Uncovering splicing-mediated disease associations in the UK Biobank and All of Us using a TWAS with Bayesian joint fine-mapping

Alex Tokolyi[1,2], Saikat Banerjee[1], David A Knowles[1,2]

[1]New York Genome Center, New York City, NY, USA.

[2]Department of Computer Science, Columbia University, New York, NY, USA.

Corresponding author:

Alex Tokolyi (atokolyi@nygenome.org)

Abstract

Alternative splicing is a key mechanism by which genetic variation contributes to human phenotypic diversity and disease risk, yet its incorporation into large-scale genetic studies remains limited. Here we utilize blood RNA-seq from 4,732 European-ancestry individuals in the INTERVAL cohort to construct genetic scores for junction-based splicing phenotypes, 13,851 of which have external $R^2 > 0.01$ in held-out individuals. These models were imputed into the UK Biobank (UKBB; n=408,590) and All of Us (AoU; n=201,958), enabling splicing transcriptome-wide association studies (sTWAS) across 1,191 harmonized disease phenotypes. We identified 4,953 splicing-disease associations, of which 66% were shared across cohorts. To account for confounding by co-regulation and linkage disequilibrium, we used a novel Bayesian joint fine-mapping framework to prioritize 1,271 likely causal splicing-disease associations across 492 genes, revealing mechanisms underlying immune, vascular, and metabolic traits. Fine-mapped associations included exon-skipping events in *GSDMB* and splicing in untranslated regions in *STAT6* and *IL18RAP* for asthma, unannotated splicing events in *PIEZO1* and *PPP3R1* for vascular traits, and alternative exons in *BAK1* and *TNFRSF14* for celiac disease. Our study demonstrates that common genetic variation influencing splicing substantially shapes complex trait architecture and provides a scalable, statistically rigorous framework to uncover transcriptional disease mechanisms.

Main

Alternative splicing is a fundamental mechanism of eukaryotic gene regulation, contributing to transcriptomic and proteomic diversity by allowing a single gene to produce multiple RNA isoforms, as well as regulating gene dosage. While alternative splicing is pervasive across the human genome, with over 95% of multi-exon genes exhibiting evidence of splicing variation[1], the genetic architecture underlying this process is complex and highly context-dependent[2]. Splicing quantitative trait loci (sQTLs) capture common regulatory variants that influence splice junction usage, and are increasingly recognized for their contribution to disease risk, particularly when they modulate protein-coding potential, regulatory element inclusion, or dosage[3–5].

Large-scale efforts have catalogued sQTLs across tissues, with studies such as GTEx[2] and INTERVAL[4] identifying thousands of splicing events influenced by local (*cis*-) and distal (*trans*-) genetic variants. These insights have laid the groundwork for transcriptome-wide association studies (TWAS), which integrate genetic and transcriptomic data to assess whether the genetically regulated component of gene expression or splicing is associated with complex traits[6,7]. While expression-based TWAS have illuminated thousands of disease-associated loci[8–10], splicing has only recently begun to receive comparable attention. This gap is particularly notable given growing evidence that splicing variation is enriched for disease-associated loci and often provides independent or complementary signals to gene expression[3,11].

The incorporation of splicing into genetic prediction models and biobank-scale analyses presents both computational and methodological challenges. Splice junction phenotypes are higher-dimensional, more sparse, variable across individuals and tissues, and differ in their quantification approaches[3]**.** Additionally, predictive modeling of splicing phenotypes requires accounting for a genetic architecture which is less well studied than that of total gene expression. Despite these complexities, splicing is likely an important mediator of genetic risk, particularly for the discovery of causal genes due to those missed by expression-only analyses (false negatives) or attributed to incorrect transcriptional mechanisms (false positives).

Efforts to integrate splicing into genetic risk models have been limited in scale, particularly across diverse populations and biobanks. Given the scale and heterogeneity of human disease phenotypes, biobank resources such as the UK Biobank[12] (UKBB) and All of Us[13] (AoU) offer an unprecedented opportunity to explore the role of genetically regulated splicing in health outcomes across populations. To enable such

analyses, robust splicing genetic scores are needed that are accurate, generalize across cohorts, and are sufficiently scalable to handle tens of thousands of traits in hundreds of thousands of individuals.

Here, we develop and benchmark splicing genetic scores across 111,937 splicing phenotypes (both known and *de novo* detected) derived from whole blood RNA-seq in the INTERVAL[4] cohort (n=4,732 individuals). Using Bayesian ridge regression we generate polygenic scores for 13,851 splicing events demonstrating withheld predictive accuracy ($R^2 > 0.01$), and impute these into both the UKBB and AoU cohorts. Leveraging harmonized PheCodes[14,15] across biobanks, we perform a splicing TWAS on 1,191 disease phenotypes using case-control logistic regression, identifying thousands of splicing-disease associations and highlighting shared signals across genetically distinct populations. We then adapt a Bayesian joint fine-mapping framework[16] to resolve pleiotropic effects, integrating both our splicing and external gene expression scores[10] to prioritize likely causal splicing events **(Figure 1)**. Our results demonstrate that common variation modulating splicing contributes broadly to disease risk, revealing new insights into the functional basis of complex traits.

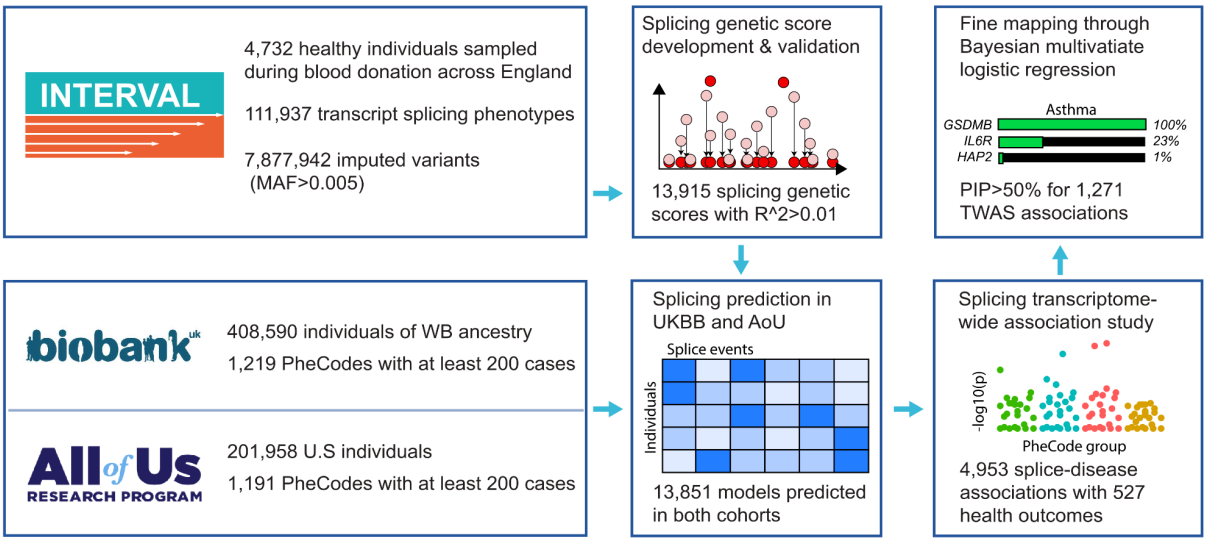

Figure 1. Overview of the data used for this study, along with the main analytical approaches.

Results

Creation of splicing genetic scores and biobank imputation

We obtained 111,937 splicing phenotypes and associated *cis*- and *trans*-sQTLs assayed from 4,732 European-ancestry individuals and associated imputed genotypes from the INTERVAL RNA-seq analysis[4]. We retained splicing events that were wholly contained within a gene body, were outside the extended HLA region[17], and had at least one significant sQTL for a non-ambiguous SNP. We utilized Bayesian ridge regression to train splicing genetic scores (**Methods**) as it has been shown to be robust for transcriptomic traits, as well as computationally scalable to the large number of traits and individuals selected here[10]. We observed large variation in predictive accuracy (R^2), likely reflective of varying heritability, across splicing genetic scores (**Figure 2A, Table S1**), with 14,783 splicing phenotypes with mean internal R^2>0.01, 3797 with R^2>0.1, 1118 with R^2>0.3, and 422 with R^2>0.5 (**Figure 2B**). Additionally, the inclusion of *trans*-sQTLs led to an additional 322 splicing events with an internal R^2>0.01 **(Figure 2C)**, and an increase in R^2 for 739 others. Using a withheld 20% subset of the INTERVAL cohort, and the combined *cis* and *trans* models, 13,851 splicing genetic scores (in 4,137 unique genes) had a mean withheld test-set R^2>0.01 and passed our inclusion criteria (**Methods**): we chose these to predict in the UK Biobank and All of Us (**Figure 2D**). Imputed splicing events were predominantly within protein coding genes and mostly previously annotated, though 30.7% had a novel donor or acceptor site (**Figure 2E, Table S2**). Just over a third (35.3%) excised a protein coding portion of the transcript, leading to a higher likelihood of downstream proteomic consequences. Disease phenotypes were mapped to PheCodes[15] (**Methods**), selecting for those with a minimum of 200 cases (n=1,191 PheCodes) in both the UK Biobank and All of Us cohorts (**Figure 2F, Table S3)**. Due to differences in health systems and EHR records, there was a large variation in case/control ratio, with that of the All of Us cohort being a median 5.5x higher than that of the UK Biobank **(Figure S1)**, consistent with previous observations[18].

