Multimodal analysis of RNA sequencing data powers discovery of complex trait genetics

Daniel Munro[1,2,3], Nava Ehsan[3], Seyed Mehdi Esmaeili-Fard[2], Alexander Gusev[4,*], Abraham A. Palmer[1,5,*], Pejman Mohammadi[2,3,6,*]

1 Department of Psychiatry, UC San Diego, La Jolla, CA, USA
2 Center for Immunity and Immunotherapies, Seattle Children's Research Institute, Seattle, WA, USA
3 Department of Integrative Structural and Computational Biology, Scripps Research, La Jolla, CA, USA
4 Division of Population Sciences, Dana-Farber Cancer Institute and Harvard Medical School, Boston, MA, USA
5 Institute for Genomic Medicine, UC San Diego, La Jolla, CA, USA
6 Department of Pediatrics, University of Washington School of Medicine, Seattle, WA, USA
* Corresponding authors: pejmanm@uw.edu, aap@ucsd.edu, alexander_gusev@dfci.harvard.edu

Abstract

Transcriptome data is commonly used to understand genome function via quantitative trait loci (QTL) mapping and to identify the molecular mechanisms driving genome wide association study (GWAS) signals through colocalization analysis and transcriptome-wide association studies (TWAS). While RNA sequencing (RNA-seq) has the potential to reveal many modalities of transcriptional regulation, such as various splicing phenotypes, such studies are often limited to gene expression due to the complexity of extracting and analyzing multiple RNA phenotypes. Here, we present Pantry (Pan-transcriptomic phenotyping), a framework to efficiently generate diverse RNA phenotypes from RNA-seq data and perform downstream integrative analyses with genetic data. Pantry currently generates phenotypes from six modalities of transcriptional regulation (gene expression, isoform ratios, splice junction usage, alternative TSS/polyA usage, and RNA stability) and integrates them with genetic data via QTL mapping, TWAS, and colocalization testing. We applied Pantry to Geuvadis and GTEx data, and found that 4,768 of the genes with no identified expression QTL in Geuvadis had QTLs in at least one other transcriptional modality, resulting in a 66% increase in genes over expression QTL mapping. We further found that QTLs exhibit modality-specific functional properties that are further reinforced by joint analysis of different RNA modalities. We also show that generalizing TWAS to multiple RNA modalities (xTWAS) approximately doubles the discovery of unique gene-trait associations, and enhances identification of regulatory mechanisms underlying GWAS signal in 42% of previously associated gene-trait pairs. We provide the Pantry code, RNA phenotypes from all Geuvadis and GTEx samples, and xQTL and xTWAS results on the web.

Introduction

RNA sequencing is used to quantify transcriptomic activity, and can be combined with genotyping to detect heritable differences in gene regulation. This quantification often includes only total gene expression and, less often, some form of alternative splicing phenotype, such as intron excision rates, resulting in expression quantitative trait loci (**eQTLs**) and splice QTLs (**sQTLs**) or predicted expression models. These molecular phenotypes can provide evidence for mechanisms by which heritable differences in gene regulation serve as the molecular intermediates between GWAS association signals and complex traits[1,2]. Other forms of transcriptomic variation, such as alternative transcription start site (**TSS**), alternative polyadenylation (**polyA**), and splice isoform ratios, have been found to explain an additional portion[3,4]. Importantly, all of these phenotypes are based on RNA-seq data but require multiple different analytic methods.

Methods and resources already exist to identify genetically driven associations between these RNA phenotypes and complex traits[5–8]. However, it is common that only gene expression is examined because of the significant extra effort needed to obtain other RNA phenotypes. Difficulties include data formatting issues, software dependencies, post-processing, computational resources, lack of field expertise, and other practical considerations. Another challenge is the statistical complexity of interpreting these correlated RNA phenotypes and downstream results in aggregate. For example, it is not straightforward to distinguish a case where two genetic association signals in different RNA modalities reveal two biological mechanisms from a case where a single mechanism is reflected in two related RNA modalities.

We present Pantry, a framework for pan-transcriptomic phenotyping that streamlines the quantification of multiple RNA modalities and their use in downstream applications, including molecular QTL (**xQTL**) mapping and transcriptome-wide association studies (**TWAS**). We apply all of this to 50 human tissue datasets and demonstrate that when TWAS is generalized to include multiple RNA modalities (**xTWAS**) there is a substantial increase in the number of significant gene-trait associations, and improved specification of the most relevant RNA phenotype.

Results

We developed Pantry, an end-to-end framework for multimodal analysis of RNA-seq data from populations for genomic interpretation (**Figure 1A**). Currently, Pantry encompasses six modalities of regulatory variation. Two of these, total gene expression and RNA stability, result in one phenotype per gene, while the other four can produce multiple molecular phenotypes per gene, such as relative abundance of each unique transcript isoform. We generated data on these six modalities of transcriptome regulation for 445 lymphoblastoid cell line (LCL) samples in Geuvadis[9] and all 17,350 samples across 54 tissues in the GTEx Project V8 release[1]. We limited our analysis to protein-coding genes and lncRNAs. We generated 204,273 phenotypes

per sample, spanning 25,657 genes in Geuvadis data (**Table 1**, **Figure 1B, C**), and similar figures in individual GTEx tissues (**Supp Figure S1**).

Modality	**Method**	**No. phenotypes produced**	**No. genes with phenotypes**	**No. phenotypes per gene (mean ± S.D.)**
Total gene expression	kallisto[10]	25,311	25,311	1
Isoform ratio	kallisto, per-isoform count over sum of isoform counts per gene	34,163	12,344	2.8 ± 1.5
Intron excision ratio	Count intron junctions using regtools[11], cluster using leafcutter[12]	62,870	11,901	5.3 ± 3.8
Alternative TSS	Generate alternative TSS annotations with txrevise[3], quantify with kallisto	39,632	10,214	3.9 ± 2.1
Alternative polyA	Generate alternative polyA annotations with txrevise, quantify with kallisto	32,411	9,100	3.6 ± 1.9
RNA stability	featureCounts[13], constitutive exon read count to intron read count ratio per gene[14]	9,886	9,886	1
Total	N/A	204,273	25,657	8.0 ± 8.3

Table 1. Default RNA modalities included in Pantry and statistics for its application to the Geuvadis dataset. Genes include protein-coding genes and lncRNAs.

