QAA_report

2025-09-07

NOTE: For the remainder of the assignment, I reference SRR25630304 as 'rhy49' and SRR25630399 as 'rhy106' in plots and tables.

Downloading the Dataset:

```
## Download your data
prefetch SRR25630304
fasterq-dump SRR25630304 #42,642,172 reads each file
gzip SRR25630304_1.fastq
gzip SRR25630304_2.fastq
mkdir Campylomormyrus_rhynchophorus_rhy49_electric_organ_adult
mv SRR25630304_1.fastq.gz Campylomormyrus_rhynchophorus_rhy49_electric_organ_adult/3.1Gb_SAMN36981042_1
mv SRR25630304_2.fastq.gz Campylomormyrus_rhynchophorus_rhy49_electric_organ_adult/3.2Gb_SAMN36981042_2

prefetch SRR25630399
fasterq-dump SRR25630399 #42,944,774 reads each file
gzip SRR25630399_1.fastq
gzip SRR25630399_2.fastq
mkdir Campylomormyrus_rhynchophorus_rhy106_electric_organ_adult
mv SRR25630399_1.fastq.gz Campylomormyrus_rhynchophorus_rhy106_electric_organ_adult/3.1Gb_SAMN36982003_
mv SRR25630399_2.fastq.gz Campylomormyrus_rhynchophorus_rhy106_electric_organ_adult/3.1Gb_SAMN36982003_
```

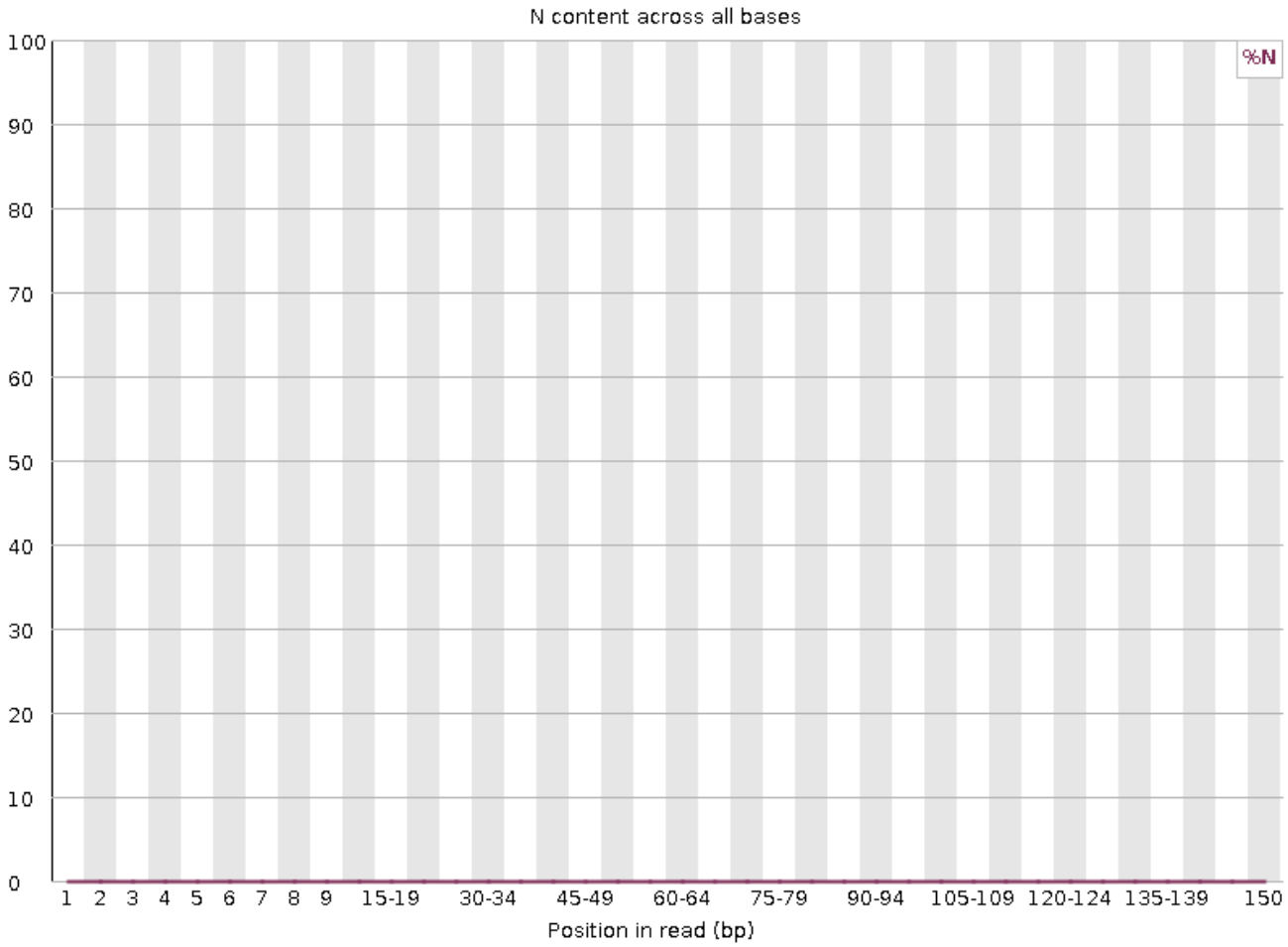
Conda Environment Setup:

```
conda create -n QAA
conda install FastQC
conda install cutadapt
conda install Trimmomatic
conda install Star
conda install Picard=2.18
conda install Samtools
conda install Numpy
conda install Matplotlib
conda install HTSeq
```

Part 1 – Read quality score distributions

Fastqc:

```
fastqc Campylomormyrus_rhynchophorus_rhy49_electric_organ_adult/3.1Gb_SAMN36981042_1.fastq.gz Campylomo
```



Overall, the plots for per-base sequence quality and N-content for all files indicates good quality, but per base sequence content greatly varies in the 1-9bp region. This is likely caused by random primer sequencing biases, so it shouldn't be a big deal. There is also a high number of duplication reported for all files, but this is to be expected because the data was generated from PCR, which leads to a large number of duplicate artifacts.

