Pipeline to identify alternative splicing using MISO program

by Tatiana Garcia & Meena Sudhakaran

Required programs

-Fastqc
-Trimmomatic
-MISO
-Mapping program. Hisat2 is suggested
-The Integrative Genomics Viewer (IGV)
-Samtools
-Picard

Before start

-Verify the correct operation of the programs.
-Remember to use the latest version of the programs and verify the parameters; these may change according to the version.

Files	Description
hg19HumanGenome.fa	Reference genome in fasta format
hg19HumanGenome.fa.fai	Genome index generated by the samtools (faidx command)
hg19Human.gene.gff3.gz	Gene annotation in GFF3 format
hollywood.mit.edu/burgelab/miso/annotations/miso_annotations_hg19_v1.zip	MISO annotation v1
http://hollywood.mit.edu/burgelab/miso/annotations/ver2/	MISO annotation v2

Download the data using AWS

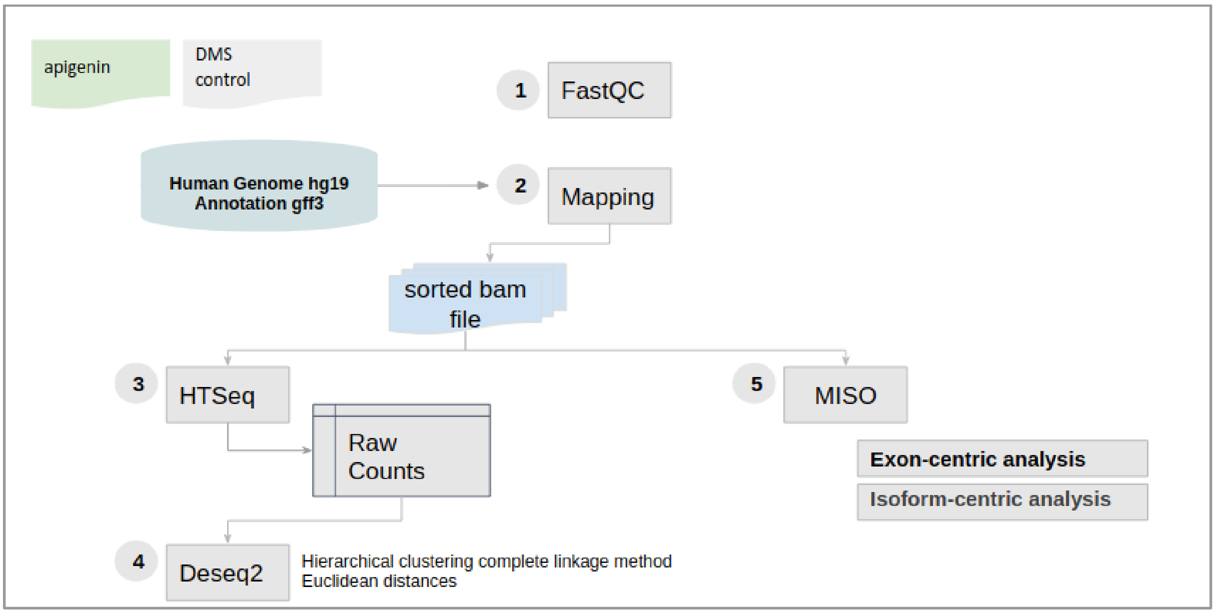
To download the data from the sequencing company, you need :
Install AWS on your computer or the server https://docs.aws.amazon.com/cli/latest/userguide/getting-started-install.html

- You must enter the command aws configure. This command will ask you for the data provided by the sequencing company as the path of the data and the key.
- Then you put in the command line:

```
aws s3 cp s3://bucketname/path /localfolder --recursive.
```

If everything is configured correctly, the download should start.

The company sends a .txt file with codes for each sequencing file. When the download is complete, remember to verify the integrity of the files with the command md5sum. it is advantageous to confirm that a file has not changed due to a faulty file transfer, a disk error, or non-malicious meddling.



Step 1. Quality control and removal of low-quality reads

An initial evaluation of the reads obtained from the sequencing company should be performed to identify possible adapters and low-quality reads. For this, **script_quality.bash** that uses the **Fastqc** program should be run. Remember to change the samplesID.txt file with the name of your samples and path.
Additional information about the modules that are evaluated is found in the following link:
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

According to the quality assessment results, sometimes data cleaning is required. There are different tools for this process, such as cutAdapt or trimmomatic. In the cleaning.bash script file, there is an example of clean data with trimmomatic
In the **script_cleaning.bash** file.
Remember that the parameters must be adjusted according to the quality analysis result. After cleaning the data, recheck the files by running fastqc program.
all parameters for trimmomatic are available in:
http://www.usadellab.org/cms/?page=trimmomatic

