CAPSTONE PROJECT REPORT - GROUP 3

Participants:

Makonk Najah: NMIMR, Ghana

Stella E. Nabirye: UVRI, Uganda

Isaac Adison: NPHL, South Sudan

OUTLINE

1. Raw Reads Quality Control and Statistics
2. Bash Pipeline tool: eboVar.sh
3. Multiqc result form pipeline
4. Variant filtering and Extraction with Bash
5. Analysis and Visualisation of Filtered Variants with R
6. Pipeline Concatineristion
7. Container Testing
8. Documentation and Pipeline Sharing
9. Acknowledgments

AfricaCDC
Centres for Disease Control and Prevention

RAW READS QUALITY CONTROL AND STATISTICS

file	num_seqs	sum_len	min_len	avg_len	max_len	Q20(%)	Q30(%)
SRR12689946_1.fas	903384	122926890	35	136.1	151	94.73	92.81
SRR12689946_2.fas	903384	123163864	35	136.3	151	92.11	89.58
SRR12689947_1.fas	848772	110154073	35	129.8	151	94.26	92.39
SRR12689947_2.fas	848772	110325227	35	130	151	91.11	88.62
SRR12689969_1.fas	1286097	184013746	35	143.1	151	96.53	95.23
SRR12689969_2.fas	1286097	183927974	35	143	151	95.01	93.26
SRR12689978_1.fas	30349	4125709	35	135.9	151	92.89	90.95
SRR12689978_2.fas	30349	4122799	35	135.8	151	82.05	78.56
SRR12689981_1.fas	189920	26863001	35	141.4	151	96.41	95.05
SRR12689981_2.fas	189920	26881238	35	141.5	151	92.59	90.54
SRR3153067_1.fasto	1051119	133055767	70	126.6	131	89.23	76.94
SRR3153067_2.fasto	1051119	123087280	70	117.1	131	84.3	71.19

- Number of sequences in reference file is 1
- Length of reference = 18959 bases

```
seqkit stats -a -T *.gz > sample_stats.tsv
```

AfricaCDC
Centres for Disease Control and Prevention

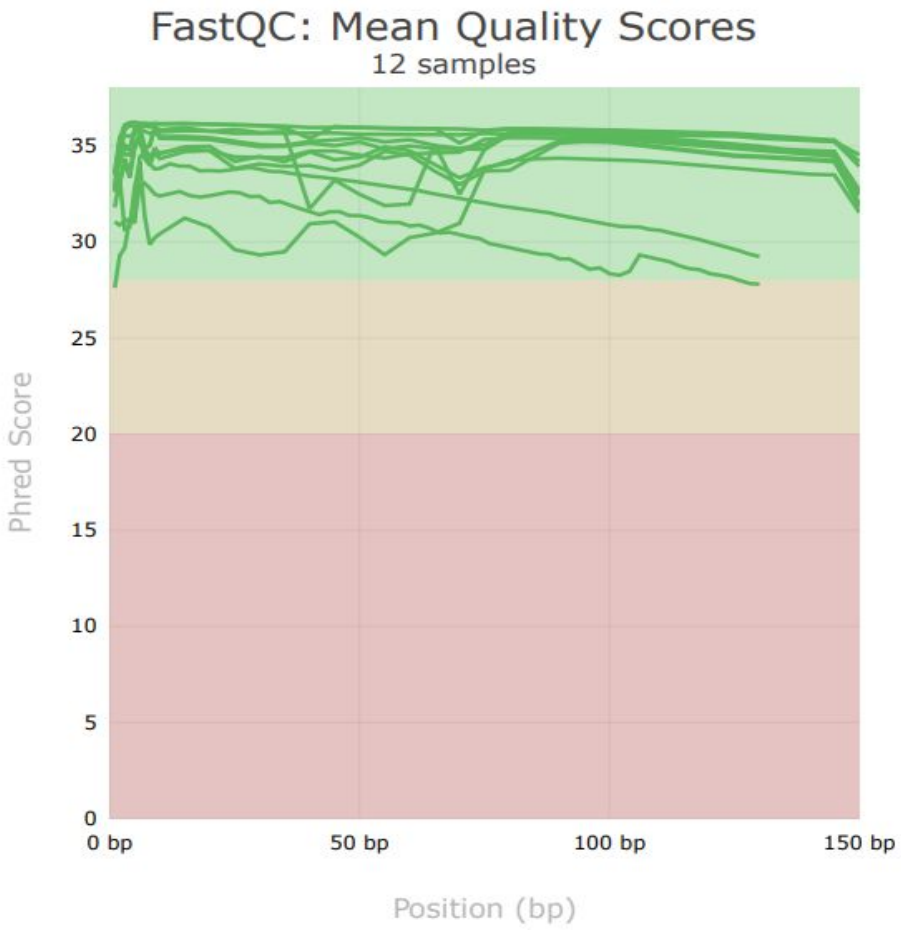
RAW READS QUALITY CONTROL AND STATISTICS

