H3ABioNet
Pan African Bioinformatics Network for H3Africa

WELLCOME GENOME CAMPUS
CONNECTING SCIENCE
ADVANCED COURSES+ SCIENTIFIC CONFERENCES

Next Generation Sequencing Bioinformatics Course 2021

Module 3: NGS Data Pre-processing & QC

FastQ Quality Control

Fatma Guerfali

Learning Objectives

- Identify and locate FastQC step in the general NGS Bioinformatics analysis workflow
- Summarize the principles of FastQC
- Describe the output of a FastQ Quality Control
- Interpret different QC metrics, graphs and reports (good data set and bad data set)

Session Plan

01

General Overview of NGS Bioinformatics

From FastQ (raw data file) to variant calling (VCF)

02

Introduction to Quality Control (QC) and FastQC

Why we need a FastQ Quality Control

03

Execute FastQC

Use the FastQC tool

04

Interpret QC metrics, graphs and reports

Guidelines on understanding Data Quality through examples of Good and Bad data sets

Session Plan

01

General Overview of NGS Bioinformatics

From FastQ (raw data file) to variant calling (VCF)

02

Introduction to Quality Control (QC) and FastQC

Why we need a FastQ Quality Control

03

Execute FastQC

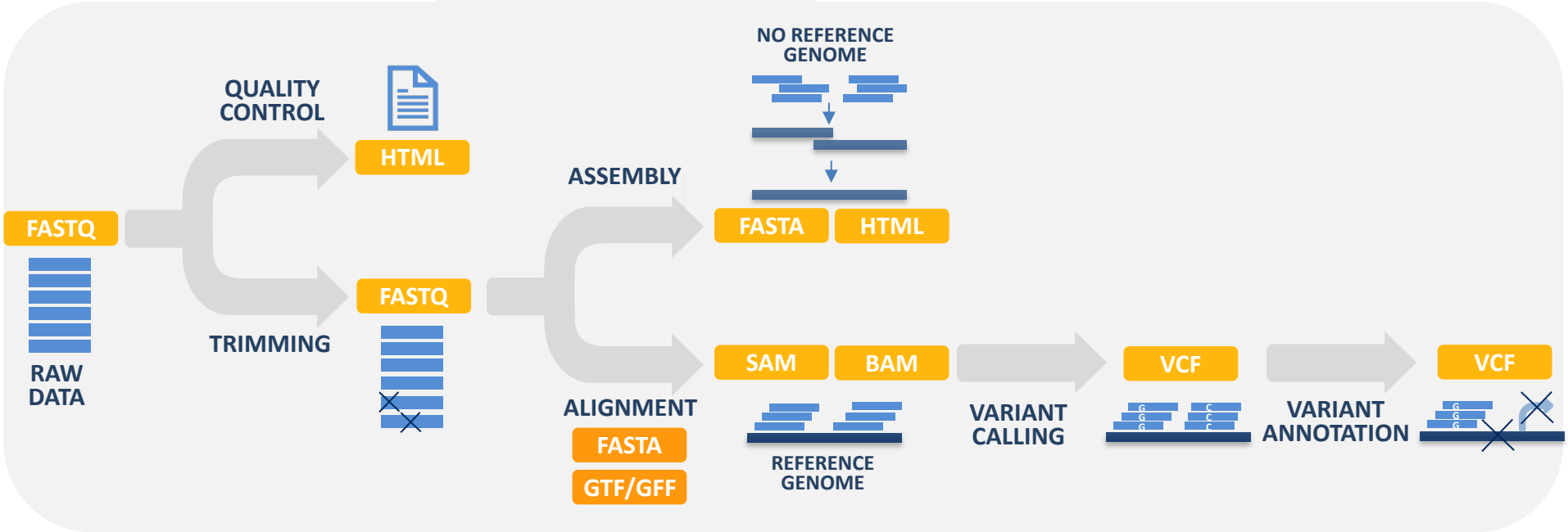
Use the FastQC tool

04

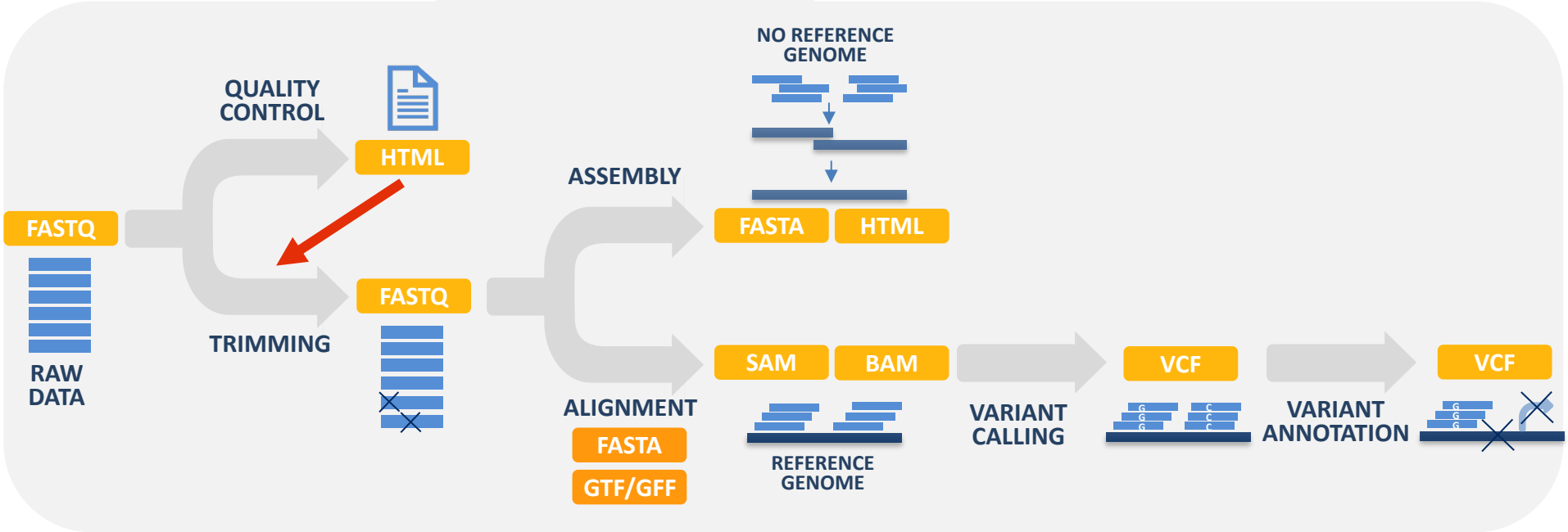
Interpret QC metrics, graphs and reports

Guidelines on understanding Data Quality through examples of Good and Bad data sets

Overview of NGS data analysis



Overview of NGS data analysis



Session Plan

01

General Overview of NGS Bioinformatics

From FastQ (raw data file) to variant calling (VCF)

02

Introduction to Quality Control (QC) and FastQC

Why we need a FastQ Quality Control

03

Execute FastQC

Use the FastQC tool

04

Interpret QC metrics, graphs and reports

Guidelines on understanding Data Quality through examples of Good and Bad data sets

FastQC

- FastQC → quality control on raw sequence data (FASTQ)
- What data is contained in a FASTQ file?

FASTQ

The FASTQ file is the most common raw data output provided by NGS sequencing platforms (text-based file)

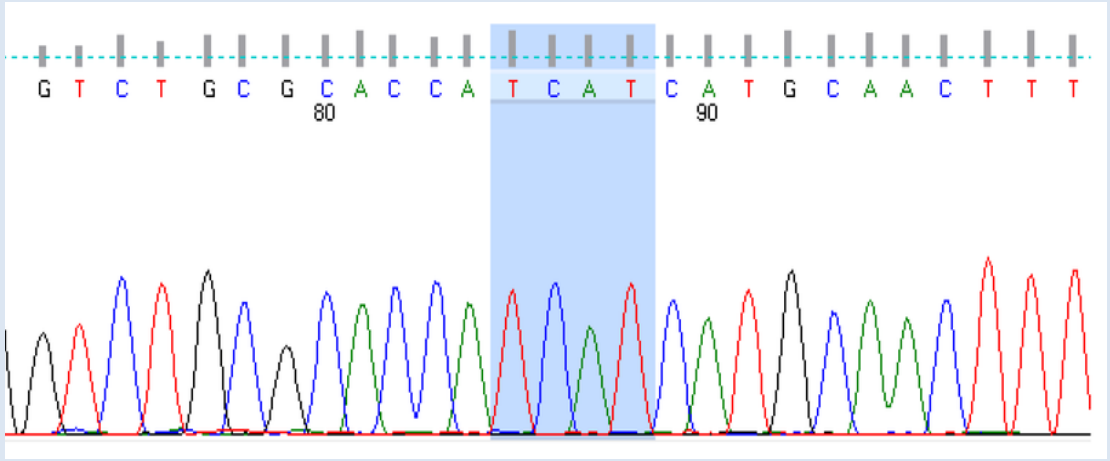
NB: Additional files exist depending on the technology used, e.g.:

- **PacBio**
  - basecall File Format (bas.h5/bax.h5) → HDFView

- **Illumina** :
  - The NextSeq, HiSeq, and NovaSeq 6000 Sequencing Systems generate raw data files in binary base call (BCL) format → bcl2fastq Conversion.
  - FASTQ ORA is a binary compressed file format of FASTQ → fastq.ora files are up to 5x smaller than their corresponding fastq.gz files

FASTQ

Reminder: Sanger Sequencing



← Sequence

← Quality

→ **a FATSQ file is a file containing :**

- **Reads sequences**
- **a Quality score associated to each Read**

FastQ

A FATSQ file defines each read by 4 lines :

- **Line 1**: Read sequence identifier (encoded descriptions of instrument, lane…)
- **Line 2**: Read sequence
- **Line 3**: « + » sign (optional: « + » followed by seq identifier)
- **Line 4**: a string of ASCII characters = Quality score associated to each Read

http://drive5.com/usearch/manual/fastq_fig.jpg

FastQC

- FastQC → quality control on raw sequence data (FASTQ)

  - Import your data from FASTQ file
  - Analyze the sequence and quality provided in the FASTQ
  - Summary metrics and graphs to visually assess your data QC
  - Export of results in the form of a HTML based report

- You can use this report to have a quick overview of the quality of your data

- This helps you know if your data has any problems that needs to be taken into consideration before doing any further analysis.

Session Plan

01
General Overview of NGS Bioinformatics

From FastQ (raw data file) to variant calling (VCF)

02
Introduction to Quality Control (QC) and FastQC

Why we need a FastQ Quality Control

03
Execute FastQC

Use the FastQC tool

04
Interpret QC metrics, graphs and reports

Guidelines on understanding Data Quality through examples of Good and Bad data sets

Execute FastQC

- The FASTQ Quality Control step can be done using a tool called **FastQC**
- At this stage, you must be able to use the command line to:
  - run this tool on multiple .fq files using the * wildcard (files will be processed serially and separately)

```
$ fastqc *.fq
```

  - get help on arguments for commands/tools

```
$ fastqc --help
```

  - Optional: use MultiQC to aggregate results from multiple FastQC (and other) runs into one single html report

Session Plan

01
General Overview of NGS Bioinformatics

From FastQ (raw data file) to variant calling (VCF)

02
Introduction to Quality Control (QC) and FastQC

Why we need a FastQ Quality Control

03
Execute FastQC

Use the FastQC tool

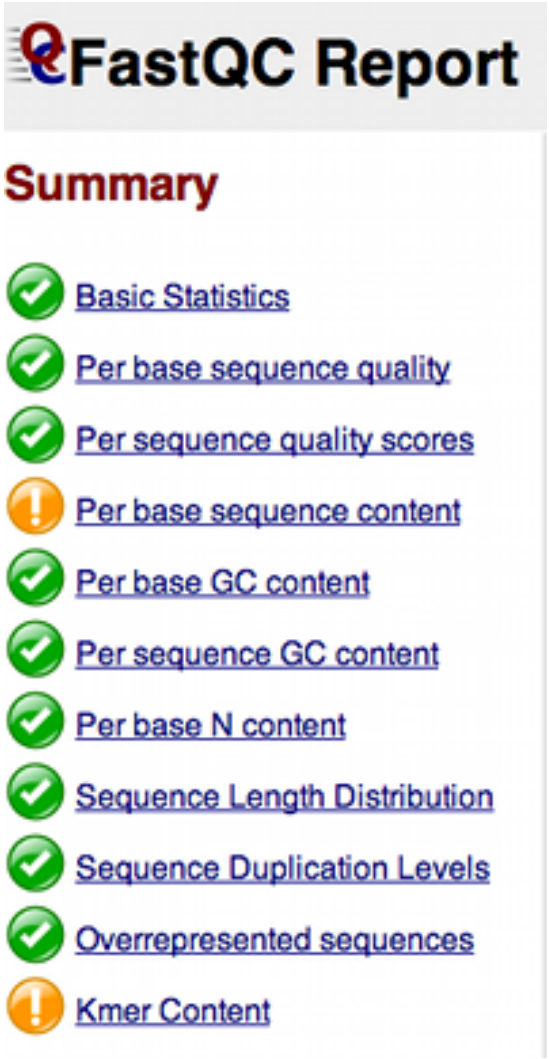
04
Interpret QC metrics, graphs and reports

Guidelines on understanding Data Quality through examples of Good and Bad data sets

FastQC Run Report

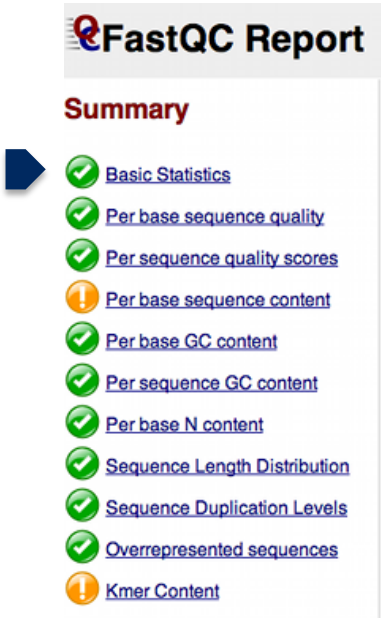
FASTQC run Report

- a FASTQC run report file is a file containing informations about the quality control made on the FASTQ file
- Metrics and Graphs (.html)



https://wiki.hpcc.msu.edu/download/attachments/15434467/fastqc-1.png?version=1&modificationDate=1365623346000&api=v2

FastQC Run Report



Basic Statistics

The Basic Statistics module generates some simple composition statistics for the file analyzed

Basic Statistics

Measure	Value
Filename	good_sequence_short.fastq
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	250000
Filtered Sequences	0
Sequence length	40
%GC	45

- ASCII encoding of quality values in the fastq file
- Total number of sequences processed.
- Read length

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

FastQC Run Report

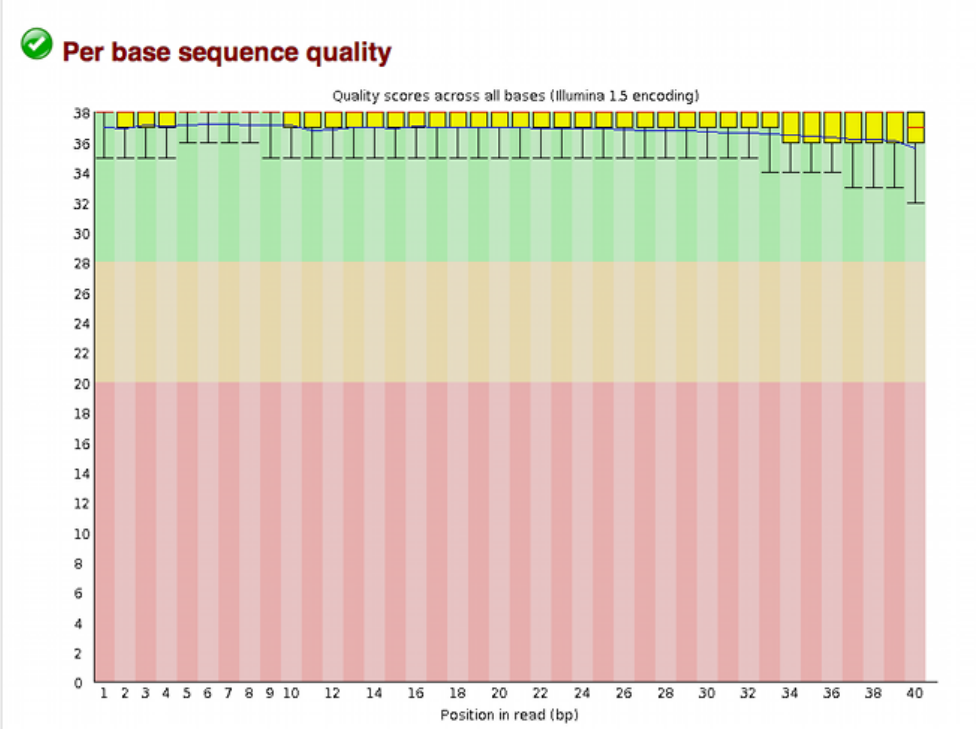
FastQC Report

Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content

Per Base Sequence Quality

Range of quality values across all bases of the read at each position



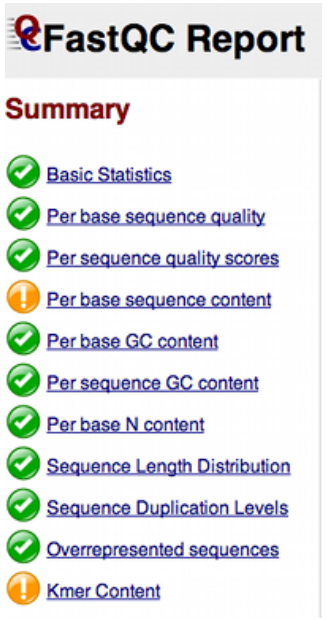
For each position a Boxplot is drawn (BoxWhisker type plot).