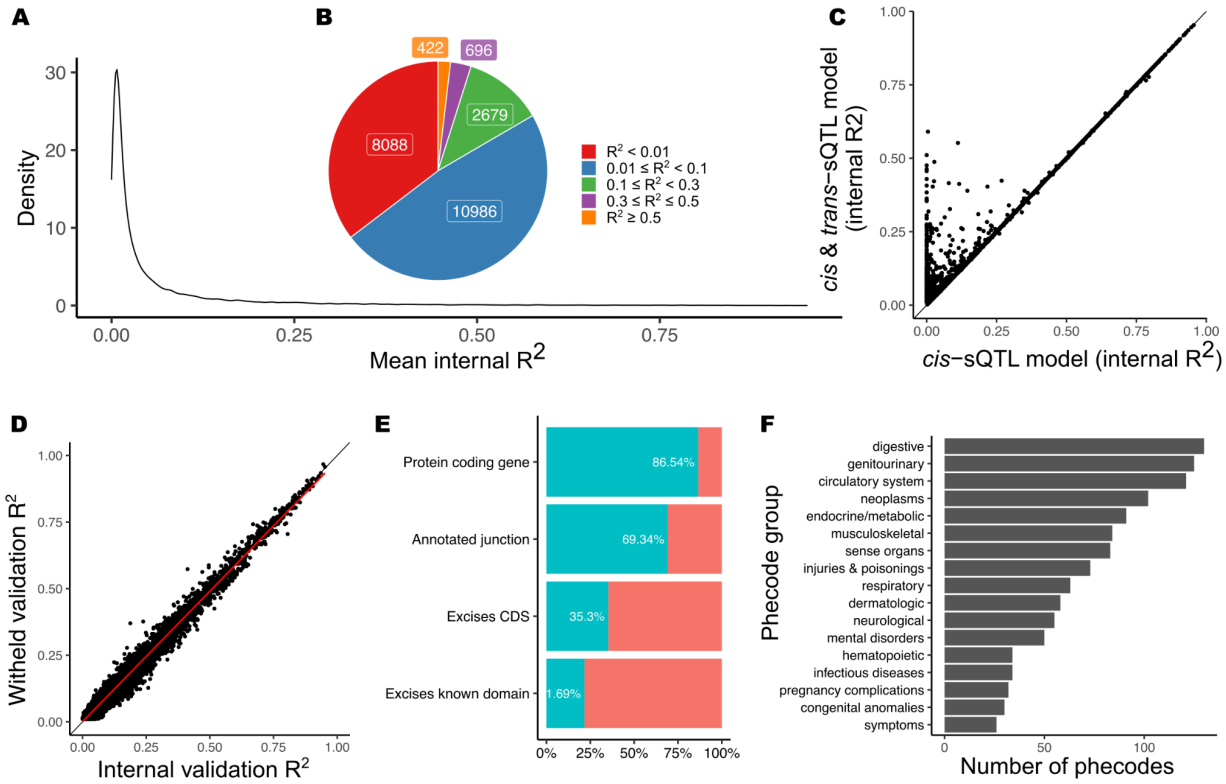

Figure 2. Validation of genetically predicted splicing and selected PheCodes for sTWAS analysis. A) Density plot showing the mean internal R^2 for all tested splicing genetic scores, with an inset **B)** pie chart showing the quantity of splicing scores at various ranges of R^2. **C)** The relationship between internal R^2 scores between a *cis*-sQTL only model (x) and a model which additionally includes *trans*-sQTLs (y). **D)** Alignment of internal validation R^2 (x) with that computed on the withheld test subset of the INTERVAL cohort (y), black line equates to y=x, red line is the slope of y~x. **E)** Overlap of the selected splicing events with main transcriptomic and proteomic features (overlap in blue, percent inset in white). **F)** Quantity of filtered PheCodes by group classification.

Splicing TWAS reveals splicing-disease associations

Using our 13,851 splicing genetic scores, we performed a TWAS on 1,191 PheCodes in both the UK Biobank and All of Us cohorts using a case/control logistic regression (**Methods**). At an FDR of 0.05 we detected 2,374 significant associations between 1094 splicing events and 256 PheCodes in the UK Biobank (n=408,590), and 594 significant associations between 337 predicted splicing events and 155 PheCodes in All of Us (n=201,958), likely owing to the reduced sample size. To assess the sharing of sTWAS results between cohorts we performed multivariate adaptive shrinkage (mash) between the

summary statistics of both cohorts (**Methods**), revealing that 66.3% of associations between the two cohorts were shared (within a factor of 0.5) (**Figure 3B, Table S4)**, with the remainder predominantly possessing a larger effect in the UKBB. This level of sharing is remarkably high, given that there were both differences in sampled populations and measurement/diagnosis between the two biobanks.

Given the high levels of sharing, we performed a fixed-effects meta analysis on the sTWAS results from both cohorts, yielding 4,953 significant associations (FDR<0.05), between 2,118 splicing events (in 1012 genes), with 527 diseases **(Figure 3A, Table S5)**. Neoplasms made up a large proportion of the significant sTWAS results, so we assessed the number of associations in different broad cancer types (**Figure 3C**). We observed that while female genitourinary cancers were the most commonly tested neoplasm group, those relating to skin and soft tissue cancers were the most likely out of the cancer PheCodes to have a significant association (91.7%). We compared our sTWAS to the current largest breast cancer sTWAS[19] which included breast cancer RNA-seq samples from 652 individuals and a case-control cohort of 426,834 individuals (n=178,534 cases). Our sTWAS revealed associations with breast cancer for splicing in 23 genes, 5 of which (*KANSL1, LSP1, CASP8, FDPS, SCAMP3*) were also observed in their analyses (p=2.21e-06), and 19 of which were collated by the authors as previous expression TWAS associations for breast cancer (p=2.68e-25).

We then sought to analyse the characteristics of splicing events that had a significant sTWAS association, using the 13,851 tested events as a background set. These were annotated using the Leafidex pipeline **(Methods)**, which detects overlaps with transcriptomic features from GENCODE[20], and proteomic features from UniProt[21] and Pfam[22](**Table S2)**. Of the 15 characteristics tested, 7 were significant, relating to features of the encoded protein domains, as well as excision events within the 5' UTR, potentially corresponding to effects on translation efficiency **(Figure 3D).** The three strongest enrichments were for UniProt localization domains (extracellular, cytoplasmic, and transmembrane), suggesting that disruption of such domains is a recurring mechanism by which splicing changes impact protein function. To determine potential global mechanisms for the genetic variants underlying disease risk through splicing, we analysed SNPs present in the models of at least one significant splicing-disease association with ENCODE RNA binding protein (RBP) data[23]. SNPs that contributed to splicing phenotypes that were significant in our TWAS were 26% more likely to be contained within an RBP peak (p=1.03e-8, OR=1.26, 95% CI: 1.17-1.37, Fisher's exact test). We additionally assessed VEP variant consequences for these SNPs, observing 8 other significantly enriched variant consequences (FDR<0.05, **Figure S2).**

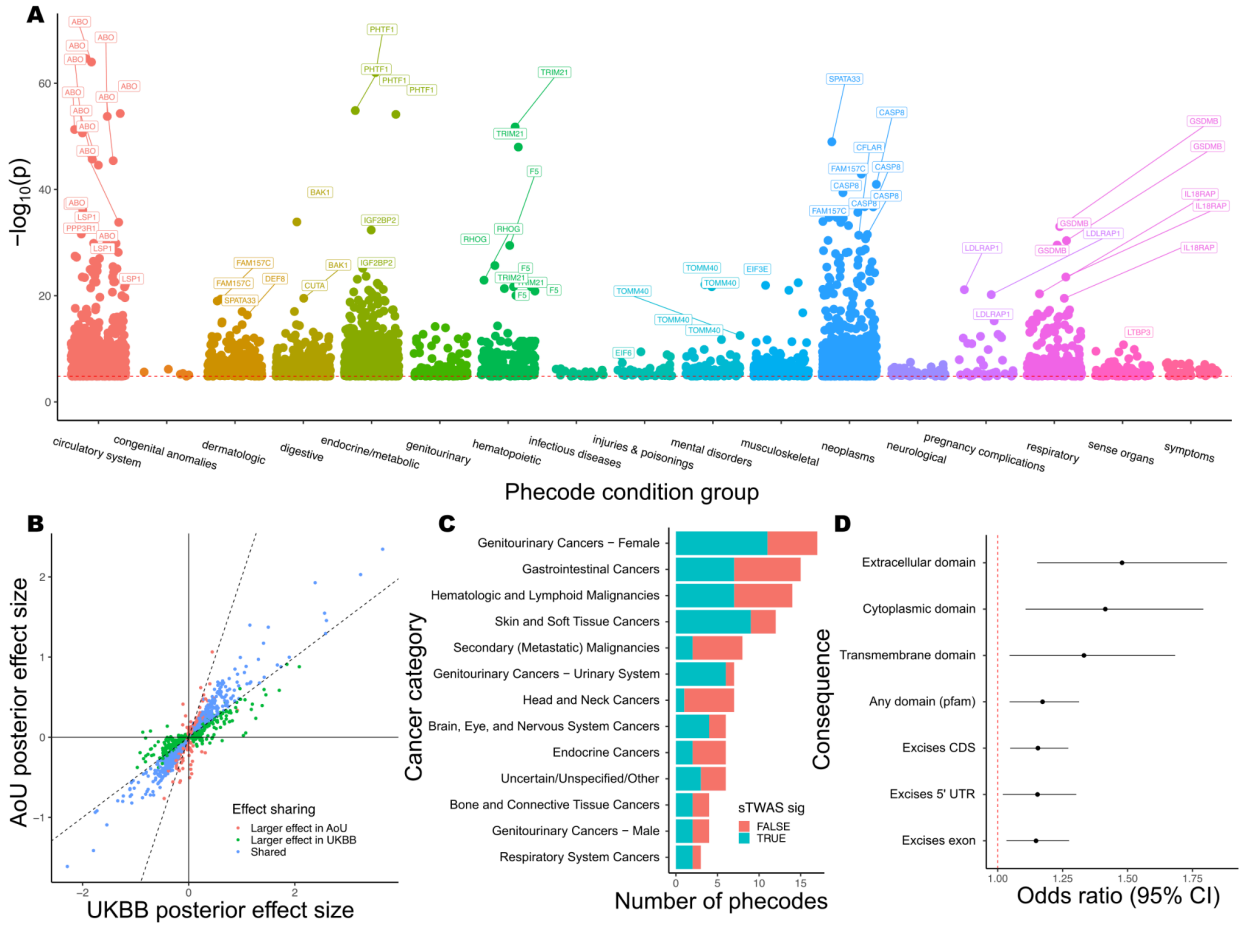

Figure 3. Summary of splicing TWAS results between UKBB and AoU. A) Manhattan-style plot of the meta-analysis sTWAS between UKBB and AoU, grouped and coloured by PheCode group (x) and significance (y), with the gene names of the most significant splicing events labelled. **B)** mash analysis of the effect sharing of sTWAS results, showing the UKBB (x) and AoU (y) posterior effect sizes. These are coloured by the effect overlap between the cohorts. **C)** Quantity of cancer PheCode groups which have a significant sTWAS association. **D)** Enrichment of annotations for splicing events with a significant sTWAS association.

