Assessing quality of illumina data

The Illumina HiSeq, Genome Analyser and MiSeq platforms produce single-end or paired-end sequence reads of a defined length, and due to the chemistry there are a few well known issues:

- The quality of the read will dip towards the end of the read
- The quality of the second read will generally be worse than the first read

One **really** important issue is to be able to assess the quality of the raw data so you can make some decision on what to do about it. There are many tools for doing this, but the standard tool is FastQC. This tool produces many plots and statistics which can help interpret the data, and can be run from the command-line, or as an interactive window.

1) Getting some data

If you haven't already, get some microRNA sequencing data from the EBI ENA:

wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR026/SRR026762/SRR026762.fastq.gz

Information about this dataset can be found here: http://www.ebi.ac.uk/ena/data/view/SRR026762

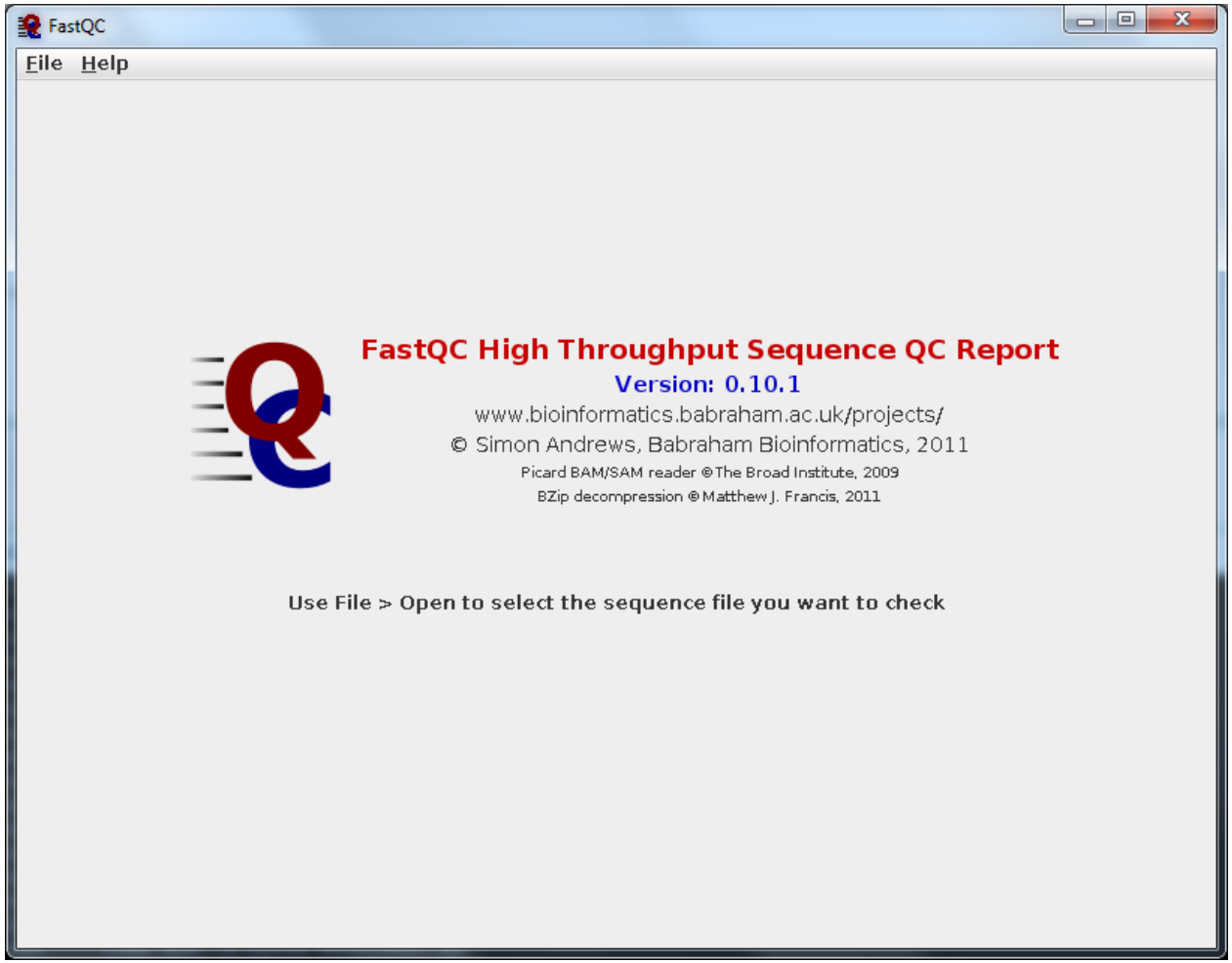
2) Running FastQC

As we are running these windows over the internet (the server the software is running on is in Virginia, USA; you are sat where you are) then this session is going to be slow. Please be patient!

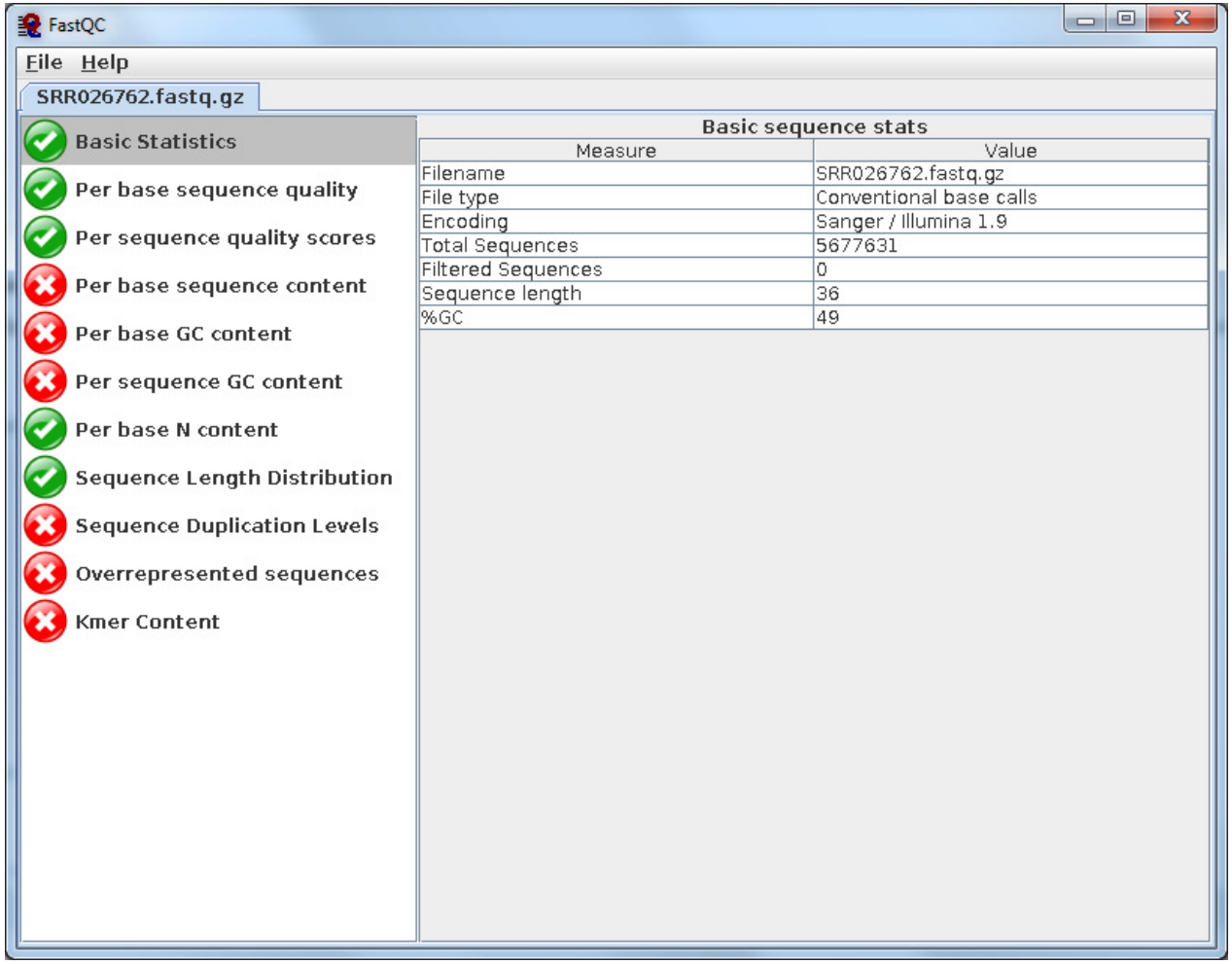
FastQC can be run in interactive mode by typing:

