Quantification method affects replicability of eQTL analysis, colocalization, and TWAS

Nolan Cole[1*], William Wu[2*], S. Taylor Head[3], Sean T. Bresnahan[3], Arjun Bhattacharya[3,4]

Affiliations:

1. Department of Biostatistics, University of Washington, Seattle, WA, USA
2. Department of BioSciences, Rice University, Houston, TX, USA
3. Department of Epidemiology, University of Texas MD Anderson Cancer Center, Houston, TX, USA
4. Institute for Data Science in Oncology, University of Texas MD Anderson Cancer Center, Houston, TX, USA

*These authors contributed equally

Address all correspondence to Arjun Bhattacharya (abhattacharya3@mdanderson.org)

ABSTRACT

eQTL mapping and TWAS are widely used to contextualize GWAS, yet the impact of RNA-seq processing choices remains unexplored. We find that RNA-seq quantification method and transcriptomic reference substantially affect eQTL detection and gene expression prediction with significant downstream impact on colocalization and TWAS results. Our findings demonstrate that seemingly minor methodological decisions substantially affect these common analyses, highlighting the need for standardized practices to ensure reproducible genetic association studies.

MAIN

Contextualizing variant-to-trait associations identified through genome-wide association studies (GWAS) often relies on expression quantitative trait loci (eQTL) mapping, colocalization and transcriptome-wide association studies (TWAS)[1,2]. Reference eQTL datasets for these integrative analyses are often based on short-read RNA-sequencing (RNA-seq) data. Two critical decisions in short-read RNA-seq analysis workflows involve (1) the choice of transcriptomic annotation that defines gene, exon, and transcript boundaries for expression quantification and (2) the method itself for quantification. GENCODE annotations are regularly updated to incorporate newly discovered transcripts and refined gene models[3]. Quantification methods also vary considerably, from alignment using STAR[4] followed by featureCounts, htseq-count, or RSEM[5–7] to alignment-free methods such as Salmon and kallisto[8,9], each with distinct assumptions about transcript structure and quantification accuracy.

Short-read RNA-seq-based eQTL colocalization and TWAS continue to be commonly used to identify mechanisms underlying GWAS-identified loci. As of June 2025, we found 125 pre-printed or published studies that included gene-level eQTL colocalization or TWAS with short-read RNA-seq indexed in PubMed Central. Of these studies, over 90% quantify expression through an alignment-based approach. These studies also used a variety of transcriptomic annotations ranging from GENCODE v17 to GENCODE v38, with a few using *de novo* transcriptomes or combining datasets quantified based on multiple annotations. Of note, only 10 of these studies directly report the quantification method or reference transcriptome either in the main text or supplemental methods, while the others either omit these details or rely on citations that describe the RNA-seq processing.

While previous studies have compared quantification methods on simulated data[10,11], their relative performance for genetic association analyses using real human tissue data remains unclear. Here, we extracted raw RNA-seq fastq files from 48 tissues from the Genotype-Tissue Expression Project (GTEx v8, **Methods**). We quantified gene-level expression using GENCODE v27, v38, and v45 annotations through either STAR alignment and featureCounts quantification[4,6] or Salmon pseudo-alignment[8]. We then conduct eQTL analysis using QTLtools[1], eCAVIAR colocalization with GWAS for 9 cancers[12–22], and TWAS[2] (**Supplemental Table S1, Methods**). Across these three annotations, the number of genes remained relatively similar, but the number of annotated transcripts increased (**Supplemental Figure S1**).

We find significant discordance in genes with significant eQTLs (eGenes) and those with predicted expression at $R^2 > 0.01$, both across quantification method and across annotation. Across 48 tissues, eGene calls are ~20-25% discordant across method and ~20-40% discordant across annotation (**Figure 1A**). Overall, we find that the number of discordant genes reflect chromosome length and relative abundance of gene biotypes (**Supplemental Figure S2-3**). However, eGenes uniquely discovered in more recent GENCODE annotations have more transcripts and exons, probably reflecting the increase in annotated transcripts across version (**Supplemental Figure S4**). Multi-SNP predictions of gene expression exhibit far more discordance: ~35-40% discordance across method and ~40-45% discordance across annotation (**Figure 1B**), owing potentially to uncertainty in variable selection during penalized regression. In fact, the number of SNPs included in expression prediction models for the same gene in the same tissue vary greatly, when comparing across method and annotation (**Supplemental Figure S5**), which lead to differences in interpretation of the polygenicity of the gene's expression. These discordance rates decrease with increasing sample size but stabilize in the tissues with the largest samples sizes (**Figure 1C-D**, **Supplemental Tables S2-5, Supplemental Figures S6-13**), suggesting that these discrepancies could potentially be mitigated in large datasets.

