RNA-SEQ DATA ANALYSIS

Prof.ssa Rosalba Giugno

FastQC

- FastQC is a software used to quality control checks on raw sequence data coming from high throughput sequencing pipelines

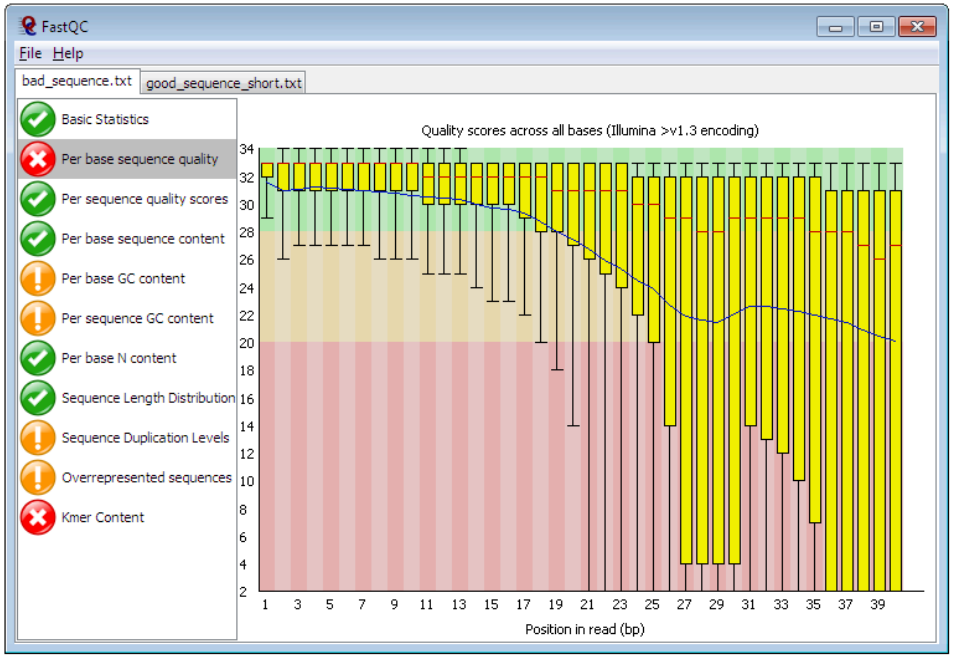
- Main features:
  - Import of data from FastQ files
  - Providing a quick overview to tell you in which areas there may be problems
  - Summary graphs and tables to quickly assess your data
  - Export of results to an HTML based permanent report
  - Offline operation to allow automated generation of reports without running the interactive application

- Download link:

  http://www.bioinformatics.babraham.ac.uk/projects/fastqc

FastQC

- You can run FastQC in interactive mode or in command-line mode
- To run FastQC on the command line you simply have to specify a list of files to process:

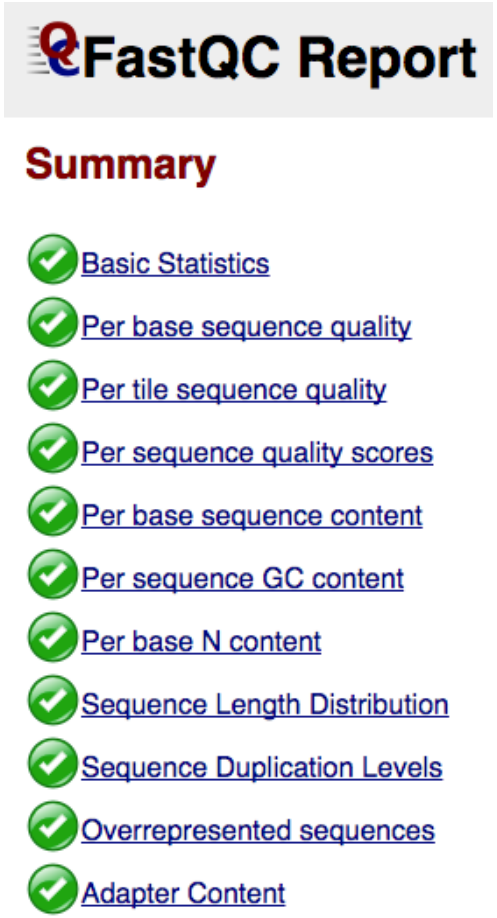



```
fastqc somefile.fq someotherfile.fq [--outdir=/some/other/dir/]
```