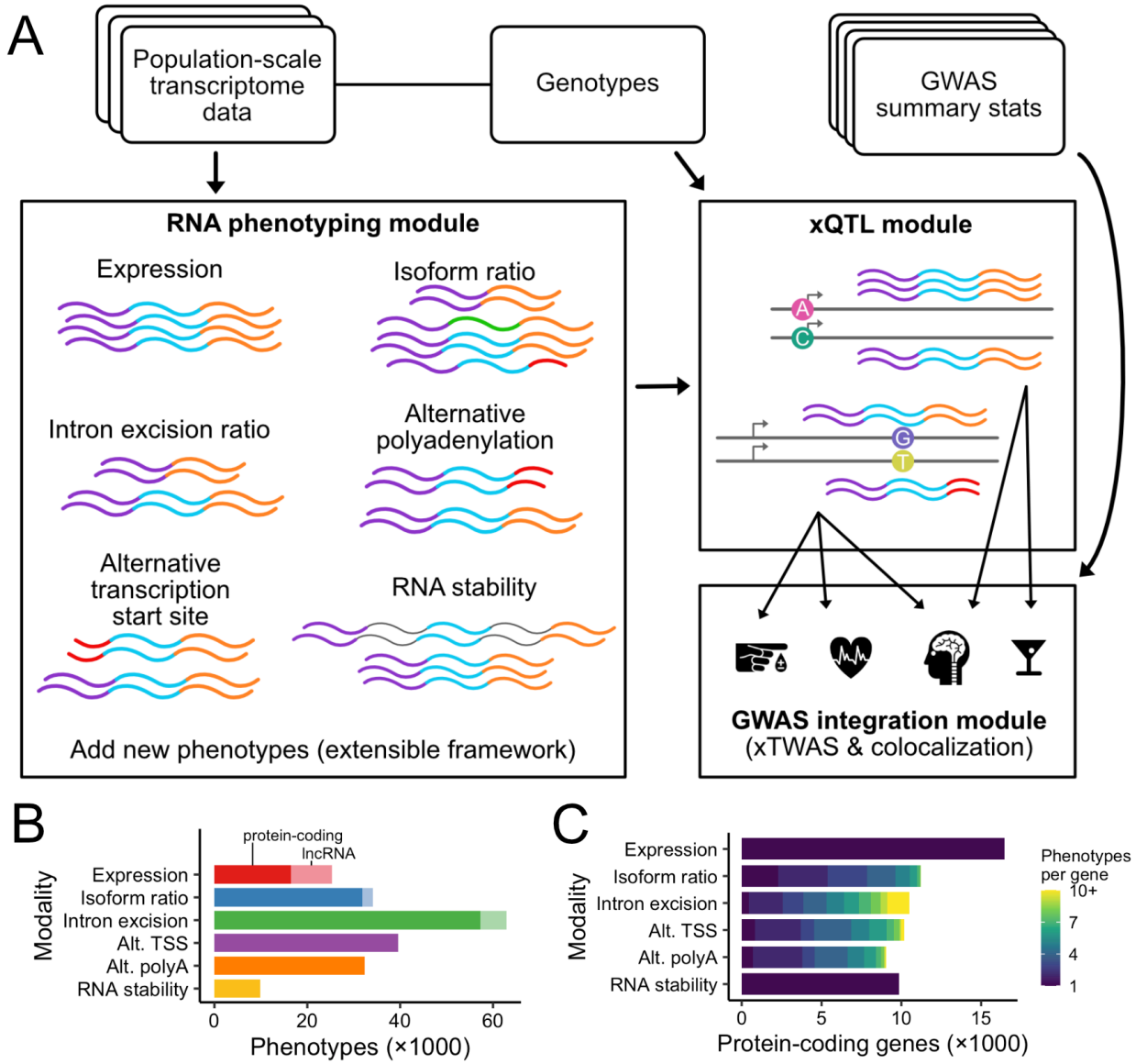

Figure 1. Pantry's multimodal RNA phenotyping. A) Flow chart of the Pantry procedure. B) Number of phenotypes extracted per modality for the Geuvadis dataset. The shading indicates the gene type: protein-coding gene or lncRNA. C) Number of protein-coding genes represented by the phenotypes in B, and the number of phenotypes extracted per gene in each modality.

Applying RNA phenotypes to genetic analyses

Mapping xQTLs across multiple modalities increases xGene discovery

To identify genetic determinants of individual transcriptome phenotypes generated by Pantry, we developed Pheast (PHEnotype Application STreamlined). Pheast uses an approach previously used for splice QTL mapping to simultaneously map cis-QTLs across all six transcriptome

modalities; we refer to this as cross-modality mapping. Specifically, a stepwise regression procedure that is used to identify conditionally independent cis-QTLs can be applied to grouped phenotypes, such as multiple splice phenotypes per gene[15,16]. But phenotypes of different modalities could also be correlated and produce redundant xQTLs, such as when alternative splicing (measured as intron excision ratio) alters the isoform ratios or total gene expression estimates. Pantry Pheast combines the sets of phenotypes across all modalities and maps cis-QTLs with stepwise regression, considering all phenotypes per gene as a single group. All xQTL results hereafter refer to those from this cross-modality cis-xQTL mapping strategy unless otherwise noted.

Using the 445 Geuvadis samples, we identified 21,045 conditionally independent xQTLs for 11,983 genes across the six studied modalities. Expression QTLs were the most abundant, with eQTLs found for 7,215 genes, which was more than 3.2 times greater than isoform ratio, which was the second-most abundant xQTL group (**Figure 2A**). However, for 4,768 genes with no identified eQTL, we found xQTLs in at least one of the other modalities. This represents a 66% increase in the number of xQTL genes (xGenes), highlighting the utility of analyzing multiple modalities of transcriptional regulation. Multiple conditionally independent xQTLs were found for 42.7% of xGenes (**Figure 2B**). The xQTLs for each gene were ranked in order of detection by stepwise regression. The proportion from each modality varied across ranks such that stronger xQTLs were most likely to be for expression, and subsequent xQTLs were more likely to be for isoform ratio or intron excision ratio (**Figure 2C**). This trend could be influenced by the relative strength of the true genetic signals in each modality, the power to detect the signals with each method, and differences in the number of phenotypes per gene.