At these sample sizes, these discrepancies in QTL calling and expression prediction have stark effects on downstream colocalization and TWAS. Colocalization analysis reveals substantial discordance in gene identification across both quantification methods and reference annotations. While the independent GWAS loci tagged by colocalized genes show mild concordance of ~60-70% (**Figure 2A**), the specific genes identified through colocalization at these loci exhibit marked discordance. Method-based comparisons between STAR and Salmon show ~25-30% discordance in colocalized gene identification, with the remainder showing concordance between methods (**Figure 2B**). Even more pronounced discordance is observed across GENCODE annotations, where up to 40-50% of colocalized genes are identified in only one or two of the three annotation versions, with substantial portions being annotation-specific rather than consistently identified across all three versions (**Figure 2B, Supplemental Figures S14-21, Supplemental Tables S6-9**). These results reflect findings from previous studies

demonstrating that annotation and methodological choices substantially impact RNA-seq quantification accuracy[23,24].

TWAS mimics the discrepancies in results we observed in colocalization. Across both quantification method and reference annotation, the independent GWAS loci that are tagged by TWAS-identified genes within 1 Mb are generally the same with ~80-90% concordance (**Figure 2C**, **Supplemental Figures S22-25, Supplemental Tables S10-13**). However, the specific genes identified at these loci across method and across annotation are quite different, showing up to 45% and 50% discordance across method and annotation, respectively (**Figure 2D**, **Supplemental Figures S26-29**). TWAS identifies a distinct set of genes per locus, depending on quantification method and annotation (**Supplemental Figure S30**). Although this discrepancy could imply that GWAS variants potentially affect transcription of multiple genes at the locus, we believe, rather, it raises a concern for replicability, especially since the underlying GWAS and eQTL datasets are the same.

Our analysis shows that seemingly minor methodological choices in quantification method or transcript annotation substantially affect QTL results. These effects are notable even at the gene level and likely will be more pronounced for transcript- and exon-level analyses. As recent QTL studies are increasingly focusing on isoform expression and splicing, annotation choice becomes critical: newer GENCODE versions introduce more isoforms and refined exon boundaries that can dramatically shift quantification and association outcomes. Additionally, as annotations vary by tissue and cell type, these choices have major implications for studying context-specific regulation. Our results show that methodological discrepancies in eQTL detection—such as those arising from different quantification approaches—diminish with increasing sample size, highlighting the value of larger molecular phenotype datasets. Recent work further reinforces this need, showing that robust QTL-GWAS colocalization may require far greater sample sizes than currently available[25]. Together, these findings support the need for a strong commitment to expanding transcriptomic datasets. However, sample size alone is not sufficient: short-read data lack the resolution to capture complex and tissue-specific transcript structures. To address this, long-read sequencing must play a complementary role in future eQTL studies, enabling more accurate quantification of nuanced, context-specific transcript structures.

We conclude by advocating the field of functional genomics to adopt stronger methodological standards to ensure the reproducibility, transparency, and relevance of eQTL and TWAS analyses. First, rigorous reporting practices are imperative. Details about the quantification methods and the transcriptomic annotations used should be treated as foundational, on par with reporting genome builds, to enable accurate interpretation and cross-study comparisons. Second, the community must establish frameworks for regularly updating foundational resources. Widely used pre-computed datasets such as GTEx eQTL maps and TWAS models from FUSION or PredictDB[26] are often treated as static references, despite rapid improvements in gene annotations and analytic methods. To reflect the current state of genomic knowledge and avoid the propagation of outdated findings, we advocate for community-supported systems to re-quantify and refresh these key resources as methods and annotations evolve. As complex deep learning methods become increasingly central to transcriptome analysis[27], the need for dynamic model updating will become even more imperative, ensuring that these models are trained on the most accurate and up-to-date representations of transcriptomic complexity. We consider these challenges presented here as a methodological opportunity to ensure reproducibility of computational inference from transcriptomic datasets.