Quality Score Distributions:

The FastQC plots reflect mine, however run time differs tremendously. This is likely because FastQC is built to handle large files and supports multithreading which is more efficient than my Python script which only runs on one node.

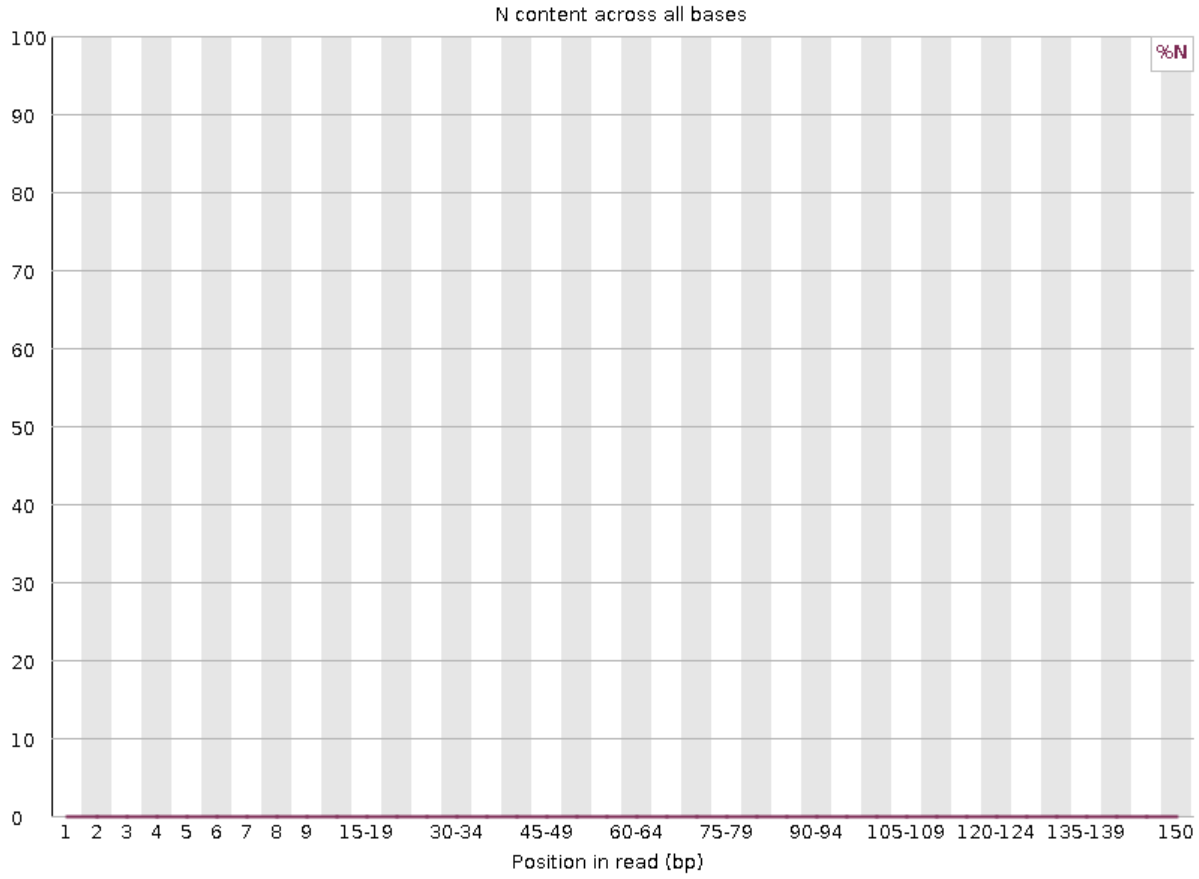


Figure 1: Figure 1.2: Plot of Per base N-Content for Rhy49 Read 2. X-axis represents the position in sequence for a read. Y-axis represents the proportion of N (unidentified bases) for a single base position. Plot indicates a low proportion of N, indicating good quality.

