SpliceScape: A Pipeline for genome-wide splicing event detection.

Beatriz Rodrigues Estevam   **Diego M. Riaño-Pachón**

University of São Paulo (CENA/USP)

Introduction and objectives

A genome-wide process. Alternative splicing is a central process in gene expression. It is nearly ubiquitous in eukaryotes, affecting over 95% of human genes and similarly high rates in major plant species.

A Narrow Research Focus. Despite its prevalence, the vast majority of splicing research focuses on a very limited set of well-characterized gene isoforms. This narrow scope restricts our understanding of splicing products at the genome-wide scale.

Bias in genome-wide analysis. Moreover, existing large-scale splicing studies are built using heterogeneous computational methods, introducing significant bias, making it difficult to reliably compare splicing data across different studies, species, or conditions.
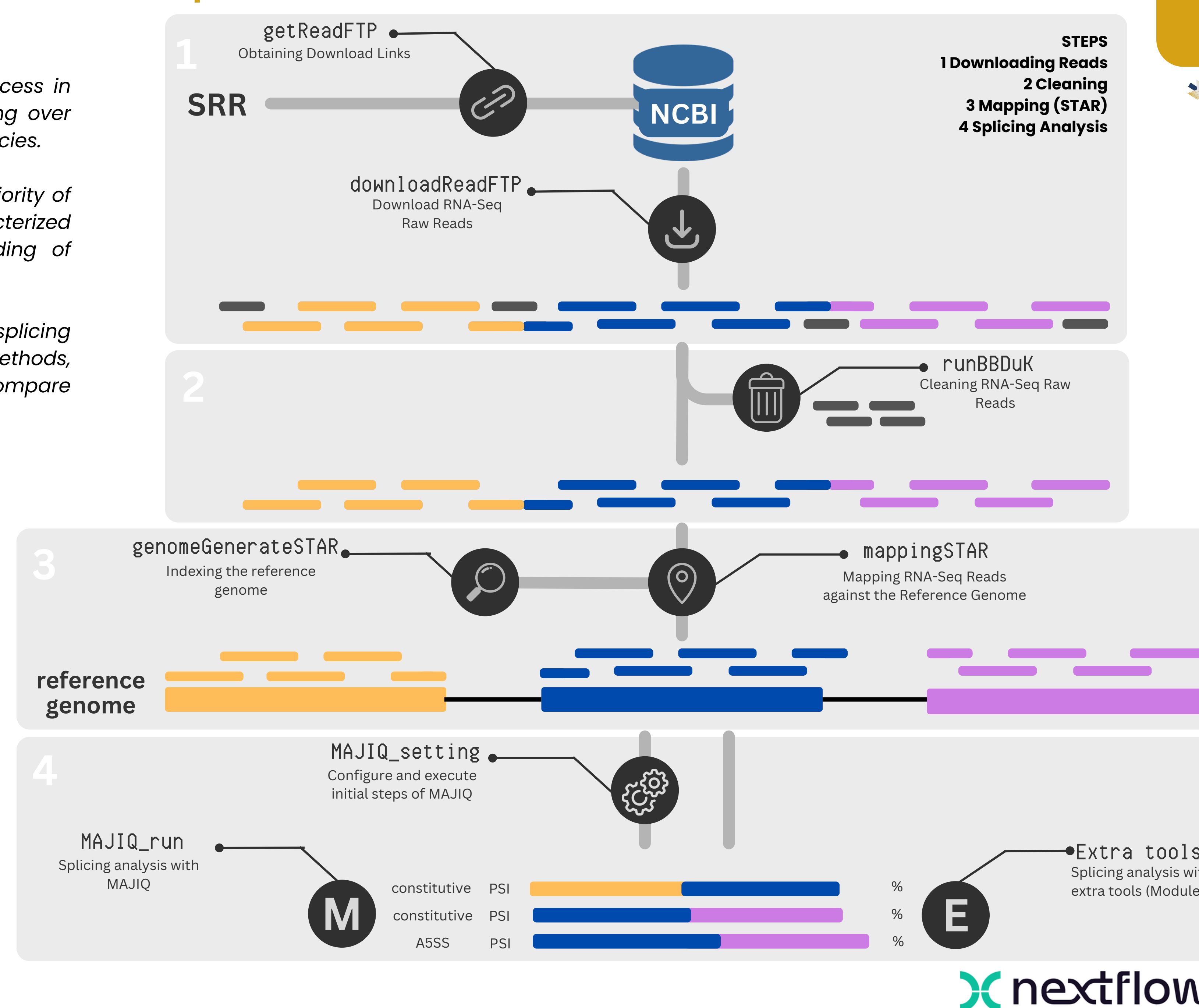
Possible Solution. An standardized computational pipeline offer a direct solution to this challenge. It could provide a robust framework for performing homogeneous, scalable, and reproducible analyses, paving the way for trustworthy discoveries in comparative splicing analysis.

