Quality Control

Requirements

Before diving into this slide deck, we recommend you to have a look at:

- Galaxy introduction

? Questions

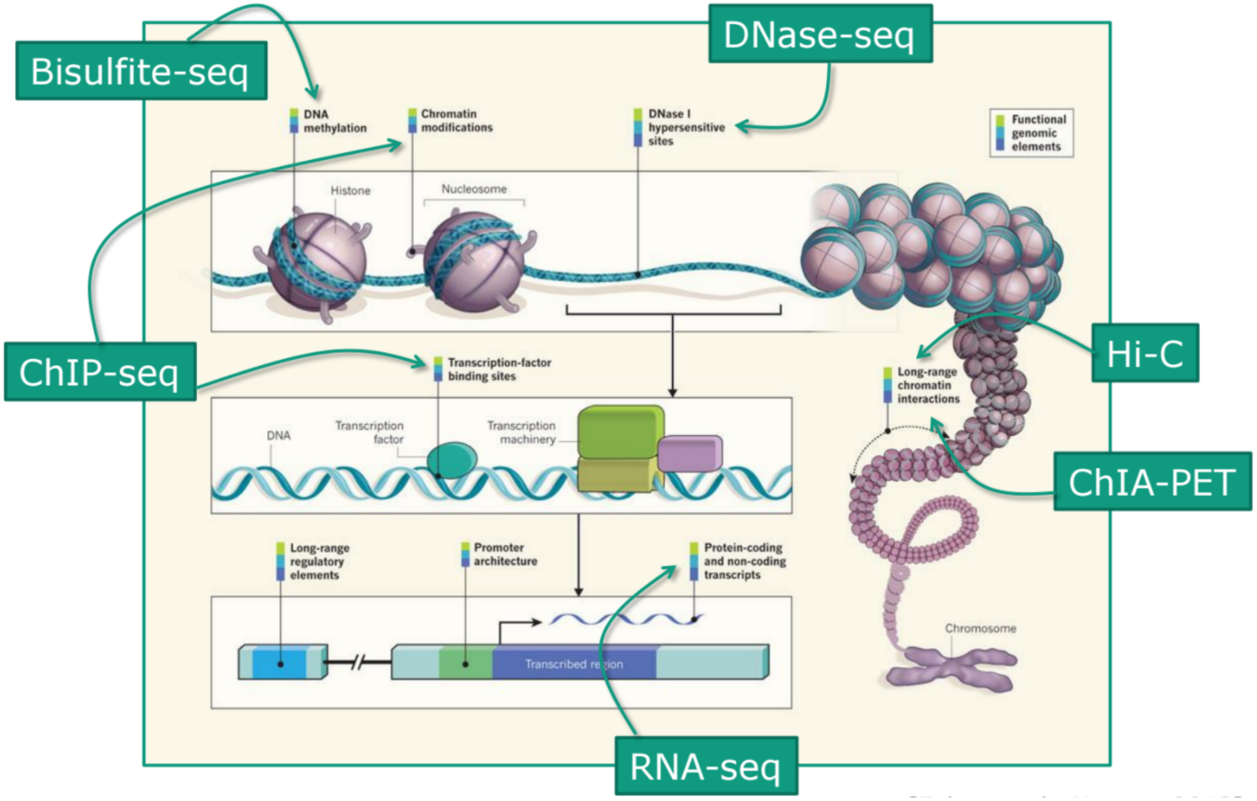
- How to control quality of NGS data?
- What are the quality parameters to check for each dataset?
- How to improve the quality of a sequence dataset?

🎯 Objectives

- Manipulate FastQ files
- Control quality from a FastQ file
- Use FastQC tool
- Understand FastQC output
- Use tools for quality correction

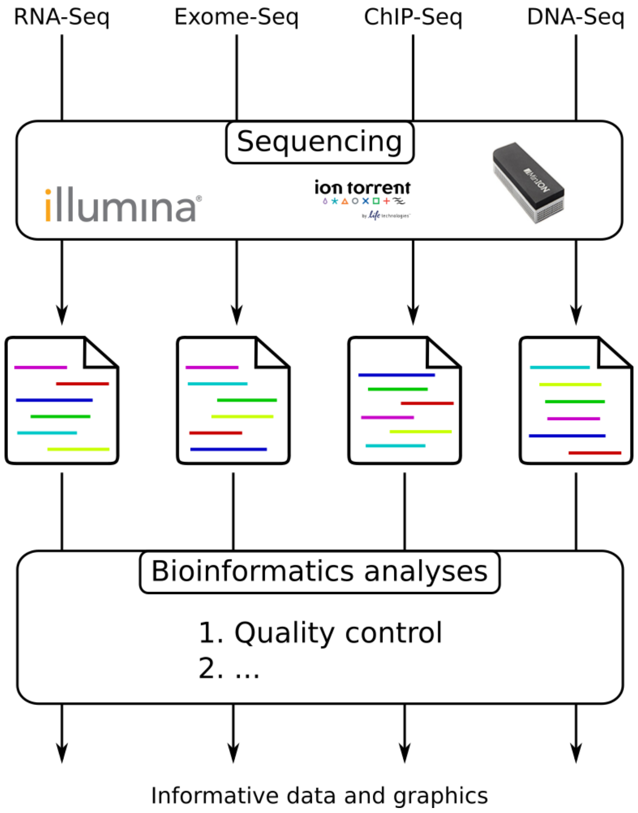
Why Quality Control?

Where is my data coming from?



