

Integrating whole genome and transcriptome sequencing to characterize the genetic architecture of isoform variation

Received: 26 November 2024

Accepted: 15 September 2025

Cite this article as: Liu, C., Joehanes, R., Ma, J. *et al.* Integrating whole genome and transcriptome sequencing to characterize the genetic architecture of isoform variation. *Nat Commun* (2025). <https://doi.org/10.1038/s41467-025-64336-8>

Chunyu Liu, Roby Joehanes, Jiantao Ma, Jiuyong Xie, Jian Yang, Mengyao Wang, Tianxiao Huan, Shih-Jen Hwang, Jia Wen, Quan Sun, Cumhur Y. Demirkale, Nancy L. Heard-Costa, Peter Orchard, April P. Carson, Jeffrey W. Haessler, Laura M. Raffield, Alex P. Reiner, Nora Franceschini, Paul L. Auer, Charles Kooperberg, Yun Li, George O'Connor, Joanne M. Murabito, Peter Munson & Daniel Levy

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

Integrating Whole Genome and Transcriptome Sequencing to Characterize the Genetic Architecture of Isoform Variation

Chunyu Liu^{1,2,*,#}, Roby Joehanes^{3,*,#}, Jiantao Ma^{4,*}, Jiuyong Xie⁵, Jian Yang¹, Mengyao Wang¹, Tianxiao Huan³, Shih-Jen Hwang³, Jia Wen⁶, Quan Sun^{7,8,9}, Cumhuri Y. Demirkale¹⁰, Nancy L. Heard-Costa^{2,11}, Peter Orchard¹², April P. Carson¹³, Jeffrey W. Haessler¹⁴, Laura M. Raffield⁶, Alex P. Reiner^{14,15}, Nora Franceschini^{6,16}, Paul L. Auer¹⁷, Charles Kooperberg¹⁴, Yun Li⁷, George O'Connor^{2,18}, Joanne M. Murabito^{2,19}, Peter Munson³, Daniel Levy^{2,3,#}

¹Department of Biostatistics, School of Public Health, Boston University, Boston, MA 02118, USA

²Framingham Heart Study, Framingham, MA 01702, USA

³Population Sciences Branch, Division of Intramural Research, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD 20892, USA

⁴Nutrition Epidemiology and Data Science, Friedman School of Nutrition Science and Policy, Tufts University, Boston, MA 02111, USA

⁵Physiology and Pathophysiology, University of Manitoba, Winnipeg, Canada

⁶Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27514, USA

⁷Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599, USA

⁸Center for Computational and Genomic Medicine, Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA

⁹Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, PA, 19104, USA

¹⁰Critical Care Medicine Department, Clinical Center, National Institutes of Health, Bethesda, MD, USA

¹¹Departments of Neurology, Boston University Schools of Medicine, Boston, MA 02118, USA

¹²Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA

¹³Department of Medicine, University of Mississippi Medical Center, Jackson, MS 39216, USA.

¹⁴Division of Public Health Sciences, Fred Hutchinson Cancer Center, Seattle, WA 98109, USA.

¹⁵Department of Epidemiology, University of Washington, Seattle, WA 98195, USA.

¹⁶Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, NC.

¹⁷Division of Biostatistics, Data Science Institute and Cancer Center, Medical College of Wisconsin, Milwaukee, WI, USA

¹⁸Pulmonary Center, Boston University School of Medicine, Boston, MA, USA

¹⁹Section of General Internal Medicine, Department of Medicine, Boston University Chobanian and Avedisin School of Medicine, Boston Medical Center, Boston, MA, USA.

*These authors contributed equally

#Corresponding authors:

Chunyu Liu, PhD: liuc@bu.edu

Roby Joehanes, PhD: roby.joehanes@nih.gov

Daniel Levy, MD: levyd@nih.gov

ABSTRACT

We present a whole-blood isoform ratio QTL (irQTL) resource by analyzing genome-wide isoform-to-gene expression ratios using sequencing data. In Framingham Heart Study (FHS, $n = 2,622$) discovery, we identify over 1.1 million *cis*-irQTLs (minor allele frequency [MAF] ≥ 0.01 , ± 1 Mb of 10,883 isoform transcripts, $P < 5 \times 10^{-8}$) across 4,971 genes. Among 11,425 sentinel *cis*-irQTLs, 72% replicate ($P < 1 \times 10^{-4}$) in the Women's Health Initiative (WHI; $n = 2,005$). Notably, 20% of *cis*-irQTLs have no significant association with overall gene expression, indicating isoform-specific regulation. These variants are enriched at splice donor/acceptor sites and genome-wide association study loci ($P < 1 \times 10^{-10}$). We also identify 1,870 sentinel *trans*-irQTLs (MAF ≥ 0.01 , $P < 1.5 \times 10^{-13}$) for 1,084 isoforms across 590 genes, and 2,327 rare *cis*-irQTLs ($0.003 < \text{MAF} < 0.01$) for 2,467 isoforms of 1,428 genes in FHS, with external replication rates of 61% and 41% in WHI, respectively. We highlight rs12898397 in *ULK3*, which alters splice site usage and reduces expression of a full-length isoform. Mendelian randomization supports a causal role between this isoform shift and reduced diastolic blood pressure. These findings highlight the power of irQTL mapping to uncover transcript-specific regulatory mechanisms underlying complex traits.

KEY WORDS, irQTL analysis, RNA alternative splicing, isoform, GWAS

INTRODUCTION

Isoform variation results from several mechanisms including alternative splicing, alternative transcription start sites, and alternative polyadenylation (polyA) sites.¹⁻³ These mechanisms generate distinct messenger RNAs (mRNAs) from the same gene, expanding proteomic diversity beyond the ~20,000 protein-coding genes in the human genome.⁴ Alternative splicing, which joins exons in different combinations to produce mature RNA isoforms,⁵ can be both a cause and consequence of disease.^{6,7} Recent advances in sequencing technologies have revealed that up to 95% of multi-exon genes undergo alternative splicing in humans.⁴

While it is a dynamic process, isoform variation is under tight regulation. Many factors, including *cis*- and *trans*-regulatory elements and DNA methylation patterns regulate isoform variation.⁸⁻¹⁰ Several previous studies have provided crucial insights into the genetic regulation of transcriptomic isoform variation, contributing to our understanding of the genetic architecture underlying complex diseases.¹¹⁻²¹ Two studies from the Genotype-Tissue Expression (GTEx) consortium used whole-genome (WGS) paired with whole-transcriptome sequencing (RNA-seq) to identify single nucleotide polymorphisms (SNPs) affecting splice variation throughout the human genome across multiple tissues.^{17,22} These studies established the widespread influence of genetic variation on transcriptional regulation, including splice variation.²² In addition to SNPs that influence the overall gene expression levels alone (i.e., expression quantitative trait loci or eQTLs), SNPs that influence splicing (with or without effects on overall gene expression) may exert a pronounced impact on complex traits.¹⁷ The replication rate for splice variant-associated SNPs identified in these two studies, however, was low, possibly reflecting methodological differences in isoform characterization and sample size limitations.^{17,22}

Comprehensive analyses of SNP associations with isoform variation in whole blood and their relevance to human health remain limited. In prior work, we identified exon-level expression variants in 2,650 genes using imputed SNPs and hybridization-based exon array data from 5,257 Framingham Heart Study (FHS) participants.²¹ Here, we present a new study leveraging WGS and RNA-seq data from 3,716 FHS participants (2,622 for discovery, 1,094 for internal replication) to map isoform ratio QTLs (irQTLs) as described originally by Yamaguchi et al. (2022)¹⁸ (Supplementary Data 1). We further evaluate these findings in two external cohorts with WGS and RNA-seq data.^{23,24} Beyond identifying irQTLs, we distinguish those with little or no impact on gene-level expression, assess overlap with genome-wide association study (GWAS) loci,^{25,26} and use Mendelian randomization (MR)²⁷ to evaluate potential causal effects on cardiovascular disease (CVD) risk factors (Figure 1). We highlight three irQTL-transcript pairs to illustrate how this resource can reveal candidate causal variants influencing disease through transcriptomic regulation.

RESULTS

We analyzed 12,887,093 SNPs (Hardy-Weinberg equilibrium, HWE $P \geq 1 \times 10^{-10}$) and isoform ratios of 25,642 transcripts (Trimmed Mean of M-values [TMM] ≥ 2 and $< 20\%$ missing values) from 10,150 genes having more than one transcript in the FHS discovery sample ($n = 2,622$) (see Methods) to identify irQTLs, which capture how genetic variants influence the relative abundance of different transcript isoforms from the same gene¹⁸ (see Methods). Replication was performed internally in an additional FHS sample ($n = 1,094$) and externally in two cohorts: WHI²⁴ ($n = 2,005$), using WGS and whole blood RNA-seq data, and the JHS²³ ($n = 1,010$), using WGS

and RNA-seq from peripheral blood mononuclear cells (PBMCs) (Table 1, Supplementary Data 2).

Discovery of transcriptome-wide common *cis*-irQTLs

A *cis*-irQTL was defined as a variant within 1 Mb of the isoform's transcription start site (TSS). We identified over 1.1 million common *cis*-irQTLs (minor allele frequency [MAF] ≥ 0.01) associated with 10,883 isoform transcripts, representing more than 3.5 million significant *cis*-irQTL-isoform pairs ($P < 5 \times 10^{-8}$) across 4,971 genes. Of these, we prioritized 11,425 sentinel variants forming 14,056 lead pairs with the same isoforms (Supplementary Data 3). A *cis*-irQTL is defined as a variant located within 1 Mb of an isoform's transcription start site (TSS), and a sentinel *cis*-irQTL is the variant most strongly (i.e., with the lowest p-value) associated with an isoform. An isoform may be associated with multiple sentinel *cis*-irQTLs if they are in mutual perfect linkage disequilibrium (LD $r^2 = 1$). The top *cis*-irQTL-isoform pairs are listed in Table 2. The most significant sentinel *cis*-irQTL, rs4841, was positively associated with the ratio of isoform, ENST00000519690.1, of *RPS14* and explained 96% of variance (R^2) of the isoform ratio in the discovery sample. *RPS14* encodes ribosomal protein 14, which plays a crucial role in ribosome biogenesis and protein synthesis as well as cell proliferation and nuclear stress.^{28,29}

The irQTL method, introduced by Yamaguchi et al. (2022),¹⁸ may identify QTLs that affect splicing, as well as QTLs associated with changes in isoform ratios resulting from other regulatory mechanisms. To better interpret the molecular consequences of each variant, we integrated multiple annotation layers from the following databases: GENCODE,³⁰ SpliceAI,³¹ SnpEff,³² and Homer.³³ Each database offers distinct strengths and limitations, capturing different aspects of variant annotation. SnpEff³² is a variant annotation tool that primarily uses Ensembl-

based gene models, but it also supports RefSeq, GENCODE, UCSC Genome Browser annotations. According to SnpEff, 279 variants are related to splicing (e.g., as splice acceptor, splice donor, or splice region variants). SpliceAI,³¹ a tool for predicting splice site alterations from genetic variants, predicted that 215 variants had SpliceAI prediction scores >0.2, suggesting potential splice-altering effects, and 82 of these had SpliceAI prediction scores >0.5, indicating a higher likelihood of splicing disruption. A total of 86 variants were identified by both SnpEff and SpliceAI as potentially splice-altering (Supplementary Data 3).

Discovery of transcriptome-wide common *trans*-irQTLs

A *trans*-irQTL was defined as being located more than 1 Mb away from the TSS of an isoform or on a separate chromosome. Using a significance threshold of $P < 1.5 \times 10^{-13}$, we identified over 130,000 *trans*-irQTLs associated with 1,084 isoforms from 590 genes. For each isoform, we selected the most significant *trans*-irQTL per chromosomes (22 autosomes and X chromosome), yielding 1,870 sentinel *trans*-irQTLs in the discovery sample (Supplementary Data 4). Notably, the largest fraction of these pairs (31%) involved transcripts from genes on chromosome 6 that were significantly associated with irQTLs across the genome, with more than half of these genes located in the HLA (Human Leukocyte Antigen)³⁴ region. Additionally, over 75% of these pairs involved irQTLs and gene transcripts located on different chromosomes (Supplementary Figure 1).