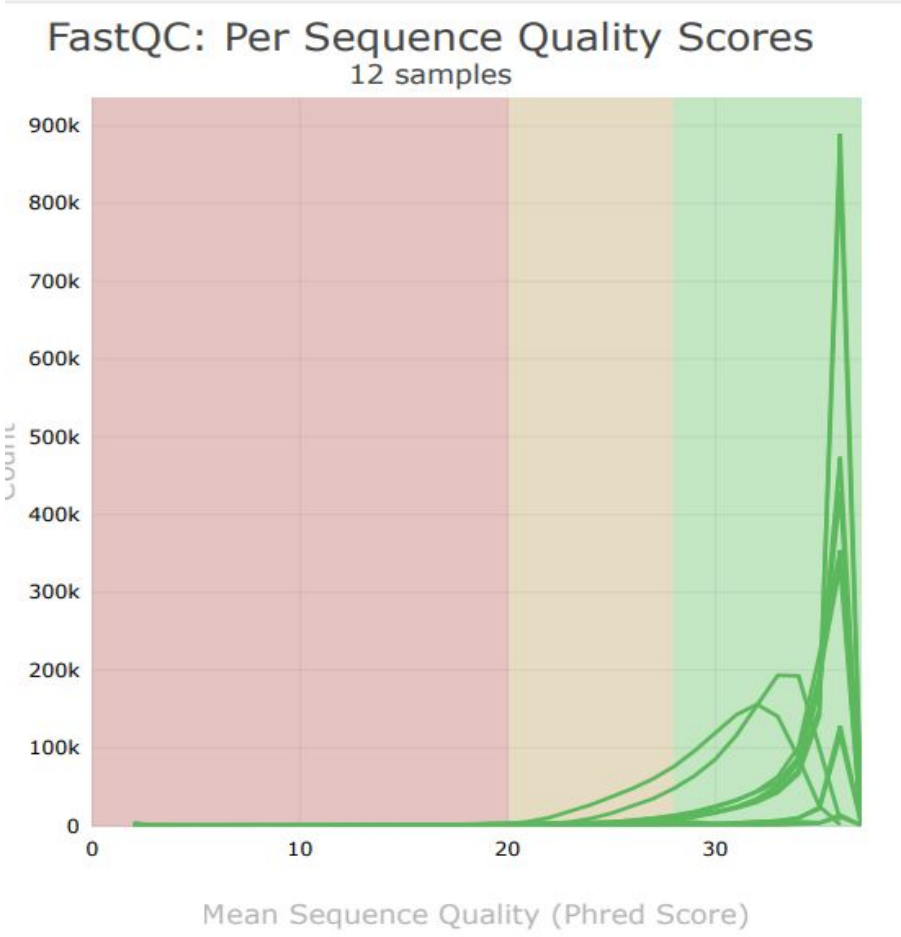
General sample statistics

Sample Name	Dups	GC	Median len	Seqs
SRR3153067_1	53.6 %	42.0 %	130 bp	1.1 M
SRR3153067_2	38.0 %	42.0 %	104 bp	1.1 M
SRR12689946_1	82.4 %	45.0 %	151 bp	0.9 M
SRR12689946_2	79.1 %	46.0 %	151 bp	0.9 M
SRR12689947_1	81.4 %	48.0 %	147 bp	0.8 M
SRR12689947_2	77.2 %	48.0 %	147 bp	0.8 M
SRR12689969_1	87.2 %	45.0 %	151 bp	1.3 M
SRR12689969_2	87.0 %	45.0 %	151 bp	1.3 M
SRR12689978_1	61.0 %	61.0 %	151 bp	0.0 M
SRR12689978_2	44.7 %	64.0 %	151 bp	0.0 M
SRR12689981_1	75.7 %	48.0 %	151 bp	0.2 M
SRR12689981_2	71.9 %	49.0 %	151 bp	0.2 M

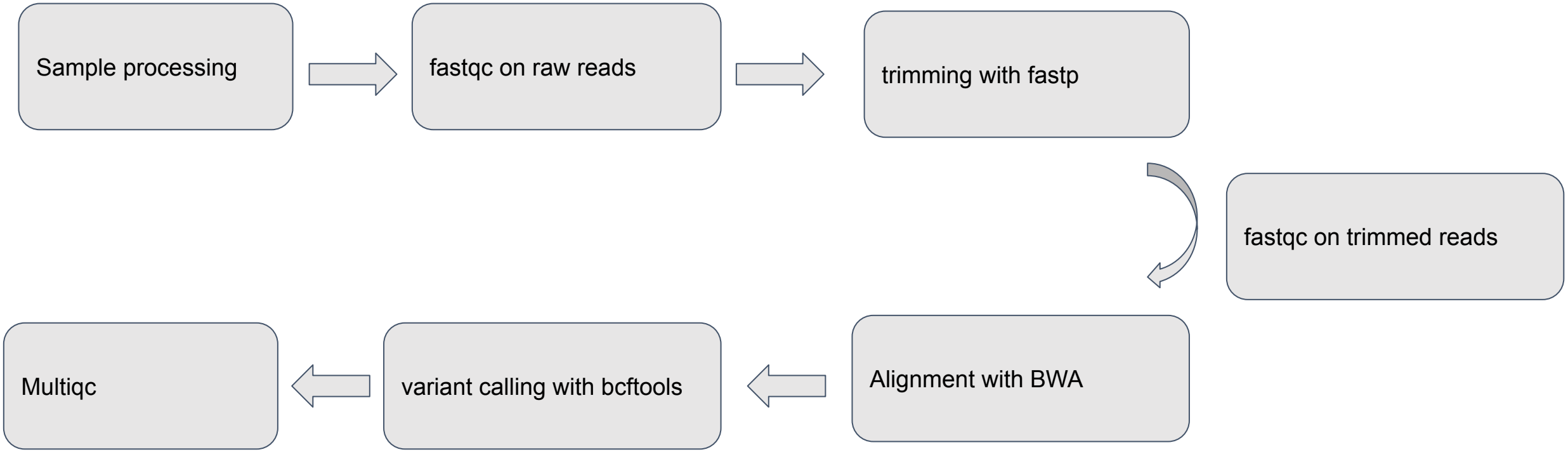
AfricaCDC
Centres for Disease Control and Prevention

