Introduction to Bulk RNAseq data analysis

QC of raw reads with FastQC - Solutions

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
```

SRR7657883.sra_1.fastq.gz SRR7657883.subset_2M.sra_1.fastq.gz
SRR7657883.sra_2.fastq.gz Test_adapter_contamination.gq.gz.
SRR7657883.subset_2M.sra_2.fastq.gz

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.

3. Run fastqc on one of the fastq files:

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

⇒ *SRR7657883.sra_1_fastqc.html*
⇒ *SRR7657883.sra_1_fastqc.zip*

4. Open the html report in a browser and see if you can answer these questions:
   A) What is the read length? **150**
   B) Does the quality score vary through the read length?
   Yes, the first few bases and the last few bases are typically of lower quality.
   C) How is the data's quality?
   Overall, pretty good.