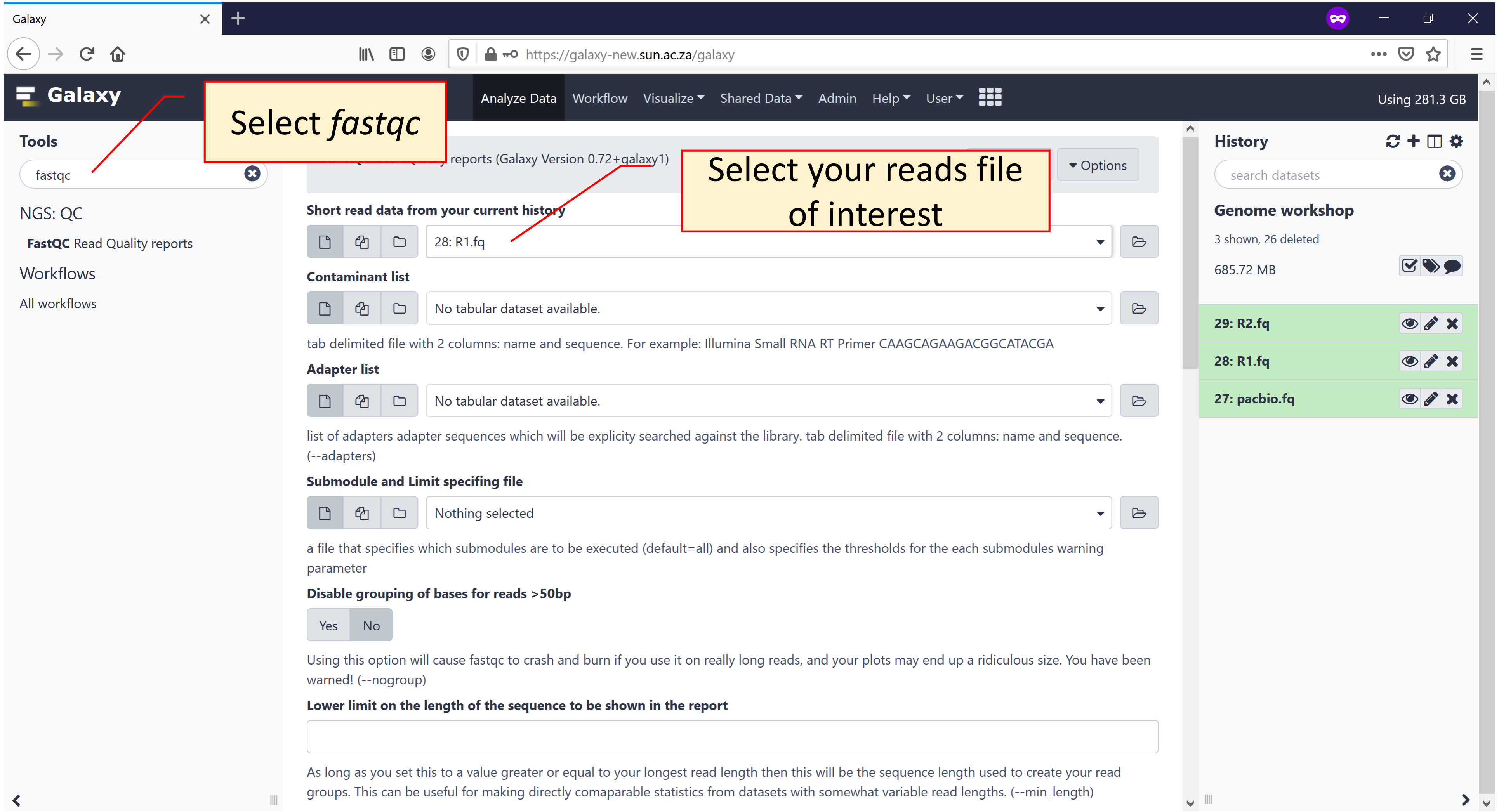
# FastQC introduction

- FastQC tutorial: https://www.youtube.com/watch?v=bz93ReOv87Y
- Help on each function in FastQC:
- http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/

# What is the quality of my reads?

**Read length**
Will be important in setting maximum k-mer size value for assembly
**Quality encoding type**
Important for quality trimming software
**% GC**
High GC organisms don’t tend to assemble well and may have an uneven read coverage distribution.
**Total number of reads**
Gives you an idea of coverage..
**Dips in quality near the beginning, middle or end of the reads**
Determines possible trimming/cleanup methods and parameters and may indicate technical problems with the sequencing process/machine run.
**Presence of highly recurring k-mers**
May point to contamination of reads with barcodes, adapter sequences etc.
**Presence of large numbers of N’s in reads**
May point to poor quality sequencing run. You need to trim these reads to remove N’s.



Select *fastqc*

Select your reads file of interest

reports (Galaxy Version 0.72+galaxy1)

**Short read data from your current history**

28: R1.fq

**Contaminant list**

No tabular dataset available.

tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAGACGGCATACGA

**Adapter list**

No tabular dataset available.

list of adapters adapter sequences which will be explicity searched against the library. tab delimited file with 2 columns: name and sequence. (--adapters)

**Submodule and Limit specifing file**

Nothing selected

a file that specifies which submodules are to be executed (default=all) and also specifies the thresholds for the each submodules warning parameter

**Disable grouping of bases for reads >50bp**

Yes [✓ No]

Using this option will cause fastqc to crash and burn if you use it on really long reads, and your plots may end up a ridiculous size. You have been warned! (--nogroup)

**Lower limit on the length of the sequence to be shown in the report**

As long as you set this to a value greater or equal to your longest read length then this will be the sequence length used to create your read groups. This can be useful for making directly comaparable statistics from datasets with somewhat variable read lengths. (--min_length)