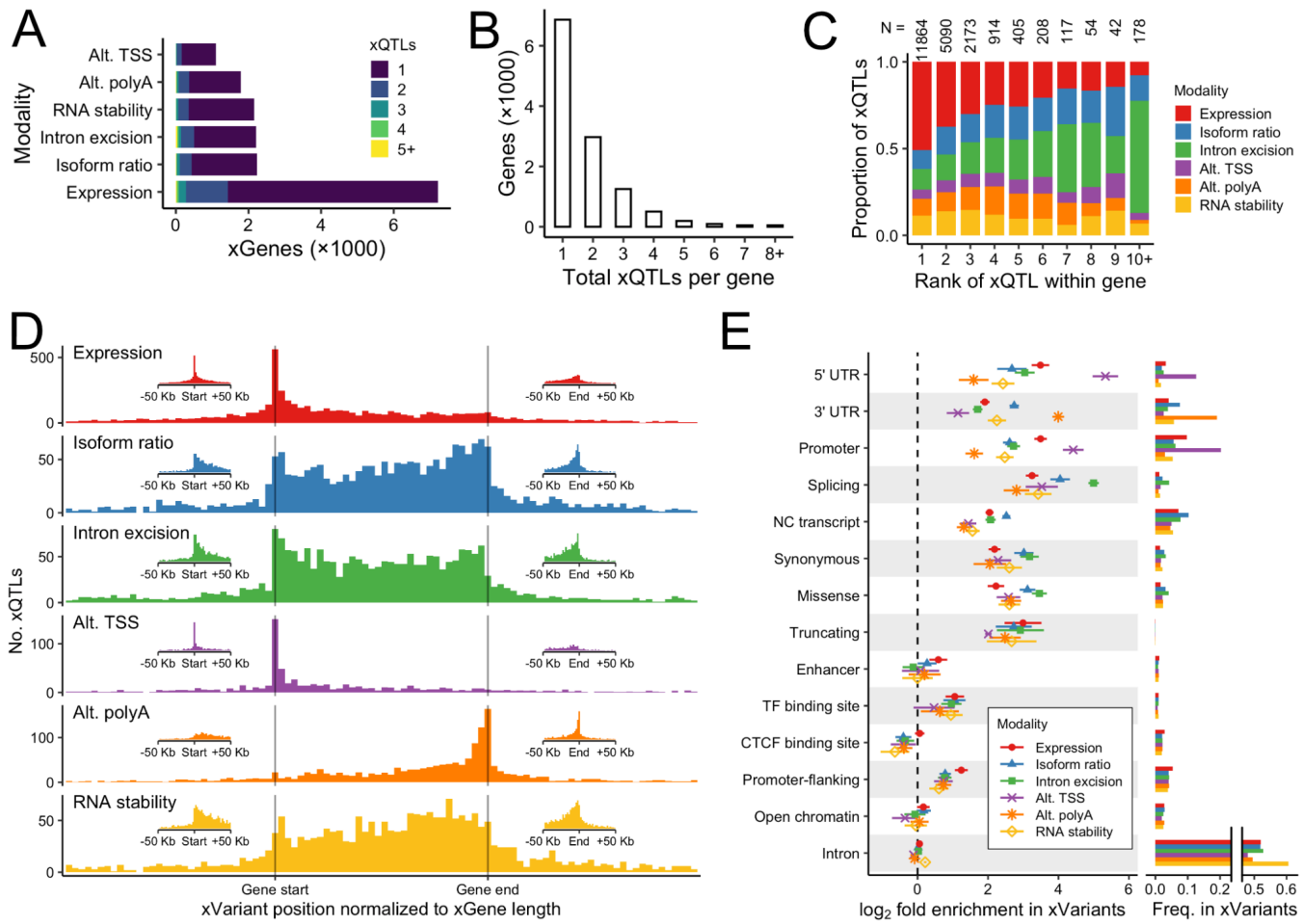

Figure 2. Multimodal xQTL mapping. A) For each modality, the number of xGenes found in Geuvadis, colored according to the number of xQTLs found for the gene. These xQTLs were produced from a single cross-modality mapping, and are grouped by modality here for visual comparison. B) Total xQTLs per xGene in Geuvadis, summed over modalities. C) For each xGene, Geuvadis xQTLs were ranked by association strength, and the proportion of xQTLs per rank that belong to each modality is shown. The number of xQTLs in each column are shown above the columns. D) Location of xQTLs in Geuvadis, relative to their xGene. The genomic coordinates of each xGene and that gene's xQTLs were linearly transformed such that the gene starts and ends are aligned on the x-axis. Only the 65% of xQTLs within one gene's length of the xGene start or end are shown. The insets show the distribution of xQTL positions within 50 Kb of the xGene start or end, without normalizing by gene length. The two insets per modality have the same y-axis scale. E) Left: enrichment of functional annotations in the top variants of each xQTL for each RNA modality, relative to all variants tested for xQTLs for each modality, for all GTEx tissues. Points and horizontal segments show mean and standard deviation across tissues, respectively. Annotation categories are ordered by decreasing variance of their six $\log_2$ enrichment means. Right: Proportion of xVariants in each modality that have each annotation. Variants can be assigned more than one annotation.

We similarly mapped xQTLs for each of 49 GTEx tissues, separately per modality (**Supp Table S1**) and with cross-modality mapping (**Supp Table S2**). We discovered comparable numbers of xQTLs as for Geuvadis, which varied across tissues due to factors such as sample size, but generally found non-expression xQTLs in thousands of genes per tissue for which no eQTLs were found in our data, resulting in a 71% increase of xGenes over eGenes alone on average (**Supp Figure S2**).

To measure concordance of xQTLs between independent datasets, for each modality we identified the strongest xQTL per xGene in Geuvadis, and extracted the associations for the same variant-RNA phenotype pairs, if tested, in GTEx EBV-transformed lymphocytes (LCL). We found that the regression slopes were consistent in both direction and magnitude between the two datasets, with Pearson's correlation coefficients ranging from 0.80 to 0.89 per modality and mean Deming regression slope of 1.005 (**Supp Figure S3**).

Location and functional effect of xVariants reflect their associated modality

While we used the same cis- window of ±1Mb from the transcription start site to map xQTLs for all six modalities, we found that the location of the mapped xQTLs relative to their xGene varies strongly depending on the modality (**Figure 2D**, **Supp Figure S4**). As expected, the distributions of expression and alternative TSS xQTLs peak around the start site, while the distribution of alternative polyA xQTLs peak around the end. Isoform ratio, intron excision ratio, and RNA stability xQTLs are more uniformly distributed across the length of their genes.

We examined functional annotations for each xQTL top variant (xVariant) to identify which annotations were most enriched in each RNA modality. Results were largely in line with expectations. For example, splicing annotations were most enriched in the intron excision ratio xVariants, 5' UTR variants and promoters were most enriched in alternative TSS xVariants, and 3' UTR variants were most enriched in alternative polyA xVariants (**Figure 2E**). Expression was the second-most enriched modality in 5' UTR and promoter variants, and the most enriched modality for promoter-flanking variants. These enrichment levels were largely consistent across GTEx tissues.

Cross-modality mapping improves quality and interpretation of xQTLs

We analyzed the impact of Pantry's cross-modality xQTL mapping strategy compared to the conventional method of mapping conditionally independent QTLs separately per modality. Cross-modality mapping resulted in fewer total xQTLs per gene on average (1.76 in Geuvadis) compared to 2.94 for separate-modality mapping (**Figure 3A**). This general trend is expected and desirable because the goal of cross-modality mapping is to eliminate correlated signals. Notably, we only observe a slight decrease (10.4%) in the total number of xGenes in spite of the 46.4% decrease in the total number of xQTLs (**Figure 3B**). Looking at individual modalities, however, we see a drastic drop (median 39.2%) in the number of xGenes (**Figure 3B**). This pattern points to deconvolution of confounding xQTL effects observed in multiple modalities by the cross-modality mapping strategy. To this end, we looked specifically at the consistency of expression QTL effect sizes. Using data from GTEx subcutaneous adipose tissue, we measured

allelic fold change (**aFC**) from gene expression data and again from allele-specific expression (**ASE**) data. These two measurements of cis-regulatory effect size are largely affected by independent sources of noise and as such allow us to gauge the quality of mapped cis-eQTLs[17]. The Pearson correlation between the two aFC measures was slightly higher for cross-modality mapping (r = 0.69, 95% CI [0.673, 0.697]) than for separate-modality mapping (r = 0.64, 95% CI [0.631, 0.651]), suggesting a refinement of eQTL signals (**Supp Figure S5**).