Key Objectives

To develop an automated pipeline for homogeneous large-scale detection of splicing that:

1. is adaptable to computational resources.
2. can capture different splice types.
3. can capture de novo events.
4. can capture conserved splicing-related sequences.
5. allows the addition of new splice detection modules
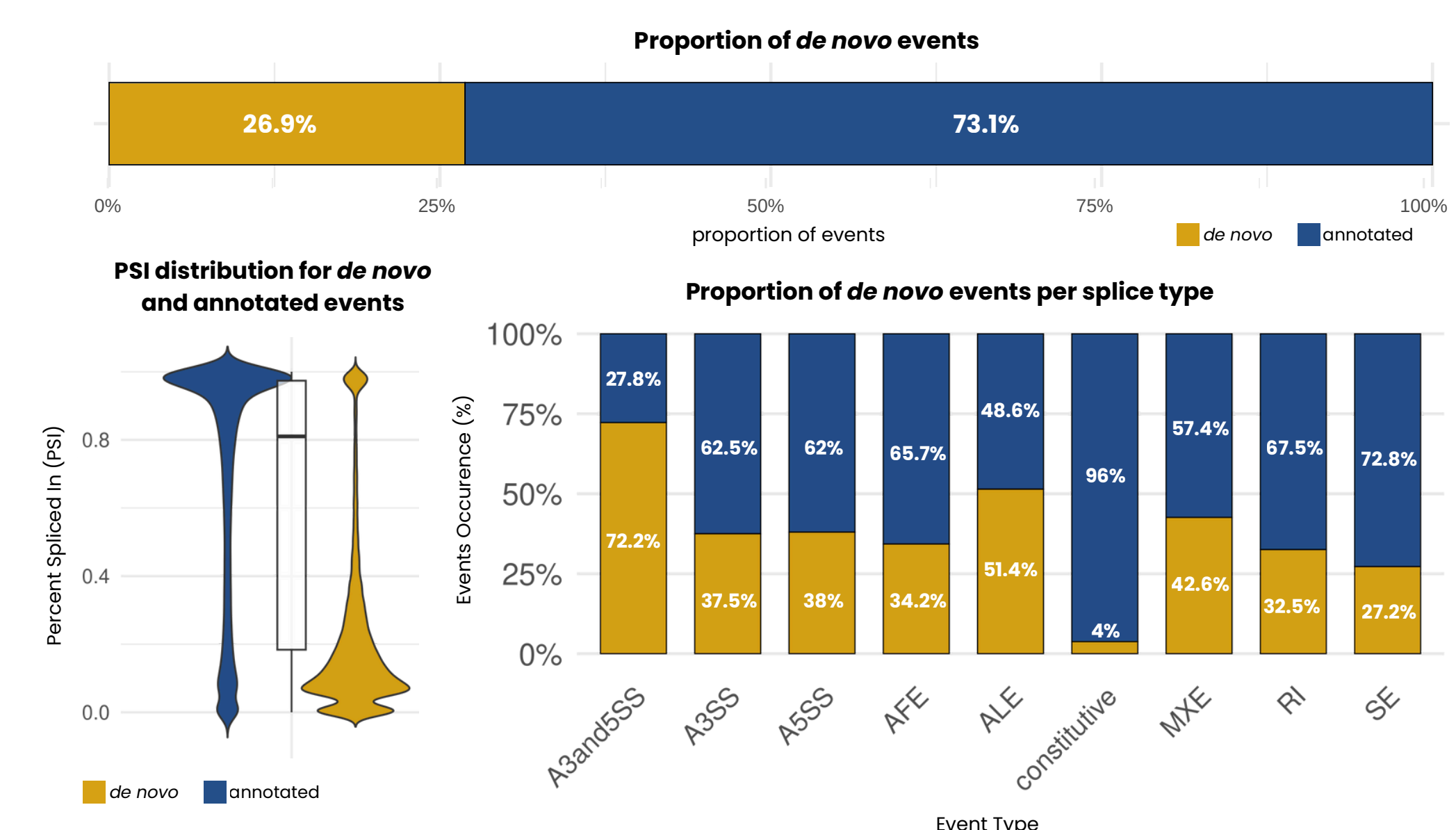
Material and Methods

