LeafCutter: Annotation-free quantification of RNA splicing

Yang I Li[1,6], David A Knowles[2,3,6], Jonathan K Pritchard[1,4,5]

[1] Department of Genetics, Stanford University, Stanford, CA

[2] Department of Computer Science, Stanford University, Stanford, CA

[3] Department of Radiology, Stanford University, Stanford, CA

[4] Department of Biology, Stanford University, Stanford, CA

[5] Howard Hughes Medical Institute, Stanford University, CA.

[6] These authors contributed equally to this work.

Correspondence should be addressed to Y.I.L. (yangili@stanford.edu), D.A.K. (dak33@stanford.edu), or J.K.P. (pritch@stanford.edu)

Abstract

The excision of introns from pre-mRNA is an essential step in mRNA processing. We developed LeafCutter to study sample and population variation in intron splicing. LeafCutter identifies variable intron splicing events from short-read RNA-seq data and finds alternative splicing events of high complexity. Our approach obviates the need for transcript annotations and overcomes the challenges in estimating relative isoform or exon usage in complex splicing events. LeafCutter can be used both for detecting differential splicing between sample groups, and for mapping splicing quantitative trait loci (sQTLs). Compared to contemporary methods, we find over three times more sQTLs, many of which help us ascribe molecular effects to disease-associated variants. LeafCutter is fast, easy to use, and available at `https://github.com/davidaknowles/leafcutter`*.*

Background

The regulated removal of introns during mRNA maturation is essential for major biological processes in eukaryotes including cellular differentiation, response to environmental stress, and proper gene regulation. Nevertheless, our ability to draw novel insights into the regulation and function of splicing is hindered by the challenge of estimating transcript abundances from short-read RNA-seq data.

Most popular approaches for studying alternative splicing from RNA-seq estimate isoform ratios (Trapnell et al., 2013; Leng et al., 2013; Patro et al., 2014; Bray et al., 2015) or exon inclusion levels (Katz et al., 2010; Anders et al., 2012). Quantification of isoforms or exons is intuitive because RNA-seq reads generally detect mature mRNA molecules from which introns have already been removed. However, estimation of isoform abundance from conventional short-read data is statistically challenging, as each read samples only a small part of the transcript, and alternative transcripts often have substantial overlap. Similarly, when estimating exon expression levels from RNA-seq read depths, read depths are often overdispersed due to technical effects, and there may be ambiguity about which version of an exon is supported by a read if there are alternative 5' or 3' splice sites.

Further, both isoform- and exon-quantification approaches generally rely on transcript models, or pre-defined splicing events, both of which may be inaccurate or incomplete (Vaquero-Garcia et al., 2016). Pre-defined transcript models are particularly limiting when comparing splicing profiles of healthy versus disease samples, as aberrant transcripts may be disease-specific; or when studying genetic variants that generate splicing events in a subset of individuals only (Stein et al., 2015). Even when transcript models are complete, it is difficult to estimate isoform or exon usage of complex alternative splicing events (Vaquero-Garcia et al., 2016).

An alternative perspective is to focus on what is *removed* in each splicing event. Excised introns may be inferred directly from reads that span exon-exon junctions. Thus, there is generally little ambiguity about the precise intron that is cut out, and quantification of usage ratios is often quite accurate (Vaquero-Garcia et al., 2016). A recent method, MAJIQ/VOILA (Vaquero-Garcia et al., 2016), also proposed to estimate local splicing variation using split-reads and identified complex splicing events, however it has not been adapted to map splicing QTLs (sQTLs). At present, there are several software programs for sQTL mapping: GLiMMPS (Zhao et al., 2013), sQTLseekeR (Monlong et al., 2014) and Altrans (Ongen and Dermitzakis, 2015). However, they rely on existing isoform annotations and reported modest numbers of sQTLs in example analyses. Powerful and streamlined tools for sQTL mapping are therefore critical in helping us interpret signal from genome-wide association studies.

Here we describe LeafCutter, a suite of novel methods that allow identification and quantification of novel and existing alternative splicing events by focusing on intron excisions. We show LeafCutter's utility by applying it to two important applications in genomics: (1) identification of differential splicing across conditions, and (2) identification of sQTLs in multiple tissues or cell types. Using LeafCutter, we recently found that alternative splicing is an important mechanism through which genetic variants contribute to disease risk (Li et al., 2016). We now show that LeafCutter dramatically increases the number of detectable associations between genetic variation and pre-mRNA splicing, thus enhancing our understanding of disease-associated loci.

Results

Overview of LeafCutter

LeafCutter uses short read RNA-seq data to detect intron excision events at base-pair precision by analyzing split-mapped reads. LeafCutter focuses on alternative splicing events including skipped exons, 5' and 3' alternative splice site usage and additional complex events that can be summarized by differences in intron excision (Vaquero-Garcia et al., 2016) (Supplementary Figure 1). LeafCutter's intron-centric view of splicing is motivated by the observation that mRNA splicing predominantly occurs through the step-wise removal of introns from nascent pre-mRNA (Tilgner et al., 2012). (Unlike isoform quantification methods such as Cufflinks (Trapnell et al., 2013), alternative transcription start sites, and alternative polyadenylation are not measured by LeafCutter as they are not directly affected by variation in intron excision.)

To identify alternatively excised introns, LeafCutter pools all mapped reads from a study and finds overlapping introns demarcated by split reads. LeafCutter then constructs a graph that connects all overlapping introns that share a donor or an acceptor splice site (Figure 1a). The connected components of this graph form clusters, which represent alternative intron excision events. Finally, LeafCutter iteratively applies a filtering step to remove spurious introns, which are defined based on the proportion of reads supporting an intron compared to other introns in the same cluster, and re-clusters leftover introns (Methods, Supplementary Methods 1). In practice, we found that this filtering is important to avoid arbitrarily large clusters when read depth increases to a level at which noisy splicing events are supported by multiple reads.

