Monday, 14th January 2013

Initiation to Galaxy

"Variant detection into chromosome 4 of the human genome"

N. Choisne, S. Gallina, O. Inizan, A. Lermine, D. Naquin, J. van Helden, M. Zytnicki

CNRS UPMC INSU
Station Biologique
Roscoff

avieSan
alliance nationale
pour les sciences de la vie et de la santé

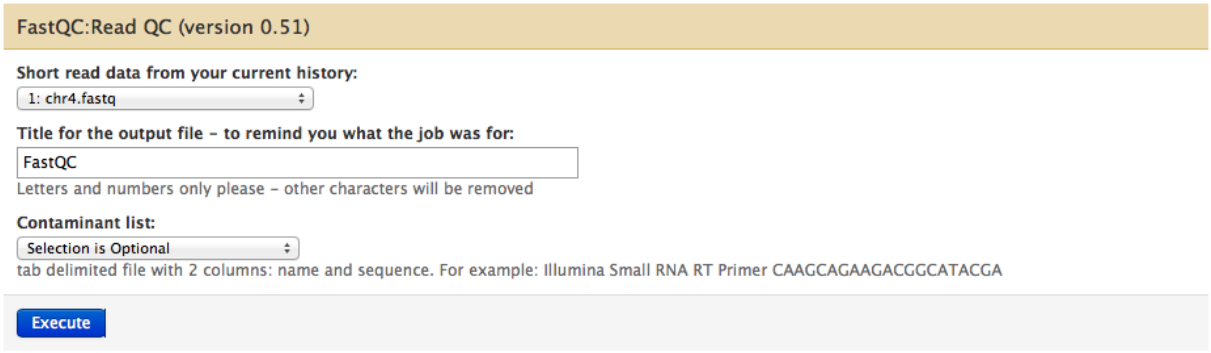
1. Get data

- Go to Shared Data/Data Libraries
- Go the library named “TP Initiation”
- Select chr4.fastq and chr4.fa and import them to the current History
- Clic on customTrack.txt and select “Download this dataset”

2. Quality control on raw data

2.1 **QC**

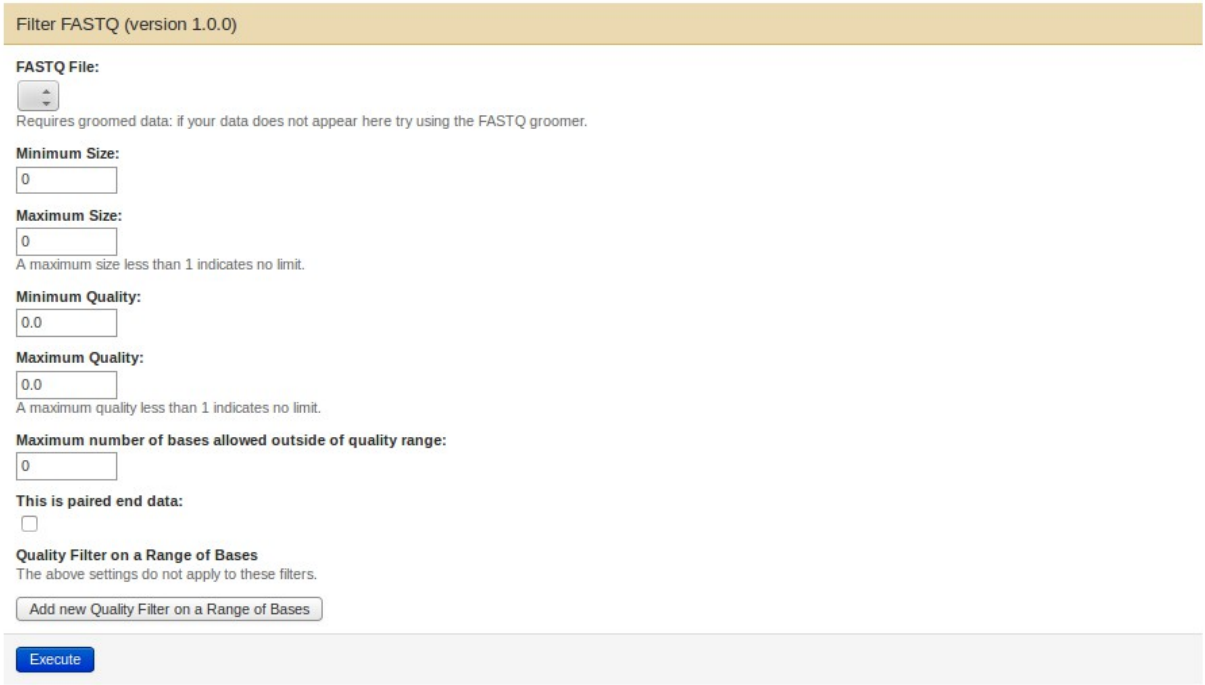
- Run **Fastqc** tool on chr4.fastq file



