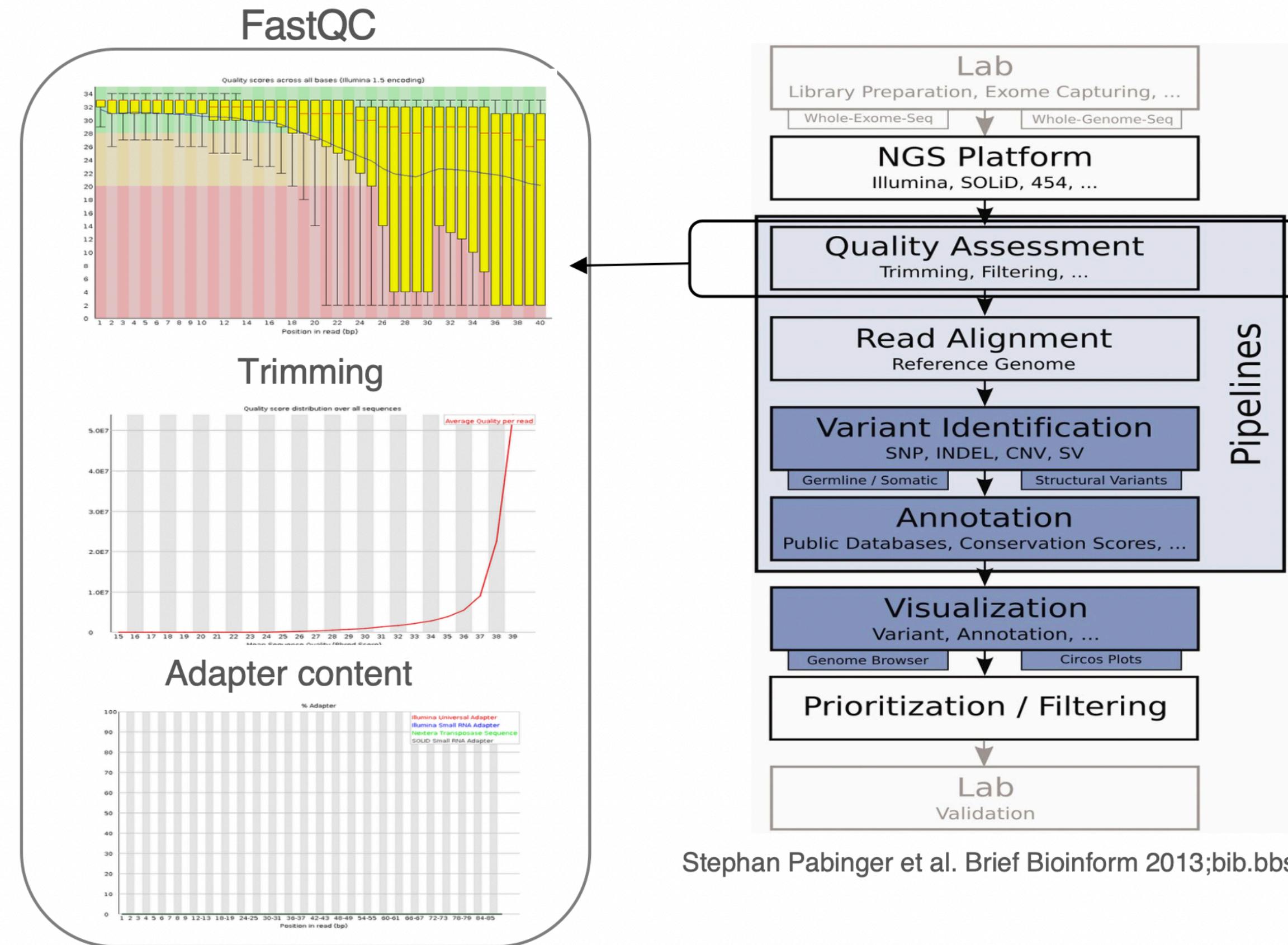


PO21: Precision Oncology Course

Quality Control

Remember



Quality Control (QC) of raw sequencing data is the first step in our bioinformatics workflow. It is done to make sure that data is OK.

Software for QC

FastQC and MultiQC

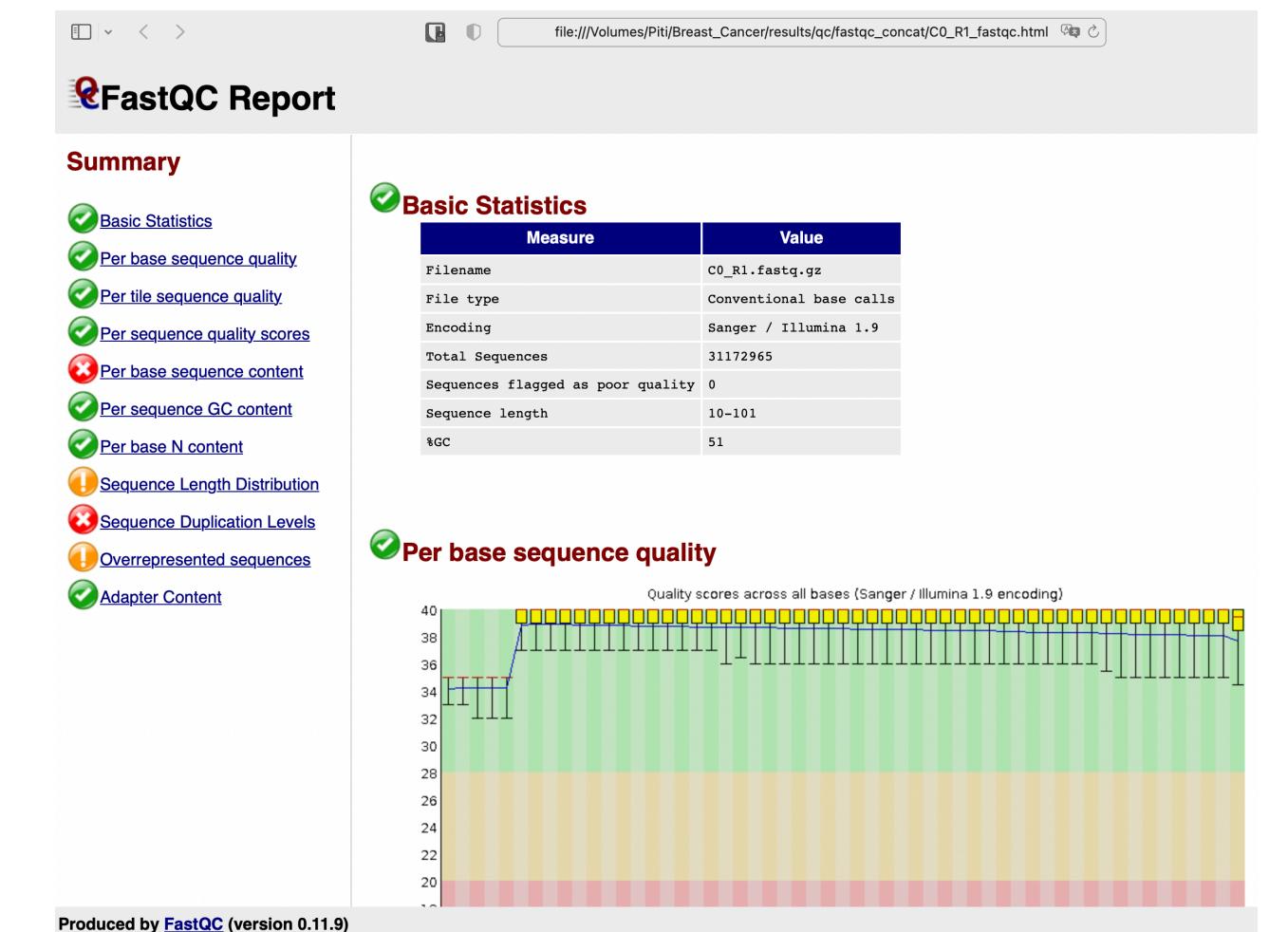
Software: FastQC



The QC is performed by **FastQC**, which takes a FASTQ as input and returns an HTML **web-like report with plots** as output.

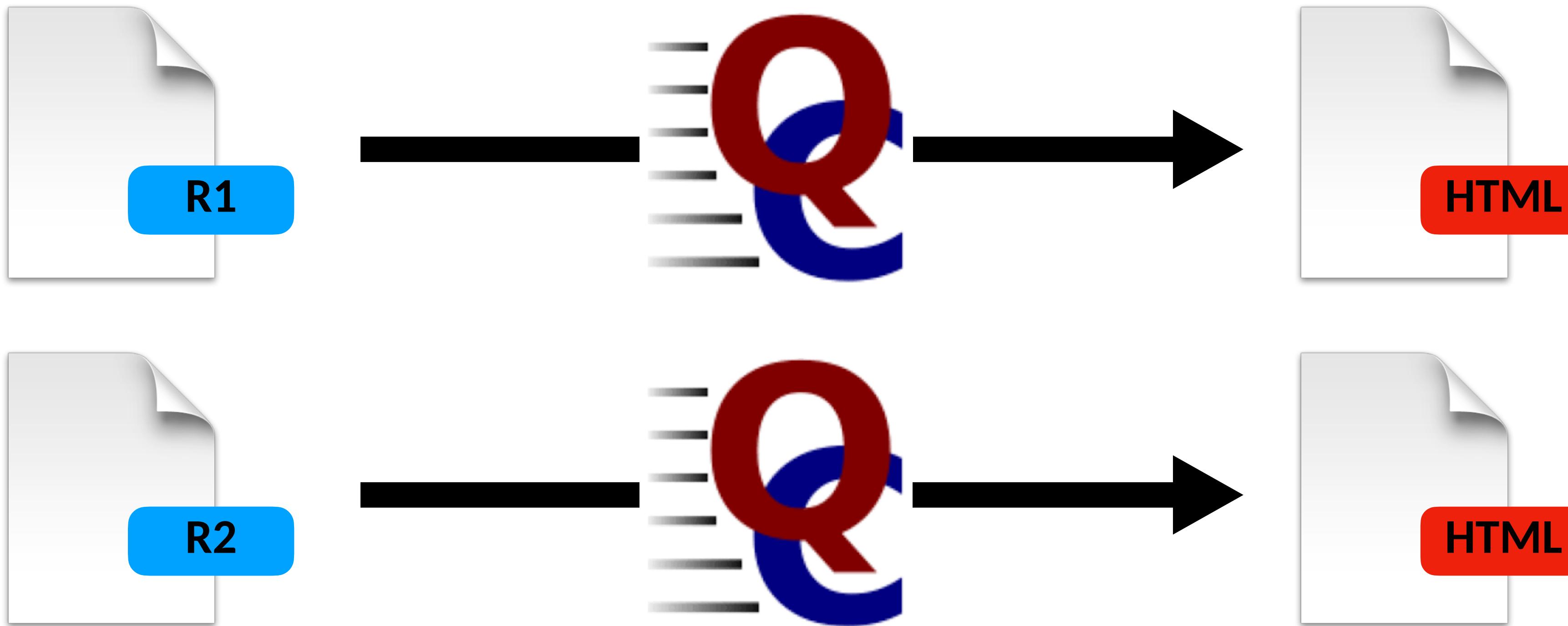
Software: FastQC

```
@SRR6671104.1 SN7001303:484:hwkwybcxx:1:1103:1482:2094 length=101
NCCAAGCGTTCATAGCGACGTCGCTTTGATCCTCGATGTCGGCTTCCATCATTGTGAAGCAG
AATTACCAAGCGTTGATTGTTACCCACTAA
+SRR6671104.1 SN7001303:484:hwkwybcxx:1:1103:1482:2094 length=101
#<DDDIHHIIIIIGHHTHIIIIIIIIIIIIIIIIIIIIIIIIIFHIIIIIIIGHHTHIIII
IIIIIIIIIIIIIGIIIIIIIIIIIIHIIIIIE
@SRR6671104.2 SN7001303:484:hwkwybcxx:1:1103:1407:2185 length=101
CACAAAACCGTGAAGAAGGCAGGCCGGTCATCATAGAAAAGTACTACACGCGCTGGCAACGACT
TCCACACGAACAAGCGCGTGTGCGAGGGAGATCG
+SRR6671104.2 SN7001303:484:hwkwybcxx:1:1103:1407:2185 length=101
DDDDDIHHIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
IIIHIIIIIIIIIIIIIIIIIIIIICHHIGIG
@SRR6671104.3 SN7001303:484:hwkwybcxx:1:1103:1363:2198 length=101
AAAGGTTGGCAGGCTGCTGGGAGCACTGCAGGGTGCAGGTCCCCACGGTGGACGAGCTGGAA
CCGTGGAGGCAGTGGCTGTGGGCATAGTGGCA
+SRR6671104.3 SN7001303:484:hwkwybcxx:1:1103:1363:2198 length=101
DDDDDIHHIIIHIGHIIIECHFIIIIIGEHG@EIIIIDHHHHHHIGGCDCHHHHIIHIGFEHH
HEHIHHGHIIHFHHHEHGGHIIIIHIIHE@
@SRR6671104.4 SN7001303:484:hwkwybcxx:1:1103:1399:2209 length=101
GTCAGTGTAGCGCGCGTTAGCCACCCAGATTGAGCAATAACAGGTCTGTGATGCCCTAGATGTCC
GGGGCTGCACCGCGCTACACTGACTGGCTCAG
+SRR6671104.4 SN7001303:484:hwkwybcxx:1:1103:1399:2209 length=101
DDDDDHIIHHIIHHHDHIHEHIIHHHDCCGHIIIGHIIHHIIHHHEHHEHCHHHIIH
HHHHIIGIH?CEHCHIIIIHIIII?F1FGH
@SRR6671104.5 SN7001303:484:hwkwybcxx:1:1103:1344:2225 length=101
CTGGAGACAGATTGTAGGACCGAGCGCGGGCAGGGAGGCAACGGAGCTACCAGCCGCTCCTCT
GCTATATGAAATATGGGAGACGACAGACCGTT
```



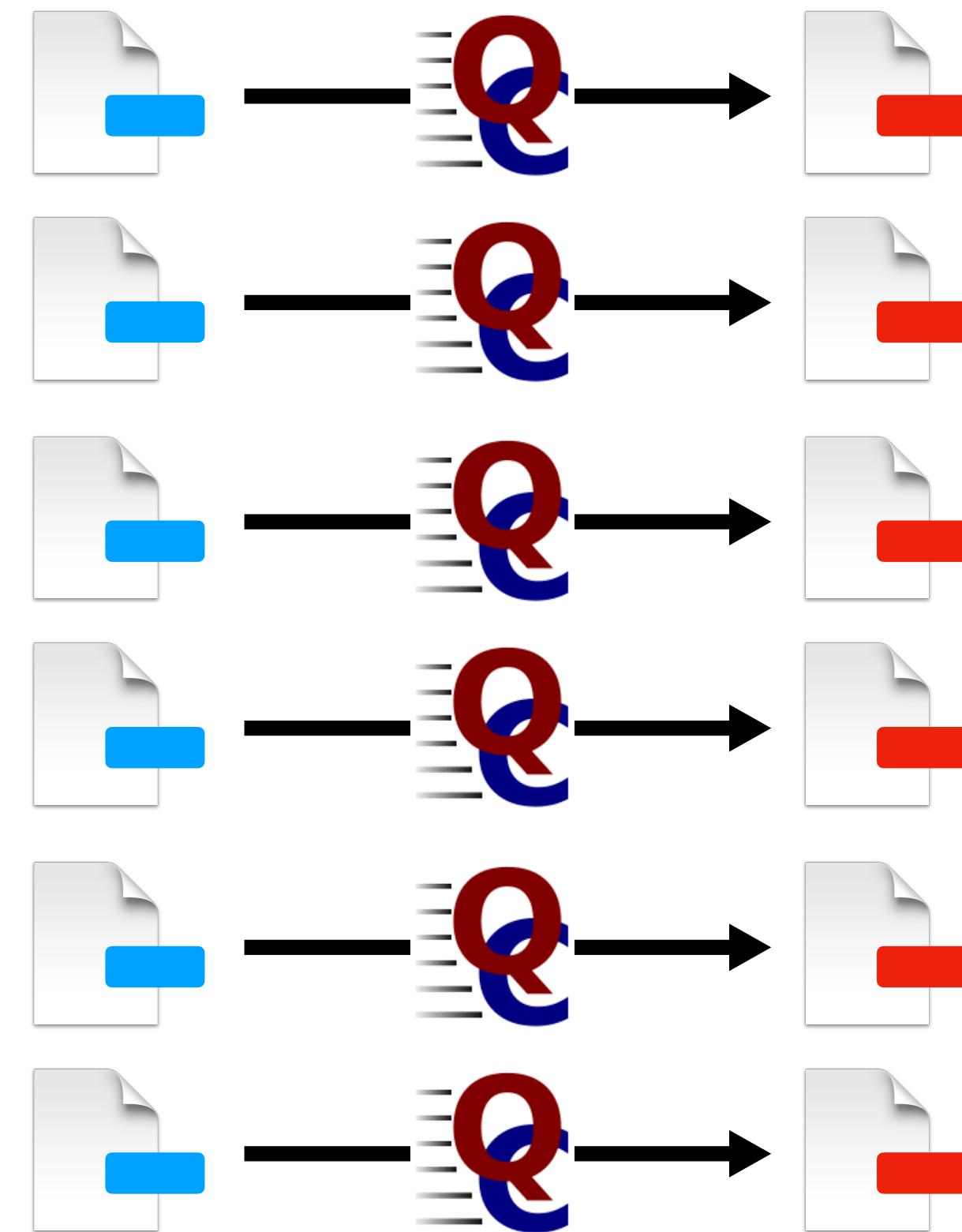
The QC is performed by **FastQC**, which takes a FASTQ as input and returns an HTML web-like report with plots as output.