Testing Data and Tools

resources config. · .fasta · .gff · SRR ids · metadata

SpliceScape Processes

Data and Tools

19 leaf *A. thaliana* samples
Over 100 nucleotides in length
Sequenced by Illumina
Bioproject PRJNA1040277

Downloading reads — ffq v0.3.0, wget

Cleaning — BBDuK v35.85

Mapping — STAR v2.7.10

Splicing analysis — MAJIQ

Extra tools (Modules) — 5

Post processing and Event Parsing — sql.db, events.bed

Python, SQLite, R, X

LabBCES Cluster
2/4 nodes
48 and 80 cores
500GB RAM each

Pipeline Overview

STEPS
1 Downloading Reads
2 Cleaning
3 Mapping (STAR)
4 Splicing Analysis

1
getReadFTP — Obtaining Download Links
SRR → NCBI
downloadReadFTP — Download RNA-Seq Raw Reads

2
runBBDuK — Cleaning RNA-Seq Raw Reads

3
genomeGenerateSTAR — Indexing the reference genome
mappingSTAR — Mapping RNA-Seq Reads against the Reference Genome
reference genome

4
MAJIQ_setting — Configure and execute initial steps of MAJIQ
MAJIQ_run — Splicing analysis with MAJIQ
constitutive PSI %
constitutive PSI %
A5SS PSI %
Extra tools — Splicing analysis with extra tools (Modules)

nextflow

Performance Analysis

Is SpliceScape adaptable to computational resources?

Resource Usage and Pipeline Execution Timeline
Arabidopsis thaliana

Resource — CPU Usage, Memory Usage

Process — ALTERNATIVE_RUN_BBDUK, GENOME_GENERATE_STAR, MAJIQ_RUN, MAJIQ_SETTING, MAPPING_STAR, SGSEQ, WGET_DOWNLOADER

Total run time: 13h 45s

Resource Usage by Process
Arabidopsis thaliana

Nextflow Process — ALTERNATIVE_RUN_BBDUK, GENOME_GENERATE_STAR, MAJIQ_RUN, MAJIQ_SETTING, MAPPING_STAR, SGSEQ, WGET_DOWNLOADER