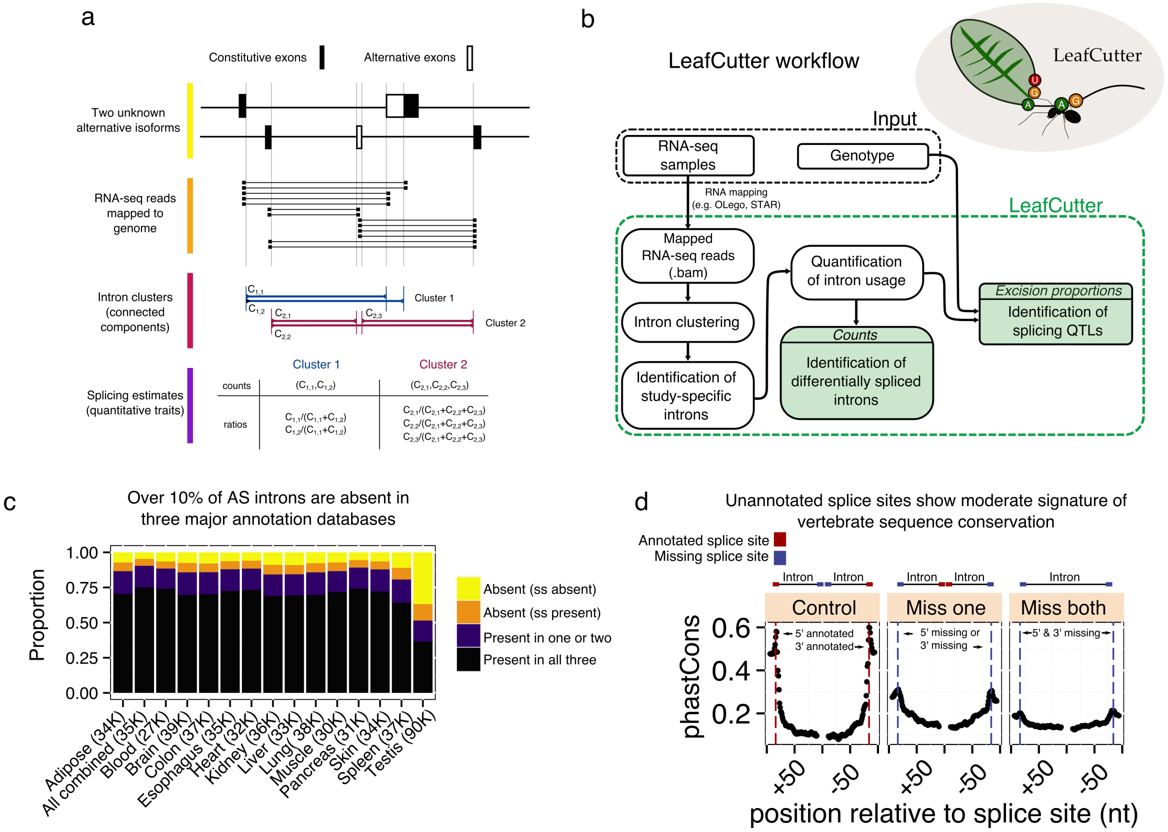

Figure 1: Overview of LeafCutter. **(a)** LeafCutter uses split reads to uncover alternative choices of intron excision by finding introns that share splice sites. In this example, LeafCutter identifies two clusters of variably excised introns. **(b)** LeafCutter workflow. First, short reads are mapped to the genome. When SNP data are available, WASP (van de Geijn et al., 2015) should be used to filter allele-specific reads that map with a bias. Next, LeafCutter extracts junction reads from `.bam` files, identifies alternatively excised intron clusters, and summarizes intron usage as counts or proportions. Lastly, LeafCutter identifies intron clusters with differentially excised introns between two user-defined groups using a Dirichlet-multinomial model or maps genetic variants associated with intron excision levels using a linear model. **(c)** Using LeafCutter to discover novel introns, we find that for any given tissue, over 10% of alternatively excised introns are unannotated. Remarkably, 48.5% of testis alternatively excised introns are unannotated. Different colors denote the proportion of introns when one or more splice sites are unannotated "(ss absent)", both splice sites are annotated but the intron is not part of any transcript "(ss present)", or when the intron is annotated in some but not all databases. **(d)** The unannotated splice sites of novel introns show moderate signature of sequence conservation as determined by vertebrate phastCons scores. Miss one: conservation of the unannotated splice site of an intron for which the cognate splice site is annotated. Miss both: conservation of splice sites of introns with both splice sites unannotated.

LeafCutter detects spliced introns *de novo*

We tested LeafCutter's novel intron detection method by mapping RNA-seq short read data from 2,192 samples (Supplementary Methods 2) across 14 tissues from the GTEx consortium (Ardlie et al., 2015) using OLego (Wu et al., 2013). We then searched introns predicted to be alternatively excised by LeafCutter in three commonly used annotation databases (GENCODE v19, Ensembl, and UCSC). To ensure that the identified introns were indeed alternatively excised, we restricted this analysis to introns that were excised at least 20% of the time as compared to other overlapping introns, in at least one fourth of the samples, considering each tissue separately. We found that between 10.8% and 19.3% (Pancreas and Spleen, respectively) of alternatively spliced introns are unannotated, with the exception of testis in which 48.5% of alternatively spliced introns are novel (Figure 1c). The latter observation is compatible with the "out-of-testis" hypothesis, which proposes that transcription is more permissive in testis and allows novel genes to be selected for if beneficial (Kaessmann, 2010). Combining all tissues but testis, 70,722 introns met our criteria, of which 22,278 were unannotated. Thus 31.5% of the alternatively excised introns we detected are unannotated (Supplementary Methods 3), consistent with a recent study that identified a similar proportion of novel splicing events in 12 mouse tissues (Vaquero-Garcia et al., 2016).

We next asked whether these novel introns show evidence of functionality as determined by sequence conservation. When we averaged phastCons scores over unannotated splice sites of introns that were absent in annotation databases, we found a moderate, but significant, signature of sequence conservation (Figure 1d). This supports the notion that noisy splicing is widespread in human and can drive transcript diversity (Pickrell et al., 2010). Nevertheless, we found that a significant number (4,616 or 15–25%) of novel splice sites are conserved across vertebrates (ave. phastCons ≥ 0.6, Supplementary Figure S3), indicating that the alternative excision of thousands of introns may be functional (Supplementary Methods 3).

Differential intron excision across sample groups

To evaluate LeafCutter's ability to identify differentially spliced introns, we searched for intron clusters that show differential splicing between tissue pairs collected by the GTEx consortium, using all tissues to identify intron clusters. LeafCutter uses counts from the clustering step (Figure 1b) to identify introns with differential splicing between user-defined groups. Read counts in an intron cluster are jointly modeled using a Dirichlet-multinomial generalized linear model (GLM), which offers superior sensitivity relative to a beta-binomial GLM that tests each intron independently (Supplementary Figure S4). The implicit normalization of the multinomial likelihood avoids the estimation of library size parameters required by methods such as DEXSEQ (Anders et al., 2012).

Combining all pairwise comparisons, we found 13,404 tissue-regulated splicing clusters in 8,335 genes at 10% FDR, which includes 4,179 clusters differentially spliced in 3,334 genes at 10% FDR between brain tissues and testis. As expected, GTEx samples mostly grouped by organ/tissue when hierarchically clustered according the excision ratios of the five hundred most differentially spliced introns among all tissue pairs

(Figure 2a, Supplementary Methods 5).

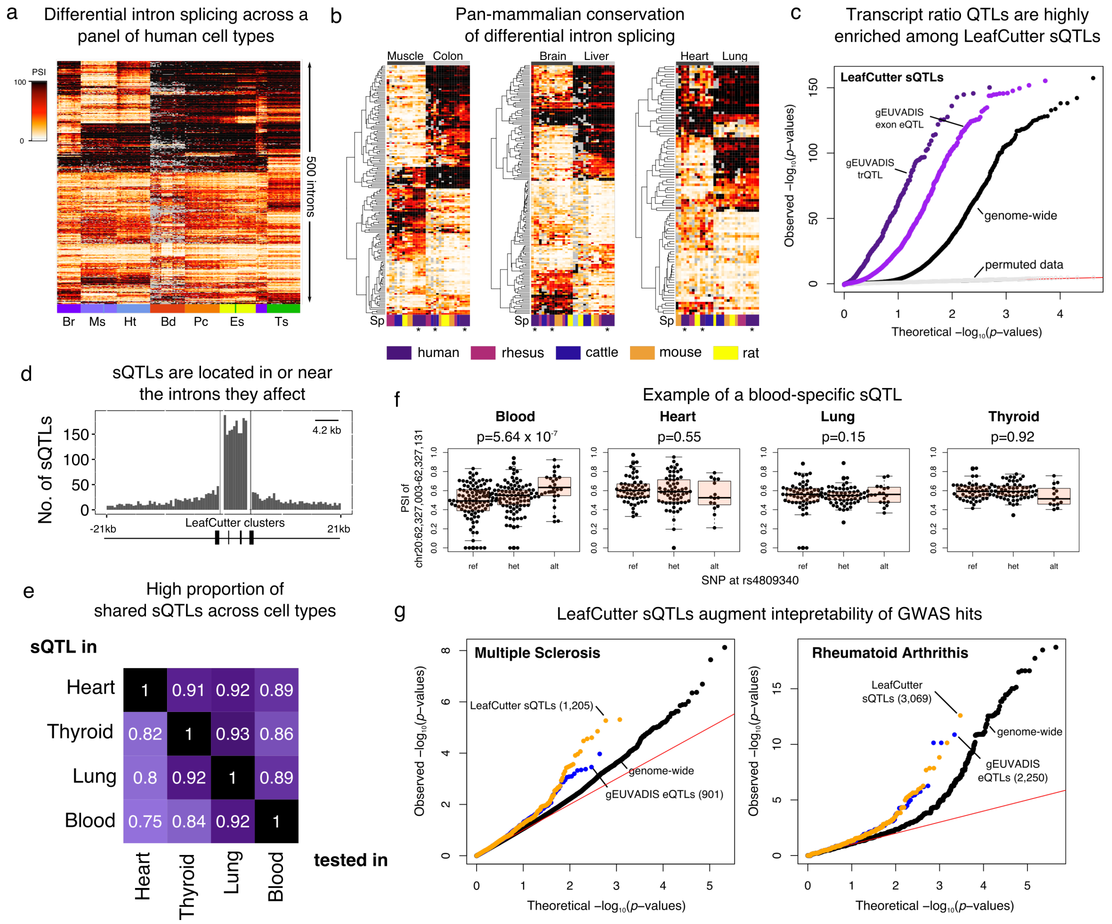

Figure 2: **(a)** LeafCutter identifies tissue-regulated intron splicing events from GTEx organ samples. Heatmap of the intron excision ratios of the top 500 introns that were found to be differentially spliced between at least one tissue pair. Tissues include brain (Br), muscle (Ms), heart (Ht), blood (Bd), pancreas (Pc), esophagus (Eg), and testis (Ts). **(b)** Tissue-dependent intron excision is conserved across mammals. Heatmap showing intron exclusion ratios of introns differentially spliced between pairs of tissues (Muscle vs Colon, Brain vs Liver, and heart vs Lung). Heatmap shows 100 random introns (97 for the heart vs lung comparison) that were predicted to be differentially excised in human with p-value $< 10^{-10}$ (LR-test) and that had no more than 5 samples where the excision rate could not be determined due to low count numbers. Heatmap of all introns that pass our criteria can be found in Supplementary Figure S6. **(c)** QQ-plot showing genome-wide sQTL signal in LCLs (black), sQTL signal conditioned on exon eQTLs (purple) and conditioned on transcript ratio QTLs (dark purple) from (Lappalainen et al., 2013). Signal from permuted data in light grey shows that the test is well-calibrated. **(d)** Positional distribution of sQTLs across LeafCutter-defined intron clusters. 1,421 of 4,543 sQTLs lie outside the boundaries (Supplementary Figure S7 for all sQTLs). **(e)** High proportion of shared sQTLs across four tissues from (Ardlie et al., 2015). **(f)** Example of a SNP associated to the excision level of an intron in blood but not in other tissues. **(g)** Enrichment of low p-value associations to multiple sclerosis and rheumatoid arthritis among LeafCutter sQTL and gEUVADIS eQTL SNPs. The numbers of top sQTLs and eQTLs that are tested in each GWAS are shown in parentheses.

