Published online 14 May 2021

NAR Genomics and Bioinformatics, 2021, Vol. 3, No. 2 **1**
doi: 10.1093/nargab/lqab041

Quantifying splice-site usage: a simple yet powerful approach to analyze splicing

Craig I. Dent[1], Shilpi Singh[1], Sourav Mukherjee[2], Shikhar Mishra[1], Rucha D. Sarwade[1], Nawar Shamaya[1], Kok Ping Loo[1], Paul Harrison[3], Sridevi Sureshkumar[1], David Powell[3] and Sureshkumar Balasubramanian[1,*]

[1]School of Biological Sciences, Monash University, VIC 3800, Australia, [2]Independent Scholar, India and [3]Monash Bioinformatics Platform, Monash University, VIC 3800, Australia

Received December 18, 2020; Revised March 24, 2021; Editorial Decision April 19, 2021; Accepted April 28, 2021

ABSTRACT

RNA splicing, and variations in this process referred to as alternative splicing, are critical aspects of gene regulation in eukaryotes. From environmental responses in plants to being a primary link between genetic variation and disease in humans, splicing differences confer extensive phenotypic changes across diverse organisms (1–3). Regulation of splicing occurs through differential selection of splice sites in a splicing reaction, which results in variation in the abundance of isoforms and/or splicing events. However, genomic determinants that influence splice-site selection remain largely unknown. While traditional approaches for analyzing splicing rely on quantifying variant transcripts (i.e. isoforms) or splicing events (i.e. intron retention, exon skipping etc.) (4), recent approaches focus on analyzing complex/mutually exclusive splicing patterns (5–8). However, none of these approaches explicitly measure individual splice-site usage, which can provide valuable information about splice-site choice and its regulation. Here, we present a simple approach to quantify the empirical usage of individual splice sites reflecting their strength, which determines their selection in a splicing reaction. Splice-site strength/usage, as a quantitative phenotype, allows us to directly link genetic variation with usage of individual splice-sites. We demonstrate the power of this approach in defining the genomic determinants of splice-site choice through GWAS. Our pilot analysis with more than a thousand splice sites hints that sequence divergence in *cis* rather than *trans* is associated with variations in splicing among accessions of *Arabidopsis thaliana.* This approach allows deciphering principles of splicing and has broad implications from agriculture to medicine.

INTRODUCTION

During mRNA formation, certain sections of the transcribed RNA (introns) are removed with joining of the adjacent regions (exons) in a process known as splicing. Splicing is a fundamental process in eukaryotic gene regulation (9–11). Introns removed during the splicing typically harbor canonical sequences at the 5′ beginning (GU) and the 3′ end (AG), which facilitate their identification as splice sites (12–14). Selection of splice sites determines the products of splicing. Differential selection of splice sites gives rise to Alternative Splicing (AS), conferring both proteome diversity and phenotypic plasticity (9,15). Changes in splicing can modulate a range of phenotypes across diverse organisms such as sex-determination in flies, stress response in plants and genetic diseases in humans (1,3,9,15–17).

The pattern of splice site selection governs the type of mRNA transcripts produced (18,19). Splice-sites are recognized by a large protein complex, the spliceosome, which contains splicing factors that identify the splice-sites and form the splicing complex (13). Maniatis and Reed defined 'the affinity of a splice site for splicing factors and/or the ability of a splice site to participate in splicing complex formation' as splice-site strength (20). The difference in splice-site strengths determines which of the competing splice-sites participate in a splicing reaction. Therefore, if one has to understand the regulation of splicing, there is a need to measure the strength of individual splice sites. While there has been efforts to develop tools to *'predict'* the affinity of splice sites based on sequence features as proxy for splice-site strength, e.g. MaxEnt Score, COSSMO, SpliceAI (21–23), there has not been a systematic effort to explicitly *'measure'* the 'ability' of any given splice site to be utilized; the empirically observed splice-site usage, which is a direct measure of its strength.

Current bioinformatic approaches to analyze splicing involve exploration of the transcriptome from high-throughput short-read RNA-seq data (5,24). One approach is to measure the abundance of various isoforms (25,26). However, variation in isoforms can occur due to differen-

*To whom correspondence should be addressed. Tel: +61 3 99051373; Email: mb.suresh@monash.edu

tial splicing, alternative transcriptional start sites as well as alternative polyadenylation. In addition, this approach also suffers from the absence of accurate descriptions of all potential isoforms. Another approach measures the coverage of exons and introns (27,28) or the relative abundance of mutually exclusive splicing patterns. These are measured either through the lens of descriptive categories (i.e. exon skipping, intron retention, often assuming that these events cannot co-occur) (29,30) or more recently as local splicing variations (8), intron clusters (6) or splice graphs (7). We reason that regulatory decisions on splicing occur at the level of individual splice sites rather than at the level of the products of these decisions, namely transcript isoforms or splicing events. However, to date, bioinformatic tools to specifically quantify the observed strength/usage of individual splice sites have been lacking.

Here, we describe our idea of analyzing splicing from first principles, keeping individual splice-sites and their usage at the center of splicing decisions. We present SpliSER (Splice-site Strength Estimate from RNA-seq), a bioinformatic tool to derive a quantitative measure of the utilization of individual splice sites from short-read RNA-seq data, while the idea can be easily applicable to long-read sequences as well. Splice-site Strength Estimates (SSE) increase the power to analyze regulation of splicing variation across the genome at a fine-scale resolution. We show an implementation of SpliSER to carry out GWAS analysis that allows us to detect SpliSE-QTLs, using the SSE as a quantitative phenotype. As a proof-of-principle, using a pilot dataset with 1430 sites that represent genes known to undergo alterantive splicing and nonsense mediated mRNA decay (31,32), we map and show that *cis*-regulatory variation and competition between splice-sites are among key genomic determinants of variation in splicing in these genes among natural accessions of *Arabidopsis thaliana*. We also present diffSpliSER, which allows detecting differentially used splice-sites across the genome between samples, which increases the power to detect differential splicing in comparison to existing cutting-edge methods. Our strategy provides a powerful approach to decipher genomic determinants of splicing and is widely applicable across diverse organisms.

MATERIALS AND METHODS

Plant material/DNA/RNA Analyses

Seeds of the 1001 genome project accessions were obtained from European Arabidopsis Stock Centre. DNA and RNA extractions were done as described previously (33). For gene expression studies DNAse I (Roche)-treated 1μg of total RNA was used for cDNA synthesis using the First strand cDNA synthesis kit (Roche) and the resulting cDNA was diluted and used for RT-PCR experiments. Primers used in RT-PCR analysis are given in Supplementary Table S5.

Splice-site usage/strength estimation

SpliSER requires a BAM file of mapped RNA-seq reads and BED file that contains a list of splice junctions detected in the alignment (such as those produced during mapping by TopHat2 (34) or HISAT2 (35), or directly from a BAM file using Regtools (36). First, SpliSER uses the junctions BED file of each RNA-seq sample to define a list of splice sites observed in all samples; the read count of each junction is concurrently added to the α-read count for each of the two sites forming the junction. The list of 'partners' (sites with which a given site has been observed to form a junction) is also recorded for each splice site. Second, SpliSER uses Samtools (37) view command to retrieve reads whose mapping covers each nucleotide either side of the splice-site, the CIGAR string of these reads are then traversed to identify reads which map on either side of the splice site; these are counted as β_1-reads for each site. Third, SpliSER identifies reads with an intron spanning across the splice site, or otherwise showing non-usage of the site along with usage of a competitor; these are counted as β_2-reads. Thus Splice-site Strength Estimate is defined as

$$SSE = \frac{\alpha}{\alpha + \beta_1 + \beta_2}$$

In each sample, for each splice site; SpliSER filters those with too few reads (sum of α, β_1, and β_2: default 10). These parameters could be adjusted to increase the sample size and we found that decreasing the sum count to as low as 3 still provided similar signals in SpliSE-QTL analysis.