Transcriptome fine-mapping of splicing TWAS results with Bayesian joint analysis

The basic TWAS approach above, where we test each splicing event-trait pair separately, is known to be confounded by co-regulation/pleiotropy and linkage disequilibrium[24,25]. As the sTWAS used a univariate logistic regression model testing only a single predicted splicing event and PheCode, these were regularly (308/527 PheCodes) associated with multiple predicted splicing events within the same linkage

disequilibrium (LD) block. Total gene expression is also tightly linked with splicing, due to the specific upregulation of individual transcript isoforms or splicing-induced nonsense mediated decay. Thus splicing and expression phenotypes for a gene may be co-regulated/show pleiotropy. To untangle these effects and prioritise individual splicing events, we adapted a Bayesian joint logistic regression model previously developed for GWAS fine-mapping (B-LORE[16]) to TWAS traits (**Methods**). We emphasize that the covariates are splicing events, rather than SNPs as in typical "fine-mapping", and the resulting posterior inclusion probabilities (PIP) represent the evidence for a causal link from splicing event to trait. To allow for prioritization between splicing and total gene expression, including that of nearby (and potentially *cis*-associated) genes, we imputed the gene expression genetic scores developed by Xu *et. al.*[10] into the UK Biobank and All of Us cohorts using the same methodology as for splicing. For each significant splicing-PheCode association, we aggregated all predicted splicing and gene expression scores located within the same LD block (**Methods**) and fit a Bayesian joint logistic regression model for each significant PheCode. A Bayesian meta-analysis was then performed across the two cohorts to compute a meta PIP per splicing-PheCode pair. In total, we observed 1,271 pairs with PIP>50% for 775 unique splicing events (in 492 genes) and 192 disease PheCodes (**Table S5**). The PIPs exhibited a bimodal distribution, with most values clustering near 0 or 1, indicating that the model confidently distinguished between non-causal and likely causal features **(Figure 4A)**.

To investigate the causal source of the detected sTWAS signals when the associated splicing event was non causal (PIP<0.5), we collected the gene expression and splicing events within the LD block that were allocated causal status (PIP>0.5). This was stratified into two gene expression categories (expression of the spliced gene, or another gene in the LD block), and 3 splicing categories (splicing event in same cluster, same gene, or in another gene in the LD block). Of the 4,720 events for which B-LORE converged, 1,271 (26.9%) had a PIP>0.5 for the original significant sTWAS splicing event, 1,380 (29.2%) had no detected causal source in the LD block, and for 2069 (43.8%) B-LORE suggests a different transcriptomic event as the causal mechanism **(Figure 4B)**. Of the latter category, the estimated cause was most commonly a different splicing event within another gene in LD (94.3%) and interestingly least often to be expression of the containing gene (10.2%, **Figure 4C**).

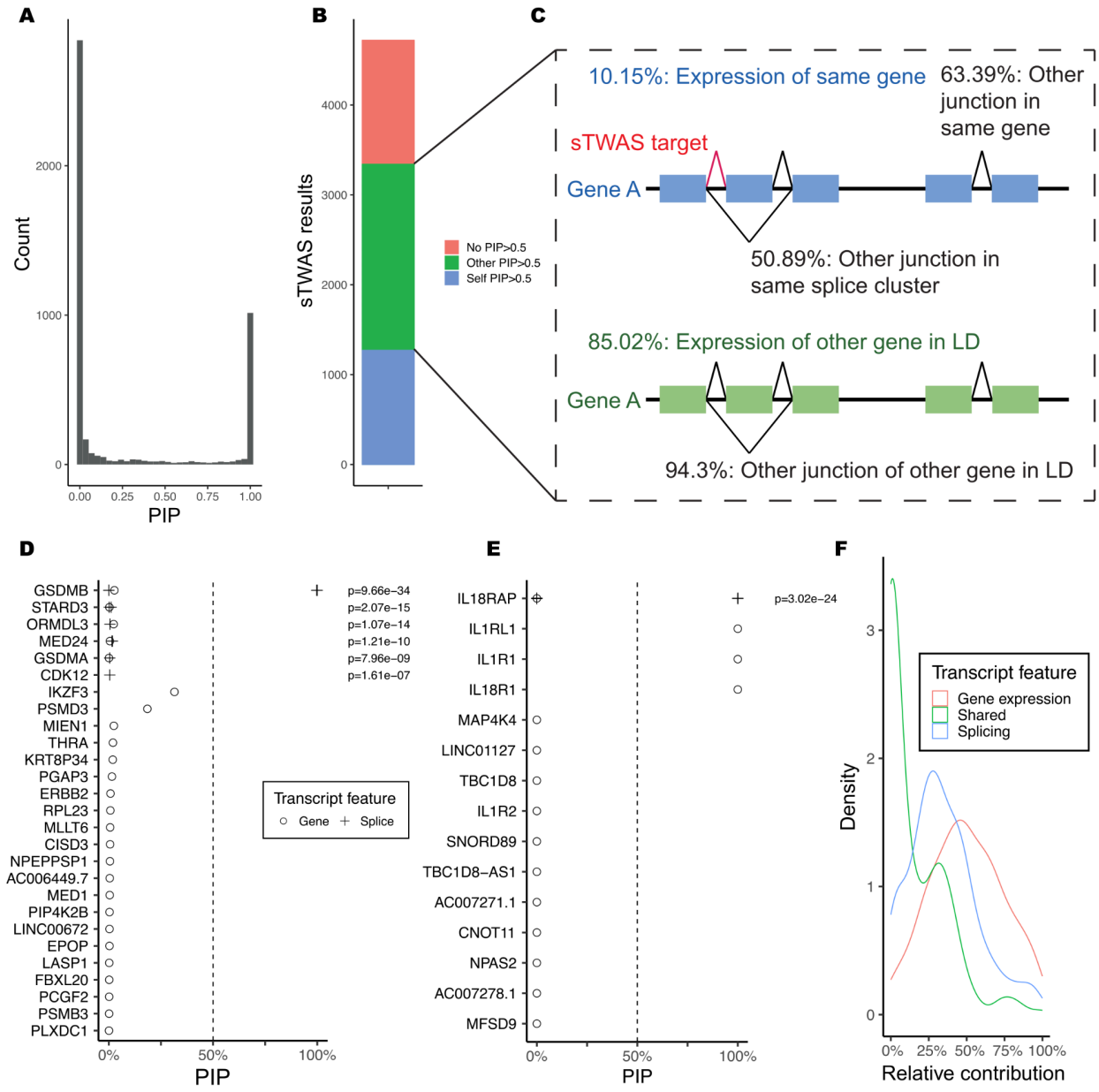

Figure 4. Bayesian fine-mapping of splicing TWAS results reveal causal transcriptomic sources. A) Distribution of PIP scores for significant sTWAS splicing events. **B)** Amount of sTWAS results with a causal PIP for the tested splicing event ("self PIP", blue), for another transcriptomic source in the LD block ("other PIP", green), or no detected causal source ("no PIP", red). **C)** Schematic of other causal transcriptomic sources within the LD block of fine-mapped sTWAS results, either within the same gene (blue) or another gene in the LD block (green). **D)** Tower plot showing the sTWAS fine-mapping PIPs for transcriptomic events (splicing events as "+", gene expression as "◯") for the LD block containing *GSDMB* with asthma risk, with p-values annotated for sTWAS significant splicing events. **E)** Tower plot

as previously for the LD block containing *IL18RAP* with asthma risk. **F)** The relative independent and shared contribution (green) of significant splicing events (blue) and gene expression (red) to tested PheCodes, with clamping (negative fractions set to zero and sum renormalized to one).

Splicing mechanisms underlying immune, vascular and metabolic traits

Fine-mapped splicing-disease associations revealed mechanisms across immune-mediated, vascular, and metabolic traits among others, consistent with the broad relevance of the blood origin of our transcriptome data. For asthma, we identified 30 splicing events across 18 genes with meta PIP>0.5. An exon 5 skipping event in *GSDMB*, which abolishes the pyroptosis-inducing domain, was associated with reduced risk ($p=9.7e10^{-34}$, OR=0.81, PIP=1), and the associated locus was causal only for splicing **(Figure 4D)**, consistent with prior functional evidence[26]. In *STAT6*, opposing effects were observed for splicing at the untranslated regions: excision of the 5′ UTR reduced risk ($p=4.6e10^{-18}$, OR=0.81, PIP=1), whereas a 3′ UTR event increased risk ($p=1.44e^{-12}$, OR=1.26, PIP=1). At the *IL18RAP* locus, previously implicated in asthma without a straightforward mechanistic explanation[27], we identified a 5′ UTR splicing event that decreased risk ($p=3.0e10^{-24}$, OR=0.86, PIP=1), providing a putative mechanistic interpretation of the GWAS signal. We additionally observed causal signals for total gene expression in 3 nearby interleukin genes (*IL1RL1*, *IL1R1* and *IL18R1*), suggesting a mechanism that interacts through multiple causal signals in both splicing and total expression at this locus **(Figure 4E)**.