Ecker et al, Nature, 2012

From experiments to data



RNA-Seq

Exome-Seq

ChIP-Seq

DNA-Seq

Sequencing

Bioinformatics analyses

1. Quality control
2. ...

Informative data and graphics

Quality control = First step of the bioinformatics analyses

My sequences?
Fasta

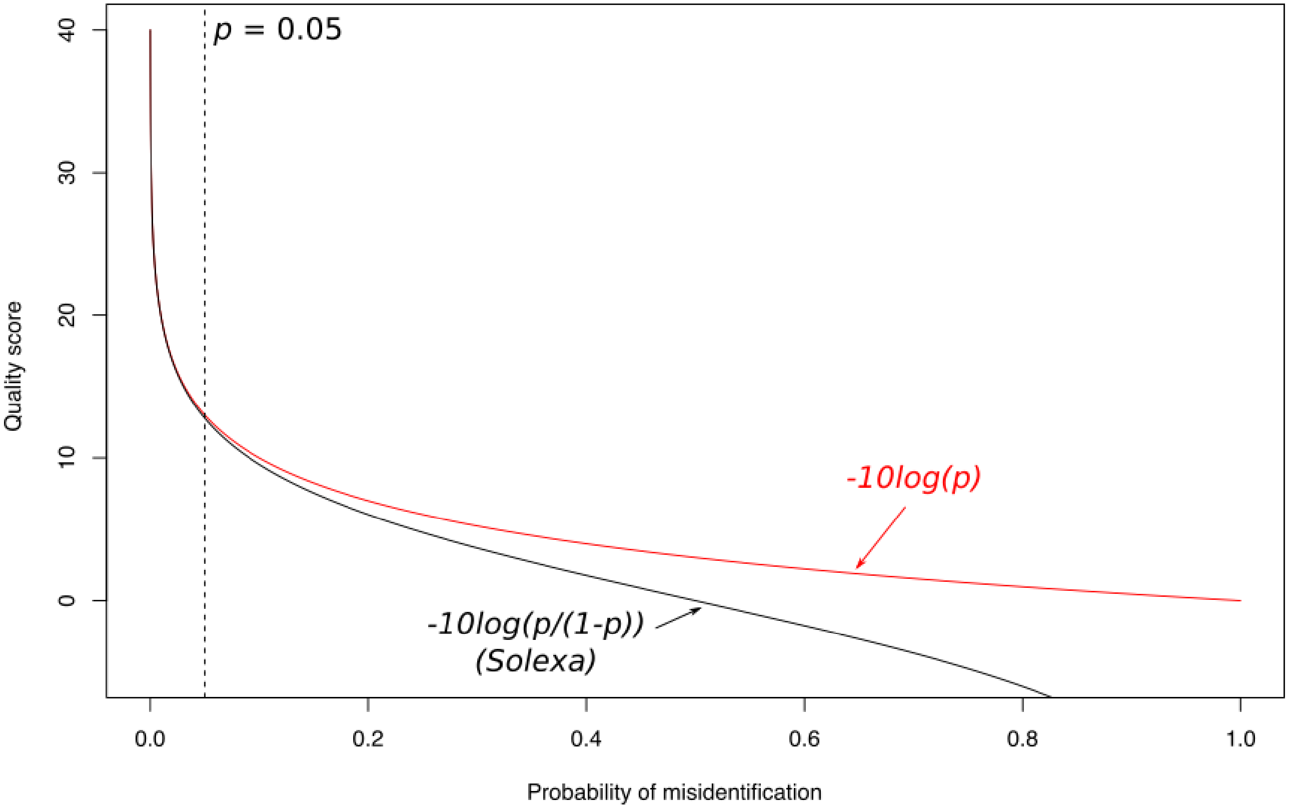
```
> Identifier1 (comment)
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
> Identifier2 (comment)
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XX
```

Quality score

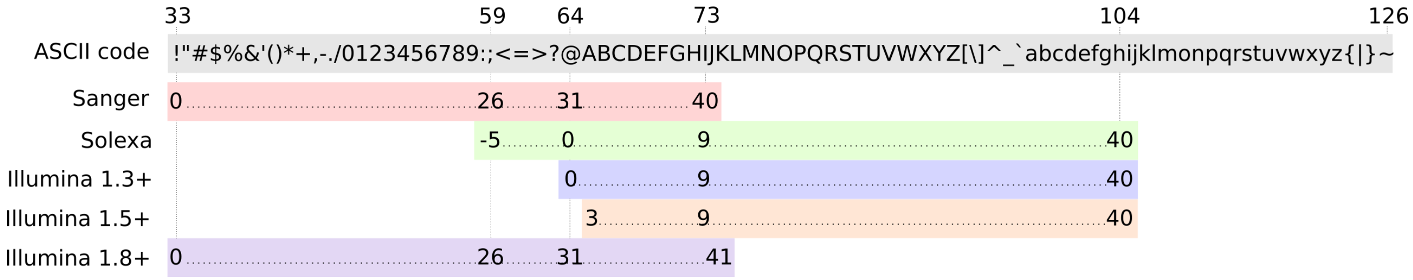
Measure of the quality of the identification of the nucleobases generated by automated DNA sequencing

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 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

Quality score



Quality score encoding



My sequences?