// **--outdir** if you want to redirect the output directory

FastQC Report

- Most sequencers will generate a QC report as part of their analysis pipeline.
- Fastqc aims to provide a QC report which can spot problems which originate either in the sequencer or in the starting library material.



FastQC – Basic Statistics

- The Basic Statistics module generates some simple composition statistics for the file analysed

Basic Statistics

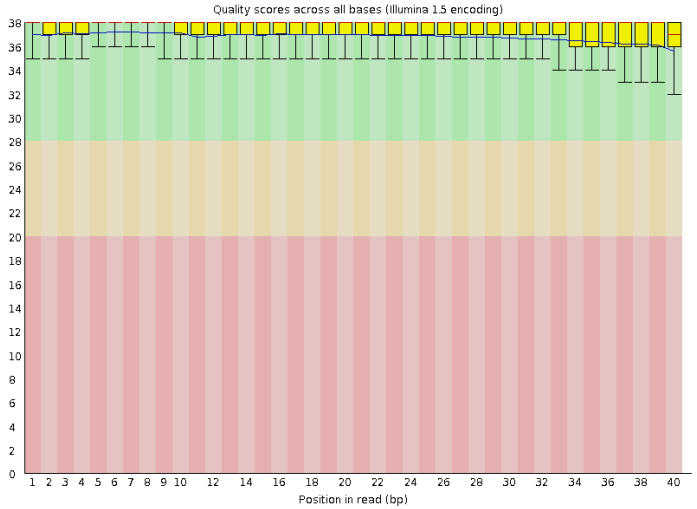
Measure	Value
Filename	good_sequence_short.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	250000
Sequences flagged as poor quality	0
Sequence length	40
%GC	45

FastQC – Per Base Sequence Quality

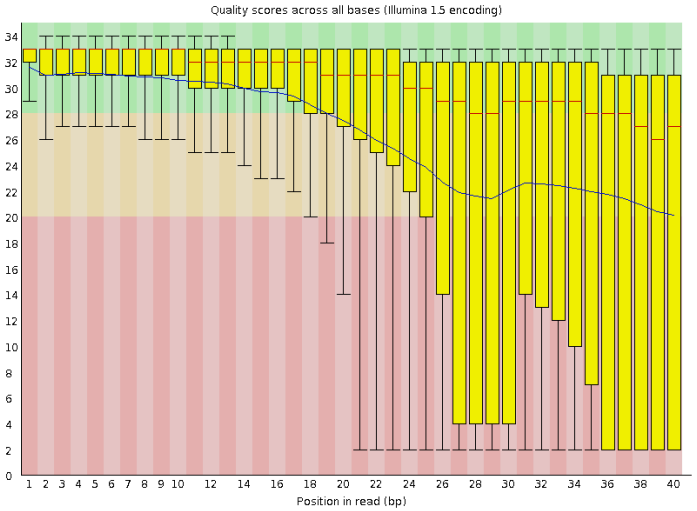
- This view shows an overview of the range of quality values across all bases at each position in the FastQ file.