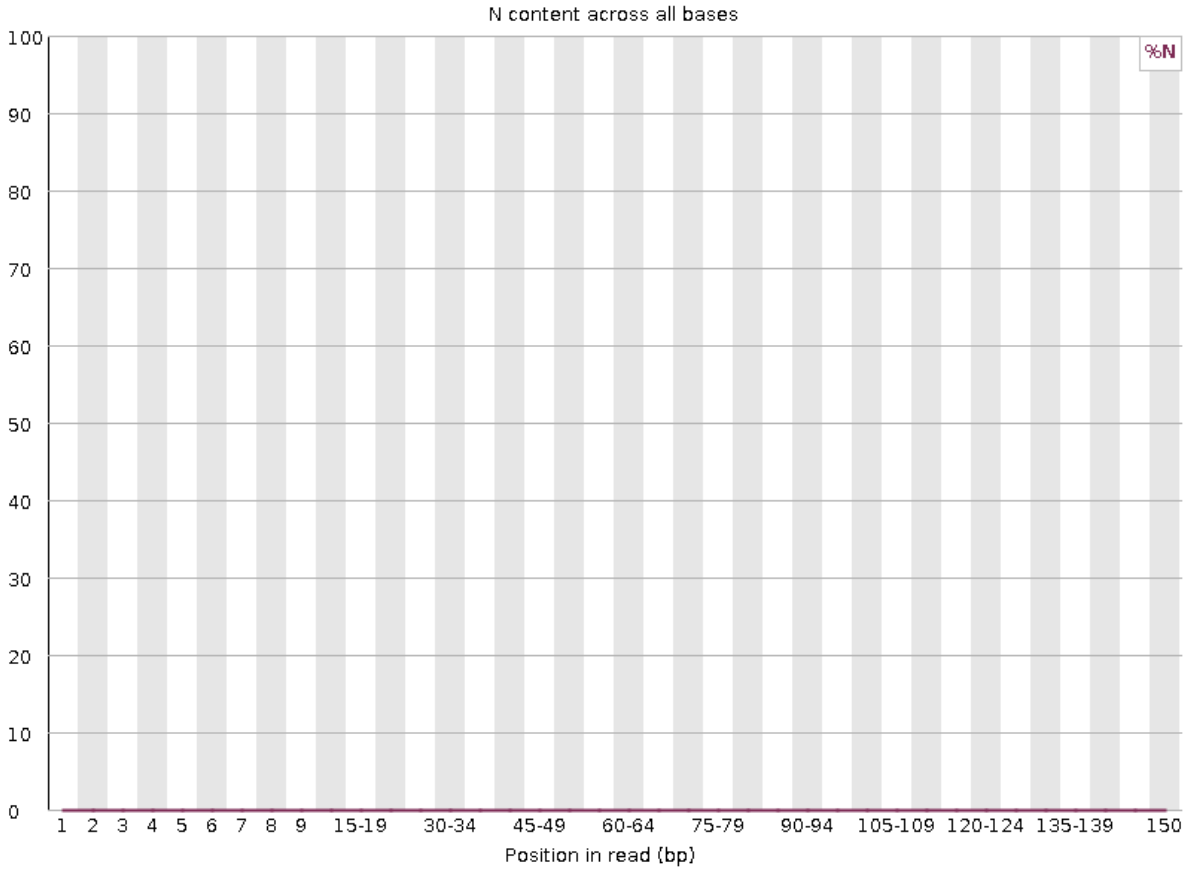


Figure 2: Figure 1.3: Plot of Per base N-Content for Rhy106 Read 1. X-axis represents the position in sequence for a read. Y-axis represents the proportion of N (unidentified bases) for a single base position. Plot indicates a low proportion of N, indicating good quality.

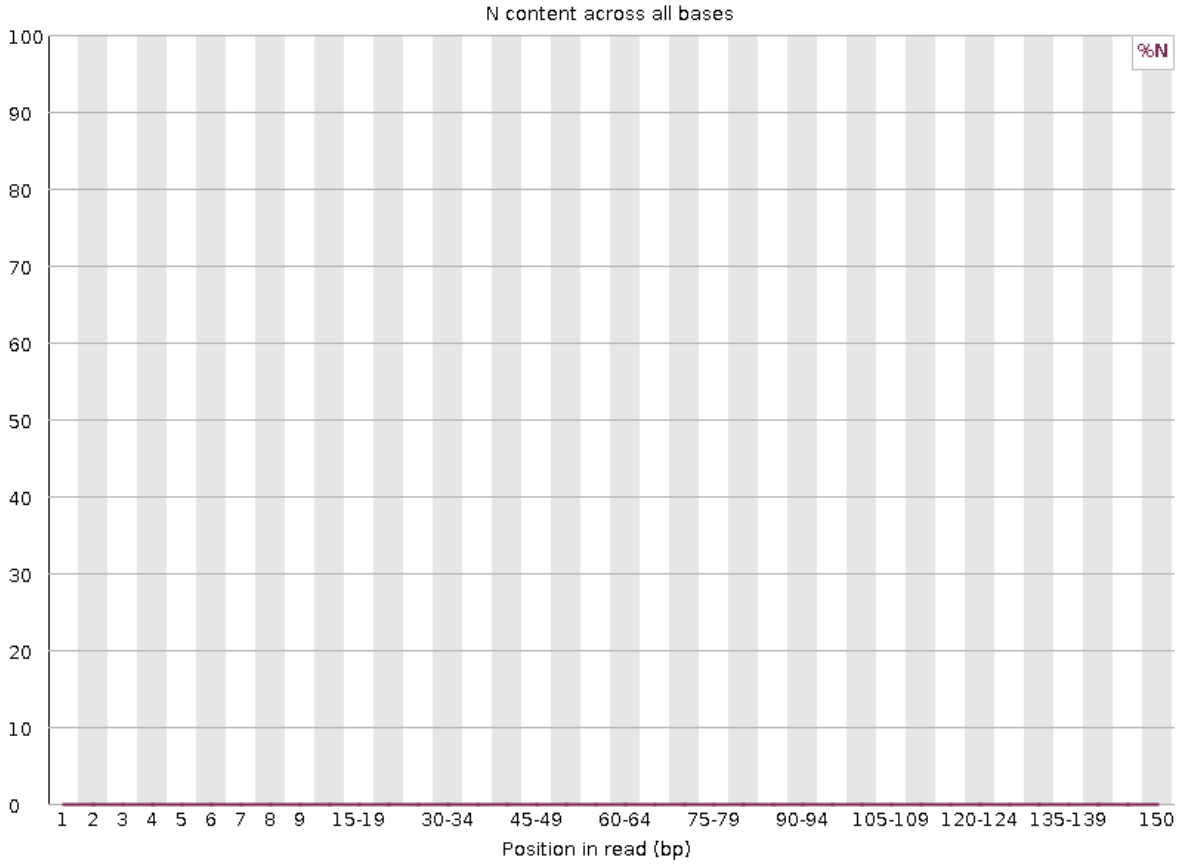


Figure 3: Figure 1.4: Plot of Per base N-Content for Rhy106 Read 2. X-axis represents the position in sequence for a read. Y-axis represents the proportion of N (unidentified bases) for a single base position. Plot indicates a low proportion of N, indicating good quality.

