

Tutorial 3 – FastQC

MICB405 – BIOINFORMATICS – 2021W-T1

17 SEPTEMBER 2021

AXEL HAUDUC

What is **fastqc**?

Program that generates an HTML report of FASTQ files that you provide it

Generally used for whole-genome shotgun sequencing output

- Less reliable for other kinds of sequencing output like mRNA-Seq, RNA-Seq, ChIP-seq, etc...

Why **fastqc**?

Quick, informal statistics on FASTQ files coming out of a sequencing environment

- Can indicate generally if something is drastically wrong with your FASTQ read output

Output should be taken with a grain of salt

- Thresholds for warnings are not going to be calibrated to your experiment necessarily
 - Will take into account Illumina vs other sequencing technologies from the read names

Where is **fastqc**?

It's already in your \$PATH (yay!)

- Bash already knows where to look when you call **fastq**
- Try **echo \$PATH** to see the directories (in order of priority) where bash looks for programs to run, as well as other programs that you can run!
- This is just for fun

Wait, what's a **.fastq** file again?

Text file that contains the sequence data from the clusters that pass filter on a flow cell (Illumina)

1. A sequence identifier with information about the sequencing run and the cluster. The exact contents of this line vary by based on the BCL to FASTQ conversion software used.
2. The sequence (the base calls; A, C, T, G and N).
3. A separator, which is simply a plus (+) sign with optional addition information after.
4. The base call quality scores. These are Phred +33 encoded, using ASCII characters to represent the numerical quality scores.

.fastq Read Name (@ Line) Formatting

Element	Requirements	Description
@	@	Each sequence identifier line starts with @.
<instrument>	Characters allowed: a–z, A–Z, 0–9 and underscore	Instrument ID.
<run number>	Numerical	Run number on instrument.
<flowcell ID>	Characters allowed: a–z, A–Z, 0–9	
<lane>	Numerical	Lane number.
<tile>	Numerical	Tile number.
<x_pos>	Numerical	X coordinate of cluster.
<y_pos>	Numerical	Y coordinate of cluster.
<UMI>	Restricted characters: A/T/G/C/N	Optional, appears when UMI is specified in sample sheet. UMI sequences for Read 1 and Read 2, seperated by a plus [+].
<read>	Numerical	Read number. 1 can be single read or Read 2 of paired-end.
<is filtered>	Y or N	Y if the read is filtered (did not pass), N otherwise.
<control number>	Numerical	0 when none of the control bits are on, otherwise it is an even number. On HiSeq X and NextSeq systems, control specification is not performed and this number is always 0.
<index>	Restricted characters: A/T/G/C/N	Index of the read.

Using FASTQC

```
ahauduc_mb20@orca01:/projects/micb405/data/mouse/chip_tutorial$ pwd
/projects/micb405/data/mouse/chip_tutorial
ahauduc_mb20@orca01:/projects/micb405/data/mouse/chip_tutorial$ fastqc --help
```

FastQC - A high throughput sequence QC analysis tool

SYNOPSIS

```
fastqc seqfile1 seqfile2 .. seqfileN
```

```
fastqc [-o output dir] [--(no)extract] [-f fastq|bam|sam]
      [-c contaminant file] seqfile1 .. seqfileN
```

DESCRIPTION

FastQC reads a set of sequence files and produces from each one a quality control report consisting of a number of different modules, each one of which will help to identify a different potential type of problem in your data.

If no files to process are specified on the command line then the program will start as an interactive graphical application. If files are provided on the command line then the program will run with no user interaction required. In this mode it is suitable for inclusion into a standardised analysis pipeline.

Grabbing the files for viewing with **scp**

You need to get your html files into an environment with a graphical user interface

- In other words, out of the command line interface?

Details vary depending on operating system but usually involve SCP

- Must almost always be in your machine's terminal

```
scp ahauduc_mb20@orca1.bcgsc.ca:/home/ahauduc_mb20/fqc_output/* .
```


Where to find files on your computer after SCPing from your home folder

Windows Command Prompt

- C:\Users\windows_username (a.k.a. %HOMEPATH%)

Windows WSL

- \\wsl\$\Ubuntu\home\linux_username (a.k.a. ~)

macOS

- /Users/macOS_username (a.k.a. ~)

Linux

- /home/linux_username (a.k.a. ~)

How to interpret the FASTQC output

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](#)

✓ Basic Statistics

Measure	Value
Filename	F01_R1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	6433077
Sequences flagged as poor quality	0
Sequence length	32-250
%GC	67