Step 2. Mapping process

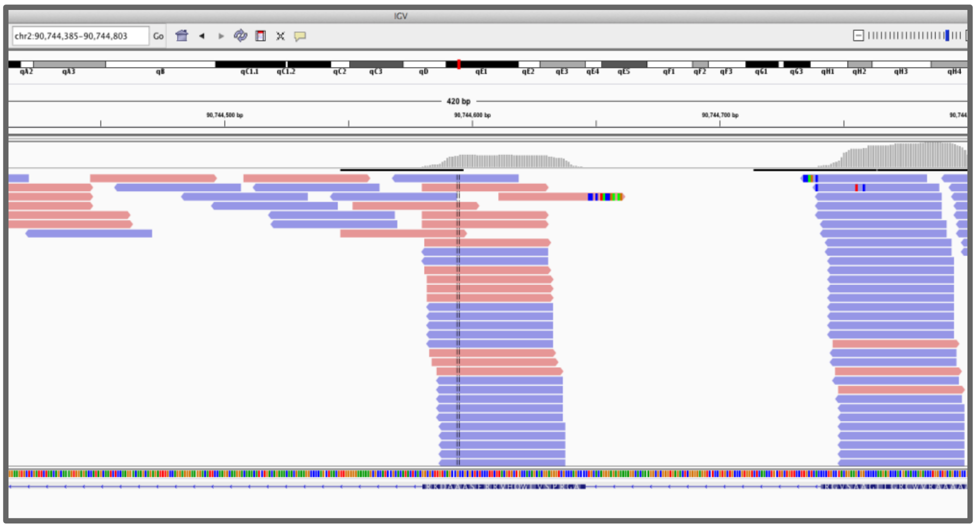
When you clean all sequencing files, move on to the mapping process. But first, you should Index the reference genome. This step will generate a data structure around the reference genome to align efficiently reads. Each program generates its specific index. For data management, it is a good practice to use the basename of the index according to the reference genome. Please check script_index_Hisat.bash for this step. When the program is finished, several files will be generated with the extension *.ht2.

Now, is it possible to Mapping the reads using the script_mapping.bash. This process will take several hours depending on the size of the sequencing files and the assigned threads. When getting the bam files remember to sort them. You can use your preferred tool like samtools or picard.

Another example of the mapping script is provided for hisat2 (see script_mapping_hisat2.bash). Note that this script has the step of ordering the reads using picard incorporated. At the end of the script you will obtain your mapping files ordered and with the mapping statistics. If you use hisat2 remember to add the spliced alignment options according to your aim. (http://daehwankimlab.github.io/hisat2/manual/)

Step 3. Alignment visualization

Different tools allow you to view alignments more graphically and interactively. We will use the Integrative Genomics Viewer (IGV) tool for this. Start IGV by clicking igv.bat (windows), igv.sh (Linux) or igv.command (MAC). Then, the reference genome must be loaded by clicking on the Genomes option, → Load Genome from file" and selecting the fasta file with the reference. To view the .bam files generated by the alignment process, select "File → Open" from the menu and choose "Apigenin1_sorted.bam" file. To see the alignments, you must zoom in on some region of the genome.



Step 4. Before running miso

When using paired-end reads in MISO, you need to know the mean and standard deviation of insert length distribution.This is used to assign reads to isoforms probabilistically. For this you need to run the script_exon_utils.bash. According to the MISO documentation if the insert length selected for in preparing the RNA-Seq library was roughly 250-300 nt, we can use constitutive exons that are at least 1000 bases long, remember to use the same GFF file used in the mapping process. This step will output a GFF file named human.min_1000.const_exons.gff into the exons directory containing only constitutive exons that are at least 1000 bases long.

