Dynamic Role of *trans* Regulation of Gene Expression in Relation to Complex Traits

Chen Yao,[1,2] Roby Joehanes,[1,2,3] Andrew D. Johnson,[1,2] Tianxiao Huan,[1,2] Chunyu Liu,[1,2] Jane E. Freedman,[4] Peter J. Munson,[5] David E. Hill,[6,7] Marc Vidal,[6,7] and Daniel Levy[1,2,*]

Identifying causal genetic variants and understanding their mechanisms of effect on traits remains a challenge in genome-wide association studies (GWASs). In particular, how genetic variants (i.e., *trans*-eQTLs) affect expression of remote genes (i.e., *trans*-eGenes) remains unknown. We hypothesized that some *trans*-eQTLs regulate expression of distant genes by altering the expression of nearby genes (*cis*-eGenes). Using published GWAS datasets with 39,165 single-nucleotide polymorphisms (SNPs) associated with 1,960 traits, we explored whole blood gene expression associations of trait-associated SNPs in 5,257 individuals from the Framingham Heart Study. We identified 2,350 *trans*-eQTLs (at $p < 10^{-7}$); more than 80% of them were found to have *cis*-associated eGenes. Mediation testing suggested that for 35% of *trans*-eQTL-*trans*-eGene pairs in different chromosomes and 90% pairs in the same chromosome, the disease-associated SNP may alter expression of the *trans*-eGene via *cis*-eGene expression. In addition, we identified 13 *trans*-eQTL hotspots, affecting from ten to hundreds of genes, suggesting the existence of master genetic regulators. Using causal inference testing, we searched causal variants across eight cardiometabolic traits (BMI, systolic and diastolic blood pressure, LDL cholesterol, HDL cholesterol, total cholesterol, triglycerides, and fasting blood glucose) and identified several *cis*-eGenes (*ALDH2* for systolic and diastolic blood pressure, *MCM6* and *DARS* for total cholesterol, and *TRIB1* for triglycerides) that were causal mediators for the corresponding traits, as well as examples of *trans*-mediators (*TAGAP* for LDL cholesterol). The finding of extensive evidence of genome-wide mediation effects suggests a critical role of cryptic gene regulation underlying many disease traits.

Introduction

Genome-wide association studies (GWASs) have identified tens of thousands of genetic variants associated with complex traits and diseases.[1,2] Genetic variants identified by GWASs, however, explain only a small proportion of phenotypic variation, even for diseases known to have a strong genetic component, such as obesity, diabetes, and schizophrenia.[3,4] This knowledge void has been termed the "missing heritability."[5] One important consideration in the search for missing heritability is that the top GWAS single-nucleotide polymorphisms (SNPs) are often not causal variants for their associated traits, but rather are in linkage disequilibrium (LD) with causal SNPs.[3] In addition, fewer than 5% of GWAS SNPs are non-synonymous substitutions, while the remainder are located within non-coding regions.[2,6] This suggests that instead of directly altering the amino acid sequence of proteins, SNPs can affect phenotypes by other mechanisms, such as regulation of gene transcription levels.

Expression quantitative trait loci (eQTLs) are genetic variants that are associated with gene transcription levels.[7] eQTLs that alter expression of nearby transcripts (*cis*-eGenes) are referred to as *cis*-eQTLs, whereas those associated with expression of remote transcripts (*trans*-eGenes), usually on different chromosomes, are referred to as *trans*-eQTLs.[8,9] When SNPs at a *trans*-eQTL locus affect the expression of multiple *trans*-eGenes, the region is defined as a *trans*-eQTL hotspot.[10] *cis*-eQTLs typically reside close to transcription start sites (TSSs), suggesting that they directly impact gene expression.[11] In contrast to *cis*-eQTLs, analysis of *trans*-eQTLs is vastly more computationally challenging and reported *trans*-eQTLs have proven to be less replicable across studies.[11,12] Therefore, many eQTL studies focus only on *cis*-eQTLs or a small subset of *trans*-eQTLs.[12,13] *trans*-eQTL hotspots are of particular interest because SNPs linked to such hotpots could serve important regulatory roles. The mechanisms by which *trans*-eQTLs alter transcription of their linked *trans*-eGenes are largely unknown and likely reflect indirect or cryptic regulation.[14,15] For example, it has been proposed that expression of *trans*-eGenes could be mediated by transcription factors residing close to the corresponding *trans*-eQTLs.[14] This phenomenon would allow *cis*-eQTLs near regulatory genes to serve as master regulators for a large number of *trans*-eGenes. We found that some eQTLs can affect expression of eGenes both in a *cis* and *trans* manner, whereby *cis*-eGenes mediate the associations between eQTLs and *trans*-eGenes.[16]

We investigated the associations of SNPs previously reported to be associated with a variety of traits in GWASs with whole blood gene expression measured in 5,257 Framingham Heart Study (FHS) participants. In total, we related genotypes for 39,165 genome-wide significant GWAS SNPs reported to be associated with 1,960 traits in GWAS databases with expression levels of 17,873 genes

[1]The Framingham Heart Study, 73 Mt. Wayte Avenue, Framingham, MA 01702, USA; [2]The Population Sciences Branch, Division of Intramural Research, National Heart, Lung, and Blood Institute, NIH, Bethesda, MD 20892, USA; [3]Hebrew Senior Life, 1200 Centre Street Room #609, Boston, MA 02131, USA; [4]Department of Medicine, University of Massachusetts Medical School, Worcester, MA 01655, USA; [5]Mathematical and Statistical Computing Laboratory, Center for Information Technology, NIH, Bethesda, MD 20817, USA; [6]Center for Cancer Systems Biology (CCSB) and Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA 02215, USA; [7]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA
*Correspondence: levyd@nhlbi.nih.gov
http://dx.doi.org/10.1016/j.ajhg.2017.02.003.