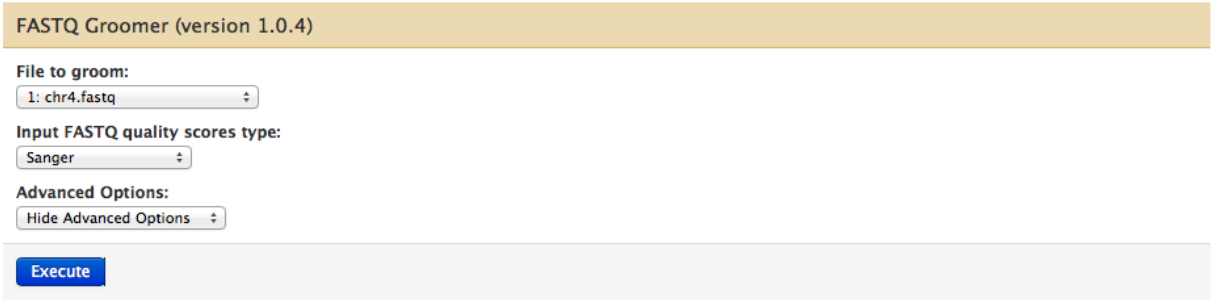
- Look at the generated html file (clic on eye)
    * How many reads do you have?
    * What is the read size?
    * What is the mean quality of most of the reads?