The most significant of these inter-chromosomal pairs was rs1458255–ENST00000418824, with the irQTL on chromosome 4 explaining 54% of variance of the isoform ratio of *RPL9P9* (ribosomal protein L9 pseudogene 9), located on chromosome 15. *RPL9P9* is a non-functional copy of the *RPL9* gene, and the latter encodes a ribosomal protein that is a component of the 60S ribosomal subunit.³⁵

Five of the sentinel variants were annotated as splicing-related by SnpEff, and one had a SpliceAI score greater than 0.2 (rs55934523 C>G variant in *FBRSL1* with a score of 0.54 for acceptor gain (Supplementary Data 4).

Discovery of transcriptome-wide rare *cis*-irQTLs

We assessed *cis*-irQTL variants ($0.003 < \text{MAF} < 0.01$) in the discovery sample. This analysis identified 2,327 rare *cis*-irQTL variants associated with 2,467 isoforms ($P < 5 \times 10^{-8}$), yielding 3,102 significant *cis*-irQTL-isoform pairs across 1,428 genes (Supplementary Data 5). Among these genes, 6% had isoforms that were not associated with common variants. Of note, some rare variants explained a larger R^2 of the corresponding isoform ratios than any common *cis*-irQTL variant for the same gene. For example, the rare variant rs1128271 ($\text{MAF} = 0.0093$) explained approximately 38% of the variance in isoform ratios of ENST00000301788.12 in *POLR2G* --much more than the strongest *cis*-irQTL, rs144760211 ($\text{MAF} = 0.03$), which explained just 9% of the variance in the isoform ratio (Supplementary Data 3 & 5). Twenty-eight sentinel variants were annotated as splicing-related by SnpEff, and seven of them had a SpliceAI score greater than 0.2 (Supplementary Data 5).

Replication of common *cis*-irQTLs

An irQTL-isoform pair was considered replicated if it was found in the replication sample with a $P < 1 \times 10^{-4}$ and showed the same direction of association as in discovery. The replication rate was defined as the proportion of discovered irQTL-isoform pairs present in the replication samples that met the predefined replication criteria.

Among the 14,056 discovered sentinel *cis*-irQTL-isoform pairs ($P < 5 \times 10^{-8}$), 10,783 (76.7%) were internally replicated in an additional 1,094 FHS samples, and 10,174 (72.4%) were externally replicated in the WHI (Figure 2, Supplementary Data

6). R^2 represents the proportion of variance in an isoform ratio explained by its corresponding irQTL. Replication rates increased with higher R^2 values, as demonstrated in Figure 2A. For example, among irQTLs in the second decile, replication rates were approximately 46% in FHS and 44% in WHI, whereas in the eighth decile, these rates increased to 98% and 93%, respectively (Figure 2A). We also compared the effect sizes and directions of association for the 10,194 *cis*-irQTL-isoform pairs present in both the discovery and WHI replication samples. Approximately 99.8% of these pairs exhibited the same direction of association and highly concordant effect sizes (Pearson coefficient $r = 0.91$) (Figure 2B, Supplementary Data 3).

The overall replication rate of *cis*-irQTLs with $MAF \geq 0.01$ in JHS ($n = 1,010$) was 32.4%, substantially lower than in WHI, though replication improved with increasing *cis*-irQTL R^2 values (Supplementary Data 3 & 6, Supplementary Figure 2). JHS isolated RNA from PBMCs,²³ whereas FHS and WHI used whole blood-derived RNA, limiting direct comparability (Supplementary Data 7). To further evaluate the effects of RNA source, ancestry, and sample size, separate analyses were conducted in African American ($n = 918$) and White American ($n = 1,000$) participants in WHI. Replication rates were lower in the White American sub-sample (60.2%), reflecting its smaller sample size, and even lower in the similarly sized African American sub-sample (49.0%), reflecting the potential influence of ancestry. This analysis suggests that the reduction in replication rate observed in the JHS cohort (32.4%) may largely reflect the use of a different RNA source.

A prior splicing QTL study in FHS identified exon-level QTLs in 2,650 genes using imputed SNPs and exon array data.²¹ That study also reported the most significant splicing QTLs within 50 kb of each gene. Only 458 (17.3%) of gene-SNP

pairs from the earlier study showed $P < 1 \times 10^{-4}$ in our discovery sample (Supplementary Data 8). This modest replication rate likely reflects differences in technology (imputed SNPs vs. WGS; exon arrays vs. RNA-seq) and methodology (SNPs associated with exon expression vs. isoform ratio changes) between the earlier and current studies.

Replication of common *trans*-irQTLs

Using the same replication criteria described earlier, the average replication rates of the 2,999 sentinel *trans*-irQTL-isoform pairs were 80.8% in FHS and 60.8% in WHI. Similarly, replication rates increased with R^2 values of the *trans*-irQTLs (Supplementary Data 4 & 9, Supplementary Figure 3). The replication rate of *trans*-irQTL-isoform pairs was again substantially lower in JHS (average rate = 20.2%) compared to WHI.

Replication of rare *cis*-irQTLs

Using the same replication criteria, replication rates among the 3,102 significant *cis*-irQTL-isoform pairs involving rare variants increased with higher R^2 values of the *cis*-irQTLs (Supplementary Data 5 & 10, Supplementary Figure 4). However, overall replication rates of rare variant *cis*-irQTL-isoform pairs remained substantially lower than those observed for common variants, only 22.2% in the additional 1,094 FHS sample and 40.7% in WHI. The lower replication rate in FHS likely reflects its smaller sample size ($n = 1,094$) compared to WHI ($n = 2,005$). JHS ($n = 1,010$) showed a markedly lower replication rate of just 9.4%.

Comparison of common *cis*-irQTLs and *cis*-eQTLs

We previously reported 6,778,286 significant *cis*-eQTL-gene pairs from 2,855,111 *cis*-eQTLs ($MAF \geq 0.01$) associated with 15,982 genes (at $P < 5 \times 10^{-8}$), using the same sample as in the current irQTL discovery analysis ($n = 2,622$).³⁶ In that study,

an expressed gene was defined as a gene with one or more significant eQTLs. Based on *cis*-eQTL and *cis*-irQTL analyses, about 80% ($n = 3,997$ of $4,971$) of genes with irQTLs also had eQTLs (Figure 3A), indicating a strong overlap between splicing regulation (irQTLs) and gene expression regulation (eQTLs). In the remaining 20% of genes significantly associated with irQTLs but not eQTLs, isoform usage changes occurred independently of total expression levels, suggesting that alternative transcript usage can play a key regulatory role without significantly affecting overall gene expression.

We analyzed 4,974 genes harboring irQTLs and their sentinel *cis*-irQTLs ($MAF \geq 0.01$) by comparing the proportion of R^2 for isoform ratios to that for overall gene expression (i.e., as *cis*-eQTLs). Based on this comparison, the SNPs were categorized into three groups: 1,946 with stronger irQTL effects (higher R^2 for isoform ratio), 2,716 with comparable irQTL and eQTL effects, and the rest with stronger eQTL effects (higher R^2 for overall gene expression). For example, rs10774671 G>A in *OAS1* was a strong *cis*-irQTL ($R^2 = 0.95$) for the isoform ratio of ENST00000202917.10, but a weak eQTL for overall *OAS1* expression ($R^2 = 0.05$) (Figure 3B).

We compared the predicted functions of sentinel *cis*-eQTLs and *cis*-irQTLs based on their physical locations (Supplementary Data 11). Compared to eQTLs, irQTLs were more enriched in splice donor sites (median fold enrichment = 25.2 vs. 9.2; two-sided t-test, $P = 4.3 \times 10^{-56}$) and splice acceptor sites (median fold enrichment = 12.6 vs. 5.9, $P = 8.8 \times 10^{-127}$) (Figure 3C). In contrast, eQTLs were more enriched in the upstream region of a gene (within 5 kb) (median fold enrichment=1.31 vs. 0.84, $P < 1 \times 10^{-300}$) or in the 5' -untranslated region (UTR, median fold enrichment=3.0 vs. 2.5, $P < 2.7 \times 10^{-188}$) (Supplementary Data 11).

Additionally, top irQTLs showed greater enrichment in RNA-binding protein (RBP) enhanced crosslinking and immunoprecipitation (eCLIP) peaks than top eQTLs (mean fold-enrichment 3.6 vs. 2.0, $P < 2 \times 10^{-16}$, see Methods) (Figure 3D).

Enrichment of *cis*-irQTLs in GWAS Catalog SNPs

We linked *cis*-irQTLs to GWAS Catalog SNPs to investigate whether *cis*-irQTLs are associated with diseases or traits.²⁶ Common *cis*-irQTL variants exhibited enrichment in GWAS Catalog variants that were associated with hundreds of traits, such as blood protein measurement (fold enrichment = 7.0, $P < 1 \times 10^{-300}$), BMI-adjusted waist-hip ratio (fold enrichment = 6.4, $P < 1 \times 10^{-300}$), and high-density lipoprotein cholesterol level (fold enrichment = 4.6, $P < 1 \times 10^{-300}$) (Supplementary Data 12). These findings were consistent with those from our prior *cis*-eQTL enrichment analysis,³⁶ reflecting a sizable proportion of overlap of *cis*-irQTLs and *cis*-eQTLs. In addition, we identified 197 GWAS traits enriched for *trans*-irQTL variants, including traits such as BMI-adjusted waist-hip ratio (fold enrichment = 16.8, $P < 1 \times 10^{-300}$) (Supplementary Data 13).

To demonstrate the utility of this irQTL resource in understanding the role of isoform regulation, particularly splicing, in relation to complex traits and diseases, we highlight three representative examples featuring top irQTLs from our analysis.

OAS1 harbors a sentinel irQTL linked to isoform regulation

The rs10774671 G>A variant was the 2nd most significant lead *cis*-irQTL in the entire analysis and is associated with the expression ratio of transcript ENST00000202917.10 (*OAS1*-201) in *OAS1* at 12q24.13 (Figure 4A and Table 2). *OAS1* produces seven detectable isoform transcripts in FHS (Supplementary Data 14). *OAS1* has long been known to play a role in the response to viral infection and diabetes risk, and has recently been implicated in COVID-19 severity.³⁷⁻⁴¹ The

rs10774671-A allele disrupts the conservative dinucleotide “AG” of the splice acceptor site of intron 5, changing it to “AA” (Figures 4B & 4C). This resulted in a lower isoform ratio of *OAS1*-201 ($R^2 = 0.95$, $\beta = -0.26$, $P = 1 \times 10^{-1713}$) and higher isoform ratios of the other isoforms including *OAS1*-203 ($\beta = 0.11$, $P = 1 \times 10^{-1359}$) (Figure 4D).

According to SpliceAI,³¹ the rs10774671-A allele has a high probability score of 0.89 for splice acceptor loss in intron 5. To directly visualize the effects of this splice acceptor loss, we analyzed RNA-seq data from a randomly selected participant for each of the three rs10774671 genotypes: G/G, A/G, and A/A. The participant with the G/G genotype (i.e., carrying the canonical 'AG' acceptor site) exhibited complete exclusion of intron 5 in the mature mRNA isoform. Conversely, the participant with the A/A genotype (i.e., with complete loss of the acceptor site) showed full retention of intron 5 in mRNA. In contrast, the participant with the A/G genotype displayed partial retention of intron 5 (Figure 4C).

In addition to its well-known role in impacting COVID-19 severity, the rs10774671-A variant was reported to be significantly associated with lower WBC count ($\beta = -0.013$, $P = 3.2 \times 10^{-11}$) in a large GWAS⁴² (Figure 4E). Of note, our MR analysis suggested that a lower isoform ratio of *OAS1*-201 was putatively causal for a lower WBC count (MR $\beta = 0.053$, $P < 6.5 \times 10^{-10}$) (Figure 4E, Supplementary Data 15). These findings support a causal role for this variant in COVID-19 severity.

