Whole genome DNA and RNA sequencing of whole blood elucidates the genetic architecture of gene expression underlying a wide range of diseases

Chunyu Liu1,2,8, Roby Joehanes3,8,*, Jiantao Ma4, Yuxuan Wang1, Xianbang Sun1, Amena Keshawarz5, Meera Sooda3, Tianxiao Huan3, Shih-Jen Hwang3, Helena Bui1, Brandon Tejada3, Peter J. Munson3, Cumhur Y. Demirkale6, Nancy L. Heard-Costa2,6, Achilleas N. Pitsillides1, Gina M. Peloso1, Michael Feolo7, Nataliya Sharopova7, Ramachandran S. Vasan2,6 & Daniel Levy2,3,8*

To create a scientific resource of expression quantitative trait loci (eQTL), we conducted a genome-wide association study (GWAS) using genotypes obtained from whole genome sequencing (WGS) of DNA and gene expression levels from RNA sequencing (RNA-seq) of whole blood in 2622 participants in Framingham Heart Study. We identified 6,778,286 cis-eQTL variant-gene transcript (eGene) pairs at $p < 5 \times 10^{-8}$ (2,855,111 unique cis-eQTL variants and 15,982 unique eGenes) and 1,469,754 trans-eQTL variant-eGene pairs at $p < 1e^{-12}$ (526,056 unique trans-eQTL variants and 7233 unique eGenes). In addition, 442,379 cis-eQTL variants were associated with expression of 1518 long non-protein coding RNAs (lncRNAs). Gene Ontology (GO) analyses revealed that the top GO terms for cis-eGenes are enriched for immune functions (FDR < 0.05). The cis-eQTL variants are enriched for SNPs reported to be associated with 815 traits in prior GWAS, including cardiovascular disease risk factors. As proof of concept, we used this eQTL resource in conjunction with genetic variants from public GWAS databases in causal inference testing (e.g., COVID-19 severity). After Bonferroni correction, Mendelian randomization analyses identified putative causal associations of 60 eGenes with systolic blood pressure, 13 genes with coronary artery disease, and seven genes with COVID-19 severity. This study created a comprehensive eQTL resource via BioData Catalyst that will be made available to the scientific community. This will advance understanding of the genetic architecture of gene expression underlying a wide range of diseases.

Over the past decade, genome-wide association studies (GWAS) have revolutionized understanding of the genetic architecture of complex traits1. To date, GWAS have reported more than 59,000 associations (at $p < 5 \times 10^{-8}$) between common genetic variants and numerous phenotypes (GWAS Catalog, v1.0.2)2. Yet, despite the clear success of GWAS, most single-nucleotide polymorphisms (SNPs) identified in GWAS reside in non-coding regions3–5 and do not illuminate causal mechanisms underlying SNP-trait associations5. We posit that many of these trait-associated non-coding SNPs are likely to be involved in the regulation of gene expression.
Expression quantitative trait locus (eQTL) analysis seeks to identify genetic variants that affect the expression of local (cis) or distant (trans) genes (eGenes). Until recently, eQTL analysis has relied on high-throughput microarray technologies and spawned a wave of genome-wide eQTL studies including a recent study from our group. These studies aided the understanding of the functional relevance of many GWAS results. Importantly, a hypothesis-free genome-wide eQTL approach permits the identification of new putatively functional loci without requiring previous knowledge of specific regulatory regions.

Most previous eQTL analyses were limited by small sample sizes and by the imprecision of microarrays. Newer technologies of RNA sequencing (RNA-seq) and whole genome sequencing (WGS) of DNA add greater precision and relevance to eQTL analyses. In conjunction with the National Heart, Lung, and Blood Institute’s (NHLBI) Trans-Omics for Precision Medicine (TOPMed) Program, the Framingham Heart Study (FHS) has obtained whole genome sequencing (WGS) in ~6100 study participants to help understand the molecular basis of heart, lung, blood, and sleep disorders and to advance precision medicine. Among FHS participants with WGS, RNA-seq was obtained in 2622 participants. We conducted genome-wide eQTL analyses using high-precision genotypes obtained via WGS and gene expression levels from RNA-seq of whole blood. The primary objectives of this study were three-fold. Firstly, it sought to provide a scientific resource of cis and trans gene-level eQTL data to facilitate understanding of the genetic architecture of gene expression traits. Secondly, it was aimed to provide eQTL data for long noncoding RNAs (lncRNAs) that were not captured in prior array-based eQTL studies. Thirdly, it attempted to demonstrate the utility of the eQTL resource in causal inference analyses.

Results

Of the 2622 FHS participants in eQTL analyses, 720 participants were from the FHS Offspring cohort (mean age 71 ± 8 years; 59% women) and 1902 were from the Third Generation cohort (mean age 47 ± 8 years; 52% women) (Supplemental Table 1). We used 19,624,299 SNPs with a minor allele count (MAC) ≥ 10 and 58,870 expression levels in association analyses to identify gene-level eQTLs. We evaluated the genomic inflation factor ($\lambda_{GC}$). The observed $\lambda_{GC}$ = 1.03, indicating that inflation was unlikely for the eQTL analyses (Supplemental Fig. 1).

Gene-level eQTL results. cis-eQTLs. Cis-eQTLs was defined as SNPs within 1 Mb of the transcription start sites (TSSs) of targeting genes. We identified 6,778,286 significant cis-eQTL variant-eGene pairs from 2,855,111 unique cis-eQTL variants and 15,982 unique eGenes (at $p < 5 \times 10^{-8}$) (Supplemental Table 2). The median number of cis-eQTL variants per gene was 183 (interquartile range = 47,463). The eGenes harboring the largest numbers of cis-eQTL variants are located in the human leukocyte antigen (HLA) or major histocompatibility complex (MHC) on chromosome 6, reflecting a large number of SNPs in strong linkage disequilibrium (LD) at the MHC locus. Owing to the computational burden, we selected the strongest cis-eQTL variant (i.e., the lead variant) as that which had the lowest p-value per eGene. If several cis-eQTL displayed the same p-value (i.e., they are in perfect LD, $r^2 = 1$), we randomly select one lead eQTL variant per eGene (Supplemental Table 3) and the top 25 pairs was displayed in Table 1. Of the 15,982 significant unique cis-eQTL variant-eGene pairs, 82.8% (n = 13,236) of SNPs were within 100 kb of the transcription start sites (TSSs) of the respective eGenes, 9.3% (n = 1486) within 101 kb–200 kb region, 5.7% (n = 909) within 201 kb–500 kb region, and 2.2% (n = 351) within 501 kb–1 Mb (Fig. 1A). Published GWAS and QTL analyses revealed that rare variants have larger effect sizes than common variants. Therefore, we compared the median, 25th percentile, and 75th percentile of the absolute values of effect sizes for lead cis-eQTL variants across four variant groups based on minor allele frequencies (MAFs). We found that rare cis-eQTL variants displayed larger effect sizes (median effect size 0.44 versus 1.77 for cis-eQTL variants with MAF in 0.1–0.5 versus cis-eQTL variants with MAF in 0.003–0.01) (trend test $P < 0.001$) (Table 2).

trans-eQTLs. Trans-eQTLs referred to the SNPs that were beyond of 1 Mb of the TSSs of the eGenes on the same chromosome or those on the different chromosomes of the eGenes. We identified 1,469,754 significant trans-eQTL variant-eGene pairs ($p < 1e$−12) from 526,056 unique trans-eQTL variants and 7233 trans-eGenes (Table 3, Supplemental Table 4). The median number of significant-eQTL variants per eGene was 11 (interquartile range = 2, 76)14. With the same method used to select the lead cis-eQTL variants, we selected the lead trans-eQTL variant based on p-values for each trans-eGene, yielding 7233 unique trans-eQTL-eGene pairs (Supplemental Table 4). We further compared the effect sizes of the lead trans-eQTL variants based on their MAF. We found that rare trans-eQTL variants (MAF in 0.003–0.01) displayed larger effect sizes (median effect size 0.42 versus 2.38 for common trans-eQTL variants (MAF in 0.1–0.5) (trend test $P < 0.001$) (Table 2).

