BIN 508: Next Generation Sequence Analysis and Informatics

Assignment 02

By : Taha Ahmad

Student Id: 2546126

Instructor : Dr Yesim Aydin Son

Question 1 A

Untrimmed

General Statistics

Copy table | Configure columns | Scatter plot | Violin plot | Export as CSV... | Showing 2/2 rows and 6/6 columns. | Summarize table

Sample Name	Dups	GC	Avg len	Median len	Failed	Seqs
sample1_1	53.5 %	44.0 %	233 bp	251 bp	18 %	0.0 M
sample1_2	55.4 %	44.0 %	233 bp	251 bp	27 %	0.0 M

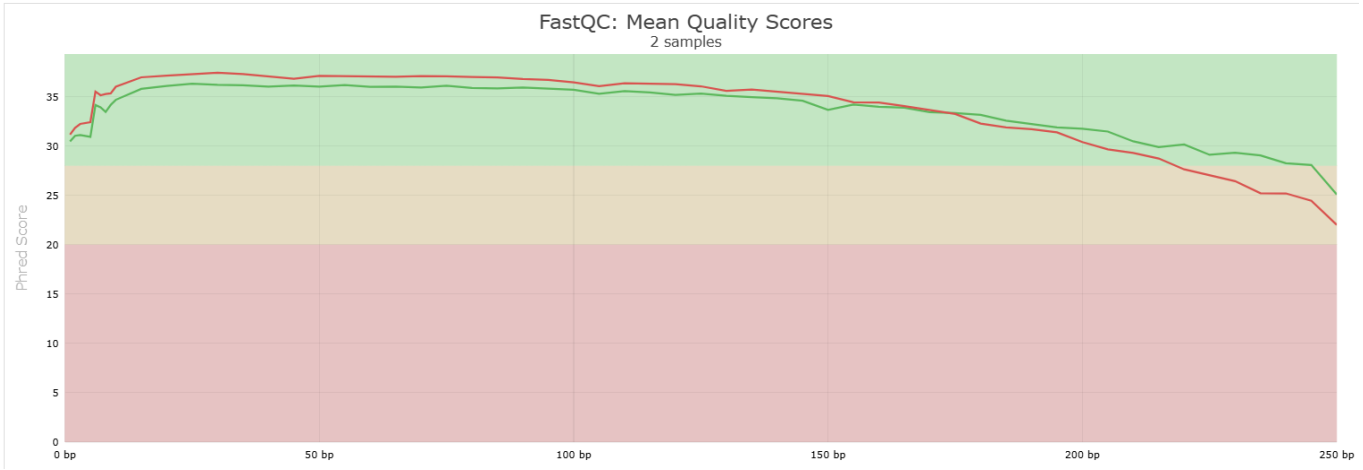
Trimmed

General Statistics

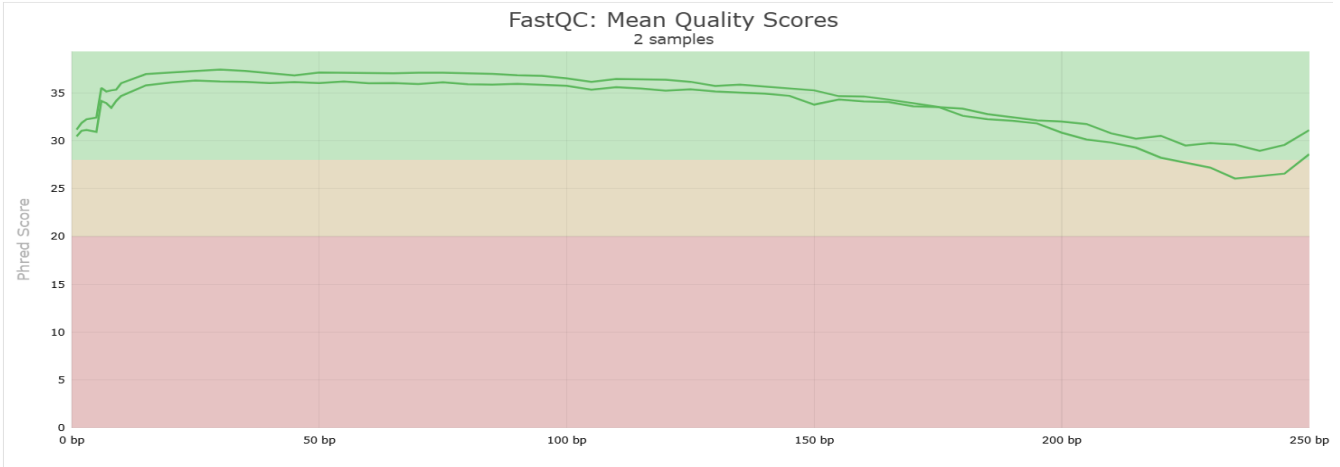
Copy table | Configure columns | Scatter plot | Violin plot | Export as CSV... | Showing 2/2 rows and 6/6 columns. | Summarize table

Sample Name	Dups	GC	Avg len	Median len	Failed	Seqs
trimmed_1	53.5 %	44.0 %	230 bp	247 bp	18 %	0.0 M
trimmed_2	55.5 %	44.0 %	228 bp	247 bp	18 %	0.0 M

The MultiQC results show clear improvements after trimming—both reads became slightly shorter due to adapter and low-quality base removal, while the reverse read failure rate dropped by 9%. This confirms trimming successfully cleaned the data, boosting reliability for downstream analysis.

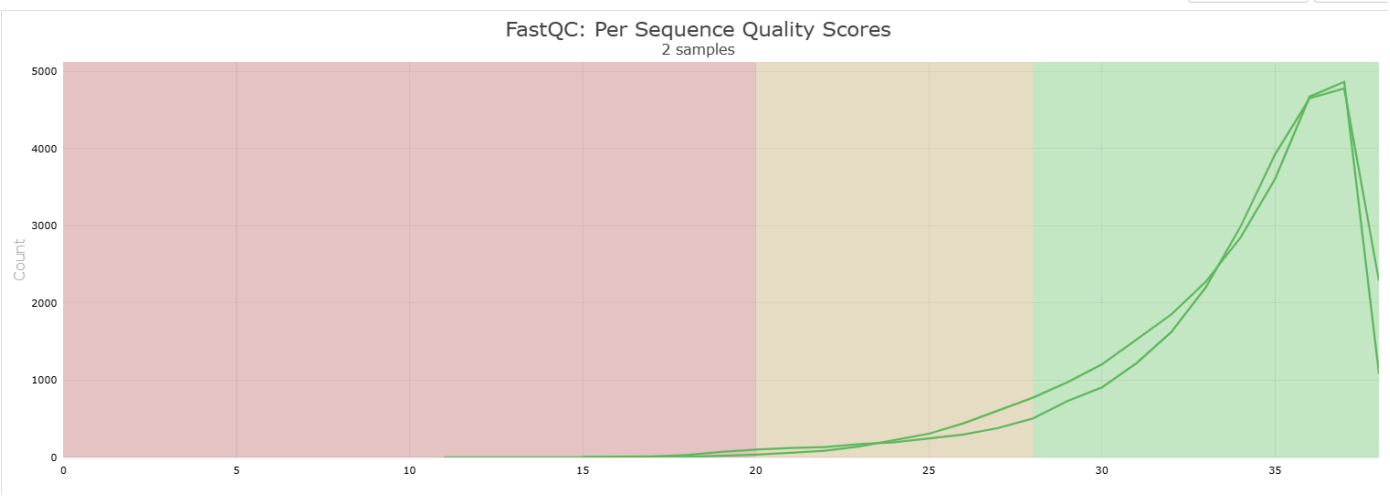


Untrimmed

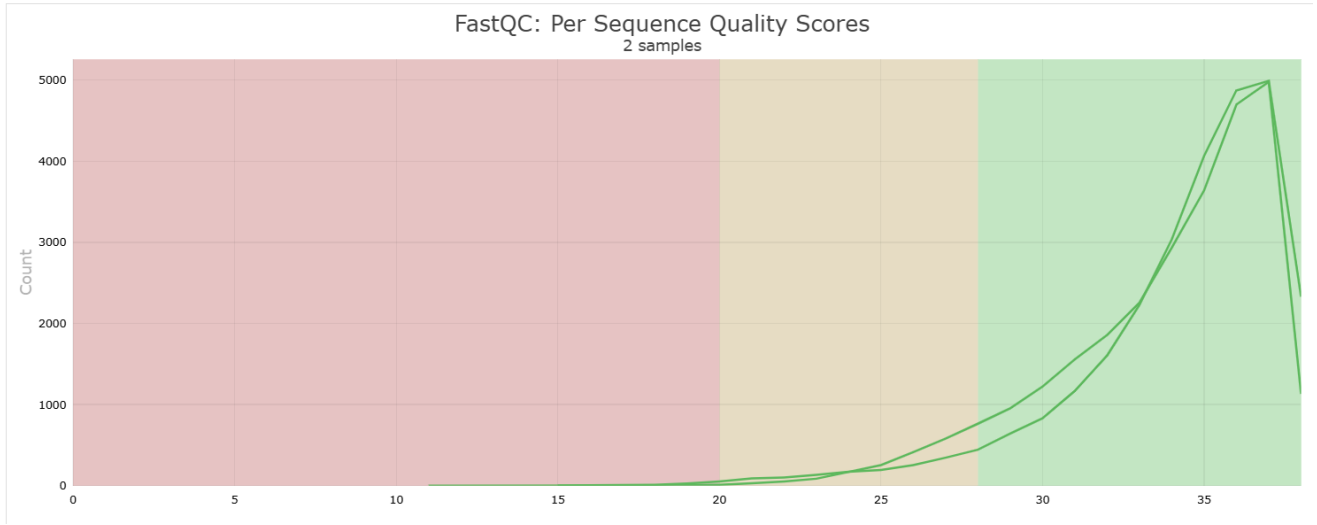


Trimmed

Untrimmed forward/reverse reads show good initial quality (Q>30) until ~200bp, with reverse degrading below Q25 at ends—typical for NGS. Post-trimming, reverse's severe drop resolves, though minor decline persists (Q26 at 230-240bp). Forward maintains high quality, confirming trimming effectively addressed end artifacts while preserving data utility.

