A Deep Dive into Statistical Modeling of RNA Splicing QTLs Reveals New Variants that Explain Neurodegenerative Disease

David Wang[1,2], Matthew R. Gazzara[1,2], San Jewell[1], Benjamin Wales-McGrath[1], Christopher D. Brown[1,†], Peter S. Choi[3,4] Yoseph Barash[1,5*]

[1]Department of Genetics, Perelman School of Medicine, University of Pennsylvania
[2]Graduate Group in Genomics and Computational Biology, Perelman School of Medicine, University of Pennsylvania
[3]Department of Pathology & Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania
[4]Division of Cancer Pathobiology, The Children's Hospital of Philadelphia
[5]Department of Computer and Information Sciences, School of Engineering, University of Pennsylvania
[†]Passed away March 18th, 2023
[*]To whom correspondence should be addressed; E-mail: yosephb@upenn.edu.

Abstract

Genome-wide association studies (GWAS) have identified thousands of putative disease causing variants with unknown regulatory effects. Efforts to connect these variants with splicing quantitative trait loci (sQTLs) have provided functional insights, yet sQTLs reported by existing methods cannot explain many GWAS signals. We show current sQTL modeling approaches can be improved by considering alternative splicing representation, model calibration, and covariate integration. We then introduce MAJIQTL, a new pipeline for sQTL discovery. MAJIQTL includes two new statistical methods: a weighted multiple testing approach for sGene discovery and a model for sQTL effect size inference to improve variant prioritization. By applying MAJIQTL to GTEx, we find significantly more sGenes harboring sQTLs with functional significance. Notably, our analysis implicates the novel variant rs582283 in Alzheimer's disease. Using antisense oligonucleotides, we validate this variant's effect by blocking the implicated YBX3 binding site, leading to exon skipping in the gene MS4A3.

Introduction

Genome-wide association studies (GWAS) have implicated thousands of genetic variants in human complex traits and disease. However, elucidating the functional mechanisms through which a variant acts on a trait remains challenging. The current prevailing hypothesis maintains that the effect of a variant on a trait is mediated by its effect on gene expression either by directly affecting trans acting elements like transcription factors or by disrupting cis regulatory elements such as transcription factor binding sites. This view is supported by the observation that GWAS associations are enriched for eQTLs in non-coding regions, leading to many studies reporting a high degree of colocalization between eQTLs and GWAS signals [1]. However, many GWAS hits still cannot be explained by eQTLs thus many alternative hypotheses have been proposed to potentially address this gap [2]. For example, several studies have suggested that eQTLs are context specific [3] while others claim that QTLs for other molecular traits could be more informative. Specifically, a comprehensive study of many molecular QTLs identified RNA splicing QTLs (sQTLs) as a key contributor to several disease phenotypes [4]. However, sQTLs that explain GWAS variants are still comparatively underrepresented in recent studies. We postulate that improved sQTL detection, and consequent overlap with GWAS signals, can be achieved by focusing on three main elements.

First, the RNA splicing representation used for sQTL discovery should ideally capture the full complexity of splicing variations in the transcriptome. Splicing can be accurately measured using short read RNA sequencing and quantified at the local "event" (e.g. cassette exon) or junction level using tools like MAJIQ [5, 6], Leafcutter [7] or rMATs [8]. It can also be quantified at the isoform level with transcript expression using tools like Salmon[9], Kalisto [10] or RSEM[11]. However, each splicing quantification approach presents only a partial view of splicing variation and joint analysis of event and isoform representations can yield complementary insight. For example, event based methods such as rMATs miss denovo junctions and exons, while methods such as Leafcutter miss intron retention. All event based methods miss alternative transcript starts and ends which can be captured by isoform based approaches. However, isoform quantification typically relies on annotations and can't handle unannotated transcripts [12]. Usage of long reads can greatly improve unannotated transcript detection and quantification but low coverage and high-error rates remain a challenge [13].

Second, statistical models for sQTL effect size inference could be improved by accounting for the discrete and heteroskedastic nature of RNA splicing data. Most modern sQTL analysis frameworks can be traced back to eQTL mapping. As such, the methods combine linear regression with various phenotype normalization procedures such as the rank inverse normal transform (RINT). However, RNA splicing quantifications are quite different from gene expression estimates. Splicing is typically quantified as percent splice-in (Ψ) which is a measure in the domain of [0,1] of relative junction or isoform usage. Moreover, Ψ values are derived from sparse discrete read counts that span splice junctions. Accounting for the consequent uncertainty in these measurements by modeling read counts can likely increase

detection power and better control false discoveries. This is especially important in contexts where variant associations with splicing are confounded by associations with expression and coverage which may introduce heteroskedastic effects. Furthermore, splicing is a multivariate phenotype with most genes containing multiple junctions which requires additional considerations in statistical model design and multiple hypothesis testing correction.

Finally, there are several covariates that affect the power of marginal SNP-junction associations which should be considered to maximize discovery power. One covariate is the proximity of a SNP to a splice site. Most variants that are associated with splicing are near splice sites and disrupt core spliceosome or RBP binding motifs [14]. Another covariate is the coverage of the junction. Junction coverage can be affected by the sequencing depth of experiments, the expression of the gene, and proximity to the 3' end. The power to detect sQTLs is greatly reduced at low coverage since the variance of statistical estimators is higher and effective sample size is smaller due to dilution caused by imputation of missing values. Currently, sQTL analysis pipelines typically handle these covariates implicitly by only considering SNPs within a fixed window around a gene's transcription start site and filtering events beyond a certain level of missing values. However, detection power could be further improved by allocating the multiple hypothesis testing budget in a way that favors sQTLs which are more likely to be detected.

Previous works have attempted to improve various aspects of sQTL modeling. Many studies acknowledge that unannotated (denovo) splice junctions are important and therefore use Leafcutter [7] for detecting such junctions. Indeed, our analysis [6] shows LeafCutter offers an efficient method for detecting splicing variations, but it is not able to capture intron retention or alternative starts and ends. Other studies consider the count-based nature of splicing using generalized linear models like GLiMMPS [15] or DRIMSeq [16] but do not account for multiple splice junctions per gene. On the other hand, methods like sQTLSeeker [17] and THISLE [18] account for the multivariate nature of splicing phenotypes through multivariate testing or tests for heterogeneity respectively, but do not model the limited count data used in RNA splicing quantification. Finally, approaches like Independent Hypothesis Weighting (IHW) [19] have been proposed to handle covariates associated with power but are designed to control false discovery rates (FDR) for independent tests. In contrast, the highly correlated nature of splicing events and SNPs in sQTL analysis requires relaxed independence assumptions [20].

In this work, we first investigate the potential limitations of existing sQTL methodologies. Based on our findings, we then propose MAJIQTL, a new statistical framework for improved sQTL discovery and analysis. Our pipeline includes a comprehensive splicing phenotype representation derived from MAJIQ quantifications, a weighted multiple hypothesis testing correction method for sGene discovery utilizing covariates, and a Beta-Binomial composite testing model for accurate and interpretable sQTL effect size inference. These components not only find more sGenes/sQTLs, but also effectively prioritize the variants with functional

significance. We demonstrate the validity and utility of our approach through extensive simulations, real data applications, and assessment of variants' effect by orthogonal methods such as splicing prediction algorithms (splicing codes) and molecular experiments. When applied to the GTEx dataset, our method finds significantly more functionally relevant sQTLs which are enriched for important functional annotations and co-localize with neurodegenerative disease GWAS variants. Notably, we find and validate rs582283, a variant that is implicated in Alzheimer's disease. We show this variant disrupts the RBP binding site of YBX3 in the gene MS4A3 leading to skipping of exon 7, providing a plausible mechanism of action. We make our results easily accessible through a dedicated web-tool which can be further explored with VIOLA-QTL, a new sQTL visualization tool. Taken together, our results illustrate the power of careful examination and consequent improvement of multiple aspects of sQTL analysis methodology, leading to novel sQTLs which explain additional GWAS signals.

Results

Analysis of GTEx Tissues Points to Potential Improvement for sQTL Modeling

We start this study by first applying a standard sQTL (denoted std-sQTL) pipeline to samples from five representative tissues in the GTEx dataset: whole blood, lung, brain - cerebellum, liver and heart (see Supplementary Note for details). This pipeline was popularized by the GTEx consortium [21] and has since been adapted by dozens of modern studies with great success. We decomposed the std-sQTL pipeline into three elements: input splicing representation, statistical models for effect size inference and power informative covariates. We then evaluated each of the three elements and investigated how it could be potentially improved. Before we turn to describe this investigation we note that although we focus on these three elements, we also identified other important areas for consideration in the std-sQTL pipeline such as mapping, filtering and confounder correction. To maintain scope, we defer those to the Supplementary Note (Figure S1).

In the first step of sQTL detection assessment, we used the std-sQTL pipeline to call sGenes and assessed the effect of using different splicing representations. For a splice junction/event level representation, we quantified splicing (Ψ) using Leafcutter's intron cluster approach (most common method used in the std-sQTL pipeline) and MAJIQ's local splice variation (LSV) approach. For an isoform level representation, we quantified transcript expression using Salmon TPM and normalized these values to obtain the percent abundance of isoforms per gene. Although several studies highlighted the limited accuracy of whole transcript quantification from short read RNASeq data, numerous studies have used this approach for transcript QTL (trQTL) [18, 22] so assessing the value of this approach is warranted. The use of MAJIQ LSVs resulted in the discovery of a similar number of sGenes compared to the use of Leafcutter intron clusters across all tissues when only considering annotated (sGenes) and denovo (dGenes) splicing events (Figure 1a). However, the std-SQL pipeline reports between 1.4 times to 2 times more sGenes (or iGenes) with intron retention events when using MAJIQ

quantification of intron retention (IR) in heart and cerebellum respectively. These genes are otherwise not significant when only considering splicing events without intron retention (Figure S2). The use of Salmon isoform quantification also identified sGenes (isoGenes) that do not overlap those identified by junction level approaches. These sGene hits are enriched for isoforms that differ at transcript starts or ends. Overall, we find that using MAJIQ with transcript quantifications in the std-SQTL pipeline can capture two fold more genetically associated splicing variations compared to using Leafcutter.

Second, we assessed the statistical methods used to detect variants associated with sGenes and sQTLs. The primary challenge with sGenes is the multivariate nature while the challenge with sQTL effect size inference lies in the data. Splicing phenotypes such as those quantified by LeafCutter or MAJIQ are derived from multiple splice junctions in a gene. Thus identifying sGenes requires detecting non-zero effect sizes between a single genetic variant and multiple splice junctions. Various models have been proposed to handle the multivariate and discrete aspects of this phenotype. First, we consider the multivariate nature of the splicing phenotype. There have been several methods proposed for multivariate testing to detect sGenes. This includes MANOVA, Anderson's pseudo F test implemented in sQTLSeekeR2 [17], sum of χ^2 test implemented in THISLE [18], ACAT[23], and exact FWER control implemented in QTLTools [24] and used in the std-sQTL pipeline. We provide a summary of each method and their mathematical relationships in the Supplementary Notes. To assess how these methods behave on splicing data, we simulate genes with correlated splice junctions and make a subset of the junctions be associated with a genetic variant. The details of this simulation study are given in the Supplementary Notes. Then we evaluate the power of each method to detect sGenes. Figure 1b shows MANOVA and χ^2 perform well compared to sQTLSeekeR2, ACAT and FWER control. However, since it is difficult to generalize MANOVA to non-Gaussian distributions, we do not consider this class of approaches any further. Note that a Dirchlet-Multinomial regression model cannot work as a generalization of MANOVA in this setting since each LSV or intron cluster is a multinomial unit and there are multiple LSVs in each gene. Interestingly, the performance of sQTLSeekeR2 eventually surpasses FWER control when 40% of the junctions in a gene are associated with the variant. Overall, these result point to potential improvement in sGene detection power by using alternatives to the commonly used FWER control.

Next, we assess whether using linear regression with RINT transformations (std-sQTL approach) for effect size inference between single SNP-junction pairs is appropriate given that splicing quantifications are derived from discrete junction spanning read counts. First, we note that Ψ values generally follow a distribution skewed toward 0 or 1 and thus residual estimates are not normal. However, given the population scale data of modern QTL studies, linear regression models used in this context are still robust because the test statistic maintains asymptotic normality at reasonable convergence rates as a consequence of the central limit theorem [25]. Nonetheless, violations of Gauss Markov assumptions, such as heteroskedastic errors, can result in miscalibrated models regardless of sample size (Supple-

mentary Note). Indeed, we show that large heteroskedastic effects are persistent in the data due to the interplay between expression and splicing (Figure 1c). Specifically, a variant that affects expression (eQTL) can cause the splice junction coverage (and consequently variance of its quantification) to vary with genotype (cQTL). The example in this figure shows a gene in which the variant affects coverage and Ψ. The variance of Ψ increases from 2.67e-4 to 6.88e-3 as the coverage decreases (Figure 1c left and middle). Furthermore, we show that the percent of top K cQTLs which are also heteroskedastic sQTLs (significant Breusch-Pagan p-value) increases as K decreases (Figure 1c right). The genotype induced differences in coverage can also lead to differences in quantification resolution which effectively renders them discrete. For example, when the total coverage is 10, there are only 11 possible values of Ψ while there are 101 possible values of Ψ when the coverage is 100. This phenomenon cannot be addressed by the commonly used RINT transformation which assumes that the phenotype is homoskedastic and continuous. Thus it becomes necessary to select a generalized linear model (GLM) with a variance function that is properly specified for this data. We compare the goodness of fit of Beta, Binomial and Beta-Binomial models as well as a previously proposed Binomial model with mixed effects [15] against a Gaussian model using residual quantiles [26] on the lead SNP-junction pairs in sGenes (Figure S3). Surprisingly, the binomial model has the worst fit while the beta binomial has the best fit, suggesting that modeling overdispersion is crucial. Furthermore, we compute the agreement between the analytical and jackknife p-values to measure calibration of the linear regression model and to assess false positive control (Figure 1d). We show that 29.8% of sQTLs have a 10 fold difference in p-value (a shift of 1 decimal place) while 1.0% have a 70 fold difference. Interestingly, the misspecified linear model generally has lower analytical p-values compared to jackknife p-values (points mostly below the x = y line) suggesting an inflation of false positives. An important point to make in this context regards the common practice to test against synthetic nulls generated through data permutations in order to assess false positive rates [27]. Such synthetic null generation pools the variance, thus removing the effects of heteroskedasticity which consequently underestimates false positives and is not appropriate here.

