

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

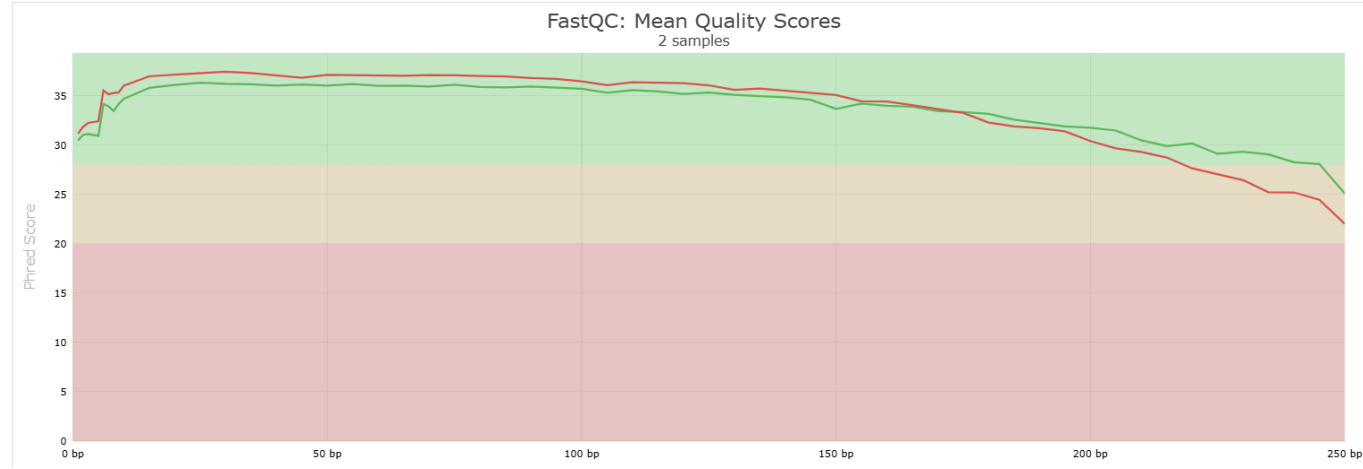
Export as CSV...

Showing 2/2 rows and 6/6 columns.

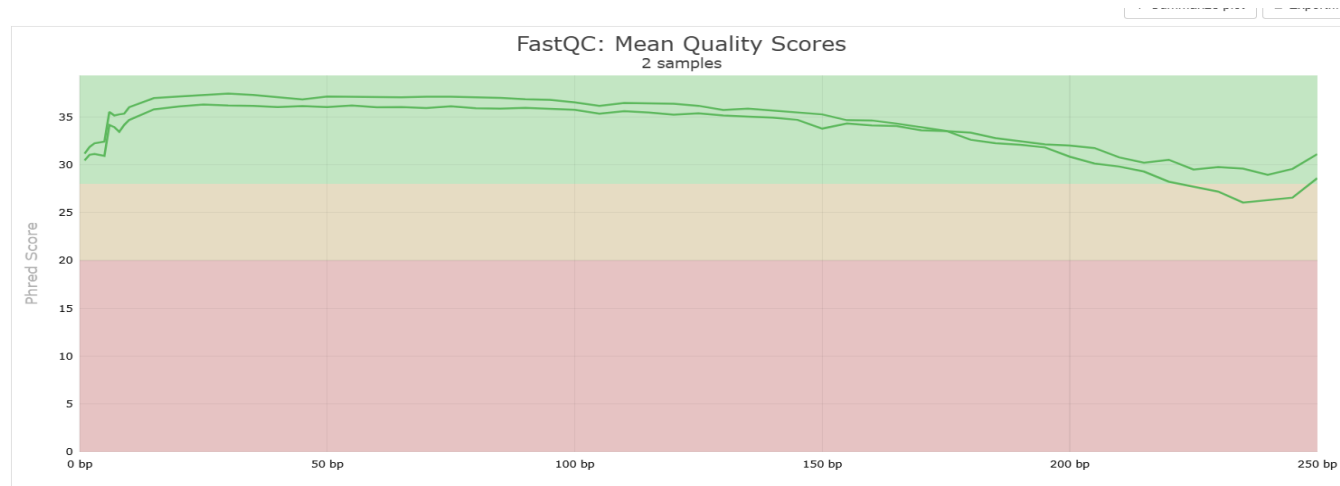
Summarize table

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.

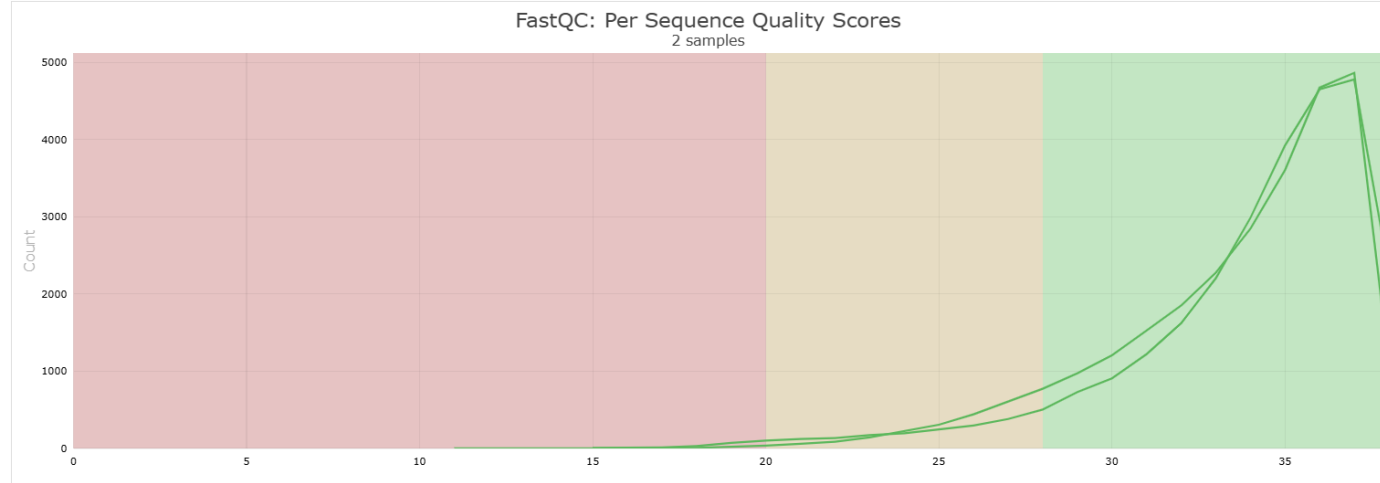


Untrimmed

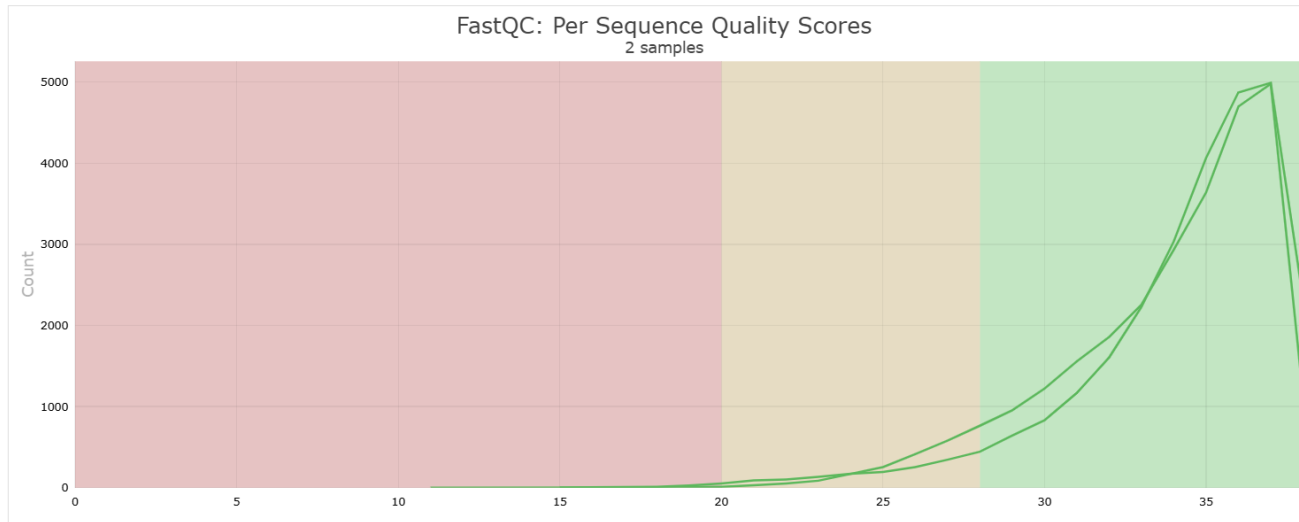


Trimmed

Untrimmed forward/reverse reads show good initial quality ($Q > 30$) until ~ 200 bp, with reverse degrading below Q_{25} at ends—typical for NGS. Post-trimming, reverse's severe drop resolves, though minor decline persists (Q_{26} at 230-240bp). Forward maintains high quality, confirming trimming effectively addressed end artifacts while preserving data utility.

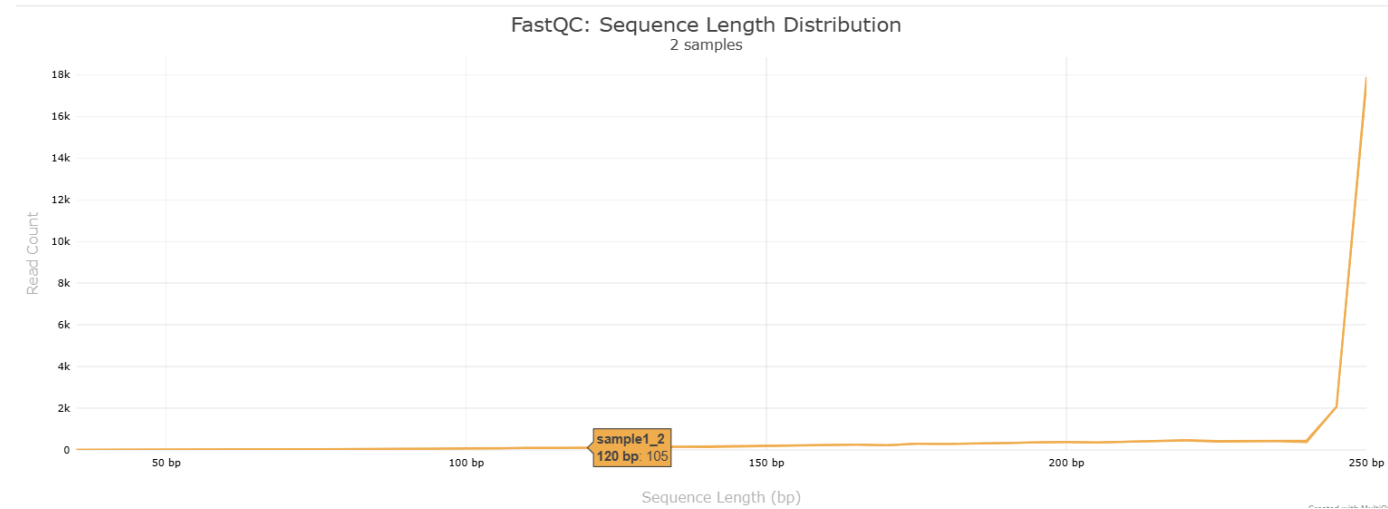


Untrimmed

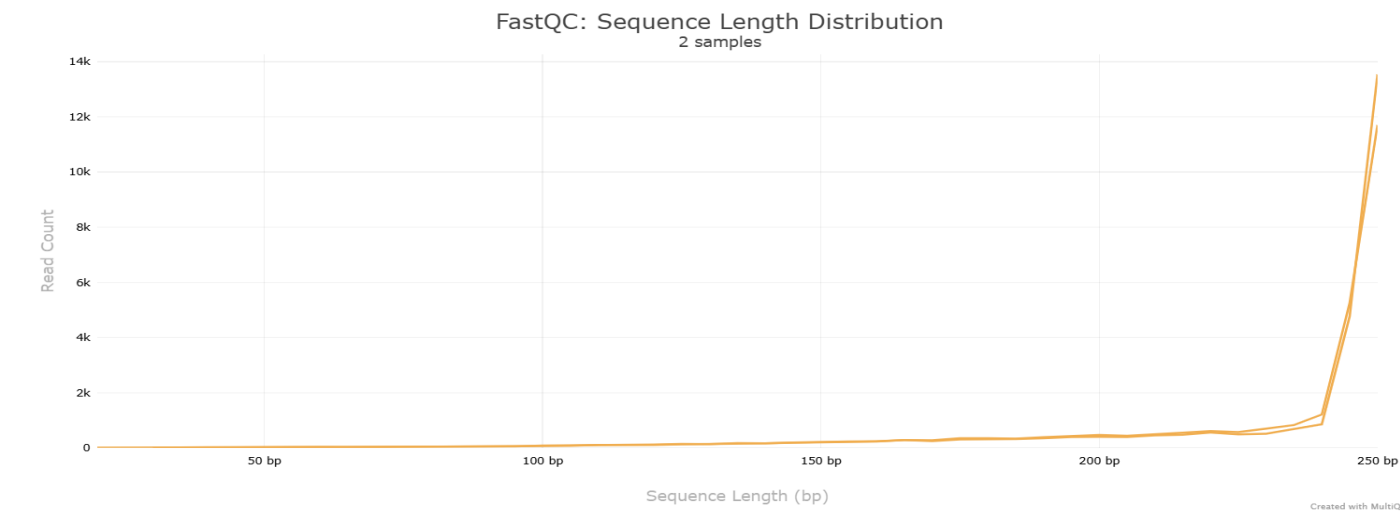


Trimmed

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/adaptor-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

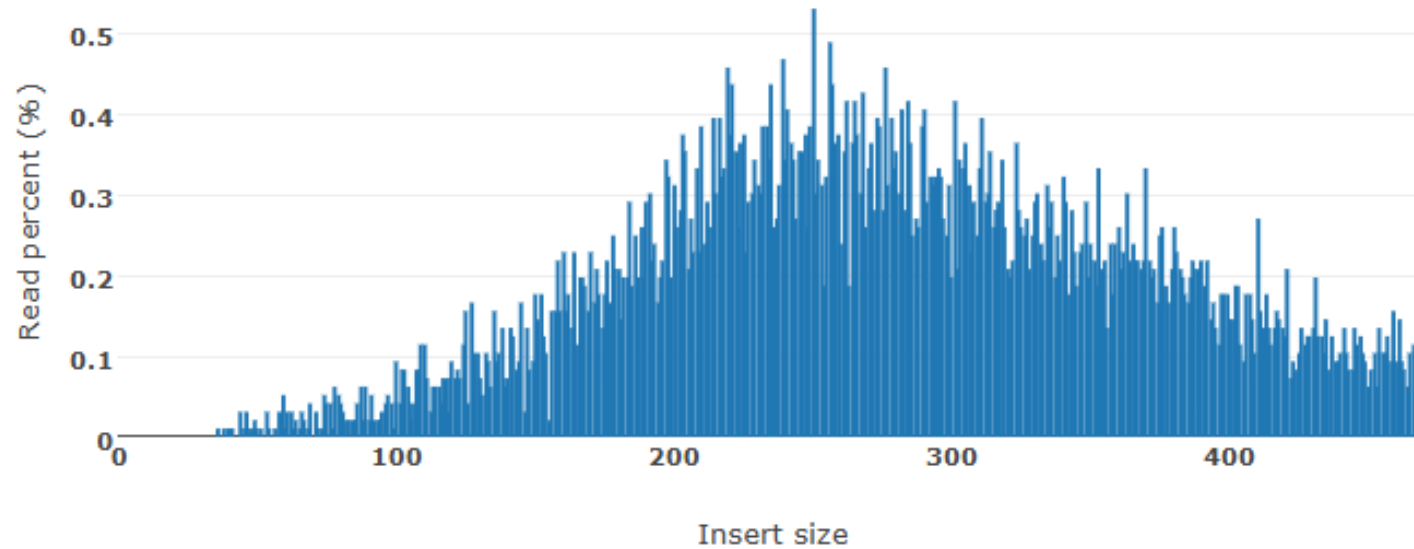
total reads:	51.718000 K
total bases:	11.990111 M
Q20 bases:	10.909680 M (90.988983%)
Q30 bases:	10.382259 M (86.590183%)
GC content:	44.522524%

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.

