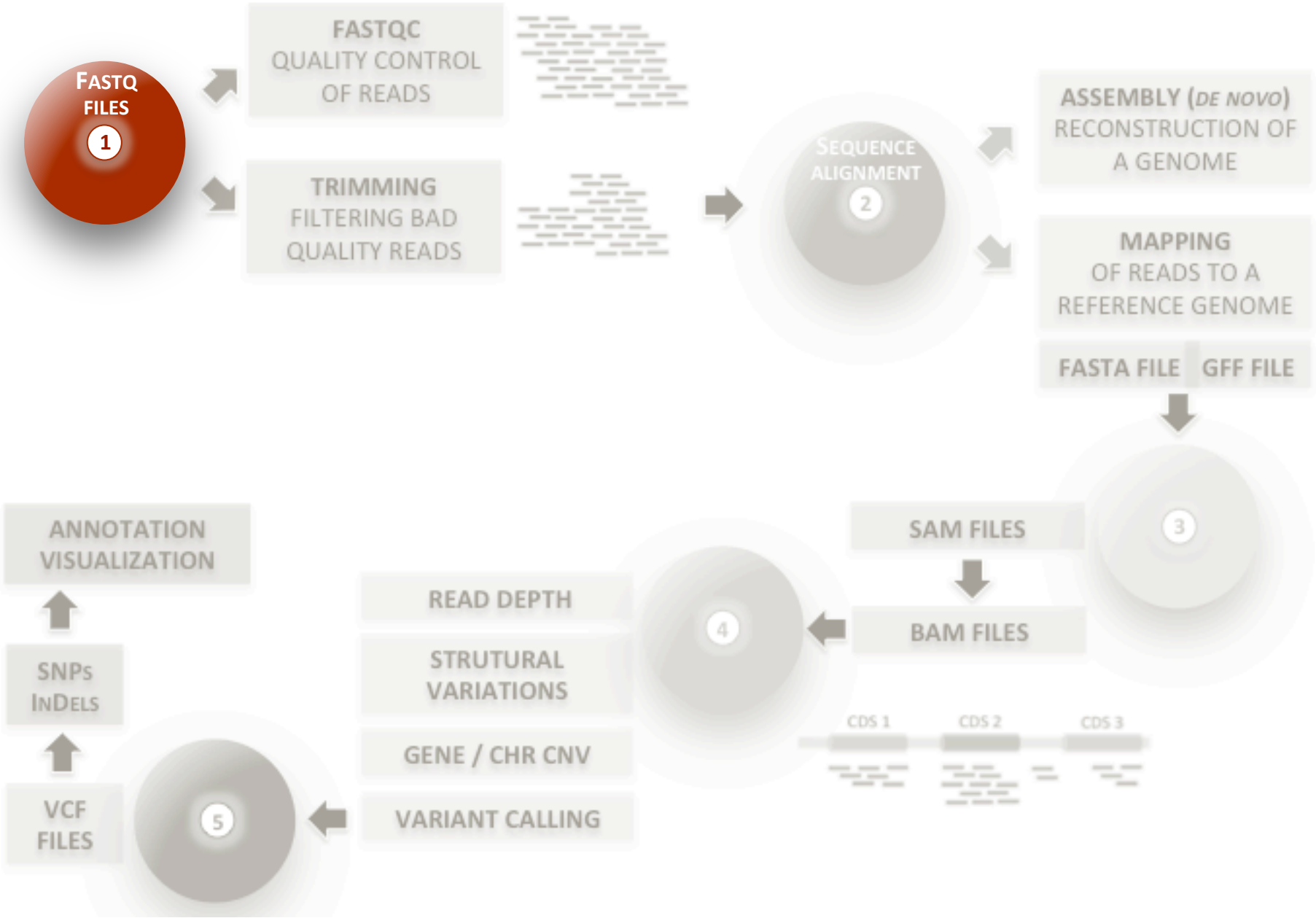
H3ABioNet
Pan African Bioinformatics Network for H3Africa

Introduction to NGS data analysis pipelines and file formats


FASTQ FILES
1
FASTQC
QUALITY CONTROL OF READS
TRIMMING
FILTERING BAD QUALITY READS
SEQUENCE ALIGNMENT
2
ASSEMBLY (DE NOVO)
RECONSTRUCTION OF A GENOME
MAPPING
OF READS TO A REFERENCE GENOME
FASTA FILE
GFF FILE
3
SAM FILES
BAM FILES
4
READ DEPTH
STRUTURAL VARIATIONS
GENE / CHR CNV
VARIANT CALLING
5
VCF FILES
SNPs InDels
ANNOTATION VISUALIZATION
CDS 1
CDS 2
CDS 3

Part I: Quality Control

FASTQ FILE FORMAT

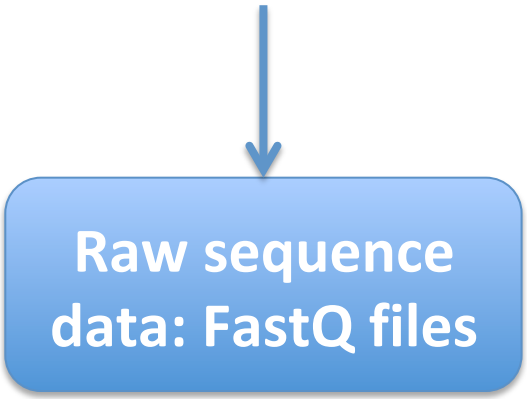
read file format

What is a FastQ file?

FASTQ= FASTA + Quality

FastQ format is a text-based format for storing both a biological sequence and the corresponding per base quality scores

-> Most common output provided by sequencing platforms

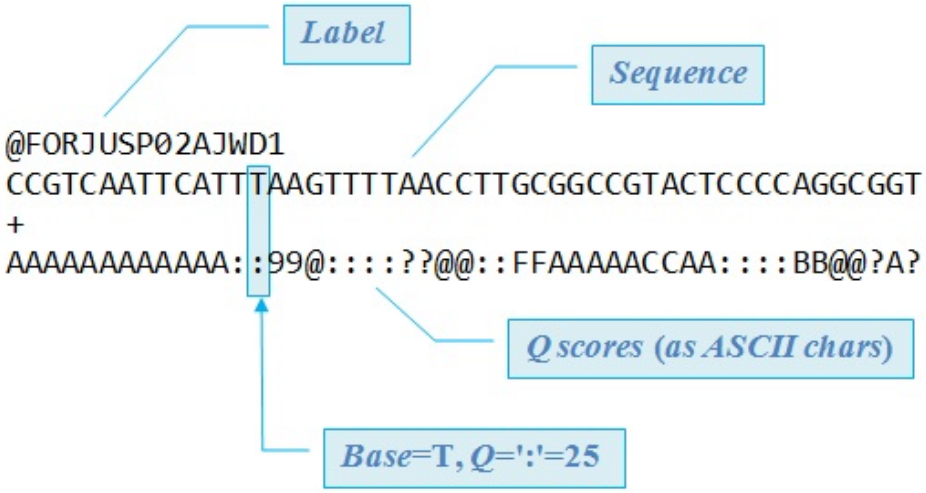


FASTQ FILE FORMAT

read file format

- A **FASTA** file uses at least 2 lines per sequence:
  - *Line 1* begins with a '>' and is followed by a sequence identifier
  - *Line 2* is the sequence letters

- A **FASTQ** file uses four lines per sequence:
  - *Line 1* begins with a '**@**' and is followed by a sequence identifier
  - *Line 2* is the **sequence letters**
  - *Line 3* begins with a '**+**' and is *rarely* followed by the sequence identifier
  - *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



FASTQ FILE FORMAT

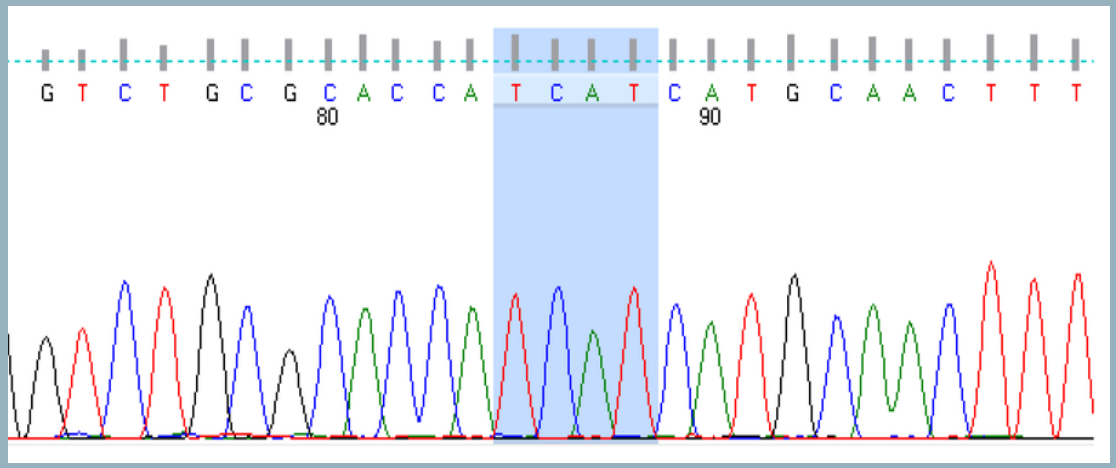
read file format

a FASTQ file is a file containing :

- reads sequences
- a quality score associated to each read position

-> Just as Sanger!