Finally, we investigate potential covariates associated with the power of marginal associations. When conducting multiple tests, it is often the case that a researcher may want to make fewer tests based on a covariate. The covariate ideally informs the a priori chance of a significant signal to enable efficient filtering of uninformative hypotheses and reduce the multiple hypothesis testing burden. It is common practice to implement basic filtering based on covariates. The std-sQTL pipeline for example, tests SNPs up to 1 Mb away from a gene's transcription start site and filters junctions which are missing in more than 50% of samples. Other studies are more conservative, filtering junctions missing in 10% or more of the samples [28]. While sensible, these filters were borrowed from eQTL studies [29] and may be less effective in the sQTL setting. In subsequent analysis, we find that proximity to splice sites (Figure 1e) and the missingness rate of a junction (Figure S4) are two covariates associated with sQTL detection power. The missingness rate is associated with coverage since missing

values are more prominent at low coverage and are MNAR (missing not at random) rather than MCAR (missing completely at random) (Figure S5). Using a naive filtering approach, we show that reducing the 1 Mb window around a gene's TSS to a 0.1 Mb window around each junction's splice site significantly increases the number of associations (Figure 1f). Furthermore, increasing the number of tested junctions up to a 50% missingness rate greatly increases total discoveries (Figure S6). Notably, such a threshold based filtering approach can be seen as a special case of weighting where the filtering criteria is not correlated with the null test statistic (Supplementary Note). With fixed thresholds, the filtered cases have their weights reduced to 0 and the remaining budget is redistributed uniformly across the remaining tests. However, it has been shown that when correcting for multiple hypothesis testing, a more nuanced approach of redistributing the test budget based on a covariate can improve power [19]. Thus, we conclude from this analysis that while optimal hypothesis weighting by covariates for sQTL testing is not known, it is evident that improvements can be made to increase power.

MAJIQTL Offers a Robust and Powerful Statistical Framework for sQTL Discovery and Analysis

Motivated by the above observations, we developed MAJIQTL, a fast and light-weight statistical framework for sQTL discovery and analysis that builds on the popular MAJIQ method for RNA splicing quantification[5, 6]. Specifically, MAJIQTL features an exhaustive input splicing representation that integrates splice junction and full isoform level information; a weighted multiple testing method which improves sGene detection power by leveraging covariates; and a robust and interpretable regression model for sQTL effect size inference and variant prioritization. MAJIQTL is tightly integrated with the MAJIQ family of tools and contains many updates to those which we expand upon below. We now turn to describe each of the MAJIQTL components.

In the first step of the analysis pipeline (first column in Figure 2), we compile a comprehensive splicing phenotype representation combining splice junction quantifications from MAJIQ and full isoform level quantifications from Salmon. This allows MAJTIQTL to capture complex splicing variations involving unannotated junctions and exons, as well as intron retention and alternate transcript starts/ends. Integration with MAJIQ also enables use of its builtin event type classifier, which can annotate both classical and complex events into interpretable units (exon skipping, Alt 3', Alt 5', etc), and MOCCASSIN, a dedicated tool for splicing confounder correction. Furthermore, we introduce various best practice guidelines for splicing phenotype processing such as filtering procedures (Supplementary Note).

In the second step of MAJTIQTL's analysis pipeline (second column in Figure 2), we introduce a new weighted multiple hypothesis testing method to improve sGene discovery power. Accounting for multiple testing is imperative given the large number of tests performed between all pairs of splice junctions and the cis SNPs in a predefined locus around each gene. However, it has been shown that the naive approach of applying a single multiple testing cor-

rection procedure (e.g. Benjamini-Hochberg) to all tests fails to control false discoveries [30]. Instead, modern sQTL studies rely on a hierarchical correction strategy [21]. This approach first controls the sGene false discovery rate (FDR) at a desired level (e.g. FDR = 0.05) by using gene level p-values computed through the max T method to learn a decision threshold α. By rejecting only sGenes with a p-value below α, the family-wise error rate (FWER) of all tests in each gene is controlled at level α. This approach is able to properly handle the complex local correlation structure and number of tests unique to each gene unlike the naive approach. However, rejecting the max T p-value at level α represents unweighted FWER control where each test is valued equally in the FWER budget (i.e. α). Our case study shows that sQTLs are more likely to appear near splice sites or in junctions with low missingness rate. Thus we instead want the p-value to reflect weighted FWER where the FWER budget is allocated in manner that is proportional to the power informative covariate.

To address this need, MAJIQTL computes a weighted p-value for each sGene that increases discovery power while controlling local FWER under the null. The inputs are the covariates and test statistics for each sQTL. Given these inputs, the method first uses an empirical Bayes approach to learn a flexible non-parametric function that maps the covariates to weights that are proportional to the posterior probability of the corresponding test being alternative $P(H_1)$ using a Guassian Process (GP) regression (Figure 2). Notably, the GP function is optimized jointly over all exons, genes, and SNPs to avoid over-fitting and information leakage. Next, our MAJIQTL computes a gene level p-value using these weights such that rejection of the p-value at level α maintains a constant FWER budget but redistributes the budget proportional to the weight of each test (bottom of second column in Figure 2). The details and theoretical guarantees are fully described in Methods. However, there are two key innovations that we highlight here. First, the estimation of $P(H_1)$ does not require any numerical optimization. Instead, we use the Kolmogorov-Smirnov distance as a "plug-in" estimator and scale this value by a learned parameter λ which maximizes sGene discoveries. We will show that this estimator is a strikingly accurate approximation for the ideal weight $P(H_1)$. Second, the null distribution of the weighted test statistic cannot be computed analytically. We will show it can be sampled quickly and accurately from a matrix normal distribution, thus completely avoiding the use of permutation procedures. These two strategies enable our approach to efficiently scale to hundreds of millions of tests and only takes minutes to run on the GTEx datasets.

In MAJIQTL's third step (third column in Figure 2), we introduce a model for variant prioritization and effect size inference. A widely recognized limitation of current sQTL pipelines is the lack of dedicated methods for selecting candidate variants in sGenes. For example, the GTEx portal only reports a single sQTL (lead SNP) with non-zero effect on splicing in each sGene. However, a researcher may be interested in all the sQTLs with the largest effect sizes (identifying pathogenic splice site disrupting variants) or may want to omit the sQTLs with small effect sizes as those may be hard to validate experimentally and may be less attractive for assessing phenotipic effects. Thus we would ideally like to

form sets of variants defined by a minimum effect size threshold such that all variants in the set exert an effect on splicing that exceeds that threshold. We can then prioritize the sets defined by larger thresholds in downstream analysis. However, as shown by our case study, effect size estimates from existing models are limited by misspecifcation and thus not comparable between sQTLs. Specifically, excessive phenotype normalizations obfuscate inference of effect sizes that can be interpreted on the same scale. Furthermore, even when using a well calibrated model, simply filtering by the effect size estimator without considering confidence in the estimate may be inappropriate [31]. Notably, even though effect sizes are considered for eQTL [32], this issue of confidence intervals around effect estimates is still not incorporated in the GTEx and similar eQTL pipelines.

To address these needs for variant prioritization and effect size inference, we propose using a composite Beta-Binomial model to identify thresholded rejection sets. These are illustrated at the bottom of the third column in Figure 2. In brief, this variant prioritization strategy starts with all the sQTLs that pass the FWER bound in each sGene which have confirmed non-zero effect size. Then we use a beta-binomial regression model on this data to infer effect size magnitude $\hat{\beta}$ (Methods). This allows us to handle uncertainty in splicing quantifications due to coverage and heteroskedastic effects and the coefficient naturally has a fold change interpretation which is the log odds of junction inclusion (Figure 2). This measure for effect size is appropriate as it has been shown that splicing dynamics are a competition for splice site usage and splicing changes induced by variants in *cis* can be modeled well using a sigmoid like switch between splice sites [33]. To define the composite rejection set, rather than rejecting variants under the null hypothesis $\hat{\beta} = 0$, we reject under the composite hypothesis $|\hat{beta}| \leq \theta$ where θ is a lower bound on tolerable effect size. This testing procedure, similar to ones previously applied for gene expression changes [34], is described in Methods.

The MAJIQTL methods are deeply integrated into the popular MAJIQ software package for RNA splicing analysis. They are designed as "plug-in" replacements for the commonly used Leafcutter and QTLTools with comparable inputs and outputs. Specifically, MAJIQTL can be run without any programmatic interface, requiring at minimum only a BED file containing splicing quantifications and a VCF file containing variant information to perform all operations. Using these standard files as input, the MAJIQTL software is equipped with several additional tools that improve speed and user accessibility. First, we implemented a new library for the Beta-Binomial model which achieves improved speed and accuracy though parallelization and use of exact gradients in optimization. Second, we include a dedicated interactive visualization tool in VOILA-QTL (bottom right illustration in Figure 2). VOILA-QTL is an extension of MAJIQ's VOILA visualization tool with a slew of new features to enable visual analysis of sQTLs (Figure 2). Finally, our results are distributed through the MAJQTL database which is easily accessible online (see Data Availability). We note that although our approach is designed to integrate and take advantage of MAJIQ's features, our pipeline is still compatible with other splicing quantification tools such as Leafcutter and rMATs.

MAJIQTL Weighted Multiple Testing Improves sGene Discovery Power

We first evaluate the power of our weighted multiple testing method to discover sGenes in the five representative GTEx tissues. We obtain a p-value for each gene by providing the model with summary statistics computed using RINT OLS between every splice junction and cis SNP within a 1 Mb window around the gene and matching covariates for each junction and SNP. The two covariates are the normalized distance of a SNP to the paired junction's splice site and the missingness rate of each splice junction across all samples. We also apply the independent filtering method to these datasets. This baseline method reports the standard max T p-value for each gene for incrementally decreasing cis window sizes and can be interpreted as a binary weighting approach (filtered tests have zero weight and the remaining tests have uniform weight). The results for brain-cerebellum are shown in Figure 3 while the results for the remaining tissues are shown in Supplementary Note. For a given FDR rate, the power of our method is significantly improved compared to the baseline (Figure 3a). At a FDR of 0.05 our method finds 12% more sGenes than independent filtering at a 0.1 Mb window size. This figure also illustrates the behavior of optimizing the scaling parameter which achieves optimal power on this dataset at $\lambda = 5$.

Next we confirm that our method still maintains false positive control under the null. To show this, we first generate synthetic null datasets using a permutation scheme that preserves the correlation between junctions and between SNPs (Supplementary Note). Then, we show that our method controls false positives by producing p-values that are uniform when applied to this null data (Figure 3b). It is reasonable to assume that false positive control is only achieved since the weights estimated by the GP trained on null data are approximately uniform and thus the weighted p-value is similar to the standard max T p-value. However, we find that our model's p-values are well calibrated regardless of whether we trained the GP on the null data or original data (non uniform weights). Furthermore, we observe the same degree of false positive control using weights randomly drawn from a standard normal distribution. These result together point to the robustness of our approach in achieving control regardless of how the weights are estimated.

The power of our method depends on how well the GP regression model is able to map covariates to $P(H_1|C)$. However, as previously described, we instead use a strategy of mapping to a plug-in estimator (KS distance between the null and observed empirical distribution of test statistics) and then scale this value to maximize sGene discoveries. Here, we explore how well this approach approximates a mapping to $P(H_1|C)$. We perform a simulation where we draw test statistics from a two component mixture of normal distributions where each component represents the null and alternative distributions. The mixing proportion, $P(H_1|C)$, is assumed to be linearly correlated to the covariate. We note that our model can theoretically handle a non-monotonic relationship between $P(H_1|C)$ and the covariate, however such behavior is not observed in real sQTL datasets and we therefore do assess such relations. The full details of the simulation are given in Supplementary Note. Using the aforementioned simulated data we then show that $P(H_1|C)$ can be approximately recov-

ered using our modeling approach. For this, , we use Spearman correlation to measure the agreement between the ground truth values of $P(H_1|C)$ and the KS statistics, as shown in Figure 3c. We used three different settings for the simulated data: Data were the mean (red bars) or variance (blue bars) of the alternative distribution was correlated to $P(H_1|C)$, or data generated such that the distribution was empirically constructed from real sQTL data (green bars). These results illustrate that our fast approximation for the weights results in near optimal power.

Next, we turn to show the quality of approximating the null distribution through sampling. The test statistics for a gene are assumed to follow a matrix normal distribution with mean zero under the null. We can sample from this distribution instead of using permutations to generate an empirical null, but the quality of this approximation is limited by the estimation of the covariance matrices. In brief, the covariance matrices can be naively estimated using the sample correlation between SNPs and between junctions. This approach has been used in similar methods like gene-MVN for eQTLs [35]. However, when the number of features (SNPs) greatly exceeds the number of samples, the quality of the covariance estimator is reduced. We remedy this using the Ledoit-Wolf shinkage estimator. We compare the naive sampling approximation and sampling approximation with shrinkage to the gold standard permutation based approach. The p-values from our approach have strong correlation with the p-values from the gold standard method which improves after using shrinkage (Figure 3d). Critically, although this approach is obviously not exact, it provides an accurate approximation while being orders of magnitudes faster than standard permutation (Supplementary Note).

Finally we show how our approach can be used to identify sQTLs within sGenes. We show the FWER bound learned by our model for an example gene in Figure 3e. This bound corresponds to the critical values that controls FWER at a level equal to the BH threshold $\alpha = 0.022$ for sGenes. For this dataset, this specific BH threshold controls FDR at a 0.05 rate. All tests in the gene with test statistics above this bound can be considered sQTLs and this procedure will maintain sGene FDR control at FDR = 0.05 and sQTL FWER control at level $\alpha = 0.022$ under the null. Importantly, the lead SNP is guaranteed to have a test statistic above this bound if the sGene is significant and below if not significant. For comparison, we also show the unweighted max T bound and Bonferroni bound. The Bonferroni bound is too conservative and finds no sQTLs. The max T bound, while less conservative, misses sQTLs near splice sites that the weighted approach is able to detect. This example clearly illustrates the advantage of our approach for sQTL detection since we can analyze more than just the lead SNP which is prohibitively limited.