We also provide a slight variation of this approach, which may be useful in certain situations. In this variant approach, we split the reads that provide information about competitive splicing into two types based on whether the evidence is direct or indirect. Reads that provide direct evidence for non-usage are referred to as ($\beta_{2\text{-SIMPLE}}$) or reads that define competitive splicing without read coverage of the target site ($\beta_{2\text{-CRYPTIC}}$). While considering the total number of β_2 reads, we apply weightings on the $\beta_{2\text{-CRYPTIC}}$ reads. Here SSE is defined as

$$SSE = \frac{\alpha}{\alpha + \beta_1 + \beta_{2\ SIMPLE} + \sum_{x=1}^{m}\left[\left(\frac{\alpha(P_x)}{\alpha}\beta_{2\ CRYPTIC}(P_x)\right)\right]}$$

where $\beta_{2\text{-}CRYPTIC}(P_x)$ is the β_2 read counts coming from partner P_x, and m the number of partners. We believe that this version may not be required for most practical purposes. Nevertheless, given that we may not have considered all exhaustive possibilities, we provide this version as well as an option for those who may be interested.

SpliSE-QTL analysis

For the SpliSE-QTL analysis, 6853 RNA-seq samples were downloaded from SRA, effectively representing 666 accessions (PRJNA319904). Each sample was aligned to the TAIR10 reference genome using Tophat (v2.1.1; paramters –minIntronLength 20, –maxIntronLength 6000, -p 6) (34). The resulting BAM files were indexed and sorted using Samtools sort v1.7 (37). The sorted BAM files and BED files for each sample were passed to SpliSER with a minimum of 10 evidencing reads required for a site to be called, and using a table adapted from the TAIR10_genes.gff3 annotation file to identify gene boundaries. The Splice-site Strength Estimate of each site in each of the 50 NMD target genes was quantified. The SSE phenotypes were further filtered using the following criteria: (i) an accession was only considered for further analysis for a given site if it had three or more individual RNA-seq samples underlying its average splicing efficiency value, and (ii) a site was

only considered for further analysis if it contained 100 or more such accessions. We calculated the broad-sense heritability (H^2) and variance (σ^2) of each splice site. Heritability was calculated as the sum of squares between genotypes divided by the total sum of squares (SS_group/SS_total) extracted from the results of a one-way ANOVA with Genotype as a factor. GWAS experiments were performed using the easyGWAS (38) and/or GWAPP (39) web portal using the EMMAX/AMM algorithm with a minimum minor allele frequency of 5%, with no transformation of phenotypes. To assess the relationship between heritability/variance and the ability to detect GWAS signals, we initially carried out GWAS with all 1430 phenotypes containing 100 or more accessions. Our findings suggested that most of the signals were detected among splice-sites with higher levels of heritable phenotypic variation. When variance is low, it resulted in spurious signals throughout the genome. Therefore, each of the mannhattan plots were individually inspected, before deciding on whether a GWAS signal that shows statistical significance is trustable. In subsequent iterations, only sites that were in the upper quartile of heritability and variance were taken for analysis. In total 186 phenotypes were analyzed through GWAS that ultimately resulted in 47 associations for 19 of the 50 analyzed genes (Figure 4, Supplementary Figure S3, Supplementary Table S1). All of these 47 were tested both through EMMAX algorithm or through AMM algorithm and both methods provided consistent signals. In instances where multiple SNPs that were in LD were giving same *P*-value of highest association, the closest SNP to the splice-site is presented in the Table.

eQTL analysis

eQTL analysis was carried out using published expression levels (40) for the RNA-Seq data using the same panel of accessions that contributed to the splicing-QTL for each of the phenotypes, using the same GWAS options described above. The data were then summarized at the gene level and if any one of the panel of accessions gave an eQTL signal, it was considered to be a positive overlapping QTL. This analysis resulted in two genes (FLM and At4g35875) having overlapping eQTL and SpliSE-QTL signals.

diffSpliSER analysis

Taking the output of SpliSER, we remove all sites containing an NA in any sample generated by insufficient read coverage; then filter all sites whose mean SSE (average of all samples) is <0.05 or >0.95, blind to experimental grouping. We utilized the EdgeR package (41), testing for significant changes in splice site strength using a generalized linear model (glmLRT() function, default parameters) with a contrast corresponding to the difference between the α and β ($\beta_1 + \beta_2$) read counts, between two samples [i.e. (alpha.group1-beta.group1)-(alpha.group2-beta.group2)]. Differentially Spliced sites were called as those with an FDR-corrected *P*-value <0.05, and an absolute change in averaged SSE ≥ 0.1 between conditions

Yan *et al.* comparison

The six RNA-seq samples described in Yan *et al.* (42) were aligned to the TAIR10 reference genome using TopHat2 v2.0.10 (34) (parameters -i 20 -I 6000 -g 10 -r 0 –mate-std-dev 50 –coverage-search, and indexed using Samtools v1.2 (37). Each resulting BAM and junction BED file were processed with SpliSER (command: process, parameters –m 6000) using an annotation file derived from the TAIR10 genes.gff3 file. The resulting SpliSER.bed files were then combined (command: combine; parameters -1 Chr1) and output (command: output; parameters –t diffSpliSER, -r 10). Differentially spliced sites were detected using the diffSpliSER R script. For comparison with MAJIQ: BAM files were processed with MAJIQ v2.1 (8). For each gene, we assessed probability of deltaPSI being above 0.1, taking all LSVs with probability >0.95 to be evidence of differential splicing. For Salmon analysis: FASTA files were processed using the Salmon v1.4.0 quant command, using a FASTA file containing TAIR10 cDNAs.

Simulated RNA-seq data

Simulated 100bp paired-end reads were generated using Polyester v1.2.6 (43). First, we identified genes with multiple annotated isoforms (*Arabidopsis:* Araport11 release 49, Human: GRCh38 release 102). To exclude isoforms that differ due to non-splicing differences (e.g. alternative transcriptional start sites and/or alternative polyadenylation) we simulated reads for transcripts that began and ended at the same position. There were instances where multiple distinct sets of transcripts met these criteria. For these genes with multiple sets of such transcripts we took only the first identified set. Reads were generated for each isoform of each gene using the formula:

$$Isoform\ Reads = \frac{x\,l}{r\,n}\ to\ nearest\ integer$$

where x in the desired isoform coverage (we took 20 as low, and 50 as high), l is the length of the isoform cDNA, r is the length of simulated reads (200 for paired-end 100 bp reads), and n is the number of isoforms for this gene. This was designed to simulate each isoform being equally transcribed, while maintaining similar coverage across genes. We selected a subset of genes at random to be differentially spliced which resulted in 1000/3379 genes in *Arabidopsis* and 500/2023 genes in Humans. We further filtered human data for genes with multiple annotations which resulted in 433/1780 genes. We simulated differential splicing by increasing the expression of isoform (we took 1.2-fold as low, 1.5-fold as moderate, and 1.7-fold as high). To ensure that there was no bias due to changes in gene expression, an equal decrease in isoform expression was evenly distributed among the remaining isoforms for each gene. For *Arabidopsis,* reads were aligned to the TAIR10 reference genome using STAR v2.5.2b (44) (–outFilterMultimapNmax 1, minimum intron size 20, maximum intron size 6000). For Human data, reads were aligned to the GRCh38 reference genome using STAR v2.5.2b (44) (–outFilterMultimapNmax 1, minimum intron size 20, maximum intron size 100 000). BAM files were then processed with the Regtools (36) *junction extract* command