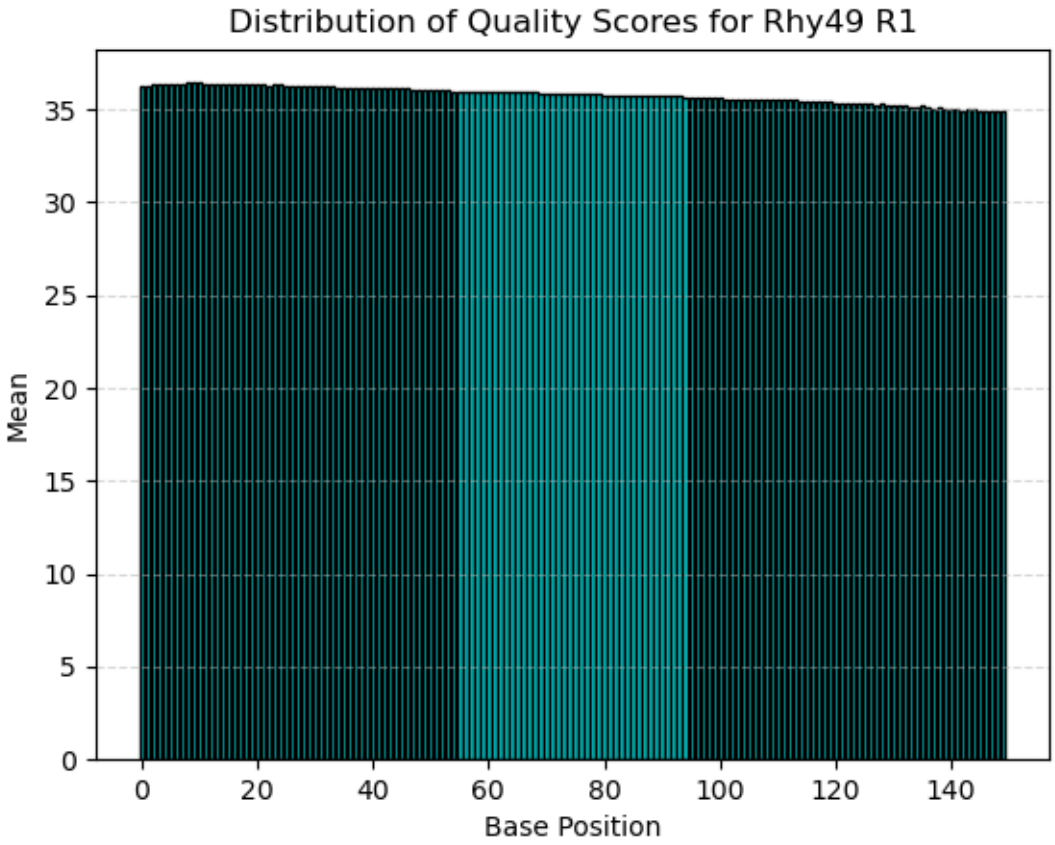


Figure 4: Figure 2.1: Plot of Per Nucleotide Quality Score Distributions for Rhy49 Read 1. X-axis represents the position in sequence for a read. Y-axis represents the mean Phred-33 encoding quality score for a single base position. Plot indicates an overall distribution of high quality scores.

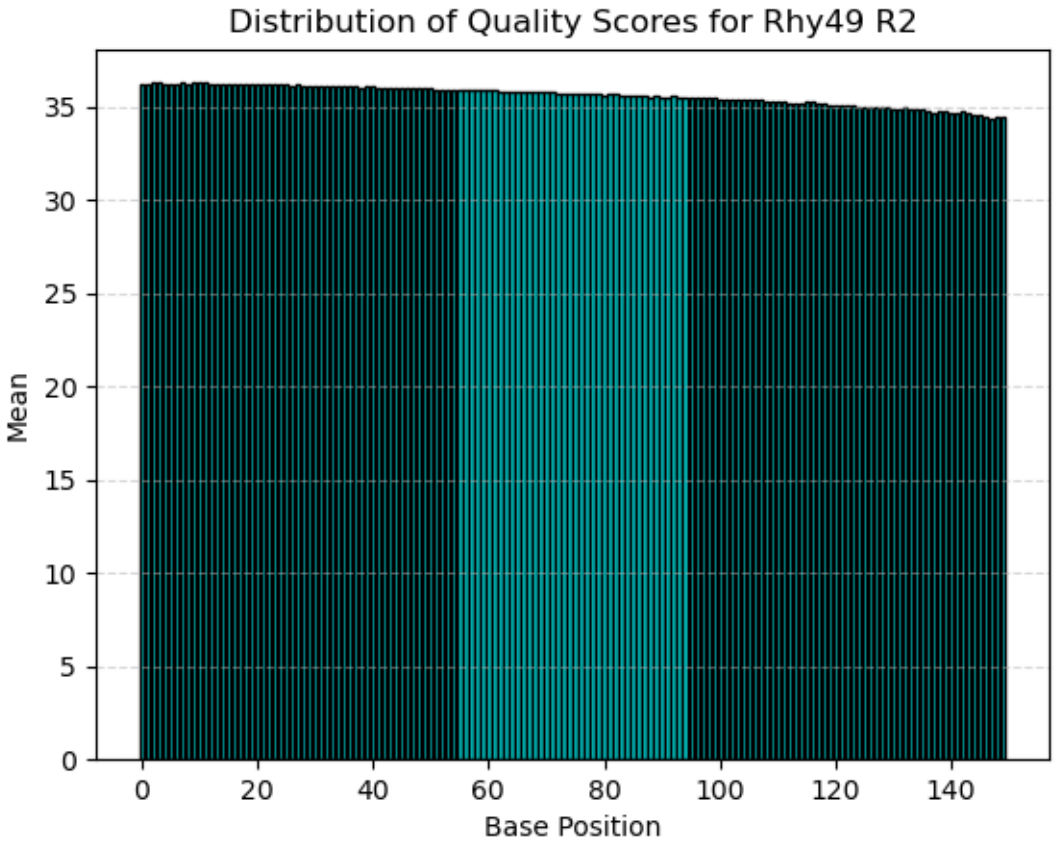


Figure 5: Figure 2.2: Plot of Per Nucleotide Quality Score Distributions for Rhy49 Read 2. X-axis represents the position in sequence for a read. Y-axis represents the mean Phred-33 encoding quality score for a single base position. Plot indicates an overall distribution of high quality scores.

