TITLE

Isoform-level analyses of 6 cancers uncover extensive genetic risk mechanisms undetected at the gene-level

AUTHORS

Yung-Han Chang[1], S. Taylor Head[2], Tabitha Harrison[3], Yao Yu[2], Chad D. Huff[2], Bogdan Pasaniuc[4], Sara Lindström[3,5], Arjun Bhattacharya[2,6]

AFFILIATIONS

1. Quantitative Sciences Program, The University of Texas MD Anderson Cancer Center UTHealth Houston Graduate School of Biomedical Sciences, Houston, TX, USA
2. Department of Epidemiology, University of Texas MD Anderson Cancer Center, Houston, TX, USA
3. Department of Epidemiology, School of Public Health, University of Washington, Seattle, WA, USA
4. Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
5. Public Health Sciences Division, Fred Hutchinson Cancer Center, Seattle, WA, USA
6. Institute for Data Science in Oncology, University of Texas MD Anderson Cancer Center, Houston, TX, USA

Corresponding author: Arjun Bhattacharya (abhattacharya3@mdanderson.org)

ABSTRACT

Integrating genome-wide association study (GWAS) and transcriptomic datasets can help identify potential mediators for germline genetic risk of cancer. However, traditional methods have been largely unsuccessful because of an overreliance on total gene expression. These approaches overlook alternative splicing, which can produce multiple isoforms from the same gene, each with potentially different effects on cancer risk.

Here, we integrate genetic and multi-tissue isoform-level gene expression data from the Genotype Tissue-Expression Project (GTEx, N = 108-574) with publicly available European-ancestry GWAS summary statistics (all N > 20,000 cases) to identify both isoform- and gene-level risk associations with six cancers (breast, endometrial, colorectal, lung, ovarian, prostate) and six related cancer subtype classifications (N = 12 total). Compared to traditional methods leveraging total gene expression, directly modeling isoform expression through transcriptome-wide association studies (isoTWAS) substantially increases discovery of transcriptomic mechanisms underlying genetic associations. Using the same RNA-seq datasets, isoTWAS identified 164% more significant unique gene associations compared to TWAS (6,163 and 2,336, respectively), with isoTWAS-prioritized genes enriched 4-fold for evolutionarily-constrained genes ($P = 6.1 \times 10^{-13}$). isoTWAS tags transcriptomic associations at 52% more independent GWAS loci compared to TWAS across the six cancers. Additionally, isoform expression mediates an estimated 63% greater proportion of cancer risk SNP heritability compared to gene expression when evaluating cis-genetic influence on isoform expression. We highlight several notable isoTWAS associations that demonstrate GWAS colocalization at the isoform level but not at the gene level, including, *CLPTM1L* (lung cancer), *LAMC1* (colorectal), and *BABAM1* (breast). These results underscore the critical importance of modeling isoform-level expression to maximize discovery of genetic risk mechanisms for cancers.

INTRODUCTION

Over the past twenty years, genome-wide association studies (GWASs) have successfully linked hundreds of genetic variants associated with increased risks for common cancers, including breast, prostate, lung, and colorectal cancers[1]. For example, Zhang and Ahearn et al. identified 32 new loci associated with breast cancer risk in European ancestry individuals[2], totaling over 200 loci identified for overall and subtype-specific breast cancer risk. Similarly, Wang and Shen et al. identified 187 novel risk associations in a multi-ancestry prostate cancer GWAS[3]. Despite these findings underscoring the potential of GWASs to uncover genetic risk factors for cancer, a significant challenge remains as most genome-wide significant variants are in non-coding regions of the human genome, making it difficult to understand the biological mechanisms underlying these genetic associations.

To address this gap, transcriptome-wide association studies (TWASs) have emerged as a powerful complementary approach, providing valuable insights into the functional impact of genetic variants[4–8]. By focusing on the regulatory effects of genetic variants on gene expression, TWASs can identify potential target genes and pathways involved in disease processes. This integrated approach enhances our ability to translate genetic findings from GWASs into a deeper molecular understanding of disease etiology, which is necessary for improved diagnostics, prevention strategies, and therapeutic interventions in oncology.

Despite the promise of TWASs and related approaches, recent work suggests that gene expression may be a poor mediator of genetic risk of complex traits, potentially owing to differences in evolutionary pressures on genetic variants with large effects on complex traits and those with large effects on molecular phenotypes[9,10]. It is possible that part of this poor mediation is due to an overreliance on total gene expression as a fundamental unit of measure of the transcriptome. Total gene expression does not account for alternative splicing that can produce multiple isoforms from the same gene, which are often under subtle genetic or environmental control. The mRNA and protein isoforms generated through alternative processing of primary RNA transcripts can vary in function and potentially affect cancer risk in different ways[11,12]. To address this limitation, we developed isoform-level TWAS (isoTWAS), a multivariate, stepwise framework that integrates genetic and isoform-level transcriptomic variation with GWAS summary statistics. This approach identifies both isoform- and gene-trait associations while controlling the false discovery rate, identifying transcriptomic associations at far more GWAS loci compared to traditional gene-level TWAS[13,14].

In this study, we utilized isoTWAS to assess the association between isoform expression and risk of six cancer types (breast[15,16], endometrial[17], colorectal[18], lung[19], ovarian[20], and prostate[21]), and their subtype classifications where applicable. We used publicly available European-ancestry cancer GWAS summary statistics and multi-tissue isoform- and gene-level expression data from the Genotype Tissue-Expression Project (GTEx, N > 100)[22]. isoTWAS reveals more associated loci than traditional TWAS, highlighting its potential to uncover novel genetic insights and enhance our understanding of cancer biology.

RESULTS

Overview of isoTWAS

We conducted isoform- (isoTWAS) and gene-level transcriptome-wide association studies (TWAS) for a total of 12 cancer outcomes: (1-3) breast (BRCA; overall, estrogen receptor (ER)+, ER-)[15,16], (4) colorectal (CRC)[18], (5) lung (LUNG; overall)[19], (6) lung adenocarcinoma (LUAD)[19], (7) lung squamous cell carcinoma (LUSC)[19] (8-9) ovarian (OVCA; overall, serous)[20], (10-11) prostate (PRCA; overall, advanced)[21], and (12) endometrial (UCEC)[17]. We integrated GWAS summary statistics (N = 63,053-228,951) with multi-tissue expression QTL data from individuals without cancer collected post-mortem from the Genotype Tissue-Expression Project (GTEx, N = 108-574). GWAS and GTEx sample sizes and assignments of relevant tissues to cancer outcomes are provided in **Supplemental Tables 1-2**. See **Methods** for further details.

We used univariate and multivariate penalized regressions (**Figure 1A**) to train gene and isoform expression models and map gene- and isoform-level risk associations for the 12 cancer outcomes using a weighted burden test (**Figure 1B**)[5,23]. In isoTWAS, we inferred gene-level associations by combining isoform associations via aggregated Cauchy association[24]. False discovery rate is controlled for via a Benjamini-Hochberg procedure, and family-wise error rates are controlled via Shaffer's modified sequentially rejective Bonferroni (MSRB) for correlated isoforms. Post-hoc analyses include permutation tests and fine-mapping to identify isoform sets driving significant associations and credible sets of genes or isoforms (**Figure 1C**). See **Methods** for further details.

Isoform-level analysis reveals cancer risk associations not detected at the gene-level

In total, isoTWAS identifies 11,078 significant isoform associations across 6,163 unique genes, representing a ~164% increase in gene identification as compared to traditional TWAS which identified 2,336 significant gene associations, (**Figure 2A; Supplemental Table S3-5, Supplemental Data S1-2**). Of the genes identified, 10% from isoTWAS and 7% from traditional TWAS are included in the OncoKB Cancer Gene List, representing a moderate enrichment of hallmark cancer genes among isoTWAS gene ($P = 0.03$)[25,26]. For both TWAS and isoTWAS, most identified gene associations are observed for BRCA, ER+ BRCA, and PRCA, which also are the three largest GWAS. Specifically, isoTWAS identifies 3.12 times as many associations for BRCA (3,889 isoforms), 2.85 times more for PRCA (1,849 isoforms), and 2.39 times as many for ER+ BRCA (2,304 isoforms) as compared to TWAS, while showing relatively similar performance for LUNG; Manhattan plots of TWAS and isoTWAS results are shown in **Supplemental Figures S1-S12**. Given that the mean of the $\chi2$ distribution is linearly related to power and sample size, the percent increase in the test statistic serves as a measure of power or effective sample size. For $\chi2 > 1$, we calculated the percent increase for isoTWAS-based associations compared to TWAS-based associations (**Figure 2B**). Across the 12 cancer outcomes, there is an average 25.3-37.4% increase in effective sample size for isoTWAS compared to TWAS, suggesting that isoTWAS may enhance discovery in real data as compared to traditional TWAS. Several key biological pathways are enriched

for cancer-specific sets of isoTWAS-prioritized genes, including cell cycle and mitosis regulation, DNA and RNA binding, immune pathways, as well as downstream targets of cancer-relevant transcription factors like ESR1[27], RUNX2[28], and YY1[29] (**Supplemental Table S6**, **Supplemental Figures S13-S18**).