- For each base there is a BoxWhisker with the following elements:
  - Central red line is the median value
  - Yellow box represents the inter-quartile range
  - Upper and lower whiskers represent the 10% and 90% points
  - Blue line represents the mean quality

- The y-axis on the graph shows the quality scores
  - Green region: good quality
  - Orange region: reasonable quality
  - Red region: poor quality



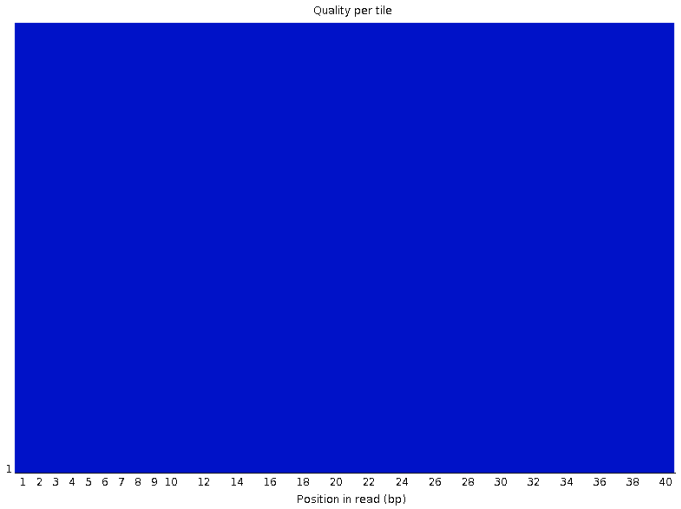
Good!



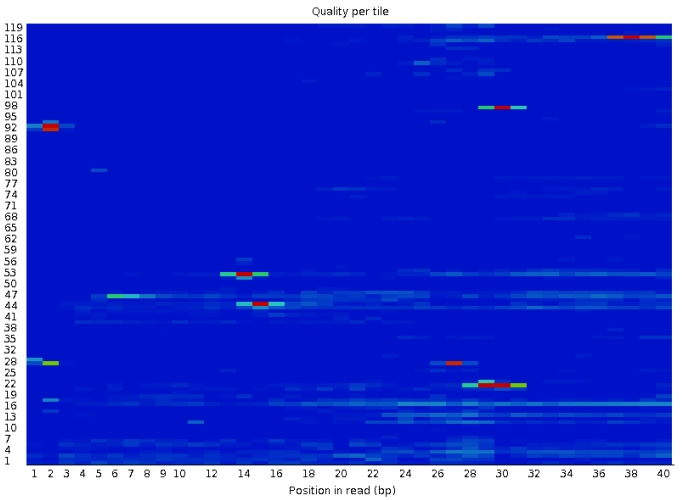
Bad!

FastQC – Per Tile Sequence Quality

- This graph will only appear if you're using an Illumina library which its original sequence identifiers
- The graph allows you to look at the quality scores from each tile across all of your bases
- Hotter colours indicate that a tile has worse quality than colder one have a better quality.



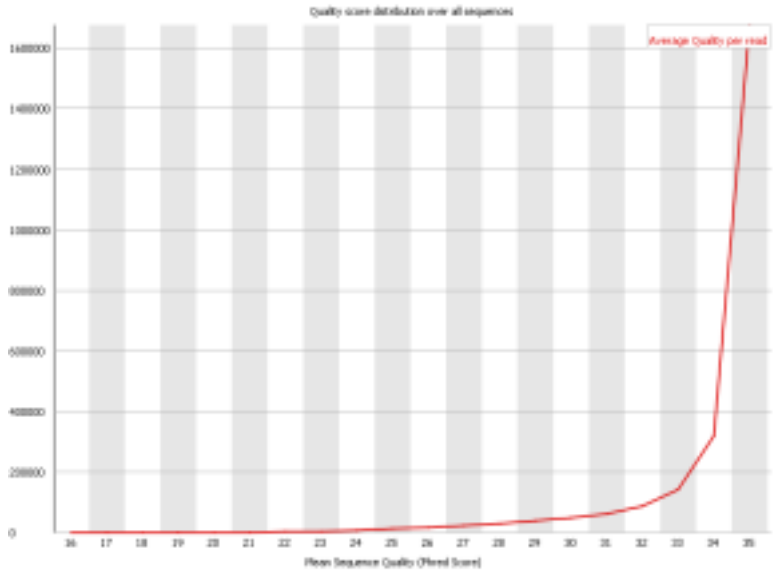
Good!