Execution Time (minutes): 10, 30, 60

Peak Memory Usage (GB)
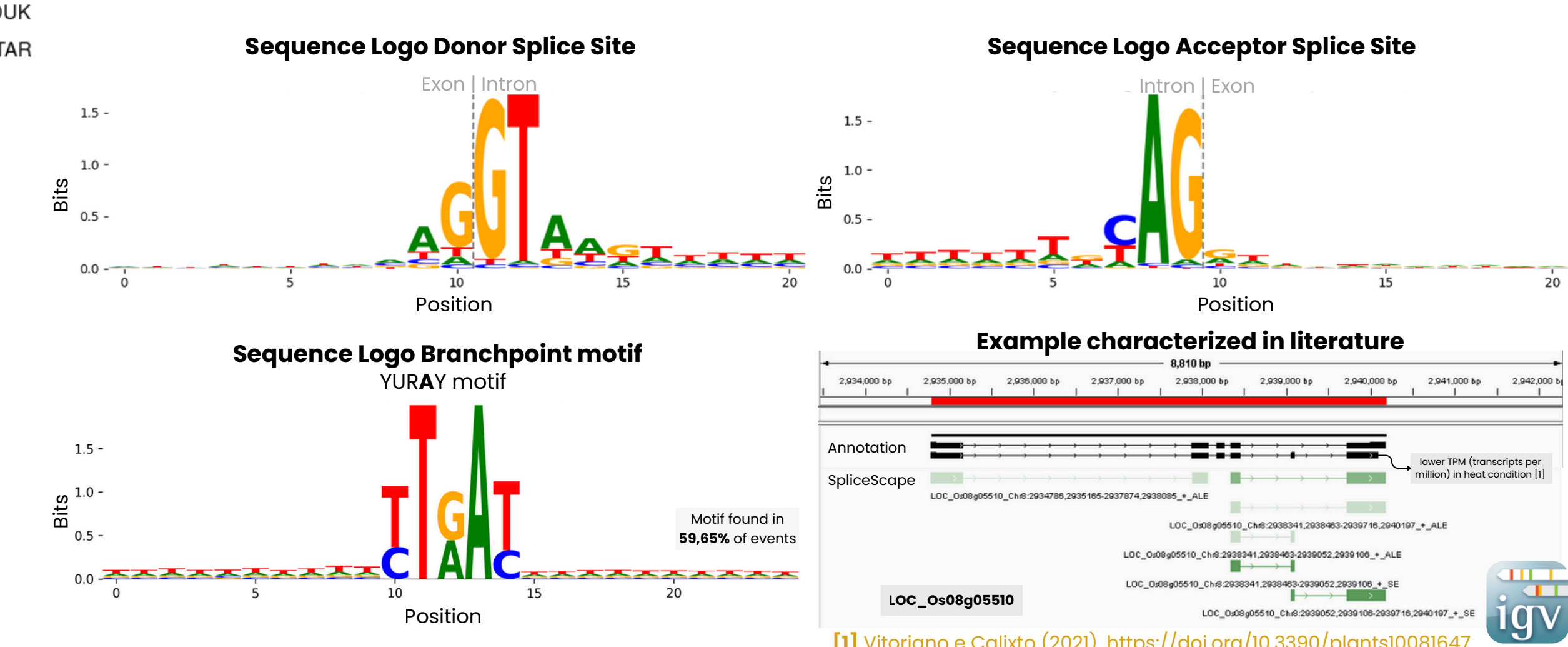
Key points

SpliceScape provides a **scalable**, **reproducible,** and **standardized** solution for genome-wide identification