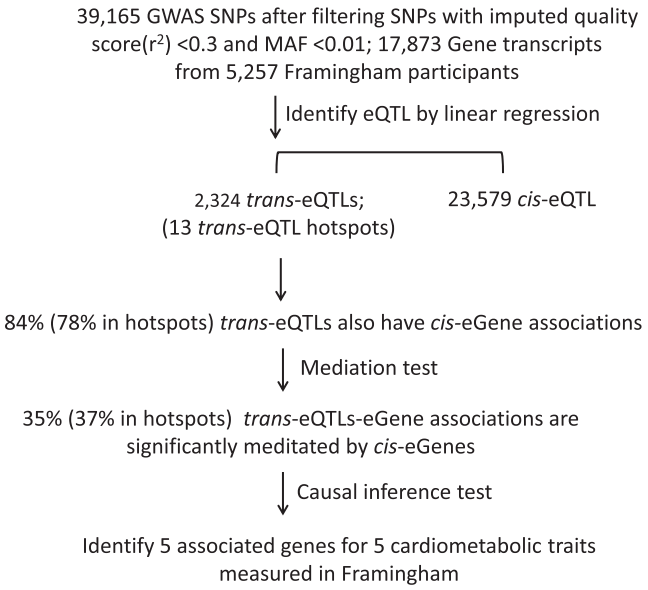

Figure 1. Work Flow and Results Summary
42,271 SNPs associated with 1,960 traits were obtained from GRASP (at $p \leq 5 \times 10^{-8}$). Whole blood samples were collected from 5,257 FHS participants. Genome-wide genotyping and mRNA expression levels were assayed. We correlated 39,165 GWAS SNPs (after filtering) with expression levels of 17,873 genes to identify expression quantitative trait loci (eQTLs). For SNPs having both local (*cis*) and remote (*trans*) regulation effects, we then tested whether the effect of *trans*-eQTLs was mediated through *cis*-eGenes. Finally, integrating genotype, gene expression, and phenotype data, we conducted causal inference testing to identify causal variants for eight cardiometabolic traits (BMI, systolic and diastolic blood pressure, LDL cholesterol, HDL cholesterol, total cholesterol, triglycerides, fasting blood glucose).

to identify *cis*- and *trans*-eQTLs and their associated eGenes.[17] Our results reveal that a large number of eQTLs regulate gene expression in both a *cis* and *trans* manner. Additionally, we identified 13 *trans*-eQTL hotspots and found that about one third of *trans* regulation is significantly mediated by the expression of *cis*-eGenes. As proof of principle, we inferred causality and directionality of SNP-transcript-trait relationships using genetic variants as instrumental variables in causal inference analyses. Specifically, we looked at eight cardiometabolic traits that were extensively characterized in the FHS, including body mass index (BMI), systolic and diastolic blood pressure, low-density lipoprotein (LDL) cholesterol, high-density lipoprotein (HDL) cholesterol, triglycerides, total cholesterol, and fasting blood glucose levels.

Material and Methods

Study Sample

In 1948, the FHS started recruiting participants (original cohort) from Framingham, MA, to begin the first round of physical examinations and lifestyle interviews to investigate cardiovascular disease (CVD) and its risk factors. In 1971 and 2002, FHS recruited offspring (and their spouses) and adult grandchildren of the original cohort participants into the Offspring and Third Generation cohorts, respectively.[18] A total of 5,257 participants from the FHS Offspring and Third Generation cohorts had gene expression profiling and genome-wide genotyping.[19] Methods for collection of whole blood samples and RNA isolation and preparation have been described previously.[19] A summary of the cardiometabolic traits used in this study can be found in Table S1. All participants provided informed consent and the protocols were approved by the institutional review board.

Genotype Data

A total of 42,271 SNPs associated with 1,960 complex traits from GWASs (at $p \leq 5 \times 10^{-8}$ in the GRASP database) were curated and matched with ~8.5 million SNPs imputed from the 1000 Genomes Project Reference Panel. GRASP v.2.0 re-annotated genotype-phenotype results from 1,390 GWASs and corresponding open-access GWAS results.[2] SNPs were input to Minimac[20] software. In brief, we combined genotype data with the HapMap CEU samples and inferred genotypes probabilistically based on shared haplotype stretches between study samples and HapMap release 22 build 36. For each genotype, imputation results were summarized as an "allele dosage" defined as the expected number of copies of the minor allele at that SNP (value between 0 and 2). SNPs with imputed quality score (r^2) < 0.3 and MAF < 0.01 were filtered out, resulting in 39,165 GWAS significant SNPs for eQTL analysis (Figure 1).

Gene Expression

Whole blood was collected in PAXgene tubes (PreAnalytiX) and frozen at −80°C. RNA was extracted using a whole blood RNA System Kit (QIAGEN) in FHS and mRNA expression profiling was assessed using the Affymetrix Human Exon 1.0 ST GeneChip platform (Affymetrix), which contains more than 5.5 million probes targeting the expression of 17,873 genes. The Robust Multi-array Average (RMA) package[21] was used to normalize gene expression values and remove any technical or spurious background variation. Linear regression models were used to adjust for technical covariates (batch, first principal component, and residual all probeset mean) and differential blood cell proportions. The pedigreemm package[22] was used to remove the effects of sex and age and accounted for familial relationships. 2,181 individuals from the Third Generation cohort had complete blood cell counts (white blood cells, neutrophils, lymphocytes, monocytes, eosinophils, and basophils). Using gene expression data, we imputed the cell counts of remaining samples by partial least square (PLS) prediction that was developed in participants with measured cell counts and expression data. We did not find a significant difference when comparing results using imputed cell counts and those using measured values. Therefore, we used measured cell counts when they were available and used imputed values when measured cell counts were not available.

Identifying eQTLs and *trans*-eQTL Hotspots

eQTL analysis was conducted on 5,257 individuals from the FHS Offspring and Third Generation cohorts using available mRNA expression data and genome-wide genotyping. Twenty PEER factors were calculated using a Bayesian framework and were used to account for hidden confounding factors in the adjusted gene expression data.[23] For each SNP-mRNA pair, a linear model was developed to identify SNP-mRNA associations, adjusting for PEER factors and familial relationship. p values were adjusted for multiple comparisons using the false discovery rate (FDR)

method.[24] eQTLs at FDR ≤ 0.05 were considered to be significant. *cis*-eQTLs were defined as SNPs that reside within 1 Mb of the transcription start site. *trans*-eQTLs were defined as SNPs that were at a distance greater than 5 Mb from the TSS of an associated transcript on the same chromosome or on a different chromosome. eGenes were defined as genes associated with eQTLs. An independent set of eQTLs was obtained by pruning eQTLs in LD ($R^2 > 0.2$) and within 250 Kb, while keeping the most significant SNPs per eGene. *trans*-eQTL hotspots were identified by an index eQTL and nearby SNPs in high LD ($R^2 > 0.8$) associated with at least ten *trans*-eGenes. We excluded from analysis eQTLs that resided on the same chromosome but were less than 5 Mb from their eGenes to avoid confounding by long-range LD patterns.

Mediation and Causal Testing

Mediation testing was conducted using the mediation package (see Web Resources) in R with eQTL as the "exposure," *cis*-eGene expression as the "mediator," and *trans*-eGene expression as the "outcome." A 100% proportion of mediation effect indicated that the entire association between an eQTL and expression of a *trans*-eGene (direct effect) is explained by effects of the eQTL on *cis*-eGene expression. Significant mediation effects were defined at a permutation threshold of $p < 0.005$ (1,000 permutations). The causal inference test (CIT) was conducted using the statistical package CIT[25] in R based on the following conditions: (1) the trait (T) is associated with the locus (L); (2) L is associated with the eGene mediator (G) after adjusting for T; (3) G is associated with T after adjusting for L; and (4) L is independent of T after adjusting for G. The p value of CIT is defined as the maximum of the four-component test p values by the intersection-union test framework (Figure S1).[26] To determine whether *cis*-eGenes or *trans*-eGenes are causal mediators for a trait, CIT was performed for *cis*-eGenes and *trans*-eGenes separately. For a *cis*-eGene, we used its *cis*-eQTL with the smallest p value as an instrumental variable. For a *trans*-eGene, we calculated its best *cis*-eQTL from ~8 million imputed SNPs residing within 1 Mb of the *trans*-eGene, based on the smallest p value.