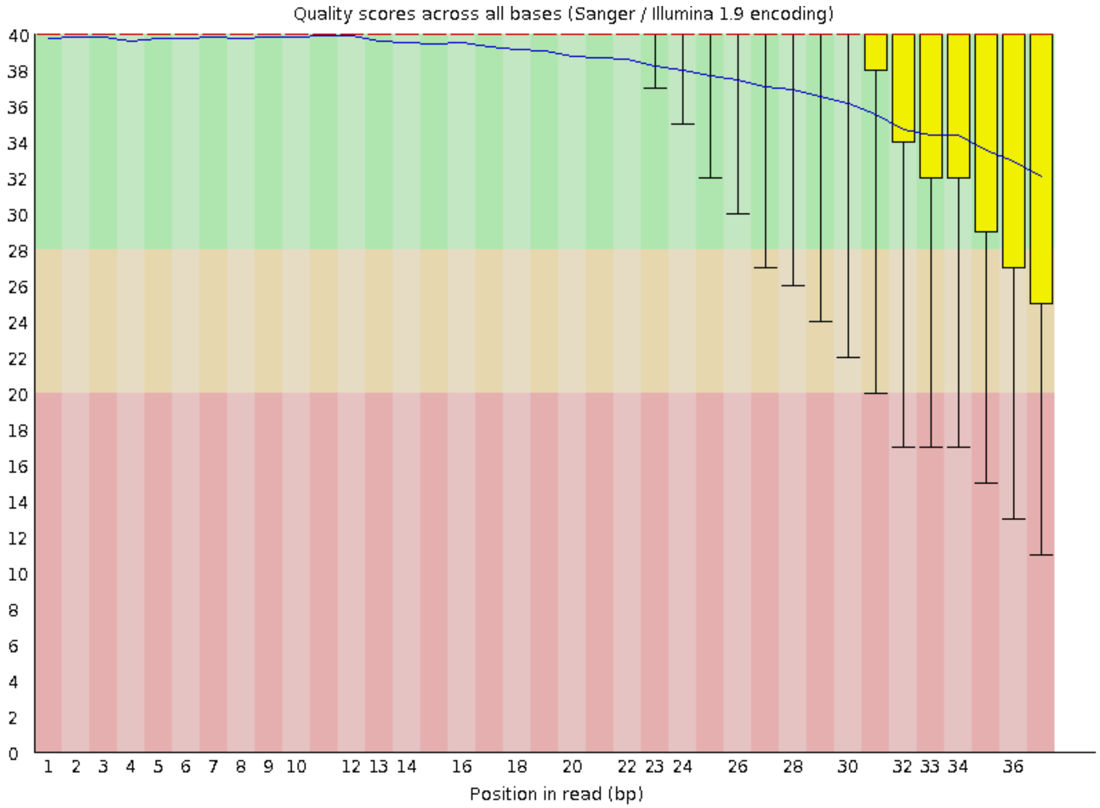
FastQ

```
@ Identifier1 (comment)
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
+
QQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQ
@ Identifier2 (comment)
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
+
QQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQ
```

How to check the quality of my sequences?

