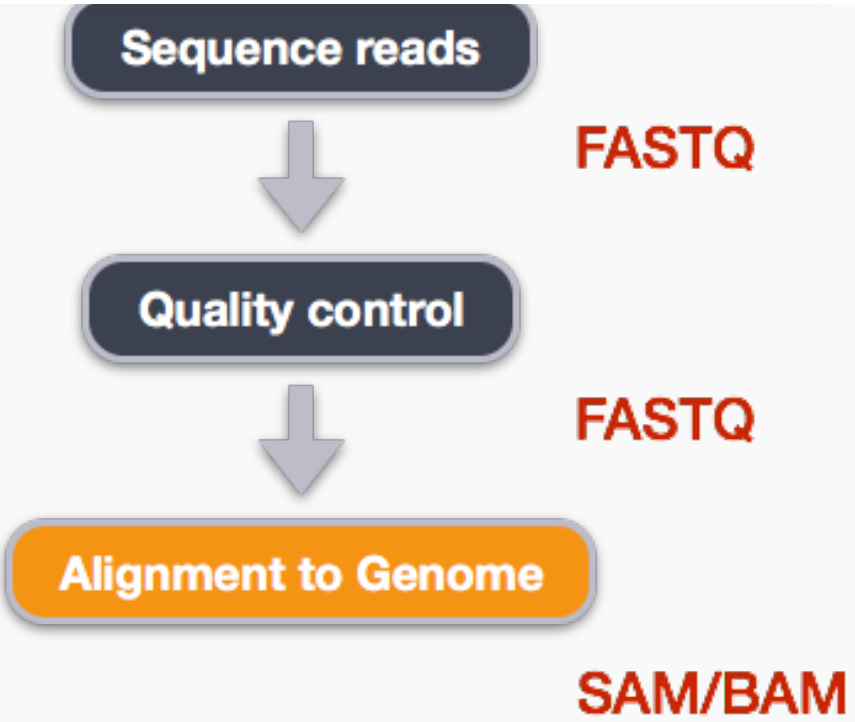
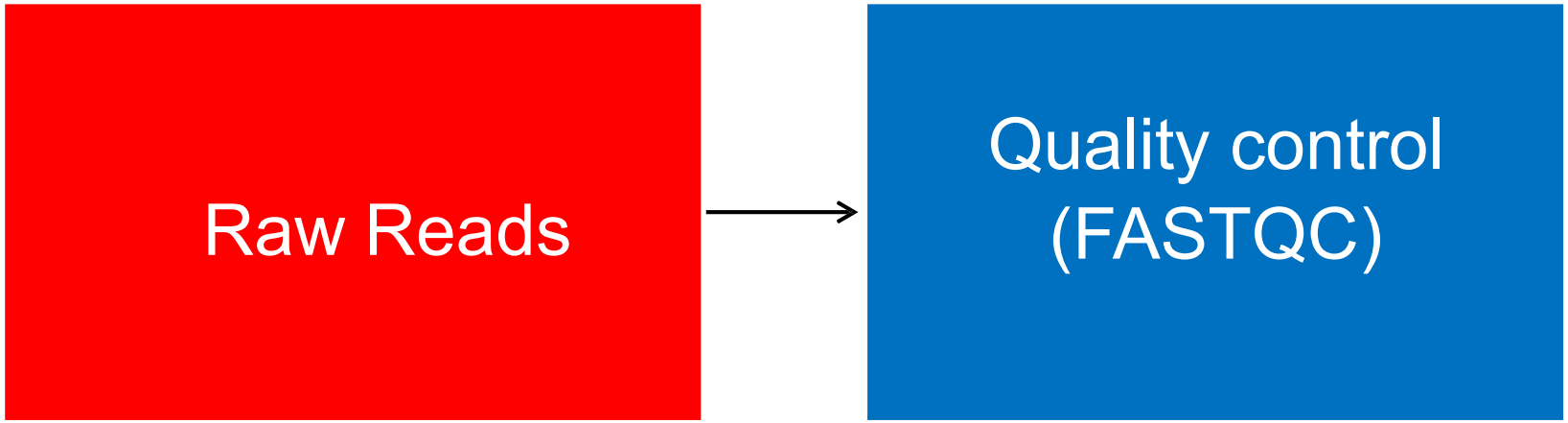
FASTQC

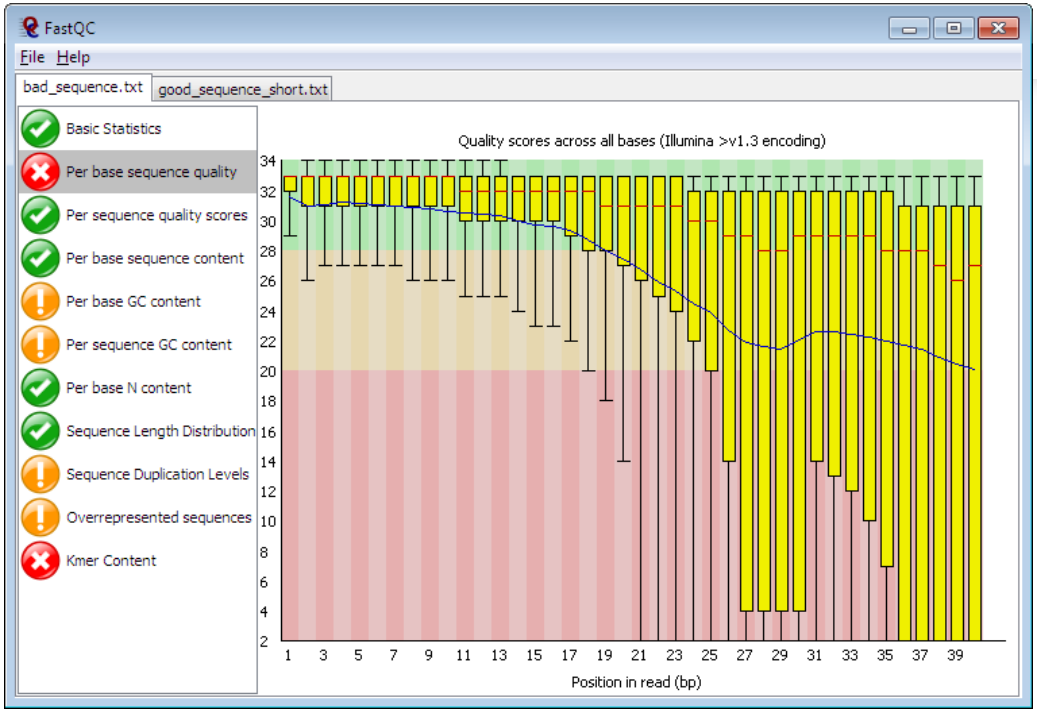
February 16th, 2023



Workflow



FastQC



Reads raw fastq files as input

Performs multiple checks

- Pass/warn/fail
- Compares to genomic library

HTML Report

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

Sequence Output Format

- FASTQ

  Line 1: Unique ID for a sequencing read

  Line 2: Sequences

  Line 3:+

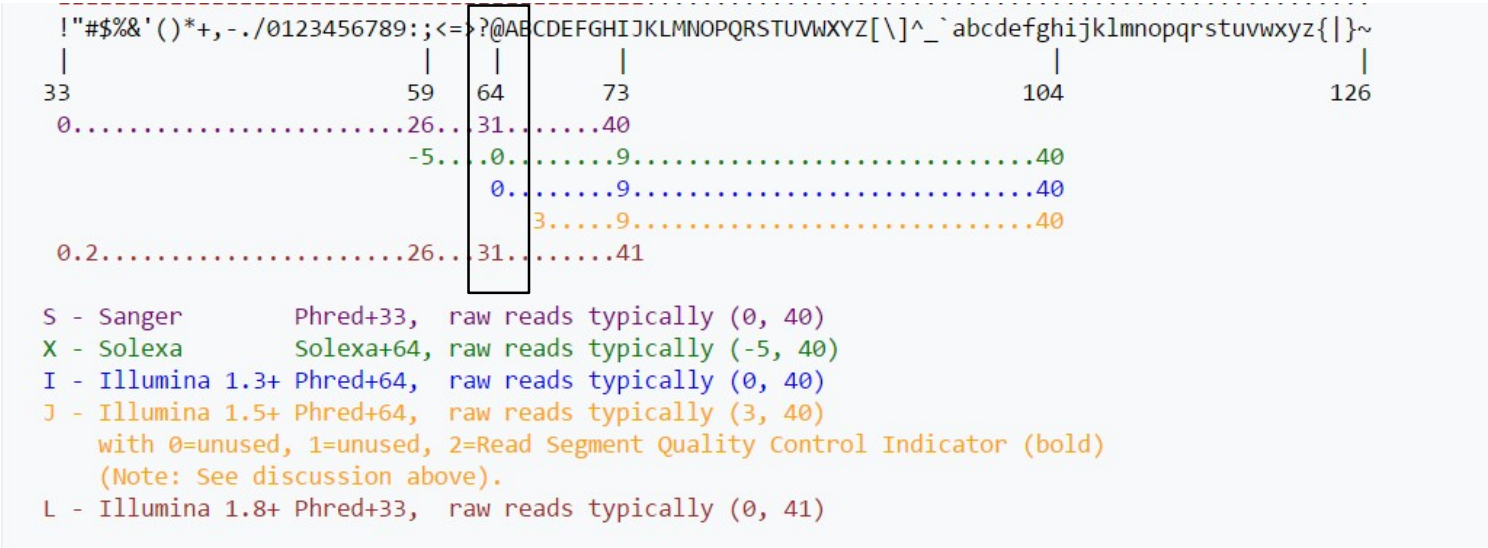
  Line 4: Base calling quality score (American Standard Code for Information Interchange *(*ASCII) value)

  Example:

```
@HISEQ:126:H14YJADXX:1:1101:1118:2101 1:N:0:ATCACG
CTCCATAGTCAGAAACTTCAGCATGACAGTACCTCATGCTGCATCAGGTGATCATGAAAAGATTACAGGCTTTCTAAAATTATCAGCAAGATATGG
+
@@?ADDDD?ADHDIIIIIIIEIIIGEFHC<?FH4C9E9BGAFIGH<DG9BD?@DGGEGHHG<DCBBCC8C>FHCGEHIGEEE>EEHEEEEC>A>;;
```

Quality Score Representation

- Quality scores are represented as ASCII characters in order to save space, so that there is one ASCII character per base.
- Converts between the ASCII value into Phred score



Quality Scores

A quality value (Q) is an integer representation of the probability p that the corresponding base call is incorrect

ASCII Quality Score	Probability of Incorrect Based Call	Base Call Accuracy	Q-score
+	1 in 10	90%	Q10
5	1 in 100	99%	Q20
?	1 in 1000	99.9%	Q30
I	1 in 10000	99.99%	Q40

$$Q\ score = -10\ \log_{10} p$$

Base Call Qualities (Phred scores)