We next investigated whether the differentially spliced clusters identified using LeafCutter are likely to be functional by assessing the pan-mammalian conservation of their splicing patterns across multiple organs.

Two previous studies analyzed the evolution of alternative splicing in mammals and, when they clustered samples using gene expression levels, saw clustering by organ as expected, however when they clustered samples using exon-skipping levels, they instead saw a clustering by species (Merkin et al., 2012; Barbosa-Morais et al., 2012). These observations indicate that a large number of alternative skipping events may lack function or undergo rapid turnover.

We initially attempted clustering using all splicing events and confirmed the previous finding that the samples mostly clustered by species (Supplementary Figure S5). We then focused on a subset of introns that LeafCutter identified as differentially excised across tissue pairs and found that this subset shows splicing patterns that are broadly conserved across mammalian organs (Figure 2b). To do this, we hierarchically clustered samples from eight organs in human and four mammals (Merkin et al., 2012) according to the orthologous intron excision proportions of differentially excised introns (p-value $< 10^{-10}$ and $\beta > 1.5$) from our pairwise analyses of human GTEx samples (Supplementary Methods 5). This revealed a striking clustering of the samples by organ, implying that hundreds of tissue-biased intron excisions events are conserved across mammals and likely have organ-specific functional roles (Reyes et al., 2013).

Mapping splicing QTLs using LeafCutter

Next, to evaluate LeafCutter's ability to map splicing QTLs, we applied LeafCutter to 372 EU lymphoblastoid cell line (LCL) RNA-seq samples from gEUVADIS, and identified 42,716 clusters of alternatively excised introns. We used the proportion of reads supporting each alternatively excised intron identified by LeafCutter and a linear model (Ongen et al., 2015) to map sQTLs (Supplementary Methods 6). We found 5,774 sQTLs at 5% FDR (compared to 620 trQTLs in the original study, i.e., 9.3 times as many) and 4,543 at 1% FDR. Compared to an sQTL mapping study that uses the same data, we found 3.2 times as many sQTLs (4,543 vs 1,427 sQTLs) at 1% FDR (Ongen and Dermitzakis, 2015).

To ensure that our sQTLs correspond to variants that affect splicing, we verified that LeafCutter finds stronger associations between intronic splicing levels and SNPs previously identified as exon eQTLs and transcript ratio QTLs in gEUVADIS (Lappalainen et al., 2013) when compared to genome-wide SNPs (Figure 2c). Importantly, 399 (81.3%) of the 491 top trQTLs tested are significantly associated to intron splicing variation, as identified by LeafCutter (compared to 4.7% when our samples are permuted, Supplementary Methods 7). Furthermore, we confirmed that the sQTLs we identified are located near splice sites and close to the introns they affect (Figure 2d).

We further used LeafCutter to identify sQTLs in four organs from the GTEx consortium. Overall, we found 442, 1,058, 1,047, and 692 sQTLs at 1% FDR in heart, lung, thyroid gland, and whole blood, respectively (Supplementary Methods 7). Using these, we estimated that 75–93% of sQTLs replicate across tissue pairs (Figure 2e, Sfigure S8, Supplementary Methods 6). This agrees with a high proportion of sharing of sQTLs across tissues (Hsiao et al., 2016); and contrasts with much lower pairwise sharing reported for these data previously (9-48%) (Ardlie et al., 2015). This leaves 7-25% of sQTLs that show tissue-specificity in our analysis. As expected we found that a large proportion of tissue-specific sQTLs arose from trivial cases where the intron is only alternatively excised, and therefore variable, in one tissue (Supplementary Figure S9). However, we also found cases in which the introns were alternatively excised in all tissues, yet show tissue-specific association with genotype (Figure 2f).

LeafCutter sQTLs link disease-associated variants to mechanism

Finally, we asked whether sQTLs identified using LeafCutter could be used to ascribe molecular effects to disease-associated variants as determined by genome-wide association studies. For example, eQTLs are enriched for disease-associated variants, and disease-associated variants that are eQTLs likely function by modulating gene expression (Lappalainen et al., 2013; Ardlie et al., 2015). We recently showed that sQTLs identified in LCLs are also enriched among autoimmune-disease-associated variants (Li et al., 2016). LeafCutter sQTLs can therefore help us characterize the functional effects of variants associated with complex diseases. Indeed, when we looked at the association signals of the top sQTLs (from LeafCutter) and eQTLs (from gEUVADIS) to multiple sclerosis and rheumatoid arthritis (Supplementary Methods 8), we found that both QTL types were enriched for stronger associations (Figure 2g) compared to genome-wide variants. Consistent with recent findings (Li et al., 2016), SNPs associated with multiple sclerosis are more highly enriched among sQTLs than eQTLs, while both eQTLs and sQTLs are similarly enriched among SNPs associated with rheumatoid arthritis (Figure 2g). These results demonstrate that by dramatically increasing the number of sQTLs detected, LeafCutter significantly enhances our ability to predict the molecular effects of disease-associated variants.

In conclusion, our analyses show that LeafCutter is a powerful approach to study variation in alternative splicing. By focusing on intron removal rather than exon inclusion rates, we can accurately measure the stepwise intron-excision process orchestrated by the splicing machinery. Our count based statistical modeling, accounting for over-dispersion, allows identification of robust variation in intron excision across conditions. Most importantly, LeafCutter allows the discovery of far more sQTLs than other contemporary methods, which improves our interpretation of disease-associated variants.

Methods

Identifying alternatively excised introns

To identify clusters of alternatively excised intron, split-reads that map with minimum 6nt into each exon are extracted from aligned `.bam` files. Overlapping introns defined by split-reads are then grouped together. For each of these groups LeafCutter constructs a graph where nodes are introns and edges represent shared splice junctions between two introns. The connected components of this graph define intron clusters. Singleton nodes (introns) are discarded. For each intron cluster, LeafCutter iteratively (1) removes introns that are supported by fewer than a number of (default 30) reads across all samples or fewer than a proportion (default 0.1%) of the total number of intronic read counts for the entire cluster, and (2) re-clustered introns according to the procedure above.

Dirichlet-multinomial generalized linear model

Intron clusters identified from LeafCutter comprise of two or more introns. More specifically, each intron clusters C identified using LeafCutter consists of J possible introns, which have counts $\mathbf{y}_{ij}$ for sample i and intron j (and cluster total $\mathbf{n}_{iC} = \sum_{j'} \mathbf{y}_{ij'}$), and N covariate column vectors $\mathbf{x}_i$ of length P. LeafCutter uses a Dirichlet-Multinomial ($\mathcal{DM}$) generalized linear model (GLM) to test for changes in intron usage across the entire cluster, instead of testing differential excision of each intron separately across conditions or genotypes.

$$\mathbf{y}_{i1}, ..., \mathbf{y}_{iJ} | \mathbf{n}_{iC} \sim \mathcal{DM}(\mathbf{n}_{iC}, \alpha_1 p_{i1}, \dots, \alpha_J p_{iJ}), \tag{1}$$

$$p_{ij} = \frac{\exp(\mathbf{x}_i \beta_j + \mu_j)}{\sum_{j'} \exp(\mathbf{x}_i \beta_{j'} + \mu_{j'})}, \tag{2}$$

where (2) corresponds to the softmax transform, which ensures $\sum_j p_{ij} = 1$. We perform maximum likelihood estimation for the outputs: the J coefficient row vectors β_j of length P, the intercepts μ_j and concentration parameters α_j. We use the following regularization to stabilize the optimization:

$$\alpha \sim \text{Gamma}(1 + 10^{-4}, 10^{-4}) \tag{3}$$

The Dirichlet-Multinomial likelihood is derived by integrating over a latent probability vector π in the hierarchy

$$\pi|a \sim \text{Dirichlet}(a) \Rightarrow P(\pi|a) = \frac{\Gamma(a_.)}{\prod_j \Gamma(a_j)} \prod_j \pi_j^{a_j - 1} \tag{4}$$

$$y_1, ..., y_J|n, \pi \sim \text{Multinomial}(n, \pi) \Rightarrow P(y|n, \pi) = \prod_j \pi_j^{y_j} \tag{5}$$

where $a_. = \sum_j a_j$, to give

$$\mathcal{DM}(y|n, a) = \frac{\Gamma(a_.) \prod_j \Gamma(a_j + y_j)}{\Gamma(a_. + y_.) \prod_j \Gamma(a_j)} \tag{6}$$

In the limit $\pi_j = e^{a_j} / \sum_{j'} e^{a_{j'}}, a_j \to \infty$ for all j, we have $\mathcal{DM}(n, a) \to \text{Multinomial}(n, \pi)$. For the GLM this means that as $\alpha_j \to \infty$ we recover a multinomial model with no overdispersion. Smaller values of α_j correspond to more overdispersion.

