

# Tutorial 3 – More Scripting & FastQC

---

25 SEPTEMBER 2020

MICB405

# 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

# 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

# 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
- PuTTY users need to use **pscp**

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

# Where to find files on your computer after SCP

---

Windows WSL -> File Explorer (to get to your Linux partition directory)

- `\\wsl$\Ubuntu\home\<your Unix username>`

Windows Git Bash -> File Explorer

- `C:\Users\<your Windows username>`

macOS -> Finder

- `/Users/<your macOS username>`

PuTTY

# Demonstration



# 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

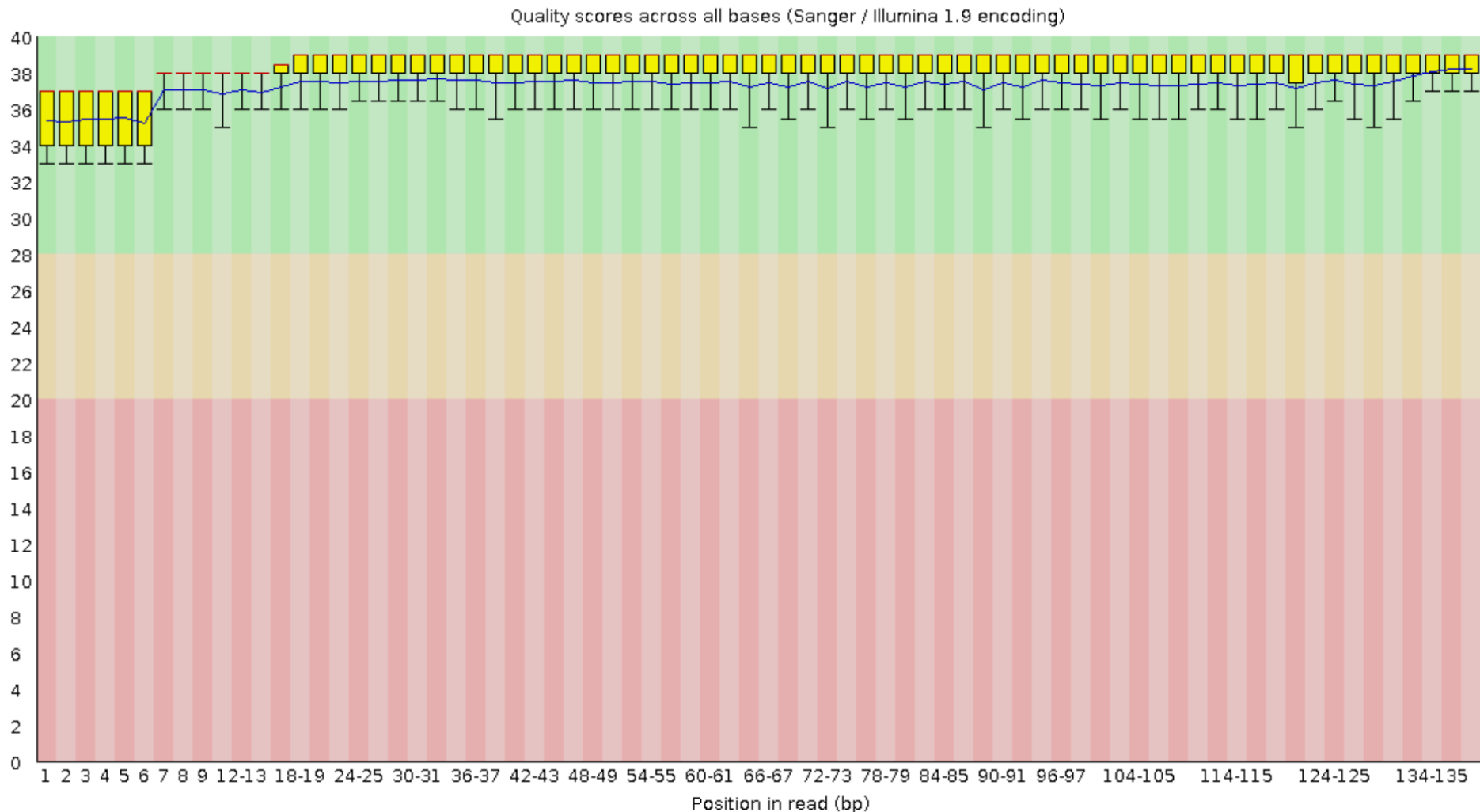
## ✓ Per base sequence quality



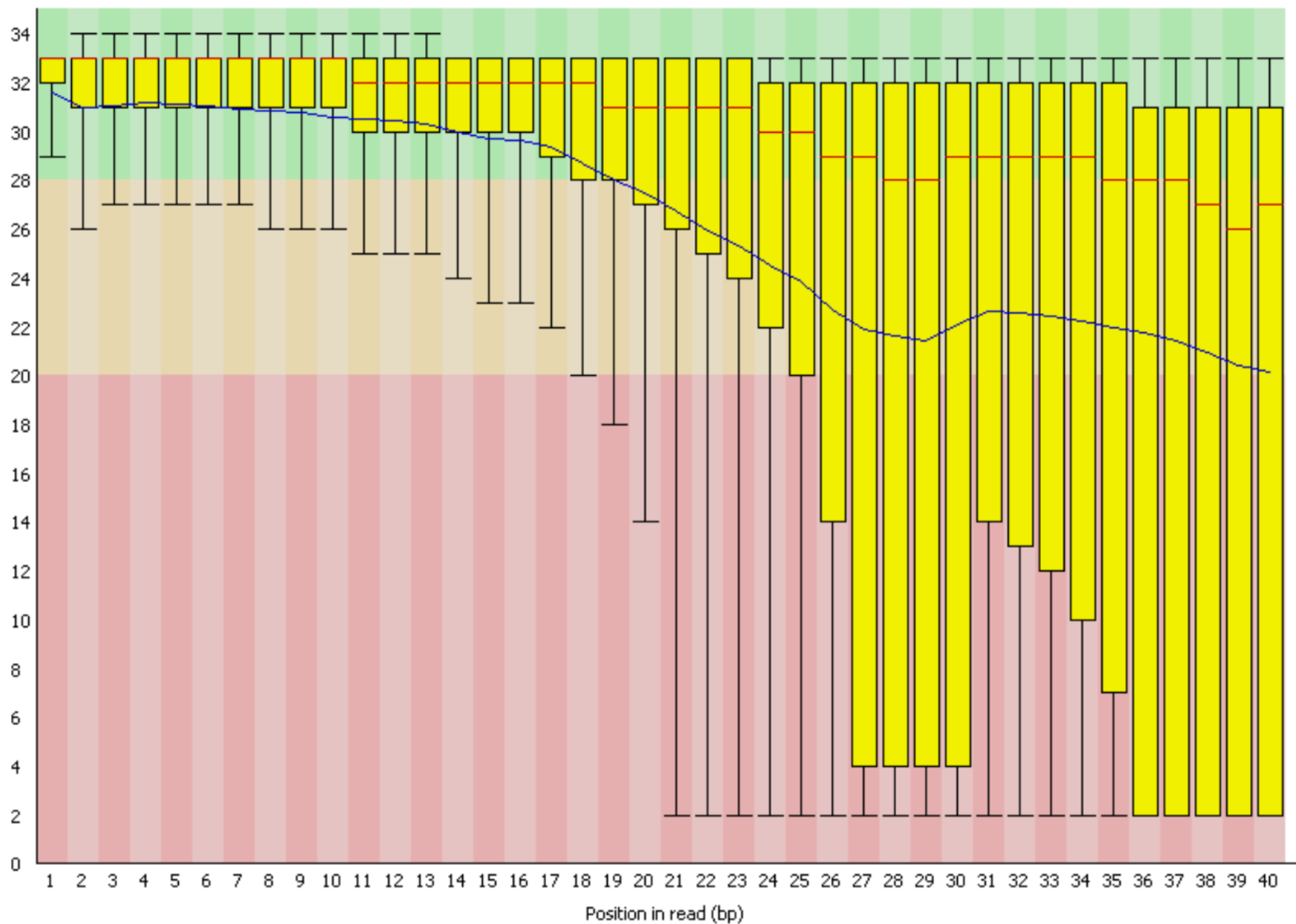
## Summary

- ✓ 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



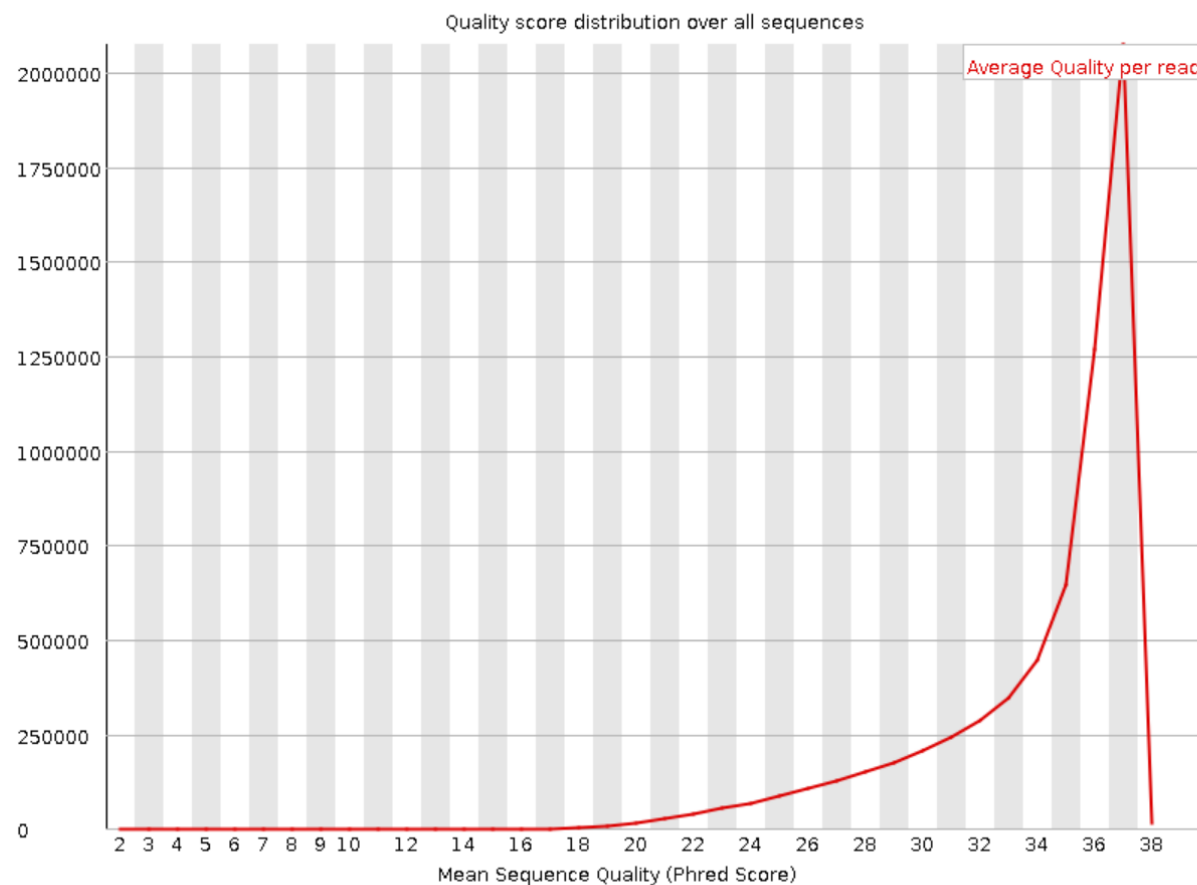
Low-quality  
example!