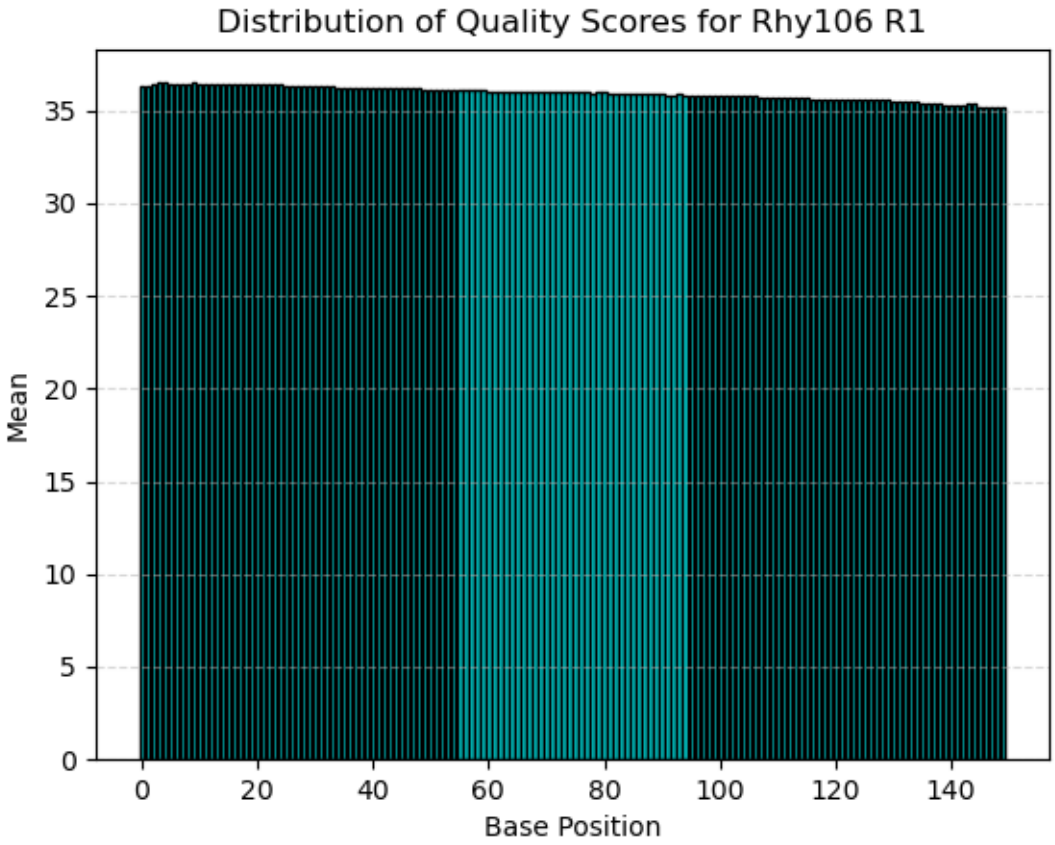


Figure 6: Figure 2.3: Plot of Per Nucleotide Quality Score Distributions for Rhy106 Read 1. X-axis represents the position in sequence for a read. Y-axis represents the mean Phred-33 encoding quality score for a single base position. Plot indicates an overall distribution of high quality scores.

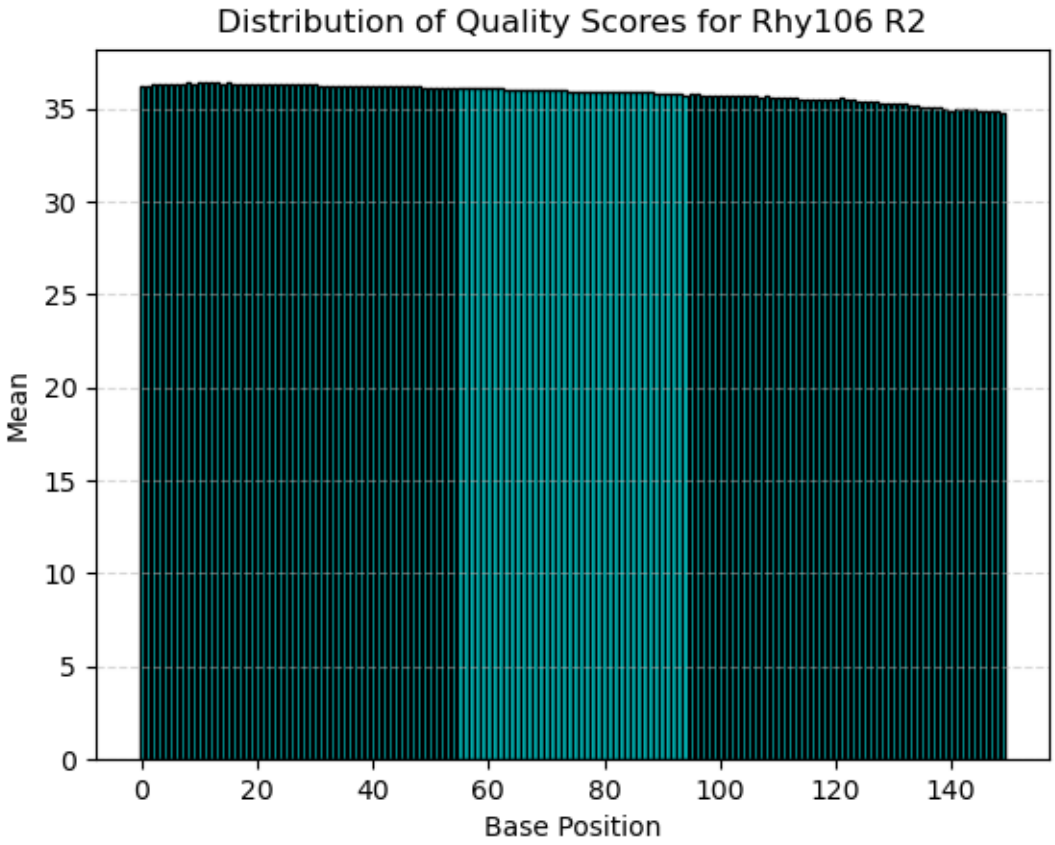


Figure 7: Figure 2.4: Plot of Per Nucleotide Quality Score Distributions for Rhy106 Read 2. X-axis represents the position in sequence for a read. Y-axis represents the mean Phred-33 encoding quality score for a single base position. Plot indicates an overall distribution of high quality scores.