An irQTL in *ULK3* alters isoform usage and protein structure

The *ULK3* gene (unc-51-like kinase 3, at 15q21.1) is a serine/threonine protein kinase gene on the negative strand of chromosome 15 (15q21.1), involved in regulating sonic hedgehog signaling and autophagy.^{43,44} It contains 16 exons and encodes 11 detectable isoform transcripts in FHS (Supplementary Data 14). Isoform

ENST00000440863.7 (*ULK3*-201) produces a full-length 472-amino acid protein, whereas ENST00000569437.5 (*ULK3*-220) produces a truncated 470-amino acids isoform, missing two amino acids (valine and lysine, “VK”) at the 5' of the 14th exon (Figure 5A, Supplementary Figure 5). The sentinel *cis*-irQTL rs12898397 (T>C), located in exon 14, was strongly associated with lower expression ratio of *ULK3*-201 ($R^2 = 0.87$, $\beta = -0.26$, $P = 1 \times 10^{-1154}$) (Table 2) and higher expression ratio of *ULK3*-220 ($R^2 = 0.63$, $\beta = 0.16$, $P = 1 \times 10^{-573}$) (Figure 5B). Since *ULK3* is transcribed from the negative strand, we converted the sequence surrounding rs12898397 to its complementary form (e.g., T>C becomes A>G at rs12898397), while preserving the 5' to 3' orientation for downstream analyses.

We sought to understand the phenotypic consequences of these isoform proportion changes. The rs12898397-G allele, previously associated with lower diastolic blood pressure (DBP) in a GWAS,⁴⁵ was also associated with a reduced expression ratio of the *ULK3*-201 isoform. Using rs12898397 as the genetic instrument in MR, we found that a lower expression ratio of *ULK3*-201 was causally associated with lower DBP (MR beta = 1.24, $P = 1.8 \times 10^{-6}$) (Figure 5C, Supplementary Data 15). Similarly, lower expression ratio of *ULK3*-201 was inferred to be causally associated with a lower level of lymphocyte percentage, consistent with the GWAS finding⁴² (MR beta = 0.054, $P = 1.4 \times 10^{-12}$). These findings support a model in which downregulation of *ULK3*-201 contributes to lower levels of DBP and lymphocyte percentage, highlighting a mechanistic connection between isoform-specific regulation and cardiovascular traits.

The *ULK3* region (± 250 kb) harbors multiple *cis*-irQTLs and *cis*-eQTLs, with rs12898397 identified as the sentinel irQTL and rs936228 as the sentinel *cis*-eQTL³⁶ using the FHS dataset (Figure 5D, Supplementary Data 16). Overall, these *cis*-

eQTLs displayed lower statistical significance than the *cis*-irQTLs. The sentinel *cis*-eQTL (rs936228) is located approximately 2 kb from the sentinel *cis*-irQTL (rs12898397), with moderate LD with each other (LD $R^2 = 0.62$). Of note, rs936228, was associated with comorbidity of COVID-19 critical illness and hypertension.⁵⁰ Additional SNPs nearby that are in strong ($R^2 > 0.9$) or moderate ($0.61 < R^2 < 0.63$) LD with the two sentinel QTLs, exhibited significant associations with several traits in GWAS (Supplementary Data 16). Future studies are needed to investigate the role of irQTLs and eQTLs in trait associations.

To assess the impact of the rs12898397 A>G variant on *ULK3* splicing and structure, we applied MaxENT to the 9 nucleotides around the variant, the exon 14 (Figure 6A),⁴⁷ which predicts splice site strength using entropy models.⁴⁷ By prediction, the rs12898397-G allele weakens the canonical 5' GT donor site (MaxENT score: 7.93→4.44) while strengthens an alternative GT site within the exon (MaxENT score: 2.98→7.20), promoting skipping of six nucleotides (GTC AAG) and loss of two amino acids ("VK") (Figure 6B). To validate this allele-dependent splicing pattern, we analyzed RNA-seq reads from three randomly selected FHS participants with A/A, A/G, and G/G genotypes (Figure 6C). As expected, the A/A individual used the canonical TG donor site to splice intron 14, while the G/G individual utilized a cryptic TG site, leading to loss of six nucleotides in the mature mRNA. The individual with A/G showed intermediate splicing, with ~50% of reads retaining the six bases . To assess structural impact, we used AlphaFold 3, an AI-powered tool with high prediction accuracy.^{48,49} The Google AlphaFold 3 analysis suggests conformational changes in the 173-amino acid C-terminal region, likely reflecting the VK loss (Figure 6D, Supplementary Figures 6 & 7).

Distinct irQTLs associated with multiple isoform transcripts of *CNN2*

CNN2 (calponin 2, 19p13.3) encodes a regulatory protein that interacts through its three well-conserved 26 amino acid residue domains, with actin, tropomyosin, and other cellular components, to modify cytoskeleton dynamics.^{51,52} *CNN2* consists of five exons (Figure 7A), producing ten detectable isoform transcripts in FHS (Supplementary Data 14). We found two significant sentinel *cis*-irQTLs (Supplementary Data 3), rs930232 (G>A) and rs5014188 (T>C) (pairwise LD, $R^2 = 0.13$, all population in LDlink⁵³), that were linked to three *CNN2* isoforms, ENST00000263097.9 (*CNN2*-201), ENST00000348419.7 (*CNN2*-202), and ENST00000568865.3 (*CNN2*-209) (Figure 7A).

The rs930232-A allele was strongly associated with a lower isoform ratio of *CNN2*-201 ($R^2 = 0.61$, $\beta = -0.032$, $P = 1 \times 10^{-532}$) and with a higher ratio of *CNN2*-202 ($R^2 = 0.65$, $\beta = 0.030$, $P = 1 \times 10^{-586}$); however, it was only weakly associated with *CNN2*-209 ($R^2 = 0.04$, $\beta = -0.00064$, $P = 1 \times 10^{-27}$) (Figure 7B, Supplementary Data 3). Despite being strong irQTL, the rs930232 (G>A) variant was not predicted to directly influence splice sites by either SpliceAI or MaxENT. The rs930232-A allele was associated with a higher neutrophil count ($\beta = 0.021$, $P = 1 \times 10^{-26}$) in 746,667 individuals in a blood-cell composition GWAS⁵⁴ (Figure 7C). In FHS participants, the *CNN2*-201 isoform ratio was negatively associated with neutrophil count ($\beta = -0.00042$, $P = 0.0015$); MR also found that the *CNN2*-201 isoform ratio was negatively associated with neutrophil count ($\beta = -0.50$, $P = 2 \times 10^{-15}$) (Figure 7C, Supplementary Data 15).

Another *cis*-irQTL of *CNN2*, rs5014188-C, was strongly associated with a lower relative level of *CNN2*-209 ($R^2 = 0.58$, $\beta = -0.0034$, $P = 1 \times 10^{-481}$) while it showed no significant association with *CNN2*-201 ($P > 5 \times 10^{-8}$) and a much weaker, positive association with *CNN2*-202 ($R^2 = 0.02$, $\beta = 0.0081$, $P = 1 \times 10^{-13}$) (Figure

7B). The effect of *CNN2* isoform variation on predicted protein sequence is depicted in Figure 7D. Predicted protein sequences were based on the standard translation of each of the mRNA sequences in the UCSC Genome Browser database⁵⁵ (<https://genome.ucsc.edu/index.html>). Three well-known calponin-like repeat motifs are indicated, superimposed on the contributions from exons 5, 6, and 7 in the three protein isoforms.⁵¹ Loss of exon 5 in *CNN2*-202 truncates the first repeat, while the gain of exon 7 in *CNN2*-209 truncates the second repeat and obliterates the third repeat entirely.⁵¹ These motifs are essential to the actin-binding properties of calmodulin 2, and thus, modifications of the relative amounts of these three isoforms, *CNN2*-201, *CNN2*-202, and *CNN2*-209, may mediate observed phenotypic consequences of *CNN2* irQTLs.

DISCUSSION

We conducted comprehensive analyses using whole-genome and whole-transcriptome sequencing to identify genetic variants associated with isoform variation. Using whole blood-derived DNA for WGS and RNA for RNA-seq in 2,622 FHS participants, we identified 10,883 isoforms across genes associated with at least one significant *cis*-irQTL ($P < 5 \times 10^{-8}$) with $MAF \geq 0.01$, and 1,084 isoforms across 590 genes associated with at least one significant *trans*-irQTL ($P = 1.5 \times 10^{-13}$) with $MAF \geq 0.01$. These associations replicated well, with *cis*-irQTL replication rates of 76.7% in FHS and 72.4% in WHI, and *trans*-irQTL replication rates of 80.8% and 64.3%, respectively. Together, these results provide a large-scale catalog of regulatory variants influencing isoform usage and establish a resource for understanding the genetic architecture of splicing regulation.

The isoform ratio method¹⁸ captures the relative expression of each isoform within a gene, allowing the detection of regulatory variation that affects full-length

splice isoforms. It is particularly useful for identifying biologically interpretable changes, such as isoform switches that impact coding potential, untranslated regions, or transcript stability. Compared to exon-level QTLs,²¹ which detect local events like exon skipping, and junction-based LeafCutter QTLs,^{16,17} which are optimized for discovering novel or cryptic intron excision, the isoform ratio method provides a clearer view of transcript-level consequences (Supplementary Data 17). However, it relies on comprehensive and accurate transcript annotations, potentially missing events in poorly annotated genes. It also has lower resolution to identify specific splice sites or exons, may overlook subtle or complex splicing changes. Additionally, the isoform ratio method introduces interdependency among isoforms of the same gene, as their proportions must sum to 1. Consequently, an increase in the expression ratio of one isoform may correspond to a decrease in one or more others. Despite these limitations, its high biological interpretability and relevance to phenotype-driven studies make it a valuable complement to other QTL approaches.

To address these limitations and better characterize isoform ratio QTLs (irQTLs), we integrated multilevel computational annotations including transcript structure, splicing predictions, variant effects, and regulatory data from established genomic databases.³⁰⁻³³ This approach enabled functional interpretation of irQTLs acting at both canonical and non-canonical splice sites. For example, the *cis*-irQTL rs10774671 (G>A) in *OAS1* disrupts a 3' splice acceptor site (predicted by SpliceAI³¹), changing expression from the antiviral isoform *OAS1*-p46 to the less effective *OAS1*-p42, potentially modulating COVID-19 severity by altering how *OAS1* splicing influences white blood cell counts.^{40,56} MR analysis, along with GWAS,⁵⁴ supports a causal role for this splicing event in immune regulation.³⁹⁻⁴¹ Similarly, rs12898397 (T>C) in *ULK3*, alters splice site usage (predicted by MaxEnt⁴⁷),

resulting in exon skipping and a truncated protein, and a causal link to DBP, supported by GWAS⁵⁷ and MR analyses. Finally, irQTLs at the *CNN2* locus showed complex exon inclusion patterns not predicted by existing tools, suggesting regulation via non-canonical mechanisms.

In addition to FHS, we analyzed community-based cohorts, JHS²³ and WHI²⁴, to enable independent replication and enhance the generalizability of our findings across diverse populations. We observed a 72% replication rate in WHI, supporting the robustness of our results. In contrast, replication in JHS was lower (32%), likely reflecting both technical and biological differences. Notably, JHS used RNA-seq from PBMCs, while FHS and WHI used whole blood. To further assess the impact of tissue source and ancestry, we stratified WHI by ancestry. Replication was highest in the full WHI cohort and lower in ancestry-specific subsets, partly reflecting reduced sample size (Supplementary Figure 2). Despite similar sample sizes, replication in JHS African American participants was lower than in WHI African American participants, suggesting that tissue source may have a greater effect on replication than ancestry. While clinical characteristics also varied across cohorts (Table 1, Supplementary Data 2), their influence on splicing could not be directly assessed. These results underscore the importance of considering tissue type, ancestry, and sample size in cross-cohort transcriptomic analyses.