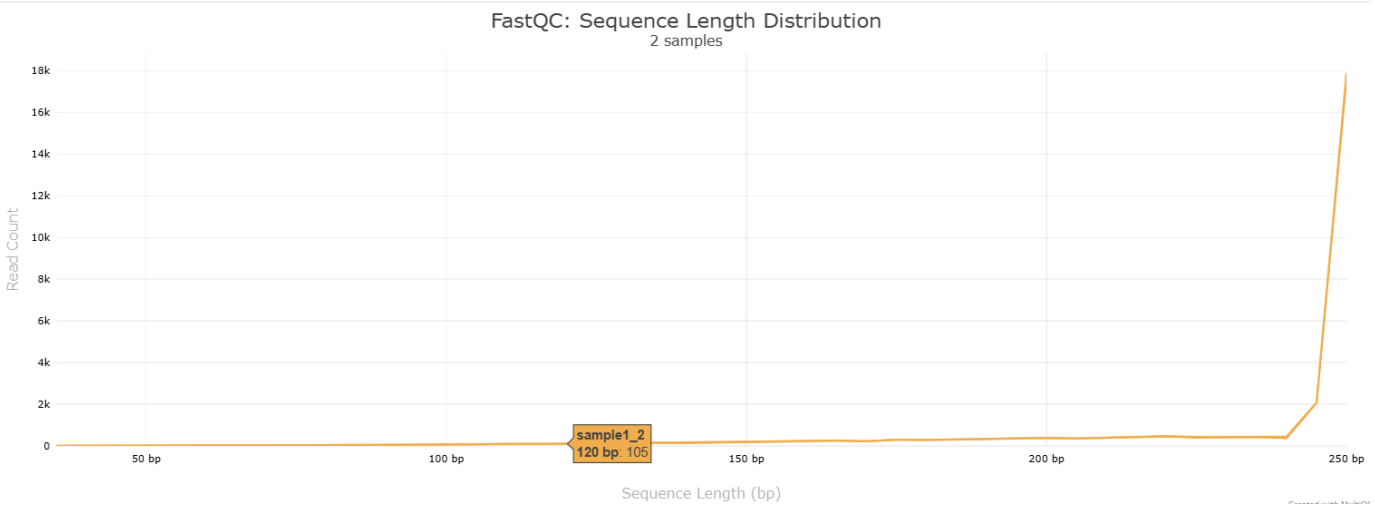


Untrimmed

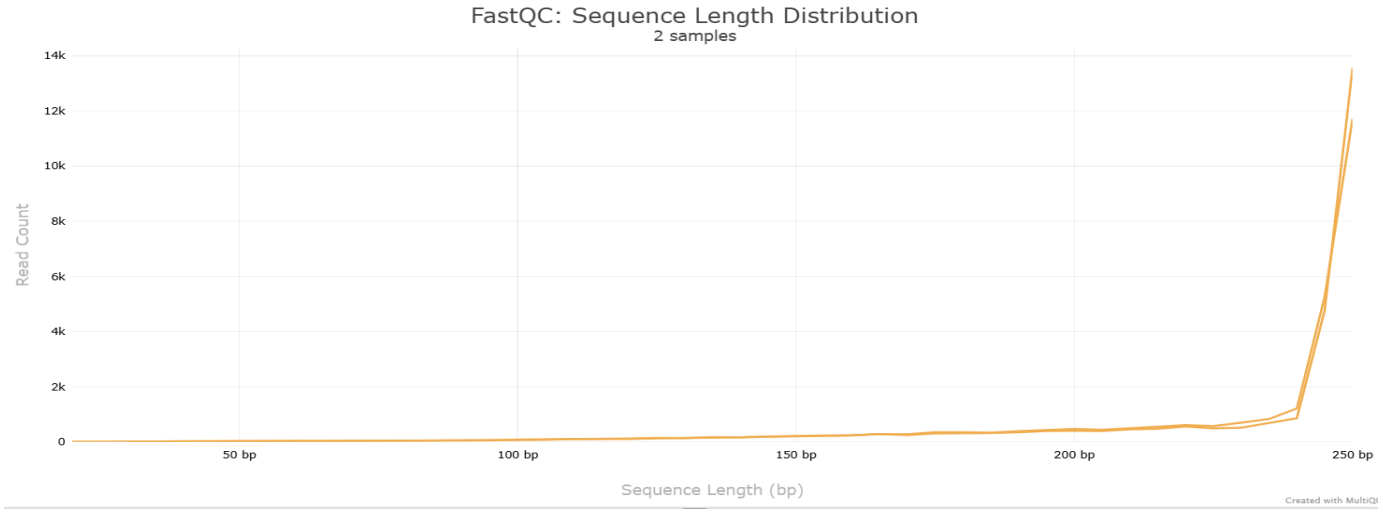


Trimmed

The per-sequence quality scores show minimal overall changes post-trimming. However, reverse reads (R2) exhibit a slight increase in high-quality sequences: for example, reads with a PHRED score of 37 rose from 4,779 to 4,978 after trimming. This indicates targeted improvement in R2 reliability without major shifts in overall quality profiles.



Untrimmed



Trimmed

While overall quality metrics show minimal changes, trimming significantly impacted read lengths. Both R1 and R2 exhibited reduced 250bp reads post-processing (from ~17.5k to 13.5k for R1 and 11.7k for R2). Trimmed data also displays greater variability in sequence lengths compared to raw reads, suggesting selective removal of low-quality/adapter-containing regions. These shifts align with expected trimming behavior, where shorter or fragmented reads are filtered to improve data reliability.

Cutadapt Report

```
(bioinfo) megatron@Taha:/mnt/d/HumanGenome$ cutadapt -q 20 -m 20 -o trimmed_1.fq.gz -p trimmed_2.fq.gz sample1_1.fq.gz sample1_2.fq.gz
This is cutadapt 4.0 with Python 3.9.21
Command line parameters: -q 20 -m 20 -o trimmed_1.fq.gz -p trimmed_2.fq.gz sample1_1.fq.gz sample1_2.fq.gz
Processing paired-end reads on 1 core ...
Done           00:00:00       27,606 reads @  28.2 µs/read;   2.13 M reads/minute
Finished in 0.78 s (28 µs/read; 2.11 M reads/minute).

=== Summary ===

Total read pairs processed:             27,606

== Read fate breakdown ==
Pairs that were too short:                  14 (0.1%)
Pairs written (passing filters):        27,592 (99.9%)

Total basepairs processed:    12,850,084 bp
  Read 1:     6,424,378 bp
  Read 2:     6,425,706 bp
Quality-trimmed:                  200,520 bp (1.6%)
  Read 1:        70,358 bp
  Read 2:       130,162 bp
Total written (filtered):     12,646,002 bp (98.4%)
  Read 1:     6,350,553 bp
  Read 2:     6,295,449 bp
```

The Cutadapt output reveals R2 required significantly more trimming (130,162bp) compared to R1 (70,358bp), totaling 200,520bp removed due to low quality. An additional 14 bases were discarded for being too short. This aligns perfectly with the MultiQC data showing R2's consistently lower quality, explaining why more aggressive trimming was needed for the reverse reads to achieve quality standards.

fastp report

Summary

General

fastp version:	0.23.2 (https://github.com/OpenGene/fastp)
sequencing:	paired end (251 cycles + 251 cycles)
mean length before filtering:	232bp, 232bp
mean length after filtering:	231bp, 231bp
duplication rate:	0.550605%
Insert size peak:	250

Before filtering

total reads:	55.212000 K
total bases:	12.850084 M
Q20 bases:	11.465327 M (89.223751%)
Q30 bases:	10.838418 M (84.345114%)
GC content:	44.618681%

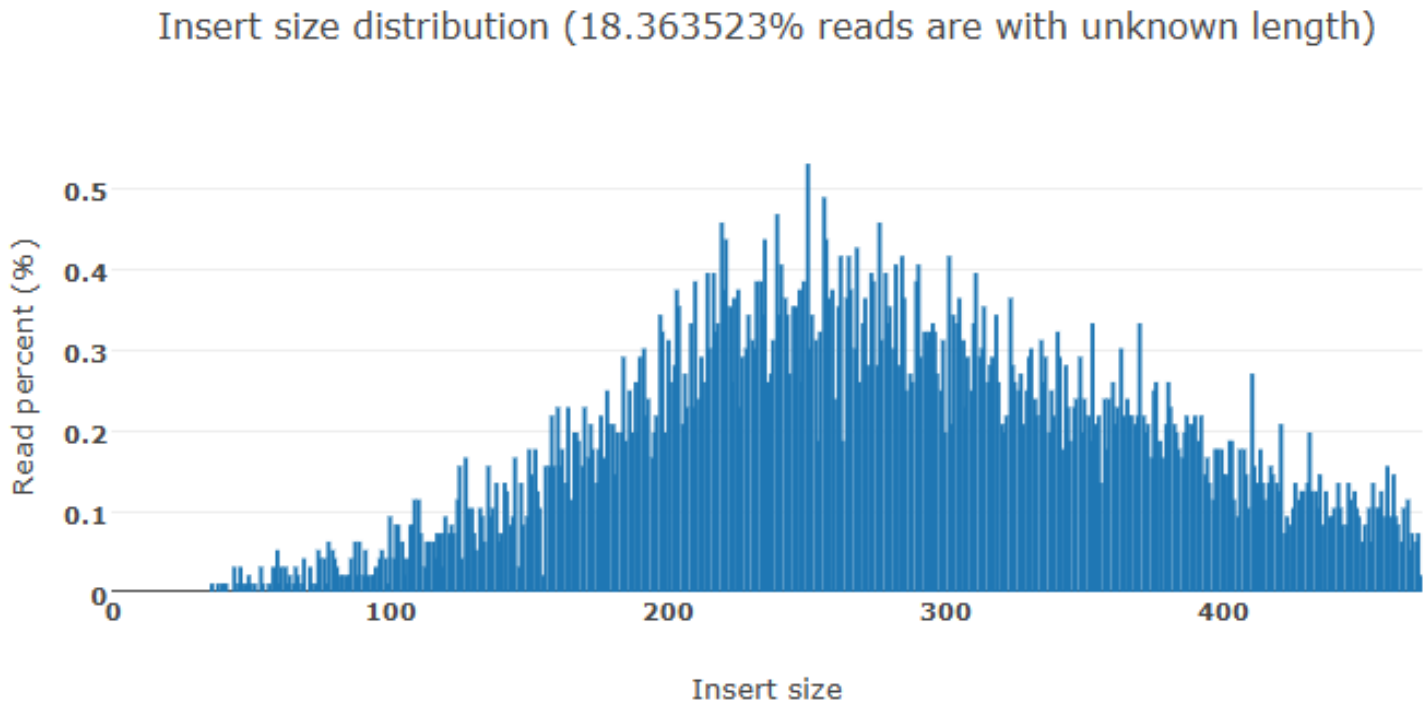
After filtering

total reads:	51.718000 K
total bases:	11.990111 M
Q20 bases:	10.909680 M (90.988983%)
Q30 bases:	10.382259 M (86.590183%)
GC content:	44.522524%