METHODS

Genotype processing

SNP genotype calls were derived from Whole Genome Sequencing data for samples 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[1].

RNA-seq processing

We quantified GTEx v8[2] RNA-seq samples for 48 tissues using two methods.

First, we used Salmon v1.10.2[3] in mapping-based mode. We first built a Salmon index for a decoy-aware transcriptome consisting of GENCODE v27, v38, and v45 transcript sequences and the full GRCh38 reference

genome as decoy sequences[4]. 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 v1.22.1[5]. Using edgeR, gene-level quantifications underwent TMM-normalization, followed by transformation into a log-space using the variance-stabilizing transformation using DESeq2 v1.44.0[6,7]. We then residualized gene expression (as log-transformed TPM) by all tissue-specific covariates (clinical, demographic, genotype principal components (PCs), and expression PEER factors) used in the original QTL analyses in GTEx.

Second, we performed read alignment using STAR v2.7.10b[8], with a two-pass-like approach optimized for spliced transcript detection. The alignment was configured for high sensitivity to splice junctions (`--alignSJoverhangMin 8`, `--alignSJDBoverhangMin 1`) and permissive mismatch thresholds (`--outFilterMismatchNmax 999`, `--outFilterMismatchNoverReadLmax 0.04`). We allowed up to 20 multimapping reads per alignment and enabled long intron detection up to 1 Mb. Output was generated as sorted BAM files with comprehensive alignment attributes. We then generated gene-level counts using featureCounts[9], using GENCODE v27, v38, and v45 annotations. Lastly, we followed the same normalization and residualization steps as before.

Before downstream analysis, we only retained only genes with 0.1 TPM for more than 75% of samples. To ensure that comparisons are fair across method and annotation, we only retained genes that were not filtered in all datasets (across method and across annotation).

eQTL mapping and colocalization

We conducted eQTL analysis using QTLtools v1.3.1 with quantile normalization of gene expression[10]. P-value correction was conducted using the Beta-distribution approximation using 1000 permutations. For eGenes with adjusted $P < 0.05$ that were within 1 Mb of a GWAS SNP with $P < 5e-8$, we then conducted colocalization using eCAVIAR, assuming up to 2 causal variants per gene[11].

Transcriptome-wide association study

We followed the standard FUSION[12] pipeline for predicting gene expression models using the most predictive in cross-validation of elastic net[13], best linear unbiased predictor from a linear mixed model[14], or Sum of Single Effects regression[15]. We then incorporated these models GWAS summary statistics with a weighted-burden test, running trait mapping only for genes predicted at cross-validation adjusted $R^2 > 0.01$. We used in-sample linkage disequilibrium from GTEx to compute the weighted burden test statistic. We aggregated P-values from multiple tissue-specific gene associations using the aggregated Cauchy association test and considered $P < 2.5 \times 10^{-6}$ as transcriptome-wide significant[16]. We define an independent GWAS locus by defining lead SNPs with minimum $P < 5\times10^{-8}$ in a 1 Mb window[17].

DATA AND CODE AVAILABILITY

GTEx genotype and short-read RNA-seq data were downloaded from dbGaP phs000424.v10.p2. Available GWAS summary statistics were accessed from the GWAS Catalog with accession numbers listed in **Supplemental Table S1**. All scripts for this analysis are available at github.com/bhattacharya-a-bt/QuantMethod_QTL_Comparison.

ACKNOWLEDGEMENTS

We thank Michael Love and Sara Lindström for discussion. This work was supported by NCI-CA293419.