Using our approach, *cis*-irQTLs showed higher external replication rates than *trans*-irQTLs in WHI ($n = 2,005$), consistent with published findings,^{58,59} likely reflecting their proximity to target isoforms and involvement in local regulation. Nevertheless, we observed more replicated *trans*-irQTL-isoform pairs than *cis* pairs in the additional FHS sample ($n = 1,094$), which may be explained by consistent experimental conditions and shared analytical pipelines across FHS discovery and

replication sample. Our study confirmed a subset of splicing QTL-gene pairs previously reported in the FHS,²¹ identifying concordant associations for 17.3% of the 2,650 pairs tested ($P < 1 \times 10^{-4}$). The limited overlap may reflect differences in study design and methodology, including genetic and transcriptomic platforms (imputed SNPs vs. WGS; exon arrays vs. RNA-seq) and the expression features tested (exon-level expression vs. isoform ratios). Some previously reported associations may also represent false positives, highlighting the importance of independent validation with high-resolution data. Our study has limited power to detect rare irQTLs reflecting constraints related to sample size and allele frequency thresholds. Larger and more diverse cohorts, along with alternative methods (e.g., SKAT⁶⁰ and SKAT-O⁶¹), will be needed to capture the full spectrum of rare splicing regulatory variation.

In summary, our study presents a comprehensive map of genetic regulation at the isoform level, offering additional insights into the complexity of transcriptomic variation. By integrating isoform ratio modeling, splicing prediction tools, and causal inference analyses, we demonstrate that irQTLs play a distinct and functionally significant role in potentially mediating genetic associations with human traits. These findings emphasize the role of transcript-level analysis in understanding the mechanisms through which genetic variation influences phenotype. Future research should prioritize systematic functional characterization^{62,63} of irQTLs and clarify their unique contributions to variant-to-trait pathways relative to eQTLs.

METHODS

Ethical compliance

All research was conducted in accordance with relevant ethical guidelines and regulations. Study protocols were approved by the Institutional Review Boards (IRBs) at participating sites: the Boston Medical Center and Boston University Medical Campus IRB (FHS); Jackson State University, Tougaloo College, and the University of Mississippi Medical Center IRBs (JHS); and the IRBs of all participating institutions, with oversight from the Fred Hutchinson Cancer Research Center IRB (WHI). All participants provided written informed consent for genetic studies.

Study participants

We included 3,716 participants ($n = 2,622$ for initial discovery and $n = 1,094$ for internal replication), mostly self-identified White American participants primarily from the FHS Offspring and Third Generation cohorts (Table 1).^{64,65} Blood samples for RNA-seq were collected at the ninth examination cycle (2011–2014) of the Offspring cohort and the second examination cycle (2008–2011) of the Third Generation cohort (Table 1). In WHI, blood samples for RNA-seq were collected during a follow-up examination (2012–2013) conducted as part of the Long Life Study (LLS), a sub-study within the larger WHI cohort.²⁴ The LLS repeated selected components of the original WHI physical examination and collected blood samples from approximately 8,000 participants across the United States. A total of 2,005 WHI participants were included in the present analysis, comprising 1,000 White American participants, 918 African American participants, and 87 participants self-identified as other race/ethnicity groups. We did not exclude any participants based on genetic similarity to reference panels such as 1000G. We also analyze 1,010 self-identified African American participants from JHS (Supplementary Information).

This study included 54.5% women in FHS, 100% in WHI, and 64.4% in JHS, based on self-reported sex (Table and Supplementary Data 2). Sex distribution was

not part of the study design, as the cohorts were previously established as population-based studies. No sex-specific analyses were performed.

Isolation of RNA from whole blood

Isolation of RNA from peripheral whole blood^{36,58} was described briefly below. PAXgene™ tubes (PreAnalytiX, Hombrechtikon, Switzerland) were used to collect peripheral whole blood samples (2.5 mL) from both FHS and WHI participants. Total RNA was extracted using a standard protocol with the PAXgene Blood RNA Kit, either at local study-specific laboratories or at the TOPMed contract laboratory at the Northwest Genomics Center. After centrifugation and washing, white blood cell pellets were collected and lysed in guanidinium-containing buffer. The extracted RNA was evaluated for quality by determining absorbance readings at 260 and 280 nm using a NanoDrop ND-1000 UV spectrophotometer. The integrity of total RNA was determined with the Agilent Bioanalyzer 2100 microfluidic electrophoresis (Nano Assay and the Caliper LabChip system). In JHS, RNA was extracted from peripheral blood mononuclear cells (PBMC) samples. Detailed blood sample collection and RNA extraction for JHS were described in Supplementary Information. Given the difference in RNA sources (i.e., whole blood vs. PBMC), we used JHS as the secondary replication cohort.

Whole genome sequencing

Details on DNA extraction, processing, and identification of variants²³ were described briefly below. Whole genome sequencing (WGS) was conducted in one of the TOPMed contract sequencing centers. Consistent sequencing and data processing criteria were used across all centers, with DNA sequence alignment to human genome build GRCh38. The resulting BAM files were sent to TOPMed's Informatics

Research Center (IRC) for re-alignment and standardized BAM file generation. This study used WGS data from the Freeze 10a, which included approximately 208 million SNPs across TOPMed cohorts.

RNA sequencing and processing

RNA sequencing was conducted at one of the NHLBI TOPMed program contract laboratories, following a standard protocol³⁶ for data processing. Base call generation was performed on the NovaSeq 6000 instrument (RTA 3.1.5), followed by the conversion of demultiplexed, unaligned BAM files to FASTQ format using SamTools bam2fq (v1.4). Sequence and base quality were assessed with the FASTX-toolkit (v0.0.13), and sequence alignment to GRCh38 with reference transcriptome GENCODE release 30 was conducted using STAR. FastQC (version 0.11.9) provided initial quality control metrics. A genome index for mapping was built by STAR to generate BAM files at the gene and transcript levels based on the GRCh38 reference build. The TOPMed RNA-seq pipeline utilized RNA-SeQC⁶⁶ to derive standard quality control metrics from aligned reads (Supplementary Information, Supplementary Figure 8). Consistent normalization was conducted across the FHS, JHS, and WHI samples.

Calculation of isoform ratio

Transcript-level expression of an isoform was quantified as Fragments Per Kilobase of transcript per Million mapped reads (FPKM), a normalization measure that considers the gene length in quantifying the expression level of a transcript. The trimmed means of M-values (TMM)⁶⁷ were derived from the FPKM values using the edgeR package.^{66,68} The TMM method estimates scale factors between samples and relative RNA production levels from RNA-seq data. The isoform expression ratio was

calculated as the transcript level of an isoform (TMM value) divided by the sum of the levels of all isoforms (the sum of the TMM values) of its parent gene. The transcript isoform ratio was considered as missing if the overall expression level of its gene was zero.¹⁸

Association analysis of genetic variants with isoform ratios

We performed linear regression analyses of genome-wide SNPs having minor allele frequencies (MAFs) ≥ 0.01 and Hardy-Weinburg Equilibrium (HWE) $p \geq 1e-10$ with isoform ratios for transcripts having the population median TMM ≥ 2 and fewer than 20% missing values. The dosage (0, 1, or 2) of the alternative allele for each SNP was used as a continuous independent variable (additive model), with the isoform ratio as the dependent variable, adjusting for age, self-reported sex (not in WHI), white blood cell counts, and additional variates described below. In FHS, the isoform ratio was adjusted for family relationships to obtain residuals, which were then used as the dependent variable, while in WHI, the isoform ratio was used directly in the regression model. To minimize confounding, regression models included five genotype-based principal components (PCs) to account for population stratification and 15 transcriptome-based PCs to adjust for unknown confounders affecting gene expression. We explored additional technical covariates, including year of blood collection, batch (sequencing machine and time, plate and well), and RNA concentration³⁶ but these covariates were not included in the final models since they explained $< 1\%$ of variance in isoform ratios. We employed the NIH-supported STRIDES cloud computing infrastructure for all association analyses. We used a Graphical Processing Unit (GPU)-based program⁵⁸ for all computations. We stored effect sizes, standard error, partial R-squared, and p-values for SNP-isoform pairs

with $P < 1e-4$ to facilitate future meta-analysis and lookups by the broad scientific community.

We identified significant isoform-ratio QTLs (irQTL). A *cis*-irQTL was defined as a polymorphism falling within 1 Mb of the transcription start site (TSS) of the isoform. In contrast, a *trans*-irQTL was defined as a variant located more than 1 Mb from the TSS or on a different chromosome. We used $P < 5 \times 10^{-8}$ to denote a significant *cis*-irQTL and $P < 1.5 \times 10^{-13}$ ($\sim 0.05/(25,642 \text{ transcripts} \times 12,887,093 \text{ SNPs})$) to define a significant *trans*-irQTL.

Internal and external replication

Association analyses were conducted in two stages: an initial discovery stage ($n = 2,622$), followed by a replication (Table 1). We conducted replication on the irQTL-isoform pairs from sentinel irQTLs (the most significant irQTL per isoform). Internal replication was performed in an additional FHS sample ($n = 1,094$), while external replication was performed in the WHI cohort ($n = 2,005$, primary external replication) and the JHS cohort ($n = 1,010$, secondary external replication, given the use of a different RNA-seq source) cohorts (Table 1, Supplementary Information, Supplementary Data 2 & 7). In WHI, we conducted analyses in the full sample, White American participants ($n = 1,000$) and African American participants ($n = 918$). Replication was assessed only for irQTL-isoform pairs that were present in both FHS and the respective replication cohort. An irQTL-isoform pair was considered successfully replicated if it had a $P < 1e-4$ (for both *cis*- and *trans*- pairs) and consistent effect direction (i.e., the same sign of beta estimates) between the discovery and replication data. The replication rate was calculated as the proportion of irQTL-isoform pairs meeting these criteria among all pairs that were present in both the discovery and the replication samples. We calculated replication rate for

irQTL-isoform pairs from sentinel *cis*-irQTLs and *trans*-irQTLs. A sentinel *cis*-irQTL was defined as the variant most strongly associated with an isoform (i.e., lowest *p*-value) within 1 Mb. For each isoform, a sentinel *trans*-irQTL was identified on each of the 23 chromosomes (22 autosomes and the X chromosome).

Comparison of functional relevance of *cis*-irQTLs and *cis*-eQTLs

We conducted several analyses to explore the functional relevance of *cis*-irQTLs in comparison to *cis*-eQTLs³⁶ in the same participants. We compared the number of genes containing *cis*-irQTLs to those containing *cis*-eQTLs. We also compared the positions of sentinel *cis*-eQTLs and *cis*-irQTLs relative to the 5' start and 3' end of their associated genes. Additionally, we evaluated the variance explained by *cis*-irQTLs compared to that explained by *cis*-eQTLs.

We assessed the enrichment of sentinel *cis*-irQTLs and *cis*-eQTLs in relevant genomic features, including splice donor/acceptor sites, intron/exon boundaries, and eCLIP peaks for binding sites of 203 RNA-binding proteins (RBPs) from ENCODE.^{69,70} To reduce bias, we generated 1,000 sets, consisting of randomly selected SNPs with MAF $\geq 1\%$ and Hardy-Weinberg equilibrium $P \geq 1 \times 10^{-10}$, one for each *cis*-irQTL or *cis*-eQTL, matched based on similar MAF and distance to the TSS for the respective *cis*-irQTLs or *cis*-eQTLs, using the following bin boundaries: MAF bins at 0.01, 0.05, 0.1, 0.2, 0.3, and 0.5, and TSS distance bins at 0kb, 1kb, 10kb, 100kb, and 1Mb. Enrichment of *cis*-irQTLs and *cis*-eQTLs in each feature was calculated as follows:

$$sQTL \text{ enrichment} = \frac{\text{proportion of observed top } sQTL \text{ with a genomic feature}}{\text{Average proportion of random SNPs with a genomic feature}}$$

$$eQTL \text{ enrichment} = \frac{\text{proportion of observed top } eQTL \text{ with a genomic feature}}{\text{Average proportion of random SNPs with a genomic feature}}$$

Enrichment analysis of *cis*-irQTLs with GWAS

Overlap of irQTLs with GWAS Catalog SNPs

We selected GWAS Catalog SNPs²⁶ using two criteria: 1) a significance threshold of $P < 5 \times 10^{-8}$, and 2) a sample size threshold reported with both discovery and replication sample sizes ≥ 5000 , or the discovery sample size $\geq 10,000$. Replicated irQTLs were those in the irQTL-isoform pairs with $P < 5 \times 10^{-8}$ in the discovery sample. We then matched the selected GWAS Catalog SNPs with the identified irQTLs.