Functional Annotation and Enrichment Testing

SNP annotations were conducted on HaploReg v.4.1,[27] which linked the SNPs with chromatin state and protein binding annotation from the Roadmap Epigenomics and ENCODE project, sequence conservation across mammals, the effect of SNPs on regulatory motifs, and the effect of SNPs on expression from eQTL studies. Regulatory motif enrichment was conducted using *cis*-eQTLs residing in *trans*-eQTL hotspots as test sets and all *cis*-eQTLs as background. The gene ontology and transcription factor target enrichment analyses were conducted by "Gene-Set Enrichment Analysis (GSEA)."[28] The transcription factors (TFs) were extracted from FANTOM,[29] the large international consortium that mapped all human TFs and the genes they regulate; it contains 1,672 human genes. The protein-protein interaction (PPI) network contained a systematically generated or literature-curated dataset of ~58,000 PPIs among 10,690 human proteins.[30] We defined hub proteins as those having no fewer than four interactions in the PPI network.

Results

eQTLs Associated with Complex Disease Traits

At a minor allele frequency > 0.01 and imputation $r^2 > 0.3$, 39,165 genome-wide significant ($p < 5 \times 10^{-8}$) SNPs reported in published GWAS databases[2] were genotyped or imputed in the FHS. At FDR < 0.05, we identified 23,579 *cis*-eQTLs (associated with expression of 2,933 *cis*-eGenes at a corresponding $p < 1 \times 10^{-4}$; Table S2) representing 5,974 independent SNPs (LD threshold < 0.2) and 2,350 *trans*-eQTLs (associated with expression of 606 *trans*-eGenes at a corresponding $p < 1 \times 10^{-7}$; Table S3) representing 486 independent SNPs (LD threshold < 0.2). Because many SNPs in high LD are associated with different traits in GWASs, we used non-pruned eQTLs in the subsequent analyses. In total, we determined that 23,951 out of 39,165 (61%) statistically significant GWAS SNPs are eQTLs, which is consistent with previous findings that GWAS SNPs are enriched for eQTLs ($p < 0.0001$ for 10,000 random sets of 39,165 SNPs at MAF > 0.01 and $r^2 > 0.3$; average eQTL number = 9,022).[13,31]

Reproducibility and Mediation Effects of *trans*-eQTLs

In accordance with previous results,[32] we found that *trans* effects on gene expression are much weaker than *cis* effects (Figure S2A, average *trans*-eQTL effect size on corresponding transcript $R^2 = 0.009$ versus average *cis*-eQTL effect size $R^2 = 0.02$, t test $p = 1.1 \times 10^{-16}$). Using the Blood eQTL Browser (meta-analysis in non-transformed peripheral blood samples from 5,311 individuals)[12] as a reference database, we found that 331 out of 1,686 (20%) *trans*-eQTL-*trans*-eGene pairs from the database were statistically significant (at $p < 1 \times 10^{-7}$) in our results. Among them, 323 pairs (98%) have concordant directions of effects (Table S4). The overlapping pairs increased to 562 (33%) when we used $p < 1 \times 10^{-4}$ as our *trans*-eQTL threshold. On the other hand, the replication rate was much higher for *cis*-eQTLs; 17,118 out of 38,608 (44%) *cis*-eQTL-*cis*-eGene pairs in the Blood eQTL Browser were statistically significant (at $p < 1 \times 10^{-4}$) in our results. Among them, 14,208 pairs (83%) had the same direction of effect. We hypothesized that the genetic effects of *trans*-eQTLs on expression of *trans*-eGenes are mediated in some cases by the expression of *cis*-eGenes (Figure 2). To test this hypothesis, we conducted mediation analyses for all 8,566 *trans*-eQTL-*trans*-eGene pairs to identify the proportion of the association between a *trans*-eQTL and *trans*-eGene that was attributable to the effect of the eQTL on *cis*-eGene expression. For *trans*-eQTLs and *trans*-eGenes on different chromosomes, we found that 1,953 out of 2,324 *trans*-eQTLs (84%) affect *cis*-eGene expression and that 2,612 *trans*-eQTL-*trans*-eGene pairs (35%) are significantly mediated by expression of *cis*-eGenes near the *trans*-eQTL. The proportion of mediation ranged from 1.4% to 100% (mean 15%). For *trans*-eQTLs and *trans*-eGenes on the same chromosome (by definition, separated by at least 5 Mb), we found that 913 out of 931 *trans*-eQTLs (98%) affect *cis*-eGene expression and that 1,011 *trans*-eQTL-*trans*-eGene pairs (90%) are significantly mediated by expression of *cis*-eGenes near the *trans*-eQTL, suggesting that *trans*-eGenes on the same chromosome are highly regulated through *cis*-eGenes (Table S5).

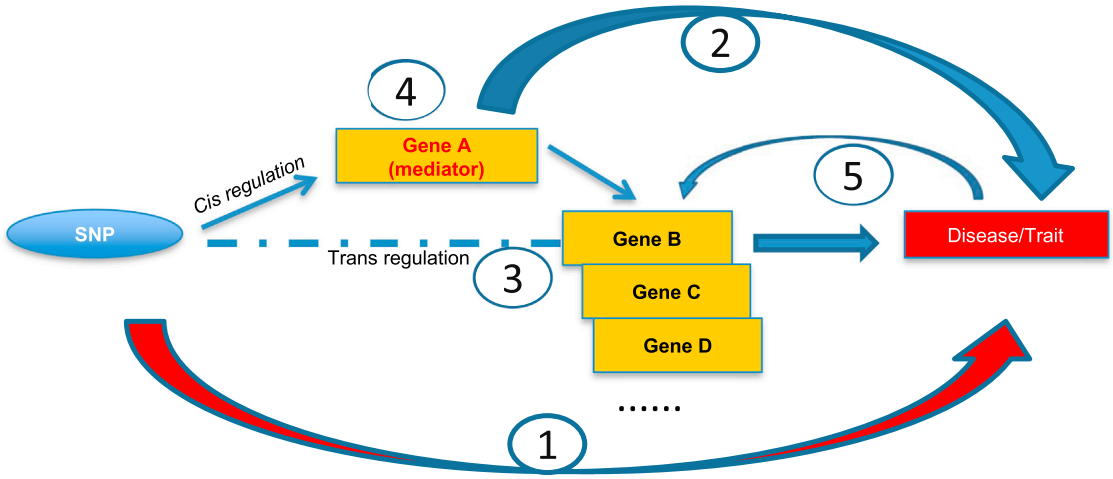

Figure 2. Mediation Mechanisms of eQTLs
Genetic variants can affect traits through the following mechanisms: (1) missense SNP affects protein structure/function; (2) non-coding SNP affects gene expression (*cis*); (3) non-coding SNP affects remote (*trans*) gene expression directly or by (4) *cis*-eGene mediation of the *trans*-eQTL-*trans*-eGene association; or (5) reverse causality (trait has feedback effect on gene expression).