Reminder: Sanger Sequencing

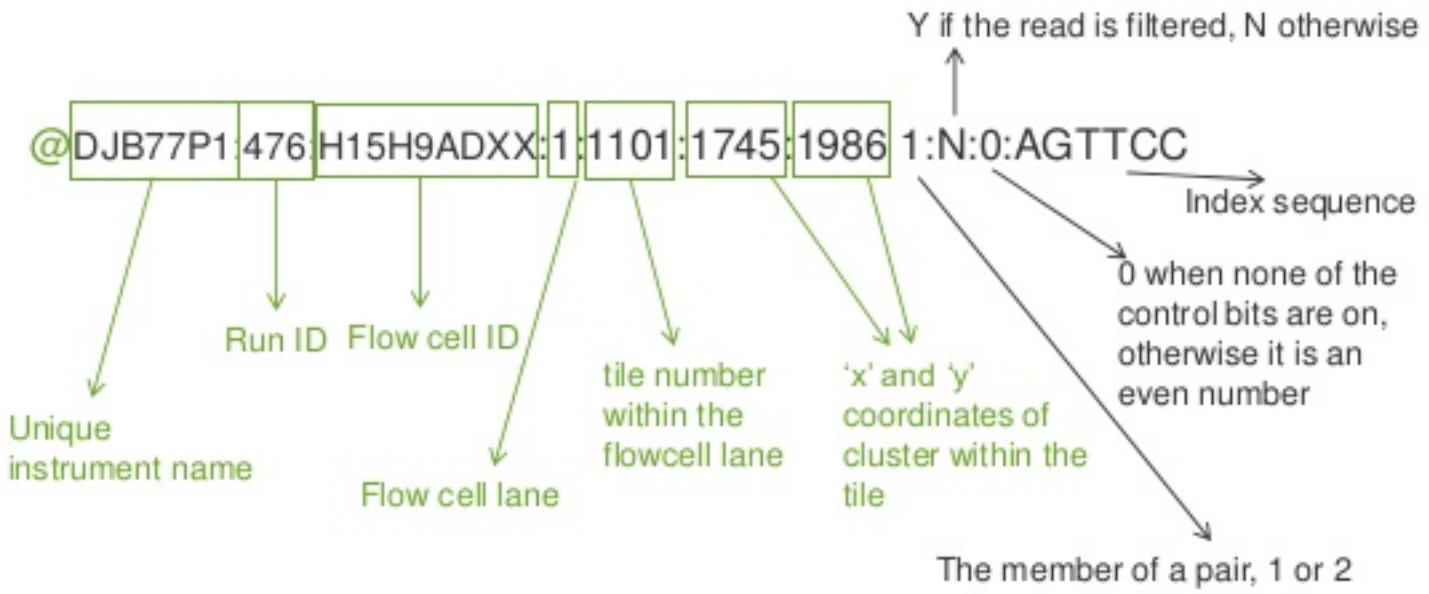


FASTQ FILE FORMAT

read file format

a FASTQ file is a file containing :

- reads sequences
- a quality score associated to each read position
- **standard read headers (very often)**



https://image.slidesharecdn.com/,data-management-for-quantitative-biology-data-sources-next-generation-technologies-apr-30-2015-dr-stefan-czemmel-27-638.jpg

FASTQ FILE FORMAT

Phred quality score

- The quality score of a base is called the **Phred score** (or **Q score**)
It is an integer value defined as a property that is logarithmically related to the base calling error probabilities (P)

$$Q = -10 \log_{10}(P) \Leftrightarrow P = 10^{-Q/10}$$

- Phred or Q scores are often represented as ASCII characters
Starting in Illumina 1.8, the quality scores have basically returned to the use of the Sanger format (Phred+33)

https://en.wikipedia.org/wiki/FASTQ_format#Quality

FASTQ FILE FORMAT

Phred quality score

FASTQ FILE FORMAT

Phred quality score

Note: Phred version is independent from the sequencing technology version

Illumina 1.8+

```
LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
  !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}
~
```

Phred score

```
|                        |    |        |                              |
|
0.........................26...31.......41
```

Table 1: Quality Scores and Base Calling Accuracy

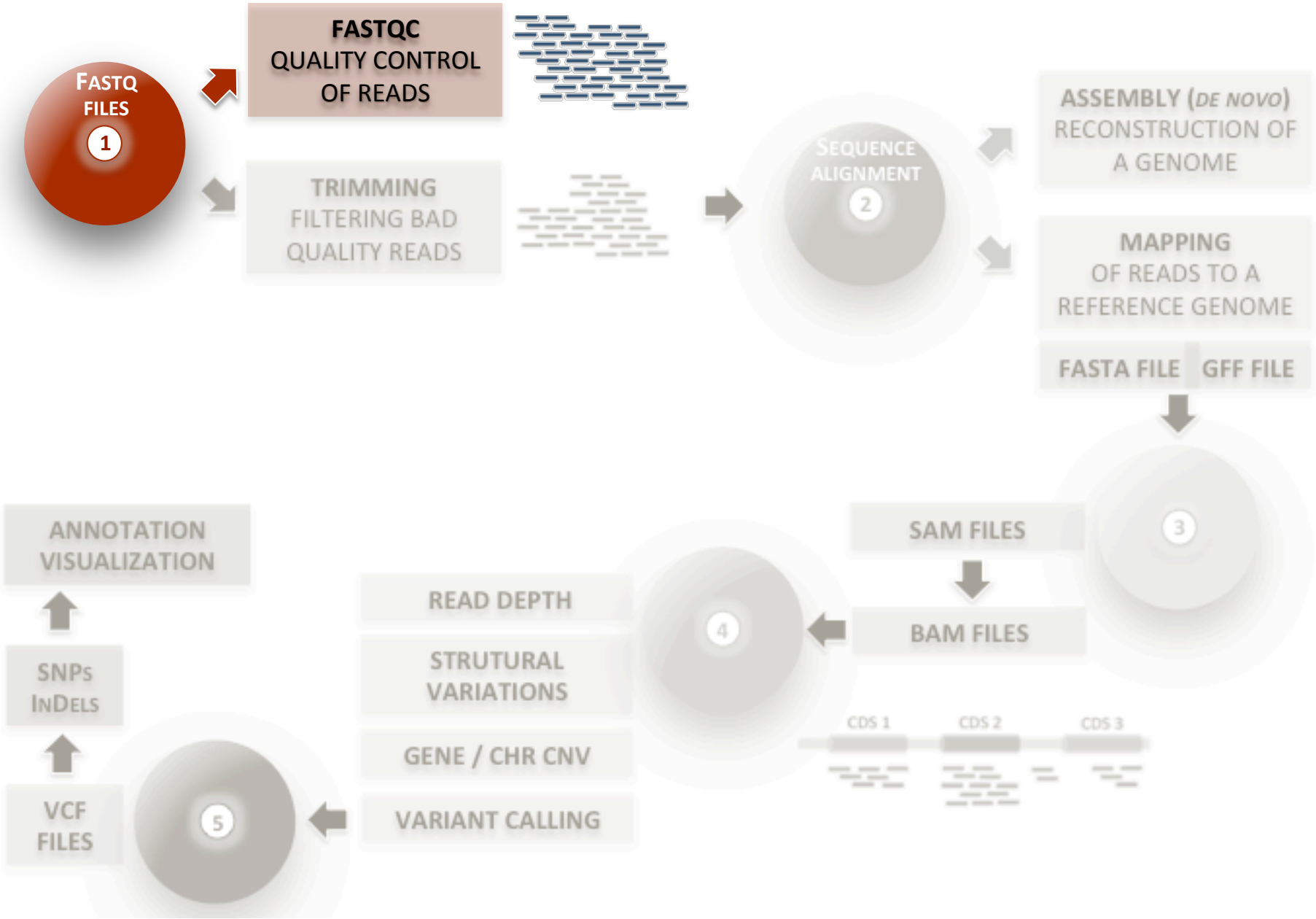
Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

Position Quality Score:

30–40: ✔ Good

<20: Discard

20 – 30: depends on overall quality


FASTQ FILES
1
FASTQC
QUALITY CONTROL OF READS
TRIMMING
FILTERING BAD QUALITY READS
SEQUENCE ALIGNMENT
2
ASSEMBLY (DE NOVO)
RECONSTRUCTION OF A GENOME
MAPPING
OF READS TO A REFERENCE GENOME
FASTA FILE
GFF FILE
3
SAM FILES
BAM FILES
4
READ DEPTH
STRUTURAL VARIATIONS
GENE / CHR CNV
VARIANT CALLING
5
VCF FILES
SNPs InDels
ANNOTATION VISUALIZATION
CDS 1
CDS 2
CDS 3

READS QUALITY CHECK

FastQC

Babraham Bioinformatics

FastQC: Quality Control for FastQ files

GUI, Command line, Available on Galaxy

Graphic reports

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

http://dnacore.missouri.edu/PDF/FastQC_Manual.pdf

READS QUALITY CHECK

FastQC

- Quality control checks on raw sequence data

Input: FastQ files (or BAM/SAM)
Output: Summary graphs allowing quick data assessment

- FastQC can be run in 2 modes:

Stand alone interactive application ← small number of samples
Non-interactive mode ← large number

FastQC can be integrated in ANALYSIS PIPELINES

READS QUALITY CHECK

FastQC

A summary of the modules which were run, and a quick evaluation of whether the results seem **entirely normal**, **slightly abnormal** or **very unusual**.

Normal

Reasonable

Unexpected

FastQC Report

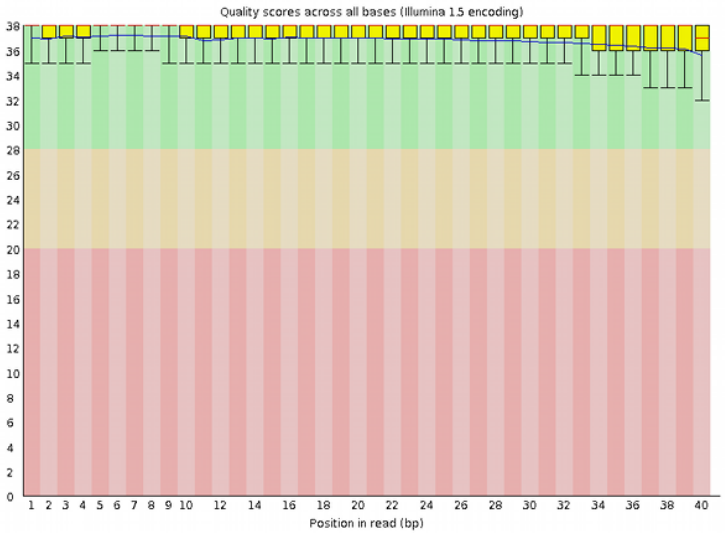
Summary

- 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

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

Per base sequence quality

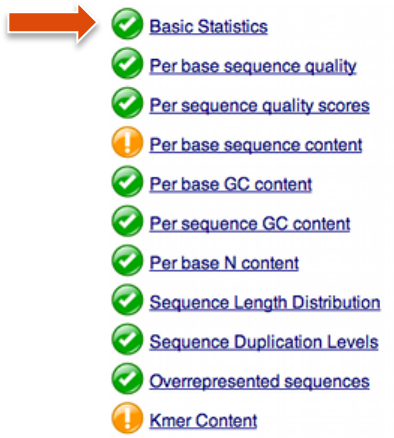


Good Quality

Poor Quality

READS QUALITY CHECK

FastQC



Basic Statistics

ASCII quality encoding format (*Phred* format)
Number of reads, Read length, …

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

READS QUALITY CHECK

FastQC

- 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

Overview of the range of quality values (*Phred* scores) across all bases from all reads