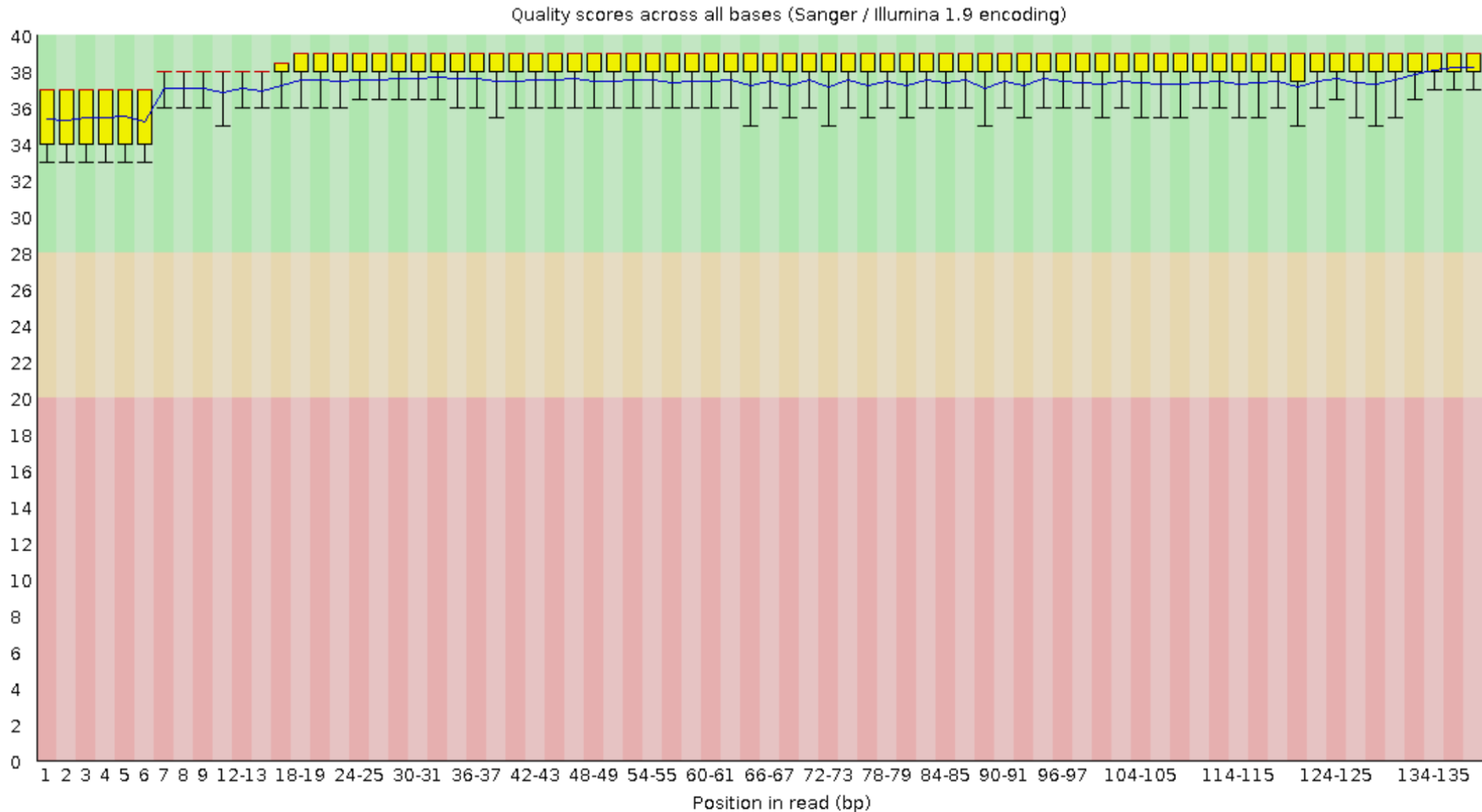
✓ Per base sequence quality



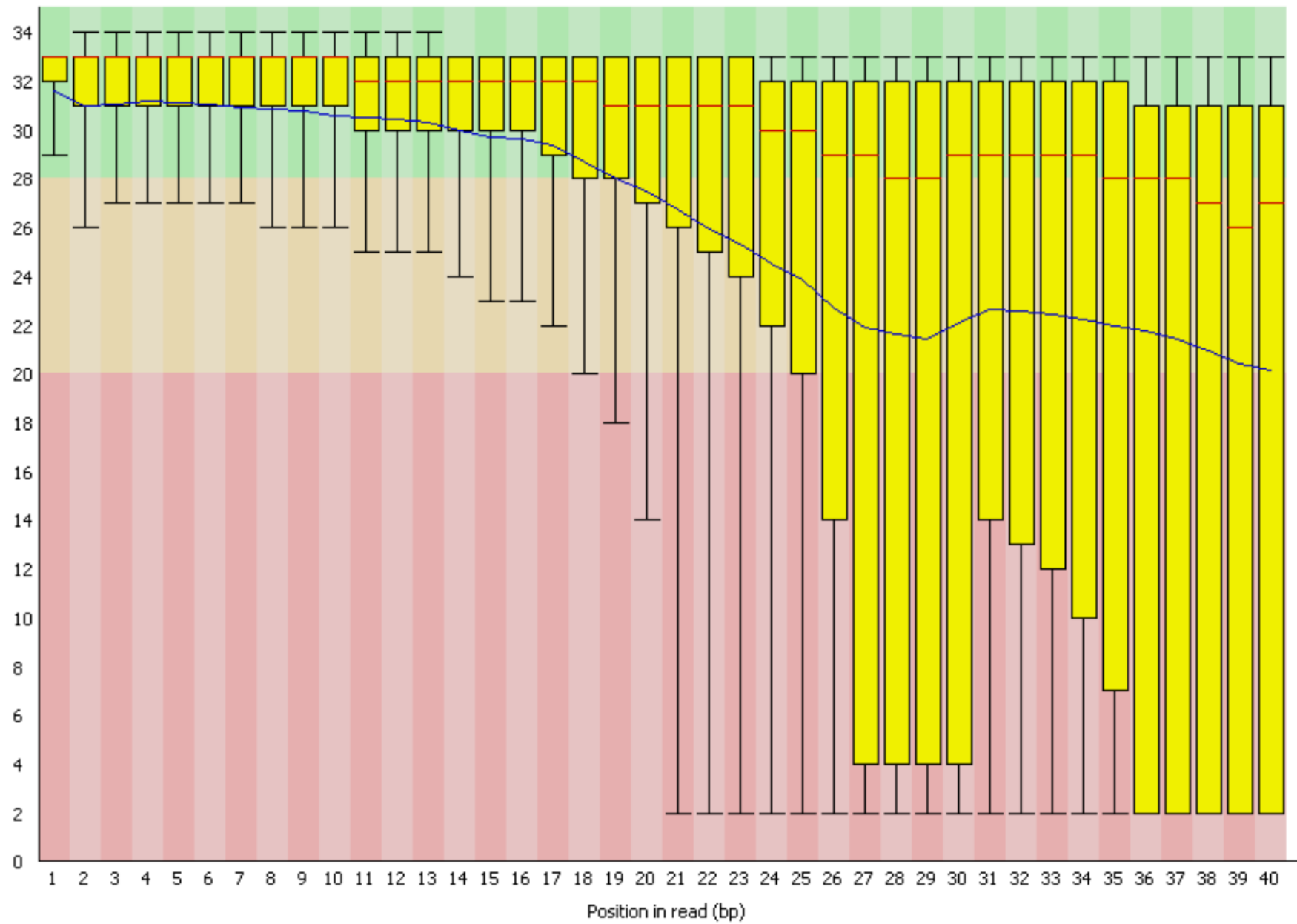
Summary

- ✓ Basic Statistics