We further explored if cancer risk genes identified by TWAS and isoTWAS capture true disease signals by identifying genes under selective constraint. If a gene is constrained, selection will act to remove variants that diminish gene function from the population, such as loss-of-function (LOF) variants. Here, we used Bayesian estimates of the s_{het} measure of constraint[30], which, unlike traditional measures of constraint, is not biased towards longer genes. In total, 19.9% of isoTWAS gene associations (1,226 of 6,163) show $s_{het} > 0.1$, compared to 12.5% of TWAS gene association (293 of 2,336) with shet > 0.1. Not only does this represent a significant enrichment of high s_{het} among isoTWAS-prioritized gene associations compared to TWAS-prioritized gene associations ($\chi 2$ test $P = 3.7 \times 10^{-15}$), isoTWAS-prioritized genes are significantly enriched compared to the genome-wide proportion of high s_{het} genes (14.7%, $\chi 2$ $P < 2.2 \times 10^{-16}$). In addition, we find significant enrichments of high s_{het} genes among transcriptome-wide significant genes for ER- BRCA, LUNG, LUSC, PRCA, Adv PRCA, and UCEC ($\chi 2$ FDR-adjusted P-value < 0.05; **Figure 2C, Supplemental Table S4**).

Lastly, isoTWAS identifies 52 genes (34 undetected by TWAS) that are associated with five or more cancer outcomes (**Figure 2D**), including multiple known oncogenes or tumor suppressor genes, such as *MYC*[31], *MUTYH*[32], *GNAI2*[33,34], *ACO2*[35], and *BMI1*[36]. The 34 genes undetected by TWAS are enriched for multiple salient pathways: regulation of CD8-positive and T cells, regulation of membrane protein complexes, and phagocytic and autolysosome function. Additionally, these 34 genes are highly enriched for downstream targets of crucial oncogenic transcription factors, like *ESR1*, *GATA4*, *YY1,* and *MYC*, all of which were determined using ChIP-Seq experiments in human or mouse cancer cells or tumors (**Figure 2E**)[37]. Altogether, not only can isoTWAS identify constrained susceptibility genes for multiple cancers, but it can also reveal pan-cancer risk signals that TWAS is unable to detect.

Isoform expression explains more GWAS loci and overall SNP heritability of cancer risk

In addition to increasing the discovery of susceptibility genes across the genome, isoTWAS can identify far more transcriptomic mechanisms within independent, high-confidence genome-wide significant loci. Across all cancer outcomes, we identify 622 risk-associated GWAS SNPs ($P < 5 \times 10^{-8}$) in independent linkage disequilibrium (LD) blocks (see **Methods**). Among these 622 loci, 288 (46.3%) are tagged by TWAS, and 439 (70.6%) are tagged by isoTWAS, with 249 (40.0%) loci identified by both TWAS and isoTWAS (**Figure 3A, Supplemental Table S7**). In total, this represents an increase in significant associations in GWAS loci of 52.4% when using isoTWAS, rather than TWAS. Additionally, isoTWAS identified 2,911 unique significant genes outside of known GWAS loci.

We also considered an orthogonal analysis to compare GWAS follow-up using both gene and isoform expression. We mapped *cis*-expression quantitative trait loci (eQTLs) for gene and isoform expression and conducted Bayesian colocalization analysis with GWAS signals using eCAVIAR[38] to estimate the CoLocalization

Posterior Probability (CLPP; see **Methods; Supplemental Table S8, Supplemental Data S3**). Generally, isoTWAS exhibits higher CLPPs compared to TWAS, indicating a stronger likelihood of colocalization for isoform-eQTLs with GWAS loci (**Figure 3B**), with the proportion of loci having isoform expression Quantitative Trait Loci (isoQTL) colocalizations with CLPP > 0.01 consistently higher across the 12 cancer outcomes (**Supplemental Figure S19**). Notably, in OVCA and UCEC, isoTWAS shows significantly higher CLPPs (median = 0.026, P = 0.013) than TWAS (median = 0.001, P = 0.001), suggesting that isoTWAS more effectively captures colocalized signals in for these cancer outcomes. Additionally, we estimated the proportion of total SNP heritability (h^2) mediated by gene- and isoform-level expression (h^2_{med}), (**Figure 3C** and **Supplemental Table S9)**. Overall, isoform-level expression explains 62.7% more of cancer risk SNP heritability (19.2 ± 10.3%) compared to gene-level expression (11.8 ± 5.7%). Wald-type tests reveal that isoform expression mediates a significant proportion of cancer h^2, with gene expression only explaining a significant portion for ER- BRCA (isoTWAS: 0.291, FDR-adjusted P = 0.005; TWAS: 0.133, P = 0.068), ER+ BRCA (isoTWAS: 0.173, P = 0.045; TWAS: 0.110, P = 0.102), LUNG (isoTWAS: 0.287, P = 0.044; TWAS: 0.181, P = 0.056), PRCA (isoTWAS: 0.314, P = 0.005; TWAS: 0.108, P = 0.164), and Adv PRCA (isoTWAS: 0.358, P = 0.014; TWAS: 0.243, P = 0.046). These results indicate that not only does isoTWAS recapitulate an overwhelming majority of TWAS signals at GWAS loci, isoTWAS substantially increases discovery of candidate GWAS mechanisms and transcriptomic features that potentially mediate genetic effects on cancer risk.

isoTWAS and isoform-eQTL colocalization prioritize undetected mechanisms at GWAS loci

A main goal of isoform-specific analyses is to nominate a more granular hypothesis of transcriptomic regulation. Post-hoc transcript-level fine-mapping reveals nine loci where isoTWAS prioritization coincides with isoform-eQTL colocalization (CLPP > 0.01) but no gene-eQTL colocalization (**Methods**; **Supplemental Data S3**), across BRCA (**Figure 4**, **Supplemental Figures S20**), ER- BRCA (**Supplemental Figure S21-S22**), CRC (**Figure 5**, **Supplemental Figure S23**), LUNG (**Figure 6**), PRCA (**Supplemental Figure S24**), and UCEC (**Supplemental Figure S25**). We highlight *CLPTM1L*, *LAMC1,* and *BABAM1* here due to previously-reported pleiotropic associations across multiple cancers. *CLPTM1L* is located near the *TERT* locus, known for its pleiotropic links to various cancers[39], the *LAMC1* locus has shown associations with PRCA[40], and *BABAM1* is a recognized GWAS hit with broad implications for cancer risk. These three genes are prioritized using isoTWAS models trained in subcutaneous adipose tissue, with additional examples presented in the **Supplemental Results**.

First, four isoforms of *CLPTM1L* (Chromosome 5p15.33, s_{het} = 0.02, 16 total isoforms in GENCODE) are significantly associated with LUNG. We also find that isoforms of *CLPTM1L* are associated with BRCA, PRCA, LUAD, and LUSC. Fine-mapping prioritizes ENST00000511268.6 in the 90% credible set with Posterior Inclusion Probability (PIP) = 0.90. Although there are many genome-wide significant SNPs within the gene body of *CLPTM1L*, no gene-eQTL signal is observed ($P < 10^{-6}$), whereas a strong ENST00000511268.6 isoform-eQTL signal colocalizes with the GWAS signal (CLPP = 0.07). Additionally, there is no strong isoform-eQTL signal for the other three isoforms identified via isoTWAS in this region. The lead isoQTL rs414965 is in high LD with

multiple genome-wide significant SNPs (**Figure 4A**). The exons comprising ENST00000511268.6 overlap with those of ENST00000503534.5 and ENST00000506641.5 with significant isoform-eQTLs flanking ENST00000511268.6 (**Figure 4B**). However, rs414965 does not have significant effects on ENST00000503534.5, ENST00000506641.5, or ENST00000503151.5 expression (**Figure 4C**), despite its strong negative associations with LUNG and ENST00000511268.6 expression.

Next, three isoforms of *LAMC1* (Chromosome 1q25.3, s_{het} = 0.10, six total isoforms in GENCODE) are associated with CRC. Fine-mapping prioritizes only ENST00000466964.1 in the 90% credible set with PIP = 1. Again, genome-wide significant CRC-associated SNPs within the gene body show no gene-eQTL signal ($P < 10^{-6}$) but a strong ENST00000466964.1 isoform-eQTL signal that colocalizes with the GWAS signal (CLPP = 0.04) is observed. For the other two isoforms, ENST00000478064.1 and ENST00000495918.1, the lead isoQTL rs20558 is in high LD with multiple genome-wide significant SNPs (**Figure 5A**). The exons of these three isoforms are generally distinct sets with significant isoform-eQTLs generally falling on the 3' end of multiple exons (**Figure 5B**). rs20558 shows a significant association with increased CRC risk, no effect on BABAM1 gene expression, and a significant decreasing effect on ENST00000466964.1 expression (**Figure 5C**). Interestingly, this same SNP has a significant increasing effect on the expression of ENST00000478064.1 and ENST00000495918.1, though only ENST00000478064.1 colocalizes with CRC risk at CLPP > 0.01.

Lastly, nine isoforms of *BABAM1* (Chromosome 19p13.11, s_{het} = 0.05, 18 total isoforms in GENCODE) are associated with BRCA. In our study, isoforms of *BABAM1* are also associated with OVCA, OVCA ser, LUSC, and ER- BRCA in isoTWAS. Fine-mapping prioritizes ENST00000599474.5 in the 90% credible set and a PIP of 1. There are multiple genome-wide significant BRCA-associated SNPs ($P < 5x10^{-8}$) within the *BABAM1* gene body showing no gene-eQTL signal ($P < 10^{-6}$). Only ENST00000599474.5 (CLPP = 0.44) and ENST00000359435.8 (CLPP=0.59) show strong isoQTL effects that colocalize with the GWAS signal. In addition, the lead isoQTL rs34084277 for ENST00000599474.5 is in high LD (LD > 0.8) with multiple genome-wide significant SNPs (**Figure 6A**). The exon structure of these nine isoforms reveals a group of exons at the 3' end of the gene body, which is flanked by the lead isoQTLs of ENST00000599474.5 (**Figure 6B**). Specifically, due to its exon structure, we followed up on this locus using a rare-variant analysis in UK Biobank whole exome sequencing data[41] using the Variant Annotation, Analysis, and Search Tool (VAAST2)[42,43]. We find three BRCA-associated isoforms in *BABAM1* enriched for risk-associated rare variants (MAF < 0.5%), with associations mainly concentrated in exons at the 5' end and first exon at the 3' end of the transcript (**Supplemental Methods; Supplemental Figure 26; Supplemental Tables S10-S11**). See **Supplemental Results** for further details. Lastly, rs34084277 has a strong protective effect on *BRCA* risk, no effect on *BABAM1* gene expression, and a significantly increasing effect on expression of ENST00000599474.5 (**Figure 6C**).