MAJIQTL Refines Variant Prioritization and Effect Size Inference

Next, we evaluate the performance and utility of our effect size inference model and variant prioritization strategy. We begin with a simulation study to analyze the behavior of the beta-binomial regression model for sQTL effect size estimation. Our aim is to understand whether

the model is well calibrated for sQTL data and thus suitable for composite testing. We simulated splice junction and genotype data under 32 parametric conditions that can affect data factors such as MAF, skewness ($E(\Psi)$), overdispersion (Φ), coverage (λ), and changes in coverage (Y/N). The genotypes were simulated under Hardy-Weinberg equilibrium ratios based on the MAF. Further details are given in Supplementary Note. We evaluate the false positive rate and power of our model under the standard null hypothesis ($H_0 : \hat{beta} = 0$) compared to two baselines: linear regression and linear regression with RINT transformed phenotypes (Figure 4a). We find that the Beta Binomial model is the uniformly most powerful and controls the false positive rate while the other models fail in some scenarios, as we detail next.

The simulation results shown in Figure 4a offer important insights on the various modeling approaches behavior. First, as expected, RINT regression and OLS both fail to control false positives in the presence of heteroskedastic effects under the null. This is reflected in the simulations where the coverage is also associated with genotype and usually manifests in real data when a SNP is also an eQTL. We observed that the problem is also exacerbated at low MAF. Since OLS assumes homoskedasticity and uses a pooled variance model, the variance estimate is dominated by the variance of the largest genotype group. When allele frequencies are imbalanced, the variance can be systematically over or under estimated. In the case of RINT regression, the model also suffers from the assumption that the phenotype data is continuous. However, at low coverage, splicing quantifications can effectively be regarded as discrete. Specifically, when the phenotype in one genotype group is discrete (e.g. the total coverage is 10 and there are only 11 possible phenotype values) but continuous in another (e.g. the coverage is much higher), the ranks of the values cannot be compared. Doing so leads to bias in the mean estimate.

Another key point regarding effect size estimation that is illustrated by our simulation analysis in Figure 4a is the differences between fold changes and changes in $\Delta\Psi$. By design, the Beta Binomial model we use has substantially higher power to detect sQTLs with large differences in fold change instead of $\Delta\Psi$. In these simulations, we used $\Delta\Psi$ as the effect size measure (set at $\Delta\Psi = 0.01$). However, when the minor allele mean is 0.5 and major allele mean is 0.52, $\Delta\Psi$ is 0.02 but the change in the log odds ratio or $\Delta log(Logit(\Psi))$ is 0.08. In contrast, when the minor allele mean is 0.95 and the major allele mean is 0.97, the effect size is still 0.02 in $\Delta\Psi$ space but 0.53 in $\Delta log(Logit(\Psi))$ space. Therefore, even through the effect size in terms of $\Delta\Psi$ is constant in all simulations, the gain in power for Beta Binomial is larger when the minor allele mean is close to 0 or 1. We argue this property is desirable for two reasons: First, many disease associated phenotype occur with low MAF where the Beta Binomial gains more power. Second, previous work showed that alternative splicing can be modeled as competition between splice sites, each with its own usage rate [33]. Consequently, genetic variants that effect the splicing rate of a specific junction results in a non linear, sigmoid shaped, effect in ΔPsi space, depending on the initial rates ratio. For example, we may observe two isoforms that exist in a 1:1 ratio and thus have a Ψ of

0.5. A large 10 fold increase in 1 isoform caused by a variant would result in a 1:10 ratio. This corresponds to $\Delta\Psi = 0.41$ and $\Delta log(Logit(\Psi)) = 2.3$. However, when the same 10 fold change is observed when the initial ratio is $1:10$ and increases to $1:100$, $\Delta\Psi = 0.08$ while $\Delta log(Logit(\Psi))$ remains the same at 2.3. For this example, it can be argued that such a gain in power for $\Delta\Psi = 0.08$ may be marginal in some biological contexts and that confidently capturing such a change is hard. We note though that researchers can still filter hits by MAJIQ's built in $\Delta\Psi$ criteria and stress that under our model the significance of such a difference is still assessed given the observed coverage level and under composite testing (see below).

Next, we sought to assess the scenarios under which the Beta Binomial and RINT OLS tend to agree or disagree when analyzing real data. We divided the tests into 4 bins based on the absolute difference between the p-values of the RINT model and Beta-Binomial model. Then in each bin, we investigated the distribution of Ψ and coverage as shown in Figure 4b. The simulation study described above suggested the Beta-Binomial model has higher power to detect associations with large isoform fold change which occurs when the mean Ψ is close to 0 or 1. Inline with this, the results on real data shown in Figure 4b demonstrate the models tend to disagree the most when coverage is low and Ψ values are skewed to 0 or 1. When investigating cases of disagreement, we find that the Beta-Binomial model is robust to low coverage outliers (see example in Figure 4b).

Having shown the regression model is well calibrated for sQTL data, we then turn to evaluate the composite testing procedure. Without composite hypothesis testing, researchers will typically construct a rejection set using associations that are significant under the standard null hypothesis and filter sQTLs with observed effect size estimates below a desired threshold θ. It is instructive to first illustrate how composite testing differs from this baseline approach by showing the decision boundaries for both approaches (Figure 4c, bottom scatter plot). Composite testing produces a non-linear decision boundary [34]. Specifically, the threshold for the effect size estimator increases as its standard error increases thus more evidence is required to reject the hypothesis that the effect size is smaller than θ when the standard error is high. The primary contributors to higher variance in estimates include low junction read coverage and minor allele frequency (Figure 4c, top box plots). The filtering approach uses the standard null hypothesis decision boundary (equivalent to composite testing with $\theta = 0$) but then applies a variance agnostic cutoff. This "mixture boundary" can completely ignore the variance of the estimator at high effect sizes ($\theta = 0.6$ in the example). Figure 4c top graph demonstrates that as a result of the aforementioned differences in decision boundaries, composite testing controls the FPR at or below the desired level. In contrast, the filtering approach can exceed the desired FPR when the filtering component begins to dominate (see simulation details in Supplementary Note).

To evaluate the composite testing approach, we compare agreement between sQTLs in the composite rejection set and high confidence effect size annotations that we treat as ground

truth labels. We cannot know the true effect size (i.e. population parameter) of each sQTL. Instead we use an information sharing strategy to combine estimates across tissues under the assumption that effect sizes are shared across tissues. This assumption has been used in many other studies [36]. Specifically, for a given sQTL, we compute the 95% confidence interval of its effect size estimate in each of the 5 representative tissues. Then we take the intersection of these intervals to be a high confidence range for the population parameter. To maintain a high quality set, sQTLs with tissue specific effect sizes in which confidence intervals across tissues do not all overlap are omitted. A sQTL is then considered to have an effect size greater than the desired threshold θ if the combined interval does not overlap and is not between θ and $-\theta$ (see Supplementary Notes for more details and illustrative examples). Using these high confidence effect size annotations as ground truth labels, we find that sQTL membership in the composite rejection set at varying FPR levels closely matches the labels as measured by AUC (Figure 4d). Importantly, for all thresholds θ, composite testing outperforms the naive filtering approach.

Finally, we show that using the composite sets to prioritize variants results in finding more sQTLs which are likely to disrupt splice sites usage, hence arguably more likely to have functional importance. For a given effect size threshold θ, we compute the rejection set at varying log10 p-value thresholds using either the composite p-values or original p-values. We then compute the enrichment of each of those rejection sets for splice site disrupting variants (Figure 4e). Here a variant was considered splice site disrupting if SpliceAI [37] predicted a change in splice site usage probability greater than 0.2. The 0.2 score is recommended by the authors for high recall. See Supplementary Note for further details of the enrichment calculation. We observe that the enrichment increases with effect size and p-value threshold. Importantly, the composite approach results in higher enrichment at all effect sizes.

MAJIQTL Improves sQTL Enrichment in Functional Annotation and Identifies Novel sQTLs that Explain GWAS Signal for Neurodegenerative Disorders

Having shown the merit of each individual component of the MAJIQTL framework, we then use the full pipeline to call sGenes and analyze the functional enrichment of their sQTLs. Specifically, we use the weighted multiple testing method to call sGenes and call sQTLs that pass the weighted FWER bound for each sGene controlled at the BH threshold. We compared our approach to the std-sQTL approach which uses the max T procedure (unweighted FWER) to discover sGenes with Leafcutter quantifications and selected sQTLs within sGenes that passed a gene level Bonferroni threshold. We analyzed enrichment of the sQTLs for two SNP functional annotations: splice site disrupting variants and RBP binding sites. A sQTL was considered splice site disrupting if it was predicted to have a SpliceAI delta score greater than 0.2 for the 5' or 3' splice site of the sQTL's splice junction. We considered a RBP to bind to a sQTL if it had an eCLIP peak that overlap the variant in the ENCODE dataset for any of the 113 RBPs in K562 cell lines. Binding was determined using an IDR threshold which is the stringent cutoff recommended by ENCODE. We find that sQTLs in MAJIQTL sGenes are enriched for these annotations compared to sQTLs in non

sGenes at a 0.05 FDR level across all 5 tissues based on Fisher's exact test (Figure 5a). The std-sQTL pipeline using Leafcutter has lower enrichment in all tissues. Notably, we found that only using lead SNP approach did not yield sufficient enrichment for either method due to the highly varied positions of lead SNPs which may not be casual.

Finally, we analyze the sQTLs discovered by the MAJIQTL pipeline to understand if they can provide new insight on disease mechanisms. We first obtained summary statistics for Alzheimer's and Parkinson's disease from the studies Kunkle et al. 2019 and Nalls et al. 2019 respectively. We find that the sQTLs identified by MAJIQTL explain 11% and 8% more GWAS variants for Alzheimer's and Parkinson's respectively compared to std-sQTL (Supplementary Note). Using our variant prioritization approach, we identify rs582283 in the composite rejection set at a 0.2 composite level as a variant of interest which may be implicated in disease (Figure 5b). This variant is associated with the splicing of exon 7 in the gene MS4A3 (p=3.55e-7) but does not appear in the GTEx catalogue as a sQTL for this splicing event. It is also a GWAS signal for Alzheimer's (p=1.27e-8) and a significant sQTL in blood but surprisingly not an eQTL (p=5.51e-2). The variant is observed to have strong LD with other variants in the gene but not variants outside of the gene. Since the variant is not significant for expression, this points to splicing disregulation as a potential mechanism underlying the disease association. Inline with this hypothesis, we find rs582283 is located on exon 7 which is in a cassette event with a denovo junction. Exon 7 is an important transmembrane domain and exclusion may have implications for loss of protein function. The variant appears to promote skipping of exon 7 with an estimated effect size of $\Delta\Psi = 0.132$ or 5.08 fold when the variant changes from a C to T. Furthermore, the variant is predicted to be a splice site disrupting variant with a SpliceAI score change for each adjacent splice site (5' and 3') of 0.2. The variant overlaps the binding motif for the YBX3 RNA Binding Protein and the position of the variant in the motif is non degenerate suggesting that the change in variant significantly reduces binding affinity. Supporting this hypothesis of rs582283 disrupting RBP binding, when we target this region with an antisense oligonucleotide (ASO) we observe a significant change in splicing (Figure 5c). Details for the experiment are describe in Supplementary Note. Furthermore, ENCODE K562 eCLIP experiments suggest YBX3 binds to this position (Log2FC = 3.79 increase compared to background), while shRNA knockdown of this RBP reduces the inclusion rate of the exon compared to controls ($\Delta\Psi = 0.118$).

Discussion

In this work, we decompose the current state of the art sQTL analysis pipelines into their fundamental components and individually evaluate each component on five representative GTEx tissues. The results of this case study set a precedent for future work by identifying three key limitations with existing pipelines and demonstrating how they could be addressed to improve sQTL discovery and analysis. First, we show that using the Leafcutter splicing representation can miss crucial splicing events like intron retention and alternate

transcript starts/ends which represent a significant number of sQTLs that explain GWAS signal. This result is inline with other works demonstrating the importance of capturing additional types splicing variations. For example, our previous work experimentally verified the effect of rs6410, a variant associated with skeletal aging, in promoting retention of intron 3 of CYP11B1 [38]. More generally, several other studies have highlighted the importance of integrated splicing representations, finding more variants explaining disease loci through full isoform and alternate transcript end (APA) QTLs [18, 39, 40].

Second, we show that the current statistical methods used for sGene/sQTL discovery are under powered and not well calibrated. It has been suggested that QTL discoveries are saturated and improvements to statistical techniques are unlikely to further close the co-localization gap [41, 42]. However, our results indicate that using multiple hypothesis testing correction based on FWER control for sGene discovery is conservative compared to alternative methods. This conclusion is concordant with the findings of a recent study that benchmarked methods for isoform QTL discovery [43]. We show that improved statistical modeling finds new sGenes harboring sQTLs which are enriched for various functional annotations and variants implicated in neurodegenerative disease. In addition, through comparison of methods for sQTL effect size inference, we show that the heteroskedastic and discrete characteristics of splicing data necessitate the use of a Beta-Binomial model. Importantly, the effect sizes of sQTLs that are also eQTLs cannot be reliably estimated by OLS due to inconsistent variance in splicing quantifications induced by differences in gene expression and coverage. The resulting inflation of false positive non-zero effects serves to illustrate that simply more hits is not always better, and carefully assessing the models' fit to the data is critical. In a broader context, we suspect that this model misspecification is partially responsbile for the widespread inconsistent reporting of overlap between eQTLs and sQTLs [4, 18, 21].

Third, we illustrate the limitations of current practice for SNP and junction selection in cis sQTL analysis. This topic is seldom discussed in recent QTL work. Instead, most sQTL studies opt to borrow the selection parameters directly from eQTL studies, using the same large window size (1Mb), discarding splicing events with even 10% missing values and imputing the rest [21, 28, 29]. We show this current standard for study design can severely decrease detection power by needlessly increasing the multiple hypothesis testing burden and potentially filter out junctions with missing values that may harbor strong sQTL signals. However, we show this problem can be easily mitigated by utilizing power informative covariates (splice site distance, missingness rate), emphasizing the importance of not discarding data.