Enrichment analyses using Fisher's exact test

We tested the null hypothesis that irQTLs are independent of GWAS SNPs (i.e., not enriched among them). A total of 12,887,093 SNPs across the genome (MAF $\geq 1\%$ and Hardy-Weinberg equilibrium $P \geq 1e-10$) served as the reference set. For example, we constructed a 2×2 contingency table including: (1) the number of SNPs overlapping between GWAS SNPs and irQTLs, (2) SNPs present in GWAS but not in irQTLs, (3) SNPs in irQTLs but not in GWAS, and (4) all remaining SNPs. Fisher's exact test was used in this analysis. We presented results with $P < 1 \times 10^{-4}$.

Prediction of SNP effects on alternative splicing

We used two methods, SpliceAI³¹ and MaxENT,⁴⁷ to predict the SNP effects for the 25 top irQTLs on transcriptomic splice variation. SpliceAI uses deep a neural network to identify splicing and helps predict non-coding cryptic splice variants in the human genome.³¹ MaxENT refers to maximum entropy models that are specified by changing a set of constraints to discriminate human DNA 5'GT (i.e., donor) and 3'AG (i.e., acceptor) splice sites from non-functional decoys.⁴⁷

Prediction of protein structures by Google AlphaFold 3.0

We used Google AlphaFold 3.0 to predict protein structures^{48,49} for the splicing variant rs12898397 [T>C] in *ULK3* [unc-51 like kinase 3]) as an example of how such variants might affect protein structure (Supplementary Information). rs12898397 [T>C] is located in the COOH-terminal region of ULK3 (NM_001411082.1). To confirm the structure, we first aligned it with the reference 4WZX crystal structure from the Protein Data Bank⁷¹ using the PyMOL Molecular Graphics System (Version 3.0.4, Schrödinger, LLC).⁴⁹ Next, we predicted the structures of the COOH-terminal region in responding to rs12898397 [T>C] using Google AlphaFold 3.0.^{48,49} The overlay of these structures was visualized using PyMOL.⁴⁹ The structural differences were evaluated using root-mean-square deviation (RMSD) values, quantified by an Angstrom units (Å). An Å of 1.0 or lower Indicates nearly identical in conformations, while Å of 3.0 or higher suggests significant structural differences.⁷²

Mendelian randomization (MR) analysis

To illustrate the utility of our irQTL resource, we conducted MR analysis to infer causality between the isoform ratios (as the exposure) of the respective transcripts for selected isoforms that harbor significant *cis*-irQTLs and CVD factors and blood composition traits (as the outcome). *cis*-irQTLs served as instrumental variables (IVs). We used the MR-base²⁷ software to conduct the MR analysis in conjunction with summary statistics of SNPs from a large GWAS database²⁶ (Supplementary Data 18). To minimize bias, we selected SNPs in MR analysis based on their linkage disequilibrium (LD) $r^2 < 0.001$.²⁷ We applied the inverse variance weighted (IVW) method to summarize results in MR analyses and used MR-Egger method to assess SNP horizontal pleiotropic effects.²⁷

All statistical analyses were conducted using R (version 4.4.0) or JMP™ (version 18.1.1). Throughout the manuscript, we reported small p-values as $-\log P$, representing $-\log_{10}$ (p-value).

Data Availability

The whole genome sequencing (WGS), RNA sequencing (RNA-seq), and phenotypic data from the Framingham Heart Study (FHS), the Jackson Heart Study (JHS), and the Women's Health Study (WHS) used in this study are available at the dbGaP database under accession codes, phs000007.v32.p13 (FHS) [https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000007.v34.p15], phs000964 (JHS) [https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000964.v5.p1], phs001237 (WHI) [https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001237.v3.p1]. The WGS, RNA-seq, and phenotypic data from all cohorts used in this study are available under restricted access to ensure confidentiality and protect participant privacy. Access can be obtained by submitting an ancillary study proposal and obtaining IRB approval. Timelines for the approval process range from 3–9 weeks.

Results reported in this study are provided in the Supplementary information/Data accompanying this paper. Data for generating Figures (Figures 2A, 2B, 3A, 3B, 3C, 3D, 4D, 5B, 5D, 7B; Supplementary Figures 1-4) have been deposited at <https://doi.org/10.6084/m9.figshare.29885138>. All summary data generated in this study for irQTLs with $P < 1 \times 10^{-4}$ have been deposited at: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001974.v8.p1. The data are accessible via dbGaP, with a

straightforward application process to ensure secure and ethical use. Researchers must first obtain an eRA Commons account, which is typically set up within 3-7 business days if the applicant's institution is already registered with NIH (institutional registration may add an additional 2-4 weeks). Once the account is active, access to the data can be requested by submitting a Data Access Request (DAR) through dbGaP. DARs are reviewed on a biweekly basis, and approvals are generally granted within 2-4 weeks, provided the documentation is complete. Following approval, researchers can seamlessly log in to the NIH BioData Catalyst (BDC) platform using the same eRA Commons credentials to download or directly analyze the data, enabling timely and secure access for downstream research.

Code availability statement

Codes for generating Figures 2A, 2B, 3A, 3B, 3C, 3D, 4D, 5B, 5D, 7B;

Supplementary Figures 1-4 are available at

<https://doi.org/10.6084/m9.figshare.29885138>.

Reference

- 1 Reyes, A. & Huber, W. Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic Acids Res* **46**, 582-592 (2018). <https://doi.org/10.1093/nar/gkx1165>
- 2 Kalsotra, A. & Cooper, T. A. Functional consequences of developmentally regulated alternative splicing. *Nat Rev Genet* **12**, 715-729 (2011). <https://doi.org/10.1038/nrg3052>
- 3 Tian, B. & Manley, J. L. Alternative polyadenylation of mRNA precursors. *Nat Rev Mol Cell Biol* **18**, 18-30 (2017). <https://doi.org/10.1038/nrm.2016.116>
- 4 Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* **40**, 1413-1415 (2008). <https://doi.org/10.1038/ng.259>
- 5 Chow, L. T., Gelinas, R. E., Broker, T. R. & Roberts, R. J. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell* **12**, 1-8 (1977). [https://doi.org/10.1016/0092-8674\(77\)90180-5](https://doi.org/10.1016/0092-8674(77)90180-5)
- 6 Scotti, M. M. & Swanson, M. S. RNA mis-splicing in disease. *Nat Rev Genet* **17**, 19-32 (2016). <https://doi.org/10.1038/nrg.2015.3>
- 7 Wang, G. S. & Cooper, T. A. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat Rev Genet* **8**, 749-761 (2007). <https://doi.org/10.1038/nrg2164>
- 8 Kornblihtt, A. R. *et al.* Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nat Rev Mol Cell Biol* **14**, 153-165 (2013). <https://doi.org/10.1038/nrm3525>
- 9 Shayevitch, R., Askayo, D., Keydar, I. & Ast, G. The importance of DNA methylation of exons on alternative splicing. *RNA* **24**, 1351-1362 (2018). <https://doi.org/10.1261/rna.064865.117>
- 10 Barash, Y. *et al.* Deciphering the splicing code. *Nature* **465**, 53-59 (2010). <https://doi.org/10.1038/nature09000>
- 11 Battle, A. *et al.* Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res* **24**, 14-24 (2014). <https://doi.org/10.1101/gr.155192.113>
- 12 Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506-511 (2013). <https://doi.org/10.1038/nature12531>
- 13 Takata, A., Matsumoto, N. & Kato, T. Genome-wide identification of splicing QTLs in the human brain and their enrichment among schizophrenia-associated loci. *Nat Commun* **8**, 14519 (2017). <https://doi.org/10.1038/ncomms14519>
- 14 Raj, T. *et al.* Integrative transcriptome analyses of the aging brain implicate altered splicing in Alzheimer's disease susceptibility. *Nat Genet* **50**, 1584-1592 (2018). <https://doi.org/10.1038/s41588-018-0238-1>
- 15 Tian, J. *et al.* CancerSplicingQTL: a database for genome-wide identification of splicing QTLs in human cancer. *Nucleic Acids Res* **47**, D909-D916 (2019). <https://doi.org/10.1093/nar/gky954>
- 16 Consortium, G. T. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318-1330 (2020). <https://doi.org/10.1126/science.aaz1776>
- 17 Garrido-Martin, D., Borsari, B., Calvo, M., Reverter, F. & Guigo, R. Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome. *Nat Commun* **12**, 727 (2021). <https://doi.org/10.1038/s41467-020-20578-2>
- 18 Yamaguchi, K. *et al.* Splicing QTL analysis focusing on coding sequences reveals mechanisms for disease susceptibility loci. *Nat Commun* **13**, 4659 (2022). <https://doi.org/10.1038/s41467-022-32358-1>
- 19 He, Y., Huang, L., Tang, Y., Yang, Z. & Han, Z. Genome-wide Identification and Analysis of Splicing QTLs in Multiple Sclerosis by RNA-Seq Data. *Front Genet* **12**, 769804 (2021). <https://doi.org/10.3389/fgene.2021.769804>
- 20 Qi, T. *et al.* Genetic control of RNA splicing and its distinct role in complex trait variation. *Nat Genet* **54**, 1355-1363 (2022). <https://doi.org/10.1038/s41588-022-01154-4>

- 21 Zhang, X. *et al.* Identification of common genetic variants controlling transcript isoform variation in human whole blood. *Nat Genet* **47**, 345-352 (2015).
<https://doi.org/10.1038/ng.3220>
- 22 GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318-1330 (2020). <https://doi.org/10.1126/science.aaz1776>
- 23 Wen, J. *et al.* Gene expression and splicing QTL analysis of blood cells in African American participants from the Jackson Heart Study. *Genetics* **228** (2024).
<https://doi.org/10.1093/genetics/iyae098>
- 24 Design of the Women's Health Initiative clinical trial and observational study. The Women's Health Initiative Study Group. *Control Clin Trials* **19**, 61-109 (1998).
[https://doi.org/10.1016/s0197-2456\(97\)00078-0](https://doi.org/10.1016/s0197-2456(97)00078-0)
- 25 Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**, D1005-D1012 (2019). <https://doi.org/10.1093/nar/gky1120>
- 26 Sollis, E. *et al.* The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res* **51**, D977-D985 (2023). <https://doi.org/10.1093/nar/gkac1010>
- 27 Hemani, G. *et al.* The MR-Base platform supports systematic causal inference across the human phenome. *Elife* **7** (2018). <https://doi.org/10.7554/eLife.34408>
- 28 Kang, J. *et al.* Ribosomal proteins and human diseases: molecular mechanisms and targeted therapy. *Signal Transduct Target Ther* **6**, 323 (2021). <https://doi.org/10.1038/s41392-021-00728-8>
- 29 Ebert, B. L. *et al.* Identification of RPS14 as a 5q- syndrome gene by RNA interference screen. *Nature* **451**, 335-339 (2008). <https://doi.org/10.1038/nature06494>
- 30 Frankish, A. *et al.* GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Res* **51**, D942-D949 (2023). <https://doi.org/10.1093/nar/gkac1071>
- 31 Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535-548 e524 (2019). <https://doi.org/10.1016/j.cell.2018.12.015>
- 32 Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80-92 (2012). <https://doi.org/10.4161/fly.19695>
- 33 Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-589 (2010). <https://doi.org/10.1016/j.molcel.2010.05.004>
- 34 Kulski, J. K., Suzuki, S. & Shiina, T. Human leukocyte antigen super-locus: nexus of genomic supergenes, SNPs, indels, transcripts, and haplotypes. *Hum Genome Var* **9**, 49 (2022).
<https://doi.org/10.1038/s41439-022-00226-5>
- 35 Hori, N. *et al.* A new human ribosomal protein sequence, homologue of rat L9. *Nucleic Acids Res* **21**, 4395 (1993). <https://doi.org/10.1093/nar/21.18.4395>
- 36 Liu, C. *et al.* Whole genome DNA and RNA sequencing of whole blood elucidates the genetic architecture of gene expression underlying a wide range of diseases. *Sci Rep* **12**, 20167 (2022). <https://doi.org/10.1038/s41598-022-24611-w>
- 37 Field, L. L. *et al.* OAS1 splice site polymorphism controlling antiviral enzyme activity influences susceptibility to type 1 diabetes. *Diabetes* **54**, 1588-1591 (2005).
<https://doi.org/10.2337/diabetes.54.5.1588>
- 38 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 (2005). <https://doi.org/10.1086/429391>
- 39 Huffman, J. E. *et al.* Multi-ancestry fine mapping implicates OAS1 splicing in risk of severe COVID-19. *Nat Genet* **54**, 125-127 (2022). <https://doi.org/10.1038/s41588-021-00996-8>
- 40 Wickenhagen, A. *et al.* A prenylated dsRNA sensor protects against severe COVID-19. *Science* **374**, eabj3624 (2021). <https://doi.org/10.1126/science.abj3624>

