Tutorial MAJIQ/Voila

Introduction

What are *MAJIQ* and *Voila* ?

MAJIQ and Voila are two software packages that together define, quantify, and visualize local splicing variations (LSV) from RNA-Seq data. Conceptually, MAJIQ/Voila can be divided into three modules:

- **MAJIQ Builder**: Uses RNA-Seq (BAM files) and a transcriptome annotation file (GFF/GTF) to define splice graphs and known/novel Local Splice Variations (LSV).
- **MAJIQ Quantifier**: Quantifies relative abundance (PSI) of LSVs and changes in relative LSV abundance (delta PSI) between conditions w/wo replicates.
- **Voila**: A visualization package that combines the output of MAJIQ Builder and MAJIQ Quantifier using interactive D3 components and HTML5. Voila creates interactive summary files with gene splice graphs, LSVs, and their quantification.

The above three modules are designed to be executed in sequence with one module's output feeding into the other. In most usage cases, the Builder will be executed only once for a given set of RNA-Seq experiments, and then the Quantifier and Voila may be executed on top of it multiple times for different analysis tasks.

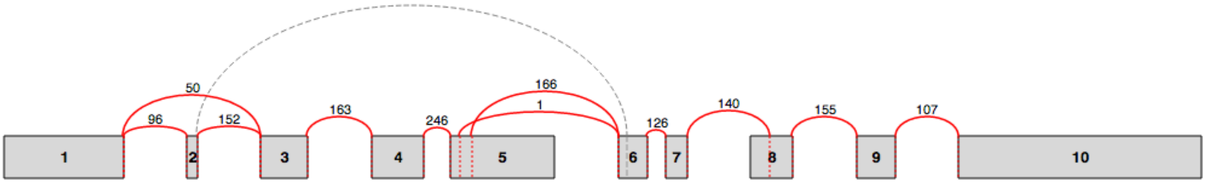
There are two main modes of executing the quantifier: Quantifying the relative inclusion levels of LSVs in a given experimental condition (also known as *"percent spliced in"*, PSI, or Ψ), and quantifying changes of LSVs inclusion levels between two experimental conditions (aka delta PSI or $\Delta\Psi$).

Voila has two main modes of visualizing the Quantifier's results, whether these are PSI or delta PSI quantifications. The first is a (possibly long) table view of LSVs that can be filtered and ordered by different attributes (columns). The second mode is gene based, in which case each gene's splice graph and matching LSVs are grouped together. In both cases, the experimental condition can be either a single experiment or a set of replicates. In all cases the output is an interactive HTML5 that can be opened in a web browser. There is also an option to dump the output as a tab delimited text file for further analysis with other tools/scripts.

Below there is more information describing what are LSVs, how are they quantified and visualized, and what MAJIQ can (and equally important - cannot) do. You can either go through those or jump directly to the Quick Start guide.

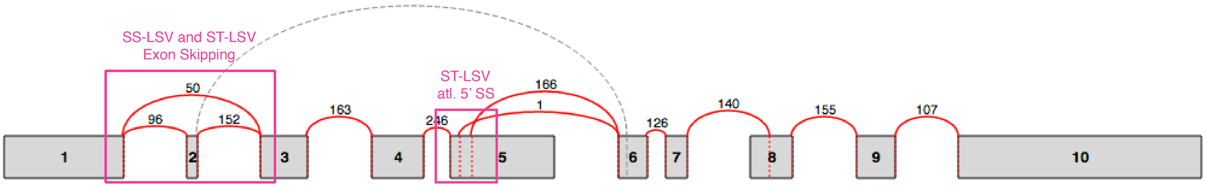
What is an LSV?

LSV stands for "local splicing variation". Briefly, exons that are spliced together can be represented using a *splice graph* Heber et al. 2002 such as this:



> In Voila Splice graphs, exons are represented by rectangles and junctions (or edges) by arcs. The raw number of reads spanning a junction is also displayed. For a more detailed description see splice graphs description.

LSVs involve an exon (or node in the splice graph) from which splits in the graph originate (aka single source LSV, or SS-LSV) or an exon into which several graph edges converge (aka single target LSV, or ST-LSV). An illustration of a splice graph is shown below with several SS-LSV and ST-LSV marked. The “local” aspect of LSVs definition stems from the fact they involve only a single source or single target exon. For a more formal definition please see XXXX(cite paper).



The terminology of LSVs generalizes that of alternative splicing (AS) “events” and AS events “types”. The most common AS types in mammals are skipped exons, alternative 3’ splice site, and alternative 5’ splice site (cf Wang et al., *Nature* 2008). These types can all be seen as specific cases of simple or binary LSVs, i.e. LSVs that involve only two way graph splits (see figure above). In the transcriptome reality though, we find many other types of LSVs that involve different combinations of 3’ and 5’ splice site choices in different exons (see figures above/below, or just run your data through MAJIQ/VOILA…). Consequently, the LSV terminology helps us define and quantitate more accurately the spectrum of local splice variations observed in the transcriptome.

Conceptually, LSVs are aimed to fill the gap between previously defined AS “types” described above, and full transcripts/isoforms. Ideally, we would like to identify and quantify all existing isoforms of each gene in a given RNA-Seq experiment. However, the complexity of gene isoforms combined with the shortness of current sequencing reads (typically ~100b long) makes isoforms quantification from RNA-Seq reads an under-determined and challenging problem. In contrast, LSVs can arguably still capture a lot of useful information about transcriptome variability while being deduced directly from RNA-Seq reads that span across splice junctions.

What is LSV quantification?

MAJIQ's LSV quantification is based on estimating the relative inclusion level of each junction in the LSV. For simple, binary, cases such as skipped exons LSV quantification is equivalent to estimating the exon's percent spliced in (PSI, or Ψ). For more complex LSVs that involve three or more splice graph edges (*i.e.*, exon joining options), *MAJIQ* computes the marginal inclusion level, or PSI, per junction. Computing only these marginals allows *MAJIQ* to handle complex LSVs, keeping computational cost linear with the number of edges while still delivering estimates for the interesting biological question of "*how much is each junction used?*".

When estimating PSI for a LSV's junctions, *MAJIQ* produces a complete posterior distribution over possible PSI values. This distribution takes into account the number of reads observed at each junction, their distribution across genomic positions, GC content bias, and some possible mapper or technical artifacts. Intuitively, the deeper and smoother the coverage of an LSV, the more concentrated the PSI posterior would be (i.e. the more "sure" *MAJIQ* is about the "true" PSI value), while lower and less even coverage would result in higher variance of the PSI estimate.