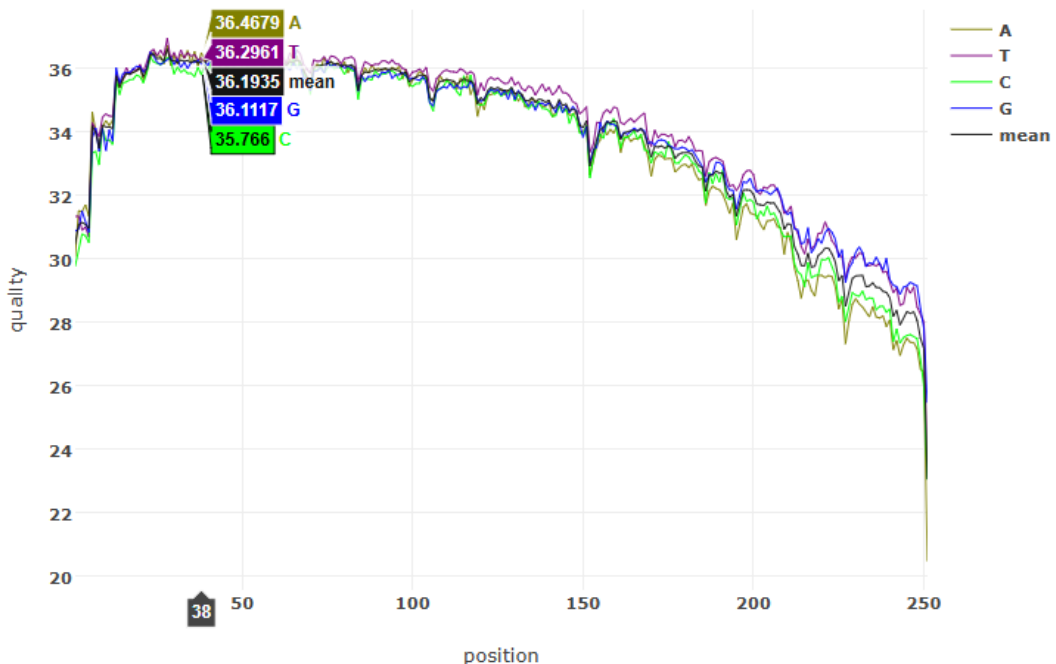
Insert size distribution (18.363523% reads are with unknown length)



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.

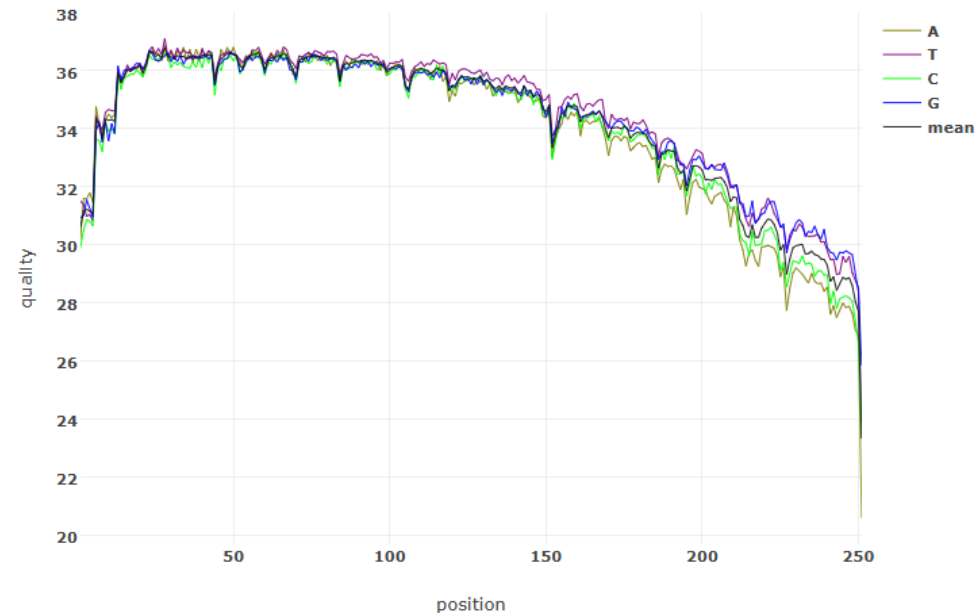
Before filtering: read1: quality

Value of each position will be shown on mouse over.



After filtering: read1: quality

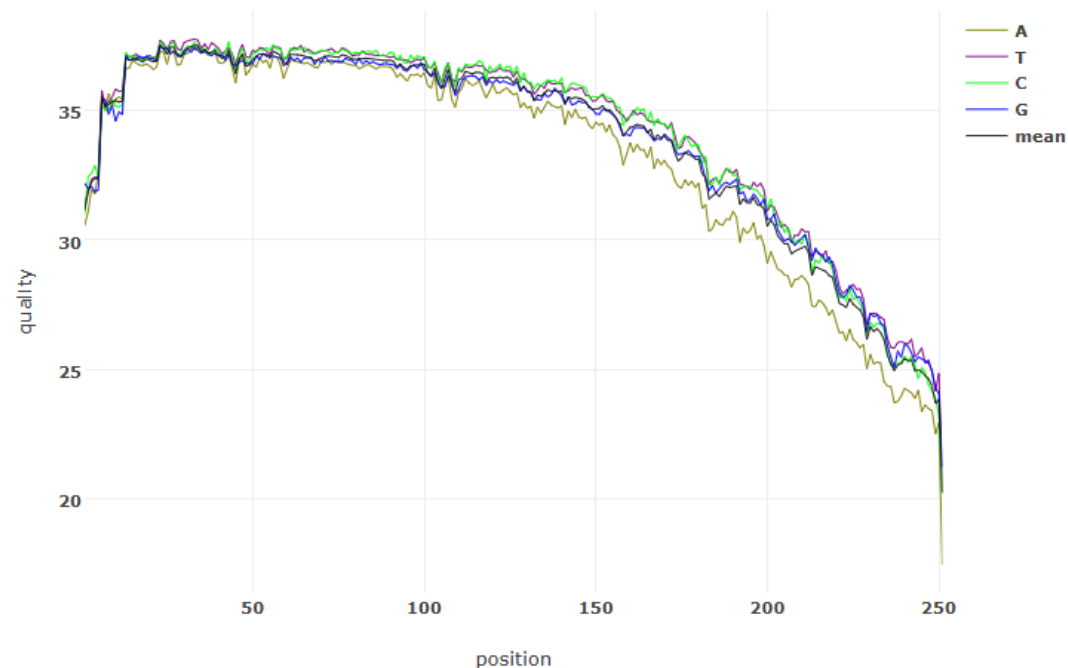
Value of each position will be shown on mouse over.



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.

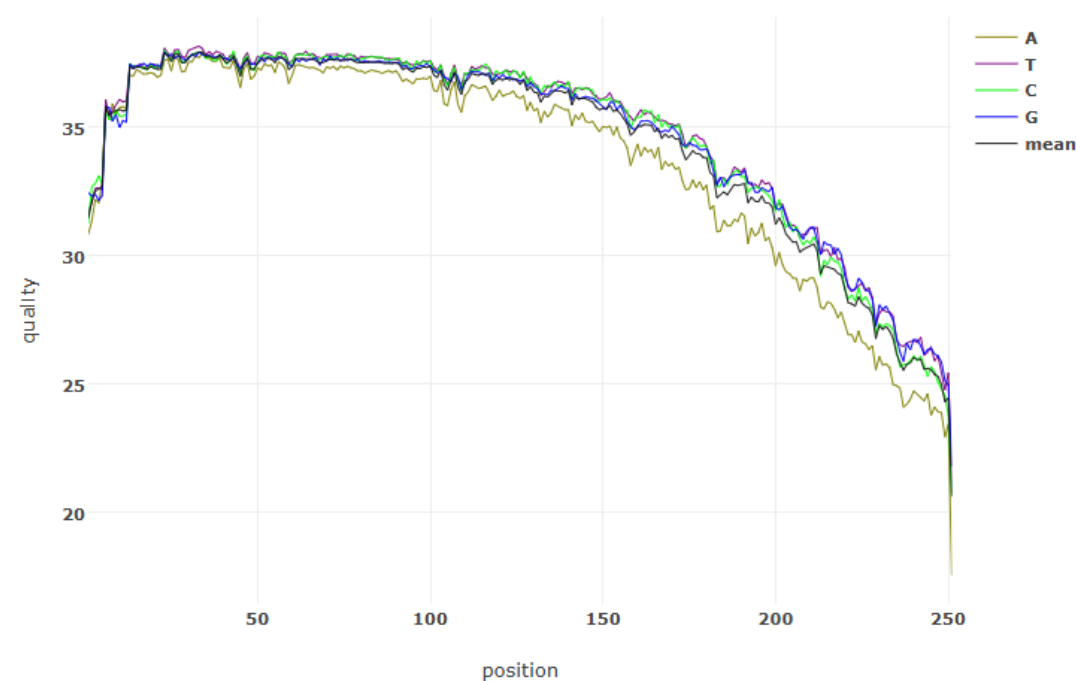
Before filtering: read2: quality

Value of each position will be shown on mouse over.



After filtering: read2: quality

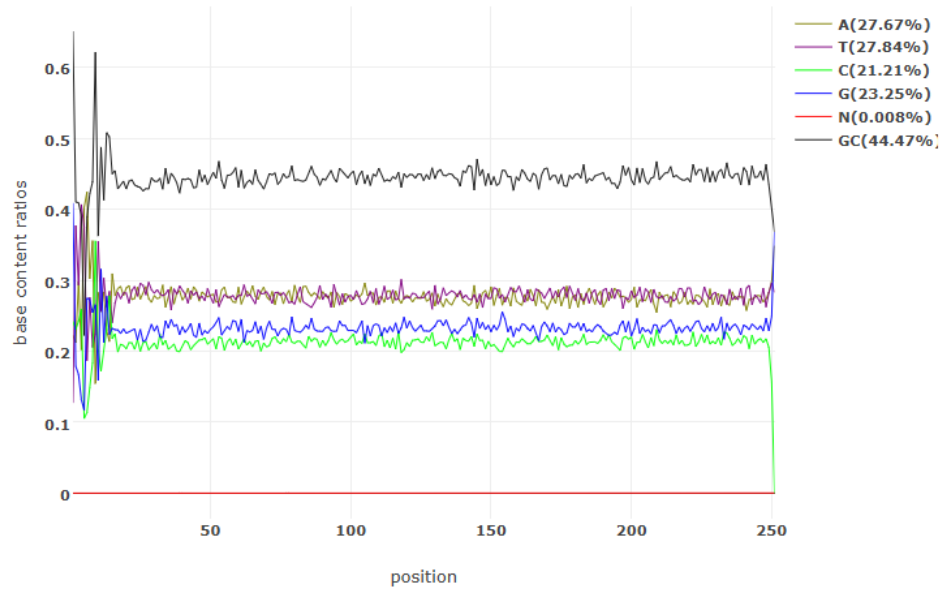
Value of each position will be shown on mouse over.



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.

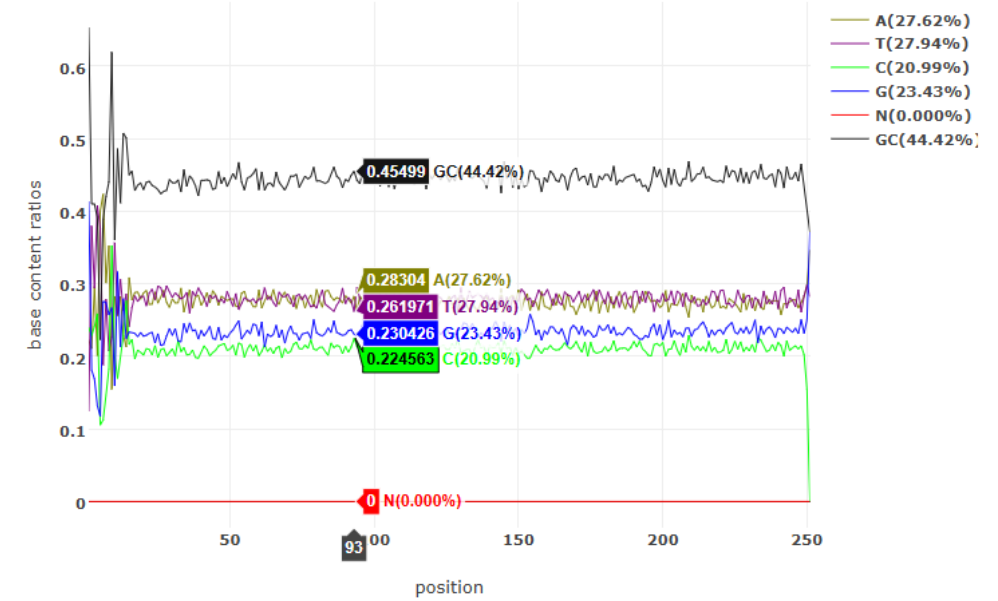
Before filtering: read1: base contents

Value of each position will be shown on mouse over.



After filtering: read1: base contents

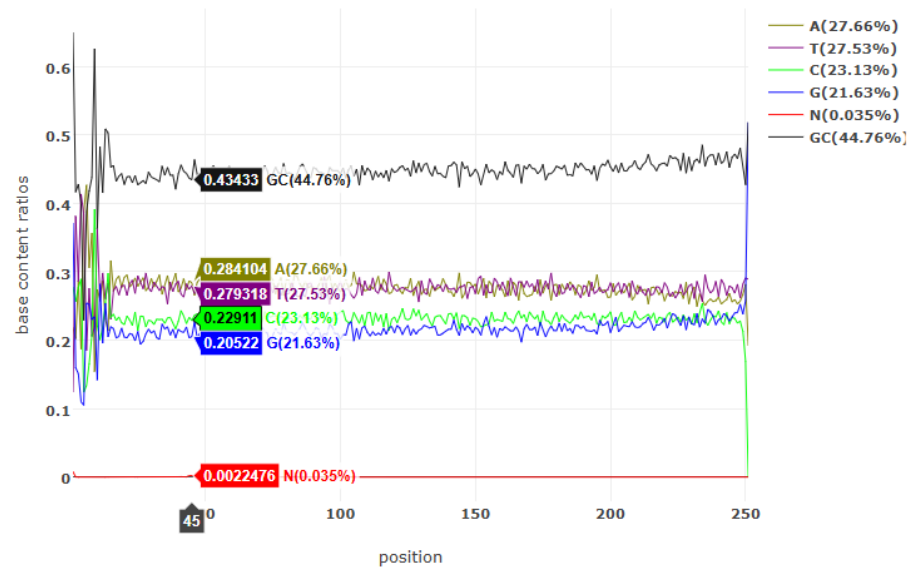
Value of each position will be shown on mouse over.