2.2 Fastq filtering by quality

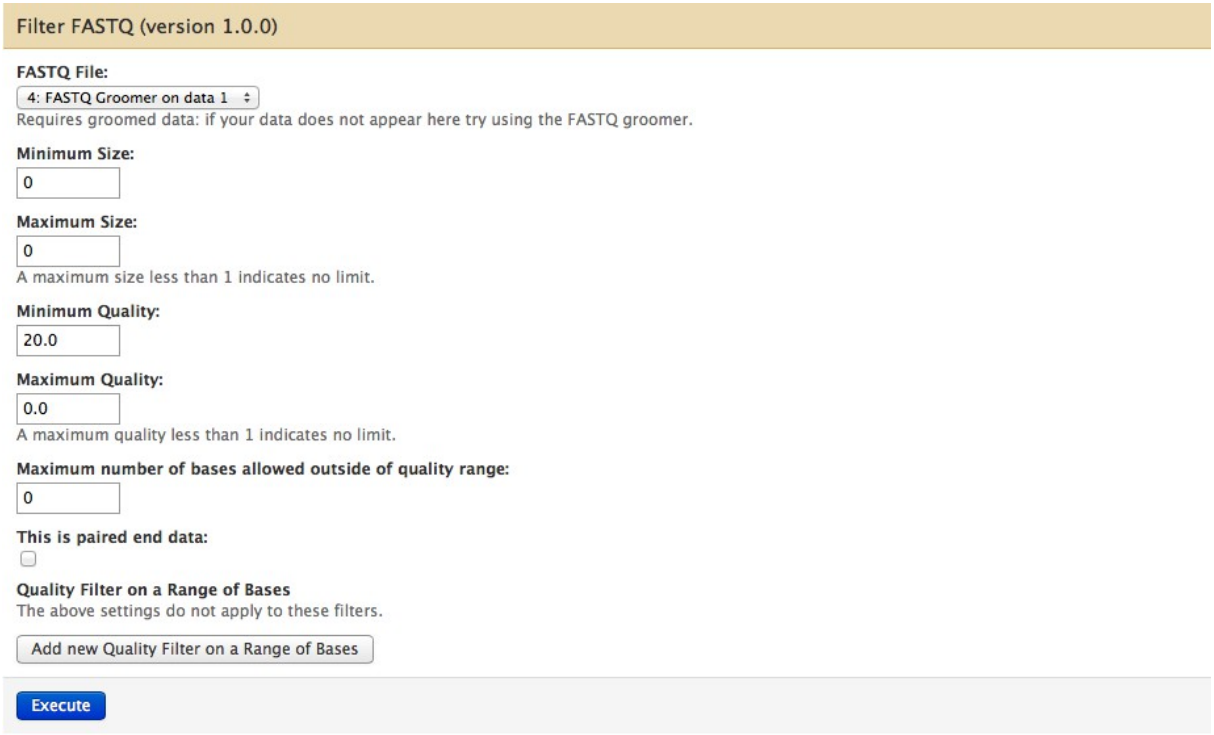
- Run **Filter FASTQ** tool



- Galaxy doesn't offered you any fastq file. It's due to the quality values which are not normalized (C.f. Fastq format presentation). In order to normalize this values, You will use the **FASTQ groomer** tool:

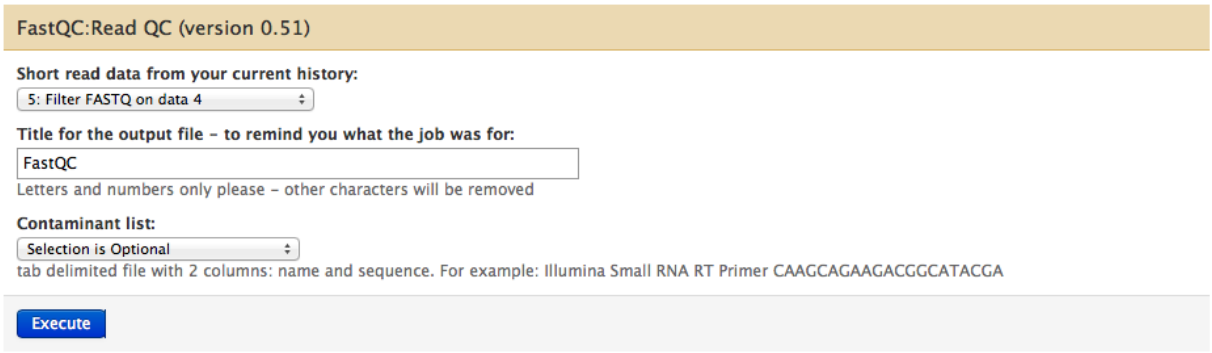


- Then rerun **Filter FASTQ** (with minimum quality at 20)



- Run **Fastqc** on groomed and filtered fastq
To rerun easily tools, just clic on the blue arrow in dataset box

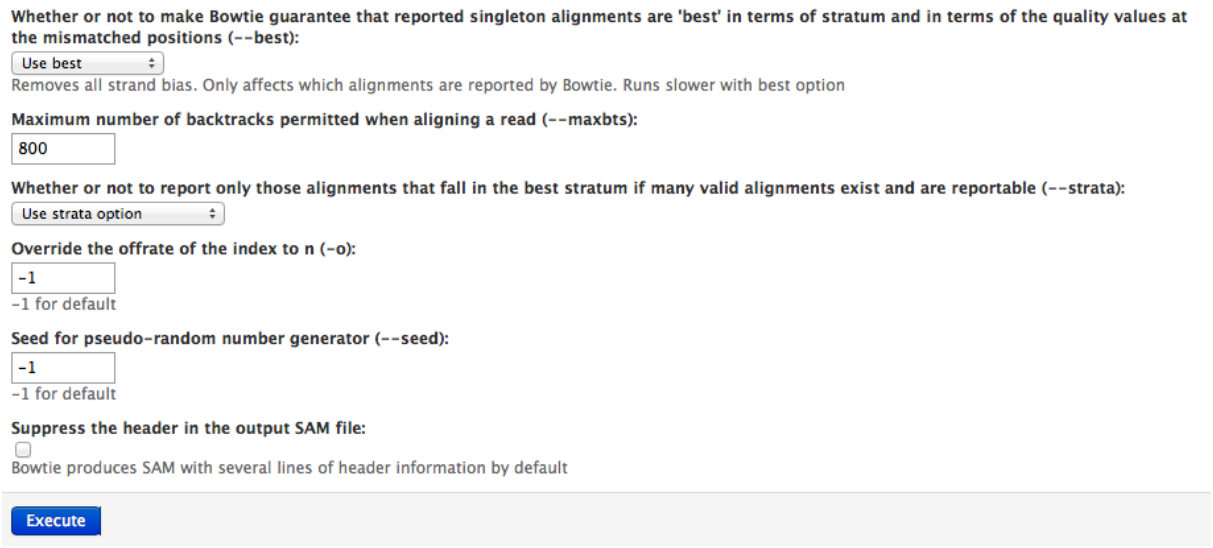
- Just change the input file by selecting the groomed and filtered fastq file



* How many reads do you have?
* What is the read size?
* What is the mean quality of most of the reads?