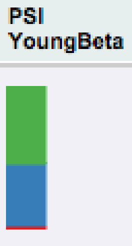
Similarly, *MAJIQ*'s quantification of LSV's differential inclusion when comparing two conditions is based on estimating a posterior distribution for the change in each junction's relative inclusion level, termed delta PSI ($\Delta\Psi$). Naturally, this distribution lies in the range of -100% to +100% (or -1 to +1 when using fractions instead of percentage points).

**Note*: For a thorough description of *MAJIQ*'s quantification algorithm for Ψ and $\Delta\Psi$ and the various parameters that control it see XXX.

Voila's visualization of LSV Quantification uses several different techniques. In all cases, colors are used to represent the different junctions in the LSV. For binary LSVs Voila uses histograms to represent the posterior probability over Ψ or $\Delta\Psi$. For more complex LSVs involving 3 or more splice graph edges *Voila* uses *violin plots*. Examples of the histogram and violin plots for single Ψ are shown below. Some users might prefer to only use violin plots, even for binary LSVs (command line option *–only-violins*.

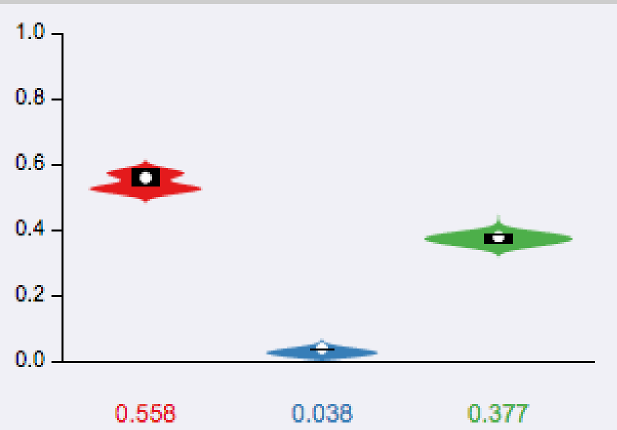
PSI

Compact view



When displaying lists of LSVs *Voila* uses a compact stacked bar chart representation. The **height** of each bar represent the **mean PSI ($E[\Psi]$)** which naturally add to 100% over all the LSV's junctions. Clicking over the bars will open a more detailed representation of the PSI distribution.

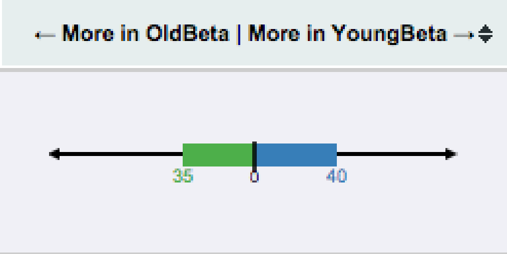
Violin plots (binary and multi-way LSV)



> Violin plots are *boxplots* plotted over the original distributions. The *box* goes from the 25 to the 75 percentile, with a white horizontal line indicating the 50% (median). The tails represents the 10 and 90 percentile. Additionally, the mean or $E(\Psi)$is marked with a white circle.

Delta PSI

Compact view



For compact visualization of $\Delta\Psi$ quantification, each colored bar represents the percentage of the expected differential inclusion for the matching edge in the splice graph. The arrows indicate the preference for one condition versus another. In this example, the blue junction is shown to have a 40% more inclusion in *YoungBeta* compared with *OldBeta*.

In contrast the green junction is expected to be more included in *OldBeta* by a 35% difference. Note that in this case the numbers do not generally sum to 1 or 100% as they reflect MAJIQ's expected change for each junction separately. In addition, users can use the more detailed violin plots to gain other information/statistics such as the confidence and probability distribution per junction.

What is MAJIQ?

MAJIQ is a software package that allows researchers to define and quantify both known and novel Local Splice Variations (LSVs) in genes from RNA-Seq data.

MAJIQ's main features

MAJIQ takes as input a set of RNA-Seq experiments (BAM files) and previous genome annotation (GFF/GTF files) and produces the following:
- Splice graph for each gene based on both known transcripts annotation and de-novo junctions detected.
- All detected (known + novel) single source and single targets LSVs per gene.
- Quantification of LSVs from a given RNA-Seq experiment (w/wo replicates).
- Changes in LSVs quantification between two conditions from RNA-Seq experiments (w/wo replicates).

What MAJIQ is not

There are many RNA-Seq analysis tasks for which *MAJIQ* was *not* designed or is currently not structured to address. Some examples include:

- Gene/isoform expression estimation: *MAJIQ* uses expression levels when it quantifies LSVs. For example, LSVs that are not present in the data will not be quantified and those with lower coverage will result in lower confidence for the relative inclusion of matching RNA segments. However, *MAJIQ* only computes *relative* inclusion of junctions in an LSV (e.g., 80% inclusion of an alternatively skipped exon). Consequently, computing the expression of genes or isoforms and comparing those in the same experiment or between experiments is not supported.

- Relative isoform abundance: *MAJIQ* only operates at the level of local splice variations (LSVs). It does not assume the full spectrum of gene isoforms is known and does not quantify those.
- Novel gene/non coding RNA detection: *MAJIQ* requires a transcriptome annotation file (GFF/GTF). It supplements by identifying both known and novel splice junctions in existing loci. Putative isoforms of new loci are not inferred during this process.
- Alternative transcription start/end
- Alternative polyadenylation (APA) identification/quantification.

What is Voila?

Voila is a package to interactively visualize splice variations in RNA-Seq data. It is written in Python and produces summary files in HTML5 that can be opened and interactively explored with any modern browser*. It has been conceived as the visual component of MAJIQ for analysis of Local Splice Variants (LSVs).

