 File). Nikpay et al. report both the SNP originally
217 identified in CAD GWAS studies and the most promising SNP in the same locus they identified
218 in a large-scale meta-analysis using 1000 genomes imputation to incorporate variants with
219 lower minor allele frequency (1). We consider both in our analysis, and for 11 of 32 loci with
220 multiple SNPs find the tier classification changes depending on the SNP considered (S3 File).
221 For example, locus 10 - rs2252641 and rs17678683 (ZEB2, AC074093.1), rs2252641, which
222 was implicated in historic GWAS studies, is annotated as ZEB2 and AC074093.1, but based on
223 QTL and position is not associated with any gene and would be considered as Tier3 while
224 rs17678683, identified by Nikpay et al. is an eQTL in skeletal muscle for ZEB2 and an sQTL in

subcutaneous adipose also for ZEB2 and would thus be considered as Tier1. The two SNPs have an $R^2 < 0.2$.

This discordance between variants in the same locus may represent a weakness of our approach that looks for overlap with an individual SNP rather than considering 'colocalization' more broadly. However, it occurs almost exclusively among SNPs that are in relatively poor LD (frequently $R^2 < 0.2$) and would not be expected to serve as good markers for one another nor be part of the same haplotype. Alternatively, this discordance may represent multiple functional variants in the locus with different target genes. Regardless, our results emphasize the complexity in elucidating functional variants from GWAS results.

Tier 1: No new candidate genes introduced – GWAS annotation supported

For 7 loci, QTL-based reprioritization and/or position supports the GWAS annotation without introducing new candidate genes, supporting all or some of the gene(s) annotated by GWAS: locus 16 - rs6903956 (ADTRP), locus 32 - rs11203042 and rs1412444 (LIPA), locus 38 - rs9319428 (FLT1), locus 42 - rs17514846 (FURIN, FES), locus 54 - rs7212798 (BCAS3), locus 57 - rs11830157 (KSR2), and locus 8 - rs6544713 (ABCG8; however ABCG5 also annotated by GWAS is not supported by QTL or position) (Table 1).

Table 1. Tier 1 CAD GWAS loci

Locus	SNP	OR	Risk Allele (Freq)	Gene	eQTL Tissue(s)	sQTL Tissue(s)	Position
16	rs6903956	1.65[a] (1.44-1.90)	A (0.08[a])	ADTRP		Testis	ADTRP (intron)

32	rs11203042	1.04 (1.02-1.06)	T (0.45)	LIPA	Adipose (subq) Adipose (visceral) Colon (transverse) Heart (atrium) Lung Pancreas Skin (sun exp) Spleen Thyroid Blood	Adipose (subq) Fibroblasts Lung	LIPA (intron)
32	rs1412444	1.07 (1.05-1.09)	T (0.37)	LIPA	Adipose (subq) Adipose (visceral) Adrenal Gland Artery (aorta) Brain (cerebellum) Colon (sigmoid) Colon (transverse) Heart (atrium) Heart (LV) Lung Skeletal Muscle Nerve Pancreas Skin (not sun exp) Skin (sun exposed) Spleen Stomach Thyroid Blood	Adipose (subq) Adipose (visceral) Adrenal Gland Artery (aorta) Artery (tibial) Brain (spinal cord) Breast Fibroblasts Lymphocytes Lung Tibial Nerve Pancreas Skin (sun exposed) Small Intestine Spleen Stomach Blood	LIPA (intron)
38	rs9319428	1.04 (1.02-1.06)	A (0.31)	FLT1	Nerve (tibial)		FLT1 (intron)
42	rs17514846	1.05 (1.03-1.07)	A (0.44)	FES	Adipose (subq) Adipose (visceral) Adrenal Gland Artery (aorta) Artery (tibial) Fibroblast Colon (transverse) Esophagus (musc.) Heart (atrium) Lung Nerve (tibial) Pancreas Pituitary Prostate Skin (not sun exp) Skin (sun exposed) Stomach Thyroid Blood	Adipose (subq) Adipose (visceral) Artery (aorta) Artery (tibial) Breast Fibroblasts Colon (sigmoid) Esophagus (GEJ) Esophagus (musc.) Heart (atrium) Heart (LV) Lung Salivary Gland Nerve (tibial) Prostate Skin (not sun exp) Skin (sun exposed) Small Intestine Spleen Thyroid Blood	FURIN (intron)
				FURIN	Artery (aorta) Artery (tibial) Esophagus		
54	rs7212798			BCAS3			BCAS3 (intron)
57	rs11830157			KSR2			KSR2

08[b]	rs6544713	1.05 (1.03-1.07)	T (0.32)	ABCG8	Colon (transverse)		(intron) ABCG8 (intron)

Tissue names in grey font indicate GWAS SNP is associated with a decrease in gene expression (eQTL) or normalized intron-excision ratio (sQTL), while those in black font are associated with increased expression/normalized intron-excision ratio.

[a] values reported from original publication (24) in Han Chinese population. rs6903956 was not significant in Nikpay et al. (1).
[b] ABCG8 and ABCG5 were both annotated by GWAS. ABCG5 was not supported by QTL or position

For four of the loci (16-ADTRP, 32-LIPA, 38-FLT1, 42-FURIN, 8-ABCG8), GWAS annotation of candidate gene assignment is supported by both QTL and position. In one instance, locus 42 - rs17514846 (FURIN, FES), more than one gene is annotated by GWAS and supported by our reprioritization. rs17514846, which falls in an intron of FURIN, serves as an eQTL and an sQTL for FES in 23 tissues and an eQTL for FURIN in 3 tissues, two of which (aorta and tibial artery) overlap with FES. In aorta and tibial artery, rs17514846 is associated with decreased expression of FES as opposed to increased expression of FURIN – a possible example of competing interactions between regulatory and promoter regions. Evidence for multiple candidate genes in a locus may represent a paradigm in which a single SNP exerts an impact through more than one gene.

In some instances the same variant in the same tissue is associated with both expression and splicing. For example, rs141244 in blood is associated with increased expression of LIPA and decreased splicing, a scenario that is consistent with greater stability of the un-spliced transcript. Thus, in considering potential mechanisms of action for the variant, it is important to evaluate not only the implications of increased levels of LIPA mRNA, but also increased levels of the un-spliced transcript.

Tier 2: New candidate genes implicated

Variants in 50 loci are associated with expression of one or more genes or physically overlap with another gene in addition to all (39 loci), some (7 loci), or none (4 loci) of the genes annotated by GWAS. Loci where additional candidate genes are introduced are classified as Tier 2 (S3 File). Candidate genes for these 50 GWAS loci are expanded by an average of 4.3 genes per locus for a total of 170 genes: 116 from eQTL based reprioritization, 17 from sQTL based reprioritization, 5 from physical position, and 32 from some combination of these features (S3 Fig).

While about a third of the loci (21) have two or fewer candidate genes, others have substantially more: e.g., locus 33 - rs12413409 and rs11191416 (CYP17A1-CNNM2-NT5C2) are associated with expression of twelve genes: ARL3, AS3MT, ATP5MD, BORCS7, CALHM2, CNNM2, CYP17A1-AS1, MARCKSL1P1, MFSD13A, NT5C2, SFXN2, and WBP1L. Importantly, these multi-gene eQTLs cannot be explained solely by co-expression between genes. These eQTLs are often associated with expression of different genes in different tissues and for those associations in the same tissue the genes are not necessarily co-expressed (25). Possible mechanisms accounting for these relationships include a co-regulated gene cluster with master regulator regions affecting all genes in that region (26,27). While we cannot exclude trans-acting effects, these are typically smaller in size; allelic expression imbalance can serve to distinguish between *cis* and *trans* effects, but robust RNA allelic ratios are often not available in GTEx.

Notably, ncRNAs are candidate genes for 33 of the 58 loci expanded from 6 loci prior to re-prioritization. For no loci are all candidate genes non-coding. The greatest percent of non-coding candidate genes (80%) is observed for locus 20 - rs12190287 and rs12202017 (TCF21). The protein-coding gene TCF21 annotated by GWAS is supported by eQTLs in 24 different tissues while 4 additional non-coding genes are introduced based on eQTL and sQTL analysis. The candidate ncRNAs are located within a 500KB region, with TARID being antisense to TCF21, and like many non-coding genes generally have lower expression levels.