trans-eQTL Hotspots

Among the 2,324 *trans*-eQTLs, we identified 13 *trans*-eQTL hotspots across eight chromosomes, with the index SNP associated with at least ten transcripts (Table 1 and Figure 3). Notably, 8 out of 13 *trans*-eQTL hotspots were also identified in the Blood eQTL Browser,[12] indicating that hotspots are more replicable than individual *trans*-eQTLs. For these *trans*-eQTL hotspots, we found that *cis*-eQTLs linked to *trans*-eQTL hotspots have smaller effects compared to *cis*-eQTLs not located within *trans*-eQTL hotspots (mean R^2 0.009 versus 0.02, t test $p < 1 \times 10^{-8}$) and have similar effect sizes as *trans*-eQTLs (mean R^2 0.009 versus 0.01, Table S6, Figure S2B). We found that eGenes associated with *trans*-eQTL hotspots have a directional bias, with 65% of *trans*-eGenes showing the same directional effect in relation to the *trans*-eQTLs,[33]

Table 1. *trans*-eQTL Hotspots

Hotspot Location (hg19)	Number of *trans*-eQTLs	Number of *trans*-eGenes Associated with Index eQTL	Directional Bias of *trans*-eGenes Associated with Index eQTL	Traits Associated in GWAS with Index eQTLs	*trans*-eGene Enrichment in TF Motifs[a]
1: 205,187,981–205,244,972	10	10	+64%	platelet count	NA
1: 248,039,451	1	12	−58%	red blood cell count	*STAT1/STAT2*
2: 60,708,597–60,725,451	14	14	−79%	fetal hemoglobin level	*NFAT/SP1*
3: 50,093,209	1	24	+100%	age at menarche	NA
3: 56,849,749–56,865,776	2	126; 84	−94%; +92%	platelet count; mean platelet volume	*TCF3/ETS2*
6: 135,411,228–13,543,5501	13	22	−55%	fetal hemoglobin	NA
6: 139,840,693–139,844,429	13	48	−70%	erythrocyte count	*SP1/TCF3*
7: 50,423,963–50,562,361	19	76	−59%	childhood acute lymphoblastic leukemia	*PAX4*
12: 54,712,308–54,736,470	2	14	+79%	mean platelet volume	*ETS2*
12: 111,884,608–112,610,714	9	13	−62%	LDL cholesterol; blood pressure; asthma	NA
16: 57,061,189–57,061,189	2	10	−100%	HDL cholesterol	*IRF8/IRF2*
17: 27,072,463–27,322,441	45	32	+55%	mean corpuscular volume	*E4F1*
17: 33,796,260–33,944,055	4	51	+75%	mean platelet volume	*ETS2/MAZ*

Plus sign (+) denotes the positive association; minus sign (−) denotes the negative association.
[a]Transcription factors whose motifs were matched with promoter regions [−2 kb, 2 kb] around transcription start site of the *trans*-eGenes; NA, no TF target enrichment.

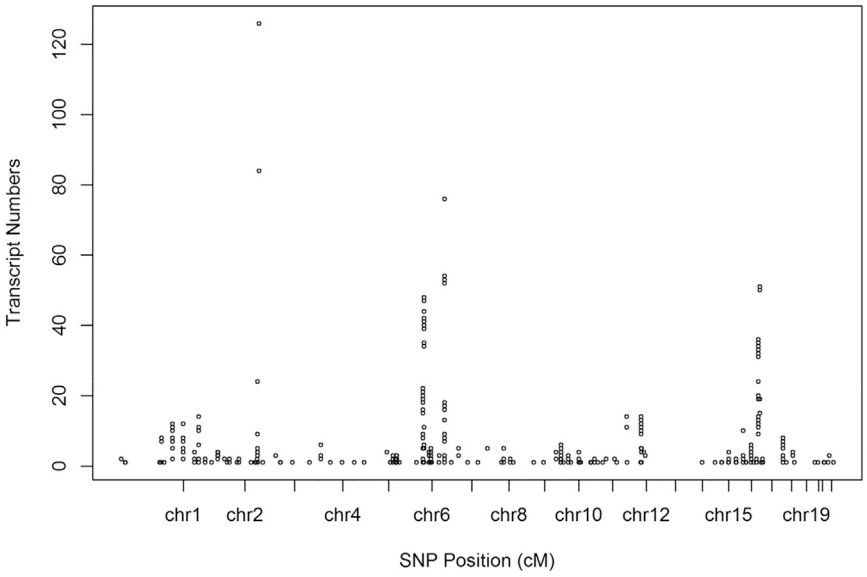

Figure 3. *trans*-eQTL Hotspots
x axis denotes the chromosomal location of SNPs. y axis denotes the number of *trans*-eGenes.

rather than equal ratios of overexpression versus underexpression, as would be expected if the *trans*-eQTLs randomly affect the direction of expression of their corresponding *trans*-eGenes. For example, for age at menarche and HDL cholesterol, the associated *trans*-eGenes show 100% directionally consistent expression in relation to the index *trans*-eQTL (Table 1). One explanation for this observation is that the *trans*-eQTL alters the activity or abundance of a transcription factor (TF, or other *trans*-acting factor), leading to concordant expression changes of all the target genes of this factor.

For all 13 *trans*-eQTL hotspots, we found that 37% of *trans*-eQTL-*trans*-eGene associations were mediated by the expression of *cis*-eGenes (at $p < 0.005$ based on 1,000 bootstrap permutations, Figure 1). The strongest mediation effect was found at the *NLRC5* locus on chromosome 16, which is associated with expression of ten eGenes at the *HLA* locus on chromosome 6. We found that 80% of the genetic effects between rs291040 and *TAP1* are mediated by the *cis* expression of *NLRC5* (Table 2). Prior studies have shown that *NLRC5* acts as a master regulator of MHC class genes in immune response[34] and interacts with the RFX transcription factor complex to induce MHC class I gene expression.[35] Using HaploReg v.4.1, we found that SNPs in *trans*-eQTL hotspots are significantly enriched for regulatory motifs (hypergeometric $p = 3.7 \times 10^{-5}$), with 121 out of 138 (88%) SNPs in ENCODE TF binding experiments (Table S7), suggesting that the expression of *trans*-eGenes is controlled by these SNPs. Among 37 *cis*-eGenes linked to *trans*-eQTL hotspots, we found that two are TFs (*NFE2* [MIM: 601490] and *IKZF1* [MIM: 603023]). Although *cis*-eGenes are not enriched for TFs, we found that *trans*-eGenes are significantly enriched for TF targets (at FDR < 0.05, Table 1) in 9 of 13 *trans*-eQTL hotspots. In addition, we found that 13 of 37 *cis*-eGenes shared the same regulatory motifs with the *trans*-eGene(s) of the same *trans*-hotspot, suggesting that both *cis*-eGene and *trans*-eGene are under the same regulatory control. The enriched functions of *trans*-eGenes are also highly consistent with the traits affected by the *trans*-eQTLs. For example, *trans*-eQTLs on chromosome 2 are associated with platelet count in GWASs. We found that the *trans*-eGenes in this hotspot are enriched for platelet degranulation (Table S8). Moreover, analyzing 25 *cis*-eGenes having binary interactions identified in a systematic screen for protein-protein interactions (PPI),[30] we found that 15 of them are hub genes in the PPI network (Figure S3, $p < 0.001$ for randomly selecting 25 proteins in a PPI network), suggesting that *cis*-eGenes linked to *trans*-eQTLs play a central regulatory role in critical biological pathways through their mediation effects on *trans*-eGenes.