The **y-axis** on the graph shows the **Quality Scores**.

The **x-axis** shows the **nucleotide position** in the read.

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
https://wiki.hpcc.msu.edu/download/attachments/15434467/fastqc-1.png

FastQC Run Report

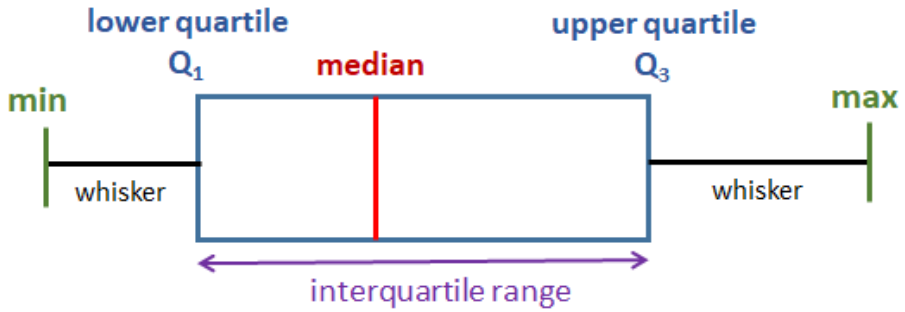


Per Base Sequence Quality

Range of quality values across all bases of the read at each position

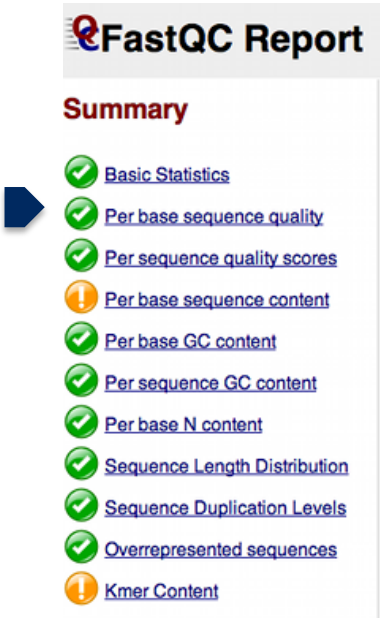
Box and Whisker Plot

A box and whisker plot (also called a box plot) shows the five-number summary of a set of data: **minimum**, **lower quartile**, **median**, **upper quartile**, and **maximum**.



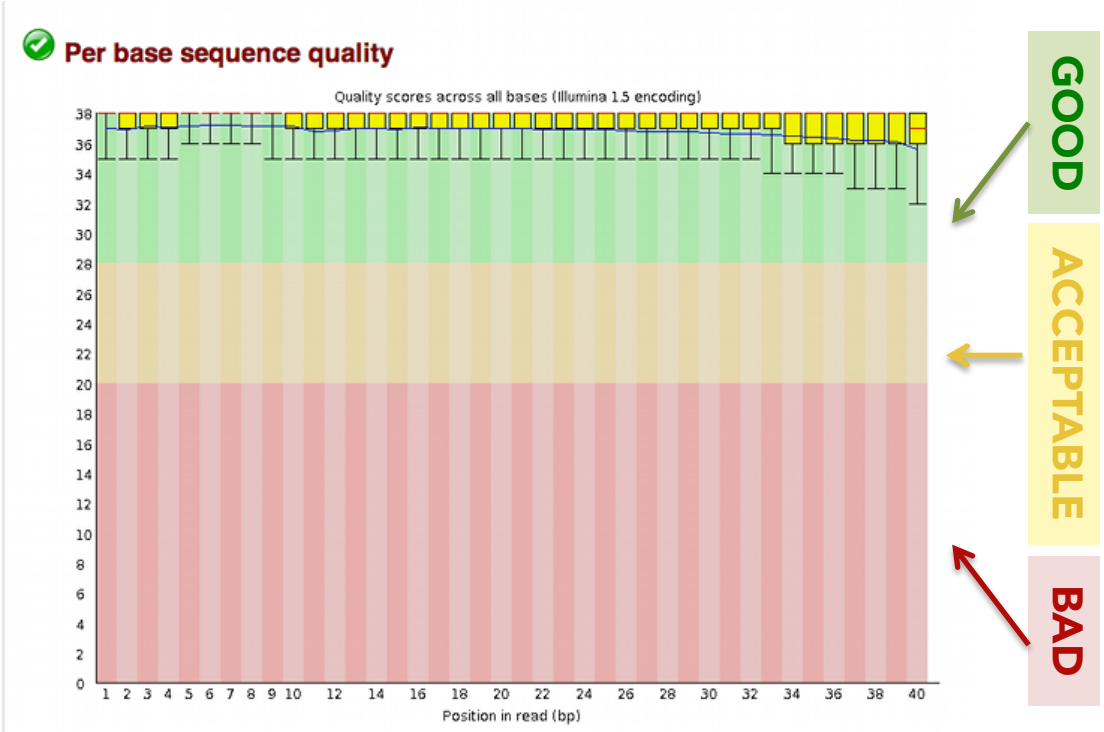
https://www.onlinemathlearning.com/image-files/box-plot.png

FastQC Run Report



Per Base Sequence Quality

GOOD DATA
MOST READS = HIGH QUALITY



The central **red** line is the median value.

The **yellow** box represents the inter-quartile range (25-75%).

The **upper and lower whiskers** represent the 10% and 90% points.

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
https://wiki.hpcc.msu.edu/download/attachments/15434467/fastqc-1.png

FastQC Run Report

FastQC Report

Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution

Per Base Sequence Quality

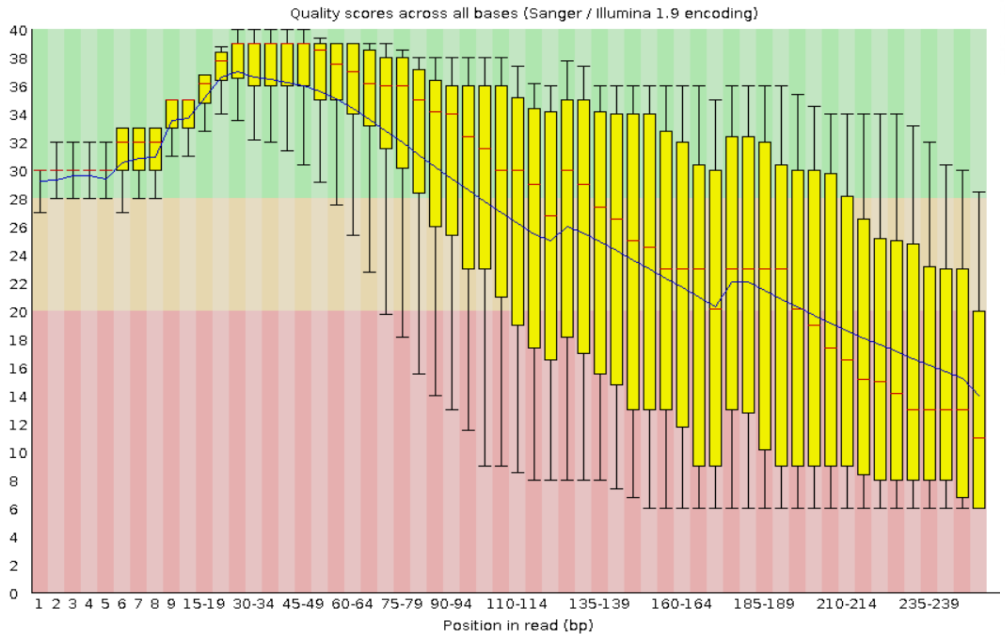
Warning if the lower quartile for any base is <10, or if median for any base is <25.

Failure if the lower quartile for any base is <5, or if median for any base is <20.