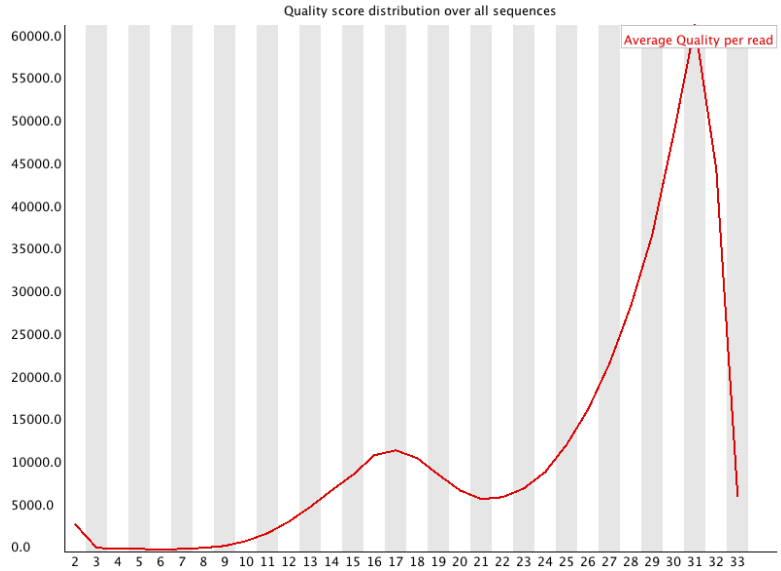
Bad!

FastQC – Per Sequence Quality scores

- This report allows you to see if a subset of your sequences have universally low quality values.



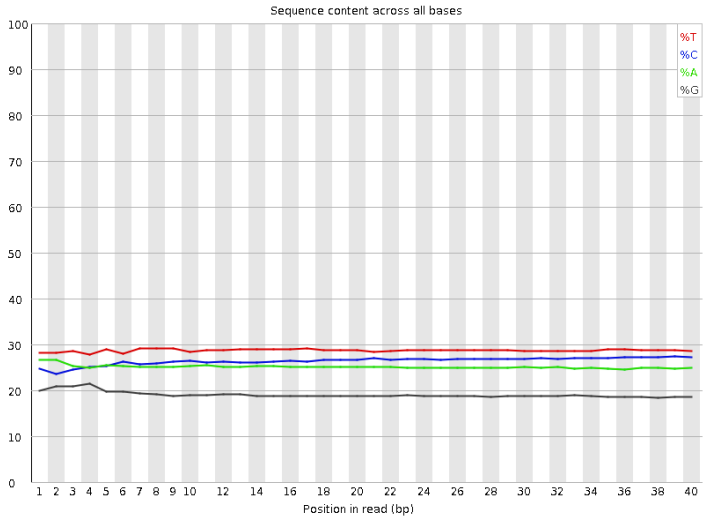
Good!



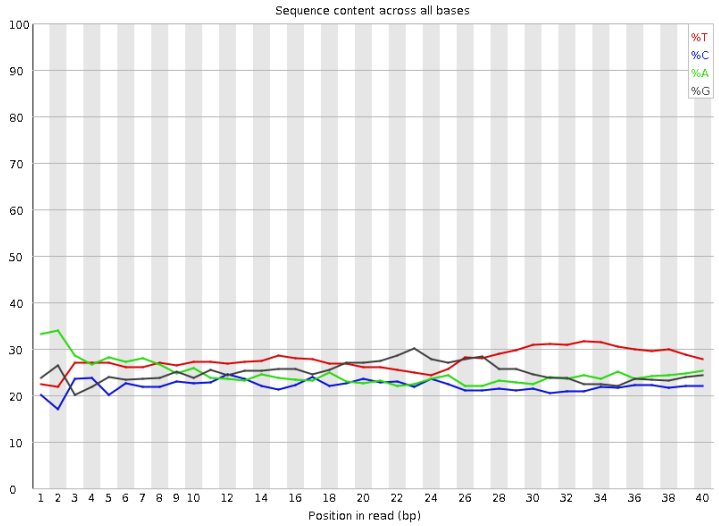
Less good!

