Bacterial Comparative Genomics Tutorial

After Edwards, Holt (2013)

1. Genome Assembly and Annotation

1.1. Downloading dataset

The tutorial is based on *E. coli* O14:H4 strain TY-2482 (ENA SRR292770). Download the FASTQ files from the FTP links.

1.2. QC

The dataset can be evaluated in term of sequencing quality using the FastQC package that has an intuitive GUI (Java based, shown below), but can also launched programmatically.
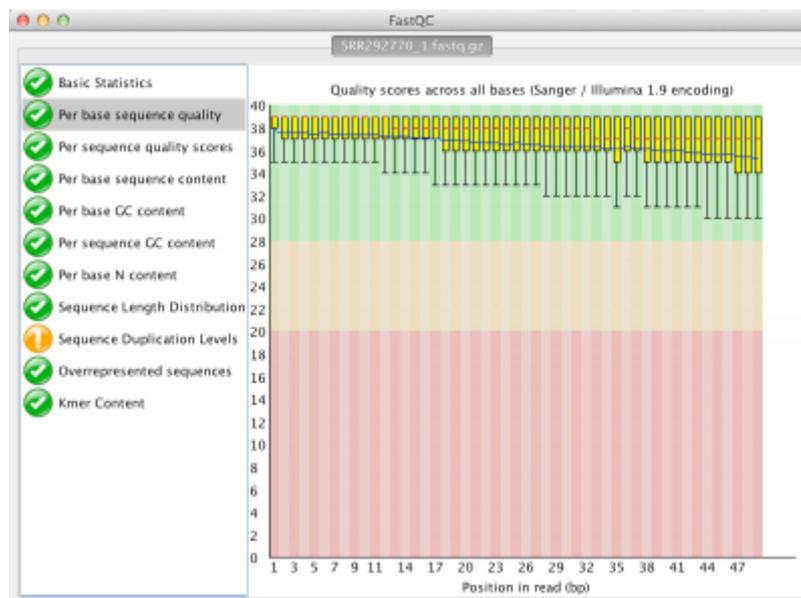
To run it from the command line type the command:

```
fastqc -o QC_output/ SRR292770_1.fastq.gz  SRR292770_2.fastq.gz
```

Where `-o Output` is the parameter telling FastQC where to save the output files, so it is a *path* that can as always be either relative (like `../QC`) or absolute (like `/tmp/Coli_QC/`). Then we should put the path to all the files we want to be analyzed. Again we can use relative or absolute paths, but also wildcards (e.g. `/path/to/reads/*.fastq.gz`). The program produces a set of HTML files with pictures of the plots, here the output I obtained:

- SRR292770_1
- SRR292770_2

1.3. Assembly

The Velvet assembler

[Velvet](#) has been on of the first reliable implementation of the De Bruijn ([slides](#)) graphs for short reads *de novo* assembly. It has not been updated in the last years, but it's worth trying, because of its simple workflow. It consists of two programs: `velveth` counts all the *k*-mer occurrences, while `velvetg` does the actual assembly. The "Output directory" of `velveth` is thus the *input* directory of `velvetg`.

```
# Example using 47 as k-mer size
velveth OutputDirectoryName 47 -fastq -shortPaired -separate
reads/Sample_R1.fastq reads/Sample_R2.fastq
velvetg OutputDirectoryName -clean yes -exp_cov auto -cov_cutoff auto -
min_contig_lgth 180
```

lesskl

The Spades assembler

2. Comparative Genome analysis

3. Typing and specialist tools

From:
https://seq.space/notes/ - **Bioinformatics Notes**

Permanent link:
https://seq.space/notes/doku.php?id=denovo2013

Last update: **2020/02/07 09:51**