While the Dirichlet-multinomial effectively accounts for overdispersion, it fails to handle extremely outlying in samples, which negatively impacts calibration. To reduce sensitivity to such outliers we developed a robust likelihood model

$$\mathbf{y}_i|\mathbf{n}_{iC} \sim (1 - \theta_C)\mathcal{DM}(\mathbf{n}_{iC}, \boldsymbol{\alpha} \circ \mathbf{p}_i) + \theta_C \mathcal{DM}(\mathbf{n}_{iC}, \mathbf{1}) \tag{7}$$

where θ_C is a per-cluster mixture proportion giving the probability that a sample comes from the outlier distribution. Using $\mathbf{1}$ as the parameter vector for the outlier distribution corresponds to the underlying Dirichlet distribution being uniform over the simplex. θ_C is learnt jointly with the other parameters, and given a prior $\text{Beta}(1.01, 10^{-4})$.

Differential intron excision across conditions

To test differential intron excision between two groups of samples, we encode $x_i = 0$ for one group and $x_i = 1$ for the other in the Dirichlet-Multinomial generalized linear model. We apply two filters to ensure we only perform reasonable tests:

- Only introns which are detected (i.e. have at least one corresponding spliced read) in at least five samples are tested.
- A cluster is only tested if each group includes at least 4 individuals with 20 spliced reads supporting introns in the cluster.

The thresholds in these filters are easily customizable as optional parameters.

Mapping splicing QTLs

To identify splicing QTLs, RNA-seq reads are mapped onto the genome using a RNA-aligner such as STAR (Dobin et al., 2013) or OLego (Wu et al., 2013). Because LeafCutter only uses reads that map across junctions to estimate intron excision rates, it is essential to remove read-mapping biases caused by allele-specific reads. This is particularly significant when a variant is covered by reads that also span intron junctions as it can lead to spurious association between the variant and intron excision level estimates. Subsequent to mapping, LeafCutter finds alternatively excised intron clusters and quantifies intron excision levels in all samples. LeafCutter outputs intron excision proportions, which are used as input for standard QTL mapping tools such as MatrixEQTL or fastQTL (Supplementary Methods 6).

Acknowledgement

We thank Xun Lan and other members of the Pritchard Lab for helpful discussions and comments. This work was supported by a CEHG Fellowship, the Howard Hughes Medical Institute, and the US National Institutes of Health (NIH grants HG007036 and MH101825)

Author Contributions

Y.I.L., D.A.K. and J.K.P. conceived of the project. Y.I.L. and D.A.K. performed the analyses and implemented the software. D.A.K. developed and performed the statistical tests and modeling. Y.I.L. and J.K.P. wrote the manuscript.

[References at the end]

Supplementary Material for

Annotation-free quantification of intron splicing for genomic studies

Yang I Li[1,6], **David A Knowles**[2,3,6], **Jonathan K Pritchard**[1,4,5].

Contents

List of Figures

1 Identifying alternatively spliced introns using LeafCutter

Starting from alignment files in `.bam` format, junctions from split-reads that map with minimum 6nt into each exon are extracted using a script we provide (1) based on two OLego helper scripts. Then, the LeafCutter clustering program (2) can be used to identify intron clusters supported by at least 30 (option `-m`) total reads (across all samples) and introns supported by more than 0.1% (option `-p`) of the total read counts for the entire cluster. The number of reads supporting each intron and cluster is then counted in all samples separately and collated in a table for downstream analyses.

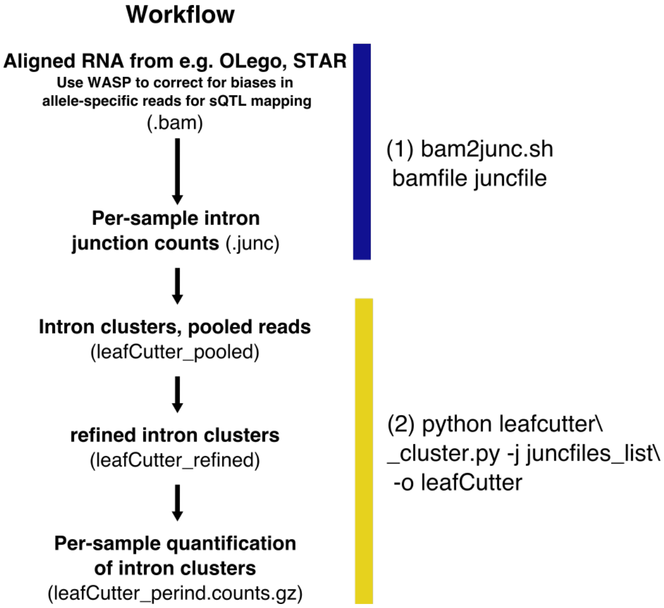
Workflow
Aligned RNA from e.g. OLego, STAR
Use WASP to correct for biases in
allele-specific reads for sQTL mapping
(.bam)
Per-sample intron
junction counts (.junc)
Intron clusters, pooled reads
(leafCutter_pooled)
refined intron clusters
(leafCutter_refined)
Per-sample quantification
of intron clusters
(leafCutter_perind.counts.gz)
(1) bam2junc.sh
bamfile juncfile
(2) python leafcutter\
_cluster.py -j juncfiles_list\
-o leafCutter

Figure S1: Helper method and LeafCutter workflow for intron clustering.

Because LeafCutter focuses on intron splicing rather than whole isoform quantification, alternative transcription start site or polyadenylation sites are not captured. However, several prevalent types of alternative splicing (Figure S2) are equivalent to specific intron excision events.

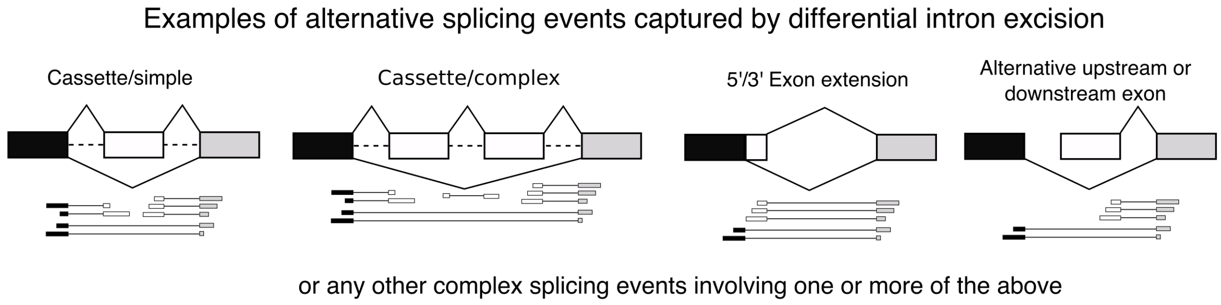
Examples of alternative splicing events captured by differential intron excision
Cassette/simple
Cassette/complex
5'/3' Exon extension
Alternative upstream or downstream exon
or any other complex splicing events involving one or more of the above

Figure S2: Several types of common alternatively splicing events are captured by the alternative excision of introns.

2 RNA-seq data processing

2.1 GTEx for intron discovery

We downloaded 2,192 RNA-seq samples from GTEx (Table S2). To analyze these, we used OLego (Wu et al., 2013) to map the RNA-seq reads to the human genome (hg19) and processed the resulting .bam files using LeafCutter.

Tissue	Sample Number
Heart	153
Testis	67
Spleen	7
Skin	340
Brain	422
Colon	86
Blood	270
Pancreas	66
Adipose Tissue	172
Lung	151
Esophagus	238
Muscle	176
Kidney	8
Liver	35

Table S2: Sample sizes of processed GTEx RNA-seq short read data by tissue type.

2.2 gEUVADIS for sQTL mapping

To control for differences in mapping procedures, we downloaded the .bam files directly from ArrayExpress (E-GEUV-3) and processed them using LeafCutter to obtain intron clusters and quantifications. We recommend the use of WASP (van de Geijn et al., 2015) to correct for biases caused by allelic reads. However, to make our comparison to other tools fair, we used the aligned reads available on ArrayExpress, and removed all clusters with an association to a SNP that overlap junction reads (see section entitled "sQTL mapping using LeafCutter"). This approach is conservative as some allelic reads do not map with a bias.

2.3 GTEx for sQTLs mapping

Again, to control for differences in mapping procedures, we used the `.bam` files provided by the GTEx consortium for sQTL mapping, and removed all clusters with an association to a SNP that overlap junction reads.