Upper and Lower whiskers: 10% and 90% points

Red line: **median** value

Blue line: **mean** quality

Yellow box: **inter-quartile** range (25-75%)

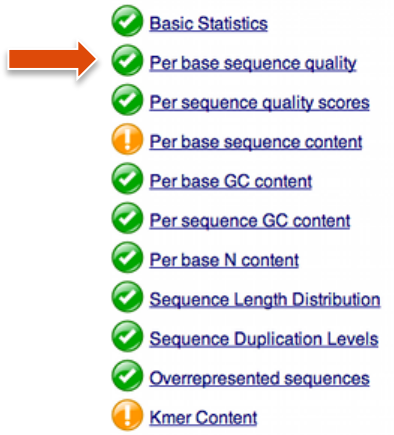
In most platforms, the quality is lower at the beginning/**end** of reads

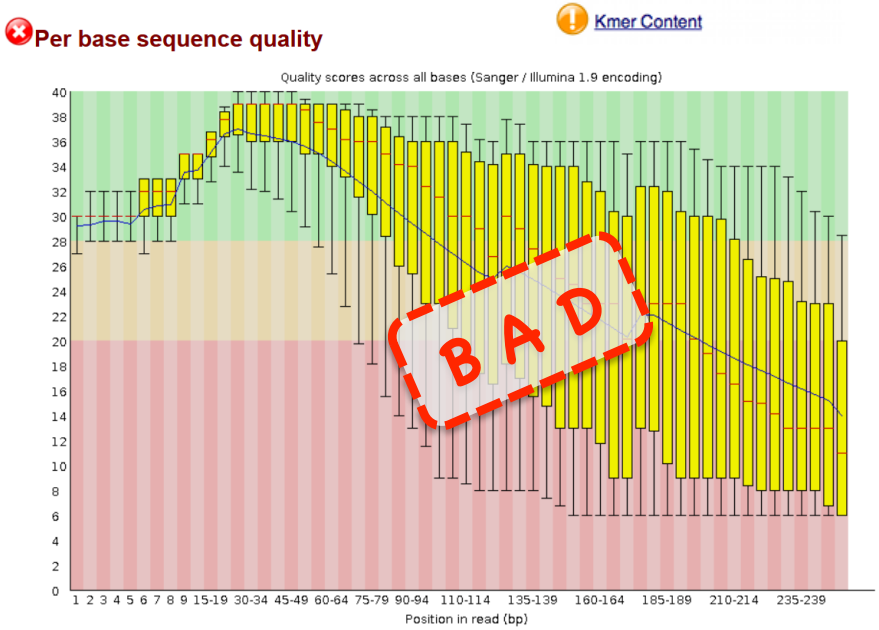
This is common ! Short loss of quality early in the run and **quality degrades as the run progresses**

READS QUALITY CHECK

FastQC

Per base sequence quality

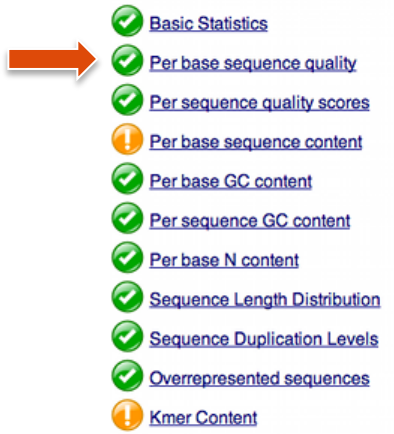




READS QUALITY CHECK

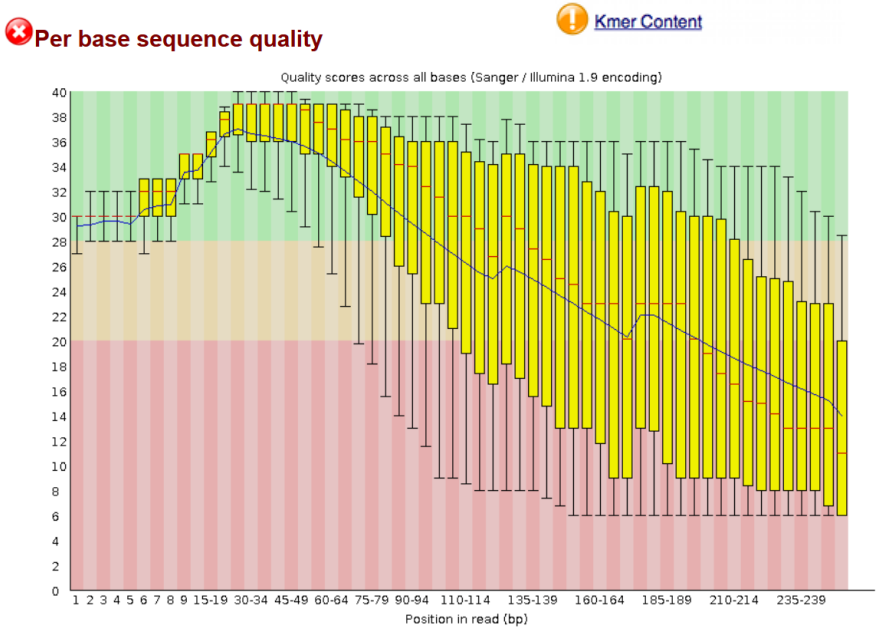
FastQC

Per base sequence quality



- High proportion of bad quality bases
- High variability
- High decrease in terms of quality towards the end of the read

