# Introduction to Bulk RNAseq data analysis

## QC of raw reads with FastQC

In our intial 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.

For this (and the next couple of sessions) we will working at the command line. If you are unfamiliar with the command line and "bash" shell command language, there is nice basic cheat sheet with the most commonly used commands here:

https://icosbigdatacamp.github.io/2018-summer-camp/slides/BASH_Cheat_Sheet.pdf (https://icosbigdatacamp.github.io/2018-summer-camp/slides/BASH_Cheat_Sheet.pdf)

FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) is a 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 an 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 the above command will generate a report for the file **my_fastq_file.fastq.gz**, which is in the folder **fastq**, and will 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` (**c**hange **d**irectory) 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 called *SRR7657883.sra_1.fastq.gz* and *SRR7657883.sra_1.fastq.gz*. These are the files for read 1 and read 2 of one of the samples we will be working with (you can ignore the other fastq files for now).

3. Run fastqc on one of the fastq files:

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

This will write the QC report to the *fastq* directory.

4. 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?

Note: In practice you would actually end up running FastQC on both read 1 and read 2 of every sample. This would usually results in large number of reports. Rather than having to open and inspect each one individually, we would normally gather them all together into a single report using the program **MultiQC**. We will cover MultiQC in the later session "QC of alignment".

---