## 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](#)

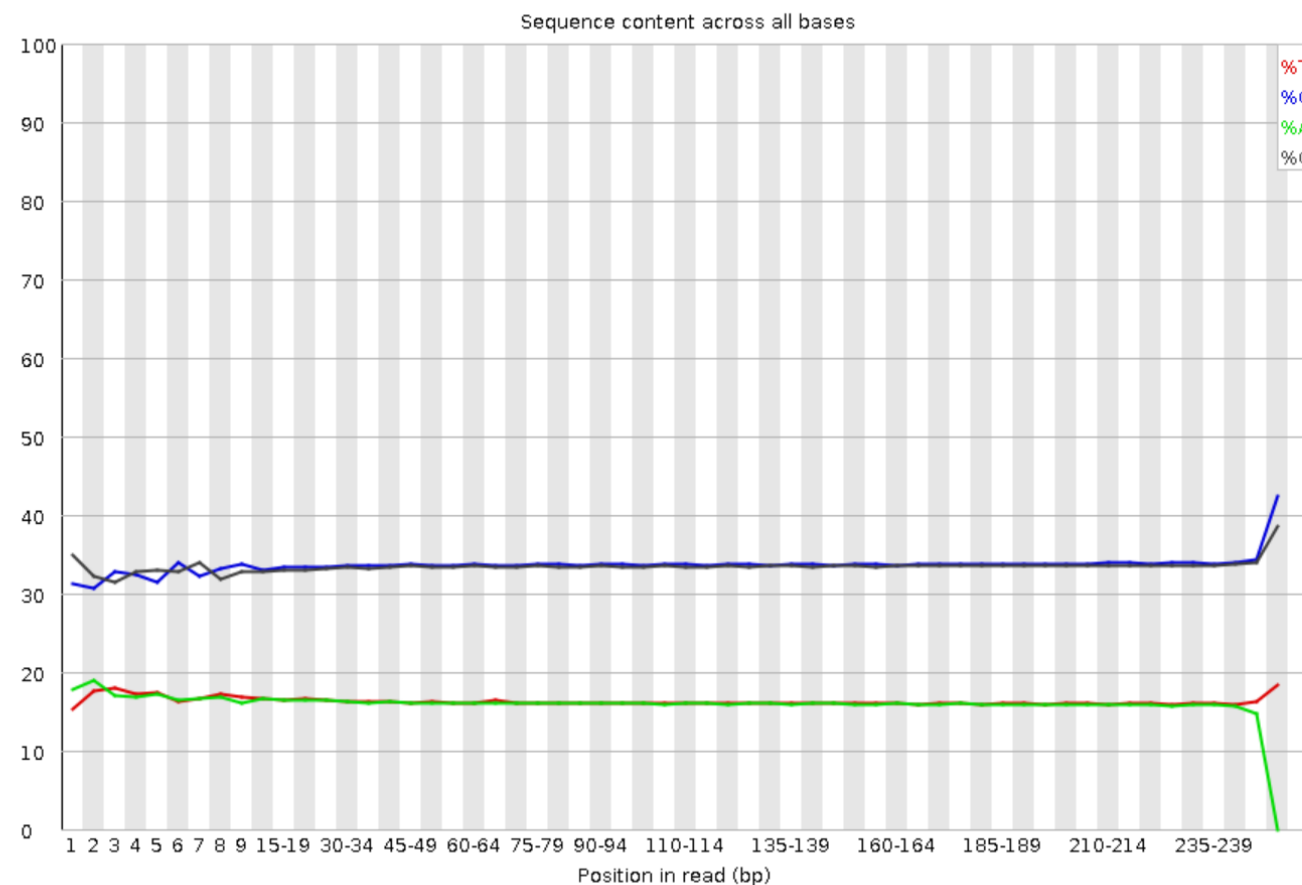
## Per sequence quality scores













## 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](#)