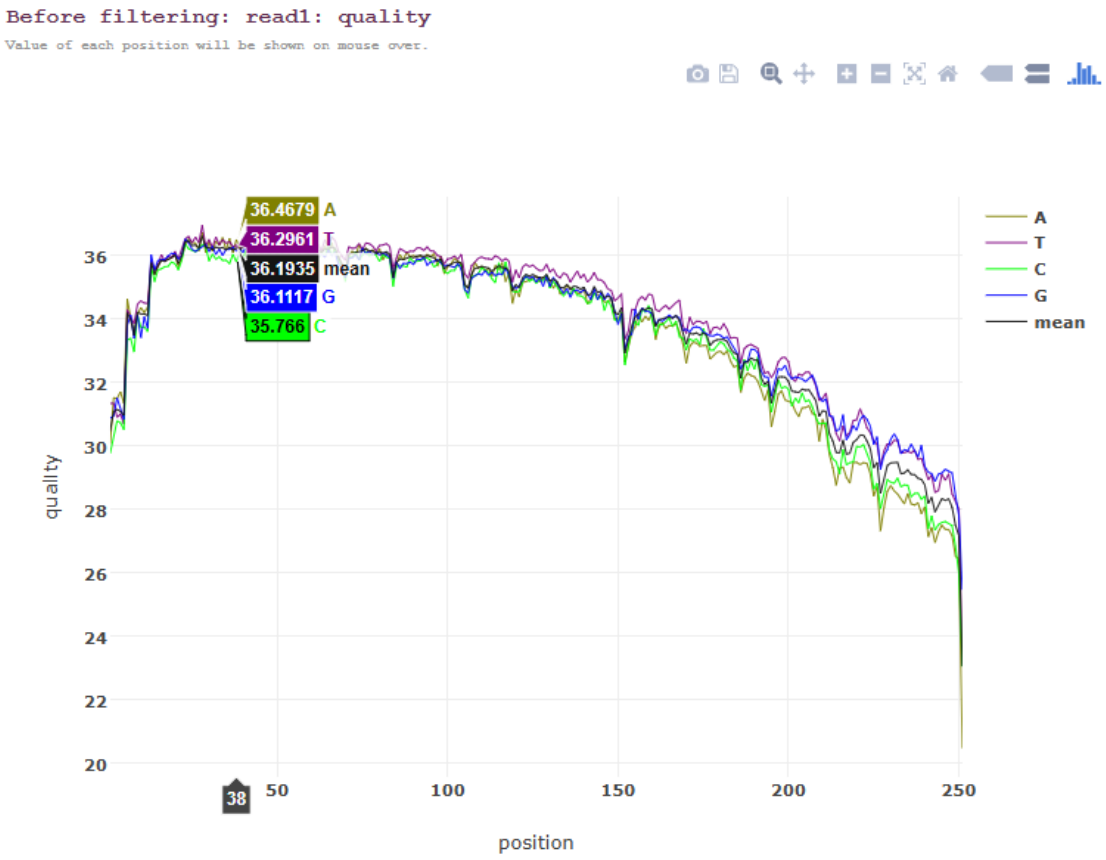
Filtering result

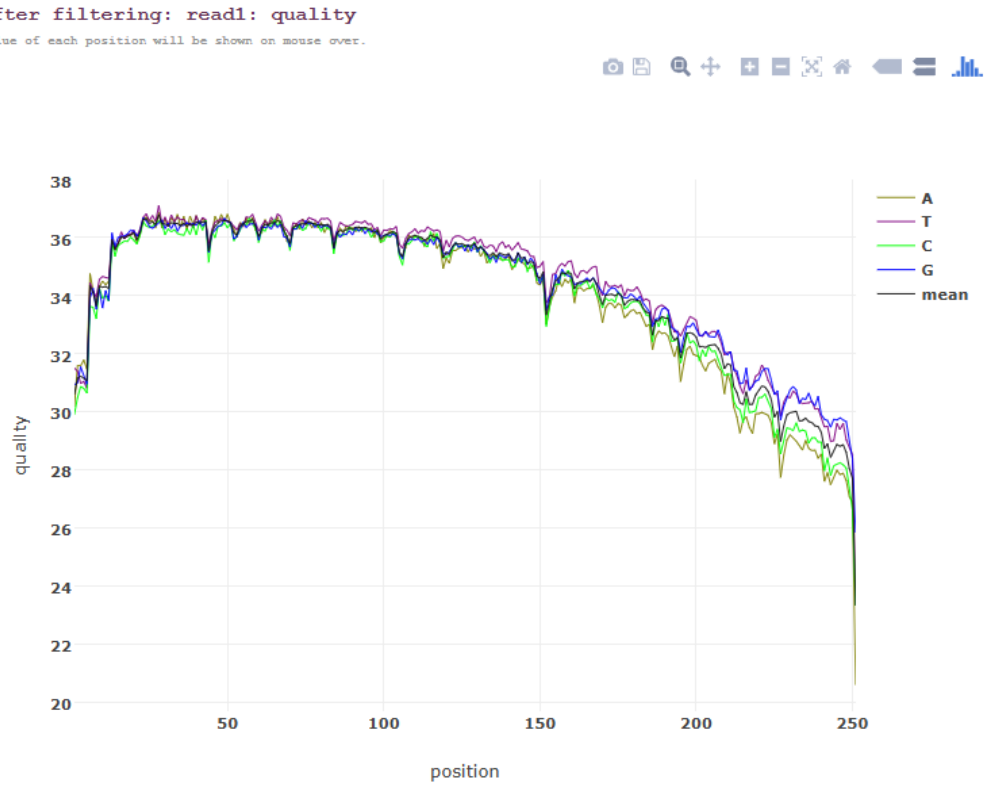
reads passed filters:	51.718000 K (93.671666%)
reads with low quality:	3.376000 K (6.114613%)
reads with too many N:	118 (0.213722%)
reads too short:	0 (0.000000%)

The fastp analysis yielded several key findings. While the total read count decreased by approximately 4,000 after processing, this reduction corresponded with measurable quality improvements - the percentage of bases with Q20+ scores increased from 89% to 90.9%, with similar gains observed for Q30 bases. Notably, read lengths remained largely unchanged and duplication rates stayed exceptionally low at just 0.55%. The filtering removed 6% of reads for failing quality thresholds, along with 0.21% containing excessive ambiguous bases (Ns), though no reads were excluded for being too short. Importantly, the GC content distribution remained stable throughout processing.