FastQC – Per Base Sequence Content

- Per report plots out the proportion of each base position in a file for which each of the DNA bases has been called
- In a normal random library you would expect to see a roughly normal distribution of GC content where the central peak corresponds to the overall GC content of the underlying genome
- Since we don't know the the GC content of the genome the modal GC content is calculated from the observed data and used to build a reference distribution



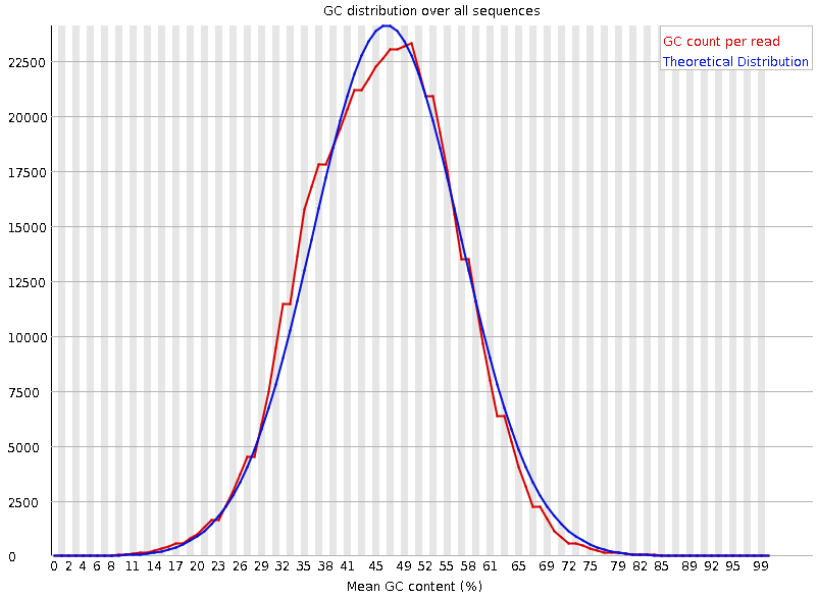
Good!



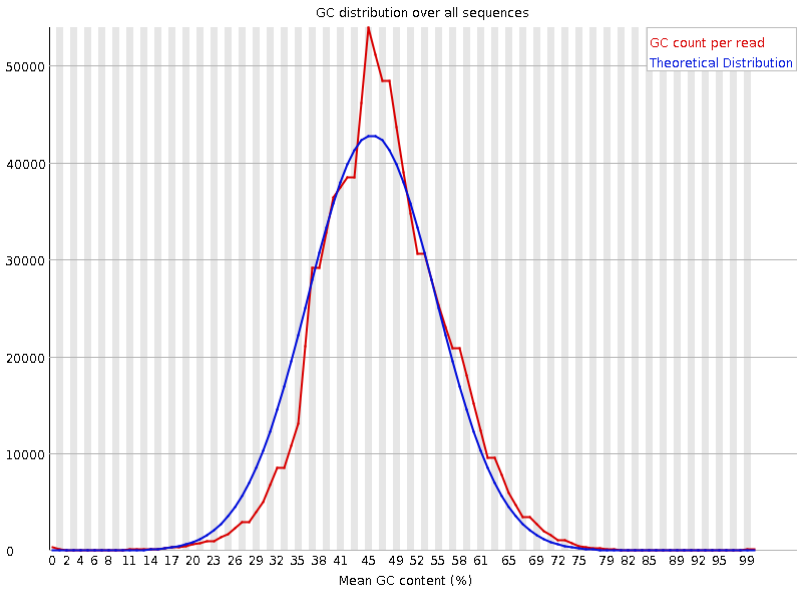
Less good!