Table 2. *cis*-eGenes in Hotspots with >10% Mediation Effects on the Relations of *trans*-eQTLs to *trans*-eGenes

SNPs	Hotspot Number	*cis*-eGene	*trans*-eGene	SNP-eGene *trans* Association (β)	SNP- *trans* eGene Association Adjusted for *cis*-eGene (β)	Proportion of Mediation[a]	p Value for Mediation
rs3811444	2	*TRIM58*	*ZER1*	−0.015	−0.011	27%	0.001
rs6762477	3	*UBA7*	*RCAN3* (MIM: 605860)	0.021	0.014	37%	0.002
rs12718597	8	*IKZF1* (MIM: 603023)	*TMEM9B* (MIM: 616877)	−0.024	−0.021	12%	0.002
rs11065987	10	*ALDH2*	*ARHGEF40*	0.017	0.013	23%	0.002
rs291040	11	*NLRC5* (MIM: 613537)	*TAP1* (MIM: 170260)	−0.023	−0.003	88%	0.001
rs10512472	13	*AP2B1* (MIM: 601025)	*TRAK2* (MIM: 607334)	0.023	0.014	40%	0.001

[a]Proportion of mediation of the *trans*-eQTL-*trans*-eGene association by the *cis*-eGene.

Table 3. Causal Inference Test Results for Cardiometabolic Traits

Trait	eQTL Annotation (hg19)			eQTL-Trait Association		eQTL-eGene Association Given Trait		eGene-Trait Association Given eQTL		
	ID	Location	Causal eGene[a]	β	p Value	B	p Value	β	p Value	p Value (CIT)
SBP	rs11065898	12: 111,634,620	*ALDH2* (c)	−0.48	0.02	0.44	1.1×10^{-23}	4.44	8.4×10^{-6}	0.02
DBP	rs11065898	12: 111,634,620	*ALDH2*(c)	0.6	0.005	−0.04	5.7×10^{-13}	2.5	0.0001	0.005
LDL-cholesterol	rs926657	6: 159,463,452	*TAGAP*(t)	2.5	0.04	0.09	8.8×10^{-23}	6.6	0.003	0.04
Total cholesterol	rs7570971	2: 135,837,906	*MCM6* (c)	3.0	8×10^{-5}	0.09	1.3×10^{-109}	7.8	0.003	0.01
	rs7570971	2: 135,837,906	*DARS* (c)	3.0	8×10^{-5}	0.03	2.3×10^{-23}	11.4	0.0002	0.01
	rs926657	6: 159,463,452	*TAGAP* (t)	3.1	0.04	0.09	2.1×10^{-22}	8.7	0.001	0.04
Triglycerides	rs4604455	8: 125,505,785	*TRIB1* (c)	0.02	0.0008	0.02	0.0005	−0.03	0.02	0.03

Abbreviations are as follows: SBP, systolic blood pressure; DBP, diastolic blood pressure.
[a]*cis*-eGenes denoted (c); *trans*-eGenes denoted (t).

Causal Effects between eQTLs and Phenotypes

To test whether expression levels of eGenes (*cis* or *trans*) associated with eQTLs might explain the observed associations between eQTLs and phenotypes, we conducted causal inference testing (CIT) using the statistical package CIT in R.[25,36] We applied this approach to the analysis of eight common cardiometabolic traits (BMI, blood lipid levels [HDL-cholesterol, LDL-cholesterol, triglycerides, and total cholesterol], fasting blood glucose, and systolic and diastolic blood pressure [SBP and DBP]) that were available along with genotype and gene expression data for 5,257 FHS participants. Among *cis*-eQTLs, we identified the *SH2B3* (MIM: 605093)/*ALDH2* (MIM: 100650) locus as having a causal effect on DBP ($p = 0.005$) and SBP ($p = 0.02$) through *ALDH2* expression (Table 3). SNPs in this locus are associated with coronary artery disease (CAD)/myocardial infarction (MI), blood pressure, LDL-cholesterol, and type 1 diabetes (Figure 4). Not only was the *cis*-locus found to be associated with risk of CAD/MI,[37] recent studies describe the *trans*-regulation of *MYADM* (MIM: 609959) and *TAGAP* (MIM: 609667) expression by the same *trans*-eQTL.[38,39] In addition, we found on average that 11% of *trans*-regulation of *trans*-eGenes for this module is mediated through expression of *ALDH2*, suggesting a new target and regulatory mechanism related to this CAD/MI module. Two additional causal loci are *RAB3GAP1* (MIM: 602536) for total cholesterol (eGenes *MCM6* [MIM: 601806] and *DARS* [MIM: 603084]) and *LOC105375745* for triglycerides (eGene *TRIB1* [MIM: 609461]). Among *trans*-eQTLs, we identified the *TAGAP* locus as having a causal effect on LDL and total cholesterol through expression of *TAGAP*. The *TAGAP* locus has been found to be significantly associated with lipoprotein (a) levels[40] and *TAGAP* was reported to be differentially expressed in CAD patients[41] and after atorvastatin treatment.[42] Another possible mechanism to explain the association of *trans*-eQTLs with the expression of their *trans*-eGenes is reverse causality, whereby an eQTL alters expression of a *trans*-eGene through its effect on phenotype. In this case, the phenotype serves as a mediator (feedback effect). CIT, however, did not identify any examples of reverse causal effects (at $p < 0.05$).

