Read QC and trimming

Conventions in this document

This is normal text

For text describing a unix command, e.g. `grep` - the command will then be in italics.

```
This is a command you need to enter on the command line
```

```
This command has one word <here> that you need to change
```

For example, <here> might be the name of the folder that will contain the output of the command

Working area

The area to do your ‘work’ and save your files is here:

```
/home/your_username
```

Where is what

All data for this part of the practical is in this folder:

`/data/qc`

You will find several fastq files in that folder. We will start the practical with these two files :

```
/data/qc/cod_read1.fastq
/data/qc/cod_read2.fastq
```

They contain 1 million randomly sampled reads from a HiSeq 2x100 bp PE (paired end) run

Part 1: Understanding reads, QC of sequence data

Learning points:

- Recognizing the fastq file format
- How to prepare and judge a QC report

A peak into the fastq files

Fastq files are very big. In order to be able to view them in a ‘page-by-page’ way, we will use the `less` command:

```
less /data/qc/cod_read1.fastq
```

This file contains the forward read (‘read 1’) dataset of the run for the sample. Use the space bar to browse through the file. Use `q` to go out of the `less` program. Make sure you recognize the fastq format, if needed use the slides from today's presentation.

Question: which of the different Illumina Sequence identifiers are used for these reads? See http://en.wikipedia.org/wiki/FASTQ_format#Illumina_sequence_identifiers.

Repeat this for the read 2 file:

```
less /data/qc/cod_read2.fastq
```

Question: do you see whether the reads in the same order in both files?

Quality control of Illumina reads

We will be using a program called **FastQC**. The program is available with a graphical user interface, or as a command-line only version. We will use the latter one. It takes a single fastq file (the file can be compressed) as input, and produces a web page (html file) with the results of a number of analyses.

Program	Options	Explanation
fastqc		Quality control of sequence data
.	-o foldername	tells the program to place the output in a folder called foldername instead of in the same folder as the input file
.	fastq file	file to be analysed by the program

Before we run the program, let’s create a new folder for the output. Do this in your home folder. First, go to your home directory. Remember you can simply type:

```
cd
```

Now, we'll make the new folder and move into it:

```
mkdir qc
cd qc
pwd
```

To run fastqc on the first MiSeq file, run the command below; <*your_username*> should be the name you used for your folder. Note that the command should be written on a *single line*. Also note where you should put spaces!

```
fastqc -o /home/<your_username>/qc /data/qc/cod_read1.fastq
```

The program will tell you how far it has come, and should finish in a minute or so. Check that it finished without error messages.

In the folder you specified after `-o`, you should now see a new folder called `cod_read1_fastqc`, and a zip-file by the same name. The *folder* contains the following:

`Icons` → folder
`Images` → folder
`fastqc_data.txt` → file
`fastqc_report.html` → file
`summary.txt` → file

Now, open a webbrowser, and, using the menu option ‘Open file’, locate the html file, choose

`your_username → qc → cod_read1_fastqc → fastqc_report.html`

Open the file called `fastqc_report.html`.
Alternatively, you could browse the file system and double-click on the file.

Study the results.

The plot called "Per base sequence quality" shows an overview of the range of quality scores across all based at each position in the fastq file. The y-axis shows quality scores and the x-axis shows the read position. For each read position, a boxplot is used to show the distribution of quality scores for all reads. The yellow boxes represent quality scores within the inter-quartile range (25% - 75%). The upper and lower whiskers represent 10% and 90% point. The central red line shows the median of the quality values and the blue line shows the mean of the quality values.

A rule of thumb is that a quality score of 30 indicates a 1 in 1000 probability of error and a quality score of 20 indicates a 1 in 100 probability of error (see the wikipedia page on the fastq format at http://en.wikipedia.org/wiki/Fastq. The higher the score the better the base call. You will see from the plots that the quality of the base calling deteriorates along the read (as is always the case with Illumina sequencing). Sometimes, a minimum requirement for Per Base Sequence Quality is that the first 36 bases should have a median and mean quality score over 20.

Now, answer these questions:

Questions

- How many reads were there in total in the `cod_read1.fastq` file?
- How many bases were there in total in the file?

- Which part(s) of the reads would you say are of low quality - if any?
- Would the run have passed the minimum requirement for Per Base Sequence Quality?

Repeat the fastqc analysis for the file `/data/qc/cod_read2.fastq`, which contains the reverse read ('read2').

Open the `fastqc_report.html` in your webbrowser.

Questions

- Are there part(s) of the reads that have a lower quality compared to the MiSeq_50x_R1.fastq file?
- Would the run have passed the minimum requirement for Per Base Sequence Quality?

NB. You can get more information about the use of the fastqc program by writing

```
fastqc -h
```

More read files

Now run fastqc on the other files in the `/data/qc` folder and evaluate the results. We'll discuss these together afterwards:

- start with the files called `more_cod_read*`. How do these compare to the cod reads you looked at before?
- then take the ChipSeq and microRNA example read files (thye only have one fastq file each)

Other programs to try

You could try the online QC program PRINSEQ on these datasets: http://edwards.sdsu.edu/prinseq/