Software: FastQC



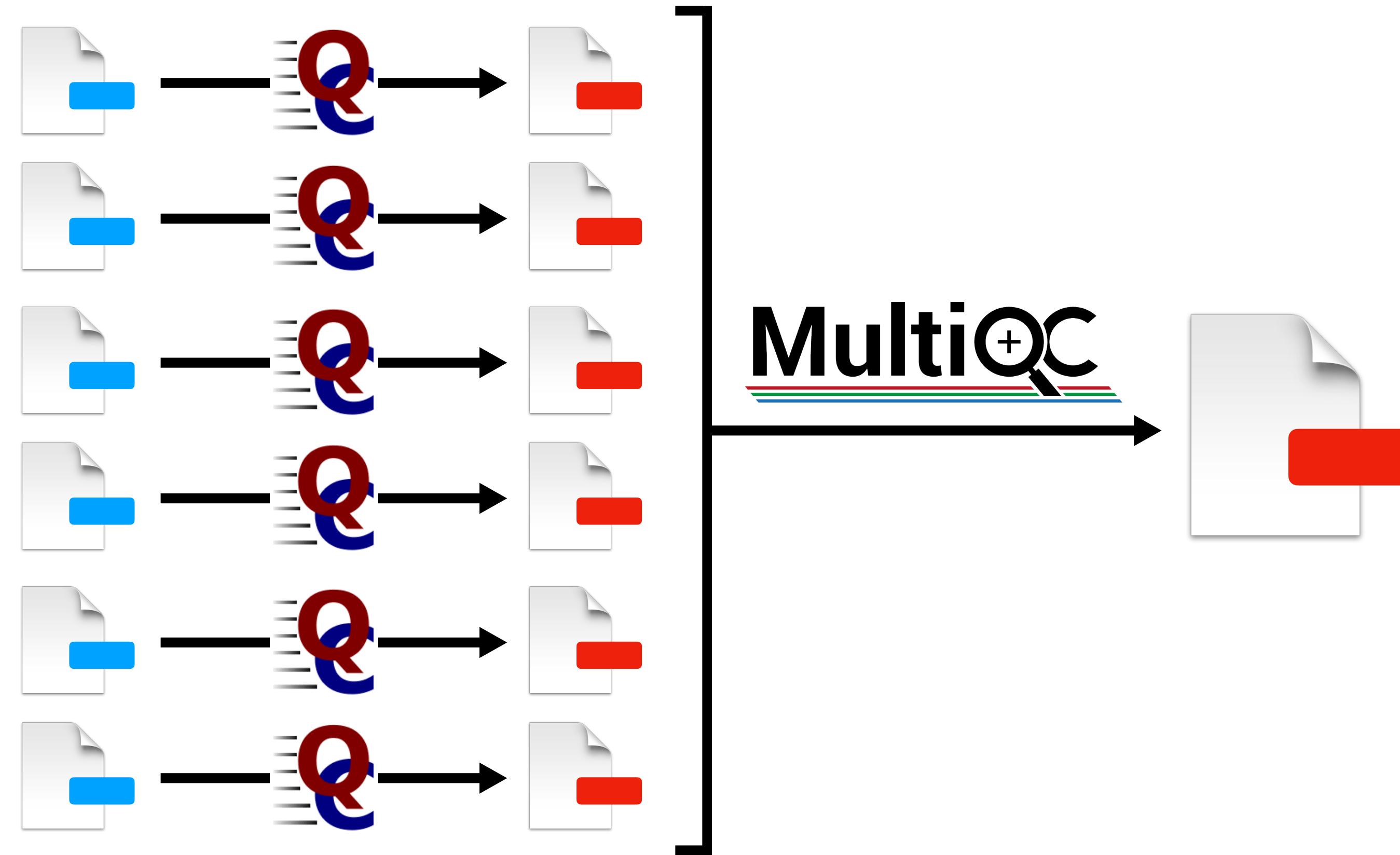
If you have **paired-end sequencing** data (two FASTQs per sample: R1.fastq and R2.fastq) you will get **two reports per sample**.

Software: FastQC



In a normal analysis you will end up with a lot of reports.

Software: MultiQC



MultiQC is a program that aggregates all FastQC reports into a single interactive one.

Report sections in FastQC and MultiQC

Examples of Good and Bad results

FastQC report sections



[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](#)



The results seem normal



The results seem slightly abnormal



The results seem unusual

Report sections



- [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](#)



- [Sequence Counts](#)
- [Sequence Quality Histograms](#)
- [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](#)
- [Status Checks](#)

Report sections



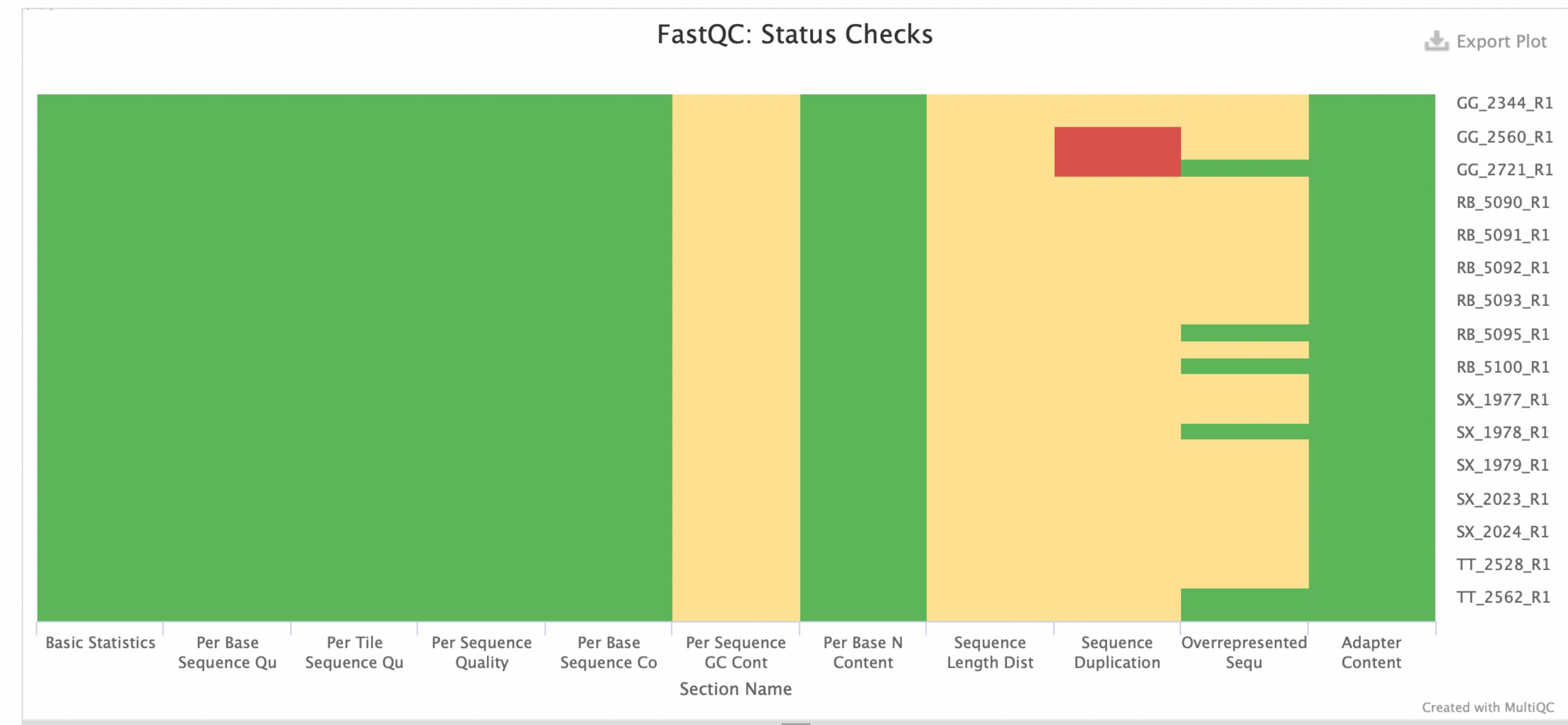
- [Basic Statistics](#)
- [Per base sequence quality](#)
- [Per tile sequence quality](#)
- [Per sequence quality scores](#)
- [Per base sequence content](#)
- [Per sequence GC content](#)
- [Per base N content](#)
- [Sequence Length Distribution](#)
- [Sequence Duplication Levels](#)
- [Overrepresented sequences](#)
- [Adapter Content](#)



Status Checks

Status for each FastQC section showing whether results seem entirely normal (green), slightly abnormal (orange) or very unusual (red).

Sort by highlight



FastQC Basic Statistics



Basic Statistics

Measure	Value
Filename	20210722_2344_AM9009_S75_R1_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	59796259
Sequences flagged as poor quality	0
Sequence length	150
%GC	45

Basic Statistics



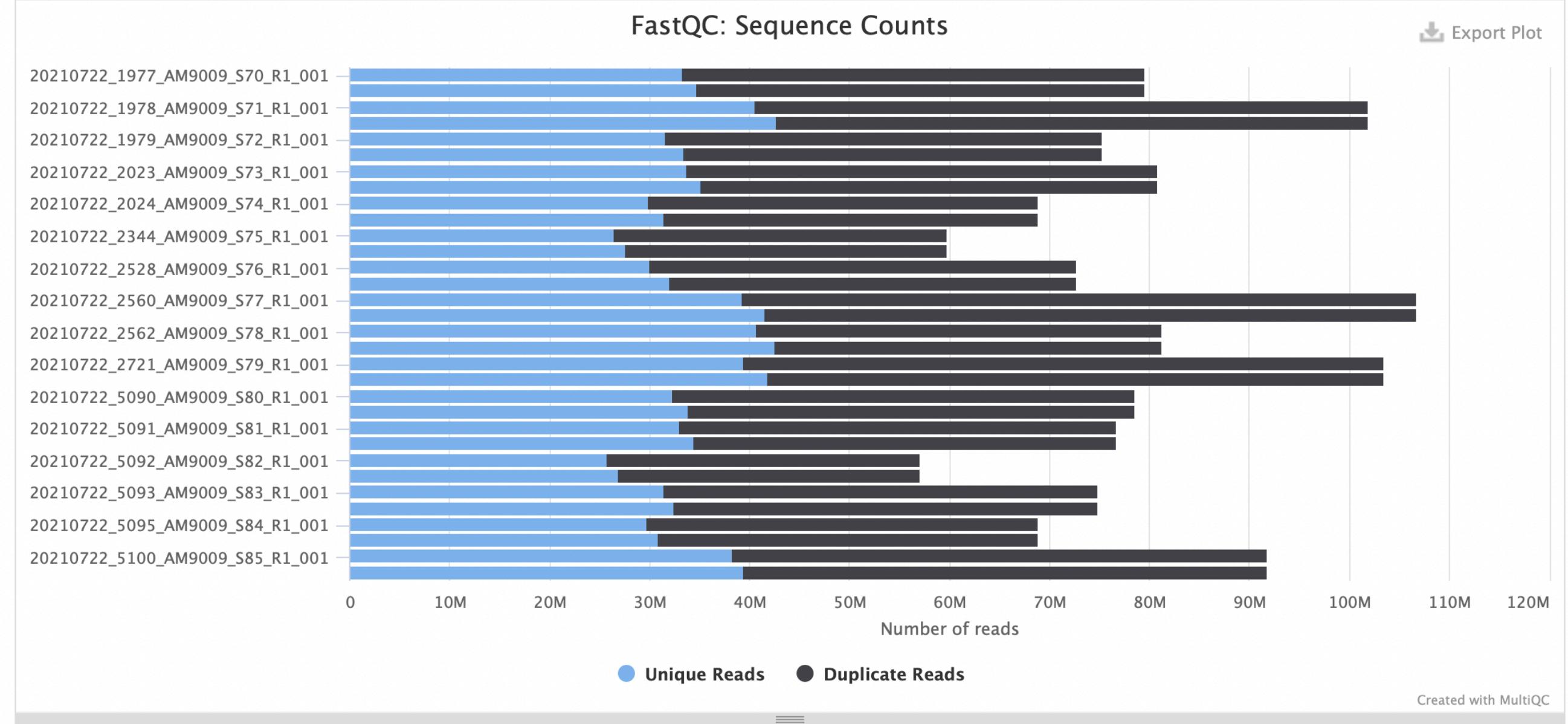
Basic Statistics

Measure	Value
Filename	20210722_2344_AM9009_S75_R1_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	59796259
Sequences flagged as poor quality	0
Sequence length	150
%GC	45

Sequence Counts

Sequence counts for each sample. Duplicate read counts are an estimate only.

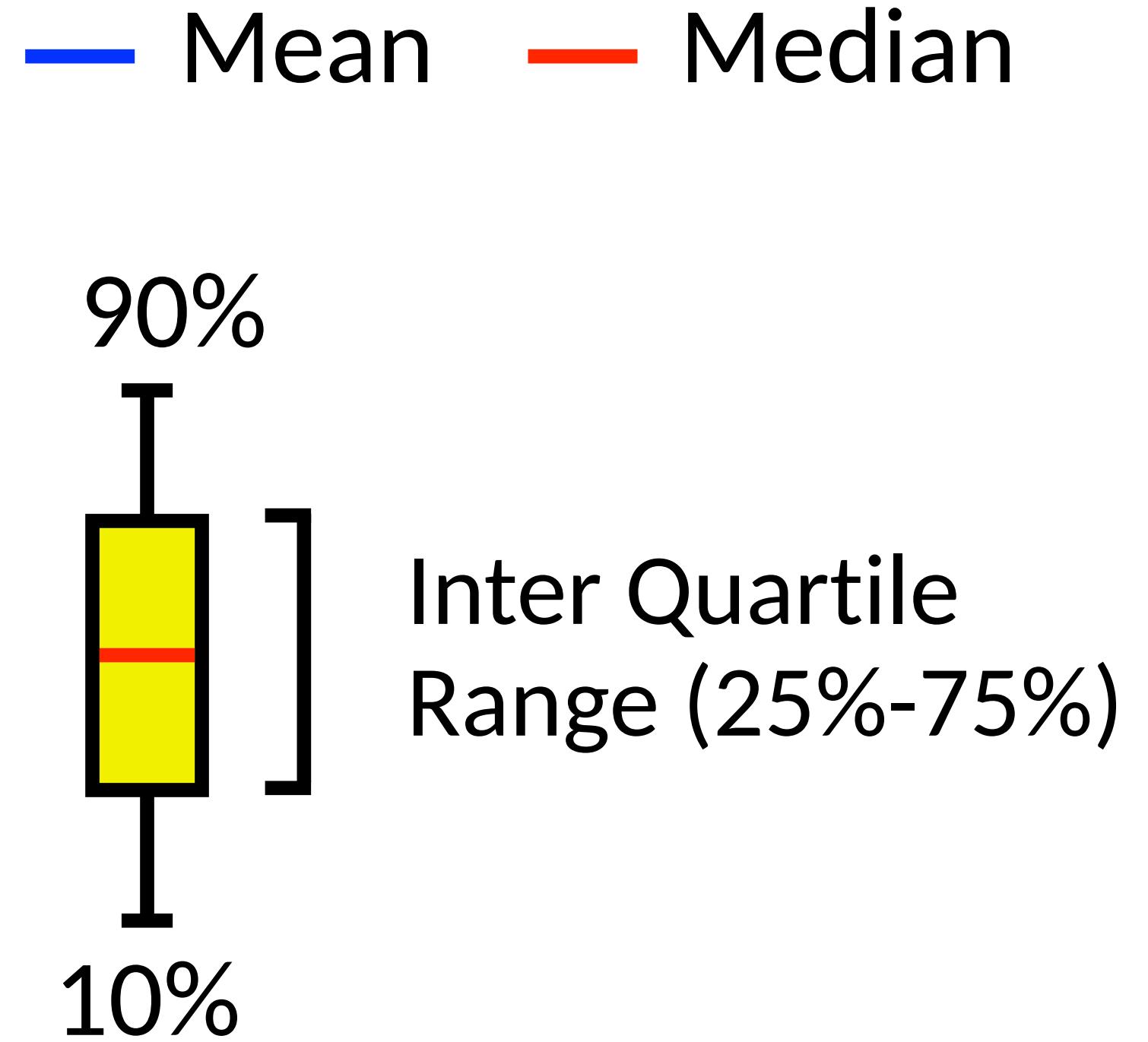
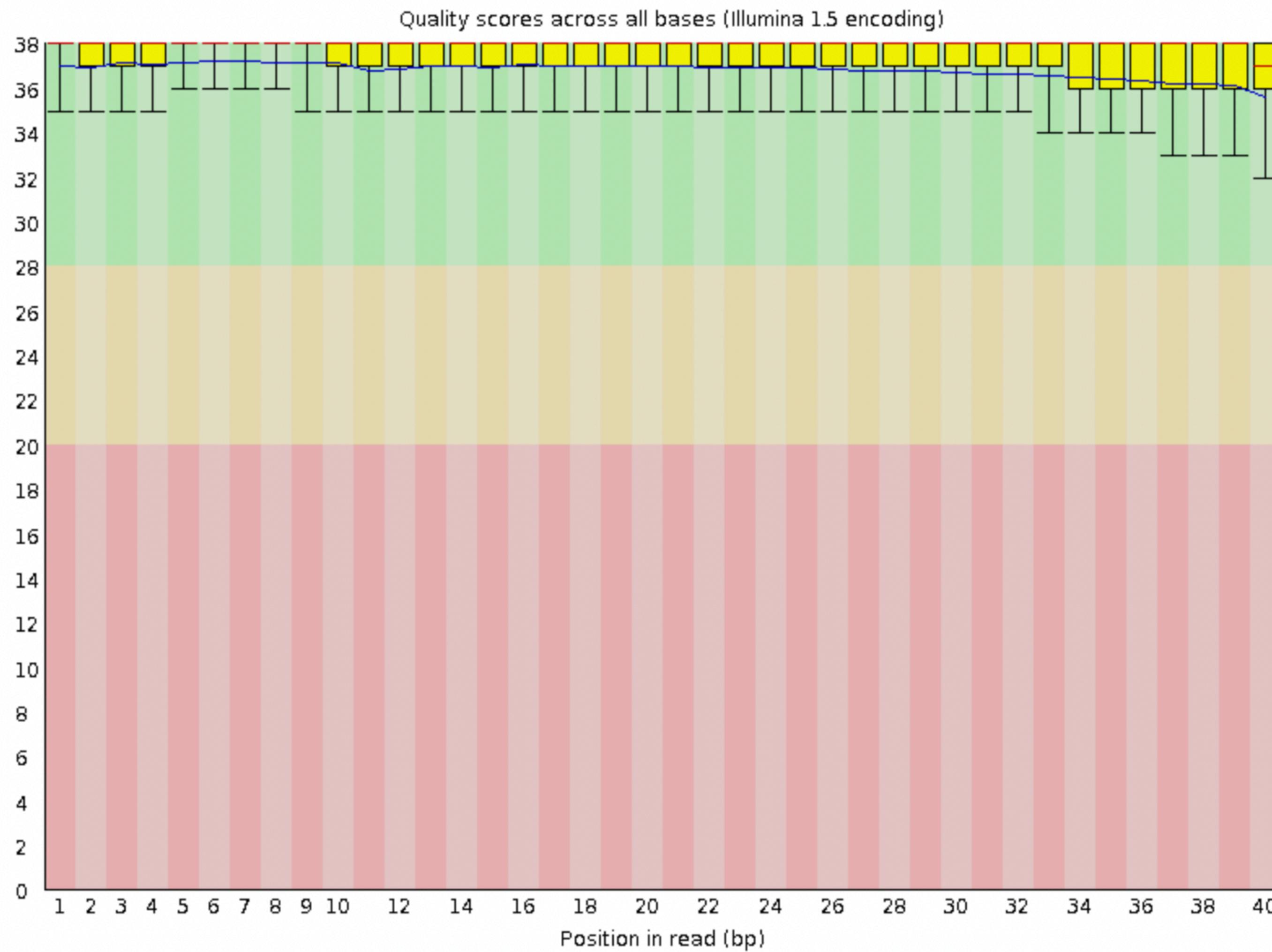
Number of reads Percentages



