H3ABioNet

Pan African Bioinformatics Network for H3Africa

16SrRNA Intermediate Bioinformatics Online Course: Int_BT_2019

16S analysis pipeline

QC and ASV picking using the dada2 pipeline

Outline

- Quality Control
- DADA2 background
- DADA2 workflow

Quality Control

- Before analyzing generated sequence to draw biological conclusions, a quality control check should be performed to make sure there is no biases in the data.

- QC gives a quick impression of whether your data has any problems of which you should be aware before doing any analysis.

Quality Control

Potential problems:

- Low confidence bases (Ns)
- Sequence specific bias
- Sequence contamination
- Adapters
- …

Quality Control

Software packages for QC:

- FastQC
- MultiQC
- FastX-Toolkit
- PRINSEQ
- TagCleaner
- NGS QC Tool-Kit
- …

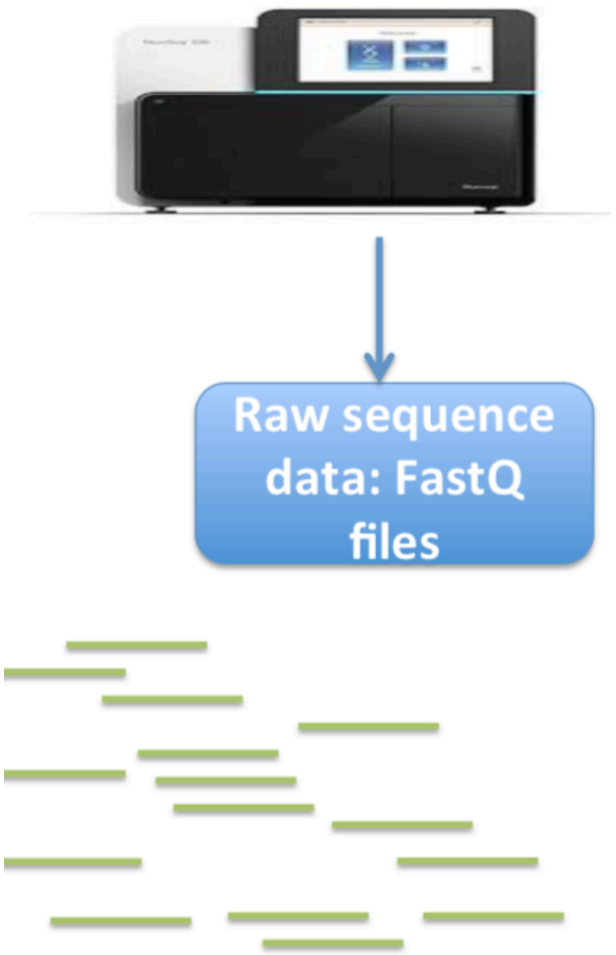
Quality Control

FASTQ format

What is a FastQ file?

FASTQ= FASTA + Quality

FastQ format is a text-based format for storing both a biological sequence and its corresponding quality scores.



Raw sequence data: FastQ files

Quality Control

FASTQ format

- Each FastQ file contains hundreds of millions of rows.
- Each block of 4 lines, starting with " @" represents a read.

Line 1 begins with a '@' character and is followed by a sequence identifier and an optional description (like a FASTA title line).

Line 2 is the raw sequence letters (ATCG).