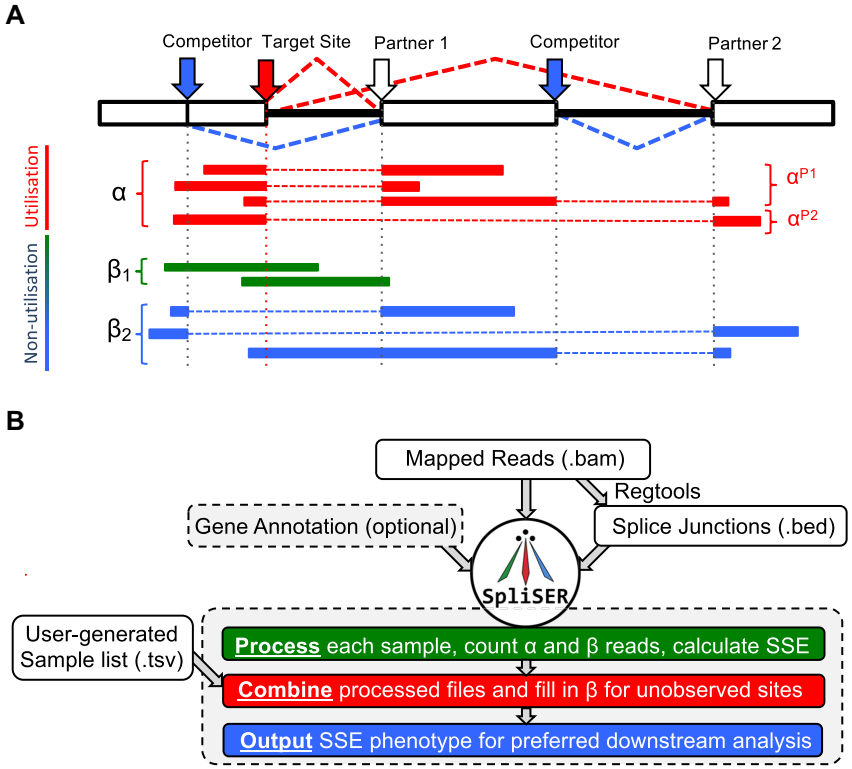

Figure 1. Overview of Splice-site Strength Estimation and SpliSER. (**A**) For a given target splice site (large red arrow) Splice-site Strength Estimate (SSE) is calculated as the proportion of reads showing of reads showing utilization of the splice site (α-reads; red rectangles) versus those showing non-utilization (β_1- and β_2-reads; green and blue rectangles). α-reads (red) are split-reads with a gap in mapping which terminates at the target splice site, showing utilization. The other splice site in this read is then considered a Partner of the target site. β_1 reads (green) map either side of the target splice site but show no gap in mapping at this position, showing the target site has not been utilized. β_2-reads (blue) show known Partners of the given site utilizing another splice site, which has therefore outcompeted the target site. β_2 reads provide direct evidence of non-utilization of the target site. (**B**) Computational workflow of SpliSER. First, mapped reads are processed with Regtools to generate a bed file containing each detected splice junction. These two files are then provided as input to SpliSER, along with an optional annotation file defining gene boundaries.

to produce a splice junction BED file. For SpliSER analysis: these two sets of files were passed to SpliSER v0.1.5 and taken through the diffSpliSER pipeline. The highest $-\log_{10}$ adjusted *P*-value (FDR) was taken as the score for each gene assessed. For Salmon analysis: FASTA files were processed using the Salmon v1.4.0 quant command, using the same FASTA file containing TAIR10 cDNAs from which the reads were originally simulated. Differential transcript utilisation was identified using the DRIMseq pipeline (45). For MAJIQ analysis: BAM files were processed with MAJIQ v2.1 (8). For each gene, the score was taken to be the highest probability of deltaPSI being above 0.2 among all LSVs in that gene, extracted from the TSV file produced by deltaPSI Analysis. For rMATS analysis: BAM files were processed with rMATS turbo v4.1.0 (29). We took the scores for each gene to be the highest $-\log_{10}$ pvalue (FDR) observed across all splicing events, for this we took the JCEC (reads that span splicing junctions and reads on target) files. We could not obtain *P*-values for 61% of the simulated genes for the Arabidopsis data. We failed to correct this even with substantial modifications of the gtf file. We did not encounter this problem with rMATS in the human data. Area under the curve (AUC) values were calculated using the *precrec* package in R; for the purposes of this calculation, genes that were not assessed by a tool were assumed to have been called non-differentially spliced.

RESULTS AND DISCUSSION

We developed the idea of quantifying the usage of a given splice site earlier (46), which we now refer to as Splice-site Strength Estimate (SSE). For any splice site (just for example, let us consider it a splice donor site), there are three types of reads that provide information about its usage (Figure 1A). First there are split reads that map perfectly with a gap between splice site and its partner sites. We call these reads as α-reads that provide evidence for the use of that splice site (Figure 1A), independent of its partners (i.e. acceptor sites). Second, there are reads that cover the exon-intron junction (β_1 reads), which provide evidence of non-usage of the target site (Figure 1A). We then use the historical concept of 'competing splice sites' (20,47) to define additional reads that provide evidence of non-usage. If two donor sites partner with the same acceptor site, obviously both events cannot occur in the same transcript and thus the evidence for utilization of one site becomes the evidence for the non-utilization of the other site. Reads that define competitive splicing and provide direct evidence for non-usage of the target site (β_2 reads). Thus, SSE is derived as the ratio of the evidence of utilization over the evidence of total possible utilization (see Materials and Methods).

We further developed this concept into a bioinformatics pipeline, which we call as SpliSER (Figure 1B). SpliSER