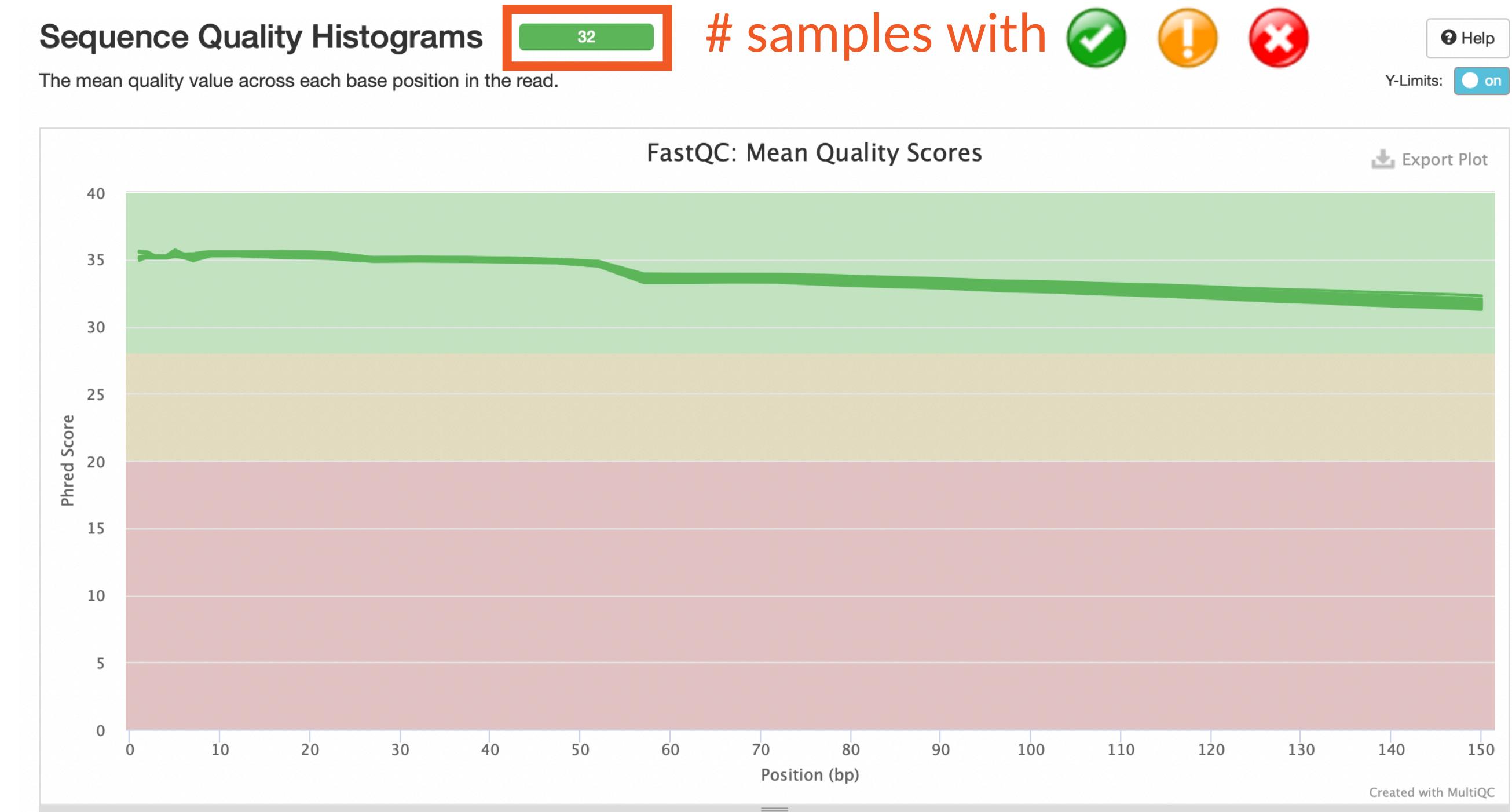
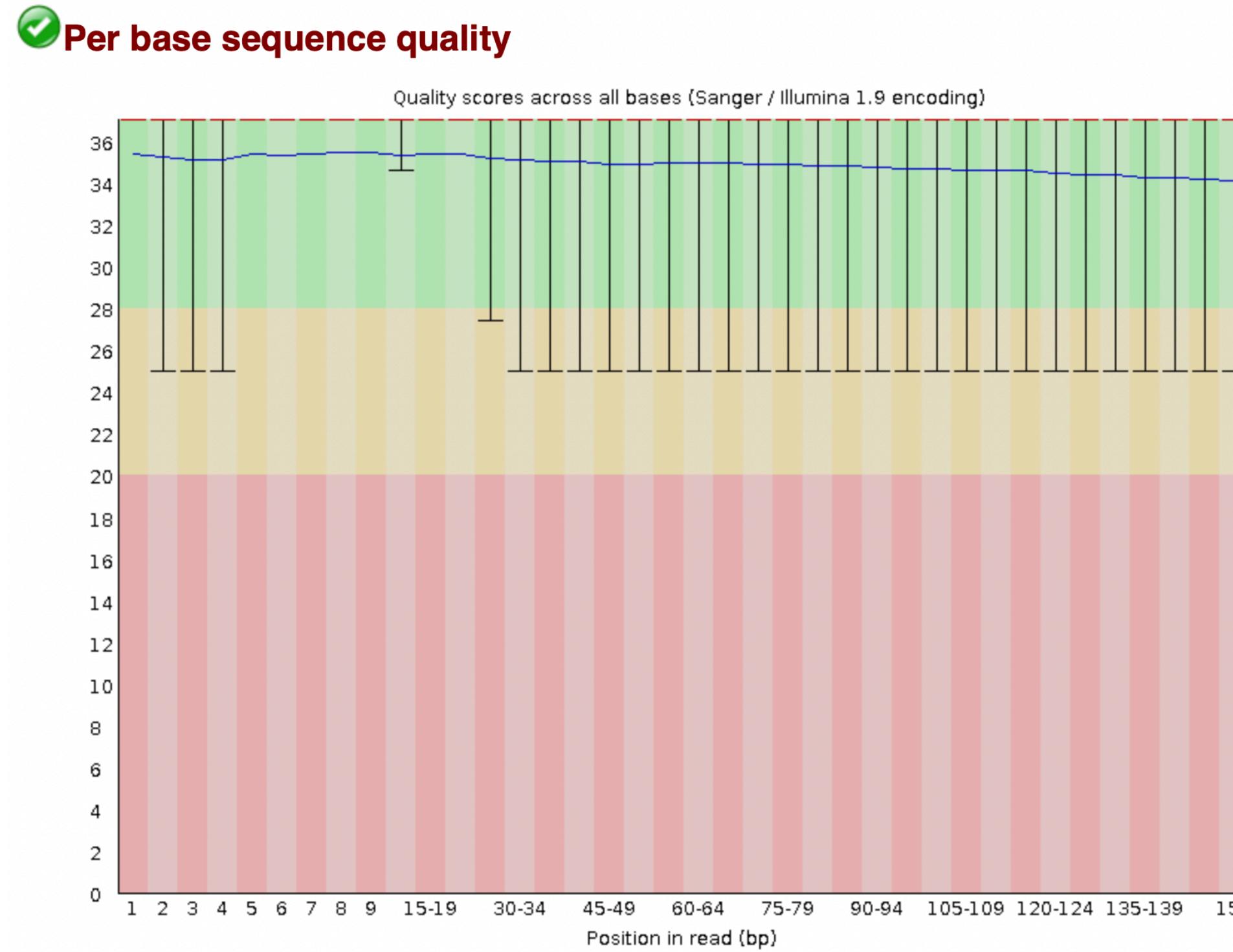
FastQC Per Base Quality



Per base sequence quality



Per Base Quality

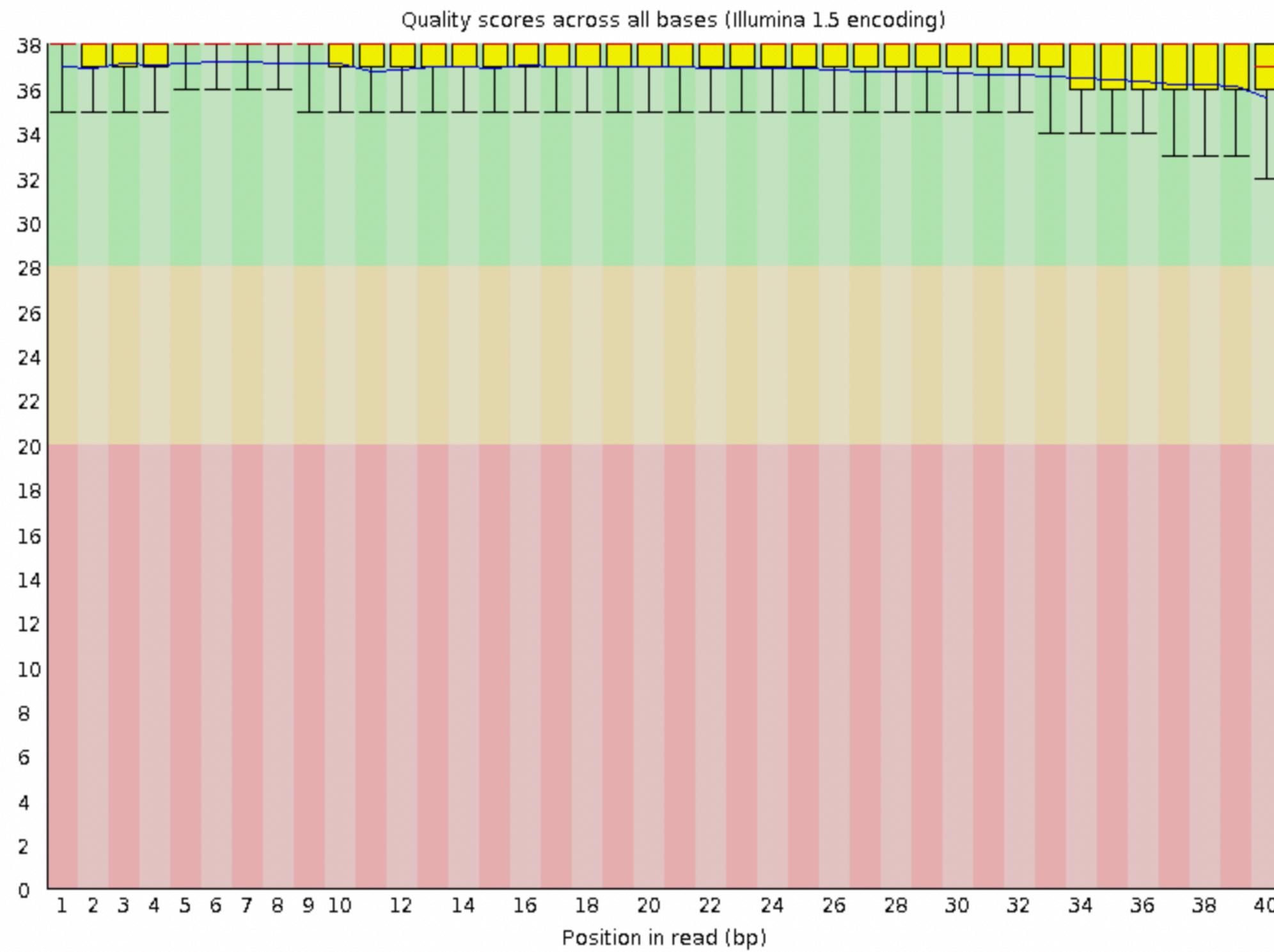


Usually, long reads lose quality at the end.

FastQC Per Base Quality

Good result

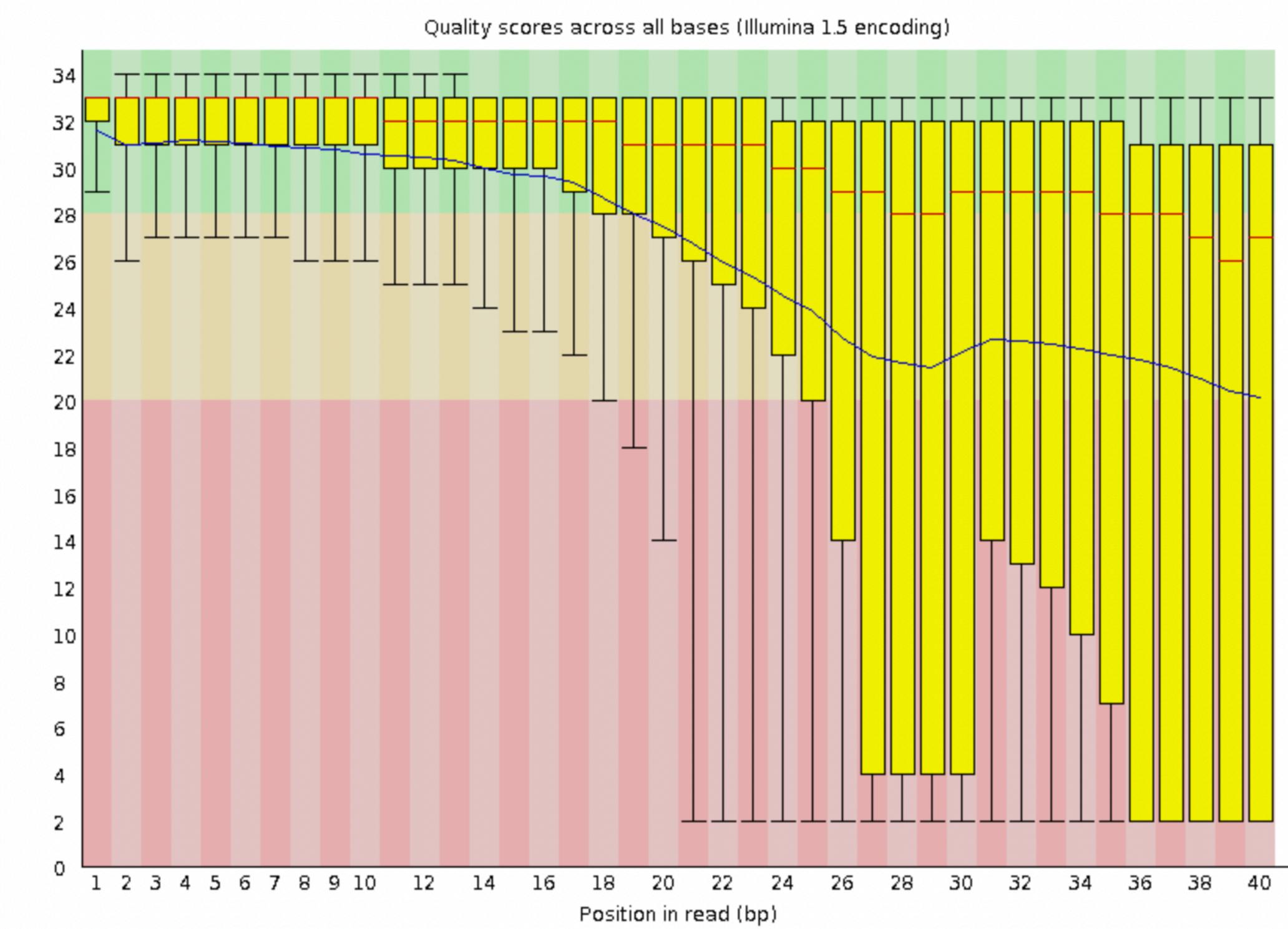
✓ **Per base sequence quality**



FastQC Example Reports: Good Illumina Data

Bad result

✗ **Per base sequence quality**



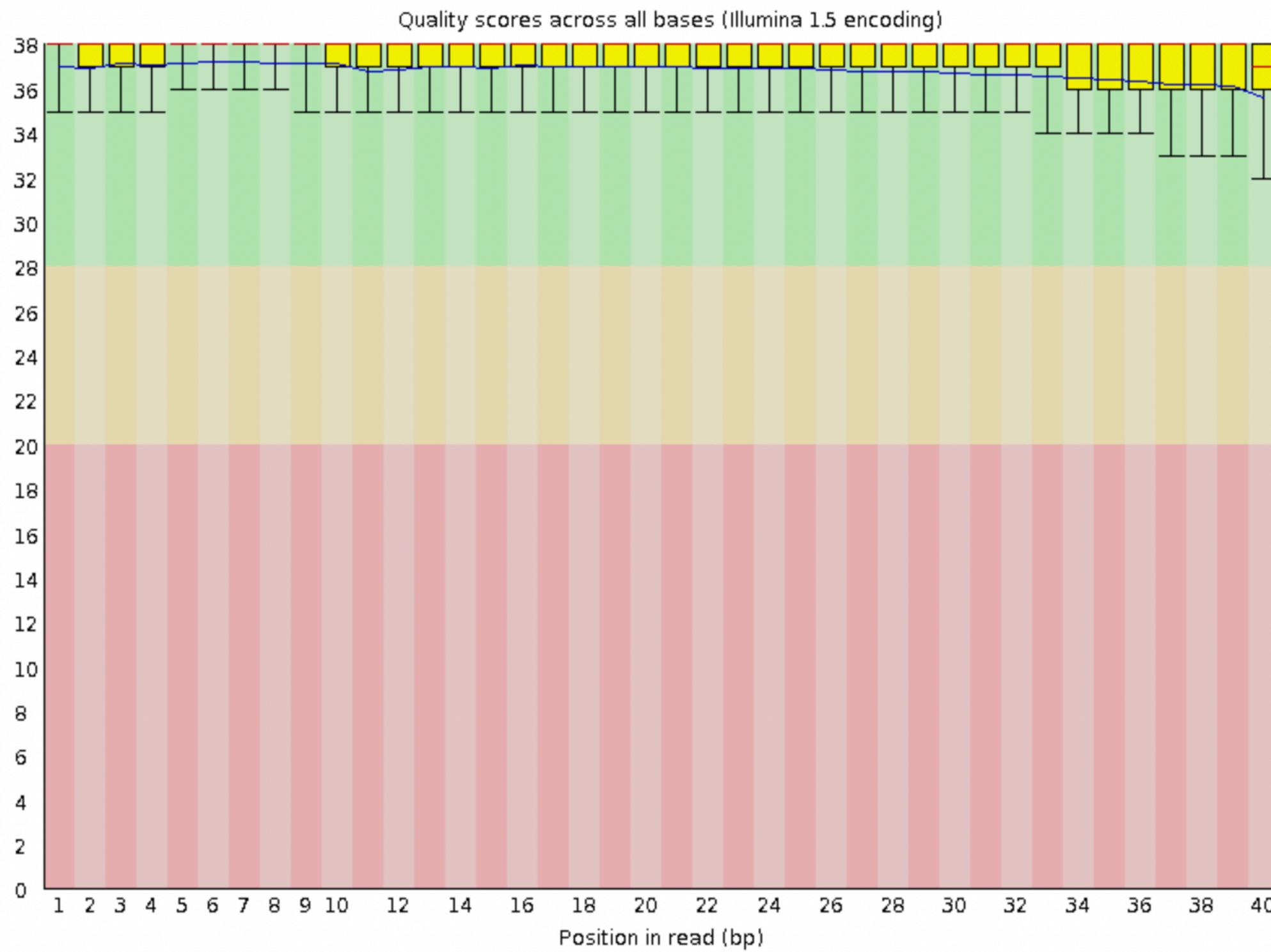
FastQC Example Reports: Bad Illumina Data

Solution: Trim where the reads are truncated based on their **average quality**.