Among vascular traits, we observed 1,188 significant splicing associations. For varicose veins, an unannotated splicing event in the 3′ UTR of *PIEZO1* decreased risk (meta $P=3.5e10^{-22}$, OR=0.53, PIP=1, **Figure S3**), complementing rare coding variant evidence at this gene[28]. At *PPP3R1*, two partially-annotated splicing events with opposing directions of effect were nominated as causal, with higher predicted expression of an annotated intronic splicing event decreasing risk ($p=5.86e^{-28}$, OR=0.77, PIP=0.97, **Figure S4**), while the other event, with a shared donor and novel acceptor, increased risk ($p=9.56e^{-28}$, OR=1.26, PIP=0.95). For coagulation traits, seven unannotated but in-frame splicing events within exon 13 of *F5*, encoding the B domain, were predicted to be causal with disorders including thrombophlebitis (min $p=1.48e^{-16}$, OR=0.84, PIP=1, **Figure S5)** and pulmonary embolism. These events all excised segments divisible by three nucleotides, consistent with isoform production that may act analogously, though more subtly, to the FV-short isoform[29,30].

For metabolic and immune traits, splicing highlighted additional mechanisms. In celiac disease, a novel splicing event truncating the final coding exon of *BAK1* increased risk (meta $P=1.3e10^{-34}$, OR=3.40,

PIP=1, **Figure S6**), consistent with modulation of pro-apoptotic activity[31]. We additionally observe a splicing event in *TNFRSF14* (encoding the pro-apoptosis protein LIGHT[32]), with an annotated donor and novel acceptor excising part of coding exon 7, containing the cytoplasmic domain, as decreasing the risk of celiac disease ($p=2.2e^{-6}$, OR=0.71, PIP=1, **Figure S7**). Hypothyroidism was causally associated with 78 splicing events across 46 genes. Two alternative 5′ UTR splicing events in *PHTF1* displayed opposite effects on hypothyroidism risk ($p=2.5e^{-61}$ & $7.3e^{-55}$; OR=1.74 & 0.53; PIP=1 & 1, **Figure S8**), potentially acting through altered translational regulation. Together, these examples illustrate how common splicing variation refines the mechanisms underlying GWAS loci across diverse trait classes, with full associations detailed in **Supplementary Table S5**.

Global comparison of expression- and splicing-based contributions to disease risk

To assess the relative contribution of splicing and gene expression to disease risk, we jointly modelled significant traits from the expression TWAS and splicing TWAS for each disease within UKBB. Across 204 PheCodes with both significant expression and splicing associations, we decomposed the unique and shared effects of each modality (**Figure 4F, Table S6)**. On average, gene expression explained 48.9% of the transcriptomic contribution, splicing accounted for 31.7%, and a shared component contributed 5.4%, though proportions varied widely across traits **(Figure S9)**. These results indicate that splicing contributes substantially and independently to disease architecture, complementing gene expression and highlighting the value of modelling both sources of transcriptomic variation.

Discussion

Our study contributes to the growing evidence that splicing is a major and previously underappreciated component of the genetic architecture of complex disease. By developing splicing genetic scores, performing a large-scale splicing TWAS across both the UK Biobank and All of Us cohorts, and applying a novel Bayesian joint fine-mapping approach, we identified thousands of splicing-disease associations with high posterior probability of causally impacting disease risk. Integrative modelling of splicing and expression further demonstrated that splicing independently accounts for on average nearly one third of the transcriptomic contribution to disease risk, independent of gene expression. These findings establish splicing as a complementary and substantial dimension of genetic regulation, extending beyond expression-based models, and refining mechanistic interpretation of GWAS loci.

Using our splicing features derived from whole blood, we were able to recapitulate both existing findings, such as *GSDMB* exon 5 splicing in epithelial cells impacting asthma risk[26], multiple genes in a recent breast-cancer sTWAS[19], as well as many new examples where splicing may explain GWAS risk loci without established mechanisms. Additionally, we note the functional importance of splicing annotations in deciphering their risk, through the observation of enrichment of the excision of localization signals and 5' UTR splicing in disease. Therapeutically, targeting of these specific splice junctions may provide more context specific effects than full gene knockouts, increasing specificity and decreasing the risk of off-target effects.

Previous TWAS studies have focussed almost exclusively on total gene expression. We demonstrate that the addition of splicing to these analyses adds additional, unique information to that of gene expression, and links to the growing recognition of post-transcriptional regulation in disease. The current study has several limitations. While we included splicing features quantified *de novo* to allow for those that are less common, our features do not represent the totality of the transcriptome. For example, the methods used did not consider alternative transcription start and end sites, which may also act as a mechanism through which genetic variation can exert its effect. Additionally while we benefitted from an increase in statistical power due to the meta-analysis of the UKBB and AoU, due to the different selection strategies of the cohorts and the difference in healthcare systems between countries, some effects may have been obscured. Our reliance on healthy whole blood to compute splicing genetic scores, while possessing many shared genetic effects with other tissues, may miss splicing effects which are tissue-specific in distal tissues, or disease-specific splicing events. Additionally due to the nature of the short-read sequencing from which our splicing features were derived, we lack the full isoform context of these excision events, particularly in the case of unannotated junctions. Future experiments with long-read sequencing will be needed to both assign the disease-associated splicing events to their effector isoforms, as well as to discern the impact of alternative 5' and 3' UTR excision on downstream molecular phenotypes such as translation efficiency and RNA stability.

While our splicing genetic scores were derived from individuals of European ancestry, we also chose to subset the UK Biobank to European-ancestry individuals, while retaining the comprehensive genetic diversity captured in the All of Us cohort, owing to its greater representation of diverse ancestries. Though we saw high sharing of associations between the cohorts, we expect our scores to translate less well across non-European genetic ancestries given that SNPs in LD with causal SNPS, rather than just those that are causal alone can be used by the Bayesian ridge model to compute the splicing scores. Increased availability of RNA-seq from diverse individuals will help to produce genetic scores that are

more generally applicable to a broader range of individuals, as well as prioritize causal variants and transcriptomic phenotypes.

In summary, our results highlight splicing as a pervasive and independent mediator of genetic risk, complementing gene expression in shaping the transcriptomic basis of complex traits. Through fine-mapping splicing-disease associations and by jointly integrating splicing with expression across large biobank cohorts, we provide mechanistic resolutions for GWAS loci and demonstrate the utility of modelling multiple transcriptomic modalities in parallel. These insights underscore the importance of incorporating splicing into future large-scale functional genomics efforts, and suggest that splicing-aware approaches will be essential for advancing disease biology, therapeutic target discovery, and precision medicine.

Data availability

The INTERVAL study data used in this paper are available to bona fide researchers from ceu-dataaccess@medschl.cam.ac.uk. The data access policy for the data has been approved by the ethics committee and is available at https://www.donorhealth-btru.nihr.ac.uk/research/blood-donors-bioresource/. This study used data from the All of Us Research Program's Controlled Tier Dataset version 7, available to authorized users on the Researcher Workbench (https://researchallofus.org/). UK Biobank data are available (http://www.ukbiobank.ac.uk/) upon application and with permission of UKBB's Research Ethics Committee. This research was conducted using the UKBB Resource under application no. 47976. Total gene expression scores were downloaded from the OmicsPred portal (https://www.omicspred.org/). ENCODE RNA binding protein eCLIP data was downloaded from the ENCODE portal (https://www.encodeproject.org/eclip/). Owing to their large size and data use agreements, unfiltered splicing TWAS summary statistics and predicted transcript splicing phenotypes and models will be returned to and made available through the respective cohort repository/workbench (UKBB and AoU) upon acceptance of the corresponding peer-reviewed manuscript.

Code availability

All original code from this analysis has been deposited to GitHub (https://github.com/atokolyi/sTWAS_paper_scripts).

Acknowledgements

We gratefully acknowledge All of Us participants for their contributions, without whom this research would not have been possible. We also thank the National Institutes of Health's All of Us Research Program for making available the participant data examined in this study. We would also like to thank the participants of the UKBB and INTERVAL cohorts for their valuable contributions. This research was conducted using the UKBB Resource under application no. 47976. This work uses data provided by patients and collected by the NHS as part of their care and support. Additionally, this work was made possible by support from the MacMillan Family and the MacMillan Center for the Study of the Non-Coding Cancer Genome at the New York Genome Center.

Author contributions

A.T conducted the statistical analyses, created figures, and wrote the paper. A.T, D.A.K, and S.B conceived the study design. S.B assisted in the implementation of B-LORE. D.A.K supervised the project. All authors reviewed and approved the final version of the paper.

Methods

Splicing data generation (INTERVAL)

The INTERVAL study consists of approximately 50,000 healthy adult blood donors recruited through England's National Health Service Blood and Transplant (NHSBT). Genotypes and splicing quantification from a subset of these individuals (n=4,732) was obtained from the INTERVAL consortium as previously described[4]. Briefly, 830,000 variants were assayed on the Affymetrix Axiom UK Biobank genotyping array, phased using SHAPEIT3, and imputed using a combined 1000 Genomes Phase 3-UK10K reference panel. Imputation was performed via the Sanger Imputation Server (https://imputation.sanger.ac.uk) resulting in 87,696,888 imputed variants and lifted over to reference build GRCh38. Whole blood RNA was sequenced with 75 bp paired-end sequencing reads (reverse stranded) on a NovaSeq 6000 system (S4 flow cell, Xp workflow; Illumina). Reads were aligned to the GRCh38 human reference genome (Ensembl GTF annotation v99) using STAR (v2.7.3a). The STAR index was built against the GRCh38 Ensembl GTF v99 using the option -jdbOverhang 75. STAR was run

in a two-pass setup with standard ENCODE options to increase mapping accuracy and detect novel junctions. Splice junctions were extracted from aligned RNA-seq BAMs for the 4,732 individuals using Regtools (v0.5.2) junctions extract. Introns represented by extracted splice junctions were then clustered into groups using the Leafcutter[33] pipeline (v0.2.9). Clustered introns were then prepared for sQTL analysis using "prepare_phenotype_table.py" to convert intron counts to normalized ratios. Total observed introns were filtered to those that were autosomal and overlapping an expressed gene body, with CPM > 0.5 in at least 24 individuals, and sufficient variance (minimum two filtered splicing event phenotypes per cluster), resulting in 111,937 filtered splicing event phenotypes in 11,016 genes. Splicing events were then annotated for transcriptomic and proteomic features using the Leafidex tool (https://github.com/atokolyi/Leafidex).