3. Alignment of raw reads with Bowtie

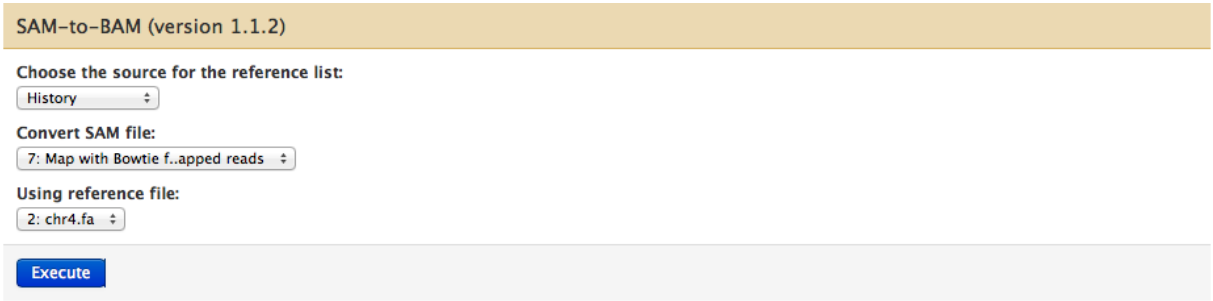
- Run the **Map with Bowtie for Illumina** tool (look at the parameters on screenshot)



* What is the output format?
* How can you distinguish the mapped reads from the unmapped reads?
* What mean 0, 4 et 16 in second row?
Visit http://picard.sourceforge.net/explain-flags.html
* How can you see that you have single end reads?