- ✓ **Per base 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

✓ Per base sequence quality



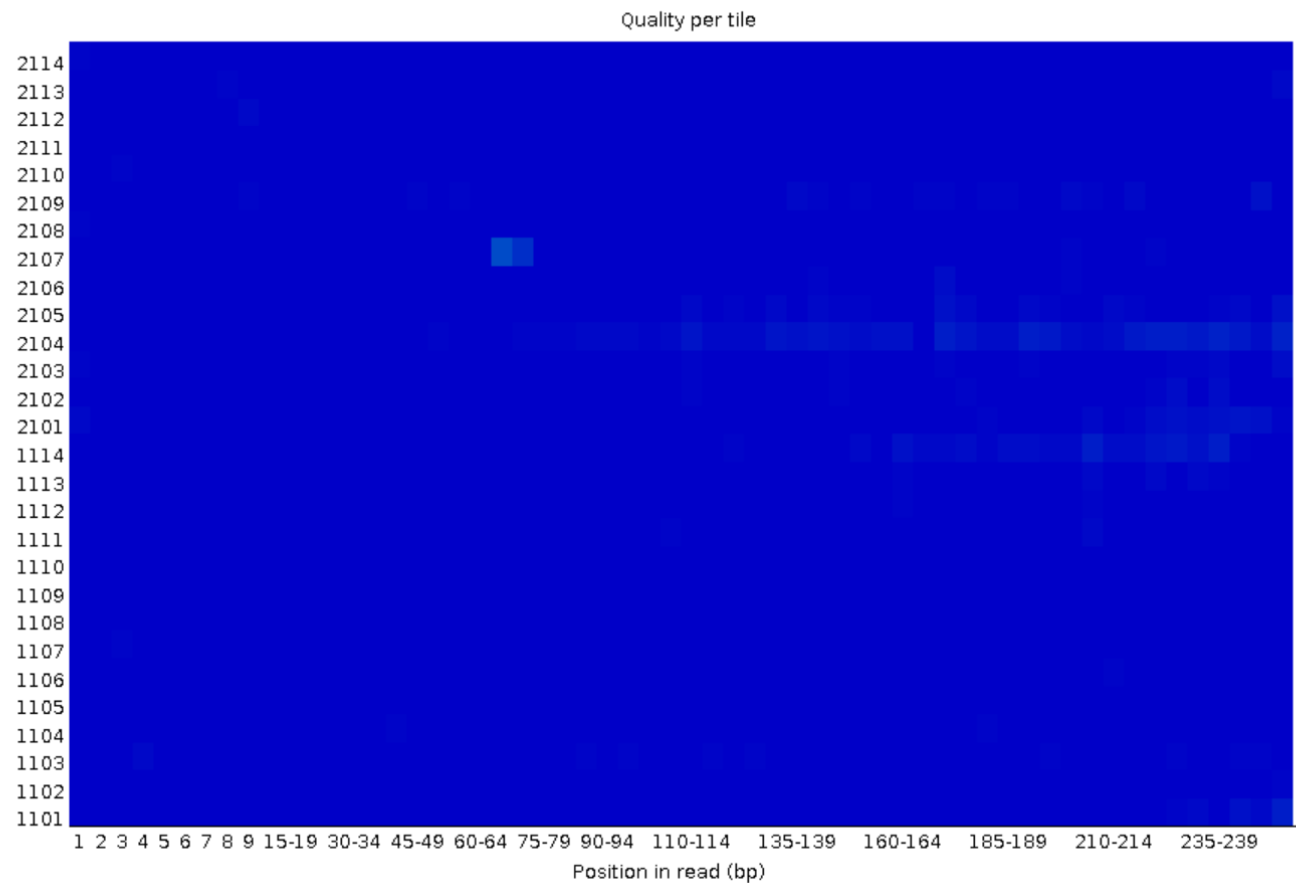
Low-quality
example!