REFERENCES

1. Delaneau, O. *et al.* A complete tool set for molecular QTL discovery and analysis. *Nat. Commun.* **8**, (2017).
2. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).
3. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).
4. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
5. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
6. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinforma. Oxf. Engl.* **30**, 923–930 (2014).
7. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
8. 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).
9. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
10. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research 2015 41521* **4**, 1521–1521 (2016).
11. Teng, M. *et al.* A benchmark for RNA-seq quantification pipelines. *Genome Biol.* **17**, 74 (2016).
12. Hormozdiari, F. *et al.* Colocalization of GWAS and eQTL Signals Detects Target Genes. *Am. J. Hum. Genet.* **99**, 1245–1260 (2016).
13. Michailidou, K. *et al.* Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**, 92–94 (2017).
14. Huyghe, J. R. *et al.* Discovery of common and rare genetic risk variants for colorectal cancer. *Nat. Genet.* **51**, 76–87 (2019).
15. Gharahkhani, P. *et al.* Genome-wide association studies in oesophageal adenocarcinoma and Barrett's oesophagus: a large-scale meta-analysis. *Lancet Oncol.* **17**, 1363–1373 (2016).

16. O'Mara, T. A. *et al*. Identification of nine new susceptibility loci for endometrial cancer. *Nat. Commun*. **9**, 3166 (2018).

17. Byun, J. *et al*. 'Cross-ancestry genome-wide meta-analysis of 61,047 cases and 947,237 controls identifies new susceptibility loci contributing to lung cancer'. *Nat. Genet*. **54**, 1167–1177 (2022).

18. Melin, B. S. *et al*. Genome-wide association study of glioma subtypes identifies specific differences in genetic susceptibility to glioblastoma and non-glioblastoma tumors. *Nat. Genet*. **49**, 789–794 (2017).

19. Law, M. H. *et al*. Genome-wide meta-analysis identifies five new susceptibility loci for cutaneous malignant melanoma. *Nat. Genet*. **47**, 987–995 (2015).

20. 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).

21. Scelo, G. *et al*. Genome-wide association study identifies multiple risk loci for renal cell carcinoma. *Nat. Commun*. **8**, 15724 (2017).

22. Phelan, C. M. *et al*. Identification of twelve new susceptibility loci for different histotypes of epithelial ovarian cancer. *Nat. Genet*. **49**, 680–691 (2017).

23. Zhao, S. & Zhang, B. A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. *BMC Genomics* **16**, 97 (2015).

24. Chisanga, D., Liao, Y. & Shi, W. Impact of gene annotation choice on the quantification of RNA-seq data. *BMC Bioinformatics* **23**, 107 (2022).

25. Rosen, J. D., Broadaway, K. A., Brotman, S. M., Mohlke, K. L. & Love, M. I. Higher eQTL power reveals signals that boost GWAS colocalization. 2025.08.05.668745 Preprint at https://doi.org/10.1101/2025.08.05.668745 (2025).

26. Gamazon, E. R. *et al*. A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet*. **47**, 1091–1098 (2015).

27. Avsec, Ž. *et al*. AlphaGenome: advancing regulatory variant effect prediction with a unified DNA sequence model. 2025.06.25.661532 Preprint at https://doi.org/10.1101/2025.06.25.661532 (2025).