To address the issues from this cases study, we developed here MAJIQTL, a novel sQTL discovery toolkit with two new statistical methods to improve sQTL mapping and facilitate downstream analysis. First, we introduce a weighted multiple hypothesis testing method (which can be seen as a sGene test) that leverages SNP and junction level covariates to

significantly increase sGene detection power. We note such weighted control procedures are not new, but previous methods have not been adapted in the QTL field due to their focus on FDR control and broad applicability [19, 44, 45]. The appeal of our method lies in its design tailored for QTL studies which requires a FWER based approach for false positive control in the presence of complex LD and junction correlation structure [30, 46]. Beyond increased power, we show that controlling the FWER of each sGene at the BH critical value allows for individual sQTLs to be called at a constant FWER while maintaining gene level FDR control. Despite the widespread use of lead-SNP analysis due to current methodological limitations [18, 40, 47], we highlight how our approach which uses multiple sQTLs per gene leads to increased enrichment for splicing relevant functional annotations.

Second, we propose using a composite Beta-Binomial model for effect size inference and variant prioritization. Unlike previous applications of similar models for sQTL detection [15], we focus on the utility of using the model parameter as an interpretable measure of sQTL effect size. Notably, there is currently no consensus definition for sQTL effect sizes reported on the GTEx portal despite extensive use of the allelic fold change metric for eQTLs [32]. For RNA splicing, two common measures are fold change and $\Delta\Psi$. The latter is arguably the most commonly used unit for studying splicing regulation though several studies have advocated for using fold change to represent the effect of genetic variants on splicing [33, 48]. In our case, the fold change arises as the natural effect size from the Beta-Binomial model, though we note that any identified sQTL can also be filtered by MAJIQ's $\Delta\Psi$ as well. We show that the fold change effect size estimates of our model are both accurate and informative of functional annotation. We also use a composite testing method to create minimum effect size sets as a way to prioritize variants. A related approach using fine mapping showed that variants in fine mapped credible sets for QTLs are enriched for disease heritability [49]. Similarly, we show our composite sets are enriched for important functional annotations such as splice site disruption and RBP binding. This provides a complementary set of annotations which we demonstrate are useful for variant prioritization. We show that the naive approach of selecting significant sQTLs (with non-zero effect size) followed by filtering for a desired effect size results in lower accuracy and variant enrichment. Although we only investigate the benefits of minimum effect size sets, we emphasize that composite testing can also be used to create maximum effect size sets. Identifying such sets may have potential application for developing therapeutics since variants exerting small effect sizes on splicing with a strong phenotype are more likely to be efficiently blocked by antisense oligos (ASO).

Beyond sQTL, we believe our effect size model and annotations can be used for other applications that involve predicting genetic effects on splicing. For example, splicing TWAS is an alternative approach for identifying risk variants that are mediated by splicing. However, current models still rely heavily on multivariate linear regression for imputing splicing phenotypes [50]. It is not clear whether focusing on a multivariate strategy is better than a parametric strategy since our results suggests that using the correct error model is crucial for estimating the genetic component of splicing effect size. Our results can also be used

to train and validate deep learning models that predict tissue specific splicing from DNA sequence. Similar ideas have been used for expression prediction models with eQTLs [51]. It is generally believed that cis-QTLs are shared across tissues but may have tissue specific effect sizes [36]. We identified many instances of tissue specific effect sizes by using composite testing and confidence intervals which can be used to train tissue specific prediction models. Although several methods have been developed to predict tissue specific Ψ from genomic sequence [52–55], the most successful methods for predicting the effect of genetic variants can still only predict splice site location which is effect size and tissue agnostic [37]. This work thus has the potential to lead to improvements in tissue specific splicing code models.

Finally, through applications to real GTEx data, we show that our method discovers many novel sQTLs which are also significant in Alzheimer's and Parkinson's GWAS. We highlight a new variant, rs528823, which may be linked to Alzheimer's through splicing rather than gene expression. Specifically, the variant appears to disrupt the inclusion of an exon in the gene MS4A3 by perturbing a YBX3 binding motif, a hypothesis supported by our ASO targeting of this site as well as YBK3 CLIP binding and shRNA KD. Although our result is only significant in blood, other genes in the MS4A gene cluster have been previously implicated in Alzheimers through immune cell related functions which may act through blood [56].

There are several limitations with our approach that are important to highlight. First, isoform quantification from short read data are generally considered to be inaccurate. There is no guarantee that isoform level sQTLs discovered are real due to errors in the quantification. However, at the pace that long read technology is improving, we believe these quantifications can be easily replaced with long read data and our approach supports that substitution. Second, we don't address trans-QTLs in this work. While MAJIQTL can recover many common and cis variants associated with splicing, many variants that colocalize with GWAS signal can be rare and trans-acting. Detection of these variants requires complementary approaches. Specifically, splicing code models can reveal rare variant effect on splicing using a single patient's DNA sequence. Such models have been effective in finding cryptic splice site resulting from variants with low frequency in populations [57]. Trans-QTLs are generally mediated by cis effects. For example, a distal regulatory splice factor could have cis QTLs which are trans QTLs for distal splicing targets. Modern approaches have used CRISPR or Perturb-Seq screens to knock out regulatory factors and associate cis-QTLs of these factors with splicing of their target genes[58]. Finally, our study is focused on methodological advancement rather than exhaustive and robust sQTL mapping. As such, it does not include a replication study. Assessing replication is an important component which helps land confidence in both the methods and consequent results. However, we caution that replication alone does not necessarily indicate correctness. Specifically, several of our results point to the issue that misspecified models can consistently produce recurrent false hits.

In conclusion, we believe that MAJIQTL will significantly advance sQTL discovery given the many improvements it offers over existing methods. Furthermore, while the pipeline

integrates with the MAJIQ tool set, it is compatible with existing tools such as Leafcutter, rMATs and QTLTools. We hope the genetics community will take full advantage of this new method for RNA splicing analysis and potentially modeling of other QTL.

Methods

A Weighted FWER Control Model for sGene Discovery

Here, we present a method for cis sGene discovery that leverages splice junction and SNP covariates to improve detection power. For a gene $g \in [G]$ with J_g splice junctions and K_g SNPs in a predefined genomic window around the gene, consider the set of $J_g \times K_g$ marginal hypothesis $\{H^g_{jk} | j \in [J_g], k \in [K_g]\}$ with corresponding test statistics z^g_{jk}. We assume that these test statistics are computed using linear regression between each pair of junctions and SNPs but do not assume they are independent under the null. Let a sGene be defined as a gene with at least one true alternative marginal hypothesis. Then to test for sGenes in the setting without covariates, we wish to reject the complete null hypothesis H_0. Formally,

$$H_0 : \{H^g_{jk} \in \mathcal{H}_0 \ | \forall j, k\}$$
$$H_1 : \{H^g_{jk} \notin \mathcal{H}_0 \ | \exists j, k\}$$

This hypothesis can be tested for each gene using $t_g = \max_{\forall j,k}(|z^g_{jk}|)$ as the test statistic, as described in Westfall and Young 1993 [59], to compute an unweighted gene level p-value. The p-value is given as

$$p_g = 1 - \hat{F}^{-1}_{t_g}(t_g)$$

where $\hat{F}^{-1}_{t_g}$ is the CDF of the null distribution of t_g which is typically estimated using a permutation based approach (we describe the details in a later section). Then a false discovery rate (FDR) control procedure such as Benjamini Hochberg (BH) is applied to control the sGene FDR at a desired level across all genes by determining the sGene p-value decision threshold α. This approach for genome-wide FDR control is equivalent to controlling the family-wise error rate (FWER) for each gene at level α. Specifically,

$$\alpha = \bigcup_{\forall j,k} P(|z^g_{jk}| \geq C^g(\alpha) | H_0)$$
$$\alpha = P(t_g \geq C^g(\alpha) | H_0)$$

where $C^g(\alpha)$ is the uniform critical value for FWER α. We emphasize that $C^g(\alpha)$ is different for each gene. Now consider the setting where we have covariates $\gamma_1^g, \ldots, \gamma_J^g$ for each junction and $\delta_1^g, \ldots, \delta_K^g$ for each SNP in gene g. In this study, we treat the junction covariates as the missing value rate $\gamma \in [0, 1]$ and the SNP covariates as the normalized proximity to the splice site which is normalized w.r.t to the maximum allow window size in bases $\delta \in [0, 1]$. Critically, these covariates must be independent of the p-value under the null [60]. However, if the covariates are correlated to $P(H_{jk}^g = 1)$, we can specify weighted critical values $C_{jk}^g(\alpha)$ that allocates more of the FWER budget to tests where the marginal alternative hypothesis is more likely to be true subject to the constraint that the FWER is still controlled at exact rate α. To derive these weighted critical values, we factor $C_{jk}^g(\alpha)$ into a covariate dependent weight component w_{jk} and a gene level constant $\Delta^g(\alpha)$ dependent on α that is used to maintain the FWER budget.

$$\begin{aligned}
\alpha &= \bigcup_{\forall j,k} P(|z_{jk}^g| \geq C_{jk}^g(\alpha)|H_0) \\
\alpha &= \bigcup_{\forall j,k} P(|z_{jk}^g| \geq w_{jk}^g + \Delta^g(\alpha)|H_0) \\
\alpha &= \bigcup_{\forall j,k} P(|z_{jk}^g| - w_{jk}^g \geq \Delta^g(\alpha)|H_0) \\
\alpha &= P(\max_{\forall j,k}(|z_{jk}^g - w_{jk}^g|) \geq \Delta^g(\alpha)|H_0)
\end{aligned}$$

The proof for this decomposition is given in Supplementary Note X. Then, in contrast to the unweighted p-value, the weighted p-value per gene can be computed as

$$\begin{aligned}
t_g &= \max_{\forall j,k}(|z_{jk}^g - w_{jk}^g|) \\
p_g &= 1 - \hat{F}_{t_g}^{-1}(t_g)
\end{aligned}$$

Assuming the weights w_{jk}^g are known, the weighted FWER rejection bound $CV_{jk}^g(\alpha)$ can then be recovered by choosing $\Delta^g(\alpha)$ such that

$$\alpha = 1 - \hat{F}_{t_g}^{-1}(\Delta^g(\alpha))$$

Individual sQTLs within sGenes can be identified if $|z_{jk}^g| \leq CV_{jk}^g(\alpha)$.

In the remaining sections, we discuss 1) how to estimate w^g_{jk} to increase the number of sGene discoveries and 2) how to estimate $\hat{F}^{-1}_{t_g}$. It is important to emphasize that the FWER under the complete null is controlled for this procedure for any $w^g_{jk} \in \mathcal{R}$, even if w^g_{jk} is not chosen optimally.

Weight Optimization

Now we describe a fast approach for choosing the weights w^g_{jk} that increases the number of sGene discoveries. This method provides an approximately optimal solution that scales to hundreds of millions of tests. Our goal is to learn some function g that maps covariates γ^g_j and δ^g_k to their corresponding weight w^g_{jk} such that the number of sGene discoveries is maximized. It has been shown that the theoretical optimal quantity for the weights is the probability of the alternative hypothesis $P(H^g_{jk} = 1)$. This is often called the local false discovery rate (lFDR) [61]. Consider the two group model for the observed test statistics as described in Efron 2001 [62]. Under this model, we assume that the test statistics follow a mixture of two normal distributions. One mixture component represents the null distribution $f_0(z)$ while the other represents the alternative distribution $f_1(z)$. In the setting with covariates, the mixing proportions π and alternative distribution depend on the covariates. Formally,

$$
\begin{aligned}
z^g_{jk} &\sim \pi(\gamma^g_j, \delta^g_k) f(z^g_{jk})_0 + (1 - \pi(\gamma^g_j, \delta^g_k)) f(z^g_{jk})_1 \\
f_0(z^g_{jk}) &= N(0, 1) \\
f_1(z^g_{jk}) &= \int_{-\infty}^{\infty} N(\mu, \sigma^2) P(\mu | \gamma^g_j, \delta^g_k) d\mu
\end{aligned}
$$

The lFDR can then be formulated as the posterior probability of the alternative hypothesis conditioned on the covariates.

$$
\widehat{lFDR}(\boldsymbol{Z}^g_{jk}, \gamma^g_j, \delta^g_k) = P(H^g_{jk} = 1 | \boldsymbol{Z}^g_{jk}, \gamma^g_j, \delta^g_k) = \prod_{z^g_{jk} \in \boldsymbol{Z}^g_{jk}} \frac{\pi(\gamma^g_j, \delta^g_k) f_0(z^g_{jk})}{f(z^g_{jk})}
$$

$$
f(z^g_{jk}) = \pi(\gamma^g_j, \delta^g_k) f_0(z^g_{jk}) + (1 - \pi(\gamma^g_j, \delta^g_k)) f_1(z^g_{jk})
$$

If we could successfully perform inference for the above model, we would have our mapping function f_w and could stop here. However, maximum likelihood inference of the model parameters is challenging and generally not feasible at scale [44, 62]. Instead, tractable solutions often use a data driven approach to derive weights. In brief, these methods search over the space of all possible weights subject to a budget constraint to find the values that maximizes hypothesis rejections in the dataset. Although this objective appears to be NP

hard, various approximations have been proposed [19, 45]. We use a simple 3 step procedure to approximate the optimal weights conditioned on covariates. We outline the steps and advantages of this approach below.

1. Compute the pseudo lFDR estimator $\hat{\omega}^g_{jk}$ which has the property $\hat{\omega}^g_{jk} \propto \widehat{lFDR}(\boldsymbol{Z}^g_{jk}, \gamma^g_j, \delta^g_k)$. This is a cheap plug-in estimator that does not require numerical optimization.

2. Learn the function $g(\gamma, \delta)$ that maps the covariates γ^g_j and δ^g_k to ω^g_{jk} using a Guassian Process (GP) regression model trained on the estimators computed in (1).

3. Estimate the global scaling factor λ such that $w_{jk} = h(\lambda\omega_{jk})$ maximizes the number of sGenes discoveries for some function h.