RAW READS QUALITY CONTROL AND STATISTICS



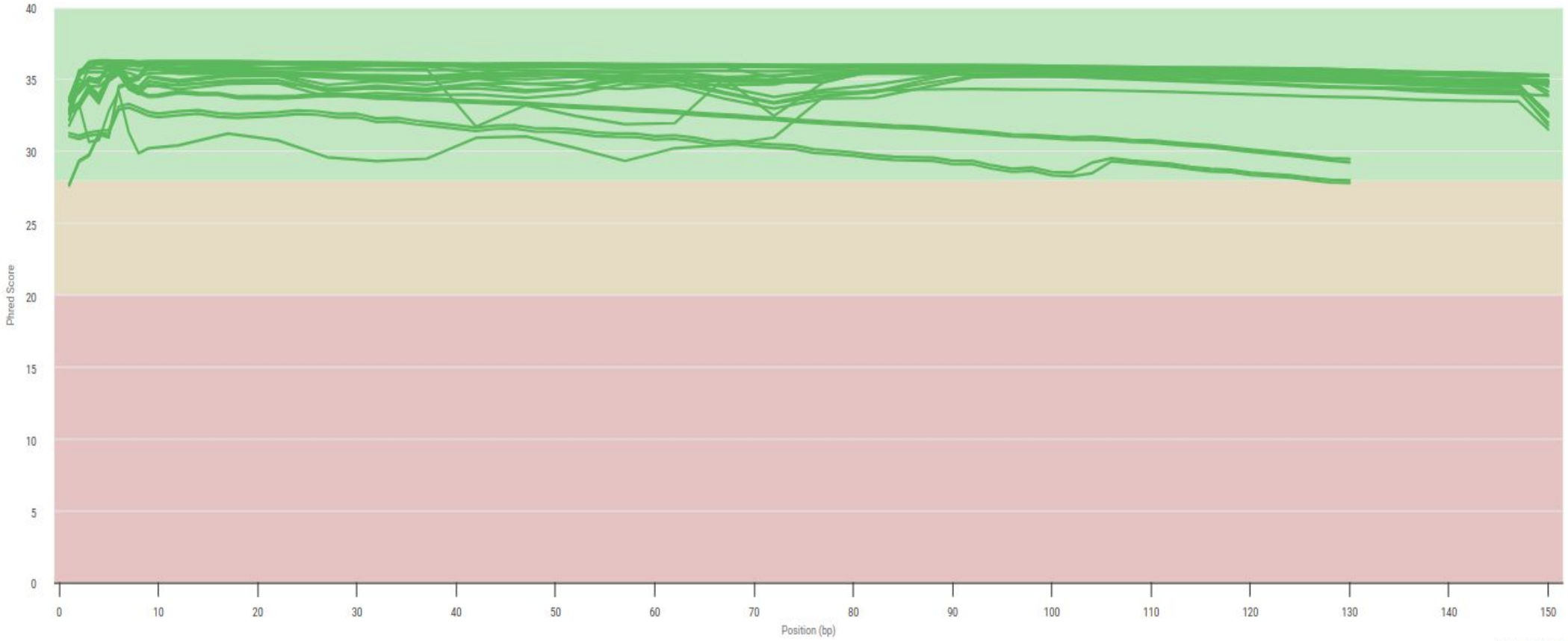


BASH PIPELINE TOOL: eboVar.sh



AfricaCDC
Centres for Disease Control and Prevention

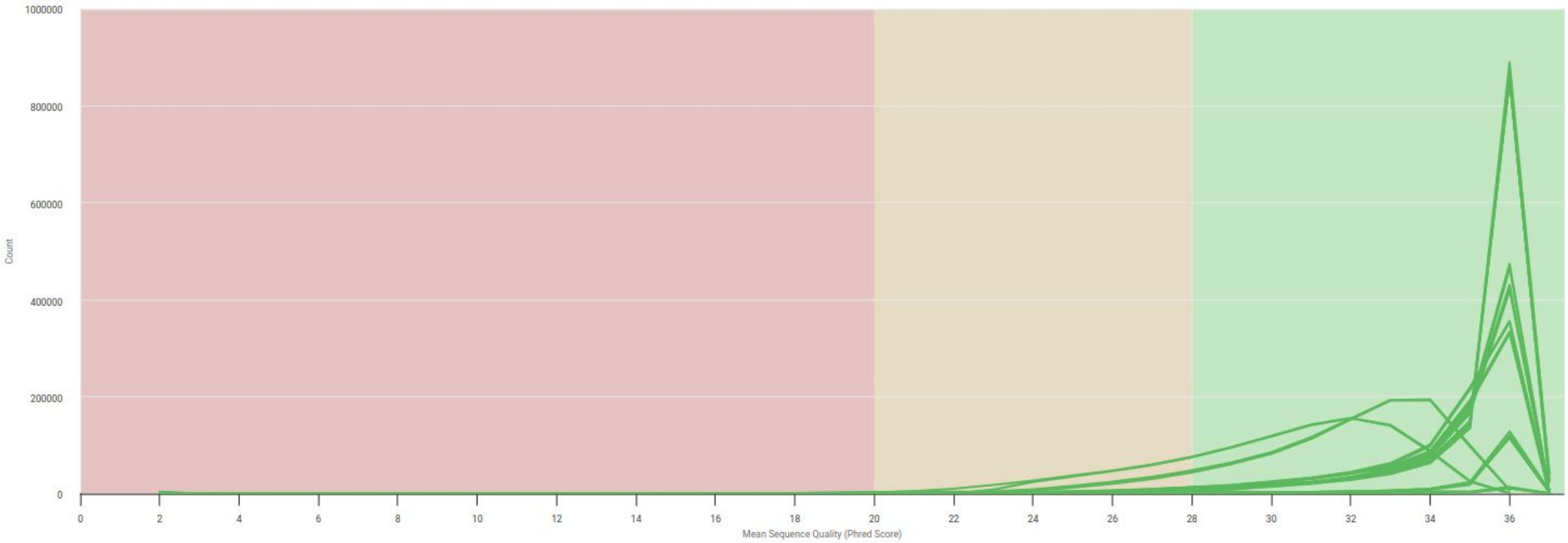
MULTIQC RESULTS FROM PIPELINE eboVar.sh



Sequence quality histogram: mean quality value across each base position in the read

AfricaCDC
Centres for Disease Control and Prevention

MULTIQC RESULTS FROM PIPELINE eboVar.sh



Per sequence quality scores: The number of reads with average quality scores. Shows if a subset of reads has poor quality

AfricaCDC
Centres for Disease Control and Prevention

MULTIQC RESULTS FROM PIPELINE eboVar.sh

General Statistics

Copy table | Configure Columns | Plot | Showing 24/24 rows and 9/13 columns.

