Bioinformatics, 34(2), 2018, 300–302
doi: 10.1093/bioinformatics/btx565
Advance Access Publication Date: 11 September 2017
Applications Note

Gene expression

MAJIQ-SPEL: web-tool to interrogate classical and complex splicing variations from RNA-Seq data

Christopher J. Green[1,†], Matthew R. Gazzara[1,2,†] and Yoseph Barash[1,3,*]

[1]Department of Genetics, [2]Department of Biochemistry and Biophysics, Perelman School of Medicine, Philadelphia, PA 19104, USA and [3]Department of Computer and Information Science, School of Engineering, University of Pennsylvania, Philadelphia, PA 19104, USA

*To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Ivo Hofacker

Received on April 18, 2017; revised on August 23, 2017; editorial decision on September 4, 2017; accepted on September 9, 2017

Abstract

Summary: Analysis of RNA sequencing (RNA-Seq) data have highlighted the fact that most genes undergo alternative splicing (AS) and that these patterns are tightly regulated. Many of these events are complex, resulting in numerous possible isoforms that quickly become difficult to visualize, interpret and experimentally validate. To address these challenges we developed MAJIQ-SPEL, a web-tool that takes as input local splicing variations (LSVs) quantified from RNA-Seq data and provides users with visualization and quantification of gene isoforms associated with those. Importantly, MAJIQ-SPEL is able to handle both classical (binary) and complex, non-binary, splicing variations. Using a matching primer design algorithm it also suggests to users possible primers for experimental validation by RT-PCR and displays those, along with the matching protein domains affected by the LSV, on UCSC Genome Browser for further downstream analysis.

Availability and implementation: Program and code will be available at http://majiq.biociphers.org/majiq-spel.

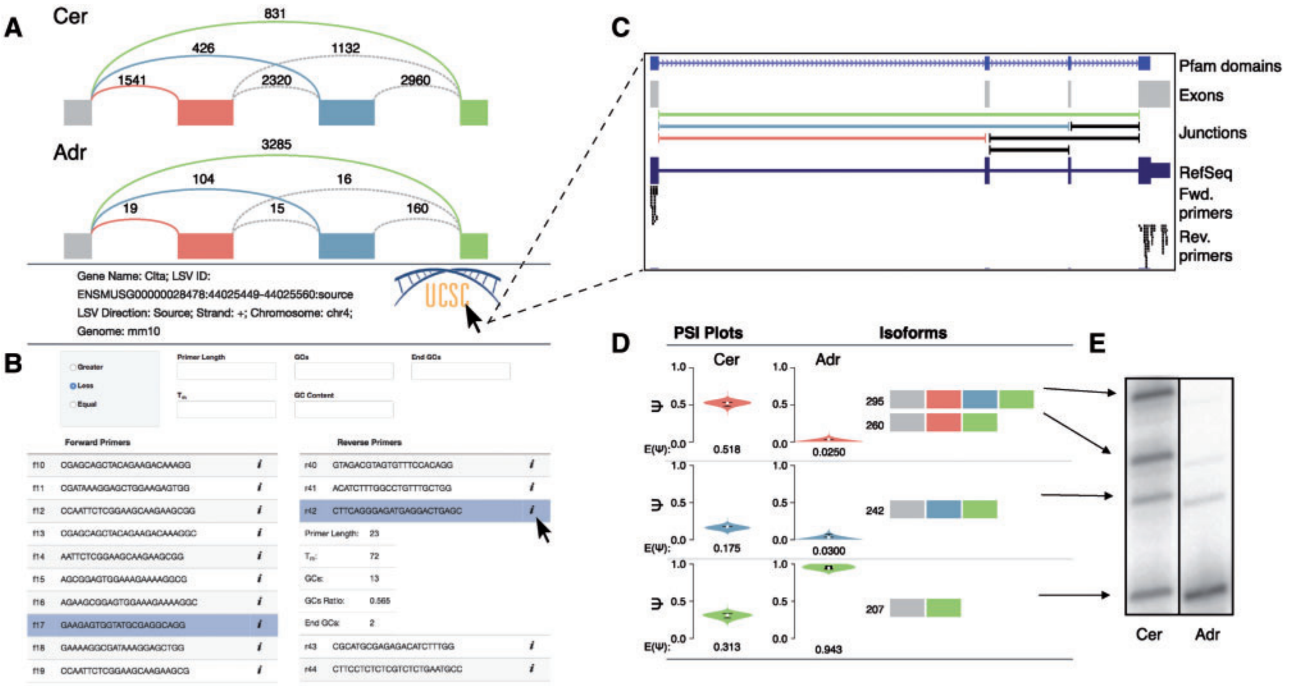
Contact: yosephb@upenn.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Advances in RNA-Seq technology have led to improved detection and quantification of splicing variations through the use of short reads that span across spliced junctions. Most commonly used AS analysis tools focus exclusively on classical, binary AS events (e.g. cassette exon, alternative 5′ or 3′ splice sites, intron retention, etc.). Recently, we formulated local splicing variations (LSVs) that capture both classical as well as complex splicing patterns (i.e. involving three or more junctions). Briefly, LSVs can be thought of as splits in a gene's splice graph where exons are nodes and splicing of pre-mRNA segments are edges. In this formulation LSVs capture several optional (alternative) pre-mRNA segments that the spliceosome may splice to a reference exon up or downstream. Figure 1A illustrates such an LSV with the reference exon marked in gray and several downstream alternative exons along with the matching LSV edges colored in red, blue and green. Such an LSVs is considered complex as it involves more than two alternative junctions. Importantly, we found that over 30% of splicing variations in extensive human and mouse RNA-Seq experiments we interrogated are complex (Vaquero-Garcia *et al.*, 2016).