FastQC – Per Sequence GC Content

- Per Base GC Content plots out the GC content of each base position in a file
- In a random library you would expect that there would be no difference between the different bases of a sequence run, so the lines in this plot should run parallel with each other



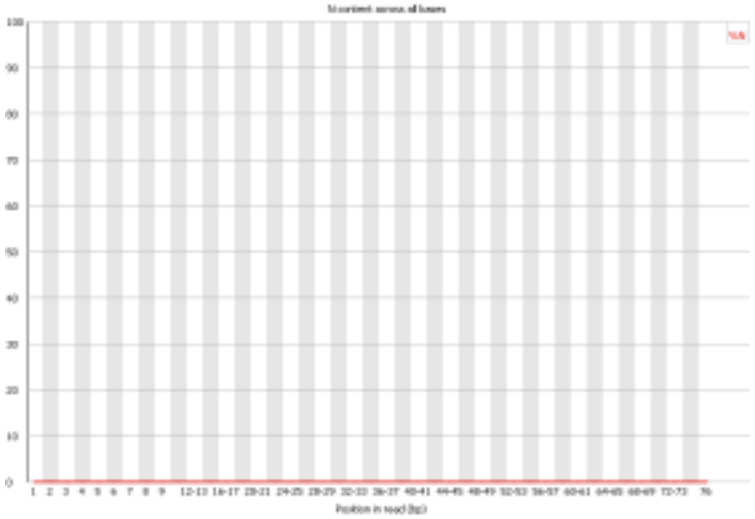
Good!



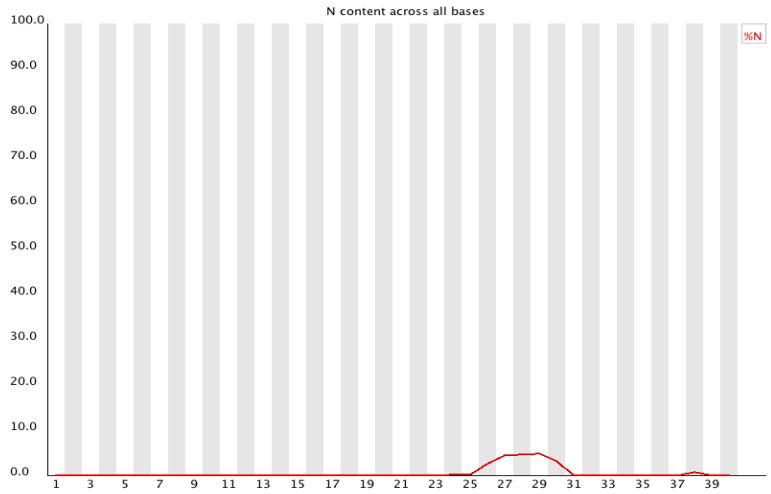
Less good!

FastQC – Per Base N Content

- If a sequencer is unable to make a base call with sufficient confidence then it will normally substitute an N rather than a conventional base call
- This module plots out the percentage of base calls at each position for which an N was called