In step 1, we describe how to compute $\hat{\omega}^g_{jk}$. We propose using the Kolmogorov-Smirnov (KS) distance $\hat{D}^g_{jk}$ between f_0 and f as the estimator for ω^g_{jk}. The primary advantage of this approach is that it does not require any numeric optimization. Following Efron 2001, it can be shown that $\hat{D}^g_{jk}$ is proportional to the upper bound on the maximum likelihood estimate for lFDR and is thus a reasonable choice of estimator. We assume that the density of $f(z)$ (the observed mixture distribution of test statistics) is typically smooth and approximately follows a normal distribution with a heavy tail. This is supported by empirical evidence across multiple studies, including our own [62, 63]. Given this approximation for $f(z)$,

$$\begin{aligned} log(\widehat{lFDR}(\boldsymbol{Z}^g_{jk}, \gamma^g_j, \delta^g_k)) &= log(\pi(\gamma_j, \delta_k)) + \sum_{z^g_{jk} \in \boldsymbol{Z}^g_{jk}} log(f_0(z^g_{jk})) - log(f(z^g_{jk})) \\ &\propto log(\pi(\gamma_j, \delta_k)) + \hat{D}^g_{jk} \end{aligned}$$

In other words, $\hat{D}^g_{jk}$ is proportional to the the log likelihood ratio of the null distribution and observed test statistic mixture distribution. We can consider $\pi(\gamma_j, \delta_k)$ as a nuisance parameter since the ratio $f(z^g_{jk})/f_0(z^g_{jk})$ provides the upper bound on $\pi(\gamma_j, \delta_k)$. Specifically, given the approximate density of $f(z^g_{jk})$,

$$\pi(\gamma_j, \delta_k) \leq argmin_z(\frac{f(z)}{f_0(z)})$$

since $\pi(\gamma_j, \delta_k)$ must satisfy the constraints $0 \leq \pi(\gamma_j, \delta_k) \leq 1$ and $0 \leq \pi(\gamma_j, \delta_k) \leq argmin_z(\frac{f(z)}{f_0(z)}) \leq 1$. Thus it remains that when

$$\widehat{lFDR}(\boldsymbol{Z}^g_{jk}, \gamma^g_j, \delta^g_k) \leq \widehat{lFDR}(\boldsymbol{Z}'^g_{jk}, \gamma'^g_j, \delta'^g_k)$$
$$\hat{D}^g_{jk} \leq \hat{D}'^g_{jk}$$

since $D^g_{jk} \leq D'^g_{jk}$ implies the upper bound of $\pi(\gamma_j, \delta_k) \leq \pi(\gamma'_j, \delta'_k)$. Here, the prime (') notation indicates any other subscript j and k.

To compute, $\hat{D}^g_{jk}$, we use a pooling strategy across genes and nearby covariates. For a local pooling area $\nu = 0.025$, let $\boldsymbol{Z}^g_{jk} = \{z^g_{jk} | g \in [G], \gamma^g_j - \nu \leq \gamma^g_j \leq \gamma^g_j + \nu, \delta^g_j - \nu \leq \delta^g_j \leq \delta^g_j + \nu\}$. The empirical CDF $\hat{F_z}(z)$ and $\hat{D}^g_{jk}$ can then be computed as

$$\hat{F}_z(z) = \frac{1}{|\boldsymbol{Z}^g_{jk}|} \sum_{z^g_{jk} \in \boldsymbol{Z}^g_{jk}} I(z^g_{jk} < z)$$
$$\hat{D}^g_{jk} = sup_z |\Phi(z) - \hat{F}_z(z)|$$

where Φ is the standard normal CDF. This gives us the the estimator under the assumptions that the lFDR conditioned on the covariates is similar across genes and nearby covariates. Then we can obtain bootstrap samples for the estimator by sampling from $\boldsymbol{Z}^g_{jk}$ with replacement and computing $\hat{D}^g_{jk}$ on the bootstrapped samples.

In step 2, we learn a function mapping from covaraties γ^g_{jk} and δ^g_{jk} to ω^g_{jk}. We model this function g with a Gaussian Process (GP) prior. The normal error model is used to approximate the the variance of the estimator across genes.

$$\omega^g_{jk} = g(\gamma^g_j, \delta^g_k) + \epsilon$$
$$g \sim GP(m, k)$$
$$\epsilon \sim N(0, \sigma^2)$$

Here, m is the zero mean function and k is the RBF kernel. To train the model, we generate training datasets $\{(\gamma_i, \hat{D}_{\gamma_i}) | i \in [N]\}$ and $\{(\delta_i, \hat{D}_{\delta_i} | i \in [N]\})$ using the pooling strategy described in (1) for up to N samples. For γ_i and δ_i, we sample covariates uniformly from [0,1]. Then for the given γ_i, generate generate set $\boldsymbol{Z}_{\gamma_i} = \{z^g_{jk} | \gamma_i - 0.025 \leq \gamma^g_j \leq \gamma_i + 0.025\}$ and $\boldsymbol{Z}_{\delta_i} = \{z^g_{jk} | \delta_i - 0.025 \leq \delta^g_k \leq \delta_i + 0.025\}$. We then bootstrap estimators $\hat{D}_{\gamma_i}$ and $\hat{D}_{\delta_i}$ from sets $\boldsymbol{Z}_{\gamma_i}$ and $\boldsymbol{Z}_{\delta_i}$ to form covariate pairs $(\gamma_i, \hat{D}_{\gamma_i})$ and $(\delta_i, \hat{D}_{\delta_i})$. Then we train the model on

these datasets using 10 fold cross validation to avoid over fitting. For inference, we use the mean of the posterior predictive distribution as the weight. Optimization is preformed using the GP regression function in scikit-learn [64].

$$\omega_{jk}^{g} = E(P(\hat{D} *_{\gamma_j^g} |\boldsymbol{Z}, \gamma_j^g)) + E(P(\hat{D} *_{\delta_j^g} |\boldsymbol{Z}, \delta_j^g))$$

In step 3, once we have predictions for ω_{jk}^{g}, we can then perform optimization of λ to maximize sGene discoveries. This is obtained by choosing λ such that for some FDR threshold α

$$argmax_{\lambda} \sum_{g \in [G]} I(p_g(\boldsymbol{\omega}^g, \lambda)) < \alpha$$

Here, $p_g(\boldsymbol{\omega}^g, \lambda)$ is the weighted p-value of gene g computed using $w_{jk}^{g} = h(\lambda \omega_{jk}^{g})$. We choose h to be the exponential function which empirically produced the best results on our datasets.

Estimation of the Null Distribution

Here we show how to estimate the CDF of the null distribution $\hat{F}_{t_g}^{-1}$. We first describe an approach for sampling from this null distribution wtihout using permutations. This approach can decrease the run time by an order of magnitude for population level studies and can be used with or without weighting.

Assume the setting where we have J splice junctions and a single fixed SNP k. We perform J association tests and obtain the test statistic vector $\boldsymbol{z}_k = (z_{1k}, z_{2k}, \ldots, z_{J_g k})$. Under the null, $\boldsymbol{z}_k \sim MVN(\boldsymbol{0}, \Sigma_{z_k})$ and the marginals $z_{jk} \sim N(0, 1)$. Similarly, considerthe setting where we have K SNPs and a fixed splice junction j. In this setting $\boldsymbol{z}_j = (z_{j1}, z_{j2}, \ldots, z_{jK})$ and $\boldsymbol{z}_j \sim MVN(\boldsymbol{0}, \Sigma_{z_j})$. Taken together, we can sample the null distribution of these test statistics from a matrix normal distribution $MVN(\boldsymbol{0}_{J \times K}, \Sigma_{z_k}, \Sigma_{z_j})$. A sensible approach to estimation is to use the standardized sample covariance estimator (i.e. correlation) of the junction quantifications $\hat{\Sigma}_y = \frac{Cov(y_j, y_{j'})}{\sigma_{y_j} \sigma_{y_{j'}}}$ and genotypes $\hat{\Sigma}_x = \frac{Cov(x_k, x_{k'})}{\sigma_{x_j} \sigma_{x_{j'}}}$ to approximate Σ_{z_j} and Σ_{z_k} respectively. It can be proven than $\Sigma_y = \Sigma_{z_k}$ or $\Sigma_x = \Sigma_{z_j}$ for OLS (Supplementary Note). With the parameters of the matrix normal estimated, we can then sample from this distribution N times and calculate $\hat{F}_{t_g}^{-1}$ as follows

$$\{z^g_{jk}|j \in [J], k \in [K]\} \sim MVN(\mathbf{0}_{J\times K}, \Sigma_{z_k}, \Sigma_{z_j})$$
$$t_i = \max_{\forall j,k}(|z^g_{jk}|)$$
$$\hat{F}^{-1}_{t_g}(t_g) = \frac{1}{N}\sum I(t_g > t_i)$$

Briefly, sampling from a multivariate normal requires an affine transform of random independent standard normals Z. Specifically:

$$MVN(0, \Sigma) = AZ \tag{0.1}$$
$$Z = A^T A \tag{0.2}$$

The matrix A is a symmetric matrix found using Cholesky decomposition. However, when Choloesky fails due to a matrix not being full rank, we resort to SVD to obtain decomposition USV such that $A = U\sqrt{S}$.

To improve the accuracy of the covariance estimators, we use two adjustments. First, since it is often the case that $K >> N$, we use the Ledoit Wolf shrinkage estimator [65] instead of sample covariance estimator $\hat{\Sigma}_k$ which is more accurate in this scenario. Specifically, it can be shown that the Frobenius norm between this LW estimator and population covariance matrix is minimized.

$$\hat{\Sigma_{LW}} = \pi\hat{\Sigma} + (1-\pi)\frac{Tr(\hat{\Sigma})}{K}I \tag{0.3}$$

Effect Size Inference Model for sQTL

sQTL Model

For a splice junction $j \in [1 \dots J]$ in sample $i \in [1 \dots M]$, let y_i be the number of reads mapped to that splice junction and n_i be the total number of reads mapped to the splicing event (LSV, intron cluster, etc). For each sample i, we also observe genotype $x_{i1} \in [0, 1, 2]$ which is encoded as the minor allele count. We also have $K-2$ covariates $x_{i2\dots i(K-1)}$ and the model intercept $x_{iK} = 1$. The covariates often include known confounding factors that affect the genotype (e.g. population structure) and splicing data (e.g. RNA-seq batch) and unknown confounding factors which are inferred using an external model (e.g. MOCASSAIN[66], PEERS[67]). The association between the genotype and splice junction in each sample is modeled as a beta binomial regression. Note that our model is not a GLM (and thus does

not benefit from generic solutions for GLMs) because the beta binomial distribution is not in the exponential family. The model likelihood is defined as

$$P(y_i|n_i,\mu_i,\phi) = \frac{\Gamma(n_i+1)}{\Gamma(y_i+1)\Gamma(n_i-y_i+1)} \frac{\Gamma(y_i+\mu_i\phi)\Gamma(n_i-y_i+(1-\mu_i)\phi)}{\Gamma(n_i+\phi)} \frac{\Gamma(\phi)}{\Gamma(\mu_i\phi)\Gamma((1-\mu_i)\phi)} \tag{0.4}$$

Here, μ_i is the mean of the distribution and ϕ is the dispersion. The gamma function is denoted as $\Gamma(.)$. We choose a logit link function for our model. Thus μ_i is a linear function of x_i transformed by the inverse logit or logistic function such that

$$\mu_i = \frac{1}{1+e^{-\hat{y}_i}} \tag{0.5}$$

and

$$\hat{y}_i = \sum_{k=1}^{K} x_{ik}\beta_k \tag{0.6}$$

where β_k is the k_{th} regression coefficient.

To compute the p-value for β_1 (recall that the index 1 refers to the genotype index), we use Wald's method. The Wald test statistic and asymptotic distribution is given by

$$t_{wald} = \frac{\hat{\beta}_1}{se(\hat{\beta})_1}$$
$$t_{wald} \sim N(0,1)$$

We compute the standard error as the following where H is the Hessian matrix.

$$se(\hat{\beta}_1) = \sqrt{diag(H(\ell)^{-1})_1}$$

Note that our computation of the Hessian when evaluated at the MLE parameters produces standard errors that are more precise than the approximate Hessian calculated by black box solvers (Supplementary Note).

Composite Testing for sQTL Effect Size

For composite testing, we follow the approach of Love et al. 2016. The null and alternative hypothesis are given as follows.

$$H_0 : |\beta| \leq \theta$$
$$H_1 : |\beta| > \theta$$

Let θ be the desired effect size threshold. Then the composite p-value is given by

$$p_{\hat{\beta}}(\theta) = max(1, (1 - F_\theta(|\hat{\beta}|)) * 2)$$

where F_θ is the CDF of $N(\theta, se(\hat{\beta}))$. When $\theta = 0$, this reduces to the null hypothesis $H_0 : \beta = 0$.

Data and code availability

Code and data will be released upon publication.

Acknowledgements:

We would like to acknowledge Dr. Caleb Radens, Dr. Iain Mathesion and Dr. Bogdan Pasaniuc for helpful discussion and advice. CDB passed away during the work on this project. We miss him dearly and believe he would have been proud of the final product. We dedicate this work to his memory and contributions to science.

Funding:

DW was supported by NIH fellowship 1F31CA265218-01. DW, SJ, BWM were supported by NIH R01GM128096 to YB and CDB, NIH R01 LM013437 and CureBRCA grant to YB.

Author Contributions: DW and YB conceived the project. DW and YB developed the methods and planned the experiments with input from MG and CB. DW wrote the code and carried out the experiments and analysis. SJ handled all aspects of the software related to MAJIQ integration, including developing the visualization tools. DW and YB wrote the manuscript. All authors read and approved the final manuscript.

Competing Interests: The authors declare that they have no competing interests.

References

1. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337.** ISBN: 0036-8075 Publisher: American Association for the Advancement of Science, 1190–1195 (2012).

2. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461.** ISBN: 0028-0836 Publisher: Nature Publishing Group UK London, 747–753 (2009).

3. Kim-Hellmuth, S. *et al.* Cell type–specific genetic regulation of gene expression across human tissues. *Science* **369.** ISBN: 0036-8075 Publisher: American Association for the Advancement of Science, eaaz8528 (2020).

4. Li, Y. I., Van De Geijn, B., Raj, A., Knowles, D. A., Petti, A. A., Golan, D., Gilad, Y. & Pritchard, J. K. RNA splicing is a primary link between genetic variation and disease. *Science* **352.** ISBN: 0036-8075 Publisher: American Association for the Advancement of Science, 600–604 (2016).

