INTRODUCTION TO UNIX-LINUX – QUALITY CONTROL

Klebsiella Workshop

Sep 2024

ASLM AFRICAN SOCIETY FOR LABORATORY MEDICINE

UNIVERSITY of the WESTERN CAPE

SANBI South African National Bioinformatics Institute

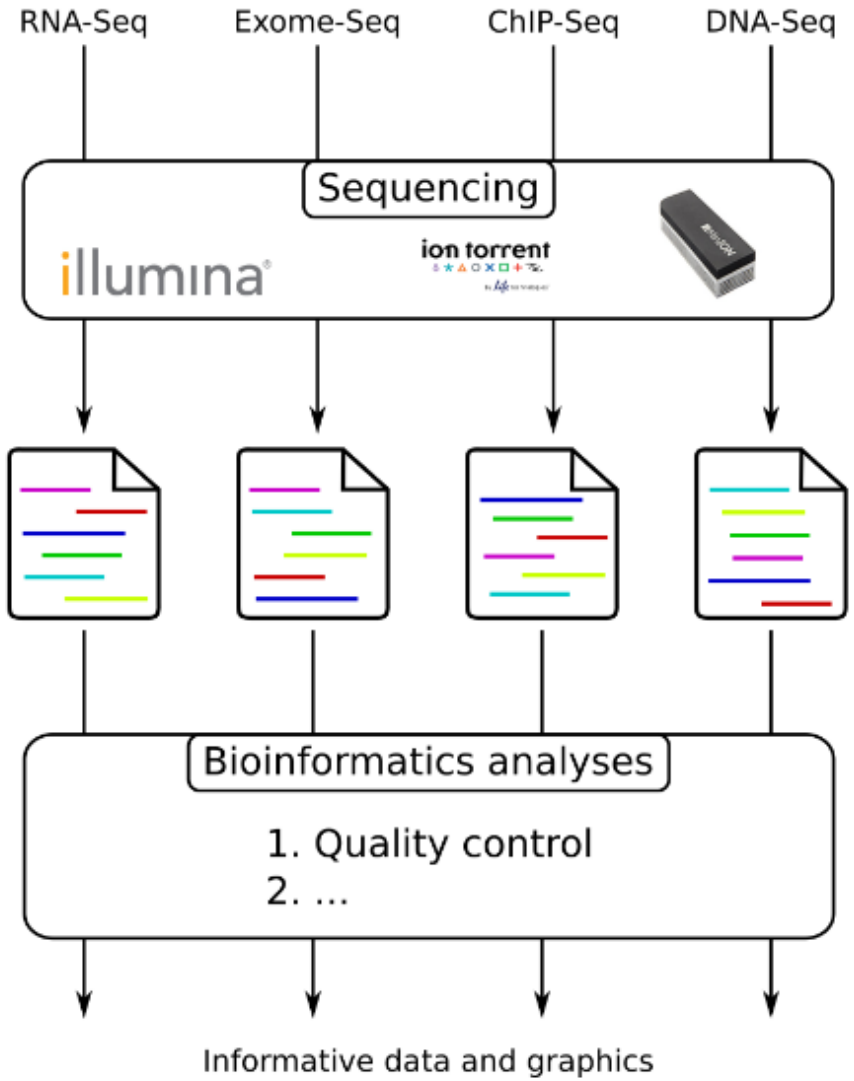
BIG DATA INSTITUTE UNIVERSITY OF OXFORD

African Union

AFRICA CDC Centres for Disease Control and Prevention Safeguarding Africa's Health

PUBLIC HEALTH ALLIANCE FOR GENOMIC EPIDEMIOLOGY

From experiments to data



Quality control = First step of the bioinformatics analyses

we are here

Quality control
(raw reads)

Adaptor / Trimming
(adapters/low quality reads)

Alignment /mapping
(QC)

Variant calling
(QC)

Tree Building
(QC/caveats)

Different Quality Control tools

we are here

Illumina - ref

- FastQC (most widely used)
- MultiQC (compresses dataset)
- Fastp (all-in-one)

DeNovo – no ref

- Metaplan
- BLAST

Quality control

Adaptor / Trimming

Alignment /mapping

Variant calling

Tree Building

Quality Control: FastQC

- A quality control tool for high throughput sequence data.
- FastQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis.