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](#)

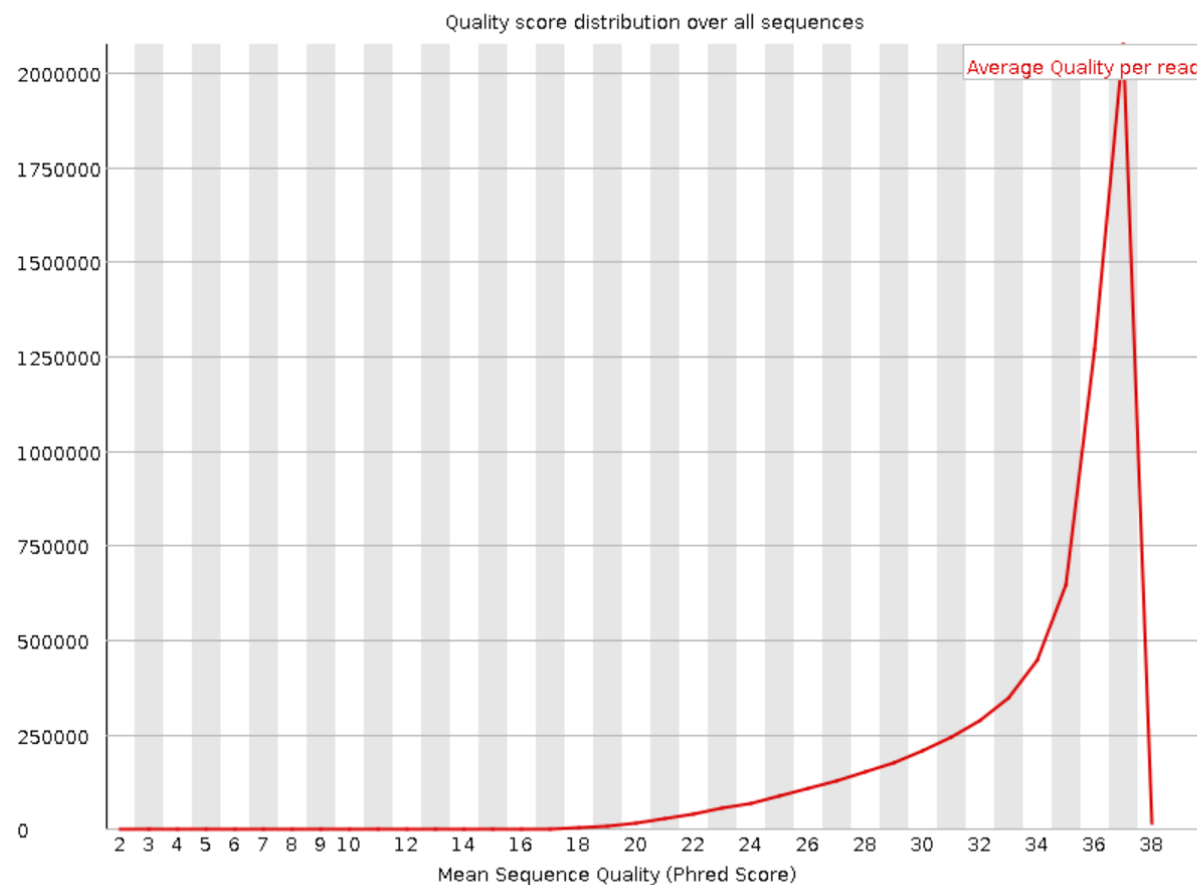
✓ Per tile sequence quality



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](#)