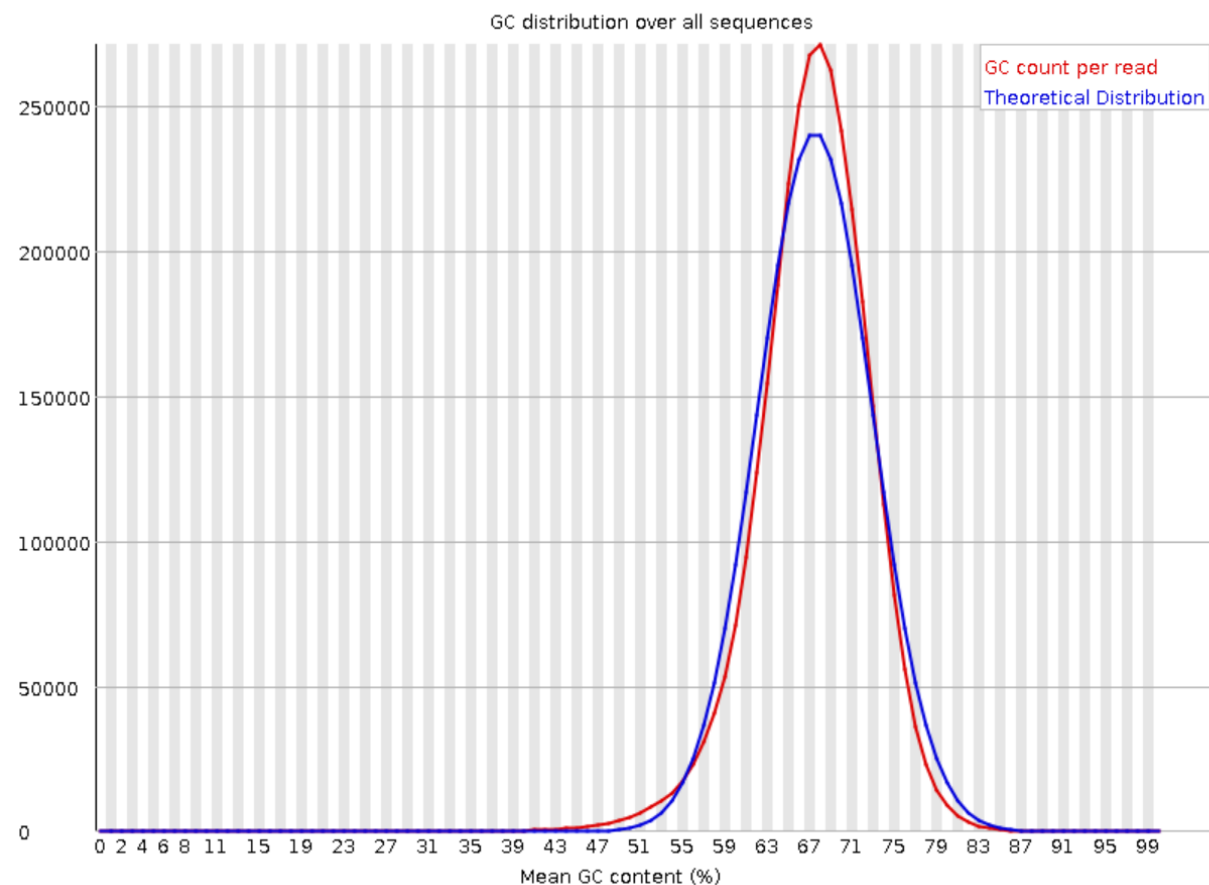
## ! Per base sequence content



## 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](#)

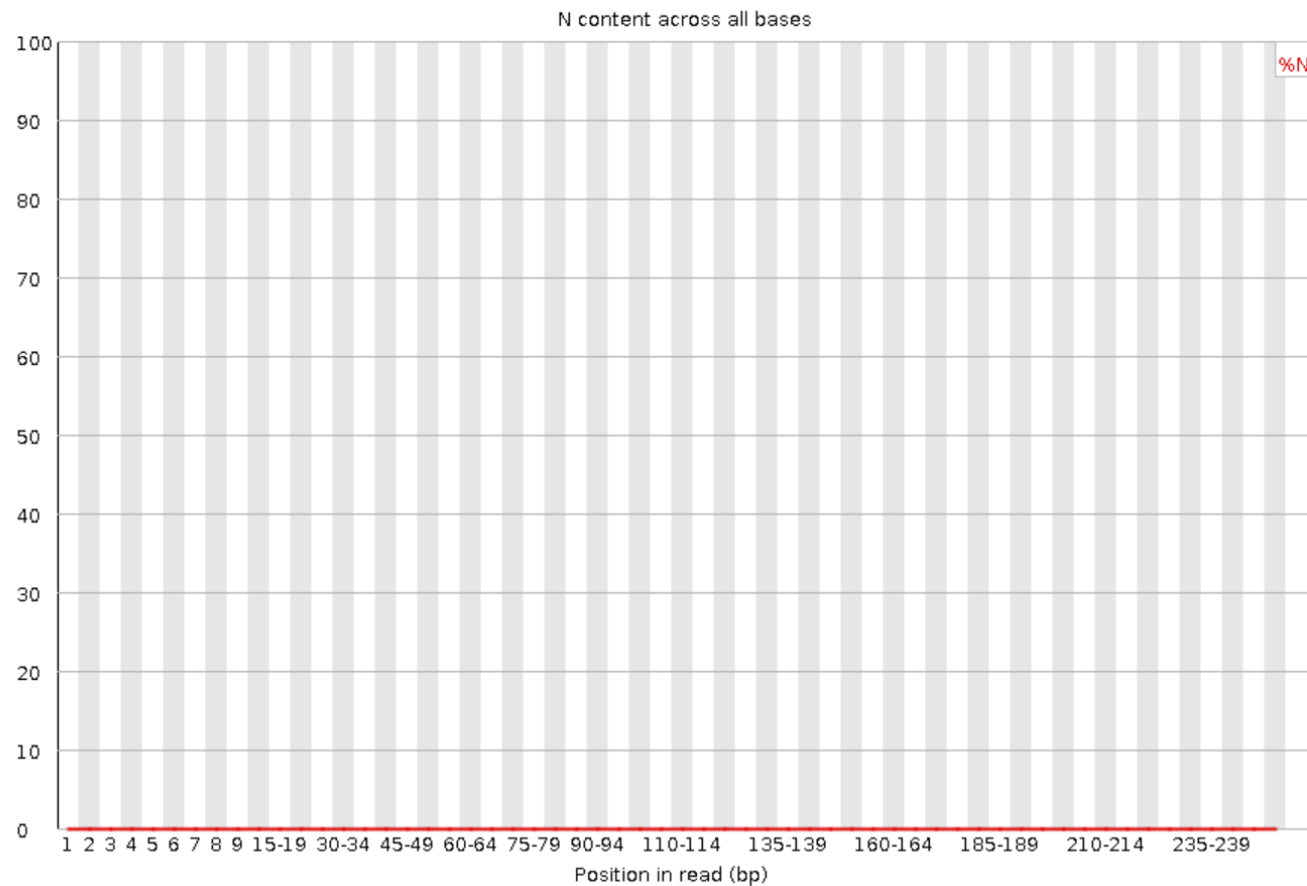