The main functions of FastQC are:

- Import of data from BAM, SAM or FastQ files (any variant)
- Providing a quick overview to tell you in which areas there may be problems
- Summary graphs and tables to quickly assess your data
- Export of results to an HTML based permanent report
- Offline operation to allow automated generation of reports without running the interactive application

Quality control

Adaptor / Trimming

Alignment /mapping

Variant calling

Tree Building

Quality Scoring

we are here

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 control

Adaptor / Trimming

Alignment /mapping

Variant calling

Tree Building

Sequence: FASTA

we are here

```
>Identifier1 (comment)
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
>Identifier2 (comment)
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XX
```

Quality control

Adaptor / Trimming

Alignment / mapping

Variant calling

Tree Building

Sequence: FASTQ

we are here

```
@Identifier1 (comment)
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
+
QQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQ
@Identifier2 (comment)
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
+
QQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQ
```

Quality control

Adaptor / Trimming

Alignment / mapping

Variant calling

Tree Building

we are here

Quality control

Adaptor / Trimming

Alignment / mapping

Variant calling

Tree Building

FastQC Report

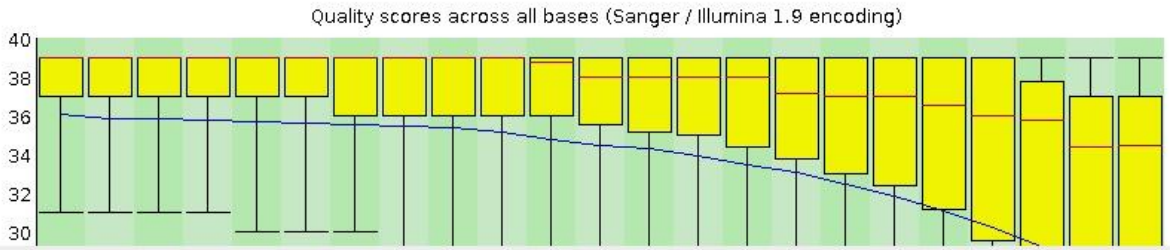
Summary

- 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

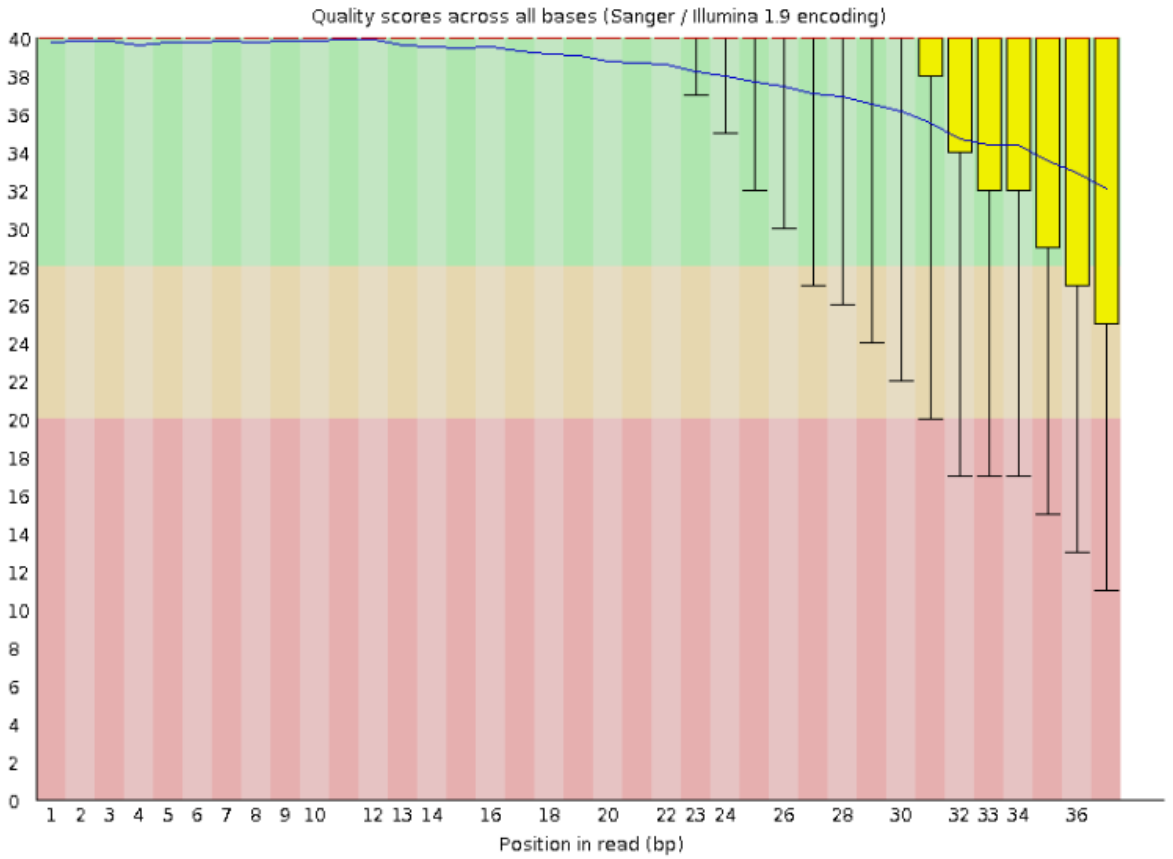
Measure	Value
Filename	WES_human_Illumina.pe_1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	4942814
Filtered Sequences	0
Sequence length	76
%GC	47

