
# 02. Reads Quality Control

Erik Dassi & Davide Bressan

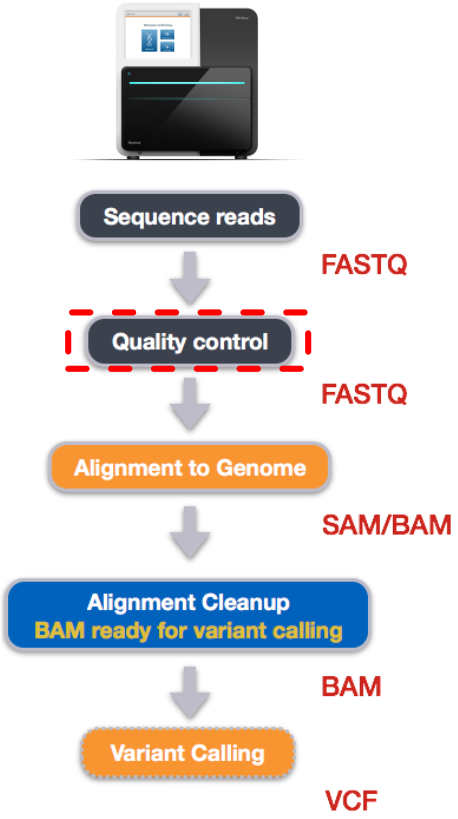
Genomics Technologies Lab

# Raw sequencing data

- Once your biological sample has been sequenced, you end up with millions of short DNA fragments (raw sequencing data). To identify variants in your sequencing data, you need several analysis steps.

- The **FIRST** step is to check the **quality** of your sequencing data, and we refer to this as QUALITY CONTROL.

- The reads are stored in a fastq file, which is usually compressed in a fastq.gz to reduce storage size.



# FASTA files

- **FASTA** format is a text-based format for representing either nucleotide sequences or peptide sequences, in which base pairs or amino acids are represented using single-letter codes

- A sequence in FASTA format begins with a single-line description, that starts with the greater-than (">") symbol, followed by lines of sequence data

Example:

```
>P01013 GENE X PROTEIN (OVALBUMIN-RELATED)
QIKDLLVSSSTDLDTTLVLVNAIYFKGMWKTAFNAEDTREMPFHVTKQESKPVQMMCMNNSFNVATLPAE
KMKILELPFASGDLSMLVLLPDEVSDLERIEKTINFEKLTEWTNPNTMEKRRVKVYLPQMKIEEKYNLTS
VLMALGMTDLFIPSANLTGISSAESLKISQAVHGAFMELSEDGIEMAGSTGVIEDIKHSPESEQFRADHP
FLFLIKHNPTNTIVYFGRYWSP
```

| | | | | | |
|---|---|---|---|---|---|
| A | adenosine | C | cytidine | G | guanine |
| T | thymidine | N | A/G/C/T (any) | U | uridine |
| *K* | *G/T (keto)* | *S* | *G/C (strong)* | *Y* | *T/C (pyrimidine)* |
| *M* | *A/C (amino)* | *W* | *A/T (weak)* | *R* | *G/A (purine)* |
| *B* | *G/T/C* | *D* | *G/A/T* | *H* | *A/C/T* |
| *V* | *G/C/A* | - | *gap of indeterminate length* | | |

| | | | |
|---|---|---|---|
| A | alanine | P | proline |
| B | aspartate/asparagine | Q | glutamine |
| C | cystine | R | arginine |
| D | aspartate | S | serine |
| E | glutamate | T | threonine |
| F | phenylalanine | U | selenocysteine |
| G | glycine | V | valine |
| H | histidine | W | tryptophan |
| I | isoleucine | Y | tyrosine |
| K | lysine | Z | glutamate/glutamine |
| L | leucine | X | any |
| M | methionine | * | translation stop |
| N | asparagine | - | *gap of indeterminate length* |

# FASTQ files

FASTQ format is a text-based format for storing both a biological sequence (usually nucleotide sequence) and its corresponding quality scores (FASTA with quality)