BAD DATA

GENERAL DEGRADATION OF QUALITY WITH INCREASING READ LENGTH.
(CHEMISTRY LIMIT)

Per base sequence quality



http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
https://wiki.hpcc.msu.edu/download/attachments/15434467/fastqc-1.png

FastQC Run Report

FastQC Report

Summary

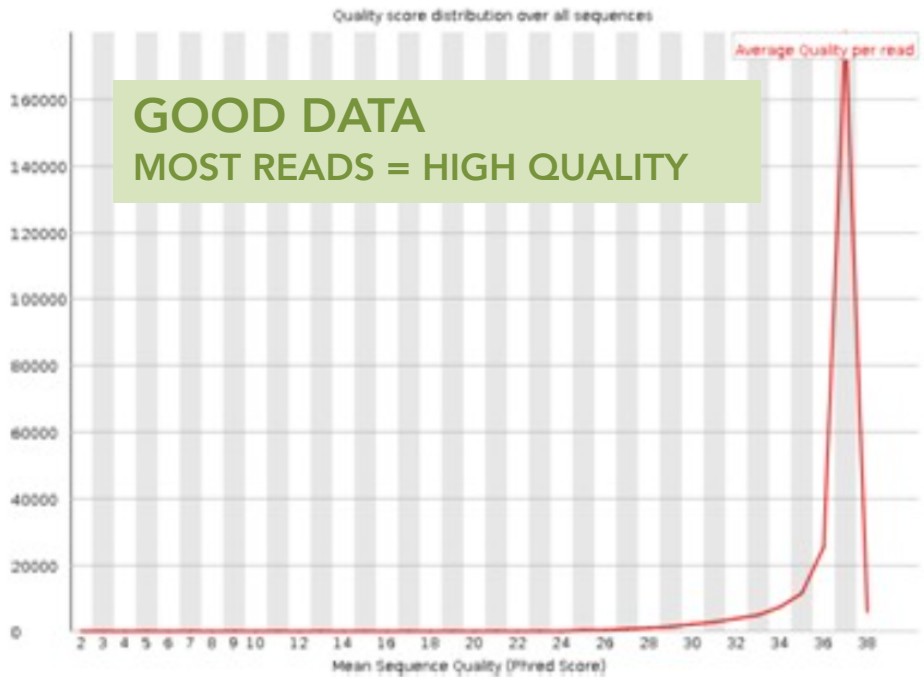
- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content

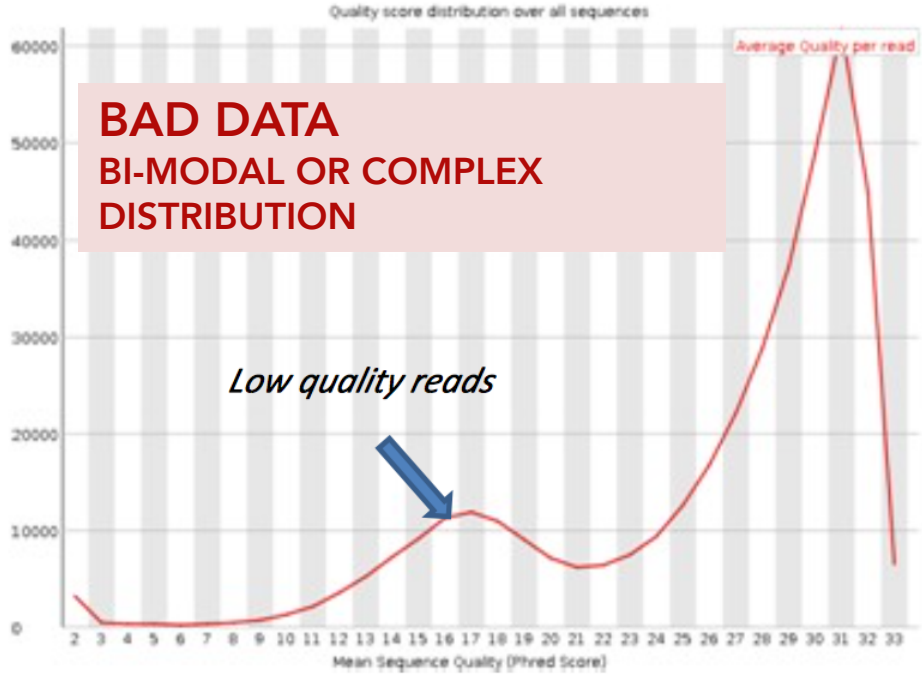
Per Sequence Quality Scores

Shows if a subset of sequences are of low quality

Warning if most frequently observed mean quality < 27 (0.2% ER).

Failure if most frequently observed mean quality < 20 (1% ER).





FastQC Run Report

FastQC Report

Summary

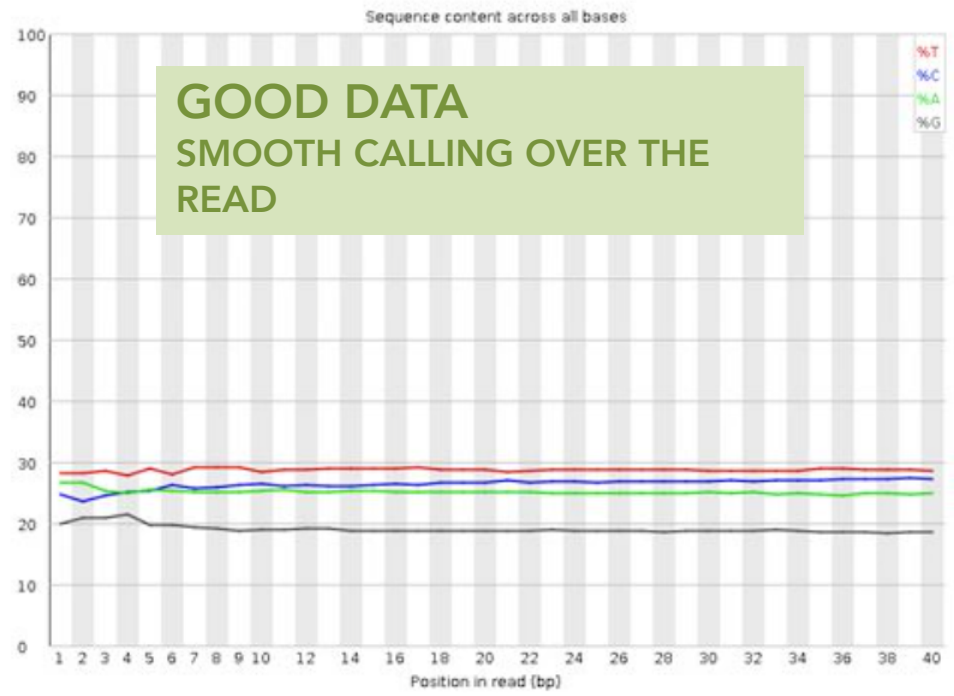
- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences

Per Base Sequence Content

Proportion of each base position for which each of the four normal DNA bases has been called.

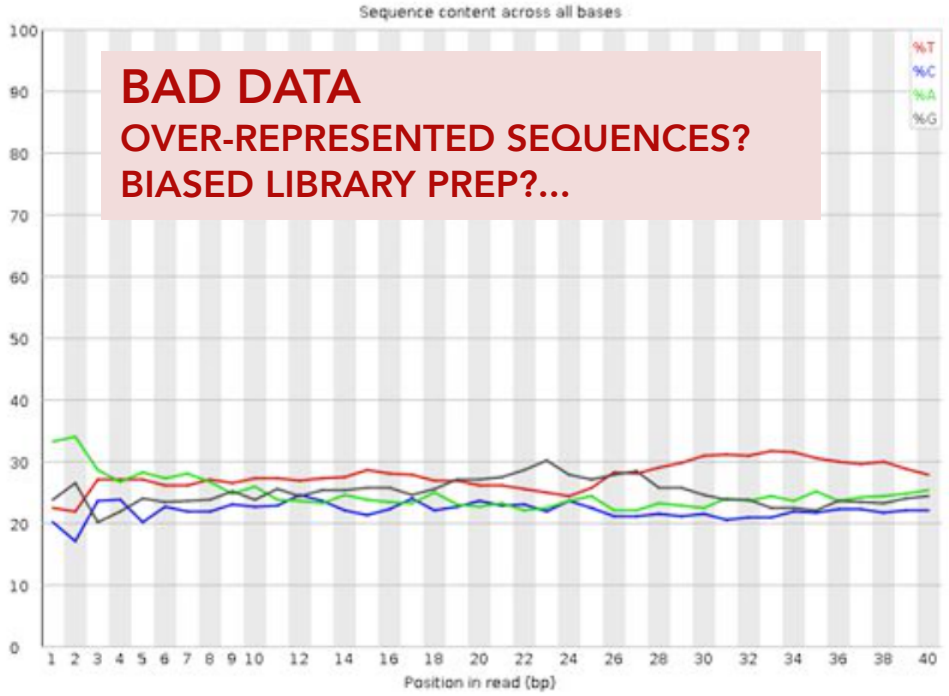
Warning if difference between A and T, or G and C >10% in any pos.

Failure if difference between A and T, or G and C >20% in any pos.