**Genome workshop**

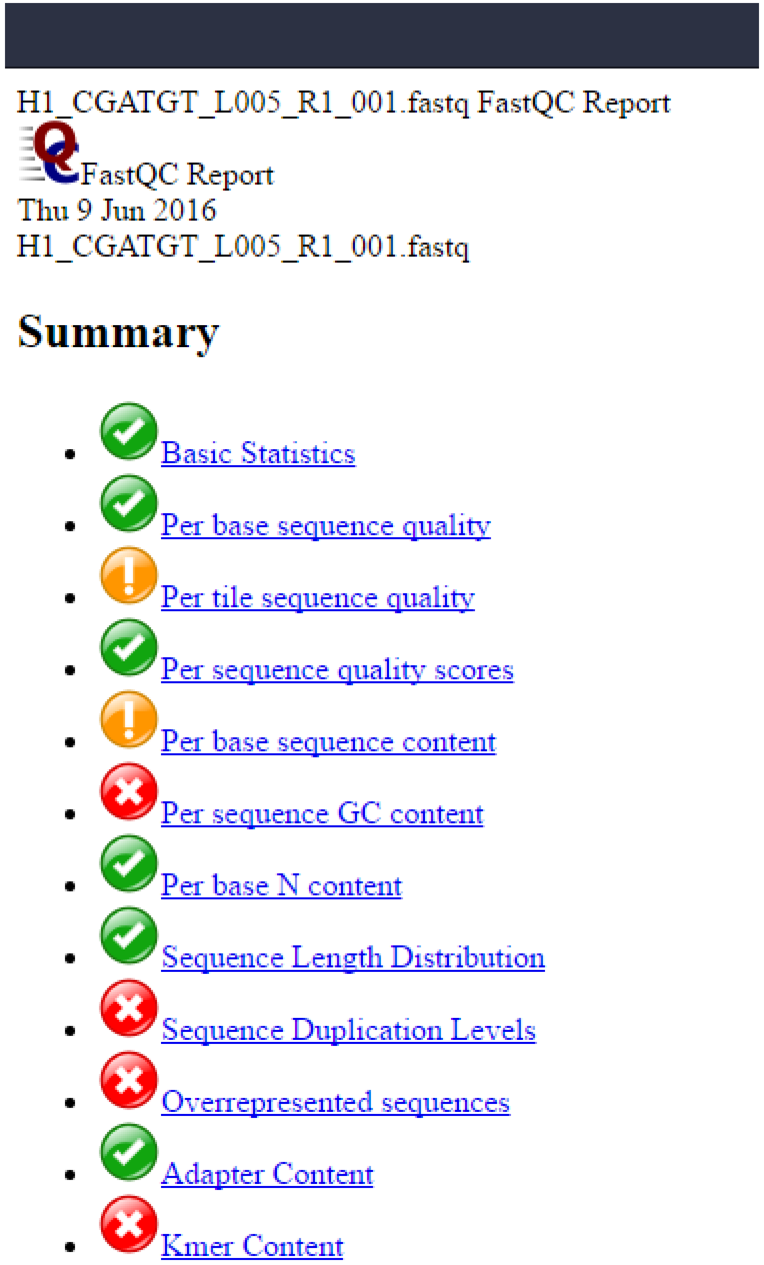
3 shown, 26 deleted

685.72 MB

**29: R2.fq**

**28: R1.fq**

**27: pacbio.fq**



H1_CGATGT_L005_R1_001.fastq FastQC Report

FastQC Report

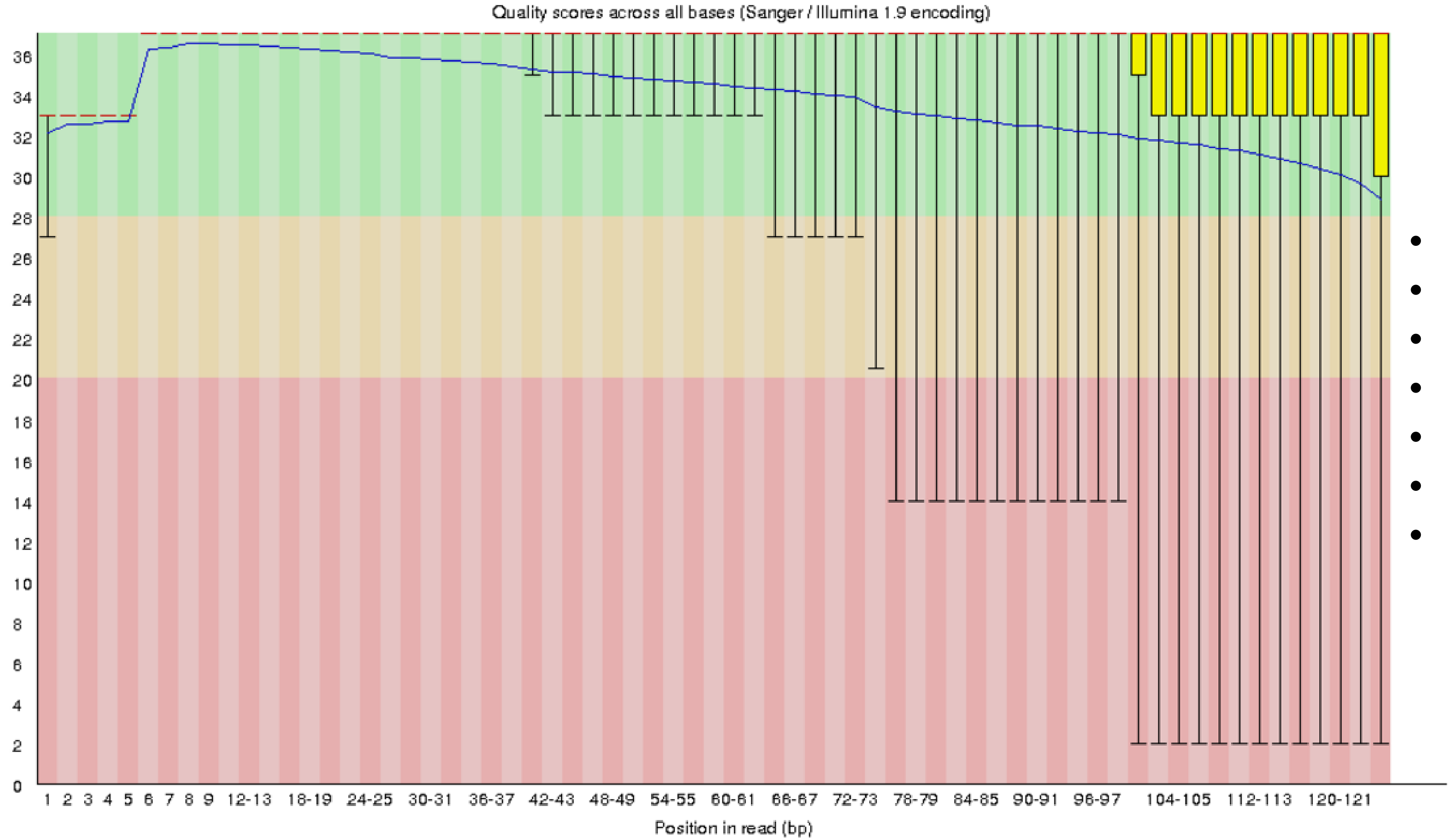
Thu 9 Jun 2016

H1_CGATGT_L005_R1_001.fastq

## Summary

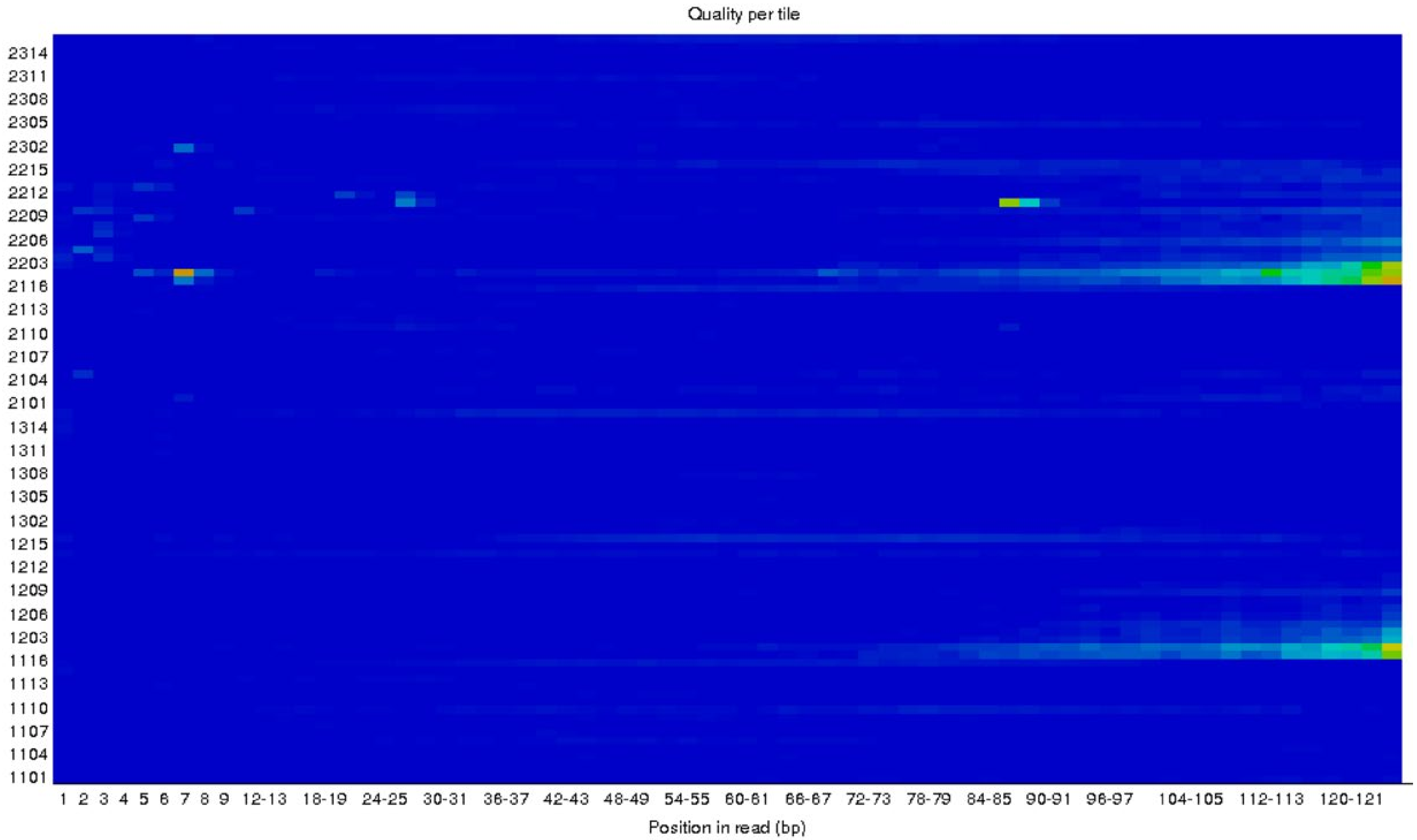
- ✓ Basic Statistics
- ✓ Per base sequence quality
- ! Per tile sequence quality
- ✓ Per sequence quality scores
- ! Per base sequence content
- ✗ Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✗ Sequence Duplication Levels
- ✗ Overrepresented sequences
- ✓ Adapter Content
- ✗ Kmer Content