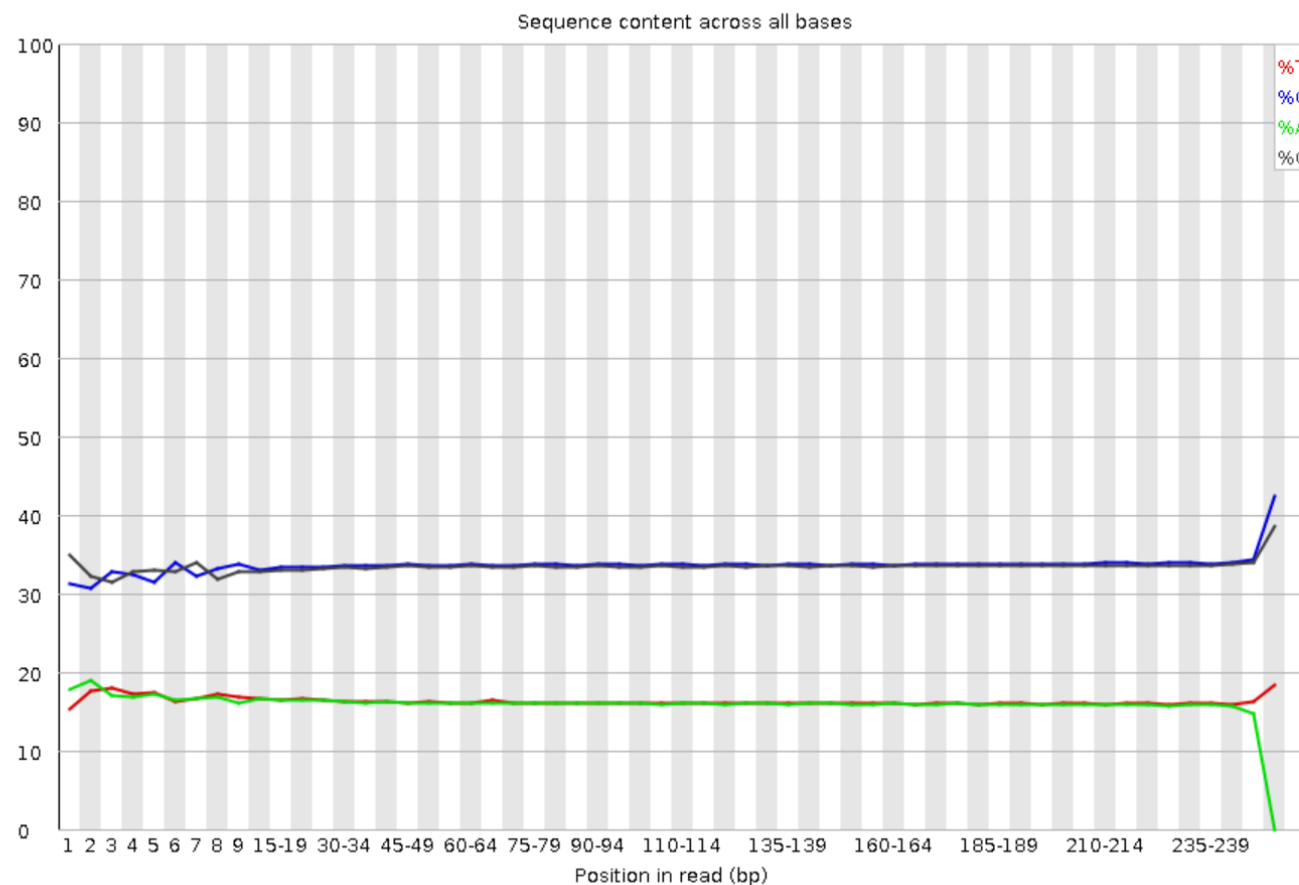
Per sequence quality scores













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](#)

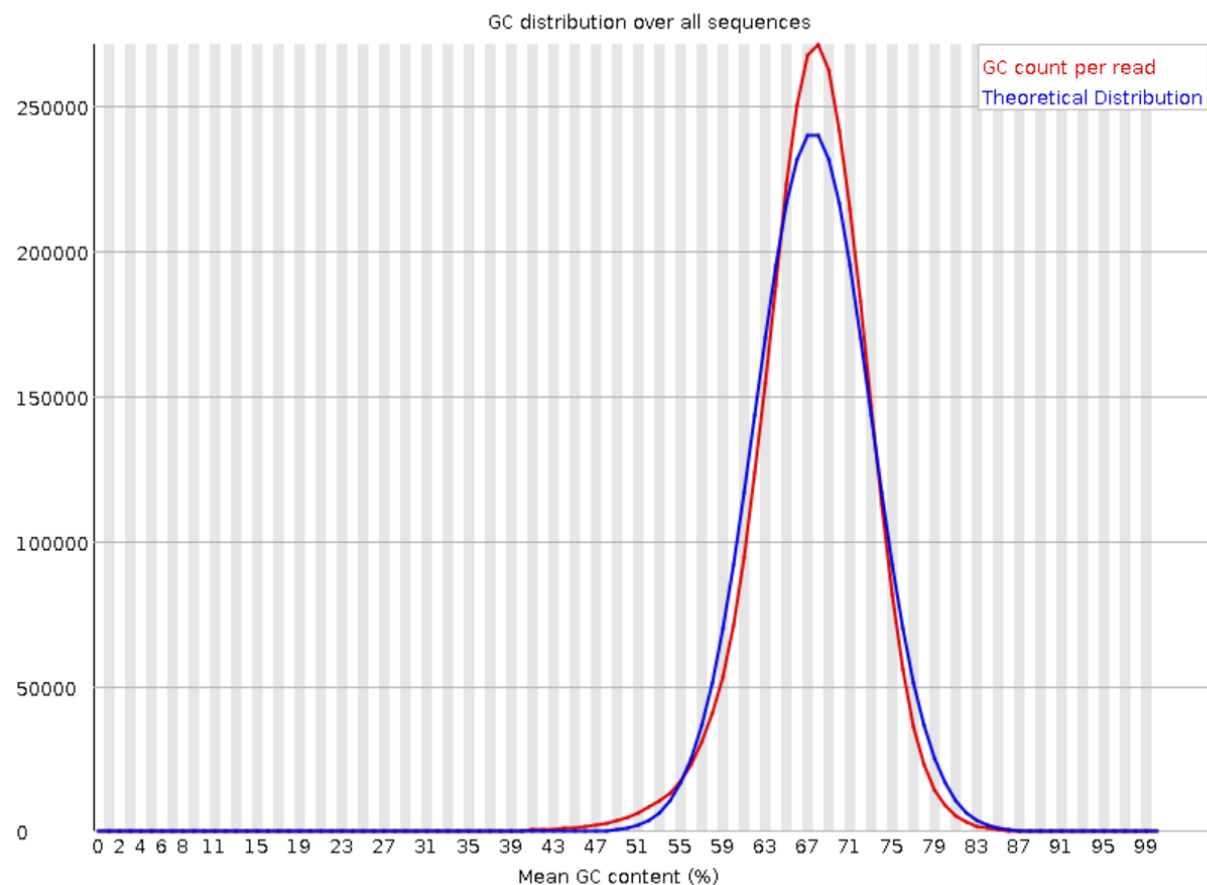
! Per base sequence content














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](#)