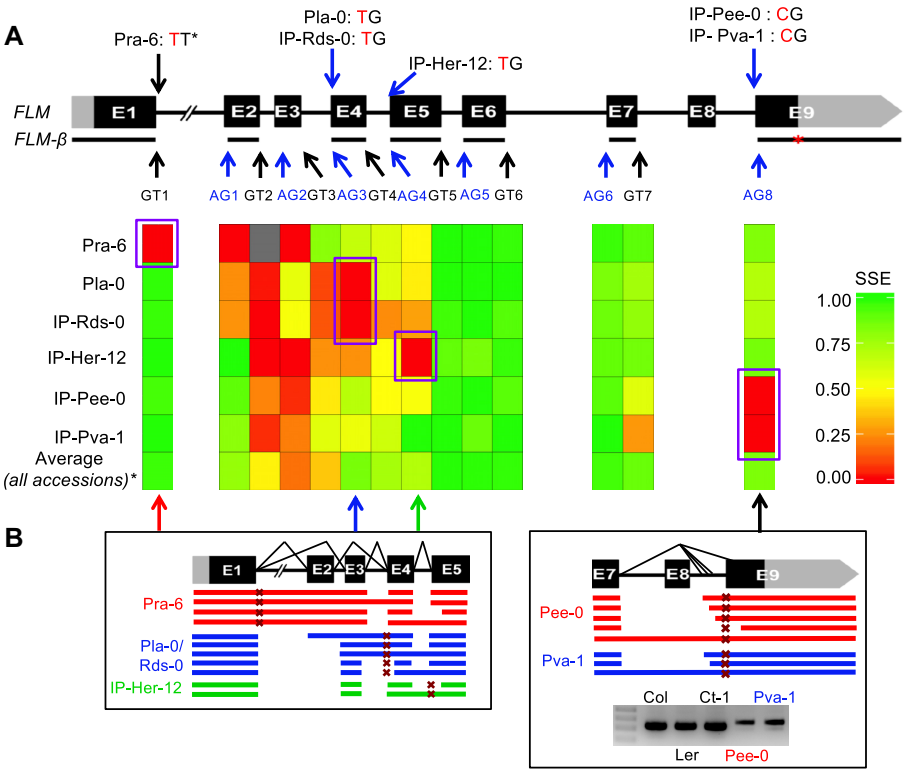

Figure 2. SpliSER estimates SSEs accurately. (**A**) Gene model of FLM locus with splice site mutations across accessions. Heatmap shows SSEs for the splice sites at the FLM locus in various genotypes, purple boxes indicate slice site mutations. (**B**) The splice variants observed in the RNA-seq data are shown in the schematic. Dark red crosses indicate position of the splice site mutations. Experimental analysis with accessions harboring splice site mutation at AG8 shows change in splicing pattern.

is annotation-independent yet allows the use of annotation when available. SpliSER uses the list of junctions to identify the splice sites from the RNA-Seq data, it then leverages the alignment file to assess all reads overlapping each site; finally, it uses all of these metrics, applies filters (see detailed methods) and produces the SSE for all splice sites detected in the RNA-seq data. In any given experiment, the user will *process* each RNA-seq sample, *combine* the samples together (to ensure that each site has a value for each sample), then *output* this information in a format ready for downstream analysis.

To assess the accuracy of the SSE, we exploited natural variation in *Arabidopsis thaliana*. We reasoned that if our estimates are accurate and specific, for genotypes with mutations in splice sites, the SSE would be close to 0 for that site in that genotype and will differ from other sites in the same gene, and with others genotypes for the same site. Exploiting the 1001 genome project data (40,48), we identified genotypes with splice-site mutations at *FLOWERING LOCUS M (FLM)*, a gene which is known to undergo alternative splicing (49). After confirming the mutations by Sanger sequencing, we downloaded the RNA-seq data for these accessions from the 1001 genome project (40) and ran SpliSER across the *FLM* locus. Consistent with our predictions, we found the SSE of only mutated sites to be close to 0 unlike other sites known to be spliced with full efficiency (e.g. GT5, GT6, AG5, AG6). In addition, the effect at a particular site was exclusive to accessions with mutations, confirming the specificity of our estimates (Figure 2A). RT-PCR experiments confirmed alternative splicing, consistent with our estimates (Figure 2B). Although total *FLM* expression level is substantially reduced in some of these accessions (Supplementary Figure S1), we were able to estimate SSE accurately, which indicates that we could pick up splicing differences despite differences in gene expression.

SSE is a quantitative measure that can be used as a phenotype in GWAS to identify potential regulatory variation. To explore this possibility, we downloaded 6583 RNA-Seq datasets representing 728 accessions from the 1001 genome project (40). As a proof-of-principle pilot-study, here we carried out SpliSE-QTL analysis for all splice sites across 50 genes that are known to undergo alternative splicing and nonsense-mediated mRNA decay (31,32). SpliSE-QTL analysis detected ∼2000 splice sites across these genes. As expected, majority of the splice sites displayed minimal variation in SSE among accessions. However, there were sites with higher variability (Supplementary Figure S2). To assess the genetic contribution to this variation, we calculated heritability of the SSEs. The heritability varied substantially (Supplementary Figure S2) even for sites within a gene (Supplementary Table S3), which indicated that there are splice sites with genetically controlled variability in SSEs. The heritability/variance pattern was similar for both donor and acceptor sites. We focused on splice sites that were: (i) in the upper quartile for heritability and variance

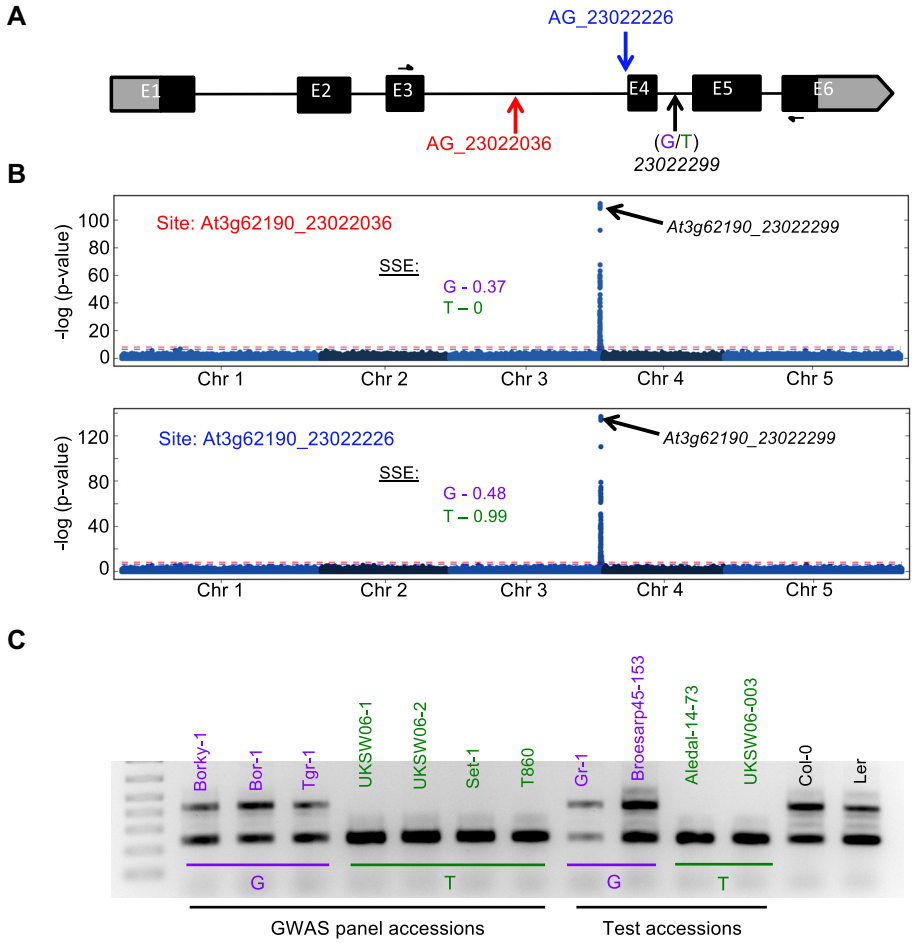

Figure 3. SpliSE-QTL analysis detects experimentally viable associations for splice-site strength. (**A**) Gene model for At3G62190. Positions of two competing splice acceptor sites and the associated SNP are shown. (**B**) GWAS analysis for SSE of both competing splice-sites identifies At3g62190_23022299 as the highest associated SNP in GWAS. The allelic effects on both sites are shown; in the presence of the alternate T allele, the relative strength of the site 23022036 (red) drops to zero. (**C**) Experimental verification of the SSE differences associated with 23022299-SNP. Common allele is shown in purple and the alternative allele is shown in green. In the presence of the alternate allele, the top band (representing utilization of the site 23022036) is absent. Col and Ler harbor common alleles at this SNP.

and (ii) with minimum 100 accessions. This filtering identified a total of 186 sites that were taken for GWAS.