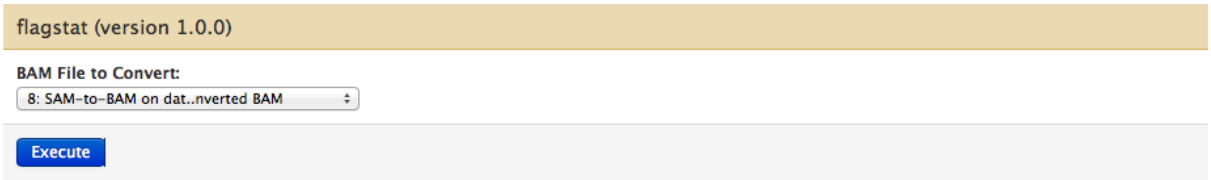
4. SAM to BAM conversion

- Run **SAM-to-BAM** tool



5. Mapping statistics

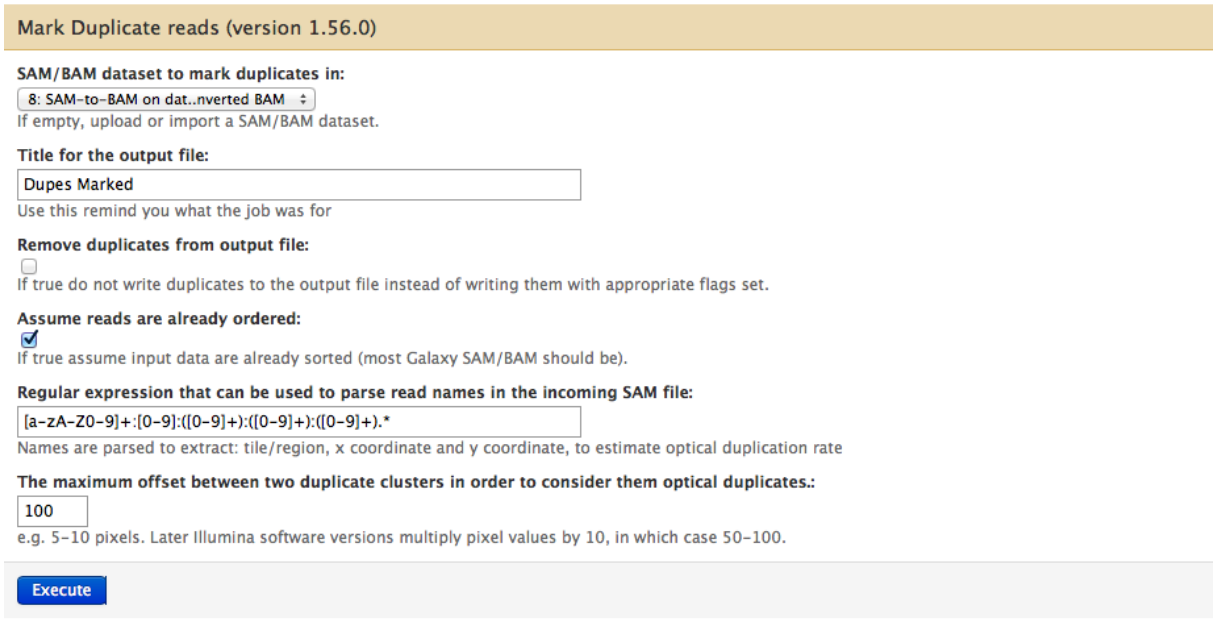
- Run the **flagstat** tool



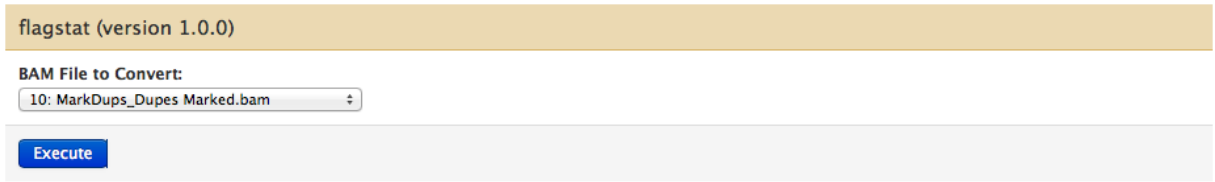
* How many reads are present ?
* How many reads are mapped?