fastqc &

The resulting window will look a little like this:



If you are **very patient** then you can go through the menu and choose **File -> Open** and select **SRR026762.fastq.gz.** However, as mentioned above, this may be very slow due to the connection over the internet. If you persevere, then you will see this:



Each of the options on the left can be clicked to display more information.

3) Running on the command line

Alternatively, FastQC can be run on the command line, also known as "batch mode". Here, we give **fastqc** the file on the command line:

fastqc *filename.tar.gz*

This time fastqc will write the data to disk. Try this:

cd

ls -l

fastqc SRR026762.fastq.gz

ls -l

What has changed? What has **fastqc** created? There is an HTML report that has been generated, and this can be viewed using an internet browser:

firefox SRR026762_fastqc/fastqc_report.html &

However, again this will be very slow due to the internet connection problems.

To avoid issues with accessing these data over the internet, we have created a copy of the fastq output HERE

How many reads are in the file?

What is the percentage GC of the entire dataset?

What is the sequence length of the reads?

What is the top over-represented sequence? (Hint: you may wish to take the sequence and BLAST it here.

I do not want to simply recreate the documentation for FastQC and so I have placed links to the information here. Please read each link and interpret the information in our fastqc report:

4) Per-base sequence quality

Shows the average and range of the sequence quality values across the read: per-base sequence quality

---

5) Per sequence quality

A graph showing the distribution of quality scores: per sequence quality

---

6) Per base sequence content

The average percentage of A, G, C and T across the read length: per base sequence content

---

7) Per-base GC content

The average GC content across the read: per-base GC content

---

8) Per-sequence GC content

Actual and theoretical distributions of GC content: per-sequence GC content

---

9) Per-base N content

The locations of N's in your dataset, which may reflect cycles in the sequencing which have problems: per-base N content

---

10) Sequence length distribution

The distribution of sequence lengths in the dataset. For illumina, this will generall be uniform: sequence length distribution

---

11) Duplicate sequences

An analysis of the level of duplicate sequences in the dataset: duplicate sequences

A more in-depth discussion of how to interpret this plot can be found http://proteo.me.uk/2011/05/interpreting-the-duplicate-sequence-plot-in-fastqc/

---

12) Over-represented sequences

An analysis of sequences that appear more than they should - can reflect adapter contamination: over-represented sequences

---

13) Over-represented kmers

An anlysis of sequence enrichment at the level of the kmer - over-represented kmers

---

The authors of FastQC have also provided examples of a good illumina dataset and a bad illumina dataset.

Does SRR026762.fastq.gz *look like* a bad dataset?

Does *FastQC* think SRR026762.fastq.gz is a bad dataset?

Do *you* think SRR026762.fastq.gz a bad dataset?

Does SRR026762.fastq.gz have over-represented sequences? What are they? Why might it have those sequences?

Does that make it a bad dataset?

Many of FastQC's tests assume a random distribution of data throughout the genome. What kind of datasets may violate this assumption?

If you have time, you may wish to run FastQC on other datasets within the **training** directory

---

Next - trimming data