Per base sequence quality



Produced by FastQC (version 0.10.1)

Quality score: Per-base



Good quality score

we are here

Quality control

Adaptor / Trimming

Alignment / mapping

Variant calling

Tree Building

we are here

Quality control

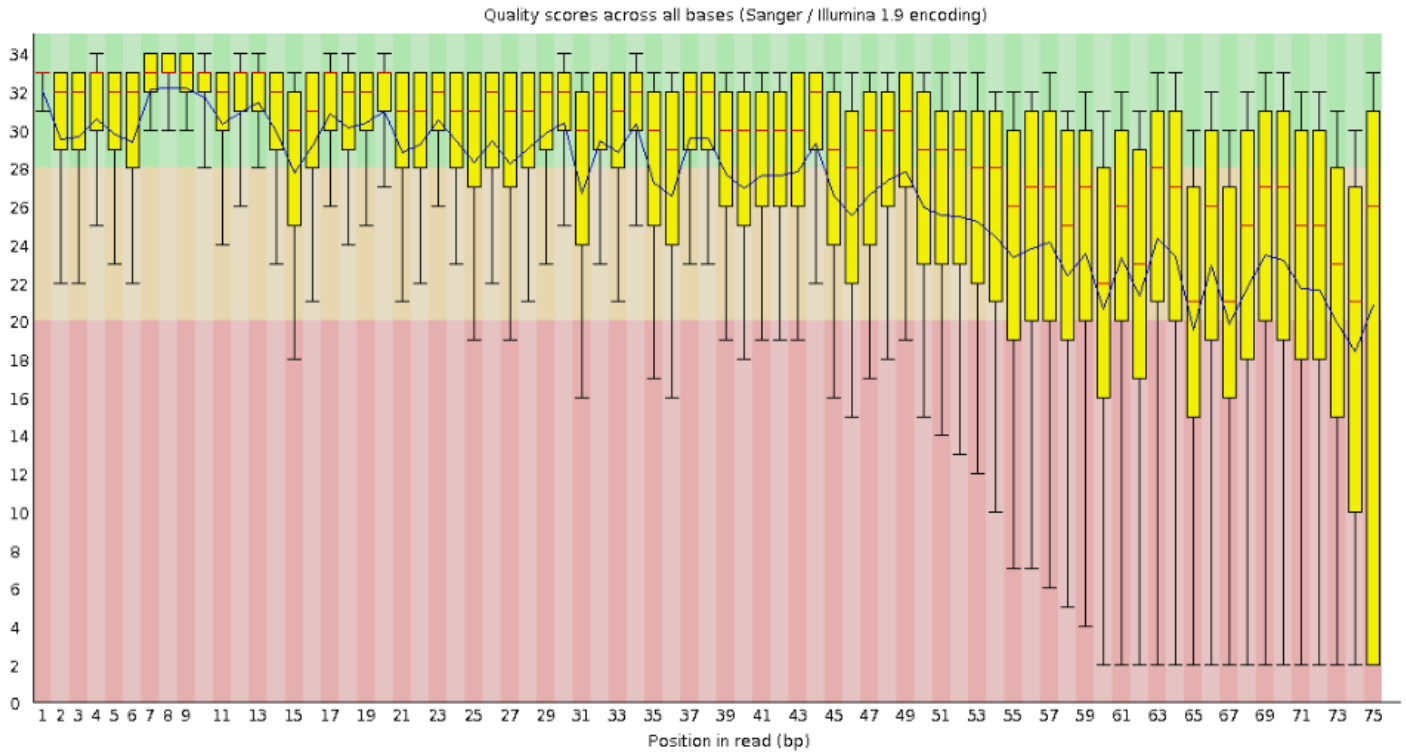
Adaptor / Trimming

Alignment / mapping

Variant calling

Tree Building

Per-base Quality



Bad quality score

Improving Quality

- 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

Quality control

Adaptor / Trimming

Alignment / mapping

Variant calling

Tree Building

Installation & Usage: FastQC

we are here

- Installation:
  - Wget https://github.com/s-andrews/FastQC/archive/refs/tags/v0.12.1.zip
- Usage:
  - Run either interactive graphical application in which you can dynamically load FastQ files and view their results. OR
  - In a non-interactive mode where you specify the files you want to process on the command line and FastQC will generate an HTML report for each file without launching a user interface.

Installation of fastQC using conda

```
conda install bioconda::fastqc
conda install
bioconda/label/broken::fastqc
conda install
bioconda/label/cf201901::fastqc
```

Quality control

Adaptor / Trimming

Alignment /mapping

Variant calling

Tree Building

MultiQC

we are here

- MultiQC is a tool to create a single report with interactive plots for multiple bioinformatics analyses across many samples.
- MultiQC provides clear and customizable visualizations, including graphs, boxplots, and heatmaps that help in interpreting the data quickly and effectively.

General Statistics

Copy table | Configure Columns | Sort by highlight | Plot | Showing 8/8 rows and 9/11 columns.