GOOD DATA

SMOOTH CALLING OVER THE READ



BAD DATA

OVER-REPRESENTED SEQUENCES?
BIASED LIBRARY PREP?...

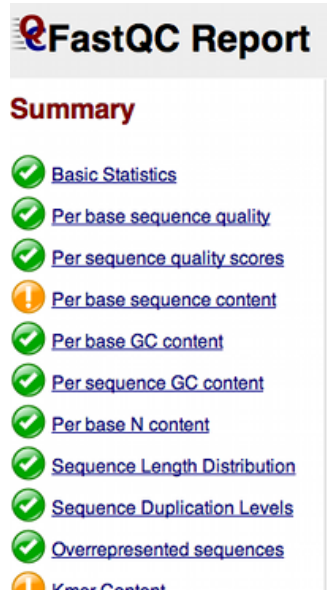
FastQC Run Report

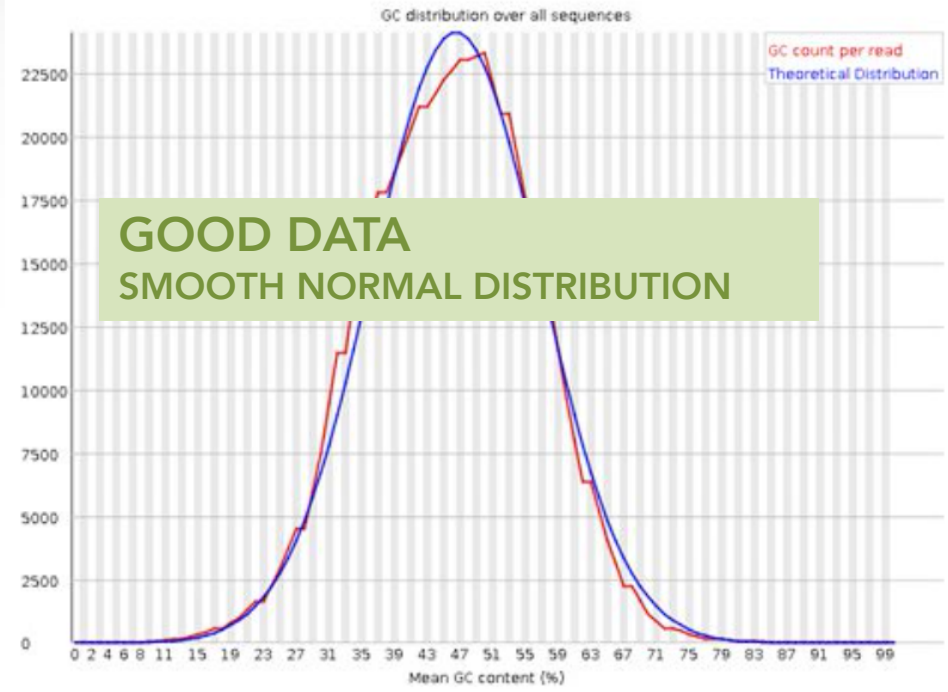
Per Sequence GC Content

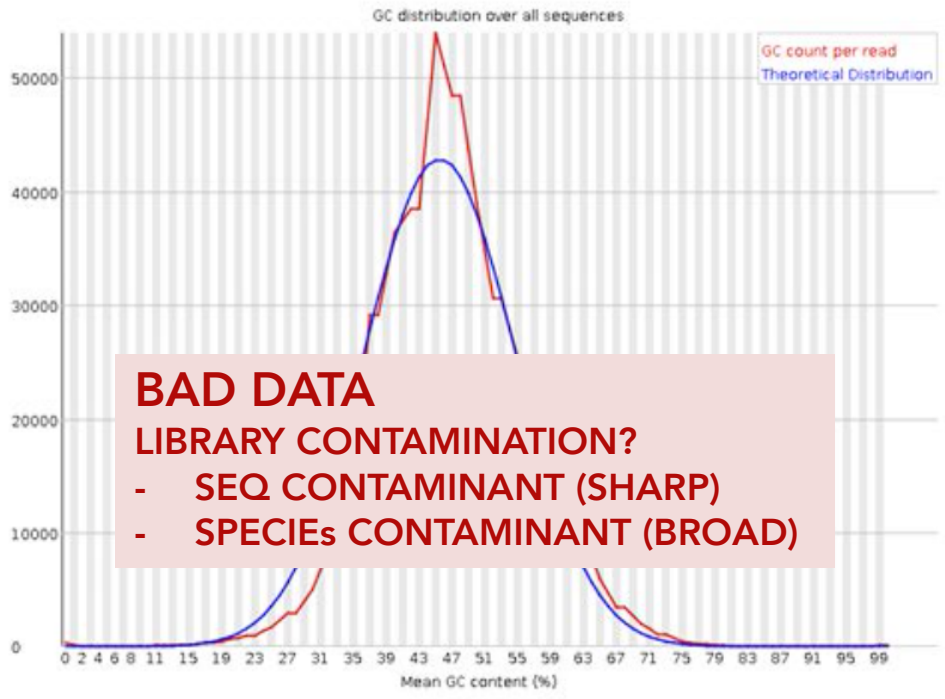
GC content across the whole length compared to a modelled normal distribution (ND) of GC content.

Warning if sum of deviations from the ND >15% of reads.

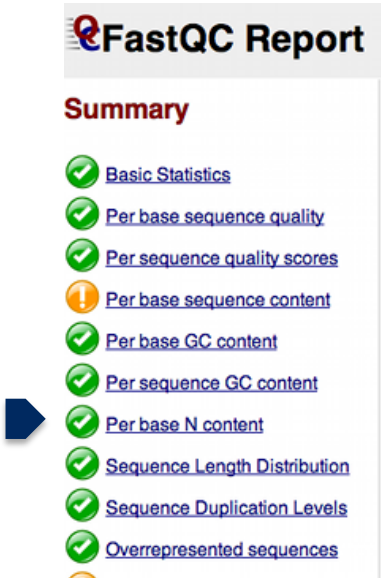
Failure if sum of deviations from the ND >30% of reads.