3 Identification of unannotated introns in tissues from GTEx

To obtain a comprehensive set of annotated introns, we downloaded the GENCODE (v19), UCSC, and RefSeq annotation databases in `.gtf` format. We classified introns as annotated if their 5' and 3' splice sites correspond to the end and start, respectively, of two consecutive exons in at least one transcript. As such it is possible that both 5' and 3' splices sites of a novel intron are annotated. We note that although a large proportion of annotated introns are present in all three databases, we found that the GENCODE annotation has the most comprehensive list of introns.

To estimate the number of unannotated alternatively excised (AE) introns, we first mapped 2,192 RNA-seq samples from 14 tissues (GTEx) to the human genome (hg19) using OLego, allowing *de novo* splice junction predictions. We then used LeafCutter to identify alternatively excised introns by pooling all junction reads. We then restricted our analyses to AE introns that were supported by at least 20% of the total number of reads that support introns from the clusters they belong to in at least 25% of all samples, considering each tissue separately. Although there is no minimum read count (an intron supported by 20 reads, 20% of 100, is less likely to be the outcome of noisy splicing than one supported by 2 reads out of 10), we reasoned that requiring 20% percent-splicing in 25% of all samples will filter out most sequencing technical artifact and noisy splicing. Importantly, using different cutoffs does not alter qualitatively our conclusions. This resulted in 70,722 AE introns that met these criteria, of which 22,278 (31.5%) AE introns were absent from all three annotation databases.

To investigate the functionality of these unannotated introns, we asked whether the unannotated splice sites of the 22,278 AE novel introns show signature of sequence conservation across vertebrates. To do this, we divided splice sites into three classes: (1) control splice sites, which are annotated in one or more databases, but whose cognate splice site is unannotated, (2) the cognate splice site itself, and (3) splice sites of introns, for which both splice sites are unannotated. To compute sequence conservation, we average the phastCons score of the predicted splice sites (over 96% of which are AG/GT) plus 2 flanking bases. Interestingly, we find that the average sequence conservation of unannotated splice sites is higher if its cognate splice site is annotated (Figure 1d, Figure S3).

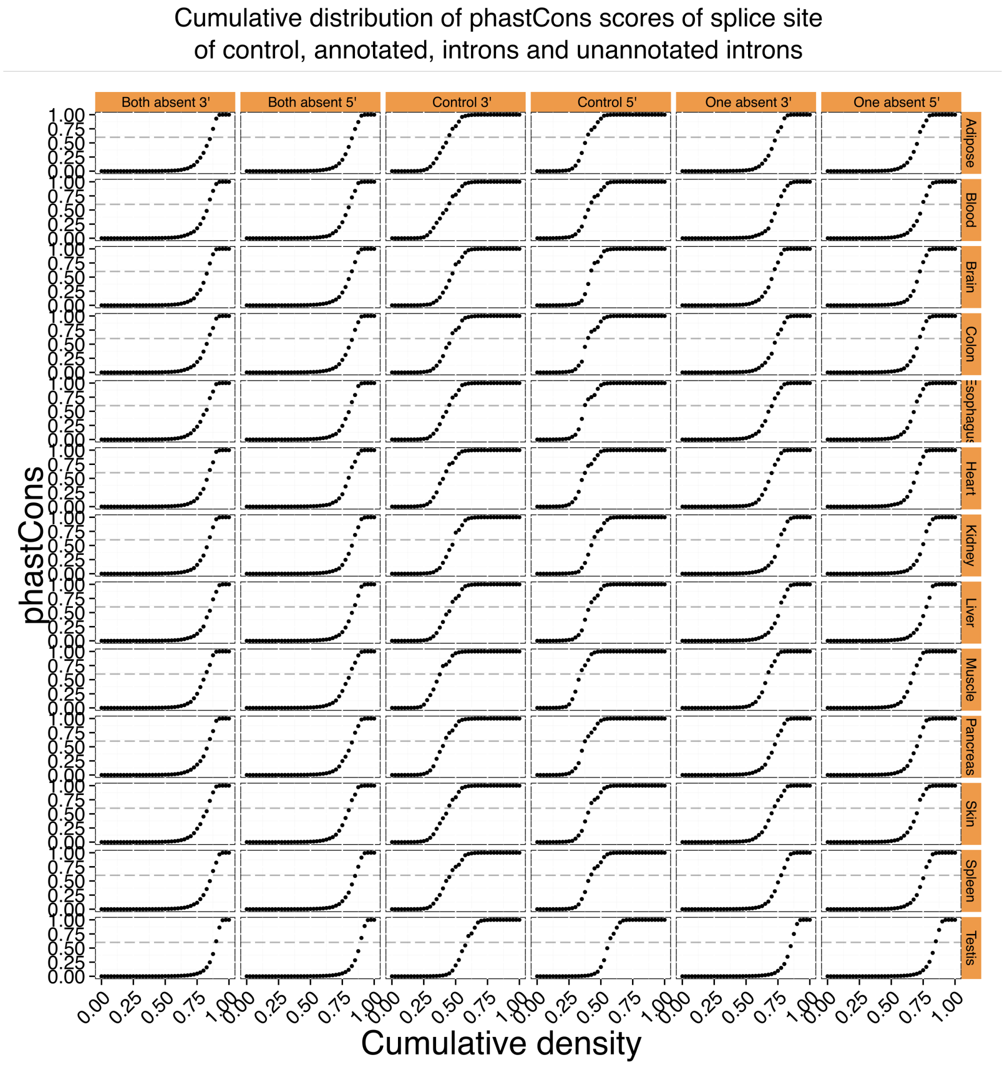

Figure S3: PhastCons score distribution of splice site of novel introns. While ∼60% of annotated splice sites have local phastCons score >0.6, only 15-25% of unannotated splice sites do. Thus ∼80% of novel splice sites may represent noisy intron excision events.

4 Statistical models

For cluster C containing J possible introns, let $\mathbf{y}_{ij}$ denote the count for sample i and intron j (and cluster total $\mathbf{n}_{iC} = \sum_{j'} \mathbf{y}_{ij'}$), and $\mathbf{x}_i$ denote a P-vector of covariates.

4.1 Beta-binomial GLM.

Our initial approach was to test each specific intron j of a cluster using

$$\mathbf{y}_{ij} | \mathbf{n}_{iC} \sim \mathcal{BB}(\mathbf{n}_{iC}, \alpha p_i, \alpha(1 - p_i)), \tag{8}$$

$$p_i = \sigma(\mathbf{x}_i \beta + \mu) \tag{9}$$

where $\mathcal{BB}$ is the beta-binomial distribution and $\sigma(x) = 1/(1 + e^{-x})$ is the logistic function. Here the parameters to be learnt are the P-vector β, intercept μ and concentration parameter α. Higher values of α correspond to the underlying beta distribution concentrating around p_i, and therefore to less count overdispersion. In particular as $\alpha \to \infty$ the $\mathcal{BB}$ likelihood converges to a multinomial likelihood, recovering a logistic regression model.

Optimization. For both the beta-binomial and Dirichlet-multinomial models we use the Bayesian probabilistic programming language Stan (Carpenter et al., 2015) to define the model, generate efficient C++ code for likelihood and gradient calculation, and to perform optimization using LBFGS.

Regularization. For some cases the likelihood as a function of the overdispersion parameter can be extremely flat, leading to numerical instability. In order to stabilize the optimization we use very weak regularization in the form of the prior

$$\alpha \sim \text{Gamma}(1 + 10^{-4}, 10^{-4}) \tag{10}$$

We experimented with two different versions of the $\mathcal{DM}$ GLM. The first uses a shared concentration parameter $\alpha_j = \alpha$ for all introns j in a cluster (the beta-binomial GLM is a special case of this model). The second allows a different α_j for each intron in the cluster.

Identifiability. The $\mathcal{DM}$ GLM shares with the more standard Multinomial GLM that the form in Equation 2 has a spurious degree of freedom: in particular, adding a constant to the input of the softmax does not