Quality score

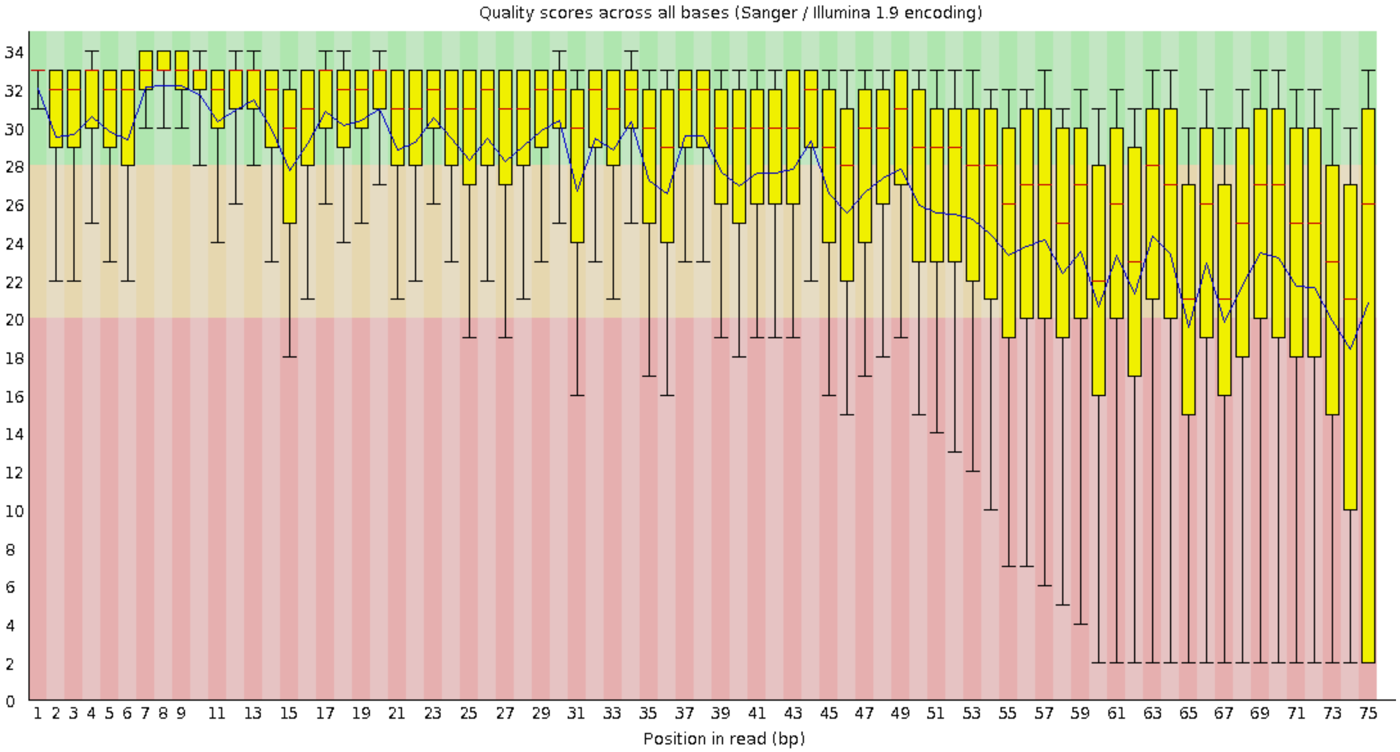
Per-base sequence quality



👍 Good quality score

Quality score

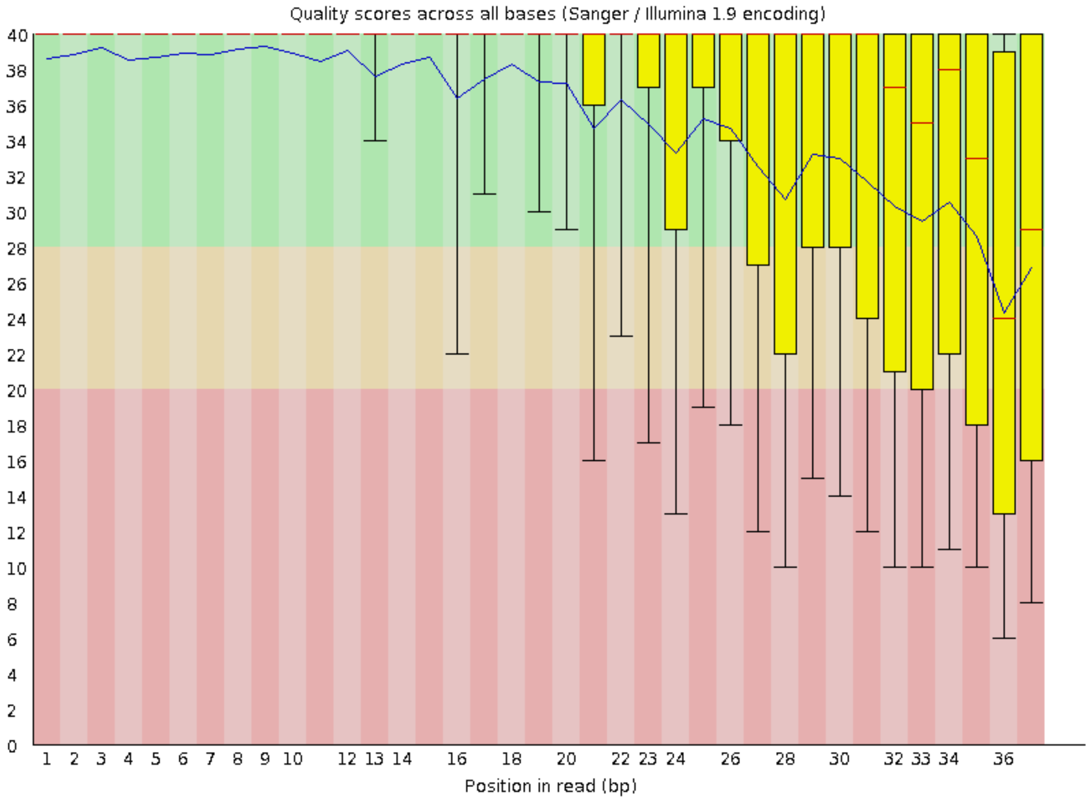
Per-base sequence quality



👎 Bad quality score

Quality score

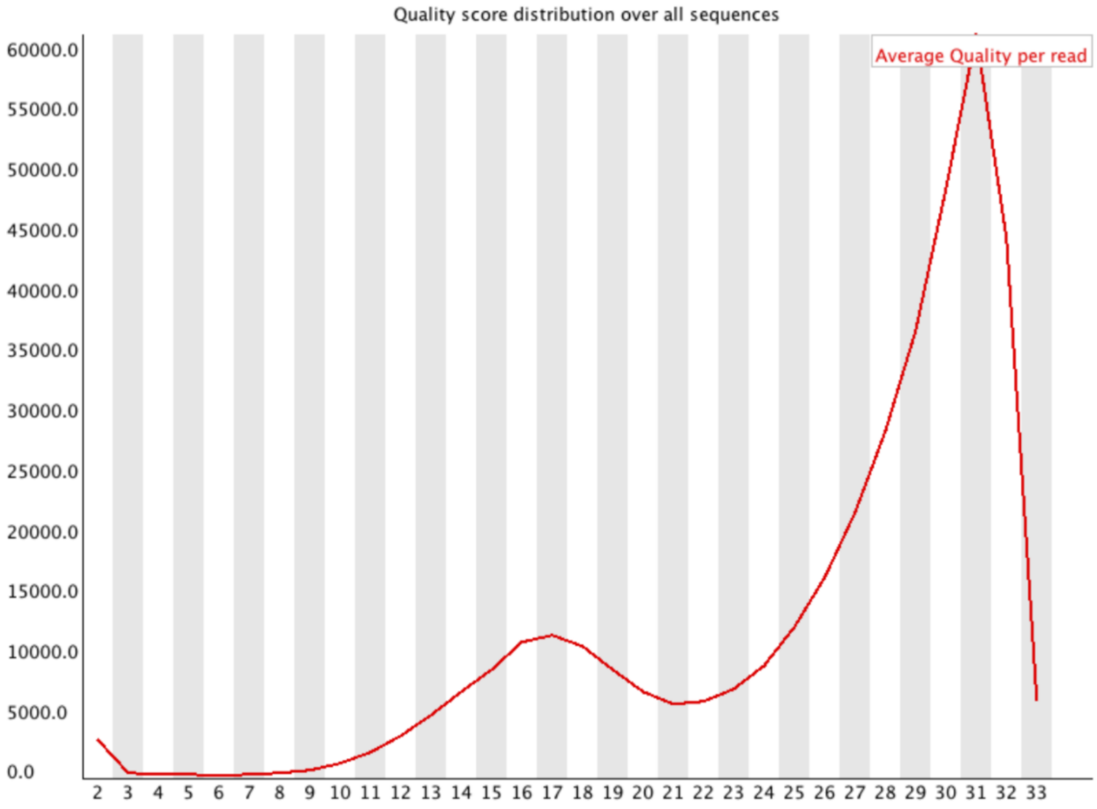
Per-base sequence quality



👍 👎 Intermediate quality score

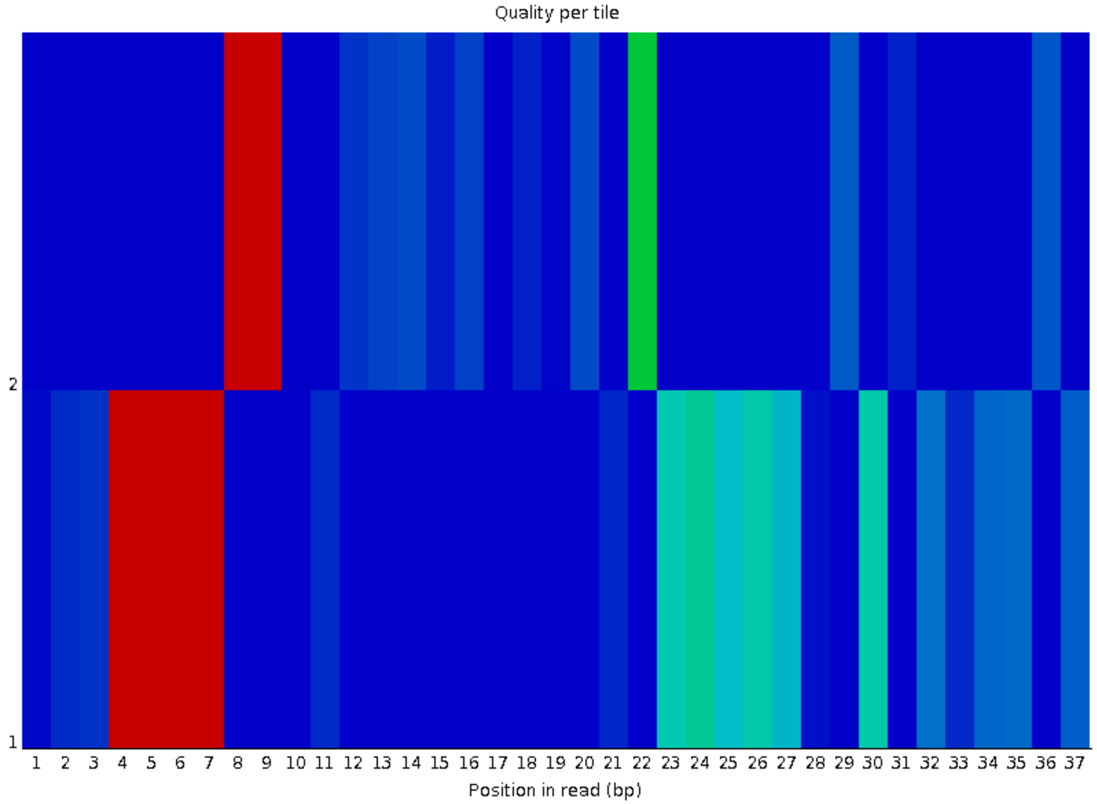