* How many reads are unmapped?
* How many reads are duplicates?

6. Duplicates detection

- Run the **Mark Duplicates reads tool** (don't use Remove duplicates from output file)



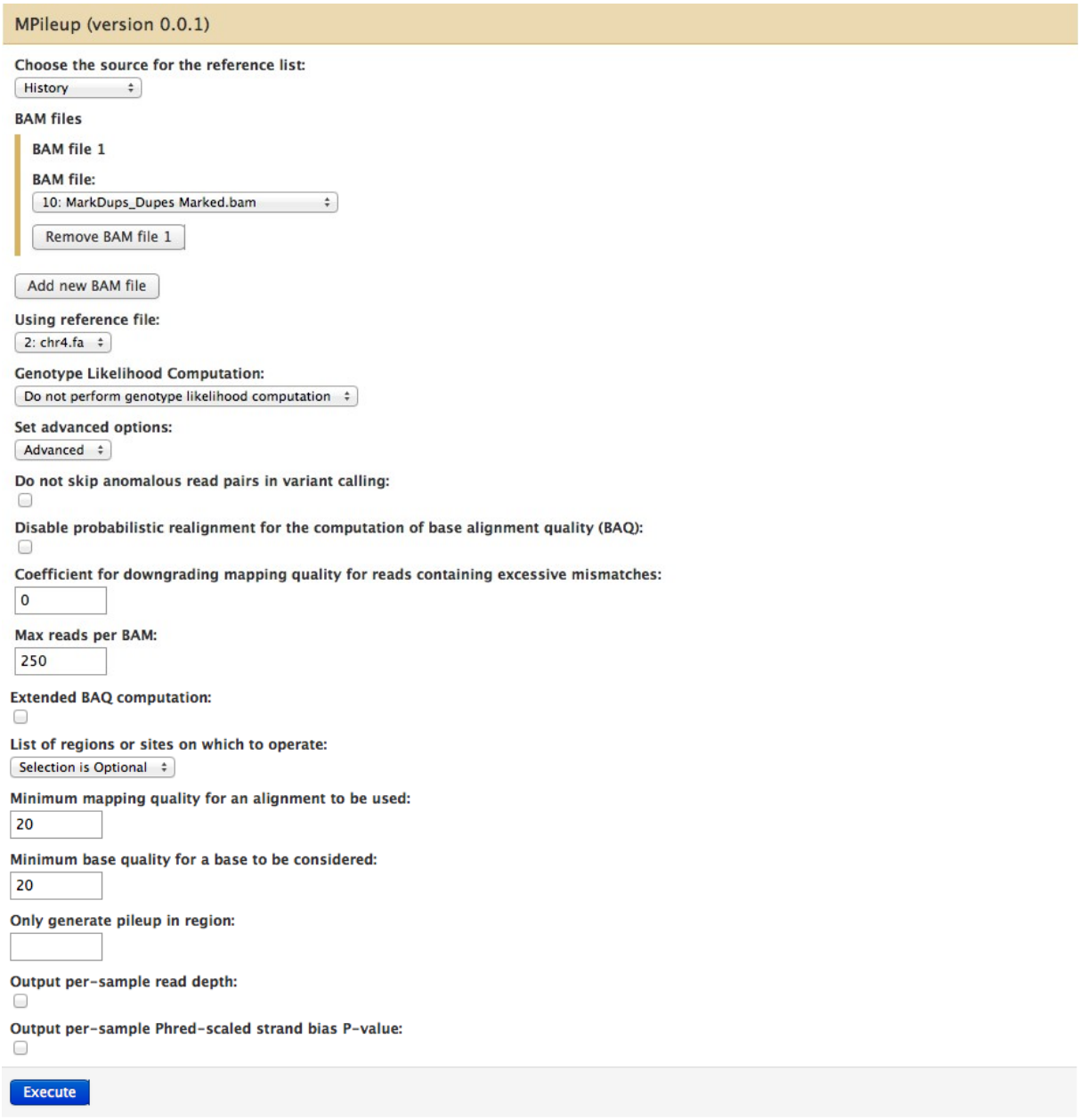
- Run the **flagstat** tool on the output bam file from last step