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

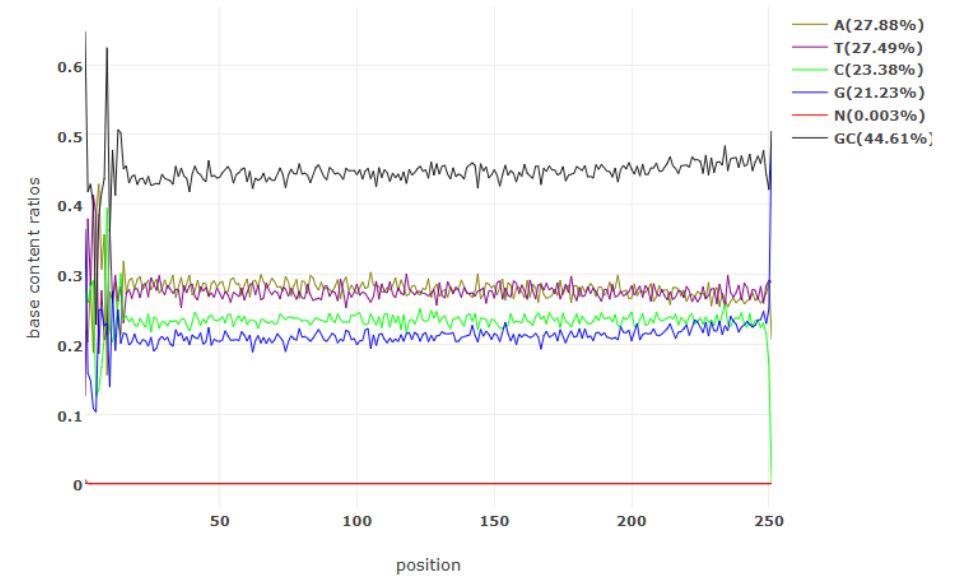
Before filtering: read2: base contents

Value of each position will be shown on mouse over.



After filtering: read2: base contents

Value of each position will be shown on mouse over.



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

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

	AA	AT	AC	AAAG	TA	TT	TC	TG	CA	CT	CC	CG	GA	GT	GC	GG
AAA	AAATA	AAATAT	AAAC	AAAGT	AAATA	AAATT	AAATC	AAAGA	AAACA	AAACT	AAACC	AAAGC	AAAGA	AAATG	AAACG	AAAGG
AAT	AAATA	AAATAT	AAATAC	AAATAG	AAATTA	AAATTT	AAATTC	AAATGC	AAATCA	AAATCT	AAATCC	AAATCG	AAATGA	AAATGT	AAATGC	AAATGG
AAC	AAATA	AAATAT	AAACAT	AAACAC	AAACGA	AAACTA	AAACTC	AAACGC	AAACCA	AAACCT	AAACCC	AAACCG	AAACGA	AAACGT	AAACGC	AAACGG
AGA	AAATA	AAATAT	AAAGT	AAAGC	AAAGTA	AAAGTT	AAAGTC	AAAGGC	AAAGCA	AAAGCT	AAAGCC	AAAGCG	AAAGGA	AAAGGT	AAAGGC	AAAGGG
ATA	AAATA	AAATAT	ATAAC	ATAGC	ATAATA	ATAATT	ATAATC	ATAGAG	ATACGA	ATACAT	ATACCC	ATAGCG	ATAGGA	ATAGGT	ATAGGC	ATAGGG
ATT	AAATA	AAATAT	ATTAC	ATTAG	ATTATA	ATTATT	ATTATC	ATTAGC	ATTACA	ATTACT	ATTACC	ATTAGC	ATTAGA	ATTAGT	ATTAGC	ATTAGG
ACA	AAATA	AAATAT	ACAAC	ACAAG	ACAATA	ACAATT	ACAATC	ACAAGA	ACAACA	ACAACCT	ACAACC	ACAAGC	ACAAGA	ACAAGT	ACAAGC	ACAAGG
ATG	AAATA	AAATAT	ATGAC	ATGAG	ATGATA	ATGATT	ATGATC	ATGAGC	ATGACA	ATGACT	ATGACC	ATGAGC	ATGAGA	ATGAGT	ATGAGC	ATGAGG
ACT	AAATA	AAATAT	ACTAC	ACTAG	ACTATA	ACTATT	ACTATC	ACTAGC	ACTACA	ACTACT	ACTACC	ACTAGC	ACTAGA	ACTAGT	ACTAGC	ACTAGG
ACC	AAATA	AAATAT	ACCAC	ACCAG	ACCATA	ACCATT	ACCATC	ACCAGA	ACCACA	ACCACCT	ACCACC	ACCAGC	ACCAGA	ACCAGT	ACCAGC	ACCAGG
ACG	AAATA	AAATAT	ACGAC	ACGAG	ACGATA	ACGATT	ACGATC	ACGAGC	ACGACA	ACGACT	ACGACC	ACGAGC	ACGAGA	ACGAGT	ACGAGC	ACGAGG
AGA	AAATA	AAATAT	AGAAC	AGAGC	AGATA	AGATT	AGATC	AGAGC	AGACA	AGACT	AGACC	AGAGC	AGAGA	AGAGT	AGAGC	AGAGG
AGT	AAATA	AAATAT	AGTAC	AGTAG	AGTATA	AGTATT	AGTATC	AGTAGC	AGTACA	AGTACT	AGTACC	AGTAGC	AGTAGA	AGTAGT	AGTAGC	AGTAGG
AGC	AAATA	AAATAT	AGCAC	AGCAG	AGCATA	AGCATT	AGCATC	AGCAGC	AGCACA	AGCACT	AGCACC	AGCAGC	AGCAGA	AGCAGT	AGCAGC	AGCAGG
ACA	AAATA	AAATAT	ACAAC	ACAAG	ACAATA	ACAATT	ACAATC	ACAAGA	ACAACA	ACAACCT	ACAACC	ACAAGC	ACAAGA	ACAAGT	ACAAGC	ACAAGG
TAA	TAAATA	TAAATAT	TAAAC	TAAAG	TAAATA	TAAATT	TAAATC	TAAAGA	TAAACA	TAAACT	TAAACC	TAAAGC	TAAAGA	TAAAGT	TAAAGC	TAAAGG
TAT	TAAATA	TAAATAT	TAAATC	TAAATG	TAAATA	TAAATT	TAAATC	TAAATG	TAAATA	TAAATC	TAAATC	TAAATG	TAAATA	TAAATG	TAAATG	TAAATG
TAC	TAAATA	TAAATAT	TAAAC	TAAAG	TAAATA	TAAATT	TAAATC	TAAAGC	TAAACA	TAAACT	TAAACC	TAAAGC	TAAAGA	TAAAGT	TAAAGC	TAAAGG
TAG	TAAATA	TAAATAT	TAAATC	TAAATG	TAAATA	TAAATT	TAAATC	TAAATG	TAAATA	TAAATC	TAAATC	TAAATG	TAAATA	TAAATG	TAAATG	TAAATG
TTA	TAAATA	TAAATAT	TTAAC	TTAGC	TTAATA	TTAATT	TTAATC	TTAGAG	TTACGA	TTACAT	TTACCC	TTAGCG	TTAGGA	TTAGGT	TTAGGC	TTAGGG
TTT	TTAATA	TTAATAT	TTTAC	TTTAG	TTTATA	TTTATT	TTTATC	TTTAGC	TTTACA	TTTACT	TTTACC	TTTAGC	TTTAGA	TTTAGT	TTTAGC	TTTAGG
TTC	TTAATA	TTAATAT	TTTAC	TTTAG	TTTATA	TTTATT	TTTATC	TTTAGC	TTTACA	TTTACT	TTTACC	TTTAGC	TTTAGA	TTTAGT	TTTAGC	TTTAGG
TTG	TTAATA	TTAATAT	TTTAC	TTTAG	TTTATA	TTTATT	TTTATC	TTTAGC	TTTACA	TTTACT	TTTACC	TTTAGC	TTTAGA	TTTAGT	TTTAGC	TTTAGG
TCT	TTAATA	TTAATAT	TCTAC	TCTAG	TCTATA	TCTATT	TCTATC	TCTAGC	TCTACA	TCTACT	TCTACC	TCTAGC	TCTAGA	TCTAGT	TCTAGC	TCTAGG
TCT	TTAATA	TTAATAT	TCTAC	TCTAG	TCTATA	TCTATT	TCTATC	TCTAGC	TCTACA	TCTACT	TCTACC	TCTAGC	TCTAGA	TCTAGT	TCTAGC	TCTAGG
TCC	TTAATA	TTAATAT	TCCAC	TCCAG	TCCATA	TCCATT	TCCATC	TCCAGA	TCCACA	TCCACCT	TCCACC	TCCAGC	TCCAGA	TCCAGT	TCCAGC	TCCAGG
TCC	TTAATA	TTAATAT	TCCAC	TCCAG	TCCATA	TCCATT	TCCATC	TCCAGA	TCCACA	TCCACCT	TCCACC	TCCAGC	TCCAGA	TCCAGT	TCCAGC	TCCAGG
TGG	TTAATA	TTAATAT	TGGAC	TGGAG	TGGATA	TGGATT	TGGATC	TGGAGC	TGGACA	TGGACT	TGGACC	TGGAGC	TGGAGA	TGGAGT	TGGAGC	TGGAGG
TGT	TTAATA	TTAATAT	TGTAC	TGTAG	TGTATA	TGTATT	TGTATC	TGTAGC	TGTACA	TGTACT	TGTACC	TGTAGC	TGTAGA	TGTAGT	TGTAGC	TGTAGG
TGC	TTAATA	TTAATAT	TGCAC	TGCAG	TGCATA	TGCATT	TGCATC	TGCAGA	TGCACA	TGCACCT	TGCACC	TGCAGC	TGCAGA	TGCAGT	TGCAGC	TGCAGG
TGG	TTAATA	TTAATAT	TGGAC	TGGAG	TGGATA	TGGATT	TGGATC	TGGAGC	TGGACA	TGGACT	TGGACC	TGGAGC	TGGAGA	TGGAGT	TGGAGC	TGGAGG
CAA	CAATA	CAATAT	CAAC	CAAGT	CAATA	CAATT	CAATC	CAAGA	CAACA	CAACT	CAACC	CAAGC	CAAGA	CAAGT	CAAGC	CAAGG
CAT	CAATA	CAATAT	CAATAC	CAATAG	CAATTA	CAATTT	CAATTC	CAATGC	CAATCA	CAATCT	CAATCC	CAATCG	CAATGA	CAATGT	CAATGC	CAATGG
CAC	CAATA	CAATAT	CAACAT	CAACAC	CAACGA	CAACTA	CAACTC	CAACGC	CAACCA	CAACCT	CAACCC	CAACCG	CAACGA	CAACGT	CAACGC	CAACGG
CAG	CAATA	CAATAT	CAAGT	CAAGC	CAAGTA	CAAGTT	CAAGTC	CAAGGC	CAAGCA	CAAGCT	CAAGCC	CAAGCG	CAAGGA	CAAGGT	CAAGGC	CAAGGG
CTA	CTATA	CTATAT	CTAAC	CTAGC	CTATA	CTATT	CTATC	CTAGAG	CTACGA	CTACAT	CTACCC	CTAGCG	CTAGGA	CTAGGT	CTAGGC	CTAGGG
CTT	CTATA	CTATAT	CTTAC	CTTAG	CTTATA	CTTTT	CTTTTC	CTTAGC	CTTACA	CTTACT	CTTACC	CTTAGC	CTTAGA	CTTAGT	CTTAGC	CTTAGG
CTC	CTATA	CTATAT	CTCAC	CTCAG	CTCATA	CTCATT	CTCATC	CTCAGC	CTCACA	CTCACT	CTCACC	CTCAGC	CTCAGA	CTCAGT	CTCAGC	CTCAGG
CTG	CTATA	CTATAT	CTGAC	CTGAG	CTGATA	CTGATT	CTGATC	CTGAGC	CTGACA	CTGACT	CTGACC	CTGAGC	CTGAGA	CTGAGT	CTGAGC	CTGAGG
CCA	CCATA	CCATAT	CCAAC	CCAGC	CCATA	CCATT	CCATC	CCAGAG	CCACGA	CCACAT	CCACCC	CCAGCG	CCAGGA	CCAGGT	CCAGGC	CCAGGG
CCT	CCATA	CCATAT	CCTAC	CCTAG	CCTATA	CCTTT	CCTTTC	CCTAGC	CCTACA	CCTACT	CCTACC	CCTAGC	CCTAGA	CCTAGT	CCTAGC	CCTAGG
CCG	CCATA	CCATAT	CCGAC	CCGAG	CCGATA	CCGATT	CCGATC	CCGAGC	CCGACA	CCGACT	CCGACC	CCGAGC	CCGAGA	CCGAGT	CCGAGC	CCGAGG
CCG	CCATA	CCATAT	CCGAC	CCGAG	CCGATA	CCGATT	CCGATC	CCGAGC	CCGACA	CCGACT	CCGACC	CCGAGC	CCGAGA	CCGAGT	CCGAGC	CCGAGG
CCA	CCATA	CCATAT	CCAAC	CCAGC	CCATA	CCATT	CCATC	CCAGAG	CCACGA	CCACAT	CCACCC	CCAGCG	CCAGGA	CCAGGT	CCAGGC	CCAGGG
CCT	CCATA	CCATAT	CCTAC	CCTAG	CCTATA	CCTTT	CCTTTC	CCTAGC	CCTACA	CCTACT	CCTACC	CCTAGC	CCTAGA	CCTAGT	CCTAGC	CCTAGG
CCT	CCATA	CCATAT	CCTAC	CCTAG	CCTATA	CCTTT	CCTTTC	CCTAGC	CCTACA	CCTACT	CCTACC	CCTAGC	CCTAGA	CCTAGT	CCTAGC	CCTAGG
CCT	CCATA	CCATAT	CCGAC	CCGAG	CCGATA	CCGATT	CCGATC	CCGAGC	CCGACA	CCGACT	CCGACC	CCGAGC	CCGAGA	CCGAGT	CCGAGC	CCGAGG
CCT	CCATA	CCATAT	CCGAC	CCGAG	CCGATA	CCGATT	CCGATC	CCGAGC	CCGACA	CCGACT	CCGACC	CCGAGC	CCGAGA	CCGAGT	CCGAGC	CCGAGG
CCT	CCATA	CCATAT	CCGAC	CCGAG	CCGATA	CCGATT	CCGATC	CCGAGC	CCGACA	CCGACT	CCGACC	CCGAGC	CCGAGA	CCGAGT	CCGAGC	CCGAGG
CCT	CCATA	CCATAT	CCGAC	CCGAG	CCGATA	CCGATT	CCGATC	CCGAGC	CCGACA	CCGACT	CCGACC	CCGAGC	CCGAGA	CCGAGT	CCGAGC	CCGAGG
CCT	CCATA	CCATAT	CCGAC	CCGAG	CCGATA	CCGATT	CCGATC	CCGAGC	CCGACA	CCGACT	CCGACC	CCGAGC	CCGAGA	CCGAGT	CCGAGC	CCGAGG
CCT	CCATA	CCATAT	CCGAC	CCGAG	CCGATA	CCGATT	CCGATC	CCGAGC	CCGACA	CCGACT	CCGACC	CCGAGC	CCGAGA	CCGAGT	CCGAGC	CCGAGG
CCT	CCATA	CCATAT	CCGAC	CCGAG	CCGATA	CCGATT	CCGATC	CCGAGC	CCGACA	CCGACT	CCGACC	CCGAGC	CCGAGA	CCGAGT	CCGAGC	CCGAGG
CCT	CCATA	CCATAT	CCGAC	CCGAG	CCGATA	CCGATT	CCGATC	CCGAGC	CCGACA	CCGACT	CCGACC	CCGAGC	CCGAGA	CCGAGT	CCGAGC	CCGAGG
CCT	CCATA	CCATAT	CCGAC	CCGAG	CCGATA	CCGATT	CCGATC	CCGAGC	CCGACA	CCGACT	CCGACC	CCGAGC	CCGAGA	CCGAGT	CCGAGC	CCGAGG
CCT	CCATA	CCATAT	CCGAC	CCGAG	CCGATA	CCGATT	CCGATC	CCGAGC	CCGACA	CCGACT	CCGACC	CCGAGC	CCGAGA	CCGAGT	CCGAGC	CCGAGG
CCT	CCATA	CCATAT	CCGAC	CCGAG	CCGATA	CCGATT	CCGATC	CCGAGC	CCGACA	CCGACT	CCGACC	CCGAGC	CCGAGA	CCGAGT	CCGAGC	CCGAGG
CCT	CCATA	CCATAT	CCGAC	CCGAG	CCGATA	CCGATT	CCGATC	CCGAGC	CCGACA	CCGACT	CCGACC	CCGAGC	CCGAGA	CCGAGT	CCGAGC	CCGAGG
CCT	CCATA	CCATAT	CCGAC	CCGAG	CCGATA	CCGATT	CCGATC	CCGAGC	CCGACA	CCGACT	CCGACC	CCGAGC	CCGAGA	CCGAGT	CCGAGC	CCGAGG
CCT	CCATA	CCATAT	CCGAC	CCGAG	CCGATA	CCGATT	CCGATC	CCGAGC	CCGACA	CCGACT	CCGACC	CCGAGC	CCGAGA	CCGAGT	CCGAGC	CCGAGG
CCT	CCATA	CCATAT	CCGAC	CCGAG	CCGATA	CCGATT	CCGATC	CCGAGC	CCGACA	CCGACT	CCGACC	CCGAGC	CCGAGA	CCGAGT	CCGAGC	CCGAGG
CCT	CCATA	CCATAT	CCGAC	CCGAG	CCGATA	CCGATT	CCGATC	CCGAGC	CCGACA	CCGACT	CCGACC	CCGAGC	CCGAGA	CCGAGT	CCGAGC	CCGAGG
CCT	CCATA	CCATAT	CCGAC	CCGAG	CCGATA	CCGATT	CCGATC	CCGAGC	CCGACA	CCGACT	CCGACC	CCGAGC	CCGAGA	CCGAGT	CCGAGC	CCGAGG
CCT	CCATA	CCATAT	CCGAC	CCGAG	CCGATA	CCGATT	CCGATC	CCGAGC	CCGACA	CCGACT	CCGACC	CCGAGC	CCGAGA	CCGAGT	CCGAGC	CCGAGG
CCT	CCATA	CCATAT	CCGAC	CCGAG	CCGATA	CCGATT	CCGATC	CCGAGC	CCGACA	CCGACT	CCGACC	CCGAGC	CCGAGA	CCGAGT	CCGAGC	CCGAGG
CCT	CCATA	CCATAT	CCGAC	CCGAG	CCGATA	CCGATT	CCGATC	CCGAGC	CCGACA	CCGACT	CCGACC	CCGAGC	CCGAGA	CCGAGT	CCGAGC	CCGAGG
CCT	CCATA	CCATAT	CCGAC	CCGAG	CCGATA	CCGATT	CCGATC	CCGAGC	CCGACA	CCGACT	CCGACC	CCGAGC	CCGAGA	CCGAGT	CCGAGC	CCGAGG
CCT	CCATA	CCATAT	CCGAC	CCGAG	CCGATA	CCGATT	CCGATC	CCGAGC	CCGACA	CCGACT	CCGACC	CCGAGC	CCGAGA	CCGAGT	CCGAGC	CCGAGG
CCT	CCATA	CCATAT	CCGAC	CCGAG	CCGATA	CCGATT	CCGATC	CCGAGC	CCGACA	CCGACT	CCGACC	CCGAGC	CCGAGA	CCGAGT	CCGAGC	CCGAGG
CCT	CCATA	CCATAT	CCGAC	CCGAG	CCGATA	CCGATT	CCGATC	CCGAGC	CCGACA	CCGACT	CCGACC	CCGAGC	CCGAGA	CCGAGT	CCGAGC	CCGAGG
CCT	CCATA	CCATAT	CCGAC	CCGAG	CCGATA	CCGATT	CCGATC	CCGAGC	CCGACA	CCGACT	CCGACC	CCGAGC	CCGAGA	CCGAGT	CCGAGC	CCGAGG
CCT	CCATA	CCATAT	CCGAC	CCGAG	CCGATA	CCGATT	CCGATC	CCGAGC	CCGACA	CCGACT	CCGACC	CCGAGC	CCGAGA	CCGAGT	CCGAGC	CCGAGG
CCT	CCATA	CCATAT	CCGAC	CCGAG	CCGATA	CCGATT	CCGATC	CCGAGC	CCGACA	CCGACT	CCGACC	CCGAGC	CCGAGA	CCGAGT	CCGAGC	CCGAGG
CCT	CCATA	CCATAT	CCGAC	CCGAG	CCGATA	CCGATT	CCGATC	CCGAGC	CCGACA	CCGACT	CCGACC	CCGAGC	CCGAGA	CCGAGT	CCGAGC	CCGAGG
CCT	CCATA	CCATAT	CCGAC	CCGAG	CCGATA	CCGATT	CCGATC	CCGAGC	CCGACA	CCGACT	CCGACC	CCGAGC	CCGAGA	CCGAGT	CCGAGC	CCGAGG
CCT	CCATA	CCATAT	CCGAC	CCGAG	CCGATA	CCGATT	CCGATC	CCGAGC	CCGACA	CCGACT	CCGACC	CCGAGC	CCGAGA	CCGAGT	CCGAGC	CCGAGG
CCT	CCATA	CCATAT	CCGAC	CCGAG	CCGATA	CCGATT	CCGATC	CCGAGC	CCGACA	CCGACT	CCGACC	CCGAGC	CCGAGA	CCGAGT	CCGAGC	CCGAGG
CCT	CCATA	CCATAT	CCGAC	CCGAG	CCGATA	CCGATT	CCGATC	CCGAGC	CCGACA	CCGACT	CCGACC	CCGAGC	CCGAGA	CCGAGT	CCGAGC	CCGAGG
CCT	CCATA	CCATAT	CCGAC	CCGAG	CCGATA	CCGATT	CCGATC	CCGAGC	CCGACA	CCGACT	CCGACC	CCGAGC	CCGAGA	CCGAGT	CCGAGC	CCGAGG
CCT	CCATA	CCATAT	CCGAC	CCGAG	CCGATA	CCGATT	CCGATC	CCGAGC	CCGACA	CCGACT	CCGACC	CCGAGC	CCGAGA	CCGAGT	CCGAGC	CCGAGG
CCT	CCATA	CCATAT	CCGAC	CCGAG	CCGATA	CCGATT	CCGATC	CCGAGC	CCGACA	CCGACT	CCGACC	CCGAGC	CCGAGA	CCGAGT	CCGAGC	CCGAGG
CCT	CCATA	CCATAT	CCGAC	CCGAG	CCGATA	CCGATT	CCGATC	CCGAGC	CCGACA	CCGACT	CCGACC	CCGAGC	CCGAGA	CCGAGT	CCGAGC	CCGAGG
CCT	CCATA	CCATAT	CCGAC	CCGAG	CCGATA	CCGATT	CCGATC	CCGAGC	CCGACA	CCGACT	CCGACC	CCGAGC	CCGAGA	CCGAGT	CCGAGC	CCGAGG
CCT	CCATA	CCATAT	CCGAC	CCGAG	CCGATA	CCGATT	CCGATC	CCGAGC	CCGACA	CCGACT	CCGACC	CCGAGC	CCGAGA	CCGAGT	CCGAGC	CCGAGG
CCT	CCATA	CCATAT	CCGAC	CCGAG	CCGATA	CCGATT	CCGATC	CCGAGC	CCGACA	CCGACT	CCGACC	CCGAGC	CCGAGA	CCGAGT	CCGAGC	CCGAGG
CCT	CCATA	CCATAT	CCGAC	CCGAG	CCGATA	CCGATT	CCGATC	CCGAGC	CCGACA	CCGACT	CCGACC	CCGAGC	CCGAGA	CCGAGT	CCGAGC	CCGAGG
CCT	CCATA	CCATAT	CCGAC	CCGAG	CCGATA	CCGATT	CCGATC	CCGAGC	CCGACA	CCGACT	CCGACC	CCGAGC	CCGAGA	CCGAGT	CCGAGC	CCGAGG
CCT	CCATA	CCATAT	CCGAC	CCGAG	CCGATA	CCGATT	CCGATC	CCGAGC	CCGACA	CCGACT	CCGACC	CCGAGC	CCGAGA	CCGAGT	CCGAGC	CCGAGG
CCT	CCATA	CCATAT	CCGAC	CCGAG	CCGATA	CCGATT	CCGATC									