- It uses four lines for sequence:
- The first begins with a '@' character and is followed by a sequence identifier and an optional description
- The second is the raw sequence letters (or processed sequence coming from trimming)
- The third begins with a '+' character and is *optionally* followed by the same sequence identifier (and any description) again
- The fourth encodes the base call quality values for the sequence in line 2 (these are Phred +33 encoded, using ASCII characters), and must contain the same number of symbols as letters in the sequence

Example

```
@K00193:38:H3MYFBBXX:4:1101:10003:44458/1
TTCCTTATGAAACAGGAAGAGTCCCTGGGCCCAGGCCTGGCCCACGGTTGTCAAGGCACATCATTGCCAGCA
+
AAFFFKKKKKKKKKKKKKKKKKKKKKKKKKFKKFKKKKF<AAKKKKKKKKKKKKKKKKKFKKKFKKKKKKKKKKKKFKA
```

# Reads Quality Control

- Raw reads (FASTQ):
  a. Quality Control (FASTQC)
  b. Trimming – Adapter removal
  c. Quality Control (FASTQC)

# FASTQC

A quality control tool for high throughput sequence data.

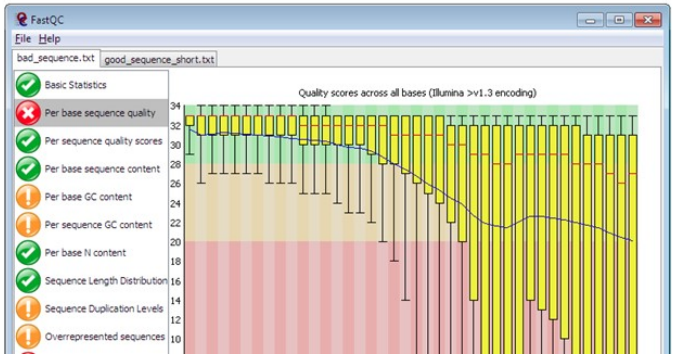
Babraham Bioinformatics

About | People | Services | Projects | Training | Publications

**FastQC**

| | |
|---|---|
| **Function** | A quality control tool for high throughput sequence data. |
| **Language** | Java |
| **Requirements** | A suitable Java Runtime Environment<br>The Picard BAM/SAM Libraries (included in download) |
| **Code Maturity** | Stable. Mature code, but feedback is appreciated. |
| **Code Released** | Yes, under GPL v3 or later. |
| **Initial Contact** | Simon Andrews |