Splicing genetic score training & evaluation

The full INTERVAL cohort was subset in to an 80% train subset (n=3,786) and a 20% withheld test subset (n=946), and sQTLs re-called with tensorQTL[34] on the train subset in *cis* (+/-500kb from the center of the splice site, minor allele frequency (MAF)>0.005, FDR<0.05) and *trans* (genome-wide, MAF>0.005, Bonferroni corrected). Both *cis* and *trans* analyses used previously defined covariates[4], namely demographic variables, technical variables, and genotype and splicing PCs. For genetic score training *cis*- or *trans*-significant sQTLs were filtered to those that were regulating splice junctions outside the extended MHC region[17] and were strictly unambiguous, bi-allelic single-nucleotide polymorphisms, resulting in 1,017,820 unique SNPs. To minimise redundant information and speed up computation, these SNPs were pruned with plink2[35] indep-pairwise using a window size of 1000kb and an R^2 threshold of 0.8, resulting in 234,889 pruned SNPs associated with 22,871 splice junctions for model training.

Splicing phenotypes for the 22,871 splice junctions were residualized using the above sQTL covariates. Bayesian Ridge (BR) regression was used in python (3.13.1) with sklearn[36] (1.6.1) "BayesianRidge" to re-estimate sQTL effect betas, using uninformative priors (10e-5) as these have been previously demonstrated as optimal for the prediction of transcriptomic traits[10]. These residualized splicing genetic models were validated using 5-fold cross validation on the 80% subset (internal validation), with and without the addition of trans-sQTL SNPs to evaluate the added benefit of their inclusion. The combined *cis* & *trans* sQTL models were then applied to the 20% withheld subset to evaluate the trained models on unseen samples (withheld test subset validation). Of the 22,871 tested splice junctions, 13,915 had a mean withheld test set R^2>0.1, and of these 13,851 were imputable in both the UKBB and AoU due to SNP overlap.

UK Biobank data processing & imputation

To provide a comparable genetic background to that of INTERVAL, we restricted UKBB to the white-British (WB) subset, and individuals whose reported sex matched the genetically-inferred sex (n=408,590) to prevent ambiguity in sex-specific PheCode analysis, using the post-quality controlled genotype data[12]. Health PheCodes were extracted from the UKBB using the University of Michigan Center for Precision Health Data Science "createUKBphenome" pipeline (https://github.com/umich-cphds/createUKBphenome). PheCodes allow for the collapsing ICD codes into phenotypes of interest, to define individuals with a shared health phenotype, as well as provide related phenotypes to exclude from the control population for phenome-wide association testing[14,15]. We restricted the analysis to PheCodes with at least 200 cases, resulting in 1,219 tested phenotypes. To impute the residualized splicing phenotypes into UKBB, we retrained the BR models for the 13,915 phenotypes passing withheld test set validation on the full INTERVAL cohort (n=4,732) using the subset of SNPs overlapping with UKBB (99.51%) and re-pruned as above. These trained models were then used for prediction with the UKBB genotypes, and the output predicted splicing phenotypes were residualized against the top 17 UKBB genotype PCs to minimize the effects of differing population structure on subsequent results.

All of Us data processing & imputation

Although our models were trained in European-ancestry individuals, we applied them to the full All of Us v7 cohort[13] to assess the broader generalizability of our results. The analysis was restricted to individuals who had both electronic health records (EHR) and short-read WGS data, and individuals whose reported sex matched the genetically-inferred sex (n=201,958). PheCodes were mapped from ICD9-CM and ICD10-CM using a custom pipeline referencing the PheWAS catalog database[15] as above (version 1.2), using the 1,219 qualifying PheCodes from the UKBB filtering, and further subsetted for at least 200 cases in AoU (final n=1,191). Splicing phenotypes were imputed using BR models retrained on the full INTERVAL cohort using all SNPs (as AoU used short-read WGS, 100% of INTERVAL SNPs were retained) and re-pruned as above. Predicted splicing events were then residualized against the top 17 genotype PCs calculated from the short-read WGS data to minimize the effects of differing population structure on the results. As shown in **Figure S10**, the resulting p-values in the full cohort are generally lower (more significant) than when restricting to European-ancestry individuals, reflecting increased sample size combined with shared predictive signal across ancestries.

Splicing transcriptome-wide association study

Next, a splicing TWAS was performed independently in UKBB and AoU due to their separate research analysis platforms by associating the 1,191 PheCodes with 13,851 predicted residualized splicing phenotypes which could be imputed in both cohorts, using a logistic regression model with statsmodels[37] (0.14.4) in python (3.13.1), with sex and age as covariates, resulting in 16,496,541 tests. Analyses were restricted to be sex-specific when required, as denoted by the PheWAS catalog database. Test statistics and p-values were extracted, and p-values FDR corrected. A fixed effects meta-analysis was then conducted in R (4.4.2) to combine the results from both cohorts.

Joint Bayesian fine-mapping of sTWAS results with total gene expression

As TWAS effects may be in-part driven by linkage-disequilibrium and pleiotropy effects, we conducted a joint fine-mapping of splicing events by adapting a Bayesian multivariate logistic regression framework (B-LORE)[16]. We opted to adapt B-LORE over existing tools such as FOCUS[25] which rely on summary statistics, as our sTWAS approach uses the full matrices of predicted splicing and gene features, and PheCode case/control status, to avoid the unnecessary computation of full summary statistics for each PheCode. Previously published total gene expression genetic scores[10] were used to impute gene expression features in the UKBB and AoU samples in the same methodology as for splicing events. Splicing and gene expression features were assigned to previously defined LD blocks[38]. A joint B-LORE model was run separately for each PheCode and cohort, for all significant PheCodes from the sTWAS meta-analysis, using all gene expression and splicing features contained in an LD-block together as a joint group, with age and sex as covariates and default parameters. A B-LORE meta-analysis on the results from both the UKBB and AoU was then performed to compute posterior inclusion probabilities (PIPs) for each splicing-PheCode effect using zmax=10 (maximum number of causal features per LD block), nominating causal results as those with a meta PIP>0.5. The majority (79%) of LD block-PheCode pairs had less than 10 nominated causal transcriptomic features, providing an appropriate upper bound to balance excessive computation and prioritise stronger effects downstream.

References

1. Frankish, A. *et al.* GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Res.* **51**, D942–D949 (2023).
2. GTEx Consortium *et al.* A vast resource of allelic expression data spanning human tissues. *Genome Biol.* **21**, 234 (12/2020).
3. Li, Y. I. *et al.* RNA splicing is a primary link between genetic variation and disease. *Science* **352**, 600–604 (2016).
4. Tokolyi, A. *et al.* The contribution of genetic determinants of blood gene expression and splicing to molecular phenotypes and health outcomes. *Nat. Genet.* 1–10 (2025) doi:10.1038/s41588-025-02096-3.
5. Najar, C. F. B. A. *et al.* Genetic and functional analysis of unproductive splicing using LeafCutter2. *Genomics* (2025).
6. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).
7. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).
8. Rowland, B. *et al.* Transcriptome-wide association study in UK Biobank Europeans identifies associations with blood cell traits. *Hum. Mol. Genet.* **31**, 2333–2347 (2022).
9. Qi, G. *et al.* Transcriptome-wide association studies at cell state level using single-cell eQTL data. *Genetic and Genomic Medicine* (2025).
10. Xu, Y. *et al.* An atlas of genetic scores to predict multi-omic traits. *Nature* **616**, 123–131 (2023).
11. 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).
12. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

13. All of Us Research Program Genomics Investigators. Genomic data in the All of Us Research Program. *Nature* **627**, 340–346 (2024).
14. Denny, J. C. *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* **31**, 1102–1110 (2013).
15. Bastarache, L. Using phecodes for research with the electronic health record: From PheWAS to PheRS. *Annu. Rev. Biomed. Data Sci.* **4**, 1–19 (2021).
16. Banerjee, S., Zeng, L., Schunkert, H. & Söding, J. Bayesian multiple logistic regression for case-control GWAS. *PLoS Genet.* **14**, e1007856 (2018).
17. Horton, R. *et al.* Gene map of the extended human MHC. *Nat. Rev. Genet.* **5**, 889–899 (2004).
18. Zeng, C. *et al.* Comparison of phenomic profiles in the All of Us Research Program against the US general population and the UK Biobank. *J. Am. Med. Inform. Assoc.* **31**, 846–854 (2024).
19. Ping, J. *et al.* Multi-ancestry transcriptome-wide association studies uncover new insights into breast cancer genetics and biology. *medRxiv* 2025.08.12.25333509 (2025) doi:10.1101/2025.08.12.25333509.
20. Mudge, J. M. *et al.* GENCODE 2025: reference gene annotation for human and mouse. *Nucleic Acids Res.* **53**, D966–D975 (2025).
21. UniProt Consortium. UniProt: The universal protein knowledgebase in 2025. *Nucleic Acids Res.* **53**, D609–D617 (2025).
22. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).
23. Van Nostrand, E. L. *et al.* A large-scale binding and functional map of human RNA-binding proteins. *Nature* **583**, 711–719 (2020).
24. Wainberg, M. *et al.* Opportunities and challenges for transcriptome-wide association studies. *Nat. Genet.* **51**, 592–599 (2019).
25. Mancuso, N. *et al.* Probabilistic fine-mapping of transcriptome-wide association studies. *Nat. Genet.* **51**, 675–682 (2019).

26. Panganiban, R. A. *et al.* A functional splice variant associated with decreased asthma risk abolishes the ability of gasdermin B to induce epithelial cell pyroptosis. *J. Allergy Clin. Immunol.* **142**, 1469–1478.e2 (2018).

27. Reijmerink, N. E. *et al.* Association of IL1RL1, IL18R1, and IL18RAP gene cluster polymorphisms with asthma and atopy. *J. Allergy Clin. Immunol.* **122**, 651–4.e8 (2008).