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

AAA	AA	AT	AC	AG	TA	TT	TC	TG	CA	CT	CC	CG	GA	GT	GC	GG
AAT	AAA	AAAT	AAAC	AAAG	AAATA	AAATT	AAATC	AAATG	AAACA	AAACT	AAACC	AAACG	AAAGA	AAAGT	AAAGC	AAAGG
AAT	AAATA	AAATAT	AAATAC	AAATAG	AAATTA	AAATT	AAATTC	AAATGT	AAATCA	AAATCT	AAATCC	AAATCG	AAATGA	AAATGT	AAATGC	AAATGG
AAC	AAACA	AAACAT	AAACAC	AAACAG	AAACTA	AAACCT	AAACCTC	AAACCTG	AAACCA	AAACCT	AAACCC	AAACCG	AAACGA	AAACGT	AAACGC	AAACGG
AAG	AAAGA	AAAGAT	AAAGAC	AAAGAG	AAAGTA	AAAGTT	AAAGTC	AAAGTG	AAAGCA	AAAGCT	AAAGCC	AAAGCG	AAAGGA	AAAGGT	AAAGGC	AAAGGG
ATA	AAATA	AAATAT	AAATAC	AAATAG	AAATTA	AAATT	AAATTC	AAATGT	AAATCA	AAATCT	AAATCC	AAATCG	AAATGA	AAATGT	AAATGC	AAATGG
ATA	AAATA	AAATAT	AAATAC	AAATAG	AAATTA	AAATT	AAATTC	AAATGT	AAATCA	AAATCT	AAATCC	AAATCG	AAATGA	AAATGT	AAATGC	AAATGG
ATT	AAATA	AAATAT	AAATAC	AAATAG	AAATTA	AAATT	AAATTC	AAATGT	AAATCA	AAATCT	AAATCC	AAATCG	AAATGA	AAATGT	AAATGC	AAATGG
ATC	AAACA	AAACAT	AAACAC	AAACAG	AAACTA	AAACCT	AAACCTC	AAACCTG	AAACCA	AAACCT	AAACCC	AAACCG	AAACGA	AAACGT	AAACGC	AAACGG
ATC	AAACA	AAACAT	AAACAC	AAACAG	AAACTA	AAACCT	AAACCTC	AAACCTG	AAACCA	AAACCT	AAACCC	AAACCG	AAACGA	AAACGT	AAACGC	AAACGG
ACT	AAACA	AAACAT	AAACAC	AAACAG	AAACTA	AAACCT	AAACCTC	AAACCTG	AAACCA	AAACCT	AAACCC	AAACCG	AAACGA	AAACGT	AAACGC	AAACGG
ACC	AAAGA	AAAGAT	AAAGAC	AAAGAG	AAAGTA	AAAGTT	AAAGTC	AAAGTG	AAAGCA	AAAGCT	AAAGCC	AAAGCG	AAAGGA	AAAGGT	AAAGGC	AAAGGG
ACC	AAAGA	AAAGAT	AAAGAC	AAAGAG	AAAGTA	AAAGTT	AAAGTC	AAAGTG	AAAGCA	AAAGCT	AAAGCC	AAAGCG	AAAGGA	AAAGGT	AAAGGC	AAAGGG
AGC	AAAGA	AAAGAT	AAAGAC	AAAGAG	AAAGTA	AAAGTT	AAAGTC	AAAGTG	AAAGCA	AAAGCT	AAAGCC	AAAGCG	AAAGGA	AAAGGT	AAAGGC	AAAGGG
AGC	AAAGA	AAAGAT	AAAGAC	AAAGAG	AAAGTA	AAAGTT	AAAGTC	AAAGTG	AAAGCA	AAAGCT	AAAGCC	AAAGCG	AAAGGA	AAAGGT	AAAGGC	AAAGGG
AGG	AAAGA	AAAGAT	AAAGAC	AAAGAG	AAAGTA	AAAGTT	AAAGTC	AAAGTG	AAAGCA	AAAGCT	AAAGCC	AAAGCG	AAAGGA	AAAGGT	AAAGGC	AAAGGG
AGA	AAAGA	AAAGAT	AAAGAC	AAAGAG	AAAGTA	AAAGTT	AAAGTC	AAAGTG	AAAGCA	AAAGCT	AAAGCC	AAAGCG	AAAGGA	AAAGGT	AAAGGC	AAAGGG
TAA	TAATA	TAAAT	TAAAC	TAAAG	TAATA	TAAAT	TAAATC	TAAATG	TAAACA	TAAACT	TAAACC	TAAACG	TAAAGA	TAAAGT	TAAAGC	TAAAGG
TAT	TAATA	TAAAT	TAAAC	TAAAG	TAATA	TAAAT	TAAATC	TAAATG	TAAACA	TAAACT	TAAACC	TAAACG	TAAAGA	TAAAGT	TAAAGC	TAAAGG
TAT	TAATA	TAAAT	TAAAC	TAAAG	TAATA	TAAAT	TAAATC	TAAATG	TAAACA	TAAACT	TAAACC	TAAACG	TAAAGA	TAAAGT	TAAAGC	TAAAGG
TAA	TAATA	TAAAT	TAAAC	TAAAG	TAATA	TAAAT	TAAATC	TAAATG	TAAACA	TAAACT	TAAACC	TAAACG	TAAAGA	TAAAGT	TAAAGC	TAAAGG
TTA	TTATA	TTAAT	TTAAC	TTAAG	TTATA	TTAAT	TTAATC	TTAATG	TTTACA	TTTACT	TTTACC	TTTACG	TTTAGA	TTTAGT	TTTAGC	TTTAGG
TTA	TTATA	TTAAT	TTAAC	TTAAG	TTATA	TTAAT	TTAATC	TTAATG	TTTACA	TTTACT	TTTACC	TTTACG	TTTAGA	TTTAGT	TTTAGC	TTTAGG
TTT	TTATA	TTAAT	TTAAC	TTAAG	TTATA	TTAAT	TTAATC	TTAATG	TTTACA	TTTACT	TTTACC	TTTACG	TTTAGA	TTTAGT	TTTAGC	TTTAGG
TTT	TTATA	TTAAT	TTAAC	TTAAG	TTATA	TTAAT	TTAATC	TTAATG	TTTACA	TTTACT	TTTACC	TTTACG	TTTAGA	TTTAGT	TTTAGC	TTTAGG
TTG	TTACA	TTACAT	TTACAC	TTACAG	TTACTA	TTACTT	TTACTC	TTACTG	TTTACA	TTTACT	TTTACC	TTTACG	TTTAGA	TTTAGT	TTTAGC	TTTAGG
TTG	TTACA	TTACAT	TTACAC	TTACAG	TTACTA	TTACTT	TTACTC	TTACTG	TTTACA	TTTACT	TTTACC	TTTACG	TTTAGA	TTTAGT	TTTAGC	TTTAGG
TGA	TTACA	TTACAT	TTACAC	TTACAG	TTACTA	TTACTT	TTACTC	TTACTG	TTTACA	TTTACT	TTTACC	TTTACG	TTTAGA	TTTAGT	TTTAGC	TTTAGG
TGA	TTACA	TTACAT	TTACAC	TTACAG	TTACTA	TTACTT	TTACTC	TTACTG	TTTACA	TTTACT	TTTACC	TTTACG	TTTAGA	TTTAGT	TTTAGC	TTTAGG
TGT	TTACA	TTACAT	TTACAC	TTACAG	TTACTA	TTACTT	TTACTC	TTACTG	TTTACA	TTTACT	TTTACC	TTTACG	TTTAGA	TTTAGT	TTTAGC	TTTAGG
TGC	TCACA	TCACAT	TCACAC	TCACAG	TCACTA	TCACTT	TCACTC	TCACTG	TCGACA	TCGACT	TCGACC	TCGACG	TCGAGA	TCGAGT	TCGAGC	TCGAGG
TGG	TCAGA	TCAGAT	TCAGAC	TCAGAG	TCAGTA	TCAGTT	TCAGTC	TCAGTG	TCAGCA	TCAGCT	TCAGCC	TCAGCG	TCAGGA	TCAGGT	TCAGGC	TCAGGG
CAA	CAATA	CAATAT	CAATAC	CAATAG	CAATTA	CAATT	CAATTC	CAATGT	CAATCA	CAATCT	CAATCC	CAATCG	CAATGA	CAATGT	CAATGC	CAATGG
CAA	CAATA	CAATAT	CAATAC	CAATAG	CAATTA	CAATT	CAATTC	CAATGT	CAATCA	CAATCT	CAATCC	CAATCG	CAATGA	CAATGT	CAATGC	CAATGG
CAT	CAACA	CACAT	CACAC	CACAG	CACATA	CACAT	CACATC	CACATG	CACACA	CACACT	CACACC	CACACG	CACAGA	CACAGT	CACAGC	CACAGG
CAT	CAACA	CACAT	CACAC	CACAG	CACATA	CACAT	CACATC	CACATG	CACACA	CACACT	CACACC	CACACG	CACAGA	CACAGT	CACAGC	CACAGG
CCT	CTATA	CTAAT	CTAAC	CTAAG	CTATA	CTAAT	CTAATC	CTAATG	CTTACA	CTTACT	CTTACC	CTTACG	CTTAGA	CTTAGT	CTTAGC	CTTAGG
CCT	CTATA	CTAAT	CTAAC	CTAAG	CTATA	CTAAT	CTAATC	CTAATG	CTTACA	CTTACT	CTTACC	CTTACG	CTTAGA	CTTAGT	CTTAGC	CTTAGG
CTC	CTACA	CTACAT	CTACAC	CTACAG	CTACTA	CTACTT	CTACTC	CTACTG	CTTACA	CTTACT	CTTACC	CTTACG	CTTAGA	CTTAGT	CTTAGC	CTTAGG
CTC	CTACA	CTACAT	CTACAC	CTACAG	CTACTA	CTACTT	CTACTC	CTACTG	CTTACA	CTTACT	CTTACC	CTTACG	CTTAGA	CTTAGT	CTTAGC	CTTAGG
CTG	CTACA	CTACAT	CTACAC	CTACAG	CTACTA	CTACTT	CTACTC	CTACTG	CTTACA	CTTACT	CTTACC	CTTACG	CTTAGA	CTTAGT	CTTAGC	CTTAGG
CTG	CTACA	CTACAT	CTACAC	CTACAG	CTACTA	CTACTT	CTACTC	CTACTG	CTTACA	CTTACT	CTTACC	CTTACG	CTTAGA	CTTAGT	CTTAGC	CTTAGG
CCC	CCATA	CCATAT	CCATAC	CCATAG	CCATTA	CCATT	CCATTC	CCATGT	CCATCA	CCATCT	CCATCC	CCATCG	CCATGA	CCATGT	CCATGC	CCATGG
CCC	CCATA	CCATAT	CCATAC	CCATAG	CCATTA	CCATT	CCATTC	CCATGT	CCATCA	CCATCT	CCATCC	CCATCG	CCATGA	CCATGT	CCATGC	CCATGG
CCG	CCAGA	CCAGAT	CCAGAC	CCAGAG	CCAGTA	CCAGTT	CCAGTC	CCAGTG	CCAGCA	CCAGCT	CCAGCC	CCAGCG	CCAGGA	CCAGGT	CCAGGC	CCAGGG
CCG	CCAGA	CCAGAT	CCAGAC	CCAGAG	CCAGTA	CCAGTT	CCAGTC	CCAGTG	CCAGCA	CCAGCT	CCAGCC	CCAGCG	CCAGGA	CCAGGT	CCAGGC	CCAGGG
CGA	CGATA	CGATAT	CGATAC	CGATAG	CGATTA	CGATT	CGATTC	CGATGT	CGATCA	CGATCT	CGATCC	CGATCG	CGATGA	CGATGT	CGATGC	CGATGG
CGA	CGATA	CGATAT	CGATAC	CGATAG	CGATTA	CGATT	CGATTC	CGATGT	CGATCA	CGATCT	CGATCC	CGATCG	CGATGA	CGATGT	CGATGC	CGATGG
CGT	CGACA	CGACAT	CGACAC	CGACAG	CGACATA	CGACAT	CGACATC	CGACATG	CGACACA	CGACACT	CGACACC	CGACACG	CGACAGA	CGACAGT	CGACAGC	CGACAGG
CGT	CGACA	CGACAT	CGACAC	CGACAG	CGACATA	CGACAT	CGACATC	CGACATG	CGACACA	CGACACT	CGACACC	CGACACG	CGACAGA	CGACAGT	CGACAGC	CGACAGG
GAA	GAATA	GAATAT	GAATAC	GAATAG	GAATTA	GAATT	GAATTC	GAATGT	GAATCA	GAATCT	GAATCC	GAATCG	GAATGA	GAATGT	GAATGC	GAATGG
GAA	GAATA	GAATAT	GAATAC	GAATAG	GAATTA	GAATT	GAATTC	GAATGT	GAATCA	GAATCT	GAATCC	GAATCG	GAATGA	GAATGT	GAATGC	GAATGG
GAT	GAACA	GACAT	GACAC	GACAG	GACATA	GACAT	GACATC	GACATG	GACACA	GACACT	GACACC	GACACG	GACAGA	GACAGT	GACAGC	GACAGG
GAT	GAACA	GACAT	GACAC	GACAG	GACATA	GACAT	GACATC	GACATG	GACACA	GACACT	GACACC	GACACG	GACAGA	GACAGT	GACAGC	GACAGG
GAC	GAAGA	GAGAT	GAGAC	GAGAG	GAGTA	GAGTT	GAGTC	GAGTG	GAGCA	GAGCT	GAGCC	GAGCG	GAGGA	GAGGT	GAGGC	GAGGG
GAC	GAAGA	GAGAT	GAGAC	GAGAG	GAGTA	GAGTT	GAGTC	GAGTG	GAGCA	GAGCT	GAGCC	GAGCG	GAGGA	GAGGT	GAGGC	GAGGG
GTA	GTATA	GTATAT	GTATAC	GTATAG	GTATTA	GTATT	GTATTC	GTATGT	GTATCA	GTATCT	GTATCC	GTATCG	GTATGA	GTATGT	GTATGC	GTATGG
GTT	GTATA	GTATAT	GTATAC	GTATAG	GTATTA	GTATT	GTATTC	GTATGT	GTATCA	GTATCT	GTATCC	GTATCG	GTATGA	GTATGT	GTATGC	GTATGG
GTT	GTATA	GTATAT	GTATAC	GTATAG	GTATTA	GTATT	GTATTC	GTATGT	GTATCA	GTATCT	GTATCC	GTATCG	GTATGA	GTATGT	GTATGC	GTATGG
GTA	GTACA	GTACAT	GTACAC	GTACAG	GTACTA	GTACTT	GTACTC	GTACTG	GTTACA	GTTACT	GTTACC	GTTACG	GTTAGA	GTTAGT	GTTAGC	GTTAGG
GTA	GTACA	GTACAT	GTACAC	GTACAG	GTACTA	GTACTT	GTACTC	GTACTG	GTTACA	GTTACT	GTTACC	GTTACG	GTTAGA	GTTAGT	GTTAGC	GTTAGG
GCA	GCATA	GCATAT	GCATAC	GCATAG	GCATTA	GCATT	GCATTC	GCATGT	GCATCA	GCATCT	GCATCC	GCATCG	GCATGA	GCATGT	GCATGC	GCATGG
GCA	GCATA	GCATAT	GCATAC	GCATAG	GCATTA	GCATT	GCATTC	GCATGT	GCATCA	GCATCT	GCATCC	GCATCG	GCATGA	GCATGT	GCATGC	GCATGG
GCC	GCAGA	GCAGAT	GCAGAC	GCAGAG	GCAGTA	GCAGTT	GCAGTC	GCAGTG	GCAGCA	GCAGCT	GCAGCC	GCAGCG	GCAGGA	GCAGGT	GCAGGC	GCAGGG
GCC	GCAGA	GCAGAT	GCAGAC	GCAGAG	GCAGTA	GCAGTT	GCAGTC	GCAGTG	GCAGCA	GCAGCT	GCAGCC	GCAGCG	GCAGGA	GCAGGT	GCAGGC	GCAGGG
GCG	CGATA	CGATAT	CGATAC	CGATAG	CGATTA	CGATT	CGATTC	CGATGT	CGATCA	CGATCT	CGATCC	CGATCG	CGATGA	CGATGT	CGATGC	CGATGG
GCG	CGATA	CGATAT	CGATAC	CGATAG	CGATTA	CGATT	CGATTC	CGATGT	CGATCA	CGATCT	CGATCC	CGATCG	CGATGA	CGATGT	CGATGC	CGATGG
GCG	CGATA	CGATAT	CGATAC	CGATAG	CGATTA	CGATT	CGATTC	CGATGT	CGATCA	CGATCT	CGATCC	CGATCG	CGATGA	CGATGT	CGATGC	CGATGG
GCG	CGATA	CGATAT	CGATAC	CGATAG	CGATTA	CGATT	CGATTC	CGATGT	CGATCA	CGATCT	CGATCC	CGATCG	CGATGA	CGATGT	CGATGC	CGATGG
GCG	CGATA	CGATAT	CGATAC	CGATAG	CGATTA	CGATT	CGATTC	CGATGT	CGATCA	CGATCT	CGATCC	CGATCG	CGATGA	CGATGT	CGATGC	CGATGG
GCG	CGATA	CGATAT	CGATAC	CGATAG	CGATTA	CGATT	CGATTC	CGATGT	CGATCA	CGATCT	CGATCC	CGATCG	CGATGA	CGATGT	CGATGC	CGATGG
GCG	CGATA	CGATAT	CGATAC	CGATAG	CGATTA	CGATT	CGATTC	CGATGT	CGATCA	CGATCT	CGATCC	CGATCG	CGATGA	CGATGT	CGATGC	CGATGG
GCG	CGATA	CGATAT	CGATAC	CGATAG	CGATTA	CGATT	CGATTC	CGATGT	CGATCA	CGATCT	CGATCC	CGATCG	CGATGA	CGATGT	CGATGC	CGATGG
GCG	CGATA	CGATAT	CGATAC	CGATAG	CGATTA	CGATT	CGATTC	CGATGT	CGATCA	CGATCT	CGATCC	CGATCG	CGATGA	CGATGT	CGATGC	CGATGG
GCG	CGATA	CGATAT	CGATAC	CGATAG	CGATTA	CGATT	CGATTC	CGATGT	CGATCA	CGATCT	CGATCC	CGATCG	CGATGA	CGATGT	CGATGC	CGATGG
GCG	CGATA	CGATAT	CGATAC	CGATAG	CGATTA	CGATT	CGATTC	CGATGT	CGATCA	CGATCT	CGATCC	CGATCG	CGATGA	CGATGT	CGATGC	CGATGG
GCG	CGATA	CGATAT	CGATAC	CGATAG	CGATTA	CGATT	CGATTC	CGATGT	CGATCA	CGATCT	CGATCC	CGATCG	CGATGA	CGATGT	CGATGC	CGATGG
GCG	CGATA	CGATAT	CGATAC	CGATAG	CGATTA	CGATT	CGATTC	CGATGT	CGATCA	CGATCT	CGATCC	CGATCG	CGATGA	CGATGT	CGATGC	CGATGG
GCG	CGATA	CGATAT	CGATAC	CGATAG	CGATTA	CGATT	CGATTC	CGATGT	CGATCA	CGATCT	CGATCC	CGATCG	CGATGA	CGATGT	CGATGC	CGATGG
GCG	CGATA	CGATAT	CGATAC	CGATAG	CGATTA	CGATT	CGATTC	CGATGT	CGATCA	CGATCT	CGATCC	CGATCG	CGATGA	CGATGT	CGATGC	CGATGG
GCG	CGATA	CGATAT	CGATAC	CGATAG	CGATTA	CGATT	CGATTC	CGATGT	CGATCA	CGATCT	CGATCC	CGATCG	CGATGA	CGATGT	CGATGC	CGATGG
GCG	CGATA	CGATAT	CGATAC	CGATAG	CGATTA	CGATT	CGATTC	CGATGT	CGATCA	CGATCT	CGATCC	CGATCG	CGATGA	CGATGT	CGATGC	CGATGG
GCG	CGATA	CGATAT	CGATAC	CGATAG	CGATTA	CGATT	CGATTC	CGATGT	CGATCA	CGATCT	CGATCC	CGATCG	CGATGA	CGATGT	CGATGC	CGATGG
GCG	CGATA	CGATAT	CGATAC	CGATAG	CGATTA	CGATT	CGATTC	CGATGT	CGATCA	CGATCT	CGATCC	CGATCG	CGATGA	CGATGT	CGATGC	CGATGG
GCG	CGATA	CGATAT	CGATAC	CGATAG	CGATTA	CGATT	CGATTC	CGATGT	CGATCA	CGATCT	CGATCC	CGATCG	CGATGA	CGATGT	CGATGC	CGATGG
GCG	CGATA	CGATAT	CGATAC	CGATAG	CGATTA	CGATT	CGATTC	CGATGT	CGATCA	CGATCT	CGATCC	CGATCG	CGATGA	CGATGT	CGATGC	CGATGG
GCG	CGATA	CGATAT	CGATAC	CGATAG	CGATTA	CGATT	CGATTC	CGATGT	CGATCA	CGATCT	CGATCC	CGATCG	CGATGA	CGATGT	CGATGC	CGATGG
GCG	CGATA	CGATAT	CGATAC	CGATAG	CGATTA	CGATT	CGATTC	CGATGT	CGATCA	CGATCT	CGATCC	CGATCG	CGATGA	CGATGT	CGATGC	CGATGG
GCG	CGATA	CGATAT	CGATAC	CGATAG	CGATTA	CGATT	CGATTC	CGATGT	CGATCA	CGATCT	CGATCC	CGATCG	CGATGA	CGATGT	CGATGC	CGATGG
GCG	CGATA	CGATAT	CGATAC	CGATAG	CGATTA	CGATT	CGATTC	CGATGT	CGATCA	CGATCT	CGATCC	CGATCG	CGATGA	CGATGT	CGATGC	CGATGG
GCG	CGATA	CGATAT	CGATAC	CGATAG	CGATTA	CGATT	CGATTC	CGATGT	CGATCA	CGATCT	CGATCC	CGATCG	CGATGA	CGATGT	CGATGC	CGATGG
GCG	CGATA	CGATAT	CGATAC	CGATAG												

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.