# Per base sequence quality



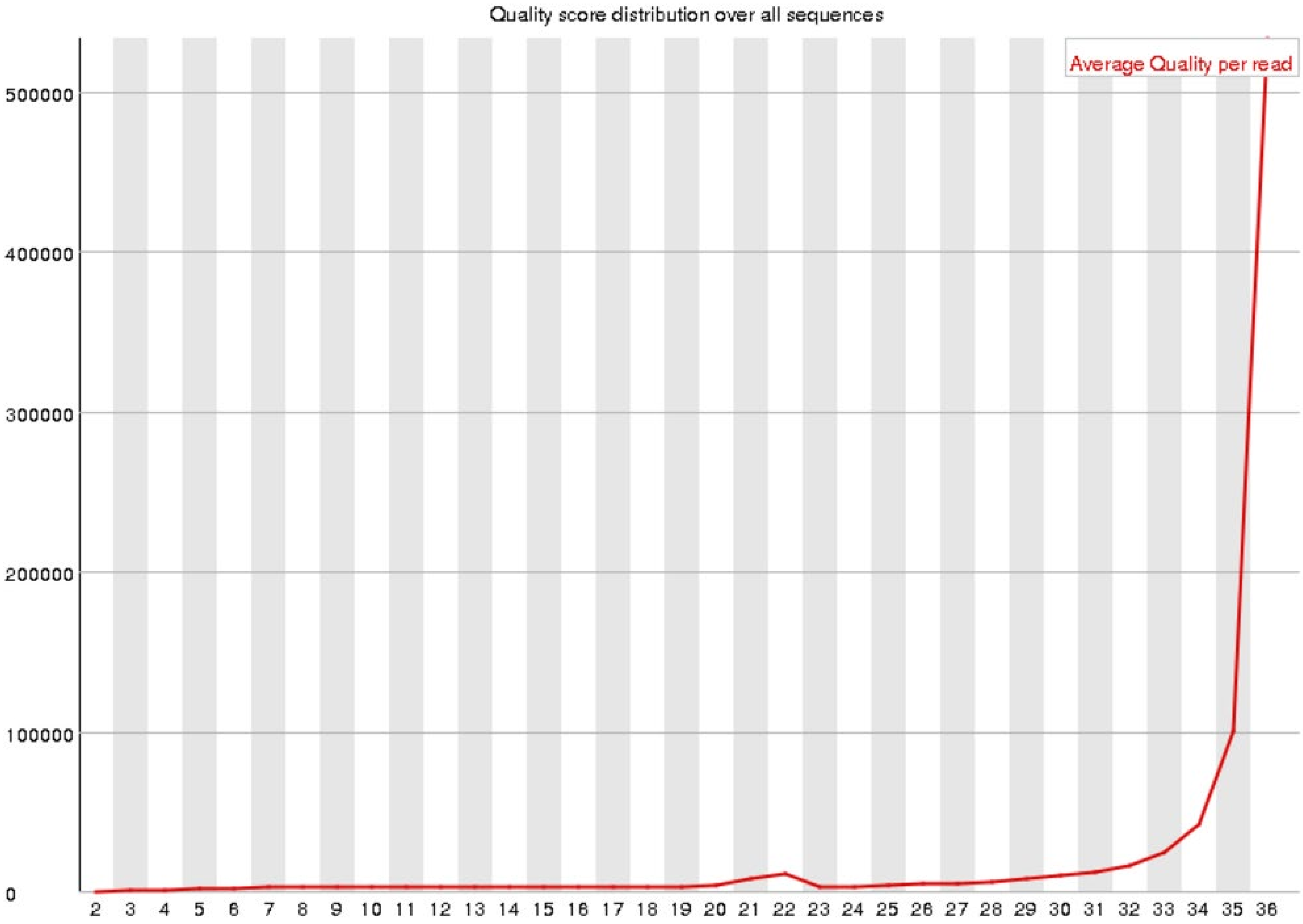
- X-axis: base position
- Y-axis: quality score (Q)
- Yellow box: 25-75th percentile
- Whiskers: 10-90th percentile
- Red line: median
- Blue line: mean
- Ideally, Q > 30

# Per tile sequence quality



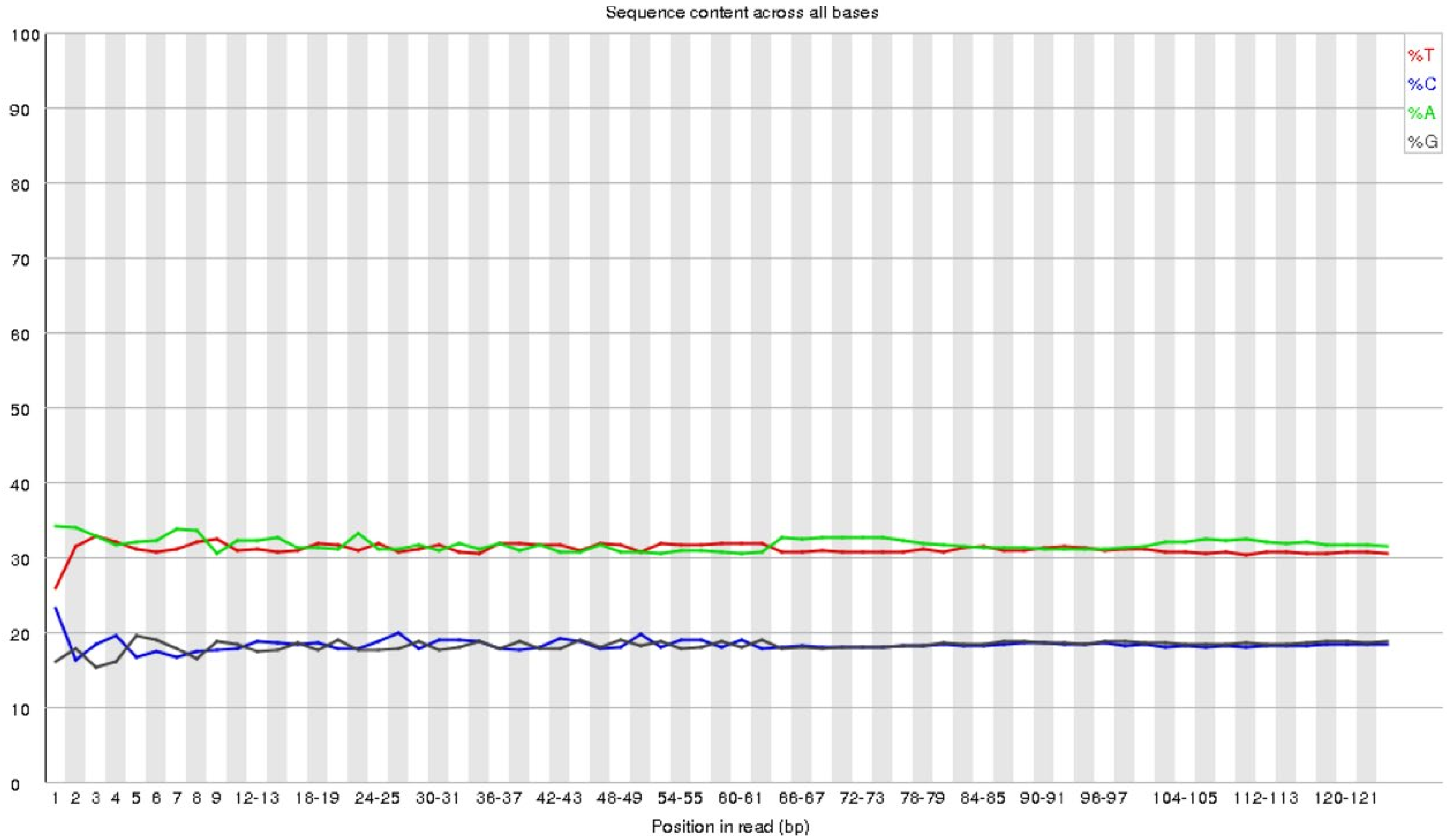
- Heatmap of read quality at a given position in sequence for each tile in a Illumina flowcell
- Colours displayed from cold (blue) to hot (red)
- Warning issued if any tile has mean Phred score 2 less than flowcell average for that position
- Error issued if any tile has Phred score 5 less than flowcell average at that position