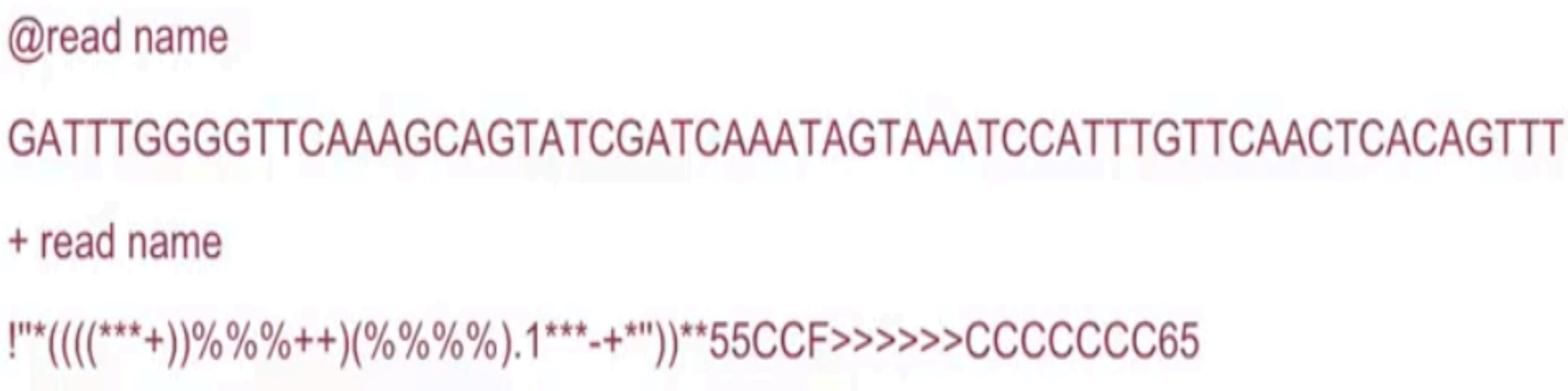
Line 3 begins with a '+' character and is optionally followed by the same sequence identifier (and any description).

Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence.

Quality Control

FASTQ format

A FastQ file containing a single sequence might look like this:



```
@read name
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+ read name
!"*((((***+))%%%++)(%%%%).1***-+*"))**55CCF>>>>>>CCCCCCC65
```

The character '!' represents the lowest quality while '~' is the highest.

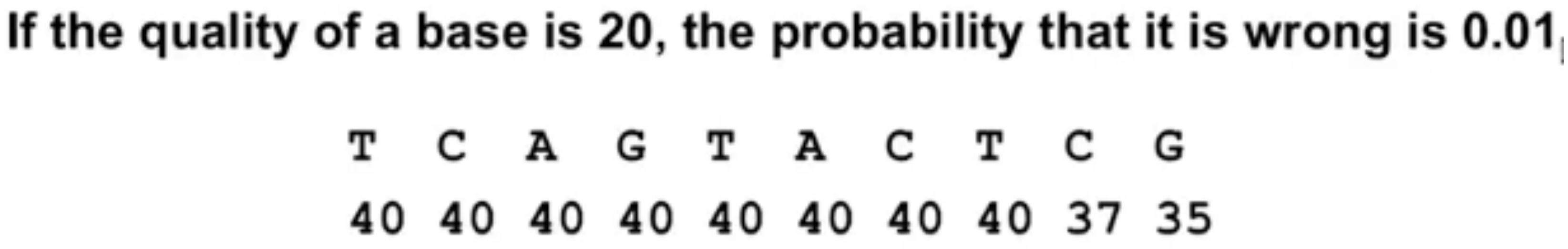
Quality Control

Quality measurements

Base-calling error probabilities are reported by sequencers.
Usually in Phred (quality) score.
Usually coded by ASCII characters