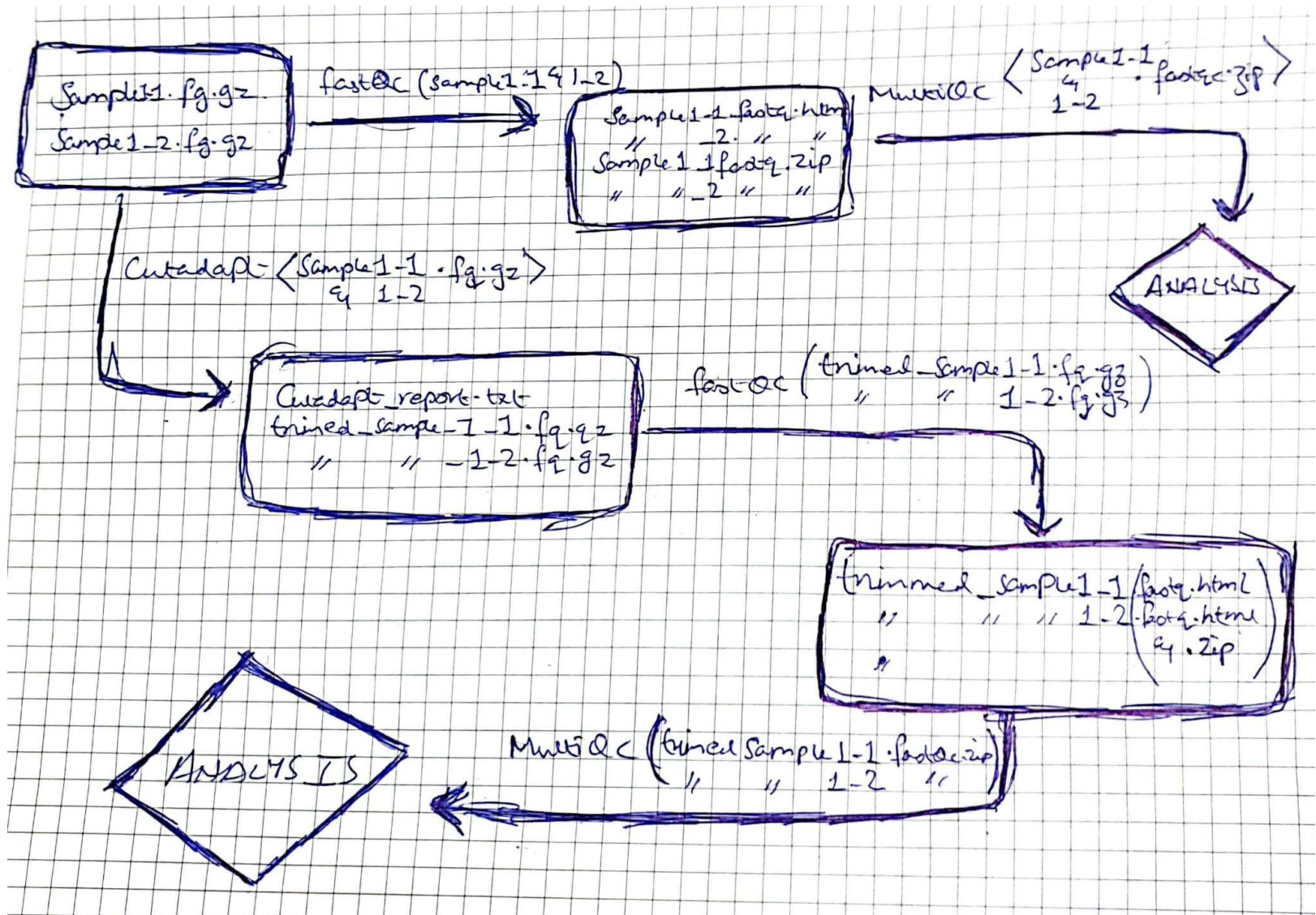
Darker background means larger counts. The count will be shown on mouse over.