Discussion

Identifying disease-causal genes and variants within GWASs results is an enormous challenge that simple association analysis cannot address.[43] Unlike GWASs, where the association between a genetic variant and trait is unidirectional, in transcriptome-wide association studies (TWASs) the direction of association between transcript and phenotype is not clear and causal inference must be drawn with caution.[44] eQTL studies hold the promise of revealing biological mechanisms of SNP-phenotype associations; integrating GWASs with TWASs may help prioritize genes and variants for functional studies.[44] In this study,

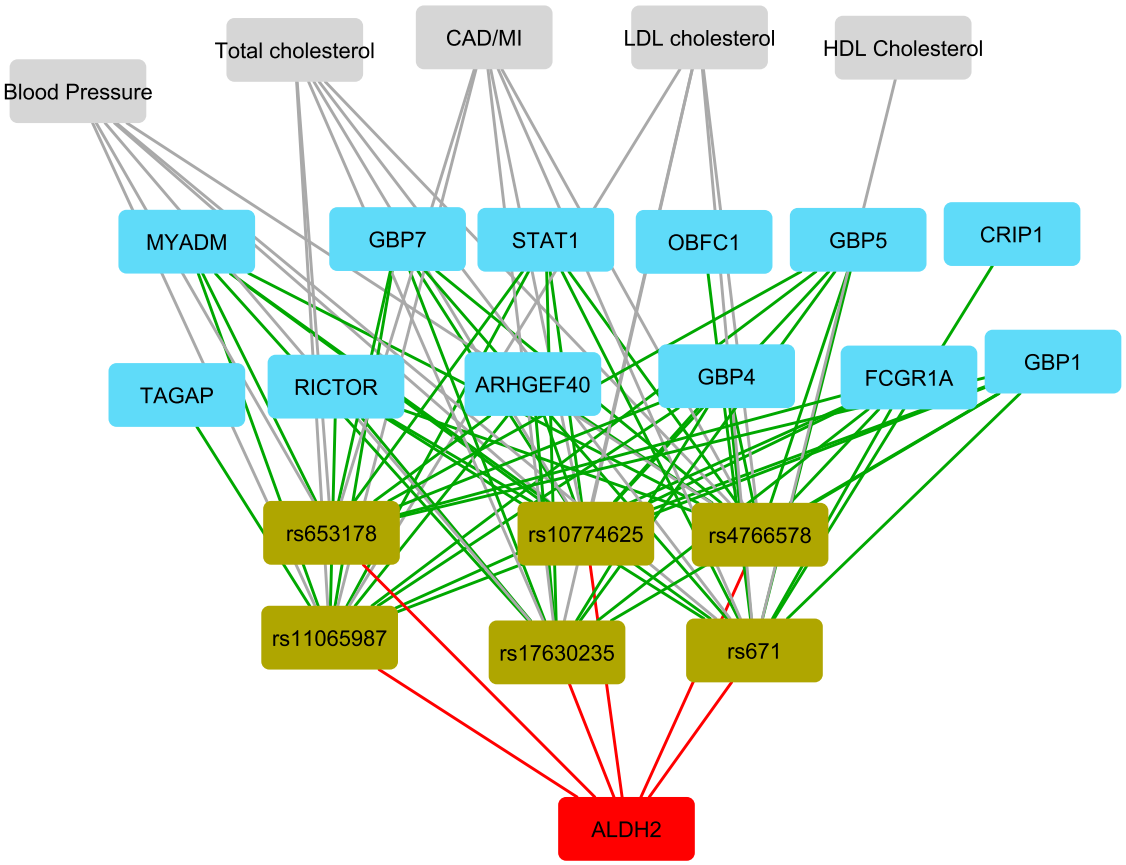

Figure 4. *cis*- and *trans*-eQTLs in the *ALDH2* Causal Module
Gray boxes list traits associated with SNPs from GWASs, green boxes list SNPs, red box lists the *cis*-eGene, and blue boxes list *trans*-eGenes. Red edges represent *cis*-associations; green edges represent *trans*-associations.

we used a causal inference approach to infer causal relations and their directionality by integrating SNPs from GWASs with gene expression and phenotype data predicated on the assumption that if a gene is causally related to a phenotype, a nearby genetic variant (i.e., a *cis*-eQTL) that explains a large proportion of its expression should be associated with the same phenotype.

We discovered that many *trans*-eQTL-*trans*-eGene associations are mediated by *cis*-eGene expression, reflecting a complex regulatory mechanism. An intuitive explanation for hidden regulation of *trans*-eGenes is TFs that directly influence gene transcription. Although we found no enrichment for TFs among *trans*-eGenes, we found that more than one-third of *cis*-eGenes shared a common motif with *trans*-eGenes from the same *trans*-hotspot, indicating that there may exist indirect relations of *cis*-eGenes to TFs. For example, we identified *trans*-eGenes in the chromosome 7 hotspot that were enriched for the targets of TF *PAX4* (MIM: 167413). The *cis*-eGene at this hotspot is TF *IKZF1* and although *PAX4* and *IKZF1* are different TFs, they share common motifs.

Using different cell types and populations, Pierce et al.[15] also reported a similar proportion (~20%) of *trans*-eQTLs that act through *cis*-mediation, indicating that the mechanism of *cis*-eGene mediation of *trans*-eGene expression may be a common feature genome wide. To extend this concept, we explored how this phenomenon affects disease pathways. For example, we observed that rs174538, which was reported to be associated with plasma phospholipids in GWASs,[45] is a *trans*-eQTL of *LDLR* (MIM: 606945) expression ($p = 3.69 \times 10^{-8}$). This association, however, was not significant after adjusting for expression of *FADS2* (a *cis*-eGene of rs174538) and the proportion of mediation of *FADS2* (MIM: 606149) on *LDLR* was 100%. *FADS2* is a key gene influencing n-3 polyunsaturated fatty acids (PUFA) levels and PUFA levels have been found to upregulate LDL receptor protein expression in fibroblasts and HepG2 cells,[46] indicating a likely pathway from PUFA to lipid metabolism. A recent study reported that *Fads1* KO mice had 40% less atheromatous plaque compared to wild-type littermates.[47] Therefore, the *FADS* gene could be a putative therapeutic target for cardiovascular disease prevention and treatment.

We found that 10 out of 13 *trans*-eQTL hotspots are blood trait related and five of them replicated in the Blood eQTL Browser.[12] Among the 227 *trans*-eGenes associated with platelet SNPs, 26 were reported as platelet eQTL-genes,[48] suggesting that *trans*-eQTLs are highly tissue specific and that SNPs might remotely affect tissue-specific eGenes. For example, some of the loci identified in GWASs for platelet traits (e.g., *ARHGEF3*) affect the expression of

hundreds of genes and may be key drivers of hematopoiesis and affect multiple blood cell lineages.[49]

This study has several limitations. First, the Blood eQTL Browser[12] is the only database that includes extensive *trans*-eQTL results in a comparable large sample size. Therefore, our results cannot readily be validated in other tissues as most other large eQTL databases provide only *cis*-eQTLs. Second, although ours is one of the largest studies to detect *trans*-eQTLs, we are still underpowered for causal inference testing, which tests the SNP-phenotype association as the first condition. Therefore, many genes were excluded from causality testing because they did not fulfill the first condition.

