

# OMICS RNA-SEQ REPORT (OMICS1 ASSIGNMENT)

*by Jude Aneke (Group: M06-310)*

This bioinformatics experiment was conducted on the sequencing data of six samples, each labeled with filenames SRR3414629\_1.fastq, SRR3414630\_1.fastq, SRR3414631\_1.fastq, SRR3414635\_1.fastq, SRR3414636\_1.fastq and SRR3414637\_1.fastq respectively. Sample SRR3414629, SRR3414630, and SRR3414631 are replicates and designated as "reprogrammed" based on information from an article titled "TRIM28 Is an Epigenetic Barrier to Induced Pluripotent Stem Cell Reprogramming," from which the data was sourced via NCBI.

Scientist has found a way to reprogram somatic cells back to its pluripotent stem cells state. They discovered a gene TRIM28 as an epigenetic modifier that acts as a barrier to this transition so they knockdown TRIM28 in some of the cell Reprogramming samples. Control replicates have no doxycycline (They are nonprogrammed) and our reprogrammed data has doxycycline. Doxycycline is some sort of antibiotic that is used in cell reprogramming. All samples are from mouse.

Sample SRR3414635, SRR3414636, and SRR3414637 replicates serve as control (Non-reprogrammed). It is important to state Illumina HiSeq 2500 was used for RNA-Seq on transcriptomic cDNA in single-end layout following standard Illumina library preparation protocols.

Quality control using FastQC was first performed on the datasets. Summary statistics details such as file types, encoding, total sequences, sequence length, and GC content are given by the FastQC report below.

## **SRR3414629\_1.fastq:**

- Total Sequences: 21,106,089
- Sequences flagged as poor quality: 0
- Sequence length: 65
- %GC: 49

## **SRR3414630\_1.fastq:**

- Total Sequences: 15,244,711
- Sequences flagged as poor quality: 0
- Sequence length: 65
- %GC: 47

## **SRR3414631\_1.fastq:**

- Total Sequences: 24,244,069
- Sequences flagged as poor quality: 0
- Sequence length: 65
- %GC: 49

#### **SRR3414635\_1.fastq:**

- Total Sequences: 20,956,475
- Sequences flagged as poor quality: 0
- Sequence length: 65
- %GC: 49

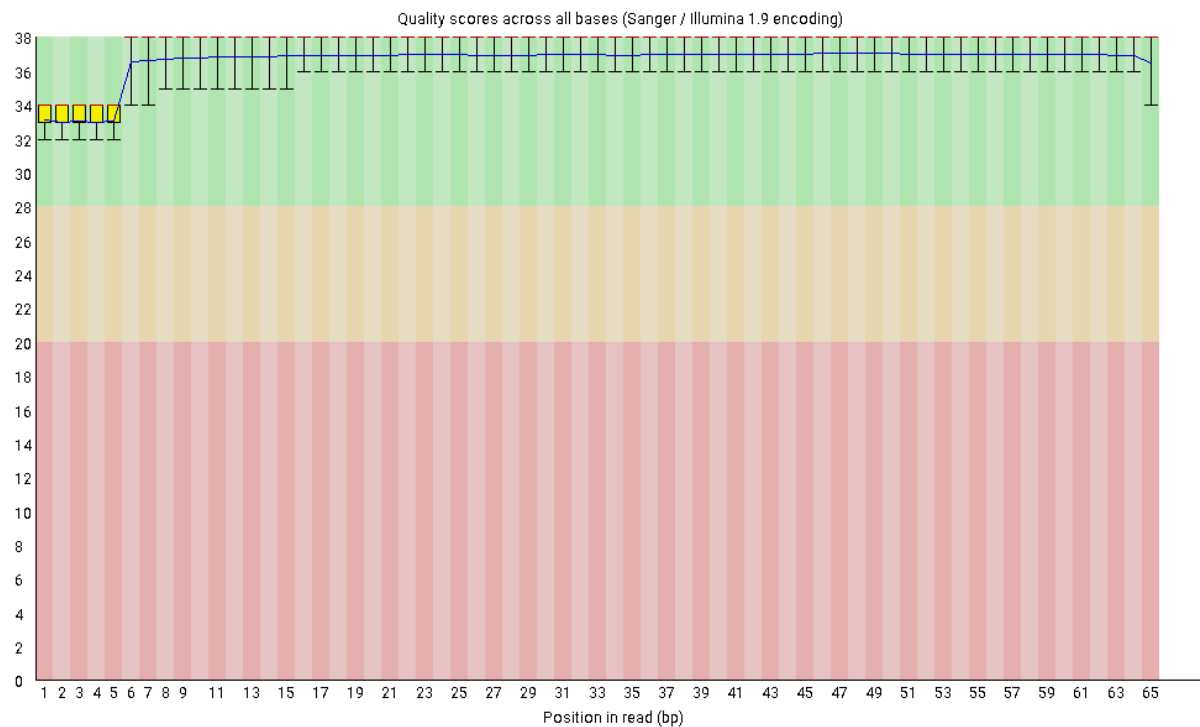
#### **SRR3414636\_1.fastq:**

- Total Sequences: 20,307,147
- Sequences flagged as poor quality: 0
- Sequence length: 65
- %GC: 49