# Per sequence quality scores



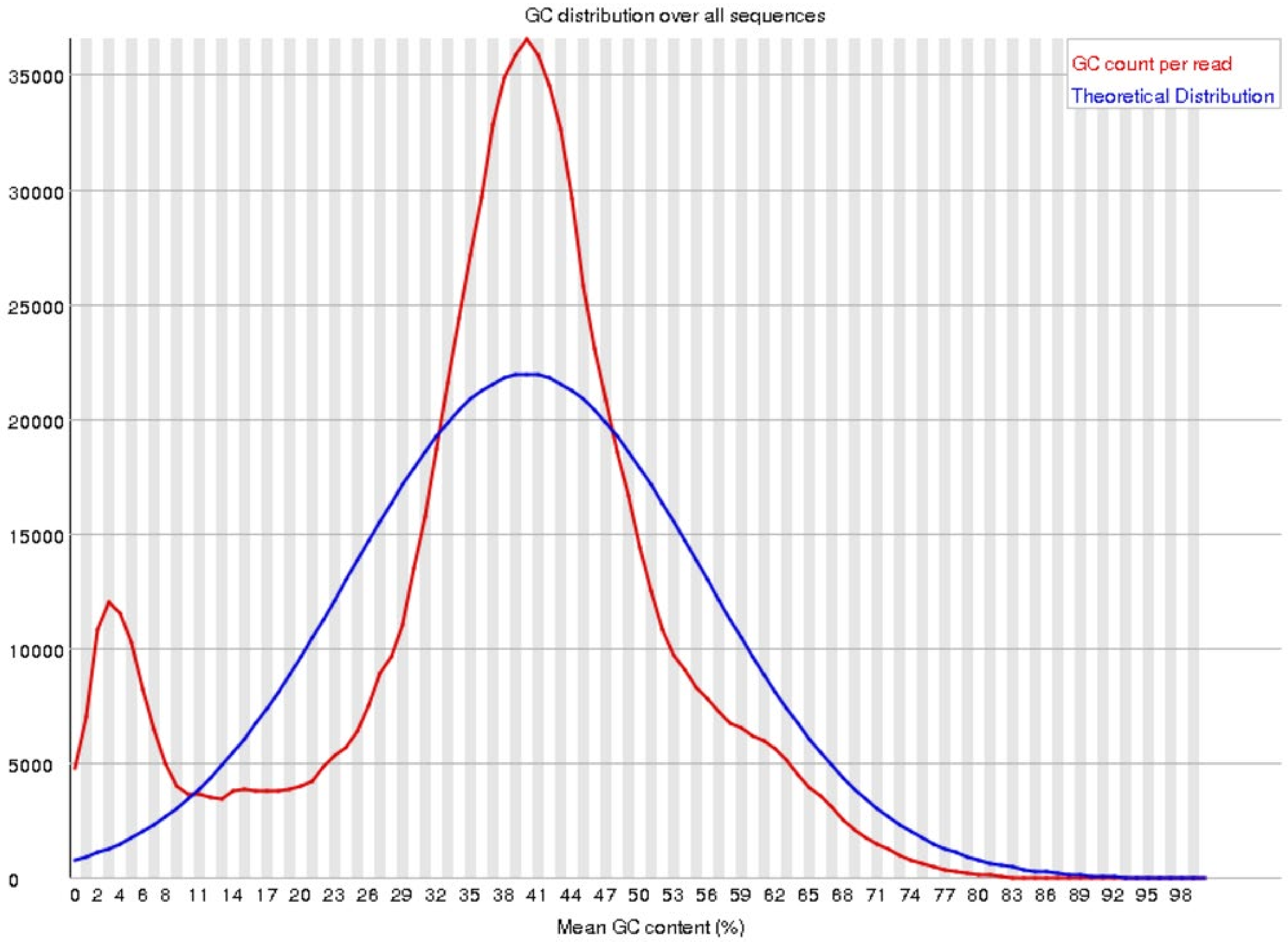
- Plot of <u>average Q</u> value for a sequence against number of sequences with the same average Q
- Distribution should have pronounced peak at the right, with few if any small bumps in the central/low Q positions