We identified significant, reliable associations for 44 sites that are spread over 19 of the 50 genes (∼38%) analyzed in this pilot-study (Supplementary Table S1; Supplementary Figure S3). To experimentally test the reliability of the observed associations, we first analyzed splicing at *At3g62190* in which two sites provided strong GWAS signals (Figure 3A and B). SSE of two splice acceptor sites (canonical 23022226 & alternative 23022036) mapped to the same SNP (23022299). Common allele (G) at this SNP (23022299) in the neighboring downstream intron was associated with reduced strength of the canonical site, and the minor allele (T) substantially increased the efficiency at the canonical site effectively competing out the alternative site (Figure 3B). Thus, accessions that differed at this associated SNP displayed distinct splicing patterns (Figure 3C). We reasoned that if the identified associations are genuine, we should observe predictable splicing patterns based on the genotype of the associated SNP, even in accessions that were not part of the GWAS panel. Consistent with our hypothesis, splicing patterns conformed with predictions (Figure 3C). We carried out similar analysis with other loci and found them to conform to predictions (Supplementary Figures S4–S6), which provided experimental support for the detected associations.

Given that these genes are known to undergo AS-NMD (31,32), changes in splice-site utilization might lead to transcripts that are degraded. Analysis of RNA-seq in general can only capture and quantify transcripts that remain in the cell. Nevertheless, any effect on transcript stability including AS-NMD would present as a corresponding decrease in gene expression in the same RNA-seq data. However, if changes in utilization of specific splice-sites is unrelated to NMD, then it would represent a change only in splicing and not in gene expression. To assess this, we carried out eQTL analysis with the same set of accessions used to detect each SpliSE-QTL. We failed to detect overlapping GWAS signals for expression in most cases, which indicates that majority of the detected SpliSE-QTLs are specific to splicing rather than gene expression (Supplementary Table S2). We did however observe overlapping eQTLs and SpliSE-QTLs

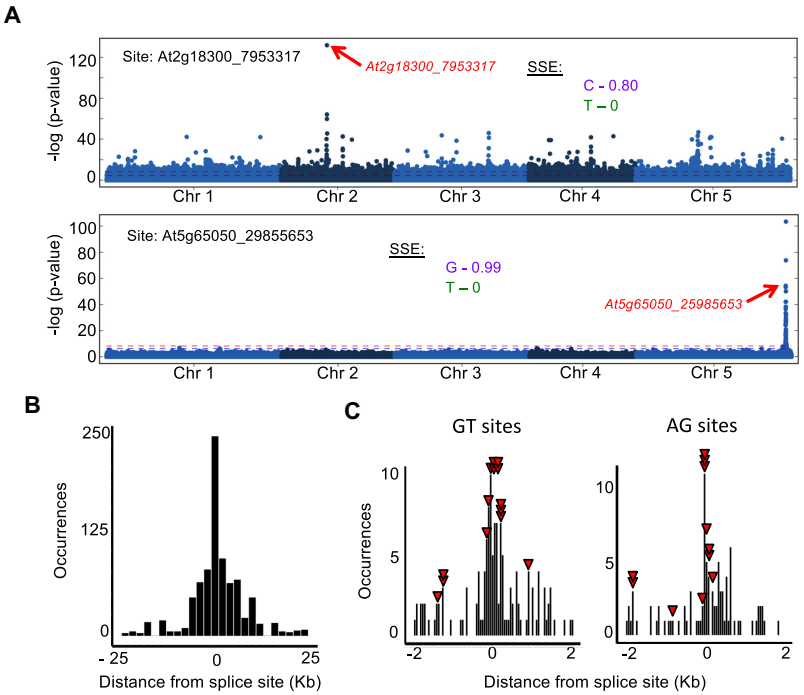

Figure 4. SpliSE-QTL can detect causal variants and *cis*-regulatory variation is among the primary determinants of variation in splicing. (**A**) Manhattan plots for the splice sites in which the GWAS identified mutations in the splice-site itself (At2g18300_7953317, At5g65050_25985653). See Supplementary Figure S3 for additional Manhattan Plots. The splice-site strength conferred by the common (purple) and the alternate (green) alleles are shown. (**B**) Distribution of distances of Top 25 highest associated SNPs for the splice-sites for which associations were detected on the same chromosome. (**C**) Distribution of the distances of the Top 10 highest associated SNPs that fell within 2 kb of the splice site. Red arrows indicate the positions of Top SNPs across associations.

for two loci (*FLM (At1g770880) & At4g357885*)). These results are consistent with the finding that *FLM* levels are regulated via AS-NMD to regulate flowering and natural variation at *FLM* modulates splice-site choice resulting in changes in expression levels (46,50). While these two examples may potentially reflect an association with NMD, we also cannot rule out the possibility that splicing and expression are correlated through some other mechanism at these loci.

Having confirmed the associations, we considered whether our analysis has the ability to pick up causal mutations. In the simplest case, a mutation of the canonical GT or AG splice site sequence should have a direct impact on the ability of the splice site to recruit spliceosomal machinery and thus reduce the splice-site strength/usage. We analyzed whether any of the detected associations actually mapped to the splice-site sequence itself. Variation in SSE for two sites (At2g18300_7953316/17 and At5g65050_25985653/54) mapped to the splice sites themselves (Figure 4a). While this is consistent with the idea that mutations at the splice site would weaken the strength of a splice site, it also argues for the potential of SpliSE-QTL analysis to identify causal mutations underlying variation in splicing.

Next, we analyzed the pattern of detected associations. We noticed that most of the GWAS signals (43 out of 47) were on the same chromosome as that of the splice site (Supplementary Table S1). To assess patterns, we compiled the distances of the associated SNPs (Top 25 SNPs for each SSE phenotype) from the splice site (Supplementary Table S1, Figure 4B). Distances clustered within 2Kb of the splice site (Figure 4B). Most of the top-associated SNPs (25 out of 43) fell within 2 kb either side of the splice site (Figure 4C) and was consistent between both splice donor and acceptor sites. In fact, for 26 out 43 associations, we detected an associated SNP within 100 bp from the splice site (Supplementary Table S1), all together suggestive of a strong *cis effect*. We also noted several instances of common associated SNPs for multiple splice-sites (Supplementary Table S1), which suggested the influence of one splice-site over the other through competition. These findings indicate that in this dataset, *cis* rather than *trans*-sequence variation and competition between splice-sites are among the primary drivers of variation in splicing.