Long noncoding RNA (lncRNA) eQTLs. lncRNAs are usually more than 200 bases in length, share no conserved sequence homology, and have variable functions. Of the 58,870 transcripts captured by RNA-seq, 7696 (13%) are lncRNAs. Of the significant cis-eQTL variant-eGene pairs (n = 6,778,286, $p < 5e$−8), 565,769 cis-eQTL variants are associated with expression of 1516 unique cis-lncRNAs (Supplemental Table 5). Of the significant trans-eQTL variant-eGene pairs ($p < 1e$−12), 164,386 trans-eQTL variants were associated with expression of 475 trans-lncRNAs (Supplemental Table 6). Three cis-eQTL-lncRNA pairs were observed among the top 25 cis-eQTL results (Table 1). The top cis-lncRNA, the MAP3K2 divergent transcript (MAP3K2-DT), is the only lncRNA that is located adjacent to a protein coding gene, the 5’-end of mitogen-activated protein kinase kinase kinase 2 (MAP3K2) on chromosome 2 (q14.3) (Supplemental Fig. 2). The correlation of expression of MAP3K2 and MAP3K2-DT was weak (Pearson correlation = 0.08; $p = 0.12$). Among the top 25 trans-eQTL pairs, we identified one trans-eQTL-lncRNA pair (Table 3). The top trans-lncRNA, AP001005.3 on chromosome 18, is not adjacent to any known genes.
Table 1. Top 25 cis-eQTLs (p < 5e−8). EA effect allele, OA the other allele, EAF effect allele frequency; R² variance explained by the lead eQTL (SNP).

| Gene Symbol | SNP | Chr | SNP position | Gene start position | R²  | Beta | log10P  | OA | EA | EAF          | Type               |
|-------------|-----|-----|--------------|---------------------|-----|------|---------|----|----|--------------|--------------------|
| PPIE        | rs7513045 | 1   | 39,738,494   | 39,692,182          | 0.84| 11.40| −1029.25| G  | T  | 0.36         | protein_coding     |
| CCDC163     | rs4660860  | 1   | 45,480,561   | 45,493,866          | 0.90| 3.30 | −1286.89| T  | A  | 0.30         | protein_coding     |
| CYP26B1     | rs13430651 | 2   | 72,215,195   | 72,129,238          | 0.81| 1.98 | 920.005 | G  | A  | 0.15         | protein_coding     |
| MAP3K2-DT   | rs2276683  | 2   | 127,389,186  | 127,389,130         | 0.88| 1.61 | −1176.37| G  | C  | 0.23         | lincRNA             |
| SLC12A7     | rs35188965  | 5   | 1,104,823    | 1,050,384           | 0.81| 29.87| 915.459 | C  | T  | 0.44         | protein_coding     |
| ENC1        | rs112722452 | 5   | 74,631,048   | 74,627,406          | 0.83| 14.53| −986.798| CA | C  | 0.11         | protein_coding     |
| ERAF2       | rs2910868  | 5   | 96,916,885   | 96,875,939          | 0.85| 36.98| 1044.91 | T  | C  | 0.43         | protein_coding     |
| BTLN3       | rs72494581 | 5   | 181,003,797  | 180,988,845         | 0.82| 13.52| 950.405 | T  | C  | 0.30         | protein_coding     |
| HLA-DRB5    | rs86176300 | 5   | 127,389,186  | 127,389,130         | 0.88| 17.13| 1003.76 | G  | T  | 0.15         | protein_coding     |
| AL512625.3  | rs112722452 | 9   | 62,906,092   | 62,856,999          | 0.83| 11.19| −993.655| T  | C  | 0.13         | lincRNA             |
| CUTALP      | rs13298616 | 9   | 120,832,525  | 120,824,828         | 0.86| 23.25| 982.88  | T  | C  | 0.40         | transcribed_unitary_pseudogene |
| LDHIC       | rs201993031 | 9   | 18,412,985   | 18,412,318          | 0.82| 0.16 | 946.833 | C  | T  | 0.12         | protein_coding     |
| ACCS        | rs2074038  | 11  | 44,066,439   | 44,065,925          | 0.83| 8.69 | 997.26  | T  | A  | 0.11         | protein_coding     |
| FAD52       | rs968567   | 11  | 61,828,092   | 61,792,980          | 0.88| 31.41| 1186.37 | C  | T  | 0.17         | protein_coding     |
| XBRAL       | rs10890051 | 11  | 74,931,506   | 74,807,739          | 0.91| 5.38 | 1327.88 | G  | A  | 0.26         | protein_coding     |
| B4GALNT3    | rs1056008  | 12  | 553,672      | 460,364             | 0.85| 6.71 | 1043.34 | T  | C  | 0.25         | protein_coding     |
| DDX11       | rs3891006  | 12  | 31,073,506   | 31,073,860          | 0.86| 13.25| 1102.08 | A  | G  | 0.44         | protein_coding     |
| RPS26       | rs1131017  | 12  | 56,042,145   | 56,041,351          | 0.81| 134.34| 929.902 | C  | G  | 0.39         | protein_coding     |
| C10or97     | rs7503725  | 17  | 510,305      | 510,325             | 0.85| 1.89 | 1055.68 | T  | C  | 0.25         | protein_coding     |
| AC126544.2  | rs2696531  | 17  | 46,278,268   | 45,586,452          | 0.86| 1.04 | 1097.79 | C  | A  | 0.21         | lincRNA             |
| SPATA20     | rs9890020  | 17  | 50,547,162   | 50,543,058          | 0.81| 1.01 | 341.73  | A  | C  | 0.37         | protein_coding     |
| CEACAMP3    | rs3745936  | 19  | 41,586,462   | 41,599,735          | 0.84| 11.11| 400.05  | A  | T  | 0.22         | transcribed_unprocessed_pseudogene |
| PWP2        | rs2277806  | 21  | 44,089,769   | 44,107,373          | 0.87| 3.16 | 139.85  | A  | C  | 0.19         | protein_coding     |
| GATD3A      | rs37881104 | 21  | 44,092,213   | 44,133,610          | 0.86| 4.25 | 104.35  | G  | A  | 0.18         | protein_coding     |
| FAM118A     | rs576259663 | 22  | 45,363,712   | 45,308,968          | 0.86| 43.45| 1108.47 | T  | TA | 0.12         | protein_coding     |