**Voila has been tested on Google Chrome [recommended], Firefox and Safari.*

Quick start

Pre MAJIQ

Select a GFF3 annotation file
The general feature format (gene-finding format, generic feature format, GFF) is a file format used for describing genes and other features of DNA, RNA, and protein sequences. The format specification for the gff version 3 can be found at gff3 format.
In our case we use some of these features in order to define genes, transcripts and exons. An example of this format is shown below

```
chr1 protein_coding gene 107399655 107452689 . + .
Name=Serpinb7;ID=ENSMUSG00000067001;Name=ENSMUSG00000067001
chr1 protein_coding mRNA 107399655 107435399 . + .
Parent=ENSMUSG00000067001;Name=Serpinb7-002;ID=ENSMUST00000154538
chr1 protein_coding exon 107399655 107399724 . + .
Parent=ENSMUST00000154538;ID=exon:ENSMUST00000154538:1
chr1 protein_coding exon 107428231 107428416 . + .
Parent=ENSMUST00000154538;ID=exon:ENSMUST00000154538:2
chr1 protein_coding exon 107434736 107434786 . + .
Parent=ENSMUST00000154538;ID=exon:ENSMUST00000154538:3
chr1 protein_coding exon 107435327 107435399 . + .
ID=exon:ENSMUST00000154538:4;Parent=ENSMUST00000154538
chr1 protein_coding five_prime_UTR 107399655 107399724 . + .
Parent=ENSMUST00000154538;ID=five_prime_UTR:ENSMUST00000154538:1
chr1 protein_coding five_prime_UTR 107428231 107428248 . + .
ID=five_prime_UTR:ENSMUST00000154538:2;Parent=ENSMUST00000154538
chr1 protein_coding start_codon 107428249 107428251 . + 0
Parent=ENSMUST00000154538;ID=start_codon:ENSMUST00000154538:1
chr1 protein_coding CDS 107428249 107428416 . + 0
ID=CDS:ENSMUST00000154538:1;Parent=ENSMUST00000154538
chr1 protein_coding CDS 107434736 107434786 . + 0
Parent=ENSMUST00000154538;ID=CDS:ENSMUST00000154538:2
…
```

In order to obtain this format, we recommend the use of some of the most well known online DB. They provide the annotation files in some format like *gtf*, and you can transform this file to *gff3* using a script, like *script*

Study configuration file

MAJIQ has a set of parameters needed for its execution. Several of them depend of the RNA-Seq study. This configuration file should include this information in order to be able to pass it the the MAJIQ Builder. Secondly,
it is useful to keep the info of the study ready and accessible.

This is an example of the configuration file, divided in two blocks, *info* and *experiments*:

```
[info]
readlen=76
samdir=/data/MGP/ERP000591/bam
genome= mm10
genome_path=/data/WASP_DATA/Genomes/goldenPath/mm10
type=strand-specific
```

```
[experiments]
Hippocampus=Hippocampus1,Hippocampus2
Liver=Liver1,Liver2
```

Info

This is the study global information needed for the analsysis. The mandatory fields are:

- *readlen*: Length of the RNA seq reads
- *samdir*: Path where the bam files are located
- *genome*: Genome assembly
- *genome_path*: Path to the genome fasta files. This path is needed for the GC content normalization procedure.
- *type=strand-specific*: When using strand specific RNASeq data with negative strand as reference. [Omit this parameter otherwise].

Experiments

This section defines the experiments and replicates that are to be analyzed. Each line defines a condition and its name can be customized in the following way:

```
<group name>=<experiment file1>[,<experiment file2>]
```

where the experiment file is the bam filename inside the samdir directory (excluding extension *.bam*).
Note: For a better performance we strongly recommend using sorted and indexed bam files. The index bam files (.bai) should be accessible in the same folder. *`<experiment file1>.bam.bai`*

MAJIQ Builder

MAJIQ Builder is the part of MAJIQ tool where RNA-Seq data is analyzed in order to detect LSV candidates.

```
majiq build <transcript list> -conf <configuration file> --nthreads NT
--output <build outdir>
```

- **Transcript list**: This is the file with the annotation database. Currently, we accept only gff3 format. For a better description, see the annotation file section.
- **Configuration file**: This is the configuration file for the study. This file should define the files and the paths for the bam files, the read length, the genome version, and some other information needed for the builder. For a more detailed information, please check the configuration file section.

- **NT**: Number of threads to use. *Disclaimer: Majiq Builder uses large amounts of memory. For example, in genome-wide for mouse, each thread uses ~3.5Gb. If you set a large number of threads that could trigger the swap mechanism, extremely slowing down the execution.*
- **Build outdir**: Directory where the output will be placed. MAJIQ builder has a set of output files *.majiq* and *.splicegraph*. For each bam file, MAJIQ builder is going to generate these two files which will be the input files in the next steps of the analysis.

MAJIQ builder has several arguments in order to tweak its analysis and performance. Please check the MAJIQ section for a more detailed explanation.

PSI Analysis

PSI quantification

MAJIQ PSI quantifies the LSV candidates given by the Builder. In order to improve its accuracy and reproducibility, it allows the use of biological replicates.

```
majiq psi <build outdir>/<replicate1>.majiq [<build
outdir>/<replicate2>.majiq ...] --nthreads NT --output <psi outdir> --
name <cond_id>
```

- ***.majiq file[s]**: the *.majiq* file(s) that were created by the MAJIQ Builder execution.
- **cond_id**: group identifier that you want to use for this execution

Visualize results with VOILA

The package VOILA allows the user to generate interactive summaries to display MAJIQ computations and quantifications in the browser. All the information is also provided in TAB-delimited files that can be easily parsed for further analysis.

```
voila psi <psi outdir>/<cond_id>.psi.pickle --genes-files <build
outdir>/<replicate1>.splicegraph [<build
outdir>/<replicate2>.splicegraph ...] -o <voila outdir>
```

- **<cond_id>.psi.pickle** is the output file from PSI computation,
- **<replicate1>.splicegraph [<replicate2>.splicegraph ...]** contains information about the genes and splice variants identified in each replicate. Note that you can specify a directory if you have more than one splice graph from the same group. In that case, all files with *.splicegraph* extension will be consider.

In the output directory **<voila outdir>** you will find:

- **index.html**: HTML file with a table containing all genes and LSVs identified and analyzed.
- **summaries/xx_<cond_id>.psi_gene.html** files: interactive HTML5 summaries with MAJIQ quantifications, where **xx** is the page counter.
- **<cond_id>._psi.txt**: A tab-delimited file with all LSV information (expected PSI value, variance, exon coordinates, junction coordinates, etc.) and genomic information (chromosome, strand and coordinates).
- **static** folder: needed for the correct visualization of the `index.html` file.
- **voila.log**: log file with the execution information of Voila.

For more information see VOILA section.