For most runs, quality should be good for most reads through the whole run

If quality deteriorates we should understand how and why

Good (Illumina) quality is generally Phred > 30

Concerning (Illumina) quality is Phred < 20

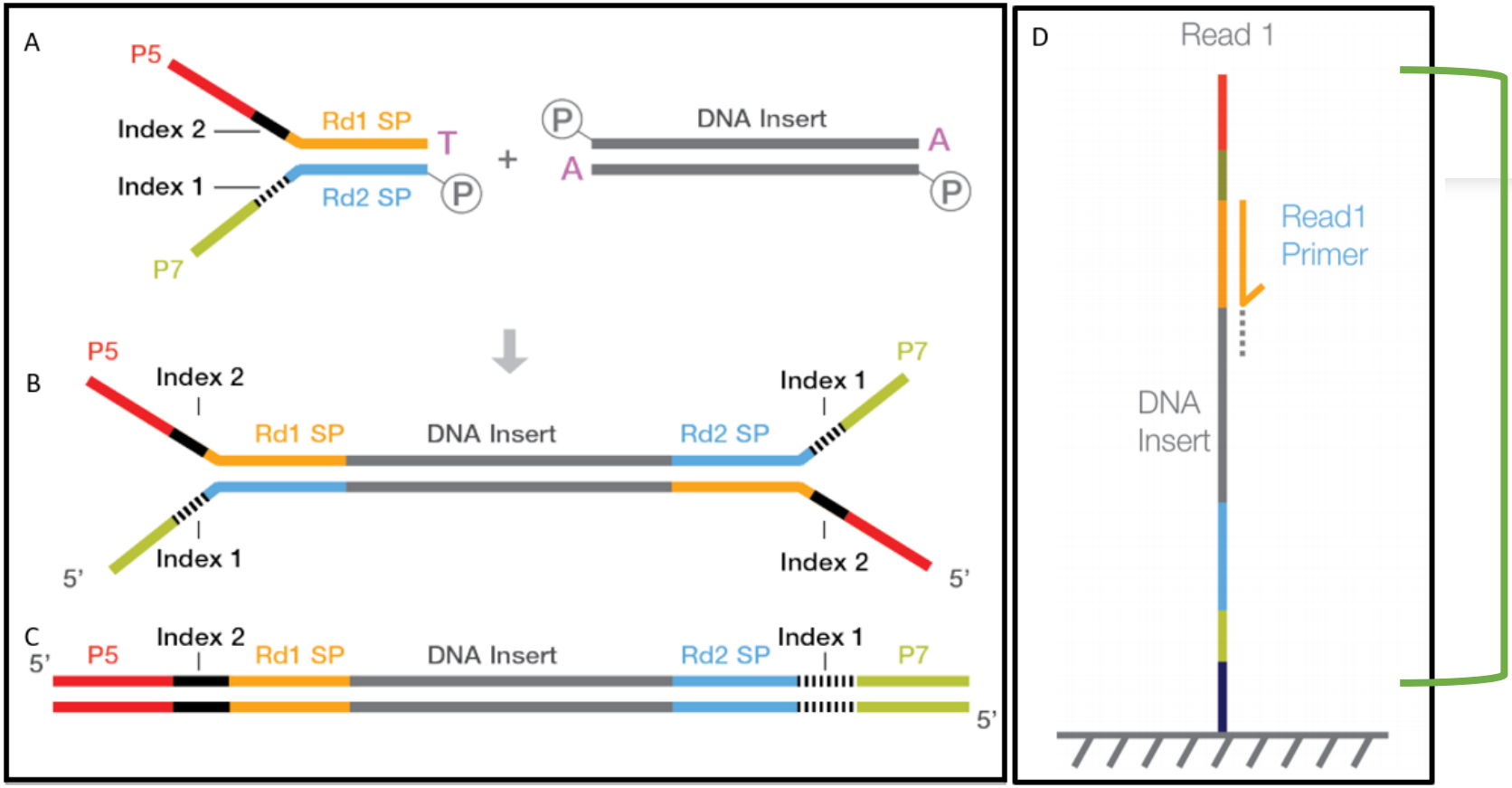
What should you look at in your libraries?

Basic Statistics

Measure	Value
Filename	Mov10_oe_1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	39971841
Filtered Sequences	0
Sequence length	100
%GC	47

Architecture of Standard Illumina NGS library

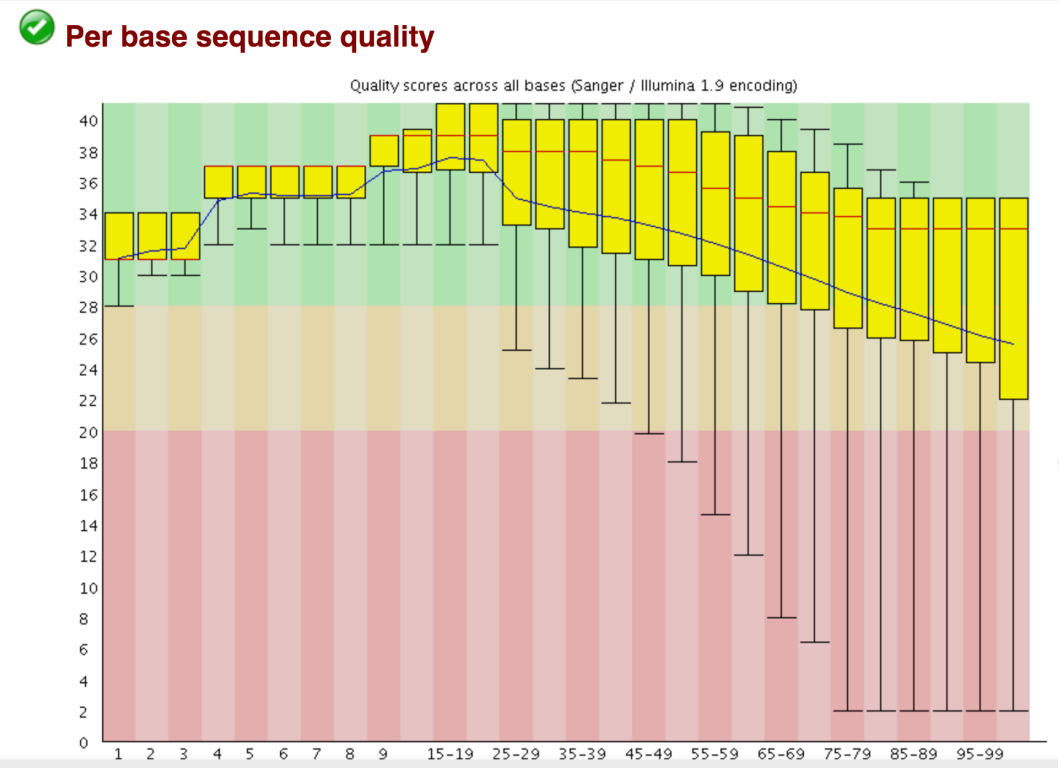
P5 and p7 sequences are required to bind the flow cell



Adaptors will serve as primer binding sites for amplification and sequencing

Indices are used to combine many samples into 1 seq. run

Phred Score



Cycles of Chemistry

For each position a BoxWhisker type plot is drawn.

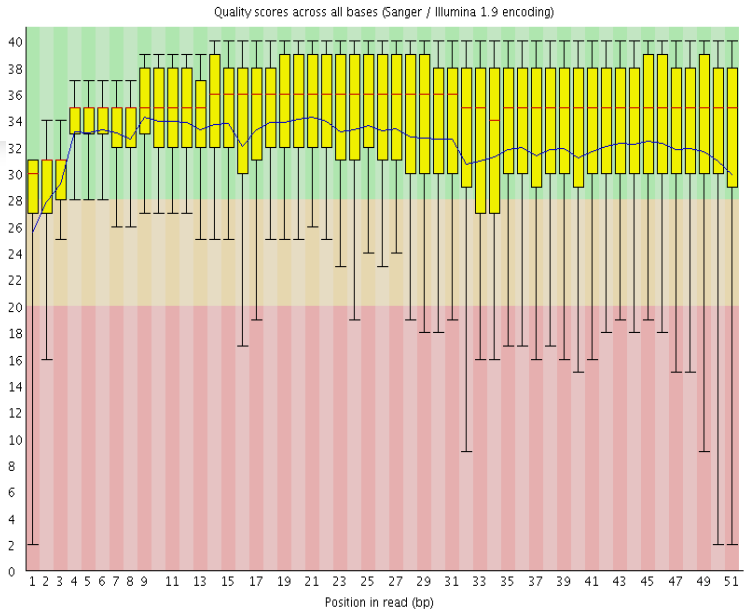
- The central red line is the median value
- The yellow box represents the inter-quartile range (25-75%)
- The upper and lower whiskers represent the 10% and 90% points
- The blue line represents the mean quality

FastQC Report

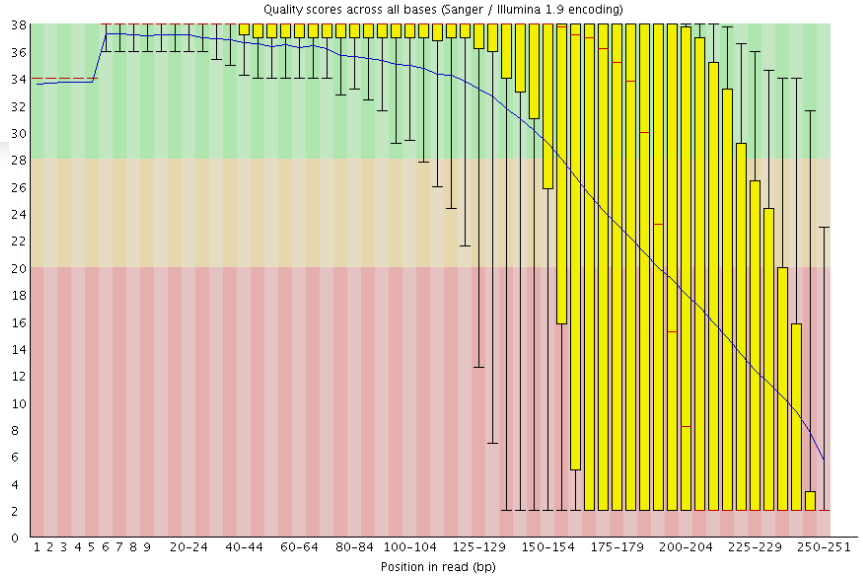
PHRED Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90 %
20	1 in 100	99 %
30	1 in 1000	99.9 %
40	1 in 10000	99.99 %
50	1 in 100000	99.999 %