297 For Tier 2C loci, there is no evidence to support the GWAS annotation. For example,
298 locus 46 - rs1122608 and rs56289821 LDLR is annotated by GWAS, a gene well-recognized for
299 its role in lipid metabolism; yet, rs1122608 falls within an intron of SMARCA4 and is both an
300 eQTL and sQTL for SMARCA4 as well as an eQTL for CARM1 and YIPF2 but not LDLR. The
301 alternative SNP identified by GWAS, rs56289821, also does not point to LDLR but rather
302 implicates RGL3, SLC44A2, and again SMARCA4. These 4 Tier 2C loci critically require future
303 work, both mechanistic and computational, to explore relevant gene targets.
304

305 ## Tier 3: No genes implicated

306 The remaining GWAS locus, locus 55 - rs663129 (MC4R, PMAIP1), classified as Tier 3,
307 did not show any association with expression of nearby genes and is not physically overlapping
308 any transcripts (S3 File). This locus and 3 others (locus 27 - rs2954029 (TRIB1), locus 54 -
309 rs7212798 (BCAS3), and locus 57 - rs11830157 (KSR2)) that are without any eQTL
310 associations may have more subtle or context-dependent effects on gene expression that
311 remain undetectable in GTEx. In particular, non-polyadenylated transcripts are not in GTEx as
312 poly-dT priming was used, leaving countless ncRNAs as additional candidates. Furthermore,
313 these SNPs may affect gene expression in *trans* (although we do not find such evidence in the
314 GTEx *trans*-eQTL dataset) or exert their effect without altering RNA levels measured by
315 RNAseq (*e.g.* by controlling the chromatin structure or co-translationally alter RNA
316 modifications). Additionally, variants affecting RNA functions and processing (structural RNA
317 SNPs) (8,28), may not be visible as eQTLs, or they may selectively affect translation by
318 changing polysomal loading (29).
319 Given GWAS variants are expected to mark functional variants rather than themselves
320 being functional, we test SNPs within a 1MB window in LD ($R^2 > 0.8$) with each of the 4 GWAS
321 variants lacking annotations, expanding the number of SNPs to 200. Using this approach, we

find significant eQTLs, but no significant sQTLs, for three of the four loci. For locus 57, we were unable to find additional candidate SNPs with an $R^2 > 0.8$ to mark the haplotype.

For locus 55, without any candidate genes identified by our original re-prioritization, eighteen variants (mean R^2 0.95 with the GWAS variant rs663129) are eQTLs in testis for MC4R, also annotated by GWAS. Seven of these variants have an R^2 of 1 with the GWAS variant. Five of these 18 SNPs fall within a non-coding RNA (AC090771.2) approximately 200kb downstream of MC4R.

For locus 27, two variants (rs10808546 and rs2954031 with R^2 of 0.87 and 0.91 with the GWAS variant rs2954029) are eQTLs in lung for a non-coding RNA (AC100858.3) almost 400kb downstream and on the opposite strand to TRIB1. These two SNPs and 15 others in high LD with the GWAS variant fall within an intron of a different non-coding RNA (AC091114.1), located immediately upstream of TRIB1.

For locus 54, we find 83 variants (mean R^2 0.96 with the GWAS variant rs7212798) are eQTLs in tibial artery and nerve for genes BCAS3, AC005884.1, and RP11-136H19.1. Five of these variants with an R^2 of 1 with the GWAS variant serve as eQTLs for AC005884.1 in tibial nerve, and are physically located within BCAS3, the protein-coding gene that has been annotated by GWAS. A different five variants are eQTLs for BCAS3; they are identified in tibial artery and have a mean R^2 with the GWAS variant of 0.81. Such long-range high LD values imply evolutionarily conserved LD blocks with biological selectable functions.

Thus, by expanding the candidate SNP list to include other variants in strong LD with that identified by GWAS over long genomic distances, we find additional candidate genes supported by eQTL and physical position.

Survey of CAD GWAS loci

346 The genomic loci for each these 245 candidate genes often harbor multiple protein-
347 coding and non-coding transcripts arranged on both the sense and antisense strands (S2 File).
348 They express an average of 9 transcripts per gene and a maximum of 189 (TEX41- locus 10 -
349 rs2252641, rs17678683), with 47% of all transcripts being non-coding (Fig 1B). More than half
350 of the gene loci (161) also contain one or more antisense genes (i.e., located on the opposite
351 strand and overlapping).
352 Using GTEx data, we find RNA expression is associated with one or more genetic
353 variant in the locus with a median of 4 eQTLs (max 32230) among 50% of the 245 candidate
354 genes in 48 different tissues (Fig 1B).
355 With a median of 26 publications and a maximum of 27,497 (APOE), only a handful of
356 these 245 candidate genes have been well studied to date (Fig 1B). Twenty percent (51) of
357 genes do not have a single paper indexed in PubMed. There are on average 20 gene ontology
358 (GO) terms, which are manually curated based on the literature, assigned to each gene;
359 however, 62 (25%) of the candidate genes have no associated GO terms. We find those genes
360 without GO terms and with limited publications do not have fewer markers of functionality
361 (eQTLs, splicing QTLs, etc.), but are almost exclusively non-coding, indicative of a recognized
362 bias in the literature toward protein-coding genes (Fig 1).
363 Each implicated locus displays an astoundingly complex architecture with multiple
364 candidate genes implicated by RNA expression and physical location, each with a number of
365 overlapping coding and non-coding transcripts including those in antisense orientation. The
366 complexity of these loci emphasizes the need for targeted molecular studies and computational
367 approaches to determine the relevant gene and transcript(s). The distribution of PubMed
368 articles and GO ids across candidate genes suggests that this targeted work has touched on
369 only a handful of genes thus far, with more recent studies beginning to focus on ‘neglected’
370 CAD candidate genes (30).
371

Allelic RNA expression imbalance reveals functional variation

To evaluate potential functionality for each of the 58 GWAS loci, we ask whether each candidate SNP is associated with allelic expression imbalance (AEI), a specific indicator of *cis*-acting regulatory variation. Whereas comparing expression of the two alleles at a heterozygous variant, various external/*trans*-acting influences on gene expression are shared and the *cis*-acting effect of the heterozygous variant can be isolated. In the absence of a functional variant altering RNA expression, the anticipated distribution between the alleles is 0.5 (ratio=1)(8,31).

Using data released by GTEx, we evaluate AEI at each of 104 candidate variants across 54 tissue types. Only 55 of the SNPs are represented in the data. The remainder likely are in intergenic regions and poorly captured by RNA sequencing, while obtaining accurate AEI ratios requires rather robust expression (>30 RPM)(32). Of the 55, many are present in only a few samples making it difficult to infer differential expression. However, several SNPs show surprisingly robust data – thousands of samples and counts for each allele in the dozens-hundreds. A majority of these SNPs fail to reveal allelic expression imbalance, with near normal distribution of deviation from the expected ratio, suggesting no correlation between the GWAS variant and allelic expression imbalance. This implies that the GWAS candidate SNPs represented in the data are actually relatively poor markers for functional *cis*-acting variants in the locus; however, splicing events generating RNA isoforms with similar turnover are one example where allelic expression imbalance would fail.

A number of SNPs do display consistent allelic expression imbalance (Fig 2). Locus 3 – rs7528419 (SORT1), which falls in the 3'UTR of CELSR2 exhibits AEI in 53/57 liver samples. Overall low expression of CELSR2 in liver tissue, however means that these ratios are for the most part based on low coverage (median total count 13). Despite the relative consistency from sample to sample, large allelic ratios derived from relatively low counts, as observed here,

raises suspicion for systemic sources of bias, *e.g.* preferential amplification of one allele. To evaluate this further, we considered allelic ratios at nearby SNPs in strong LD ($R^2 > 0.9$) and weak LD ($R^2 < 0.1$). As these SNPs are co-located, systemic sources of bias should affect all SNPs in the locus while 'true' biological AEI would be expected only for those variants in strong LD with a functional SNP. We observe AEI for those SNPs in strong LD with the GWAS marker, but not for those in the same region in weak LD, a pattern that is suggestive of 'true' biological AEI and a functional cis-acting variant.

Fig 2. Allelic expression imbalance at GWAS variants mark functional SNPs. Deviation in the observed from the expected ratio for individuals heterozygous for given GWAS variant. (A) Locus 3 – rs7528419 (SORT1) exhibits AEI in 53/57 liver samples. Subcutaneous adipose, also shown, demonstrates near normal distribution of deviation from the expected allelic ratio and is representative of the 46 other tissues with at least 5 samples. (B) Locus 14 - rs72689147 (GUCY1A3) exhibits AEI in 114/121 samples across 10 different tissues.

Importantly, even one sample without AEI suggests the variant might itself not be functional but rather in high LD with a functional variant and serving as a marker. With only a few samples not exhibiting AEI, rs7528419 can be considered an excellent marker in tight LD with the functional SNP. Furthermore, that this pattern is only found in liver suggests that the regulatory variant is tissue specific. In contrast, the bidirectional ratios observed in adipose tissue suggests that either the marker SNP is in low LD with a functional variant, or low expression is leading to ratio deviations from unity in a random fashion. The counts in adipose are relatively low (median total count 11), however it is not substantially different from liver with a median count of 13.