Sample Name	% Duplication	M Reads After Filtering	GC content
SRR12689946_1	25.1%	1.7	45.2%
SRR12689946_2			
SRR12689946_trimmed_R1			
SRR12689946_trimmed_R2			
SRR12689947_1	27.0%	1.6	48.1%
SRR12689947_2			
SRR12689947_trimmed_R1			
SRR12689947_trimmed_R2			
SRR12689969_1	30.1%	2.5	45.7%
SRR12689969_2			
SRR12689969_trimmed_R1			
SRR12689969_trimmed_R2			
SRR12689978_1	24.3%	0.0	56.7%

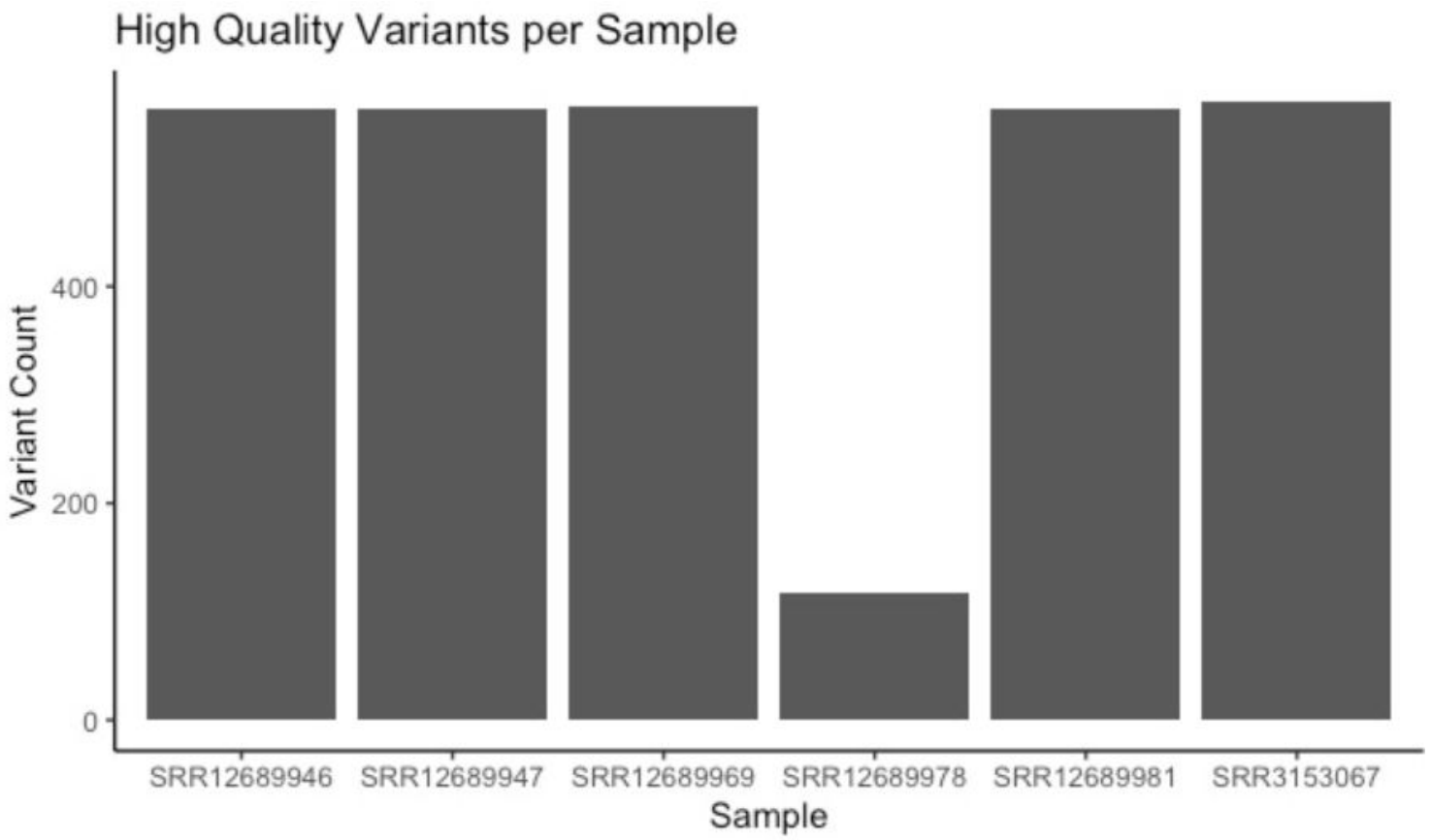
AfricaCDC
Centres for Disease Control and Prevention