DISCUSSION

We show that integrating cancer risk GWAS summary statistics with isoform-level transcriptomic variation can greatly increase discovery of susceptibility genes for six cancers and their subtype classifications. By using isoTWAS rather than traditional gene-level TWAS, we identify nearly 2.5-fold more associations. More saliently, isoTWAS-identified genes are significantly enriched for evolutionarily constrained genes, which are more likely to contain clinically-relevant *de novo* or rare variants, predict drug toxicity, and characterize transcriptional regulation[10,44]. Isoform expression, rather than total gene expression, captures more risk-associated genetic variation and exhibits stronger colocalization with GWAS signals. Additionally, though isoform expression alone does not reconcile most of the SNP heritability of cancer risk, isoform expression is estimated to mediate nearly twice as much heritability as gene expression, and nearly three times of the SNP heritability signal in the case of PRCA. Most importantly, isoTWAS can find isoform associations at 249 of 288 GWAS loci tagged by TWAS-identified genes and uniquely tag 190 additional GWAS loci that cannot be contextualized via TWAS. In aggregate, these findings underscore the utility of considering the transcriptome on the transcript-isoform, rather than gene-level. By modeling a different quantification of the same steady-state RNA-seq datasets with sample sizes of only up to ~800, isoTWAS increases discovery specifically at GWAS loci by ~52% without additional sequencing costs.

Our isoform-level fine-mapping coupled with eQTL colocalization identifies nine gene candidates with GWAS risk SNPs within the gene body that colocalize with isoQTLs despite exhibiting no gene-eQTL signal at $P < 10^{-6}$. We discuss three of these gene candidates. First, isoTWAS identifies an association between ENST00000511268.6, an isoform of *CLPTM1L* (16 total isoforms), and LUNG. Studies spanning back 15 years have identified multiple polymorphisms within the *CLPTM1L* gene body associated with LUNG, as well as pleiotropic associations with other malignancies[45–49]. In particular, an *in vitro* study provided evidence that *CLPTM1L* is oncogenic, specifically for Ras-driven lung cancers, in line with its effect on protecting tumor cells from genotoxic apoptosis[50], and is a promising therapeutic target for therapy-resistant tumors[51]. *CLPTM1L* (Chromosome 5p15.33) is also local to *TERT*, which harbors pleiotropic associations with multiple cancers[39]. Our study identifies associations between *CLPTM1L* and PRCA, consistent with previous findings of SNPs in the 5p15.33 region associated with PRCA[52,53]. However, while *TERT* is known for its pleiotropic effects, we observed only gene-level associations for *TERT* in our analyses, with no isoTWAS associations detected for its specific isoforms. Further work is needed to fully interrogate this pleiotropy and assess tissue-specific isoform expression for both *TERT* and *CLPTM1L*, especially since *TERT* shows low expression across multiple tissues in GTEx[22]. We find a similar pattern for colorectal cancer and ENST00000466964.1, an isoform of *LAMC1* (six total isoforms), a gene that regulates cell adhesion, differentiation, migration, and signaling, and lies at a pleiotropic locus for CRC, PRCA[40], and obesity risk[54]. A previous study incorporating splicing measures has prioritized *LAMC1* as a novel transcriptome-mediated CRC locus[55]. Additionally, rs34295433 in *LAMC1* was identified as a susceptibility SNP for PRCA in a Taiwanese population[40]. However, in our study, isoTWAS did not detect any associations between *LAMC1* and PRCA. *LAMC1* belongs to the laminin family of extracellular matrix proteins that are significantly involved in survival and proliferation of cancer cells, angiogenesis, migration and basement membrane breach by cancer

cells, and metastatic events[56]. Computational analyses of laminin proteins have shown their prognostic ability in colorectal cancer progression, placing higher weights for *LAMC1* compared to other constituents of the family[57].

Lastly, isoTWAS detects an association between BRCA and ENST00000599474.5, an isoform of *BABAM1* (18 total isoforms), a gene involved in checkpoint signaling, regulation of DNA repair, and mitosis. *BABAM1* has previously been implicated as a low-penetrance risk locus that interacts with *BRCA1* in both triple-negative breast cancer and ovarian cancer risk[58–60]. Consistent with previous studies, isoTWAS also identifies associations within this region for OVCA, OVCA ser, and ER- BRCA. Additionally, our VAAST analysis for *BABAM1* in UKBB indicates that a group of exons at the 3' end of the gene harbor potentially disease-causing rare variants (MAF < 0.01, only in coding regions). The lead isoQTL for our isoTWAS-prioritized isoform of *BABAM1* (MAF > 0.01, both coding and non-coding region) is upstream of these exons at the 3' end, potentially influencing the splicing patterns specifically at the 3' end of the gene. Further research is required to investigate how these isoQTLs influence splicing regulation and their specific role in tumorigenesis and cancer development, but a methodological opportunity may lie in integrating splicing- and isoQTLs with rare variants to identify transcriptomic mechanisms for cancer risk. These results underscore that the added resolution provided with isoform expression can lead to more specific mechanistic hypotheses for cancer risk and inform follow-up with functional studies, both *in silico* and experimental.

We conclude with three limitations of our work. First, the complexity of the *BABAM1* transcript structure and the sheer number of *BABAM1* isoforms with strong isoQTL signals underscores a methodological opportunity. We applied the FOCUS framework which leverages a non-informative prior that may be insufficient[61]. Fine-mapping of isoforms is challenging due to horizontal pleiotropy of SNP-isoform effects shared across exons and strong LD patterns. This horizontal pleiotropy can reduce power and increase false-positive rates. Incorporating classes of isoforms with shared exon sets may lead to improved fine-mapping, thereby increasing coverage and resolution of credible sets of isoforms within a GWAS locus.

Second, we derived isoform-level quantifications using Salmon, a method constrained by the limitations of short-read RNA-seq where maximum-likelihood estimates are based on transcriptomic annotations, introducing some uncertainty. The limited diversity of tissue-specific annotated isoforms may affect the accuracy of these estimates. In contrast, long-read RNA-seq can capture splicing and structural variation missed by short-read RNA-seq, revealing more complex and rarer transcript-isoforms. As long-read RNA-seq becomes increasingly scalable, its comprehensive view of transcriptomes will allow for more precise quantification of isoforms potentially improving isoTWAS performance. Third, as current eQTL datasets are mostly generated in populations of European ancestry, we focused the current analyses on cancer GWAS summary statistics based on individuals of European ancestry. Previous research show that expression models trained on predominantly European-ancestry cohorts often fail to predict expression accurately in individuals of different ancestries[62–65]. A critical emphasis of future data generation should be to increase diversity in molecular datasets and for

methodological research to develop computational frameworks that can model multi-ancestry samples across eQTL and GWAS datasets.

METHODS

Data collection

GTEx genotype and transcriptomic data

For 48 tissues from GTEx[22], we quantified RNA-seq data (all N > 100) using Salmon v1.5.2 in mapping-based mode[66]. We built a decoy-aware transcriptomic index in Salmon with GENCODE v38 transcript sequences and the full GRCh38 reference genome as decoy sequences[67]. Salmon was then run on FASTQ files with mapping validation and corrections for sequencing and GC bias. We then imported Salmon isoform-level quantifications and aggregated to the gene-level using tximeta v1.16.1[68]. Using edgeR, gene and isoform-level quantifications underwent TMM-normalization, followed by transformation into a log-space using the variance-stabilizing transformation using DESeq2 v1.38.3[69,70]. We then residualized isoform-level and gene-level expression (as log-transformed CPM) by all tissue-specific covariates (clinical, demographic, genotype principal components (PCs), and expression PEER factors) used in the original QTL analyses in GTEx.

SNP genotype calls were derived from Whole Genome Sequencing data from individuals of European ancestry, filtering out SNPs with minor allele frequency (MAF) less than 5% or that deviated from HWE at $P < 10^{-5}$. We further filtered out SNPs with MAF less than 1% frequency among the European ancestry samples in 1000 Genomes Project[71]. Due to limited sample sizes in other ancestry groups in GTEx, we focused solely on European ancestry for this study.

GWAS summary statistics

We obtained GWAS summary statistics for risk of 12 cancer outcomes, all from samples of European-ancestry individuals[72]: overall breast cancer (BRCA; 122,977 cases/105,974 controls)[15], estrogen-receptor positive breast cancer (ER+ BRCA; 69,501 cases/95,042 controls)[15], estrogen-receptor negative breast cancer (ER- BRCA; 30,882 cases/110,058 controls)[16], Colorectal cancer (CRC; 55,168 cases/65,160 controls)[18], overall lung cancer (LUNG; 29,266 cases/56,450 controls)[19], lung adenocarcinoma (LUAD; 11,273 cases/55,483 controls)[19], lung squamous cell Carcinoma (LUSC; 7,426 cases/55,627 controls)[19], overall ovarian cancer (OVCA; 22,406 cases/40,951 controls)[20], serous ovarian cancer (OVCA ser; 19,890 cases/68,502 controls)[20], overall prostate cancer (PRCA; 79,166 cases/61,106 controls)[21], advanced prostate cancer (Adv PRCA; 15,167 cases/58,308 controls)[21], endometrial cancer (UCEC; 12,906 cases/108,979 controls)[17].