The pervasiveness of complex splicing variations suggests that accurate interpretation of the underlying isoforms is crucial for experimentally interrogating and understanding the consequences of these splicing changes. We therefore developed MAJIQ and VOILA (http://majiq.biociphers.org) to define, quantify and visualize LSVs (Vaquero-Garcia *et al.*, 2016). LSVs visualization is



Gene Name: Clta; LSV ID:
ENSMUSG00000028478:44025449-44025560:source
LSV Direction: Source; Strand: +; Chromosome: chr4;
Genome: mm10

B

○ Greater
● Less
○ Equal

Primer Length
GCs
End GCs
Tm
GC Content

Forward Primers

f10	CGAGCAGCTACAGAAGACAAAGG	i
f11	CGATAAAGGAGCTGGAAGAGTGG	i
f12	CCAATTCTCGGAAGCAAGAAGCGG	i
f13	CGAGCAGCTACAGAAGACAAAGGC	i
f14	AATTCTCGGAAGCAAGAAGCGG	i
f15	AGCGGAGTGGAAAGAAAAGGCG	i
f16	AGAAGCGGAGTGGAAAGAAAAGGC	i
f17	GAAGAGTGGTATGCGAGGCAGG	i
f18	GAAAAGGCGATAAAGGAGCTGG	i
f19	CCAATTCTCGGAAGCAAGAAGCG	i

Reverse Primers

r40	GTAGACGTAGTGTTTCCACAGG	i
r41	ACATCTTTGGCCTGTTTGCTGG	i
r42	CTTCAGGGAGATGAGGACTGAGC	i
Primer Length:	23	
Tm:	72	
GCs:	13	
GCs Ratio:	0.565	
End GCs:	2	
r43	CGCATGCGAGAGACATCTTTGG	i
r44	CTTCCTCTCTCGTCTCTGAATGCC	i