→ Important to remove bad quality data

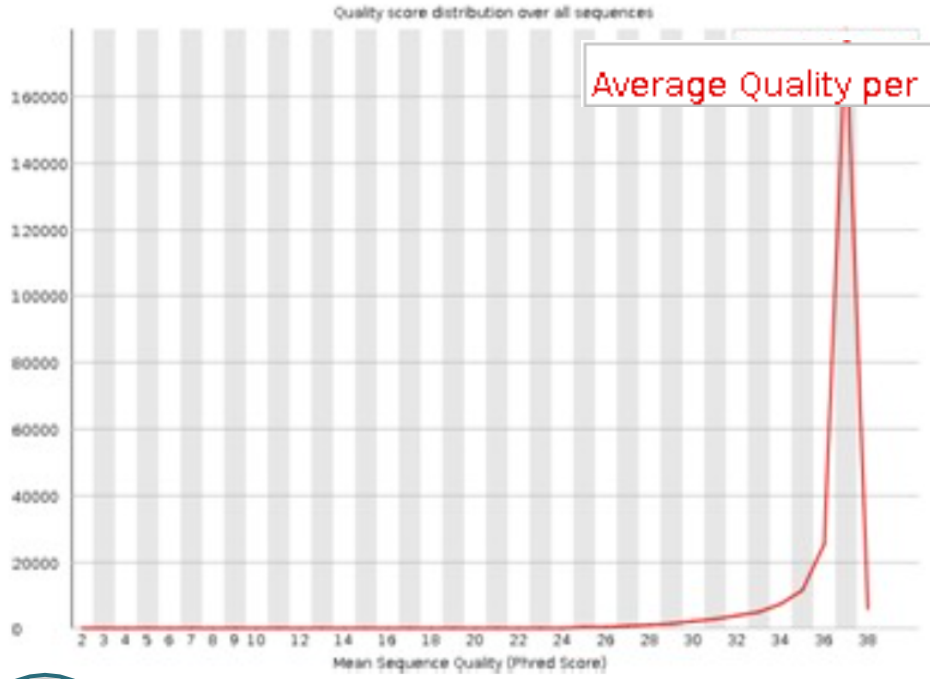


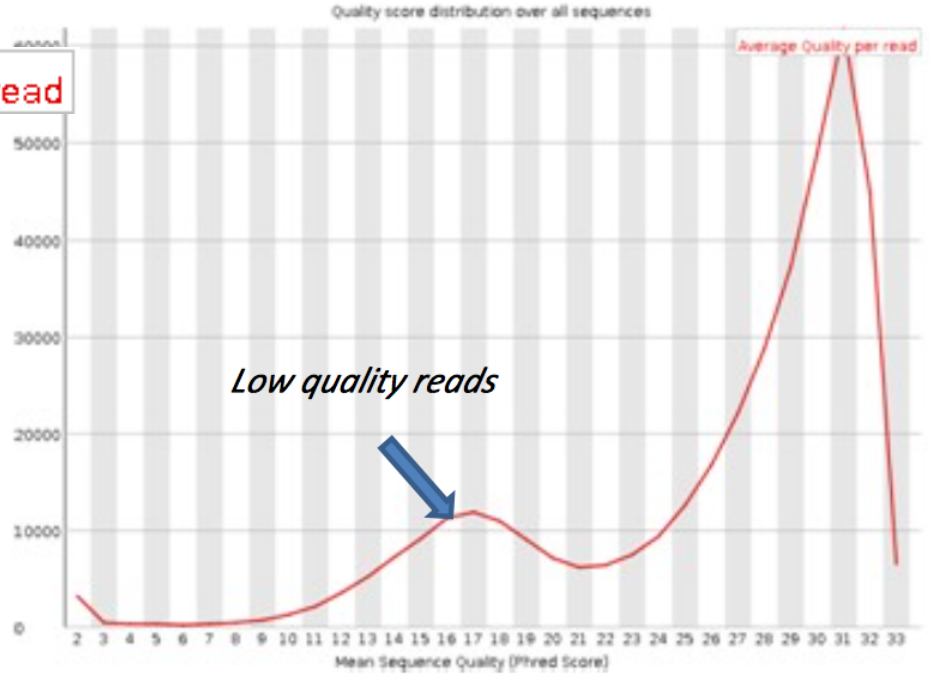
READS QUALITY CHECK

FastQC

- 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





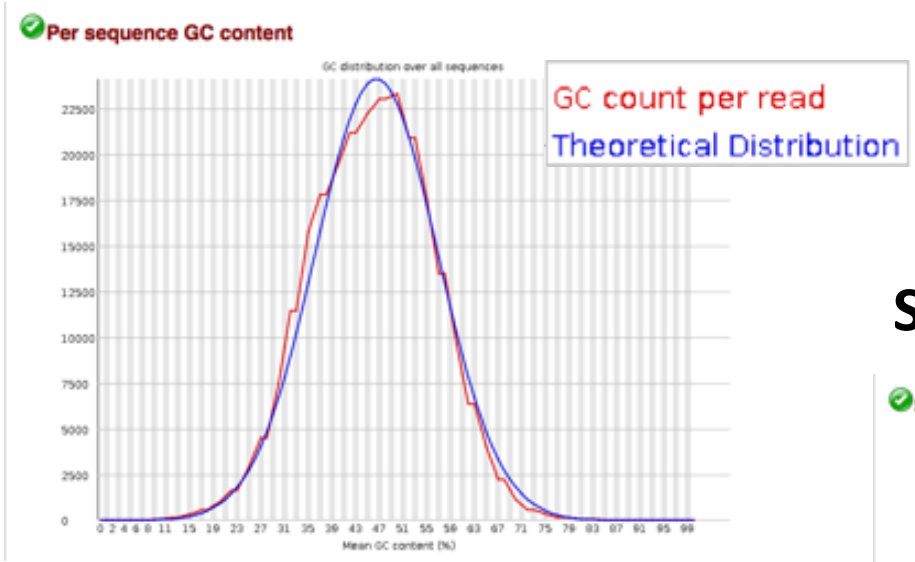
PART 2

READS QUALITY CHECK

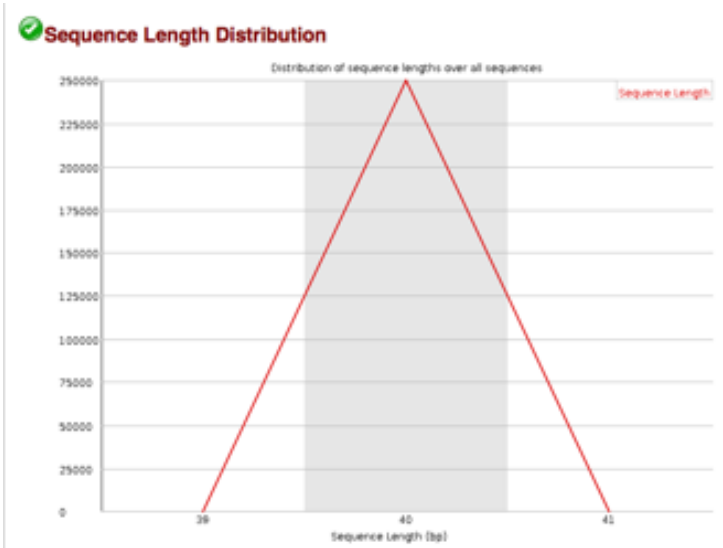
FastQC

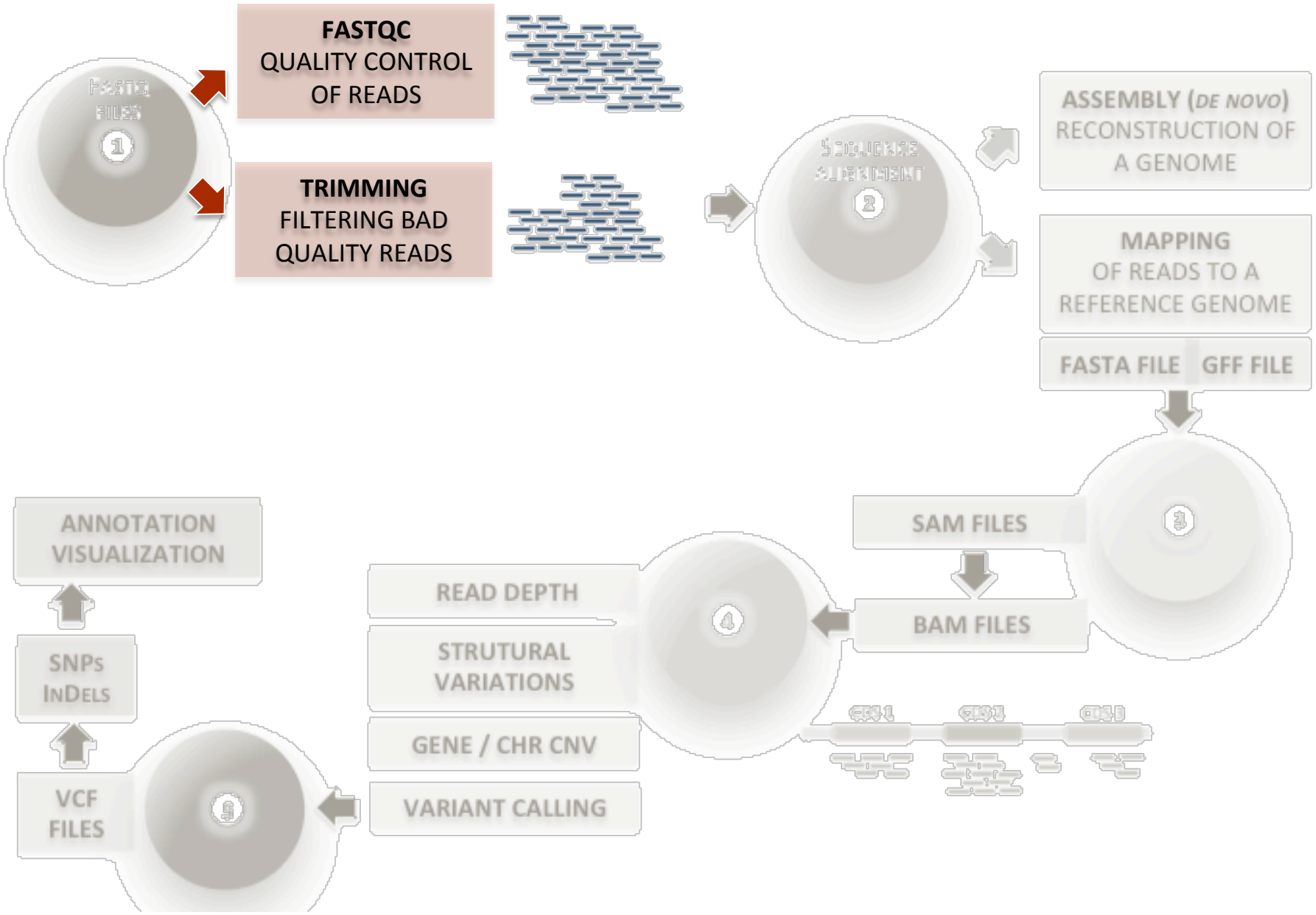
- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content

Per Sequence GC Content



Sequence Length Distribution




FASTQ
QUALITY CONTROL
OF READS
TRIMMING
FILTERING BAD
QUALITY READS
ASSEMBLY (DE NOVO)
RECONSTRUCTION OF
A GENOME
MAPPING
OF READS TO A
REFERENCE GENOME
FASTA FILE
GFF FILE
SAM FILES
BAM FILES
READ DEPTH
STRUTURAL
VARIATIONS
GENE / CHR CNV
VARIANT CALLING
VCF
FILES
SNPs
INDELS
ANNOTATION
VISUALIZATION

READ PRE-PROCESSING

Trimming low quality bases

Trimming low Quality Bases:

- Remove low quality bases, identified by the probability that they are called incorrectly.

- Widely but heterogeneously applied

- Keep in mind! Trimming might impact downstream analyses

Trimming is more important for denovo assembly, check this study

(Williams et al., 2016)