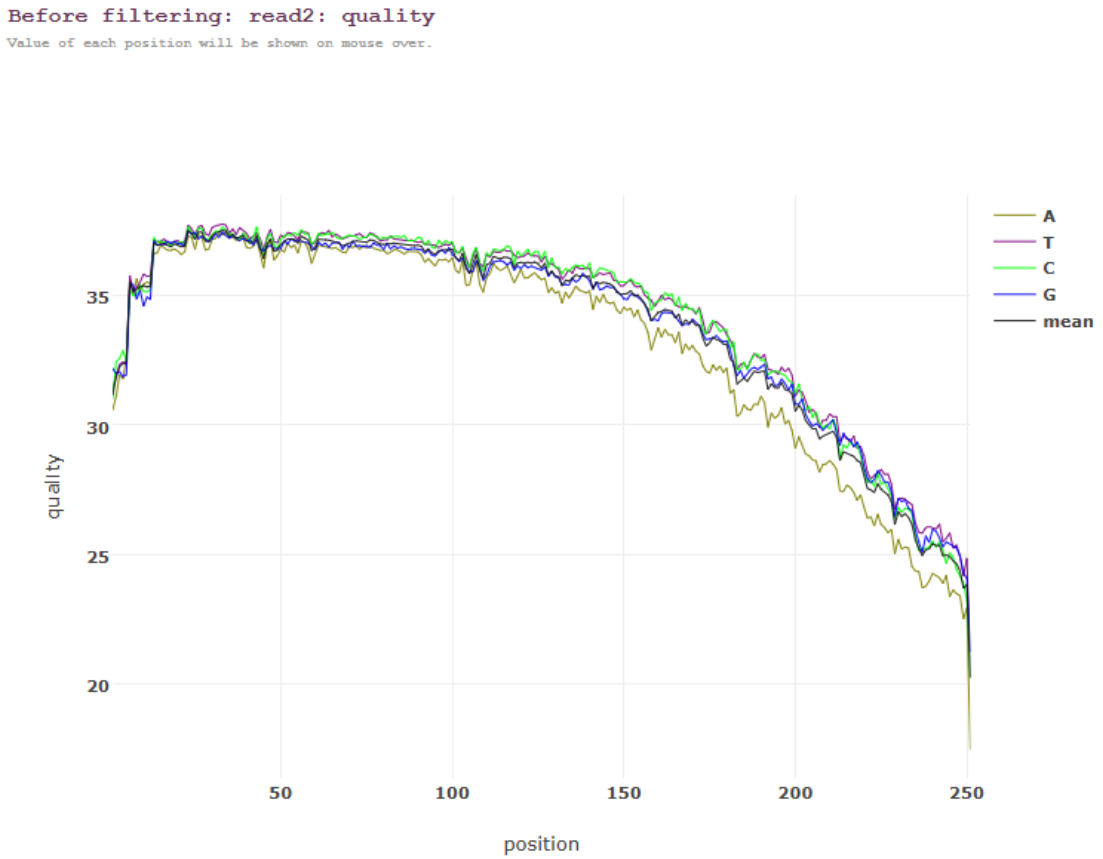


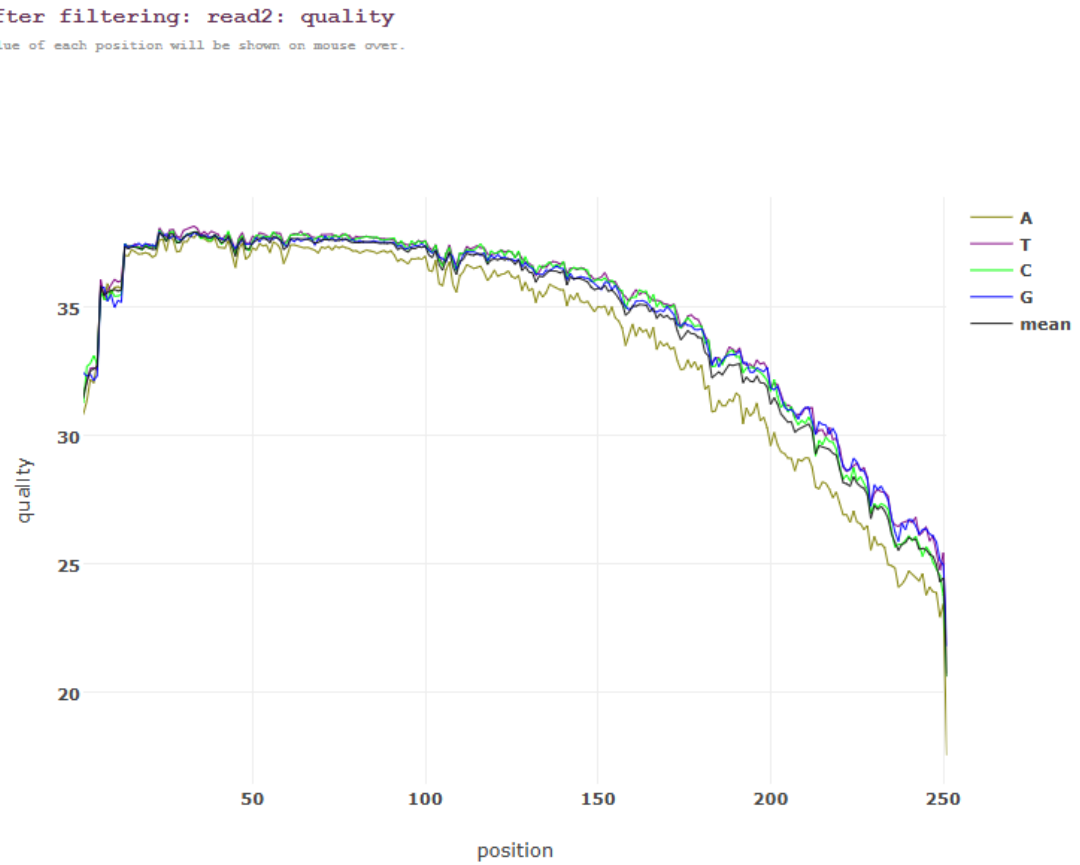
The insert size distribution peaks at 250bp (>0.5% of reads), while 18.36% couldn't be sized - likely fragments <30bp or >470bp based on the detectable range. This indicates some library preparation artifacts exist alongside the dominant fragment population.



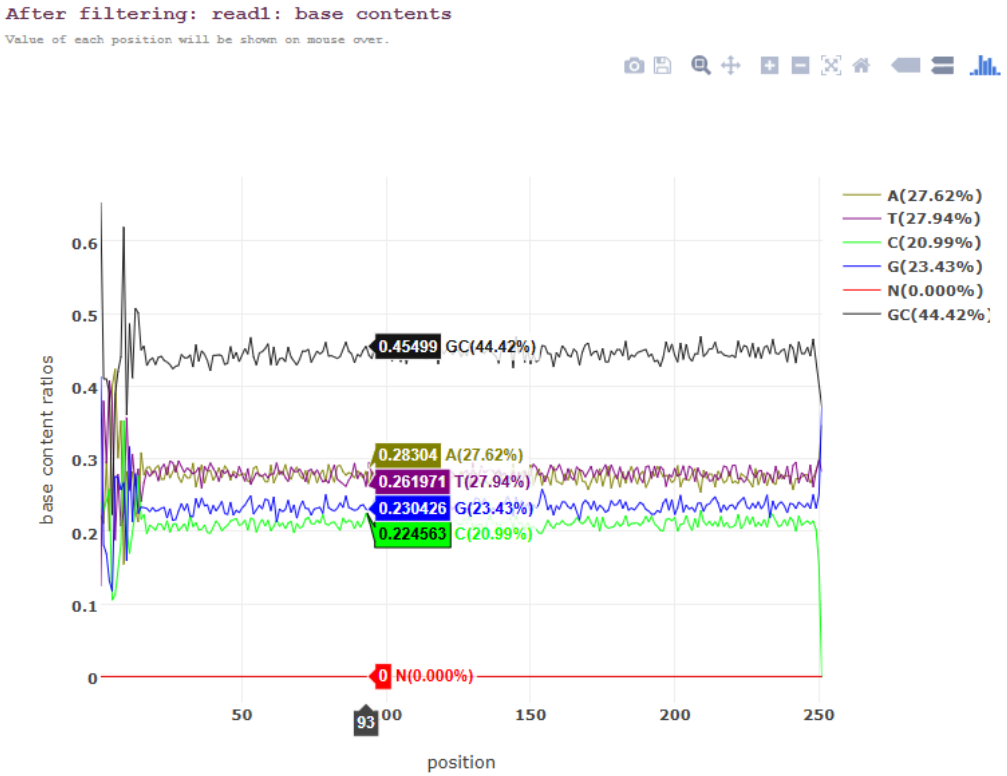


The quality scores across 250bp show nearly identical trends before and after filtering, with both datasets averaging Q27 at the end position. The consistent Y-axis range (20-30) confirms all recorded bases maintain Q20+ scores throughout. This stability indicates the filtering preserved quality while removing problematic reads without altering the remaining data's profile.

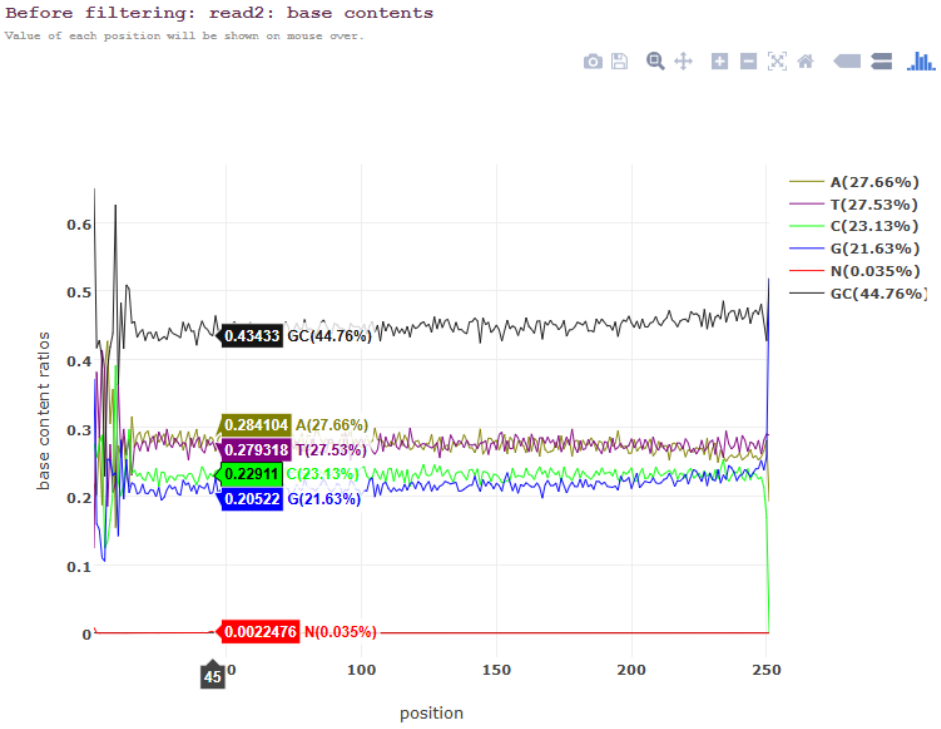


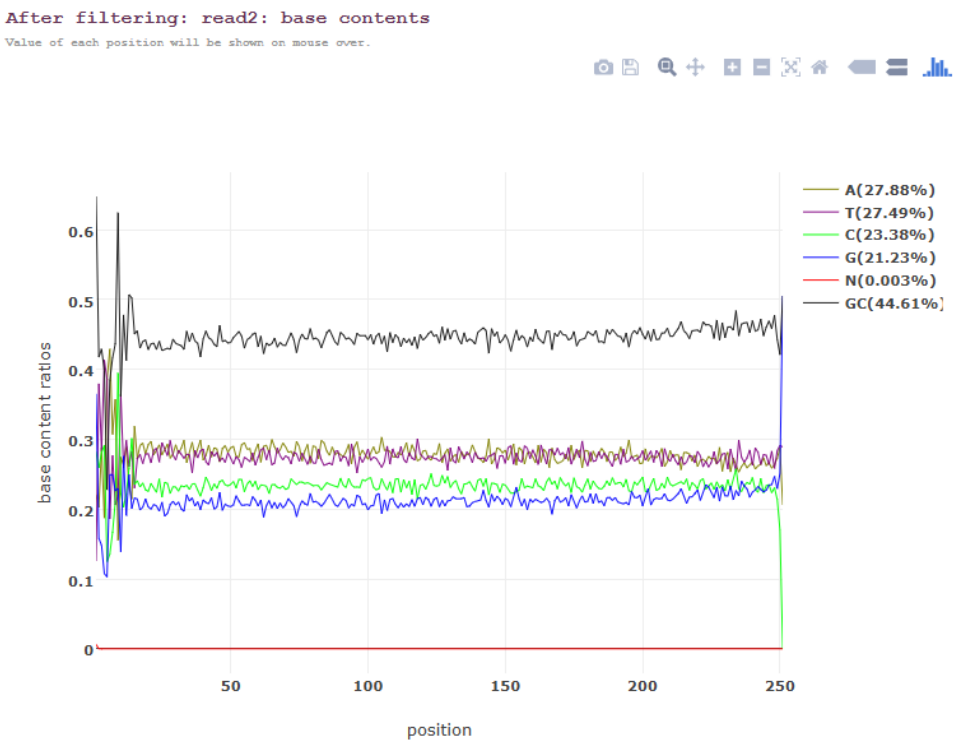


The analysis reveals minimal differences overall, though Read 2 shows a notable quality decline at the 250bp position. Specifically, the "A" nucleotide scores drop below Q20 at this endpoint, pulling down the average. This pattern differs from Read 1's more stable quality profile, highlighting Read 2's greater susceptibility to end-of-read degradation - a common phenomenon in paired-end sequencing that trimming couldn't fully resolve.