FIGURES

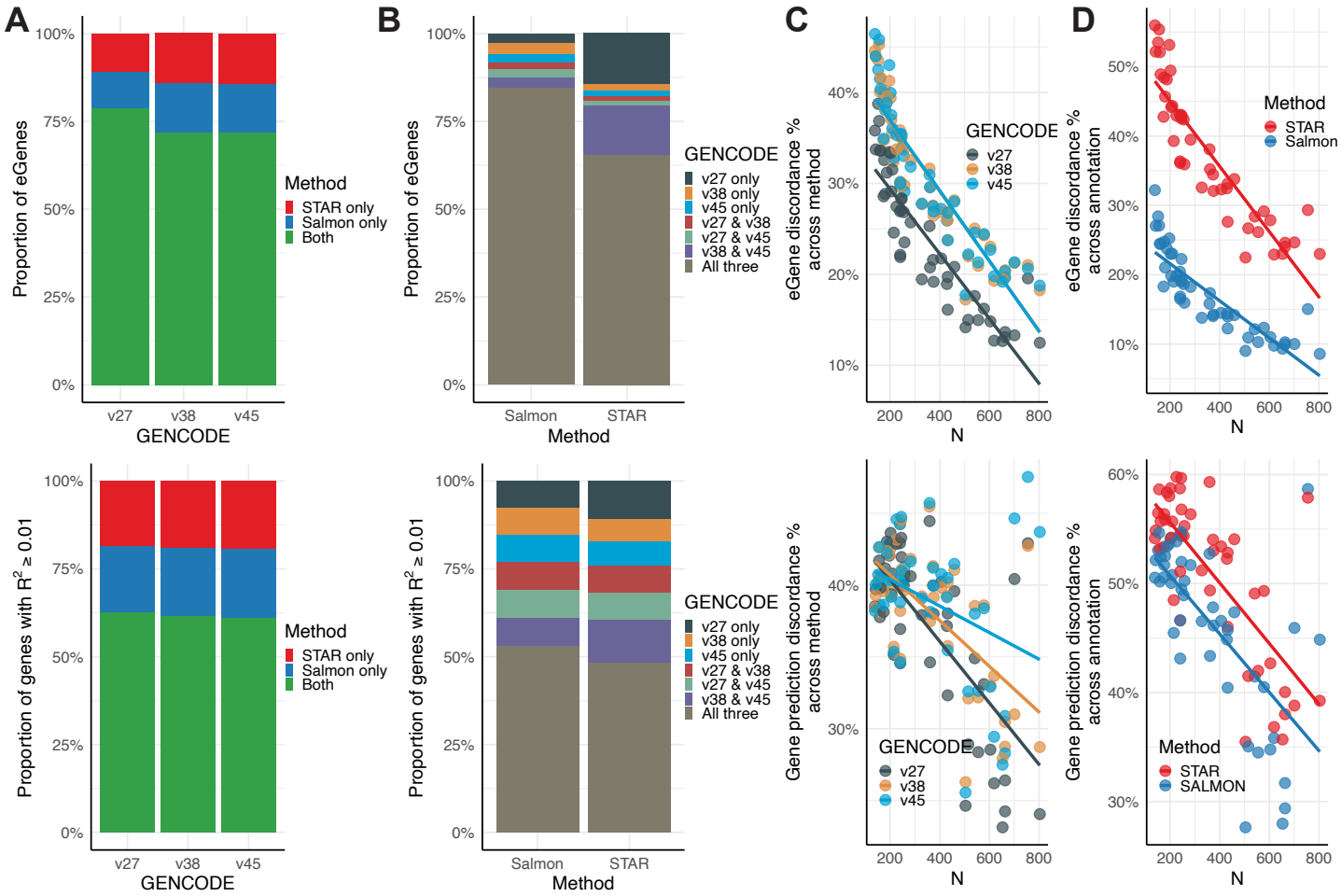

Figure 1: *Quantification method and transcriptomic annotation affect eGene identification and gene expression prediction.* **(A)** Proportion of eGenes (top) and genes with $R^2 > 0.01$ (bottom) identified by STAR only (red), Salmon only (blue), or both methods (green) across GENCODE versions. **(B)** Proportion of eGenes (top) and predicted genes (bottom) found across one or more GENCODE versions by quantification method. **(C)** Method discordance percentage vs. tissue sample size for eGene calling (top) and expression prediction (bottom). **(D)** Annotation discordance percentage vs. tissue sample size for eGene calling (top) and expression prediction (bottom).