of **splicing events from RNA-seq data.** It allows customization based on computational resources and helps standardize splicing landscapes, likely reducing methodological biases.
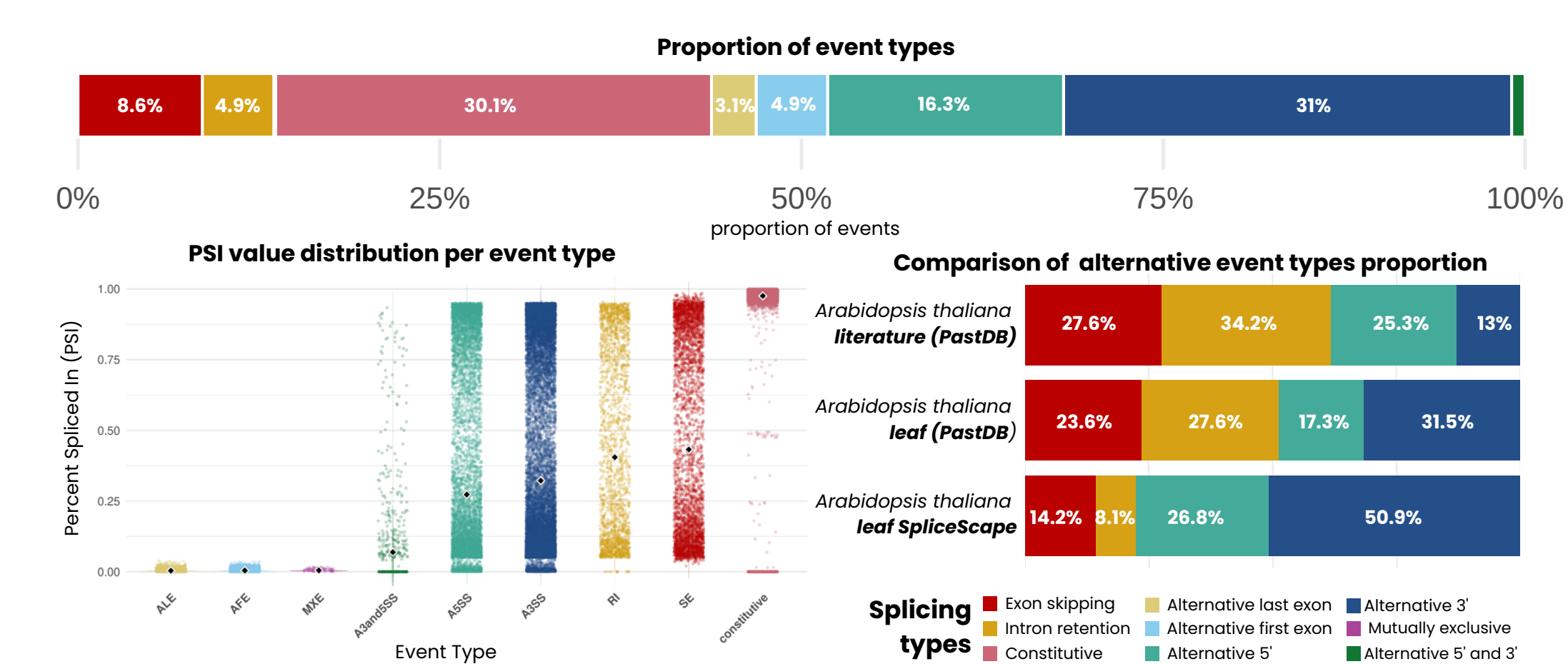
---

