ITMO Ecole de Bioinformatique
Hands-on session: smallRNA-seq
N. Servant

21rd November 2013

1. Data and objectives

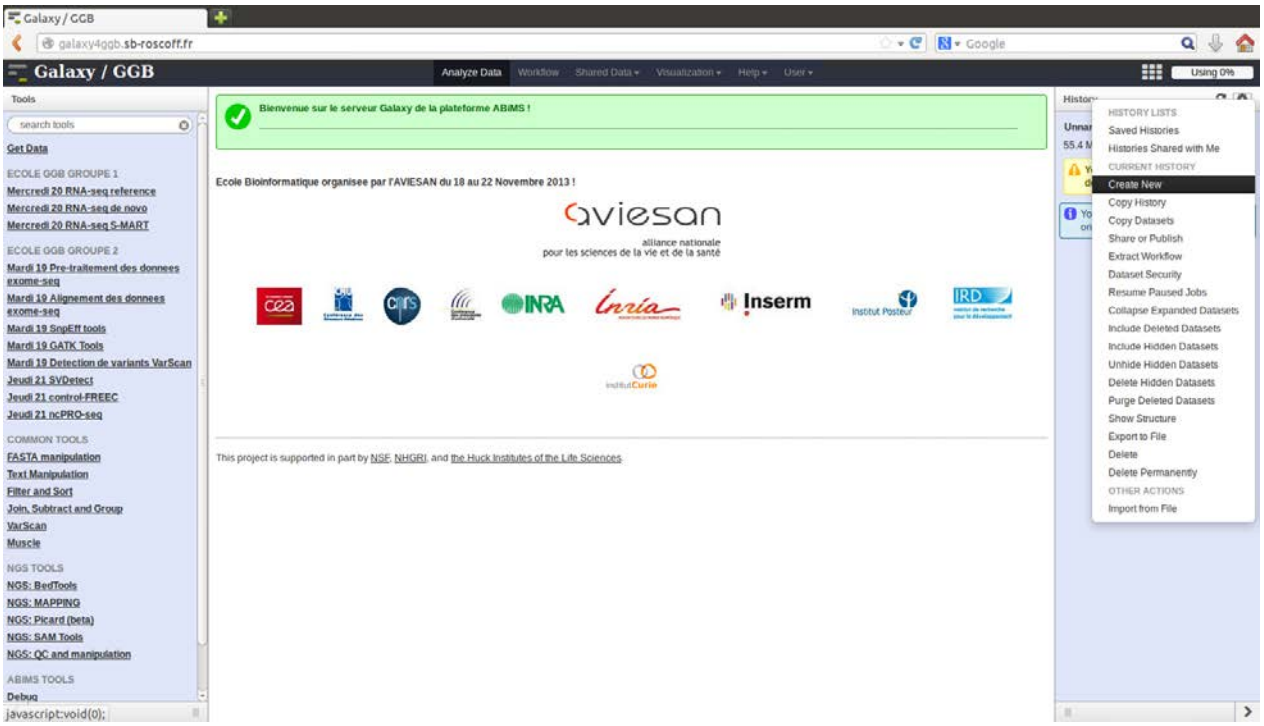
We will use the data from GEO (GSE35368, Toedling, Servant et al. 2011).
Two samples were selected, one from male ES cells and one from female ES cells. The goal of this session is to manage raw small RNA-seq data, from quality control to miRNAs quantification.
Some information about the data:

- Illumina sequencing, single-ends 1x50 bp
- 3' Adapter sequence : CTGTAGGCACCATCAATCGTA
- Random selection of 100 000 reads

The objective of the session is to quantify the miRNAs expressed in Mouse male and female ES cells.

2. Open Galaxy (http://galaxy4ggb.sb-roscoff.fr/) and load the data

→ Create a new history: choose a name (for example “miRNAseq”):
History → Create New

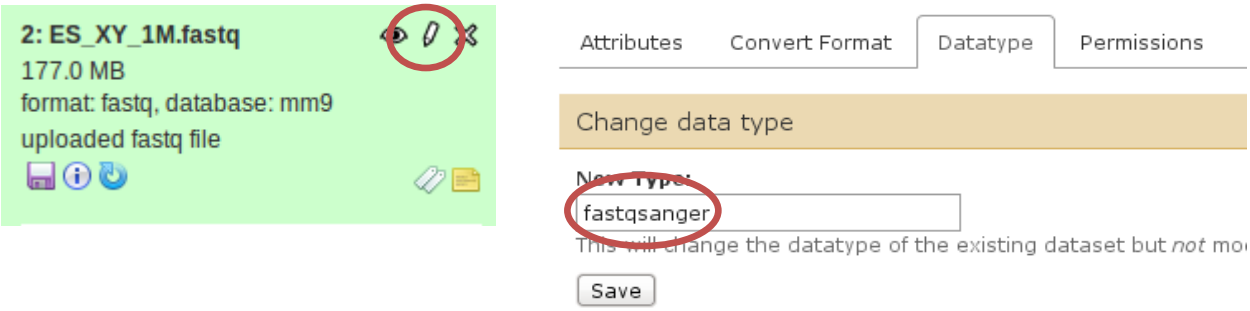


→ Go to **Shared Data → Data Libraries → tp-jeudi-ncpro-nicolas**
→ Select the following files and import them in your current history.

- ES.XX_100K.fastq
- ES.XY_100K.fastq

These files are raw reads in fastq format. We random select 100 000 reads for this training.