Part 2 – Adaptor trimming comparison

Adapter Trimming:

```
sbatch ./adapter_trim.sh
```

```
##read 1 sanity check
#adapter for barcode 1 shows up in read 1
zcat /projects/bgmp/ewi/bioinfo/Bi623/Assignments/QAA/Campylomormyrus_rhynchophorus_rhy106_electric_orga
#adapter for barcode 2 does not show up in read 1
zcat /projects/bgmp/ewi/bioinfo/Bi623/Assignments/QAA/Campylomormyrus_rhynchophorus_rhy106_electric_orga

##read 2 sanity check
#adapter for barcode 1 does not show up in read 2
zcat /projects/bgmp/ewi/bioinfo/Bi623/Assignments/QAA/Campylomormyrus_rhynchophorus_rhy106_electric_orga
#adapter for barcode 2 shows up in read 2
zcat /projects/bgmp/ewi/bioinfo/Bi623/Assignments/QAA/Campylomormyrus_rhynchophorus_rhy106_electric_orga
```

Rhy49:
*** Read 1: 8.0%.**
*** Read 2: 8.7%.**
*** Total Trimmed = 16.7%.**

Rhy106:
*** Read 1: 11.6%..**
*** Read 2: 11.2%..**
*** Total Trimmed = 22.8%.**

Quality Trim Read Length Distribution:

```
sbatch ./quality_trim.sh
./plot_trimmed.py
```

Because of the different priming sites and sequencing directions, the specific adapter sequence that is read on R1 is the reverse complement of the one read on R2. The rate at which these "read-through" events happen can vary, leading to different trimming rates.

FastQC on Trimmer:

FastQC reported a higher per-base quality for the ending of the sequences after trimming compared to before trimming.

Part 3 – Alignment and strand-specificity

Downloading Gene Models:

```
cp /projects/bgmp/shared/Bi623/PS2/campylomormyrus.fasta .
cp /projects/bgmp/shared/Bi623/PS2/campylomormyrus.gff .
gffread campylomormyrus.gff -T -o campylomormyrus.gtf
```

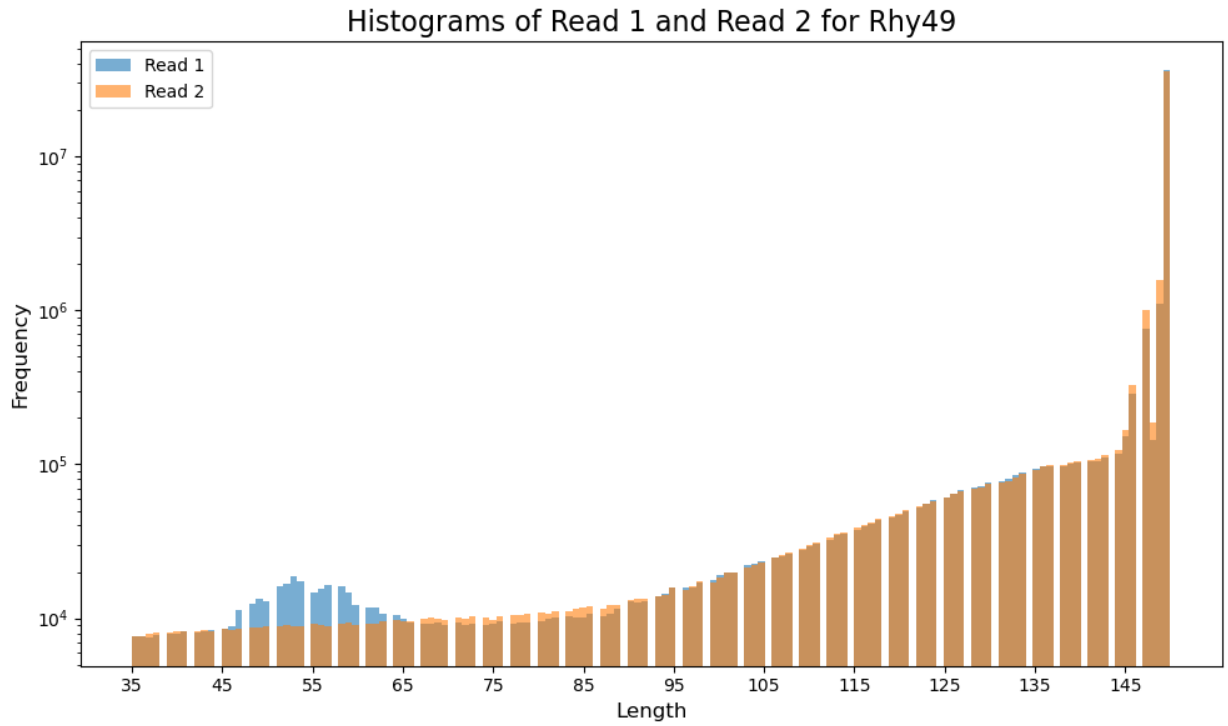


Figure 8: Figure 3.1: Plot of Sequence Length Distributions for Rhy49. X-axis represents the length. Y-axis represents the frequency of sequences at a given length. Legend denotes Read 1 vs Read 2 distributions.