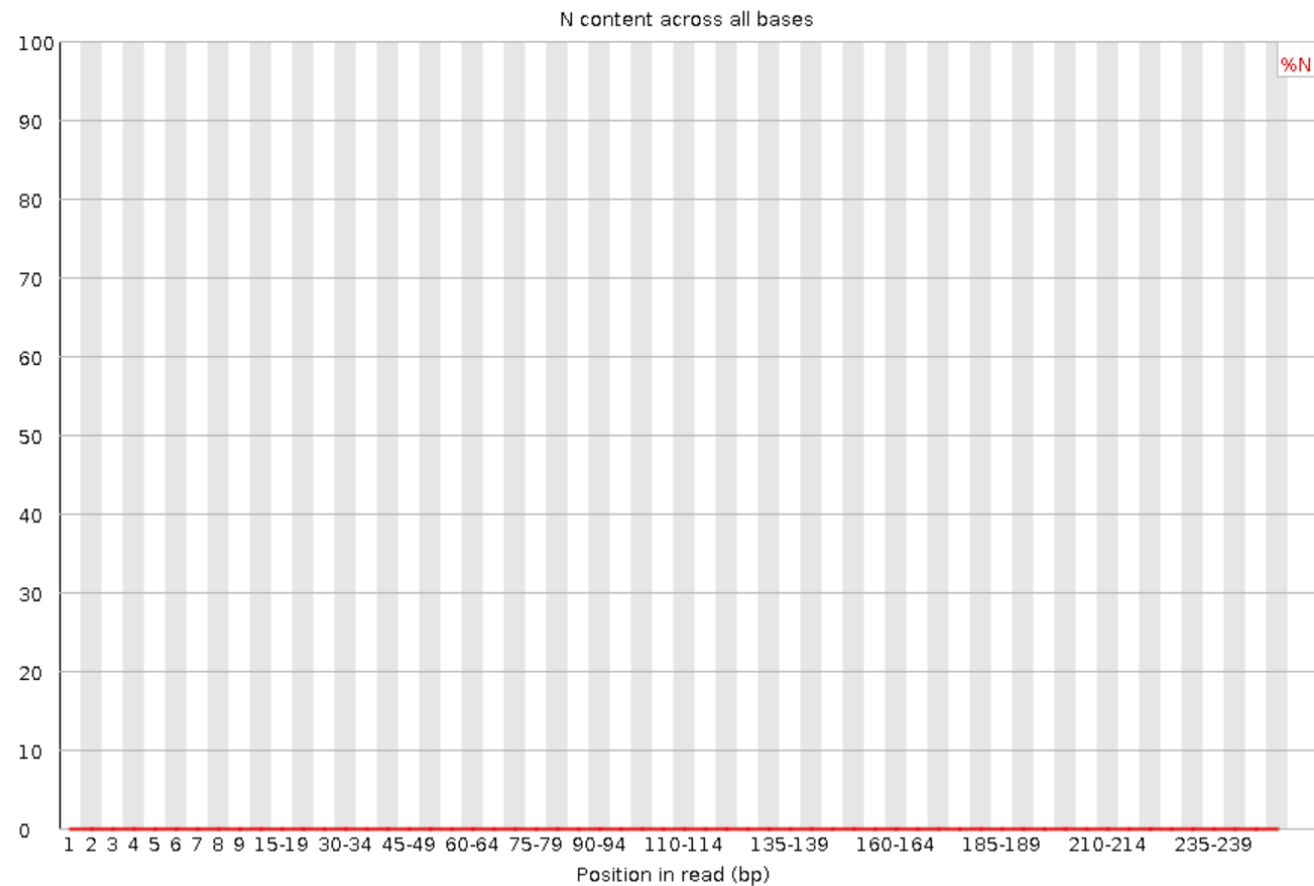
Per sequence GC content



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](#)