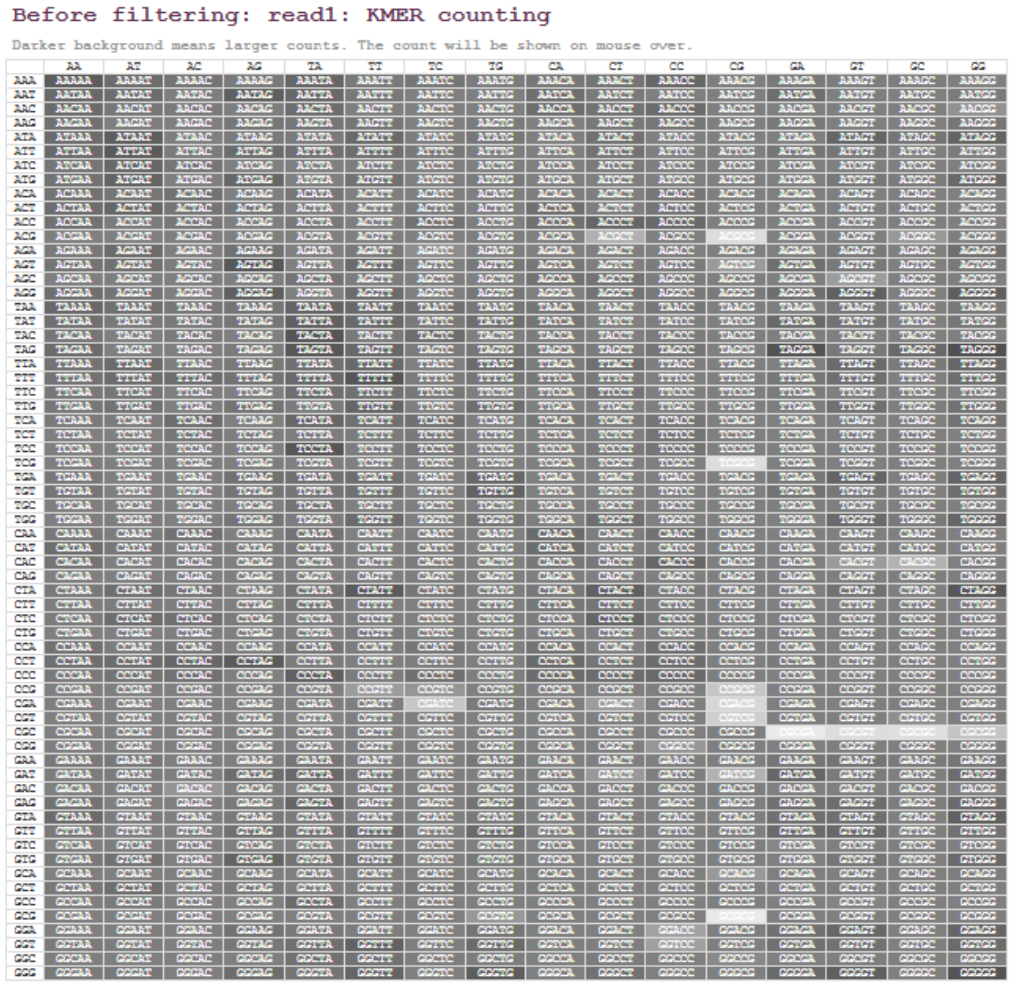


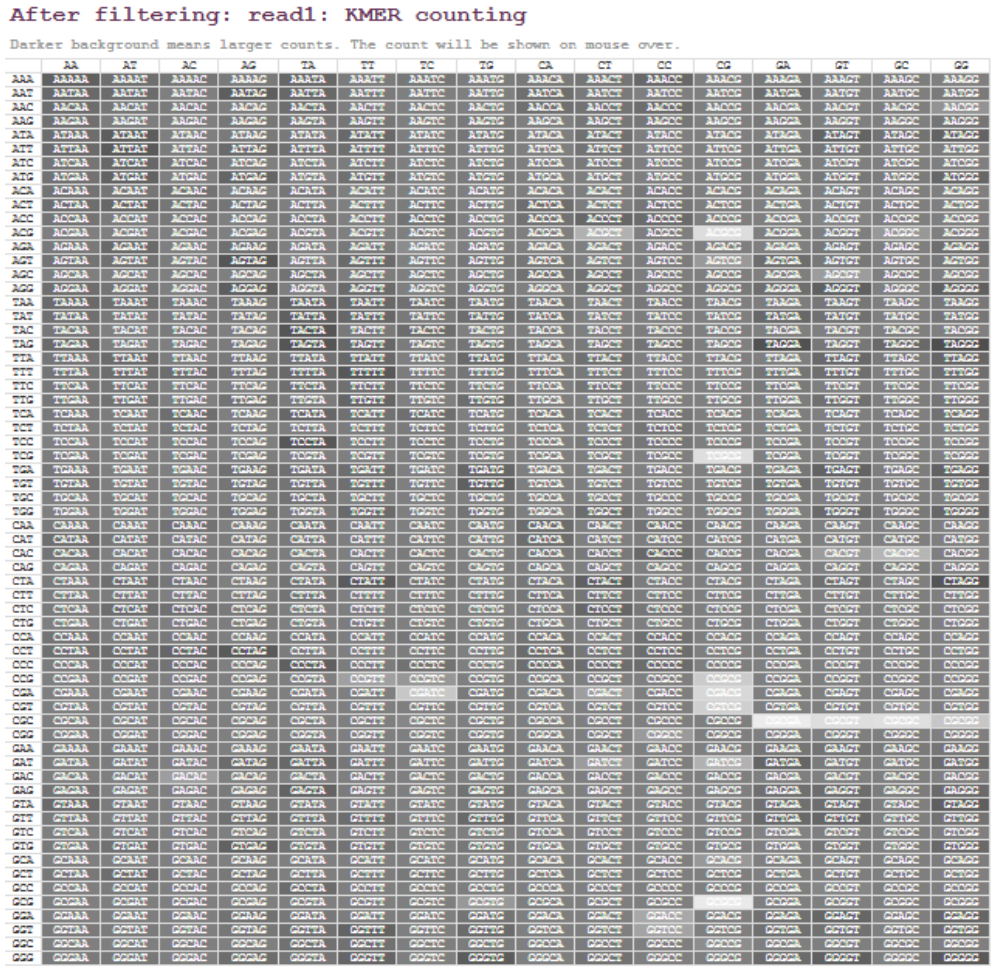
No significant differences in base contents visible in the data before and after filtering





No significant differences in base contents visible in the data before and after filtering

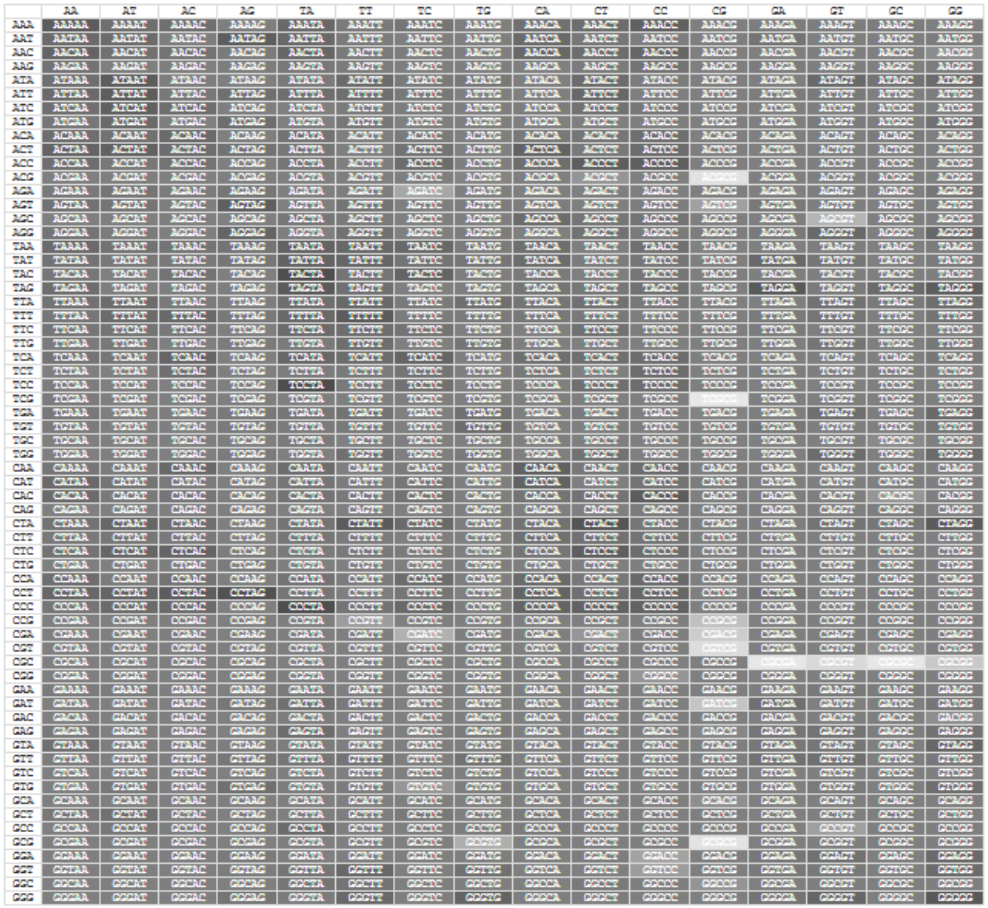




The k-mer analysis reveals consistent patterns between raw and filtered data, with no significant differences in sequence composition. The heatmap's similar shading intensity across both datasets indicates that filtering preserved the original k-mer distribution. This suggests the quality control process maintained the fundamental sequence characteristics while selectively removing problematic reads. The darkest cells correspond to the most frequent length combinations, which remain stable post-filtering.