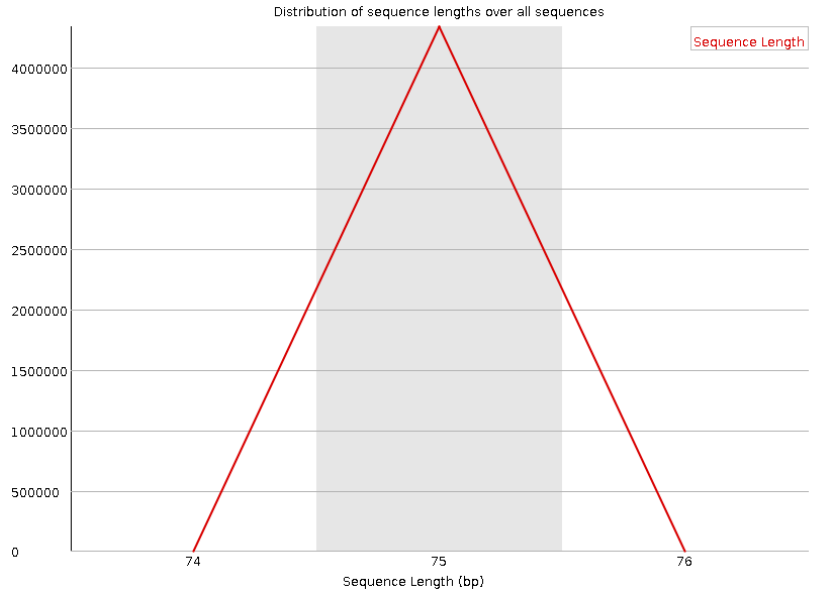
Good!



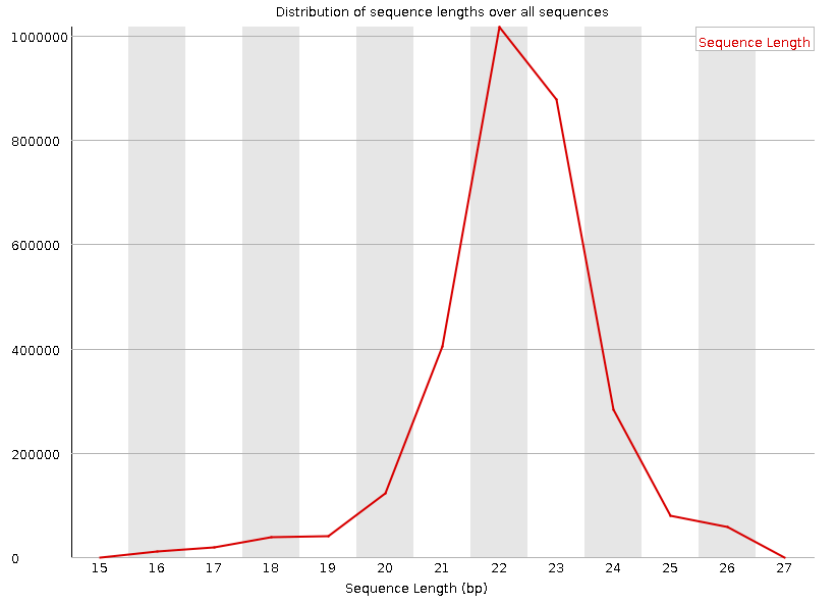
Less good!

FastQC – Sequence Length Distribution

- Some high throughput sequencers generate sequence fragments of uniform length, but others can contain reads of wildly varying lengths
- This report shows you the length distribution of your reads



All reads of length 75



Non uniform length distribution

FastQC – Overrepresented Sequences

- A normal library will contain a diverse set of sequences, with no individual sequence making up a tiny fraction of the whole