## Per sequence GC content



## 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

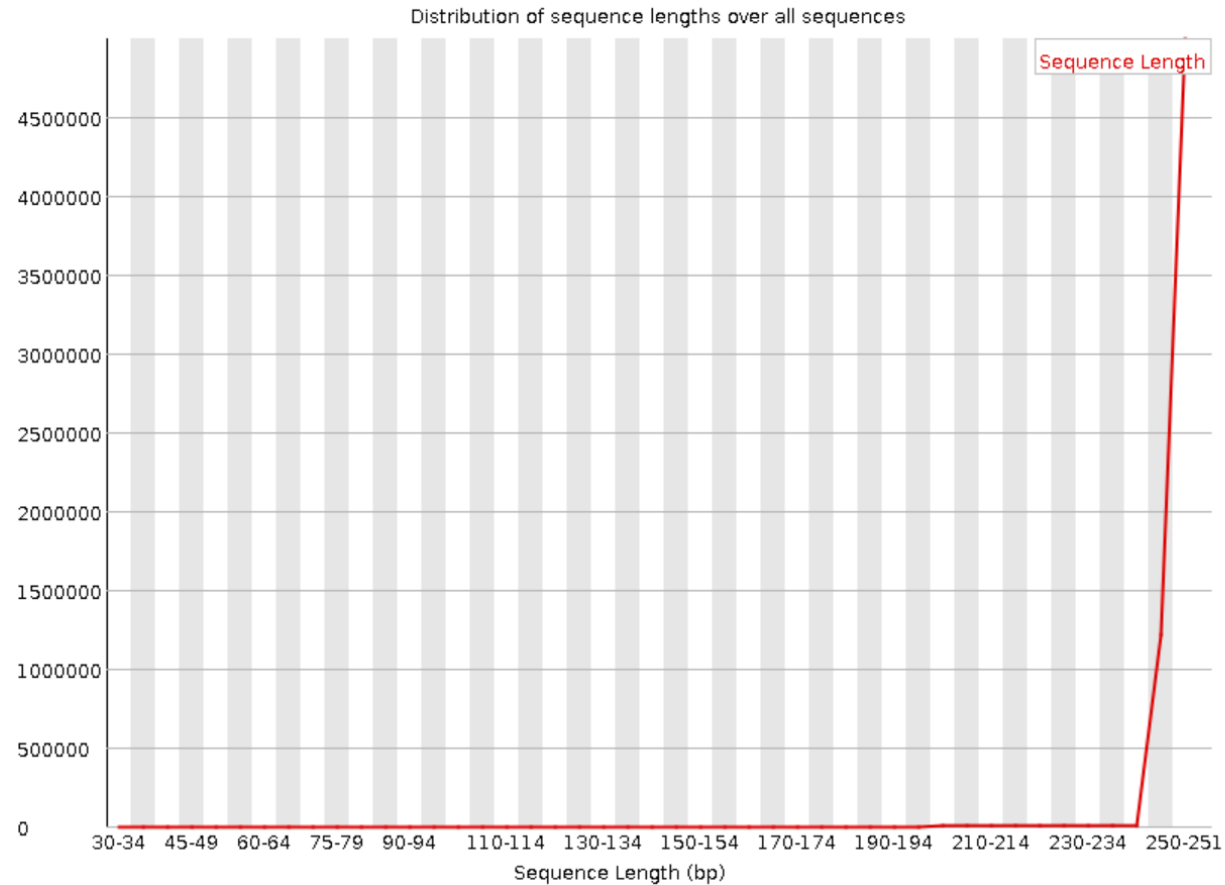




## Summary

-  [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](#)

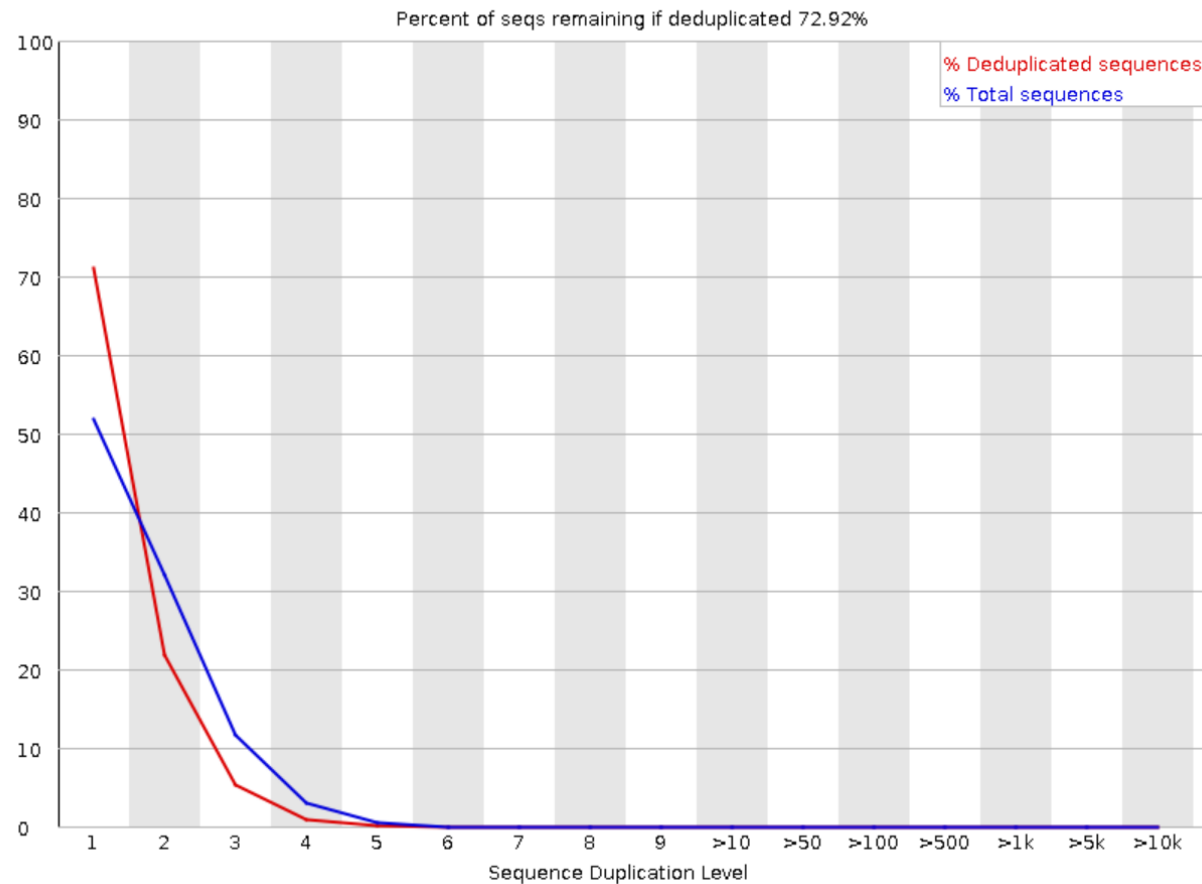
## Sequence Length Distribution



## 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
GTCCCTGCTGCGGGCCACGACAGCCGTAGATCGAGCTGCGGCAGGTCGACCC	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