A recent study (51) utilized the same RNA-Seq data from the 1001 genome project and carried out isoform-based splicing QTL analysis with sQTLseekeR (52) based UlfasQTL (53). Given that this study (51) used the same data as in our case, it provided an opportunity to directly compare the methods for analyzing splicing-QTLs. Khokhar *et al.* concluded from their analysis that majority of splicing variation in Arabidopsis is due to *trans* rather than *cis* QTL (51). Their finding contradicted earlier genetic studies (54). Our findings support the assertions from genetic studies (54), which show that *cis*-regulatory variation as a primary determinant of splicing differences. Therefore, we compared the GWAS results obtained in this study with ours for overlapping set of genes (Supplementary Table S4). We observed that the isoform-based approach failed to detect even a single SNP that we report for the same set of

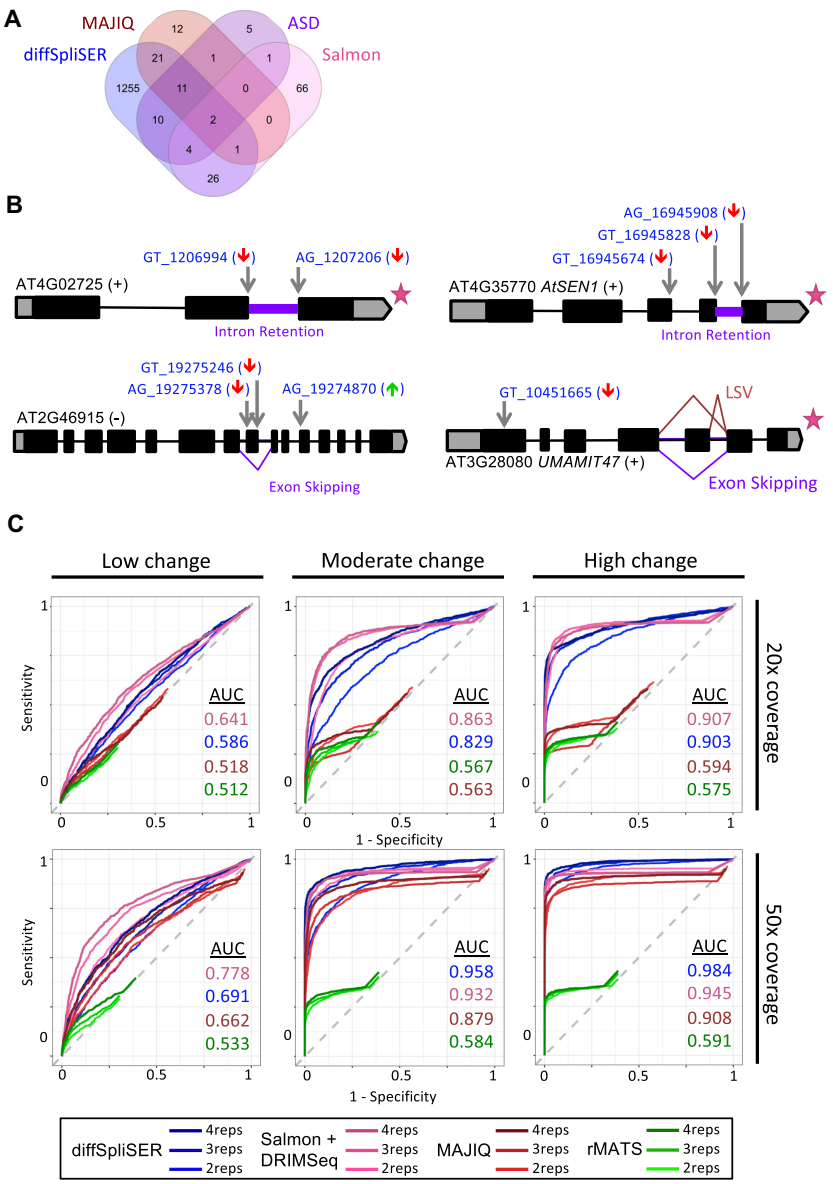

Figure 5. Performance of diffSpliSER. (**A**) The overlap of genes identified as differentially spliced by ASD, MAJIQ, Salmon and diffSpliSER from RNA-seq of the Yan *et al*. study (**B**) DiffSpliSER analysis identified significant changes in splice sites (blue text, grey arrows) underlying three of the four experimentally verified splicing events (purple) reported in the Yan *et al*. study. By contrast MAJIQ identified differential splicing at one of the four events (burgundy lines). Pink stars indicate Salmon detected differential transcript usage for that gene. Bright red and green arrows after the site label indicate a significant decrease or increase in SSE, respectively. (**C**) Receiver operating characteristic curves show performance of diffSpliSER, MAJIQ and rMATS in detecting differential splicing spiked into 1000 out of 3474 genes in simulated RNAseq experiments. Simulated experiments had varying read coverage (top to bottom) and simulated splicing changes had varying degrees of magnitude per experiment (left to right). Curves are truncated due to some genes not being assessed by each tool, either filtered or not recognized. Area Under the Curve (AUC) values indicate performance of each tool.

genes (Supplementary Table S4). In addition, there was substantial noise in Khokhar *et al*'s GWAS study such that associations were found with SNPs scattered around the entire genome, while we detected specific signals (Supplementary Table S4, Supplementary Figure S7). Thus, in addition to reducing noise, SpliSE-QTL analysis also provided specific associations with splice-sites.

Having confirmed the utility of SpliSER, we also implemented diffSpliSER, a method that would allow detection of differentially spliced sites across the genome between samples. To assess the practical efficacy of this approach, we applied SpliSER on a previously published RNA-Seq data on quintuple mutants of the *sc35-scl* in *Arabidopsis thaliana* (42). In this comparison, Yan *et al*., utilized ASD to detect differential splicing and found 34 genes (FDR<0.05) (42). We applied MAJIQ v2.1 (8), Salmon v1.4.0 (26) and SpliSER v0.1.3 on the same data; representing a suite of approaches to differential splicing detection (Figure 5A). MAJIQ detected 48 genes of which 14 overlapped with ASD (Hypergeometric probability *P*-value = 2.15e–31). Salmon

Table 1. Comparison of strengths and limitations of assessed splicing measurement tools

	Leafcutter	Salmon	rMATS	MAJIQ	SpliSER
Differential splicing – isoforms	No	Yes	No	No	**No**
Differential splicing – patterns	Yes	Yes[b]	Yes	Yes	**No**
Differential splicing – splice sites	No	No	No	No	**Yes**
Is it annotation independent?	Yes	Yes[a]	No	No	**Yes**
Is intron retention considered?	No	Yes	Yes	Yes	**Yes**
GWAS for splice-site usage	No	No	No	No	**Yes**

[a]Requires a reference set of transcript sequences.
[b]There are supplementary tools available to derive event information from transcript quantification.

detected 100 genes, of which seven overlapped with ASD (P = 2.46e–11) and 3 with MAJIQ (P = 0.0006). However, Salmon's differential isoform detection will also include isoform differences caused by differential transcriptional start and/or alternative polyadenylation. In contrast SpliSER detected a total of 1330 genes of which 27 overlapped with ASD (P = 2.73e30), 35 overlapped with MAJIQ (P = 1.30e–36) and 33 overlapped with Salmon (P = 8.71e–20). Thus, SpliSER detected substantially higher number of genes to be differentially spliced in comparison with all programs. Overall, only two genes were detected by all four programs, but SpliSER captured most of the genes (27/34 of ASD; 35/48 of MAJIQ and 33/100 of Salmon). Yan *et al*. also experimentally analyzed four genes falling below the FDR <0.05 threshold, and SpliSER identified underlying splice-sites in three of the experimentally analyzed genes (Figure 5B). Thus, diffSpliSER analysis can provide splice-site based details of real differential splicing.