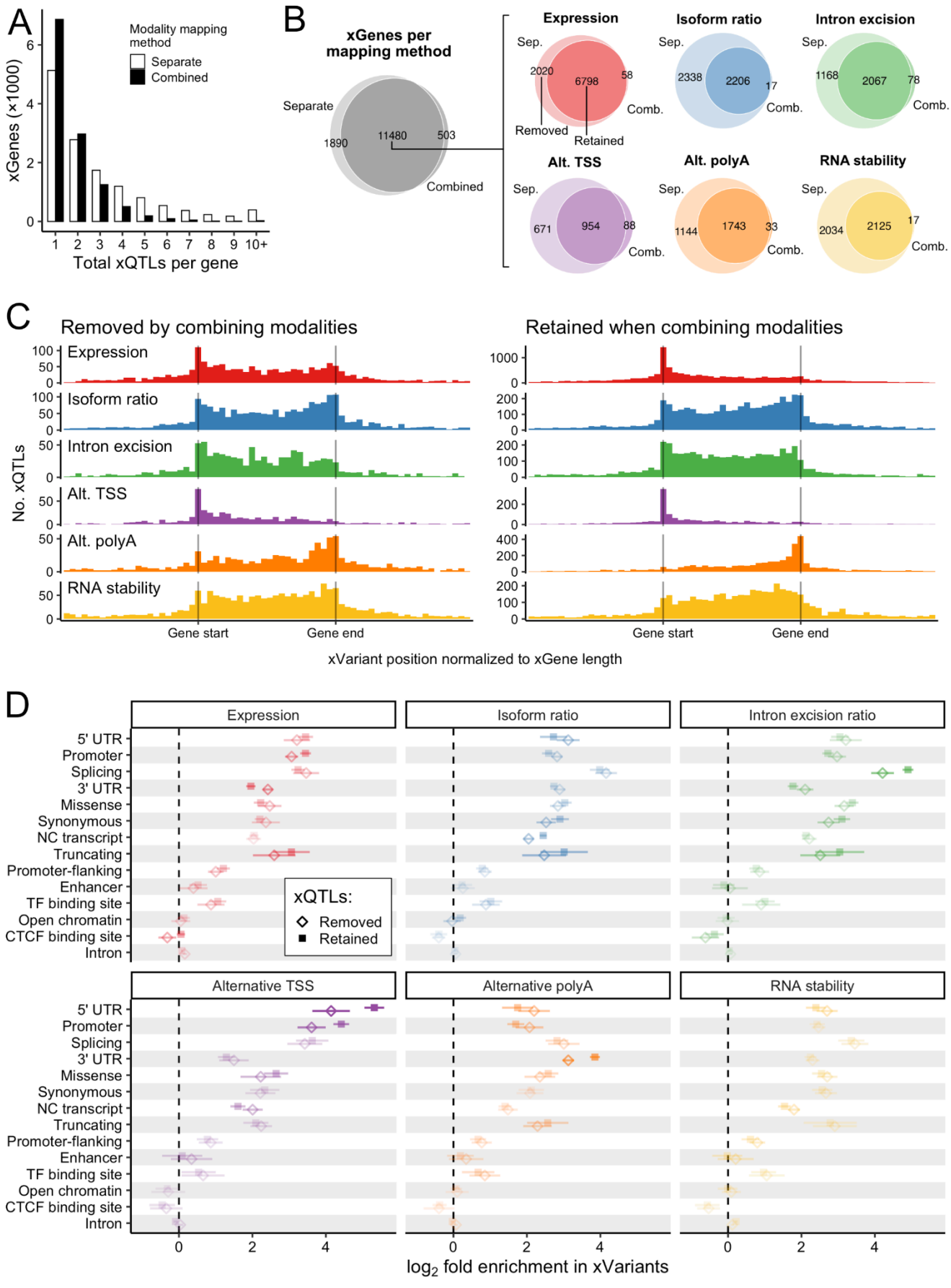

Figure 3. Comparison of separate-modality and cross-modality xQTL mapping. A) Total xQTLs per xGene in Geuvadis, summed over modalities, when testing for conditionally independent xQTLs separately per modality and when testing across all modalities. B) Left: Venn Diagram for xGenes from separate-modality mapping and from cross-modality mapping in Geuvadis. Right: For the genes with xQTLs from both mapping methods, the same comparison for the subsets of xGenes for each modality. C) The same type of relative xQTL position plots as Figure 2D, but for the two subsets of xGenes per modality, "removed" and "retained", indicated in B. D) Enrichment of functional annotations in the xVariants in GTEx tissues, similar to Figure 2E but for the "removed" and "retained" xGene subsets determined for each tissue. Opacity of the points and segments is proportional to the distance between each pair of points.

Next, we looked at how cross-modality mapping affects the overall functional characteristics of the resulting set of xVariants. For genes with Geuvadis xQTLs from both mapping methods, we examined, for each modality, the subset of genes that had significant xQTLs using separate-modality mapping but no significant xQTLs when using cross-modality mapping ("removed"; **Figure 3B**) and the subset that were also found with cross-modality mapping ("retained"; **Figure 3B**). We hypothesize that the "removed" xQTLs were found in multiple modalities and were better characterized with a phenotype of a different modality. We observed sharper modality-specific distributions of xVariant positions in the "retained" xQTLs (**Figure 3C**). Specifically, there were relatively fewer retained expression xVariants within the gene body and especially near the transcription end site (**TES**), compared to the peak at the TSS, and likewise fewer alternative TSS and polyA xVariants within the gene body relative to the peak at the TSS or TES, respectively. These observations indicate that cross-modality mapping results are more biologically plausible. We also compared functional annotation enrichment in removed and retained xQTLs and observed similar characteristic differences (**Figure 3D**). These include stronger enrichment of promoter annotations and weaker enrichment of 3' UTR annotations in expression QTLs; stronger enrichment of splicing annotations in intron excision ratio QTLs; stronger enrichment of 5' UTR and promoter annotations in alternative TSS QTLs and the opposite in alternative polyA QTLs; and stronger enrichment of 3' UTR annotations in alternative polyA QTLs.

xTWAS doubles the discovery of trait-associated genes

While TWAS is most commonly applied to gene expression data, the underlying principles and models are largely applicable to any of the RNA phenotypes provided by Pantry. We trained TWAS models on all RNA phenotypes (xTWAS), training one model per phenotype in the same way as the conventional method of training one expression model per gene. We performed xTWAS on a published collection of harmonized data for 114 traits, including cardiometabolic, psychiatric-neurologic, anthropometric, immune, blood, and other trait categories[18]. For Geuvadis, we found 10,065 significant hits across 80 traits involving 4,304 unique RNA phenotypes for 1,934 genes. Of the 4,487 unique trait-gene pairs among these hits, 51.3% involved only non-expression RNA phenotypes, and thus would not have been identified in a typical expression-only TWAS analysis (**Figure 4A**). While xTWAS produced a dramatic increase in findings compared to TWAS, expression phenotypes produced the single largest number of TWAS hits, and the most top hits per gene, of any modality.