The Locus 14 SNP rs72689147 (*GUCY1A3*), which falls within an intron of GUCY1A3, exhibits AEI in 114/121 samples across 10 different tissues. Again, this SNP does not appear to

be functional as not all samples display AEI, but it is a robust marker. While located in an intron, expression is sufficient to extract allelic ratios; as these are consistently below unity, this results suggests a gain of function.

Resolving number of signals in a locus using LD

Focusing on the 7 loci where eQTL-based reprioritization pointed to a single gene as well as the two examples of AEI discussed above, we find dozens of other significant eQTLs for each gene. To determine whether these eQTLs represent one or more functional variants, we plot the effect size (beta) of the variant on RNA expression for each eQTL against its LD (R^2) with the top scoring (most significant p-value) eQTL in each tissue where eQTLs are detectable. Assuming one functional variant in the locus, the beta for each eQTL should correlate with its R^2 relative to the highest scoring SNP (33).

This approach reveals that the observed eQTLs for a gene often represent more than one regulatory variant, with the exception of FLT1 in Tibial Nerve – represented by only one cluster of variants marked by the GWAS SNP (Fig 3). This result is critical to the correct interpretation of GWAS that would otherwise rely on a single variant rather than considering the combined effect of more than one causative variant.

Fig 3. Number of eQTL signals. Correlation plots show absolute value of beta for variant effects on RNA expression versus R^2 with the top eQTL (most significant p-value), including all significant eQTLs in the given gene-tissue combination. Blue dots represent the top eQTL (most significant p-value), red dots represent GWAS variant(s). (A) FLT1 in Tibial Nerve: eQTLs are accounted for by a single eQTL marked by the GWAS variant (all eQTLs display a linear correlation with R^2). CELSR2 (liver), GUCY1A3 (tibial artery), and LIPA (blood), correlation between beta and R^2 suggests multiple functional variants. (B) At least three distinct

LD blocks represented by LIPA eQTLs in whole blood. Correlations are shown left to right between the absolute value of beta and R^2 with rs142444 (GWAS SNP), rs1051338, or rs2250781. Tightly linked SNPs (D' > 0.9; R^2 > 0.9) are shown in the same color.

As an example, we consider the number of distinct eQTLs needed to maximally account for LIPA expression in blood. The most significant eQTL consists of a group of SNPs in high LD marked by the GWAS variant (red dot in Fig 3), while two additional clusters of SNPs (marked by rs1051338 and rs2250781) have equally or even more robust beta and p-values but show relatively poor linkage with the GWAS cluster (R^2 ~ 0.5) (Fig 3B). These SNPs are more significant eQTLs than predicted by their LD with the trait-associated variant and may mark additional functional variants in the locus. To test the significance of any additional regulatory variants, we used a separate dataset (CATHGEN) to evaluate whether including an additional marker variant in a regression model improves the ability to account for LIPA expression in blood. Including additional markers improved the eQTL model, while adding a marker in strong LD with the original variant did not (Table 2), indicating there are likely multiple functional variants, incompletely represented by the GWAS variant alone, that contribute to LIPA expression in blood.

Table 2. Assessing multiple regulatory variants for LIPA

Variable of interest	ANOVA p-value	Model 1	Model 2
rs1412444	8.8e-16	XP ~ sex + age	XP ~ rs1412444 + sex + age
rs13332328	8.8e-16	XP ~ sex + age	XP ~ rs13332328 + sex + age
rs1051338	8.8e-16	XP ~ sex + age	XP ~ rs1051338 + sex + age
rs2250781	8.8e-16	XP ~ sex + age	XP ~ rs2250781 + sex + age
rs1412444 *in context of* rs13332328	1.0	XP ~ rs1412444 + sex + age	XP ~ rs1412444 + rs13332328+ sex + age

rs1412444 *in context of* rs1051338	0.23	XP ~ rs1412444 + sex + age	XP ~ rs1412444 + rs1051338 + sex + age
rs1412444 *in context of* rs2250781	0.04	XP ~ rs1412444 + sex + age	XP ~ rs1412444 + rs2250781 + sex + age
rs1412444 & rs2250781 *in context of* rs1051338	0.19	XP ~ rs1412444 + rs2250781 + sex + age	XP ~ rs1412444 + rs2250781 + rs1051338 + sex + age
rs1412444 & rs1051338 *in context of* rs2250781	0.04	XP ~ rs1412444 + rs1051338 + sex + age	XP ~ rs1412444 + rs1051338 + rs2250781 + sex + age
rs1412444	1e-3	MI ~ covariates	MI ~ rs1412444 + covariates
rs13332328	1e-3	MI ~ covariates	MI ~ rs13332328 +covariates
rs1051338	6e-4	MI ~ covariates	MI ~ rs1051338 + covariates
rs2250781	4e-4	MI ~ covariates	MI ~ rs2250781 + covariates
rs1412444 *in context of* rs13332328	0.79	MI ~ rs1412444 + covariates	MI ~ rs1412444 + rs13332328+ covariates
rs1412444 *in context of* rs1051338	0.36	MI ~ rs1412444 + covariates	MI ~ rs1412444 + rs1051338 + covariates
rs1412444 *in context of* rs2250781	0.56	MI ~ rs1412444 + covariates	MI ~ rs1412444 + rs2250781 + covariates
rs1412444 & rs2250781 *in context of* rs1051338	0.11	MI ~ rs1412444 + rs2250781 + covariates	MI ~ rs1412444 + rs2250781 + rs1051338 + covariates
rs1412444 & rs1051338 *in context of* rs2250781	0.17	MI ~ rs1412444 + rs1051338 + covariates	MI ~ rs1412444 + rs1051338 + rs2250781 + covariates

ANOVA comparing ability of models with different SNP combinations to account for LIPA expression and MI. Covariates in CATHGEN include sex, age, hypercholesterolemia, smoking, and number of diseased vessels.

Testing these additional variants with MI instead of LIPA expression did not yield significant associations (Table 2). However, LIPA expression itself is not associated with MI except when rs1412444 is homozygous minor, which may explain the discrepancy. In looking separately at the associations between the GWAS hit and LIPA expression and the GWAS hit and MI, we find that rs1412444 is associated with increased risk of MI and increased expression of LIPA, but counterintuitively those with two minor alleles and MI exhibit lower rather than

478 higher expression, a pattern that also holds in GTEx although it is only statistically significant in
479 CATHGEN (Fig 4, S4 Figure).
480
481 **Fig 4. LIPA expression, MI, and genotype**. Comparison of LIPA expression in
482 CATHGEN for those with and without MI based on rs142444 genotype. LIPA exhibits higher
483 expression only in those without MI in the homozygous minor group (p-value = 0.02).
484

485 Context – Tissue & Transcript specific eQTLs

486 Genetic variation exists and functions within a context – the surrounding sequence, the
487 tissue type and its preferred transcription factors, etc. In an effort to resolve the functional
488 variation behind statistical associations observed in GWAS, it is essential to consider these
489 contexts. As highlighted by the tissue specific AEI patterns above, if these relationships are not
490 considered in a context specific manner (*e.g.*, on a tissue by tissue basis), many robust effects
491 will remain hidden. In an effort to evaluate some of these contextual features, we consider
492 tissue and transcript specific eQTLs.
493 eQTL analysis may focus the search on a relevant tissue. However, eQTLs are
494 detectable only where expression and sample size are sufficiently high; accordingly tissue-
495 specific differences in eQTLs reflect overall patterns of tissue selective expression and sample
496 size, in addition to the influence of genetic variation.
497 To consider how eQTLs for a given gene compare across different tissues, we cluster
498 genome-wide significant eQTLs reported by GTEx for LIPA in a heatmap organized by their
499 pairwise LD (R^2), using a colored bar at the top of the heatmap to denote tissue type (Fig. 5A).
500 eQTL SNPs generally cluster by tissue, suggesting distinct regulatory variants in different
501 tissues. However, there are two LD blocks that contain eQTLs in more than half of tissues
502 indicative of genetic variation that acts across different tissue types. Variants detected by

GWAS for LIPA appear as a significant eQTLs in a subset of tissues (Table 1), some of which fit with our understanding of CAD pathology (heart, aorta, adipose), others suggest as yet unexplained biological consequences (spleen, pancreas).