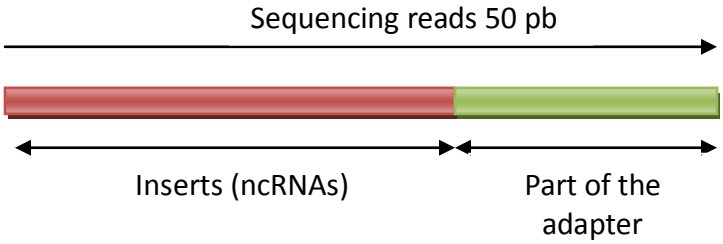
→ Change "fastq" format to "fastqsanger" for each file .fastq.
Click on Edit attributes → Datatype



This is related to the quality format of the raw reads. Some tools require a specific quality format to analyze data …

2. Adapter removal (cutadapt)

At this stage of the analysis, our reads have a size of 50nt whereas most of the ncRNAs have a size smaller than 50nt. This is explained by the fact that part of your sequence represents the 3' Adapter.



Thus, the first task of a smallRNA-seq analysis is to remove this adapter sequence in order to directly work with the inserts (of variable size).

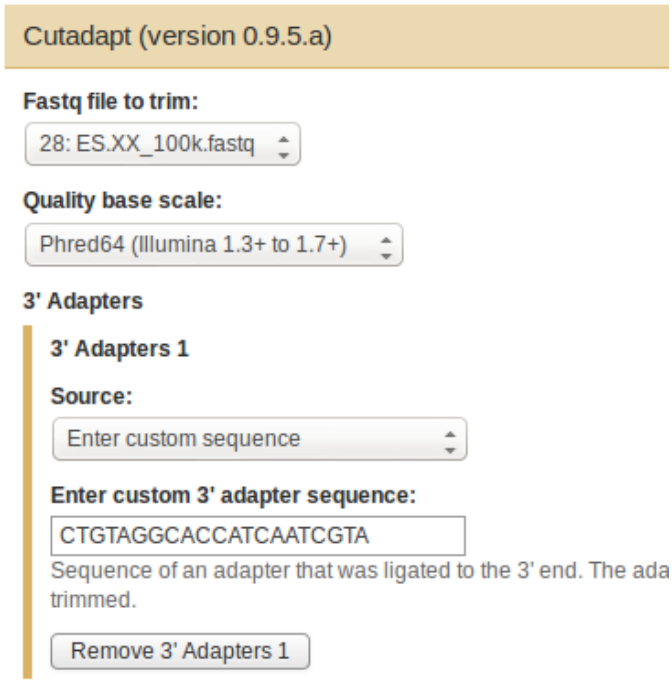
→ Mercredi 20 RNA-seq de novo→ Cutadapt

Cutadapt aims at removing an adapter sequence from 5' and/or 3' ends.
A few parameters have to be set:

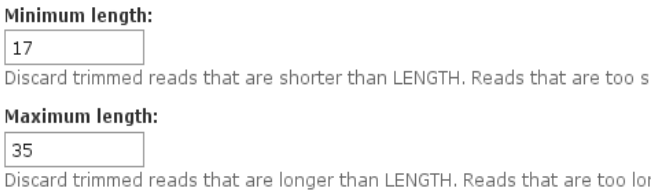
- Change the quality format to Phred64 (Illumina 1.3 format)

→ Add new 3' Adapters

- Enter a custom sequence of 3’ adapter to trim (CTGTAGGCACCATCAATCGTA).



- Set the minimum and maximum insert size to report (Minimum length = 17, Maximum length = 35). This has to be set according to your project and the ncRNAs population you want to focus on.



Note: In any case, do not keep the default parameter (min size = 0). Otherwise, your trimmed fastq file will contain empty lines for adapter reads …

- Look at the cutadapt report to have an idea of the number of trimmed reads. It gives you a first quality control of your dataset.

Key Points:

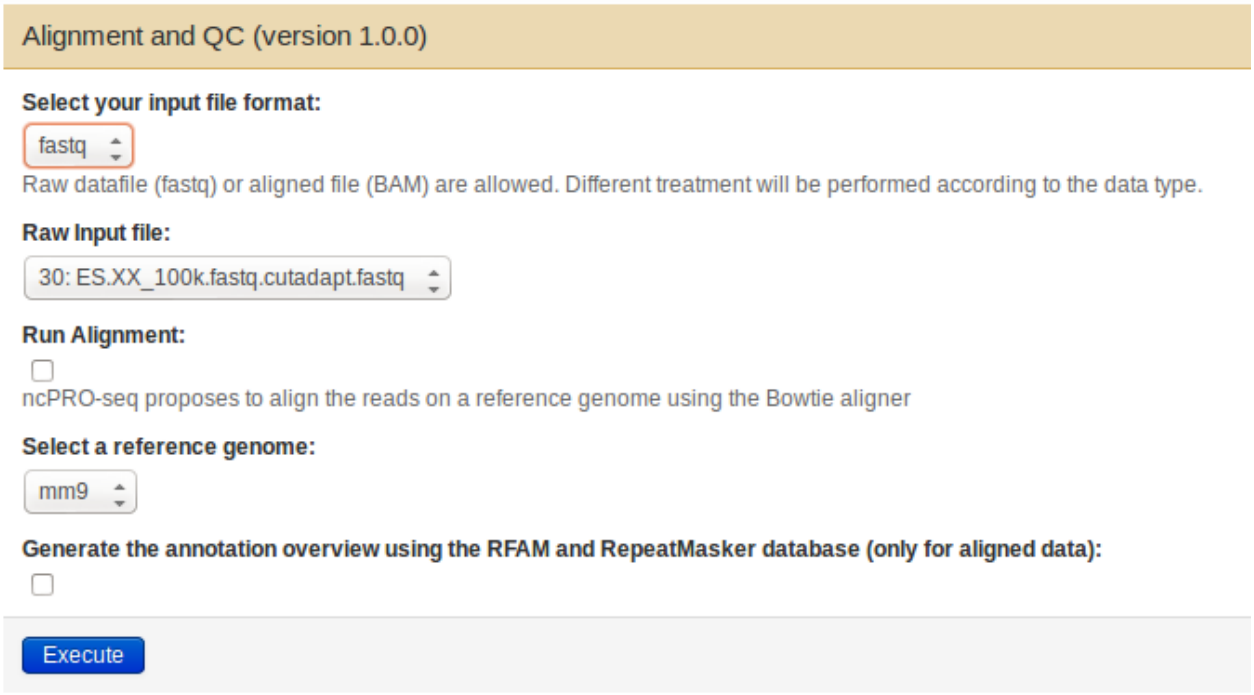
- Check the number of trimmed reads from the cutadapt report