Delta PSI Analysis

Delta PSI quantification

Majiq Delta PSI quantifies the differential splicing between two different groups (or conditions). Like PSI, Delta PSI is able to use replicates for each group in order to improve its accuracy and reproducibility.

```
majiq deltapsi -grp1 <build outdir>/<cond1_rep1>.majiq [<build outdir>/<cond1_rep2>.majiq ...] -grp2 <build outdir>/<cond2_rep1>.majiq [<build outdir>/<cond2_rep2>.majiq ...] --nthreads NT --output <dpsi outdir> --names <cond1_id> <cond2_id>
```

- **-grp1 *.majiq* file[s]**: Set of *.majiq* file[s] for the first condition,
- **-grp2 *.majiq* file[s]**: Set of *.majiq* file[s] for the second condition,
- **–name cond_id1 cond_id2**: group identifiers for *grp1* and *grp2* respectively.

Visualize results with VOILA

To visualize deltapsi quantification with Voila execute:

```
voila deltapsi <dpsi outdir>/<cond1_id>_<cond2_id>.deltapsi.pickle --genes-exp1 <build outdir>/<cond1>_dir --genes-exp2 <build outdir>/<cond2>_dir -o <voila outdir>
```

- `<dpsi outdir>/<cond1_id>_<cond2_id>.deltapsi.pickle` is the output file from delta PSI computation,
- `<build outdir>/<cond1>_dir` and `<build outdir>/<cond2>_dir` are the folders containing splicegraph files. Note that all files with *.splicegraph* extension will be consider. Alternatively a list of splicegraph files can be specified.

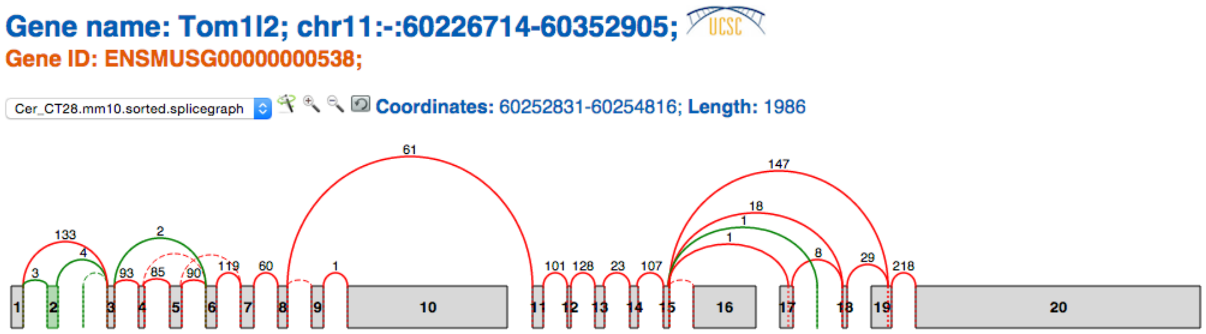
In the output directory `<voila outdir>` you will find:

- `summaries/xx_<cond1_id>_<cond2_id>_deltapsi_gene.html` files: interactive HTML5 summaries with MAJIQ quantifications.
- `<cond1_id>_<cond2_id>_deltapsi.txt` : A tab-delimited file with all the genes and LSV quantfication and genomic information.

VOILA

HTML5 Summaries

Splice Graphs



Gene name: Tom1l2; chr11:-:60226714-60352905;

Gene ID: ENSMUSG00000000538;

Cer_CT28.mm10.sorted.splicegraph

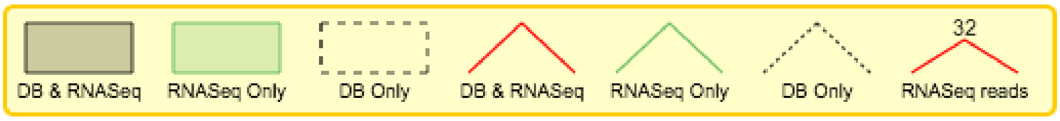
Coordinates: 60252831-60254816; Length: 1986

The splice graph gadget included in VOILA summarizes all the splice variants found in a gene by MAJIQ. Splice Graphs include the following features:

- Easy differentiation between exons and junctions annotated (red) and *de novo* detected (green) in RNA-Seq data,
- Contextual information about the coordinates for each exon and intron (hovering over exons/junctions).
- Raw reads counts for each junction.
- Scaled view of the gene and the splice graph . By default, introns are trimmed to obtain a more compact representation of the splice graph. Switching between scaled and default view is accomplished by clicking on the wand.
- Zoom in/out to explore complex splice graphs.

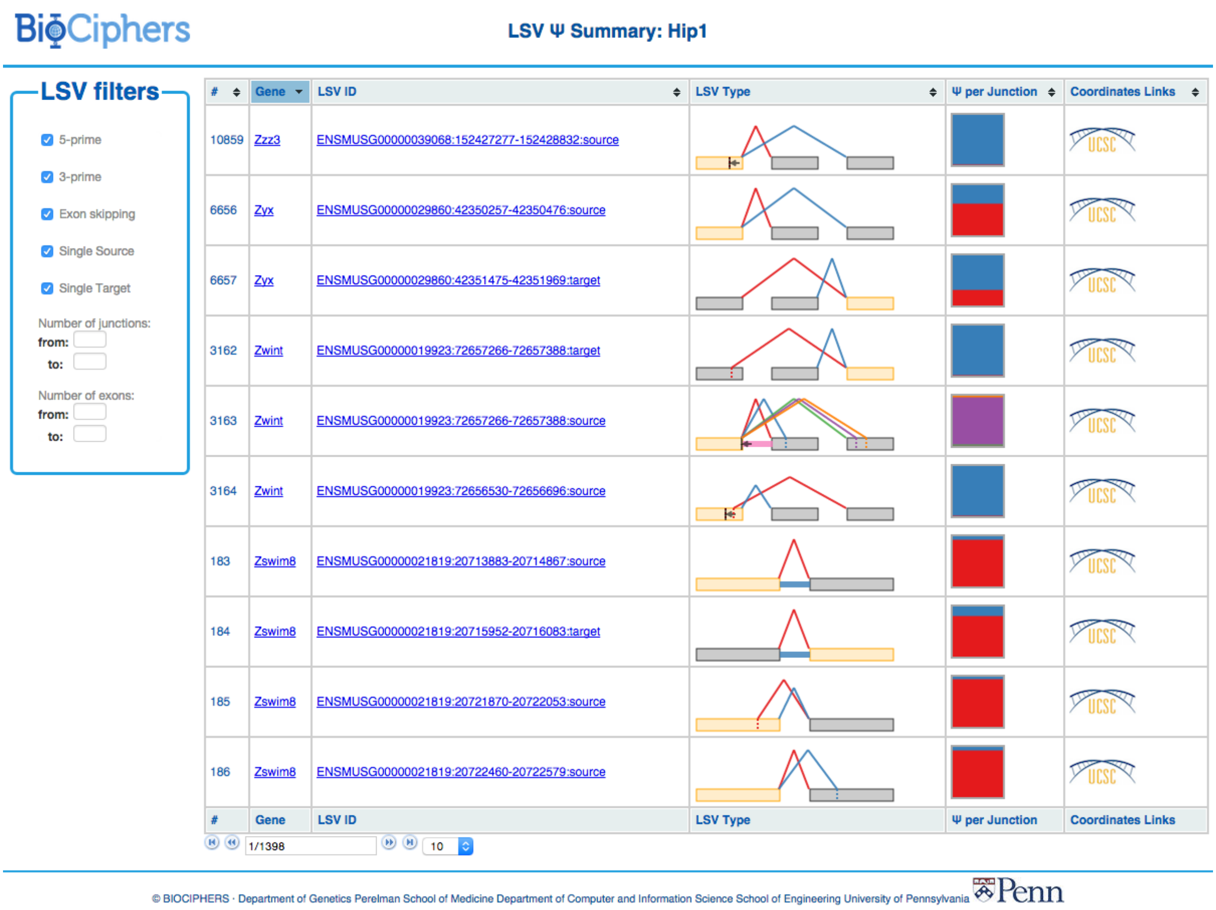
- Possibility to switch between replicates (condition members) when more than one splice graph is available.