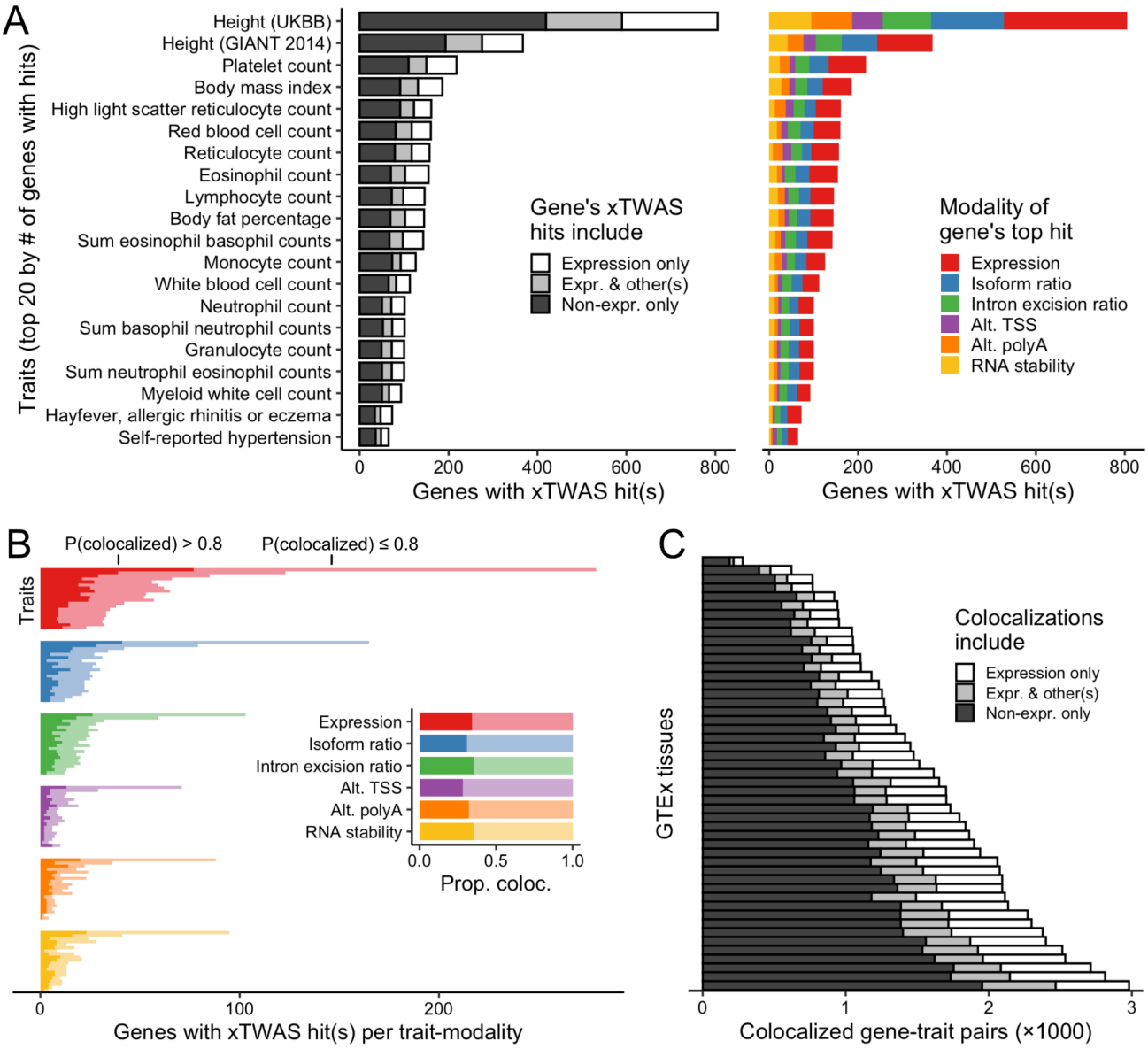

Figure 4. Multimodal TWAS (xTWAS). A) Left: For each trait, the number of genes with at least one xTWAS hit using Geuvadis RNA phenotypes are shown, shaded by whether each gene's hits(s) were for its expression phenotype, one or more other RNA phenotypes, or both. Only the top 20 traits in terms of gene count are shown. Right: The same traits and genes, colored by the modality of each gene's hit with the lowest xTWAS p-value. In both plots, each gene is represented at most once per trait, and genes in the "Expression only" category on the left overlap with, but are not the same set as, genes in the "Expression" top hit category on the right. B) For the same 20 traits, the number of TWAS hits per modality which also have strong evidence of single-variant-level colocalization is indicated with shading. The inset shows mean colocalizing proportions per modality. C) Colocalized gene-trait pairs for each GTEx tissue, shaded by whether the colocalization(s) involved an expression phenotype, one or more other RNA phenotypes, or both.

For each xTWAS hit, we sought more stringent evidence for mediation by testing for colocalization of the RNA phenotype and trait genetic associations using COLOC[19], which is a more conservative test than TWAS[18]. Approximately one-third of the xTWAS hits exhibited strong evidence of colocalization at a shared variant (posterior probability of association > 0.8), ranging from 32.1% to 36.3% per modality (**Figure 4B**). That is, no modality was especially depleted of colocalizations among its TWAS associations. We also ran xTWAS on each GTEx tissue (**Supp Table S3**), identifying colocalizing hits for 50,442 more trait-tissue-gene triplets than would be found using expression alone, a 2.73-fold change (**Figure 4C**).

GWAS loci are often provisionally attributed to the nearest gene, although it is generally understood that the nearest gene may or may not have a mediating role. We identified the two nearest genes to each GWAS locus for all traits and matched those trait-gene pairs with Geuvadis colocalizing xTWAS hits. Across the 7,071 loci, 566 (8%) could be potentially explained by an xTWAS hit matching one of the two nearest genes. Of those loci, 333 (59%) matched only non-expression hits. We repeated this analysis with colocalizing xTWAS hits from all 49 GTEx tissues after applying a more stringent Bonferroni threshold for TWAS p-values that accounts for the number of tissues in addition to the number of modalities. We found that 1,906 loci (27%) could be potentially explained by a hit in any tissue. Of those, 651 (34%) matched only non-expression hits. Compared to the single-tissue Geuvadis data, xTWAS hits across 49 tissues provided more contexts in which to detect potential mediators. While this resulted in more loci having at least one matching expression hit, this multi-tissue analysis still resulted in 95% more loci potentially explained by exclusively non-expression colocalizing xTWAS hits.

We also examined GTEx xTWAS hits for which the gene was originally reported in the GWAS study as a potential mediator based on its proximity to an association locus. For example, from the colocalizing xTWAS hits for neuroticism in UK Biobank, we identified several genes known previously to be relevant to behavior: *ORC4*, *CRHR1*, and *DRD2*. All three were reported in a GWAS on neuroticism in UK Biobank as being within associated regions[20]. In our xTWAS, their associated modalities included only isoform ratio and/or intron excision ratio across all 13 brain tissues for *ORC4* (**Figure 5**); expression, isoform ratio, intron excision ratio, alternative TSS, and/or RNA stability in four brain tissues for *CRHR1*; and expression in one brain tissue, cerebellum, for *DRD2*.