In summary, we provide evidence of a *cis*-mediated mechanism that explains distal regulation of *trans*-eGenes by their *trans*-eQTLs. Importantly, the causal loci, especially the *trans*-eQTLs identified from our integrative genomic approach, could not be detected from traditional GWASs by searching SNPs around the GWAS signal. Our next steps are to explore eQTL data from more disease-related tissues and to incorporate whole-genome sequence data to identify more causal eQTLs. We speculate that it may be worthwhile to apply this approach across eQTL databases and across multiple phenotypes as a means of identifying plausible targets for therapeutic intervention.

Supplemental Data

Supplemental Data include three figures and eight tables and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2017.02.003.

Acknowledgments

We thank all of the study participants who helped to create this valuable resource and supported this work. We thank the data management group of FHS for organizing and providing these data. We thank the NIH Fellows Editorial Board members for their valuable edits and comments. This study used the high-performance computational capabilities of the Biowulf Linux cluster at the NIH. The FHS is funded by NIH contract N01-HC-25195. The laboratory work for this investigation was funded by the Division of Intramural Research, National Heart, Lung, and Blood Institute, NIH. The analytical component of this project was funded by American Heart Association (AHA) Cardiovascular Genome-Phenome Study (CVGPS) grant 15CVGPS23430000. The views expressed in this manuscript are those of the authors and do not necessarily represent the views of the National Heart, Lung, and Blood Institute; the NIH; or the U.S. Department of Health and Human Services.

Received: November 1, 2016
Accepted: February 1, 2017
Published: March 9, 2017

Web Resources

Bioconductor, http://www.bioconductor.org
Blood eQTL Browser, http://genenetwork.nl/bloodeqtlbrowser/
GRASP, downloaded June 2016, http://grasp.nhlbi.nih.gov/Overview.aspx
HaploReg v.4.1, http://www.broadinstitute.org/mammals/haploreg/haploreg.php
mediation: Causal Mediation Analysis, https://cran.r-project.org/web/packages/mediation/index.html
OMIM, http://www.omim.org/
R statistical software, http://www.r-project.org/
SNiPA, http://snipa.helmholtz-muenchen.de/snipa/

References

1. Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., and Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. *42*, D1001–D1006.
2. Eicher, J.D., Landowski, C., Stackhouse, B., Sloan, A., Chen, W., Jensen, N., Lien, J.P., Leslie, R., and Johnson, A.D. (2015). GRASP v2.0: an update on the Genome-Wide Repository of Associations between SNPs and phenotypes. Nucleic Acids Res. *43*, D799–D804.
3. Visscher, P.M., Brown, M.A., McCarthy, M.I., and Yang, J. (2012). Five years of GWAS discovery. Am. J. Hum. Genet. *90*, 7–24.
4. Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature *511*, 421–427.
5. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.
6. Schaub, M.A., Boyle, A.P., Kundaje, A., Batzoglou, S., and Snyder, M. (2012). Linking disease associations with regulatory information in the human genome. Genome Res. *22*, 1748–1759.
7. Rockman, M.V., and Kruglyak, L. (2006). Genetics of global gene expression. Nat. Rev. Genet. *7*, 862–872.
8. Michaelson, J.J., Loguercio, S., and Beyer, A. (2009). Detection and interpretation of expression quantitative trait loci (eQTL). Methods *48*, 265–276.
9. Atanasovska, B., Kumar, V., Fu, J., Wijmenga, C., and Hofker, M.H. (2015). GWAS as a driver of gene discovery in cardiometabolic diseases. Trends Endocrinol. Metab. *26*, 722–732.
10. Breitling, R., Li, Y., Tesson, B.M., Fu, J., Wu, C., Wiltshire, T., Gerrits, A., Bystrykh, L.V., de Haan, G., Su, A.I., and Jansen, R.C. (2008). Genetical genomics: spotlight on QTL hotspots. PLoS Genet. *4*, e1000232.
11. Stranger, B.E., Montgomery, S.B., Dimas, A.S., Parts, L., Stegle, O., Ingle, C.E., Sekowska, M., Smith, G.D., Evans, D., Gutierrez-Arcelus, M., et al. (2012). Patterns of cis regulatory variation in diverse human populations. PLoS Genet. *8*, e1002639.
12. Westra, H.J., Peters, M.J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., Christiansen, M.W., Fairfax, B.P., Schramm, K., Powell, J.E., et al. (2013). Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat. Genet. *45*, 1238–1243.
13. Zhang, X., Gierman, H.J., Levy, D., Plump, A., Dobrin, R., Goring, H.H., Curran, J.E., Johnson, M.P., Blangero, J., Kim, S.K., et al. (2014). Synthesis of 53 tissue and cell line expression QTL datasets reveals master eQTLs. BMC Genomics *15*, 532.