Figure 1. Variance in eGenes explained by lead e-QTL variants. (A) Variance in eGenes explained by lead cis-eQTLs in relation to the distance in mega base pairs of the lead cis-eQTLs to the transcription start site (TSS) of the cis-gene. (B) Comparison of median variance in the expression levels of eGenes explained by the lead eQTL variants. The median value of variance explained in cis-eGenes was significantly higher than that of trans-eGenes (4.8% versus 2.8%, p < 2e−16). For cis-eGenes, the protein-coding-eGenes had slightly higher median variance explained in the expression level than cis-lncRNA genes (5.1% versus 4.3%, p = 0.0004). No significant difference in variance explained in trans-eGenes of protein-coding genes versus trans-lncRNA genes (2.8% versus 2.7%, p = 0.41).
Variance in expression explained by lead eQTL variants. The narrow-sense heritability is the proportion of phenotypic variance explained by additive genetic effects\textsuperscript{17}. We estimated the proportion of variance ($R^2$) in the expression of a gene that was explained by the lead cis-eQTL or trans-eQTL variant. We found that the median $R^2$ value of lead cis-eQTLs was significantly higher than that of lead trans-eQTLs ($R^2$ 4.8% versus 2.8%, $p < 2.2\times10^{-16}$) (Fig. 1B). For the majority of cis-eGenes (85%) and trans-eGenes (96%), the lead eQTL variants only explained a small proportion of variance in expression ($R^2 < 0.2$). The lead e-QTLs explained a large proportion of variance in expression ($R^2 > 0.8$) of an extremely small proportion of eGenes (0.17% of cis-eGenes and five trans-eGenes). We further compared the median values of cis-eQTL $R^2$ versus trans-eQTL $R^2$ between protein-coding genes and lncRNA genes. For cis-eGenes, protein-coding-eGenes had slightly higher median variance.

Table 2. Comparison of effect sizes of eQTLs with different minor allele frequency ranges. MAF minor allele frequency, $n$ the number of eQTLs, effect size the effect size of association between eQTL and eGene.

| MAF range | $n$ | Median effect size | 1st Quartile effect size | 3rd Quartile effect size |
|-----------|-----|---------------------|-------------------------|-------------------------|
| Cis-eQTL variants |
| [0.1, 0.5) | 1,960,534 | 0.44 | 0.12 | 1.41 |
| [0.05, 0.1) | 390,362 | 0.64 | 0.17 | 2.02 |
| [0.01, 0.05) | 424,024 | 0.98 | 0.29 | 2.84 |
| (0.003–0.01) | 80,191 | 1.77 | 0.50 | 5.03 |
| Trans-eQTL variants |
| [0.1, 0.5) | 321,275 | 0.42 | 0.12 | 1.01 |
| [0.05, 0.1) | 81,083 | 0.69 | 0.17 | 1.36 |
| [0.01, 0.05) | 94,234 | 0.89 | 0.38 | 1.71 |
| (0.003–0.01) | 29,464 | 2.38 | 0.80 | 8.13 |

Table 3. Top 25 top trans-eQTLs ($p < 1\times10^{-12}$). EA effect allele, OA the other allele, EAF effect allele frequency, $R^2$ variance explained by the lead eQTL (SNP).

| Gene symbol | SNP | Gene Chr | SNP Chr | SNP Pos | Gene Start Pos | R$^2$ | Beta | t value | log10P | OA | EA | EAF Gene type |
|-------------|-----|----------|---------|--------|---------------|------|------|---------|-------|----|----|-----|
| EMBP1       | rs4549528 | 1 | 5 | 50,372,700 | 121,519,112 | 0.70 | 1.63 | 78.04 | −677.53 | T | C | 0.48 transferred_unprocessed_pseudogene |
| AL365357.1  | rs8481   | 1 | 5 | 150,446,963 | 178,411,616 | 0.64 | 2.79 | 67.64 | −570.81 | C | T | 0.25 processed_pseudogene |
| AL591846.1  | rs13161099 | 1 | 5 | 150,442,799 | 206,695,837 | 0.62 | 1.87 | 65.67 | −549.988 | G | A | 0.25 processed_pseudogene |
| AC004057.1  | rs1131017 | 4 | 12 | 56,042,145 | 113,214,046 | 0.61 | −0.42 | 63.03 | −521.823 | C | G | 0.39 transcribed_processed_pseudogene |
| RPL10P9     | rs6655287 | 5 | X | 154,396,528 | 168,616,352 | 0.64 | 4.11 | 68.52 | −580.03 | T | C | 0.10 processed_pseudogene |

Variance in expression explained by lead eQTL variants. The narrow-sense heritability is the proportion of phenotypic variance explained by additive genetic effects\textsuperscript{17}. We estimated the proportion of variance ($R^2$) in the expression of a gene that was explained by the lead cis-eQTL or trans-eQTL variant. We found that the median $R^2$ value of lead cis-eQTLs was significantly higher than that of lead trans-eQTLs ($R^2$ 4.8% versus 2.8%, $p < 2.2\times10^{-16}$) (Fig. 1B). For the majority of cis-eGenes (85%) and trans-eGenes (96%), the lead eQTL variants only explained a small proportion of variance in expression ($R^2 < 0.2$). The lead e-QTLs explained a large proportion of variance in expression ($R^2 > 0.8$) of an extremely small proportion of eGenes (0.17% of cis-eGenes and five trans-eGenes). We further compared the median values of cis-eQTL $R^2$ versus trans-eQTL $R^2$ between protein-coding genes and lncRNA genes. For cis-eGenes, protein-coding-eGenes had slightly higher median variance.
explained in their expression level than for lncRNA genes (5.1% versus 4.3%, p = 0.0004). However, we did not see this trend for trans-eGenes of protein-coding versus lncRNA genes (2.8% versus 2.7%, p = 0.41) (Fig. 1B).

**Gene Ontology analyses.** We identified 100 significant GO terms for the top 1000 cis-eGenes at FDR < 0.05. Of these GO terms, there were 58 for Biological Process, 31 for Cellular Component, and 11 for Molecular Function (Supplemental Table 7). Of note, the top GO terms appeared to be related to immune functions. For example, the top two Biological Processes are "leukocyte degranulation" (FDR = 1e−6) and "myeloid leukocyte mediated immunity" (FDR = 2e−6) and the top two Cellular Components are cytoplasm (FDR = 3e−6) and MHC protein complex (FDR = 6e−6). The top 1000 trans-eGenes gave rise to 75 significant (FDR < 0.05) GO terms including 37 for Biological Process, 32 for Cellular Component, and 6 for Molecular Function. The top GO terms for the top 1000 trans-eGenes were enriched in pathways and molecular functions related to immune functions (Supplemental Table 7).

**GWAS enrichment analyses.** We linked 1,855,111 cis-eQTL variants (P < 5e−8) to GWAS Catalog variants. At FDR < 0.05, the cis-eQTL variants were enriched with GWAS SNPs associated with 815 traits, representing 28% of the traits in the GWAS Catalog. The top traits identified in enrichment analyses include several cardiovascular disease risk factors. For example, cis-eQTL variants are enriched with BMI-associated SNPs (fold enrichment = 84, FDR = 3.3e−267), total cholesterol (fold enrichment = 98, FDR = 7.3e−162) (Supplemental Table 8). We identified 193 GWAS traits enriched for the trans-eQTL variants (Supplemental Table 9). The top traits in the trans enrichment analysis included neuroticism measurement (fold enrichment = 3, FDR = 1.9e−89) and BMI-adjusted waist circumference (fold enrichment = 2, 6.4e−87).