19 samples   Total run time: 13h 45s   41,953 events identified

Different event types
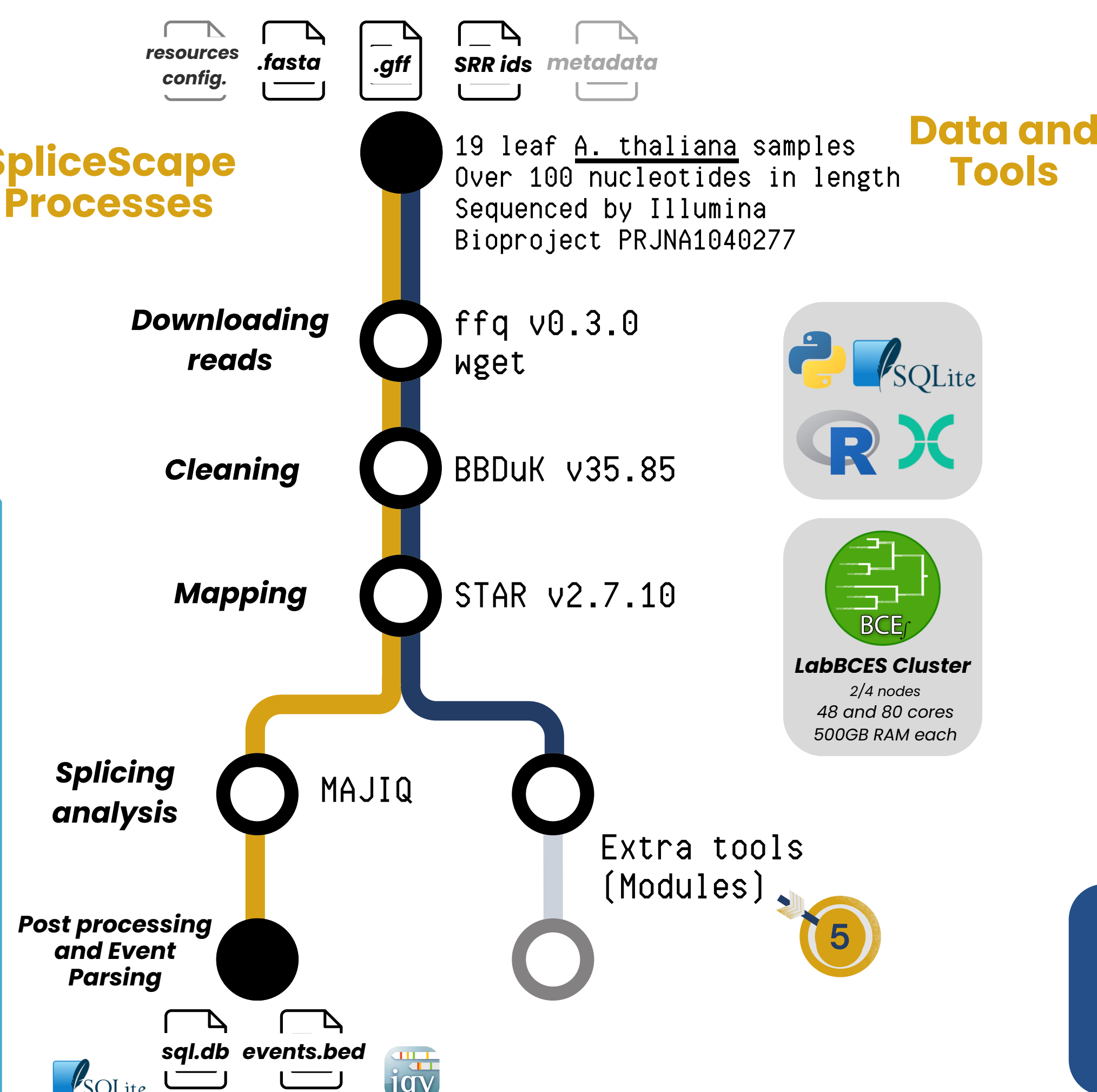**Does SpliceScape capture different splicing types?**

Proportion of event types
8.6% | 4.9% | 30.1% | 3.1% | 4.8% | 16.3% | 31%
0% — 25% — 50% — 75% — 100%
proportion of events

PSI value distribution per event type
Percent Spliced In (PSI)

Comparison of alternative event types proportion

Arabidopsis thaliana literature (PastDB)	27.6%	34.2%	25.3%	13%
Arabidopsis thaliana leaf (PastDB)	23.6%	27.6%	17.3%	31.5%
Arabidopsis thaliana leaf SpliceScape	14.2%	8.1%	26.8%	50.9%

Splicing types: Exon skipping · Intron retention · Constitutive · Alternative last exon · Alternative first exon · Alternative 5' · Alternative 3' · Mutually exclusive · Alternative 5' and 3'

De novo events
Does SpliceScape capture de novo splicing events?

Proportion of de novo events
26.9% | 73.1%
0% — 25% — 50% — 75% — 100%
proportion of events
de novo · annotated

PSI distribution for de novo and annotated events
Percent Spliced In (psi)
de novo · annotated

Proportion of de novo events per splice type
Events Occurrence (%)

	A3andSS	A3SS	A5SS	AFE	ALE	constitutive	MXE	RI	SE
de novo (top)	27.8%	62.5%	62%	65.7%	48.6%	96%	57.4%	67.5%	72.8%
annotated (bottom)	72.2%	37.5%	38%	34.2%	51.4%	4%	42.6%	32.5%	27.2%

Event Type

Splicing-related sequences
Does SpliceScape capture events with splicing-related sequences?

Sequence Logo Donor Splice Site
Exon | Intron
Bits — Position

Sequence Logo Acceptor Splice Site
Intron | Exon
Bits — Position

Sequence Logo Branchpoint motif
YURAY motif
Bits — Position
Motif found in 59,65% of events

Example characterized in literature
Annotation
SpliceScape
LOC_Os08g05510
igv

[1] Vitoriano e Calixto (2021) https://doi.org/10.3390/plants10081547

SpliceScape

Lab Web page · Lab instagram