All *Gene Summaries* include a descriptive legend of what you might find in the Splice Graphs:



> *DB* refers to exons and junctions annotated in the GFF file and *RNASeq* to exons and junctions found in RNA-Seq data. *RNASeq reads* alludes to the raw reads found in RNA-Seq data. Please, note that retained introns (narrow rectangles connecting two exons) do not appear currently in the legend.

PSI Summary



LSV Ψ Summary: Hip1

LSV filters

☑ 5-prime

☑ 3-prime

☑ Exon skipping

☑ Single Source

☑ Single Target

Number of junctions:
from:
to:

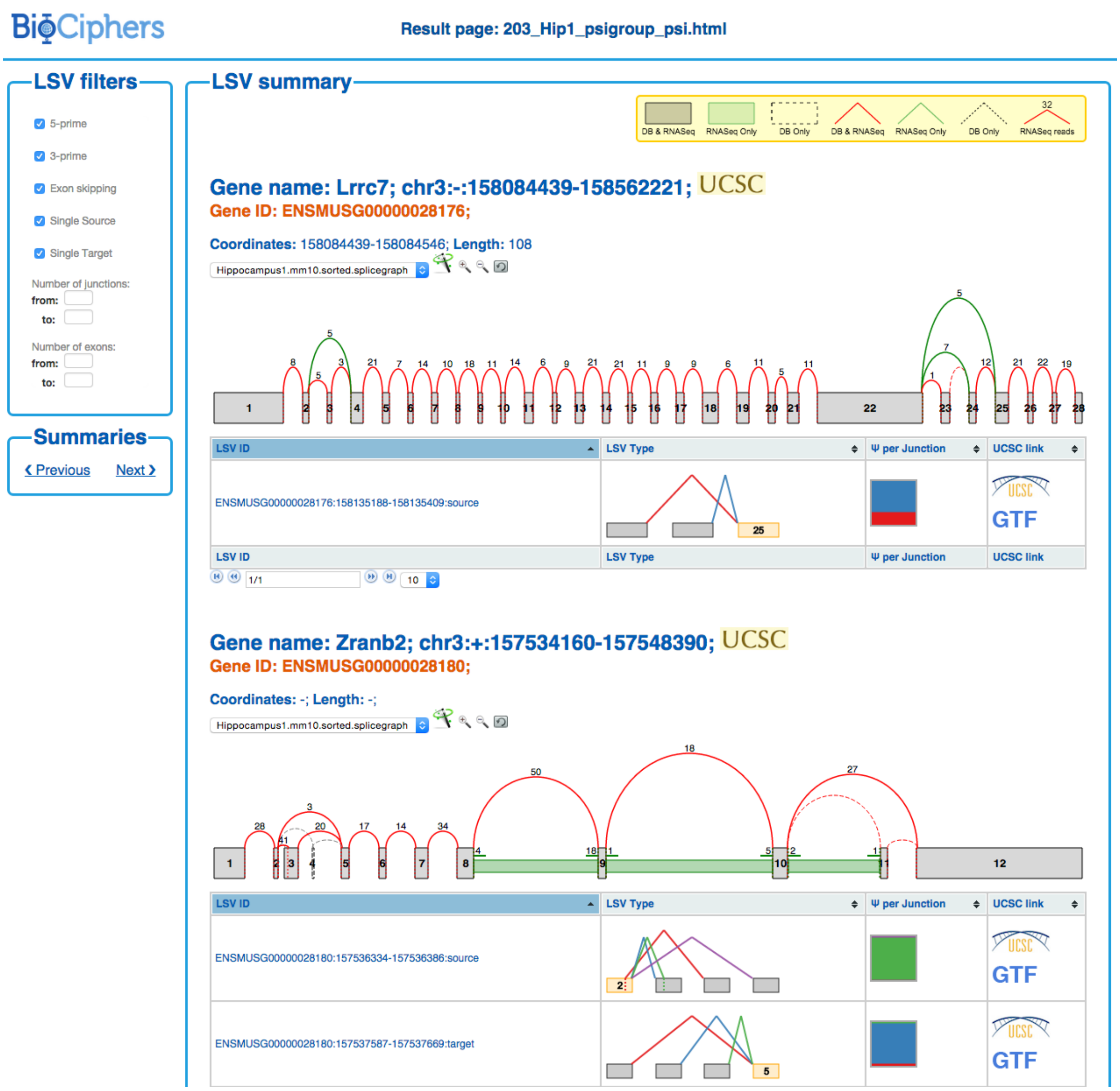
Number of exons:
from:
to:

#	Gene	LSV ID	LSV Type	Ψ per Junction	Coordinates Links
10859	Zzz3	ENSMUSG00000039068:152427277-152428832:source		🟦	UCSC
6656	Zyx	ENSMUSG00000029860:42350257-42350476:source		🟦 🟥	UCSC
6657	Zyx	ENSMUSG00000029860:42351475-42351969:target		🟦 🟥	UCSC
3162	Zwint	ENSMUSG00000019923:72657266-72657388:target		🟦	UCSC
3163	Zwint	ENSMUSG00000019923:72657266-72657388:source		🟪	UCSC
3164	Zwint	ENSMUSG00000019923:72656530-72656696:source		🟦	UCSC
183	Zswim8	ENSMUSG00000021819:20713883-20714867:source		🟥	UCSC
184	Zswim8	ENSMUSG00000021819:20715952-20716083:target		🟥	UCSC
185	Zswim8	ENSMUSG00000021819:20721870-20722053:source		🟥	UCSC
186	Zswim8	ENSMUSG00000021819:20722460-20722579:source		🟥	UCSC
#	Gene	LSV ID	LSV Type	Ψ per Junction	Coordinates Links

1/1398 10

© BIOCIPHERS · Department of Genetics Perelman School of Medicine Department of Computer and Information Science School of Engineering University of Pennsylvania Penn

VOILA PSI Index file offers an overview of all the LSVs detected in a table, providing links to detailed summaries of LSVs and genes. Clicking over a gene or LSV ID opens up a new tab with a summary containing interactive splice graphs, distributions of PSIs per junction and links to the UCSC. Below is an example of PSI summaries for *Lrrc7*:



Result page: 203_Hip1_psigroup_psi.html

LSV filters

☑ 5-prime
☑ 3-prime
☑ Exon skipping
☑ Single Source
☑ Single Target

Number of junctions:
from:
to:

Number of exons:
from:
to:

Summaries

‹ Previous Next ›

LSV summary

Gene name: Lrrc7; chr3:-:158084439-158562221; UCSC
Gene ID: ENSMUSG00000028176;
Coordinates: 158084439-158084546; Length: 108
Hippocampus1.mm10.sorted.splicegraph

LSV ID	LSV Type	Ψ per Junction	UCSC link
ENSMUSG00000028176:158135188-158135409:source	25	🟦🟥	UCSC GTF
LSV ID	LSV Type	Ψ per Junction	UCSC link

1/1 10

Gene name: Zranb2; chr3:+:157534160-157548390; UCSC
Gene ID: ENSMUSG00000028180;
Coordinates: -; Length: -;
Hippocampus1.mm10.sorted.splicegraph

LSV ID	LSV Type	Ψ per Junction	UCSC link
ENSMUSG00000028180:157536334-157536386:source	2	🟩	UCSC GTF
ENSMUSG00000028180:157537587-157537669:target	5	🟦	UCSC GTF

Tip: PSI Summaries can be navigated through the *Previous* and *Next* links, without having to go back to the *index* file.

The information is broken into genes (10 per page), each of them with an interactive splice graph and an associated table with LSV quantification data. The table has the following information about the LSV:

- LSV ID: a unique identifier for the LSV.
- LSV type: a thumbnail representing the splicing event. Each junction has a different color.
- PSI per junction: the expected Percentage of Splice Inclusion (PSI) per junction. It has

two *views* that can be switched between by clicking on them:

  - ***Compact view***. Initially, the PSI per junction is represented by the height of a colored box. Each color refers to the expected inclusion of a particular junction. The bigger the rectangle, the more included the junction is expected to be. Clicking on the rectangle (zoom in pointer) will open the *expanded view*.
  - ***Expanded view***. Distributions of probabilities of PSI per LSV junction (violin boxplots). The *white* dot represents the expected PSI, whereas the box plot indicates the 10, 25, 50, 75 and 90 percentile of the distribution. Consistently with the Compact view, each color refers to a particular junctions of the LSV.

- LSV links:
  - **GTF**. Link to the GTF file associated with the LSV.
  - **UCSC**. Coordinates link to explore the LSV on UCSC Genome Browser (when available).

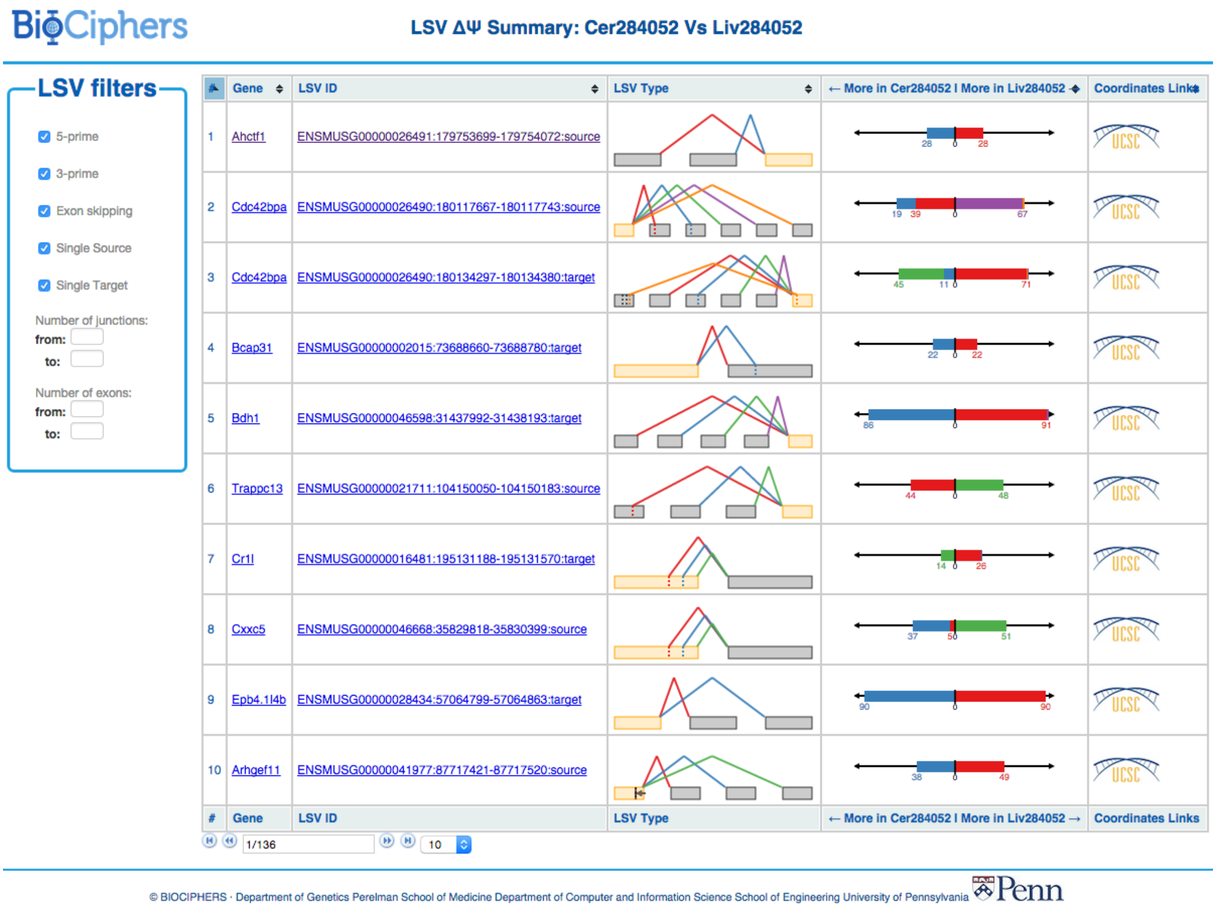
Lastly, the panel *LSV filters* allows the user to screen out LSVs with certain properties like having alternative 5-prime splice sites, involve Exon Skipping or contain a certain amount of exons and junctions.