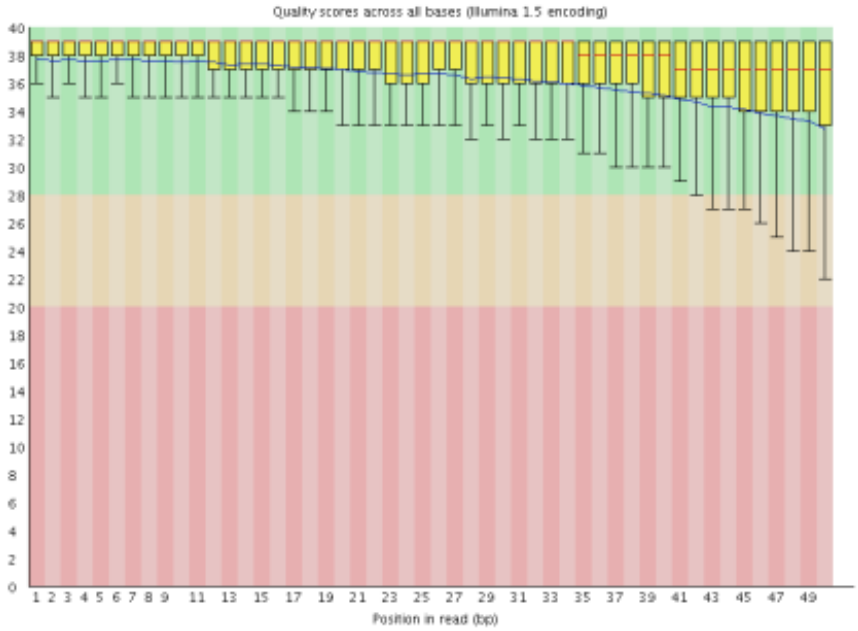
**Download Now**



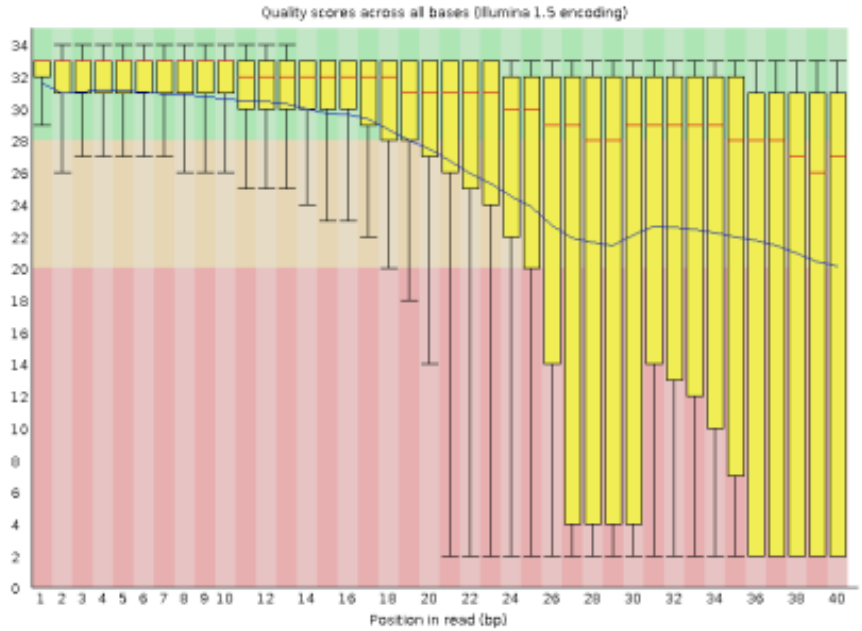

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

# FASTQC

## Single-base quality assessment




Good reads quality!




Poor reads quality...

# FASTQC

In addition to the quality of each sequenced base, FastQC will give you an idea of :