14. Bryois, J., Buil, A., Evans, D.M., Kemp, J.P., Montgomery, S.B., Conrad, D.F., Ho, K.M., Ring, S., Hurles, M., Deloukas, P., et al. (2014). Cis and trans effects of human genomic variants on gene expression. PLoS Genet. *10*, e1004461.
15. Pierce, B.L., Tong, L., Chen, L.S., Rahaman, R., Argos, M., Jasmine, F., Roy, S., Paul-Brutus, R., Westra, H.J., Franke, L., et al. (2014). Mediation analysis demonstrates that trans-eQTLs are often explained by cis-mediation: a genome-wide analysis among 1,800 South Asians. PLoS Genet. *10*, e1004818.
16. Yao, C., Chen, B.H., Joehanes, R., Otlu, B., Zhang, X., Liu, C., Huan, T., Tastan, O., Cupples, L.A., Meigs, J.B., et al. (2015). Integromic analysis of genetic variation and gene expression identifies networks for cardiovascular disease phenotypes. Circulation *131*, 536–549.
17. Joehanes, R., Zhang, X., Huan, T., Yao, C., Ying, S.X., Nguyen, Q.T., Demirkale, C.Y., Feolo, M.L., Sharopova, N.R., Sturcke, A., et al. (2017). Integrated genome-wide analysis of expression quantitative trait loci aids interpretation of genomic association studies. Genome Biol. *18*, 16.
18. Mahmood, S.S., Levy, D., Vasan, R.S., and Wang, T.J. (2014). The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. Lancet *383*, 999–1008.
19. Joehanes, R., Johnson, A.D., Barb, J.J., Raghavachari, N., Liu, P., Woodhouse, K.A., O'Donnell, C.J., Munson, P.J., and Levy, D. (2012). Gene expression analysis of whole blood, peripheral blood mononuclear cells, and lymphoblastoid cell lines from the Framingham Heart Study. Physiol. Genomics *44*, 59–75.
20. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., and Abecasis, G.R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat. Genet. *44*, 955–959.
21. Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., and Speed, T.P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics *4*, 249–264.
22. Bates, D., Machler, M., Bolker, B.M., and Walker, S.C. (2015). Fitting linear mixed-effects models using lme4. J. Stat. Softw. *67*, 1–48.
23. Stegle, O., Parts, L., Piipari, M., Winn, J., and Durbin, R. (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nat. Protoc. *7*, 500–507.
24. Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate - a practical and powerful approach to multiple testing. J. R. Stat. Soc. B *57*, 289–300.
25. Millstein, J., Zhang, B., Zhu, J., and Schadt, E.E. (2009). Disentangling molecular relationships with a causal inference test. BMC Genet. *10*, 23.
26. Orozco, L.D., Morselli, M., Rubbi, L., Guo, W., Go, J., Shi, H., Lopez, D., Furlotte, N.A., Bennett, B.J., Farber, C.R., et al. (2015). Epigenome-wide association of liver methylation patterns and complex metabolic traits in mice. Cell Metab. *21*, 905–917.
27. Ward, L.D., and Kellis, M. (2012). HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res. *40*, D930–D934.
28. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550.
29. Lizio, M., Harshbarger, J., Abugessaisa, I., Noguchi, S., Kondo, A., Severin, J., Mungall, C., Arenillas, D., Mathelier, A., Medvedeva, Y.A., et al. (2017). Update of the FANTOM web resource: high resolution transcriptome of diverse cell types in mammals. Nucleic Acids Res. *45* (D1), D737–D743.
30. Rolland, T., Taşan, M., Charloteaux, B., Pevzner, S.J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R., et al. (2014). A proteome-scale map of the human interactome network. Cell *159*, 1212–1226.
31. Nicolae, D.L., Gamazon, E., Zhang, W., Duan, S., Dolan, M.E., and Cox, N.J. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet. *6*, e1000888.
32. McKenzie, M., Henders, A.K., Caracella, A., Wray, N.R., and Powell, J.E. (2014). Overlap of expression quantitative trait loci (eQTL) in human brain and blood. BMC Med. Genomics *7*, 31.
33. Thompson, D., Regev, A., and Roy, S. (2015). Comparative analysis of gene regulatory networks: from network reconstruction to evolution. Annu. Rev. Cell Dev. Biol. *31*, 399–428.
34. Kobayashi, K.S., and van den Elsen, P.J. (2012). NLRC5: a key regulator of MHC class I-dependent immune responses. Nat. Rev. Immunol. *12*, 813–820.
35. Meissner, T.B., Liu, Y.J., Lee, K.H., Li, A., Biswas, A., van Eggermond, M.C., van den Elsen, P.J., and Kobayashi, K.S. (2012). NLRC5 cooperates with the RFX transcription factor complex to induce MHC class I gene expression. J. Immunol. *188*, 4951–4958.
36. Millstein, J., Chen, G.K., and Breton, C.V. (2016). cit: hypothesis testing software for mediation analysis in genomic applications. Bioinformatics *32*, 2364–2365.
37. Deloukas, P., Kanoni, S., Willenborg, C., Farrall, M., Assimes, T.L., Thompson, J.R., Ingelsson, E., Saleheen, D., Erdmann, J., Goldstein, B.A., et al.; CARDIoGRAMplusC4D Consortium; DIAGRAM Consortium; CARDIOGENICS Consortium; MuTHER Consortium; and Wellcome Trust Case Control Consortium (2013). Large-scale association analysis identifies new risk loci for coronary artery disease. Nat. Genet. *45*, 25–33.
38. Hunt, K.A., Zhernakova, A., Turner, G., Heap, G.A., Franke, L., Bruinenberg, M., Romanos, J., Dinesen, L.C., Ryan, A.W., Panesar, D., et al. (2008). Newly identified genetic risk variants for celiac disease related to the immune response. Nat. Genet. *40*, 395–402.
39. Huan, T., Esko, T., Peters, M.J., Pilling, L.C., Schramm, K., Schurmann, C., Chen, B.H., Liu, C., Joehanes, R., Johnson, A.D., et al.; International Consortium for Blood Pressure GWAS (ICBP) (2015). A meta-analysis of gene expression signatures of blood pressure and hypertension. PLoS Genet. *11*, e1005035.
40. Lu, W., Cheng, Y.C., Chen, K., Wang, H., Gerhard, G.S., Still, C.D., Chu, X., Yang, R., Parihar, A., O'Connell, J.R., et al. (2015). Evidence for several independent genetic variants affecting lipoprotein (a) cholesterol levels. Hum. Mol. Genet. *24*, 2390–2400.
41. Arvind, P., Jayashree, S., Jambunathan, S., Nair, J., and Kakkar, V.V. (2015). Understanding gene expression in coronary artery disease through global profiling, network analysis and independent validation of key candidate genes. J. Genet. *94*, 601–610.

42. Won, H.H., Kim, S.R., Bang, O.Y., Lee, S.C., Huh, W., Ko, J.W., Kim, H.G., McLeod, H.L., O'Connell, T.M., Kim, J.W., and Lee, S.Y. (2012). Differentially expressed genes in human peripheral blood as potential markers for statin response. J. Mol. Med. *90*, 201–211.
43. Wang, K., Dickson, S.P., Stolle, C.A., Krantz, I.D., Goldstein, D.B., and Hakonarson, H. (2010). Interpretation of association signals and identification of causal variants from genome-wide association studies. Am. J. Hum. Genet. *86*, 730–742.
44. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W., Jansen, R., de Geus, E.J., Boomsma, D.I., Wright, F.A., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. Nat. Genet. *48*, 245–252.
45. Lemaitre, R.N., Tanaka, T., Tang, W., Manichaikul, A., Foy, M., Kabagambe, E.K., Nettleton, J.A., King, I.B., Weng, L.C., Bhattacharya, S., et al. (2011). Genetic loci associated with plasma phospholipid n-3 fatty acids: a meta-analysis of genome-wide association studies from the CHARGE Consortium. PLoS Genet. *7*, e1002193.
46. Yu-Poth, S., Yin, D., Kris-Etherton, P.M., Zhao, G., and Etherton, T.D. (2005). Long-chain polyunsaturated fatty acids upregulate LDL receptor protein expression in fibroblasts and HepG2 cells. J. Nutr. *135*, 2541–2545.
47. Powell, D.R., Gay, J.P., Smith, M., Wilganowski, N., Harris, A., Holland, A., Reyes, M., Kirkham, L., Kirkpatrick, L.L., Zambrowicz, B., et al. (2016). Fatty acid desaturase 1 knockout mice are lean with improved glycemic control and decreased development of atheromatous plaque. Diabetes Metab. Syndr. Obes. *9*, 185–199.
48. Simon, L.M., Chen, E.S., Edelstein, L.C., Kong, X., Bhatlekar, S., Rigoutsos, I., Bray, P.F., and Shaw, C.A. (2016). Integrative multi-omic analysis of human platelet eQTLs reveals alternative start site in mitofusin 2. Am. J. Hum. Genet. *98*, 883–897.
49. Gieger, C., Radhakrishnan, A., Cvejic, A., Tang, W., Porcu, E., Pistis, G., Serbanovic-Canic, J., Elling, U., Goodall, A.H., Labrune, Y., et al. (2011). New gene functions in megakaryopoiesis and platelet formation. Nature *480*, 201–208.