FastQC Per Base Quality

Good result

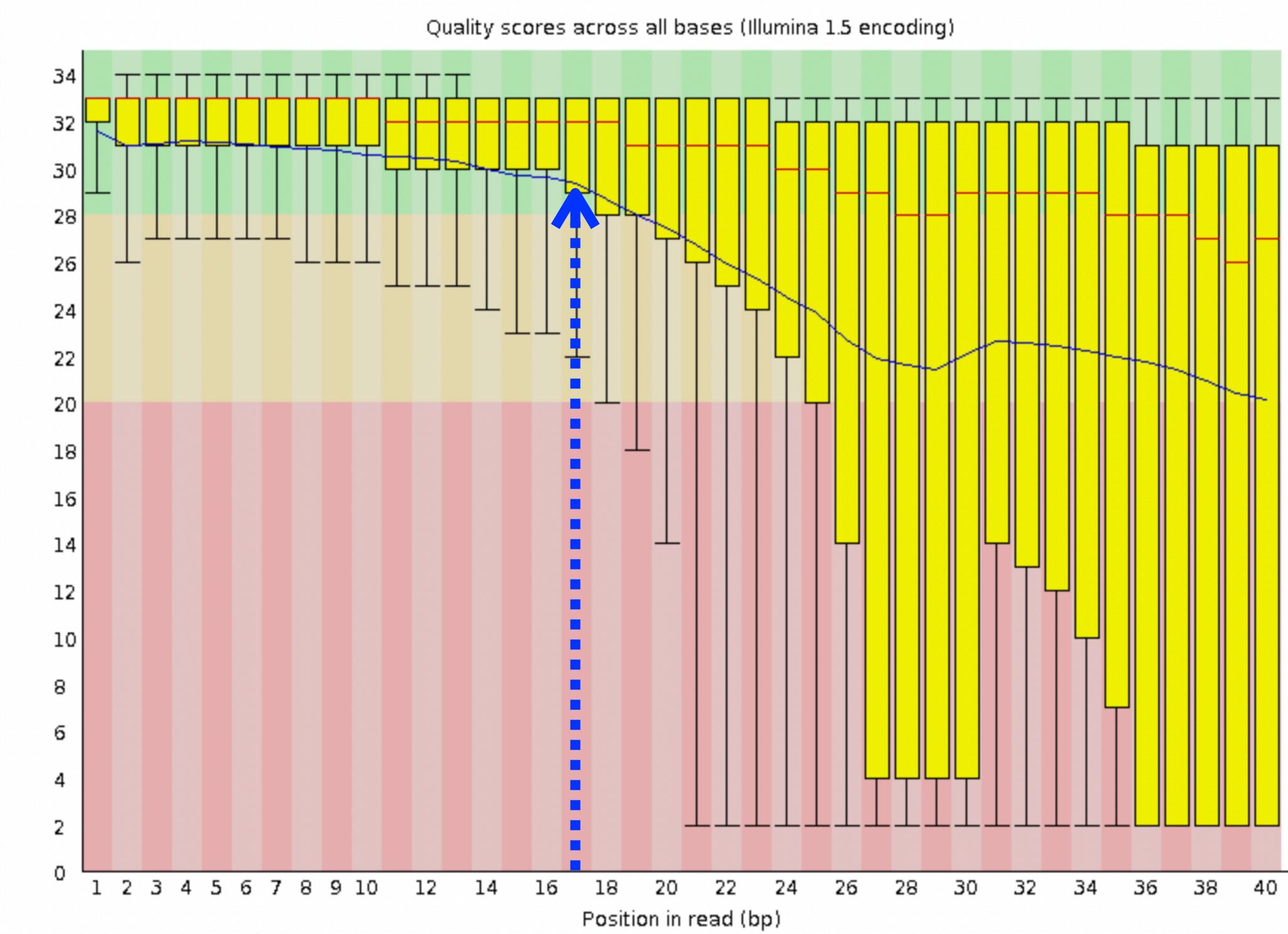
✓ Per base sequence quality



FastQC Example Reports: Good Illumina Data

Bad result

✗ Per base sequence quality



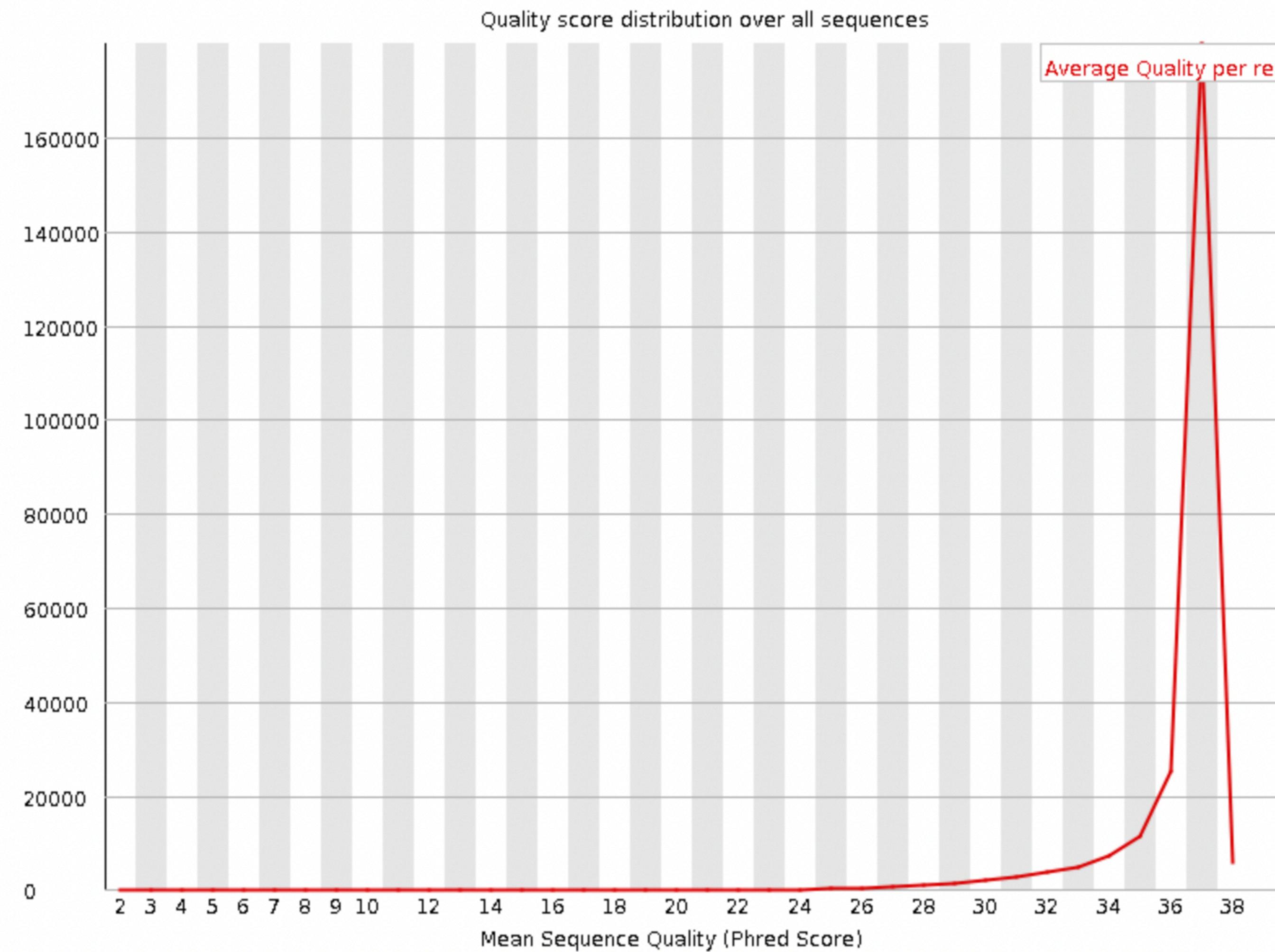
FastQC Example Reports: Bad Illumina Data

Solution: Trim where the reads are truncated based on their **average quality**.