After filtering: read2: KMER counting

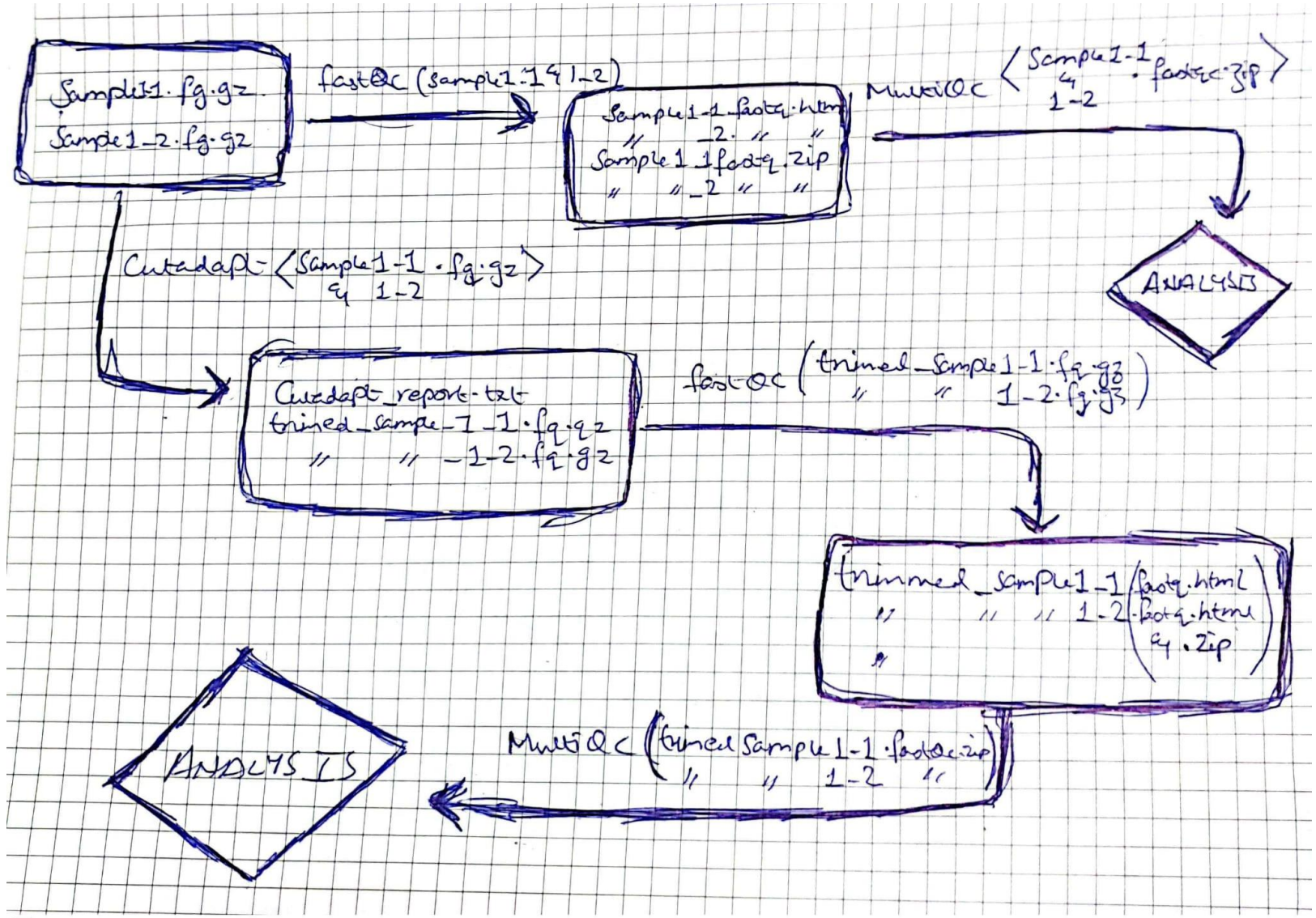
Darker background means larger counts. The count will be shown on mouse over.



Before filtering: read2: KMER counting

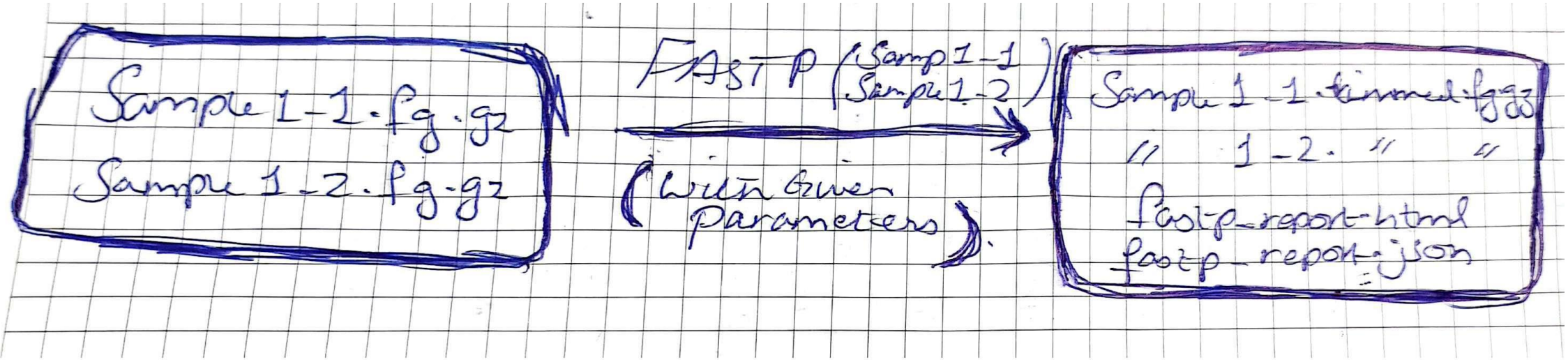
Darker background means larger counts. The count will be shown on mouse over.

Same as previous analysis



WORKFLOW 1A

WORKFLOW 1B



When working with clean data, fastp is convenient because it handles trimming and basic quality checks all at once, which saves time and keeps things tidy. But if I'm aiming for a more in-depth look at quality metrics or need to compare multiple runs in detail, I'd lean toward using FastQC followed by Cutadapt and summarized with MultiQC. This combination gives better insights into subtle differences, especially when data quality isn't ideal. In our case, since duplication levels were already low, fastp didn't highlight much variation, but with noisier data, those differences would likely stand out more according to my understanding.

Cutadapt Report

Question 2 A

```
(bioinfo) megatron@Taha:/mnt/d/HumanGenome$ cutadapt -q 20 -m 20 -o sample2_trimmed_1.fq.gz -p sample2_trimmed_2.fq.gz sample2_1.fastqsanger sample2_2.fastqsanger
This is cutadapt 4.0 with Python 3.9.21
Command line parameters: -q 20 -m 20 -o sample2_trimmed_1.fq.gz -p sample2_trimmed_2.fq.gz sample2_1.fastqsanger sample2_2.fastqsanger
Processing paired-end reads on 1 core ...
Done           00:00:07       339,276 reads @  21.5 µs/read;   2.80 M reads/minutee
Finished in 7.29 s (21 µs/read; 2.79 M reads/minute).

=== Summary ===

Total read pairs processed:            339,276

== Read fate breakdown ==
Pairs that were too short:              10,175 (3.0%)
Pairs written (passing filters):       329,101 (97.0%)

Total basepairs processed:    33,927,600 bp
  Read 1:    16,963,800 bp
  Read 2:    16,963,800 bp
Quality-trimmed:                 807,738 bp (2.4%)
  Read 1:       234,433 bp
  Read 2:       573,305 bp
Total written (filtered):     32,600,871 bp (96.1%)
  Read 1:    16,310,200 bp
  Read 2:    16,290,671 bp
```

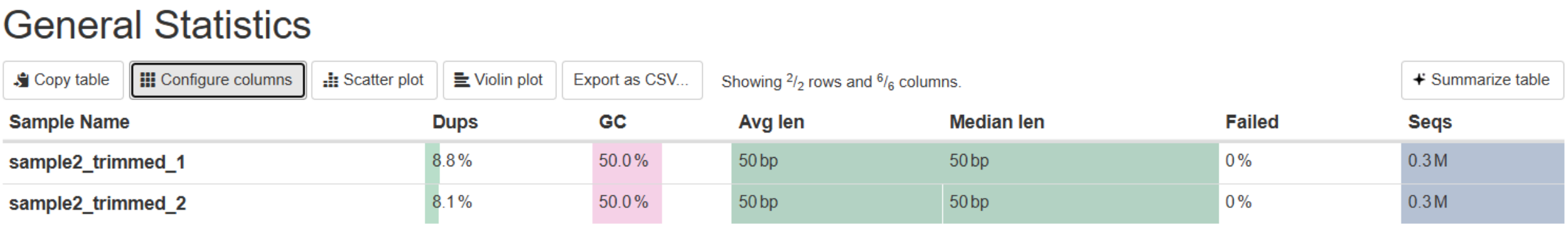
The Cutadapt trimming removed 3% of reads for being too short and trimmed 2.4% due to low quality. These modest percentages indicate the original data was relatively clean, with the majority of reads meeting length and quality thresholds.

Untrimmed

Copy table | Configure columns | Scatter plot | Violin plot | Export as CSV... | Showing 2/2 rows and 6/6 columns. | Summarize table

Sample Name	Dups	GC	Avg len	Median len	Failed	Seqs
sample2_1	9.3%	50.0%	50 bp	50 bp	0%	0.3 M
sample2_2	8.5%	50.0%	50 bp	50 bp	0%	0.3 M

Trimmed



General Statistics

Copy table | Configure columns | Scatter plot | Violin plot | Export as CSV... | Showing 2/2 rows and 6/6 columns. | Summarize table

Sample Name	Dups	GC	Avg len	Median len	Failed	Seqs
sample2_trimmed_1	8.8%	50.0%	50 bp	50 bp	0%	0.3 M
sample2_trimmed_2	8.1%	50.0%	50 bp	50 bp	0%	0.3 M

The general statistics comparison shows that other than duplication and 3.9% trimmed bases, the GC % and length has not changed