5. Vaquero-Garcia, J., Barrera, A., Gazzara, M. R., Gonzalez-Vallinas, J., Lahens, N. F., Hogenesch, J. B., Lynch, K. W. & Barash, Y. A new view of transcriptome complexity and regulation through the lens of local splicing variations. *Elife* **5.** ISBN: 2050-084X Publisher: eLife Sciences Publications Limited, e11752 (2016).

6. Vaquero-Garcia, J. *et al.* RNA splicing analysis using heterogeneous and large RNA-seq datasets. *Nature communications* **14.** ISBN: 2041-1723 Publisher: Nature Publishing Group UK London, 1230 (2023).

7. Li, Y. I., Knowles, D. A., Humphrey, J., Barbeira, A. N., Dickinson, S. P., Im, H. K. & Pritchard, J. K. Annotation-free quantification of RNA splicing using LeafCutter. *Nature genetics* **50.** ISBN: 1061-4036 Publisher: Nature Publishing Group US New York, 151–158 (2018).

8. Shen, S., Park, J. W., Lu, Z.-x., Lin, L., Henry, M. D., Wu, Y. N., Zhou, Q. & Xing, Y. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proceedings of the national academy of sciences* **111.** ISBN: 0027-8424 Publisher: National Acad Sciences, E5593–E5601 (2014).

9. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods* **14.** ISBN: 1548-7105 Publisher: Nature Publishing Group, 417–419 (2017).

10. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nature biotechnology* **34.** ISBN: 1546-1696 Publisher: Nature Publishing Group, 525–527 (2016).

11. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *Bmc bioinformatics* **12.** ISBN: 1471-2105 Publisher: Springer, 1–16 (2011).

12. Teng, M. *et al.* A benchmark for RNA-seq quantification pipelines. *Genome biology* **17.** Publisher: Springer, 1–12 (2016).

13. Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E. & Gouil, Q. Opportunities and challenges in long-read sequencing data analysis. *Genome biology* **21.** ISBN: 1474-760X Publisher: Springer, 30 (2020).

14. Wang, Z. & Burge, C. B. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *Rna* **14.** ISBN: 1355-8382 Publisher: Cold Spring Harbor Lab, 802–813 (2008).

15. Zhao, K., Lu, Z.-x., Park, J. W., Zhou, Q. & Xing, Y. GLiMMPS: robust statistical model for regulatory variation of alternative splicing using RNA-seq data. *Genome biology* **14.** Publisher: Springer, 1–15 (2013).

16. Nowicka, M. & Robinson, M. D. DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics. *F1000research* **5.** Publisher: Faculty of 1000 Ltd (2016).

17. 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. *Nature communications* **12.** ISBN: 2041-1723 Publisher: Nature Publishing Group UK London, 727 (2021).

18. Qi, T., Wu, Y., Fang, H., Zhang, F., Liu, S., Zeng, J. & Yang, J. Genetic control of RNA splicing and its distinct role in complex trait variation. *Nature genetics* **54.** ISBN: 1061-4036 Publisher: Nature Publishing Group US New York, 1355–1363 (2022).

19. Ignatiadis, N., Klaus, B., Zaugg, J. B. & Huber, W. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature methods* **13.** ISBN: 1548-7091 Publisher: Nature Publishing Group US New York, 577–580 (2016).

20. Huang, C., Clayton, E. A., Matyunina, L. V., McDonald, L. D., Benigno, B. B., Vannberg, F. & McDonald, J. F. Machine learning predicts individual cancer patient responses to therapeutic drugs with high accuracy. *Scientific reports* **8.** ISBN: 2045-2322 Publisher: Nature Publishing Group, 1–8 (2018).

21. Consortium, G. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369.** ISBN: 0036-8075 Publisher: American Association for the Advancement of Science, 1318–1330 (2020).

22. Hayer, K. E., Pizarro, A., Lahens, N. F., Hogenesch, J. B. & Grant, G. R. Benchmark analysis of algorithms for determining and quantifying full-length mRNA splice forms from RNA-seq data. *Bioinformatics* **31.** ISBN: 1367-4811 Publisher: Oxford University Press, 3938–3945 (2015).

23. Liu, Y., Chen, S., Li, Z., Morrison, A. C., Boerwinkle, E. & Lin, X. ACAT: a fast and powerful p value combination method for rare-variant analysis in sequencing studies.

The american journal of human genetics **104.** ISBN: 0002-9297 Publisher: Elsevier, 410–421 (2019).

24. Delaneau, O., Ongen, H., Brown, A. A., Fort, A., Panousis, N. I. & Dermitzakis, E. T. A complete tool set for molecular QTL discovery and analysis. *Nature communications* **8.** ISBN: 2041-1723 Publisher: Nature Publishing Group UK London, 15452 (2017).

25. McCaw, Z. R., Lane, J. M., Saxena, R., Redline, S. & Lin, X. Operating characteristics of the rank-based inverse normal transformation for quantitative trait analysis in genome-wide association studies. *Biometrics* **76.** ISBN: 0006-341X Publisher: Oxford University Press, 1262–1272 (2020).

26. Dunn, P. K. & Smyth, G. K. Randomized quantile residuals. *Journal of computational and graphical statistics* **5.** ISBN: 1061-8600 Publisher: Taylor & Francis, 236–244 (1996).

27. Li, Y., Ge, X., Peng, F., Li, W. & Li, J. J. Exaggerated false positives by popular differential expression methods when analyzing human population samples. *Genome biology* **23.** ISBN: 1474-760X Publisher: Springer, 79 (2022).

28. Zhang, Y., Yang, H. T., Kadash-Edmondson, K., Pan, Y., Pan, Z., Davidson, B. L. & Xing, Y. Regional variation of splicing QTLs in human brain. *The american journal of human genetics* **107.** ISBN: 0002-9297 Publisher: Elsevier, 196–210 (2020).

29. Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464.** ISBN: 0028-0836 Publisher: Nature Publishing Group UK London, 768–772 (2010).

30. Huang, Q. Q., Ritchie, S. C., Brozynska, M. & Inouye, M. Power, false discovery rate and Winner's Curse in eQTL studies. *Nucleic acids research* **46.** ISBN: 0305-1048 Publisher: Oxford University Press, e133–e133 (2018).

31. Nygaard, V., Rødland, E. A. & Hovig, E. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics* **17.** ISBN: 1468-4357 Publisher: Oxford University Press, 29–39 (2016).

32. 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 research* **27.** ISBN: 1088-9051 Publisher: Cold Spring Harbor Lab, 1872–1884 (2017).

33. Baeza-Centurion, P., Miñana, B., Schmiedel, J. M., Valcárcel, J. & Lehner, B. Combinatorial genetics reveals a scaling law for the effects of mutations on splicing. *Cell* **176.** ISBN: 0092-8674 Publisher: Elsevier, 549–563. e23 (2019).

34. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15.** Publisher: Springer, 1–21 (2014).

35. Sul, J. H., Raj, T., De Jong, S., De Bakker, P. I., Raychaudhuri, S., Ophoff, R. A., Stranger, B. E., Eskin, E. & Han, B. Accurate and fast multiple-testing correction in

eQTL studies. *The american journal of human genetics* **96.** ISBN: 0002-9297 Publisher: Elsevier, 857–868 (2015).

36. Urbut, S. M., Wang, G., Carbonetto, P. & Stephens, M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nature genetics* **51.** ISBN: 1061-4036 Publisher: Nature Publishing Group US New York, 187–195 (2019).

37. Jaganathan, K. *et al.* Predicting splicing from primary sequence with deep learning. *Cell* **176.** ISBN: 0092-8674 Publisher: Elsevier, 535–548. e24 (2019).

38. Grgic, O. *et al.* CYP11B1 variants influence skeletal maturation via alternative splicing. *Communications biology* **4.** ISBN: 2399-3642 Publisher: Nature Publishing Group UK London, 1274 (2021).

39. Yamaguchi, K. *et al.* Splicing QTL analysis focusing on coding sequences reveals mechanisms for disease susceptibility loci. *Nature communications* **13.** ISBN: 2041-1723 Publisher: Nature Publishing Group UK London, 4659 (2022).

40. Li, L. *et al.* An atlas of alternative polyadenylation quantitative trait loci contributing to complex trait and disease heritability. *Nature genetics* **53.** ISBN: 1061-4036 Publisher: Nature Publishing Group US New York, 994–1005 (2021).

41. Umans, B. D., Battle, A. & Gilad, Y. Where are the disease-associated eQTLs? *Trends in genetics* **37.** ISBN: 0168-9525 Publisher: Elsevier, 109–124 (2021).

42. Mostafavi, H., Spence, J. P., Naqvi, S. & Pritchard, J. K. Systematic differences in discovery of genetic effects on gene expression and complex traits. *Nature genetics* **55.** ISBN: 1061-4036 Publisher: Nature Publishing Group US New York, 1866–1875 (2023).

43. LaPierre, N. & Pimentel, H. Accounting for isoform expression increases power to identify genetic regulation of gene expression. *Plos computational biology* **20.** ISBN: 1553-734X Publisher: Public Library of Science San Francisco, CA USA, e1011857 (2024).

44. Scott, J. G., Kelly, R. C., Smith, M. A., Zhou, P. & Kass, R. E. False discovery rate regression: an application to neural synchrony detection in primary visual cortex. *Journal of the american statistical association* **110.** ISBN: 0162-1459 Publisher: Taylor & Francis, 459–471 (2015).

45. Zhang, M. J., Xia, F. & Zou, J. Fast and covariate-adaptive method amplifies detection power in large-scale multiple hypothesis testing. *Nature communications* **10.** ISBN: 2041-1723 Publisher: Nature Publishing Group UK London, 3433 (2019).

46. Aguet, F., Alasoo, K., Li, Y. I., Battle, A., Im, H. K., Montgomery, S. B. & Lappalainen, T. Molecular quantitative trait loci. *Nature reviews methods primers* **3.** ISBN: 2662-8449 Publisher: Nature Publishing Group UK London, 4 (2023).

47. Pickrell, J. K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *The american journal of human genetics* **94.** ISBN: 0002-9297 Publisher: Elsevier, 559–573 (2014).

48. Wagner, N., Çelik, M. H., Hölzlwimmer, F. R., Mertes, C., Prokisch, H., Yépez, V. A. & Gagneur, J. Aberrant splicing prediction across human tissues. *Nature genetics* **55.** ISBN: 1061-4036 Publisher: Nature Publishing Group US New York, 861–870 (2023).

49. Hormozdiari, F. *et al.* Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits. *Nature genetics* **50.** ISBN: 1061-4036 Publisher: Nature Publishing Group US New York, 1041–1047 (2018).

50. Bhattacharya, A., Vo, D. D., Jops, C., Kim, M., Wen, C., Hervoso, J. L., Pasaniuc, B. & Gandal, M. J. Isoform-level transcriptome-wide association uncovers extensive novel genetic risk mechanisms for neuropsychiatric disorders in the human brain. *Medrxiv.* Publisher: Cold Spring Harbor Laboratory Press, 2022.08. 23.22279134 (2022).

51. Kelley, D. R., Reshef, Y. A., Bileschi, M., Belanger, D., McLean, C. Y. & Snoek, J. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome research* **28.** ISBN: 1088-9051 Publisher: Cold Spring Harbor Lab, 739–750 (2018).

52. Cheng, J., Çelik, M. H., Kundaje, A. & Gagneur, J. MTSplice predicts effects of genetic variants on tissue-specific splicing. *Genome biology* **22.** Publisher: Springer, 1–19 (2021).

53. Barash, Y., Calarco, J. A., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B. J. & Frey, B. J. Deciphering the splicing code. *Nature* **465.** ISBN: 0028-0836 Publisher: Nature Publishing Group UK London, 53–59 (2010).

54. Zeng, T. & Li, Y. I. Predicting RNA splicing from DNA sequence using Pangolin. *Genome biology* **23.** ISBN: 1474-760X Publisher: Springer, 103 (2022).

55. Jha, A., Gazzara, M. R. & Barash, Y. Integrative deep models for alternative splicing. *Bioinformatics* **33.** ISBN: 1367-4803 Publisher: Oxford University Press, i274–i282 (2017).

56. You, S.-F. *et al.* MS4A4A modifies the risk of Alzheimer disease by regulating lipid metabolism and immune response in a unique microglia state. *Medrxiv.* Publisher: Cold Spring Harbor Laboratory Preprints (2023).

57. Lord, J. *et al.* Predicting the impact of rare variants on RNA splicing in CAGI6. *Human genetics.* ISBN: 0340-6717 Publisher: Springer, 1–9 (2024).

58. Yao, D. *et al.* Scalable genetic screening for regulatory circuits using compressed Perturb-seq. *Nature biotechnology.* ISBN: 1087-0156 Publisher: Nature Publishing Group US New York, 1–14 (2023).

59. Westfall, P. H. & Young, S. S. *Resampling-based multiple testing: Examples and methods for p-value adjustment* (John Wiley & Sons, 1993).

60. Bourgon, R., Gentleman, R. & Huber, W. Independent filtering increases detection power for high-throughput experiments. *Proceedings of the national academy of sciences* **107.** ISBN: 0027-8424 Publisher: National Acad Sciences, 9546–9551 (2010).

61. Efron, B. Microarrays, empirical Bayes and the two-groups model (2008).

62. Efron, B., Tibshirani, R., Storey, J. D. & Tusher, V. Empirical Bayes analysis of a microarray experiment. *Journal of the american statistical association* **96.** ISBN: 0162-1459 Publisher: Taylor & Francis, 1151–1160 (2001).

63. Van Zwet, E., Gelman, A., Greenland, S., Imbens, G., Schwab, S. & Goodman, S. N. A new look at p values for randomized clinical trials. *Nejm evidence* **3.** ISBN: 2766-5526 Publisher: Massachusetts Medical Society, EVIDoa2300003 (2023).

64. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *The journal of machine learning research* **12.** ISBN: 1532-4435 Publisher: JMLR. org, 2825–2830 (2011).

65. Ledoit, O. & Wolf, M. Honey, I shrunk the sample covariance matrix. *Upf economics and business working paper* (2003).