Phred score

$$Q = -10log_{10}P$$

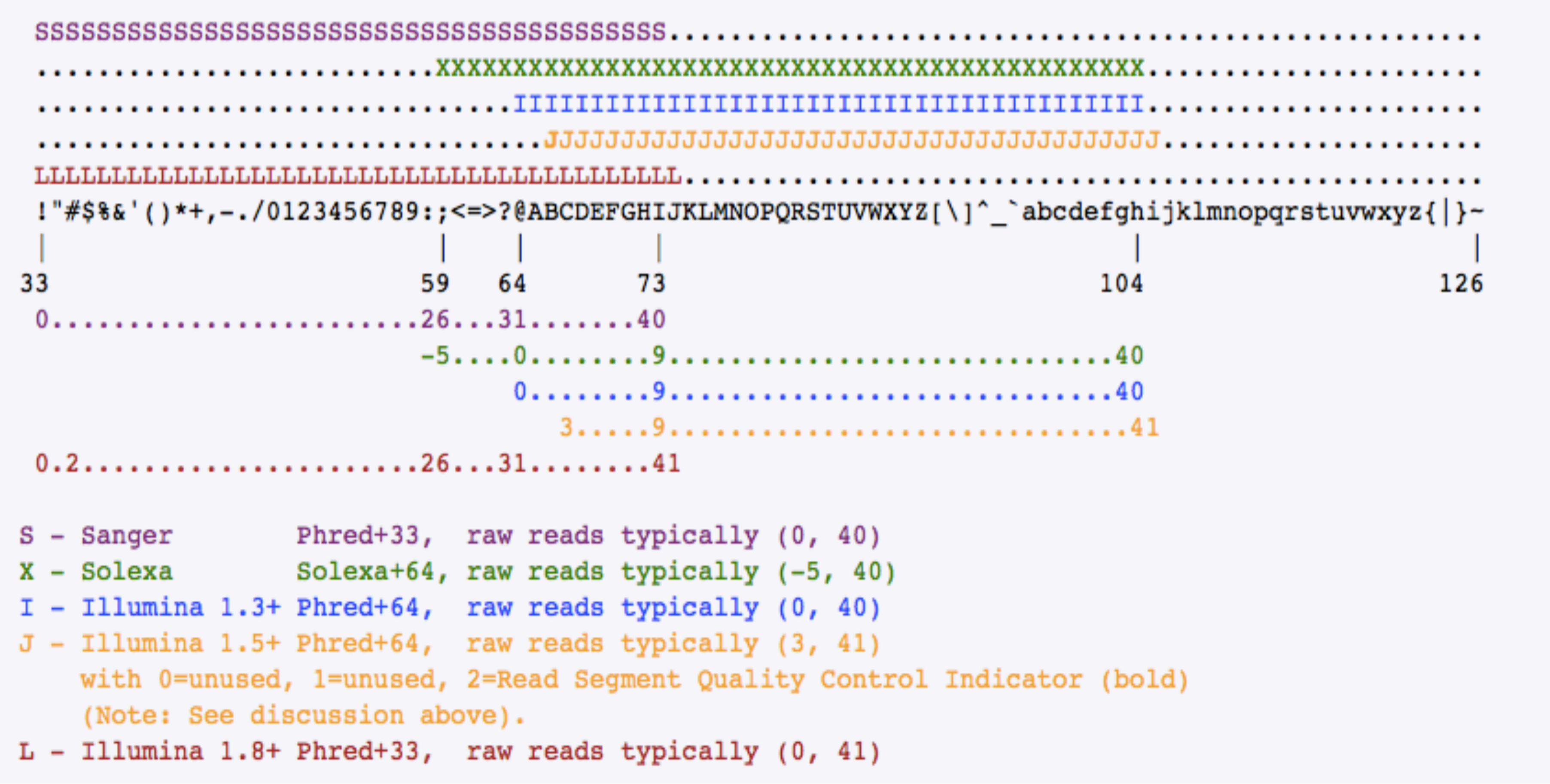


If the quality of a base is 20, the probability that it is wrong is 0.01

T	C	A	G	T	A	C	T	C	G
40	40	40	40	40	40	40	40	37	35

Quality Control

Quality measurements



```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
  ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
  ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
  .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
  LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
  !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
  |                         |    |        |                              |                     |
 33                        59   64       73                            104                   126
  0........................26...31.......40
                           -5....0........9.............................40
                                 0........9.............................40
                                    3.....9..............................41
  0.2......................26...31........41

 S - Sanger        Phred+33,  raw reads typically (0, 40)
 X - Solexa        Solexa+64, raw reads typically (-5, 40)
 I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
 J - Illumina 1.5+ Phred+64,  raw reads typically (3, 41)
     with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
     (Note: See discussion above).
 L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

H3ABioNet
Pan African Bioinformatics Network for H3Africa

Quality Control

What is FastQC?

FastQC aims to provide a QC report which can spot problems which originate either in the sequencer or in the library material.

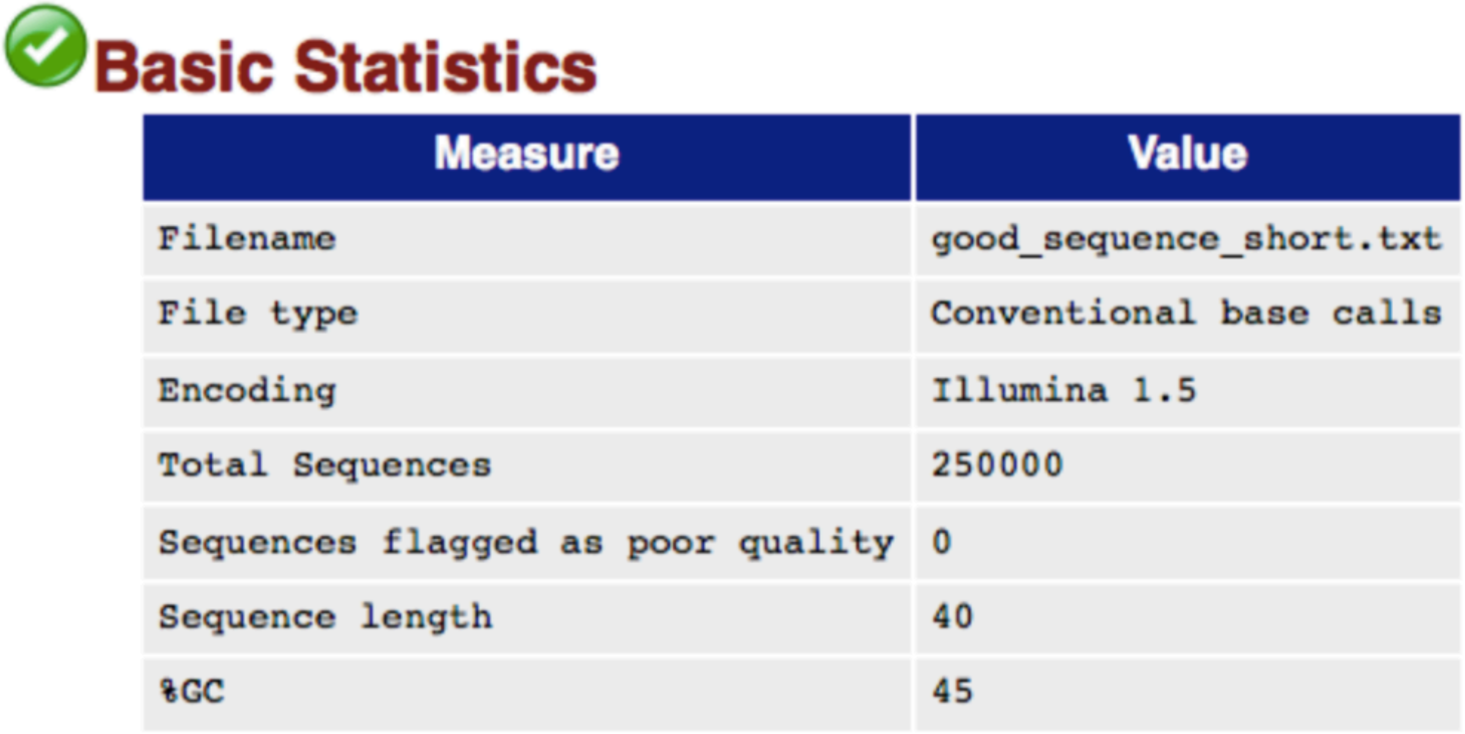
Quality Control

FastQC reports

🟩	Normal
🟧	Slightly abnormal
🟥	Unexpected

Quality Control

FastQC reports



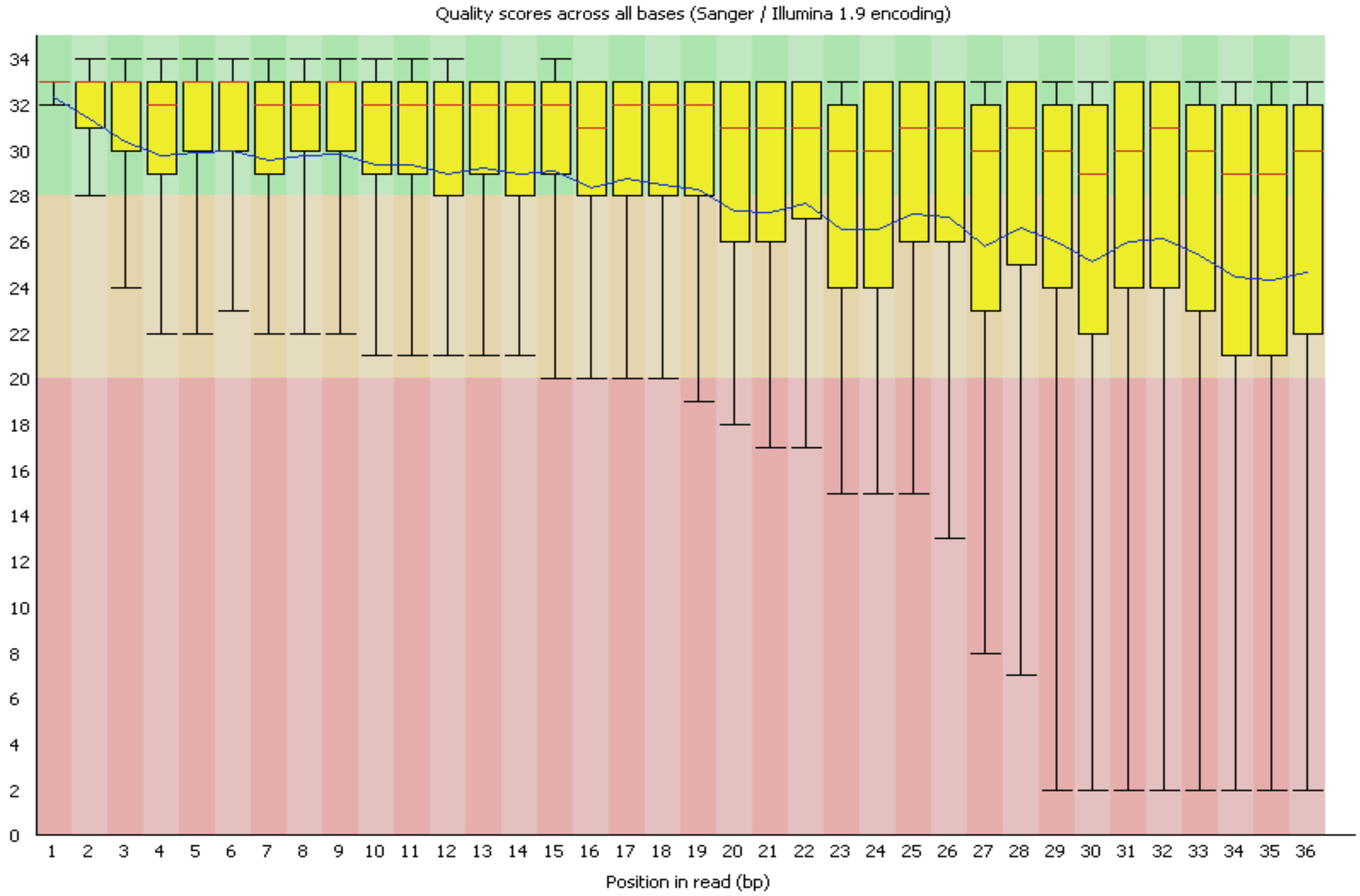