3. Raw reads quality control (ncPRO-seq)

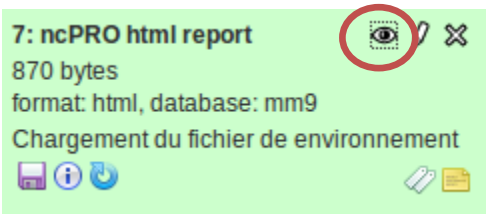
ncPRO-seq (Chen C., Servant N. et al 2012) is a comprehensive and flexible ncRNA analysis pipeline (Non-Coding RNA PROfiling in sRNA-seq, http://ncpro.curie.fr/), which is able to interrogate and perform detailed analysis on small RNAs derived from annotated non-coding regions in miRBase, Rfam and repeatMasker. The ncPRO-seq pipeline supports input read sequences in fastq, fasta and color space format, as well as alignment results in BAM format, meaning that small RNA raw data from the 3 current major platforms (Roche-454, Illumina-Solexa and Life technologies-SOLiD) could be analyzed with this pipeline. Finally, the ncPRO-seq pipeline can be used to analyze data based on genome from metazoan to plants.

Now that we have removed our adapter sequence, we can go deeper in the quality control of our samples.

→ Jeudi 21 ncPRO-seq → ncPRO QC
Select the trimmed reads and run the quality control.



A quality report is then available at the end of the analysis. To see it, just click on the eyes icon.



The following information is available:

- **Base Composition Information**

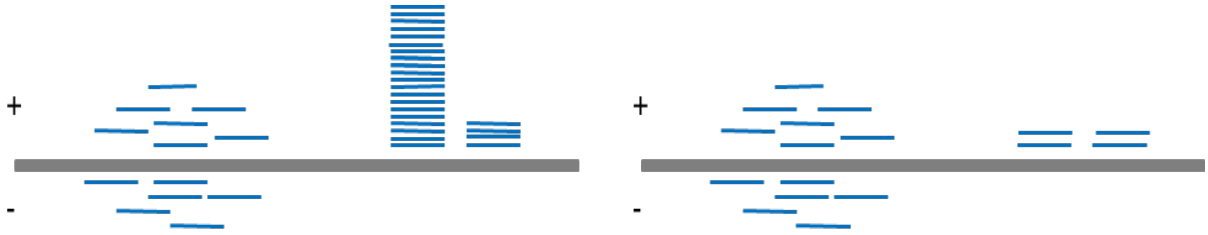
Display the proportion of each base position for which each of the four normal DNA bases has been called (or GC content). If you see strong biases which change in different bases then this usually indicates an overrepresented sequence which is contaminating your library. A bias which is consistent across all bases either indicates that the original library was sequence biased, or that there was a systematic problem during the sequencing of the library.

- **Quality Score**

This view presents the quality values across all bases at each position in the FastQ file.
The y-axis on the graph shows the mean quality scores. The higher the score the better the base call. The quality of calls on most platforms will degrade as the run progresses, so it is common to see base calls falling into the orange area towards the end of a read.
We usually consider as good quality, the data with a mean quality higher than 20.

- **Reads Length Distribution**

The insert size distribution is the most important quality control in sRNA-seq data. ncPRO-seq provides two types of information, i.e. the abundant versus the distinct reads length distribution.



Abundant Distinct

The abundant distribution considers all reads as they are described in the fastq file. The distinct distribution merges all duplicated sequence as one. This view usually decreases the importance of miRNAs to highlight other population-based ncRNAs.

Key Points:
- Look at the reads quality
- Look at inserts size distribution (abundant vs distinct)

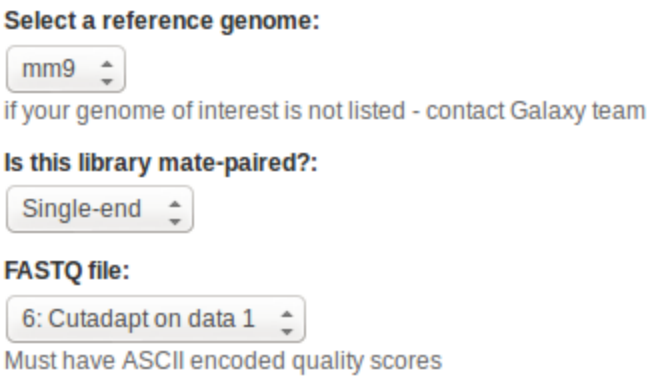
Bonus:
- Run the FastQC tool and compare the results (**→ Mercredi 20 RNA-seq de novo → FastQC**)