66. Slaff, B., Radens, C. M., Jewell, P., Jha, A., Lahens, N. F., Grant, G. R., Thomas-Tikhonenko, A., Lynch, K. W. & Barash, Y. MOCCASIN: A method for correcting for known and unknown confounders in RNA splicing analysis. *Nature communications* **12.** ISBN: 2041-1723 Publisher: Nature Publishing Group, 1–9 (2021).

67. Stegle, O., Parts, L., Durbin, R. & Winn, J. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *Plos computational biology* **6.** ISBN: 1553-734X Publisher: Public Library of Science San Francisco, USA, e1000770 (2010).

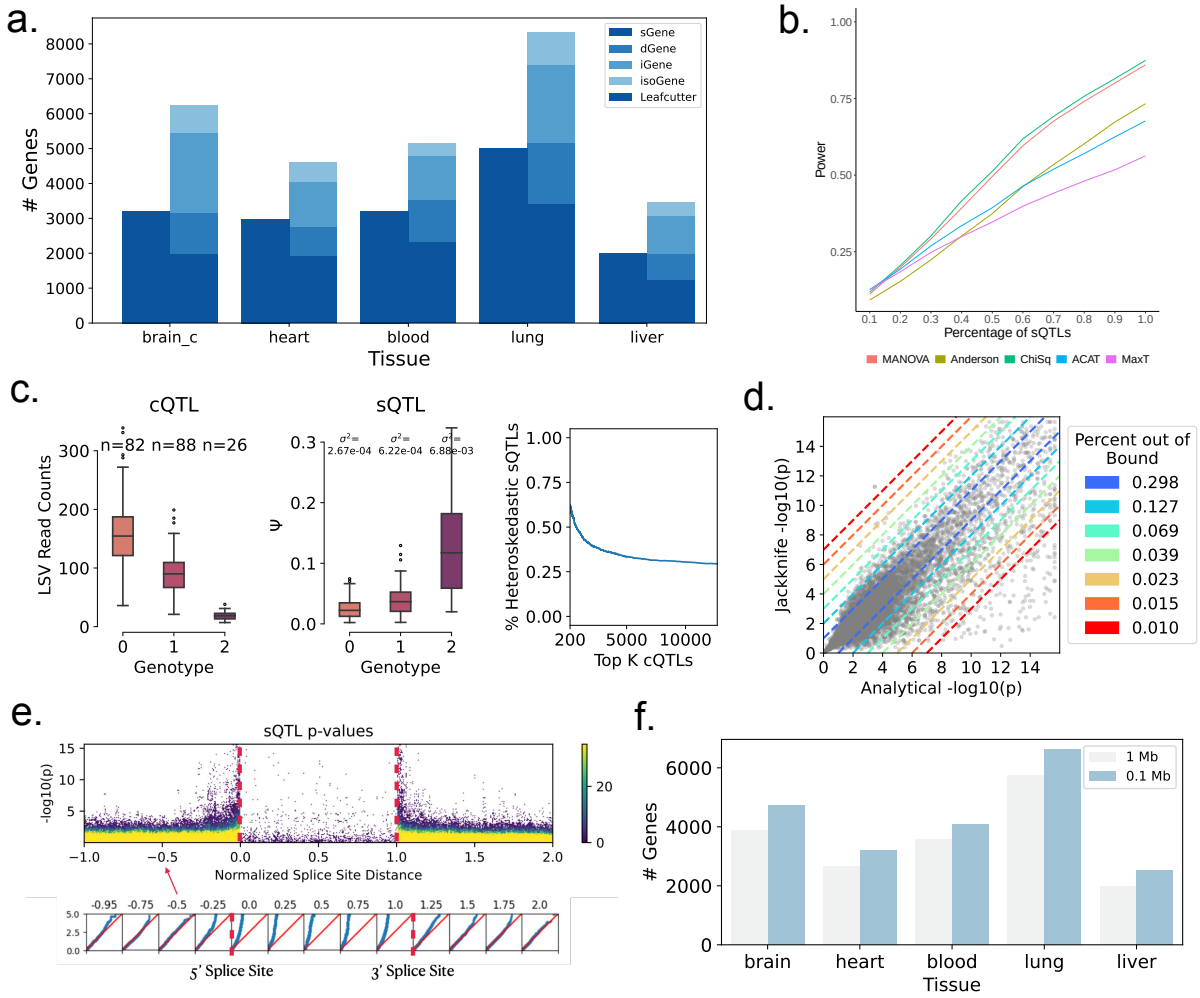

Figure 1: GTEx Case Study.

(A) Barplot showing the number of sGenes found by MAJIQ and Leafcutter across 5 representative tissues. The unique sGenes are stratified by whether they are called due to annotated events (sGenes), denovo events (dGenes), intron retention (iGenes) or isoforms (isoGenes). Using Leafcutter alone fails to find iGenes and isoGenes. **(B)** A comparison of the statistical methods for sGene discovery. The x-axis represents the percentage of tests in a locus that are sQTLs. For example, 0.2 indicates that there are 2 junctions out 10 where the alternative hypothesis is true. The y-axis shows the power. **(C)** An example of heteroskedasticity introduced by coverage. The left plot shows that this event is a cQTL in which the genotype is correlated with coverage. The middle plot shows that the event is an sQTL in which the genotype is correlated with Ψ. However, the variance is highest when the coverage is low and the variance is lowest when the coverage is high. The right plot shows that the percent of top K cQTLs that are also heteroskedastic sQTLs (based on Bruesch-Page test) increases as K decreases. **(D)** A comparison of jackknife p-values (y-

axis) and OLS analytical p-values (x-axis) when applying linear regression to splicing data. Due to model misspecification (i.e. heteroskedasticity), the values do not agree well. The colors indicate bounds which represent minimum fold differences. For example, the bound at 1 (blue) represents at least a 10 fold difference in -log10 p-value between the jackknife and analytical values for all points to the left and right of the bound. The numbers in the legend incidate the percentage of points that fall beyond the bound. **(E)** P-value inflation as a function of the SNP's distance to the 5' and 3' splice sites. The normalized distance is shown for a 1Mb window. The QQ plots below show that the distribution of the observed p-values is far from uniform for sQTLs closer to slice sites. The colors represents the density of data points. **(F)** The number of sGenes discovered in each tissue depending on whether a 1 Mb (blue) or 0.1 Mb (grey) window was used for the std-SQTL pipeline.

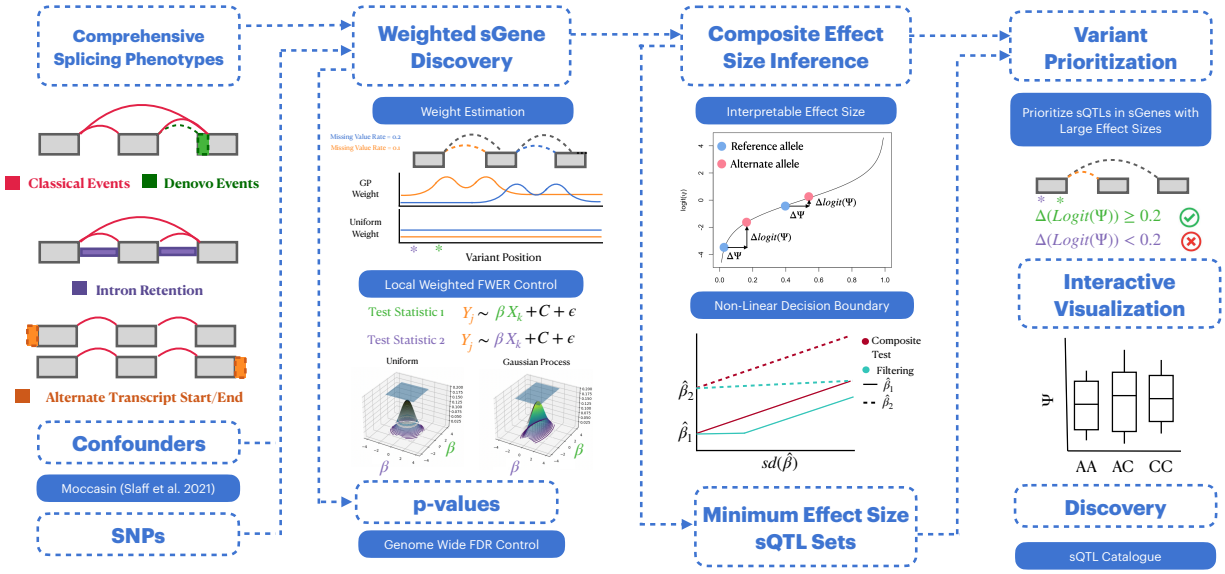

Figure 2: MAJIQTL Pipeline.

An overview of the MAJIQTL pipeline. First (column 1), we quantify splicing using MAJIQ and combine these quantifications with isoform ratios, creating a comprehensive set of splicing events which includes classical, denovo, intron retention and alternate transcript start/ends events. Confounding effects are corrected using MOCCASIN. Second (column 2), we use our weighted multiple testing method to discover sGenes. This approach uses a GP regression model to learn a mapping to weights from covariates based on the missingness rate of the junction (orange/blue) and the proximity of the variant to the splice site (purple/green). The baseline uniform weight model is shown for comparison. Then we use these weights to compute a gene level p-value that controls the FWER of sQTL discoveries in each gene. Intuitively, FWER control is achieved by assigning a threshold to each sQTL's test statistic (represented by the edges of the blue square) subject to a constant budget constraint (volume of the density plot directly under the square). Under a uniform model, the thresholds are all equal (left). However, our weighting model assigns thresholds proportional the weights (right). Third (column 3), we use a composite Beta-Binomial regression model to estimate effect sizes of sQTLs identified in the previous step. The effect size estimate has a fold change interpretation which is related to the $\Delta\Psi$ measure of splicing change (x-axis) through the logit function (y-axis). Then we then use a composite testing approach to create sets of sQTLs defined by a minimum effect size. An advantage of this approach (red) is the non-linear decision boundary for assigning sQTLs to the set which accounts for the variance of estimates unlike filtering by observed effect sizes (blue). Finally (column 4), we can use these sets to prioritize sQTLs with large effect sizes. Our results can be visualized using the VOILA-QTL package in MAJIQ and we report a catalogue of our sQTLs in the MAJIQlopedia database.

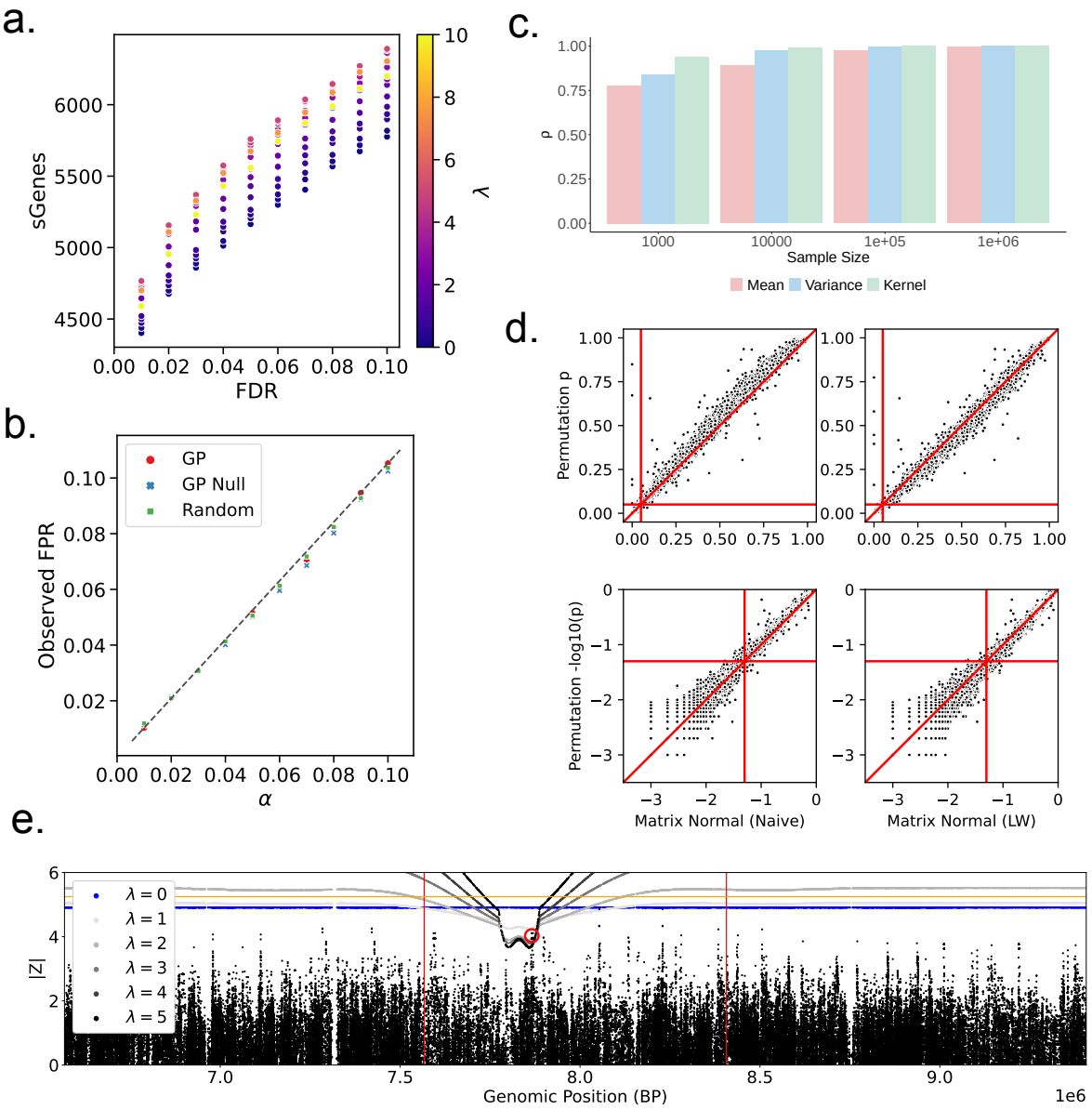

Figure 3: Weighted Multiple Testing Evaluation (A) The number of sGenes (y-axis) discovered by MAJIQTL's weighted multiple testing method in brain - cerebellum at varying FDR levels (x-axis). For a given FDR level, the color of each point indicates the value of the model parameter λ which was optimized to maximize sGene discoveries ($\lambda = 5$). When $\lambda = 0$, the behavior of the method is equivalent to the unweighted max T sGene discovery method used by GTEx (baseline). **(B)** The observed false positive rate (y-axis) of the method at varying p-value cutoffs α (x-axis). The method was applied to synthetic null data generated using permutations and the observed FPR matches the expected FPR (x=y line).