FastQC Per Sequence Quality Scores



Per sequence quality scores

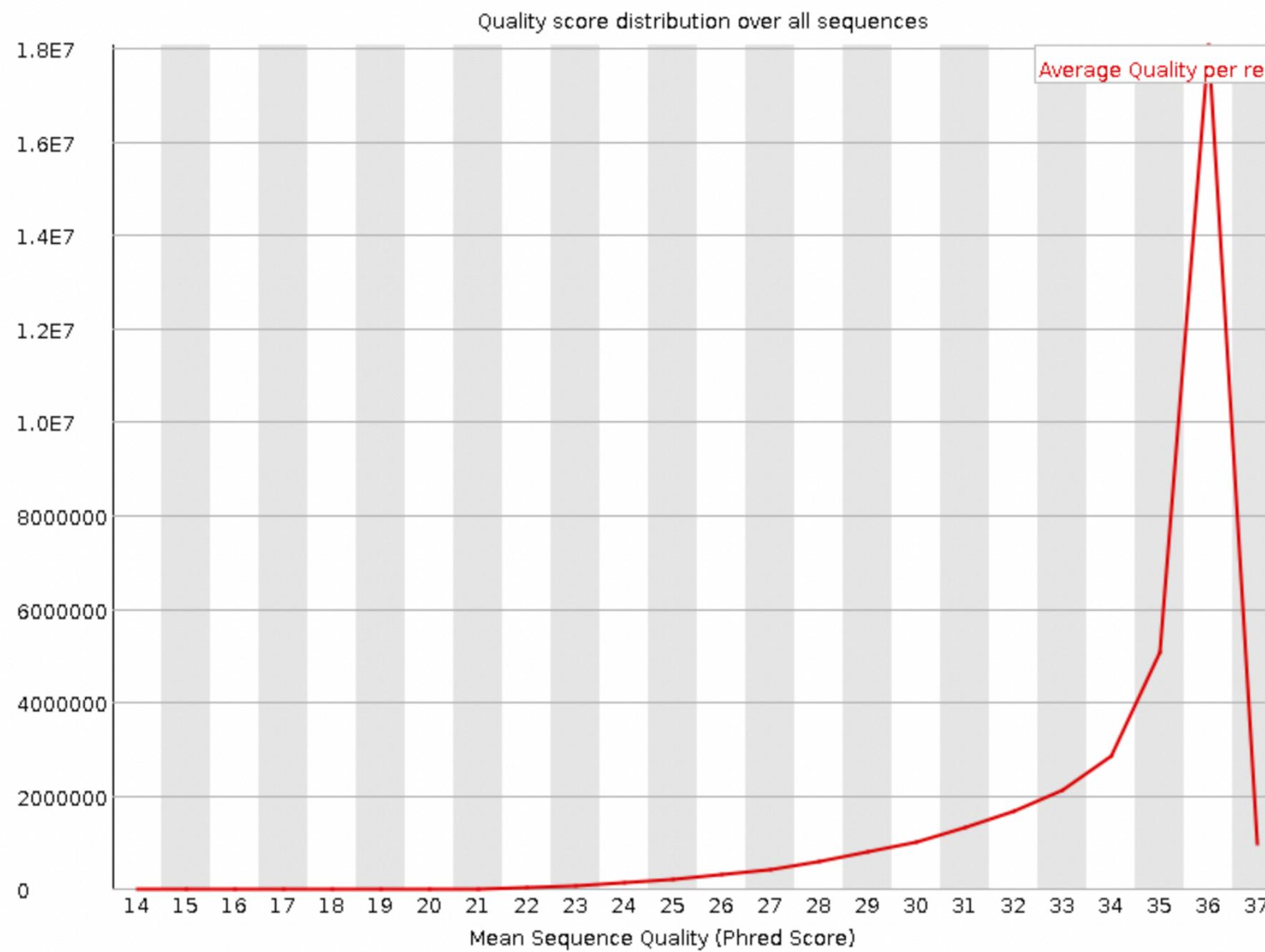


FastQC Example Reports: Good Illumina Data

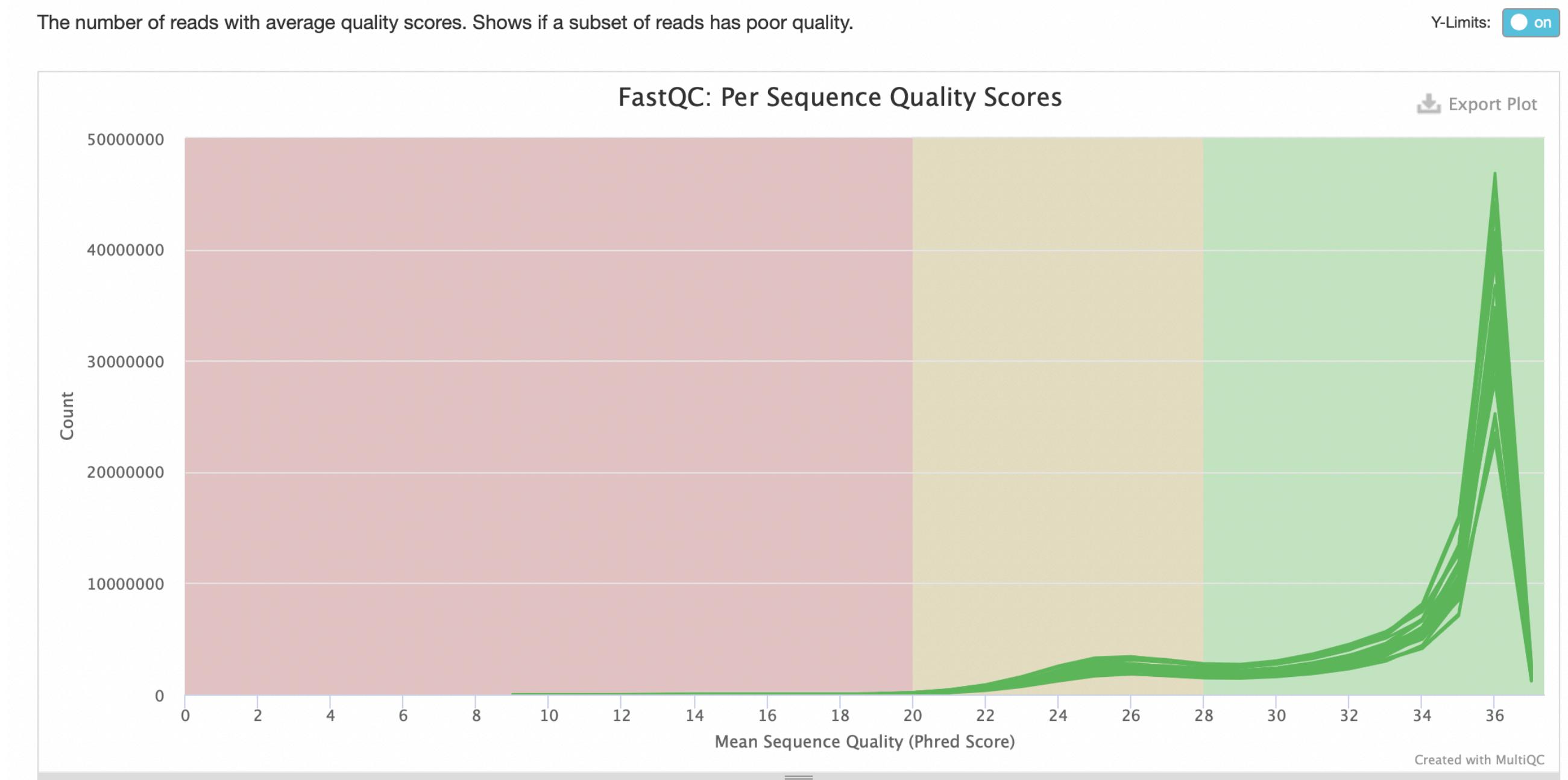
Per Sequence Quality Scores



Per sequence quality scores



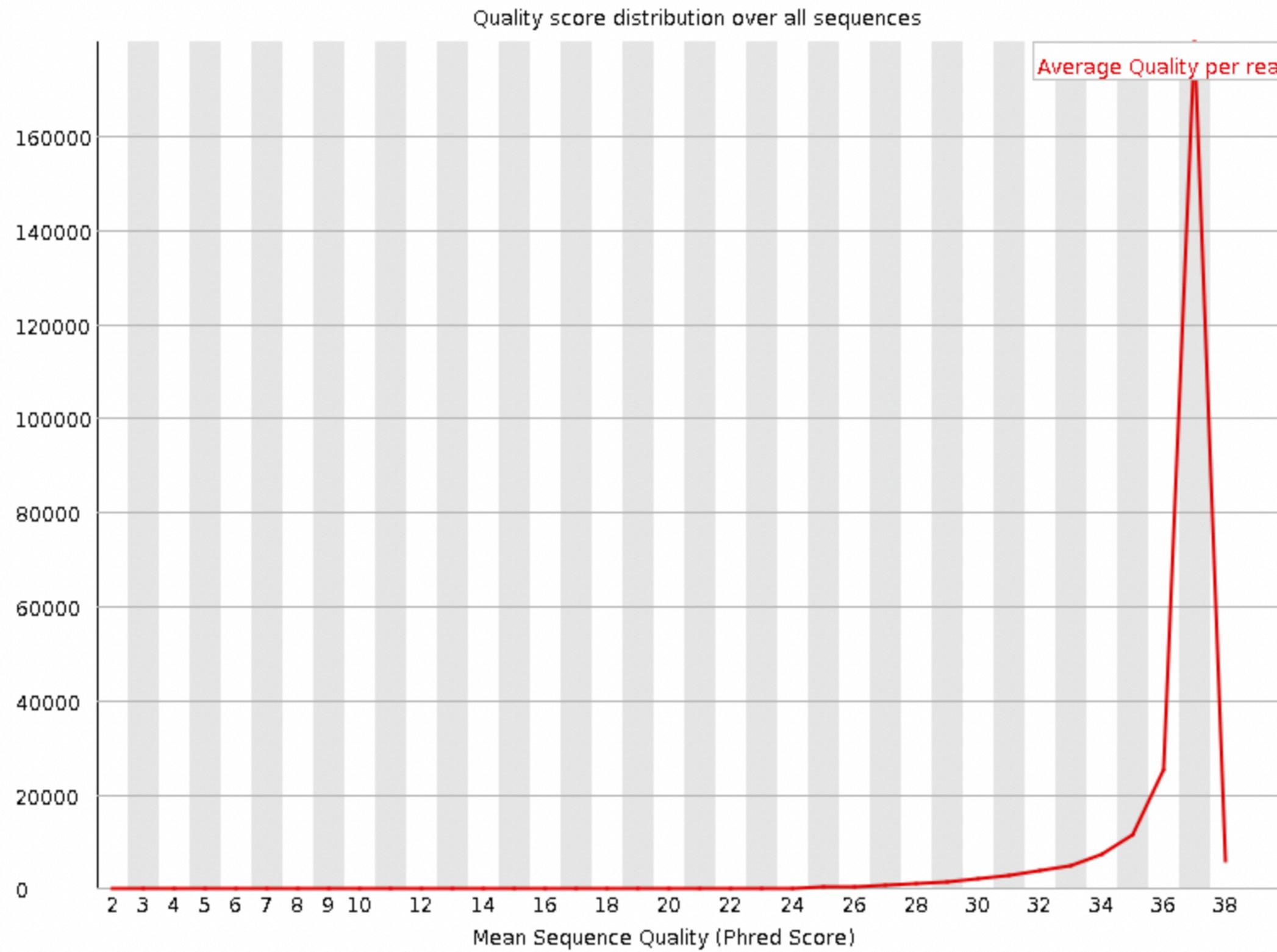
Per Sequence Quality Scores



FastQC Per Sequence Quality Scores

Good result

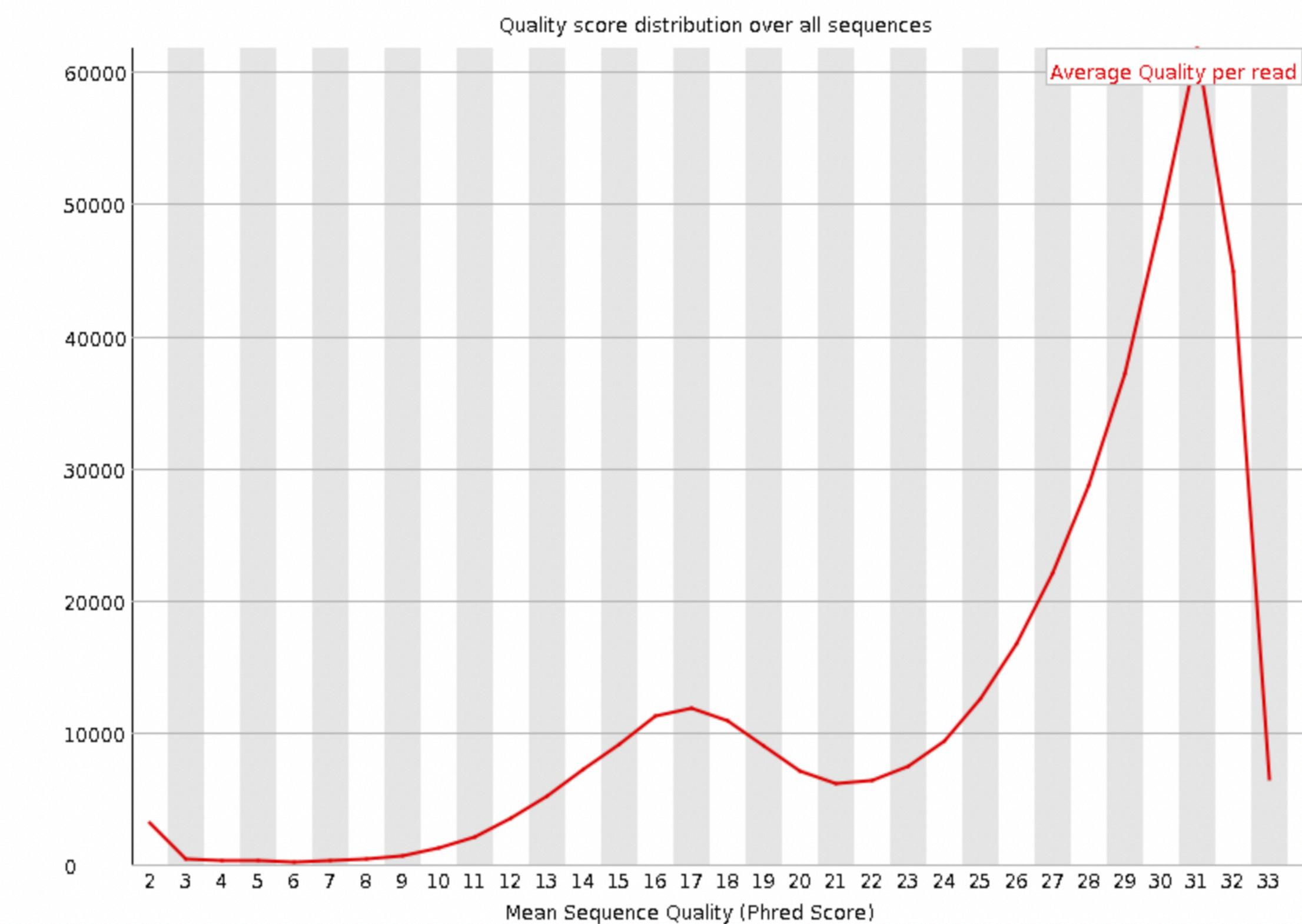
Per sequence quality scores



FastQC Example Reports: Good Illumina Data

Bad result

Per sequence quality scores

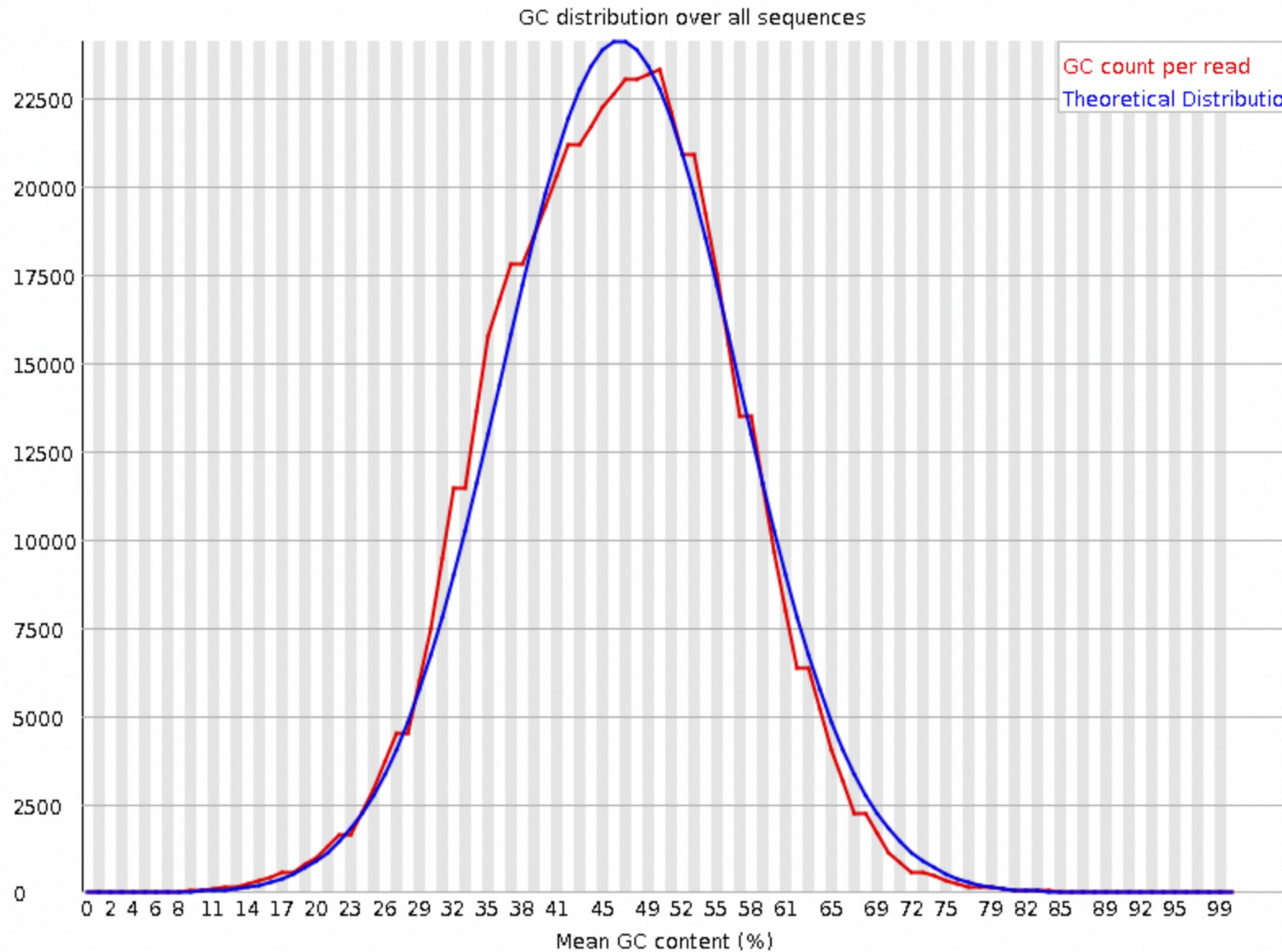


FastQC Example Reports: Bad Illumina Data

FastQC Per Sequence GC Content



Per sequence GC content

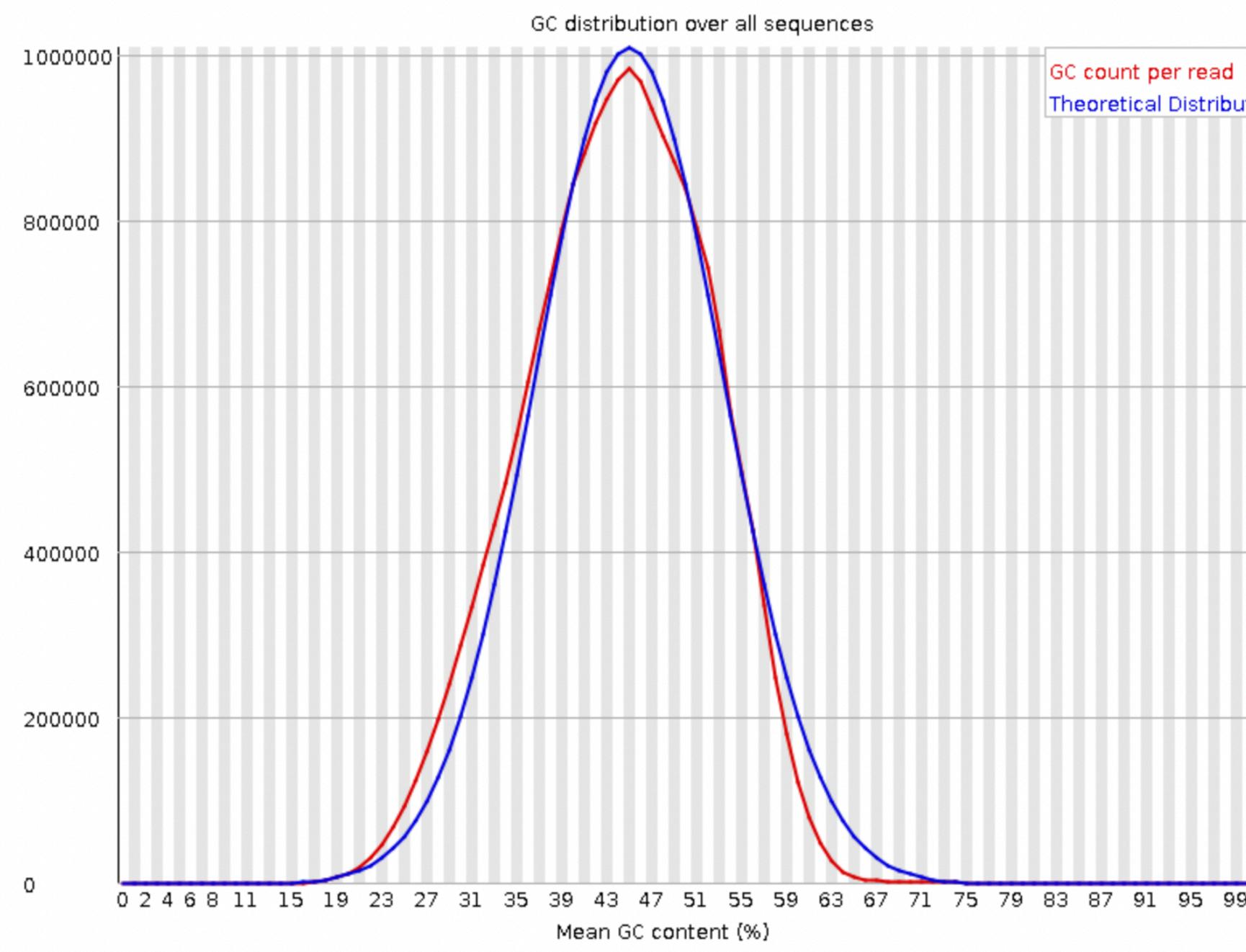


FastQC Example Reports: Good Illumina Data

Per Sequence GC Content



Per sequence GC content

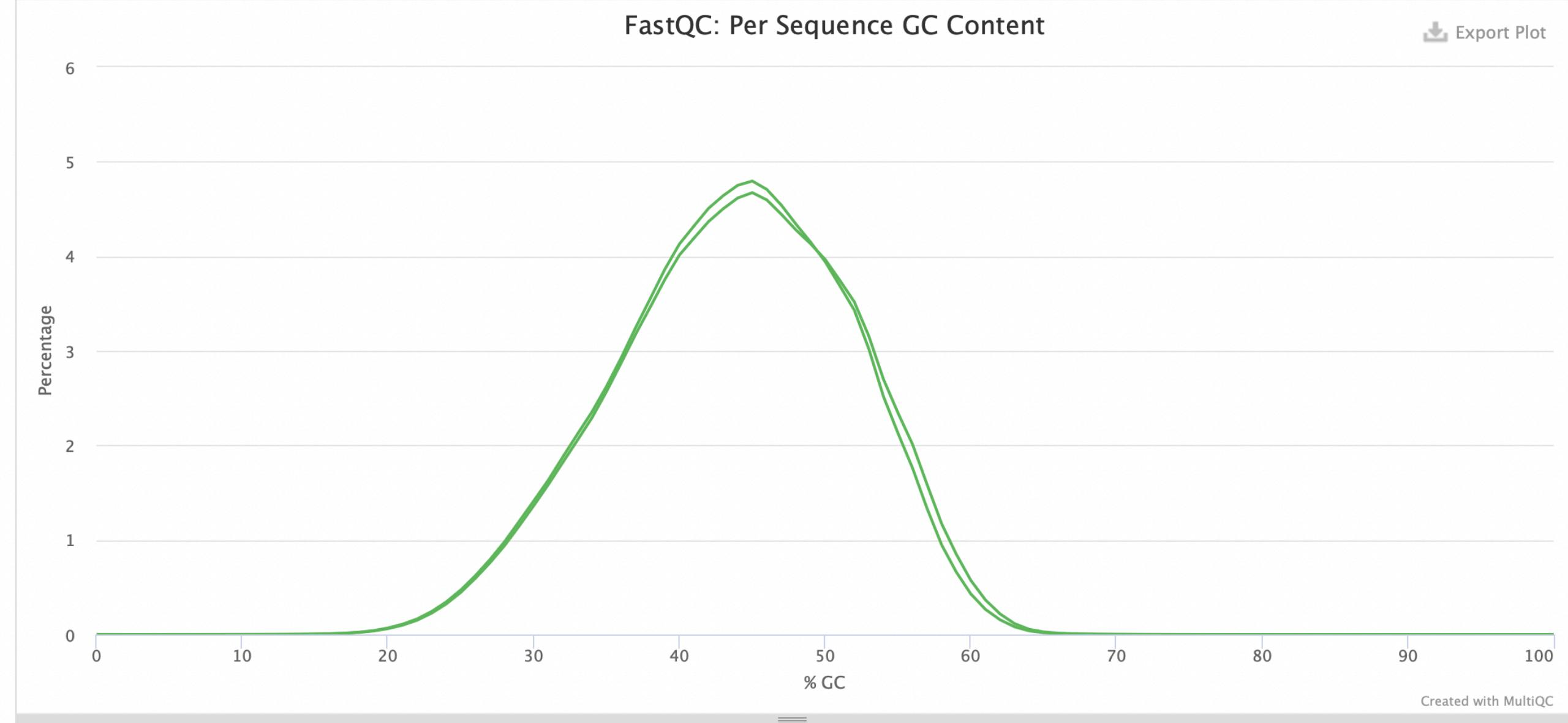


Per Sequence GC Content

The average GC content of reads. Normal random library typically have a roughly normal distribution of GC content.

ⓘ Warning: 24 samples hidden. See toolbox.