28. Smelser, D. T. *et al.* Association of varicose veins with rare protein-truncating variants in PIEZO1 identified by exome sequencing of a large clinical population. *J. Vasc. Surg. Venous Lymphat. Disord.* **10**, 382–389.e2 (2022).

29. Camire, R. M. Rethinking events in the haemostatic process: role of factor V and TFPI. *Haemophilia* **22 Suppl 5**, 3–8 (2016).

30. Vincent, L. M. *et al.* Coagulation factor V(A2440G) causes east Texas bleeding disorder via TFPIα. *J. Clin. Invest.* **123**, 3777–3787 (2013).

31. Cherñavsky, A. C. *et al.* Evidences of the involvement of Bak, a member of the Bcl-2 family of proteins, in active coeliac disease. *Autoimmunity* **35**, 29–37 (2002).

32. Wang, J. & Fu, Y.-X. LIGHT (a cellular ligand for herpes virus entry mediator and lymphotoxin receptor)-mediated thymocyte deletion is dependent on the interaction between TCR and MHC/self-peptide. *J. Immunol.* **170**, 3986–3993 (2003).

33. Li, Y. I. *et al.* Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* **50**, 151–158 (1/2018).

34. Taylor-Weiner, A. *et al.* Scaling computational genomics to millions of individuals with GPUs. *Genome Biol.* **20**, 228 (12/2019).

35. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).

36. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *arXiv [cs.LG]* 2825–2830 (2012) doi:10.5555/1953048.2078195.

37. Seabold, S. & Perktold, J. Statsmodels: Econometric and statistical modeling with python. in

Proceedings of the Python in Science Conference 92–96 (SciPy, 2010). doi:10.25080/majora-92bf1922-011.

38. MacDonald, J. W., Harrison, T., Bammler, T. K., Mancuso, N. & Lindström, S. *An Updated Map of GRCh38 Linkage Disequilibrium Blocks Based on European Ancestry Data*. http://biorxiv.org/lookup/doi/10.1101/2022.03.04.483057 (2022).

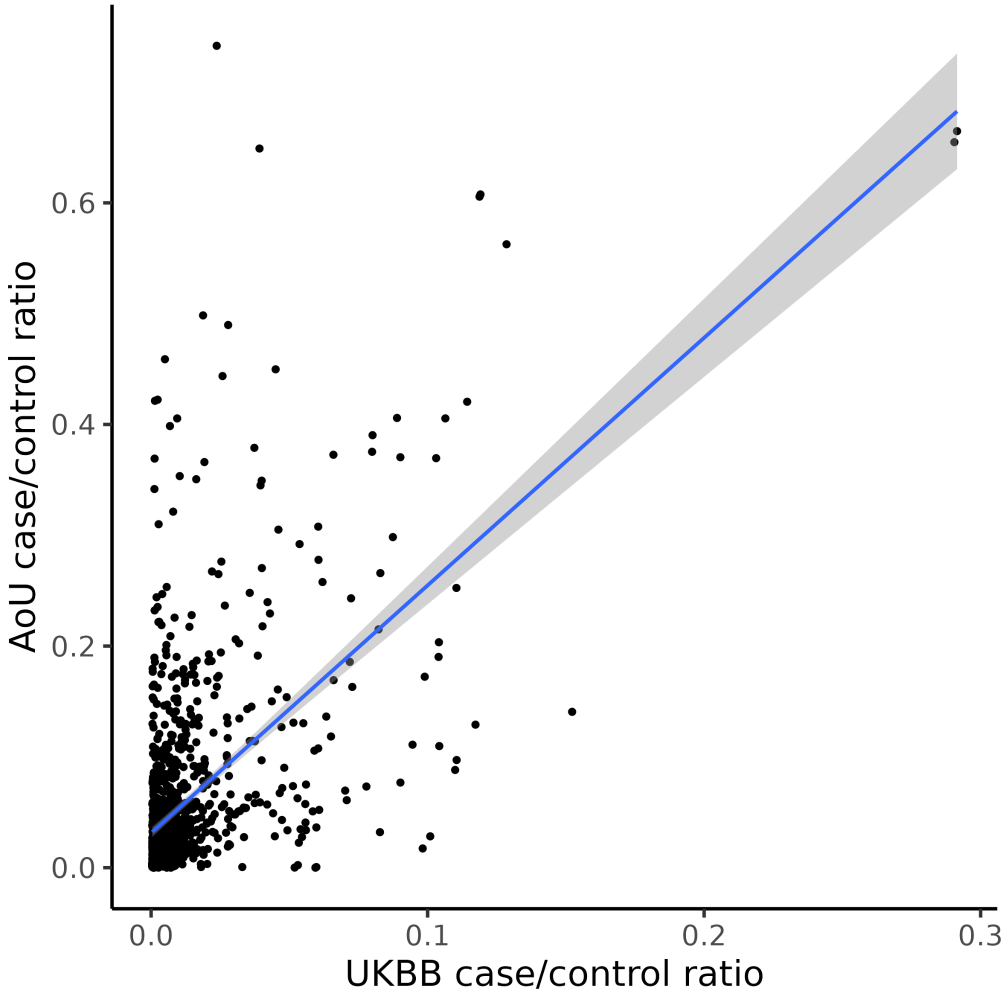

Figure S1. Case/control ratios in the UK Biobank and All of Us. Blue line represents a line of best fit, with a shaded 95% confidence interval.

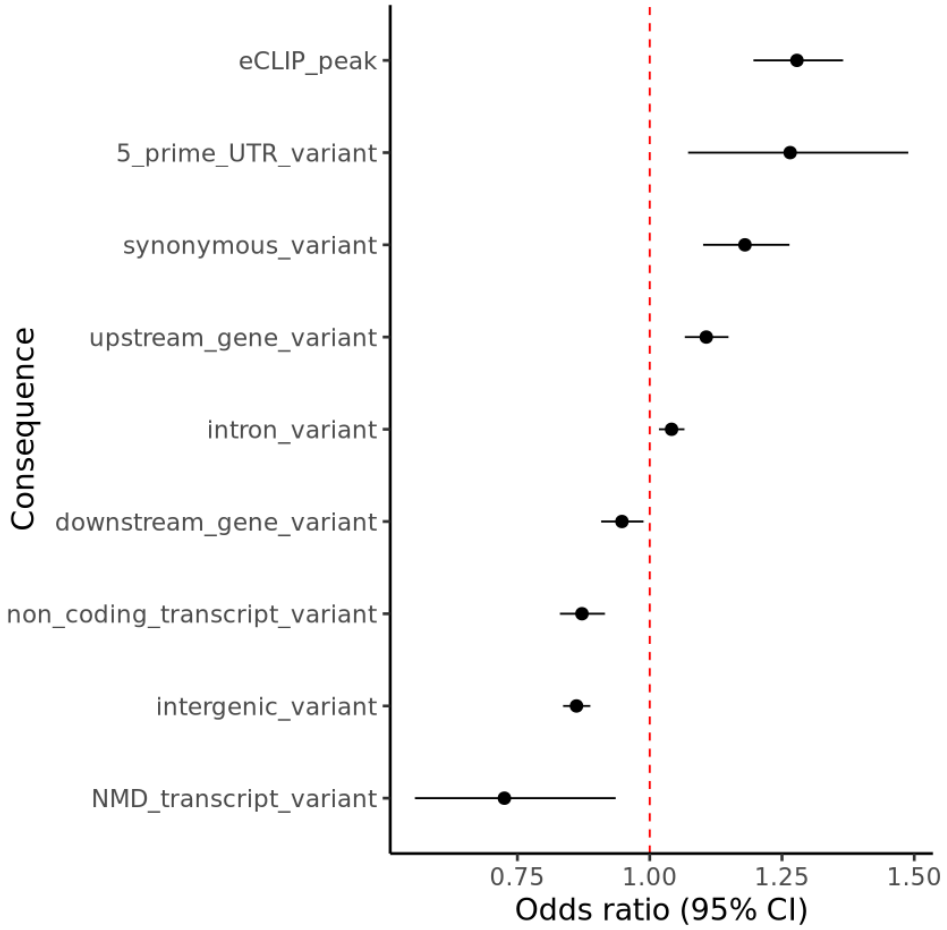

Figure S2. Enrichment of SNP consequences for SNPs of splicing event models with a significant sTWAS association. Points represent odds ratio, with extending lines showing the 95% confidence interval.

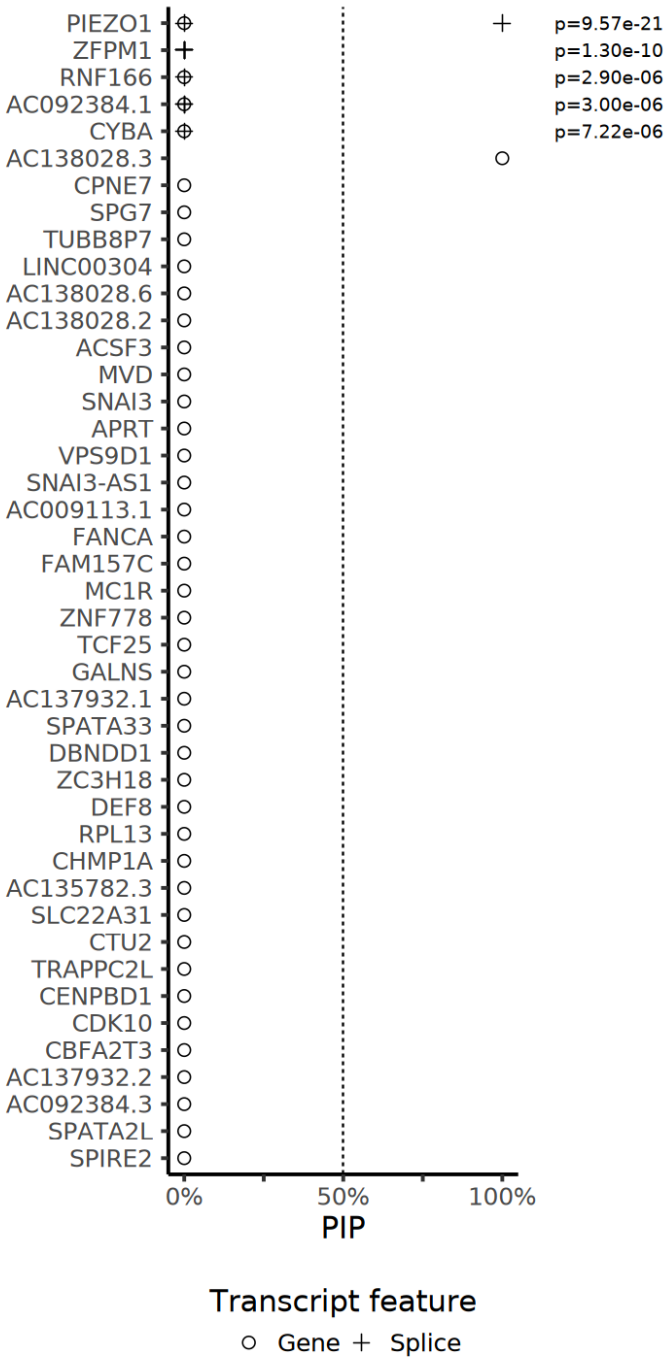

Figure S3. Tower plot showing the sTWAS fine-mapping PIPs for transcriptomic events for the LD block containing *PIEZO1* with risk of varicose veins. Splicing events represented as a plus, gene expression as a circle), with p-values annotated for other sTWAS significant splicing events.