Base Call Qualities – Per Cycle

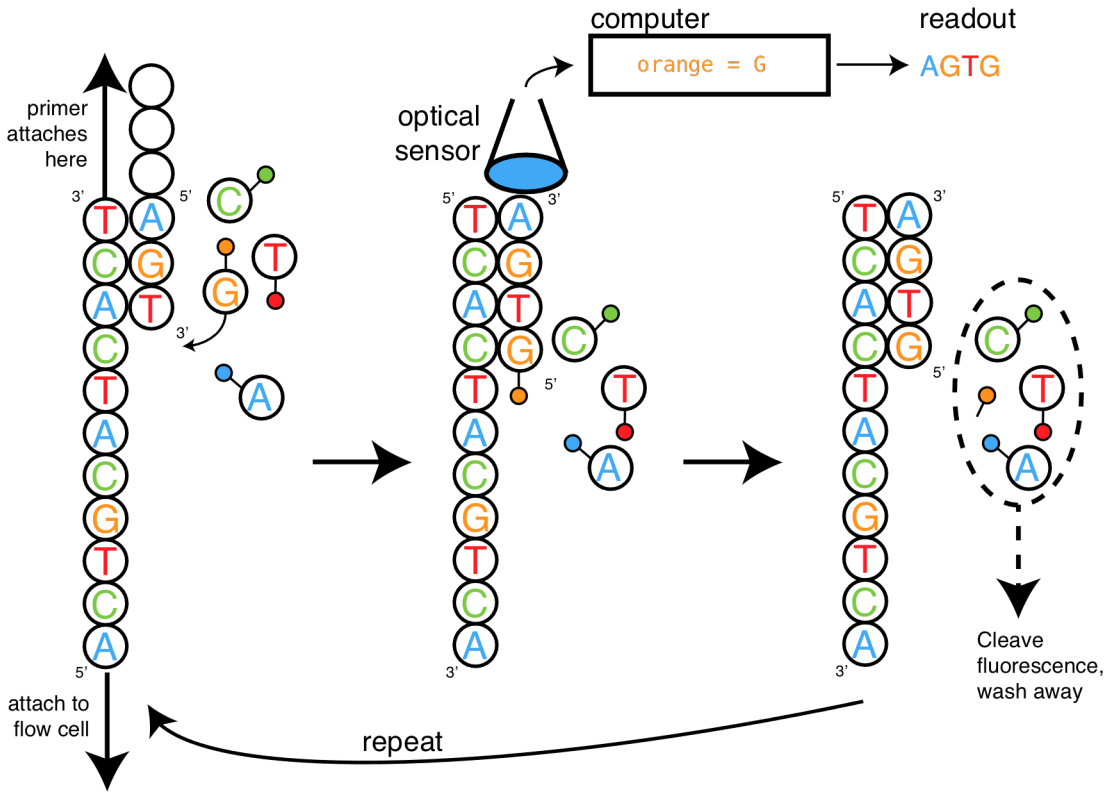
Read 1



Read 2



The quality of base call tend to degrade as the run progresses



Then sequencing begins!

- Have fluorescent-labeled nucleotides
- Each dNTP has a corresponding color
- During each **cycle**, a labeled dNTP is added to the growing chain
- An image is then taken
- Then the fluorescent dye is cleaved to allow for the next nucleotide to be incorporated in the next **cycle**

Problem

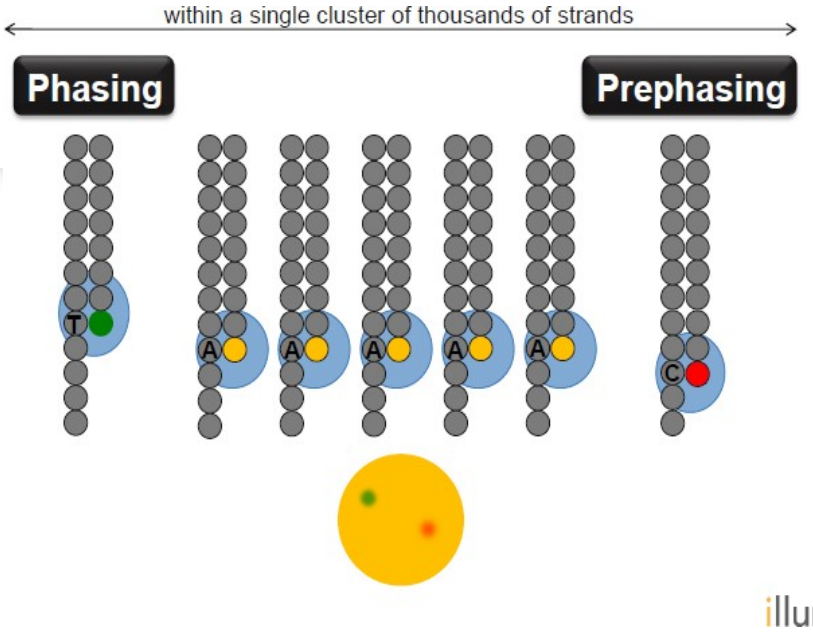
As the chain grows, signal quality deteriorates

At each cycle, accumulates noise

Solution

Sequence only (up to 300bp)

Enables robust base-calling across the genome including in repetitive sequence regions



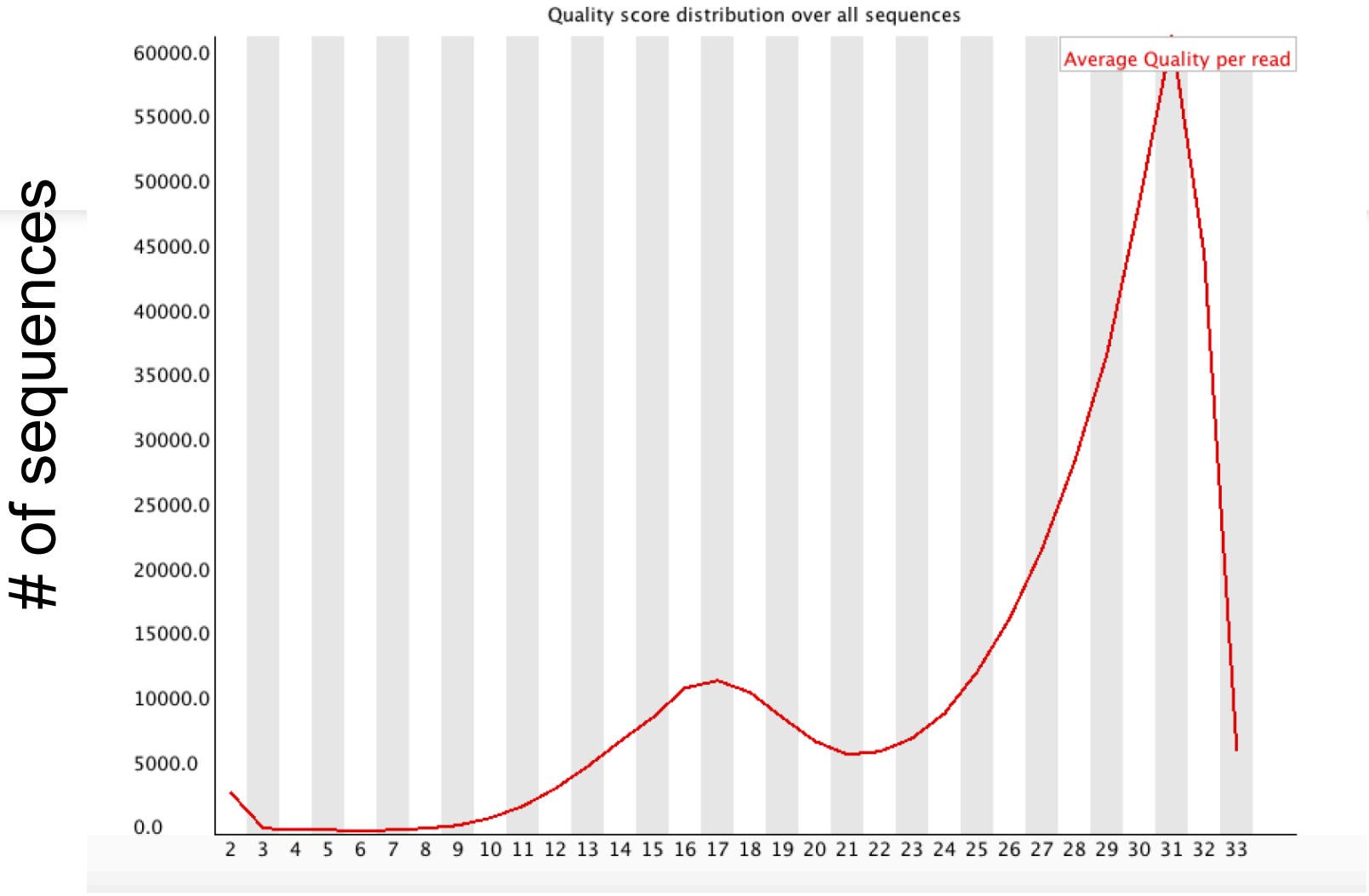
Trimming

If the quality of the library falls to a low level then the most common remedy is to perform quality trimming where reads are truncated based on their average quality.

Trimming must be performed on all samples - Before committing to any action, check in