AAA	AA	AT	AC	AA	TA	TT	TC	TC	CA	CT	CC	CG	CA	CT	CC	CG
AAA	AAAAT	AAATAT	AAATAC	AAATAG	AAATTA	AAATTT	AAATTC	AAATC	AAATCA	AAATCT	AAATCC	AAATCG	AAATCA	AAATCT	AAATCC	AAATCG
AAT	AAMTAA	AATATAT	AATATAC	AATATAG	AATATTA	AATATTT	AATATTC	AATATC	AATATCA	AATATCT	AATATCC	AATATCG	AATATCA	AATATCT	AATATCC	AATATCG
AAC	AAACAA	AAACAT	AAACAT	AAACAC	AAACAT	AAACCT	AAACCT	AAACCT	AAACCA	AAACCT	AAACCC	AAACCG	AAACCA	AAACCT	AAACCC	AAACCG
AAG	AAAGAA	AAAGAT	AAAGAC	AAAGAG	AAAGTA	AAAGTT	AAAGTC	AAAGTC	AAAGCA	AAAGCT	AAAGCC	AAAGCG	AAAGCA	AAAGCT	AAAGCC	AAAGCG
ATA	ATAATA	ATAATAT	ATAATAC	ATAATAG	ATAATTA	ATAATTT	ATAATTC	ATAATC	ATAATCA	ATAATCT	ATAATCC	ATAATCG	ATAATCA	ATAATCT	ATAATCC	ATAATCG
ATA	ATAATA	ATAATAT	ATAATAC	ATAATAG	ATAATTA	ATAATTT	ATAATTC	ATAATC	ATAATCA	ATAATCT	ATAATCC	ATAATCG	ATAATCA	ATAATCT	ATAATCC	ATAATCG
ATC	ATCAAA	ATCATAT	ATCATAC	ATCATAG	ATCATTA	ATCATTT	ATCATTC	ATCATC	ATCATCA	ATCATCT	ATCATCC	ATCATCG	ATCATCA	ATCATCT	ATCATCC	ATCATCG
ATG	ATGAAA	ATGATAT	ATGATAC	ATGATAG	ATGATTA	ATGATTT	ATGATTC	ATGATC	ATGATCA	ATGATCT	ATGATCC	ATGATCG	ATGATCA	ATGATCT	ATGATCC	ATGATCG
ACA	ACAAAA	ACAAAT	ACAAAC	ACAAAG	ACAAAT	ACAACT	ACAACT	ACAACT	ACAACA	ACAACT	ACAACC	ACAACG	ACAACA	ACAACT	ACAACC	ACAACG
ACT	ACTAAA	ACTATAT	ACTATAC	ACTATAG	ACTATTA	ACTATTT	ACTATTC	ACTATC	ACTATCA	ACTATCT	ACTATCC	ACTATCG	ACTATCA	ACTATCT	ACTATCC	ACTATCG
ACC	ACCAAA	ACCATAT	ACCATAC	ACCATAG	ACCATTA	ACCATTT	ACCATTC	ACCATC	ACCATCA	ACCATCT	ACCATCC	ACCATCG	ACCATCA	ACCATCT	ACCATCC	ACCATCG
ACG	ACGAAA	ACGATAT	ACGATAC	ACGATAG	ACGATTA	ACGATTT	ACGATTC	ACGATC	ACGATCA	ACGATCT	ACGATCC	ACGATCG	ACGATCA	ACGATCT	ACGATCC	ACGATCG
ACT	ACGAAA	ACGATAT	ACGATAC	ACGATAG	ACGATTA	ACGATTT	ACGATTC	ACGATC	ACGATCA	ACGATCT	ACGATCC	ACGATCG	ACGATCA	ACGATCT	ACGATCC	ACGATCG
AGT	AGTAAA	AGTATAT	AGTATAC	AGTATAG	AGTATTA	AGTATTT	AGTATTC	AGTATC	AGTATCA	AGTATCT	AGTATCC	AGTATCG	AGTATCA	AGTATCT	AGTATCC	AGTATCG
AGC	AGCAAA	AGCATAT	AGCATAC	AGCATAG	AGCATTA	AGCATTT	AGCATTC	AGCATC	AGCATCA	AGCATCT	AGCATCC	AGCATCG	AGCATCA	AGCATCT	AGCATCC	AGCATCG
AGG	AGGAAA	AGGATAT	AGGATAC	AGGATAG	AGGATTA	AGGATTT	AGGATTC	AGGATC	AGGATCA	AGGATCT	AGGATCC	AGGATCG	AGGATCA	AGGATCT	AGGATCC	AGGATCG
TAA	TAAAA	TAAAT	TAAAC	TAAAG	TAAAT	TAACT	TAACT	TAACT	TAACT	TAACT	TAACT	TAACT	TAACT	TAACT	TAACT	TAACT
TAT	TATAAA	TATATAT	TATATAC	TATATAG	TATATTA	TATATTT	TATATTC	TATATC	TATATCA	TATATCT	TATATCC	TATATCG	TATATCA	TATATCT	TATATCC	TATATCG
TAC	TACAAA	TACATAT	TACATAC	TACATAG	TACATTA	TACATTT	TACATTC	TACATC	TACATCA	TACATCT	TACATCC	TACATCG	TACATCA	TACATCT	TACATCC	TACATCG
TAG	TAGAAA	TAGATAT	TAGATAC	TAGATAG	TAGATTA	TAGATTT	TAGATTC	TAGATC	TAGATCA	TAGATCT	TAGATCC	TAGATCG	TAGATCA	TAGATCT	TAGATCC	TAGATCG
TIA	TIAAAA	TIATAT	TIACAT	TIAGAT	TIATTA	TIATTT	TIATTC	TIATTC	TIATCA	TIATCT	TIATCC	TIATCG	TIATCA	TIATCT	TIATCC	TIATCG
TTT	TTTAAA	TTTATAT	TTTATAC	TTTATAG	TTTATTA	TTTATTT	TTTATTC	TTTATC	TTTATCA	TTTATCT	TTTATCC	TTTATCG	TTTATCA	TTTATCT	TTTATCC	TTTATCG
TTG	TTGAAA	TTGATAT	TTGATAC	TTGATAG	TTGATTA	TTGATTT	TTGATTC	TTGATC	TTGATCA	TTGATCT	TTGATCC	TTGATCG	TTGATCA	TTGATCT	TTGATCC	TTGATCG
TCA	TCAAAA	TCATAT	TCACAT	TCAGAT	TCATTA	TCATTT	TCATTC	TCATTC	TCATCA	TCATCT	TCATCC	TCATCG	TCATCA	TCATCT	TCATCC	TCATCG
TCG	TCGAAA	TCGATAT	TCGATAC	TCGATAG	TCGATTA	TCGATTT	TCGATTC	TCGATC	TCGATCA	TCGATCT	TCGATCC	TCGATCG	TCGATCA	TCGATCT	TCGATCC	TCGATCG
TCC	TCCTAA	TCCTAT	TCCTAC	TCCTAG	TCCTTA	TCCTTT	TCCTTC	TCCTTC	TCCTCA	TCCTCT	TCCTCC	TCCTCG	TCCTCA	TCCTCT	TCCTCC	TCCTCG
TCC	TCCTAA	TCCTAT	TCCTAC	TCCTAG	TCCTTA	TCCTTT	TCCTTC	TCCTTC	TCCTCA	TCCTCT	TCCTCC	TCCTCG	TCCTCA	TCCTCT	TCCTCC	TCCTCG
TCG	TCGAAA	TCGATAT	TCGATAC	TCGATAG	TCGATTA	TCGATTT	TCGATTC	TCGATC	TCGATCA	TCGATCT	TCGATCC	TCGATCG	TCGATCA	TCGATCT	TCGATCC	TCGATCG
TGG	TCGAAA	TCGATAT	TCGATAC	TCGATAG	TCGATTA	TCGATTT	TCGATTC	TCGATC	TCGATCA	TCGATCT	TCGATCC	TCGATCG	TCGATCA	TCGATCT	TCGATCC	TCGATCG
TGT	TGTAAA	TGTATAT	TGTATAC	TGTATAG	TGTATTA	TGTATTT	TGTATTC	TGTATC	TGTATCA	TGTATCT	TGTATCC	TGTATCG	TGTATCA	TGTATCT	TGTATCC	TGTATCG
TGC	TGCAAA	TGCATAT	TGCATAC	TGCATAG	TGCATTA	TGCATTT	TGCATTC	TGCATC	TGCATCA	TGCATCT	TGCATCC	TGCATCG	TGCATCA	TGCATCT	TGCATCC	TGCATCG
TGG	TGCAAA	TGCATAT	TGCATAC	TGCATAG	TGCATTA	TGCATTT	TGCATTC	TGCATC	TGCATCA	TGCATCT	TGCATCC	TGCATCG	TGCATCA	TGCATCT	TGCATCC	TGCATCG
TGA	TGAAAA	TGATAT	TGACAT	TGAGAT	TGATTA	TGATTT	TGATTC	TGATTC	TGATCA	TGATCT	TGATCC	TGATCG	TGATCA	TGATCT	TGATCC	TGATCG
TGA	TGAAAA	TGATAT	TGACAT	TGAGAT	TGATTA	TGATTT	TGATTC	TGATTC	TGATCA	TGATCT	TGATCC	TGATCG	TGATCA	TGATCT	TGATCC	TGATCG
TGT	TGTAAA	TGTATAT	TGTATAC	TGTATAG	TGTATTA	TGTATTT	TGTATTC	TGTATC	TGTATCA	TGTATCT	TGTATCC	TGTATCG	TGTATCA	TGTATCT	TGTATCC	TGTATCG
TGC	TGCAAA	TGCATAT	TGCATAC	TGCATAG	TGCATTA	TGCATTT	TGCATTC	TGCATC	TGCATCA	TGCATCT	TGCATCC	TGCATCG	TGCATCA	TGCATCT	TGCATCC	TGCATCG
TGG	TGCAAA	TGCATAT	TGCATAC	TGCATAG	TGCATTA	TGCATTT	TGCATTC	TGCATC	TGCATCA	TGCATCT	TGCATCC	TGCATCG	TGCATCA	TGCATCT	TGCATCC	TGCATCG
TGA	TGAAAA	TGATAT	TGACAT	TGAGAT	TGATTA	TGATTT	TGATTC	TGATTC	TGATCA	TGATCT	TGATCC	TGATCG	TGATCA	TGATCT	TGATCC	TGATCG
TGA	TGAAAA	TGATAT	TGACAT	TGAGAT	TGATTA	TGATTT	TGATTC	TGATTC	TGATCA	TGATCT	TGATCC	TGATCG	TGATCA	TGATCT	TGATCC	TGATCG
TGT	TGTAAA	TGTATAT	TGTATAC	TGTATAG	TGTATTA	TGTATTT	TGTATTC	TGTATC	TGTATCA	TGTATCT	TGTATCC	TGTATCG	TGTATCA	TGTATCT	TGTATCC	TGTATCG
TGC	TGCAAA	TGCATAT	TGCATAC	TGCATAG	TGCATTA	TGCATTT	TGCATTC	TGCATC	TGCATCA	TGCATCT	TGCATCC	TGCATCG	TGCATCA	TGCATCT	TGCATCC	TGCATCG
TGG	TGCAAA	TGCATAT	TGCATAC	TGCATAG	TGCATTA	TGCATTT	TGCATTC	TGCATC	TGCATCA	TGCATCT	TGCATCC	TGCATCG	TGCATCA	TGCATCT	TGCATCC	TGCATCG
TGA	TGAAAA	TGATAT	TGACAT	TGAGAT	TGATTA	TGATTT	TGATTC	TGATTC	TGATCA	TGATCT	TGATCC	TGATCG	TGATCA	TGATCT	TGATCC	TGATCG
TGA	TGAAAA	TGATAT	TGACAT	TGAGAT	TGATTA	TGATTT	TGATTC	TGATTC	TGATCA	TGATCT	TGATCC	TGATCG	TGATCA	TGATCT	TGATCC	TGATCG
TGT	TGTAAA	TGTATAT	TGTATAC	TGTATAG	TGTATTA	TGTATTT	TGTATTC	TGTATC	TGTATCA	TGTATCT	TGTATCC	TGTATCG	TGTATCA	TGTATCT	TGTATCC	TGTATCG