VARIANT FILTERING AND EXTRACTION WITH BASH

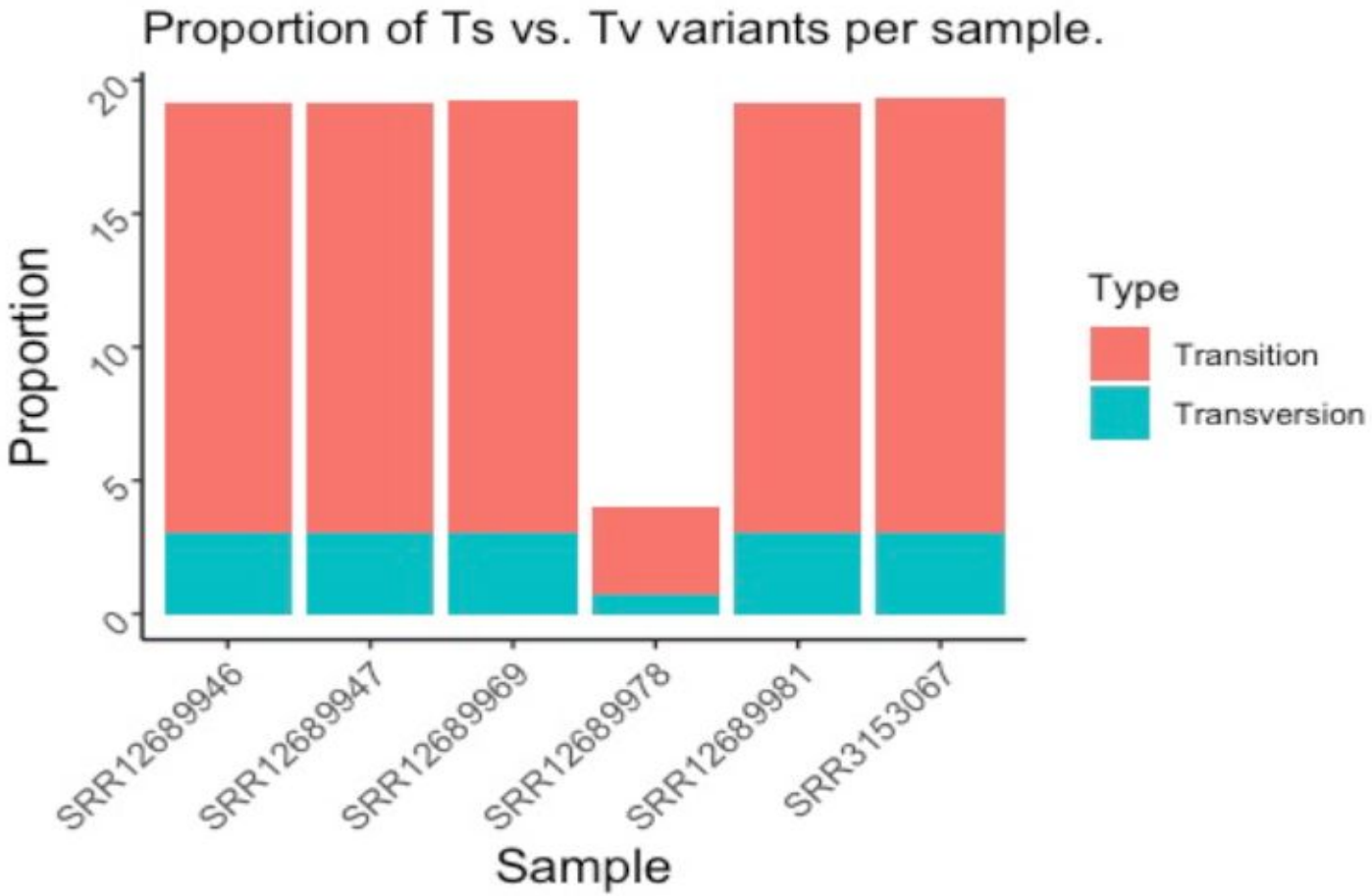
- Phred score ≥ 30
- Read Depth ≥ 10
- Allele Frequency > 0.05

File	CHROM	POS	ID	REF	ALT	QUAL	DP	AC	AN	AF
SRR12689946.vcf.gz	NC_002549.1	127	.	C	T	225.422	333	2	2	1
SRR12689946.vcf.gz	NC_002549.1	149	.	C	T	225.417	349	2	2	1
SRR12689946.vcf.gz	NC_002549.1	155	.	A	C	225.417	347	2	2	1
SRR12689946.vcf.gz	NC_002549.1	182	.	A	G	225.422	246	2	2	1
SRR12689946.vcf.gz	NC_002549.1	187	.	A	G	225.417	247	2	2	1
SRR12689946.vcf.gz	NC_002549.1	230	.	C	T	225.417	247	2	2	1
SRR12689946.vcf.gz	NC_002549.1	236	.	T	C	225.417	251	2	2	1
SRR12689946.vcf.gz	NC_002549.1	257	.	A	G	228.393	237	2	2	1
SRR12689946.vcf.gz	NC_002549.1	261	.	C	T	225.417	235	2	2	1
SRR12689946.vcf.gz	NC_002549.1	263	.	G	A	228.387	246	2	2	1
SRR12689946.vcf.gz	NC_002549.1	295	.	A	C	225.417	250	2	2	1
SRR12689946.vcf.gz	NC_002549.1	356	.	C	T	225.417	252	2	2	1
SRR12689946.vcf.gz	NC_002549.1	360		G	A	228.406	250	2	2	1

ANALYSIS AND VISUALISATION OF FILTERED VARIANTS WITH R



ANALYSIS AND VISUALISATION OF FILTERED VARIANTS WITH R



AfricaCDC
Centres for Disease Control and Prevention

PIPELINE CONTAINERISATION

```
Bootstrap: docker
From: ubuntu:22.04

%labels
    Maintainer Makonk Najah, Stella E Nabirye, Isaac Adison
    Version v1.0
    Description "EBOV Variant Analysis Pipeline Container (Mamba-based)"

%help
    ----------------------------------------------------------------------
    EBOV Variant Analysis Pipeline: QC → Trimming → Alignment → Variant Calling

    This container runs:
      1. FastQC         - quality check on raw and trimmed reads
      2. fastp          - trimming, filtering, and adapter removal
      3. BWA-MEM        - alignment to EBOV reference genome
      4. samtools       - BAM processing and indexing
      5. bcftools       - variant calling and VCF sorting/indexing
      6. MultiQC        - combined reporting for all QC steps

    📄 Usage:
      singularity run ebovar.sif -i <input_folder> -o <output_folder> -r <reference.fa> [-t <threads>]

    📂 Mount input/output folders using --bind

    All tools are installed via Mamba (Miniforge) using environment.yml

    Authors:
      Makonk Najah
      Stella E Nabirye
      Isaac Adison
    ----------------------------------------------------------------------
```

```
# ----------------------------------------------------------------------
# 📦 Build the container image (requires root privileges):
#     sudo apptainer build ebovar.sif ebovar.def
#
# 🚀 Run the pipeline from your current working directory with bind mount:
#     apptainer run --bind $(pwd):/data ebovar.sif \
#         -i /data/rawreads \
#         -o /data/results_container \
#         -r /data/ebov_ref.fa \
#         -t 4
# ----------------------------------------------------------------------
```