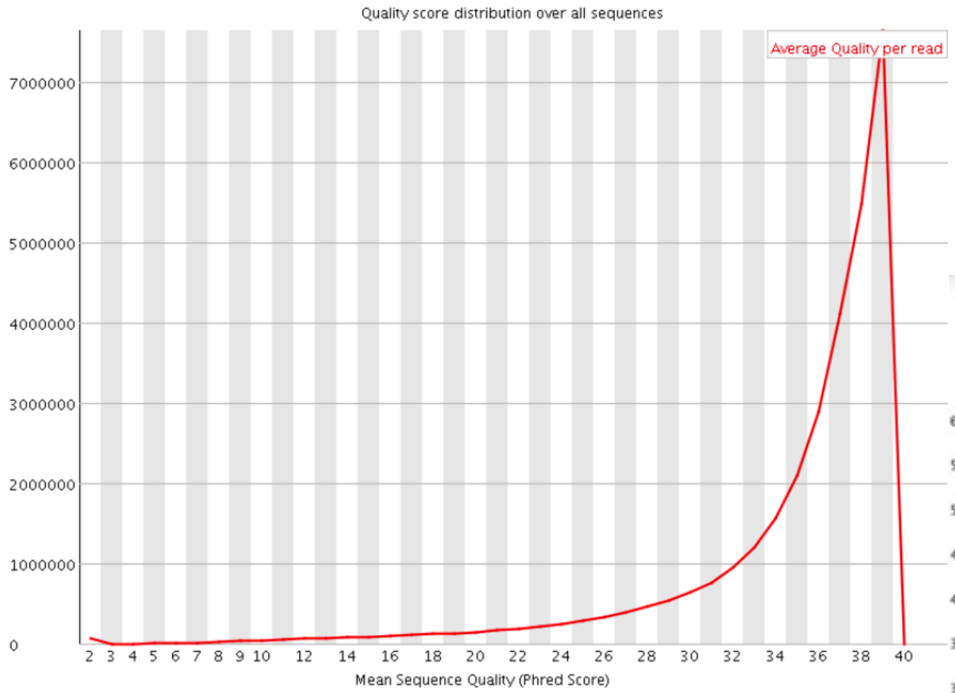
Per Sequence Quality Score

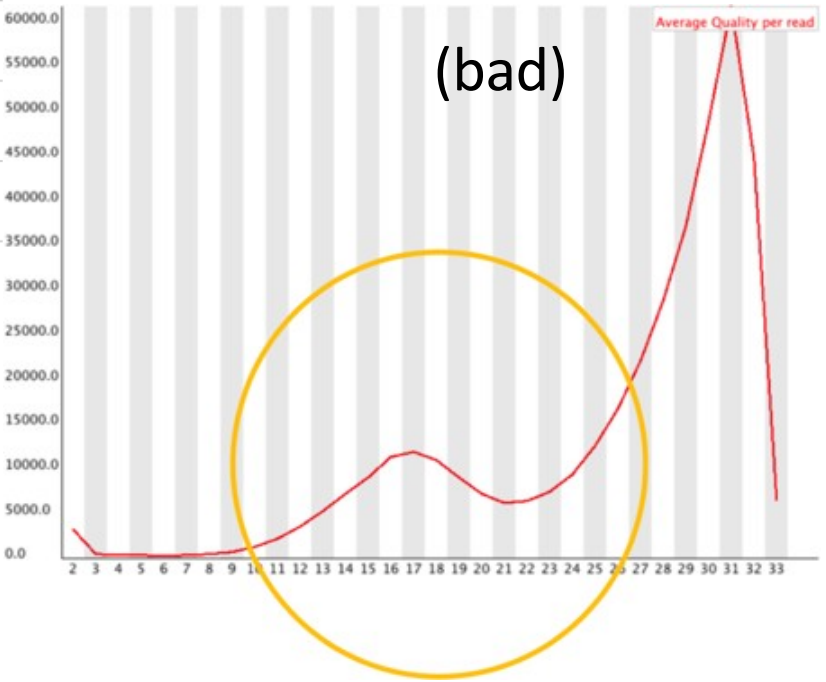
Is it a subset of the sequences that have low scores or is it the majority of them?



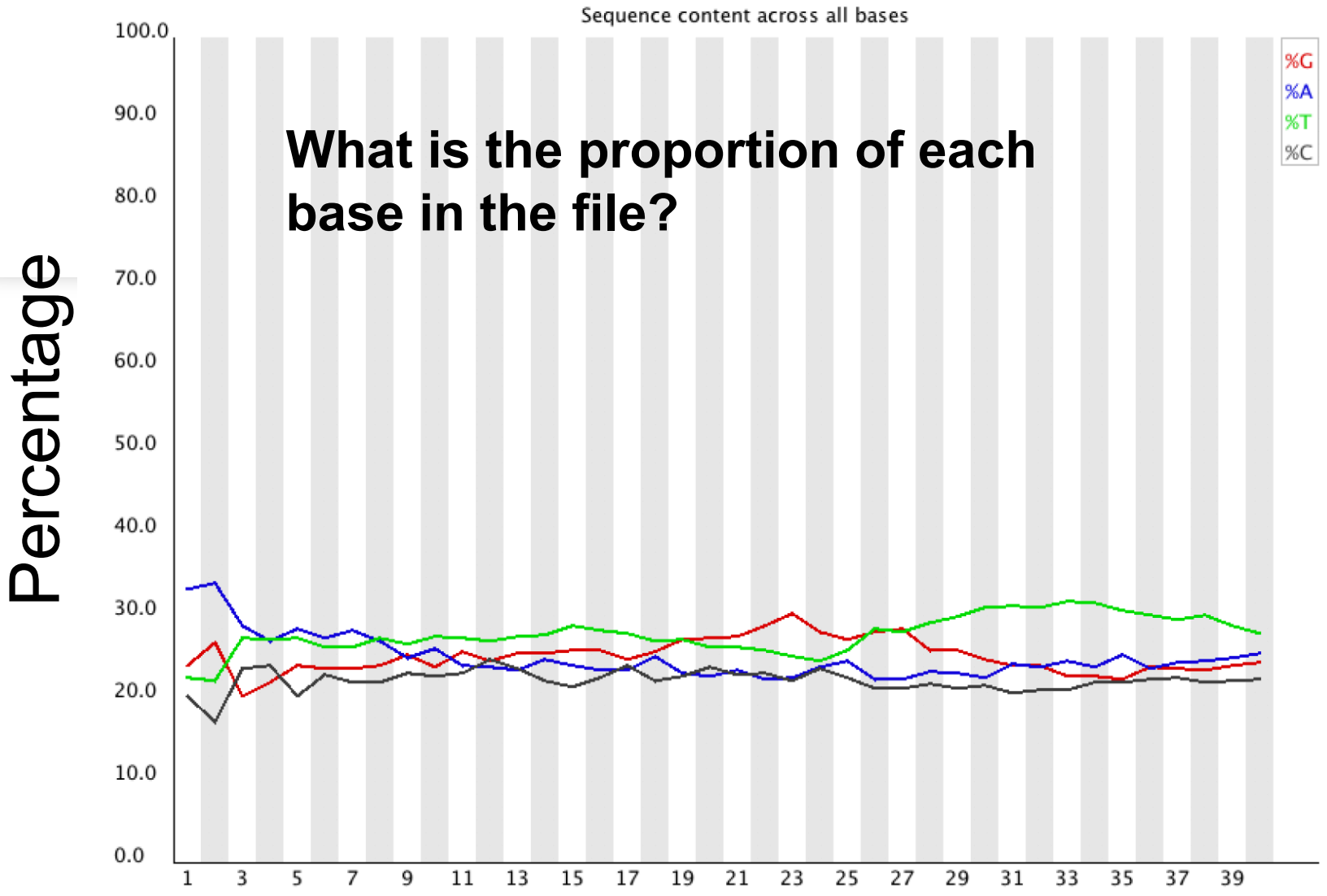
FastQC: per-read mean base qualities

Per sequence quality scores



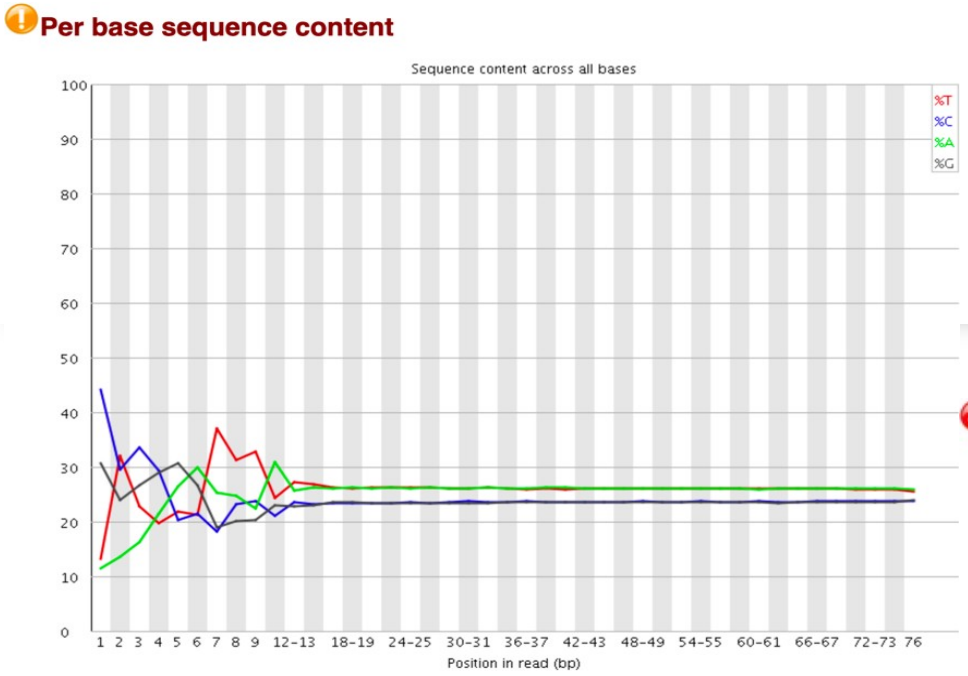


Per Base Sequence content



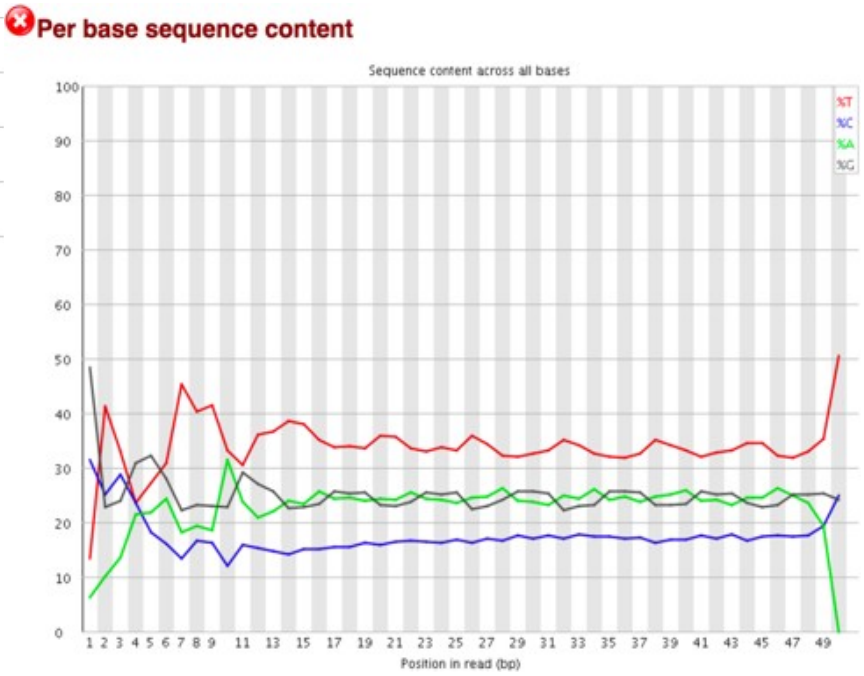
If there are 4 bases out of 100% what is the gold standard?

