Quality Assessment (FastQC)

It is always necessary to assess the quality of the sequence reads in FASTQ files from the sequencing facility. FastQC is a quality control application for high-throughput sequencing data. By using FastQC, we could be aware of any problems in raw sequence data before moving on to the next step.

Install FastQC

```
# FastQC requires a suitable 64-bit Java Runtime Environment (JRE) installed and
in the path. Check the version of Java:
$ java -version

# For Linux, download and extract the latest version of FastQC from [project
website]
(http://www.bioinformatics.babraham.ac.uk/projects/download.html#fastqc):
$ wget
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/fastqc_v0.11.9.zip
$ unzip fastqc_v0.11.9.zip

# Append the path to your PATH environment variable:
$ export PATH=$PATH:/path/to/FastQC

# Verify installation:
$ fastqc -help
```

Run FastQC

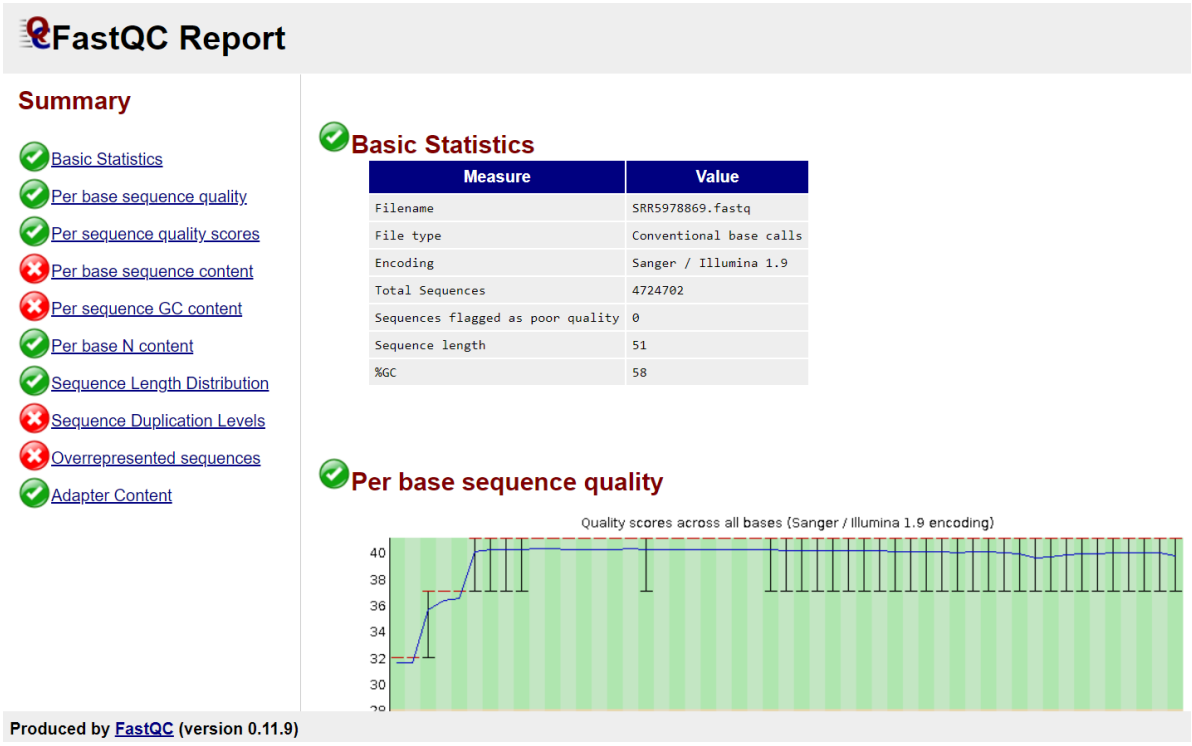
```
# Examine the quality of one FastQ file:
$ fastqc -o /path/to/fastqc_result/ /path/to/raw_data/homo/SRR5978869.fastq

# or examine the quality of multiple FastQ files:
$ fastqc -o /path/to/fastqc_result/ -t 6 /path/to/raw_data/homo/*.fastq
```

Note: `-o` (or `--outdir`) will create all output files in the specified output directory. `-t` specifies the number of files / threads that can be processed in parallel.

FastQC Results

FastQC produces two output files for each FastQ file: an HTML report ("SRR5978869_fastqc.html") and a packed file ("SRR5978869_fastqc.zip").

You could transfer the HTML file to local place by *FileZilla* (mac) or *WinSCP* (win), and open the file in browser. A screenshot of part of the HTML file is shown below.



Note that two of the most important analysis modules in FastQC are **"Per base sequence quality"** plot and the **"Overrepresented sequences"** table. The "Per base sequence quality" plot provides the distribution of quality scores across all bases at each position in the reads. The "Overrepresented sequences" table displays the sequences (at least 20 bp) that occur in more than 0.1% of the total number of sequences, which aids in identifying contamination. You could also refer to Analysis Modules in FastQC documentation for the interpretation of the HTML report.

The other output is a zip file for each sample. You could unpack the zip files and have a look at the summary.

```
# Unpack a .zip file in the result directory:
$ unzip SRR5978869_fastqc.zip

# or unpack .zip files in the result directory:
$ for zip in *.zip
do
unzip $zip
done

# To see the content of a single summary file:
$ cat SRR5978827_fastqc/summary.txt

# or cat all summary files into one text file and have a look at it:
$ cat */summary.txt > ~/all/fastqc_summaries.txt
$ cat ~/all/fastqc_summaries.txt
```

For paired-end data, since the two reads of the pair are generated separately, trying to get statistics like per base sequence quality on the combined forward and reverse reads would make no sense. In this case, you may check quality by inputting BAM files.