READ PRE-PROCESSING

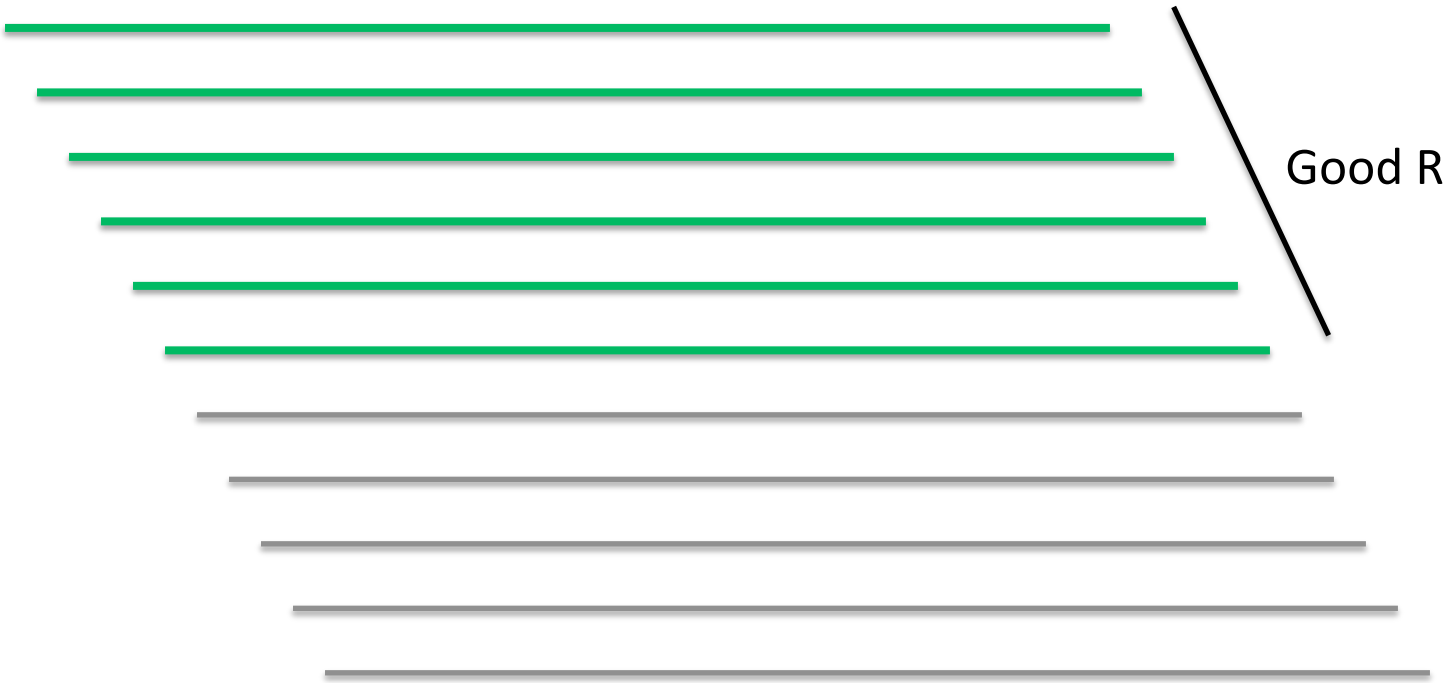
Trimming low quality bases

Raw reads sequences

READ PRE-PROCESSING

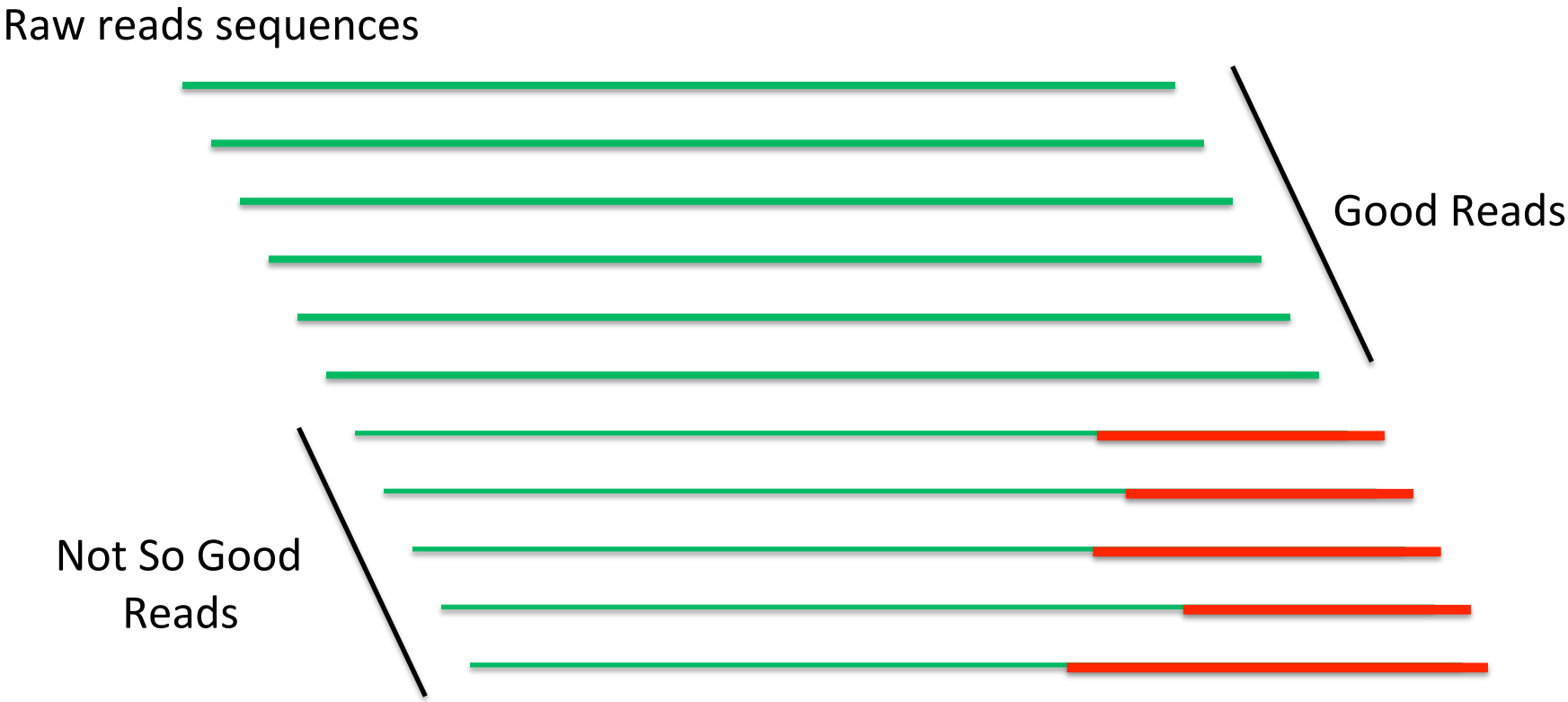
Trimming low quality bases

Raw reads sequences



READ PRE-PROCESSING

Trimming low quality bases

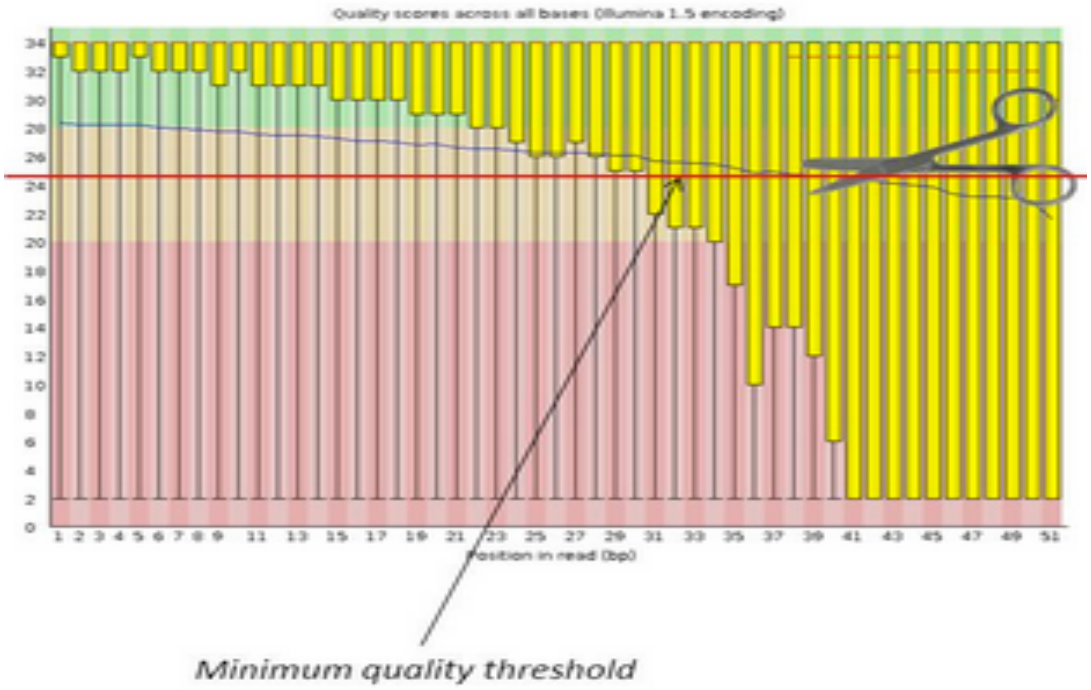


READ PRE-PROCESSING

Trimming low quality bases

Our example sample analysed using FastQC:

- reads of 51bp
- average quality = 26



=> Poor quality in the ends of reads can be removed using reads **filtering** or reads **trimming** tools

READ PRE-PROCESSING

Trimming low quality bases

1) **READS FILTERING**

2) READS TRIMMING

3) ADAPTIVE TRIMMING

4) ADAPTIVE TRIMMING FOLLOWED BY **READ FILTERING**

READ PRE-PROCESSING

Trimming low quality bases

Reads with poor quality ends can be discarded using **FILTERING** tools

READ PRE-PROCESSING

Trimming low quality bases

Reads with poor quality ends can be discarded using **FILTERING** tools

AFTER FILTERING:

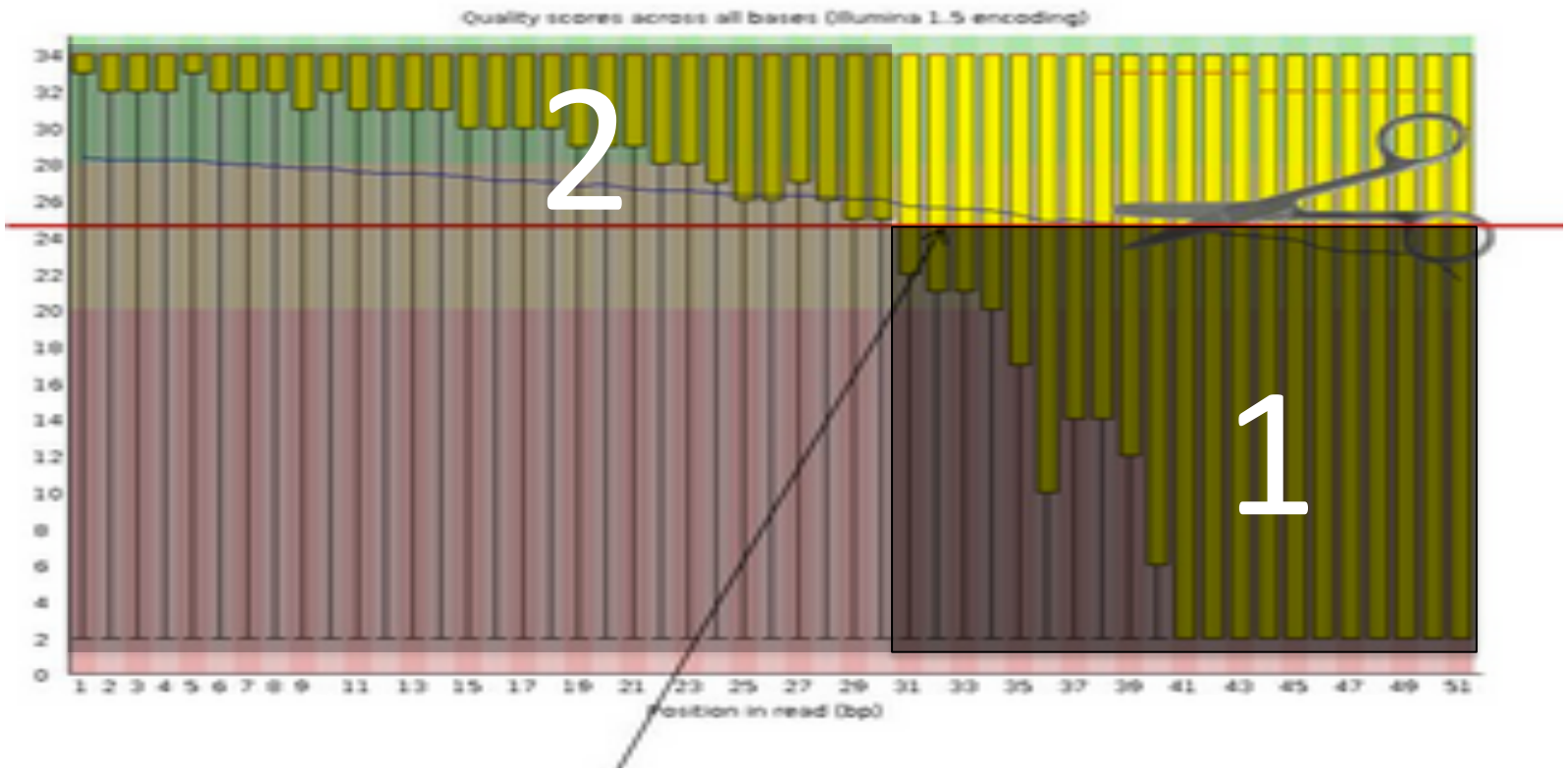
Sample Overall Quality ↑

Number of Reads ↓

READ PRE-PROCESSING

Trimming low quality bases

FastQC graph



Reads with poor quality ends are removed from the dataset (1)