4. Reads alignment (Bowtie)

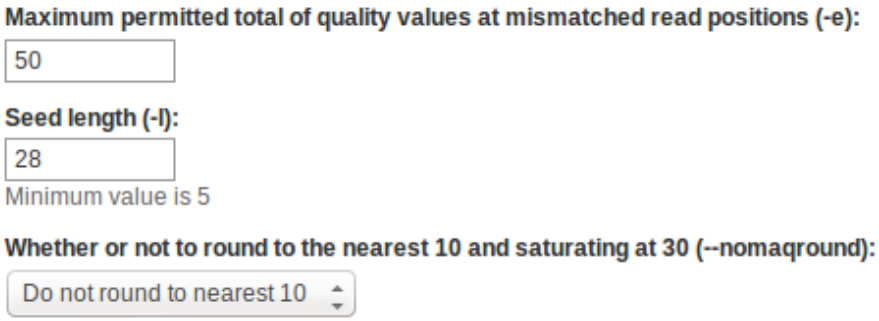
The ncPRO-seq pipeline integrates a mapping step. However, this is still useful to do it by yourself to understand which parameter you are using. Bowtie is a widely used aligner for NGS data. It is extremely fast for common mapping as for the sRNA-seq data.

→ NGS: Mapping → Map with Bowtie for Illumina
Look at the manual available on the Bowtie's Galaxy page for details and run the alignment.

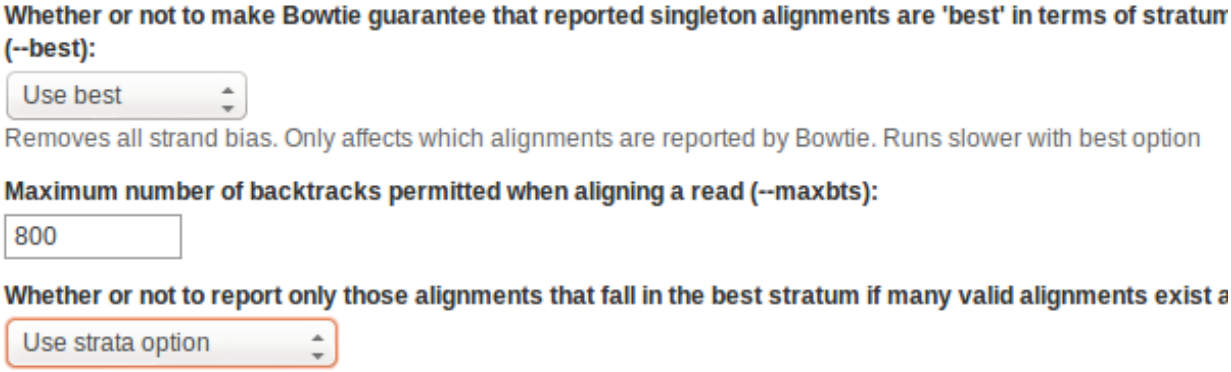
- **Common parameters**: select your input file, your reference genome, and your library type.



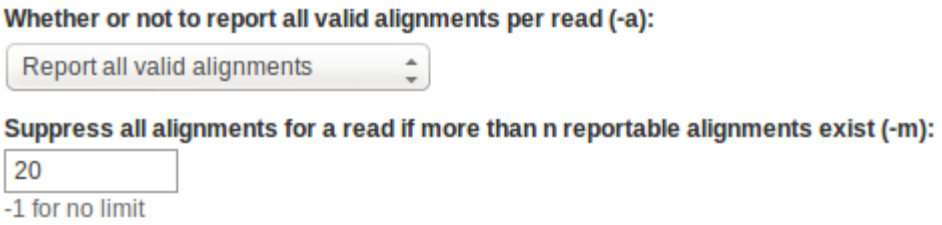
- **Number of mismatch** (-e 50 –nomaqround): Error rate allowed during the mapping procedure.



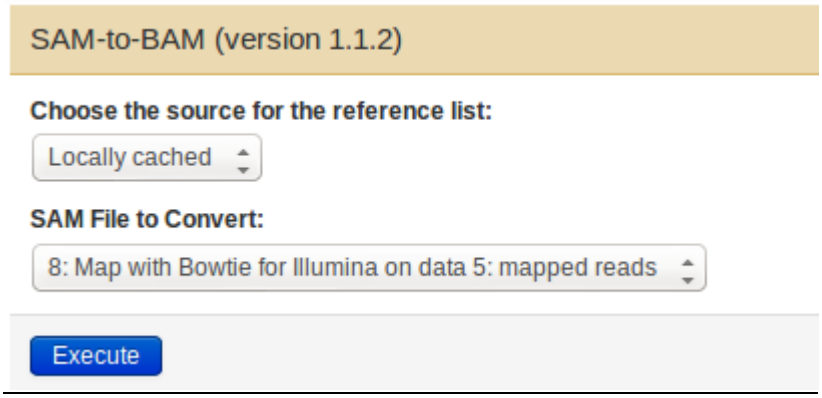
- **Reporting** (-- best -- strata): report only the best alignment for a given read.



- **Multiple hits** (-a –m 20): report reads that align up to 20 times on the genome.