- Per base sequence content

- Presence and abundance of contaminating sequences

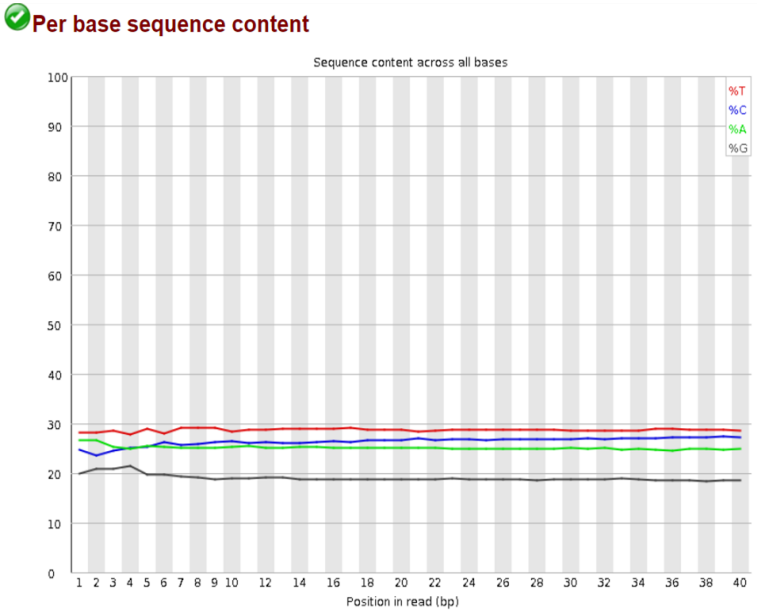
- Average read length

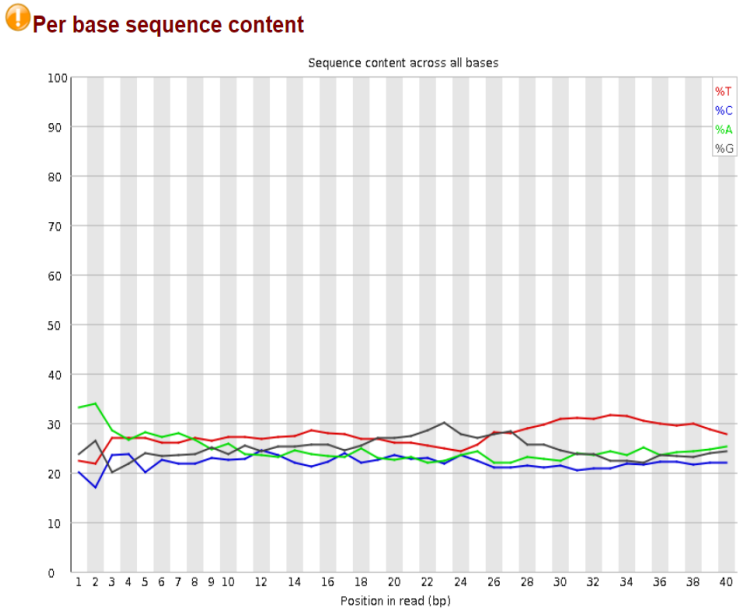
- GC content and N content

- Adapter content

# FASTQC

- Per base sequence content

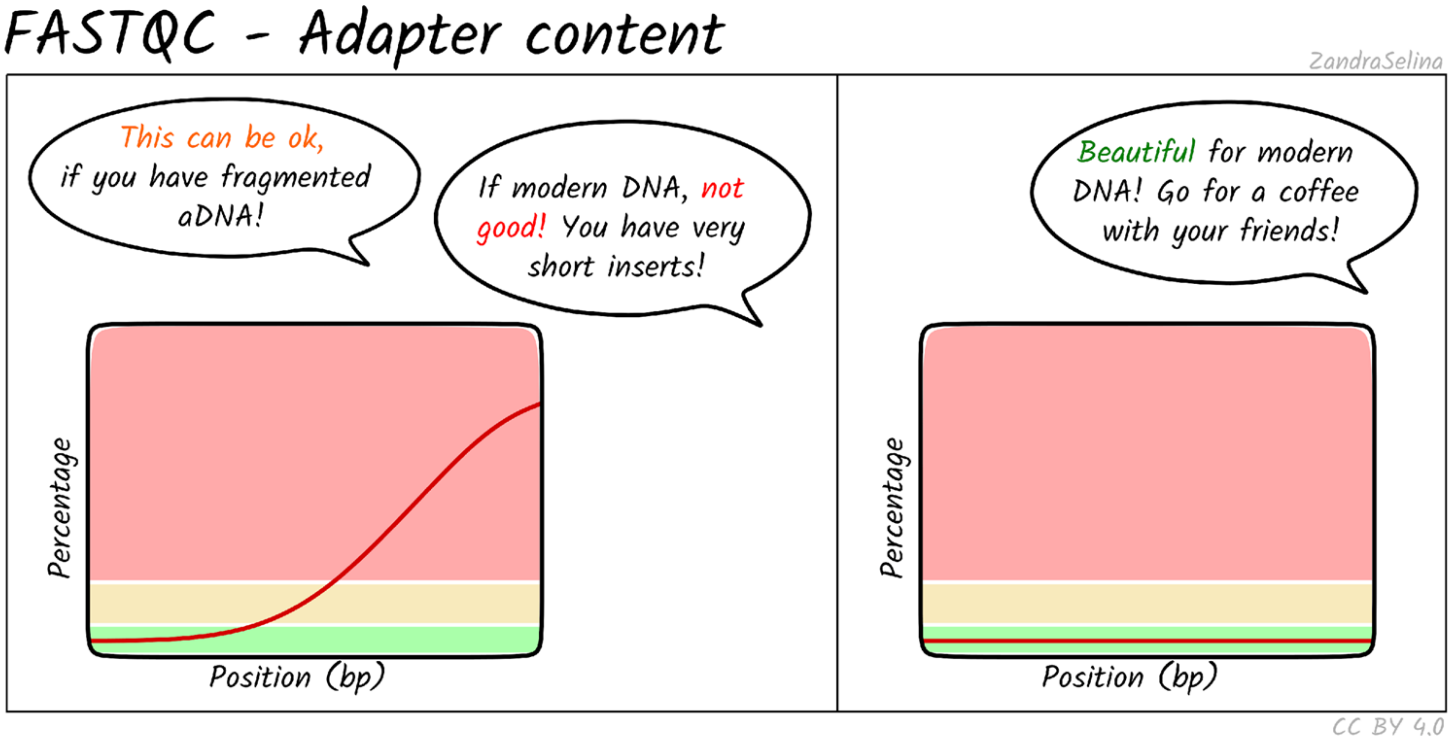



https://www.bioinformatics.babraham.ac.uk/projects/fastqc/good_sequence_short_fastqc.html#M0
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html#M0