FastQC: %ACGT over read length

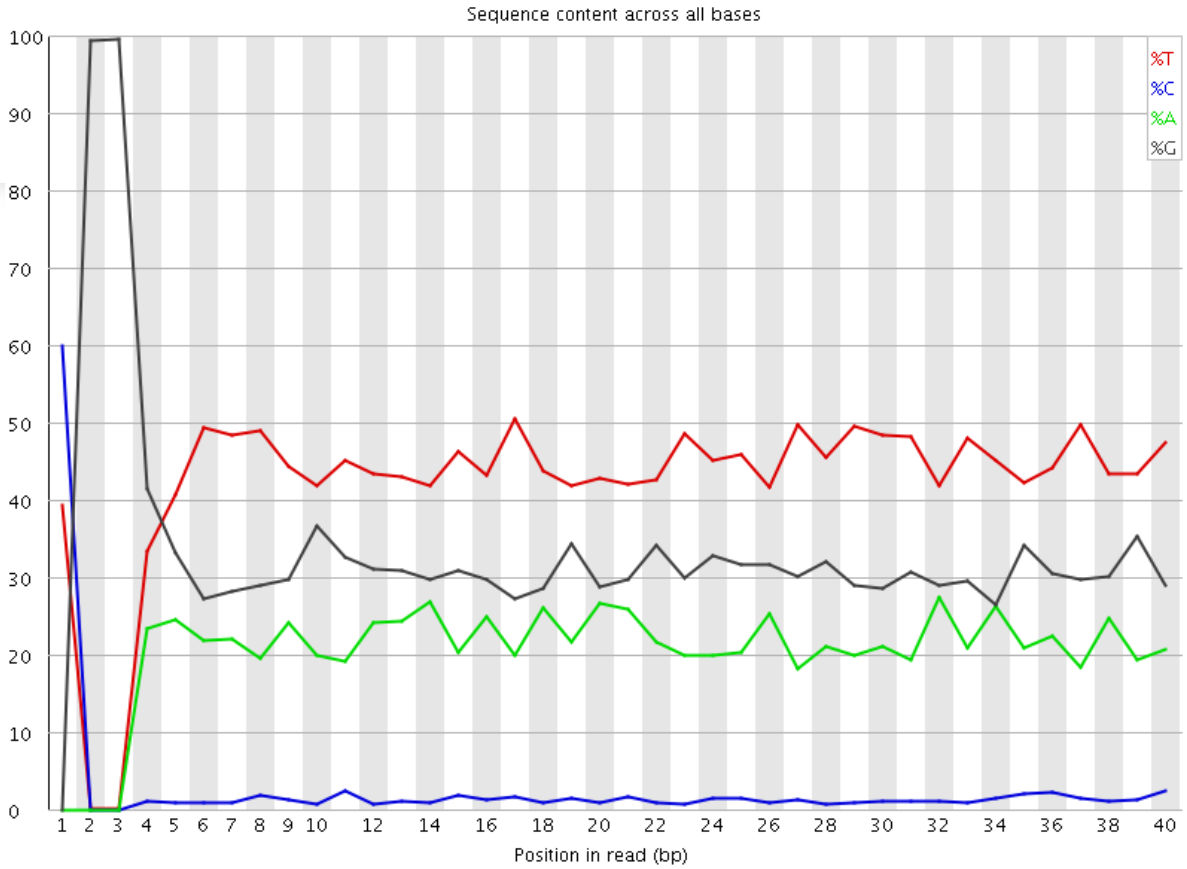


If there is a strong preference, this indicates an overrepresented sequence

(bad)



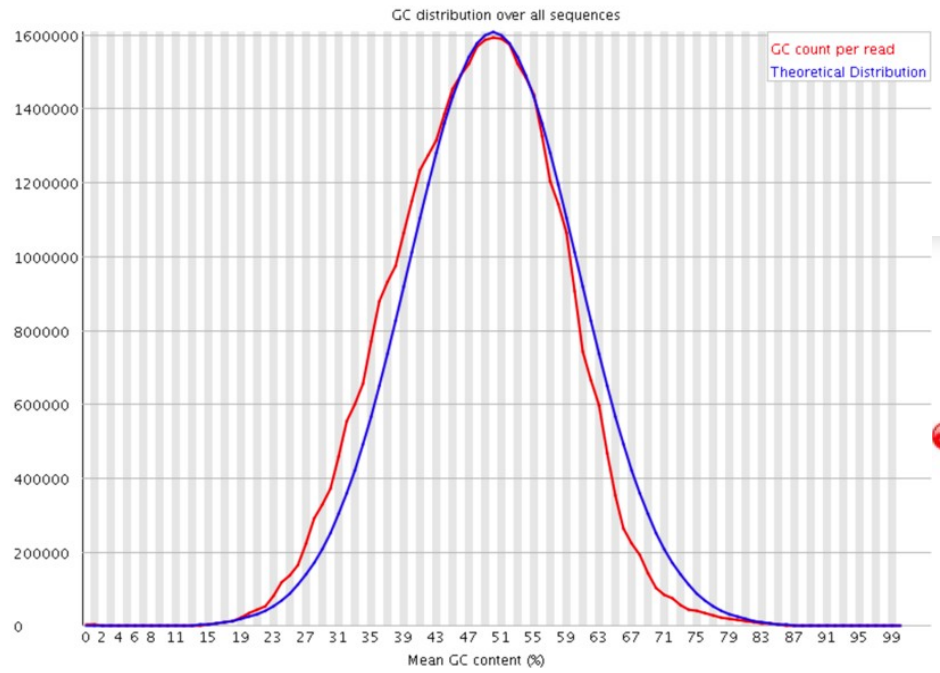
Library Base Composition



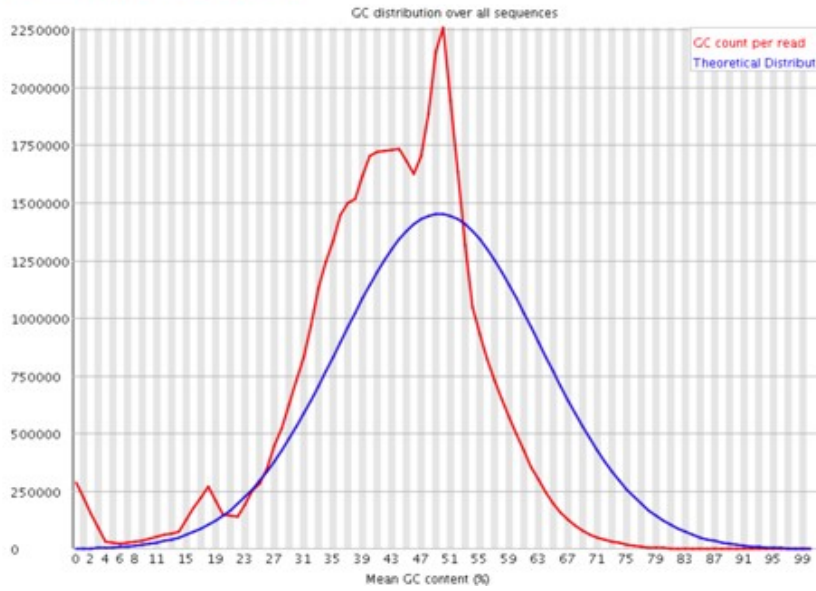
Bisulphite treated – C is converted to T

FastQC: per-read %GC

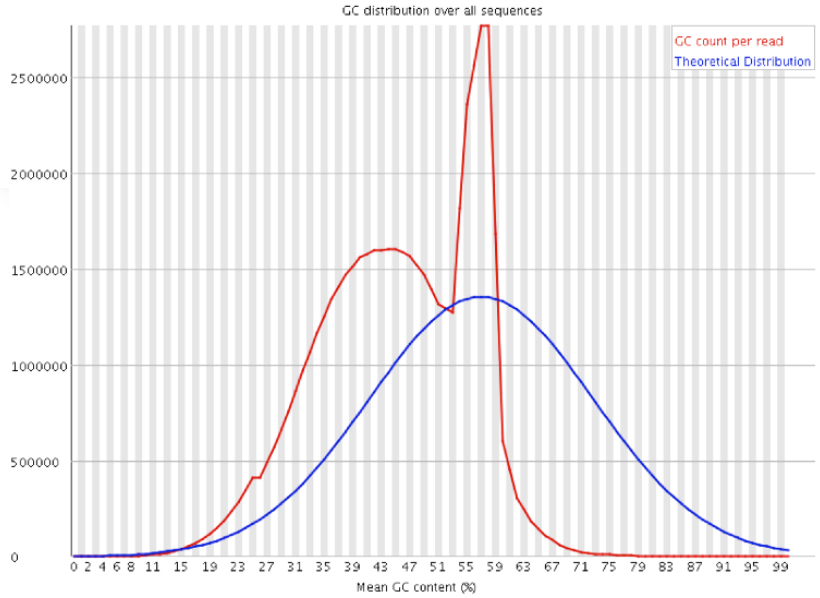
Per sequence GC content



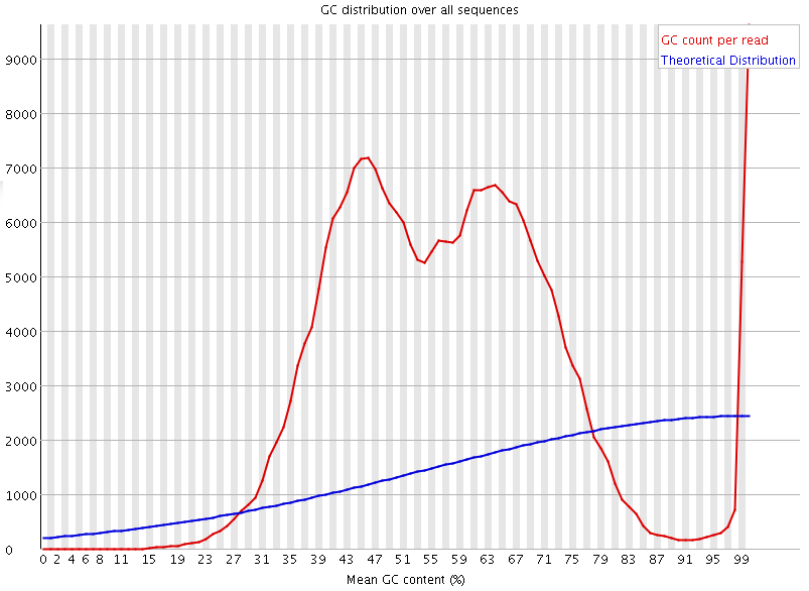
Per sequence GC content (bad)