**Mendelian randomization analysis.** We performed two-sample MR to test for potential causal association of the cis-eGenes with SBP, CAD, and COVID-19 severity. We found 1558 genes containing at least one eQTL variant (median 29; interquartile range [IQR] 6, 88) that coincided with variants from GWAS of SBP (P < 5e−8)18. After Bonferroni correction for multiple testing, MR identified putative causal associations for 60 genes with SBP (i.e., p < 0.05/1558) (Table 4, Supplemental Table 10). Of these 60 genes, six lncRNAs (AC066612.1, AC069200.1, AC092747.4, AC100810.3, AL590226.2, and LY6E-DT) showed putative causal associations with SBP. Six lncRNA genes (5.1% versus 4.3%, p = 0.41) (Fig. 1B).

| Exposure | Chr | Gene type | Outcome | INV MR\(^1\) | Beta | SE  | p       | N SNPs |
|----------|-----|-----------|---------|--------------|------|-----|---------|--------|
| PSRC1    | 1   | Protein coding | CHD | −0.084 | 0.0075 | 4.8E−29 | 7       |
| LTA      | 6   | Protein coding | CHD | −0.069 | 0.011  | 1.3E−09 | 5       |
| MIR6991  | 6   | lncRNA     | CHD    | 1.72    | 0.28   | 2.0E−09 | 25      |
| LIPA     | 10  | Protein coding | CHD  | 0.0033  | 0.00039| 2.9E−17 | 18      |
| PHETA1   | 12  | Protein coding | CHD  | −0.078  | 0.013  | 4.7E−09 | 3       |
| AC3L6    | 5   | Protein coding | COVID-19 | 0.19   | 0.064  | 0.0025^  | 4       |
| DPP9     | 19  | Protein coding | COVID-19 | −0.044 | 0.017  | 0.0078^  | 3       |
| HLA-DRB1 | 6   | Protein coding | COVID-19 | 0.00999 | 0.00018| 1.9E−08  | 35      |
| IFNAR2   | 21  | Protein coding | COVID-19 | −0.023  | 0.0037 | 1.8E−06  | 11      |
| OAS1     | 12  | Protein coding | COVID-19 | −0.0086 | 0.0022 | 1.6E−04  | 1       |
| SLC22A31 | 12  | Protein coding | COVID-19 | 0.32    | 0.11   | 0.0029  | 13      |
| TYK2     | 21  | Protein coding | COVID-19 | 0.011   | 0.0021 | 2.8E−08  | 3       |
| AC006460.2 | 2 | Bidirectional promoter lncRNA | SBP | −5.60 | 0.55 | 2.3E−24 | 3       |
| MAP4     | 3   | Protein coding | SBP | 0.092  | 0.0086 | 4.6E−27  | 4       |
| PHETA1   | 12  | Protein coding | SBP | −0.92  | 0.058  | 1.9E−58  | 3       |
| SLC3A11  | 16  | Protein coding | SBP | −0.82  | 0.066  | 5.3E−35  | 21      |
| ACADVL   | 17  | Protein coding | SBP | −0.035 | 0.0030 | 1.5E−31  | 3       |

Table 4. Top results in Mendelian randomization analyses. \(^1\)Beta/SE and p-value were obtained by inverse variance weighted MR method. \(^2\)Heterogeneity was observed in MR analyses. Sensitivity analyses were performed with median-based and mode-based MR methods in Supplemental Table 9. \(^3\)MR analysis was performed at gene level. At splice variation level (rs10774671), the MR p = 4E−06.

**GWAS enrichment analyses.** We linked 1,855,111 cis-eQTL variants (P < 5e−8) to GWAS Catalog variants. At FDR < 0.05, the cis-eQTL variants were enriched with GWAS SNPs associated with 815 traits, representing 28% of the traits in the GWAS Catalog. The top traits identified in enrichment analyses include several cardiovascular disease risk factors. For example, cis-eQTL variants are enriched with BMI-associated SNPs (fold enrichment = 84, FDR = 3.3e−267), total cholesterol (fold enrichment = 98, FDR = 7.3e−162) (Supplemental Table 8). We identified 193 GWAS traits enriched for the trans-eQTL variants (Supplemental Table 9). The top traits in the trans enrichment analysis included neuroticism measurement (fold enrichment = 3, FDR = 1.9e−89) and BMI-adjusted waist circumference (fold enrichment = 2, 6.4e−87).

**Mendelian randomization analysis.** We performed two-sample MR to test for potential causal association of the cis-eGenes with SBP, CAD, and COVID-19 severity. We found 1558 genes containing at least one eQTL variant (median 29; interquartile range [IQR] 6, 88) that coincided with variants from GWAS of SBP (P < 5e−8)18. After Bonferroni correction for multiple testing, MR identified putative causal associations for 60 genes with SBP (i.e., p < 0.05/1558) (Table 4, Supplemental Table 10). Of these 60 genes, six lncRNAs (AC066612.1, AC069200.1, AC092747.4, AC100810.3, AL590226.2, and LY6E-DT) showed putative causal associations with SBP. Six lncRNA genes (5.1% versus 4.3%, p = 0.41) (Fig. 1B).
Replication analyses. The Battle study only provided \( p \) values for eQTL analyses. Of the reported 10,914 cis-eQTL-eGene pairs from the study by Battle et al.\(^{23} \) (FDR < 0.05)\(^{23} \), 6782 (62%) pairs displayed \( p < 5 \times 10^{-8} \) in the present study. The average proportion of variance explained by these 6782 cis-eQTL variants in respective genes was 0.11 (Supplemental Table 12). Of the 269 trans-eQTL-eGene pairs (FDR < 0.05) reported by Battle et al.\(^{23} \), 47 (18%) pairs displayed \( p < 1 \times 10^{-12} \) in the current study. The average proportion of variance explained by these 47 trans-eQTL variants in respective genes was 0.076. Of note, all 47 trans-eQTL variants and respective trans-eGenes are located on the same chromosomes (Supplemental Table 13). The average distance between these trans-eQTL variants and respective trans-eGenes is within 22 Mb.

We conducted additional replication analysis for the cis-eQTL variant-eGene pairs generated from 8,372,247 SNPs and 20,188 gene transcripts that were common to our study (\( n = 2622 \) participants) and to GTEx(6) (\( n = 755 \) participants). At \( p < 5 \times 10^{-8} \), we identified 1,080,485 cis-eQTL variant-eGene pairs in GTEx and 3,852,182 pairs in our study; of these, 951,085 pairs (88% of pairs in GTEx) displayed the same effect direction as in our larger study. (Supplemental Fig. 3). At \( p < 1 \times 10^{-4} \), we identified 1,815,208 cis-eQTL variant-eGene pairs in GTEx and 6,364,173 pairs in this study; of these, 1,797,977 (99% of pairs in GTEx) displayed the same effect directionality with our study. As can be seen in the figure, there is considerable concordance between the Framingham Heart Study (FHS) and GTEx eQTL effect sizes, although the FHS has a larger sample size than GTEx whole blood samples (2622 vs 755), which results in a smaller standard error and larger t-statistics (Supplemental Fig. 4).