Bowtie creates a SAM file in output with your aligned reads.
You should now transform your SAM file in BAM file which is the common format for most NGS tools.
→ NGS: SAM Tools → SAM-to-BAM

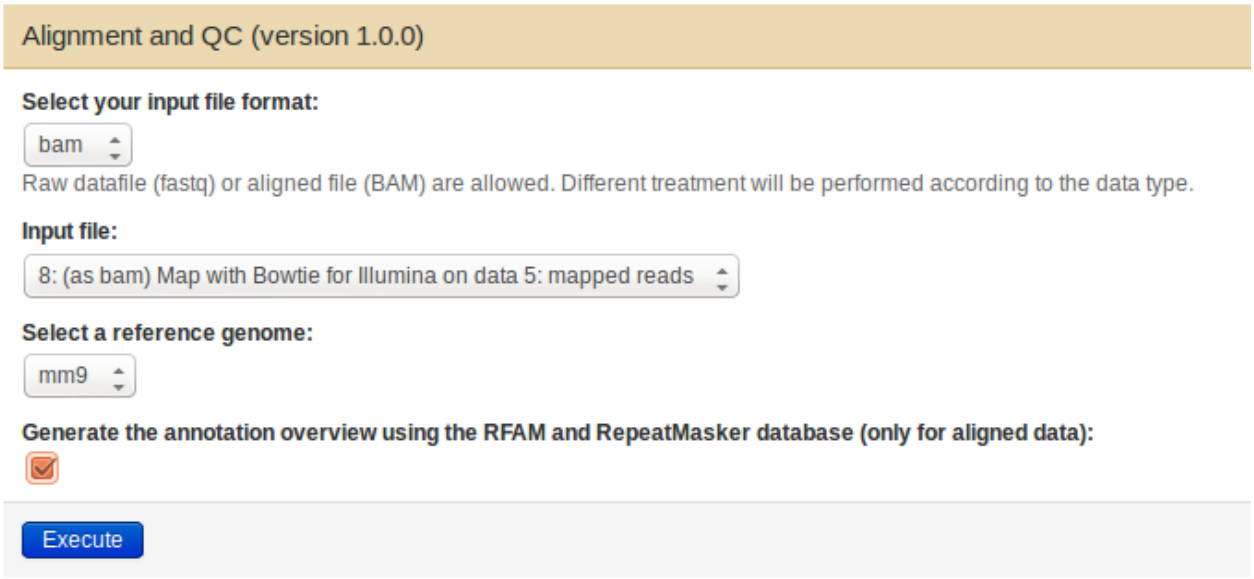


Key Points:
- Read the Bowtie manual

Bonus:
- Look at your BAM file using the IGV browser

5. Quality control of aligned reads

Go back to the ncPRO-seq quality control and now run it on mapped reads. In the case of aligned reads, check the annotation overview in order to have an idea of the ncRNAs population from your sample.
→ Jeudi 21 ncPRO-seq → ncPRO QC



Additional quality reports are available with the following information:

- **Mapping statistics**

The proportions of reads with unique, multiple mapping sites in the genome, and unmapped reads is plotted. For sRNA-seq data, we usually expect to have a large proportion of unique hits.

- **Annotation overview**

The reads annotation family is the most general overview, and counts the reads based on the following annotations: coding genes, ncRNAs from Rfam, smallRNAs from repeated regions, rRNAs, and precursor miRNAs from miRBase.

- **miRNA reads proportion (miRBase)**

A dedicated plot is available for pre-miRNAs. In this step, abundant reads mapped in mature miRNA regions are counted, and plotted as the proportion of all mapped reads in the genome. The annotation file of mature miRNA is generated using files from miRBase. Each miRNA count is calculated using the intersection of the reads alignment with the precursor position.
In a classical sRNA-seq experiment, we usually expect to have a high level of miRNAs (around 70%). This information can be used as a quality control for mammals. If a small proportion of miRNAs is observed, it means that another population of ncRNA predominates. This can be real biological information, or a contamination (tRNA, rRNA, etc.)

- **ncRNA annotation (RFAM)**

To compare the read expression in different repeat/Rfam families, we count the number of abundant reads in each family and plot the relative proportion.
We catalogue non-coding RNA genes in Rfam annotation into five big classes: tRNA, rRNA, snRNA,

snoRNA and others. Note that miRNA annotations are excluded in the Rfam noncoding RNA analyses to be replaced by the miRBase annotation.

- **Repeats annotation (RepeatMasker)**

ncPRO-seq uses repeat annotations from RepeatMasker database. We classify different repeats based on the name of repeat family.