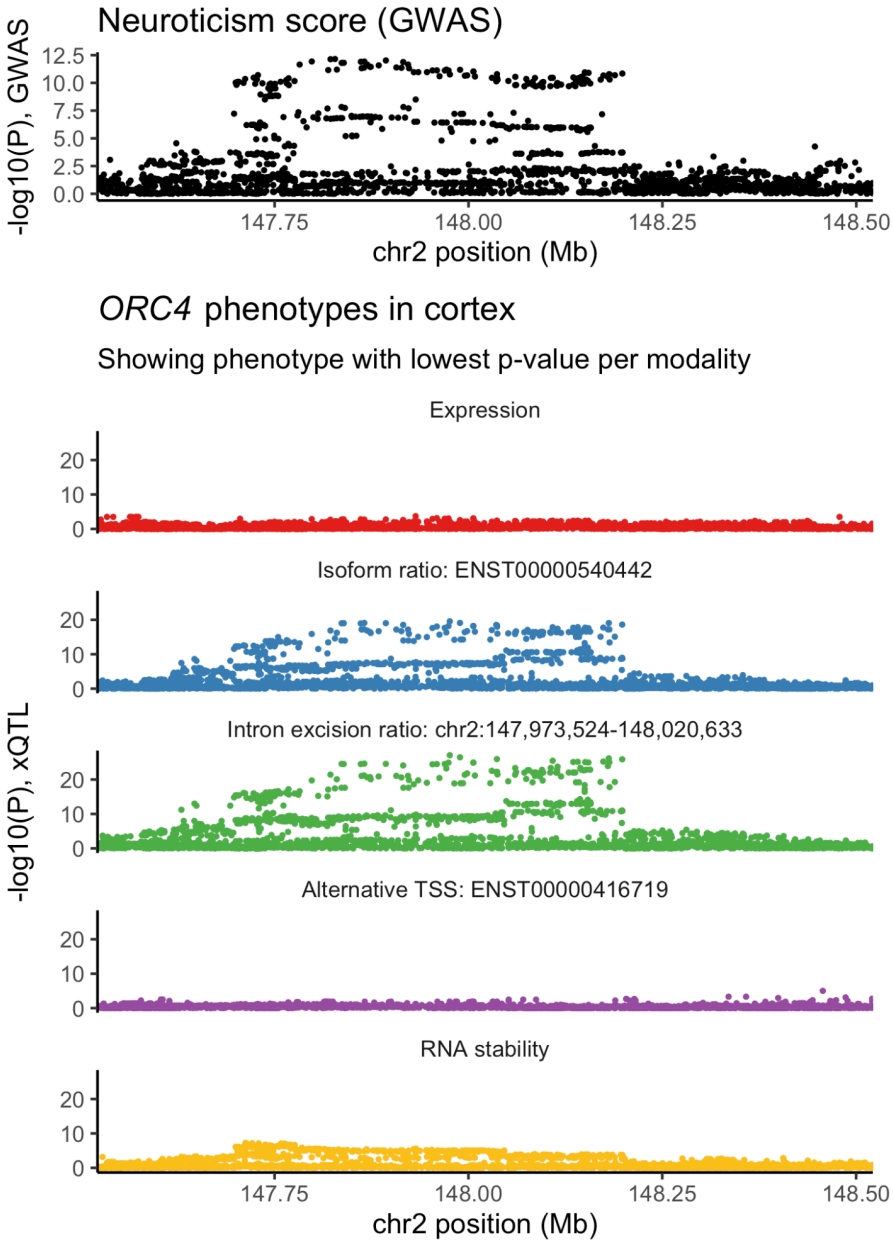

Figure 5. Example of a non-expression xTWAS hit. Top: plot showing GWAS p-values for the neuroticism trait within the *ORC4* cis-window on Chromosome 2, the region for which we performed xTWAS. Below: nominal p-values from xQTL mapping in GTEx cortex tissue, for four *ORC4* RNA phenotypes, in the same Chromosome 2 interval. Significant xTWAS hits for neuroticism were only found for isoform ratio and intron excision ratio modalities. Only the *ORC4* phenotype with the lowest cis-QTL p-value per modality is shown.

We also examined biologically relevant genes that had been reported based on colocalizing eQTLs rather than proximity alone. For example, in the PGC schizophrenia GWAS, *CYP2D6*, which encodes a pharmacologically important P450 enzyme[21], was included among blood eQTLs, but not brain eQTLs, that fell within a GWAS locus credible set[22]. We found colocalizing

xTWAS hits for this schizophrenia trait for *CYP2D6* in five GTEx brain tissues, liver, and seven other tissues, all for isoform ratio, intron excision ratio, or RNA stability phenotypes, and none for expression phenotypes.

A GWAS for sleep duration in UK Biobank found *PER1*, a well-characterized circadian rhythm gene, within an associated locus[23]. We found colocalizing xTWAS hits for circadian rhythm for *PER1* in thyroid, coronary artery, and sigmoid colon, all of which were for the alternative TSS modality. For a related trait, “morning/evening person” chronotype, a GWAS in UK Biobank followed by pathway analysis identified *RELN*, a gene previously linked to schizophrenia but not circadian rhythm[24]. We found colocalizing xTWAS hits for *RELN* exclusively in cerebellar hemisphere, cerebellum, and tibial nerve, for morning/evening person chronotype, and for alternative polyA modality, and did not find any other hits for other tissues, traits, or modalities.

RNA modalities harbor largely consistent proportions of genetic regulation across tissues

We examined the proportion of xQTLs and xTWAS hits coming from each RNA modality for each tissue to identify trends or outliers that could have biological significance. The proportions were fairly consistent across GTEx tissues, with no strong relationship to sample size (**Supp Figure S6**). A notable deviation was in Testis, which had the highest proportion of intron excision ratio phenotypes in both xQTLs (28.3%) and xTWAS hits (29.1%) of any tissue. This observation is consistent with existing knowledge that alternative splicing is especially prevalent in the testis[25]. Another strong deviation was cultured fibroblasts having a relatively high fraction of xQTL hits for RNA stability (13.4%, compared to mean 8.7% across tissues).

Discussion

We have introduced Pantry, a framework for multimodal analysis of RNA-seq data and its application to xQTL discovery and GWAS interpretation. Pantry dramatically increases the number of genomic discoveries when used to reanalyze previously generated datasets. Notably, for more than two-fifths of the gene-trait pairs with previous TWAS hits from gene expression analysis, we identified at least one additional regulation modality. While these genes are not completely new discoveries, the association with the new modality may facilitate the identification of the biological mechanism driving the association. We have shown that the systematic analysis of multiple RNA modalities reveals complementary biological information and genetic signals, improving the number and the specificity of genetic discoveries as compared to the conventional gene expression-based analysis using the same data. Finally, we share all the tools, methods and generated data with the community, including the RNA phenotypes, xQTLs, xTWAS weights, and xTWAS associations generated from the GTEx project and Geuvadis data.