Fig 5. Tissue and transcript specific eQTLs for LIPA. (A) Heatmap of LD for those SNPs reported by GTEx as genome-wide significant eQTLs for LIPA. Lighter-colored squares in the heatmap represent LD blocks, with SNPs clustered by R^2 and not by genomic position. Colored bars at top eQTLs in each tissue with more significant p-values denoted by darker color. (B) Local Manhattan plot for *LIPA* in blood. Position is denoted on the x-axis; black bar indicates gene start and stop. Negative log(p-value) for association with RNA expression is denoted on the y-axis. Each transcript and the gene overall are marked by a different color (see legend). Inset zooms in on the regions carrying rs1412444, rs1051338, rs2250781 identified as distinct top scoring eQTLs each contributing to *LIPA* gene expression (see Fig 3).

To consider transcript-specific eQTLs more specifically, for 1mb upstream and downstream of LIPA, we calculated the association between each SNP and expression of both overall mRNA and individual isofoms. The resulting local Manhattan plot displays the association p-values for LIPA transcripts in blood (Fig 4B). Five transcripts with no expression (zero-values) in 90% or more of individuals were excluded (ENST00000428800, ENST00000282673, ENST00000487618, ENST00000463623, ENST00000489359). Whereas the overall eQTL patterns are similar between the gene and individual transcripts, divergent SNPs indicate the presence of transcript-specific eQTLs. Remarkably, robust eQTLs distribute over the entire 2 MB region, with the most distant eQTLs associated with individual transcripts rather than overall gene expression. The most significant eQTLs form a peak of multiple SNPs close to the transcribed region, topped by an eQTL for the gene level summary of expression

(ENSG00000107798) rather than any specific isoform. Tissue selectivity and functionality of each transcript requires separate studies.

Conclusions

We consider each of 58 loci implicated in CAD by GWAS to better understand the biological meaning of the underlying statistical associations. In evaluating each of these loci, we find numerous candidate genes that were not included in the original annotation by GWAS. Many of these are non-coding. Non-coding RNAs, now well-recognized for their role as regulators, have historically been dismissed and continue to be difficult to study, a trend that is apparent in their poor representation in the literature, among GO annotations, and as annotated by GWAS (26). We find no evidence to suggest these non-coding RNAs are less likely to account for the observed associations in GWAS and would advocate for their inclusion in further mechanistic and computational work examining these loci. In addition to broadening candidate gene lists to include non-coding transcripts, we would urge reconsideration of current assignments, especially for those loci categorized as Tier2C where expression, splicing, and physical position do not support the gene annotated by GWAS. LDLR is a particularly prominent example. Given our understanding of the critical role lipid metabolism plays in CAD, it is counterintuitive not to assign a CAD GWAS variant to LDLR when it lies within 15kb of the *LDLR* locus (34). However, RNA expression and splicing data do not support this annotation, instead supporting the notion that such genetic variation affects the function of other nearby genes including SMARCA4, CARM1, YIPF2, RGL3, SLC44A2 (29).

Using allelic ratios built from tissue-specific RNA sequencing data available through GTEx, we were able to identify two loci where the GWAS variant served as a robust marker for a functional *cis*-acting regulatory variant. Locus 3 – rs7528419 (SORT1) falls in the 3'UTR of CELSR2, exhibits AEI exclusively in liver, and is in nearly perfect LD with rs12740374 which

was shown by Musunuru et al. through a series of molecular experiments to create a C/EBP binding site increasing expression of SORT1, a multiligand sorting receptor which they concomitantly showed to be associated with LDL-C and VLDL levels (35). This work revealed a single functional variant for a single target gene with a substantial effect size, the authors estimated a ~40% difference in MI risk. Our work suggests additional eQTLs not explained by their LD with the LD block marked by GWAS variant rs7528419. As we begin to identify functional variation behind GWAS associations, an important next step will be resolving additional functional variants within the loci that may modify these associations and better account for disease risk (36).

This work emphasizes that the linear presentation of GWAS results as a single variant tied to a single gene fails to capture the complexity of these loci. Many loci contain several SNPs identified by GWAS, and for each of these, multiple candidate genes are implicated by RNA expression and splicing associations as well as physical proximity. LD alone rarely accounts for the observed eQTLs, suggesting multiple functional variants within these loci. Extending the search for functional variants implicated by GWAS over a larger genomic region (1 Mb) can further reveal novel candidate loci. Although some GWAS associations may ultimately implicate single variants that alter expression of individual genes, this work indicates that true genetic effect size of a gene locus is accounted for by a multi-factorial system that allows for multiple functional variants regulating one or more genes. The approach presented here must be expanded to include functional variants that are undetectable by RNAseq of whole tissues, including cell type specific expression, effect on RNA-protein interactions, distribution in sub-cellular domains, alteration of translational processes, and of course variants that change protein functions.

Acknowledgements

This study was supported by National Institutes of Health National Institute of General Medical Science Pharmacogenetics Research Network [Grant U01 GM092655] and the National Center for Advancing Translational Sciences [TL1 TR001069].

The GTEx Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from: the GTEx Portal and dbGaP accession number phs000424.

For CATHGEN, clinical data originated from the Duke Databank for Cardiovascular Disease (DDCD) and biological samples originated from the Duke Cardiac CATHeterization (CATHGEN) study. Funding support for the Genetic Mediators of Metabolic CVD Risk was provided by NHLBI grant RC2 HL101621 (William E. Kraus). The data used for the analyses described in this manuscript were obtained from the dbGaP accession number phs0000703.v1.p1.

Computing time provided by the Ohio Supercomputer Center, GRANT #: PAS0885-2 and the Prometheus Cyfronet AGH.

References

1. Nikpay M, Goel A, Won H-H, Hall LM, Willenborg C, Kanoni S, et al. A comprehensive 1000 Genomes–based genome-wide association meta-analysis of coronary artery disease. Nat Genet [Internet]. 2015 Sep 7 [cited 2016 Jun 28];47(10):1121–30. Available from: http://www.nature.com/doifinder/10.1038/ng.3396
2. Inouye M, Abraham G, Nelson CP, Wood AM, Sweeting MJ, Dudbridge F, et al. Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults: Implications for Primary Prevention. J Am Coll Cardiol [Internet]. 2018 Oct 16 [cited 2020 Oct 22];72(16):1883–93. Available from: https://doi.org/10.1016/j.jacc.2018.07.079

603 3. Musunuru K, Kathiresan S. Genetics of Common, Complex Coronary Artery Disease. Vol.
604 177, Cell. Cell Press; 2019. p. 132–45.
605 4. Nurnberg ST, Zhang H, Hand NJ, Bauer RC, Saleheen D, Reilly MP, et al. From Loci to
606 Biology. Circ Res [Internet]. 2016 Feb 19 [cited 2017 Jul 1];118(4):586–606. Available
607 from: http://www.ncbi.nlm.nih.gov/pubmed/26892960
608 5. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential
609 etiologic and functional implications of genome-wide association loci for human diseases
610 and traits. Proc Natl Acad Sci U S A. 2009;106:9362–7.
611 6. Smemo S, Tena JJ, Kim K-H, Gamazon ER, Sakabe NJ, Gómez-Marín C, et al. Obesity-
612 associated variants within FTO form long-range functional connections with IRX3. Nature
613 [Internet]. 2014 Mar 20 [cited 2015 Jun 29];507(7492):371–5. Available from:
614 http://dx.doi.org/10.1038/nature13138
615 7. Won HH, Natarajan P, Dobbyn A, Jordan DM, Roussos P, Lage K, et al. Disproportionate
616 Contributions of Select Genomic Compartments and Cell Types to Genetic Risk for
617 Coronary Artery Disease. Snyder M, editor. PLoS Genet [Internet]. 2015 Oct 28 [cited
618 2016 Jun 28];11(10):e1005622. Available from:
619 http://dx.plos.org/10.1371/journal.pgen.1005622
620 8. Johnson AD, Zhang Y, Papp AC, Pinsonneault JK, Lim J-E, Saffen D, et al.
621 Polymorphisms affecting gene transcription and mRNA processing in pharmacogenetic
622 candidate genes: detection through allelic expression imbalance in human target tissues.
623 Pharmacogenet Genomics [Internet]. 2008 Sep [cited 2016 Jul 15];18(9):781–91.
624 Available from: http://www.ncbi.nlm.nih.gov/pubmed/18698231
625 9. Barrie ES, Hartmann K, Lee S-H, Frater JT, Seweryn M, Wang D, et al. The
626 CHRNA5/CHRNA3/CHRNB4 Nicotinic Receptor Regulome: Genomic Architecture,
627 Regulatory Variants, and Clinical Associations. Hum Mutat [Internet]. 2016 Nov [cited
628 2016 Dec 9]; Available from: http://doi.wiley.com/10.1002/humu.23135