Quality score

Per-sequence quality scores



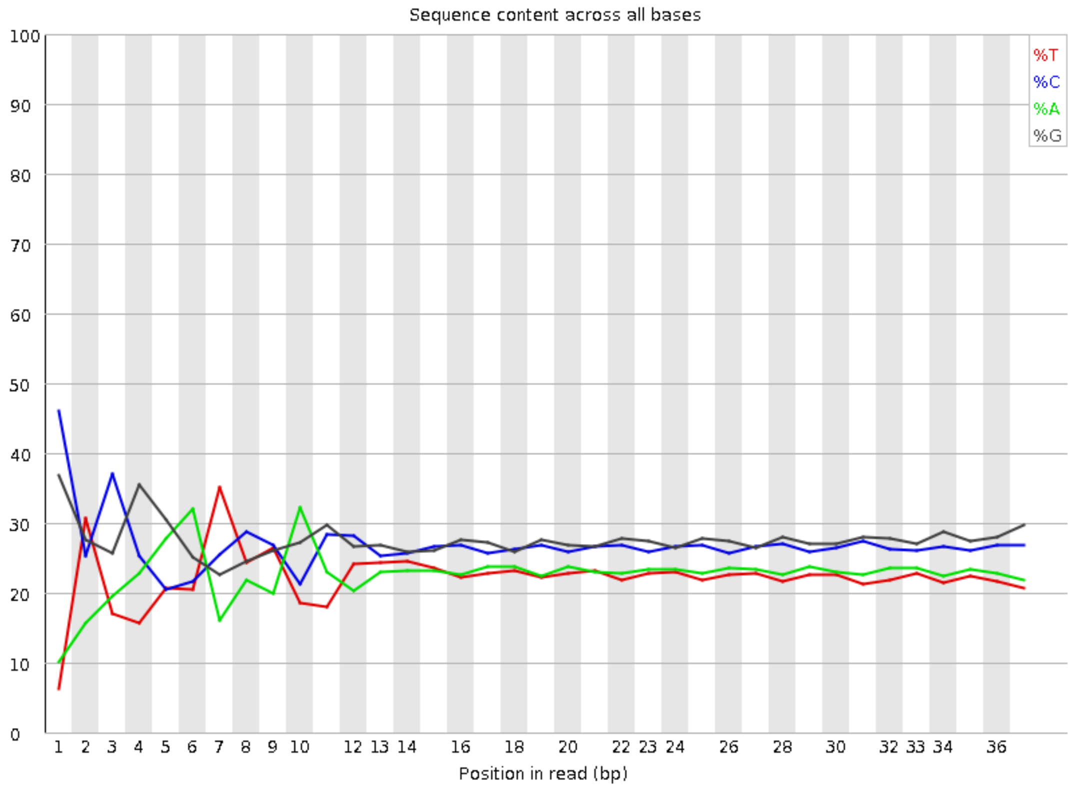
Quality score

Per-tile sequence quality



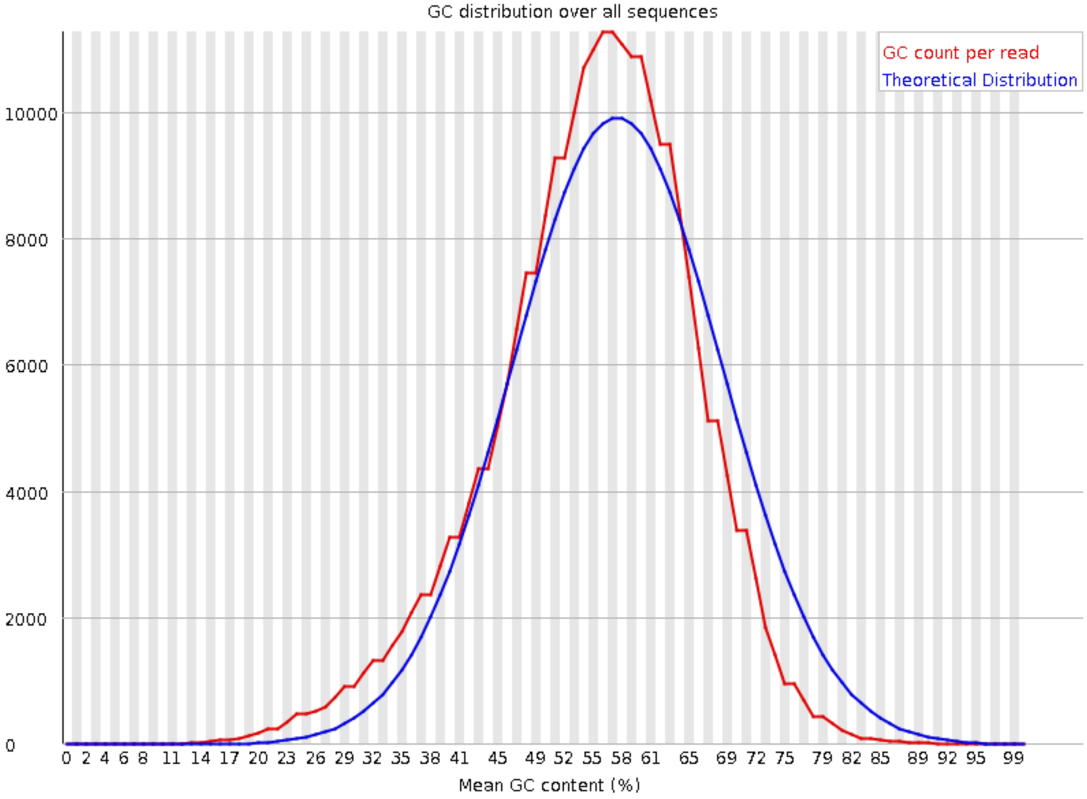
Also to check: Sequence content

Per-base sequence content



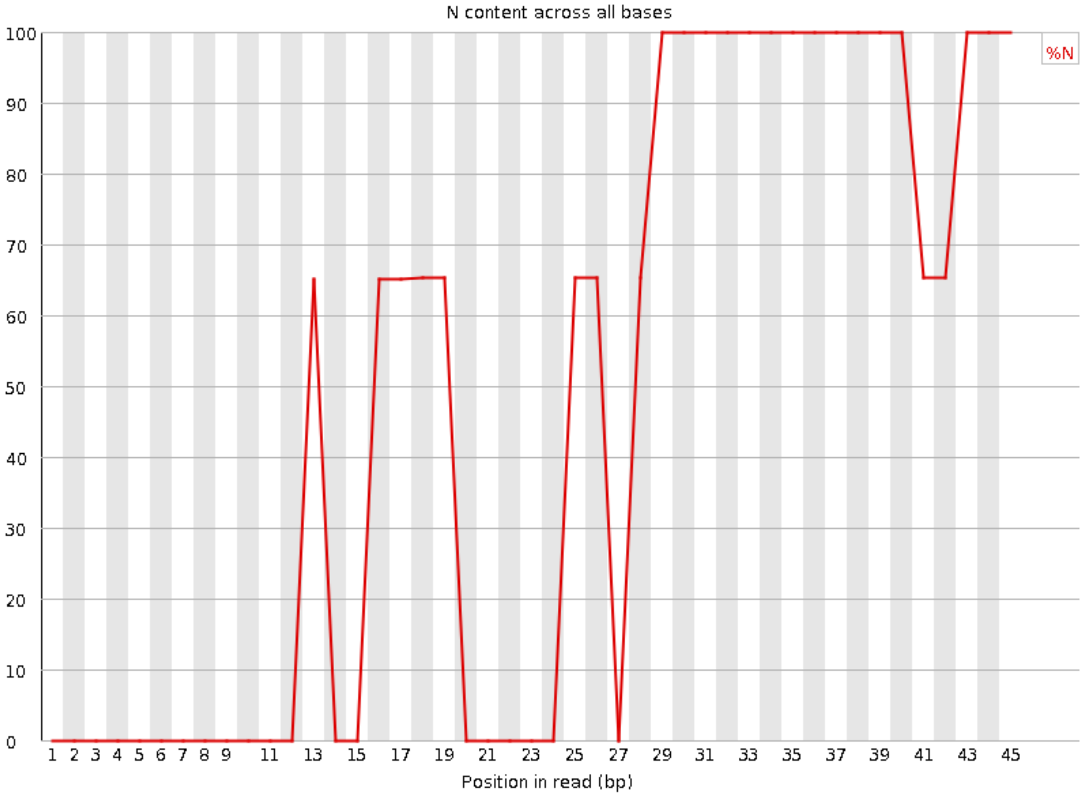
Also to check: Sequence content

Per-sequence GC content



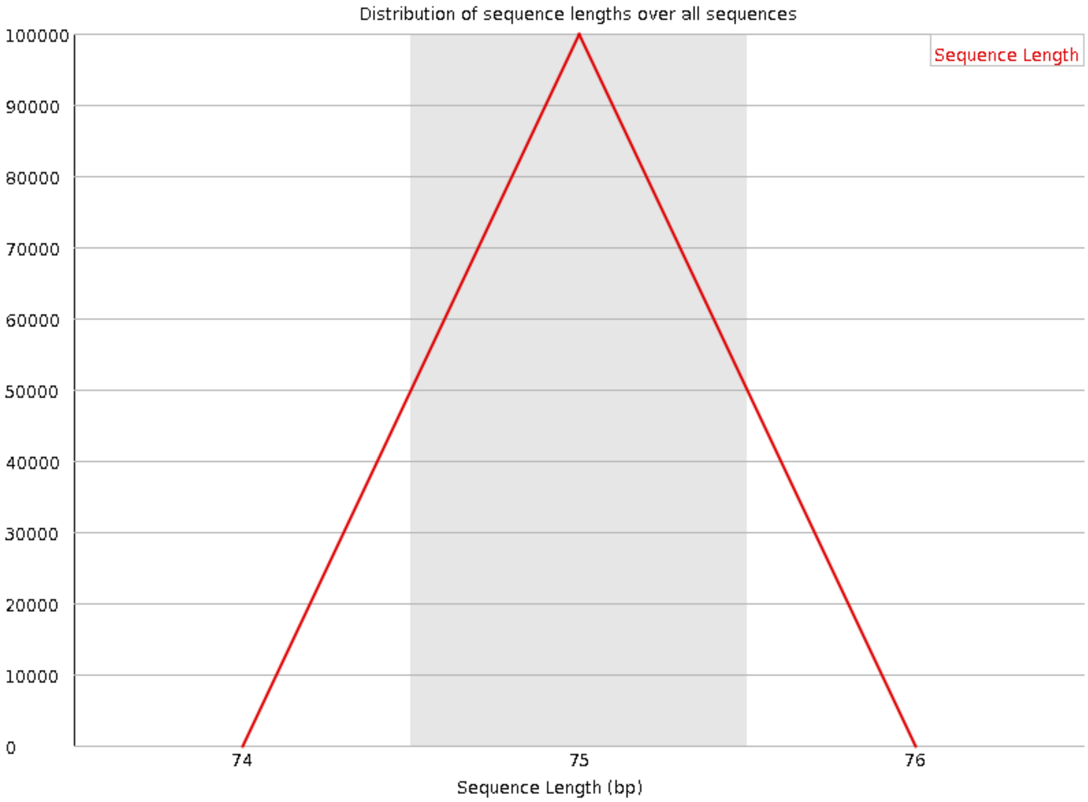