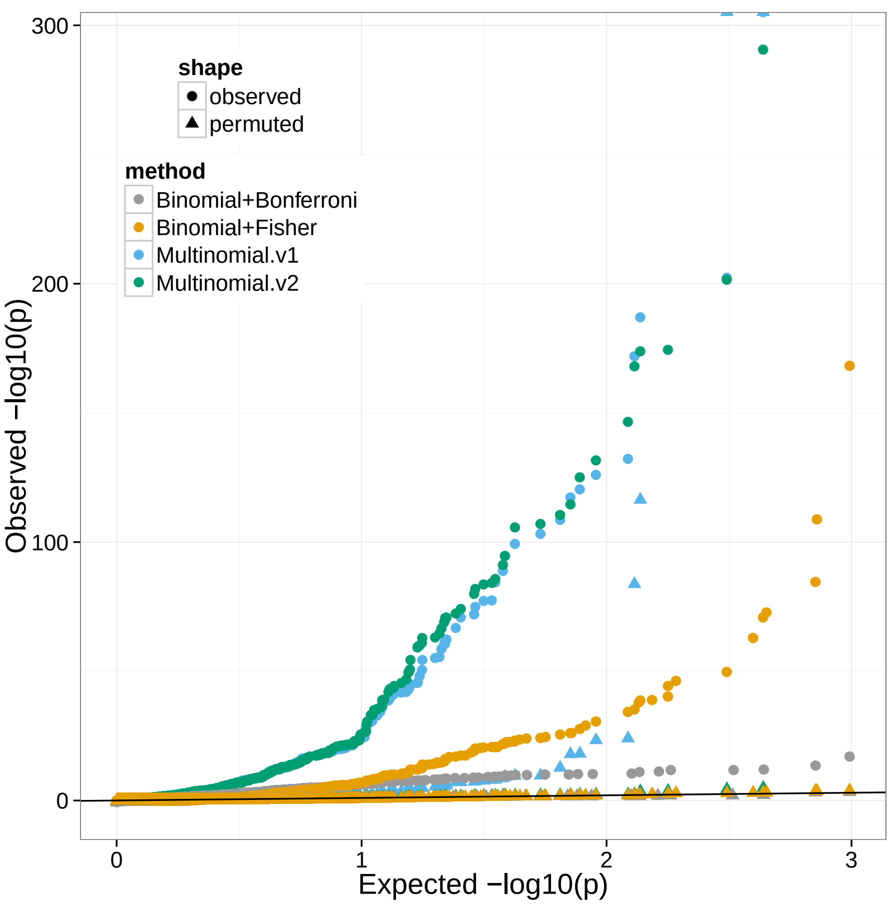

Figure S4: Comparison between beta-binomial and Dirichlet-multinomial models for differential splicing analyses, performed on 10 male brain vs. heart samples from GTEx. Two approaches for combining per-intron p-values into cluster level introns are compared: Bonferroni correction and Fisher's combined test. Bonferroni is very conservative, as expected. Fisher's combined test has considerably lower power than the multinomial approaches. However, only v2 of the Dirichlet-multinomial (which uses a per intron concentration/overdispersion parameter) is well calibrated under permutations.

change its output. To remove this degree of freedom from the model we parameterize each β_j as

$$\beta_{jp} := \bar{\beta}_p(\tilde{\beta}_{jp} - \frac{1}{J}) \tag{11}$$

where $\tilde{\beta}_{1p}, ..., \tilde{\beta}_{Jp}$ is constrained to lie on the J-simplex, i.e. $\tilde{\beta}_{jp} \geq 0, \sum_j \tilde{\beta}_{jp} = 1$, a constraint Stan naturally handles using a change of variables.

4.2 Likelihood ratio tests

Likelihood ratio tests are generally better calibrated than alternatives such as Wald statistics for testing for the significance of covariates, especially for modest sample sizes. We optimize wrt to β, μ, α separately for the null and alternative models (excluding and including the group indicator x respectively) to obtain log likelihoods λ_0 and λ_1 (for efficiency we initialize the optimization for the alternative model using the null model parameters) and then perform a likelihood ratio test: under the null $2(\lambda_1 - \lambda_0) \sim \chi^2_\rho$ where ρ is the appropriate degrees of freedom. For the beta-binomial GLM $\rho = P_1 - P_0$ where P_0 and P_1 are the number of covariates in the null and alternative models respectively. For the Dirichlet-multinomial GLM we have $\rho = (J-1)(P_1 - P_0)$ where J is the number of introns in the cluster.

5 Differential intron excision analyses

5.1 Identification of tissue-dependent intron excision levels

We used LeafCutter's Dirichlet-multinomial GLM to identify intron clusters with at least one differentially excised intron. We searched for intron excision level differences between all tissue pairs. However, we should note that owing to sample size differences, we will have different power to detect differential splicing of varying magnitude between pairs (we can detect splicing differences of small magnitude only in comparisons with large sample sizes). When we hierarchically clustered all samples according to the intron excision levels of introns that were present (i.e. were supported by reads) in all species, we saw a mix between tissue and species clustering (Figure S5). However, when we conditioned on introns that were differentially excised across human tissue pairs according to LeafCutter, we saw a clear clustering by tissue (Figure 2a).

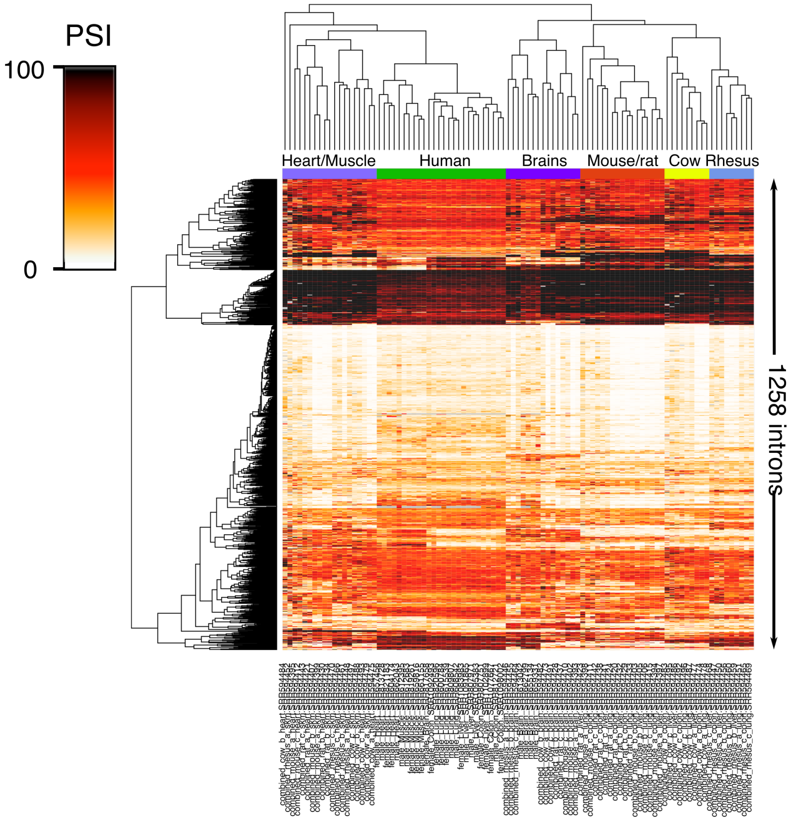

Figure S5: Hierarchical clustering on all 1,258 introns that had no missing values in any of the samples.

5.2 Pan-mammalian tissue clustering of intron excision profiles

To evaluate the conservation of intron excision profiles across mammalian tissues, we used OLego to map RNA-seq data (Merkin et al., 2012) from eight organs (testes, heart, kidney, liver, lung, brain, colon, and spleen) in four mammals (mouse, rat, cow, and rhesus macaque) to their respective genomes. We then projected all introns supported by RNA-seq reads onto the human genome using liftOver and clustered projected introns from all four mammals and human GTEx samples using LeafCutter. We then focused on four disjoint pairwise comparisons (Testis vs Kidney, Muscle vs Colon, Heart vs Lung, and Brain vs Liver, Figure S6).

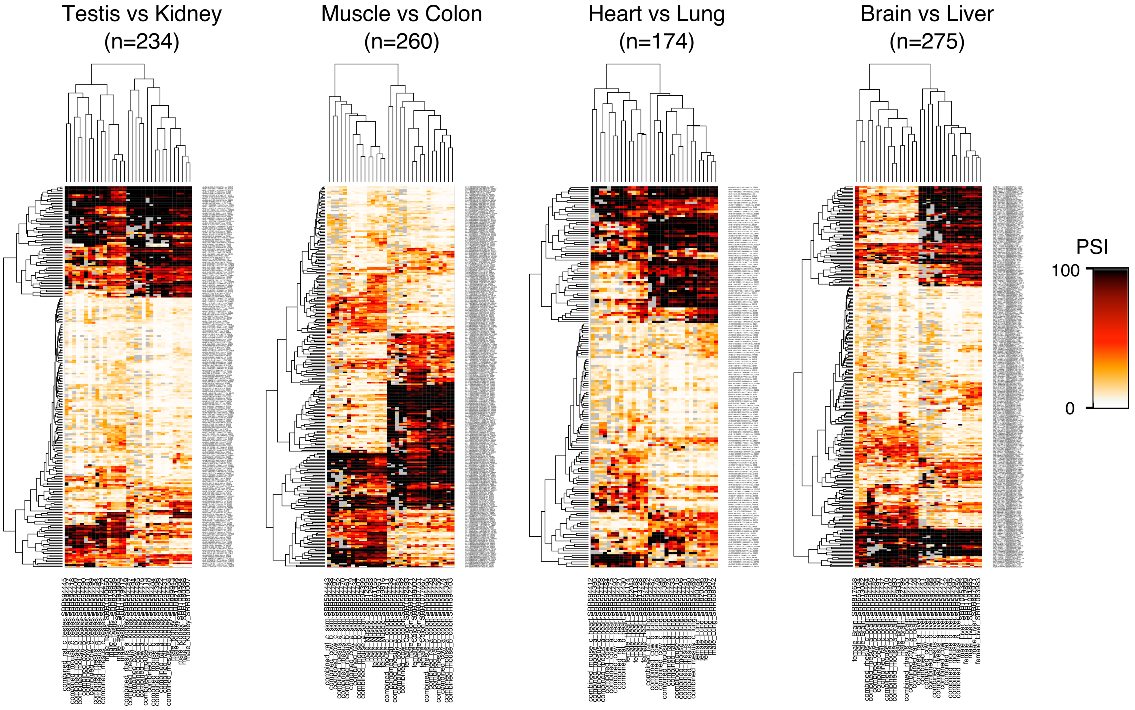

Figure S6: We restricted to introns that were found to be differentially excised between human tissues (p-value $< 10^{-10}$ and effect size > 1.0)