10. Halley P, Kadakkuzha BM, Faghihi MA, Magistri M, Zeier Z, Khorkova O, et al. Regulation of the apolipoprotein gene cluster by a long noncoding RNA. Cell Rep [Internet]. 2014 Jan 16 [cited 2015 Jan 16];6(1):222–30. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3924898&tool=pmcentrez&rendertype=abstract

11. Tada H, Won HH, Melander O, Yang J, Peloso GM, Kathiresan S. Multiple associated variants increase the heritability explained for plasma lipids and coronary artery disease. Circ Cardiovasc Genet [Internet]. 2014 Oct 1 [cited 2016 Jun 28];7(5):583–7. Available from: http://circgenetics.ahajournals.org/cgi/doi/10.1161/CIRCGENETICS.113.000420

12. Wang D, Hartmann K, Seweryn M, Sadee W. Interactions Between Regulatory Variants in CYP7A1 (Cholesterol 7α-Hydroxylase) Promoter and Enhancer Regions Regulate CYP7A1 Expression. Circ Genomic Precis Med [Internet]. 2018 [cited 2020 Jul 12];11(10):e002082. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30354296

13. Lippert C, Listgarten J, Davidson RI, Baxter S, Poon H, Poong H, et al. An exhaustive epistatic SNP association analysis on expanded Wellcome Trust data. Sci Rep [Internet]. 2013 Jan [cited 2015 Jan 12];3:1099. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3551227&tool=pmcentrez&rendertype=abstract

14. Hemani G, Knott S, Haley C, Li Z, Wang S. An Evolutionary Perspective on Epistasis and the Missing Heritability. Mackay TFC, editor. PLoS Genet [Internet]. 2013 Feb 28 [cited 2017 Aug 8];9(2):e1003295. Available from: http://dx.plos.org/10.1371/journal.pgen.1003295

15. Hartmann K, Seweryn M, Handleman SK, Rempała GA, Sadee W. Non-linear interactions between candidate genes of myocardial infarction revealed in mRNA expression profiles. BMC Genomics [Internet]. 2016 [cited 2016 Oct 12];17(1):738. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27640124

16. Skwark MJ, Croucher NJ, Puranen S, Chewapreecha C, Pesonen M, Xu YY, et al. Interacting networks of resistance, virulence and core machinery genes identified by genome-wide epistasis analysis. Gojobori T, editor. PLOS Genet [Internet]. 2017 Feb 16 [cited 2020 Oct 22];13(2):e1006508. Available from: https://dx.plos.org/10.1371/journal.pgen.1006508

17. Nie SF, Zha LF, Fan Q, Liao YH, Zhang HS, Chen QW, et al. Genetic regulation of the thymic stromal lymphopoietin (TSLP)/TSLP receptor (TSLPR) gene expression and influence of epistatic interactions between IL-33 and the TSLP/TSLPR axis on risk of coronary artery disease. Front Immunol [Internet]. 2018 Aug 3 [cited 2020 Oct 22];9(AUG):3. Available from: /pmc/articles/PMC6085432/?report=abstract

18. Rio S, Mary-Huard T, Moreau L, Bauland C, Palaffre C, Madur D, et al. Disentangling group specific QTL allele effects from genetic background epistasis using admixed individuals in GWAS: An application to maize flowering. Springer NM, editor. PLOS Genet [Internet]. 2020 Mar 4 [cited 2020 Oct 22];16(3):e1008241. Available from: https://dx.plos.org/10.1371/journal.pgen.1008241

19. Wallace C, Rotival M, Cooper JD, Rice CM, Yang JHM, McNeill M, et al. Statistical colocalization of monocyte gene expression and genetic risk variants for type 1 diabetes. Hum Mol Genet [Internet]. 2012 Jun 15 [cited 2020 Oct 22];21(12):2815–24. Available from: http://cran.r-project.org

20. 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. Williams SM, editor. PLoS Genet [Internet]. 2014 May 15 [cited 2020 Oct 22];10(5):e1004383. Available from: https://dx.plos.org/10.1371/journal.pgen.1004383

21. Hormozdiari F, van de Bunt M, Segrè A V., Li X, Joo JWJ, Bilow M, et al. Colocalization of GWAS and eQTL Signals Detects Target Genes. Am J Hum Genet. 2016 Dec

1;99(6):1245–60.

22. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. Nat Genet [Internet]. 2013 Jun [cited 2014 Jul 10];45(6):580–5. Available from: http://dx.doi.org/10.1038/ng.2653

23. C G, J ZL, W Z, M H, H S, J B, et al. A Bayesian Framework for Multiple Trait Colocalization From Summary Association Statistics. Bioinformatics [Internet]. 2018 [cited 2020 Jun 12];34(15). Available from: https://pubmed.ncbi.nlm.nih.gov/29579179/?from_term=COLOC&from_pos=2

24. F W, CQ X, Q H, JP C, XC L, D W, et al. Genome-wide Association Identifies a Susceptibility Locus for Coronary Artery Disease in the Chinese Han Population. Nat Genet [Internet]. 2011 [cited 2020 Jun 12];43(4). Available from: https://pubmed.ncbi.nlm.nih.gov/21378986/

25. Wainberg M, Sinnott-Armstrong N, Mancuso N, Barbeira AN, Knowles DA, Golan D, et al. Opportunities and challenges for transcriptome-wide association studies. Nat Genet. 2019;51(4).

26. Mongelli A, Martelli F, Farsetti A, Gaetano C. The dark that matters: Long noncoding RNAs as master regulators of cellular metabolism in noncommunicable diseases [Internet]. Vol. 10, Frontiers in Physiology. Frontiers Media S.A.; 2019 [cited 2020 Oct 22]. p. 369. Available from: www.frontiersin.org

27. Lee DP, Tan WLW, Anene-Nzelu CG, Lee CJM, Li PY, Luu TDA, et al. Robust CTCF-Based Chromatin Architecture Underpins Epigenetic Changes in the Heart Failure Stress–Gene Response. Circulation [Internet]. 2019 Apr 16 [cited 2020 Oct 22];139(16):1937–56. Available from: https://www.ahajournals.org/doi/10.1161/CIRCULATIONAHA.118.036726

28. Sadee W, Hartmann K, Seweryn M, Pietrzak M, Handelman SK, Rempala GA. Missing heritability of common diseases and treatments outside the protein-coding exome. Hum

Genet [Internet]. 2014 Oct [cited 2015 Jan 16];133(10):1199–215. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25107510

29. Mascarenhas R, Pietrzak M, Smith RM, Webb A, Wang D, Papp AC, et al. Allele-Selective Transcriptome Recruitment to Polysomes Primed for Translation: Protein-Coding and Noncoding RNAs, and RNA Isoforms. PLoS One [Internet]. 2015 Jan 2 [cited 2016 Feb 6];10(9):e0136798. Available from: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0136798

30. Jones PD, Webb TR. From GWAS to new biology and treatments in CAD. Aging (Albany NY) [Internet]. 2019 [cited 2020 Jun 12];11(6):1611. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6461175/

31. Lu R, Smith RM, Seweryn M, Wang D, Hartmann K, Webb A, et al. Analyzing allele specific RNA expression using mixture models. BMC Genomics [Internet]. 2015 Jan 1 [cited 2016 Jan 29];16(1):566. Available from: http://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-015-1749-0

32. Smith RM, Webb A, Papp AC, Newman LC, Handelman SK, Suhy A, et al. Whole transcriptome RNA-Seq allelic expression in human brain. BMC Genomics [Internet]. 2013 Jan [cited 2015 Jan 2];14:571. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3765493&tool=pmcentrez&rendertype=abstract

33. Ni G, Moser G, Ripke S, Neale BM, Corvin A, Walters JTR, et al. Estimation of Genetic Correlation via Linkage Disequilibrium Score Regression and Genomic Restricted Maximum Likelihood. Am J Hum Genet. 2018;102(6).

34. Roberts R. Genetics of coronary artery disease: an update. Methodist Debakey Cardiovasc J [Internet]. 2014 Jan [cited 2015 Jul 28];10(1):7–12. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4051327&tool=pmcentrez&rendertype=abstract

35. Musunuru K, Strong A, Frank-Kamenetsky M, Lee NE, Ahfeldt T, Sachs K V., et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature [Internet]. 2010 Aug 5 [cited 2020 Oct 22];466(7307):714–9. Available from: https://pubmed.ncbi.nlm.nih.gov/20686566/

36. Crouch DJM, Bodmer WF. Polygenic inheritance, GWAS, polygenic risk scores, and the search for functional variants. 2020;117(32):18924–33.

Supporting information

S1 File. Fig 1 Gene Names. Gene names corresponding to bar plot presented in Fig 1 B.

S2 File. Example locus. Example of a locus (*LIPA*) implicated by GWAS taken from ensemble.org. There are numerous annotated protein-coding and non-coding transcripts in close proximity and overlapping one another.