Statistical analysis

Code and data availability

Sample scripts for analyses are available from github.com/bhattacharya-a-bt/MultiCancerIsoTWAS. Links to GWAS summary statistics are provided in **Supplemental Table S1**. isoTWAS and TWAS models are available

from https://zenodo.org/records/11048201[73]. GTEx genotyping and expression data were accessible from dbGaP accession number phs000424.v9. Supplemental Data Tables S1-S3 can be downloaded from https://zenodo.org/records/14010391.

Isoform- and gene-level transcriptome-wide association studies

We trained predictive models of gene and isoform expression using all *cis*-SNPs within 1 Mb of the gene body (**Figure 1A**). For gene expression, we implemented three methods: (1) elastic net regression[74], (2) linear mixed models using a ridge regression approximation[75], and (3) Sum of Single Effects (SuSiE) regression[76] (commonly used in FUSION[5]) and then selected the model with the best prediction based on 10-fold cross-validation[5]. For isoform expression, we considered the matrix of expression for all isoforms of a gene and implemented four multivariate methods: (1) multivariate elastic net regression, (2) multivariate LASSO with covariance estimation, (3) multivariate elastic net with a stacked generalization, and (4) multivariate sparse partial least squares. If a gene only had one isoform, we trained univariate penalized regressions, as done for gene expression. Again, we selected the best isoform model through a 10-fold cross-validation. We only retained genes and isoforms that could be predicted at cross-validation $R^2 > 0.01$.

Using these predictive models, we employed a weighted burden test to estimate the association between the genetically-predicted component of a gene or isoform with cancer risk (**Figure 1B**). We denoted $\hat{w}$ as the prediction weights for a gene or isoform model. We denoted Z as the vector of standardized effect sizes from the GWAS and Σ as the LD matrix between the SNPs represented in Z. We estimated the standardized effect size of the genetically-predicted component of a gene or isoform's expression on cancer risk as $T = \frac{\hat{w}Z}{\hat{w}^{\mathrm{T}}\Sigma\,\hat{w}}$. As one gene is likely to carry multiple isoform associations, isoform-level mapping has an increased testing burden (**Figure 1C**). Accordingly, we used the aggregated Cauchy association to combine the t test statistics $T_1, \ldots, T_t$ for isoforms of a single gene to estimate the isoTWAS gene-level association. For both TWAS and isoTWAS, we controlled for false discovery across all genes via the Benjamini-Hochberg procedure. For isoTWAS, we first identified isoforms for genes with an FDR-adjusted ACAT $P < 0.05$, and then employed Shaffer's modified sequentially rejective Bonferroni procedure to control the within-gene family-wide error rate (FWER). At the end of these two steps, isoTWAS identified a set of genes and their isoforms that were associated with the trait.

For significant genes (from TWAS) and isoforms (from isoTWAS), we tested whether the SNP-gene/isoform effects from the predictive models add additional information beyond the SNP-risk effects from the GWAS through a conservative permutation test. We permuted the SNP-gene/isoform effects in the predictive models 10,000 times and generated a null distribution for the gene/isoform test statistic. We used this null distribution to generate a permutation-based P-value for the original test statistic for each gene/isoform. We only conducted this permutation test for genes from TWAS with FDR-adjusted $P < 0.05$ and isoforms from isoTWAS with FDR-adjusted $P < 0.05$ and FWER-adjusted $P < 0.05$.

After permutation testing, we conducted isoform- and gene-level fine-mapping for isoTWAS and TWAS associations that overlap in a 1 Mb window using methods from the FOCUS framework[61]. We accounted for the correlation between genetically-predicted isoform or gene expression induced by LD and shared prediction weights and control for certain pleiotropic effects. Through Bayesian methods, we estimated the Posterior Inclusion Probability (PIP) for each isoform or gene and defined a credible set of isoforms or genes to explain the signal with 90% confidence.

Gene-set enrichment analysis

We conducted gene-set enrichment analysis using enrichR[37] to investigate the gene ontologies enriched among genes identified by isoTWAS. The analysis queried multiple databases, including GO Biological Process 2023, GO Cellular Component 2023, GO Molecular Function 2023, KEGG 2021 Human, Reactome 2022, and ChEA 2022. ChEA is a curated database of transcription factor targets across multiple ChIP-chip, ChIP-seq, ChIP-PET and DamID assays across a host of tissues and cell-types[77]. We extracted the top 10 most significant terms from each database (FDR-adjusted $P < 0.1$) and calculated the odds ratios to identify over-represented gene functions.

Quantitative trait locus mapping and Bayesian colocalization

For the same GTEx tissues that were used in TWAS/isoTWAS, we mapped *cis*-eQTLs for all genes and isoforms using all SNPs within a 1 Mb interval around the transcription start site. We used ordinary least squares to estimate the allelic effect of the SNP (coded as 0, 1, or 2 copies of the risk allele) on gene or isoform expression, adjusted for the same variables used in the original GTEx analyses (five principal components of the genotype matrix, 30-60 PEER factors depending on sample size for the given tissue, age at death, sex). We estimated the eQTL effect sizes, Wald-type standard errors, and P-values.

Next, using GWAS summary statistics for each of the 12 cancer outcomes, gene- and isoform-eQTL summary statistics, and reference LD for European ancestry individuals from the 1000 Genomes Project[71], we conducted Bayesian colocalization using eCAVIAR[38] to estimate the CoLocalization Posterior Probability (CLPP) that the same SNP is causal for both cancer risk and the gene or isoform expression. We considered a gene or isoform to colocalize with a GWAS-significant locus if three conditions were met: a SNP has a (1) strong effect on cancer risk (GWAS $P < 5 \times 10^{-8}$), (2) strong effect on gene or isoform expression (eQTL $P < 10^{-6}$), and (3) CLPP > 0.01, as is proposed in the original eCAVIAR paper.

Estimation of expression-mediated SNP heritability

We employed mediated expression score regression (MESC)[9] to estimate the proportion of total SNP heritability (h^2) mediated by the *cis*-genetic component of gene or isoform expression levels (h^2_{med}). First, using the genotypes of all SNPs, expression of all genes or isoforms across all 48 tissues in GTEx, and the same covariates used in the eQTL analysis above, we estimated eQTL effect sizes in each tissue using LASSO

regression (`./run_mesc.py --compute-expscore-indiv …`). Next, we meta-analyzed expression scores across the 49 tissues and estimated h^2_{med} for each of the 12 cancer outcomes by using GWAS summary statistics munged to the LD score regression `.sumstats` format. This analysis provided estimates of h^2, h^2_{med}, and h^2_{med}/h^2, each with standard errors and P-values for the hypothesis test comparing to the null value of no heritability.

AUTHORS' DISCLOSURES

No disclosures are reported by the authors.

AUTHORS' CONTRIBUTIONS

YHC: Formal analysis, validation, investigation, visualization, methodology, writing–original draft, writing–review and editing; STH: Formal analysis, writing–review and editing; TH: Formal analysis, writing–review and editing; YY: Formal analysis, validation, visualization, methodology, writing–review and editing; CDH: Methodology, writing–review and editing, funding acquisition; BP: Writing–review and editing, resources, funding acquisition; SL: Writing–review and editing, funding acquisition; AB: Conceptualization, resources, supervision, funding acquisition, investigation, methodology, writing–review and editing

ACKNOWLEDGEMENTS

Funding was provided by NIH CA293419 and CA194393. The funder had no role in study design, data collection, analysis, or interpretation, or writing of the article.

REFERENCES

1. Sud, A., Kinnersley, B. & Houlston, R. S. Genome-wide association studies of cancer: current insights and future perspectives. *Nat Rev Cancer* **17**, 692–704 (2017).
2. Zhang, H. *et al.* Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. *Nat Genet* **52**, 572–581 (2020).
3. Wang, A. *et al.* Characterizing prostate cancer risk through multi-ancestry genome-wide discovery of 187 novel risk variants. *Nat Genet* **55**, 2065–2074 (2023).
4. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet* **47**, 1091–1098 (2015).
5. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet* **48**, 245–252 (2016).
6. Wainberg, M. *et al.* Opportunities and challenges for transcriptome-wide association studies. *Nat Genet* **51**, 592–599 (2019).
7. Mancuso, N. *et al.* Large-scale transcriptome-wide association study identifies new prostate cancer risk regions. *Nat Commun* **9**, 4079 (2018).
8. Gao, G. *et al.* A joint transcriptome-wide association study across multiple tissues identifies candidate breast cancer susceptibility genes. *The American Journal of Human Genetics* **110**, 950–962 (2023).
9. Yao, D. W., O'Connor, L. J., Price, A. L. & Gusev, A. Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nat Genet* **52**, 626–633 (2020).
10. Mostafavi, H., Spence, J. P., Naqvi, S. & Pritchard, J. K. Systematic differences in discovery of genetic effects on gene expression and complex traits. *Nat Genet* **55**, 1866–1875 (2023).
11. Zhang, Y., Qian, J., Gu, C. & Yang, Y. Alternative splicing and cancer: a systematic review. *Sig Transduct Target Ther* **6**, 1–14 (2021).
12. Wang, E. & Aifantis, I. RNA Splicing and Cancer. *Trends in Cancer* **6**, 631–644 (2020).
13. Bhattacharya, A. *et al.* Isoform-level transcriptome-wide association uncovers genetic risk mechanisms for neuropsychiatric disorders in the human brain. *Nat Genet* **55**, 2117–2128 (2023).
14. O'Connell, K. S. *et al.* Genomics yields biological and phenotypic insights into bipolar disorder. 2023.10.07.23296687 Preprint at https://doi.org/10.1101/2023.10.07.23296687 (2024).