# FASTQC

- Adapter content



# FASTQC

**NOTE** – FastQC is good, but it is very strict and will not hesitate to call your dataset bad on one of the many metrics it tests the raw data for.
Always remember what kind of data are you looking at and use logic, read the explanation for each metrics, and decide if it is acceptable or not in your specific experimental setting.

https://pharmafeatures.com/decoding-the-genetic-frontier-primary-ngs-processing-for-cutting-edge-genomics/

# FASTQC – How to run

From the shell, you can see a description of FASTQC options with:

**fastqc** --help

Then, you can run it on your fastq file with:

**fastqc** -o <output_directory> -f fastq <filename.fastq.gz>

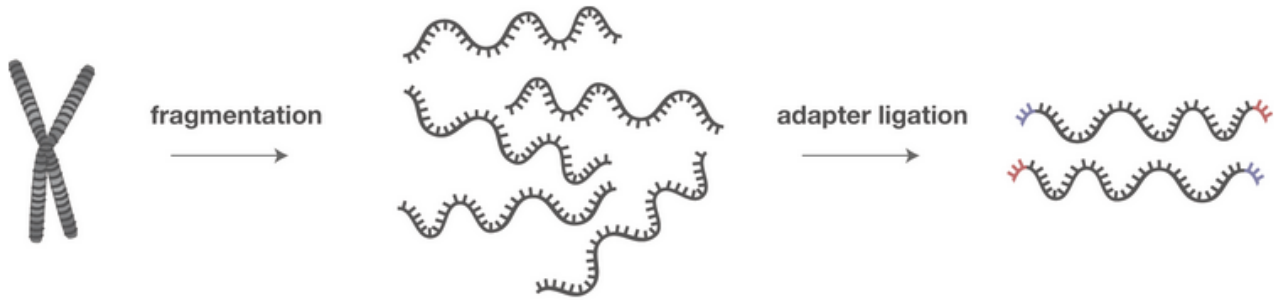
You can also run it on multiple fastq files in the current directory with:

**fastqc** -o <output_directory> -f fastq *.fastq.gz

The output is a zip file with the name of the fastq file.
Unzip it and open fastqc_report.html

# Adapters

Often your raw reads will contain **adapter** sequences. In Illumina sequencing, 5' and 3' adapters are added to the DNA fragments. These adapters can be used as sequence barcode, primers for paired-end sequencing, and sequences for attaching DNA to the flow cell, crucial for bridge amplification.



# Trimming

**Trimming** is the process of:

1) Removing adapters from the reads

2) Discard low-quality regions at the end of a read

1) Remove reads that are too short

**Several tools are available:**

Trimmomatic, Cutadapt, Trim Galore, fastp, ...

# Trimmomatic

**Trimmomatic** is a fast tool for trimming sequences, and it works particularly well with paired-end data. Below is a generalized command line for running Trimmomatic:

**java –jar /usr/local/Trimmomatic-0.39/trimmomatic-0.39.jar** PE <filename_R1.fastq.gz> <filename_R2.fastq.gz> <filename_paired_R1.fastq.gz> <filename_unpaired_R1.fastq.gz> <filename_paired_R2.fastq.gz> <filename_unpaired_R2.fastq.gz> ILLUMINACLIP:TruSeq3-PE.fa:2:30:10:2:True LEADING:3 TRAILING:30 MINLEN:36

Note that for each sample you have two fastq.gz files:

- R1 is the fastq for the forward strand
- R2 is the reverse strand

For a detailed description of the parameters see:
http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/TrimmomaticManual_V0.32.pdf

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics, btu170.

