TD : From raw data to SNPs

In this formation, we will use Illumina 100bp pair-ends sequences coming from the transcriptome of different samples of cultivated African rice *Oryza glaberrima.*

1. **Introduction to *Galaxy* :**

In this training, we will use *Galaxy*. This tool allows you to use many currently used bioinformatics tools through an user friendly web interface

- Log in to the IRD *Galaxy* server with this URL : http://bioinfo-inter.ird.fr:8080
- Use the individual account that was provided

There are many possibilities to add a dataset to your history. Here we will use datasets that are shared in galaxy.

- Add the RC1 sequences to your history *(Shared Data/Data Libraries/Formation/Pre-processing and Mapping)*

The *RC1_1.fastq* contains the *forward* sequences and *RC1_2.fastq* contains the *reverse* sequences. These files are paired, this means that the first sequence of the *forward* file is linked to the first sequence of the *reverse* file.

2. **Checking sequence quality :**

Before starting our analysis, we need to check the quality of our data. In this purpose, we will use the software *FASTQC* (freely available at http://www.bioinformatics.bbsrc.ac.uk/projects/download.html#fastqc and implemented in *Galaxy*). This software can be used on all operating systems and generate an HTML report of your sequencing data qualities.

- Check your sequence quality with *FASTQC*, available in *NGS : QC and manipulation.*
- Do not add a contaminant list, the software will use its default list.
- What are the important criteria to check ?
- What can you say about these data ?

3. **5' and adapters trimming**

In order to generate our sequences, we passed through an amplification step using random hexaprimers. However, this protocol can lead to a strong error rate on the first seven bases of our reads. To avoid false positive SNPs due to these errors, we will remove these bases of our sequences.

In the same step, we will remove the adapters/primers. These sequences are used to create the library and to sequence the polonies. They should not be sequenced, but there often are remnants of them in our data.

In this purpose, we will use the software *CutAdapt* who can trim the adapters sequences given by the user in input.

- Use the software *NGS QC and manipulation/Generic Fastq manipulation. Fastq Trimmer by column*
- Remove the 7 first bases of 5' end
- Clean your sequences with *Cutadapt (NGS Cleaning/Cutadapt)*, specifying the specific adapter for your file in add new 5' or 3' adapters, a minimum overlap of 7, a quality cutoff and a minimum length of 20.

Those criteria will remove remnants of adapters from the multiplexing step without removing too much of relevant sequences. The quality threshold of 20 will trim bad quality bases from the end of the reads, which can also lead to false SNP discovery.

- What are the risks on data integrity in this step ?
- Why don't we look for every possible known adapters?

4. Filter on mean quality :

In order to keep only good qualities sequences, we will filter them on mean qualities.

- Use the software *NGS Cleaning/Filter Fastq* to remove sequences with a mean of qualities below 30 and a minimum length of 35.
- What is the problem about keeping bad qualities sequences?
- After these steps, what modification did we make on data structure and on data quality ?