FastQC Run Report

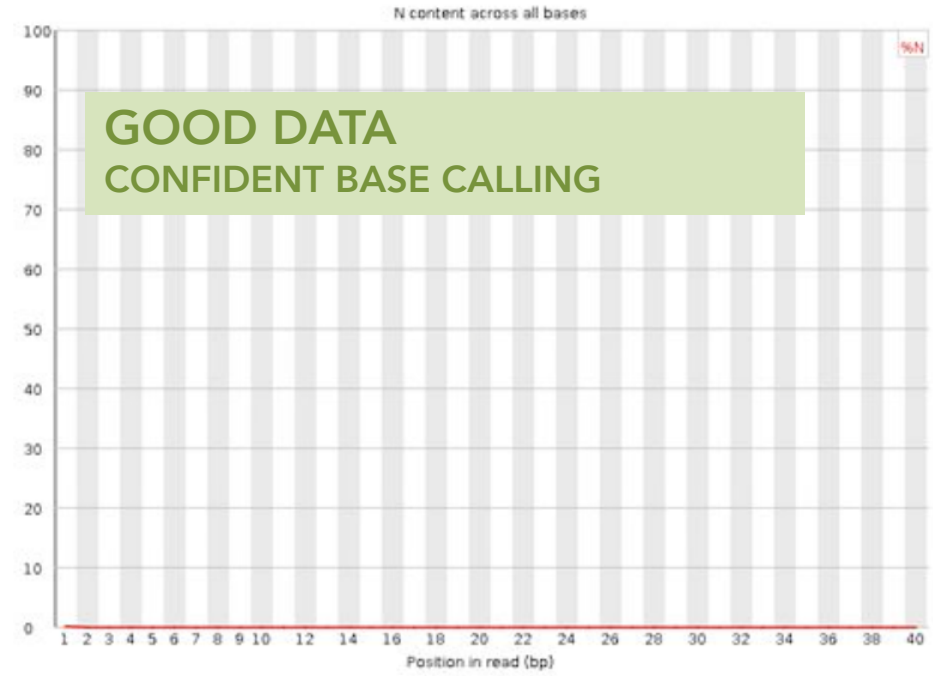


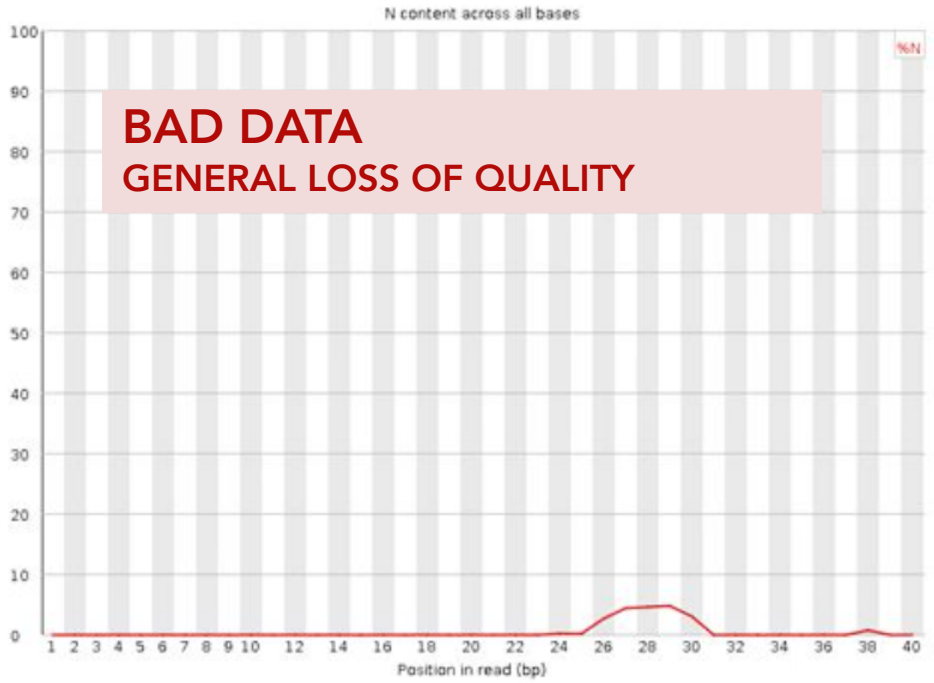
Per Base N Content

Percentage of base calls at each position for which an N was called.

Warning if any position shows an N content of >5%.

Failure if any position shows an N content of >20%.





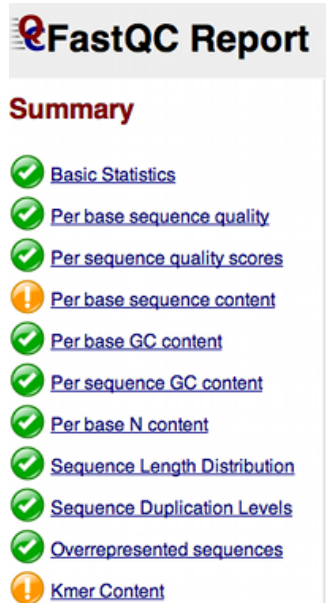
FastQC Run Report

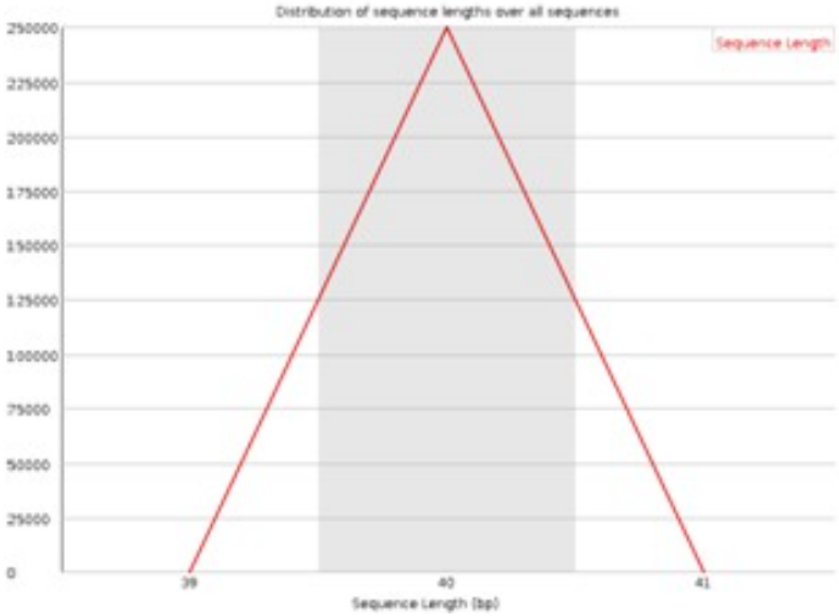
Sequence Length distribution

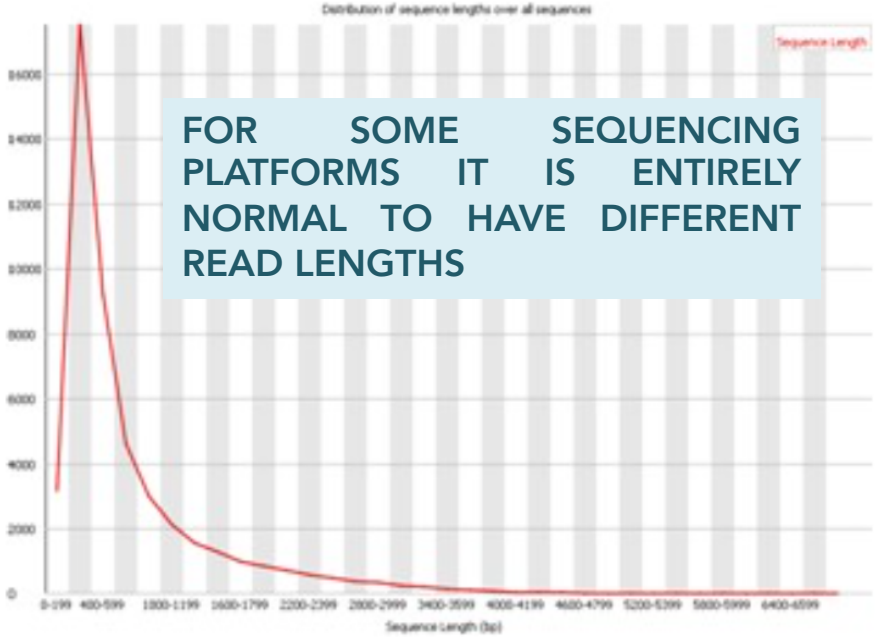
Distribution of fragment sizes in the file (some Libraries can generate sequence fragments of uniform length, others vary).

Warning if all sequences are not of the same length.

Failure if any of the sequences have length = 0.







FastQC Run Report

FastQC Report

Summary

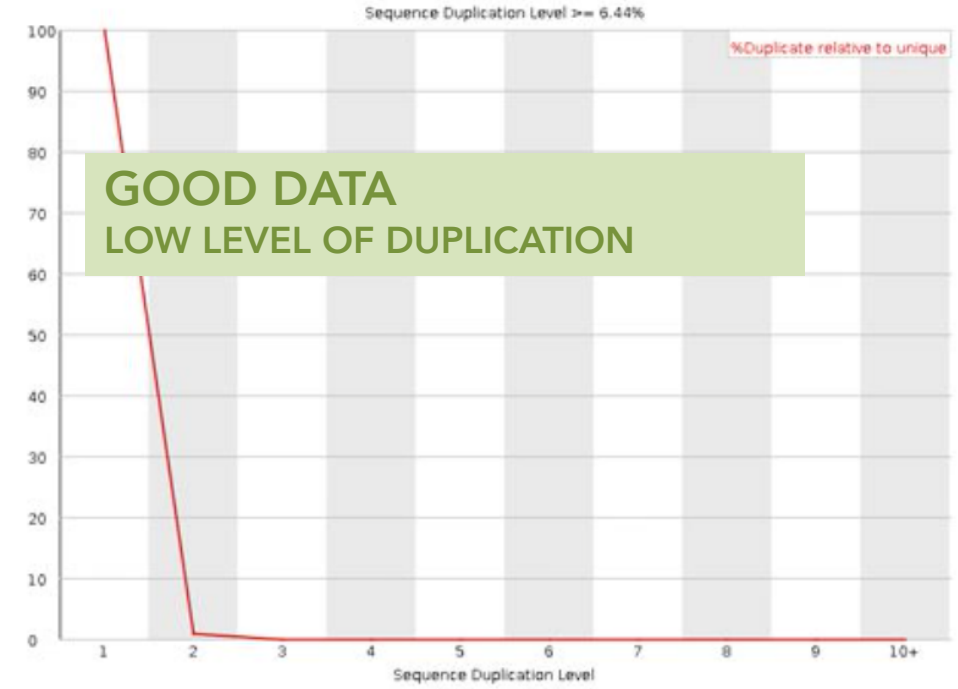
- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content