- 41 Zhou, S. *et al.* A Neanderthal OAS1 isoform protects individuals of European ancestry against
COVID-19 susceptibility and severity. *Nat Med* **27**, 659-667 (2021).
<https://doi.org/10.1038/s41591-021-01281-1>
- 42 Vuckovic, D. *et al.* The Polygenic and Monogenic Basis of Blood Traits and Diseases. *Cell* **182**,
1214-1231 e1211 (2020). <https://doi.org/10.1016/j.cell.2020.08.008>
- 43 Rubin, L. L. & de Sauvage, F. J. Targeting the Hedgehog pathway in cancer. *Nat Rev Drug*
Discov **5**, 1026-1033 (2006). <https://doi.org/10.1038/nrd2086>
- 44 Luo, S., Zheng, N. & Lang, B. ULK4 in Neurodevelopmental and Neuropsychiatric Disorders.
Front Cell Dev Biol **10**, 873706 (2022). <https://doi.org/10.3389/fcell.2022.873706>
- 45 Keaton, J. M. *et al.* Genome-wide analysis in over 1 million individuals of European ancestry
yields improved polygenic risk scores for blood pressure traits. *Nat Genet* **56**, 778-791
(2024). <https://doi.org/10.1038/s41588-024-01714-w>
- 46 Ciccarelli, F. D. *et al.* The identification of a conserved domain in both spartin and spastin,
mutated in hereditary spastic paraplegia. *Genomics* **81**, 437-441 (2003).
[https://doi.org/10.1016/s0888-7543\(03\)00011-9](https://doi.org/10.1016/s0888-7543(03)00011-9)
- 47 Yeo, G. & Burge, C. B. Maximum entropy modeling of short sequence motifs with
applications to RNA splicing signals. *J Comput Biol* **11**, 377-394 (2004).
<https://doi.org/10.1089/1066527041410418>
- 48 Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**,
583-589 (2021). <https://doi.org/10.1038/s41586-021-03819-2>
- 49 Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold
3. *Nature* **630**, 493-500 (2024). <https://doi.org/10.1038/s41586-024-07487-w>
- 50 Guo, H., Li, T. & Wen, H. Identifying shared genetic loci between coronavirus disease 2019
and cardiovascular diseases based on cross-trait meta-analysis. *Front Microbiol* **13**, 993933
(2022). <https://doi.org/10.3389/fmicb.2022.993933>
- 51 Liu, R. & Jin, J. P. Calponin isoforms CNN1, CNN2 and CNN3: Regulators for actin
cytoskeleton functions in smooth muscle and non-muscle cells. *Gene* **585**, 143-153 (2016).
<https://doi.org/10.1016/j.gene.2016.02.040>
- 52 Plagnol, V., Smyth, D. J., Todd, J. A. & Clayton, D. G. Statistical independence of the
colocalized association signals for type 1 diabetes and RPS26 gene expression on
chromosome 12q13. *Biostatistics* **10**, 327-334 (2009).
<https://doi.org/10.1093/biostatistics/kxn039>
- 53 Machiela, M. J. & Chanock, S. J. LDlink: a web-based application for exploring population-
specific haplotype structure and linking correlated alleles of possible functional variants.
Bioinformatics **31**, 3555-3557 (2015). <https://doi.org/10.1093/bioinformatics/btv402>
- 54 Chen, M. H. *et al.* Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667
Individuals from 5 Global Populations. *Cell* **182**, 1198-1213 e1114 (2020).
<https://doi.org/10.1016/j.cell.2020.06.045>
- 55 Nassar, L. R. *et al.* The UCSC Genome Browser database: 2023 update. *Nucleic Acids Res* **51**,
D1188-D1195 (2023). <https://doi.org/10.1093/nar/gkac1072>
- 56 Soveg, F. W. *et al.* Endomembrane targeting of human OAS1 p46 augments antiviral activity.
Elife **10** (2021). <https://doi.org/10.7554/eLife.71047>
- 57 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 (2018).
<https://doi.org/10.1038/s41588-018-0205-x>
- 58 Joeanes, R. *et al.* Integrated genome-wide analysis of expression quantitative trait loci aids
interpretation of genomic association studies. *Genome Biol* **18**, 16 (2017).
<https://doi.org/10.1186/s13059-016-1142-6>
- 59 Tian, C. *et al.* Single-cell RNA sequencing of peripheral blood links cell-type-specific
regulation of splicing to autoimmune and inflammatory diseases. *Nat Genet* **56**, 2739-2752
(2024). <https://doi.org/10.1038/s41588-024-02019-8>

- Wu, M. C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* **89**, 82-93 (2011).
<https://doi.org/10.1016/j.ajhg.2011.05.029>
- Lee, S. *et al.* Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet* **91**, 224-237 (2012). <https://doi.org/10.1016/j.ajhg.2012.06.007>
- Liu, L., Das, U., Ogunisola, S. & Xie, J. Transcriptome-Wide Detection of Intron/Exon Definition in the Endogenous Pre-mRNA Transcripts of Mammalian Cells and Its Regulation by Depolarization. *Int J Mol Sci* **23** (2022). <https://doi.org/10.3390/ijms231710157>
- Khodor, Y. L., Menet, J. S., Tolan, M. & Rosbash, M. Cotranscriptional splicing efficiency differs dramatically between *Drosophila* and mouse. *RNA* **18**, 2174-2186 (2012).
<https://doi.org/10.1261/rna.034090.112>
- Splansky, G. L. *et al.* The Third Generation Cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: design, recruitment, and initial examination. *Am J Epidemiol* **165**, 1328-1335 (2007). <https://doi.org/10.1093/aje/kwm021>
- Kannel, W. B., Feinleib, M., McNamara, P. M., Garrison, R. J. & Castelli, W. P. An investigation of coronary heart disease in families. The Framingham offspring study. *Am J Epidemiol* **110**, 281-290 (1979). <https://doi.org/10.1093/oxfordjournals.aje.a112813>
- Lun, A. T., Chen, Y. & Smyth, G. K. It's DE-licious: A Recipe for Differential Expression Analyses of RNA-seq Experiments Using Quasi-Likelihood Methods in edgeR. *Methods Mol Biol* **1418**, 391-416 (2016). https://doi.org/10.1007/978-1-4939-3578-9_19
- Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* **11**, R25 (2010). <https://doi.org/10.1186/gb-2010-11-3-r25>
- Chen, Y., Lun, A. T. & Smyth, G. K. From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. *F1000Res* **5**, 1438 (2016). <https://doi.org/10.12688/f1000research.8987.2>
- Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012). <https://doi.org/10.1038/nature11247>
- Luo, Y. *et al.* New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res* **48**, D882-D889 (2020). <https://doi.org/10.1093/nar/gkz1062>
- Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res* **28**, 235-242 (2000).
<https://doi.org/10.1093/nar/28.1.235>
- Magala, P., Klevit, R. E., Thomas, W. E., Sokurenko, E. V. & Stenkamp, R. E. RMSD analysis of structures of the bacterial protein FimH identifies five conformations of its lectin domain. *Proteins* **88**, 593-603 (2020). <https://doi.org/10.1002/prot.25840>

Acknowledgement

The Framingham Heart Study (FHS) was supported by NIH contracts N01-HC-25195, HHSN268201500001I, and 75N92019D00031. Molecular data for the Trans-Omics in Precision Medicine (TOPMed) program were supported by the National Heart, Lung, and Blood Institute (NHLBI). Whole genome sequencing for NHLBI TOPMed: Whole Genome Sequencing and Related Phenotypes in the Framingham Heart Study (phs000974.v1.p1) was performed at the Broad Institute Genomics Platform under grants 3R01HL092577-06S1 and 3U54HG003067-12S2. RNA-seq for the same project was performed at the Northwest Genomics Center under contract HHSN268201600032I. RNA-seq assays were supported in part by the Division of Intramural Research (PI D. L.) and an NIH Director's Challenge Award (PI D. L.). Centralized genomic read mapping, genotype calling, and variant quality control were provided by the TOPMed Informatics Research Center 3R01HL117626-02S1; HHSN268201800002I. The analytical component of this project was funded in part by the NHLBI Division of Intramural Research (D.L., R.J., T.H., S-J.H, C.Y.D),

The Jackson Heart Study (JHS) is supported and conducted in collaboration with Jackson State University HHSN268201800013I, Tougaloo College HHSN268201800014I, the Mississippi State Department of Health HHSN268201800015I/HHSN26800001, and the University of Mississippi Medical Center HHSN268201800010I, HHSN268201800011I, and HHSN268201800012I, under contracts from the National Heart, Lung, and Blood Institute (NHLBI) and the National Institute on Minority Health and Health Disparities (NIMHD). We thank the staff and participants of the JHS. Genome sequencing for the JHS (phs000964.v1.p1) was performed at the Northwest Genomics Center HHSN268201100037C as part of the NHLBI Trans-Omics for Precision Medicine (TOPMed) program. Centralized genomic read mapping, genotype calling, and variant quality control were provided by the TOPMed Informatics Research Center 3R01HL117626-02S1; HHSN268201800002I. Phenotype harmonization, data management, sample identity quality control (QC), and program coordination were supported by the TOPMed Data Coordinating Center (R01HL120393; U01HL120393; HHSN268201800001I), and further supported through R01HL129132 (Y.L.), R01AG075884 (L.M.R.), and R01HL146500 (A.P.R.).

The Women's Health Initiative (WHI) program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts 75N92021D00001, 75N92021D00002, 75N92021D00003, 75N92021D00004, 75N92021D00005. The TOPMed WHI (phs001237) were supported by HHSN268201500014C and HHSN268201600034I (Broad Institute, WGS, RNAseq and metabolomics), HHSN268201600038I (Keck MGC, Methylomics). The RNAseq data from WHI was also supported through TOPMed X01 HL139376 and X01HL153408. Scientific Computing Infrastructure at Fred Hutch was funded by ORIP grant S10OD02868, and partly supported by R01HL152439 (J.W.H., C.K., A.P.R.), R01HL175681 (N.F.), R01HG013163 (N.F.), R01HL163972 (N.F.), K26DK138425 (N.F.).

No other authors received project-specific funding related to this study.

Author Contributions Statement

C.L., R.J., J.M., P.M., and D.L. conceived the project. R.J., C.L., J.M., and P.M. performed statistical analyses. J.Y., M.W., and J.X. contributed to functional annotation analyses. N.L.H. and J.W.H. assisted with data acquisition. L.M.R., A.R., and Y.L. provided funding for sample processing, RNA sequencing and analysis for JHS; G.O., J.M.M., and D.L. for FHS; and C.K. for WHI. C.L. wrote the manuscript. R.J., J.M., P.M., and D.L. reviewed and edited the manuscript. L.M.R., T.H., S.-J.H., J.W., Q.S., D.Y.C., P.O., N.F., P.L.A., and A.P.C. provided critical feedback. P.M. and D.L. supervised the project.

Conflict of interest

LMR is a consultant for the TOPMed Administrative Coordinating Center (through Westat). The remaining authors declare no competing interests.

