Quality control of sequencing reads

Course material for the INF-BIOx121 'High Throughput Sequencing technologies and bioinformatics analysis ' course at the University of Oslo, Fall 2015

Conventions in this document

This is normal text

For text describing a unix command, e.g. `grep` - the command will then be look like this:

```
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

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
```

Followed by the 'enter' key.

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

```
mkdir qc
cd qc
pwd
```

We will be using the *module* system to 'activate' programs (technically, to add them to your environment). To be able to use fastqc, run this command:

```
module load fastqc
```

To check what modules we have loaded, type

```
module list
```

You should see

```
Currently Loaded Modulefiles:
  1) fastqc/0.11.2
```

(for more technical information on the module system, see http://modules.sourceforge.net/).

To run fastqc on the first 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 ./ /data/qc/cod_read1.fastq
```

Note that we use '`-o ./`' here, which specifies the current folder '`./`' as location for the output.

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 zip file called `cod_read1_fastqc.zip`, and an html file called `cod_read1_fastqc.html`

Download the html file to the local hard disk of the PC/Mac you are using, see the instructions on the course wiki. Open a webbrowser, and, using the menu option ‘Open file’, locate the html file. 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).

The plot 'Per tile sequence quality' shows the deviation from the average quality for each tile, i.e. part of the flowcell. The graph allows you to look at the quality scores from each tile across all of your bases to see if there was a loss in quality associated with only one part of the flowcell. The colours are on a cold to hot scale, with cold colours being positions where the quality was at or below the average for that base in the run, and hotter colours indicate that a tile had worse qualities than other tiles for that base. A good run should show a plot that is blue all over.

Now, answer these questions:

Questions

- What quality encoding did fastqc determine the quality scores to be in? See also the wikipedia page on the fastq format again
- 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 you have accepted this data if you were given it by your sequencing provider?

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

Open the `cod_read2_fastqc.html` in your webbrowser.

Questions

- Are there part(s) of the reads that have a lower quality compared to the `cod_read1.fastq` file?
- Would you have accepted this data if you were given it by your sequencing provider?

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 (they only have one fastq file each)

Question: which of the different Illumina Sequence identifiers are used for these reads?

Question: discuss the results with your neighbour, and try to explain the fastqc results for these files.

Other programs to try

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