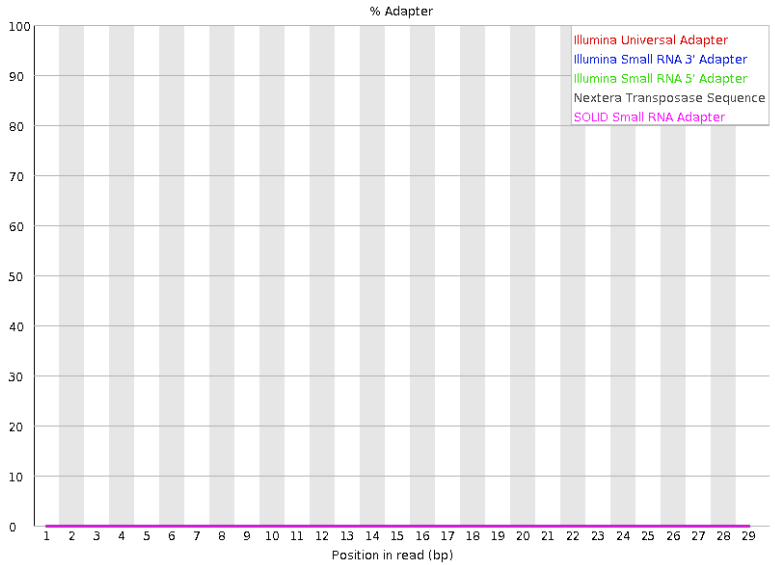
- If a single sequence is very overrepresented in the set either means that it is highly biologically significant, or indicates that the library is contaminated, or not as diverse as you expected
  - In RNA-Seq experiments sequences may naturally be present in a significant proportion
  - Adapters presence

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACTCGTCTGAATCTCGTAT	98856	3.977007572953645	TruSeq Adapter, Index 16 (97% over 38bp)
CGGCGCGCGGCGCGGGCTCCGGGGCGGCGGGTCCAACCCCGCGGGGGTTC	4502	0.18111685778746167	No Hit
CGGGTATCTGGCTTCCTCGGCCCCGGGATTCGGCGAAAGCTGCGGCCGGA	4237	0.17045582550987898	No Hit
CGGGGGTCGGCGCGCGGCGCGGGCTCCGGGGCGGCGGGTCCAACCCCGCG	3783	0.1521912645513033	No Hit
CTCGTCGCGGCGTAGCGTCCGCGGGGCCCGACGCCGCGGGGGCGAAACCC	3726	0.149898136853861	No Hit
CTCCTACTCGTCGCGGCGTAGCGTCCGCGGGGCCCGACGCCGCGGGGGCG	2754	0.11079427506589724	No Hit
CTCGCGTCCAGAGTCGCCGCCGCCGCCGGCCCCCCCGAGTGTCCGGGCCC	2567	0.10327120700586719	No Hit

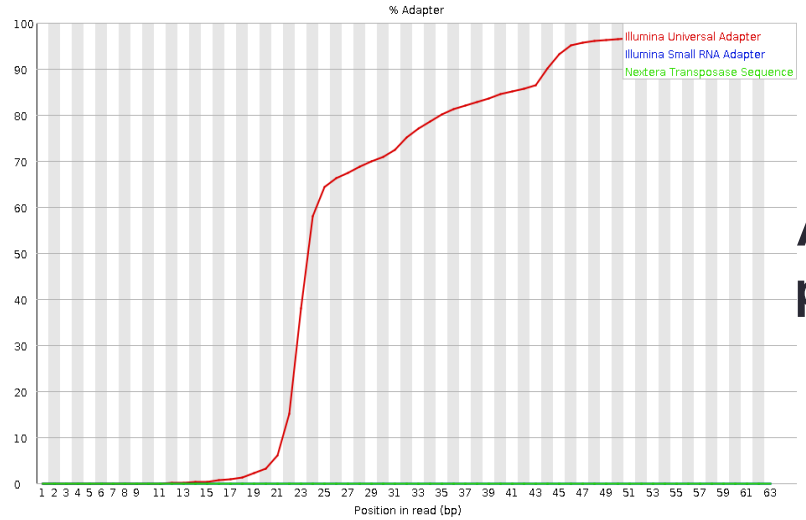
FastQC – Adapter Content

- A class of overrepresented sequences which you might want to analyse are adapter sequences. It is useful to know if your library contains a significant amount of adapter in order to be able to assess whether you need to adapter trim or not

- FastQC looks for a list of contaminants in your data and shows a cumulative percentage count of the proportion of your library which has seen each of the adapter sequences at each position



No Adapter!



Adapter present!