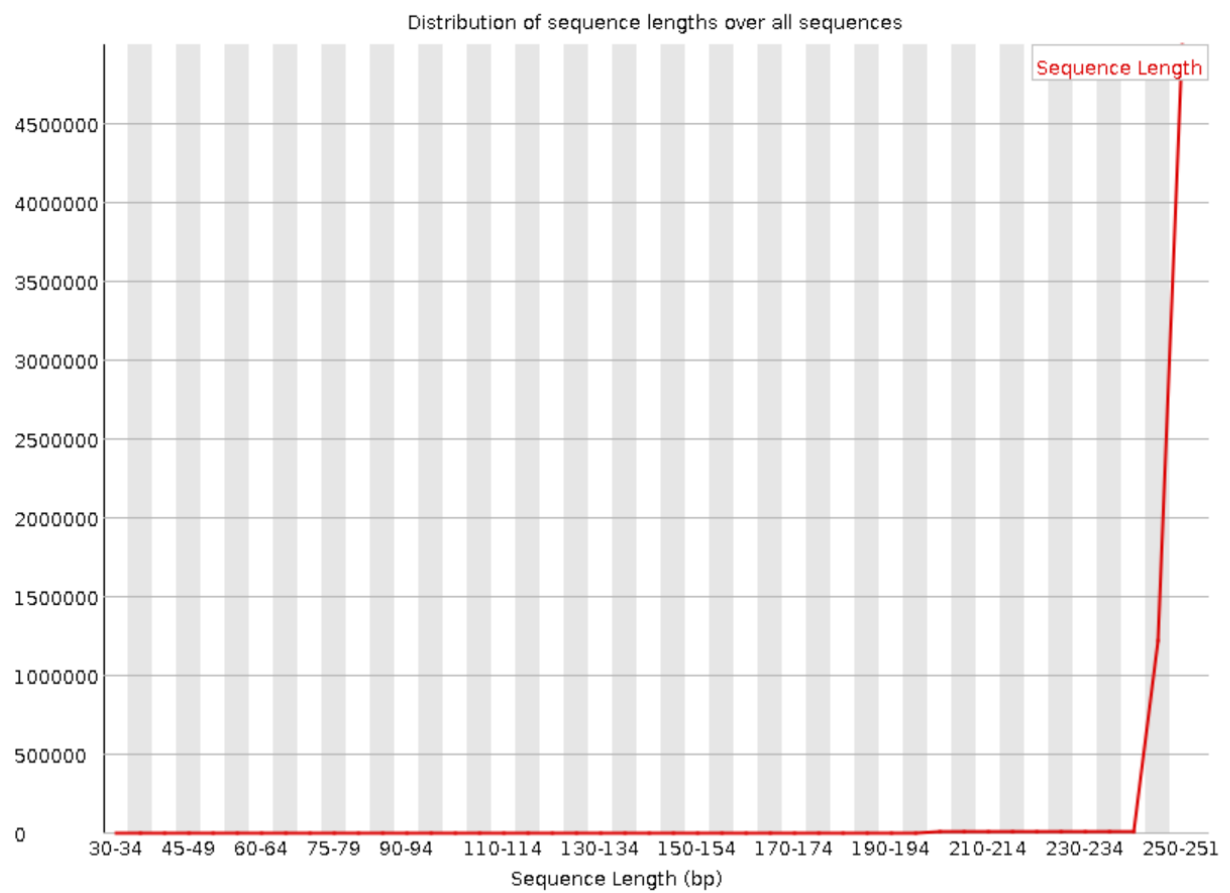
Per base N content














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](#)

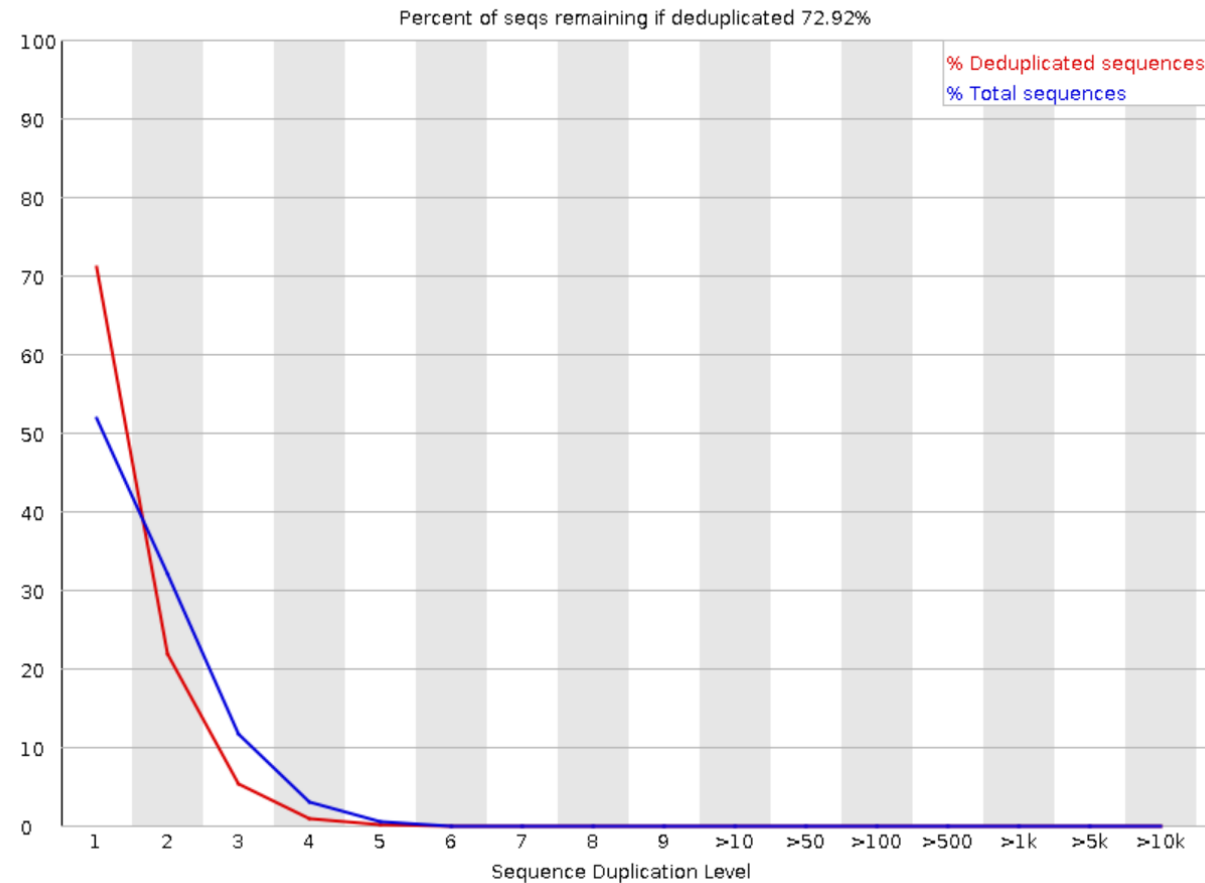
! Sequence Length Distribution



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](#)