Command line arguments:

```
positional arguments:
  majiq_output.pickle   Pickle file with the bins produced by
Majiq.
optional arguments:
  -h, --help            show this help message and exit
  -o output_dir, --output output_dir
                        Output directory where the files will be
placed.
  -c 0.95, --confidence 0.95
                        Percentage of confidence required (by
default, 0.95).
  --logger LOGGER        Path for the logger. Default is output
directory
  --silent              Silence the logger.
  --lsv-types LSV_TYPE1 [LSV_TYPE2 ...]
                        LSV type to filter the results. (If no gene
list is
                        provided, this option will display only
genes
                        containing LSVs of the specified type).
  --genes-exp1 Hippocampus1.splicegraph [Hippocampus2.splicegraph
...]
                        Splice graph information file(s) or
directory with
```

```
                        *.splicegraph file(s).
  --gene-names-file GENE_NAMES
                        File with gene names to filter the results
(one gene
                        per line). Use - to type in the gene names.
```

Delta PSI Summary



BioCiphers

LSV ΔΨ Summary: Cer284052 Vs Liv284052

LSV filters

☑ 5-prime
☑ 3-prime
☑ Exon skipping
☑ Single Source
☑ Single Target

Number of junctions:
from:
to:

Number of exons:
from:
to:

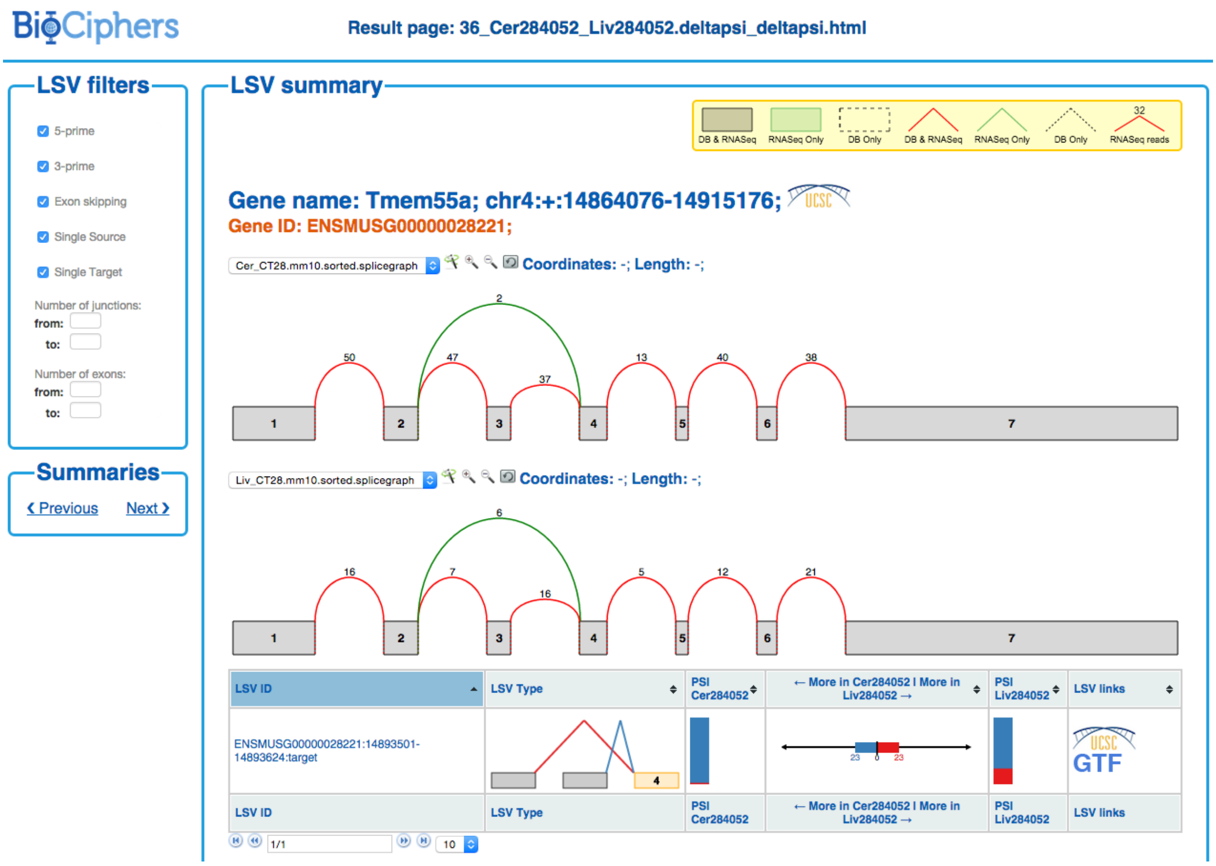
#	Gene	LSV ID	LSV Type	← More in Cer284052 I More in Liv284052	Coordinates Links
1	Ahctf1	ENSMUSG00000026491:179753699-179754072:source		28 0 28	UCSC
2	Cdc42bpa	ENSMUSG00000026490:180117667-180117743:source		19 39 0 67	UCSC
3	Cdc42bpa	ENSMUSG00000026490:180134297-180134380:target		45 11 0 71	UCSC
4	Bcap31	ENSMUSG00000002015:73688660-73688780:target		22 0 22	UCSC
5	Bdh1	ENSMUSG00000046598:31437992-31438193:target		86 0 91	UCSC
6	Trappc13	ENSMUSG00000021711:104150050-104150183:source		44 0 48	UCSC
7	Cr1l	ENSMUSG00000016481:195131188-195131570:target		14 0 26	UCSC
8	Cxxc5	ENSMUSG00000046668:35829818-35830399:source		37 50 51	UCSC
9	Epb4.1l4b	ENSMUSG00000028434:57064799-57064863:target		90 0 90	UCSC
10	Arhgef11	ENSMUSG00000041977:87717421-87717520:source		38 0 49	UCSC
#	Gene	LSV ID	LSV Type	← More in Cer284052 I More in Liv284052 →	Coordinates Links

