

RNA-seq analysis in R

QC of raw reads with FastQC

In our initial QC of the raw fastq file we will be interested in gathering various metrics, such as the total number of reads, sequence length, or GC content. We will also want to summarise such things as base quality scores and make assessments of the contamination of the reads with adapter sequence.

FastQC is quality control tool for high throughput sequence data that is maintained by the Babraham Institute. It is free to download and use. It runs a number QC analyses on sequencing data (in various formats, not just fastq) and summarises the results in a easy to read report.

The basic command to run FastQC is simply `fastqc`.

Access the help page to find the basic usage and other options:

```
fastqc --help
```

The **Usage** is:

```
fastqc seqfile1 seqfile2 .. seqfileN

fastqc [-o output_dir] [--(no)extract] [-f fastq|bam|sam]
      [-c contaminant_file] seqfile1 .. seqfileN
```

The simplest way to use it is just to type `fastqc` followed by all the sequence files that you wish to QC. It will then run through as many files as you provide generating a report for each one.

There are many additional options that you can provide to modify the behaviour of the programme. The most common one is `-o output_directory`. By default the report is written to the same directory as the fastqc file, however, if you would like to gather the QC in a different directory, you can specify this using the `-o` flag followed by the name of the directory, e.g:

```
fastqc -o QC fastq/my_fastq_file.fastq.gz
```

In this case we wish to generate a report for the file `my_fastq_file.fastq.gz`, which is in the folder `fastq`, and to have the report written into a directory called `QC`.

Note that the output directory must already exist, FastQC will not create it.

Exercise

1. a) Check the location of the current directory using the command `pwd`
b) If the current directory is not `Course_Materials`, then navigate to the `Course_Materials` directory using the `cd` (change directory) command:

```
cd ~/Course_Materials
```

2. a) Use `ls` to list the contents of the directory. There should be directory called `fastq`
b) Use `ls` to list the contents of the `fastq` directory:

```
ls fastq
```

You should see two fastq files. These are the files for read 1 and read 2 of one of the samples we will be working with.

3. Create a new directory for the QC results called **QC** using the `mkdir` command:

```
mkdir QC
```

4. Run fastqc on one of the fastq files:

```
fastqc fastq/SRR7657883.sra_1.fastq.gz
```

5. The previous command has written the report to the **fastq** directory - the default behaviour for fastqc. We want it in the **QC** directory.

- a) Use the `rm` (remove) command to delete the report:

```
rm SRR7657883.sra_1_fastqc.html
```

- b) Also delete the associated zip file (this contains all the figures and the data tables for the report)
6. Run the FastQC again, but this time:
 - a) have FastQC analyse both fastq files at the same time. You will need to add `-t 2` before the sequence file names. See `fastqc --help` to find out about this option.
 - b) try to use the `-o` option to have the reports written to the **QC** directory.
 7. Open the html report in a browser and see if you can answer these questions:
 - A) What is the read length?
 - B) Does the quality score vary through the read length?
 - C) How is the data's quality?
-