Discussion

We leveraged WGS and RNA-seq data from 2622 FHS participants to create a powerful scientific resource of eQTLs. We identified significant unique cis-eQTL variants-eGene pairs (\( n = 2,855,111 \) unique variants with cis-15,982 eGenes) and 526,056 unique trans-eQTL variants-eGene pairs (526,056 unique variants and unique 7233 trans-eGenes). A large proportion of reported cis-eQTL variant-eGene pairs were replicated with directionally concordant in our study including 88% of cis-variant-eGene pairs from GTEx. Consistent with our previous study and others\(^{7,22,24,25,90} \), 90% of eQTL variants identified in the present study are located within 1 Mb of the corresponding cis-eGene and 83% are within 100 kb of the TSSs of the corresponding eGene. While the majority of (85% of cis- and 96% of trans-) lead eQTL variants explained only a small proportion (\( R^2 < 0.2 \)) of interindividual variation in expression of the corresponding eGenes, 15% of lead cis-eQTL variants and 4% of lead trans variant explained 20% or more of interindividual variation in expression of the corresponding eGenes\(^{26} \). Additionally, eQTL variants were enriched (\( p < 0.0001 \)) in disease-associated SNPs identified by GWAS. We further demonstrated the utility of our eQTL resource for conducting causal inference testing. Our MR analyses revealed putatively causal relations of gene expression to several disease phenotypes including SBP, CAD, and COVID-19 severity. Taken together, the comprehensive eQTL resource we provide can advance understanding of the genetic architecture of gene expression underlying a wide variety of diseases. The interactive and browsable eQTL resource will be posted to the National Heart, Lung, and Blood Institute's BioData Catalyst site and will be freely accessible to the scientific community.

Our study expands current knowledge by creating an accessible and browsable resource of eQTLs based on WGS and RNA-seq technologies. It also includes eQTLs for lncRNAs that were not reported in prior eQTL studies that used array-based expression profiling. Over the past decade, accumulating evidence shows that lncRNAs are widely expressed and have key roles in gene regulation\(^{32,38} \). It is estimated that the human genome contains 16,000 to 100,000 lncRNAs\(^{37} \). We identified 447,598 cis-eQTL variants for 1518 lncRNAs and 121,241 trans-eQTLs for 475 trans-lncRNAs (Supplemental Tables 5 and 6). In addition, we identified six lncRNAs that showed putative causal associations with SBP. However, the functions of these six lncRNAs remain to be determined. Thus, our novel eQTL database may also help in the study of non-protein-coding RNAs in relation to health and disease.

As a proof of concept of the application of the eQTL resource, we performed MR analyses on a small number of cardiovascular traits and COVID-19 severity and demonstrated that the eQTL database can identify promising candidate genes with evidence of putatively causal relations to disease that may merit functional studies. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has spread across the globe and caused millions of deaths since it emerged in 2019. Recent GWAS of COVID-19 susceptibility and severity\(^{29,30} \) have identified SNPs associated with Immunodeficiency and measles virus susceptibility and play an essential and a narrow role in human antiviral immunity\(^{33} \). A recent study further showed that loss-of-function mutations in IFNAR2 are associated with severe COVID-19\(^{9,14} \). These studies, considered alongside our MR results provide evidence of a causal role of IFNAR2 expression in severe COVID-19 infection. This study identified cis-eQTL in whole blood and used them as IVs in MR analyses. Therefore, caution is needed in interpreting the causal relations of several genes to several disease traits. Blood tissue is more easily accessible than other tissues, e.g., kidney or heart, for large association studies. Several previous studies of omics
data have shown that findings in whole blood were comparable to other tissues\textsuperscript{15,16}. Several eQTL studies have also demonstrated \textit{cis}- and \textit{trans}-eQTLs gene regulation across tissues. The GTEx Consortium found that permissive \textit{cis}-eQTLs affect the majority of human genes across tissues\textsuperscript{27}. Another study also showed that the identified cardiometabolic genetic loci share downstream \textit{cis}- and \textit{trans}-gene regulation across tissues and diseases\textsuperscript{40}.

This study has several noteworthy limitations. This study included White participants of European ancestry who were middle-aged and older; therefore, the eQTLs identified may not be generalizable to other races or age ranges. The current RNA-seq platform included ~7700 lncRNAs, which is a modest subset of all lncRNAs in the human genome\textsuperscript{27}. We used MR analyses to infer causal relation of genes to disease traits. MR analysis is predicated on a set of critical assumptions that may not be testable in the setting of eQTL analysis\textsuperscript{39,40}. Replication of our eQTL findings is warranted in studies with larger sample sizes and more diverse populations. In addition, our study found that the \textit{cis}-eQTL variants were enriched for GWAS-associated SNPs. It is possible that we may have underestimated signal regions when local LD in the GWAS region was not considered. An exhaustive investigation of colocalization\textsuperscript{41,42} is not feasible in this study due to computational burden and storage limitations.

Our study also has several strengths. The advent of high-throughput gene sequencing technology provides an unparalleled opportunity to accelerate understanding of the genetic architecture of gene expression. Our study extends and expands the existing literature by identifying novel eQTLs based on WGS and RNA-seq. We demonstrate the potential applications of a vast eQTL resource by analyzing the concordance of eQTL variants with SNPs from GWAS of several disease phenotypes followed by causal inference analyses that identified promising disease-related genes that may merit functional studies. We created an open and freely accessible eQTL repository that can serve as a promising scientific resource to better understand the genetic architecture of gene expression and its relations to a wide variety of diseases.

**Methods**

**Study participants.** This study included participants from the FHS Offspring\textsuperscript{19} and Third Generation cohorts\textsuperscript{11}. Blood samples for RNA seq were collected from Offspring participants who attended the ninth examination cycle (2011–2014) and the Third Generation participants who attended the second examination cycle (2008–2011). Protocols for participant examinations and collection of genetic materials were approved by the Institutional Review Board at Boston Medical Center. All participants provided written, informed consent for genetic studies. All research was performed in accordance with relevant guidelines/regulations.

**Isolation of RNA from whole blood and RNA-seq.** Peripheral whole blood samples (2.5 mL) were collected from FHS participants (Offspring participants at the ninth examination cycle and the Third Generation participants at the second examination cycle) using PAXgene\textsuperscript{TM} tubes (PreAnalytix, Hombrechtikon, Switzerland), incubated at room temperature for 4 h for RNA stabilization, and then stored at ~80 °C until use. Total RNA was isolated using a standard protocol using a PAXgene Blood RNA Kit at the FHS Genetics Laboratory (FHS Third Generation cohort) and the TOPMed contract laboratory at Northwest Genomics Center (Offspring cohort). Tubes were allowed to thaw for 16 h at room temperature. White blood cell pellets were collected after centrifugation and washing. Cell pellets were lysed in guanidinium-containing buffer. The extracted RNA was tested for its quality by determining absorbance readings at 260 and 280 nm using a NanoDrop ND-1000 UV spectrophotometer. The Agilent Bioanalyzer 2100 microfluidic electrophoresis (Nano Assay and the Caliper LabChip system) was used to determine the integrity of total RNA.