But some good quality positions are lost (2)

READ PRE-PROCESSING

Trimming low quality bases

1) READS FILTERING

2) **READS TRIMMING**

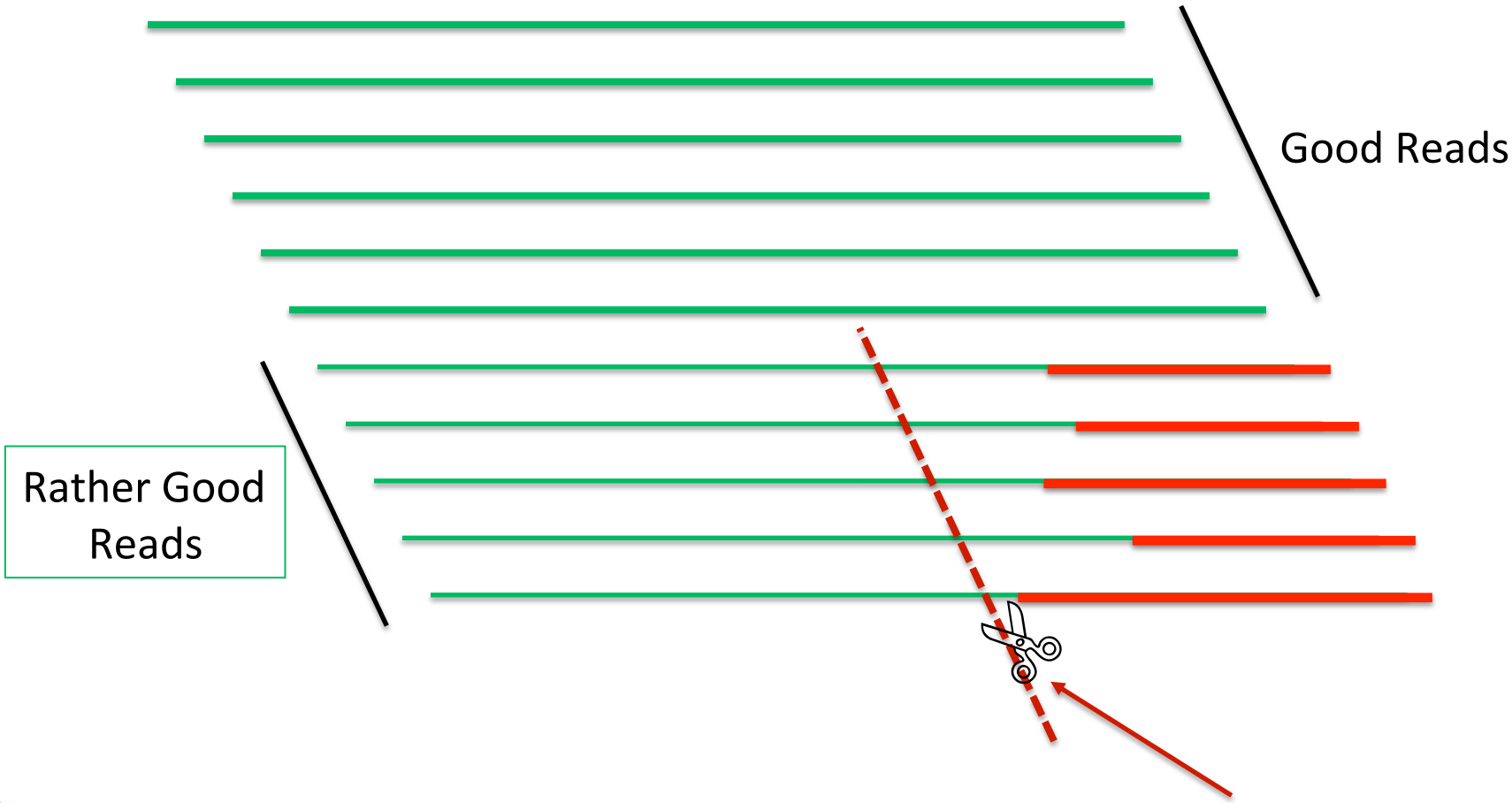
3) ADAPTIVE TRIMMING

4) ADAPTIVE TRIMMING FOLLOWED BY **READ FILTERING**

READ PRE-PROCESSING

Trimming low quality bases

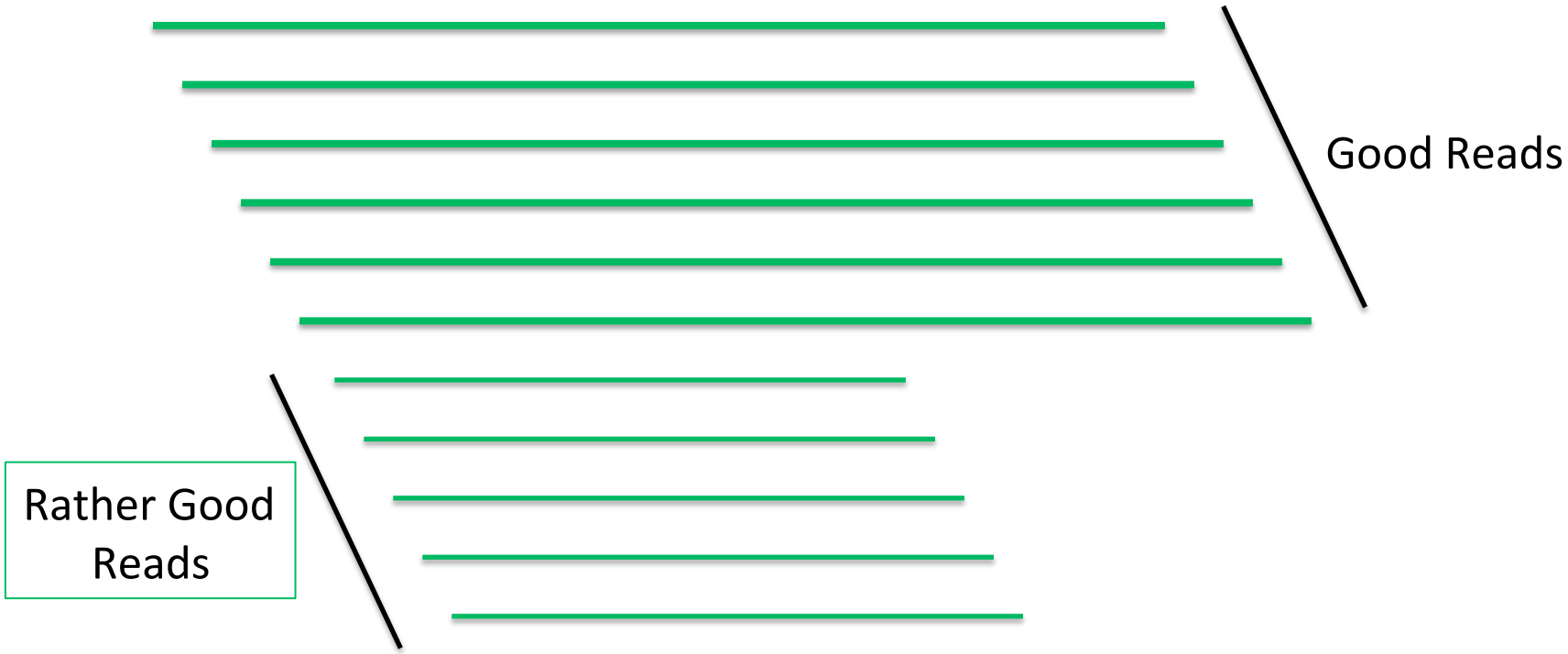
Poor quality ends of reads can be cut using **TRIMMING** tools



READ PRE-PROCESSING

Trimming low quality bases

Poor quality ends of reads can be cut using **TRIMMING** tools

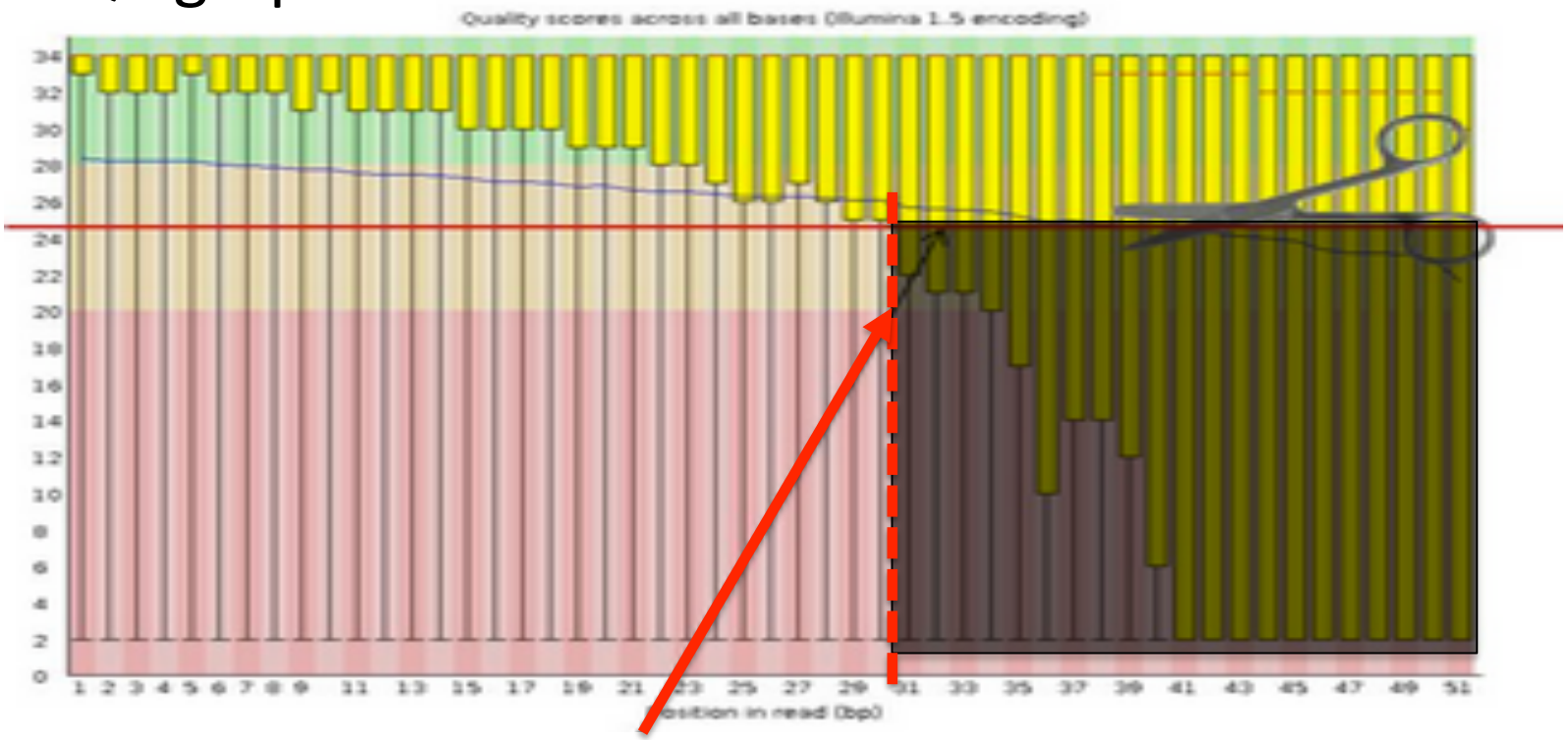


Trimmed reads are of better quality than raw reads but they are shorter!