ARTICLE IN PRESS

Table 1. Participant characteristics of the discovery and replication samples in the Framingham Heart Study (FHS) and Women's Health Initiative (WHI)

Variable mean (SD) or %	FHS		WHI external replication (n = 2,005)
	Discovery (n = 2,622)	Internal replication (n = 1,094)	
Women	53.7	56.4	100
Age, years	52.9 (14.0)	52.1 (12.7)	79.2 (6.7)
BMI, kg/m ²	28.0 (5.7)	28.0 (5.8)	28.7 (6.0)
SBP, mmHg	118.6 (15.6)	119.0 (15.0)	126.9 (15.1)
DBP, mmHg	73.4 (9.6)	73.8 (9.2)	72.7 (9.1)
Fasting glucose,	98.6 (20.8)	97.1 (15.3)	100.4 (30.4)
TC, mg/dL	185.9 (35.9)	188.9 (37.0)	195.7 (39.6)
HDL, mg/dL	60.1 (18.2)	60.9 (18.5)	60.4 (15.0)
Trig, mg/dL	114.2 (77.9)	114.6 (74.4)	104.9 (52.4)
LDL, mg/dL	103.1 (30.8)	105.1 (31.1)	114.4 (35.0)
Current smoking	8.6	10.7	2.6
Hypertension	64.3	49.4	82.1
Diabetes	7.5	5.7	34.3
HRX	26.7	24.9	42.6
LIPIDRX	24.8	23.0	40.3
DMRX	5.7	4.0	14.1

BMI, body mass index; SBP/DBP, systolic/diastolic blood pressure; TC, total cholesterol; HDL, high density lipoprotein; Trig, triglyceride; LDL, low-density lipoprotein; HRX, treatment for hypertension; LIPIDRX, treatment for high lipid level; DMRX, treatment for diabetes

981 **Table 2.** Top *cis*-irQTL-isoform pairs identified in the Framingham Heart Study and replication in Women's Health Initiative

Sentinel SNP	A:EA	Isoform	Chr	SNP Pos	Gene Symbol	FHS (n = 2,622)				WHI (n = 2,005)			
						EAF	R ²	Beta	-logP	EAF	R ²	Beta	-logP
rs4841	C:T	ENST00000519690.1	5	150446963	<i>RPS14</i>	0.25	0.96	0.045	1859	0.23	0.90	0.02	1004
rs10774671	G:A	ENST00000202917.10	12	112919388	<i>OAS1</i>	0.62	0.95	-0.258	1714	0.55	0.95	-0.26	1249
rs73191241	A:G	ENST00000388962.4	12	108619543	<i>SELPLG</i>	0.19	0.92	0.095	1443	0.14	0.77	0.11	637
rs62477626	T:C	ENST00000475668.6	7	142065610	<i>MGAM</i>	0.19	0.92	-0.352	1402	0.12	0.72	-0.26	549
rs7513045	G:T	ENST00000495526.5	1	39738494	<i>PPIE</i>	0.36	0.91	0.219	1314	0.30	0.88	0.08	893
rs2318043	A:G	ENST00000381605.8	20	1615827	<i>SIRPB1</i>	0.21	0.9	-0.261	1297	0.20	0.73	-0.21	560
rs41280971	G:T	ENST00000493091.5	7	100356834	<i>PILRB</i>	0.18	0.9	0.15	1284	0.15	0.81	0.07	702
rs2425068	T:C	ENST00000437340.5	20	35626801	<i>CPNE1</i>	0.08	0.88	0.152	1169	0.05	0.84	0.12	788
rs12898397	T:C	ENST00000440863.7	15	74837752	<i>ULK3</i>	0.6	0.87	-0.263	1154	0.39	0.73	-0.21	561
rs7309653	A:G	ENST00000541272.1	12	124910282	<i>UBC</i>	0.79	0.86	0.012	1103	0.54	0.87	0.05	875
rs13320	C:T	ENST00000368814.8	1	151874041	<i>THEM4</i>	0.58	0.86	0.165	1082	0.53	0.70	0.11	523
rs14434	C:T	ENST00000460851.6	3	150584223	<i>EIF2A</i>	0.86	0.85	-0.291	1045	0.74	0.80	-0.17	681
rs28374262	A:G	ENST00000465395.5	2	218234030	<i>ARPC2</i>	0.49	0.85	0.008	1044	0.32	0.67	0.01	483
rs2724360	C:T	ENST00000357714.5	1	207769813	<i>CD46</i>	0.8	0.84	-0.149	1012	0.78	0.80	-0.15	686
rs2241976	A:G	ENST00000418251.6	2	113185893	<i>PSD4</i>	0.41	0.84	0.101	1011	0.55	0.73	0.08	568
rs7063	A:T	ENST00000296754.7	5	96774507	<i>ERAP1</i>	0.3	0.83	0.165	1000	0.22	0.79	0.17	676
rs7171507	T:C	ENST00000564570.1	15	75444946	<i>MAN2C1</i>	0.28	0.83	0.038	1000	0.38	0.80	0.07	688
rs9616714	G:A	ENST00000405854.5	22	49657211	<i>*C22orf34</i>	0.63	0.83	0.298	995	0.45	0.74	0.29	571
rs452717	T:G	ENST00000493242.1	19	54277224	<i>LILRB2</i>	0.85	0.83	-0.129	990	0.73	0.78	-0.10	645
rs10493821	C:T	ENST00000489444.6	1	89009452	<i>GBP3</i>	0.34	0.82	0.29	963	0.21	0.74	0.25	571
rs72494581	T:C	ENST00000514448.5	5	181003797	<i>BTNL8</i>	0.3	0.82	0.135	959	0.27	0.63	0.12	426
rs372699745	TA:T	ENST00000343139.9	6	32593216	<i>HLA-</i>	0.29	0.82	-0.208	951	0.25	0.56	-0.15	357
rs7724759	G:A	ENST00000348386.7	5	96740783	<i>CAST</i>	0.33	0.82	0.1	947	0.22	0.78	0.09	658
rs11382548	C:CA	ENST00000334888.9	11	61398259	<i>TMEM216</i>	0.85	0.81	0.147	938	0.82	0.58	0.14	369
rs34634498	G:A	ENST00000376228.9	6	31268623	<i>HLA-C</i>	0.11	0.81	-0.081	924	0.12	0.58	-0.11	375
rs930232	G:A	ENST00000263097.9	19	1036019	<i>CNN2</i>	0.45	0.61	-0.032	532	0.43	0.43	-0.04	244
rs930232	G:A	ENST00000348419.7	19	1036019	<i>CNN2</i>	0.45	0.64	0.030	586	0.43	0.39	0.03	213
rs5014188	T:C	ENST00000568865.3	19	1037109	<i>CNN2</i>	0.88	0.58	-0.0034	481	0.83	0.58	-0.01	375

982 A:EA, allele:effective allele; EAF, effective allele frequency; R², variance in an isoform ratio that was explained by the sentinel irQTL; -logP, -log₁₀ (p-value).
 983 FHS, Framingham Heart Study; WHI, Women's Health Initiative. Linear regression was performed between the irQTL and genotype dosage, with significance
 984 assessed using a two-sided *t*-test for the null hypothesis, beta = 0. Significance was defined as -log₁₀(P) > 7.301 to account for multiple testing. A random
 985 sentinel *cis*-irQTL was selected per gene if more than one *cis*-irQTL was associated with that gene. *All genes, except for *C22orf34* (a long intergenic non-
 986 coding RNA), are protein-coding genes.

Figure Legend

Figure 1. Schematic Outline. (A) RNA sequencing was performed on whole blood samples from the Framingham Heart Study (FHS, discovery cohort) and the Women's Health Initiative (WHI, replication cohort). These stored samples from community-based populations enabled genome-wide analysis of both sequence and isoform-specific expression across thousands of genes. (B) Expression quantitative trait loci (eQTL) and isoform ratio QTL (irQTL) analyses were performed to identify genetic variants associated with overall gene expression or the relative abundance of specific transcript isoforms. Variant allele dosage (0, 1, or 2 copies) may correspond to low, intermediate, or high expression levels of particular isoforms, reflecting distinct regulatory patterns. Box plots illustrate the distribution of expression across genotype groups, showing the median (central line), interquartile range (box), and range of non-outlier values (whiskers), with individual outliers plotted beyond 1.5×interquartile range (IQR). This panel is schematic and intended for illustrative purposes only; no statistical tests were performed. (C) Functional consequences of genetically driven isoform regulation were evaluated using tools such as Gene Set Enrichment Analysis and external resources including the GWAS Catalog, ENCODE, and SpliceAI, which predicts the impact of genetic variants on RNA splicing. (D) The clinical relevance of irQTLs was assessed by integrating GWAS data and applying Mendelian randomization to identify causal links between genetic variants, alternative splicing, and cardiovascular or hematological traits. This framework connects genotype to splicing variation and ultimately to disease risk or trait expression.

Figure 2. Replication of sentinel *cis*-irQTLs in relation to their associated isoforms. We identified 11,425 common variants ($MAF \geq 0.01$) as sentinel *cis*-irQTLs for 10,883 isoforms across 4,971 genes in the FHS discovery sample ($n = 2,622$). Internal replication was performed in an independent FHS sample ($n = 1,094$), and external replication in the WHI cohort ($n = 2,005$). A sentinel *cis*-irQTL-isoform ratio association was considered replicated if its $P < 1 \times 10^{-4}$ in the replication sample and showed the same direction of effect as in the discovery sample. Replication rate was defined as the proportion of discovered pairs that met these criteria in the replication sample. (A) Replication rates (y-axis) increased with R^2 (x-axis), proportion of isoform ratio variance explained by the irQTL. Average replication rates were 76.7% in FHS and 72.4% in WHI. (B) Comparison of effect sizes between the FHS discovery (x-axis) and WHI replication (y-axis) samples showed strong consistency (Pearson $r = 0.91$; 95%CI: 0.90, 0.91), with 99.8% of *cis*-irQTL-isoform pairs sharing the same direction of association. MAF, minor allele frequency.

Figure 3. Comparison of *cis*-irQTLs and *cis*-eQTLs. (A) Among 2,622 Framingham Heart Study participants, more genes had *cis*-eQTLs³⁶ ($n = 15,707$) than *cis*-irQTLs ($n = 4,971$). (B) Association strength (R^2) of *cis*-eQTLs versus *cis*-irQTLs across 4,971 irQTL genes in 2,622 participants. Dashed lines separate SNPs into stronger irQTLs (cyan, $n = 1,946$), overlapping signals (middle, $n = 2,716$), and stronger eQTLs (red). Highlighted genes (*OAS1*, *ULK3*, *CNN2*) are discussed in the main text. (C) Fold enrichment of sentinel *cis*-eQTLs and *cis*-irQTLs in selected

genomic features based on illustrative regions. Enrichment was calculated using 1,000 matched SNP sets based on their MAFs and TSS distance. Fold enrichment reflects the observed overlap relative to the mean across matched null sets and exhibits a non-normal distribution. Box plots show the distribution of fold enrichment across replicates; white boxes represent the interquartile range, with the horizontal line indicating the median. Based on median (IQR; range), irQTLs showed greater enrichment than eQTLs at splice donor (25.2 [12.6–25.2; 4.2–25.2] versus 9.2 [5.5–13.8; 2.8–27.6]; $P = 4 \times 10^{-56}$), splice acceptor (12.6 [6.3–12.6; 3.2–12.6] versus 5.9 [4.0–11.9; 1.5–11.9]); $P = 9 \times 10^{-127}$), splice region (6.5 [5.6–7.7; 3.5–15.5] versus 3.2 [2.9–3.4; 2.2–5.2]; $P < 2.2 \times 10^{-308}$), and missense sites (3.5 [3.2–3.8; 2.5–5.9] versus 1.9 [1.8–2.0; 1.6–2.6]; $P < 2.2 \times 10^{-308}$). Two-sided Wilcoxon rank-sum tests compared fold enrichment distributions between irQTLs and eQTLs within genomic regions. (D) Fold enrichment of sentinel *cis*-eQTLs and *cis*-irQTLs in RNA-binding protein (RBP) binding sites from 203 in eCLIP datasets profiled by ENCODE, showing stronger enrichment of irQTLs in functional RBP regions. Fold enrichment analysis was performed using the same method described above. Median enrichment (IQR; range) was 1.9 (1.6, 2.4; 0, 5.7) for *cis*-eQTLs and 3.7 (2.6, 4.8; 0, 7.4) for *cis*-irQTLs. MAF, minor allele frequency; HWE, Hardy-Weinberg Equilibrium; IQR, interquartile range; TSS, transcript start site.