- Definition file
- Pipeline script
- Environment file

CONTAINER TESTING: COMPARING RESULTS

```
eboVar_results/
├── bam
├── logs
├── multiqc
│   └── multiqc_data
├── qc_raw
├── qc_trimmed
├── trimmed
└── vcf
```

```
container_results/
├── bam
├── logs
├── multiqc
│   └── multiqc_data
├── qc_raw
├── qc_trimmed
├── trimmed
└── vcf
```

AfricaCDC
Centres for Disease Control and Prevention

CONTAINER TESTING: COMPARING RESULTS

```
├── multiqc
│   ├── multiqc_data
│   │   ├── multiqc_citations.txt
│   │   ├── multiqc_data.json
│   │   ├── multiqc_fastp.txt
│   │   ├── multiqc_fastqc.txt
│   │   ├── multiqc_general_stats.txt
│   │   ├── multiqc.log
│   │   ├── multiqc_software_versions.txt
│   │   └── multiqc_sources.txt
│   └── multiqc_report.html
```

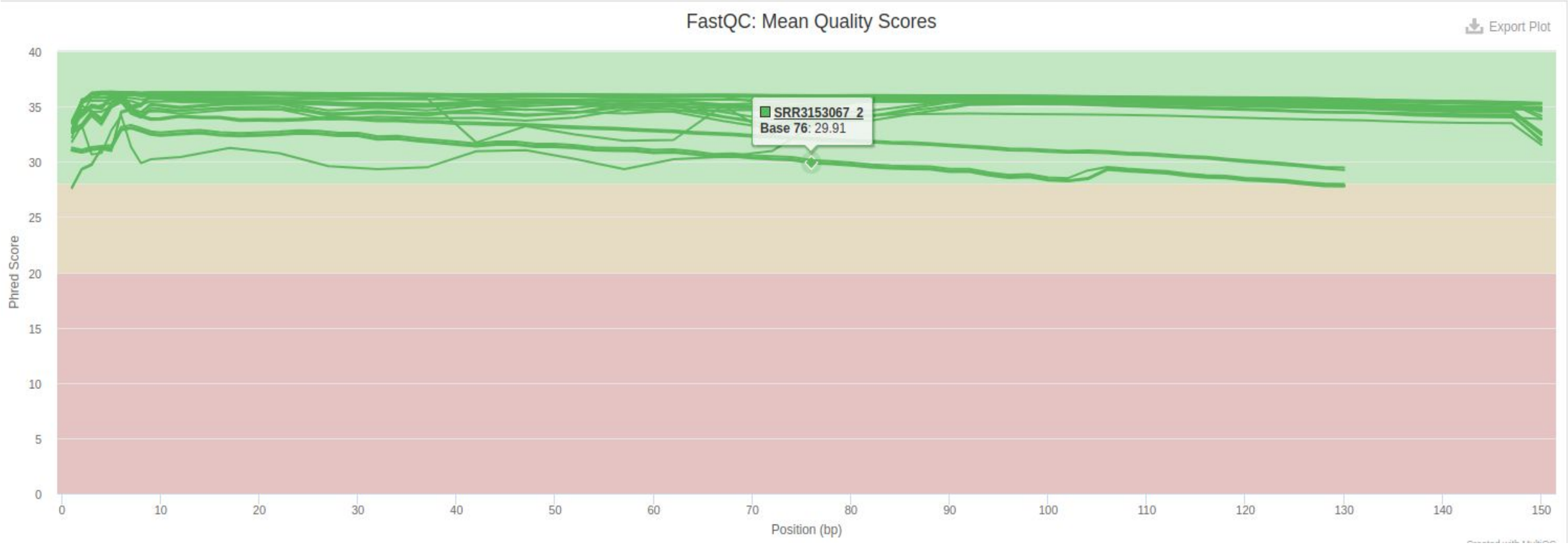
eboVar.sh

```
├── multiqc
│   ├── multiqc_data
│   │   ├── multiqc_citations.txt
│   │   ├── multiqc_data.json
│   │   ├── multiqc_fastp.txt
│   │   ├── multiqc_fastqc.txt
│   │   ├── multiqc_general_stats.txt
│   │   ├── multiqc.log
│   │   └── multiqc_sources.txt
│   └── multiqc_report.html
```