* How many reads are duplicates?

7. Variant Calling

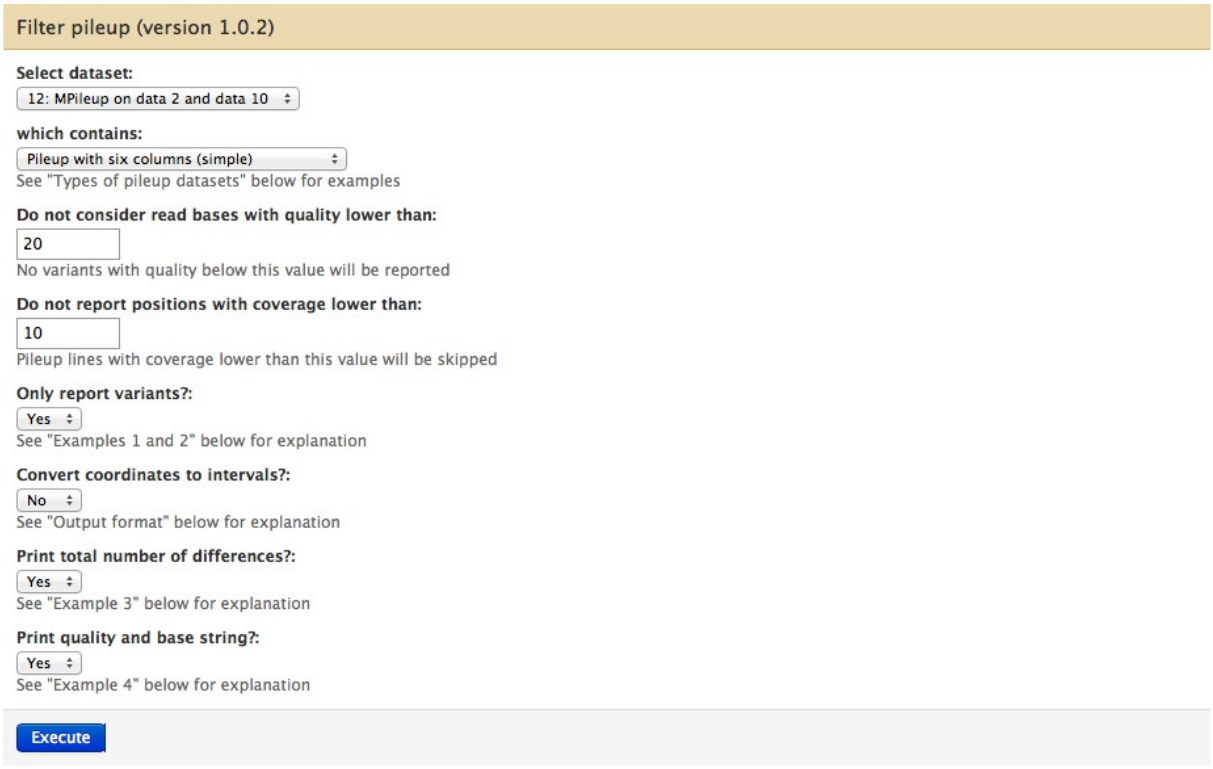
- Run the **Mpileup** tool



* What is the meaning of row 3, 4 et 5?
* How can we read the content of row 5?
Visit http://samtools.sourceforge.net/pileup.shtml

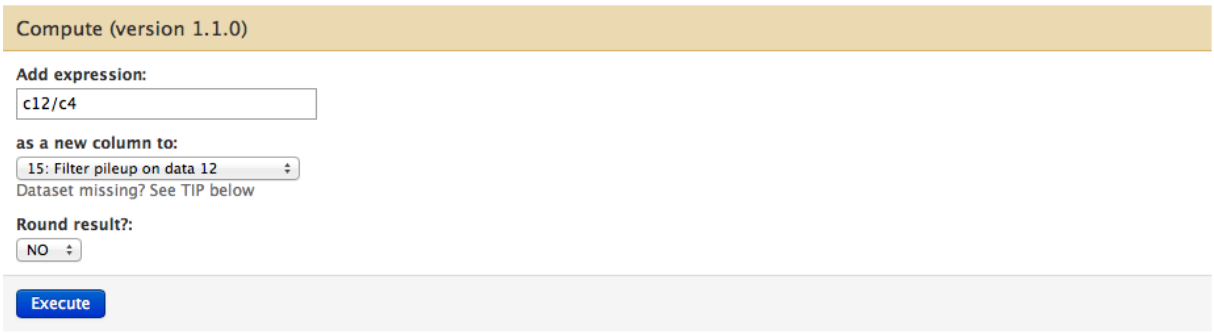
8. Pileup file filtering

- Run the **Filter Pileup** tool



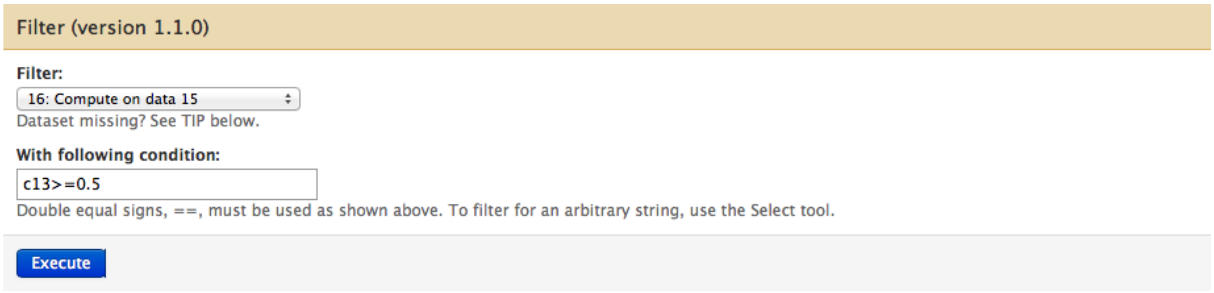
* How many variant have you?
* What is the meaning of row 4 et 12?

- Run the **Compute** tool



* Whats the meaning of the calculated value?
* How can we use this value to select the most likely variants?

- Run the **Filter** tool



* How many good variants do you have?
* What are the characteristics of this(ese) variant(s)?
    - Chromosome position
    - Nucleotide from the reference?
    - Nucleotide from the variant?
    - Variation frequency?

9. Identified variant visualization

- Go to this page
- Clic on choose file and select the customTrack.txt file downloaded at the beginning
- Clic on go to genome browser
- Go to the variant chromosome and position

10. Construct the corresponding workflow

- Clic on the history parameters button
- Select Extract Workflow
- Give it a name and clic on Create Workflow
- You can edit your workflow by clicking on Workflow/You workflow name/Edit

11. Results publication

- Use the galaxy pages system to publish your results (Answers to the questions and integration of the created datasets).