Figure 4. Analyses of irQTLs Illustrated with the *OAS1* Gene. (A) *OAS1* on positive strand of chromosome 12 (hg38). Its pre-mRNA undergoes editing and alternative splicing. The sentinel irQTL, rs10774671 (G>A), lies at the splice acceptor site of intron 5. (B) Schematic of alternative splicing: introns with conserved donor (GU/GT) and acceptor (AG) sites are removed from pre-mRNA, generating isoforms with defined splice junctions. (C) IGV visualization of RNA-seq reads near rs10774671 (G>A) shows genotype-dependent splicing in three FHS participants: G/G with exon skipping, A/G with partial splicing, and A/A with full intron 5 splicing. These patterns align with high, intermediate, and low expression of isoform *OAS1*-201 (ENST00000202917.10). (D) Overall *OAS1* expression is weakly associated with rs10774671, showing a decrease with increasing A allele dosage (“as eQTL”). The box plots show median (IQR; range) expression as –0.0027 (–0.16, 0.23; –0.89, 3.0) for G/G (n = 386), –0.074 (–0.25, 0.17; –1.09, 3.93) for A/G (n = 1,172), and –0.11 (–0.33, 0.081; –1.26, 4.28) for A/A (n = 999). In contrast, isoform ratios of *OAS1*-201 and *OAS1*-203 are strongly associated with rs10774671 (G>A) as irQTLs, with *OAS1*-201 decreasing and *OAS1*-203 increasing with A allele dosage. The median (IQR; range) isoform ratio of *OAS1*-201 was 0.94 (0.93, 0.95; 0.49, 0.97) for G/G, 0.74 (0.72, 0.76; 0.60, 0.94) for A/G, and 0.44 (0.41, 0.46; 0.28, 0.92) for A/A. The median (IQR; range) isoform ratio of *OAS1*-203 was 0.018 (0.014, 0.023; 0.0015, 0.20) for G/G, 0.10 (0.089, 0.113; 0.020, 0.18) for A/G, and 0.231 (0.209, 0.252; 0.030, 0.35) for A/A. (E) The rs10774671-A allele is associated with reduced *OAS1*-201 expression in linear regression ($P = 1 \times 10^{-1714}$). MR supports a causal relationship between lower *OAS1*-201 levels and reduced WBC count ($P = 6.5 \times 10^{-10}$), consistent with GWAS findings associating rs10774671-A with lower WBC levels ($P = 3.2 \times 10^{-11}$). All tests are two-sided t-test or z-test (for MR) in regression analyses. IGV, Integrative Genomics Viewer; IQR, interquartile range; WBC, white blood cell; MR, Mendelian randomization. Gene names are italicized.

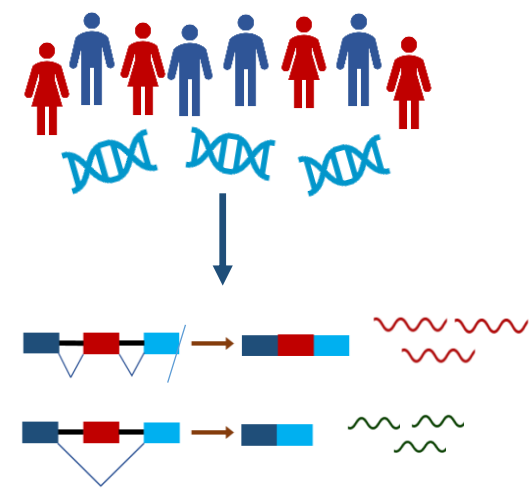
Figure 5. Genetic associations of rs12898397-G with *ULK3* splicing, expression, and blood pressure trait. (A) *unc-51* like kinase 3 (*ULK3*), containing 16 exons, is located on the negative strand of chromosome 15. To maintain 5'→3' transcriptional orientation, the positive-strand sequence around rs12898397 (e.g., T>C) was converted to its complementary negative-strand form (A>G). (B) rs12898397 (A>G) showed the strongest association with the isoform ratio of *ULK3*-201 (ENST00000440863.7), which encodes the full-length 472-aa protein ($\beta = -0.26$, $P = 1 \times 10^{-1154}$). Median (IQR; range) isoform ratios were 0.49 (0.44, 0.54; 0.25, 0.84) for A/A (n = 444), 0.34 (0.30, 0.39; 0, 0.62) for A/G (n = 1,208), and 0 (0, 0; 0, 0.34) for G/G (n = 999). In contrast, rs12898397 was positively associated with *ULK3*-220 (ENST00000569437.5) ($\beta = 0.16$, $P = 1 \times 10^{-573}$), which lacks two amino acids ('VK') at the 5' end of exon 14. Median (IQR; range) isoform ratios were 0 (0, 0; 0, 0.011) for A/A, 0 (0, 0; 0, 0.49) for A/G, and 0.26 (0.20, 0.33; 0, 0.69) for G/G. (C) MR supports a causal association ($P = 1.8 \times 10^{-6}$) between reduced *ULK3*-201 level and DBP ($P = 1 \times 10^{-1154}$), consistent with the GWAS association of rs12898397-G ($P = 4 \times 10^{-51}$). (D) The sentinel irQTL (rs12898397) and eQTL (rs936228, $P = 1.4 \times 10^{-165}$) within *CYP1A1-ULK3* region are associated with both DBP (rs12898397: $P = 9.2 \times 10^{-47}$; rs936228: $P = 1.1 \times 10^{-73}$) and lymphocyte percentage (lymph_per, rs12898397: $P = 8.1 \times 10^{-15}$; rs936228: $P = 3.6 \times 10^{-10}$). All tests are two-sided t-test or z-test (for MR) in regression analyses. IQR, interquartile range; DBP, diastolic blood pressure; MR, Mendelian randomization. Gene names are italicized.

Figure 6. Allele-dependent splice donor usage associated with rs12898397 in *ULK3*. (A) The *ULK3* exon 14 region is shown in transcriptional orientation (5'→3') as the complementary positive-strand sequence, highlighting two isoform transcripts, *ULK3*-201 and *ULK3*-220. (B) MaxENT was applied to 9-nucleotide sequences surrounding rs12898397. Lowercase letters denote intronic sequence and capital letters exon sequences. The rs12898397-G allele was predicted to weaken the canonical splice donor site (gt: 7.93 → 4.44) and strengthen a cryptic donor site (GT: 2.98 → 7.20), supporting allele-dependent splice switching. The green color highlights the canonical and cryptic splice donor sites. (C) RNA-seq from three FHS subjects showed rs12898397 (A>G) genotype-dependent splicing: A/A (green) used the canonical donor, G/G (grey) used a cryptic donor skipping two codons (GTC-AAG), resulting in the absence of the valine-lysine ('VK') dipeptide, and A/G (pink) showed mixed splicing. Lowercase letters denote intronic sequence. (D) PyMOL visualization of AlphaFold 3.0-predicted C-terminal models of *ULK3* isoforms, with (173 amino acids in green color) and without (in orange color) the 'VK' dipeptide, showed substantial structural divergence when aligned. The yellow region in the model without 'VK' highlights the area without 'VK'. The Root-Mean-Square Deviation (RMSD), which quantifies the average atomic distance between superimposed protein structures, was 5.81 Å, indicating substantial structural divergence. In general, RMSD values above 2–3 Å suggest significant conformational differences between compared models. Gene names are italicized.

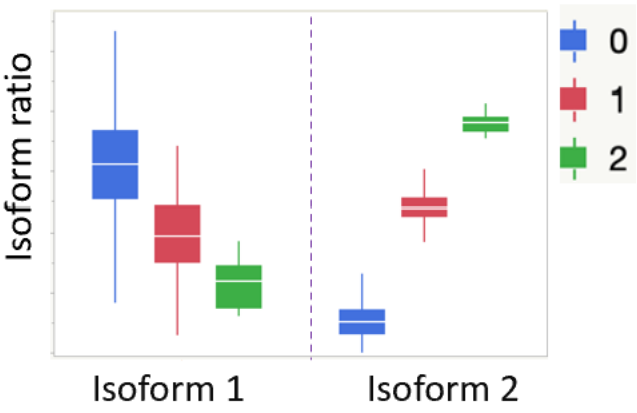
Figure 7. *CNN2* transcripts under genetic control. (A) Genome browser view of *CNN2* on the positive strand of chromosome 19 (hg38). Three *CNN2* isoforms, *CNN2*-201 (ENST00000263097.9), *CNN2*-202 (ENST00000348419.7), and *CNN2*-209 (ENST00000568865.3), differ by inclusion of Exons 5 and 6, Exon 6 alone, or

Exons 5, 6, and 7, respectively. (B) Isoform ratios of *CNN2*-201 and *CNN2*-202 are strongly associated with the sentinel *cis*-irQTL rs930232, while *CNN2*-209 is more strongly associated with a different *cis*-irQTL, rs5014188, in regression analysis. For rs930232, median (IQR; range) isoform ratios were as follows: for *CNN2*-201, 0.91 (0.90–0.92; 0.51–0.94) in G/G (n=781), 0.88 (0.86–0.89; 0.35–0.93) in A/G (n=1,309), and 0.84 (0.83–0.86; 0.40–0.90) in A/A (n=532); for *CNN2*-202, 0.015 (0.010–0.024; 0–0.33), 0.044 (0.033–0.057; 0–0.42), and 0.079 (0.065–0.095; 0.016–0.50), respectively; and for *CNN2*-209, 0.0058 (0.004–0.008; 0–0.017), 0.0052 (0.004–0.007; 0–0.019), and 0.0050 (0.004–0.006; 0–0.012). For rs5014188, median (IQR; range) isoform ratios were: for *CNN2*-201, 0.89 (0.88–0.90; 0.51–0.91) in T/T (n=52), 0.89 (0.86–0.90; 0.56–0.94) in C/T (n=534), and 0.88 (0.85–0.90; 0.35–0.94) in C/C (n=2,036); for *CNN2*-202, 0.026 (0.022–0.034; 0.012–0.33), 0.035 (0.020–0.054; 0.004–0.33), and 0.043 (0.026–0.069; 0–0.50), respectively; and for *CNN2*-209, 0.012 (0.0093–0.014; 0.0024–0.017), 0.0082 (0.0066–0.010; 0–0.019), and 0.0048 (0.0036–0.0060; 0–0.011). (C) rs930232 (G>A) is associated with reduced expression of *CNN2*-201 ($\beta = -0.032$, $P = 1 \times 10^{-532}$), which is causally linked to higher neutrophil percentage by Mendelian randomization ($\beta = -0.50$, $P = 2 \times 10^{-15}$) consistent with GWAS results showing a positive association between rs930232 and neutrophil percentage ($\beta = 0.021$, $P = 1 \times 10^{-21}$). (D) Predicted calponin-2 protein sequences show how Exons 5, 6, and 7 contribute to isoform structure. The three conserved 26-residue actin-binding domains are underlined in each isoform. All tests are two-sided t-test or z-test (for MR) in regression analyses. IQR, interquartile range; MR, Mendelian randomization. Gene names are italicized.

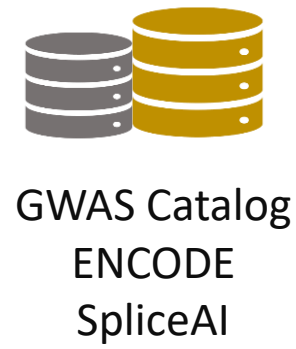
A. RNA-seq in discovery and replication samples



B. Analyses of irQTLs



C. Functional inference



D. irQTL, isoform ratio, and clinical traits: regression and MR analyses