Untrimmed

Trimmed

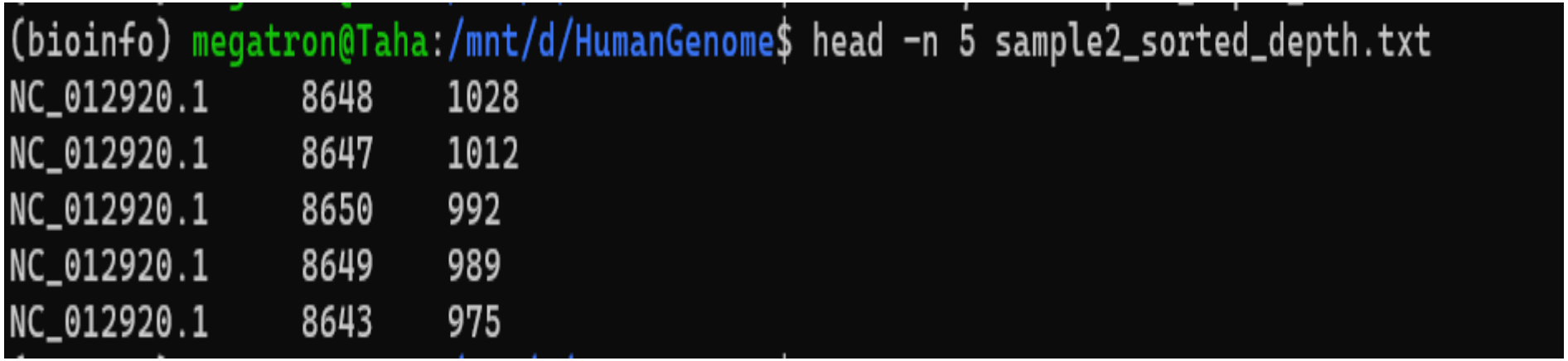
The mean quality scores confirm high-quality reads both before and after trimming, with noticeable improvement post-processing. While both datasets maintained strong baseline quality, the trimming enhanced scores further without altering the fundamental profile

```
(bioinfo) megatron@Taha:/mnt/d/HumanGenome$ head sample2_aligned.sam
@SQ     SN:NC_000001.11 LN:248956422
@SQ     SN:NT_187361.1  LN:175055
@SQ     SN:NT_187362.1  LN:32032
@SQ     SN:NT_187363.1  LN:127682
@SQ     SN:NT_187364.1  LN:66860
@SQ     SN:NT_187365.1  LN:40176
@SQ     SN:NT_187366.1  LN:42210
@SQ     SN:NT_187367.1  LN:176043
@SQ     SN:NT_187368.1  LN:40745
@SQ     SN:NT_187369.1  LN:41717
```

After QC, the alignment was performed.
Above is a screenshot of the files generated after indexing and aligning, along with a snippet of the .sam file

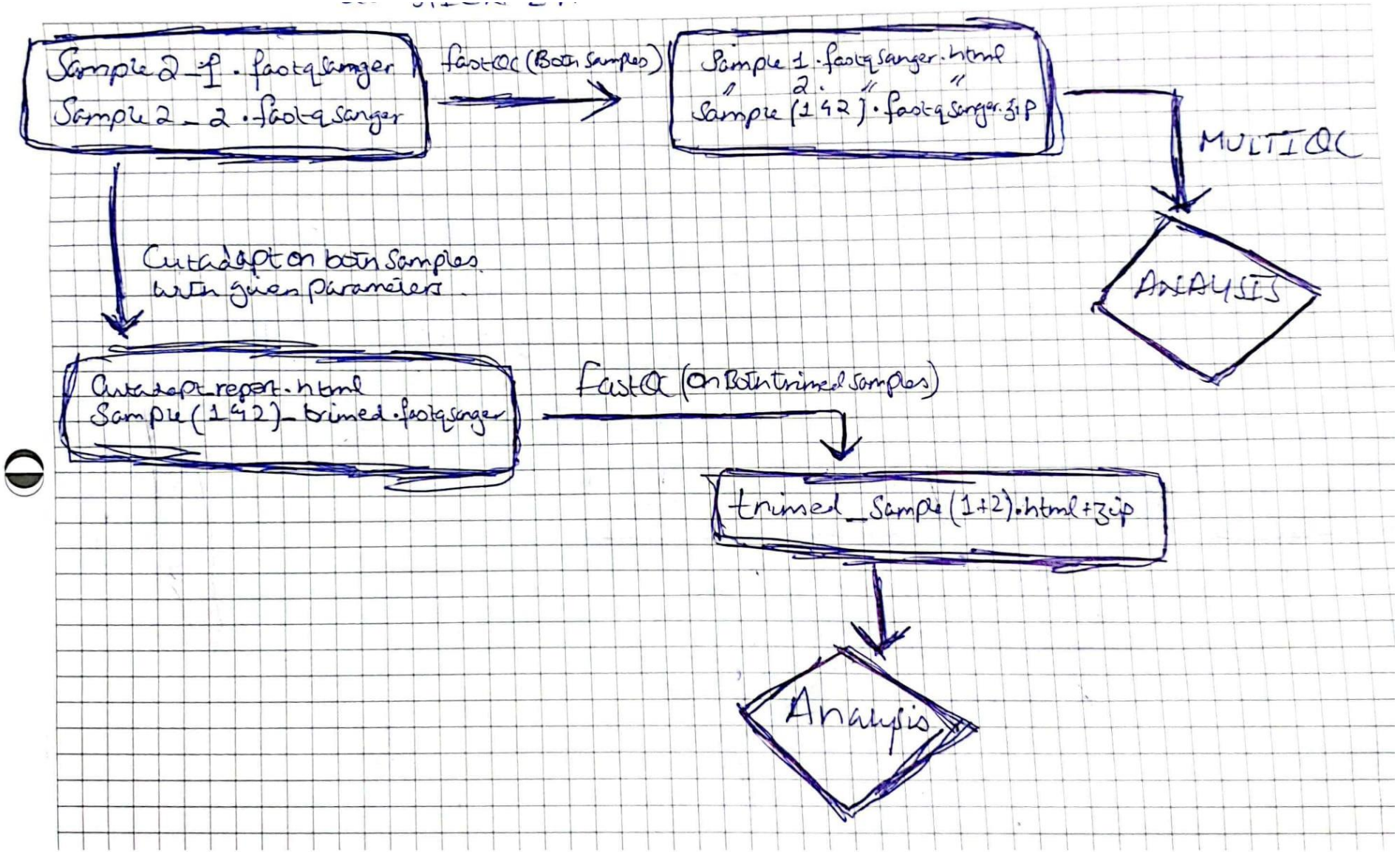
```
@SQ     SN:NC_012920.1  LN:16569
@PG     ID:bwa  PN:bwa  VN:0.7.17-r1188 CL:bwa mem -t 8 GCF_000001405.40_GRCh38.p14_genomic.fna sample2_trimmed_1.fq.gz sample2_trimmed_2.fq.gz
HWUSI-EAS1794_0001_FC61KOJ:8:49:15990:3926#0    83      NC_000012.12    120169850       60      50M     =       120169669       -231    TCCTTTTGAATTTTGTACTGTATGAATGTATAAC
CAACTGAAACAAACAG        IIIGIIIIHIIIHIIIIIIIIIIIIGHIIIIIIIIIIIIIIIIIIIII      NM:i:0  MD:Z:50 MC:Z:50M        AS:i:50 XS:i:20
HWUSI-EAS1794_0001_FC61KOJ:8:49:15990:3926#0    163     NC_000012.12    120169669       40      50M     =       120169850       231     AATTTTTGTACTTTTAGTAAAGACGGGGTTTCAC
CATGCTGGCCAGAATG        EHIHIIIIIHIIIIFIIHIHIHGIIHDIHHIIIIIGIIIIIIGIIIIEII      NM:i:1  MD:Z:38T11      MC:Z:50M        AS:i:45 XS:i:45
HWUSI-EAS1794_0001_FC61KOJ:8:106:17677:4998#0   83      NC_000002.12    175179212       21      41M9S   =       175179057       -196    AGACACACCATTCTGGGCCCCATTAAATACCGTA
GAGCCCTCTCCAGTCC        FGGGFGEGGGGEGGGEHEBHGGGGGGGGFGHHHHGHGEHHHDHHHHHHHH      NM:i:0  MD:Z:41 MC:Z:16S34M     AS:i:41 XS:i:37 XA:Z:NC_000006.12,-7331554,47M3S,2;
HWUSI-EAS1794_0001_FC61KOJ:8:106:17677:4998#0   163     NC_000002.12    175179057       9       16S34M  =       175179212       196     CTTCAGCGAAGGGTTTCTGGCATAACCAATGATA
AGGCTGCCAAAGACTG        IIIFFIGIIIGGIIIIDIIFIIIIIIGHIFIGIDEIHGIGG<FEFG@<DG      NM:i:0  MD:Z:34 MC:Z:41M9S      AS:i:34 XS:i:36 XA:Z:NC_000006.12,+7331394,9S41M,1;
HWUSI-EAS1794_0001_FC61KOJ:8:12:4514:3966#0     99      NC_000014.9     92940599        0       50M     =       92940749        200     GAGGCTCCCGCGCCTATCCTCACCTAGGTGACTT
TATGGAAAGGCAGGCT        HDFEHHHHHGGHHHHHHHHHHHGHGHHHCGEEGGGHGGGGHHDHHHHHBH      NM:i:0  MD:Z:50 MC:Z:50M        AS:i:50 XS:i:50 XA:Z:NT_187601.1,+55161,50M,0;
HWUSI-EAS1794_0001_FC61KOJ:8:12:4514:3966#0     147     NC_000014.9     92940749        0       50M     =       92940599        -200    TTGCCCAAATCACAGCACAAATCCTCAGCACAGG
CGCCATCGGCTCGGGC        HGHHFFCHEEGGHHIHEBIIHFHGEIIGIIIIIII@IGHEDIDIIIIIII      NM:i:0  MD:Z:50 MC:Z:50M        AS:i:50 XS:i:50 XA:Z:NT_187601.1,-55311,50M,0;
HWUSI-EAS1794_0001_FC61KOJ:8:38:8158:19548#0    65      NC_000003.12    101794419       60      50M     NC_000017.11    64502186        0       GGGTAGGACATTCAGGGAGTTTCTGG
CTGTGAAGAAGATGGAAGGTAGGA        GGGBGIIHIIIIIIIIIHIFIIIIIIIIIIIIIIGHIIHIGEEGGBGGGE      NM:i:0  MD:Z:50 MC:Z:40M        AS:i:50 XS:i:0
HWUSI-EAS1794_0001_FC61KOJ:8:38:8158:19548#0    129     NC_000017.11    64502186        60      40M     NC_000003.12    101794419       0       GTTGACTCTTGTCACCATGGATACCC
ATGGCAGGCCACCC  IIIIIIIIIIIIHGIIEGIGBGIIIGIBIIHHIIBD>DDD        NM:i:1  MD:Z:39T0       MC:Z:50M        AS:i:39 XS:i:0
HWUSI-EAS1794_0001_FC61KOJ:8:106:18976:17584#0  81      NC_000022.11    23310392        0       43M7S   =       23295110        -15325  ACAGACAGGGGTCTTCGGAGTCAAGATTGCTGTG
GTCACCAAGAGAGAGA        IIIIIIIIIIIGIIIIIIIIIIIIIIIIIIIHIIIIIIIIIIHIIIIIII      NM:i:0  MD:Z:43 MC:Z:46M4S      AS:i:43 XS:i:43 XA:Z:NC_000022.11,-24645704,43M7S,0;NC_000
022.11,-22631029,43M7S,0;
HWUSI-EAS1794_0001_FC61KOJ:8:106:18976:17584#0  161     NC_000022.11    23295110        60      46M4S   =       23310392        15325   GGACGGCGAGAGCACGGACAGACTCATGGGGAAG
GGCCAGGTCCAGCTGG        IIIIIIHHIIHIHIIGIIIGEBHIIFIIIIIIIIGIIIIHHBIHHIHEII      NM:i:0  MD:Z:46 MC:Z:43M7S      AS:i:46 XS:i:0
HWUSI-EAS1794_0001_FC61KOJ:8:12:5094:3287#0     99      NC_000005.10    128184805       60      50M     =       128186598       1843    GGTTAAATGAGTTATTAAAGGAACATTCAAGCAC
AGCTAATATTATTGTC        IIIIHIHIIIIIIGIIIIIIGIIIIIIIIIIIHHIIIIIHIIIIIIIII      NM:i:0  MD:Z:50 MC:Z:50M        AS:i:50 XS:i:19
HWUSI-EAS1794_0001_FC61KOJ:8:12:5094:3287#0     147     NC_000005.10    128186598       60      50M     =       128184805       -1843   GAATCATCAGAGTGTCCTTACCTTCTATTCATAA
ATGTTCTATACAGTGG        IIIIIIIIIIIIIIHDIIIIIIIHIIIIIGHHIIIIIIIHIIIIIIIII      NM:i:0  MD:Z:50 MC:Z:50M        AS:i:50 XS:i:19
```