Given that the number of genes picked up with SpliSER was vastly higher than comparable programs, we considered the specificity and sensitivity using unbiased simulated RNA-seq data in comparison to popular tools MAJIQ (8) and rMATS v.4.1.0 (29). We also included Salmon (26), which would provide a positive control given the perfect match between source isoforms and annotated isoforms of the simulated data (Figure 5C, Supplementary Table S6). We did not include Leafcutter (6) since it does not model intron retention, which is abundant in the Arabidopsis (55), and also observed in other organisms including humans (56). In our simulations, diffSpliSER performed as well as other contemporary approaches to differential splicing analysis, which indicates that the higher number of genes that we detected to be differentially spliced, could not be due to reduced sensitivity/specificity and most of these potentially reflect true splicing differences. In addition, diffSpliSER provides quantitative comparisons at the splice-site level between samples (Table 1).

Although SpliSER is developed using Arabidospis system, there is nothing inherent in our approach that is specific to a species. We were able to run differential splicing across a range of plants (e.g. *Marchantia polymorpha*, *Capsella rubella)* and animals (Drosophila and mouse). To assess the performance of SpliSER in human data, we carried out similar simulations using human transcriptome (see details in methods). Similar to the Arabidopsis simulations, in human data as well SpliSER performed pretty similar to other tools (Supplementary Figure S8, Supplementary Table S6).

To the best of our knowledge, SpliSE-QTL is the only approach that allows detecting genetic variation that is associated with changes in splicing of a specific individual splice site. Our findings indicate that SpliSER provide reproducible quantification of SSE and SpliSE-QTL analysis has the potential to detect genomic determinants of variation in splicing. While our analysis is primarily based on short read RNA seq data, in theory the same principles could be applied to long-read data with minor changes. Further, we believe that even with arbitrarily long reads; measures of splice site utilization, such as we have presented here, will remain essential for analysis of splicing regulation. Our findings in Arabidopsis suggest that *cis*-regulatory changes and competition between splice-sites based on their splice-site strength are among the key determinants of variation in splicing. While more work is needed to explain the mechanisms for each of these associations, we have demonstrated that these associated SNPs can be used as markers for tagging-splicing patterns. Splice-tagging SNPs would be of great use as markers having wide-ranging implications from agriculture to human disease. Future GWAS studies across genomes would unravel both the complexity and regulators of variation in splicing in an unprecedented fine scale.

DATA AVAILABILITY

SpliSER software, and scripts for statistical testing are available in the GitHub repository https://github.com/CraigIDent/SpliSER.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

We thank the European and North American Arabidopsis Stock Centers, and Alex Fournier-Level for the seeds. We thank Arun Konagurthu for advice on data processing. We thank Ya-Long Guo, Institute of Botany, Chinese Academy of Sciences, Beijing for critical comments and discussions on the manuscript.

FUNDING

Australian Government's Research Training Program (RTP) fellowship (to C.D.); Australian Research Council – Future Fellowship [FT190100403 to SrS]; ARC-Discovery Project [DP190101479]; Chinese Academy of Sciences President's International Fellowship for Visiting Scientists (to S.B.).
Conflict of interest statement. None declared.