Library GC Content

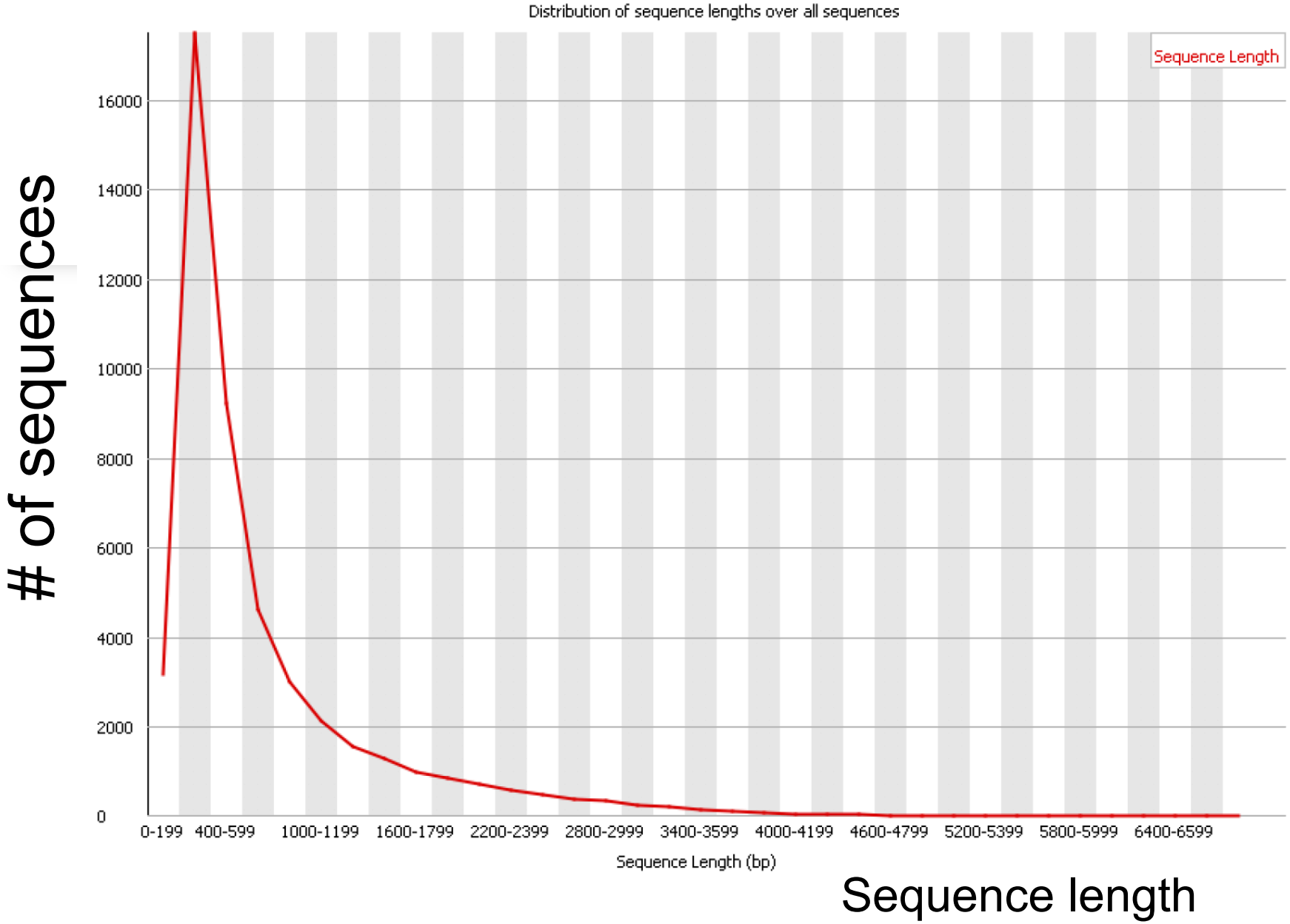


Specific Contamination



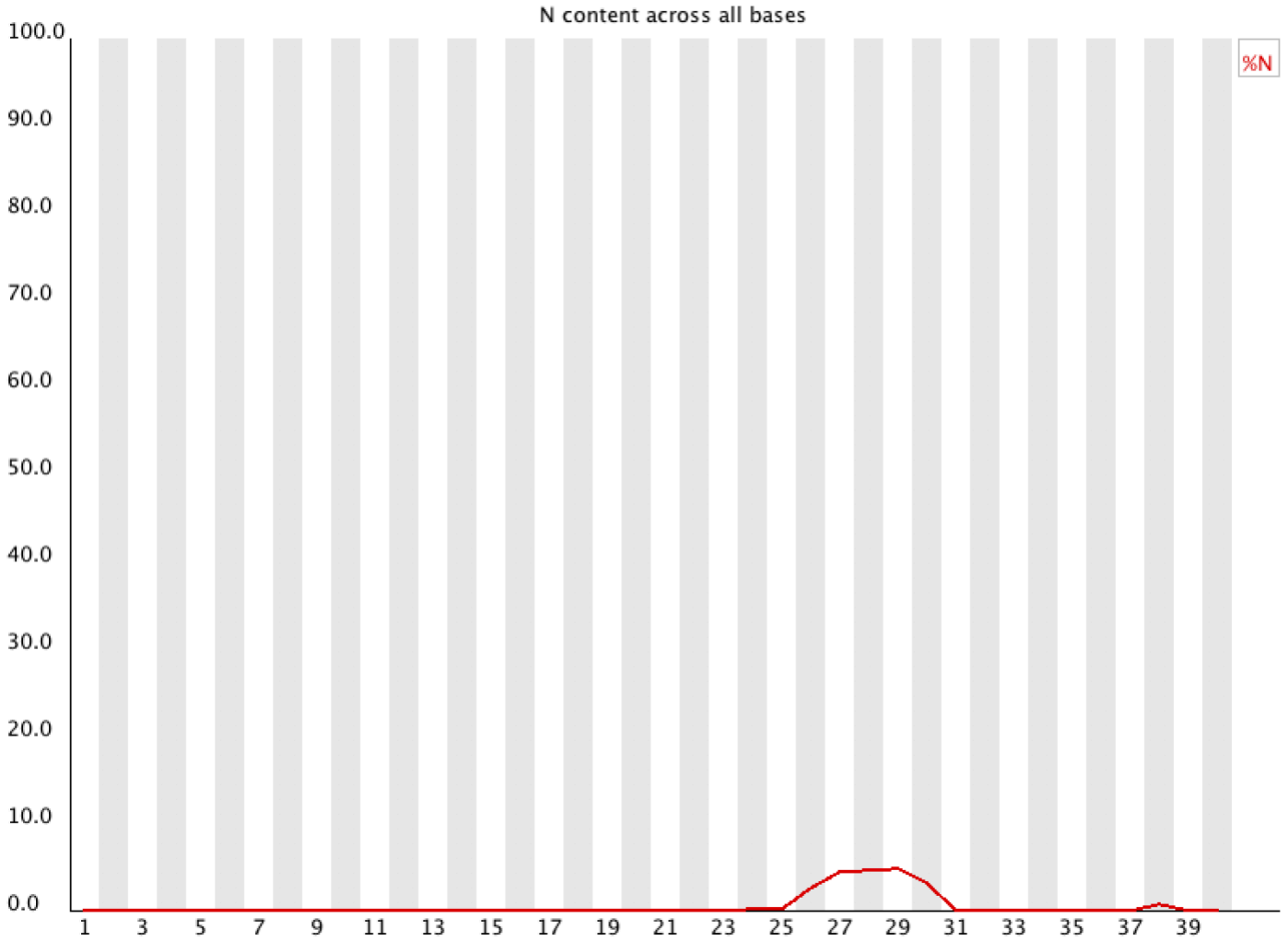
Broad Contamination

What is the average fragment length?

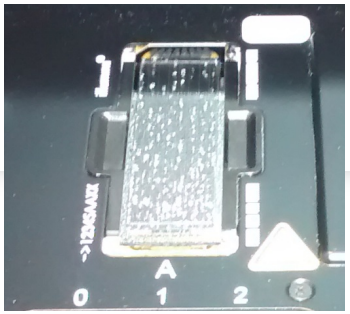


What is the percentage over each position for which an N was called?

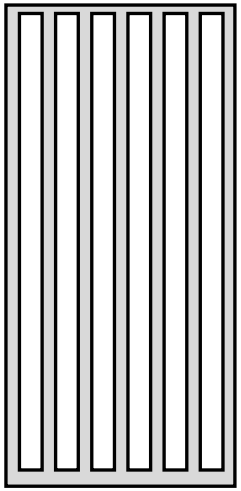
N= not enough confidence to call an A,T,C,G