All RNA samples were sequenced by an NHLBI TOPMed program\textsuperscript{13} reference laboratory (Northwest Genomics Center) following the TOPMed RNA-seq protocol. All RNA-seq data were processed by University of Washington. The raw reads (in FASTQ files) were aligned using the GRCh38 reference build to generate BAM files. RNA-SeQC\textsuperscript{43} was used for processing of RNA-seq data by the TOPMed RNA-seq pipeline to derive standard quality control metrics from aligned reads. Gene-level expression quantification was provided as read counts and transcripts per million (TPM). GENCODE 30 annotation was used for annotating gene-level expression.

**Whole blood cell counts.** Whole blood cell counts include white blood cell (WBC) count, red blood cell count, platelet count, and WBC differential percentages (neutrophil percent, lymphocyte percent, monocyte percent, eosinophil percent, and basophil percent). Contemporaneously measured blood cell counts were available in 2094 (80%) of the 2622 FHS participants used in eQTL analyses. We performed partial least squares (PLS) prediction method\textsuperscript{44} with three-fold cross-validation (2/3 samples for training and 1/3 for validation) to impute these blood cell components using gene expression from RNA-seq. Prediction accuracy (R-squared) varied across blood component: WBC: 0.58, platelet: 27%, neutrophil percentage: 82%, lymphocyte percentage: 85%, monocyte percentage: 77%, eosinophil percentage: 87%, basophil percentage: 32%. Because 80% of the participants in this study had directly measured cell count variables and only 20% received imputed variables, we used the measured (in 2094 participants) and predicted (in 528 participants) blood cell components as covariates in regression models for eQTL analyses.

**RNA-seq quality control, and data adjustment.** To minimize confounding, expression residuals were generated by regressing transcript expression level on age, sex, measured or predicted blood cell count and differential cell proportions, and genetic principal components. Principal component (PC) analysis is a technique for reducing the dimensionality in large data sets\textsuperscript{45}. It has been widely used in regression analyses to minimize unknown confounding. We included five PCs computed from FHS genotype profiles to account for population stratification. We also included 15 PCs computed from the transcriptome profile to account for unknown confounders that may affect gene expression. In addition, we adjusted for a relatedness matrix, and technical
covariates including year of blood collection, batch (sequencing machine and time, plate and well), and RNA concentration.

**Whole genome sequencing.** Whole genome sequencing of genomic DNA from whole blood was conducted in ~6000 FHS participants as part of NHLBI’s TOPMed program. Standard procedures were used to obtain DNA fragmentation and library construction. Sequencing was performed by a TOPMed reference laboratory (the Broad Institute of MIT and Harvard) using Hi Seq X with sequencing software HiSeq Control Software (HCS) version 3.3.76, then analyzed using RTA2 (Real Time Analysis). The DNA sequence reads were aligned to a human genome build GRCh38 using a common pipeline across all TOPMed WGS centers. A sample's sequence was considered complete when the mean coverage of nDNA was ≥ 30x. This analysis used genetic variants generated from TOPMed Freeze 10.

**Association analyses of expression levels with SNPs.** We performed association analyses of expression levels with genome-wide SNPs with minor allele count (MAC) ≥ 10. In a simple regression model, a SNP was used as an independent variable and the residuals of a transcript expression level was used as the dependent variable. All analyses were performed on the NIH-supported STRIDES cloud infrastructure. A graphical Processing Unit (GPU)-based program was used to facilitate computation. Effect sizes, standard error, partial R-squared, and p-values for all SNP-gene expression pairs with p < 1e–4 were stored to enable lookups and to facilitate later meta-analysis. We evaluated the genomic inflation factor for eQTL analyses. Due to storage burden, we evaluated the genomic inflation factor based on full eQTL analysis (i.e., no p value restriction) on chromosome 12 because the length of this chromosome is close to the median length of chromosome 1–22.

In this study, we defined cis-eQTLs as targeting genes within 1 Mb of their transcription start site (TSS). Trans-eQTLs referred to those that were beyond of 1 Mb of the TSSs of the eGenes on the same chromosome or those on the different chromosomes of the eGenes. A significant cis-eQTL of an eGene was identified if a SNP within 1 Mb of that gene was associated with expression of a transcript of that gene at P < 5 × 10−8. A significant trans-eQTL was defined as a SNP beyond 1 Mb that gave rise to P < 1 × 10−15 in association a gene.

**Estimation of variance in expression level explained by eQTLs.** An accurate estimation of heritability may help understand the degree to which genetic factors influence a trait. Narrow-sense heritability measures the proportion of phenotypic variance explained by additive genetic effects. We estimated the proportion of variance (RV) in the expression level of a gene that was explained by the lead cis-eQTL (cis-RV) or trans-eQTL (trans-RV) variant. We conducted Mood’s median test (median_test in the “coin” R package) to compare the median value of variance in cis-eGenes and trans-eGenes explained by cis-eQTLs versus trans-eQTLs. For cis- or trans-eGenes, we compared the median value of genetic variance in protein-coding eGenes versus lncRNA eGenes.

**Comparison of effect sizes of eQTLs with different minor allele frequencies.** Previous studies showed that rare variants showed a large effect size in QTL analysis. For significant cis-eQTLs (P < 5 × 10−8) and trans-eQTLs (P < 1 × 10−15), we compared the median (25% quartile, 75% quartile) of the absolute values of effect sizes for eQTL variants in four intervals based on their minor allele frequencies (MAFs): [0.01, 0.05), [0.05, 0.1), [0.01, 0.05), and (0.003–0.01).

**Gene Ontology analyses.** We selected the single, most significant eQTL variant (i.e. lead variant) for each eGene (for the gene level analysis) from cis- and trans-eQTL results separately. The eGenes annotated to the selected lead cis and trans eQTL variants were matched into Entrez IDs. We used the “goana” function from the “limma” package to test for over-representation of gene ontology (GO) terms or KEGG pathways applied to the top 1000 eGenes. We used FDR < 0.05 to report GO terms including Biological Process, Cellular Component, and Molecular Function.

**Enrichment analyses using GWAS Catalog.** We linked the eQTL variants with SNPs from the GWAS Catalog (data downloaded on October 22, 2021), which included 243,618 entries for 2960 mapped traits at p < 5e−8. cis- and trans-eQTL variants were analyzed separately. Unique SNP RS IDs were used for enrichment analysis with Fisher's test. FDR < 0.05 was used for significance.

**Correlation analysis of selected IncRNA and protein coding genes.** For IncRNAs that were in the top 25 cis-eQTL variant-eGene pairs, we performed partial Pearson correlation analyses between the expression level of the IncRNA and its nearby protein coding gene, adjusting for the same set of covariates that were included in eQTL analysis. We performed random sampling of 1000 genes 500 times to derive null distributions of partial Pearson correlation of these gene pairs. We calculated an empirical p-value to evaluate whether the partial Pearson correlation coefficient between the expression level of an IncRNA and its nearby protein coding gene was significantly higher than the average partial Pearson correlation coefficient from randomly selected gene pairs. The empirical p-value was calculated as the proportion of partial Pearson correlation coefficients that were more extreme than the correlation coefficient of an IncRNA and its nearby protein coding gene.