TGC	TGCAAA	TGCATAT	TGCATAC	TGCATAG	TGCATTA	TGCATTT	TGCATTC	TGCATC	TGCATCA	TGCATCT	TGCATCC	TGCATCG	TGCATCA	TGCATCT	TGCATCC	TGCATCG
TGG	TGCAAA	TGCATAT	TGCATAC	TGCATAG	TGCATTA	TGCATTT	TGCATTC	TGCATC	TGCATCA	TGCATCT	TGCATCC	TGCATCG	TGCATCA	TGCATCT	TGCATCC	TGCATCG
TGA	TGAAAA	TGATAT	TGACAT	TGAGAT	TGATTA	TGATTT	TGATTC	TGATTC	TGATCA	TGATCT	TGATCC	TGATCG	TGATCA	TGATCT	TGATCC	TGATCG
TGA	TGAAAA	TGATAT	TGACAT	TGAGAT	TGATTA	TGATTT	TGATTC	TGATTC	TGATCA	TGATCT	TGATCC	TGATCG	TGATCA	TGATCT	TGATCC	TGATCG
TGT	TGTAAA	TGTATAT	TGTATAC	TGTATAG	TGTATTA	TGTATTT	TGTATTC	TGTATC	TGTATCA	TGTATCT	TGTATCC	TGTATCG	TGTATCA	TGTATCT	TGTATCC	TGTATCG
TGC	TGCAAA	TGCATAT	TGCATAC	TGCATAG	TGCATTA	TGCATTT	TGCATTC	TGCATC	TGCATCA	TGCATCT	TGCATCC	TGCATCG	TGCATCA	TGCATCT	TGCATCC	TGCATCG
TGG	TGCAAA	TGCATAT	TGCATAC	TGCATAG	TGCATTA	TGCATTT	TGCATTC	TGCATC	TGCATCA	TGCATCT	TGCATCC	TGCATCG	TGCATCA	TGCATCT	TGCATCC	TGCATCG
TGA	TGAAAA	TGATAT	TGACAT	TGAGAT	TGATTA	TGATTT	TGATTC	TGATTC	TGATCA	TGATCT	TGATCC	TGATCG	TGATCA	TGATCT	TGATCC	TGATCG
TGA	TGAAAA	TGATAT	TGACAT	TGAGAT	TGATTA	TGATTT	TGATTC	TGATTC	TGATCA	TGATCT	TGATCC	TGATCG	TGATCA	TGATCT	TGATCC	TGATCG
TGT	TGTAAA	TGTATAT	TGTATAC	TGTATAG	TGTATTA	TGTATTT	TGTATTC	TGTATC	TGTATCA	TGTATCT	TGTATCC	TGTATCG	TGTATCA	TGTATCT	TGTATCC	TGTATCG
TGC	TGCAAA	TGCATAT	TGCATAC	TGCATAG	TGCATTA	TGCATTT	TGCATTC	TGCATC	TGCATCA	TGCATCT	TGCATCC	TGCATCG	TGCATCA	TGCATCT	TGCATCC	TGCATCG
TGG	TGCAAA	TGCATAT	TGCATAC	TGCATAG	TGCATTA	TGCATTT	TGCATTC	TGCATC	TGCATCA	TGCATCT	TGCATCC	TGCATCG	TGCATCA	TGCATCT	TGCATCC	TGCATCG
TGA	TGAAAA	TGATAT	TGACAT	TGAGAT	TGATTA	TGATTT	TGATTC	TGATTC	TGATCA	TGATCT	TGATCC	TGATCG	TGATCA	TGATCT	TGATCC	TGATCG
TGA	TGAAAA	TGATAT	TGACAT	TGAGAT	TGATTA	TGATTT	TGATTC	TGATTC	TGATCA	TGATCT	TGATCC	TGATCG	TGATCA	TGATCT	TGATCC	TGATCG
TGT	TGTAAA	TGTATAT	TGTATAC	TGTATAG	TGTATTA	TGTATTT	TGTATTC	TGTATC	TGTATCA	TGTATCT	TGTATCC	TGTATCG	TGTATCA	TGTATCT	TGTATCC	TGTATCG
TGC	TGCAAA	TGCATAT	TGCATAC	TGCATAG	TGCATTA	TGCATTT	TGCATTC	TGCATC	TGCATCA	TGCATCT	TGCATCC	TGCATCG	TGCATCA	TGCATCT	TGCATCC	TGCATCG
TGG	TGCAAA	TGCATAT	TGCATAC	TGCATAG	TGCATTA	TGCATTT	TGCATTC	TGCATC	TGCATCA	TGCATCT	TGCATCC	TGCATCG	TGCATCA	TGCATCT	TGCATCC	TGCATCG
TGA	TGAAAA	TGATAT	TGACAT	TGAGAT	TGATTA	TGATTT	TGATTC	TGATTC	TGATCA	TGATCT	TGATCC	TGATCG	TGATCA	TGATCT	TGATCC	TGATCG
TGA	TGAAAA	TGATAT	TGACAT	TGAGAT	TGATTA	TGATTT	TGATTC	TGATTC	TGATCA	TGATCT	TGATCC	TGATCG	TGATCA	TGATCT	TGATCC	TGATCG
TGT	TGTAAA	TGTATAT	TGTATAC	TGTATAG	TGTATTA	TGTATTT	TGTATTC	TGTATC	TGTATCA	TGTATCT	TGTATCC	TGTATCG	TGTATCA	TGTATCT	TGTATCC	TGTATCG
TGC	TGCAAA	TGCATAT	TGCATAC	TGCATAG	TGCATTA	TGCATTT	TGCATTC	TGCATC	TGCATCA	TGCATCT	TGCATCC	TGCATCG	TGCATCA	TGCATCT	TGCATCC	TGCATCG
TGG	TGCAAA	TGCATAT	TGCATAC	TGCATAG	TGCATTA	TGCATTT	TGCATTC	TGCATC	TGCATCA	TGCATCT	TGCATCC	TGCATCG	TGCATCA	TGCATCT	TGCATCC	TGCATCG
TGA	TGAAAA	TGATAT	TGACAT	TGAGAT	TGATTA	TGATTT	TGATTC	TGATTC	TGATCA	TGATCT	TGATCC	TGATCG	TGATCA	TGATCT	TGATCC	TGATCG
TGA	TGAAAA	TGATAT	TGACAT	TGAGAT	TGATTA	TGATTT	TGATTC	TGATTC	TGATCA	TGATCT	TGATCC	TGATCG	TGATCA	TGATCT	TGATCC	TGATCG
TGT	TGTAAA	TGTATAT	TGTATAC	TGTATAG	TGTATTA	TGTATTT	TGTATTC	TGTATC	TGTATCA	TGTATCT	TGTATCC	TGTATCG	TGTATCA	TGTATCT	TGTATCC	TGTATCG
TGC	TGCAAA	TGCATAT	TGCATAC	TGCATAG	TGCATTA	TGCATTT	TGCATTC	TGCATC	TGCATCA	TGCATCT	TGCATCC	TGCATCG	TGCATCA	TGCATCT	TGCATCC	TGCATCG
TGG	TGCAAA	TGCATAT	TGCATAC	TGCATAG	TGCATTA	TGCATTT	TGCATTC	TGCATC	TGCATCA	TGCATCT	TGCATCC	TGCATCG	TGCATCA	TGCATCT	TGCATCC	TGCATCG
TGA	TGAAAA	TGATAT	TGACAT	TGAGAT	TGATTA	TGATTT	TGATTC	TGATTC	TGATCA	TGATCT	TGATCC	TGATCG	TGATCA	TGATCT	TGATCC	TGATCG
TGA	TGAAAA	TGATAT	TGACAT	TGAGAT	TGATTA	TGATTT	TGATTC	TGATTC	TGATCA	TGATCT	TGATCC	TGATCG	TGATCA	TGATCT	TGATCC	TGATCG
TGT	TGTAAA	TGTATAT	TGTATAC	TGTATAG	TGTATTA	TGTATTT	TGTATTC	TGTATC	TGTATCA	TGTATCT	TGTATCC	TGTATCG	TGTATCA	TGTATCT	TGTATCC	TGTATCG
TGC	TGCAAA	TGCATAT	TGCATAC	TGCATAG	TGCATTA	TGCATTT	TGCATTC	TGCATC	TGCATCA	TGCATCT	TGCATCC	TGCATCG	TGCATCA	TGCATCT	TGCATCC	TGCATCG
TGG	TGCAAA	TGCATAT	TGCATAC	TGCATAG	TGCATTA	TGCATTT	TGCATTC	TGCATC	TGCATCA	TGCATCT	TGCATCC	TGCATCG	TGCATCA	TGCATCT	TGCATCC	TGCATCG
TGA	TGAAAA	TGATAT	TGACAT	TGAGAT	TGATTA	TGATTT	TGATTC	TGATTC	TGATCA	TGATCT	TGATCC	TGATCG	TGATCA	TGATCT	TGATCC	TGATCG
TGA	TGAAAA	TGATAT	TGACAT	TGAGAT	TGATTA	TGATTT	TGATTC	TGATTC	TGATCA	TGATCT	TGATCC	TGATCG	TGATCA	TGATCT	TGATCC	TGATCG
TGT	TGTAAA	TGTATAT	TGTATAC	TGTATAG	TGTATTA	TGTATTT	TGTATTC	TGTATC	TGTATCA	TGTATCT	TGTATCC	TGTATCG	TGTATCA	TGTATCT	TGTATCC	TGTATCG
TGC	TGCAAA	TGCATAT	TGCATAC	TGCATAG	TGCATTA	TGCATTT	TGCATTC	TGCATC	TGCATCA	TGCATCT	TGCATCC	TGCATCG	TGCATCA	TGCATCT	TGCATCC	TGCATCG
TGG	TGCAAA	TGCATAT	TGCATAC	TGCATAG	TGCATTA	TGCATTT	TGCATTC	TGCATC	TGCATCA	TGCATCT	TGCATCC	TGCATCG	TGCATCA	TGCATCT	TGCATCC	TGCATCG
TGA	TGAAAA	TGATAT	TGACAT	TGAGAT	TGATTA	TGATTT	TGATTC	TGATTC	TGATCA	TGATCT	TGATCC	TGATCG	TGATCA	TGATCT	TGATCC	TGATCG
TGA	TGAAAA	TGATAT	TGACAT	TGAGAT	TGATTA	TGATTT	TGATTC	TGATTC	TGATCA	TGATCT	TGATCC	TGATCG	TGATCA	TGATCT	TGATCC	TGATCG
TGT	TGTAAA	TGTATAT	TGTATAC	TGTATAG	TGTATTA	TGTATTT	TGTATTC	TGTATC	TGTATCA	TGTATCT	TGTATCC	TGTATCG	TGTATCA	TGTATCT	TGTATCC	TGTATCG
TGC	TGCAAA	TGCATAT	TGCATAC	TGCATAG	TGCATTA	TGCATTT	TGCATTC	TGCATC	TGCATCA	TGCATCT	TGCATCC	TGCATCG	TGCATCA	TGCATCT	TGCATCC	TGCATCG
TGG	TGCAAA	TGCATAT	TGCATAC	TGCATAG	TGCATTA	TGCATTT	TGCATTC	TGCATC	TGCATCA	TGCATCT	TGCATCC	TGCATCG	TGCATCA	TGCATCT	TGCATCC	TGCATCG
TGA	TGAAAA	TGATAT	TGACAT	TGAGAT	TGATTA	TGATTT	TGATTC	TGATTC	TGATCA	TGATCT	TGATCC	TGATCG	TGATCA	TGATCT	TGATCC	TGATCG
TGA	TGAAAA	TGATAT	TGACAT	TGAGAT												

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

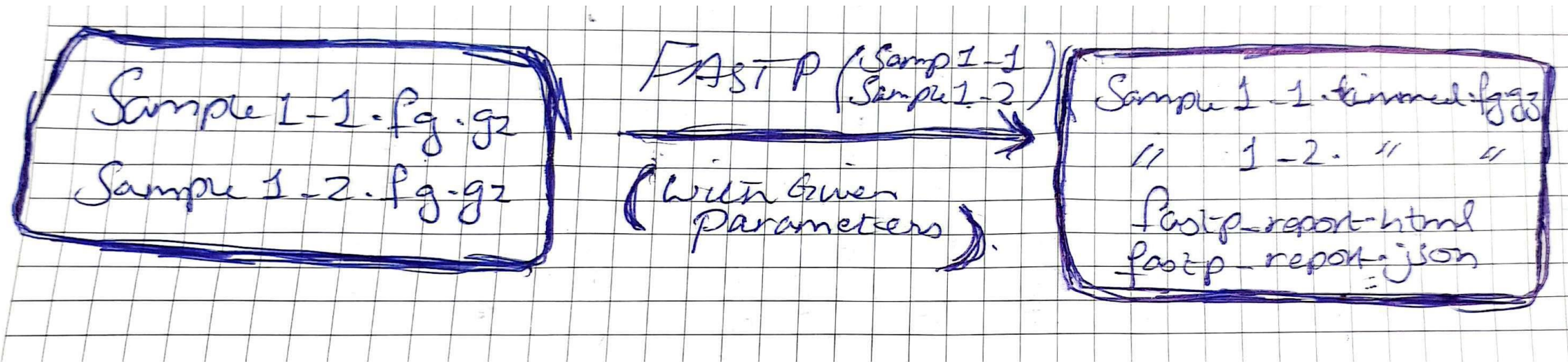
[illegible]

Same as previous analysis



WORKFLOW 1A

WORKFLOW 1B



CS Scanned with CamScanner

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/minute
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.