The Pantry framework is modular and amenable to the addition of other transcriptomic modalities not considered here to facilitate further expansion and adaptation by the genomics

community. These could include types of RNAs lacking polyA tails, which may only be sufficiently quantified in non-polyA-selected RNA-seq libraries. Alternative forms of existing modalities, such as different ways to represent isoform abundance or more abstract features that represent expression variation, could be explored using this framework.

The established technique of using stepwise regression to find conditionally independent xQTLs naturally lends itself to multi-modal RNA phenotype data. Not only can it avoid redundant xQTLs in the presence of multiple phenotypes of the same modality, such as those representing alternative splice junctions, but it also avoids redundant xQTLs across modalities. However, when two phenotypes of different modalities share an xQTL signal, one modality might reflect the underlying causal mechanism better than the other, and that phenotype may not always have the stronger association. Thus, while the multi-modal conditionally independent xQTL mapping does sharpen the modality-specific functional characteristics of the xQTLs overall when compared to separate-modality mapping, it may also remove some of the true associations to a lower powered transcriptional modality (e.g., RNA stability) in favor of a better powered one (e.g., gene expression). We therefore also provide results from each modality individually analyzed for applications that benefit from more comprehensive data and focused on a single modality of gene regulation.

There are existing methods such as isoTWAS[26] and OPERA[5] that incorporate molecular information beyond total gene expression into genetic analysis[27]. Such methods have demonstrated that many more gene-trait associations can be discovered compared to using gene expression alone. Other studies have shown that cis-regulatory variants can be mapped for various transcriptional modalities beyond gene expression[1,3,7,12,28]. Pantry's unique strength is in providing a framework that begins with the raw RNA-seq data, produces comprehensive transcriptional phenotypes, and applies them seamlessly to multimodal genetic analyses.

This study has several important limitations. Pantry would require modification to handle single cell RNA-seq data. RNA-seq datasets such as those analyzed here may not cover the developmental stage, environmental exposures, or ancestry groups in which a transcriptomic mediator would be active and detectable. Still other molecular mediation could be only detectable in other types of omics data, such as DNA methylation or proteins. For species with sparser reference transcriptome data, Pantry would produce fewer RNA phenotypes, leading to fewer discovered genetic associations. Finally, xQTL mapping and xTWAS primarily detect associations for common variants, so other techniques would need to be employed to detect most regulatory effects of rare variants.

We have reported both broad characteristics of the xQTL and xTWAS results and specific observations that demonstrate Pantry's utility. However, given the high dimensionality of these analyses (tissues, genes, modalities, and often multiple phenotypes per modality for xQTLs, and the additional dimension of traits for xTWAS), we expect many more interesting biological insights to be found in the data repository published alongside this study. Furthermore, the inclusion of intermediate data such as the RNA phenotype quantifications and TWAS models for all GTEx tissues can enhance future methods development and GWAS.

Methods

Geuvadis dataset

We downloaded the quality control-filtered Geuvadis RNA-seq dataset (N=445 lymphoblastoid cell line samples) and corresponding genotypes for 13.4 million variants. These were filtered to autosomal biallelic variants with minor allele frequency (MAF) >= 0.01, resulting in 12.9 million variants. We ran the data through the default Pantry phenotyping and Pheast pipelines.

GTEx datasets

We downloaded the RNA-seq reads for all 54 GTEx v8 tissues. We obtained corresponding genotypes for 10.7 million variants and filtered to autosomal biallelic variants with MAF >= 0.01, resulting in 10.4 million variants. We ran the data through the default Pantry and pipeline, and ran QTL and TWAS analyses on the 49 tissues originally selected for eQTL mapping in GTEx v8.

RNA phenotyping

RNA phenotypes were generated using default Pantry parameters. We used human genome reference version GRCh38 and version 106 Ensembl gene annotations. Genes were not filtered by their annotated biotype, but final processing of results included filtering to protein-coding and lncRNA genes for statistics and visualizations. For RNA stability phenotypes, we filtered annotations to those with the Ensembl pipeline as an annotation source to limit rare or speculative isoforms that would prevent any constitutive exons from being counted for many genes.

xQTL mapping

We used code included in the Pantry Pheast module to compute covariates. For each modality in each dataset (tissue), we ran principal component analysis (PCA) on the RNA phenotype table and used the first 20 principal components (PCs) as covariates. We also ran PCA on each LD-pruned genotype alternative allele count matrix and included the first 5 PCs as covariates.

We mapped conditionally-independent cis-QTLs for each modality in each dataset (tissue) using tensorQTL[16], running the default commands included in Pantry Pheast. Modalities with multiple phenotypes per gene were mapped as per-gene groups so that cis-QTLs were conditionally independent across phenotypes within each gene.

For cross-modality mapping, the RNA phenotype tables per dataset (tissue) were concatenated, and 25 total covariates were computed in the same way as for individual modalities using the combined phenotype table and the genotypes. We then mapped cis-QTLs for this combined

dataset, grouping all phenotypes per gene so that cis-QTLs were conditionally independent across all phenotypes of all modalities within each gene.

Allelic fold change validation using allele-specific expression

To compare the robustness of cis-eQTLs from different mapping strategies, we measured the effect size of cis-eQTLs in GTEx subcutaneous adipose tissue (ADPSBQ), both for eQTLs found when mapping modalities separately and those found from cross-modality mapping. We estimated aFC using the aFC-n model[29] with phased genotypes. We also calculated aFC from allele-specific expression in heterozygous individuals using phASER[30].

xTWAS

We downloaded summary statistics from a collection of 114 GWAS traits[18] (https://zenodo.org/record/3629742#.XjCh9OF7m90). We ran TWAS analysis using FUSION[31] with default parameters. First, we fit predictive models for each Geuvadis RNA phenotype using TSS ±500 kb cis-window genotypes, along with the same 25 covariates used for cis-QTL mapping, and ran FUSION's built-in comparison of 'blup', 'lasso', 'top1', and 'enet' models for each phenotype. We then used these models to test TWAS associations for each RNA phenotype against each GWAS trait. We used a genome-wide P-value threshold, Bonferroni adjusted for the number of RNA modalities, of 8.33×10^{-9} (5×10^{-8} divided by 6) to determine significant TWAS hits. We also used FUSION's built-in option to report COLOC posterior probabilities for each hit.

For the GWAS loci-based analysis, we determined loci for each of the 114 traits by extracting all genome-wide significant ($P<5\times10^{-8}$) variants and grouping them such that any two significant variants <500 Kb apart were in the same locus. We found the two nearest genes to each locus based on the distance between the variant in the locus with the lowest p-value and the nearest point in the gene's interval. For each locus, xTWAS hits matching the trait and either of the two nearest genes were assigned to the locus. For this analysis, we used only colocalizing xTWAS hits (those with COLOC posterior probability of association >0.8), and for GTEx hits, used a more stringent TWAS P-value threshold that was Bonferroni adjusted for the number of tissues in addition to the number of RNA modalities, i.e. 1.70×10^{-10}.