Also to check: Sequence content

Per-base N content

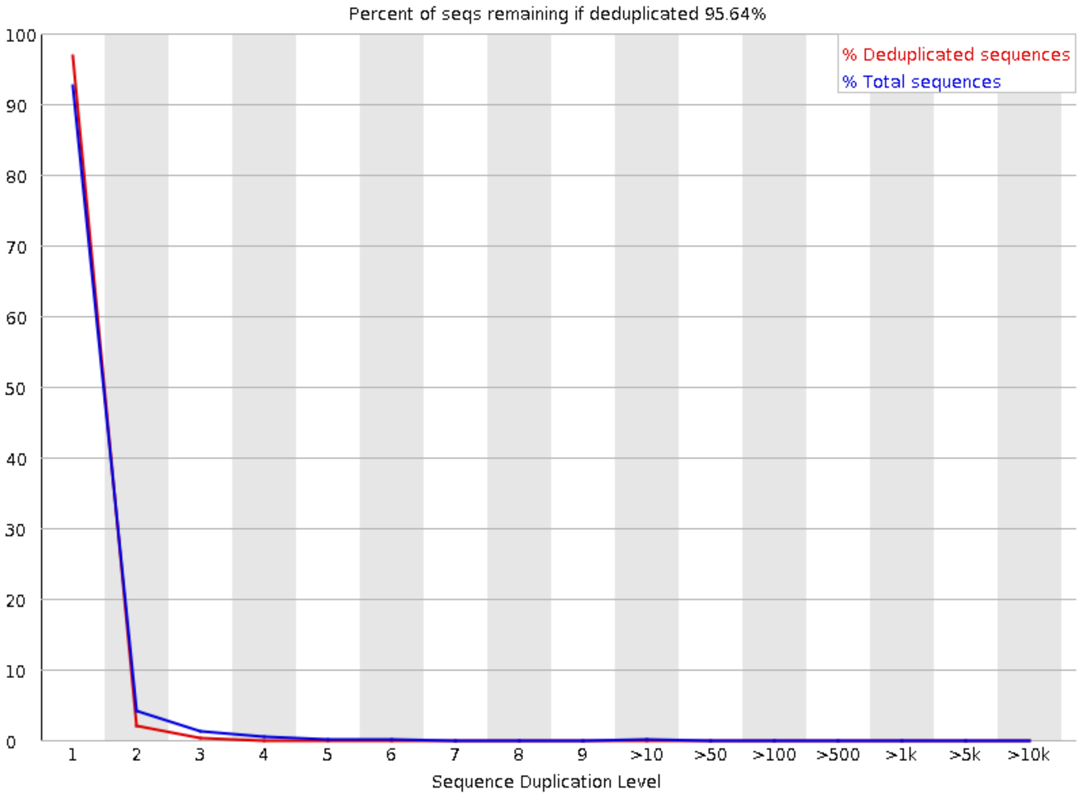


Also to check: Sequence length

Sequence length distribution

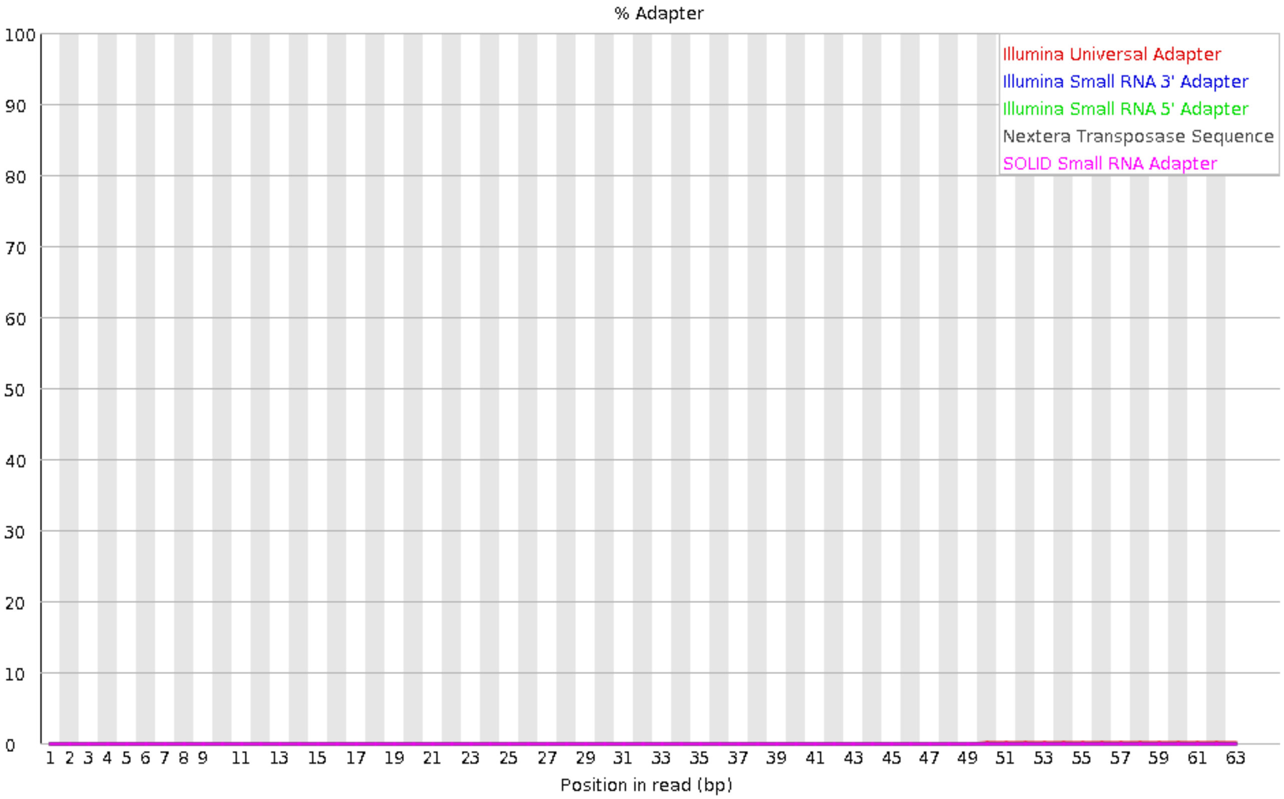


Also to check: Duplicated sequences



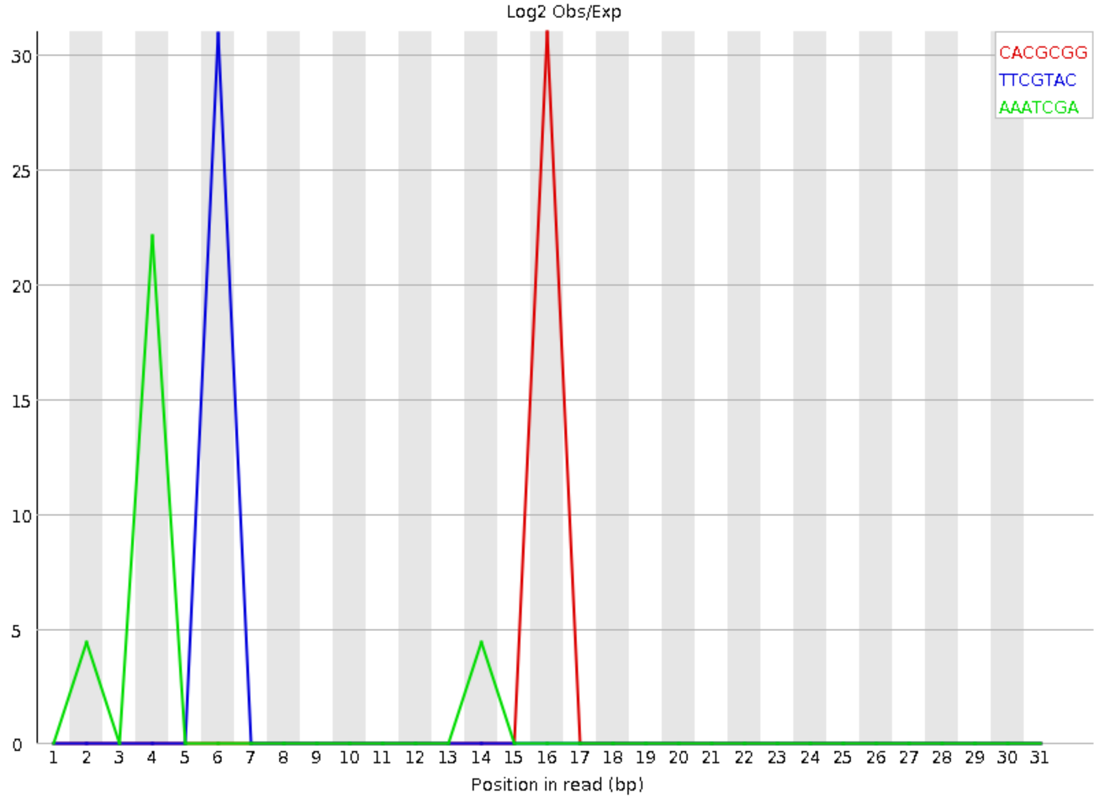
Also to check: Tag sequences

Adapter contamination



Also to check: Tag sequences

K-mer content



How to improve the quality of my sequences?

Sequence quality improvements

- Filtering of sequences
  - with small mean quality score
  - too small
  - with too many N bases
  - based on their GC content
  - ...
- Cutting/Trimming sequences
  - from low quality score parts
  - tails
  - ...

Key points

- Run quality control on every dataset before running any other bioinformatics analysis
- Take care of the parameters used to improve the sequence quality
- Re-run FastQC to check the impact of the quality control

Thank you!

This material is the result of a collaborative work. Thanks the Galaxy Training Network and all the contributors (Bérénice Batut) !

Found a typo? Something is wrong in this tutorial?
Edit it on GitHub