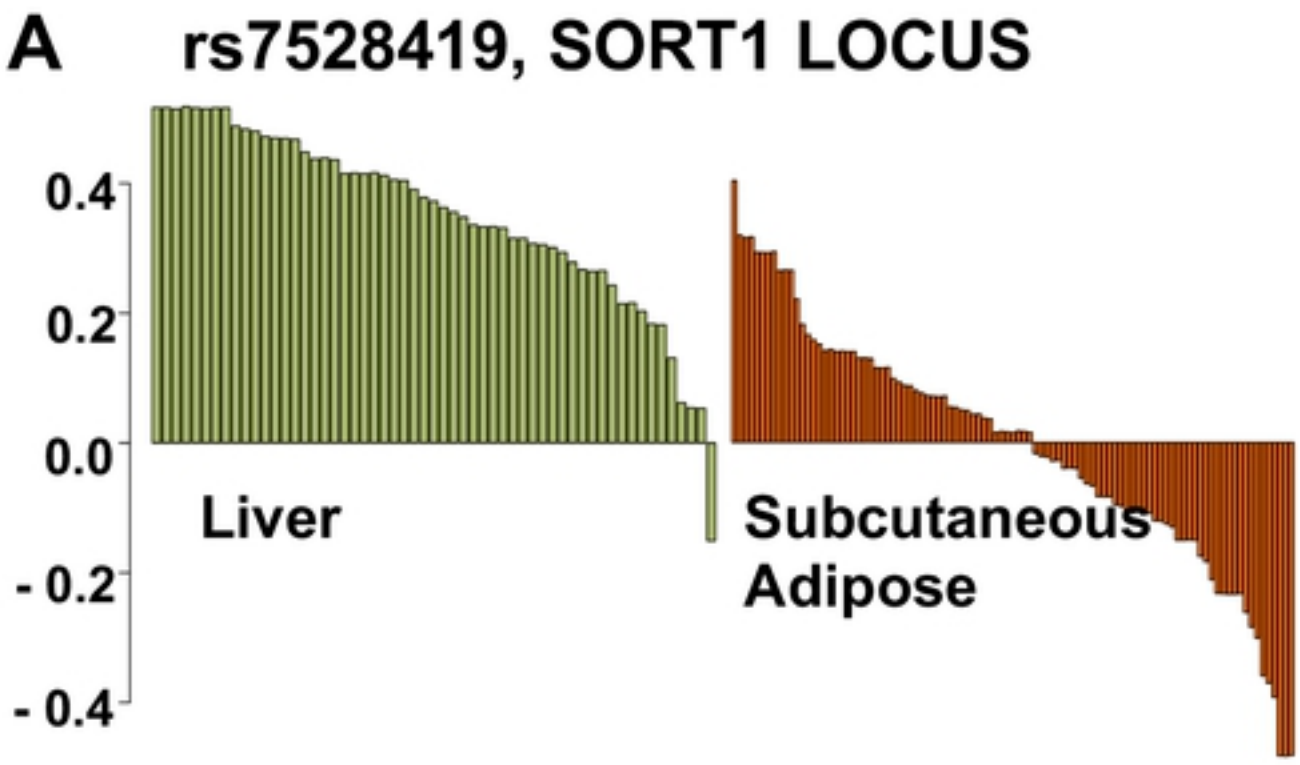
S3 File. 58 CAD GWAS loci. Table of 58 GWAS loci including tier designation, SNPs considered, GWAS annotation, and genes introduced by eQTL, sQTL, and position.

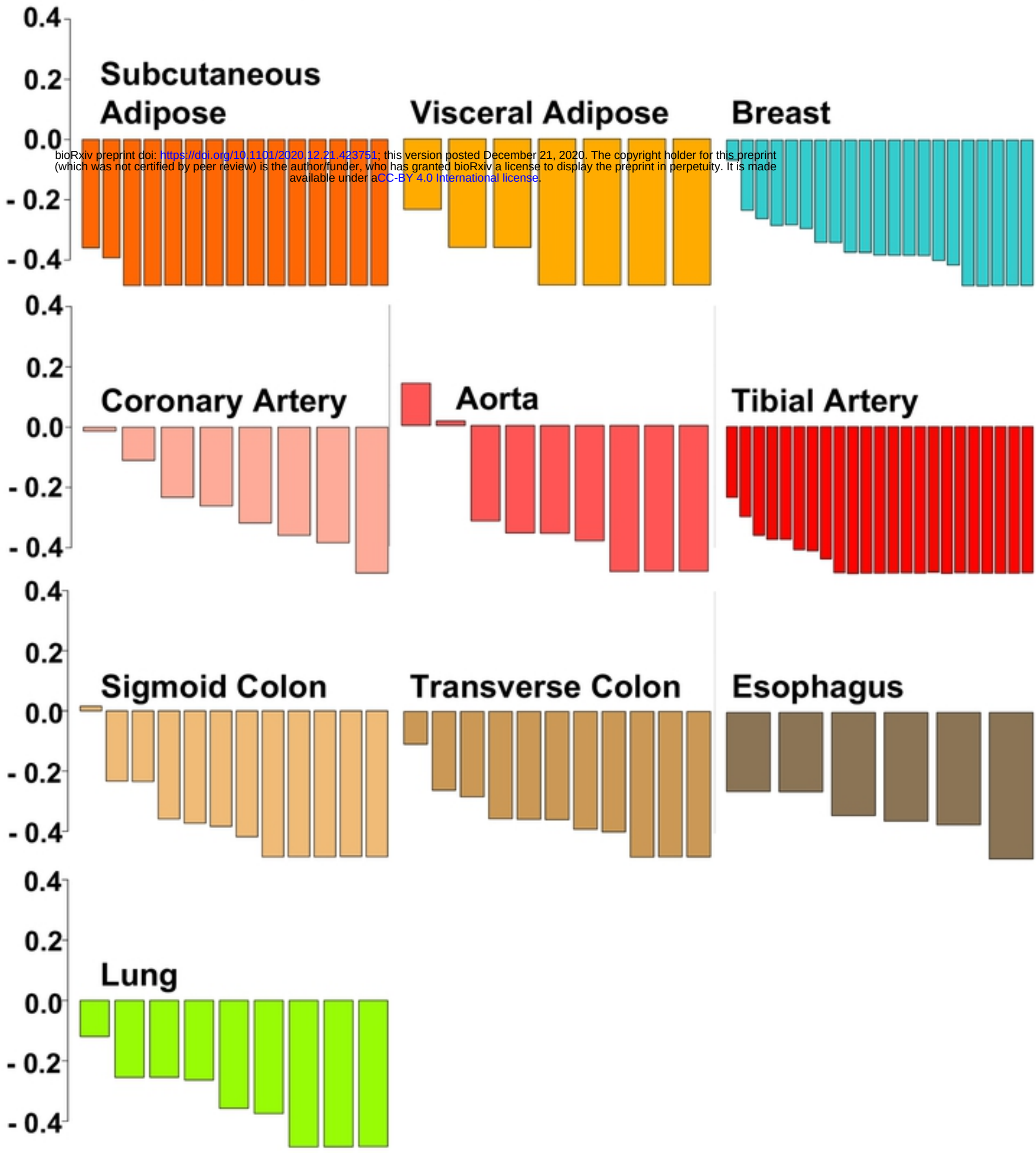
S1 Fig. Expanding candidate genes process. Flowchart portraying process of expanding candidate gene list from 75 to 245 using eQTL, sQTL, and physical position.

S2 Fig. Tier assignment process. Flowchart portraying process of assigning tiers to CAD GWAS loci.