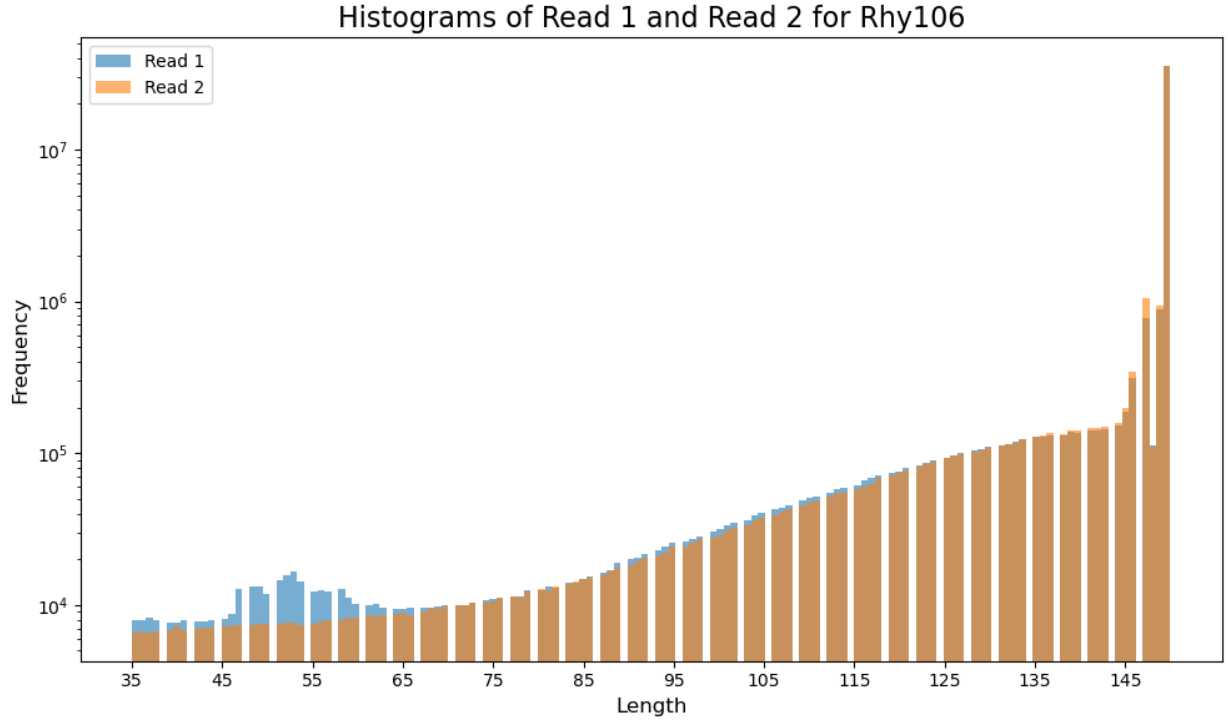


Figure 9: Figure 3.1: Plot of Sequence Length Distributions for Rhy106. X-axis represents the length. Y-axis represents the frequency of sequences at a given length. Legend denotes Read 1 vs Read 2 distributions.