READ PRE-PROCESSING

Trimming low quality bases

FastQC graph



Cut-off threshold Length = 30 bp

Reads are now of different lengths

Number of raw reads = number of raw reads

READ PRE-PROCESSING

Trimming low quality bases

1) READS FILTERING

2) READS TRIMMING

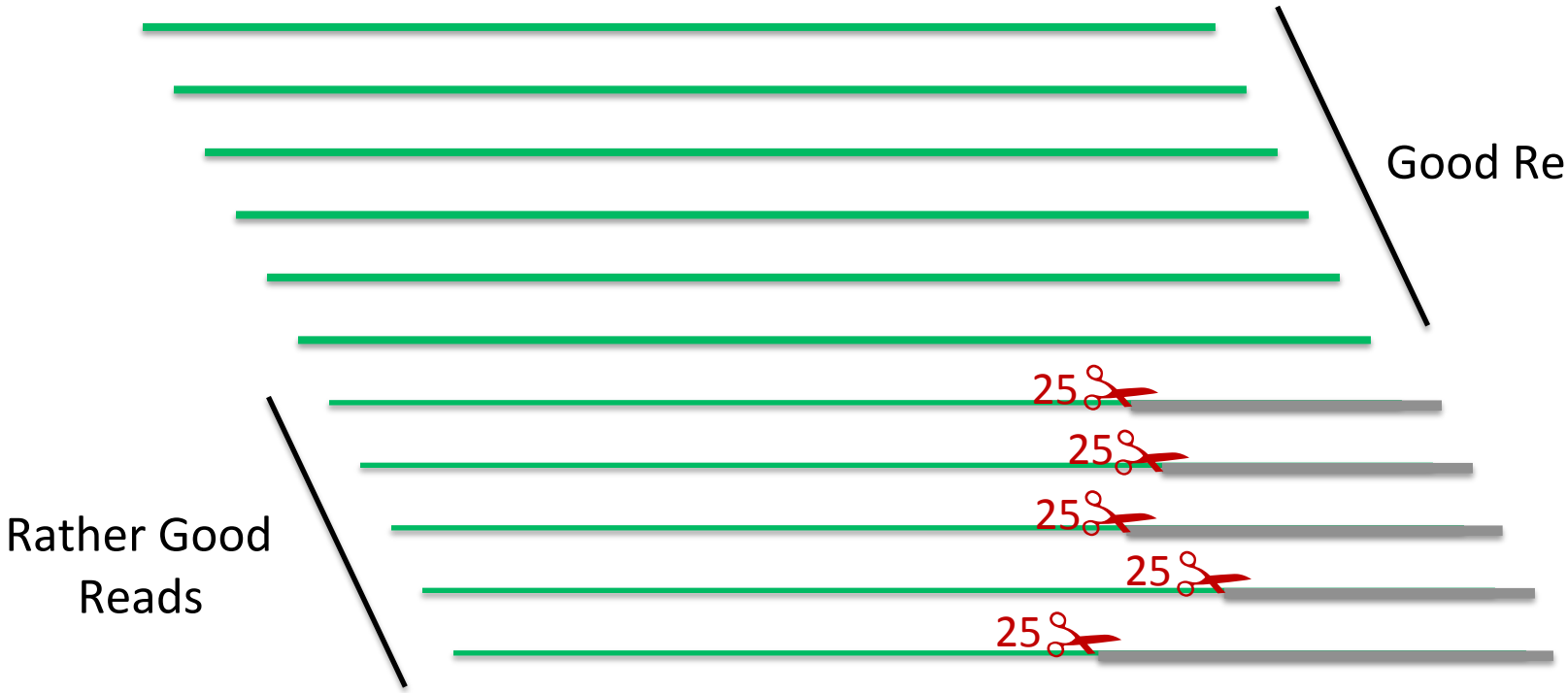
3) **ADAPTIVE TRIMMING**

4) ADAPTIVE TRIMMING FOLLOWED BY **READ FILTERING**

READ PRE-PROCESSING

Trimming low quality bases

A more precise **TRIMMING** approach



READ PRE-PROCESSING

Trimming low quality bases

A more precise **TRIMMING** approach

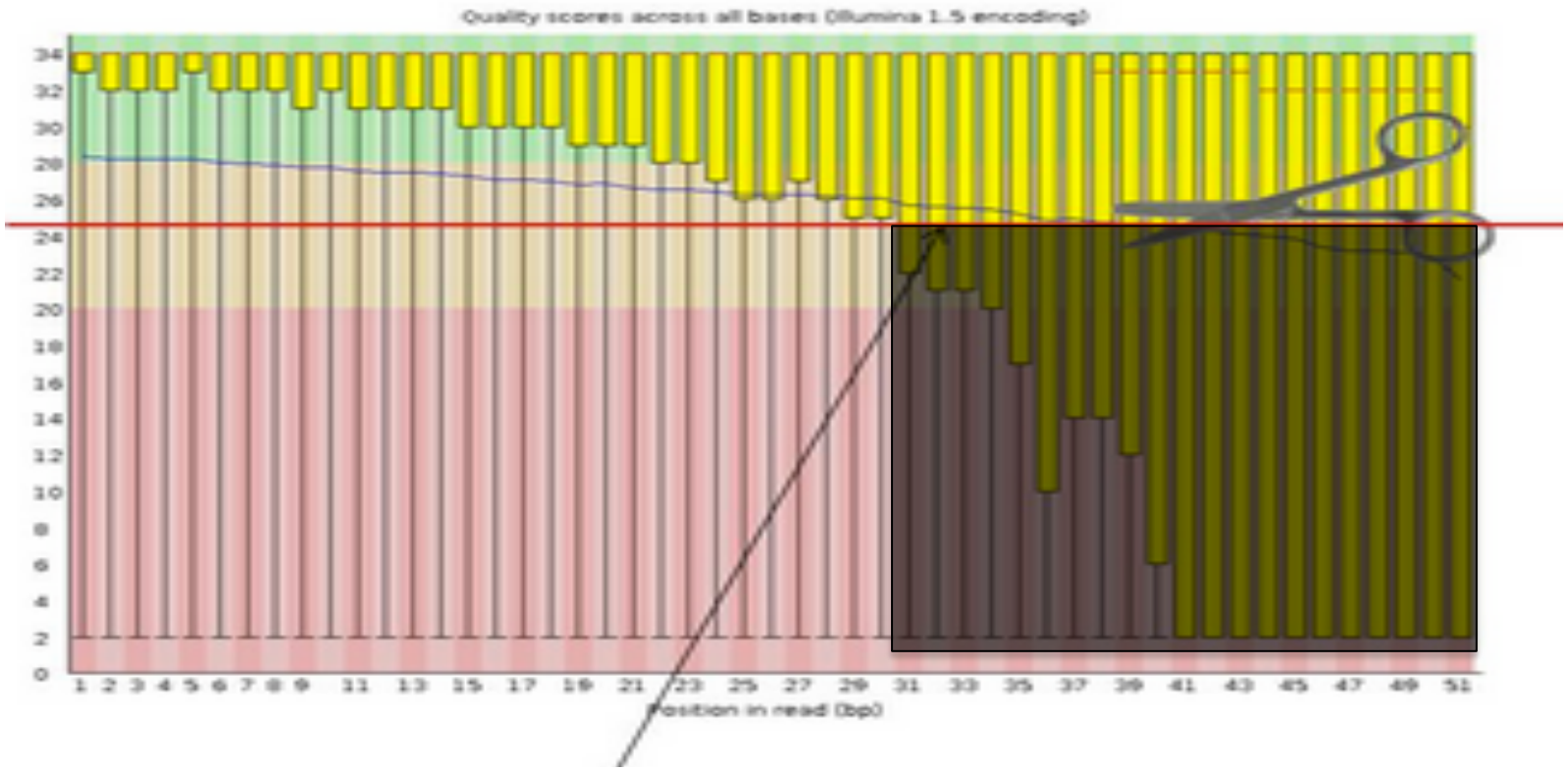
Trimmed reads are of better quality than raw reads but they are of **variable** lengths.

Some reads can be extremely short!

READ PRE-PROCESSING

Trimming low quality bases

FastQC graph



Poor quality ends are removed from the dataset

But the good quality positions are kept (yeah!)