**Mendelian randomization analysis.** We conducted Mendelian randomization (MR) to demonstrate the application of the eQTL resource in causal inference analysis. We tested for potential causal association of the cis-eGenes with SBP, coronary artery disease (CAD), and COVID-19 severity. SBP-associated SNPs were obtained
from GWAS of over 1 million people. CAD-associated SNPs were obtained from the study of 34,541 CAD cases and 261,984 controls of UK Biobank resource followed by replication in 88,192 cases and 162,544 controls from CARDIoGRAMplusC4D. COVID-19 associated SNPs were obtained from a recent GWAS including 14,134 COVID-19 cases and 1,284,876 controls of European ancestry by the COVID-19 Host Genetics Initiative. We performed two-sample MR analyses using the TwoSampleMR R package. The instrumental variables (IVs) were independent cis-eQTL variants (LD $r^2 < 0.1$) from this study. The primary analysis used the inverse variance weighted (IVW) method. We also assessed heterogeneity of the IVs in each gene and conducted sensitivity analysis using the MR-Egger method to test for potential horizontal pleiotropy. We also performed the median-based method and mode-based method when heterogeneity was present in MR analyses due to outliers among the IVs. We reported putative causal genes if Bonferroni correction $p < 0.05/n$ (n is the number of genes tested).

**Replication analyses.** A previous study reported 10,914 cis-eQTL variant-eGene pairs and 269 trans pairs (FDR < 0.05) through RNA-sequencing of 922 individuals. We performed replication analyses using the reported cis- and trans-eQTL variant-eGene pairs in conjunction with the pairs in the present study. We also used the cis-eQTL database generated from GTEx whole blood (version 8) for replication of our cis-QTL findings. Whole genome sequencing and RNA-seq were conducted in whole blood of 755 samples in GTEx. The replication was only performed using the cis-eQTL variant-eGene pairs generated by 8,372,247 SNPs and 20,188 gene transcripts that were found in common between our study and GTEx. Because this study was aimed to provide eQTL resource for the broad scientific community, we present replication results using both $p < 5e^{-8}$ and $p < 1e^{-4}$ for replicating cis-eQTL variant-eGene pairs.

**Data availability**

The datasets analyzed in the present study are available at the dbGAP repository phs000007.v32.p1. The datasets analyzed in the present study are available at the dbGAP repository phs000007.v32.p1 (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000007.v30.p11).

Received: 26 April 2022; Accepted: 17 November 2022

Published online: 23 November 2022

**References**

1. Visscher, P. M. et al. 10 Years of GWAS discovery: Biology, function, and translation. Am. J. Hum. Genet. 101, 5–22. https://doi.org/10.1016/j.ajhg.2017.06.005 (2017).
2. MacArthur, J. et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 45, D896–D901. https://doi.org/10.1093/nar/gkw1133 (2017).
3. Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74. https://doi.org/10.1038/nature11247 (2012).
4. Eicher, J. D. et al. GRASP v2.0: An update on the Genome-Wide Repository of Associations between SNPs and phenotypes. Nucleic Acids Res. 43, D799–804. https://doi.org/10.1093/nar/gku1202 (2015).
5. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: From polygenic to omicnic. Cell 169, 1177–1186. https://doi.org/10.1016/j.cell.2017.05.038 (2017).
6. The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science 369, 1318–1330. https://doi.org/10.1126/science.aaz1776 (2020).
7. Ferrmann, R. S. et al. Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. PLoS Genet. 7, e1002197. https://doi.org/10.1371/journal.pgen.1002197 (2011).
8. Westra, H. J. et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat. Genet. 45, 1238–1243. https://doi.org/10.1038/ng.2756 (2013).
9. Liang, L. et al. A cross-platform analysis of 14,177 expression quantitative trait loci derived from lymphoblastoid cell lines. Genome Res. 23, 716–726. https://doi.org/10.1101/gr.142521.112 (2013).
10. Battle, A. & Montgomery, S. B. Determining causality and consequence of expression quantitative trait loci. Hum. Genom. 133, 727–735. https://doi.org/10.1007/s10439-014-1446-0 (2014).
11. Kirsten, H. et al. Dissecting the genetics of the human transcriptome identifies novel trait-related trans-eQTLs and corroborates the regulatory relevance of non-protein coding loci. Hum. Mol. Genet. 24, 4746–4763. https://doi.org/10.1093/hmg/ddv194 (2015).
12. Johanes, R. et al. Integrated genome-wide analysis of expression quantitative trait loci aids interpretation of genomic association studies. Genome Biol. 18, 16. https://doi.org/10.1186/s13059-016-1142-6 (2017).
13. Taltavull, D. et al. Sequencing of 53,831 diverse genomes from the NHGRI TOPMed Program. Nature 590, 290–299. https://doi.org/10.1038/s41586-021-03205-y (2021).
14. Miretti, M. M. et al. A high-resolution linkage-disequilibrium map of the human major histocompatibility complex and first generation of tag single-nucleotide polymorphisms. Am. J. Hum. Genet. 76, 634–646. https://doi.org/10.1086/429393 (2005).
15. Yao, C. et al. Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. Nat Commun 9, 3268. https://doi.org/10.1038/s41467-018-05512-x (2018).
16. Wright, M. W. A short guide to long non-coding RNA gene nomenclature. Hum. Genom. 8, 7. https://doi.org/10.1186/s41479-7364-8-7 (2014).
17. Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era – concepts and misconceptions. Nat. Rev. Genet. 9, 255–266. https://doi.org/10.1038/nrg2322 (2008).
18. Evangelou, E. et al. Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. Nat. Genet. 50, 1412–1425. https://doi.org/10.1038/s41588-018-0205-x (2018).
19. van der Harst, P. & Verweij, N. Identification of 64 novel genetic loci provides an expanded view on the genetic architecture of coronary artery disease. Circ. Res. 122, 433–443. https://doi.org/10.1161/CIRCRESAHA.117.312086 (2018).
20. Kousathanas, A. et al. Whole genome sequencing reveals host factors underlying critical Covid-19. Nature https://doi.org/10.1038/s41586-022-04576-6 (2022).
21. Zhou, S. et al. A Neanderthal OAS1 isoform protects individuals of European ancestry against COVID-19 susceptibility and severity. Nat. Med. 27, 659–667. https://doi.org/10.1038/s41591-021-01281-1 (2021).
22. Bonnevie-Nielsen, V. et al. Variation in antiviral 2',5'-oligoadenylate synthetase (2'5'AS) enzyme activity is controlled by a single-nucleotide polymorphism at a splice-acceptor site in the OAS1 gene. Am. J. Hum. Genet. 76, 623–633. https://doi.org/10.1086/429391 (2005).

23. Battle, A. et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res. 24, 14–24. https://doi.org/10.1101/gr.153192.113 (2014).

24. Morley, M. et al. Genetic analysis of genome-wide variation in human gene expression. Nature 430, 743–747. https://doi.org/10.1038/nature02797 (2004).