Sequence Duplication Levels





Overrepresented sequences

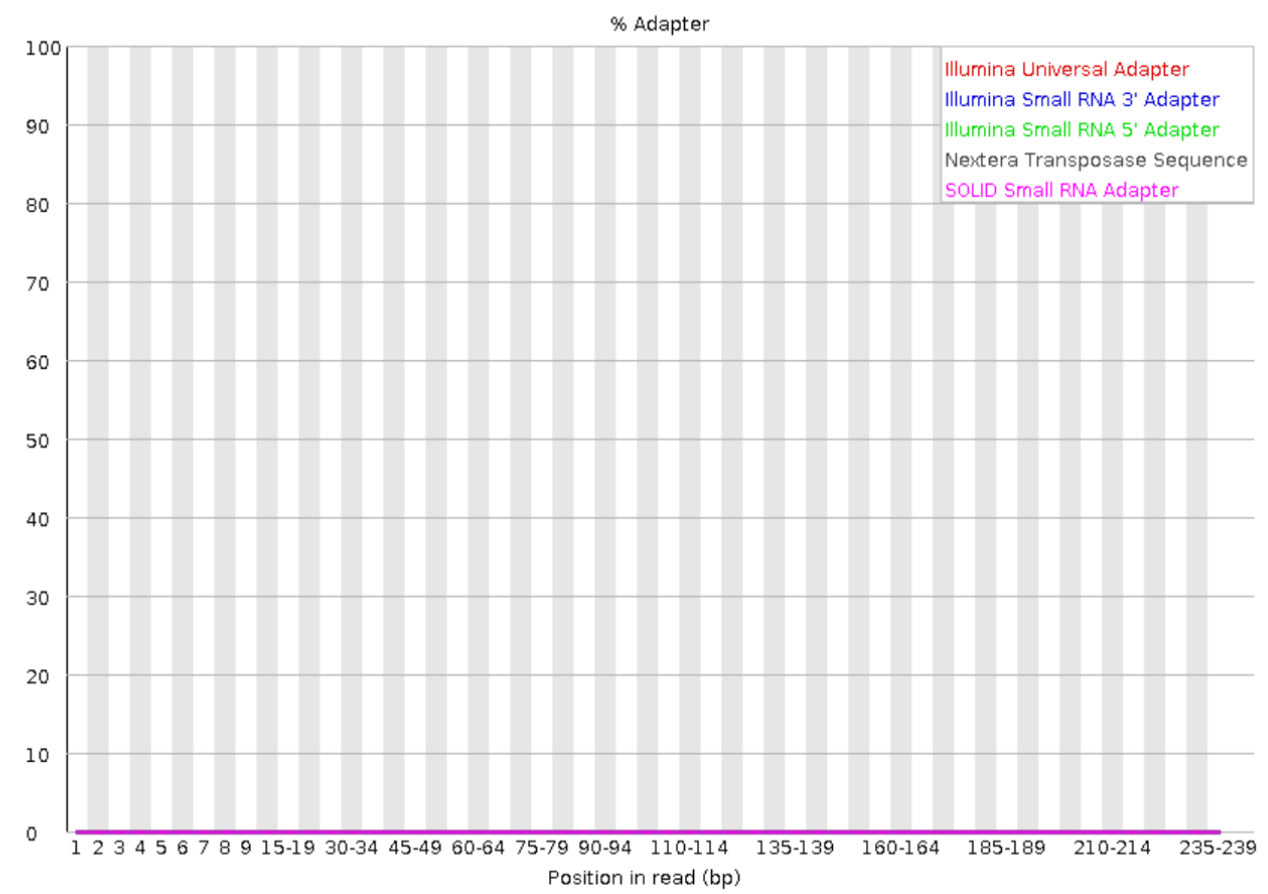
Sequence	Count	Percentage	Possible Source
CTGCTATGGCCACCAGACTCTCAGGCTCCATGCAGTGGCCAGCCTCATCG	2554	0.8349133703824779	No Hit
CAGCGGTCTAGTTTGAAGAACCTGACCCGAGTCTTGGTGACGAAGGCCAG	2463	0.8051650866296176	No Hit
GTTTGAAGAACCTGACCCGAGTCTTGGTGACGAAGGCCAGATTTGCGATC	1920	0.6276560967636483	No Hit
CCACAGGGTCCCAGGTCATGGGTACCGAGTCCAGGTCATAGTGCCGGATG	1219	0.39849624060150374	No Hit
GAAGAACCTGACCCGAGTCTTGGTGACGAAGGCCAGATTTGCGATCTTCA	1186	0.3877084014383786	No Hit
GGCAGGTGGACCCGGAGCCGCTGACAGAGGAGGTCAGCCCCTGAGTTGGA	1111	0.3631905851585486	No Hit
CACAGGGTCCCAGGTCATGGGTACCGAGTCCAGGTCATAGTGCCGGATGT	1079	0.35272965021248776	No Hit
GTCCCTGCTGCGGGCCACGACAGACCGTAGATCGAGCTGCGGCAGGTCGACCC	1036	0.3386727688787185	No Hit

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

No overrepresented sequences

✓ Adapter Content



How do I control FASTQ file quality?

fastp!

- Toolbox for filtering, & trimming your reads to control for quality
- Good all-around tool that can replace collections of previous tools such as trimmomatic, etc...
- You can play with it on the server or through <https://fastq.sandbox.bio/>