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.sra_2.fastq.gz

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
```

⇒ **QC**

4. 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*

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)

```
rm -f fastq/SRR7657883.sra_1_fastqc.zip
```

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.

```
fastqc -t 2 -o QC fastq/SRR7657883.sra_1.fastq.gz fastq/SRR7657883.sra_2.fastq.gz
```

or more simply we can use the `*` wild card:

```
fastqc -t 2 -o QC fastq/SRR7657883.sra_*.fastq.gz
```

⇒ *QC/SRR7657883.sra_1_fastqc.html*
⇒ *QC/SRR7657883.sra_1_fastqc.zip*
⇒ *QC/SRR7657883.sra_2_fastqc.html*
⇒ *QC/SRR7657883.sra_2_fastqc.zip*

7. 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.