15. K, M. *et al.* Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**, (2017).
16. RI, M. *et al.* Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. *Nature genetics* **49**, (2017).
17. O'Mara, T. A. *et al.* Identification of nine new susceptibility loci for endometrial cancer. *Nat Commun* **9**, 3166 (2018).
18. Huyghe, J. R. *et al.* Discovery of common and rare genetic risk variants for colorectal cancer. *Nat Genet* **51**, 76–87 (2019).
19. McKay, J. D. *et al.* Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. *Nat Genet* **49**, 1126–1132 (2017).
20. Phelan, C. M. *et al.* Identification of 12 new susceptibility loci for different histotypes of epithelial ovarian cancer. *Nat Genet* **49**, 680–691 (2017).
21. Schumacher, F. R. *et al.* Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nat Genet* **50**, 928–936 (2018).
22. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
23. Pasaniuc, B. *et al.* Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics* **30**, 2906–2914 (2014).
24. Liu, Y. *et al.* ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. *Am J Hum Genet* **104**, 410–421 (2019).
25. Chakravarty, D. *et al.* OncoKB: A Precision Oncology Knowledge Base. *JCO Precis Oncol* 1–16 (2017) doi:10.1200/PO.17.00011.
26. Suehnholz, S. P. *et al.* Quantifying the Expanding Landscape of Clinical Actionability for Patients with Cancer. *Cancer Discovery* **14**, 49–65 (2024).
27. Carroll, J. S. Mechanisms of oestrogen receptor (ER) gene regulation in breast cancer. *Eur J Endocrinol* **175**, R41-49 (2016).
28. Pulica, R., Cohen-Solal, K. & Lasfar, A. Evaluating the Role of RUNX2 in Cancer and Its Potential as a Therapeutic Target. in *Handbook of Cancer and Immunology* (ed. Rezaei, N.) 1–22 (Springer International Publishing, Cham, 2022). doi:10.1007/978-3-030-80962-1_254-1.

29. Agarwal, N. & Theodorescu, D. The Role of Transcription Factor YY1 in the Biology of Cancer. *Crit Rev Oncog* **22**, 13–21 (2017).
30. Zeng, T., Spence, J. P., Mostafavi, H. & Pritchard, J. K. Bayesian estimation of gene constraint from an evolutionary model with gene features. *Nat Genet* **56**, 1632–1643 (2024).
31. Chen, H., Liu, H. & Qing, G. Targeting oncogenic Myc as a strategy for cancer treatment. *Sig Transduct Target Ther* **3**, 1–7 (2018).
32. Paller, C. J. *et al.* Pan-Cancer Interrogation of MUTYH Variants Reveals Biallelic Inactivation and Defective Base Excision Repair Across a Spectrum of Solid Tumors. *JCO Precis Oncol* e2300251 (2024) doi:10.1200/PO.23.00251.
33. Gupta, S. K., Gallego, C., Johnson, G. L. & Heasley, L. E. MAP kinase is constitutively activated in gip2 and src transformed rat 1a fibroblasts. *Journal of Biological Chemistry* **267**, 7987–7990 (1992).
34. Ikezu, T. *et al.* Bidirectional regulation of c-fos promoter by an oncogenic gip2 mutant of G alpha i2. A novel implication of retinoblastoma gene product. *Journal of Biological Chemistry* **269**, 31955–31961 (1994).
35. Wang, Z. *et al.* Pan-Cancer analysis shows that ACO2 is a potential prognostic and immunotherapeutic biomarker for multiple cancer types including hepatocellular carcinoma. *Front Oncol* **12**, 1055376 (2022).
36. Xu, J., Li, L., Shi, P., Cui, H. & Yang, L. The Crucial Roles of Bmi-1 in Cancer: Implications in Pathogenesis, Metastasis, Drug Resistance, and Targeted Therapies. *Int J Mol Sci* **23**, 8231 (2022).
37. Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128 (2013).
38. Hormozdiari, F. *et al.* Colocalization of GWAS and eQTL Signals Detects Target Genes. *Am J Hum Genet* **99**, 1245–1260 (2016).
39. Chen, H. *et al.* Large-scale cross-cancer fine-mapping of the 5p15.33 region reveals multiple independent signals. *HGG Adv* **2**, 100041 (2021).
40. Bau, D.-T. *et al.* Genetic susceptibility to prostate cancer in Taiwan: A genome-wide association study. *Mol Carcinog* **63**, 617–628 (2024).
41. Backman, J. D. *et al.* Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* **599**, 628–634 (2021).

42. Yandell, M. *et al.* A probabilistic disease-gene finder for personal genomes. *Genome Res* **21**, 1529–1542 (2011).
43. Hu, H. *et al.* VAAST 2.0: Improved Variant Classification and Disease-Gene Identification Using a Conservation-Controlled Amino Acid Substitution Matrix. *Genet Epidemiol* **37**, 622–634 (2013).
44. Wang, X. & Goldstein, D. B. Enhancer Domains Predict Gene Pathogenicity and Inform Gene Discovery in Complex Disease. *Am J Hum Genet* **106**, 215–233 (2020).
45. McKay, J. D. *et al.* Lung cancer susceptibility locus at 5p15.33. *Nat Genet* **40**, 1404–1406 (2008).
46. Rafnar, T. *et al.* Sequence variants at the TERT-CLPTM1L locus associate with many cancer types. *Nat Genet* **41**, 221–227 (2009).
47. Shete, S. *et al.* A Genome-Wide Association Study Identifies Two Novel Susceptible Regions for Squamous Cell Carcinoma of the Head and Neck. *Cancer Res* **80**, 2451–2460 (2020).
48. Lesseur, C. *et al.* Genome-wide association analyses identify new susceptibility loci for oral cavity and pharyngeal cancer. *Nat Genet* **48**, 1544–1550 (2016).
49. Petersen, G. M. *et al.* A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33. *Nat Genet* **42**, 224–228 (2010).
50. James, M. A., Vikis, H. G., Tate, E., Rymaszewski, A. L. & You, M. CRR9/CLPTM1L regulates cell survival signaling and is required for Ras transformation and lung tumorigenesis. *Cancer Res* **74**, 1116–1127 (2014).
51. Parashar, D. *et al.* Targeted biologic inhibition of both tumor cell-intrinsic and intercellular CLPTM1L/CRR9-mediated chemotherapeutic drug resistance. *npj Precis. Onc.* **5**, 1–13 (2021).
52. Kim, N. W. *et al.* Specific association of human telomerase activity with immortal cells and cancer. *Science* **266**, 2011–2015 (1994).
53. Ge, M. *et al.* Functional evaluation of TERT-CLPTM1L genetic variants associated with susceptibility of papillary thyroid carcinoma. *Sci Rep* **6**, 26037 (2016).
54. Gholami, M. *et al.* Association of miRNA targetome variants in LAMC1 and GNB3 genes with colorectal cancer and obesity. *Cancer Med* **11**, 3923–3938 (2022).

55. Hazelwood, E. *et al.* Integrating multi-tissue expression and splicing data to prioritise anatomical subsite- and sex-specific colorectal cancer susceptibility genes with therapeutic potential. 2024.09.10.24313450 Preprint at https://doi.org/10.1101/2024.09.10.24313450 (2024).
56. Qin, Y., Rodin, S., Simonson, O. E. & Hollande, F. Laminins and cancer stem cells: Partners in crime? *Semin Cancer Biol* **45**, 3–12 (2017).
57. Galatenko, V. V., Maltseva, D. V., Galatenko, A. V., Rodin, S. & Tonevitsky, A. G. Cumulative prognostic power of laminin genes in colorectal cancer. *BMC Med Genomics* **11**, 9 (2018).
58. Bogdanova, N., Helbig, S. & Dörk, T. Hereditary breast cancer: ever more pieces to the polygenic puzzle. *Hereditary Cancer in Clinical Practice* **11**, 12 (2013).
59. Antoniou, A. C. *et al.* A locus on 19p13 modifies risk of breast cancer in BRCA1 mutation carriers and is associated with hormone receptor-negative breast cancer in the general population. *Nat Genet* **42**, 885–892 (2010).
60. Bolton, K. L. *et al.* Common variants at 19p13 are associated with susceptibility to ovarian cancer. *Nat Genet* **42**, 880–884 (2010).
61. Mancuso, N. *et al.* Probabilistic fine-mapping of transcriptome-wide association studies. *Nat Genet* **51**, 675–682 (2019).
62. Bhattacharya, A. *et al.* Best practices for multi-ancestry, meta-analytic transcriptome-wide association studies: Lessons from the Global Biobank Meta-analysis Initiative. *Cell Genomics* **2**, 100180 (2022).
63. Kachuri, L. *et al.* Gene expression in African Americans, Puerto Ricans and Mexican Americans reveals ancestry-specific patterns of genetic architecture. *Nat Genet* **55**, 952–963 (2023).
64. Bhattacharya, A. *et al.* A framework for transcriptome-wide association studies in breast cancer in diverse study populations. *Genome Biol* **21**, 42 (2020).
65. Keys, K. L. *et al.* On the cross-population generalizability of gene expression prediction models. *PLoS Genet* **16**, e1008927 (2020).
66. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**, 417–419 (2017).
67. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* **47**, D766–D773 (2019).