ebovar.sif

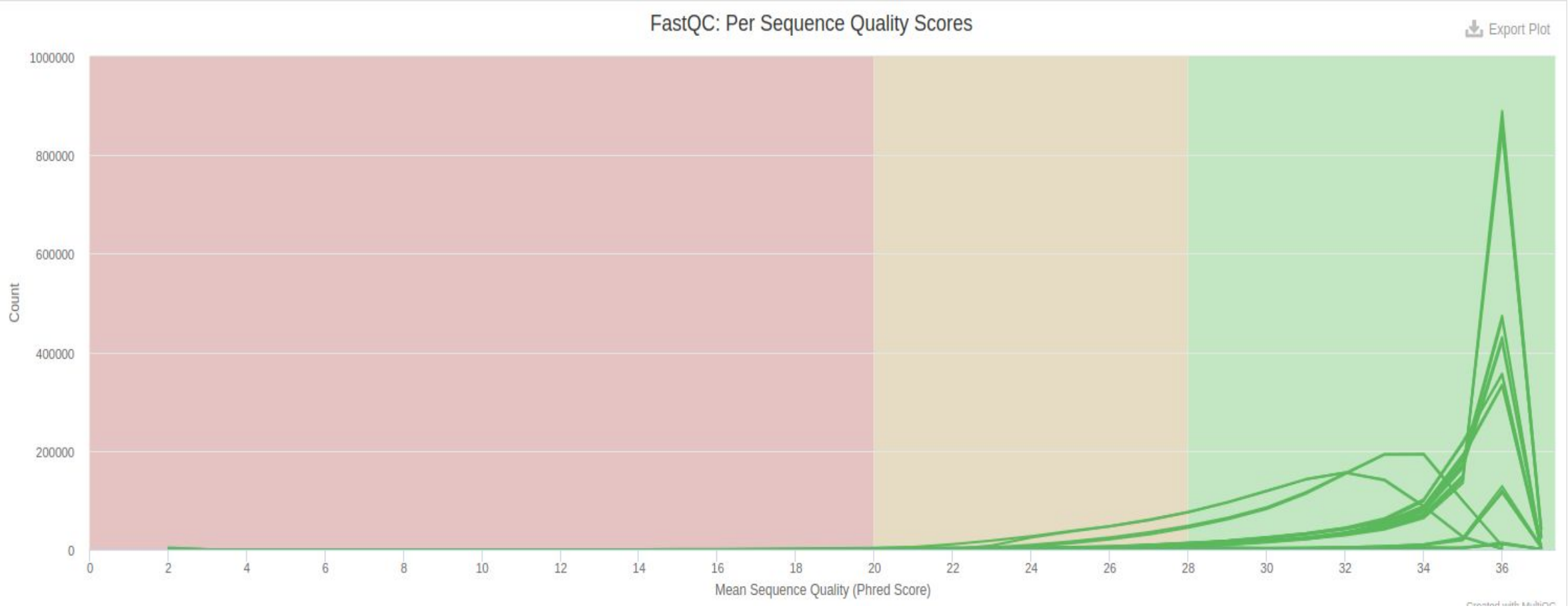
AfricaCDC
Centres for Disease Control and Prevention

CONTAINER TESTING: MULTIQC



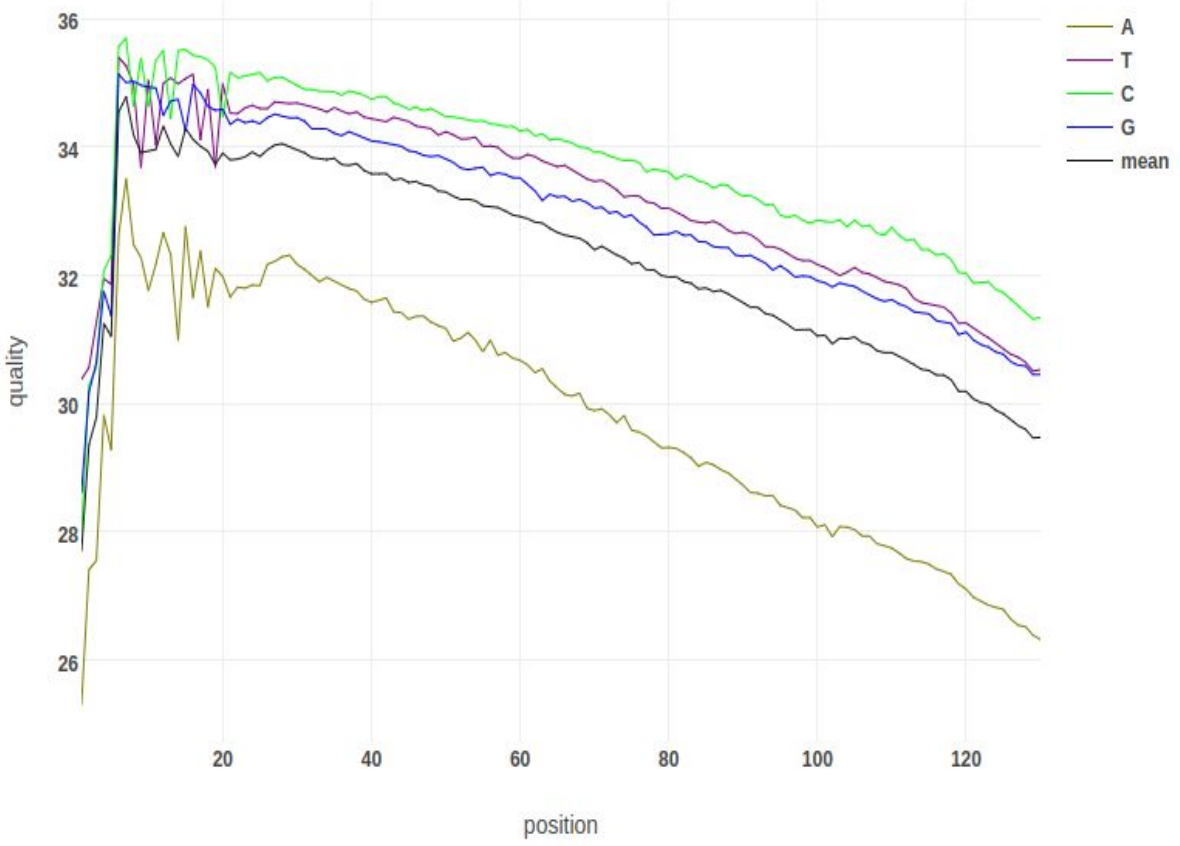
AfricaCDC
Centres for Disease Control and Prevention

