FASTQC Analysis

So, you will be working with Illumina sequence data, and you are interested in data quality, and particularly in your sequencing error expressed as base quality values. This is one way to look at the data.

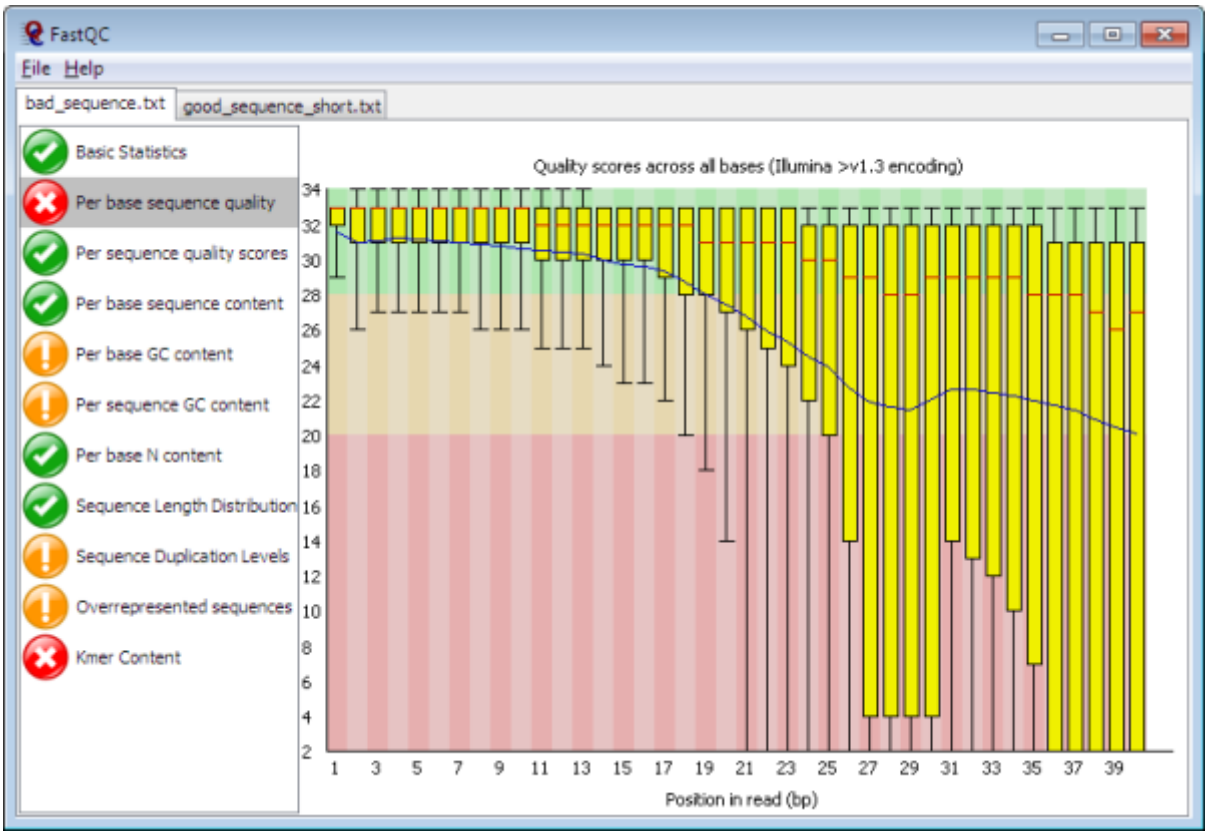

Figure 1: Example output of a FASTQC analysis taken from the software homepage

1. Identify the read set that you want to analyse. Make sure that you have a fastq format

   In the case you are having a paired end library layout, is convention that you analyse forward and reverse reads separately, since their quality typically varies considerably. You can use samtools to split your readset into forward and reverse reads in the case you are having both in one file

2. Install FASTQC using Anaconda
3. Use the fastq file as input for a FASTQC analysis

   After installing FASTQC in its own conda environment, activate the new environment with the corresponding conda command. Now you can call FASTQC by simply typing

   ```
   fastqc filename.fq -o /path/to/output/directory
   ```

   This will generate a html file which you can copy to your local computer using scp. We made a short video that explains you how to do this.

4. Open the html output files with any browser. With your results try to answer the following

questions:

1. What kind of information do you get after running FASTQC?
2. Try to make a statement about the quality of your sequencing run.
3. Take a look at the overrepresented sequences, and overrepresented Kmers report. Interpret the results and reconcile with your expectation. In case you have no expectation, make sure to discuss with the tutors.

5. Perform an end trimming of the sequencing reads using Trimmomatic. What kind of information do you need to perform this analysis step? The Trimmomatic page may provide some initial help.

- First you need to extract the sequence of the adapter from FASTQC and save as a fasta file using a texteditor like nano:

```
touch adaptator.fasta
nano adaptator.fasta
```

- It should have the format you know from fasta files with a /1 for the forward adapter and a /2 for the reverse. The prefix has to be the same in both identifiers.

```
>Truseq/1
forward adapter-sequence from FASTQC
>Truseq/2
reverse adapter-sequence from FASTQC
```

- Make sure that you are in the conda environment that contains Trimmomatic and document the command you use to run it. Discuss with the group or the tutors if you are unsure about the parameters.