68. Love, M. I. *et al.* Tximeta: Reference sequence checksums for provenance identification in RNA-seq. *PLOS Computational Biology* **16**, e1007664 (2020).
69. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47 (2015).
70. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).
71. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
72. Lindström, S. *et al.* Genome-wide analyses characterize shared heritability among cancers and identify novel cancer susceptibility regions. *JNCI: Journal of the National Cancer Institute* **115**, 712–732 (2023).
73. Bhattacharya, A. isoTWAS models for 48 GTEx models (04/2024). Zenodo https://doi.org/10.5281/zenodo.11048201 (2024).
74. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* **33**, 1–22 (2010).
75. Endelman, J. B. Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *The Plant Genome* **4**, (2011).
76. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A Simple New Approach to Variable Selection in Regression, with Application to Genetic Fine Mapping. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **82**, 1273–1300 (2020).
77. Lachmann, A. *et al.* ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics* **26**, 2438 (2010).

FIGURES

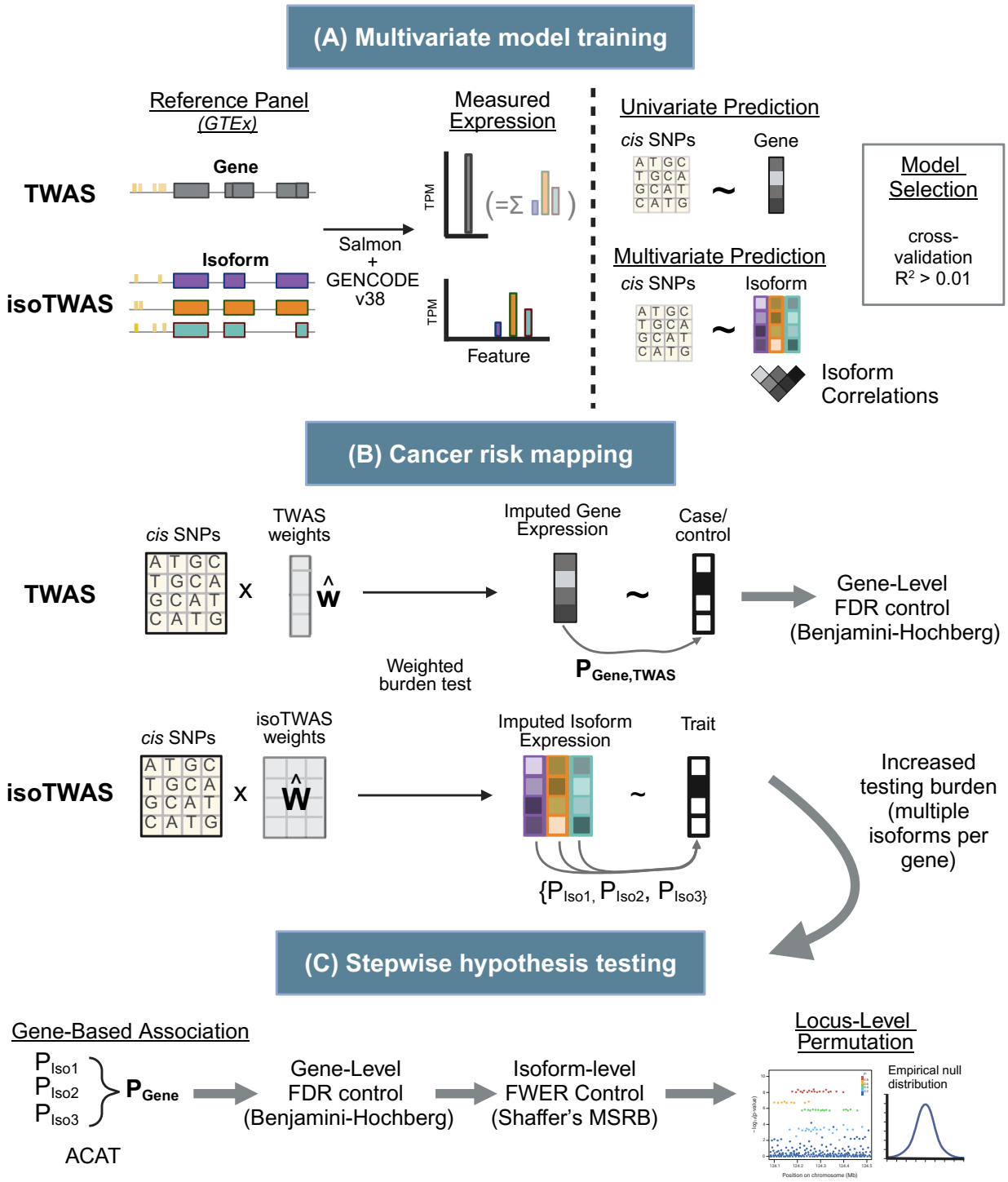

Figure 1: *Overview of isoform- and gene-level TWAS*. **(A)** Using multi-tissue GTEx data, isoform- and gene-level expression was quantified using Salmon and GENCODE v38 annotations. Univariate and multivariate predictive models for gene and isoform expression were trained. **(B)** Using the weighted burden test, predictive models of gene and isoform expression were used to map gene- and isoform-level cancer risk associations. For gene-level TWAS associations, FDR is controlled to 5% via Benjamini-Hochberg. **(C)** As isoTWAS presents an increased testing burden, stepwise hypothesis testing is conducted. Isoform P-values are aggregated to the gene-level via aggregated Cauchy aggregation. FDR is controlled at 5% via Benjamini-Hochberg procedure on gene-level P-values. For significantly-associated genes, FWER is controlled to 5% for correlated isoform associations using Shaffer's MSRB procedure. Locus-level permutation is conducted to control of local LD patterns. Modified from Bhattacharya et al 2023.

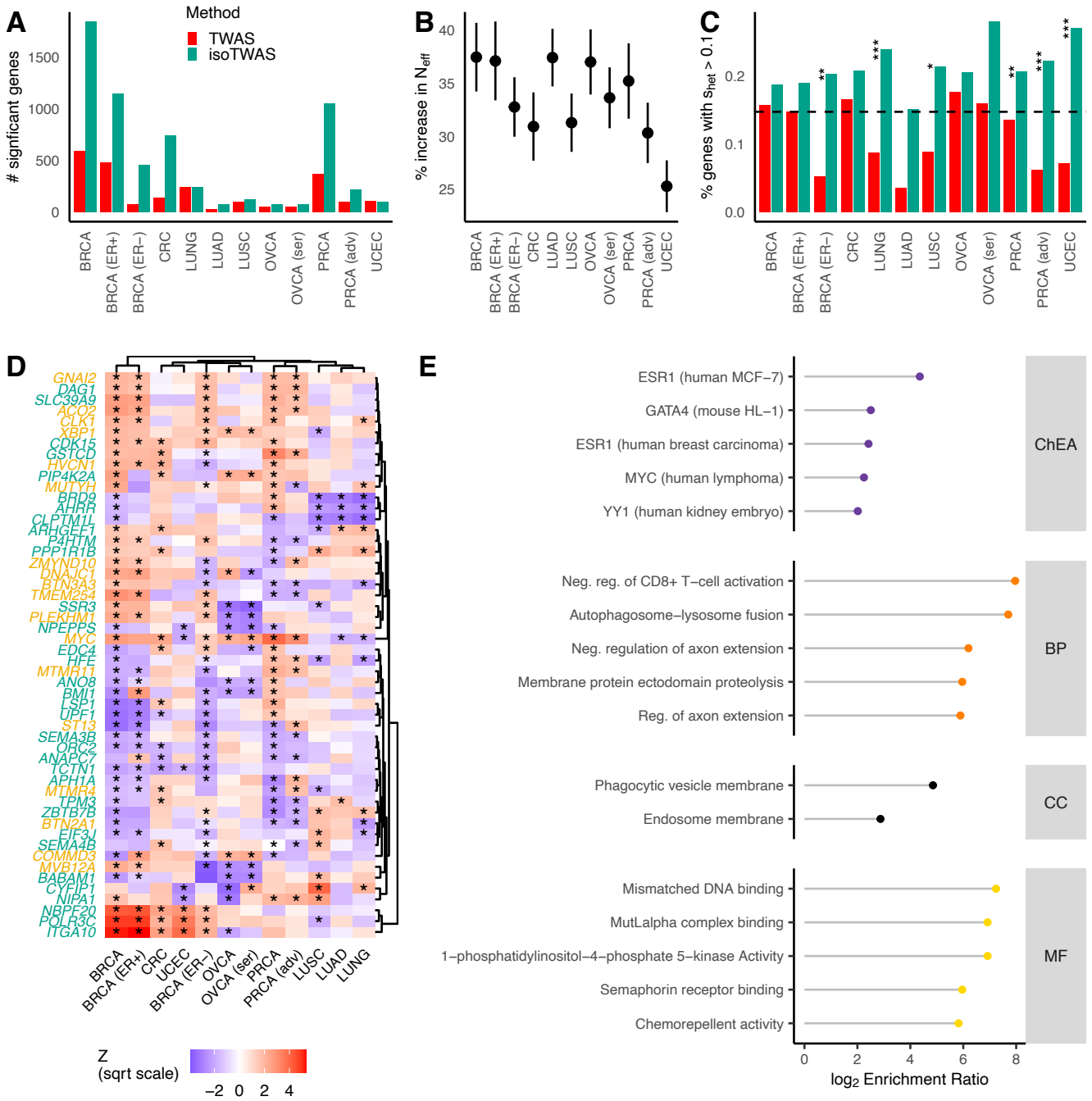

Figure 2: *Isoform-level analysis identifies substantially more gene associations with cancer risk across 12 outcomes*. **(A)** The number of unique transcriptome-wide significant genes identified with TWAS (red) and isoTWAS (green). **(B)** Percent increase in effective sample size with Wald-type 95% confidence interval using jackknife standard errors. **(C)** Proportion of transcriptome-wide significant genes identified with s_{het} > 0.1. Asterisks indicate FDR-adjusted P-value of $\chi 2$ test for enrichment ratio of high pLI genes among all transcriptome-wide significant genes across method. (*) indicates FDR-adjusted P < 0.05, (**) P < 0.01, (***) P < 0.005. Black line shows the genome-wide proportion of genes with s_{het} > 0.1. **(D)** Scaled gene-level isoTWAS Z-score for 52 genes with isoform-level risk associations with at least 5 cancer outcomes across 12 outcomes. Genes are marked in green if no gene-level risk associations with any cancer outcome, and asterisk is shown if the gene association is significant for the given cancer outcome. **(E)** Log-enrichment ratio (X-axis) of over-represented gene ontologies (Y-axis), across ChIP- seq identified transcription factor targets (ChEA), biological process (BP), cellular component (CC), or molecular function (MF) pathways for 34 genes with isoform-level risk associations with at least 5 cancer outcomes and no TWAS associations.