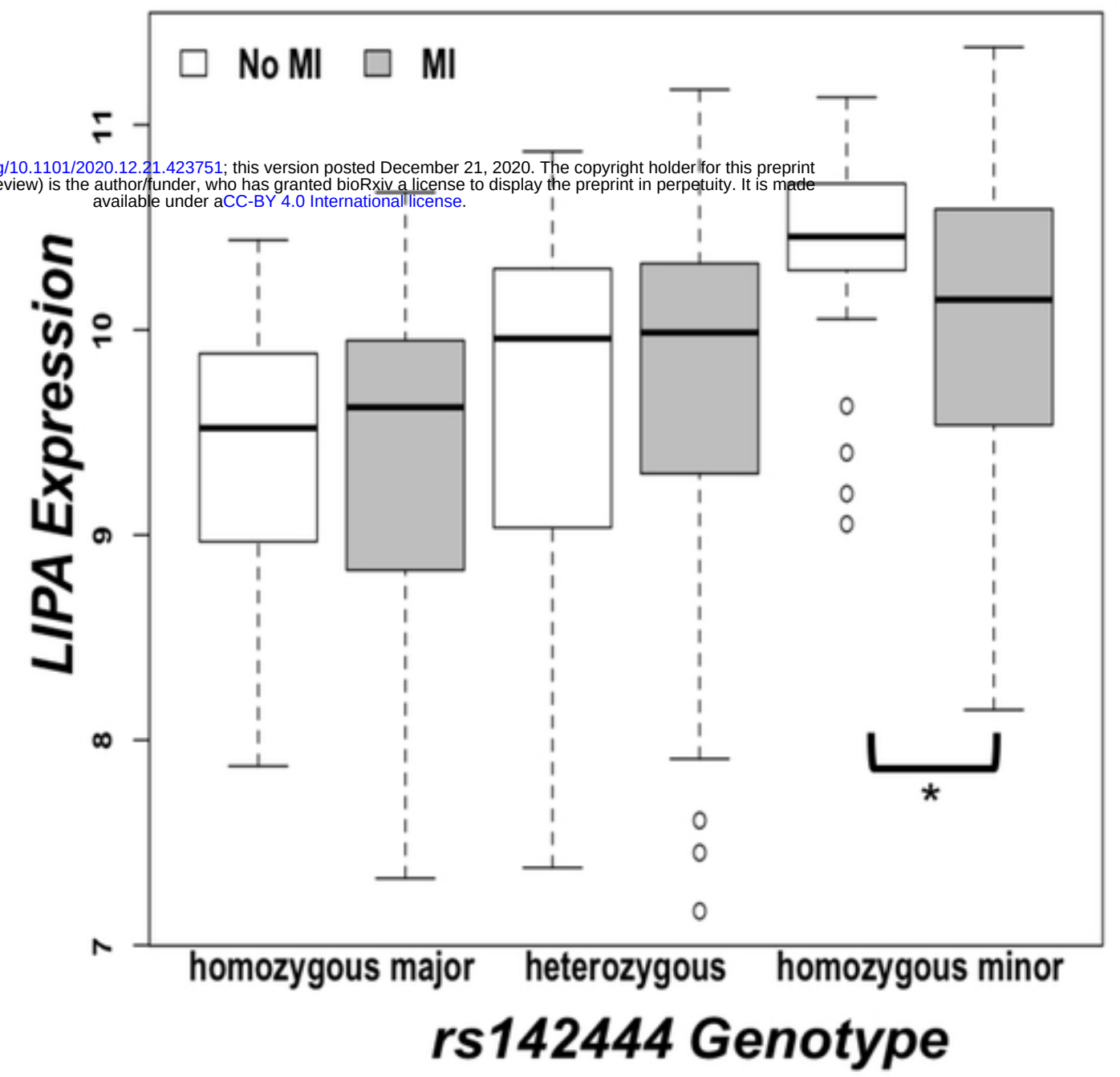
S3 Fig. eQTL, sQTL, position Venn diagram. Venn diagram showing overlap in candidate genes derived from eQTL, sQTL, and position-based re-prioritization.

S4 Fig. LIPA expression, CAD, and genotype in GTEx. Comparison of LIPA expression in GTEx for those with and without heart disease based on rs142444 genotype. LIPA exhibits higher expression in those without heart disease only in the homozygous minor group (p value = 0.22).