6 sQTL mapping using LeafCutter

6.1 Mapping sQTLs in gEUVADIS LCL samples (linear regression)

To map sQTLs in gEUVADIS LCLs samples, we restricted our analysis on 372 samples derived from European individuals. We downloaded genotype files from ArrayExpress (E-GEUV-1). We used LeafCutter to obtain read proportions for all introns within alternatively excised intron clusters. We then standardized the values across individuals for each intron and quantile normalize across introns (Degner et al., 2012) and used this as our phenotype matrix. We then used linear regression (as implemented in fastqtl) (Ongen et al., 2015) to test for associations between variants (MAF ≥ 0.05) within 100kb of intron clusters and the rows of our phenotype matrix that correspond to the introns within each cluster. As covariate, we used the first 3 principal components of the genotype matrix plus the first 15 principal components of the phenotype matrix. To estimate the number of sQTLs at any given false discovery rate (FDR), we used the correct p-values from fastqtl, and then used Bonferroni correction to control for the number of introns we test per cluster (note that this is conservative). We then use Benjamini-Hochberg to estimate the FDR (sample permutations show that our association p-values at this step are well calibrated).

For this analysis, we do not correct for biases caused by allelic reads to keep comparisons fair, we instead removed all associations that might be caused by SNPs that overlap junction reads. To do this, we removed all intron clusters that had a variant that were 70 or fewer base pairs (gEUVADIS RNA-seq read length is 75bp and at least 6nt must overlap with all exons) away from the splice sites (in the exonic part).

6.2 Mapping sQTLs in gEUVADIS LCL samples (Dirichlet-multinomial GLM)

In addition to using linear regression, we also used LeafCutter's Dirichlet-multinomial GLM to map sQTLs. This approach has two main advantages: (1) it accounts for the over-dispersion of read count data, and (2) it combines signal from changes in intron excision levels across the entire cluster instead of considering each intron independently. However, when we applied to our gEUVADIS data and controlled FDR using permutations, we found fewer sQTLs than our linear model approach, likely driven by clusters with heavy-tailed count distributions which are effectively handled by the quantile normalization in the linear approach.

6.3 Mapping sQTLs in four GTEx tissues

To identify sQTLs in GTEx tissues, we used the same strategy as in gEUVADIS LCLs (linear regression). However, we used the first 5 genotype PCs and the first 10 PCs as covariates (5+10 instead of 3+15).

Tissue	Number of individuals
Heart	95
Blood	170
Lung	128
Thyroid	118

Table S6: Sample sizes of processed GTEx `.bam` files for sQTL mapping.

7 sQTL analyses

7.1 Comparison with gEUVADIS exon eQTLs, and trQTLs

Although LeafCutter does not explicitly search for genetic variants that are associated with differences in exon level splicing or transcript ratios, we expected that these variants will also affect intron excision, which are detected by LeafCutter. To verify this, we compared the distribution of p-values from the association between LeafCutter intron excision and genome-wide SNPs to the p-values from the association between LeafCutter intron excision and SNPs that were previously classified as exon eQTLs and transcription ratio QTLs in gEUVADIS. More specifically, we downloaded the list of exon eQTLs and trQTLs from ArrayExpress (E-GEUV-3) and for each exon/gene took the SNP with the strongest association to exon level or transcript ratio. We then computed the association p-values of these SNPs with all tested LeafCutter intron excision levels, using Bonferroni correction to adjust our p-values. As expected both exon eQTL and trQTL SNPs were enriched in strong associations to intron excision levels compared to random SNPs, and trQTL SNPs were most enriched in strong associations.

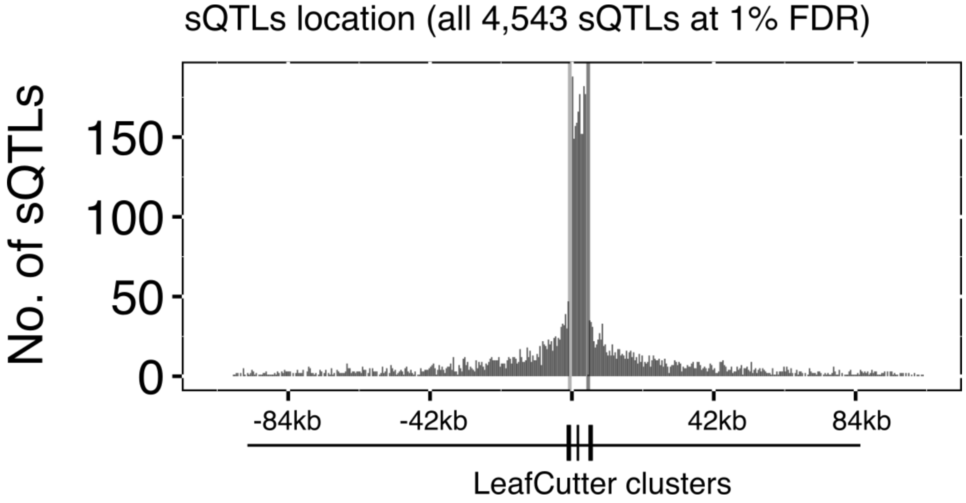

Figure S7: Meta-cluster representation of position of all 4,543 sQTLs identified at 1%FDR.

We next wished to verify that trQTLs detected in gEUVADIS were mostly identified as LeafCutter intron sQTLs. We again took the best trQTL SNP for each gene, and estimated the number that were associated with a cluster at a corrected p-value < 0.05. To correct for SNPs tested against multiple clusters, we used Bonferroni correction to adjust the p-value of the strongest association. We find that 399 (81.3%) of the 491 top trQTLs we tested are significantly associated ($p < 0.05$); this percentage is likely higher because our Bonferroni correction is conservative. Furthermore, as expected, when we use the same procedure to ask how many of the top 491 trQTLs are significantly associated to intron splicing when our sample labels are permuted, we find that only 4.7% are (our statistical tests are well calibrated; ∼5% of our tests should achieve a 0.05 significance under the null model).

7.2 Replication of sQTLs across GTEx tissues

To estimate the proportion of sQTLs that are replicable across tissue types, we took the best SNP of each sQTL-cluster pair for each tissue and asked whether the sQTL association was significant ($p < 0.05$) in another tissue. This estimate is likely to be conservative as it does not account for incomplete power. The replication is therefore likely to be even higher than our current estimates of 75-93%.

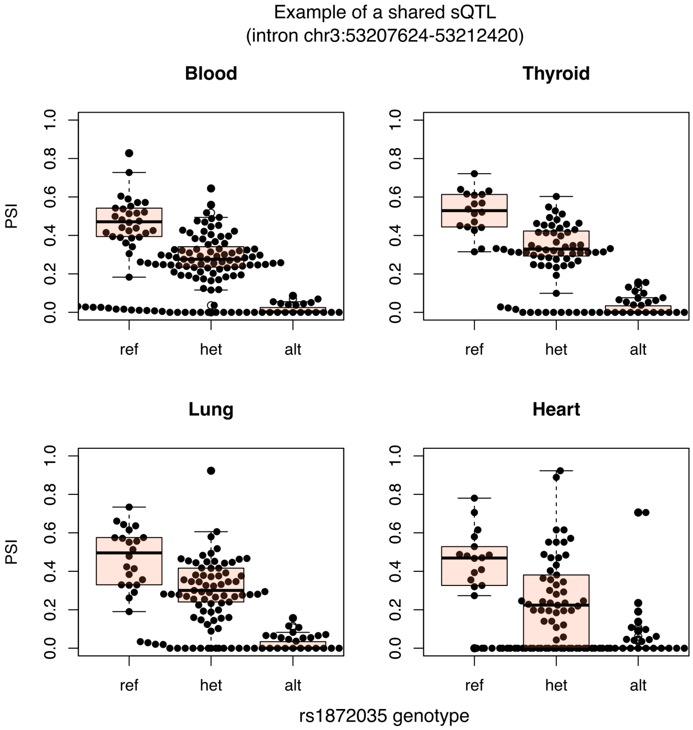

Figure S8: Example of a shared sQTL.

7.3 Tissue-specific sQTLs

To identify tissue-specific sQTLs, we searched for genetic variants that were associated significantly with intron excision levels in one tissue, but not in any of the other three tissues ($p > 0.1$), requiring all tissues to have junction reads in the intron cluster.