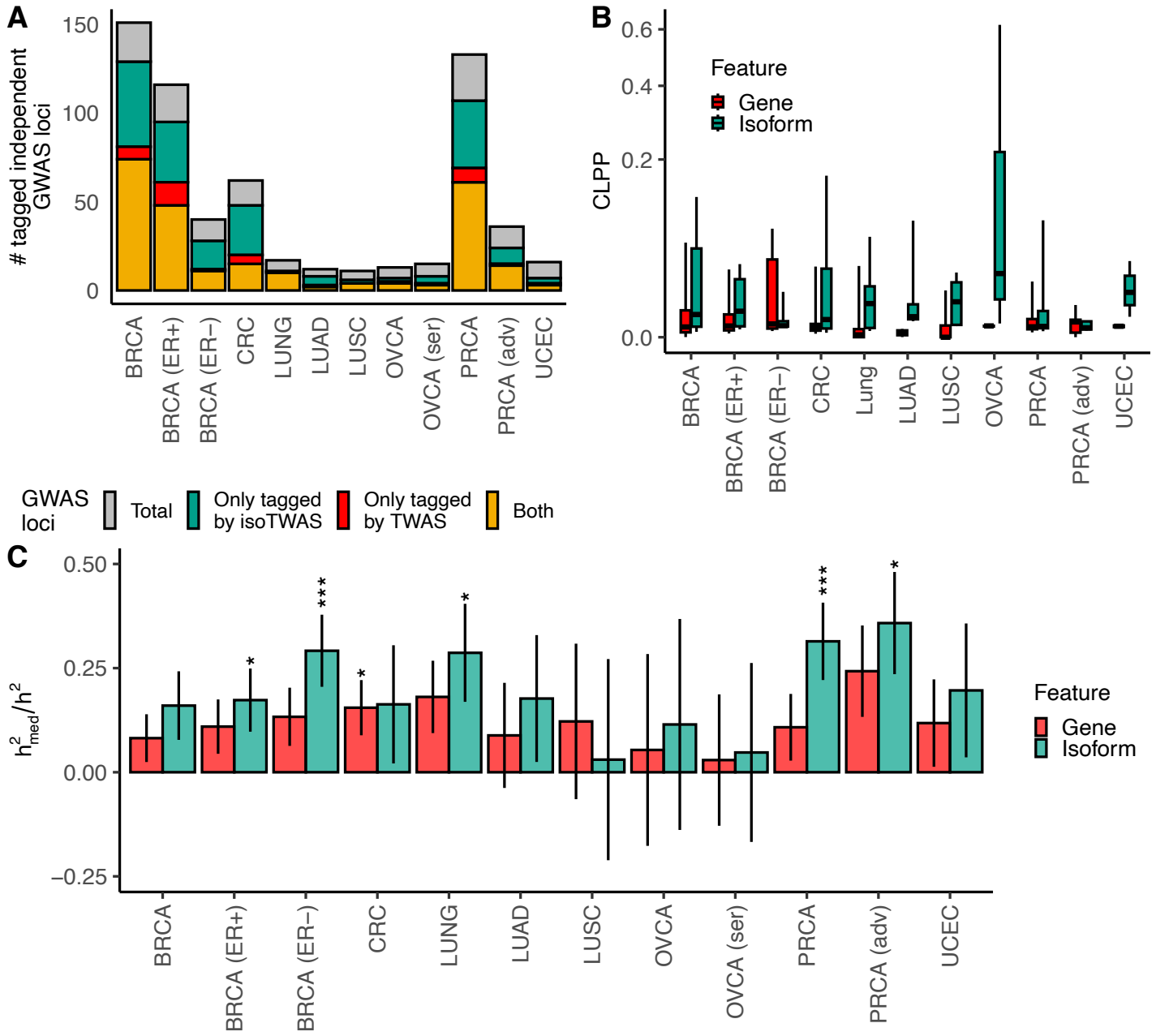

Figure 3: *Isoform expression potentially mediates far more GWAS signal than gene expression*. **(A)** The number of independent GWAS loci tagged by TWAS (red) or isoTWAS (green) or both in common (gold). **(B)** Boxplots of CoLocalization Posterior Probability (CLPP, Y-axis, square-root axis) of GWAS and gene expression QTL (red) and isoform expression QTL (green) for genomic loci with a GWAS $P < 5x10^{-8}$ and QTL $P < 10^{-6}$. **(C)** Ratio of gene- (red) and isoform-level (green) expression mediated heritability (h^2_{med}) and total SNP heritability (h^2) with standard errors. Asterisks indicate FDR-adjusted P-value of Wald-type Z tests of $h^2_{med}/h^2 = 0$. (*) indicates FDR-adjusted $P < 0.05$, (**) $P < 0.01$, (***) $P < 0.005$.

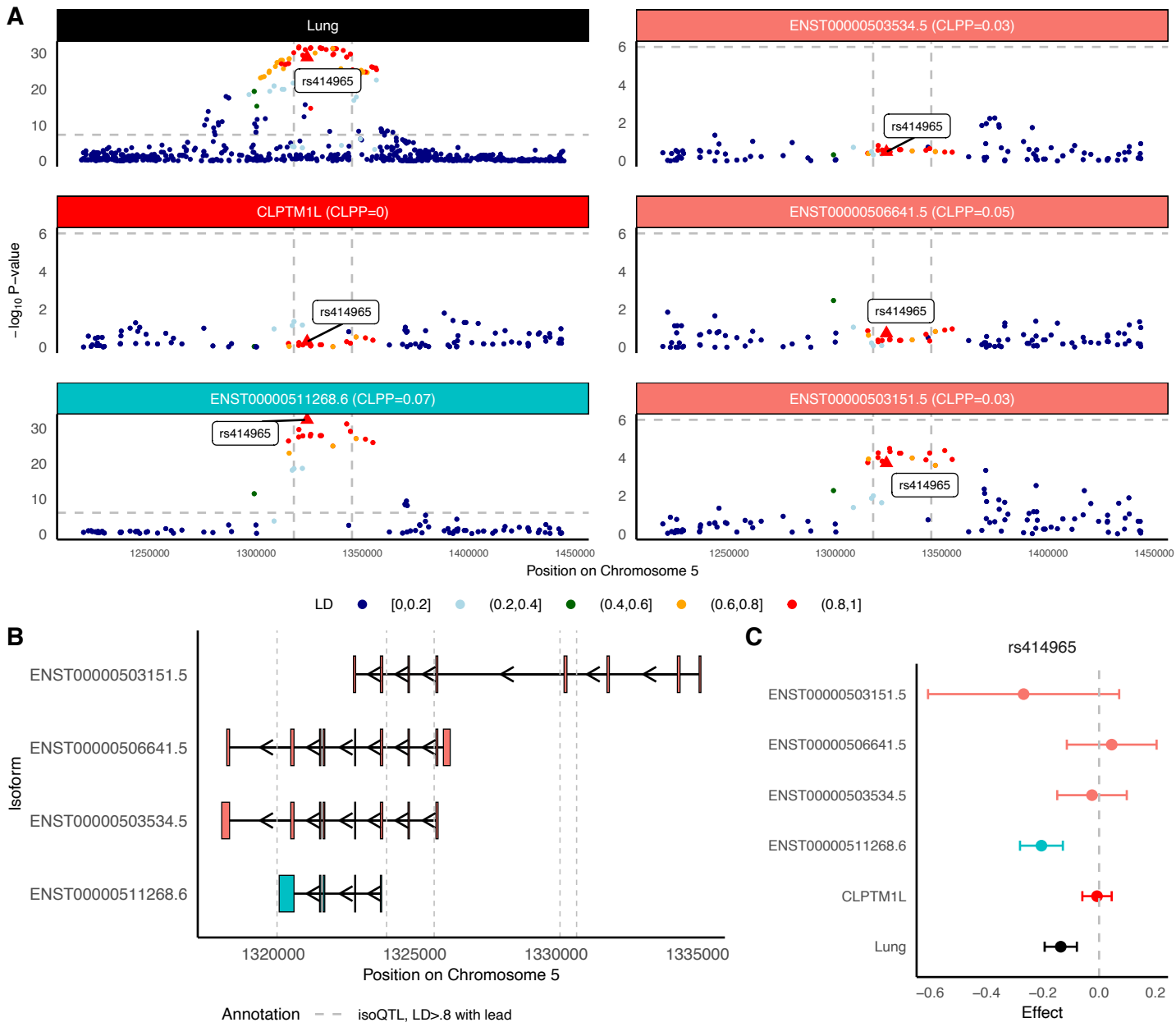

Figure 4: *CLPTM1L isoforms may mediate lung cancer risk GWAS locus at Chromosome 5p15.33.* **(A)** Manhattan plot of GWAS effects, *CLPTM1L* gene-eQTLs, and isoform-eQTLs for all significantly associated isoforms of *CLPTM1L*, colored by LD to rs414965, the lead isoQTL for ENST00000511268.6. **(B)** Transcript structure of significantly associated isoforms of *CLPTM1L*. Vertical lines indicated significant isoQTLs of LD > 0.8 to rs414965, strongest isoQTL of *CLPTM1L* isoforms. **(C)** SNP effect sizes on lung cancer risk (black), *CLPTM1L* gene expression (red), ENST00000511268.6 isoform expression (blue), and other expression of other isoforms (peach) for rs414965.