Positional Quality

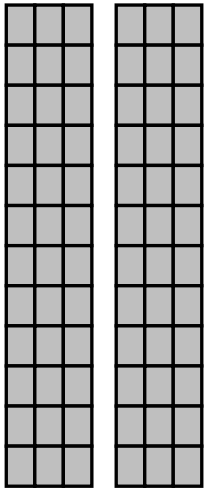


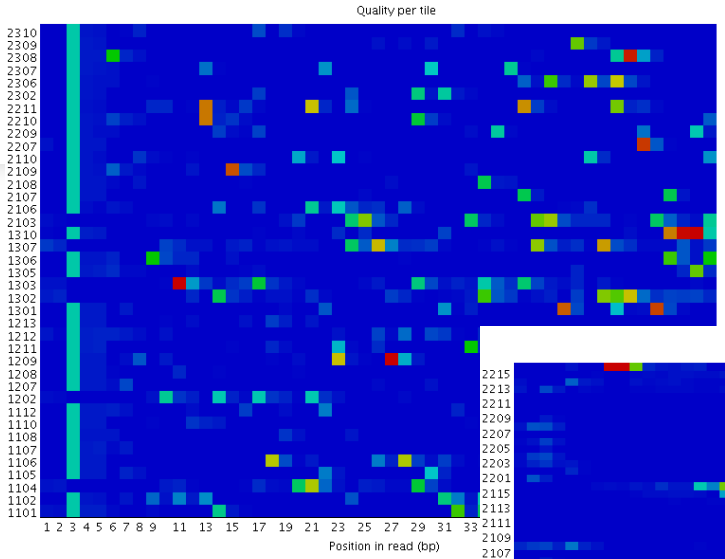
Lanes



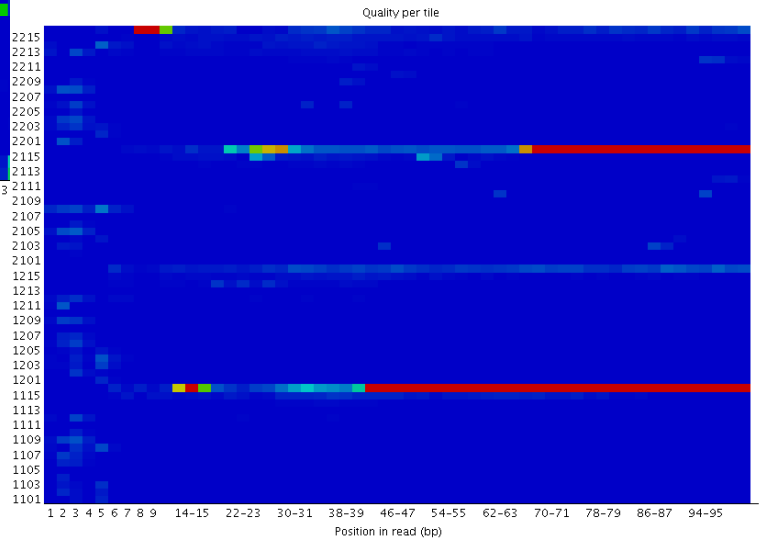
Tiles

Top Bottom





Random Patterning



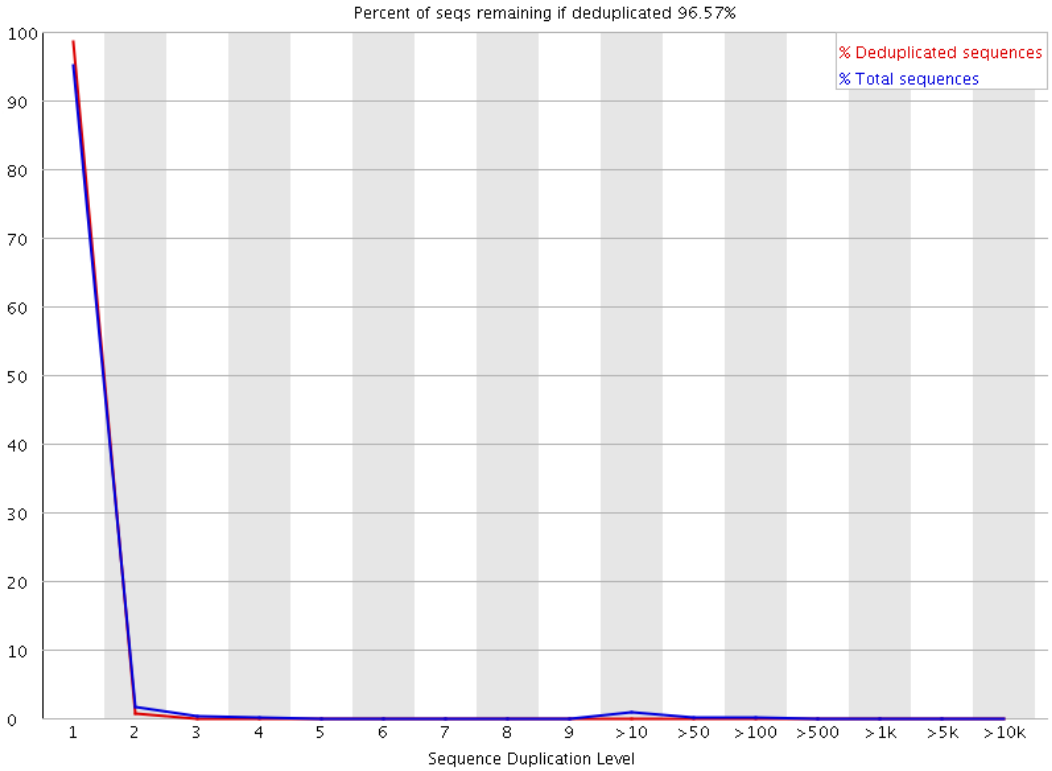
Position Specific Patterning

Duplication

The exact same sequence appears more than once in your library

- The sequences come from different biological molecules and the duplication is coincidental
  - Deep sequencing
  - Highly present sequences (repeats for example)
  - Restricted diversity libraries (amplicon sequencing, restricted libraries)
- The sequences come from the same biological and the duplication is technical
  - PCR duplicates

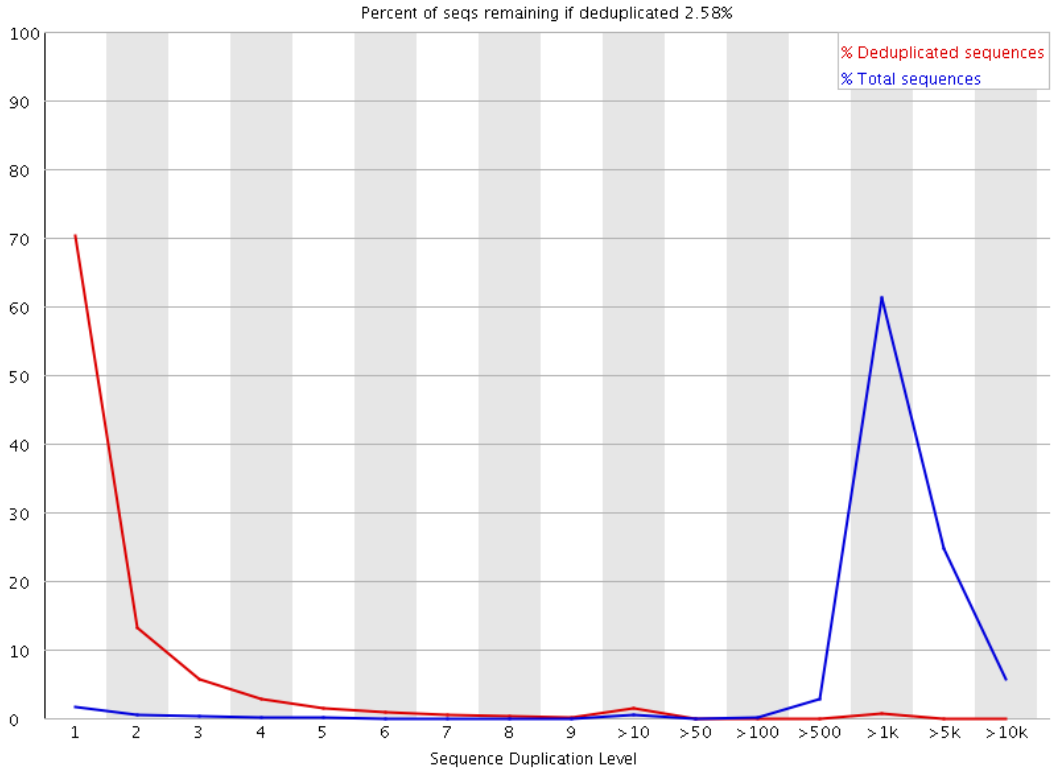
Duplication



Most sequences are unique

High diversity (or low coverage)

Duplication



Most sequences are present many times

Highly duplicated (low diversity)

Overrepresented Sequences

Sequence	Count	Obs/Exp Overall	Obs/Exp Max	Max Obs/Exp Position
AAAAA	16795015	4.1748657	6.059911	2

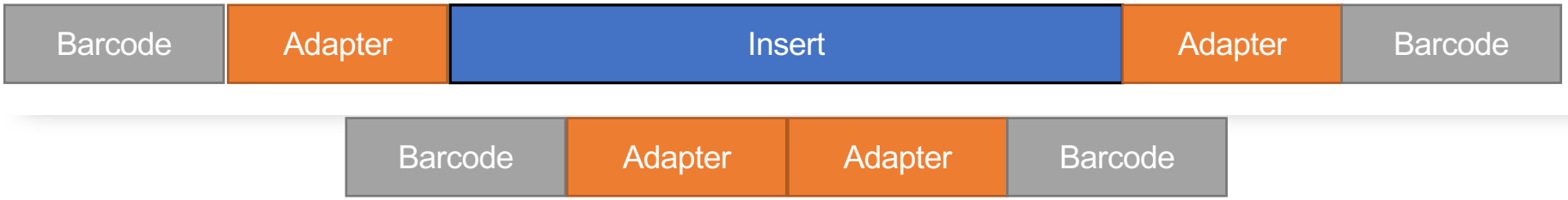
PolyA – Common in RNA-Seq

PolyG – Empty space in 2-colour chemistry

PolyN – Quality too poor to make any calls

Specific sequences – Normally Adapter Dimers

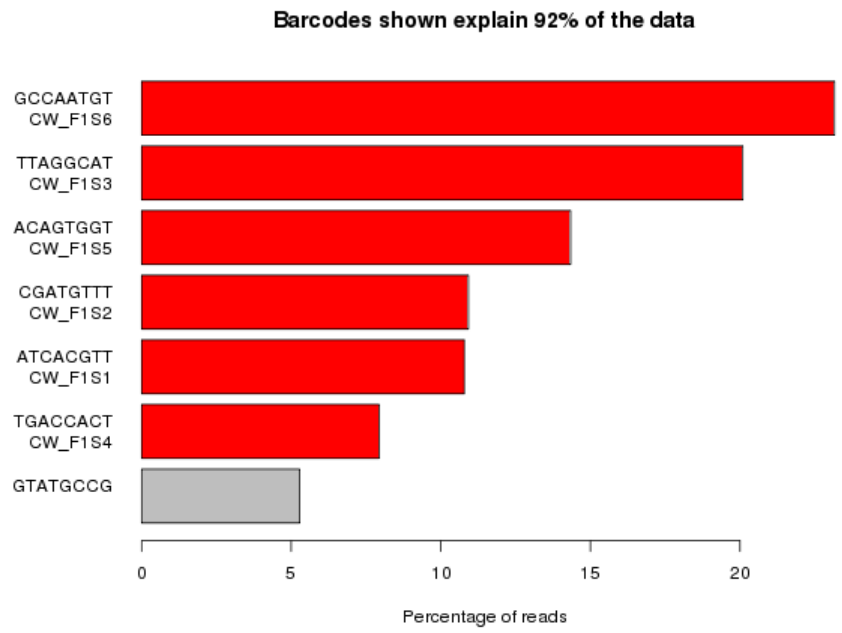
Adapter Dimers



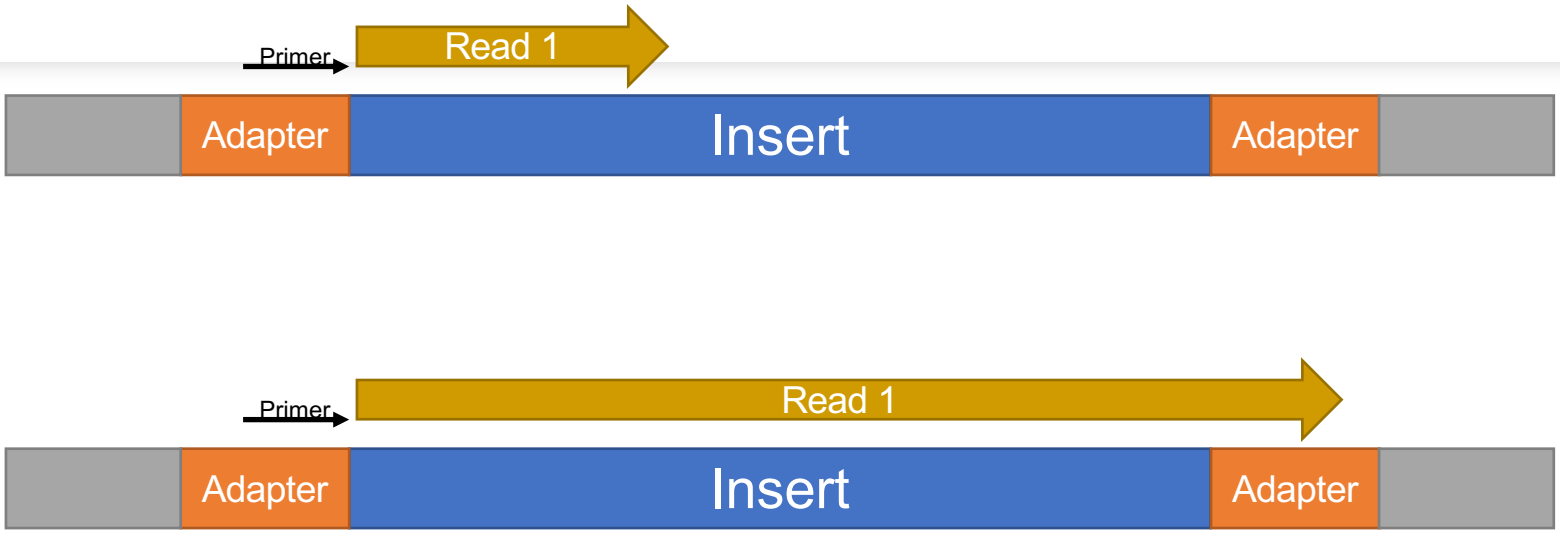
Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCTTGTAATCTCGTATGC	17957	0.14359551756800035	TruSeq Adapter, Index 12 (100% over 50bp)