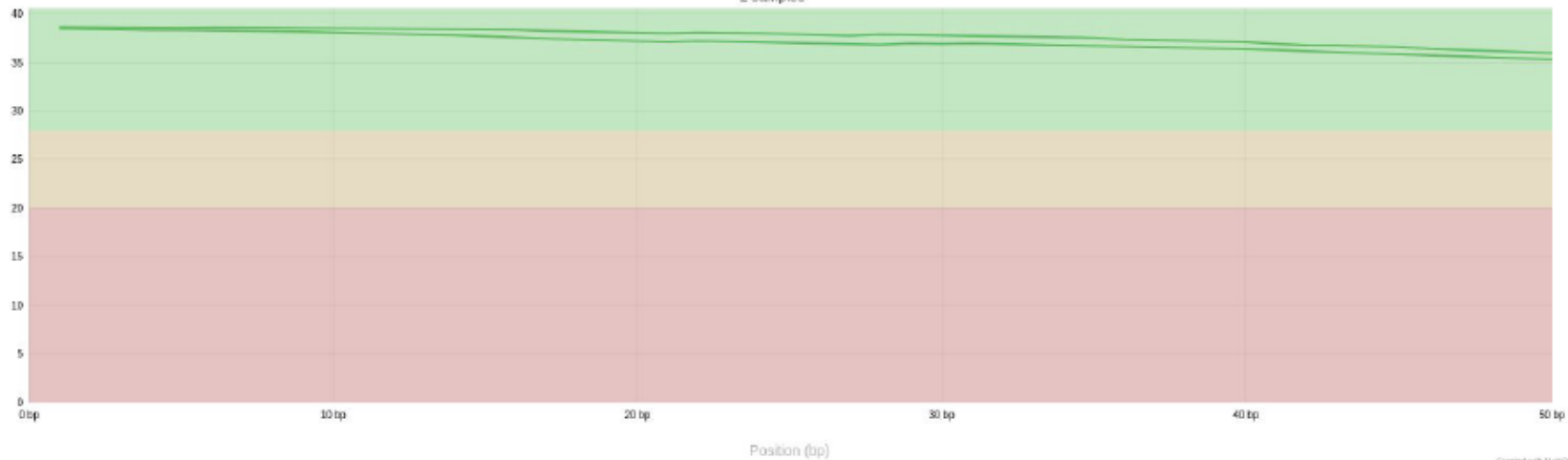
Summarize table

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

FastQC: Mean Quality Scores

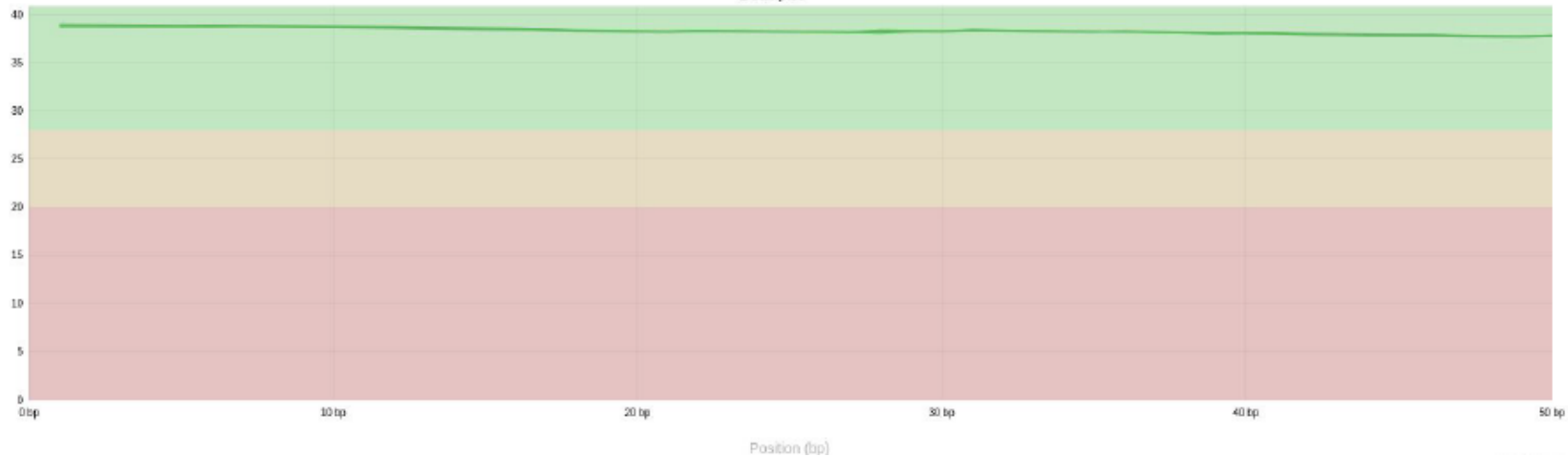
2 samples



Untrimmed

FastQC: Mean Quality Scores

2 samples



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
```

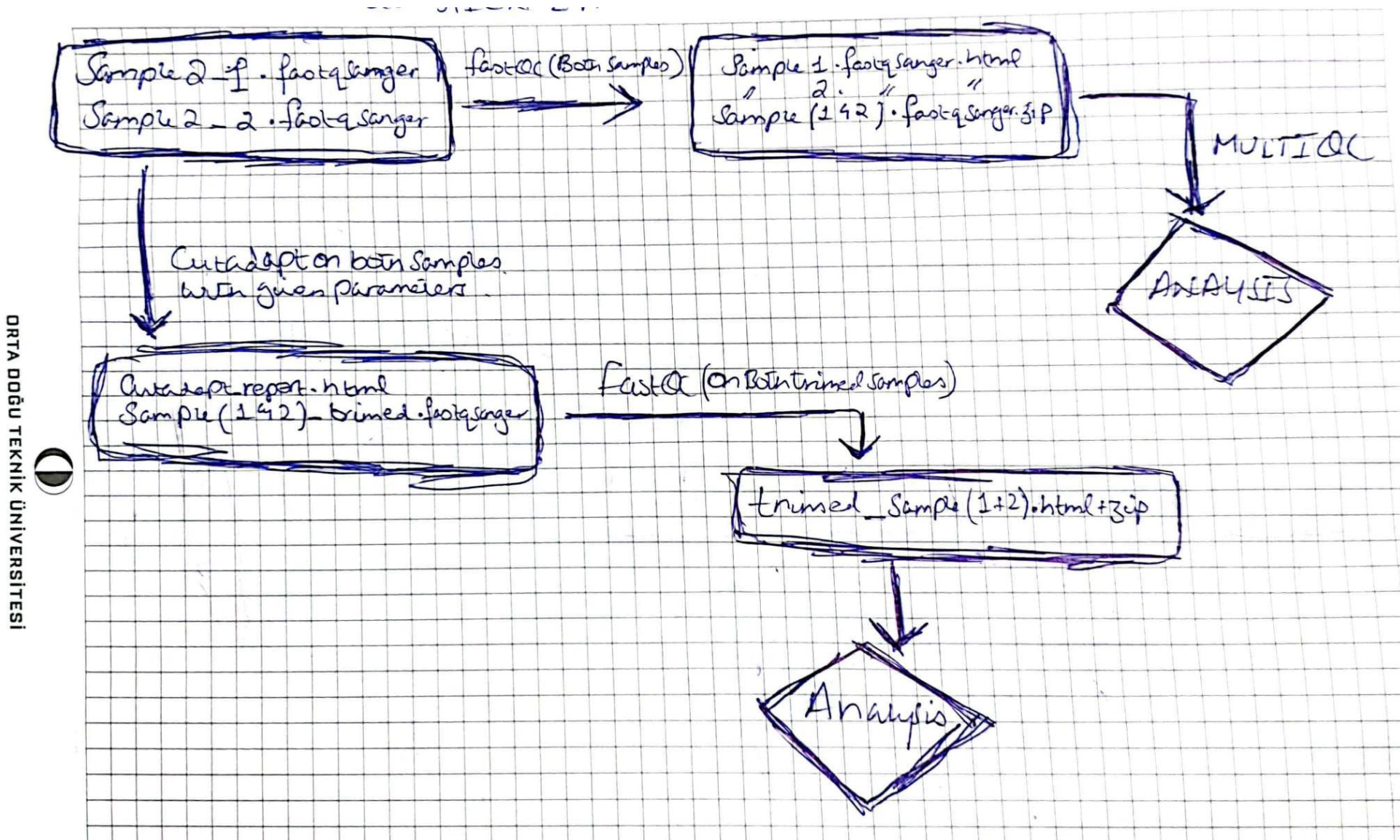
After QC, the alignment was performed.

Above is a screenshot of the files generated after indexing and aligning, along with a snippet of the .sam file


```
(bioinfo) megatron@Taha:/mnt/d/HumanGenome$ head -n 5 sample2_sorted_depth.txt
NC_012920.1      8648      1028
NC_012920.1      8647      1012
NC_012920.1      8650       992
NC_012920.1      8649       989
NC_012920.1      8643       975
```

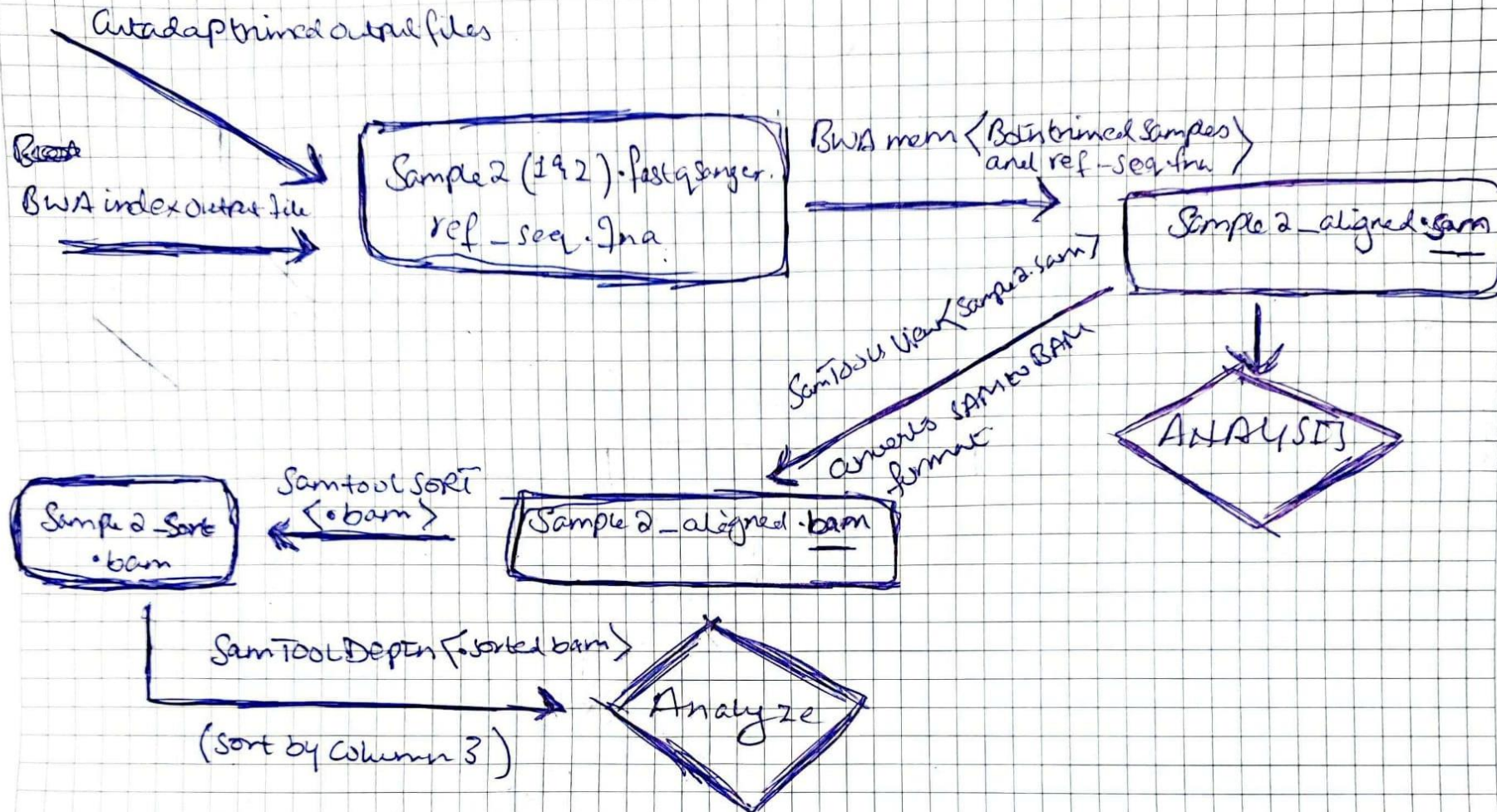
2.c) NC_012920.1 which corresponds to chrM, reads aligned the most

Q2) QC Workflow



ORTA DOĞU TEKNİK ÜNİVERSİTESİ





Linux History

```
1
2 ##### START #####
3
4 ## Setting up the environment
5
6 megatron@Taha:~$ sudo apt update
7 megatron@Taha:~$ source ~/miniconda3/bin/activate
8 (base) megatron@Taha:~$ conda activate bioinfo
9 (bioinfo) megatron@Taha:~$ conda env list
10
11
12 ## Question 1 Workflow
13
14 # Raw data QC
15
16 (bioinfo) megatron@Taha:/mnt/d/HumanGenome/bin508-03$ mkdir raw_fastqc
17 (bioinfo) megatron@Taha:/mnt/d/HumanGenome$ fastqc sample1_1.fq.gz sample1_2.fq.gz -o raw_fastqc/
18 (bioinfo) megatron@Taha:/mnt/d/HumanGenome$ multiqc raw_fastqc/ -o raw_multiqc/
19
20 # trimming with cutadapt
21
22 (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
23
24 #trimmed data QC
25 (bioinfo) megatron@Taha:/mnt/d/HumanGenome$ mkdir trimmed_fastqc
26 (bioinfo) megatron@Taha:/mnt/d/HumanGenome$ fastqc trimmed_1.fq.gz trimmed_2.fq.gz -o trimmed_fastqc/
27 (bioinfo) megatron@Taha:/mnt/d/HumanGenome$ multiqc trimmed_fastqc/ -o trimmed_multiqc/
28
29 # fastp
30
31 (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
32 | --length_required 20 --html fastp_report.html
33
34 ## Question 2 Workflow
35
36 # raw data qc
37 (bioinfo) megatron@Taha:/mnt/d/HumanGenome$ mkdir sample2_raw_fastqc
38 (bioinfo) megatron@Taha:/mnt/d/HumanGenome$ fastqc sample2_1.fastqsanger sample2_2.fastqsanger -o sample2_raw_fastqc/
39
```

```
40 # trimming
41
42 (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
43 sample2_2.fastqsanger
44
45 # trimmed data QC
46
47 (bioinfo) megatron@Taha:/mnt/d/HumanGenome$ mkdir sample2_trimmed_fastqc
48 (bioinfo) megatron@Taha:/mnt/d/HumanGenome$ fastqc sample2_trimmed_1.fq.gz sample2_trimmed_2.fq.gz -o sample2_trimmed_fastqc/
49 (bioinfo) megatron@Taha:/mnt/d/HumanGenome$ multiqc sample2_trimmed_fastqc/ -o sample2_trimmed_multiqc/
50
51 # BWA-MEM Alignment
52
53 # Align trimmed reads to reference genome using BWA-MEM
54 # -t 8: Use 8 CPU threads for faster alignment
55 # -R: Add read group information (required for downstream analysis)
56 # "@RG\tID:sample2\tSM:sample2\tPL:ILLUMINA": Read group metadata
57
58 (bioinfo) megatron@Taha:/mnt/d/HumanGenome$ bwa mem -t 8 -R "@RG\tID:sample2\tSM:sample2\tPL:ILLUMINA" /mnt/d/HumanGenome/
59 GCF_000001405.40_GRCh38.p14_genomic.fna sample2_trimmed_1.fq.gz sample2_trimmed_2.fq.gz > sample2_aligned.sam
60
61 (bioinfo) megatron@Taha:/mnt/d/HumanGenome$ head sample2_aligned.sam
62
63 # popst alignment processing
64
65 (bioinfo) megatron@Taha:/mnt/d/HumanGenome$ mkdir -p sample2_q2c_results/alignment sample2_q2c_results/depth
66 (bioinfo) megatron@Taha:/mnt/d/HumanGenome$ samtools sort sample2_aligned.sam -o sample2_q2c_results/alignment/sample2_q2c_sorted.bam
67 (bioinfo) megatron@Taha:/mnt/d/HumanGenome$ samtools depth sample2_q2c_results/alignment/sample2_q2c_sorted.bam > sample2_q2c_results/depth/
68 sample2_q2c_depth.txt
69
70 # Sort depth file by coverage (column 3) in descending order
71 # -k3,3nr: Sort numerically by 3rd column in reverse order
72 # Helps identify regions with highest read coverage
73
74 (bioinfo) megatron@Taha:/mnt/d/HumanGenome$ sort -k3,3nr sample2_q2c_results/depth/sample2_q2c_depth.txt > sample2_q2c_results/depth/
75 sample2_q2c_sorted_depth.txt
76
77 # display
78
79 (bioinfo) megatron@Taha:/mnt/d/HumanGenome$ head -n 5 sample2_q2c_results/depth/sample2_q2c_sorted_depth.txt
80
81 |
82 ##### END #####
```