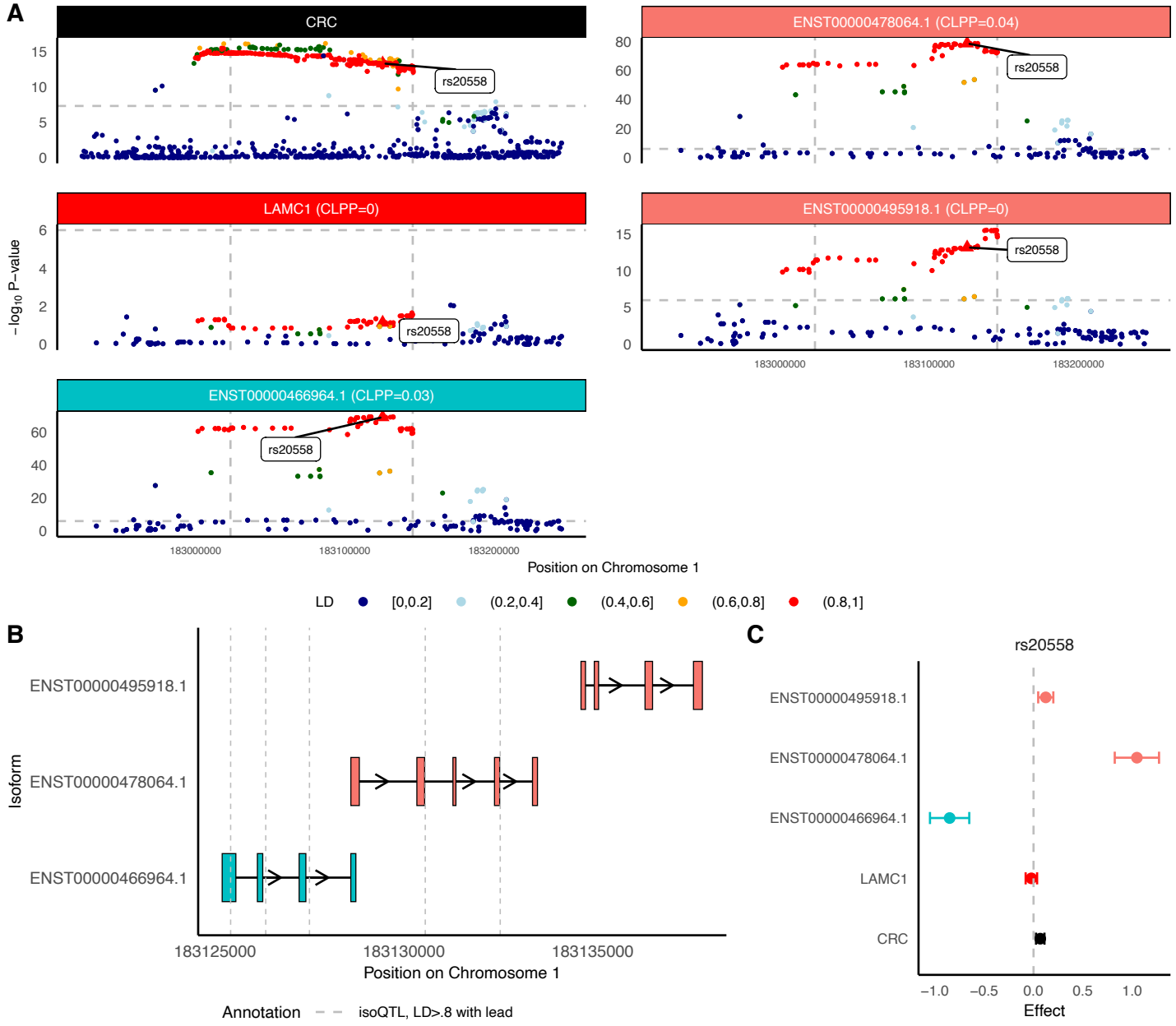

Figure 5: *LAMC1 isoforms may mediate colorectal cancer GWAS locus at Chromosome 1q25.3*. **(A)** Manhattan plot of GWAS effects, *LAMC1* gene-eQTLs, and isoform-eQTLs for all significantly associated isoforms of *LAMC1*, colored by LD to rs20558, the lead isoQTL for ENST00000466964.1. **(B)** Transcript structure of significantly associated isoforms of *BABAM1*. Vertical lines indicated significant isoQTLs of LD > 0.8 to rs20558, strongest isoQTL of *LAMC1* isoforms. **(C)** SNP effect sizes on colorectal cancer risk (black), *LAMC1* gene expression (red), ENST00000466964.1 isoform expression (blue), and other expression of other isoforms (peach) for rs20558.

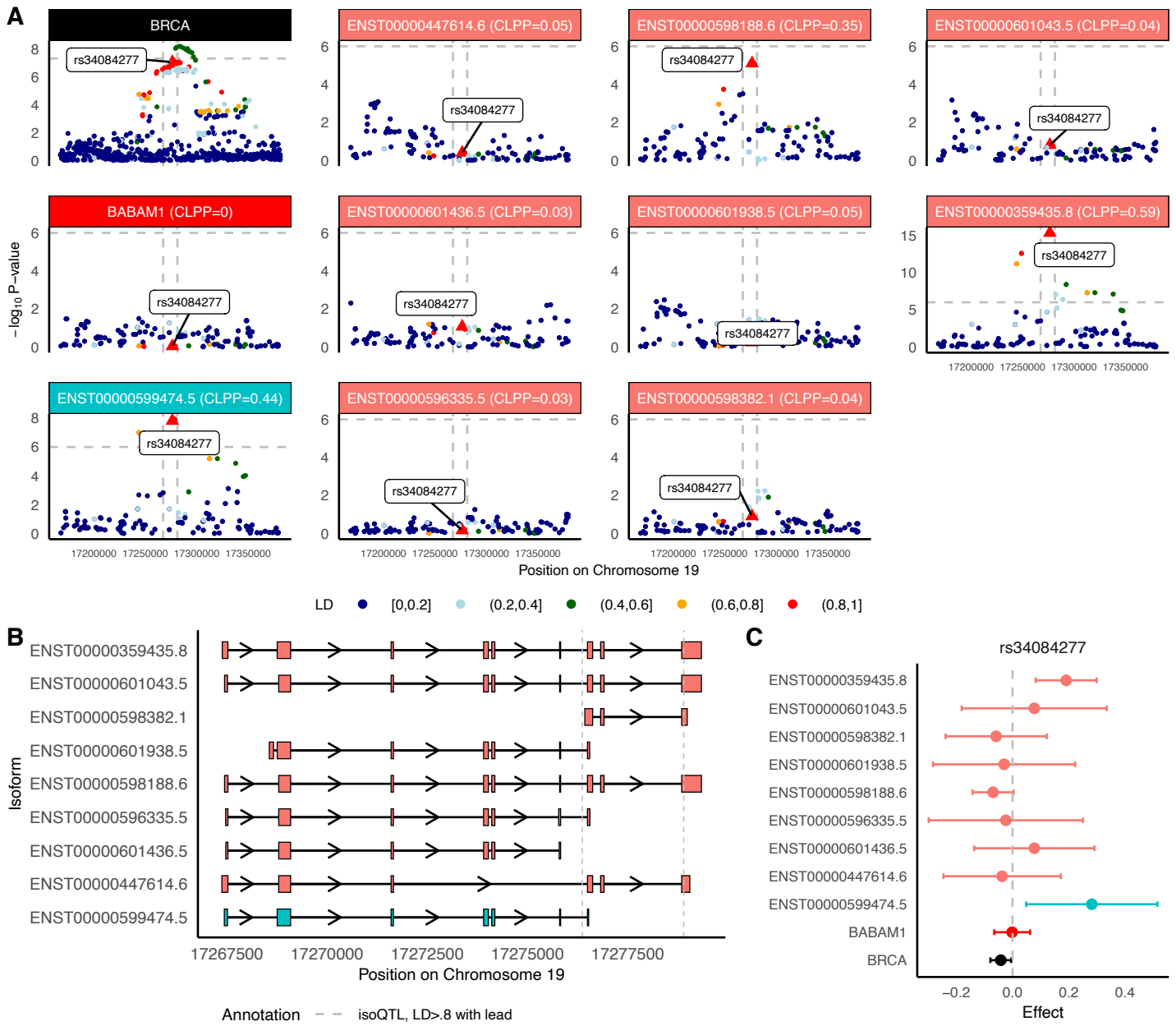

Figure 6: *BABAM1 isoforms may mediate breast cancer risk GWAS locus at Chromosome 19p13.11*. **(A)** Manhattan plot of GWAS effects, *BABAM1* gene-eQTLs, and isoform-eQTLs for isoforms of *BABAM1*, either prioritized through isoTWAS or with an isoQTL with P < 1e-6. **(B)** Transcript structure of *BABAM1*. Vertical lines indicated significant isoQTLs of LD > 0.8 to rs34084277, strongest isoQTL of *BABAM1* isoforms. **(C)** SNP effect sizes on breast cancer risk (black), *BABAM1* gene expression (red), expression of isoTWAS-prioritized isoforms (blue), and expression of other isoforms (peach) for rs34084277 (lead isoQTL).