December 2014

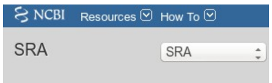
RNA-Seq de novo assembly training session Day 1

hands-on

Useful links :

public read archives :

The European Nucleotide Archive (ENA) provides a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation.
http://www.ebi.ac.uk/ena/



The Sequence Read Archive (SRA) stores raw sequencing data from the next generation of sequencing platforms including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD® System, Helicos Heliscope®, Complete Genomics®, and Pacific Biosciences SMRT®.
http://www.ncbi.nlm.nih.gov/sra

Software packages :

FastQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

FLASH (Fast Length Adjustment of SHort reads) is a very fast and accurate software tool to merge paired-end reads from next-generation sequencing experiments. FLASH is designed to merge pairs of reads when the original DNA fragments are shorter than twice the length of reads. The resulting longer reads can significantly improve genome assemblies. They can also improve transcriptome assembly when FLASH is used to merge RNA-seq data.
http://ccb.jhu.edu/software/FLASH/

cutadapt removes adapter sequences from high-throughput sequencing data. This is usually necessary when the read length of the sequencing machine is longer than the molecule that is sequenced, for example when sequencing microRNAs.
http://code.google.com/p/cutadapt/

Trim Galore! is a wrapper script to automate quality and adapter trimming as well as quality control, with some added functionality to remove biased methylation positions for RRBS sequence files (for directional, non-directional (or paired-end) sequencing).
http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

Pre-requisites : To be able to use an *nix environment :

- get connected
- move around
- create, delete, visualize files
- run commands

The examples data used in the exercises can be found at :

http://genoweb.toulouse.inra.fr/~formation/RNASeq_de_novo/Quality/

Exercise n°1: Raw data quality checking using fastQC and raw data insert size checking

For each of the fastq file :

- run fastQC in command line
- retrieve the resulting zip file on your computer
- visualize the results
- find remarkable elements which either show
  - that the file contains transcriptome reads
  - that the file contains problematic reads (needing to be cleaned before the assembly)

Using the ERR145651 fastq files :

- check the insert size using flash
- what is the average insert size?

Exercice n°2: N cleaning and adapter cleaning

For each of the fastq file :

- process the fastq file with cutadapt using the provided adapter file
- process the file with fastq_longest_subseq_without_Ns.py
- which files have to be cleaned before assembly?

At the end of the session :

If you are using a training account do not forget to clean all the data and software packages you have generated, downloaded and installed.

For further questions :

- e-mail : support.genopole@toulouse.inra.fr .
- You can check the FAQ of the genotoul website : http://bioinfo.genotoul.fr/index.php?id=11 .
- Using the following link , you can have more information about the other training sessions provided by BIOINFO GENOTOUL : http://bioinfo.genotoul.fr/index.php?id=10 .