Sample Name	5'-3' bias	M Aligned	% Aligned	M Aligned	% Aligned	M Aligned	% Dups	% GC	M Seqs
Irrel_kd_1	1.18	35.6	86.4%	31.2	92.1%	33.2	55.9%	47%	36.1
Irrel_kd_2	1.14	30.4	86.0%	26.5	92.2%	28.4	53.6%	47%	30.8
Irrel_kd_3	1.19	23.6	85.7%	20.5	92.0%	22.0	50.1%	48%	23.9
Mov10_kd_2	1.13	51.9	86.0%	45.3	91.6%	48.3	60.5%	48%	52.7
Mov10_kd_3	1.13	30.7	86.0%	26.8	91.6%	28.5	54.6%	47%	31.1
Mov10_oe_1	1.09	38.1	80.2%	32.1	88.9%	35.5	56.5%	47%	40.0
Mov10_oe_2	1.18	35.4	81.0%	30.0	88.8%	33.0	55.9%	48%	37.1
Mov10_oe_3		20.3	81.5%	17.3	90.0%	19.1	50.1%	47%	21.2

Quality control

Adaptor / Trimming

Alignment / mapping

Variant calling

Tree Building

Installation & Usage: MultiQC

we are here

- Installation:
  - `pip install multiqc`

- Usage:
  - Basic command: `multiqc .`
  - Or navigate within directory: `multiqc [path to fastqc files] data/*_fastqc.zip`
  - Other parameters / options:
    - `-o/--outdir` (desired folder name)
    - `-p/--export` (export .pdf .jpg .png files)

Quality control

Adaptor / Trimming

Alignment / mapping

Variant calling

Tree Building

Different trimming tools

Illumina – ref data

- Trimmomatic
- CutaDapt
- Fastp
- Flexar
- Trimgalore

DeNovo – no ref

- Kraken
- Filtlong
- Porechop
- Blobology
- BLAST

Quality control

Adaptor / Trimming

Alignment / mapping

Variant calling

Tree Building

Trimmomatic

- Trimmomatic performs a variety of useful trimming tasks for illumina paired-end and single ended data. The selection of trimming steps and their associated parameters are supplied on the command line.
- Requies Java installation, check java –version
- **Inputs -** Single-end or Paired-end FASTQ or FASTQ.GZ reads
- https://github.com/usadellab/Trimmomatic?tab=readme-ov-file