✓ Basic Statistics

Measure	Value
Filename	good_sequence_short.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	250000
Sequences flagged as poor quality	0
Sequence length	40
%GC	45

Quality Control

FastQC reports

Per Base Sequence Quality



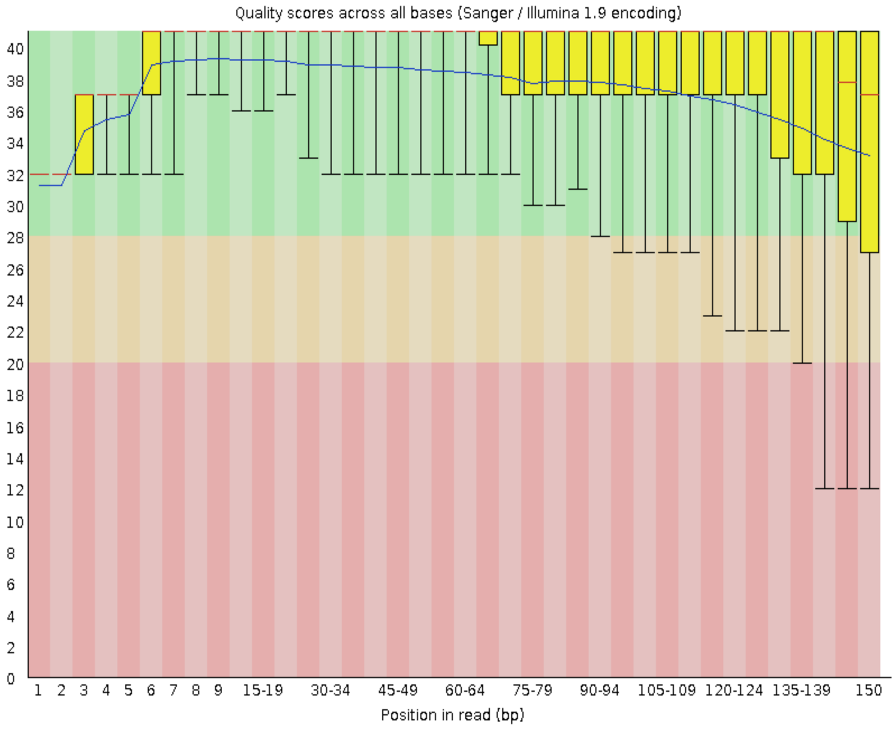
H3ABioNet
Pan African Bioinformatics Network for H3Africa

Quality Control

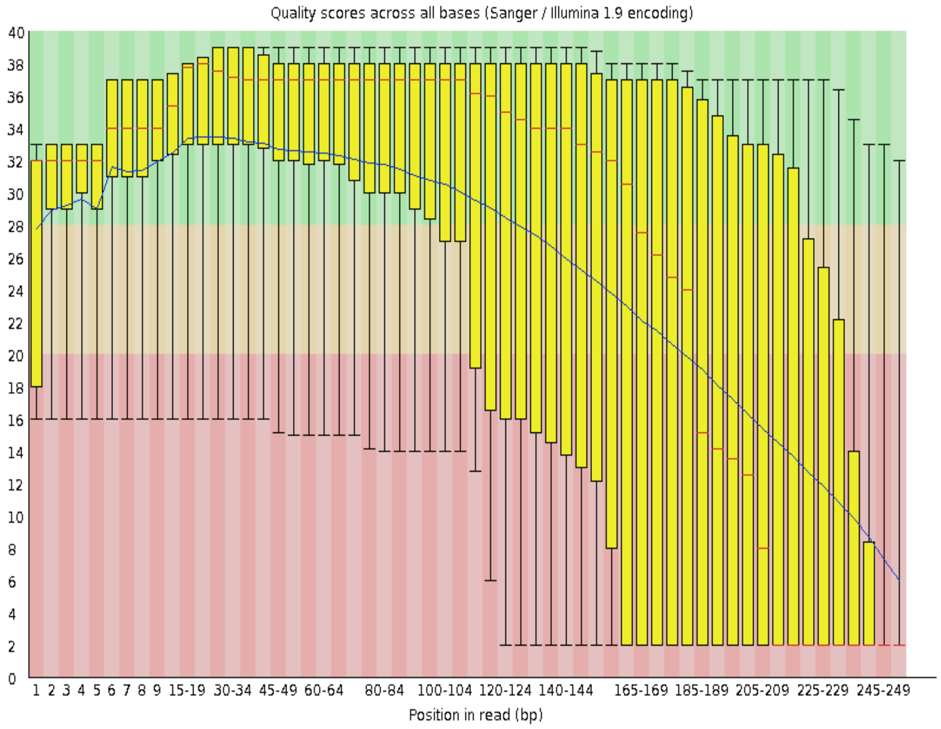
FastQC reports

Per Base Sequence Quality

Good quality FastQC report:



Bad quality FastQC report

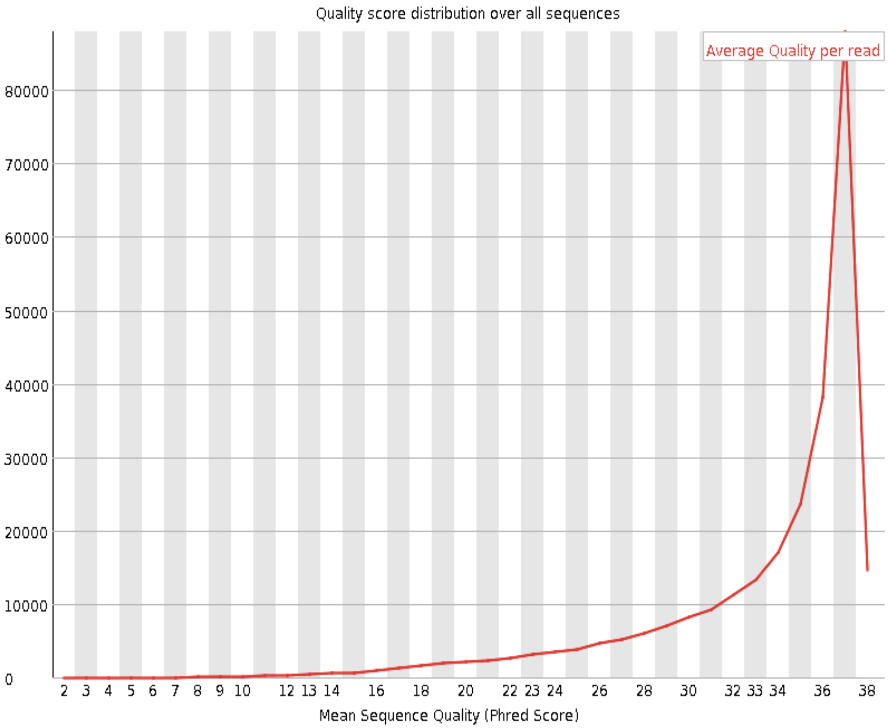


Quality Control

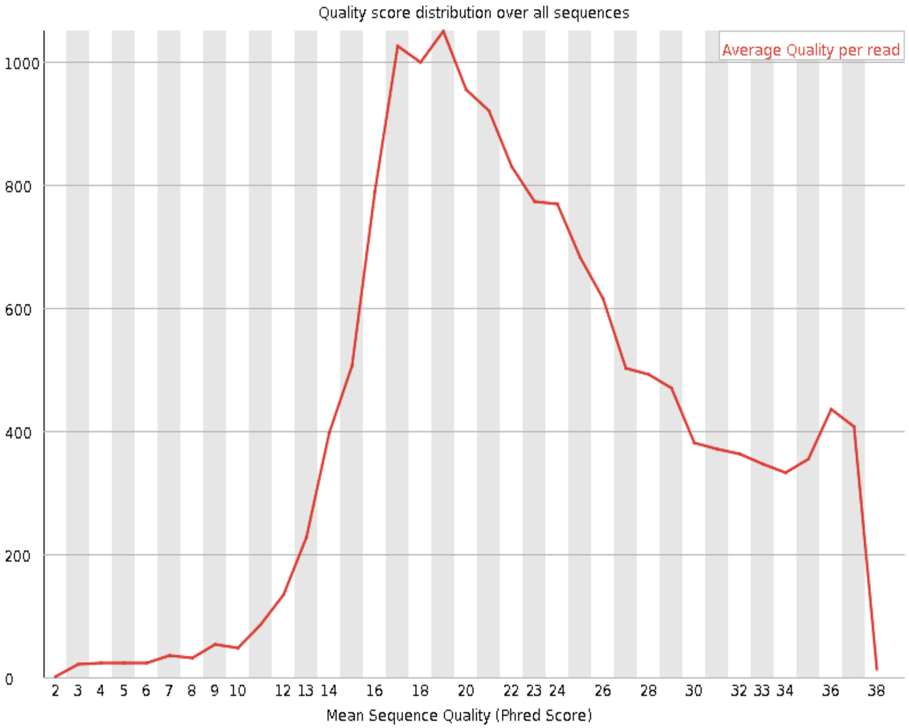
FastQC reports

Per Sequence Quality Scores

Good quality FastQC report:



Bad quality FastQC report

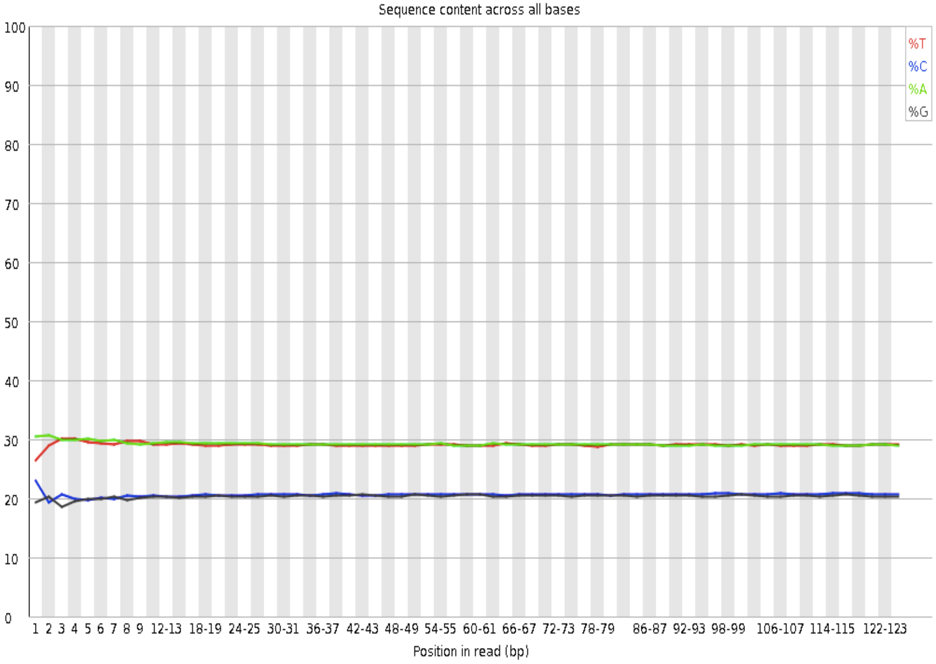


Quality Control

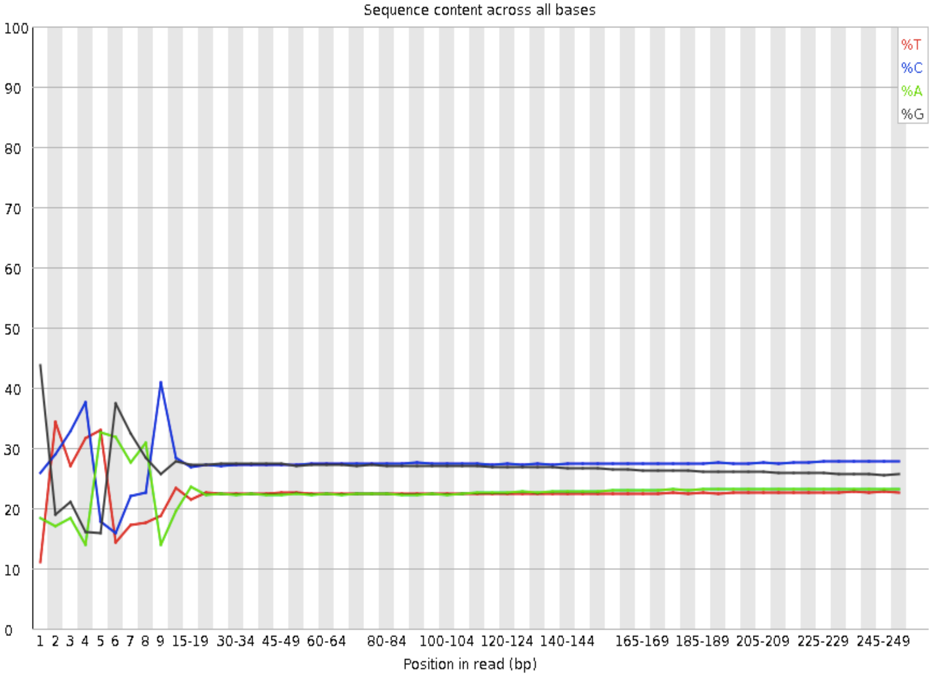
FastQC reports

Per Base Sequence Content

DNA library



RNA library


Sequence content across all bases
%T
%C
%A
%G
Position in read (bp)