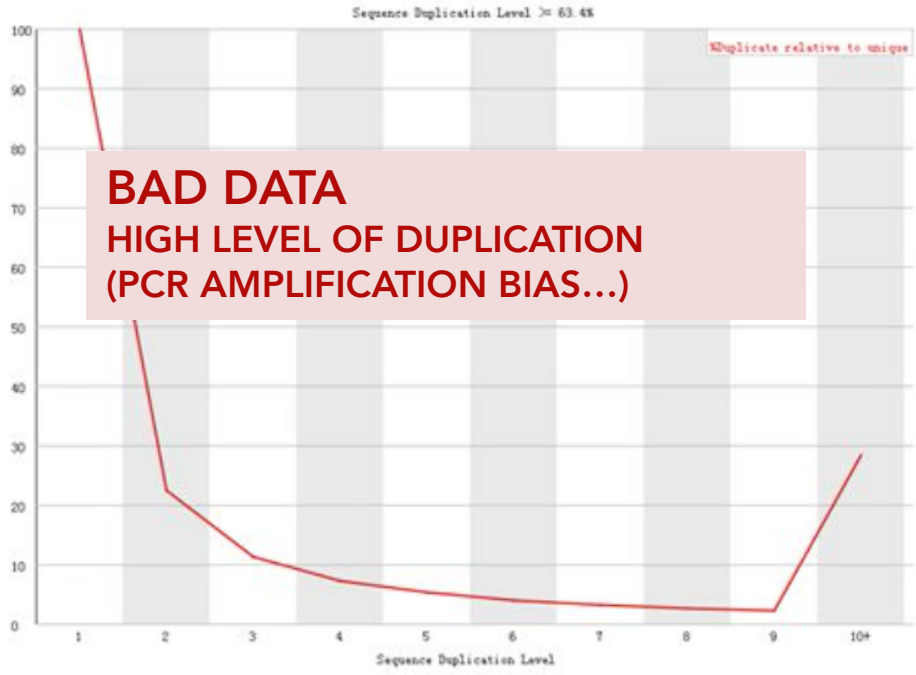
Sequence Duplication Level

In a diverse library most sequences will occur only once in the final set. Higher number of duplicated sequences is expected in Transcriptomics !

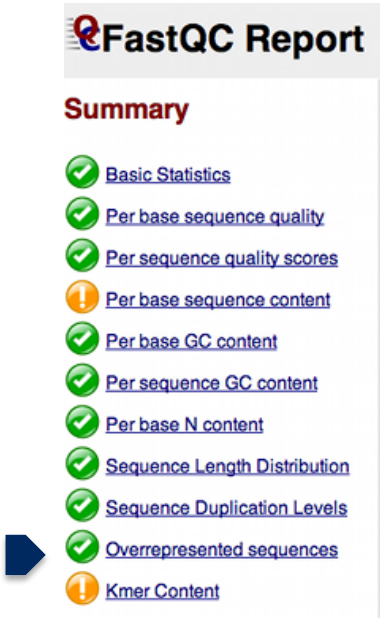
Warning if non-unique sequences > 20% of the total.

Failure if non-unique sequences > 50% of the total.





FastQC Run Report



Over-represented Sequences

If a single sequence is very overrepresented in the set = either means that it is highly biologically significant, or indicates that the library is contaminated, or not as diverse as you expected.

Warning if any sequence represent > 0.1% of the total.

Failure if any sequence represent > 1% of the total.

Sequence	Count	Percentage
AGAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTTC	2065	0.5224039181558763
GATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATG	2047	0.5178502762542754
ATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATGA	2014	0.5095019327680071
CGATAAAAATGATTGGCGTATCCAACCTGCAGAGTTTTAT	1913	0.4839509420979134
GTATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAGA	1879	0.47534961850600066
AAAAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCT	1846	0.4670012750197325

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
https://wiki.hpcc.msu.edu/download/attachments/15434467/fastqc-1.png

FastQC Run Report

FastQC Report

Summary

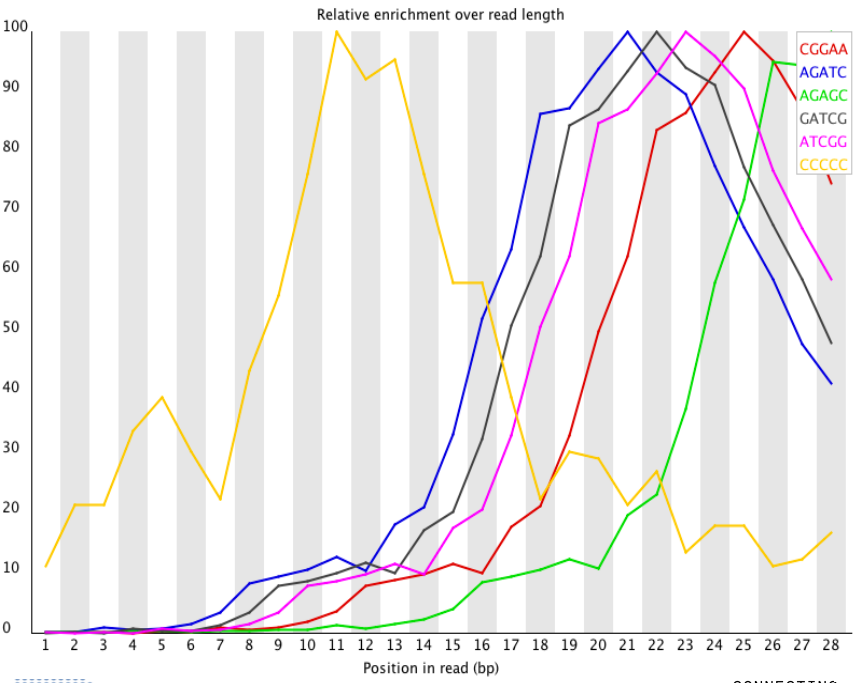
- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content

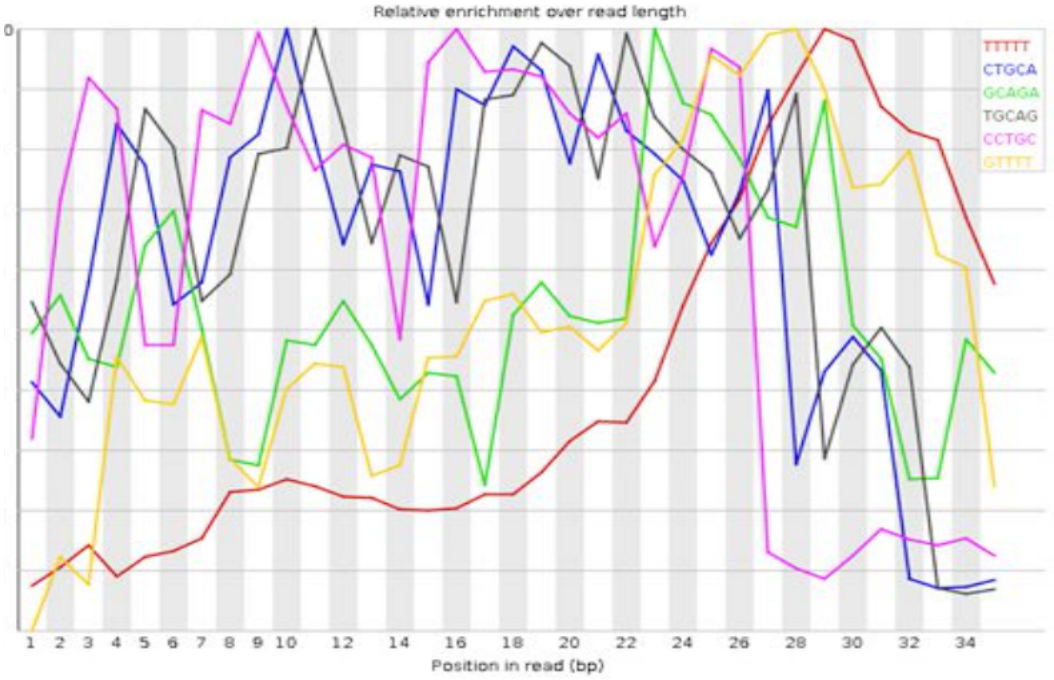
K-mer content

K-mer content of your reads sequences.

Warning if any k-mer is imbalanced with a binomial p-value <0.01.

Failure if any k-mer is imbalanced with a binomial p-value < 10^-5.