Fig. 1. (**A**) Splice graph representation of LSV within *Clta* from mouse cerebellum (top) and adrenal gland (bottom) with reads detected from RNA-Seq data displayed above each junction. Junctions quantified directly within the LSV are colored. (**B**) Primer table (bottom) that suggests possible forward (left) and reverse (right) primers. Additional information for each primer can be displayed by clicking the 'i' icon as shown below the black cursor. Various primer filters can be applied (top) for additional, on the fly, filtering. (**C**) UCSC Genome Browser snapshot with custom tracks produced by MAJIQ-SPEL as labeled. SPEL opens those when clicking the Genome Browser logo shown on the left. (**D**) Isoform table that displays PSI ($\Psi \in [0,1]$) quantifications (left) and possible isoforms associated with each LSV edge (right). Note that as illustrated here complex LSVs may have a single PSI capturing multiple isoforms and that similarly PSI captures the fraction of a splicing event (edge), not necessarily the fraction of each colored exon. Fraction of each Nucleotide sizes correspond to products produced using the selected primers from (B). (**E**) Representative RT-PCR validation of predicted product sizes and quantification using the primers selected in (B) on total RNA from mouse cerebellum (left) and adrenal gland (right)

based on segments of splice graphs as shown in Figure 1A, while quantification is based on PSI (Percent Selected Index, $\Psi \in [0,1]$) which captures the marginal fraction of each LSV edge (i.e. the fraction of isoforms that utilize this splicing junction). Similarly, changes between experimental conditions are measured by dPSI ($\Delta\Psi \in [-1,1]$). However, no current tool offers a user-friendly interface to connect LSVs, whether simple or complex, to the underlying known gene isoform and affected protein domains. Also, there is a clear need for automated design and visualization of potential primers that flank an LSV for experimental validation via RT-PCR, the gold standard in the field. Specifically, previous work only allows for design of a single primer pair and focuses on classical, binary AS events (Tokheim *et al.*, 2014).

2 Results

We developed the web-tool MAJIQ-SPEL (MAJIQ for Sampling Primers and Evaluating LSVs) to aid in the visualization, interpretation and experimental validation of both classical and complex splicing variations. Typically MAJIQ and its visualization package VOILA (Vaquero-Garcia *et al.*, 2016) are executed by users on datasets ranging from just a few to hundreds or thousands of RNASeq samples to detect local splicing variations (LSV) of interest. SPEL can then analyze LSVs of interest from such large executions which quantify PSI (quantification of a single experimental group) or delta PSI (quantification of splicing changes between two experimental groups). MAJIQ-SPEL (or SPEL for short) is implemented on a Galaxy web server (Afgan *et al.*, 2016) and takes as input the output of VOILA (Vaquero-Garcia *et al.*, 2016). Specifically, users can now click a button to copy a splice graph and LSV quantification of interest, then paste it into SPEL's Galaxy input form and run the analysis. SPEL is intended to be used primarily as a Galaxy web-tool SPEL but we also made it available as a stand-alone version. The stand-alone is light on memory and CPU, taking about 0.5 s and 24 MB of memory per job on a standard laptop.

MAJIQ-SPEL output contains several components, which we highlight in Figure 1 using a complex LSV within *Clta* generated comparing RNA-Seq from mouse cerebellum and adrenal gland (Zhang *et al.*, 2014). First, colorized representations of the LSV are displayed with junction spanning read counts for each junction quantified directly in the LSV (colored arcs in Fig. 1A). Also shown are counts for junction spanning reads that occur within the boundaries of the event, but are not part of the LSV quantified (dashed grey arcs). This visualization allows for quick interpretation of which paths are commonly utilized in each sample. We note that the ratio of the colored read counts usually correspond approximately to the expected PSI ($E[\Psi]$) but may vary from it due to various normalization factors applied during quantification (GC content, stack removal etc.).