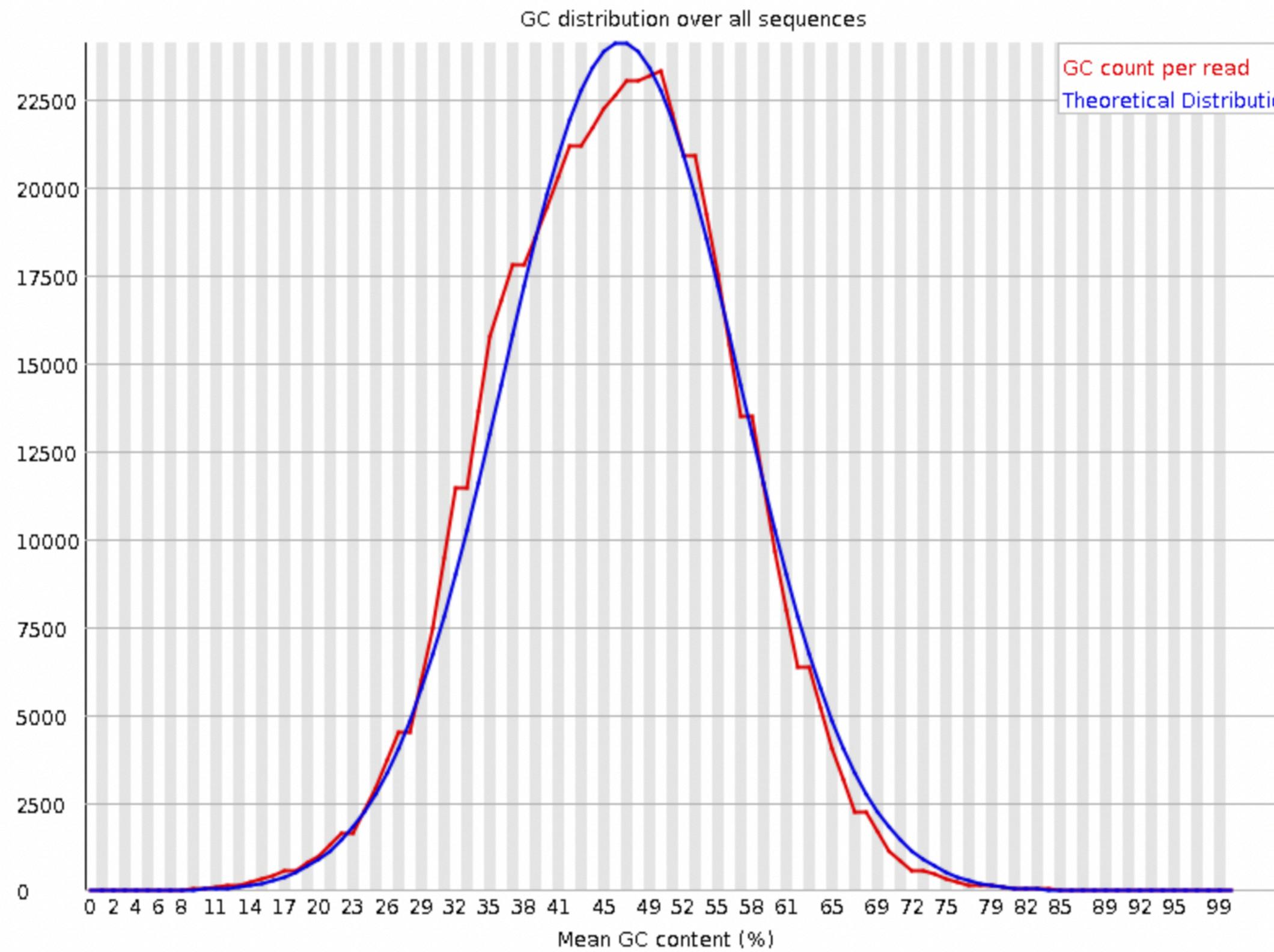
Percentages Counts



FastQC Per Sequence GC Content

Good result

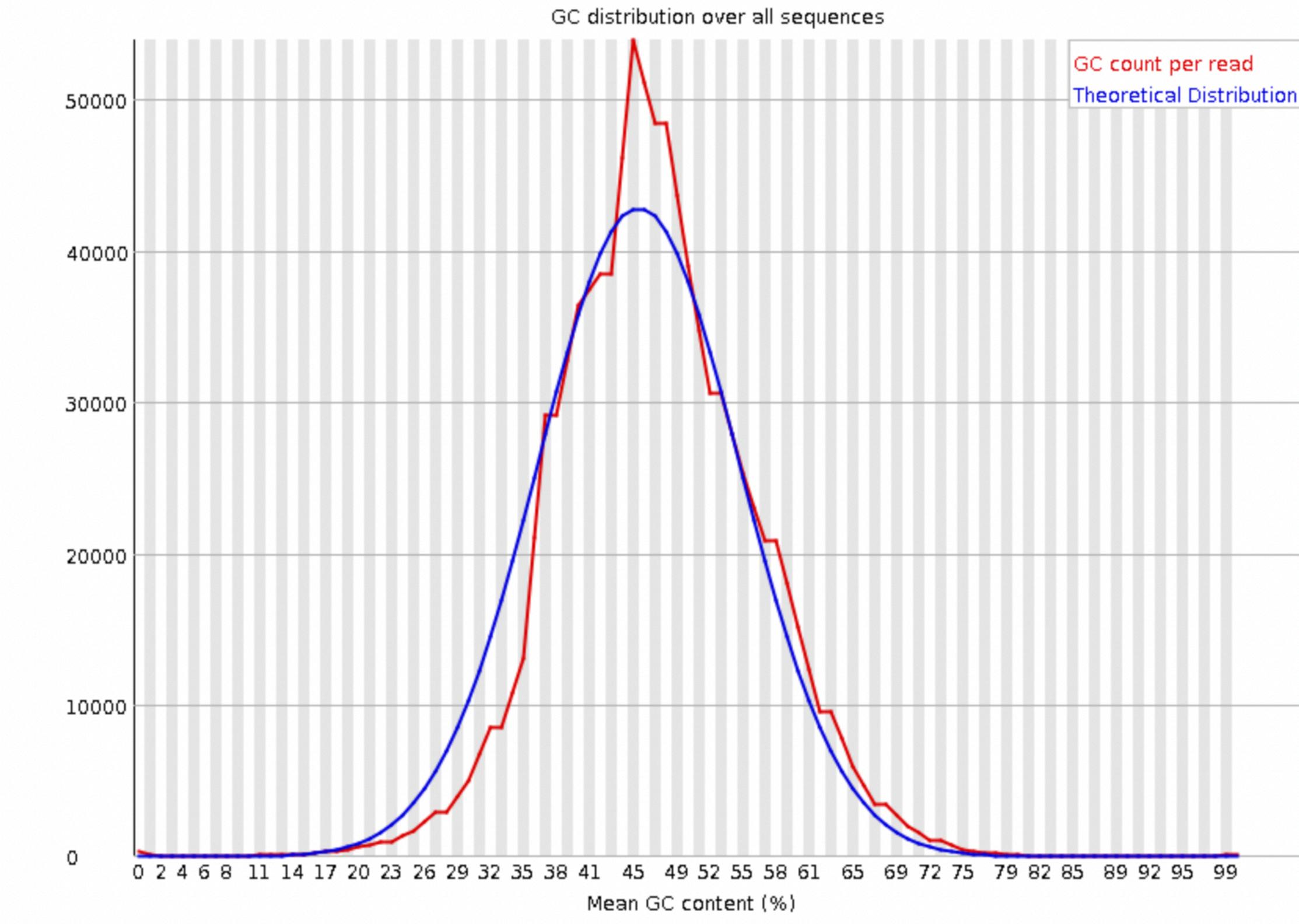
Per sequence GC content



FastQC Example Reports: Good Illumina Data

Bad result

Per sequence GC content

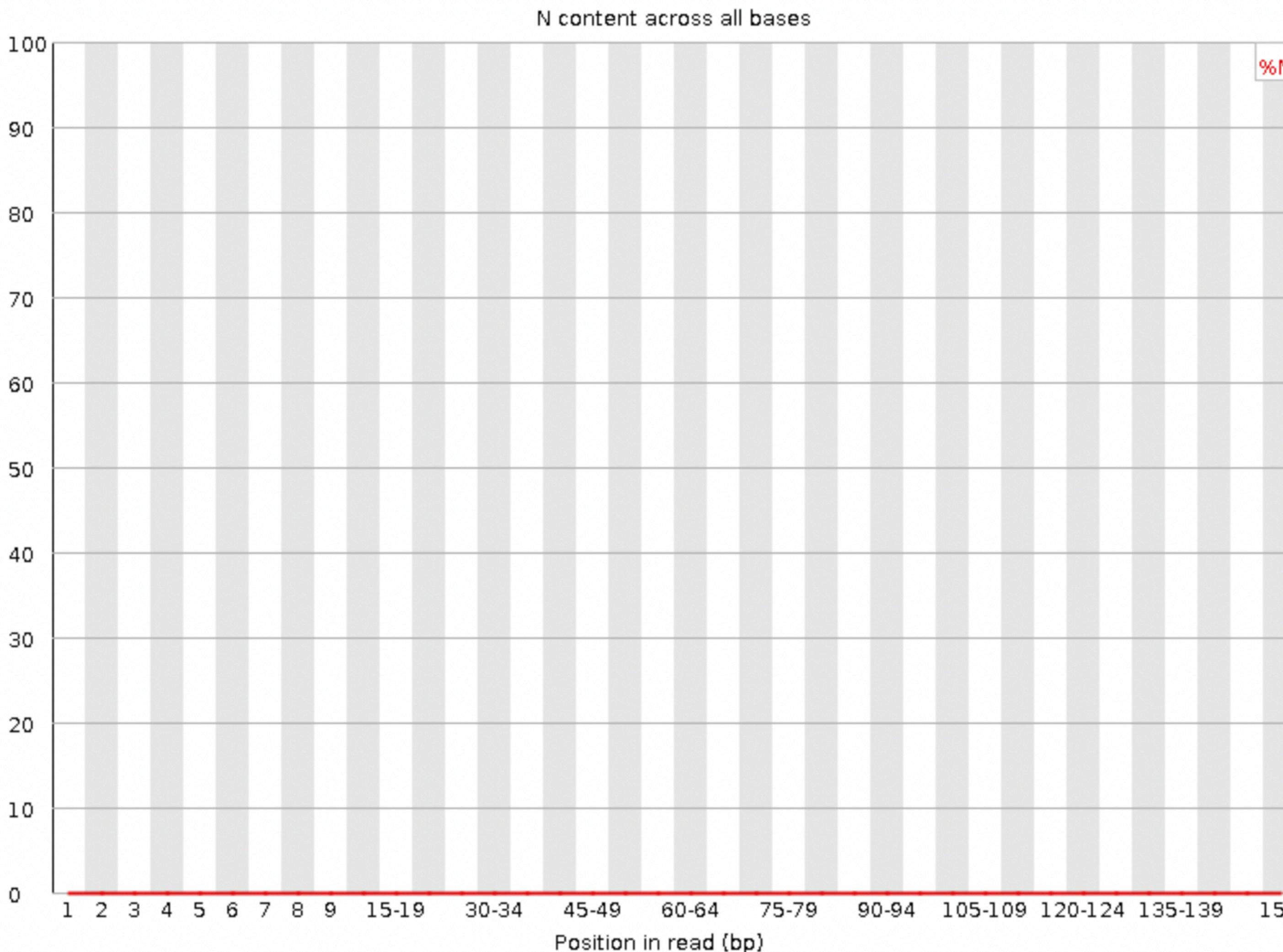


FastQC Example Reports: Bad Illumina Data

FastQC Per Base N Content



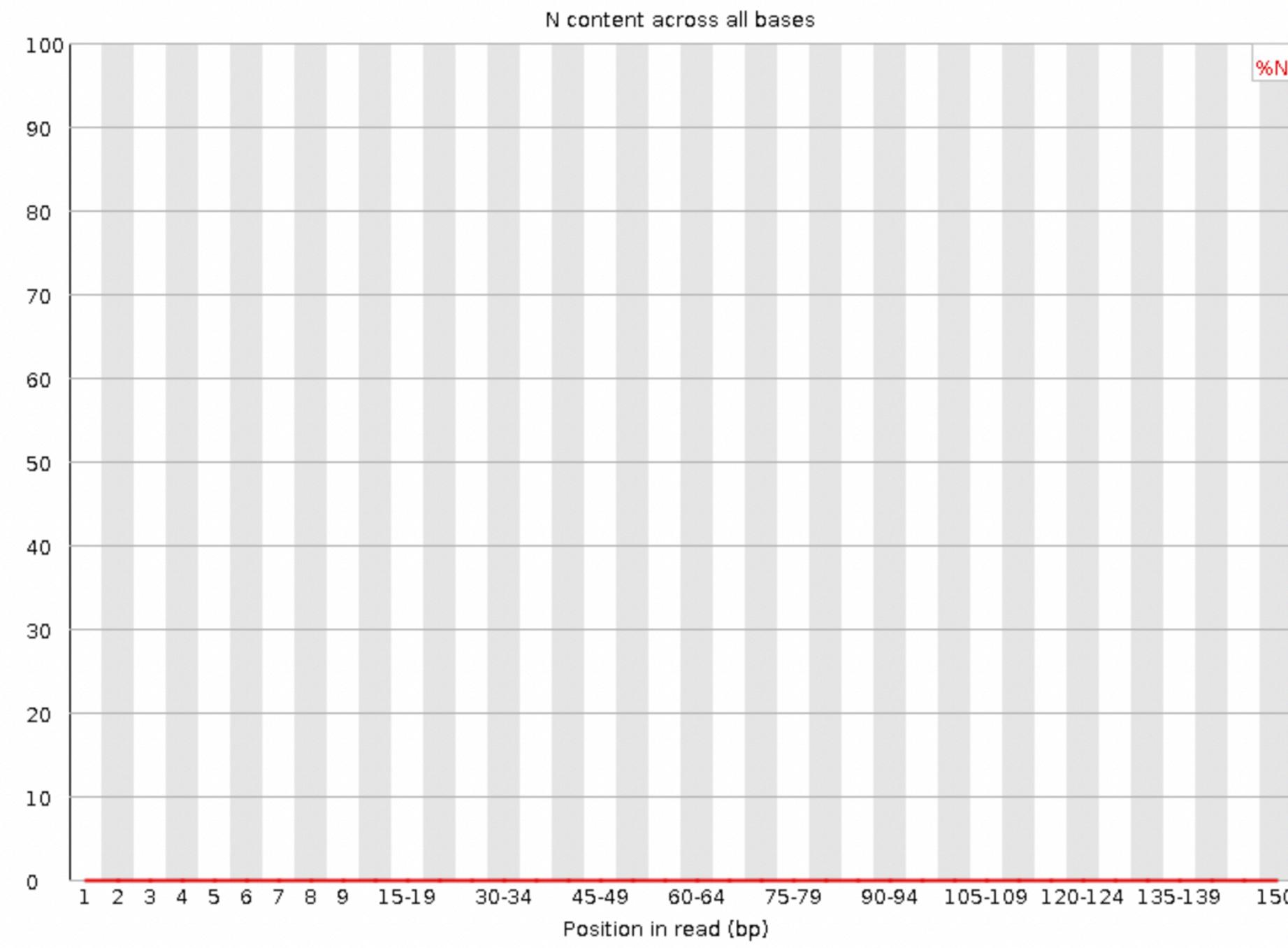
Per base N content



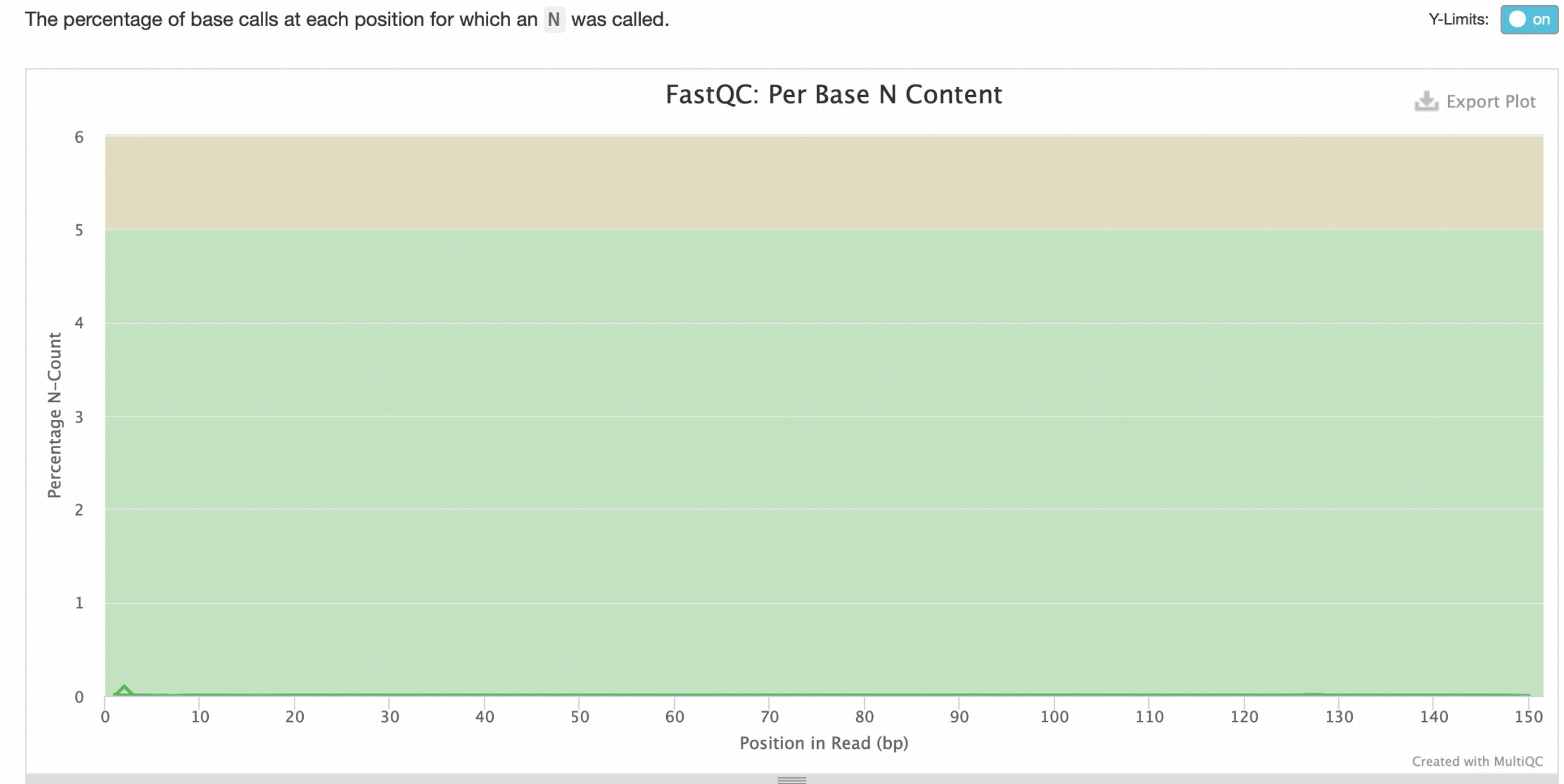
Per Base N Content



Per base N content



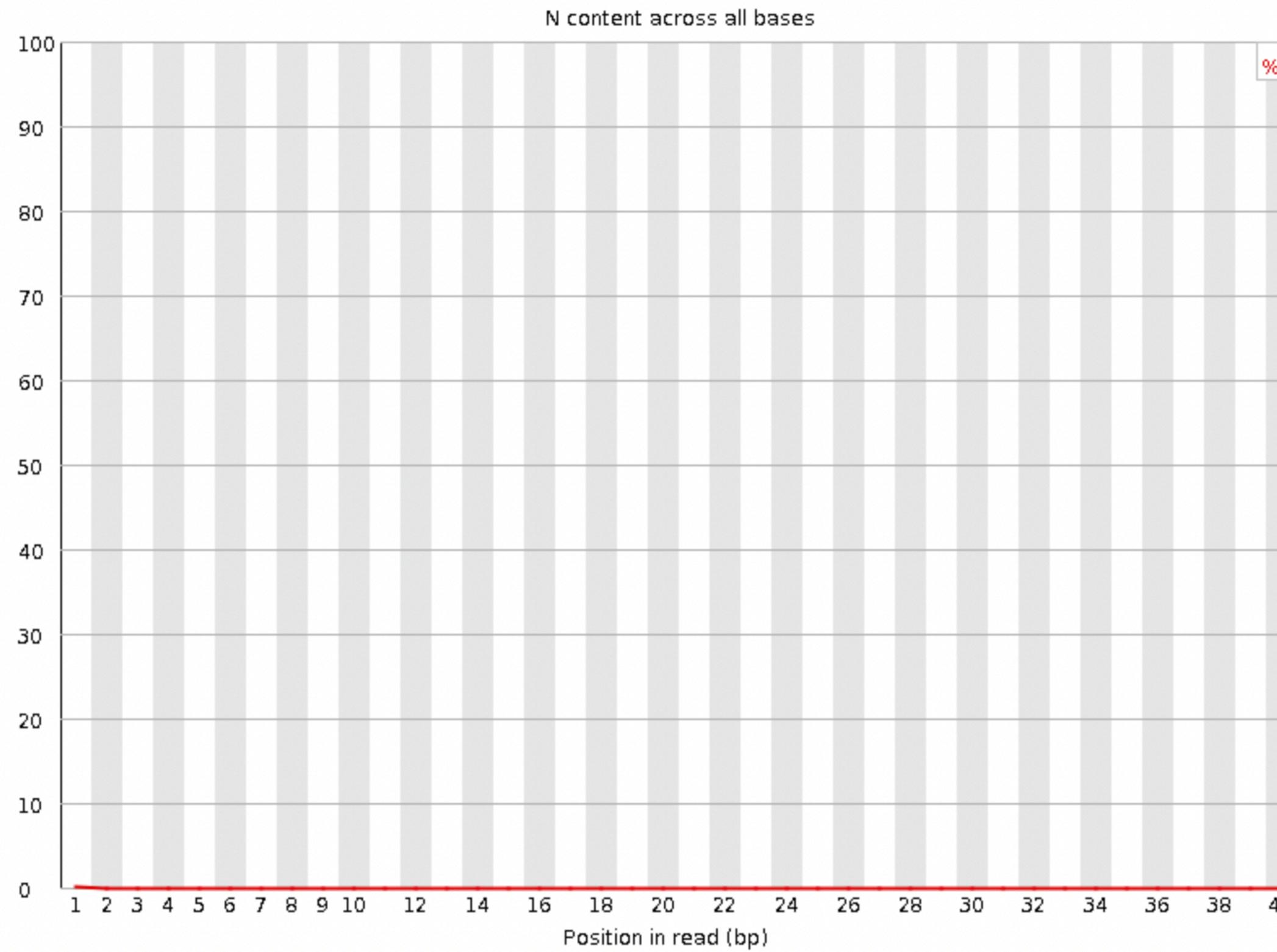
Per Base N Content



FastQC Per Base N Content

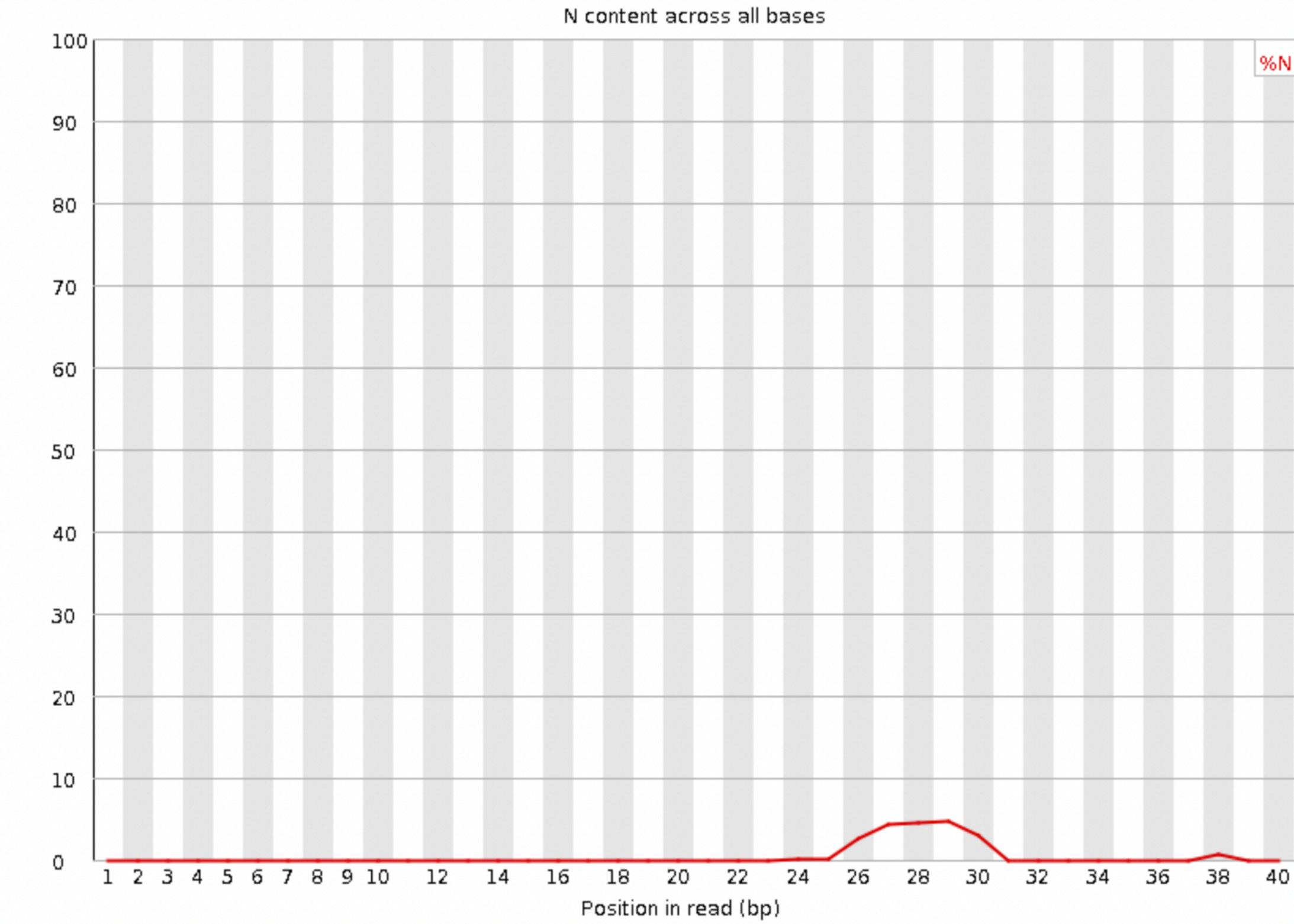
Good result

✓ **Per base N content**



Bad result

✓ **Per base N content**



FastQC Example Reports: Good Illumina Data

FastQC Example Reports: Bad Illumina Data

FastQC Overrepresented Sequences

✖ Overrepresented sequences

Sequence	Count	Percentage	Possible Source
AAGCAGTGGTATCAACGCAGAGTACTTTTTTTTTTTTTTTTT	6157956	10.298229526365521	Clontech SMART CDS Primer II A (100% over 26bp)
AGCAGTGGTATCAACGCAGAGTACTTTTTTTTTTTTTTTTT	714216	1.1944158580221549	Clontech SMART CDS Primer II A (100% over 25bp)
AAAAAAAAAAAAA AAAAAAAAAAAAA AAAAAAAAAAAAA AAAAAAAAAAAAA	544134	0.9099800039330219	No Hit
TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	104346	0.1745025554190606	No Hit

FastQC Overrepresented Sequences



Overrepresented sequences

Sequence	Count	Percentage	Possible Source
AAGCAGTGGTATCAACGCAGAGTACTTTTTTTTTTTTTTTTT	6157956	10.298229526365521	Clontech SMART CDS Primer II A (100% over 26bp)
AGCAGTGGTATCAACGCAGAGTACTTTTTTTTTTTTTTTTT	714216	1.1944158580221549	Clontech SMART CDS Primer II A (100% over 25bp)
AAAAAAAAAAAAA AAAAAAAAAAAAA AAAAAAAAAAAAA AAAAAAAAAAAAA	544134	0.9099800039330219	No Hit
TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	104346	0.1745025554190606	No Hit

Solution: It depends on the source of the overrepresented sequence:

- **Primers/Adapters:** Trim them and re-run FastQC.
- **No Hit:**
 - **polyA/T tails (in RNA-seq):** Ignore them, they won't align since they are a post-transcriptional modifications.
 - **Other:** Possible contamination (other organism, rRNA in RNA-seq,...). Remove the undesired reads.

Overrepresented Sequences



Overrepresented sequences

Sequence	Count	Percentage	Possible Source
AAGCAGTGGTATCAACGCAGAGTACTTTTTTTTTTTTTTTTT	6157956	10.298229526365521	Clontech SMART CDS Primer II A (100% over 26bp)
AGCAGTGGTATCAACGCAGAGTACTTTTTTTTTTTTTTTTT	714216	1.1944158580221549	Clontech SMART CDS Primer II A (100% over 25bp)
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	544134	0.9099800039330219	No Hit
TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	104346	0.1745025554190606	No Hit

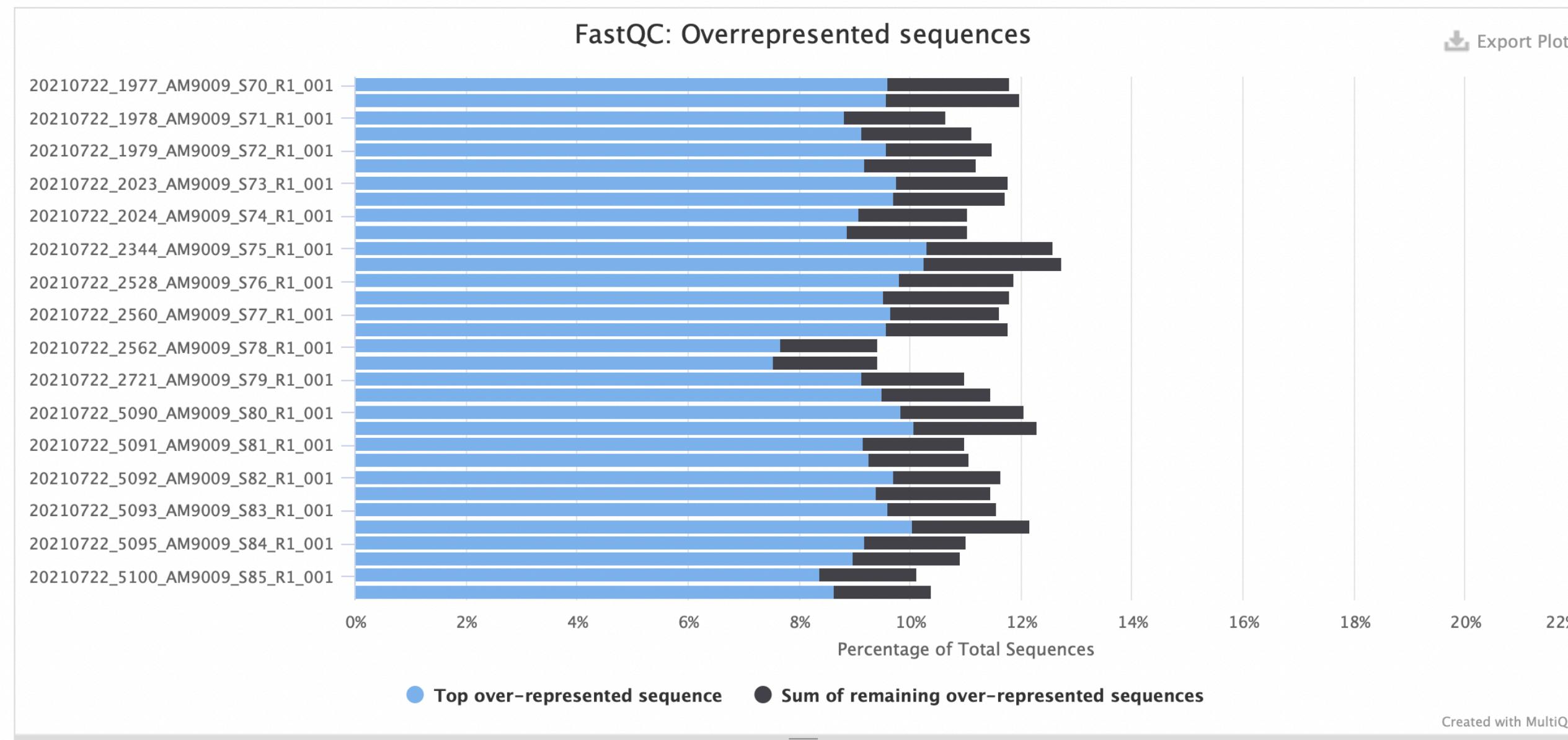


Overrepresented sequences

32

Help

The total amount of overrepresented sequences found in each library.



FastQC Overrepresented Sequences



Overrepresented sequences

No overrepresented sequences

Good result



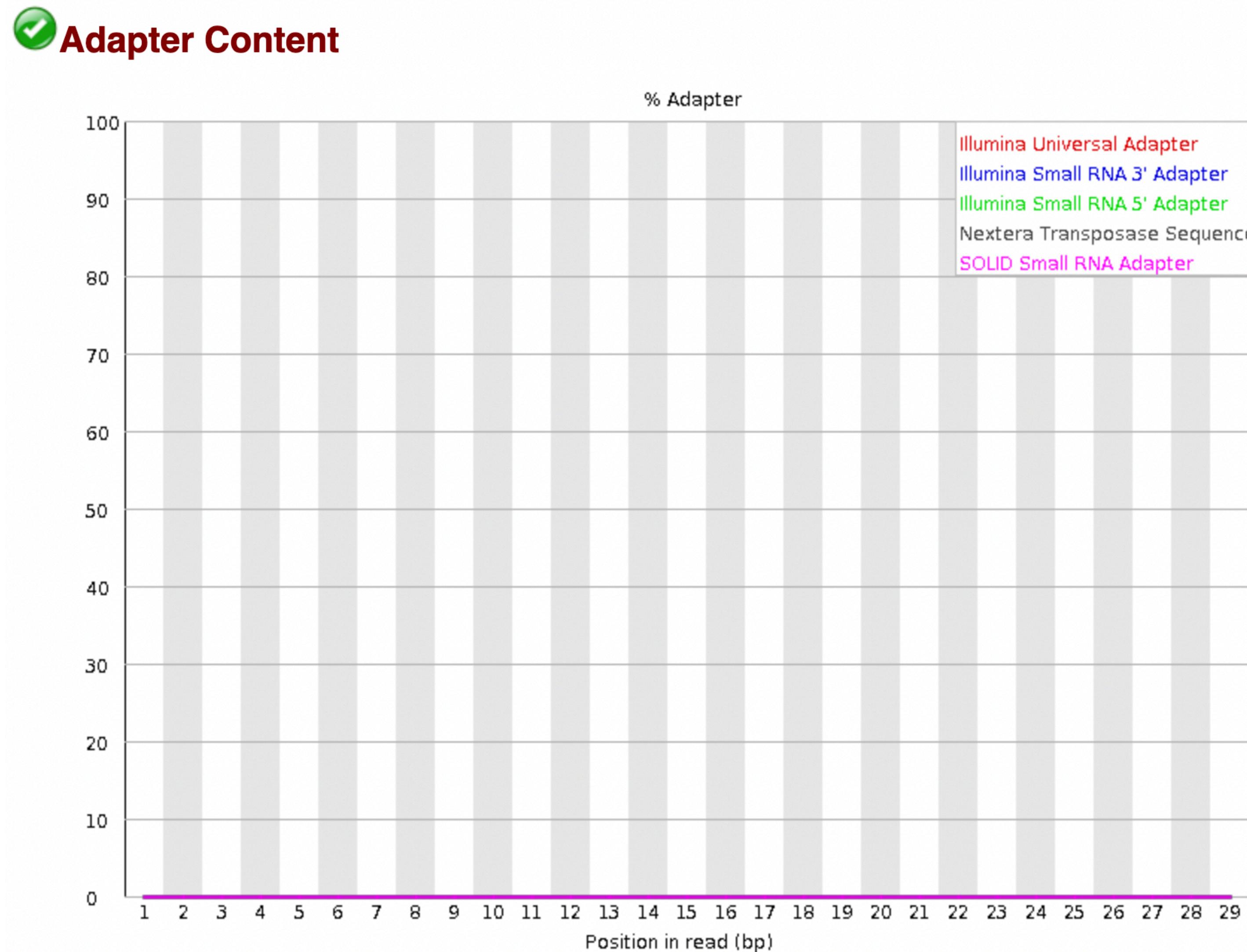
Overrepresented sequences

Bad result

Sequence	Count	Percentage	Possible Source
AGAGTTTATCGCTTCATGACGCAGAAGTTAACACTTC	2065	0.5224039181558763	No Hit
GATTGGCGTATCCAACCTGCAGAGTTTATCGCTTCATG	2047	0.5178502762542754	No Hit
ATTGGCGTATCCAACCTGCAGAGTTTATCGCTTCATGA	2014	0.5095019327680071	No Hit
CGATAAAATGATTGGCGTATCCAACCTGCAGAGTTTAT	1913	0.4839509420979134	No Hit
GTATCCAACCTGCAGAGTTTATCGCTTCATGACGCAGA	1879	0.47534961850600066	No Hit
AAAAATGATTGGCGTATCCAACCTGCAGAGTTTATCGCT	1846	0.4670012750197325	No Hit
TGATTGGCGTATCCAACCTGCAGAGTTTATCGCTTCAT	1841	0.46573637449150995	No Hit
AACCTGCAGAGTTTATCGCTTCATGACGCAGAAGTTAA	1836	0.46447147396328753	No Hit
GATAAAATGATTGGCGTATCCAACCTGCAGAGTTTATC	1831	0.4632065734350651	No Hit
AAATGATTGGCGTATCCAACCTGCAGAGTTTATCGCTTC	1779	0.45005160794155147	No Hit