25. Li, Q. et al. Expression QTL-based analyses reveal candidate causal genes and loci across five tumor types. Hum. Mol. Genet. 23, 5294–5302. https://doi.org/10.1093/hmg/ddu228 (2014).

26. Huan, T. et al. A systematic heritability analysis of the human whole blood transcriptome. Hum. Genet. 134, 343–358. https://doi.org/10.1007/s00439-014-1524-3 (2015).

27. Starello, L., Guo, C. J., Chen, L. L. & Huarte, M. Gene regulation by long non-coding RNAs and its biological functions. Nat. Rev. Mol. Cell Biol. 22, 96–118. https://doi.org/10.1038/s41580-020-00315-9 (2021).

28. He, R. Z., Luo, D. X. & Mo, Y. Y. Emerging roles of IncRNAs in the post-transcriptional regulation in cancer. Genes Dis. 6, 6–15. https://doi.org/10.1016/j.gendis.2019.01.003 (2019).

29. Initiative, C.-H.G. The COVID-19 Host Genetics Initiative, a global initiative to elucidate the role of host genetic factors in susceptibility and severity of the SARS-CoV-2 virus pandemic. Eur. J. Hum. Genet. 28, 715–718. https://doi.org/10.1038/s41431-020-0636-6 (2020).

30. Ellinghaus, D. et al. The Severe Covid-19 GWAS Group. Genomewide association study of severe Covid-19 with respiratory failure. N. Engl. J. Med. 383, 1522–1534. https://doi.org/10.1056/NEJMoa2002833 (2020).

31. Li, Y. et al. Genome-wide association study of COVID-19 severity among the Chinese population. Cell Discov. 7, 76. https://doi.org/10.1038/s41421-021-00316-6 (2021).

32. Kotelov, S. V. et al. Hypotheses and facts for genetic factors related to severe COVID-19. World J. Virol. 10, 137–155. https://doi.org/10.5501/wjv.t.14.137 (2021).

33. Duncan, C. J. et al. Human IFNAR2 deficiency: Lessons for antiviral immunity. Sci. Transl. Med. 7, 307ra154. https://doi.org/10.1126/scitranslmed.aac4227 (2015).

34. Smieszek, S. P., Polymenopoulos, V. M., Xiao, C., Polymenopoulos, C. M. & Polymenopoulos, M. H. Loss-of-function mutations in IFNAR2 in COVID-19 severe infection susceptibility. J. Glob. Antimicrob. Resist. 26, 239–240. https://doi.org/10.1016/j.jgar.2021.06.005 (2021).

35. Hillman, S. L. et al. Novel DNA methylation profiles associated with key gene regulation and transcription pathways in blood and placenta of growth-restricted neonates. Epigenetics 10, 50–61. https://doi.org/10.4161/15592294.2014.989741 (2015).

36. Yang, S. Y. et al. Blood-derived mitochondrial DNA copy number is associated with gene expression across multiple tissues and is predictive for incident neurodegenerative disease. Genome Res. 31, 349–358. https://doi.org/10.1101/gr.269381.120 (2021).

37. Consortium, G. et al. Genetic effects on gene expression across human tissues. Nature 550, 204–213. https://doi.org/10.1038/nature24277 (2017).

38. Franzen, O. et al. Cardiometabolic risk loci share downstream cis- and trans-genome regulation across tissues and diseases. Science 353, 827–830. https://doi.org/10.1126/science.aad6970 (2016).

39. Davies, N. M., Holmes, M. V. & Davey Smith, G. Reading Mendelian randomisation studies: A guide, glossary, and checklist for clinicians. BMJ 362, k601. https://doi.org/10.1093/bmj/k601 (2018).

40. Hartwig, F. P., Davies, N. M., Hemani, G. & Davey Smith, G. Two-sample Mendelian randomization: Avoiding the downsides of a powerful, widely applicable but potentially fallible technique. Int. J. Epidemiol. 45, 1717–1726. https://doi.org/10.1093/ije/dyx028 (2016).

41. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. 10, e1004383. https://doi.org/10.1371/journal.pgen.1004383 (2014).

42. Wallace, C. Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. PLoS Genet. 16, e1008720. https://doi.org/10.1371/journal.pgen.1008720 (2020).

43. Deluca, D. S. et al. RNA-SeqQC: RNA-seq metrics for quality control and process optimization. Bioinformatics 28, 1530–1532. https://doi.org/10.1093/bioinformatics/bts196 (2012).

44. Nguyen, D. V. Partial least squares dimension reduction for microarray gene expression data with a censored response. Math. Biosci. 193, 119–137. https://doi.org/10.1016/j.mbs.2004.10.007 (2005).

45. Jolliffe, I. T. & Cadima, J. Principal component analysis: A review and recent developments. Philos. Trans. A 374, 20150202. https://doi.org/10.1098/rsta.2015.0202 (2016).

46. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 43, e47. https://doi.org/10.1093/nar/gkv007 (2015).

47. Hemani, G. et al. The MR-Base platform supports systematic causal inference across the human phenome. Elife https://doi.org/10.7554/eLife.34408 (2018).

48. Bowden, J., Davey Smith, G., Haycock, P. C. & Burgess, S. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. Genet. Epidemiol. 40, 304–314. https://doi.org/10.1002/gepi.21965 (2016).

49. Hartwig, F. P., Davey Smith, G. & Bowden, J. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. Int. J. Epidemiol. 46, 1985–1998. https://doi.org/10.1093/ije/dyx102 (2017).

50. Burgess, S. et al. Guidelines for performing Mendelian randomization investigations. Wellcome Open Res. 4, 186. https://doi.org/10.12688/wellcomeopenres.15555.2 (2019).

Disclaimer
The views and opinions 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 National Institutes of Health, or the U.S. Department of Health and Human Services.

Author contributions
C.L. wrote the main manuscript text; R.J., C.L., X.S., and P.J.M. performed statistics analyses and prepared the tables and figures; Y.W., G.M.P., and A.N.P. cleaned RNASeq data; N.L.H.-C., and A.N.P. coordinate data acquisition; D.L., R.J., J.M., A.K., M.S., T.H., S-J.H., H.B., B.T., P.J.M., and R.S.V. reviewed/edited the manuscript; D.C., M.F., and N.S. performed data sharing; R.S.V. and D.L. provided funding for whole genome sequencing and RNASeq.

Funding
The Framingham Heart Study was supported by NIH contracts N01-HC-25195, HHSN26820150001I, and 75N92019D00031. DNA methylation assays were supported in part by the Division of Intramural Research.
(D. Levy, Principal Investigator) and an NIH Director’s Challenge Award (D. Levy, Principal Investigator). The analytical component of this project was funded by the NHLBI Division of Intramural Research (D. Levy, Principal Investigator). Whole genome sequencing for the TransOomics in Precision Medicine (TOPMed) program was supported by the NHLBI. Core support including centralized genomic read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Core support including phenotype harmonization, data management, sample identity QC, and general program coordination were provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I). J. Ma is supported by NIH grants, K22HL135075 and R01AA028263.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-24611-w.

Correspondence and requests for materials should be addressed to C.L., R.J. or D.L.

Reprints and permissions information is available at www.nature.com/reprints.

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

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

© The Author(s) 2022