# Per base sequence content



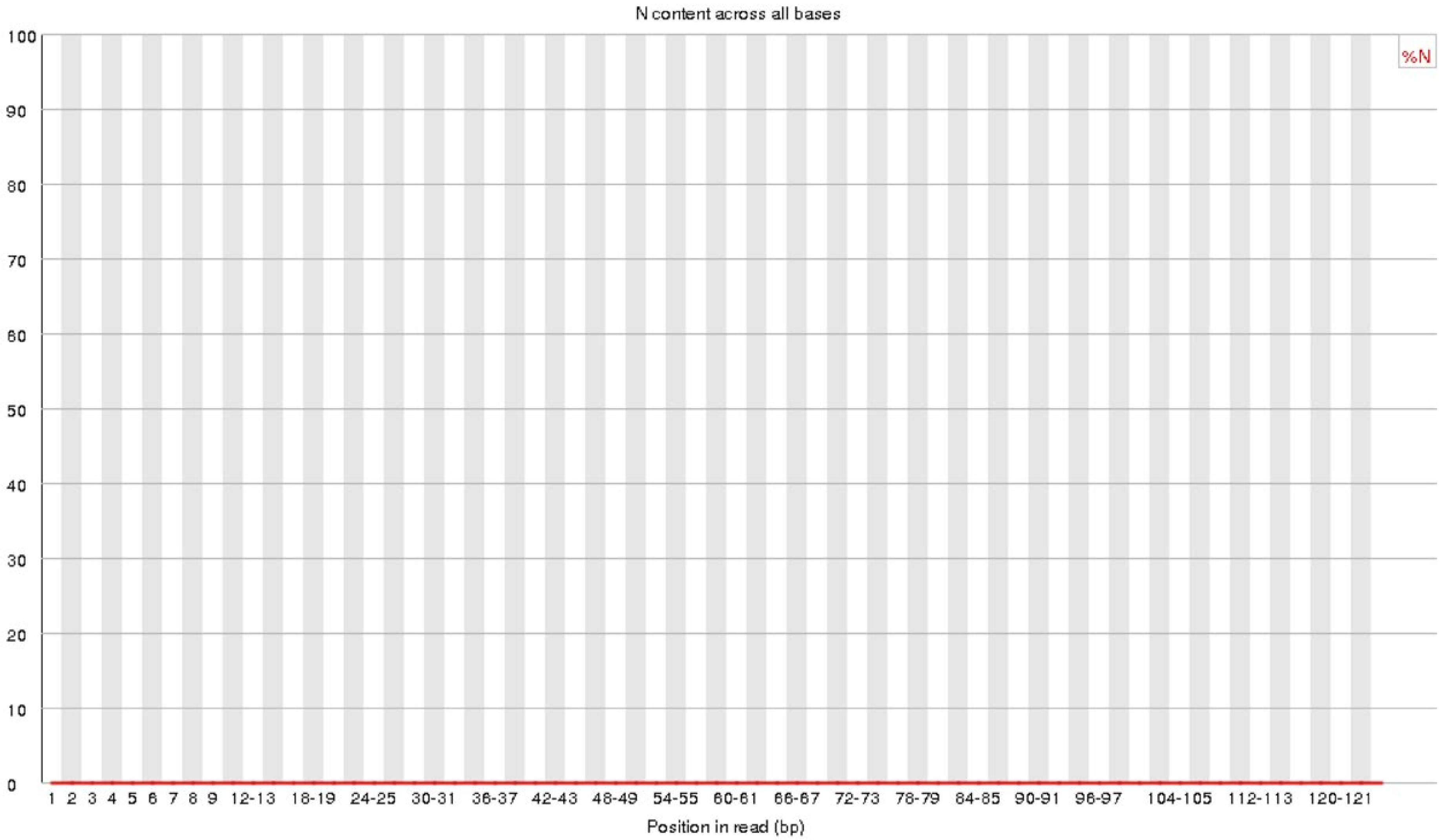
- Plot of % content for each of the 4 nucleotides G, A, T and C
- The % composition for each should be fairly constant for the sequence length
- If there is a significant deviation at one or more positions, it may indicate a significant library bias, or problem with the synthesis reaction

# Per sequence GC content



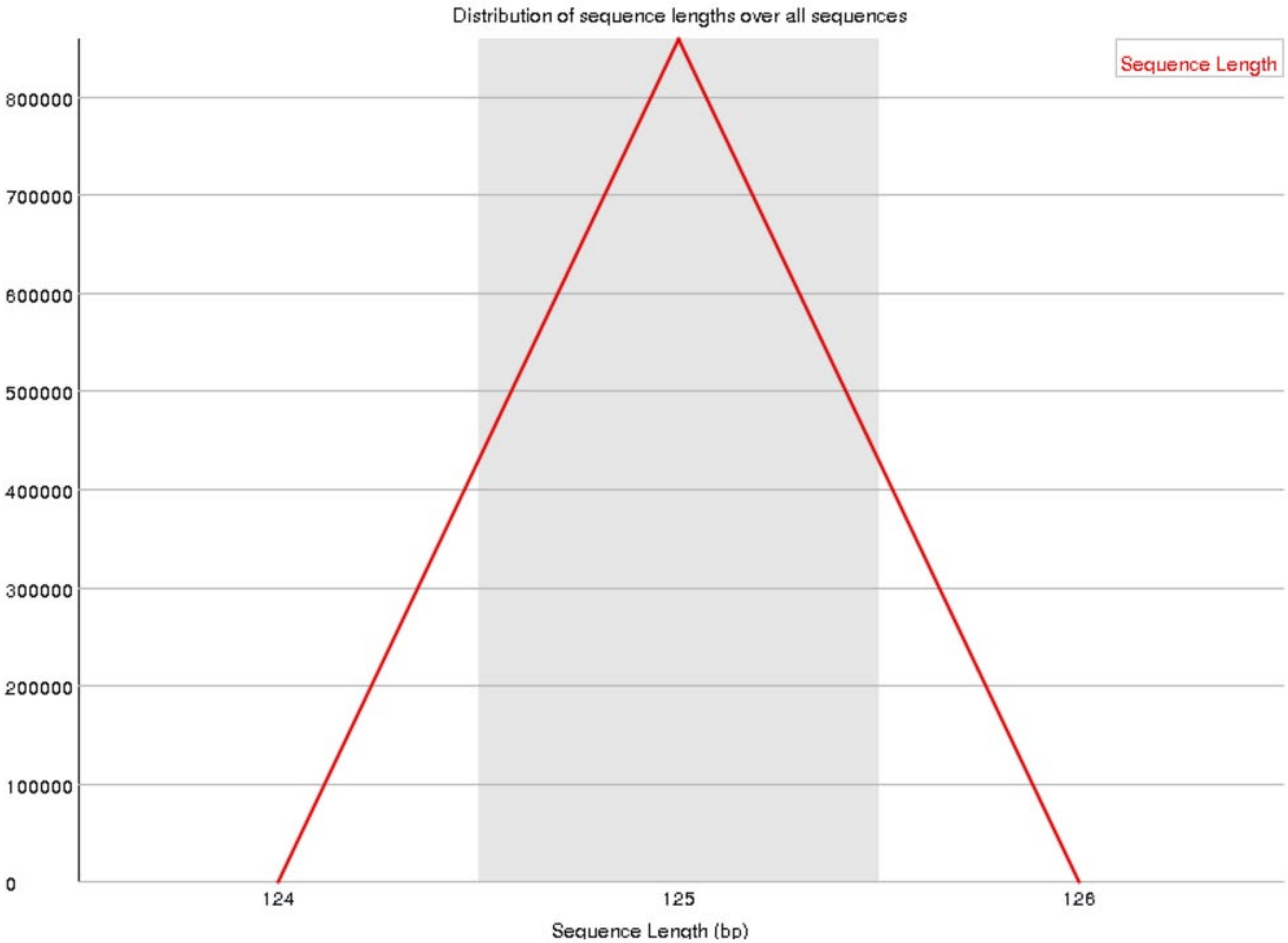
- Plot of the average GC% against number of sequences
- A theoretical, normal distribution is calculated, based on the observed GC% of all the sequences
- The practical distribution and normal distribution should be very similar
- The appearance of “bumps” to the side of the distribution may indicate adapter dimers or another library bias
- A warning is given if the sum of the deviations represent >15% of the reads
- A failure is given if the sum of the deviations represent >30% of the reads