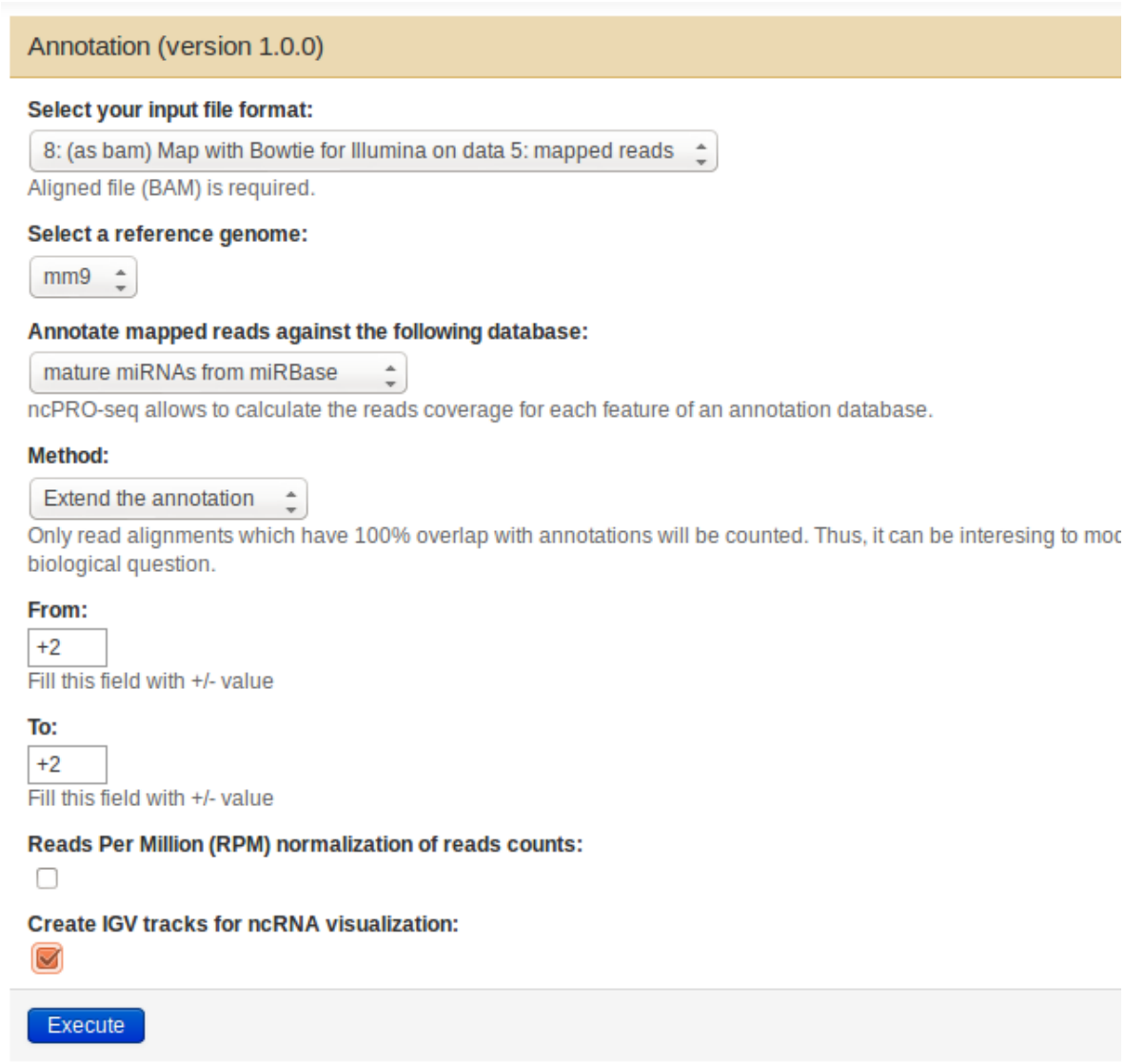
Key Points:
- Look the quality control of aligned reads
- Look at inserts size distribution (abundant vs distinct)
- Look at the mapping statistics
- Comment your annotation profile and the level of miRNAs on your sample

6. miRNA Annotation

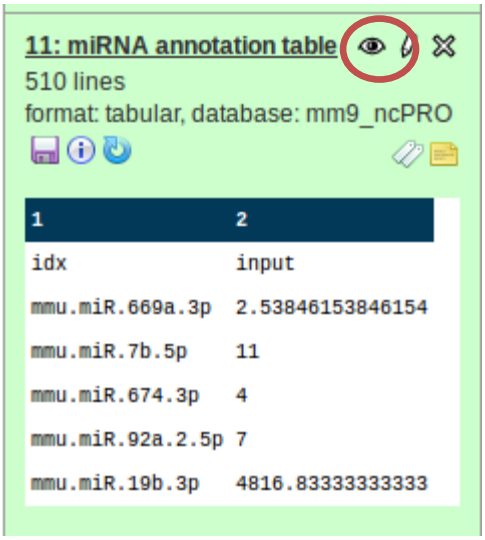
A major challenging problem using NGS sequencing data is the annotation of reads aligned at multiple locations. Most of the available frameworks resolve this situation by discarding these reads or by providing random annotations. Here, we propose to keep all the reads aligned to the genome, and to weight them by the number of mapping sites. Suppose a read can be aligned 5 times to the genome, for each mapping site, the read would be counted as 0.2, i.e. 1/5

There are four types of extended items which can be used to modify the annotation coordinates:
- **extend** [+-] *n1* bp at 5' end, [+-] *n2* bp at 3' end
- **shorten** [+-] *n1* bp at 5' end, [+-] *n2* bp at 3' end
- **Focus on the 5' end**, get coordinates for sub-region from position *n1* to *n2* indexed from 5' end
- **Focus on the 3' end**, get coordinates for sub-region from position *n1* to *n2* indexed from 3' end

→ Jeudi 21 ncPRO-seq → annotation of microRNAs



You can generate both annotations for mature and precursor miRNAs.
Two different outputs are available. For all miRNAs, ncPRO-seq creates a table file showing read coverage of each single miRNA in the sequencing library. These table files are almost compatible with R package like DESeq to identify significantly expressed family members.