2.b) position 120,169,850 on chromosome 12. (highlighted in yellow)

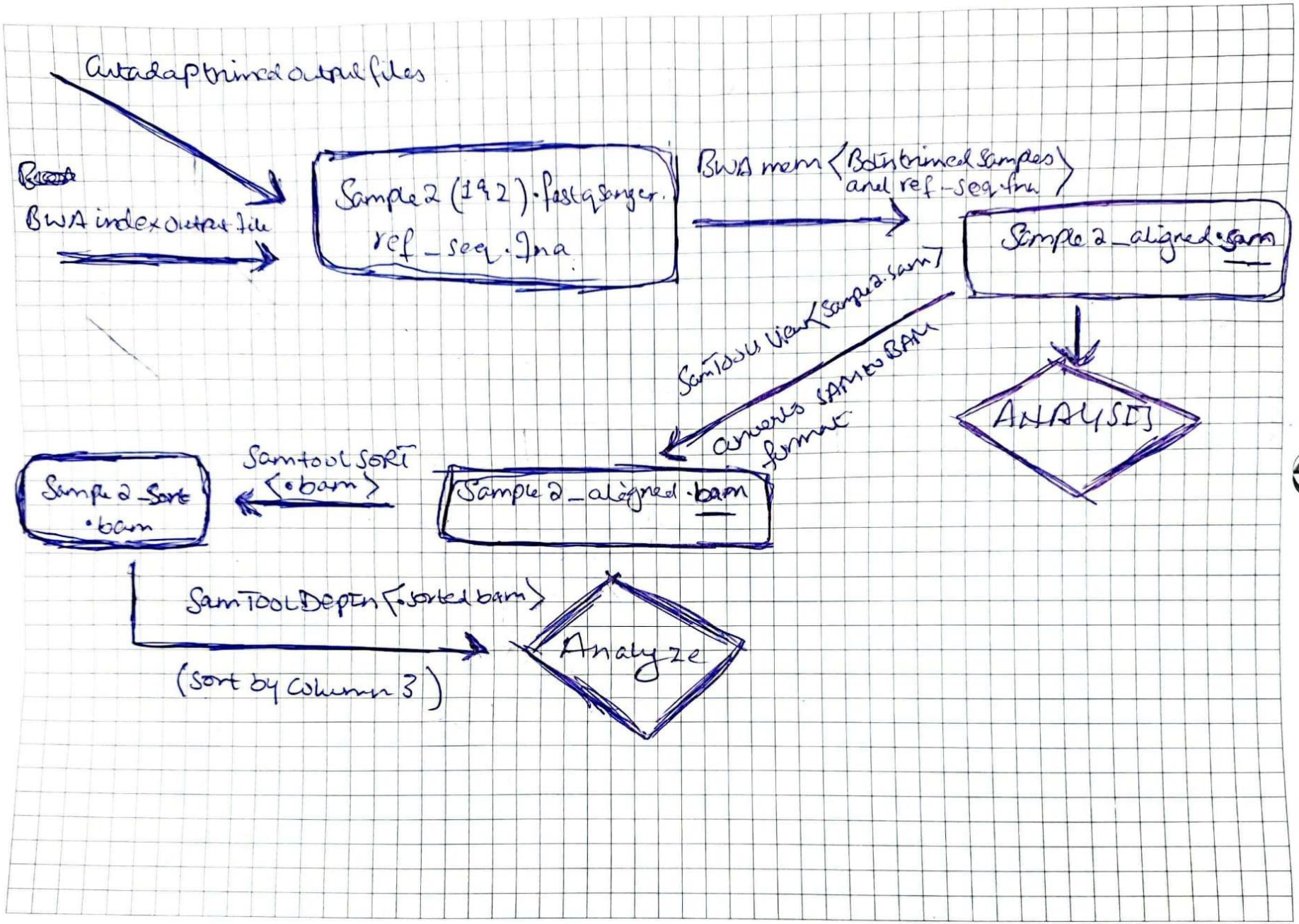


2.c) NC_012920.1 which corresponds to chrM, reads aligned the most

Q2) QC WorkFlow



ORTA DOĞU TEKNİK ÜNİVERSİTESİ



ORTA DOĞU TEKNİK ÜNİVERSİTESİ

Linux History

```
################ START ############################

## Setting up the environment

megatron@Taha:~$ sudo apt update
megatron@Taha:~$ source ~/miniconda3/bin/activate
(base) megatron@Taha:~$ conda activate bioinfo
(bioinfo) megatron@Taha:~$ conda env list


## Question 1 WorkFlow

# Raw data QC

(bioinfo) megatron@Taha:/mnt/d/HumanGenome/bin508-03$ mkdir raw_fastqc
(bioinfo) megatron@Taha:/mnt/d/HumanGenome$ fastqc sample1_1.fq.gz sample1_2.fq.gz -o raw_fastqc/
(bioinfo) megatron@Taha:/mnt/d/HumanGenome$ multiqc raw_fastqc/ -o raw_multiqc/

# trimming with cutadapt

(bioinfo) megatron@Taha:/mnt/d/HumanGenome$ cutadapt -q 20 -m 20 -o trimmed_1.fq.gz -p trimmed_2.fq.gz sample1_1.fq.gz sample1_2.fq.gz

#trimmed data QC
(bioinfo) megatron@Taha:/mnt/d/HumanGenome$ mkdir trimmed_fastqc
(bioinfo) megatron@Taha:/mnt/d/HumanGenome$ fastqc trimmed_1.fq.gz trimmed_2.fq.gz -o trimmed_fastqc/
(bioinfo) megatron@Taha:/mnt/d/HumanGenome$ multiqc trimmed_fastqc/ -o trimmed_multiqc/

# fastp

(bioinfo) megatron@Taha:/mnt/d/HumanGenome$ fastp -i sample1_1.fq.gz -I sample1_2.fq.gz -o fastp_1.fq.gz -O fastp_2.fq.gz --qualified_quality_phred 20
--length_required 20 --html fastp_report.html

## Question 2 WorkFlow

# raw data qc
(bioinfo) megatron@Taha:/mnt/d/HumanGenome$ mkdir sample2_raw_fastqc
(bioinfo) megatron@Taha:/mnt/d/HumanGenome$ fastqc sample2_1.fastqsanger sample2_2.fastqsanger -o sample2_raw_fastqc/

```

```
# trimming

(bioinfo) megatron@Taha:/mnt/d/HumanGenome$ cutadapt -q 20 -m 20 -o sample2_trimmed_1.fq.gz -p sample2_trimmed_2.fq.gz sample2_1.fastqsanger
sample2_2.fastqsanger

# trimmed data QC

(bioinfo) megatron@Taha:/mnt/d/HumanGenome$ mkdir sample2_trimmed_fastqc
(bioinfo) megatron@Taha:/mnt/d/HumanGenome$ fastqc sample2_trimmed_1.fq.gz sample2_trimmed_2.fq.gz -o sample2_trimmed_fastqc/
(bioinfo) megatron@Taha:/mnt/d/HumanGenome$ multiqc sample2_trimmed_fastqc/ -o sample2_trimmed_multiqc/

# BWA-MEM Alignment

# Align trimmed reads to reference genome using BWA-MEM
# -t 8: Use 8 CPU threads for faster alignment
# -R: Add read group information (required for downstream analysis)
# "@RG\tID:sample2\tSM:sample2\tPL:ILLUMINA": Read group metadata

(bioinfo) megatron@Taha:/mnt/d/HumanGenome$ bwa mem -t 8 -R "@RG\tID:sample2\tSM:sample2\tPL:ILLUMINA" /mnt/d/HumanGenome/
GCF_000001405.40_GRCh38.p14_genomic.fna sample2_trimmed_1.fq.gz sample2_trimmed_2.fq.gz > sample2_aligned.sam

(bioinfo) megatron@Taha:/mnt/d/HumanGenome$ head sample2_aligned.sam

# popst alignment processing

(bioinfo) megatron@Taha:/mnt/d/HumanGenome$ mkdir -p sample2_q2c_results/alignment sample2_q2c_results/depth
(bioinfo) megatron@Taha:/mnt/d/HumanGenome$ samtools sort sample2_aligned.sam -o sample2_q2c_results/alignment/sample2_q2c_sorted.bam
(bioinfo) megatron@Taha:/mnt/d/HumanGenome$ samtools depth sample2_q2c_results/alignment/sample2_q2c_sorted.bam > sample2_q2c_results/depth/
sample2_q2c_depth.txt

# Sort depth file by coverage (column 3) in descending order
# -k3,3nr: Sort numerically by 3rd column in reverse order
# Helps identify regions with highest read coverage

(bioinfo) megatron@Taha:/mnt/d/HumanGenome$ sort -k3,3nr sample2_q2c_results/depth/sample2_q2c_depth.txt > sample2_q2c_results/depth/
sample2_q2c_sorted_depth.txt

# display

(bioinfo) megatron@Taha:/mnt/d/HumanGenome$ head -n 5 sample2_q2c_results/depth/sample2_q2c_sorted_depth.txt


########### END #################
```