READ PRE-PROCESSING

Trimming low quality bases

1) READS FILTERING

2) READS TRIMMING

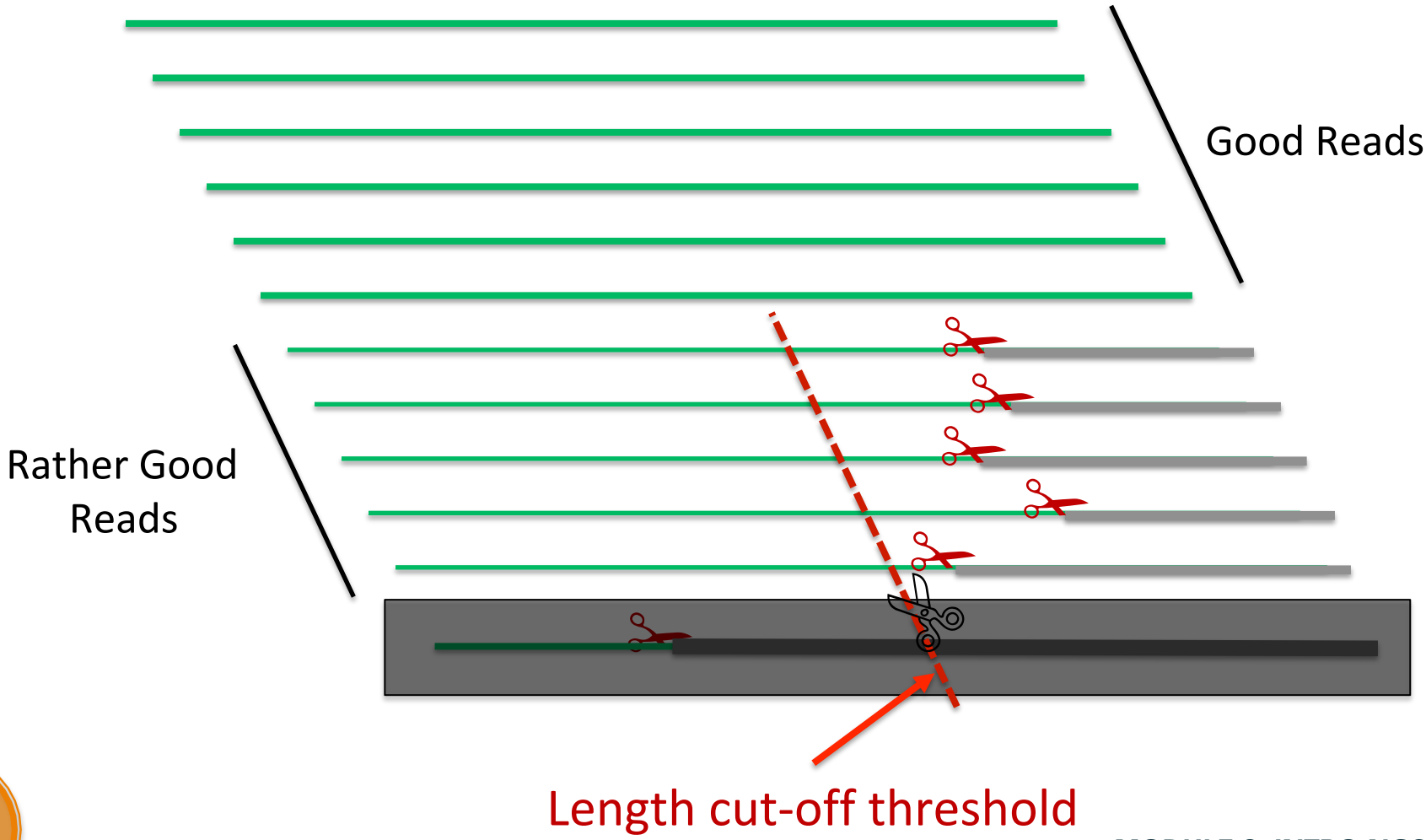
3) ADAPTIVE TRIMMING

4) **ADAPTIVE TRIMMING** FOLLOWED BY **READ FILTERING**

READ PRE-PROCESSING

Trimming low quality bases

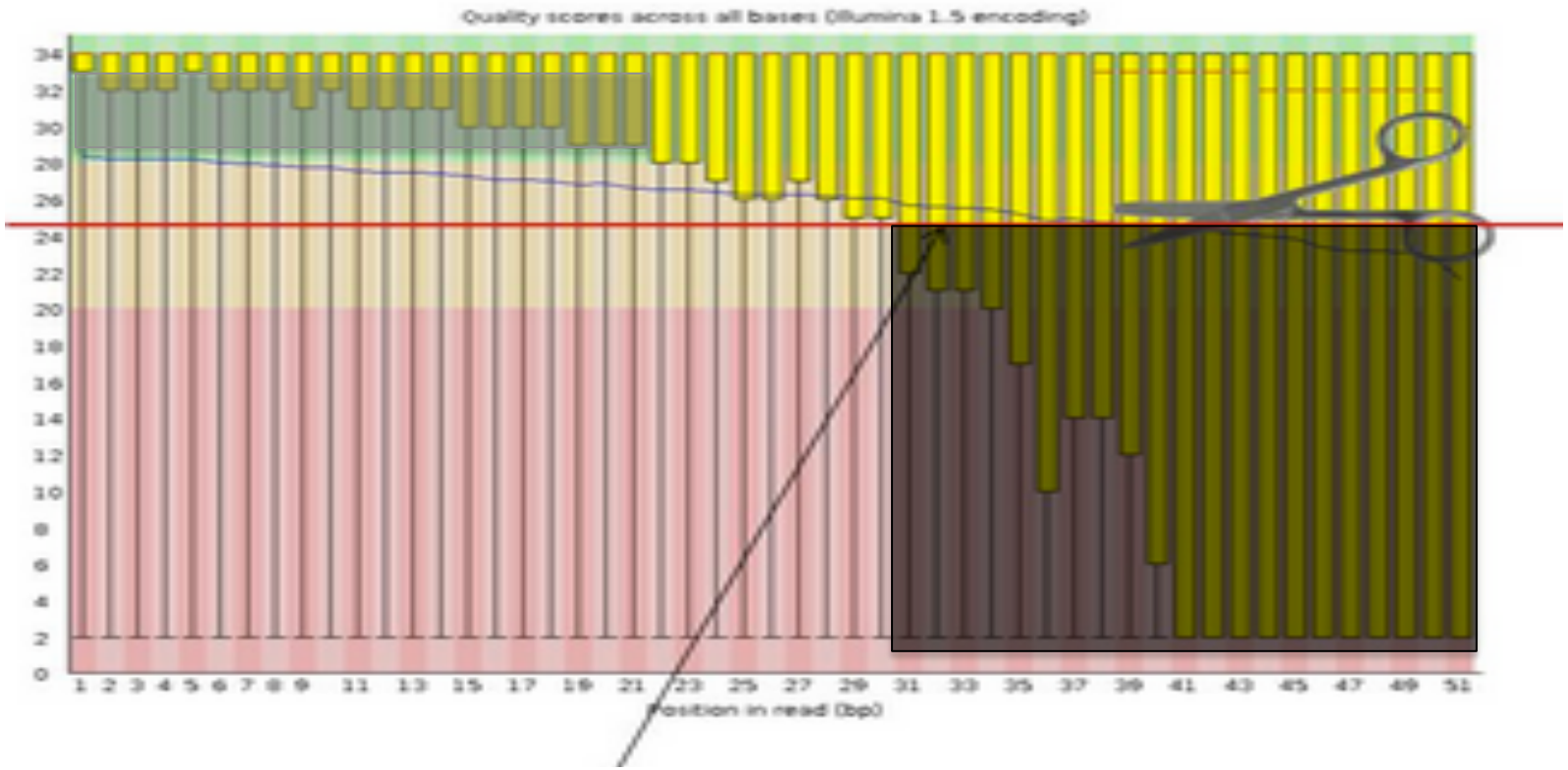
Poor quality in the end can be cut using **TRIMMING AND FILTERING** tools



READ PRE-PROCESSING

Trimming low quality bases

FastQC graph



Poor quality ends are removed from the dataset

But the good quality positions are kept (yeah!)

READ PRE-PROCESSING

Trimming low quality bases

- Trimming low Quality Bases

Low quality base reads from the sequencer can cause an otherwise mappable sequence not to align. There are a number of open source tools that can trim off 3' bases and produce a FASTQ file of the trimmed reads to use as input to the alignment program.

Manipulating tools

FASTX-Toolkit provides a set of command line tools for manipulating fasta and fastq files. The available modules include a fastx_trimmer utility for trimming fastq sequences (and quality score strings) before alignment.

```
gunzip -c Sample.fastq.gz | fastx_trimmer -l 50 -Q 33 > Sample_trimmed.fq
```

Trim down to 50 bases (last base is 50)

option that specifies how base qualities on the 4th line of each fastq entry are encoded

https://wikis.utexas.edu/display/bioiteam/Evaluating+your+raw+sequencing+data#Evaluatingyourrawsequencingdata-Trimminglowqualitybases

READ PRE-PROCESSING

Trimming adapters

- Adapters Trimming

Adaptors contaminations can lead to alignemnt errors and a big number of unaligned reads since the adaptor sequences are synthetic and don't occur in the genomic sequence.
Unlike general fixed-length trimming, adapter trimming removes differing numbers of 3' bases depending on where the adapter sequence is found.

Manipulating tools
Cutadapt program is an excellent tool for removing adapter contamination.
Ex cutadapt on on a single-end read sample:

```
cutadapt -m 22 -O 33 -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC
```

discard any sequence that is smaller than 22 bases after trimming

No trimming of 3' adapter sequences unless at least the first 10 bases of the adapter are seen at the 3' end of the read

https://wikis.utexas.edu/display/bioiteam/Evaluating+your+raw+sequencing+data#Evaluatingyourrawsequencingdata-Trimminglowqualitybases

READ PRE-PROCESSING

Tools

Trimming sequencing adapters/primers (clipping) **and poor quality ends** and **filter short reads**

Dozens of tools available!

AlienTrimmer, Cutadapt, ConDeTri, FastX, Sickle, SolexaQA, Trimmomatic…

They use different methods and algorithms and offer at lot of options

OPEN ACCESS Freely available online PLOS ONE

An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis

Cristian Del Fabbro[1☯], Simone Scalabrin[2☯], Michele Morgante[1], Federico M. Giorgi[1,3*]

1 Institute of Applied Genomics, Udine, Italy, 2 IGA Technology Services, Udine, Italy, 3 Center for Computational Biology and Bioinformatics, Columbia University, New York, New York, United States of America

If PE reads: Filter unpaired reads

READ PRE-PROCESSING

Tools

Trimming Poor Quality Ends

A few examples from Del Fabbro *et. Al.*

OPEN ACCESS Freely available online PLOS ONE

An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis

Cristian Del Fabbro[1¤], Simone Scalabrin[2¤], Michele Morgante[1], Federico M. Giorgi[1,3*]

1 Institute of Applied Genomics, Udine, Italy, 2 IGA Technology Services, Udine, Italy, 3 Center for Computational Biology and Bioinformatics, Columbia University, New York, New York, United States of America

- RNA-Seq (*Homo sapiens*)
  - "*SolexaQA* achieves the highest quality while keeping the highest amount of reads"
- Assembly (*Prunus persica*)
  - "Read trimming affects only partially genome assembly results"
  - "Stringent trimming tends to heavily remove data and decrease overall assembly quality"
- SNP Identification (*Prunus persica* and *Saccharomyces cerevisiae*)
  - "All trimmers drastically reduce the percentage of alternative allele nucleotides (…) bringing this false positive call indicator from 30% to 10%"

There is no golden method / tool

→ DEPENDS ON THE APPLICATION