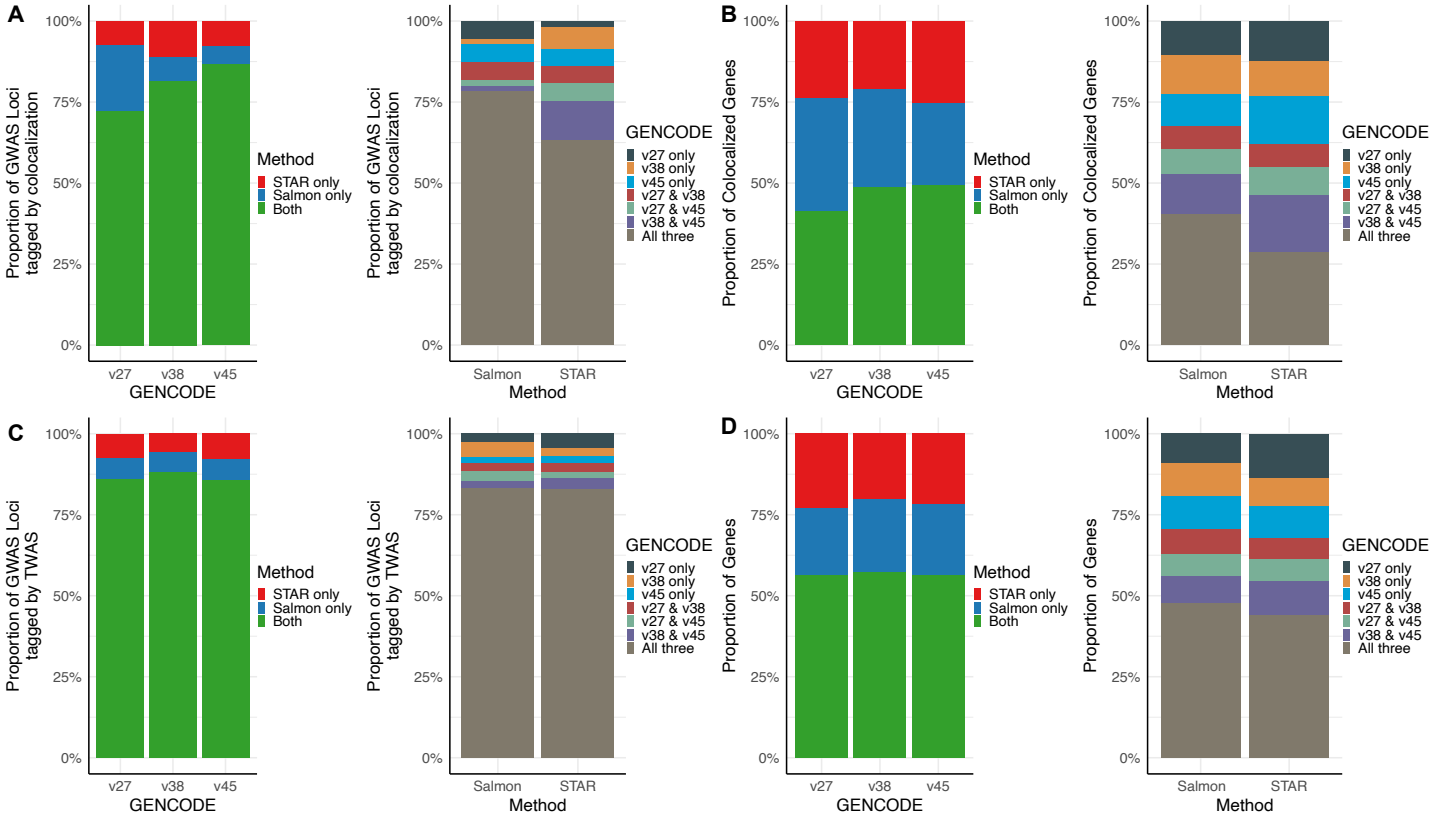

Figure 2: *Differences in quantification method and transcriptomic annotation lead to discordances in GWAS colocalization and TWAS*. Across 48 GTEx tissues, proportion (Y-axis) of all GWAS loci tagged through colocalization **(A)** or TWAS **(C)** using QTL data quantified with different methods or using different reference annotations. Across 48 GTEx tissues, proportion (Y-axis) of all colocalized (B) or TWAS-prioritized (D) genes within 1 Mb of a GWAS locus across quantification method and reference annotation.

SUPPLEMENTAL TABLE LEGENDS

Table S1: GWAS sample sizes and accession codes

Table S2: eGene numbers by method

Table S3: eGene numbers by annotation

Table S4: Gene prediction numbers by method

Table S5: Gene prediction numbers by annotation

Table S6: Number of GWAS loci tagged by a colocalization across method

Table S7: Number of GWAS loci tagged by a colocalization across annotation

Table S8: Number of colocalized genes across method

Table S9: Number of colocalized genes across annotation

Table S10: Number of GWAS loci tagged by a TWAS association across method

Table S11: Number of GWAS loci tagged by a TWAS association across annotation

Table S12: Number of TWAS-prioritized genes across method

Table S13: Number of TWAS-prioritized genes across annotation