The colors represent 3 different approaches used to select the weights: the GP trained on the original data (red), the GP trained on the null data (blue), and random weights drawn from a uniform distribution between 0 and 1 (green). **(C)** The Spearman correlation ρ between $P(H_1|C)$ and the pooled KS statistic estimator $\hat{D}$ at varying sample sizes (the number of sQTL summary statistics used to estimate the pooled KS statistic). The colors represent 3 different models for the alternate distribution. Under the mean (red) and variance (blue) models, the mean and variance of the alternate distribution vary with $P(H_1|C)$ respectively. The kernel model (green) uses the empirical kernel density of real sQTL summary statistics to approximate an alternate distribution. **(D)** The correlation between the unweighted max T p-values where the null distribution is computed using permutations (y-axis) or our matrix normal (MN) sampling approach (x-axis). The left panels show results using the naive MN sampling approach (sample covariance estimator) while the right panels show results using the Ledoit-Wolf shrinkage estimator. The top panels show the raw p-values while the bottom panels show the -log10 p-values. **(E)** An example of the weighted FWER bound computed by the method for the locus around a gene with position shown by vertical red lines. For clarity, only the bound for a single junction is shown. The FWER of all tests in the locus is controlled at the BH critical value (0.022) which controls gene level FDR at 0.05. The greyscale colors indicate the value of λ for which each bound was computed. Recall that when $\lambda = 0$, the method is equivalent to the unweighted max T method and produces a uniform bound (blue). For comparison, the Bonferroni bound is shown in orange. This gene is considered an sGene using the weighted method since it contains at least 1 sQTL with a test statistic (y-axis) above the bound (red circle).

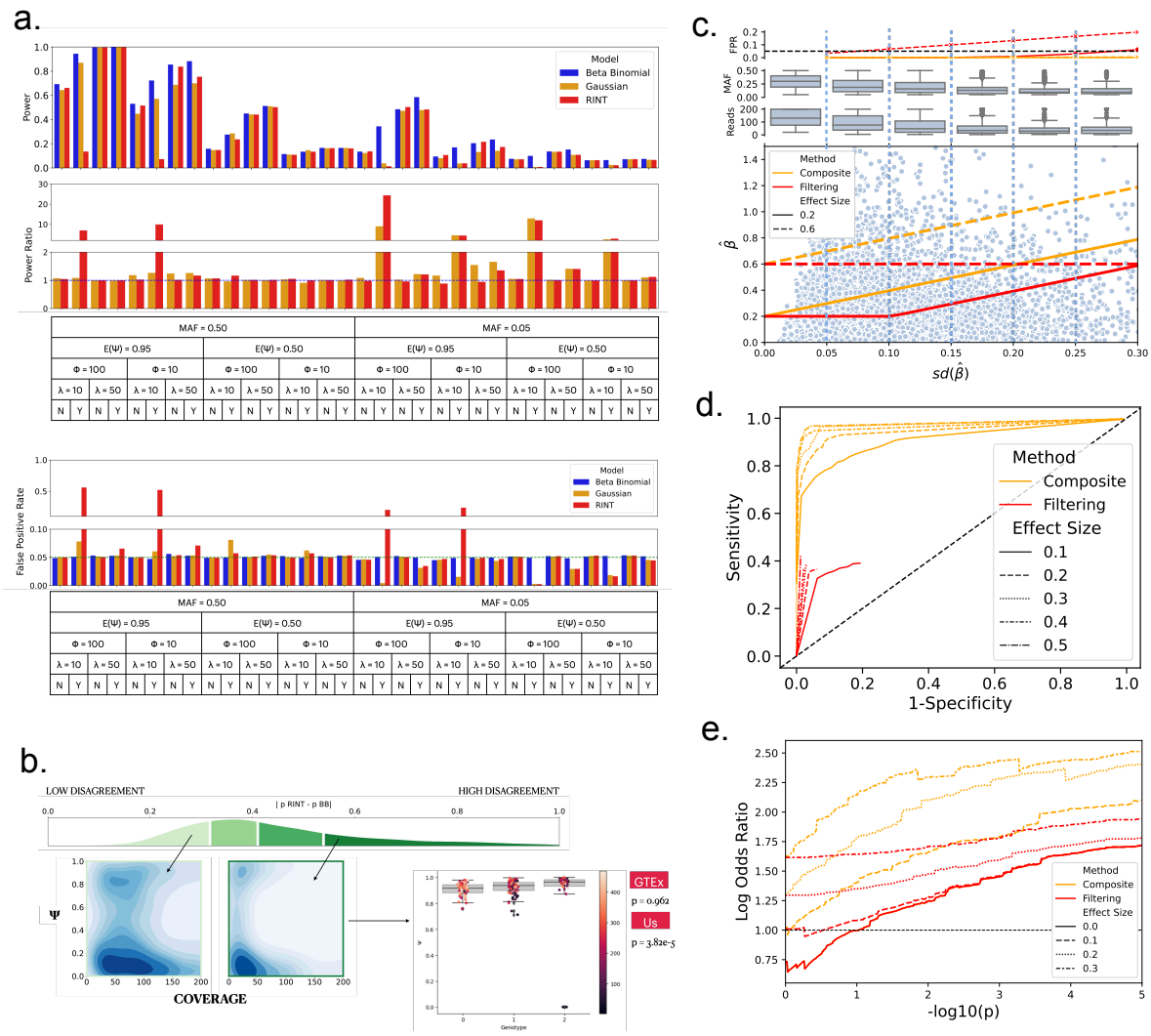

Figure 4: Evaluation of Composite Effect Size Inference

(A) Simulations comparing the power (top) and false positive rate (bottom) of the Beta-Binomial model, OLS regression and OLS regression with RINT transform across 32 parametric combinations for detecting sQTLs with non-zero effect size. The combinations are given by the chart on the x-axis. MAF is the minor allele frequency. $E(\Psi)$ is the mean Ψ for the major allele. ϕ is the overdispersion. λ is the mean coverage in reads drawn from a Poisson distribution. Y/N refers to whether the genotype is correlated with coverage (i.e. is a cQTL). All simulations use an effect size of $\Delta\Psi = 0.01$. The cQTL coverage effect size is 20 reads. **(B)** A comparison of effect size estimates between Beta-Binomial and RINT OLS in brain - cerebellum. Effect sizes were divided into 4 bins based on the quantile of the differences between their p-values (green histogram). The models tend to disagree when

coverage is low (< 50) and Ψ values are skewed to 0 or 1 (upper quantile). The models have much higher agreement when coverage is high (> 50) and Ψ values are near intermediate values (bottom quantile). The example boxplot shows a sQTL which has high disagreement between the models. However, the Beta-Binomial model is also robust to low coverage outliers and makes a more accurate effect size inference since GTEx is unable to call this sQTL due to the outliers. **(C)** A comparison of the decision boundaries of two approaches for constructing sQTL sets defined by a minimum effect size threshold: composite testing (orange) and filtering (red). For clarity, we only show decision boundaries corresponding to 0.05 FPR control for two thresholds: 0.2 (solid line) and 0.5 (dashed line). The composite approach has a non-linear boundary that increases with the standard error (x-axis) of the effect size estimator (y-axis). In contrast, the decision boundary of the filtering approach does not always account for the standard error. sQTLs with a Beta Binomial effect size estimator (blue points) greater than the decision boundary are considered to be in the minimum effect size set. The distribution of the MAF and read counts for sQTLs in each bin (blue dashed lines) on the x-axis (middle and bottom bar plots) indicates that these factors decrease when the variance increases. Composite testing also controls the FPR at the desire 0.05 level (albeit conservatively) while the filtering approach does not (top bar plot). **(D)** The ROC curve for the minimum effect size sQTL sets generated by the composite (orange) and filtering (red) approaches. Five effect size thresholds are shown from 0.1 and 0.5 and are represented by the various lines. The decision threshold for each approach is the sQTL p-value. For filtering, it eventually becomes impossible to separate the positive and negative classes by p-value since the effect size filter dominates, hence the red lines appear truncated. **(E)** Enrichment of minimum effect size sQTL sets for variants that disrupt splice sites (spliceAI Delta score > 0.2). The minimum effect size threshold for each set is indicated by the style of the line and the method used to generate the set is indicated by the color. The x-axis is the p-value (shown in -log10 space) rejection threshold for the method which can be interpreted as the FPR. The y-axis is the enrichment measured by the log-odds ratio. The composite and filtering approaches are equivalent at a threshold of 0 thus these two lines overlap.

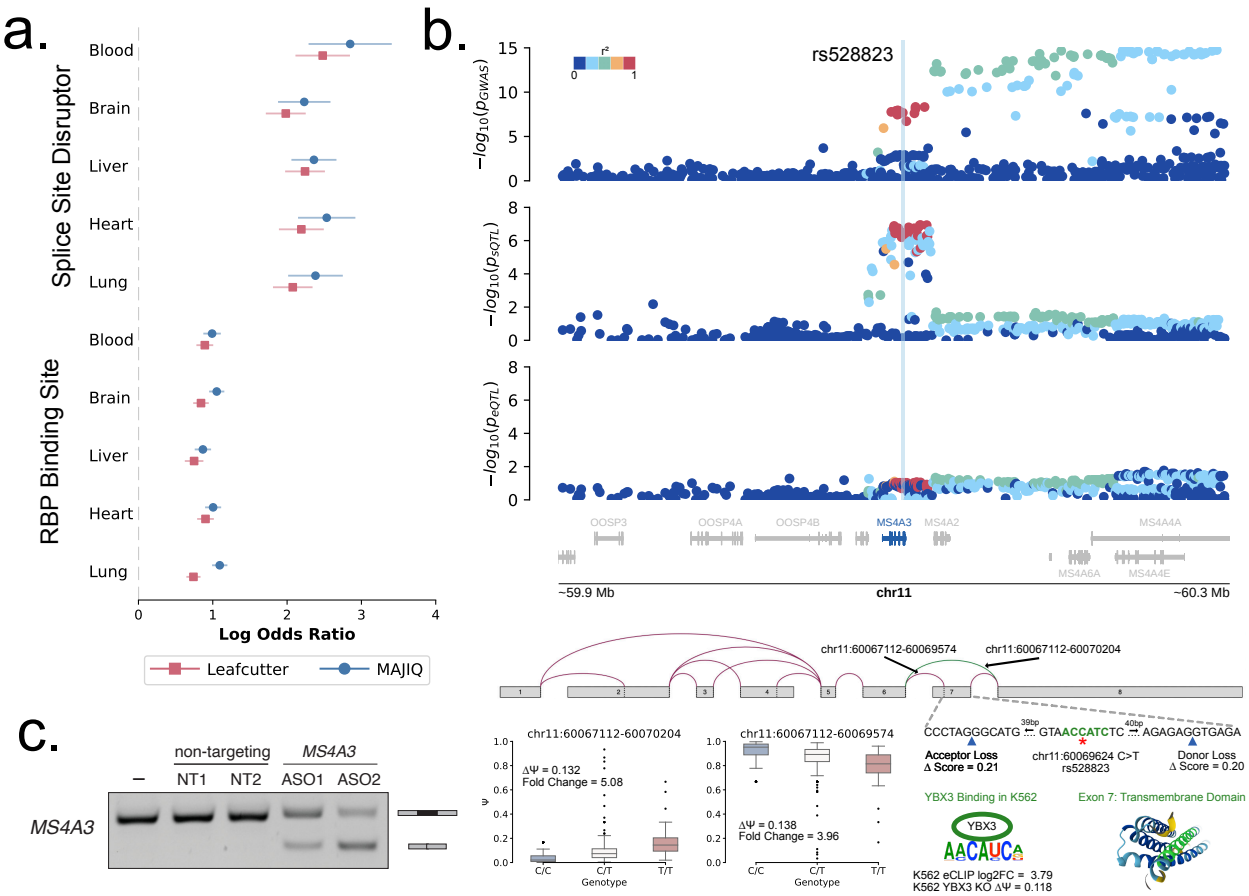

Figure 5: Functional Analysis of sQTLs

(A) Enrichment of sQTLs proximal to splice sites within sGenes discovered by MAJIQTL (blue) for two functional annotations: splice site disruption and RBP binding. Leafcutter (red) sQTLs are found using the std-SQTL pipeline and the Bonferroni threshold in each sGene. A variant is considered splice site disrupting if it has a SpliceAI Delta score > 0.2. A variant is considered a RBP binding site if binding of at least one of one RBP is observed in the ENCODE eCLIP data (K562). The x-axis shows the log-odds ratio for enrichment and the y-axis shows the tissues. The confidence interval is the 95% interval for the log odds ratio computed using Fisher's exact method. **(B)** Analysis of the variant rs528823 (vertical blue line) and its role in disease. The Manhattan plots show the Alzheimer's GWAS (top), MS4A3 exon 7 junction sQTL (middle) and MS4A3 expression eQTL (bottom) p-values for all variants in a locus around the gene (highlighted in blue) on chromosome 11. The QTL tissue is blood. The points are colored by their LD with the variant of interest. The splice graph shows all junctions (green denovo, red annotated) in the gene. The junctions of interest are marked by the black arrows. The variant rs528823 is located on exon 7 at the position indicated by the red star. The effect size of this variant on the marked junctions

is shown in the box plots (bottom left). The variant has > 0.2 spliceAI delta score at the indicated splice sites (blue arrow). The bases highlighted in green show the position of the YBX3 binding motif. YBX3 binds at this location based on ENCODE eCLIP data and alters splicing based on the splicing change observed in YBX3 KD. Exon 7 encodes a transmembrane domain of the protein. **(C)** PCR reactions in K562 cell lines visualized on a agarose gel for the MS4A3 splicing event. From left to right there are 5 conditions: control (no ASO blocking), two ASOs with non-targeting control sequences (Supplementary Note), and two ASOs that targeting exon 7 of MS4A3 (Supplementary Note). When the ASO successfully blocks the splice site on exon 7, we see isoforms where exon 7 is skipped (smaller fragment further down the gel).