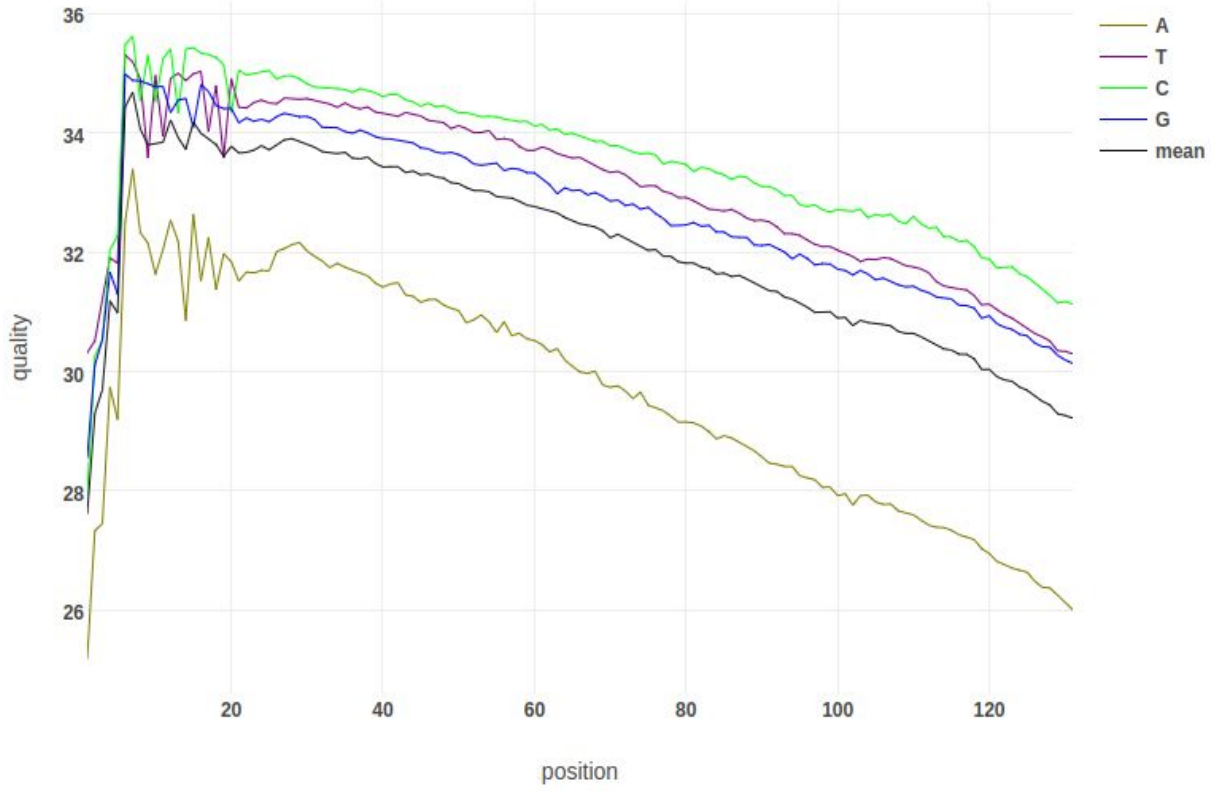
CONTAINER TESTING: MULTIQC



AfricaCDC
Centres for Disease Control and Prevention

CONTAINER TESTING: SAMPLE FASTP





SRR3153067

eboVar.sh

ebovar.sif

AfricaCDC
Centres for Disease Control and Prevention

DOCUMENTATION AND PIPELINE SHARING

Capstone_project_Africa_CDC Public

Watch 0

main | 1 Branch | 0 Tags | Go to file t | Add file | Code

StellaNabirye Update README.md 7f9d4f5 · 7 hours ago 9 Commits

README.md	Update README.md	7 hours ago
download_reads.sh	Uploading first batch of files	14 hours ago
eboVar.sh	Uploading first batch of files	14 hours ago
ebov_ref.fa	Uploading first batch of files	14 hours ago
ebovar.def	Uploading first batch of files	14 hours ago
environment.yml	Uploading first batch of files	14 hours ago
filter_and_merge_vcfs.sh	uploaded the variant analysis R script and the merged ts...	8 hours ago
hq_allvariants.tsv	Uploading first batch of files	14 hours ago
individual_vcf_filter.sh	added the script for filtering the individual vcf files	7 hours ago
variant_analysis.R	uploaded the variant analysis R script and the merged ts...	8 hours ago
variant_analysis_report.Rmd	added the Rmd file and pdf for the variant analysis	7 hours ago
variant_analysis_report.pdf	added the Rmd file and pdf for the variant analysis	7 hours ago

README

CONCLUSION

Concepts applied

- Knowledge in Unix
- Files structure
- Bash Scripting
- Genomics
- Pipeline Automation
- Containerisation
- Github Documentation & Collaboration
- R Programming

LINK TO GITHUB REPOSITORY

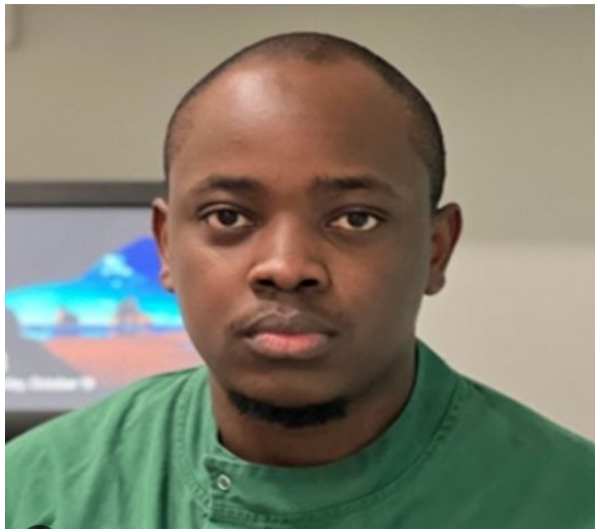
https://github.com/Yitzhak97/Capstone_project_Africa_CDC

ACKNOWLEDGMENTS

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

TRAINERS

Shahiid Kiyaga



Leonard Ndwiga

Julien A. Nguinkal,

THANK YOU

AfricaCDC
Centres for Disease Control
and Prevention