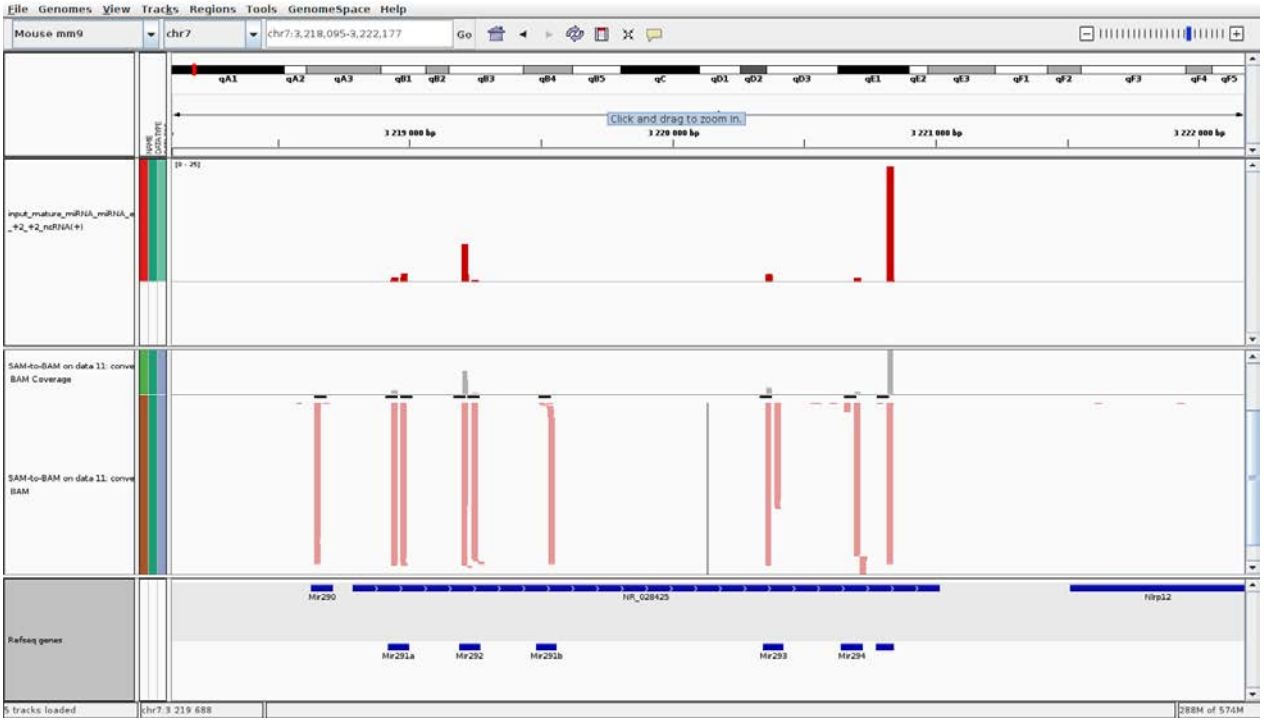
11: miRNA annotation table

510 lines

format: tabular, database: mm9_ncPRO

1	2
idx	input
mmu.miR.669a.3p	2.53846153846154
mmu.miR.7b.5p	11
mmu.miR.674.3p	4
mmu.miR.92a.2.5p	7
mmu.miR.19b.3p	4816.83333333333

The miRNA tracks in the BEDGraph format are created and can be visualized in any genome browser.



Key Points:

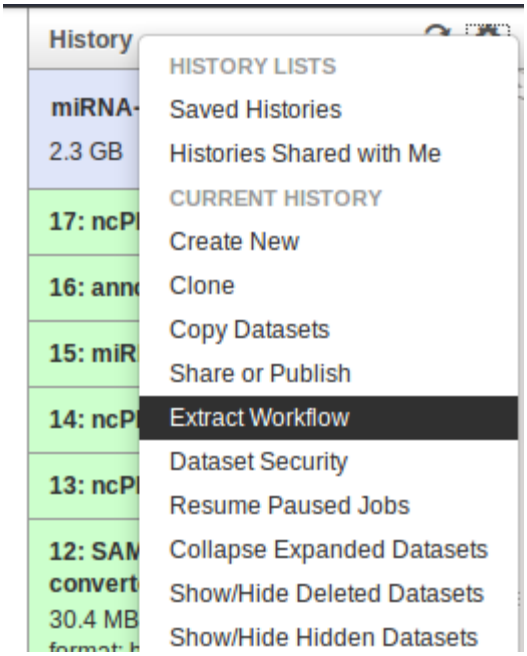
- Look at the miRNAs expression table
- Load the track in your IGV Browser

Bonus:

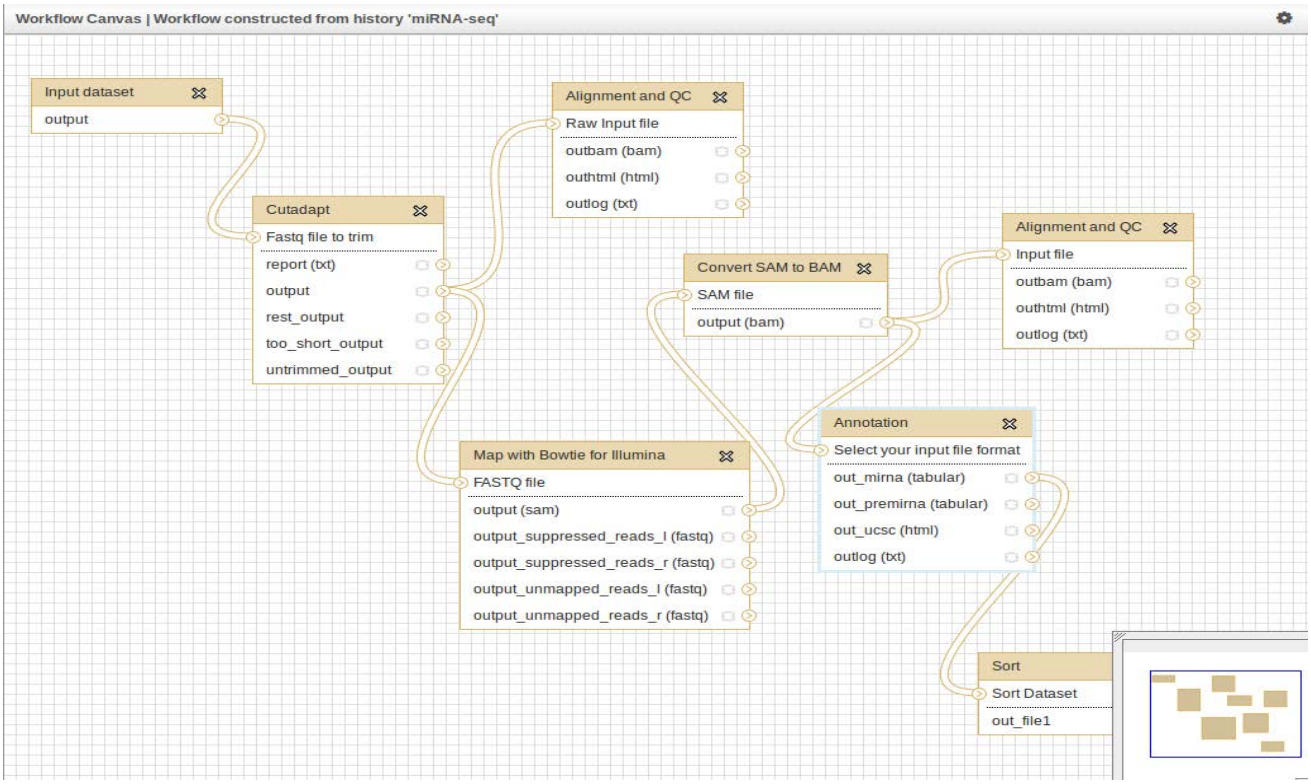
- Sort the tabular file using → **Filter and Sort** to find the most expressed miRNAs)