1/136 10

© BIOCIPHERS · Department of Genetics Perelman School of Medicine Department of Computer and Information Science School of Engineering University of Pennsylvania Penn

> **Tip**: index tables can be sorted by *gene* name, *LSV Type* and/or *most changing junction* clicking on the column header respectively. It is possible to sort the table by multiple columns holding the *shift* key.

Delta PSI index file is identical to PSI index except for the *PSI per junction* column that represents the estimated differential inclusion levels as bars leaning towards condition1 or condition2 (see *Compact view* below). Clicking over LSV IDs and genes opens up individual summaries.



The results are broken into 10-genes per page summaries, with LSV quantifications group by gene. Unlike in single PSI summaries, there are 2 splice graphs (one per condition) which facilitate quick visual inspection of possible differences. If multiple replicates were used in a condition, the splice graphs for each replicate can be switched between through the drop down box located next to the wand. The LSV information is displayed as follows:

- LSV ID: a unique identifier for the LSV.
- LSV type: a thumbnail representing the splicing event. Each junction has a different color.
- PSI *condition[1|2]*: individual Ψ in *condition[1|]2*, similar to the Ψ *per junction* column in PSI summary, with compact and expanded views.
- More in *condition 1* | More in *condition 2*: the expected delta PSI for all junctions. It has two *views* which can be switched between by clicking on them:
    - ***Compact view***. Initially, the observed delta PSI preference for a certain condition (more included) per junction. A bar going towards *condition 1* (left) with a value on *25* means that the junction is 25% more included in condition 1 than in condition 2. Each color refers to a particular junction of the LSV. In the above example, the red junction is 46% more included in hippocampus compared to liver.
    - ***Expanded view***. Distributions of probabilities of Delta PSI per LSV junction (violin boxplots). The *white* dot represents the expected PSI, whereas the box plot indicates the 10, 25, 50, 75 and 90 percentile of the distribution. Consistent with

the Compact view, each color refers to a particular junction of the LSV.

- LSV links:
    - **GTF**. Link to the GTF file associated with the LSV.
    - **UCSC**. Coordinates link to explore the LSV on UCSC Genome Browser (when available).

Command line arguments:

```
positional arguments:
  majiq_output.pickle   Pickle file with the bins produced by
Majiq.

optional arguments:
  -h, --help            show this help message and exit
  -o output_dir, --output output_dir
                        Output directory where the files will be
placed.
  -c 0.95, --confidence 0.95
                        Percentage of confidence required (by
default, 0.95).
  --logger LOGGER        Path for the logger. Default is output
directory
  --silent              Silence the logger.
  --threshold THRESHOLD
                        Filter out LSVs with no junction predicted
to change
                        over a certain value (in percentage).
  --show-all            Show all LSVs including those with no
junction with
                        significant change predicted
  --pairwise-dir PAIRWISE
                        Directory with the pairwise comparisons.
  --genes-exp1 Hippocampus1.splicegraph [Hippocampus2.splicegraph
...]
                        Experiment 1 splice graph information
file(s) or
                        directory.
  --genes-exp2 Liver1.splicegraph [Liver2.splicegraph ...]
                        Experiment 2 splice graph information
file(s) or
                        directory.
  --gene-names-file GENE_NAMES
                        File with gene names to filter the results
(one gene
                        per line). Use - to type in the gene names.
  --lsv-types LSV_TYPE1 [LSV_TYPE2 ...]
                        LSV type(s) used to filter the results. (If
no gene list is
                        provided, this option will display only
genes
                        containing LSVs of the specified type).
```

Tab-delimited file

VOILA provides a tab-delimited text file to allow users to parse MAJIQ results and further analyze particular LSVs or genes of interest. Most fields are shared between single PSI and delta PSI computations for the expected values and the confidence measures (variance in the case of single PSI and the probability of delta psi > 0.2 in delta PSI analysis). The common fields are: Gene name; LSV ID; LSV Type; LSV attributes (A5SS, A3SS, ES, Num. Junctions and Num. Exons); chromosome; strand; LSV coordinates (junctions and exons coordinates); and finally, if additional evidence is required to determine what is the start/end of an LSV, a list with all possible alternative starts and ends is provided.

Installation

The compiled (built) versions of MAJIQ of VOILA can be found in `TODO: URL to download the files???`. At this point, MAJIQ only works on Linux, whereas VOILA can be executed both in Linux and Mac.

Please note that *MAJIQ* is computationally-intensive (we recommend at least 8Gb of RAM and 4CPUs).

MAJIQ has the following dependencies:

- Numpy 1.8 (or greater),
- Matplotlib 1.1.1 (or greater),
- Scipy 0.11.0 (or greater),
- pysam 0.8.0 (or greater)

- Jinja2 2.7.2 (or greater)
- argparse 1.1 (or greater)

FAQ

In VOILA gene summaries, what is that *number at the beginning* of the HTML file?

To achieve a better performance, VOILA creates HTML files of up to 10 genes. Therefore, if MAJIQ detect and quantify LSVs from N genes, there will be N/10 pages (always rounded to the upper integer limit).
For example, let say that we executed `voila psi data/Liver1.majiq_psi.pickle --genes-exp1 data/Liver1.splicegraph -o psi_gene_out/` and 182 genes were detected. There will be 182/10=19 pages (starting with 0): 0_Liver1.majiq_psi_lsv_single_gene.html, 1_Liver1.majiq_psi_lsv_single_gene.html, …, 18_Liver1.majiq_psi_lsv_single_gene.html.