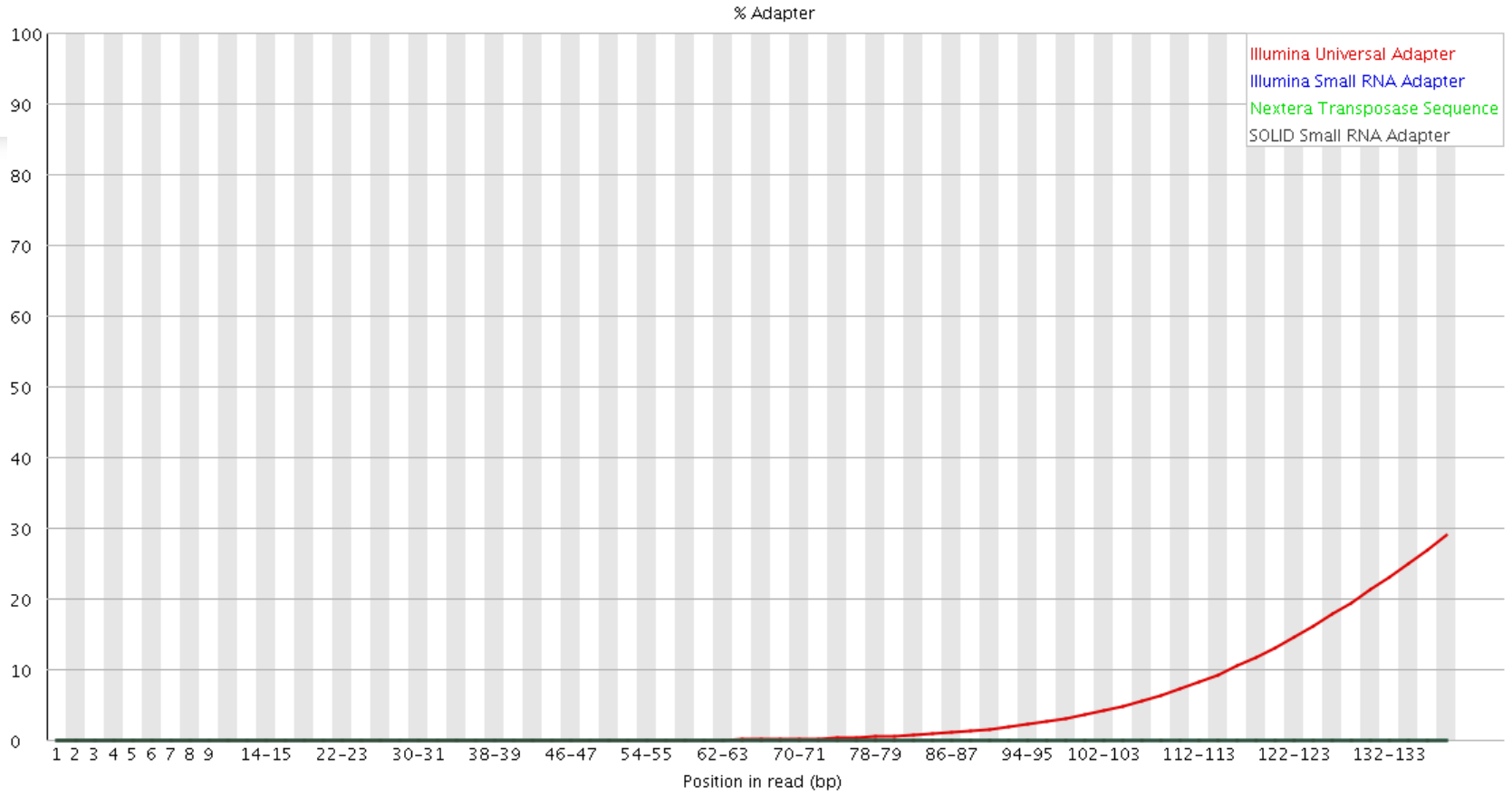
Barcode Sequences



Read-through Adapters



Measuring Read-though Adapters



Let’s run FASTQC

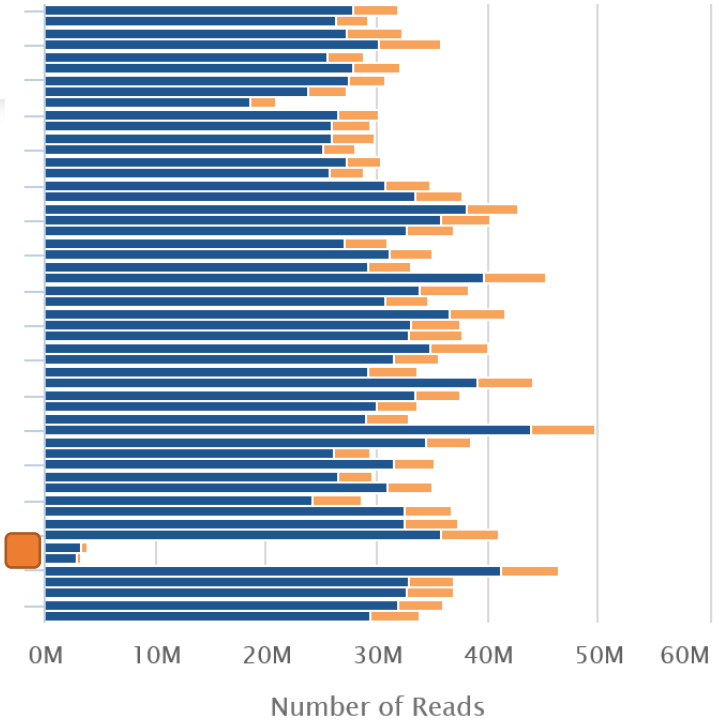
Aggregated Statistics

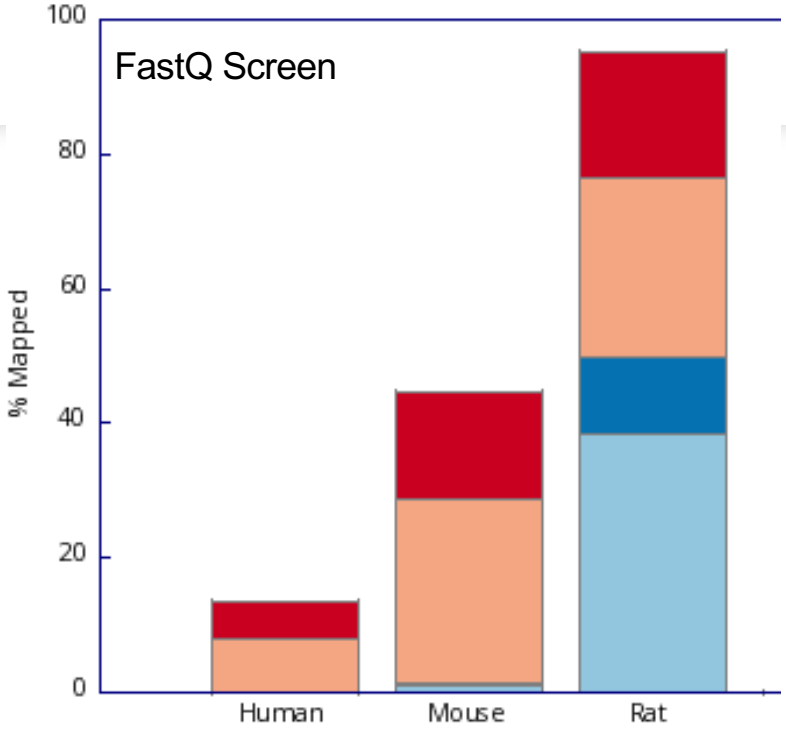
Individual QC reports are useful but can be difficult to interpret witho

The simplest way to spot a local QC problem is that one sample do
the rest

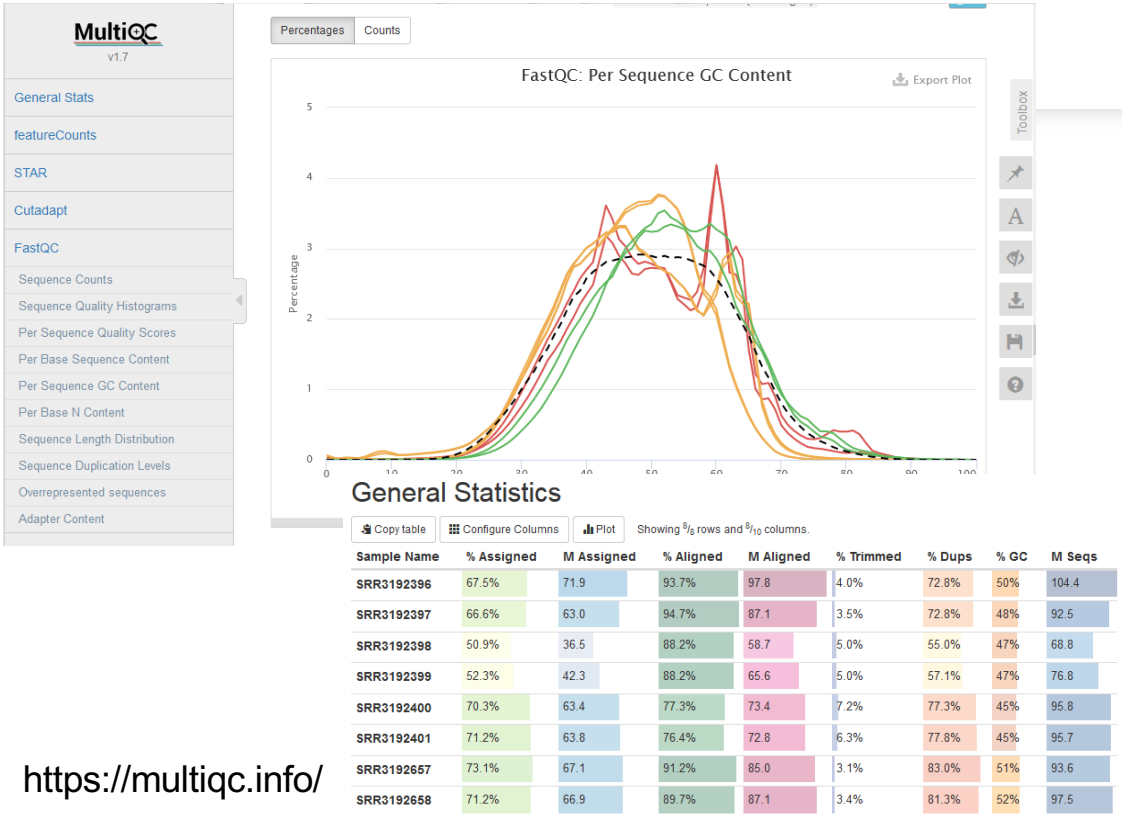
We can aggregate and plot a range of QC statistics to make this ea

Aggregated Mapping Stats





MultiQC



https://multiqc.info/

Aggregates QC information from multiple samples

Large number of programs supported

Combined HTML report