# Per base N content



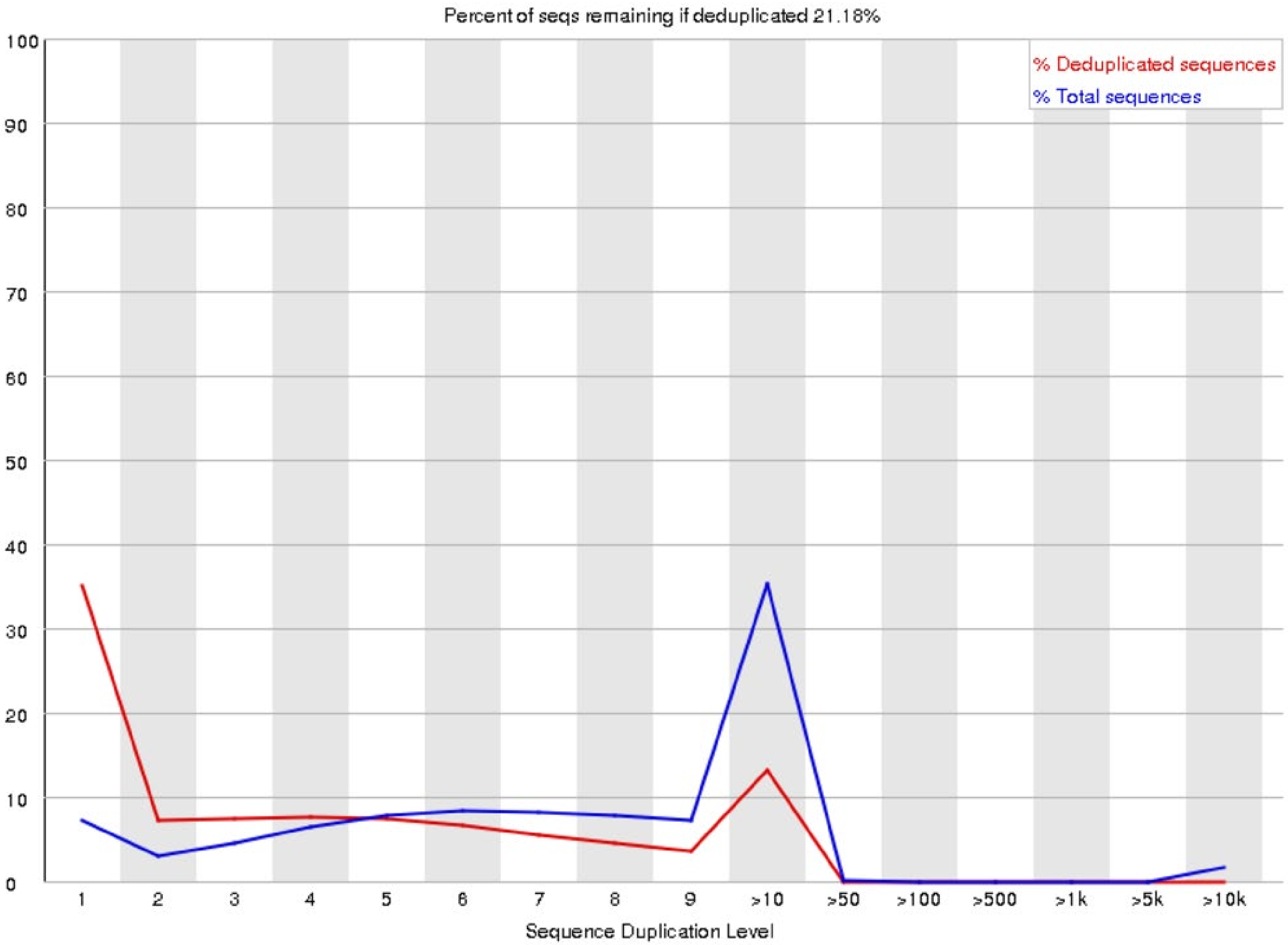
- % of uncalled bases (N) at every position in the sequence
- Indication of sequence quality and base-calling specificity
- Warning issued if N% > 5 at any position
- Failure if N% > 20 at any position