FastQC Run Report

FastQC Report

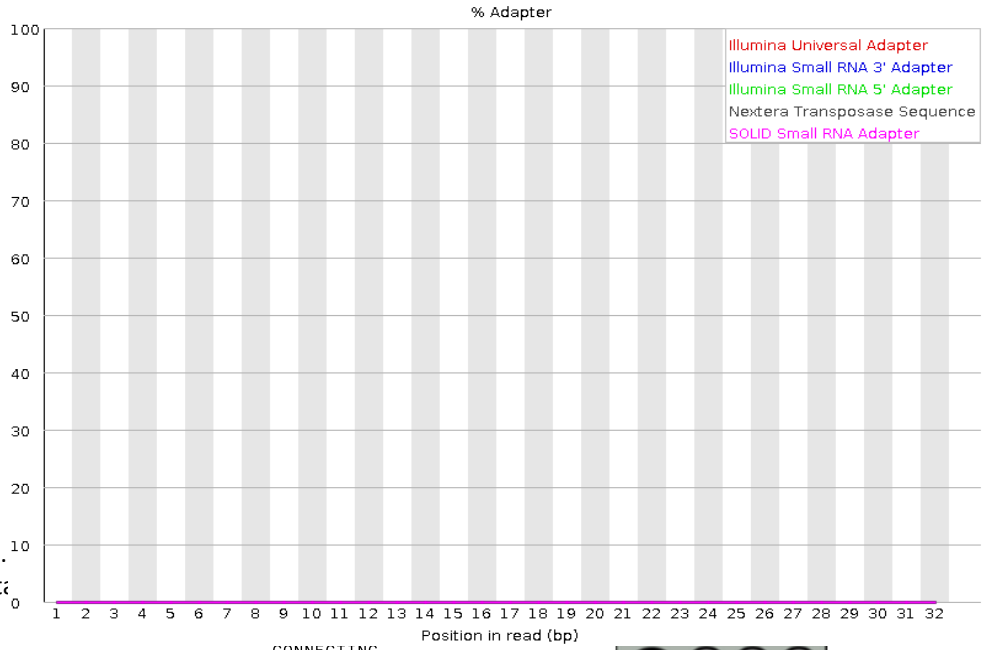
- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Adapter Content

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

Warning if any sequence is present in more than 5% of all reads.

Failure if any sequence is present in more than 10% of all reads.



http://www.bioinformatics.babraham.ac.
https://wiki.hpcc.msu.edu/download/atta

Interpreting FastQC

Warnings or Failures

Should be taken cautiously but may not be detrimental for the rest of the analysis. They might even be expected !

- **Sequence Duplication Levels:**
  - **low** number of duplicated sequences expected in Genomics (organism-dependent)
  - but **high**er number of duplicated sequences expected in Transcriptomics

- **Over-represented sequences:**
  - small RNA libraries where sequences are not subjected to random fragmentation
  - same sequence may natrually be present in a significant proportion of the library.

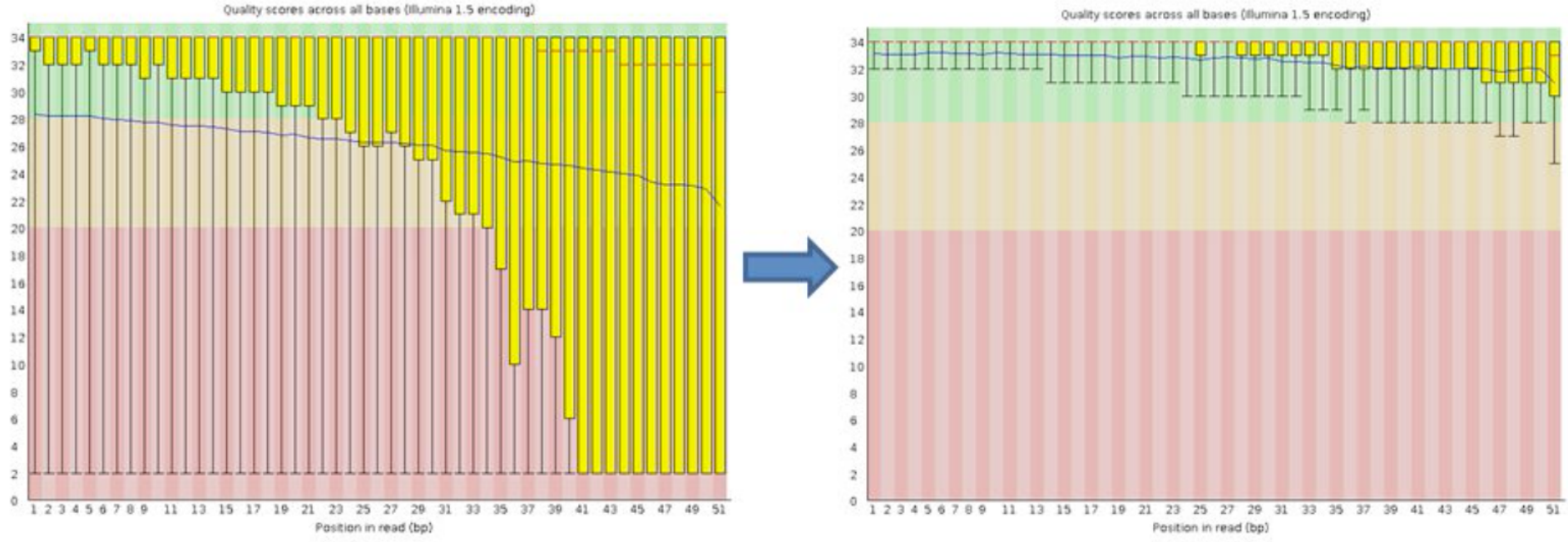
- **Kmer Content:**
  - Platform-dependent
  - Typical artifacts (libraries with random priming will nearly always show Kmer bias at the start of the library).

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
https://wiki.hpcc.msu.edu/download/attachments/15434467/fastqc-1.png

Interpreting FastQC

However if for some reason reads sequences show bad quality, it might be envisioned to perform a **Sequence Filtering / Trimming**

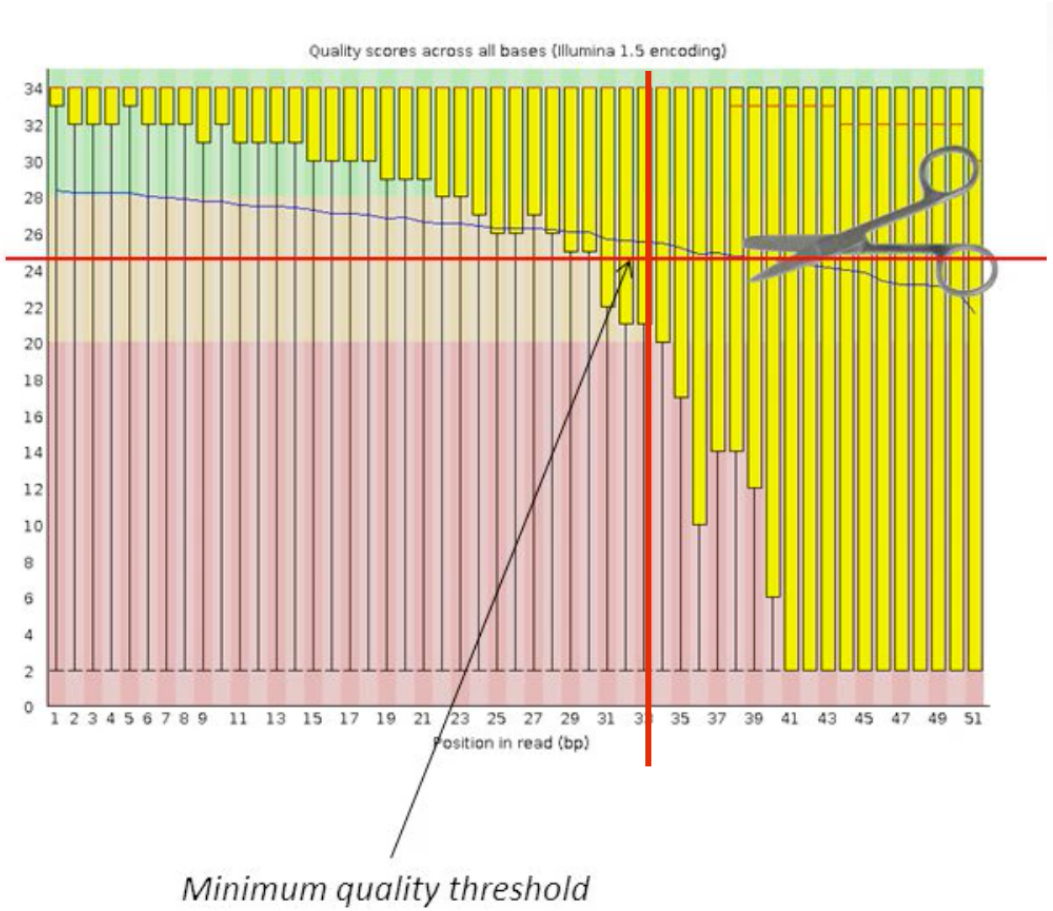
- Important to remove bad quality reads
- Downstream analysis might be improved



https://wiki.hpcc.msu.edu/download/attachments/15434467/fastqc-1.png

Interpreting FastQC

- Sequence filtering:
  - Mean quality
  - Read length
  - Read length after trimming
  - Percentage of bases above a quality threshold
  - Adapter trimming
  - Adapter reads



http://www.bioinformatics.babraham.ac.uk/
https://slideplayer.com/slide/5422676/