Quality control

Adaptor / Trimming

Alignment / mapping

Variant calling

Tree Building

Installation & Usage: Trimmomatic

we are here

- Installation:
  - wget https://github.com/usadellab/Trimmomatic.git
- Usage:
  - ```
    java -jar trimmomatic-0.39.jar PE input_forward.fq.gz
    input_reverse.fq.gz output_forward_paired.fq.gz
    output_forward_unpaired.fq.gz
    output_reverse_paired.fq.gz
    output_reverse_unpaired.fq.gz ILLUMINACLIP:TruSeq3-
    PE.fa:2:30:10:2:True LEADING:3 TRAILING:3 MINLEN:36
    ```
- The above code is for reference only

Quality control

Adaptor / Trimming

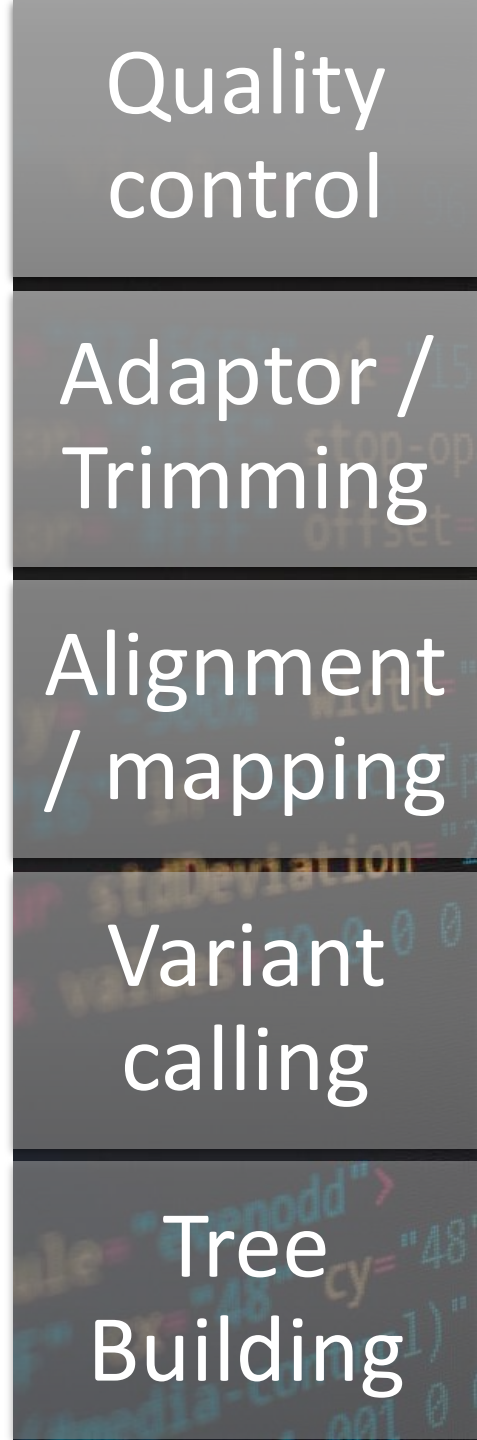
Alignment / mapping

Variant calling

Tree Building
    ```

Fastp

- fastp is a tool designed to provide fast all-in-one preprocessing for FASTQ files. This tool is developed in C++ with multithreading supported to afford high performance.
- **Inputs -** Single-end or Paired-end FASTQ or FASTQ.GZ reads
- https://github.com/OpenGene/fastp?tab=readme-ov-file#features



Installation & Usage: Fastp

- Installation:
  - wget http://opengene.org/fastp/fastp
  - chmod a+x ./fastp

- Usage:
  - SE code: `fastp -i in.fq -o out.fq`
  - PE code: `fastp -i in.R1.fq.gz -I in.R2.fq.gz -o out.R1.fq.gz -O out.R2.fq.gz`

- The above code is for reference only

Quality control

Adaptor / Trimming

Alignment / mapping

Variant calling

Tree Building