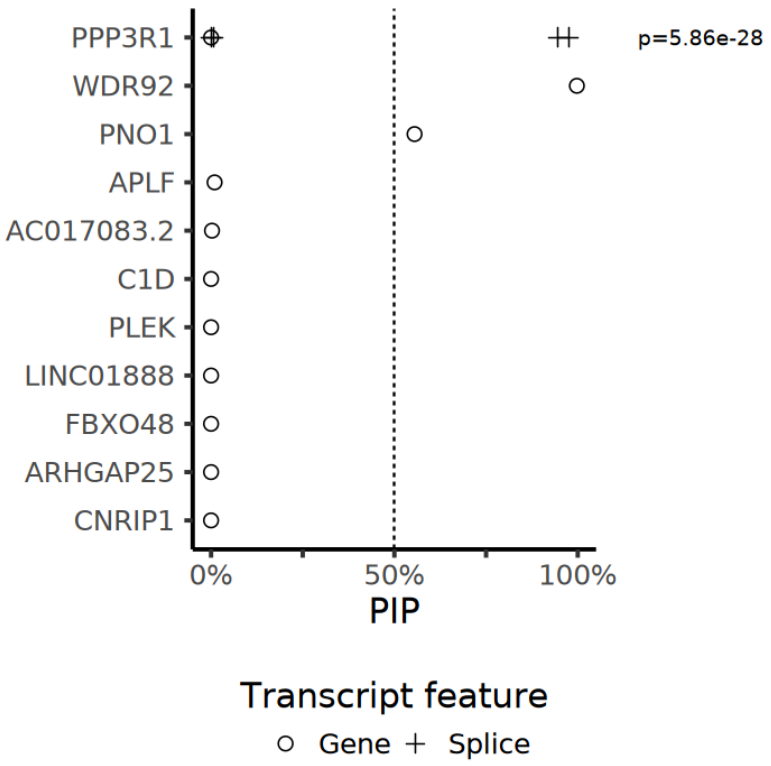

Figure S4. Tower plot showing the sTWAS fine-mapping PIPs for transcriptomic events for the LD block containing *PPP3R1* with risk of varicose veins. Splicing events represented as a plus, gene expression as a circle), with p-values annotated for other sTWAS significant splicing events.

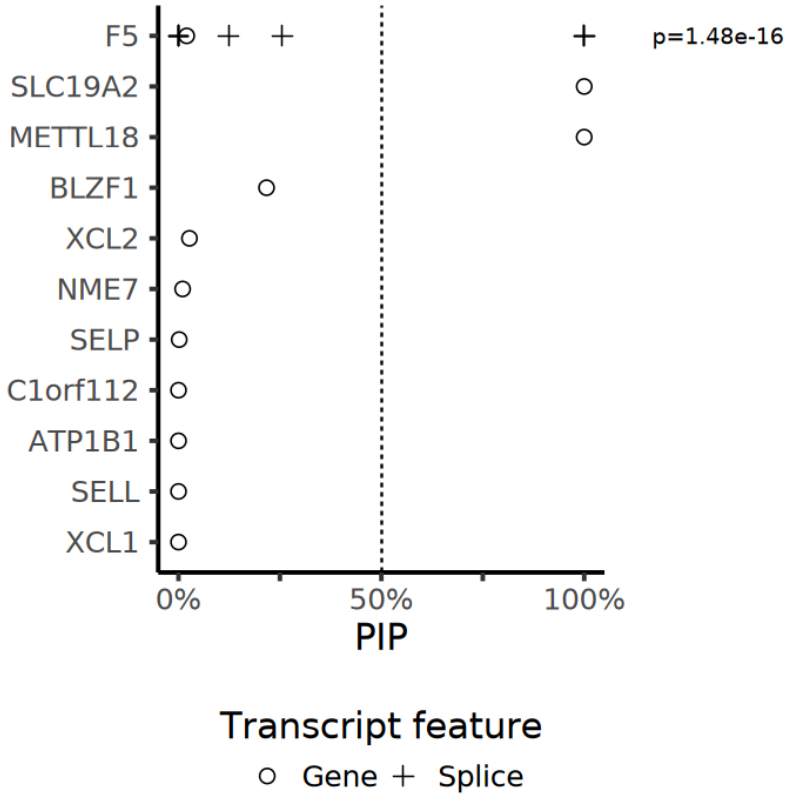

Figure S5. Tower plot showing the sTWAS fine-mapping PIPs for transcriptomic events for the LD block containing *F5* with risk of thrombophlebitis. Splicing events represented as a plus, gene expression as a circle), with p-values annotated for other sTWAS significant splicing events.

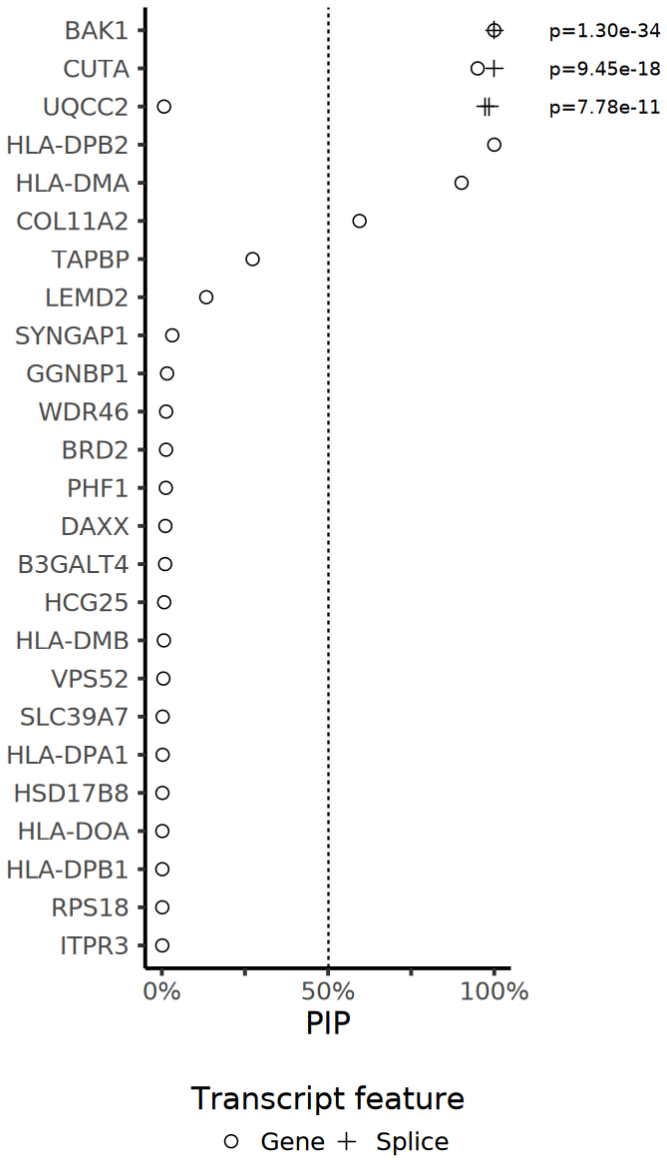

Figure S6. Tower plot showing the sTWAS fine-mapping PIPs for transcriptomic events for the LD block containing *BAK1* with risk of celiac disease. Splicing events represented as a plus, gene expression as a circle), with p-values annotated for other sTWAS significant splicing events.