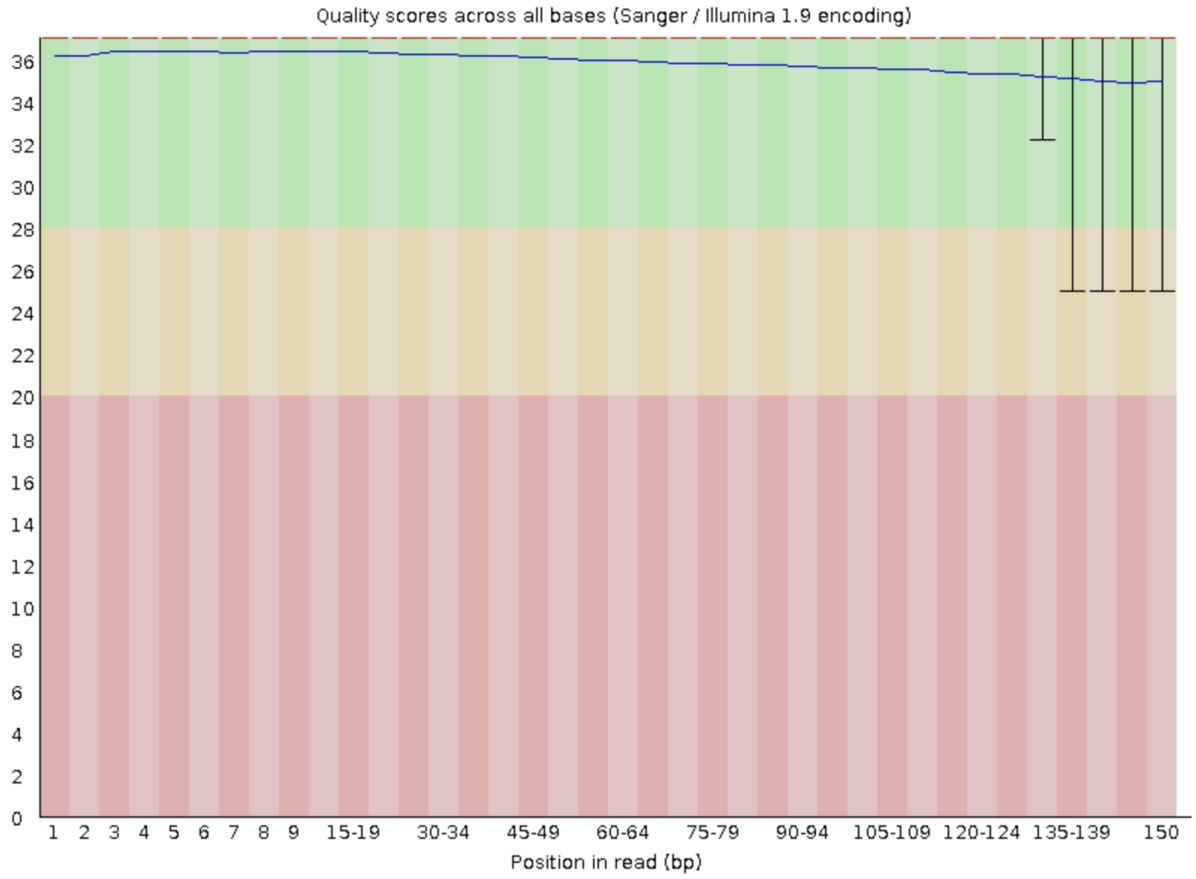


Figure 10: Figure 4.1: Quality Score of Bases before trimming

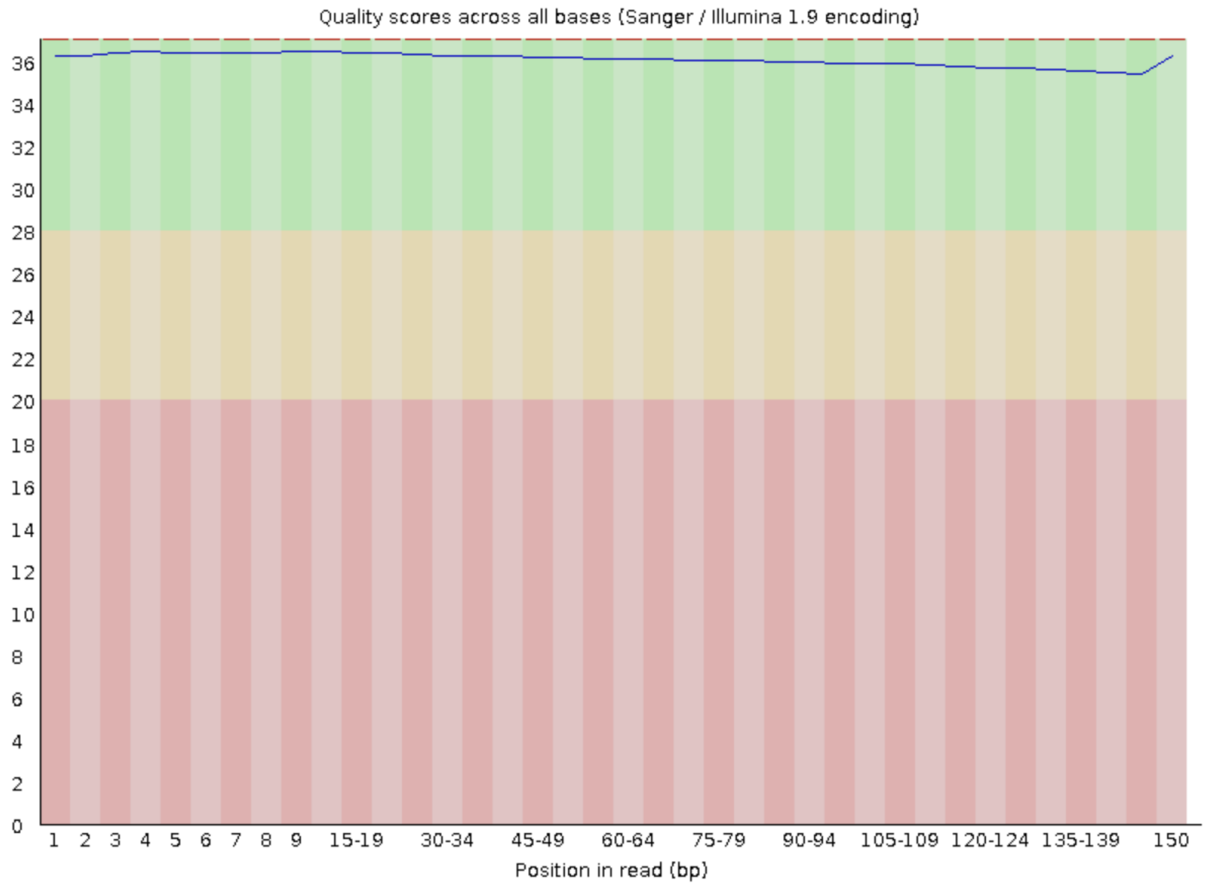


Figure 11: Figure 4.2: Quality Score of Bases after Trimming

Star Alignment:

Building the database

```
./build_star_database.sh
```

Aligning reads

```
./align_reads_star.sh
```

Remove Duplicates:

```
./remove_PCR_dup.sh
```

Mapped/Unmapped Read Counts from PS8 Script:

	Mapped	Unmapped
Rhy49	28028911	3771295
Rhy106	39312727	4670164

Count Deduplicated Reads:

```
./count_deduplicated.sh
awk '{total += $2; if (!/^__/) {mapped += $2}} END {if (total > 0) {printf "%.2f%%\n", mapped * 100 / to
awk '{total += $2; if (!/^__/) {mapped += $2}} END {if (total > 0) {printf "%.2f%%\n", mapped * 100 / to
awk '{total += $2; if (!/^__/) {mapped += $2}} END {if (total > 0) {printf "%.2f%%\n", mapped * 100 / to
awk '{total += $2; if (!/^__/) {mapped += $2}} END {if (total > 0) {printf "%.2f%%\n", mapped * 100 / to
```

	'stranded=yes' % mapped	'stranded=rev' % mapped
Rhy49	3.43%	32.82%
Rhy106	3.48%	35.19%

Based on the library kit, it is strand-specific as the product description explicitly states that the kit is a "fast, strand-specific total RNA-seq library-prep solution." More specifically, it is a first-strand library because the % mapped when 'stranded=rev' was 32.82% and 35.19% (Rhy49 and Rhy106 respectively), which is magnitudes greater than 3.43% and 3.48% when 'stranded=rev'. Therefore, 'stranded=rev' tag should be used in future differential gene expression analysis.