# Trimmomatic – warning!

If you run Trimmomatic and get on the terminal **java.io.FileNotFoundException**, run this command below and then run Trimmomamtic again:

```
cp /usr/local/Trimmomatic-0.36/adapters/TruSeq-PE.fa .
```

The dot at the end of the command means copy the file in the current directory on the terminal. Double check that you are in the right directory, e.g. **/var/tmp/your_name**

# Trimmomatic

The command in the previous slide will perform the following:

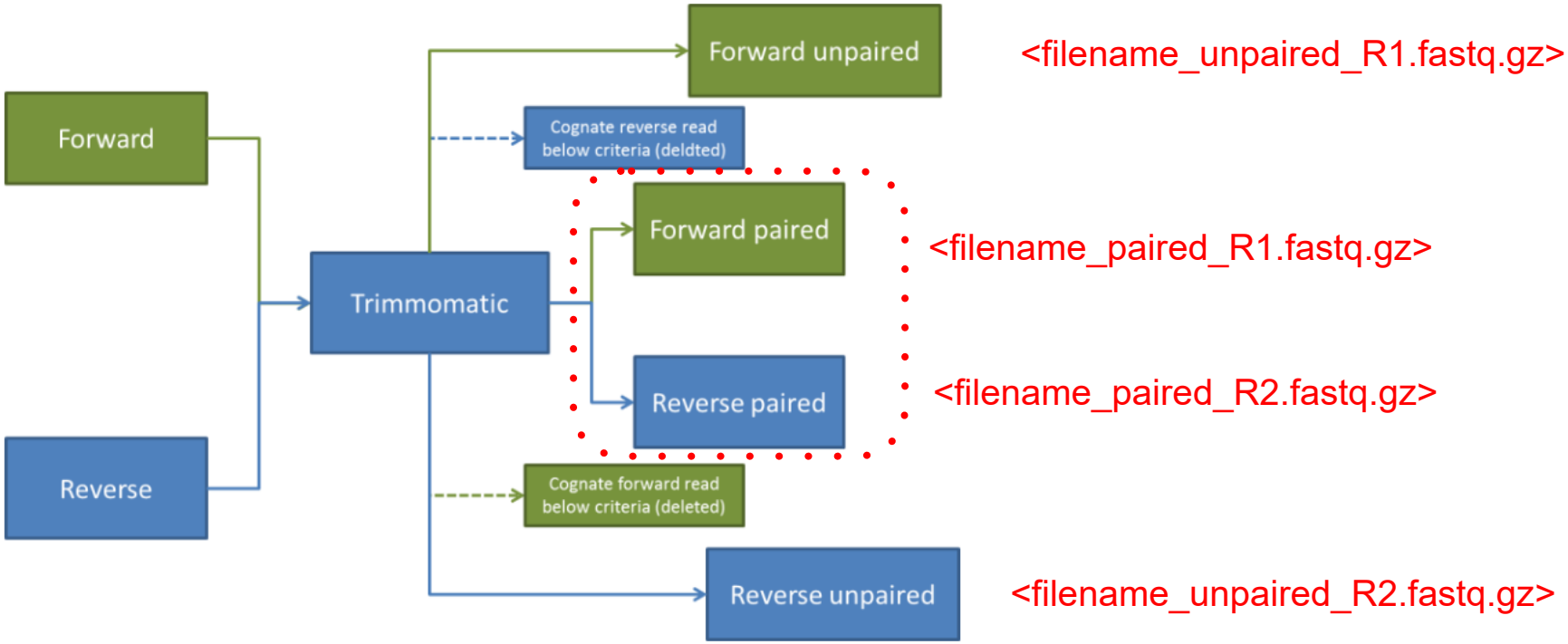
1) Remove adapters (ILLUMINACLIP:TruSeq3-PE.fa:2:30:10)

1) Remove leading low quality or N bases (below quality 3) (LEADING:3)

1) Remove trailing low quality or N bases (below quality 3) (TRAILING:3)

1) Drop reads below the 36 bases long (MINLEN:36)

# Trimmomatic



# Trimmomatic

At this point, run again FASTQC on the output files of Trimmomatic by adapting the command in the slide FASTQC – How to run

<filename_paired_R1.fastq.gz> <filename_paired_R2.fastq.gz>

Compare the output of fastqc before and after trimming to see what happened.

# Downloading and storing files

- **Your user folder maximum size is limited on the lab machines**
- But we can use the machine **local** disk to store the files
- Create a folder within the **/var/tmp/** folder (e.g. /var/tmp/myname)
- Store files there when downloading them and do the analyses from there