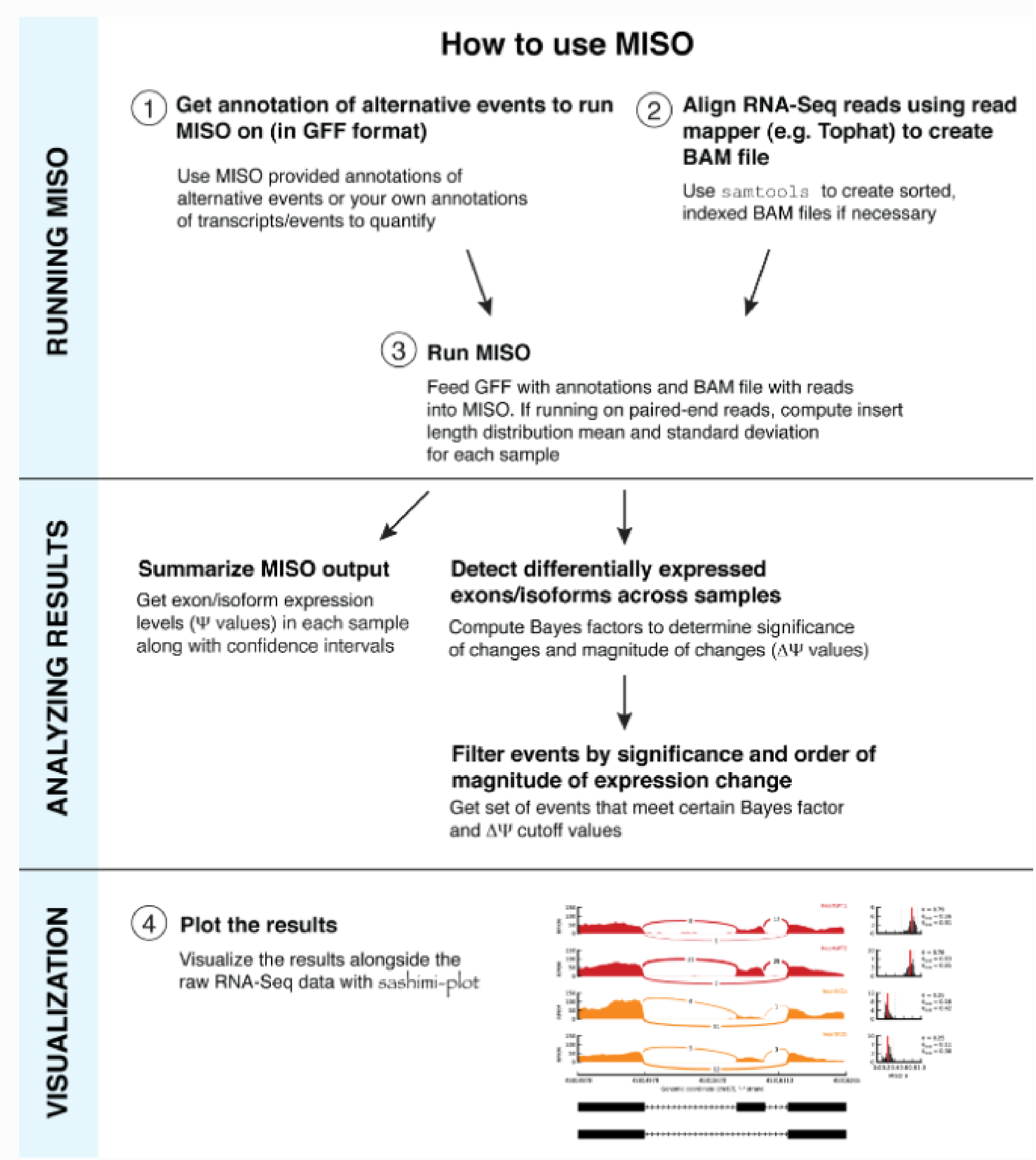
Then, this file can be used to compute the insert length distribution of all bam files with the script_pe_utils.bash. For each file you will obtain a file where the first line (header) gives the mean, standard deviation and dispersion values for the distribution, for example: #mean=129.0,sdev=12.1,dispersion=1.1,num_pairs=862148

Step 5. MISO (Mixture of Isoforms)

Miso is a probabilistic framework that quantitates the expression level of alternatively spliced genes from RNA-Seq data, and identifies differentially regulated isoforms or exons across samples.

MISO needs to prepare the alternative isoforms annotation whose expression should be estimated from data for whole mRNA transcripts (“isoform-centric” analysis).For this please use the script_miso_step1.bash. Here you should use the same GFF that can be given as input to the mapping step for incorporating known genomic junctions. For exon-centric analyses please use the GFF file for each kind of alternative splicing (SE,RI,A5SS,A3SS,AFE,ALE and MXE) to generate the indexed directory that is then used as input to MISO.

When indexing finishes, use the scriptRunmiso.bash to identify the splicing events in each case. Then, run the script script_summarize_control_api.bash to obtain a summary of each event. To finish, run the script_comparison_control_api.bash which performs paired comparisons between samples to detect differentially expressed isoforms/events. All files can be imported into R to continue the analysis.



How to use MISO

RUNNING MISO

1. **Get annotation of alternative events to run MISO on (in GFF format)**

   Use MISO provided annotations of alternative events or your own annotations of transcripts/events to quantify

2. **Align RNA-Seq reads using read mapper (e.g. Tophat) to create BAM file**

   Use `samtools` to create sorted, indexed BAM files if necessary

3. **Run MISO**

   Feed GFF with annotations and BAM file with reads into MISO. If running on paired-end reads, compute insert length distribution mean and standard deviation for each sample

ANALYZING RESULTS

Summarize MISO output

Get exon/isoform expression levels (Ψ values) in each sample along with confidence intervals

Detect differentially expressed exons/isoforms across samples

Compute Bayes factors to determine significance of changes and magnitude of changes (ΔΨ values)

Filter events by significance and order of magnitude of expression change

Get set of events that meet certain Bayes factor and ΔΨ cutoff values

VISUALIZATION

4. **Plot the results**

   Visualize the results alongside the raw RNA-Seq data with sashimi-plot