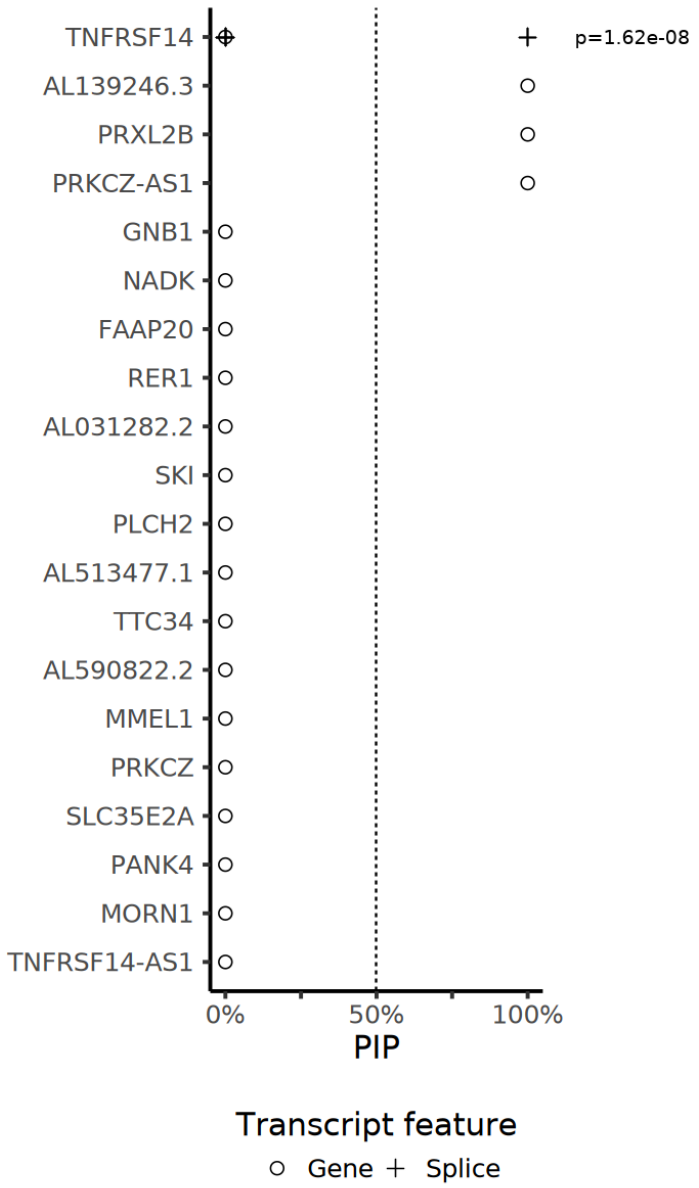

Figure S7. Tower plot showing the sTWAS fine-mapping PIPs for transcriptomic events for the LD block containing *TNFRSF14* with risk of celiac disease. Splicing events represented as a plus, gene expression as a circle), with p-values annotated for other sTWAS significant splicing events.

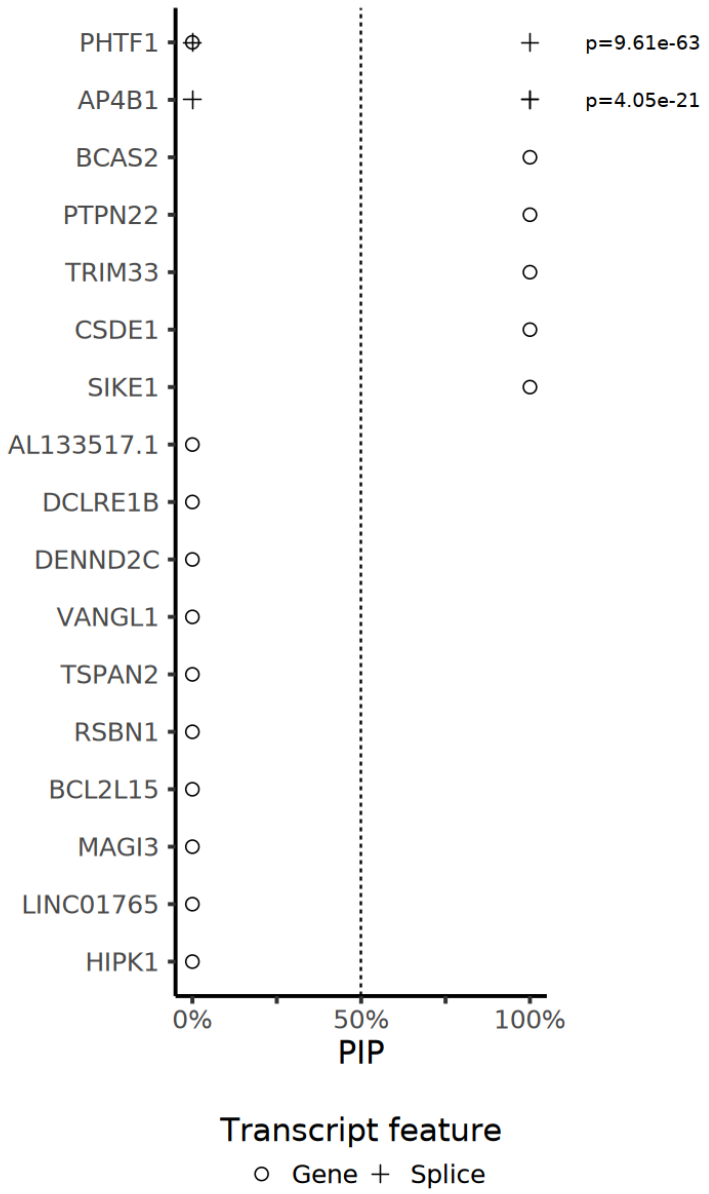

Figure S8. Tower plot showing the sTWAS fine-mapping PIPs for transcriptomic events for the LD block containing *PHTF1* with risk of hypothyroidism. Splicing events represented as a plus, gene expression as a circle), with p-values annotated for other sTWAS significant splicing events.

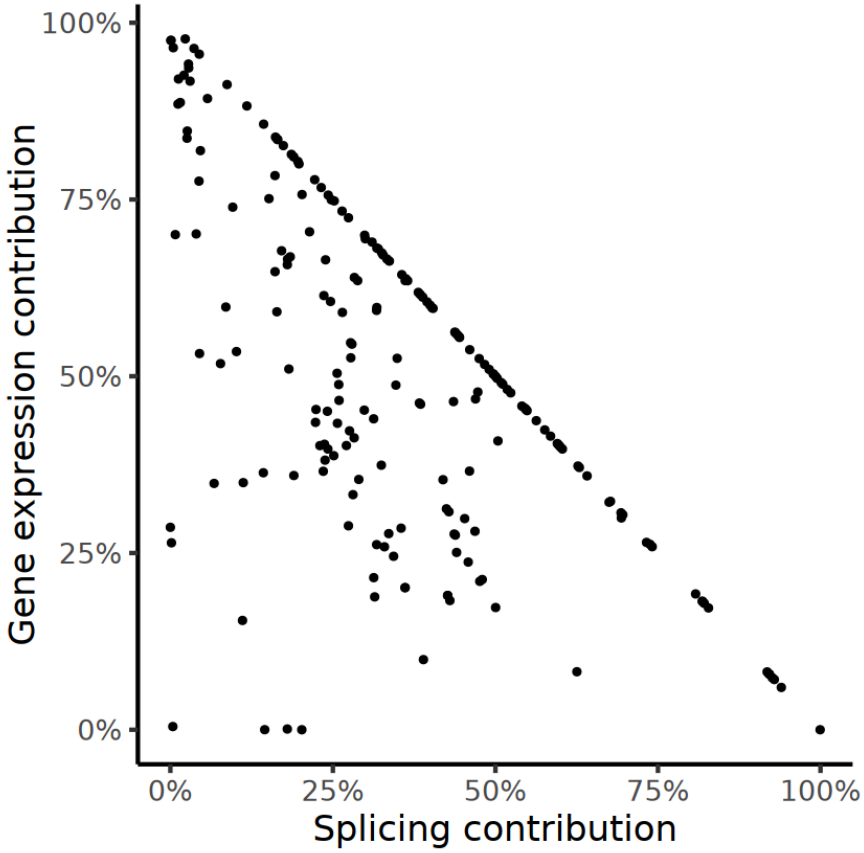

Figure S9. Scatter plot showing the relative transcriptomic proportion of the contribution of splicing and gene expression to traits.

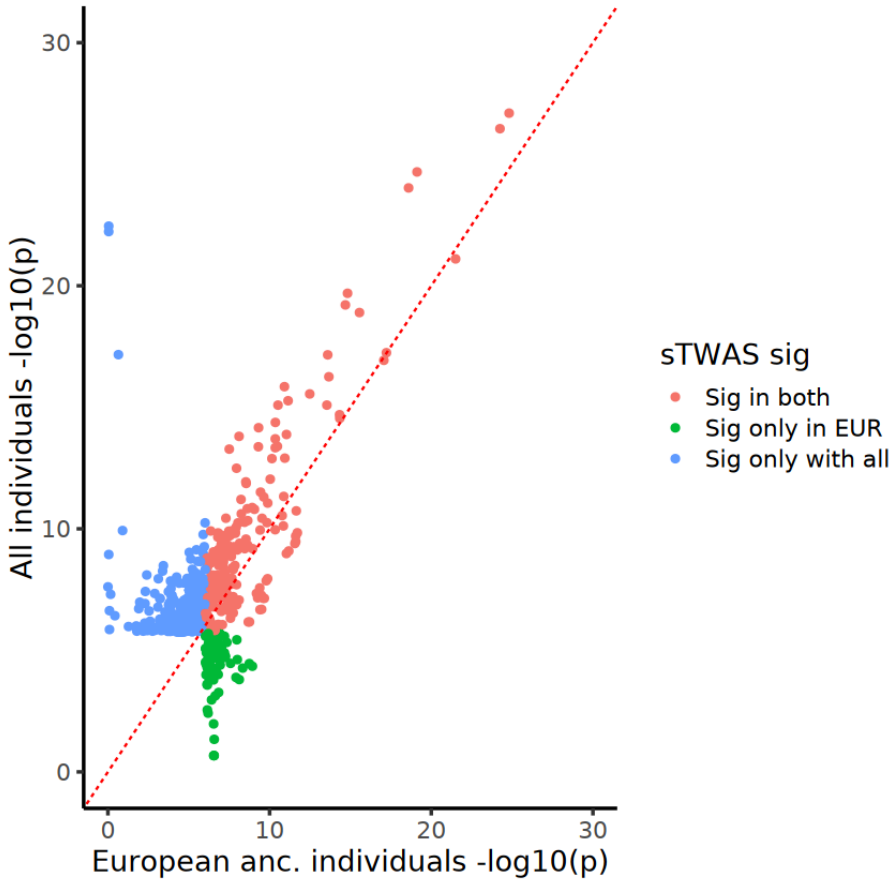

Figure S10. Comparison of TWAS association significance between European-only and full All of Us cohort. Red dashed line represents y=x, splicing event-phecode pairs which are significant in both analyses are in red, those only significant in the European-only subset are in green, and those only significant in the full AoU cohort are in blue. Pairs which are not significant in either analysis are removed.