7. Create a workflow

All the analysis in your history can be used to create a miRNA-seq workflow. On the right menu of your history, select “Extract Workfow” and give a name to your workflow.



Then go the Workflow section (bottom left), and edit your workflow (right click on its name).



To finish, run your workflow on a new sample (right click on the workflow's name).
Import a sample in a new history, and run again the complete workflow. The workflow can be saved and shared between users.

Key Points:

- Build a workflow and replay an analysis

Bonus:

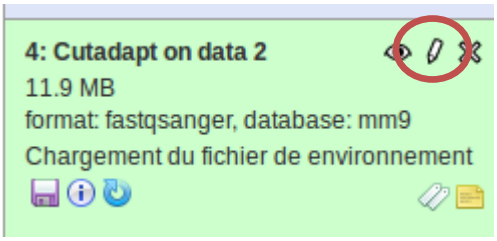
- Look at all annotation tracks on your IGV browser to find the miRNA with a different expression in ES male and female cells.

Galaxy - FAQ

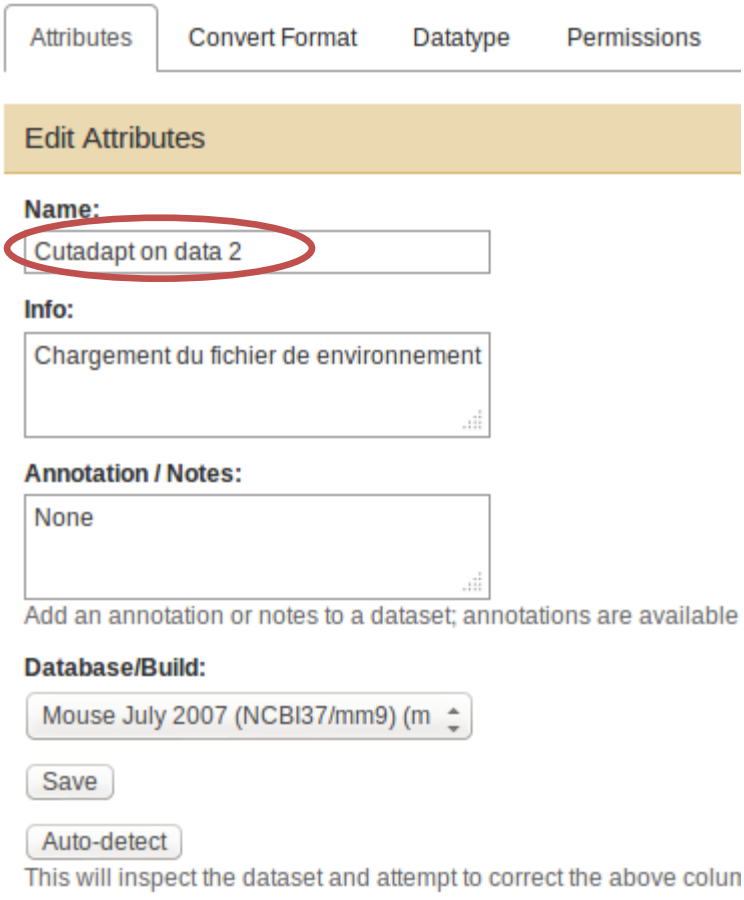
1. Change the name of a feature in your history

Galaxy usually put default name in your history. This can be confusing to try to follow several analyses. We recommend changing this name at the end of the analysis.
To do it:

- Select the “edit” item



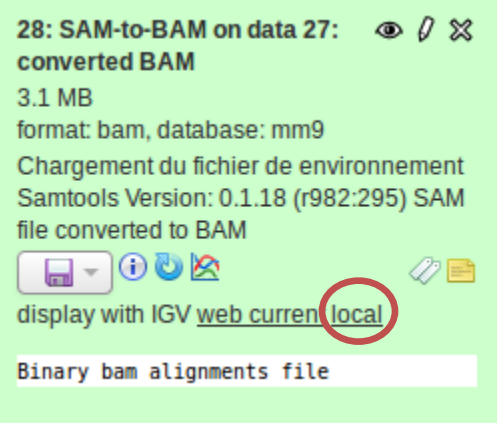
- Change the name attribute



2. Open your data with the IGV browser

IGV aims at visualizing aligned (BAM) and annotation files (BED, GFF, etc.).
To look at your data from Galaxy:

- Open IGV on your laptop
- Click on the local link to IGV. This will open a blank page in your navigator. If the IGV Web/Local link is no available, edit the file and attach an organism.



- Your file is now available through your IGV browser!