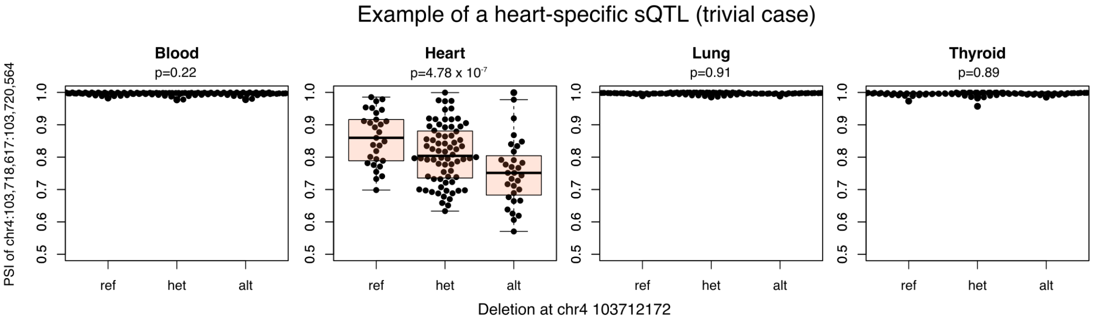

Figure S9: Example of a tissue-specific sQTL.

8 LeafCutter sQTL signals in genome-wide association studies

To verify that LeafCutter sQTLs can help us identify disease-associated variants that function by modulating splicing, we downloaded summary statistics from two autoimmune GWAS studies (multiple sclerosis (Sawcer et al., 2011) and rheumatoid arthritis (Okada et al., 2014)) and looked for enrichment of strong association *p*-values among the top LeafCutter sQTLs and gEUVADIS gene eQTLs (we removed the extended MHC region from this analysis). We found that 1,205 LeafCutter sQTL SNPs and 901 gEUVADIS eQTL SNPs (the SNP with most significant *p*-value) were also tested (with >5% MAF) in the multiple sclerosis genome-wide association study, and that 3,069 LeafCutter sQTL SNPs and 2,250 gEUVADIS eQTL SNPs were tested in the rheumatoid arthritis study. We then took the QTLs and plotted the distribution of $-\log_{10}(p\text{-value})$ of their association to each trait separately. As expected (Li et al., 2016), we found that LeafCutter sQTLs were more highly enriched in associations with low *p*-values compared to gEUVADIS eQTLs in multiple sclerosis and were similarly enriched in rheumatoid arthritis. This is notable because we considered a larger number of LeafCutter sQTLs than gEUVADIS eQTLs for both diseases. These observations suggest that LeafCutter allows us to identify as many or more disease-associated variants that *act* by affecting splicing as compared to those that *act* by affecting total expression levels.

9 Processed data availability

Data	Accession
RNA-seq and genotype (gEUVADIS)	E-GEUV-3 (ArrayExpress)
RNA-seq (Merkin et al., 2012)	GSE41637 (GEO)
RNA-seq and genotype (GTEx)	phs000424 (dbGaP)

References

Anders, S., Reyes, A., and Huber, W., 2012. Detecting differential usage of exons from RNA-seq data. *Genome Res.*, **22**(10):2008–2017.

Ardlie, K. G., Deluca, D. S., Segre, A. V., Sullivan, T. J., Young, T. R., Gelfand, E. T., Trowbridge, C. A., Maller, J. B., Tukiainen, T., Lek, M., *et al.*, 2015. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, **348**(6235):648–660.

Barbosa-Morais, N. L., Irimia, M., Pan, Q., Xiong, H. Y., Gueroussov, S., Lee, L. J., Slobodeniuc, V., Kutter, C., Watt, S., Colak, R., *et al.*, 2012. The evolutionary landscape of alternative splicing in vertebrate species. *Science*, **338**(6114):1587–1593.

Bray, N., Pimentel, H., Melsted, P., and Pachter, L., 2015. Near-optimal rna-seq quantification. *arxiv*, **1**(1):1.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., and Riddell, A., *et al.*, 2015. Stan: a probabilistic programming language. *Journal of Statistical Software*, **1**(1).

Degner, J. F., Pai, A. A., Pique-Regi, R., Veyrieras, J. B., Gaffney, D. J., Pickrell, J. K., De Leon, S., Michelini, K., Lewellen, N., Crawford, G. E., *et al.*, 2012. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature*, **482**(7385):390–394.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R., 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**(1):15–21.

Hsiao, Y. E., Bahn, J. H., Lin, X., Chan, T. M., Wang, R., and Xiao, X., 2016. Alternative splicing modulated by genetic variants demonstrates accelerated evolution regulated by highly conserved proteins. *Genome Res.*, .

Kaessmann, H., 2010. Origins, evolution, and phenotypic impact of new genes. *Genome Res.*, **20**(10):1313–1326.

Katz, Y., Wang, E. T., Airoldi, E. M., and Burge, C. B., 2010. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods*, **7**(12):1009–1015.

Lappalainen, T., Sammeth, M., Friedlander, M. R., 't Hoen, P. A., Monlong, J., Rivas, M. A., Gonzalez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P. G., *et al.*, 2013. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, **501**(7468):506–511.

Leng, N., Dawson, J. A., Thomson, J. A., Ruotti, V., Rissman, A. I., Smits, B. M., Haag, J. D., Gould, M. N., Stewart, R. M., and Kendziorski, C., *et al.*, 2013. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics*, **29**(8):1035–1043.

Li, Y. I., van de Geijn, B., Raj, A., Knowles, D. A., Petti, A. A., Golan, D., Gilad, Y., and Pritchard, J. K., 2016. RNA splicing is a primary link between genetic variation and disease. *Revise and Resubmit*, . 2016.

Merkin, J., Russell, C., Chen, P., and Burge, C. B., 2012. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science*, **338**(6114):1593–1599.

Monlong, J., Calvo, M., Ferreira, P. G., and Guigo, R., 2014. Identification of genetic variants associated with alternative splicing using sQTLseekeR. *Nat Commun*, **5**:4698.

Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., Kochi, Y., Ohmura, K., Suzuki, A., Yoshida, S., *et al.*, 2014. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*, **506**(7488):376–381.

Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T., and Delaneau, O., 2015. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics*, .

Ongen, H. and Dermitzakis, E. T., 2015. Alternative Splicing QTLs in European and African Populations. *Am. J. Hum. Genet.*, **97**(4):567–575.

Patro, R., Mount, S. M., and Kingsford, C., 2014. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat. Biotechnol.*, **32**(5):462–464.

Pickrell, J. K., Pai, A. A., Gilad, Y., and Pritchard, J. K., 2010. Noisy splicing drives mRNA isoform diversity in human cells. *PLoS Genet.*, **6**(12):e1001236.

Reyes, A., Anders, S., Weatheritt, R. J., Gibson, T. J., Steinmetz, L. M., and Huber, W., 2013. Drift and conservation of differential exon usage across tissues in primate species. *Proc. Natl. Acad. Sci. U.S.A.*, **110**(38):15377–15382.

Sawcer, S., Hellenthal, G., Pirinen, M., Spencer, C. C., Patsopoulos, N. A., Moutsianas, L., Dilthey, A., Su, Z., Freeman, C., Hunt, S. E., *et al.*, 2011. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature*, **476**(7359):214–219.

Stein, S., Lu, Z. X., Bahrami-Samani, E., Park, J. W., and Xing, Y., 2015. Discover hidden splicing variations by mapping personal transcriptomes to personal genomes. *Nucleic Acids Res.*, **43**(22):10612–10622.

Tilgner, H., Knowles, D. G., Johnson, R., Davis, C. A., Chakrabortty, S., Djebali, S., Curado, J., Snyder, M., Gingeras, T. R., and Guigo, R., *et al.*, 2012. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res.*, **22**(9):1616–1625.

Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., and Pachter, L., 2013. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.*, **31**(1):46–53.

van de Geijn, B., McVicker, G., Gilad, Y., and Pritchard, J. K., 2015. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods*, **12**(11):1061–1063.

Vaquero-Garcia, J., Barrera, A., Gazzara, M. R., Gonzalez-Vallinas, J., Lahens, N. F., Hogenesch, J. B., Lynch, K. W., and Barash, Y., 2016. A new view of transcriptome complexity and regulation through the lens of local splicing variations. *Elife*, **5**.

Wu, J., Anczukow, O., Krainer, A. R., Zhang, M. Q., and Zhang, C., 2013. OLego: fast and sensitive mapping of spliced mRNA-Seq reads using small seeds. *Nucleic Acids Res.*, **41**(10):5149–5163.

Zhao, K., Lu, Z. X., Park, J. W., Zhou, Q., and Xing, Y., 2013. GLiMMPS: robust statistical model for regulatory variation of alternative splicing using RNA-seq data. *Genome Biol.*, **14**(7):R74.