Variant effect enrichment

We downloaded variant annotations from the GTEx Portal (https://storage.googleapis.com/gtex_analysis_v8/reference/WGS_Feature_overlap_collapsed_VEP_short_4torus.MAF01.txt.gz). To reduce low-frequency annotation categories, we merged Splice acceptor, Splice donor, and Splice region categories into one "Splicing" category, and merged Frameshift and Stop gained into one "Truncating" category. For the conditionally independent cis-QTLs for each RNA modality for each GTEx tissue, enrichment of each annotation in the xVariants was computed as the $\log_2$-ratio of the proportion among the xVariants to the proportion in all variants within all the cis-windows tested for that RNA modality. To control the variance of enrichment values from infrequent annotations that result in low

annotated xVariant counts, we added a pseudocount of 0.5 to each annotated xVariant count, and added an amount to the total xVariant counts such that the added xVariants had background annotation frequency. We also omitted tissue-modality-annotation combinations with fewer than 2 annotated xVariants from enrichment analysis.

Data availability

We provide data processed with Pantry for Geuvadis and all GTEx tissues in a public repository at https://www.dropbox.com/scl/fo/c88y8wu8u6beaxar3bs7g/h?rlkey=d2atycui3uyq83igk4hipu8ln&dl=0 *(will be replaced with Zenodo accession prior to manuscript acceptance)*. These include, for all six modalities in each tissue, RNA phenotype matrices, covariates, xQTLs, xTWAS transcriptomic model weights, and xTWAS associations for 114 GWAS traits. This repository is about 42 GB when compressed.

Software availability

The Pantry code is available at https://github.com/daniel-munro/Pantry. It is structured as a two-stage pipeline using the Snakemake workflow management system[32]. The pipeline consists of existing programs, e.g., STAR[33] and samtools[34], and additional scripts to process their input and output data. The pipeline was designed for computational and storage efficiency by reducing redundant computation and large files, and is compatible with high performance computing environments. See Supplementary Information for details.

Acknowledgments

We thank Robert Vogel for helpful discussion. Funding: National Institute on Drug Abuse [P50DA037844], National Institute of General Medical Sciences [R01GM140287].

References

1. Consortium, T. Gte. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
2. Lappalainen, T., Li, Y. I., Ramachandran, S. & Gusev, A. Genetic and molecular architecture of complex traits. *Cell* **187**, 1059–1075 (2024).
3. Alasoo, K. *et al.* Genetic effects on promoter usage are highly context-specific and contribute

to complex traits. *eLife* **8**, e41673 (2019).

4. Garrido-Martín, D., Borsari, B., Calvo, M., Reverter, F. & Guigó, R. Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome. *Nat. Commun.* **12**, 727 (2021).
5. Wu, Y. *et al.* Joint analysis of GWAS and multi-omics QTL summary statistics reveals a large fraction of GWAS signals shared with molecular phenotypes. *Cell Genomics* 100344 (2023) doi:10.1016/j.xgen.2023.100344.
6. Pan, S. *et al.* COLOCdb: a comprehensive resource for multi-model colocalization of complex traits. *Nucleic Acids Res.* gkad939 (2023) doi:10.1093/nar/gkad939.
7. Kerimov, N. *et al.* eQTL Catalogue 2023: New datasets, X chromosome QTLs, and improved detection and visualisation of transcript-level QTLs. *PLOS Genet.* **19**, e1010932 (2023).
8. Gao, G. *et al.* A multi-tissue, splicing-based joint transcriptome-wide association study identifies susceptibility genes for breast cancer. *Am. J. Hum. Genet.* (2024) doi:10.1016/j.ajhg.2024.04.010.
9. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
10. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
11. Cotto, K. C. *et al.* Integrated analysis of genomic and transcriptomic data for the discovery of splice-associated variants in cancer. *Nat. Commun.* **14**, 1589 (2023).
12. Li, Y. I. *et al.* Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* **50**, 151–158 (2018).
13. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
14. Gaidatzis, D., Burger, L., Florescu, M. & Stadler, M. B. Analysis of intronic and exonic reads in RNA-seq data characterizes transcriptional and post-transcriptional regulation. *Nat.*

Biotechnol. **33**, 722–729 (2015).

15. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).

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

17. Mohammadi, P., Castel, S. E., Brown, A. A. & Lappalainen, T. Quantifying the regulatory effect size of *cis* -acting genetic variation using allelic fold change. *Genome Res.* **27**, 1872–1884 (2017).

18. Barbeira, A. N. *et al.* Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. *Genome Biol.* **22**, 49 (2021).

19. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLOS Genet.* **10**, e1004383 (2014).

20. Luciano, M. *et al.* Association analysis in over 329,000 individuals identifies 116 independent variants influencing neuroticism. *Nat. Genet.* **50**, 6–11 (2018).

21. Wang, B. *et al.* New insights into the structural characteristics and functional relevance of the human cytochrome P450 2D6 enzyme. *Drug Metab. Rev.* **41**, 573–643 (2009).

22. Ripke, S. *et al.* Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).

23. Dashti, H. S. *et al.* Genome-wide association study identifies genetic loci for self-reported habitual sleep duration supported by accelerometer-derived estimates. *Nat. Commun.* **10**, 1100 (2019).

24. Jones, S. E. *et al.* Genome-wide association analyses of chronotype in 697,828 individuals provides insights into circadian rhythms. *Nat. Commun.* **10**, 343 (2019).

25. Song, H., Wang, L., Chen, D. & Li, F. The Function of Pre-mRNA Alternative Splicing in Mammal Spermatogenesis. *Int. J. Biol. Sci.* **16**, 38–48 (2020).

26. 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).

27. Li, R. *et al.* RNA alternative splicing impacts the risk for alcohol use disorder. *Mol. Psychiatry* **28**, 2922–2933 (2023).

28. Huan, T. *et al.* Genome-wide identification of microRNA expression quantitative trait loci. *Nat. Commun.* **6**, 6601 (2015).

29. Ehsan, N. *et al.* Haplotype-aware modeling of cis-regulatory effects highlights the gaps remaining in eQTL data. *Nat. Commun.* **15**, 522 (2024).

30. Castel, S. E., Mohammadi, P., Chung, W. K., Shen, Y. & Lappalainen, T. Rare variant phasing and haplotypic expression from RNA sequencing with phASER. *Nat. Commun.* **7**, 12817 (2016).

31. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).

32. Mölder, F. *et al.* Sustainable data analysis with Snakemake. Preprint at https://doi.org/10.12688/f1000research.29032.2 (2021).

33. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

34. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).

Supplementary Material

Supplementary_Material.pdf - Supplementary Information and Supplementary Figures S1-S6

Supplementary Tables

See **Supp_Tables_1-3.xlsx**

Table S1. Counts of xGenes and xQTLs per GTEx tissue, modalities mapped separately

Table S2. Counts of xGenes and xQTLs per GTEx tissue, cross-modality mapping

Table S3. Counts of xTWAS hits and unique genes per trait-tissue pair for GTEx tissues