FastQC Example Reports: Bad Illumina Data

FastQC Adapter Content

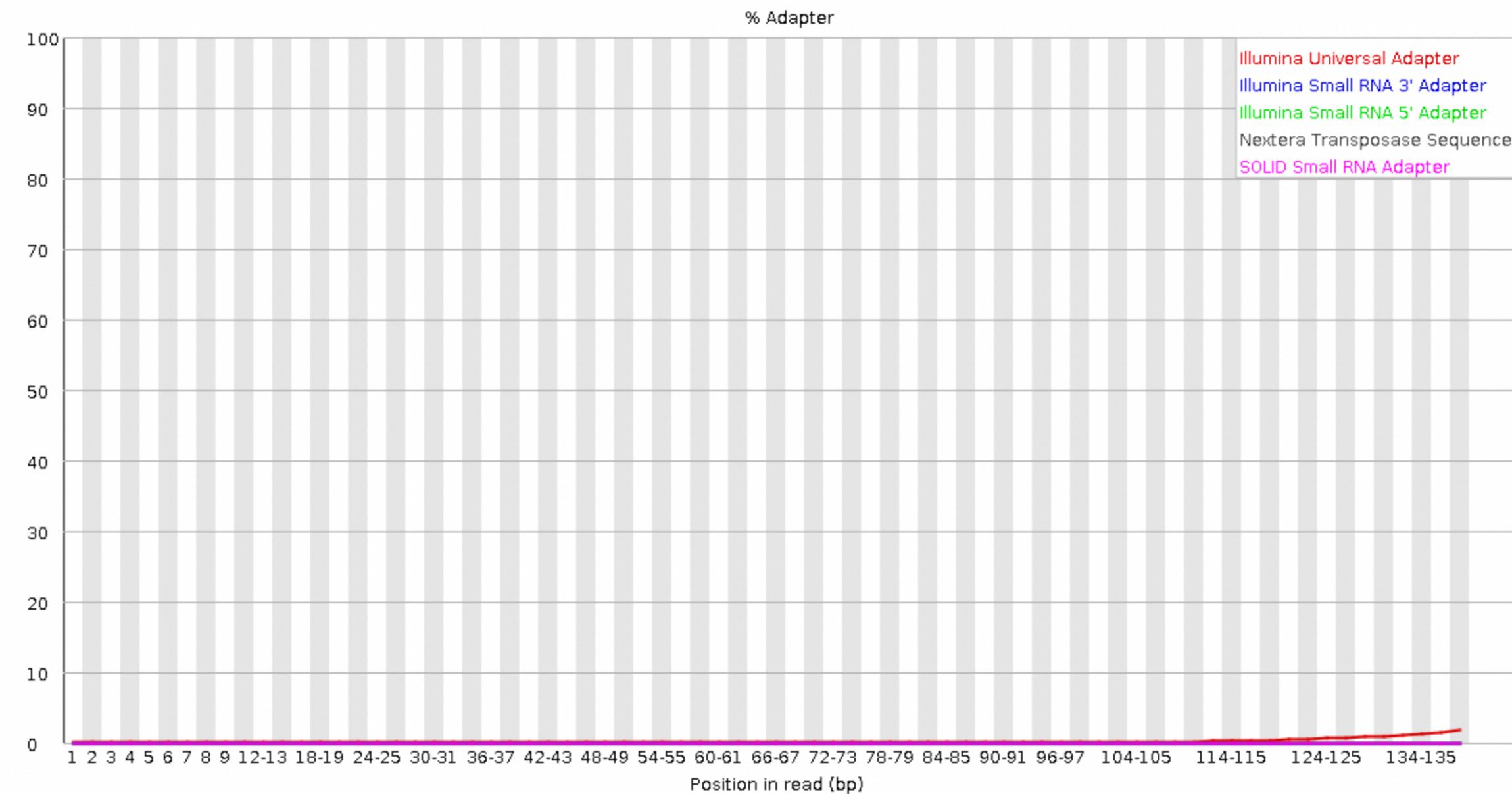


FastQC Example Reports: Good Illumina Data

Sequence Adapter Content

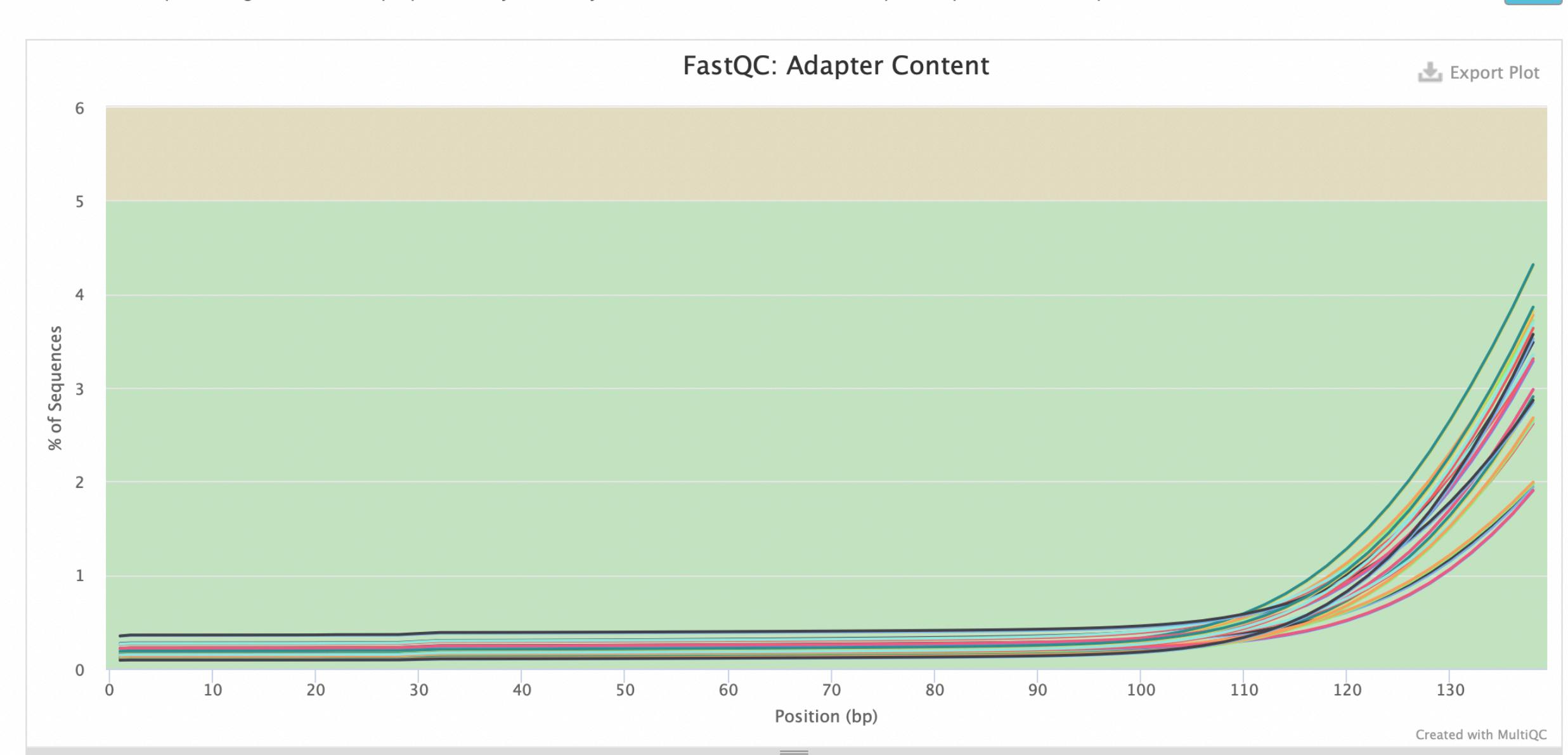


Adapter Content



Adapter Content

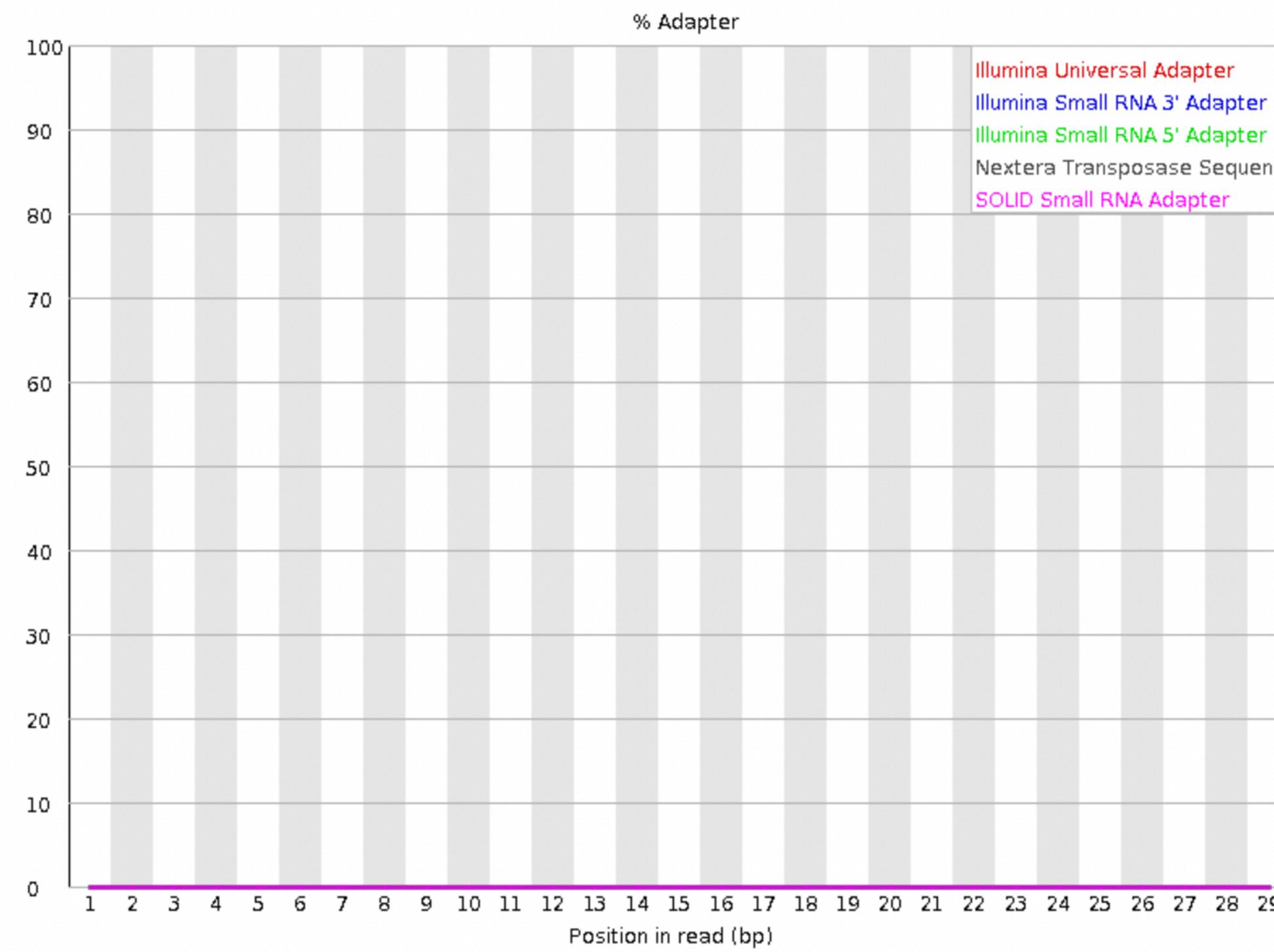
The cumulative percentage count of the proportion of your library which has seen each of the adapter sequences at each position.



FastQC Adapter Content

Good result

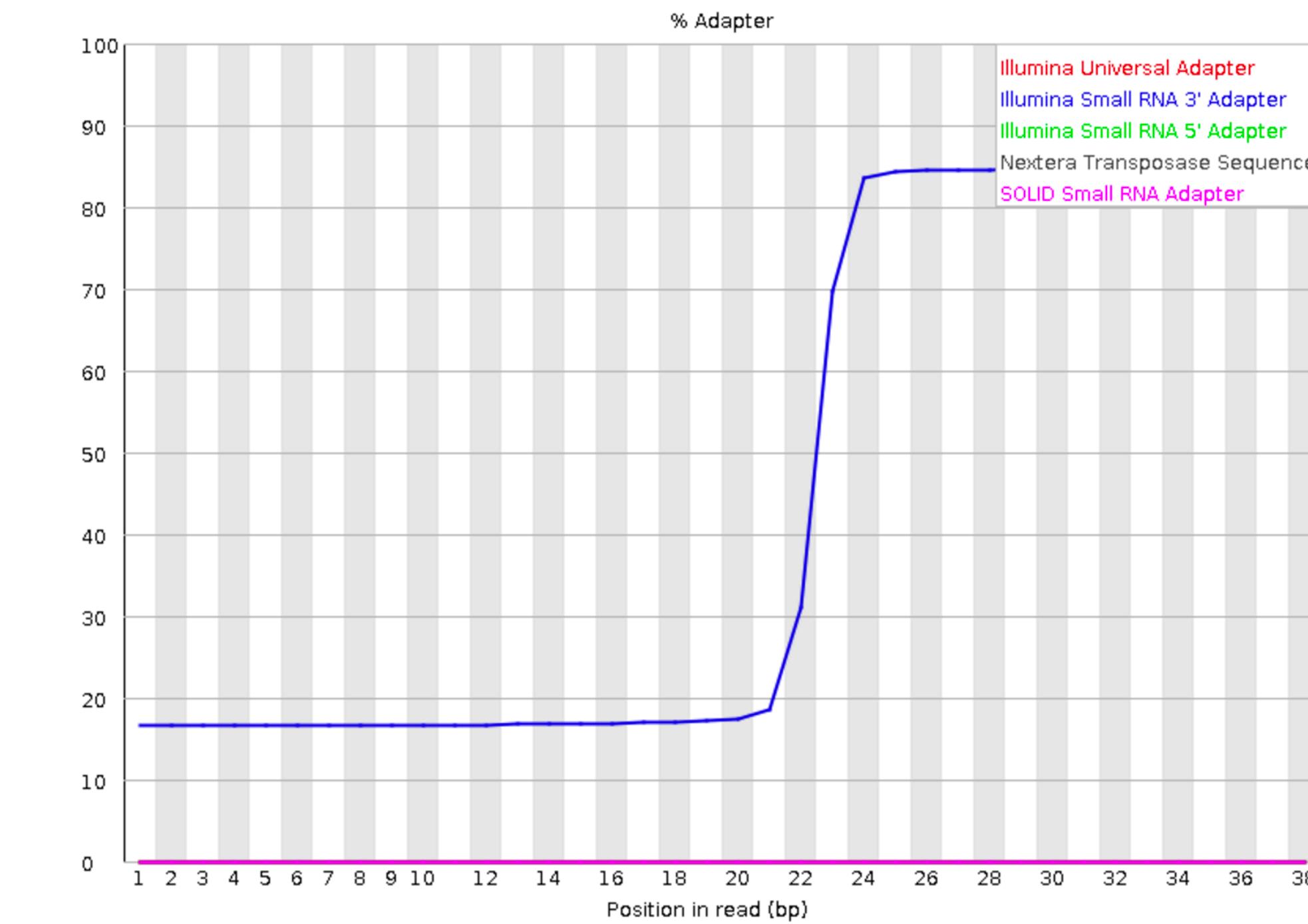
Adapter Content



FastQC Example Reports: Good Illumina Data

Bad result

Adapter Content



[Biostars post #258230](#)

Solution: Trim the adapters.

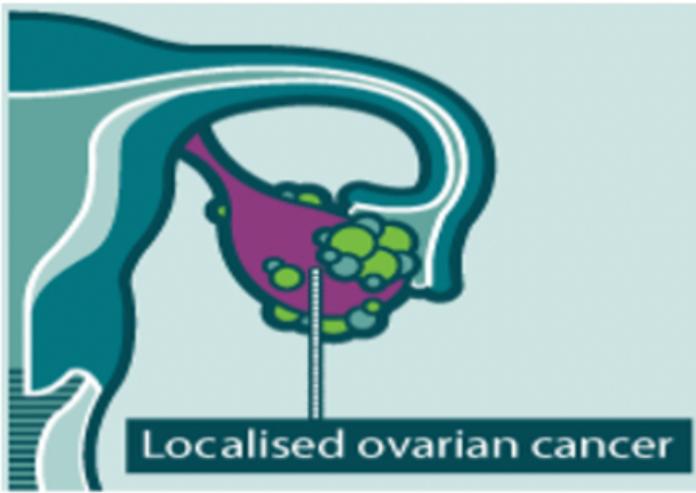
Further Info

You can check more information about these plots and examples of good and bad data at
www.bioinformatics.babraham.ac.uk/projects/fastqc

Exercise

QC of the OVCA case

Study Case - OVCA



Tumor type: Patient with Ovarian Cancer

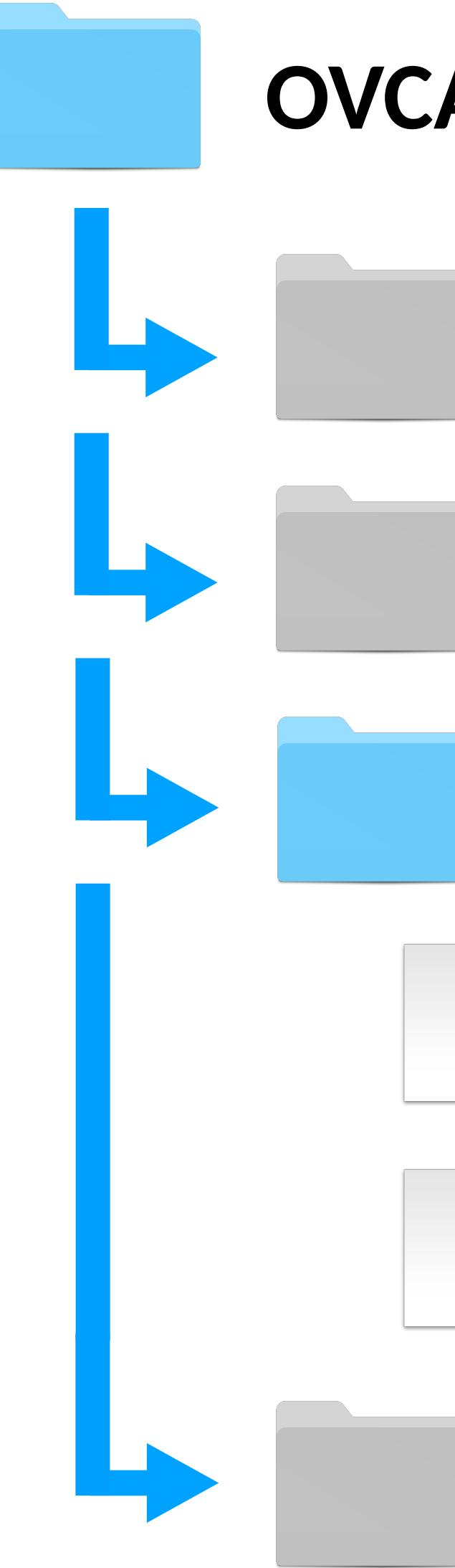
Sequencing platform: Illumina
HiSeq2000

Type of data: Whole Exome Sequencing

Samples: 2 samples of tumor with
matched healthy tissue (epithelium)

Files: [download link here](#)

NOTE: This data was simulated and
reduced



ovCA_case

Raw_data

WEx_Normal_R1.fastq
WEx_Normal_R2.fastq
WEx_Tumour_R1.fastq
WEx_Tumour_R2.fastq

Uncompress the data

1. Download the data and save it to your Desktop
2. Open a terminal (Ctrl + Alt + T) and move to the Desktop

```
$ cd /home/$USER/Desktop
```

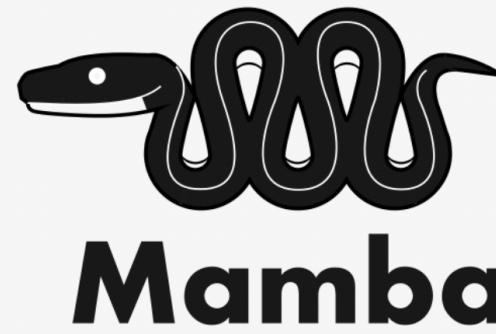
3. Untar the folder

```
$ tar xvf OVCA_case.tar
```

Install FastQC and MultiQC

1. Install FastQC and MultiQC through mamba

```
$ conda install -n base -c conda-forge mamba  
$ mamba create -n QC -c bioconda fastqc multiqc
```



Mamba is a reimplementation of conda written in C++.

<https://github.com/mamba-org/mamba>

It uses the same syntax as conda and works much faster, so it allows us to download the programs really quick.

Execution

1. Activate the conda environment

```
$ conda activate QC
```

2. Move to Raw_data folder

```
$ cd OVCA_case/Raw_data
```

3. Execute FastQC

```
$ fastqc *
```

4. Execute MultiQC

```
$ multiqc *
```

Execution

5. Deactivate the environment

```
$ conda deactivate
```

IMPORTANT: FastQC and MultiQC are executed automatically in varca, as we'll see tomorrow.

Questions

10 min

- What sequence depth was run in the experiment? (# of sequenced reads) What is the read length?
- What Phred Score encoding is detected by the algorithm?
- How is the general quality of each file?
- Is there any plot with an error/warning? Which one(s)? Any ideas why?

Answers

- Sequence depth and length: 1.4 M reads of 75 bp length.
- Phred Score encoding: Sanger/Illumina 1.9.
- Overall quality: Quite good.
- Warning/errors in GC content and Overrepresented sequences (polyNs). There are a lot of N nucleotides because this is simulated data created on purpose to raise errors/warnings.