REFERENCES

1. 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. *Science*, **352**, 600–604.
2. Bush,S.J., Chen,L., Tovar-Corona,J.M. and Urrutia,A.O. (2017) Alternative splicing and the evolution of phenotypic novelty. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **372**, 20150474.
3. Laloum,T., Martin,G. and Duque,P. (2018) Alternative splicing control of abiotic stress responses. *Trends Plant Sci.*, **23**, 140–150.
4. Song,Q.A., Catlin,N.S., Brad Barbazuk,W. and Li,S. (2019) Computational analysis of alternative splicing in plant genomes. *Gene*, **685**, 186–195.
5. Liu,R., Loraine,A.E. and Dickerson,J.A. (2014) Comparisons of computational methods for differential alternative splicing detection using RNA-seq in plant systems. *BMC Bioinformatics*, **15**, 364.
6. Li,Y.I., Knowles,D.A., Humphrey,J., Barbeira,A.N., Dickinson,S.P., Im,H.K. and Pritchard,J.K. (2018) Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.*, **50**, 151–158.
7. Sterne-Weiler,T., Weatheritt,R.J., Best,A.J., Ha,K.C.H. and Blencowe,B.J. (2018) Efficient and accurate quantitative profiling of alternative splicing patterns of any complexity on a laptop. *Mol. Cell*, **72**, 187–200.
8. 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**, e11752.
9. Kalsotra,A. and Cooper,T.A. (2011) Functional consequences of developmentally regulated alternative splicing. *Nat. Rev. Genet.*, **12**, 715–729.
10. Reddy,A.S., Marquez,Y., Kalyna,M. and Barta,A. (2013) Complexity of the alternative splicing landscape in plants. *Plant Cell*, **25**, 3657–3683.
11. Szakonyi,D. and Duque,P. (2018) Alternative splicing as a regulator of early plant development. *Front Plant Sci.*, **9**, 1174.
12. Herzel,L., Ottoz,D.S.M., Alpert,T. and Neugebauer,K.M. (2017) Splicing and transcription touch base: co-transcriptional spliceosome assembly and function. *Nat. Rev. Mol. Cell Biol.*, **18**, 637–650.
13. Matera,A.G. and Wang,Z. (2014) A day in the life of the spliceosome. *Nat. Rev. Mol. Cell Biol.*, **15**, 108–121.
14. Naftelberg,S., Schor,I.E., Ast,G. and Kornblihtt,A.R. (2015) Regulation of alternative splicing through coupling with transcription and chromatin structure. *Annu. Rev. Biochem.*, **84**, 165–198.
15. Baralle,F.E. and Giudice,J. (2017) Alternative splicing as a regulator of development and tissue identity. *Nat. Rev. Mol. Cell Biol.*, **18**, 437–451.
16. Salz,H.K. (2011) Sex determination in insects: a binary decision based on alternative splicing. *Curr. Opin. Genet. Dev.*, **21**, 395–400.
17. Xu,X., Yang,D., Ding,J.H., Wang,W., Chu,P.H., Dalton,N.D., Wang,H.Y., Bermingham,J.R. Jr, Ye,Z., Liu,F. *et al.* (2005) ASF/SF2-regulated CaMKIIdelta alternative splicing temporally reprograms excitation-contraction coupling in cardiac muscle. *Cell*, **120**, 59–72.
18. Andreadis,A., Gallego,M.E. and Nadal-Ginard,B. (1987) Generation of protein isoform diversity by alternative splicing: mechanistic and biological implications. *Annu. Rev. Cell Biol.*, **3**, 207–242.
19. Breitbart,R.E., Andreadis,A. and Nadal-Ginard,B. (1987) Alternative splicing: a ubiquitous mechanism for the generation of multiple protein isoforms from single genes. *Annu. Rev. Biochem.*, **56**, 467–495.
20. Reed,R. and Maniatis,T. (1986) A role for exon sequences and splice-site proximity in splice-site selection. *Cell*, **46**, 681–690.
21. Bretschneider,H., Gandhi,S., Deshwar,A.G., Zuberi,K. and Frey,B.J. (2018) COSSMO: predicting competitive alternative splice site selection using deep learning. *Bioinformatics*, **34**, i429–i437.
22. Yeo,G. and Burge,C.B. (2004) Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.*, **11**, 377–394.
23. Jaganathan,K., Kyriazopoulou Panagiotopoulou,S., McRae,J.F., Darbandi,S.F., Knowles,D., Li,Y.I., Kosmicki,J.A., Arbelaez,J., Cui,W., Schwartz,G.B. *et al.* (2019) Predicting splicing from primary sequence with deep learning. *Cell*, **176**, 535–548.
24. Alamancos,G.P., Agirre,E. and Eyras,E. (2014) Methods to study splicing from high-throughput RNA sequencing data. *Methods Mol. Biol.*, **1126**, 357–397.
25. Bray,N.L., Pimentel,H., Melsted,P. and Pachter,L. (2016) Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, **34**, 525–527.
26. Patro,R., Duggal,G., Love,M.I., Irizarry,R.A. and Kingsford,C. (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods*, **14**, 417–419.
27. Anders,S., Reyes,A. and Huber,W. (2012) Detecting differential usage of exons from RNA-seq data. *Genome Res.*, **22**, 2008–2017.
28. Hartley,S.W. and Mullikin,J.C. (2016) Detection and visualization of differential splicing in RNA-Seq data with JunctionSeq. *Nucleic Acids Res.*, **44**, e127.
29. Shen,S., Park,J.W., Lu,Z.X., Lin,L., Henry,M.D., Wu,Y.N., Zhou,Q. and Xing,Y. (2014) rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, E5593–E5601.
30. Trincado,J.L., Entizne,J.C., Hysenaj,G., Singh,B., Skalic,M., Elliott,D.J. and Eyras,E. (2018) SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biol.*, **19**, 40.
31. Drechsel,G., Kahles,A., Kesarwani,A.K., Stauffer,E., Behr,J., Drewe,P., Ratsch,G. and Wachter,A. (2013) Nonsense-mediated decay of alternative precursor mRNA splicing variants is a major determinant of the Arabidopsis steady state transcriptome. *Plant Cell*, **25**, 3726–3742.
32. Kalyna,M., Simpson,C.G., Syed,N.H., Lewandowska,D., Marquez,Y., Kusenda,B., Marshall,J., Fuller,J., Cardle,L., McNicol,J. *et al.* (2012) Alternative splicing and nonsense-mediated decay modulate expression of important regulatory genes in Arabidopsis. *Nucleic Acids Res.*, **40**, 2454–2469.
33. Zhu,W., Ausin,I., Seleznev,A., Mendez-Vigo,B., Pico,F.X., Sureshkumar,S., Sundaramoorthi,V., Bulach,D., Powell,D., Seemann,T. *et al.* (2015) Natural variation identifies ICARUS1, a universal gene required for cell proliferation and growth at high temperatures in arabidopsis thaliana. *PLoS Genet.*, **11**, e1005085.
34. Kim,D., Pertea,G., Trapnell,C., Pimentel,H., Kelley,R. and Salzberg,S.L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, **14**, R36.
35. Kim,D., Paggi,J.M., Park,C., Bennett,C. and Salzberg,S.L. (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.*, **37**, 907–915.
36. Feng,Y.Y., Ramu,A., Cotto,K.C., Skidmore,Z.L., Kuniaski,J., Conrad,D.F., Lin,Y., Chapman,W.C., Uppaluri,R., Govindan,R. *et al.* (2018) RegTools:Integrated analysis of genomic and transcriptomic data for discovery of splicing variants in cancer. bioRxiv doi: https://doi.org/10.1101/436634, 05 January 2021, preprint: not peer reviewed.
37. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and Genome Project Data Processing, S. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
38. Grimm,D.G., Roqueiro,D., Salome,P.A., Kleeberger,S., Greshake,B., Zhu,W., Liu,C., Lippert,C., Stegle,O., Scholkopf,B. *et al.* (2017) easyGWAS: a cloud-based platform for comparing the results of genome-wide association studies. *Plant Cell*, **29**, 5–19.
39. Seren,U., Vilhjalmsson,B.J., Horton,M.W., Meng,D., Forai,P., Huang,Y.S., Long,Q., Segura,V. and Nordborg,M. (2012) GWAPP: a web application for genome-wide association mapping in Arabidopsis. *Plant Cell*, **24**, 4793–4805.
40. Kawakatsu,T., Huang,S.C., Jupe,F., Sasaki,E., Schmitz,R.J., Urich,M.A., Castanon,R., Nery,J.R., Barragan,C., He,Y. *et al.* (2016) Epigenomic diversity in a global collection of arabidopsis thaliana accessions. *Cell*, **166**, 492–505.
41. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
42. Yan,Q., Xia,X., Sun,Z. and Fang,Y. (2017) Depletion of Arabidopsis SC35 and SC35-like serine/arginine-rich proteins affects the transcription and splicing of a subset of genes. *PLoS Genet.*, **13**, e1006663.
43. Frazee,A.C., Jaffe,A.E., Langmead,B. and Leek,J.T. (2015) Polyester: simulating RNA-seq datasets with differential transcript expression. *Bioinformatics*, **31**, 2778–2784.

44. 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**, 15–21.
45. Nowicka,M. and Robinson,M.D. (2016) DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics. *F1000Res*, **5**, 1356.
46. Sureshkumar,S., Dent,C., Seleznev,A., Tasset,C. and Balasubramanian,S. (2016) Nonsense-mediated mRNA decay modulates FLM-dependent thermosensory flowering response in Arabidopsis. *Nat. Plants*, **2**, 16055.
47. Eperon,L.P., Estibeiro,J.P. and Eperon,I.C. (1986) The role of nucleotide sequences in splice site selection in eukaryotic pre-messenger RNA. *Nature*, **324**, 280–282.
48. The 1001 Genomes Consortium. (2016) 1,135 Genomes reveal the global pattern of polymorphism in Arabidopsis thaliana. *Cell*, **166**, 481–491.
49. Scortecci,K.C., Michaels,S.D. and Amasino,R.M. (2001) Identification of a MADS-box gene, FLOWERING LOCUS M, that represses flowering. *Plant J.*, **26**, 229–236.
50. Hanemian,M., Vasseur,F., Marchadier,E., Gilbault,E., Bresson,J., Gy,I., Violle,C. and Loudet,O. (2020) Natural variation at FLM splicing has pleiotropic effects modulating ecological strategies in Arabidopsis thaliana. *Nat. Commun.*, **11**, 4140.
51. Khokhar,W., Hassan,M.A., Reddy,A.S.N., Chaudhary,S., Jabre,I., Byrne,L.J. and Syed,N.H. (2019) Genome-wide identification of splicing quantitative trait loci (sQTLs) in diverse ecotypes of *Arabidopsis thaliana*. *Front Plant Sci.*, **10**, 1160.
52. 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.
53. Yang,Q., Hu,Y., Li,J. and Zhang,X. (2017) ulfasQTL: an ultra-fast method of composite splicing QTL analysis. *BMC Genomics*, **18**, 963.
54. Wang,X., Yang,M., Ren,D., Terzaghi,W., Deng,X.W. and He,G. (2019) Cis-regulated alternative splicing divergence and its potential contribution to environmental responses in Arabidopsis. *Plant J.*, **97**, 555–570.
55. Ner-Gaon,H., Halachmi,R., Savaldi-Goldstein,S., Rubin,E., Ophir,R. and Fluhr,R. (2004) Intron retention is a major phenomenon in alternative splicing in Arabidopsis. *Plant J.*, **39**, 877–885.
56. Vaquero-Garcia,J., Norton,S. and Barash,Y. (2018) Leafcutter vs. MAJIQ and comparing software in the fast moving field of genomics. bioRxiv doi: https://doi.org/10.1101/463927, 08 November 2018, preprint: not peer reviewed.