# Sequence Length Distribution



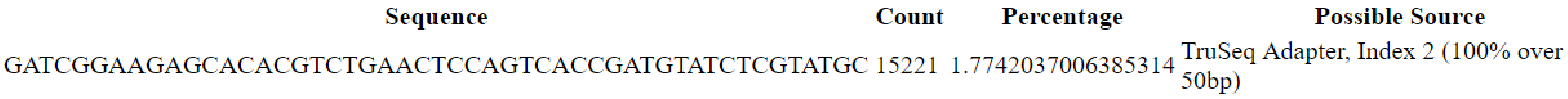
- Distribution of sequence lengths in whole dataset
- For Illumina, all sequences should be the same length
- After trimming (see `trim_galore`, later), this distribution may change

# Sequence Duplication Levels



- This plot shown a bin distribution, with an the number of sequences in each bin (blue line)
- Over-representation of sequences may to due to high sequence coverage, or contamination with low-complexity sequences
- Only the first 50 nt of the first 100,000 sequences in a file are analyzed
- The red line shown the “de-duplicated” sequences – each duplicated sequence is counted only once in each bin
- The percentage of sequence remaining after “de-duplication” is given
- Warning: non-unique sequences make up >20% of total
- Error: non-unique sequences make up >50% of total

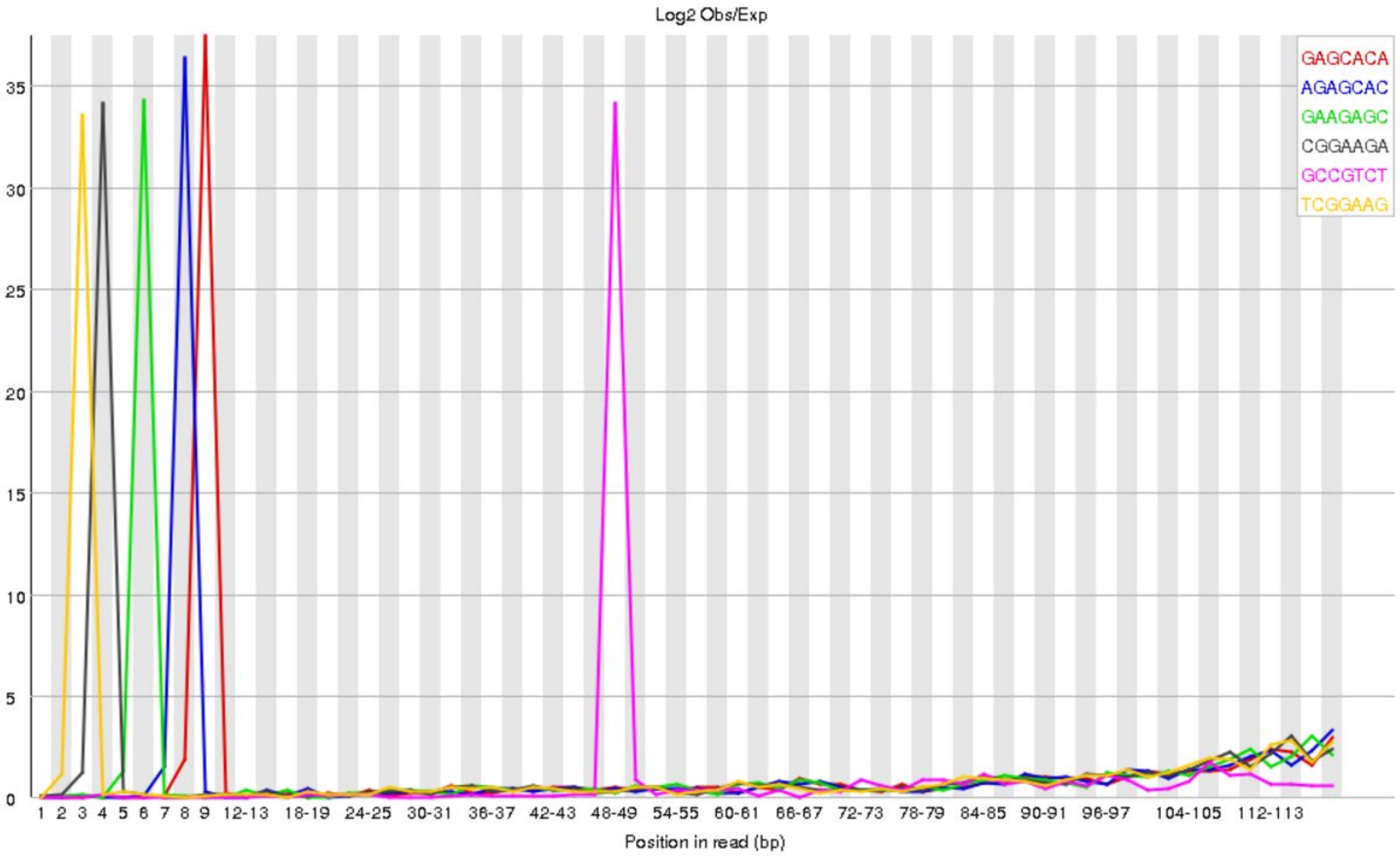
# Overrepresented sequences



| Sequence | Count | Percentage | Possible Source |
|---|---|---|---|
| GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGC | 15221 | 1.7742037006385314 | TruSeq Adapter, Index 2 (100% over 50bp) |

- The first 100,000 sequences in a file are scanned against the whole file
- A fit over 20 nt with at most one mismatch is a hit
- Hits are screened against a database of common contaminants, including adapter and primer sequences common for the given sequencing platform
- This identified contaminating sequences in the dataset can be removed by filtering the dataset
- Warning: Any one sequence representing >0.1% of dataset
- Error: Any one sequence representing >1% of dataset

# Kmer Content



- The occurrence of each possible k-mer (7-mer) is determined at every position for 2% of the dataset
- The likelihood (p) of finding a specific k-mer at each position is calculated with a binomial distribution
- Top 6 over-represented k-mers shown
- Over-represented sub-sequences are not identified in duplicated sequences or per base content analysis
- Over-represented k-mers may be due to amplification of random primer sub-population