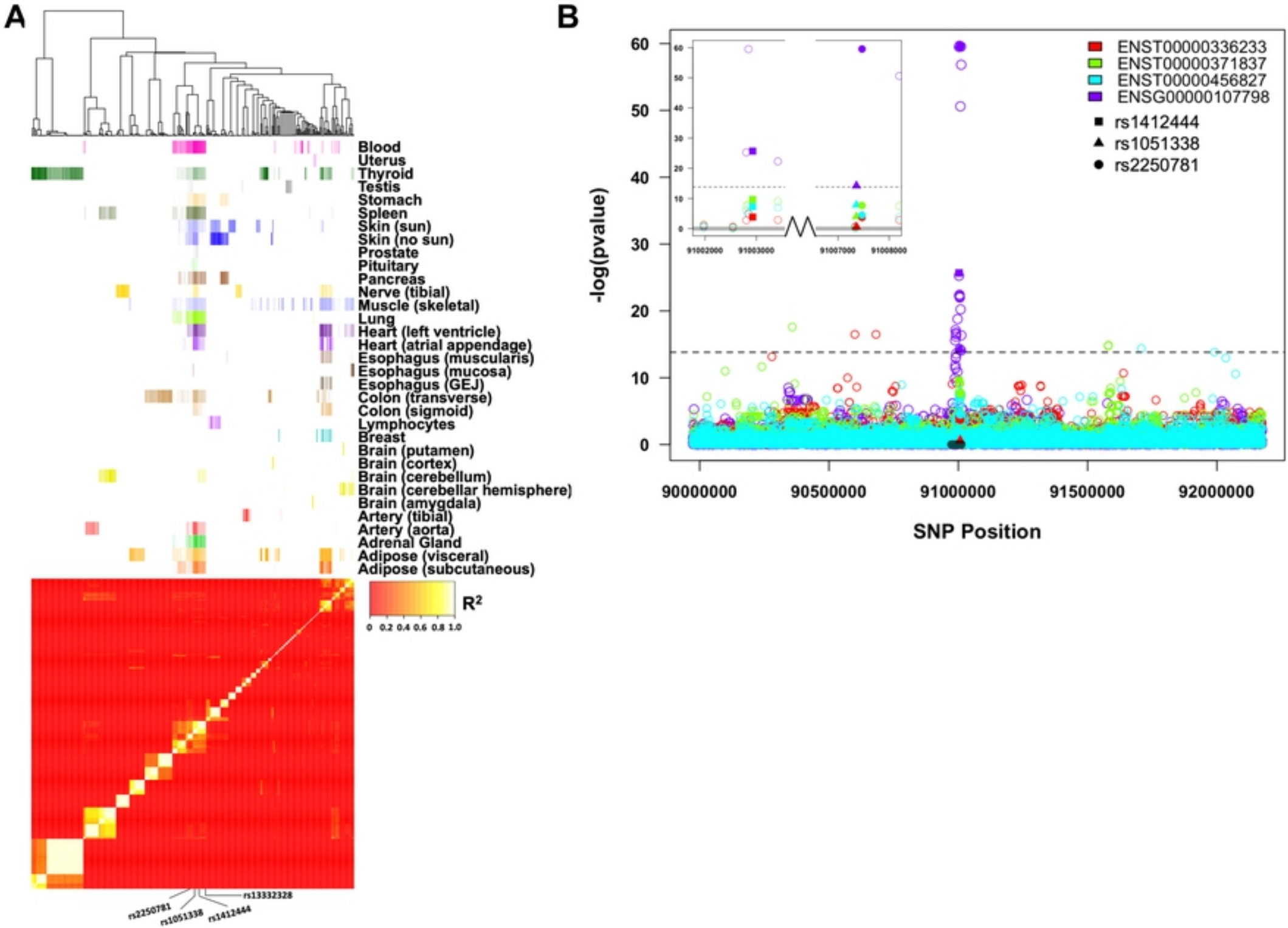




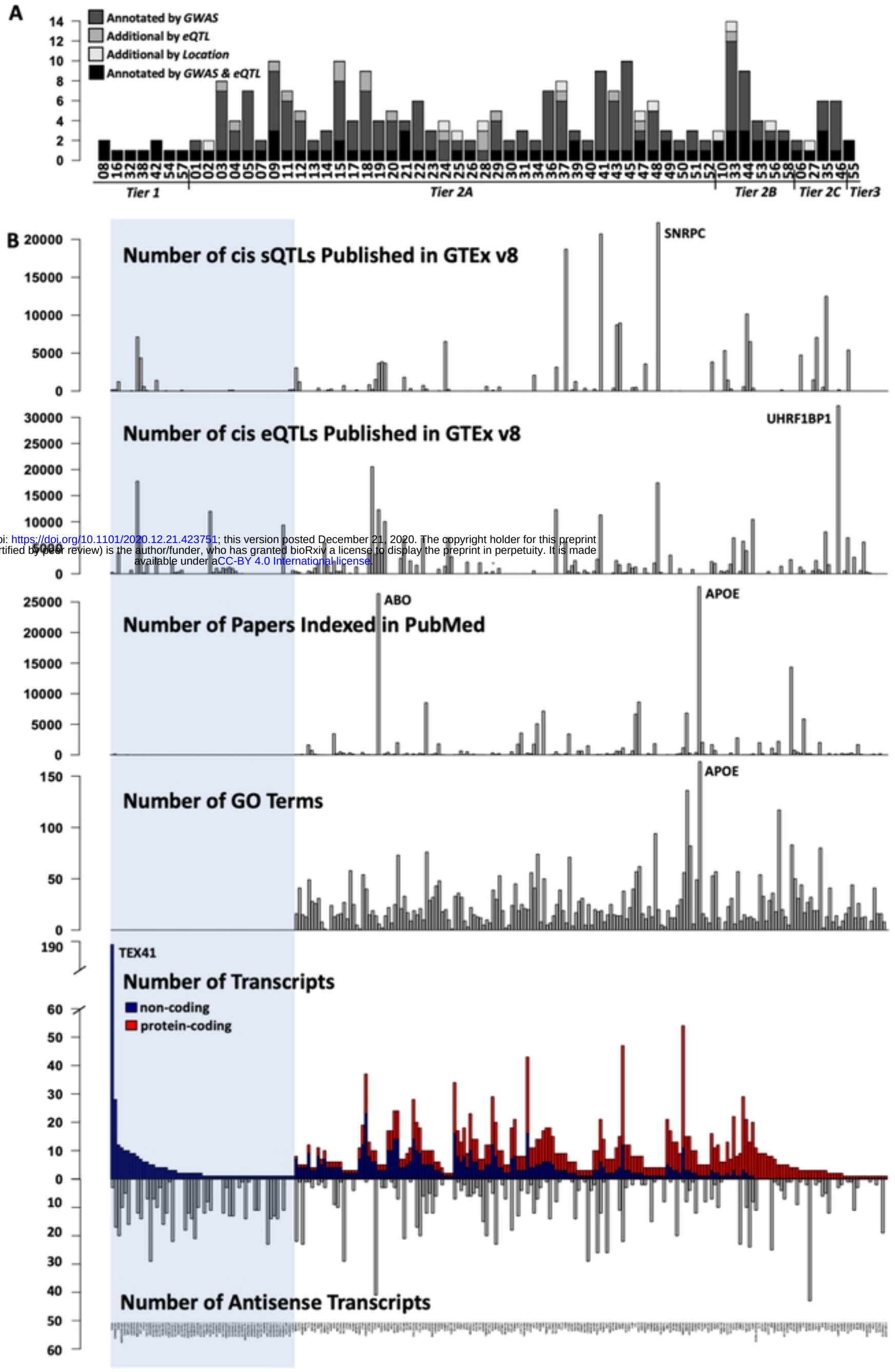
Figure



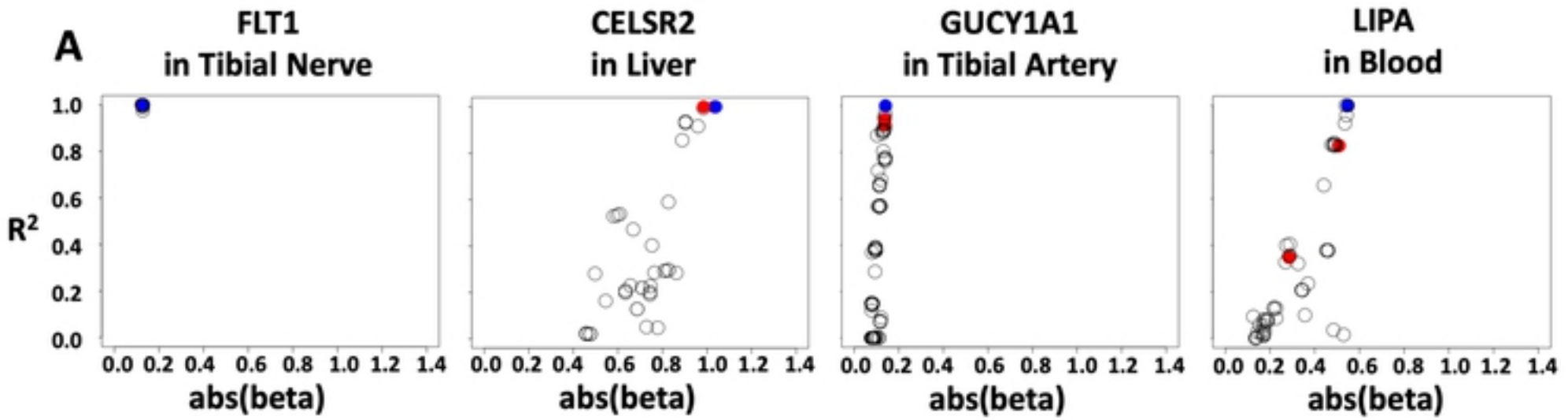
Figure

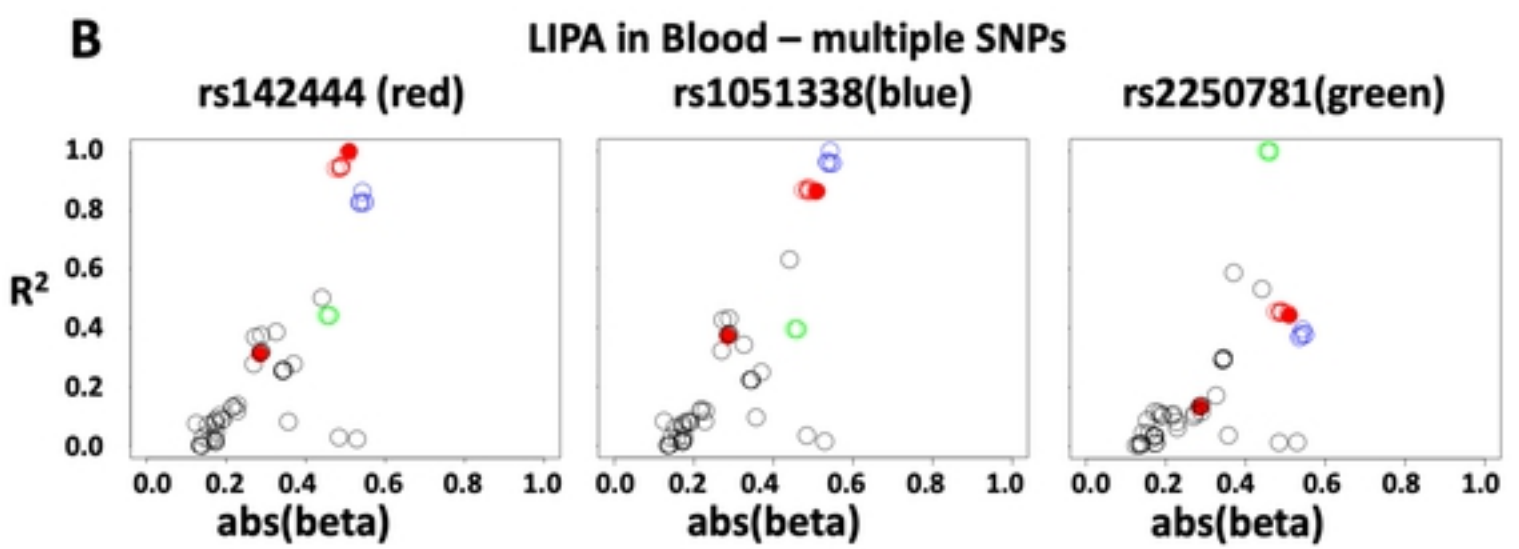

A
Blood
Uterus
Thyroid
Testis
Stomach
Spleen
Skin (sun)
Skin (no sun)
Prostate
Pituitary
Pancreas
Nerve (tibial)
Muscle (skeletal)
Lung
Heart (left ventricle)
Heart (atrial appendage)
Esophagus (muscularis)
Esophagus (mucosa)
Esophagus (GEJ)
Colon (transverse)
Colon (sigmoid)
Lymphocytes
Breast
Brain (putamen)
Brain (cortex)
Brain (cerebellum)
Brain (cerebellar hemisphere)
Brain (amygdala)
Artery (tibial)
Artery (aorta)
Adrenal Gland
Adipose (visceral)
Adipose (subcutaneous)
R²
0 0.2 0.4 0.6 0.8 1.0
rs2250781
rs1051338
rs1412444
rs13332328
B
-log(pvalue)
SNP Position
ENST00000336233
ENST00000371837
ENST00000456827
ENSG00000107798
rs1412444
rs1051338
rs2250781
90000000
90500000
91000000
91500000
92000000

Figure



Figure


A
FLT1
in Tibial Nerve
CELSR2
in Liver
GUCY1A1
in Tibial Artery
LIPA
in Blood
R2
1.0
0.8
0.6
0.4
0.2
0.0
0.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4
abs(beta)


B
LIPA in Blood – multiple SNPs
rs142444 (red)
rs1051338(blue)
rs2250781(green)
R2
1.0
0.8
0.6
0.4
0.2
0.0
0.0 0.2 0.4 0.6 0.8 1.0
abs(beta)

Figure