Second, SPEL produces a table of putative forward and reverse primers for the 5′- and 3′-most exons within the LSV (Fig. 1B). The primers are optimized for validating the given LSV via low-cycle RT-PCR, based on the experimental protocols and primer design factors described in (Smith and Lynch, 2014). In brief, to allow for a stringent RT-PCR assay each primer must have a minimal melting temperature ($T_{m)}$ of 76 °C by the Marmur formula, have a GC content of between 50 and 60%, and have between 2 and 4 G or C nucleotides at the 3′ end. Additionally, the selected primer pair should produce expected products within a certain size range to allow visualization via gel electrophoresis and to reduce bias during reverse transcription (Smith and Lynch, 2014). Importantly, design considerations such as minimum and maximum primer length, product length, GC content, T_m and T_m estimation method can be adjusted under 'Advanced Options' in the submission form. All primers that meet these criteria are displayed for users to sample and select a pair that best meets experimental needs. For ease of use, the primer table is searchable and key summary information for each primer can also be displayed. Additionally, a number of filters can be applied to the primer table to further reduce the number of primers shown, on the fly, without re-executing SPEL (Fig. 1B).

MAJIQ-SPEL also offers UCSC Genome Browser (Kent *et al.*, 2002) connectivity. Clicking the browser's logo brings up custom tracks that display the exons, junctions and locations of putative primers to aid in selection of primers for validation (Fig. 1C). These tracks also include known isoforms and annotated protein domains [Pfam (Finn *et al.*, 2016)], which can aid in examining the functions of alternative isoforms produced.

Finally, MAJIQ-SPEL traverses all possible paths within the splice graph contained in the LSV region based on observed and annotated junctions to create the isoform segments table that links the MAJIQ PSI quantification to the associated isoform(s) (Fig. 1D). Importantly, once the user selects a forward and reverse primer pair, this table updates to display the expected product size for each isoform segment for validation. Additionally, once a primer pair is chosen, the user can run In-Silico PCR through UCSC Genome Browser (Kent *et al.*, 2002) to further validate and check the specificity of the chosen pair. In the example shown, RT-PCR performed using primers generated by MAJIQ-SPEL demonstrates both accurate prediction of all four product sizes and quantification for both cerebellum and adrenal gland (Fig. 1E).

Beyond handling classic or complex splicing variations, MAJIQ-SPEL also offers researchers fast and accurate primer design for de novo splicing variations not in the annotated transcriptome. In such cases experimental validation is crucial. Such a case is shown in an event in *Fubp3* (Supplementary Fig. S1). Since this LSV involves novel exon skipping it will likely not be captured in other tools for splicing quantification and visualization packages that rely only on the annotation database.

The *Fubp3* and *Clta* splicing variations shown here also highlight how MAJIQ-SPEL can aid in functional analysis of LSVs. The combined UCSC Genome Browser tracks show the alternative exons overlap annotated protein domains, suggesting a functional effect. The cassette exon in *Fubp3* is not a multiple of three, suggesting a frameshift and the Browser tracks revealed that skipping inserts a premature termination codon (PTC). Future extensions of this work will aim to further integrate these and other functional analyses into MAJIQ-SPEL.

Funding

This work has been supported in part by the Penn Institute for Biomedical Informatics Pilot Grant and R01 AG046544 to YB.

Conflict of Interest: none declared.

References

Afgan,E. *et al.* (2016) The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.*, **44**, W3.

Finn,R.D. *et al.* (2016) The pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.

Kent,W.J. *et al.* (2002) The human genome browser at ucsc. *Genome Res.*, **12**, 996–1006.

Smith,S.A. and Lynch,K.W. (2014) Cell-based splicing of minigenes. *Spliceosomal Pre-mRNA Splicing Methods Protoc.*, **1126**, 243–255.

Tokheim,C. *et al.* (2014) Primerseq: design and visualization of rt-pcr primers for alternative splicing using rna-seq data. *Genomics Proteomics Bioinf.*, **12**, 105–109.

Vaquero-Garcia,J. *et al.* (2016) A new view of transcriptome complexity and regulation through the lens of local splicing variations. *eLife*, **5**, e11752.

Zhang,R. *et al.* (2014) A circadian gene expression atlas in mammals: implications for biology and medicine. *Proc. Natl. Acad. Sci. USA*, **111**, 16219–16224.