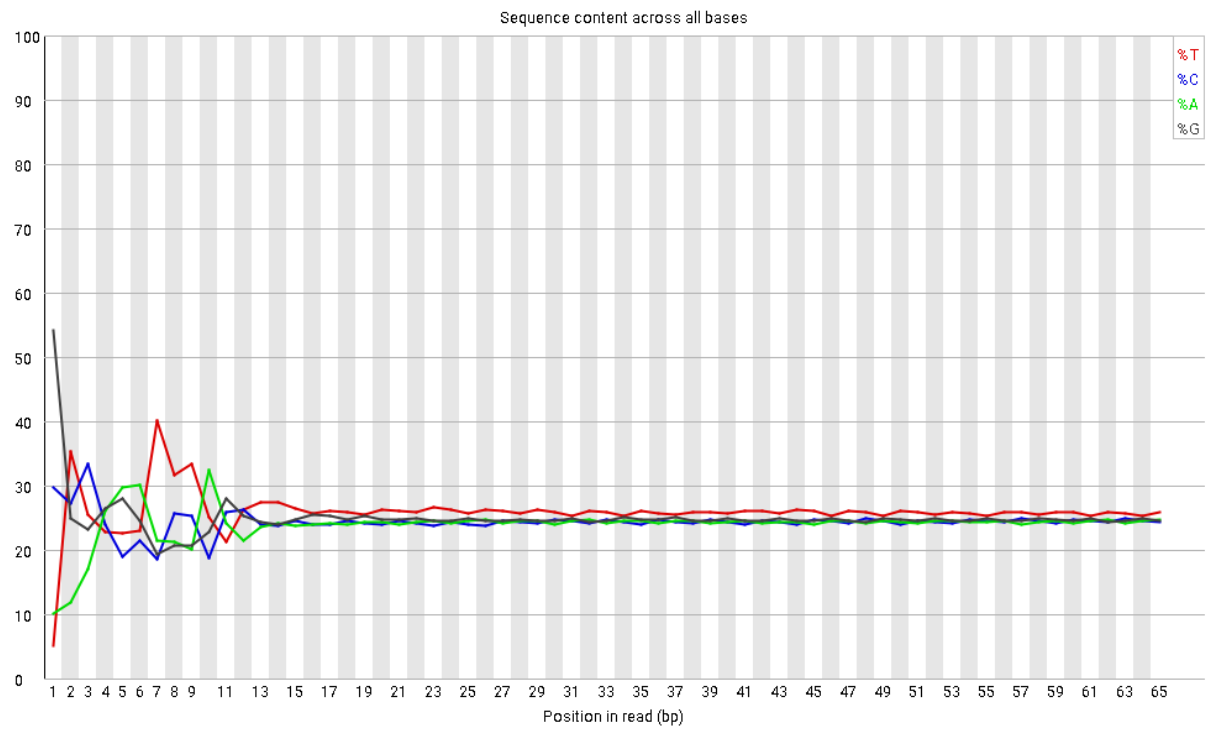
#### **SRR3414637\_1.fastq:**

- Total Sequences: 20,385,570
- Sequences flagged as poor quality: 0
- Sequence length: 65
- %GC: 47

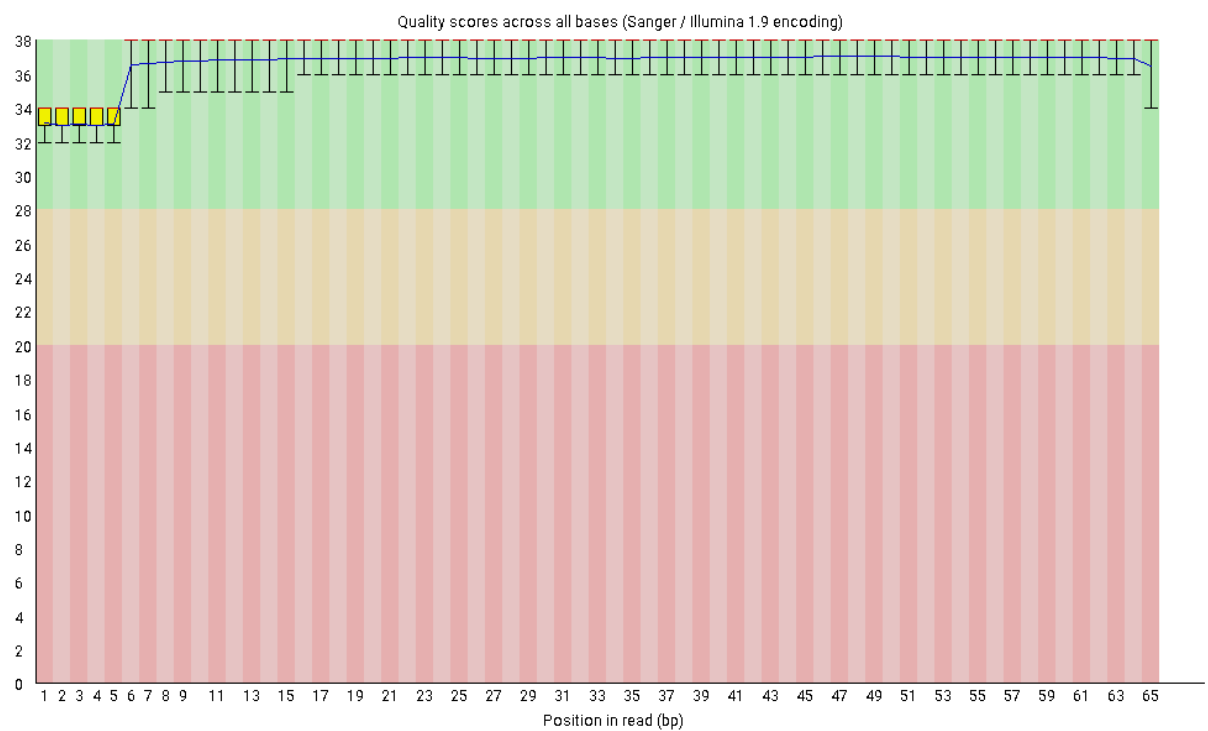
**Fig 1a: Per Base Sequence Quality of SRR3414629**



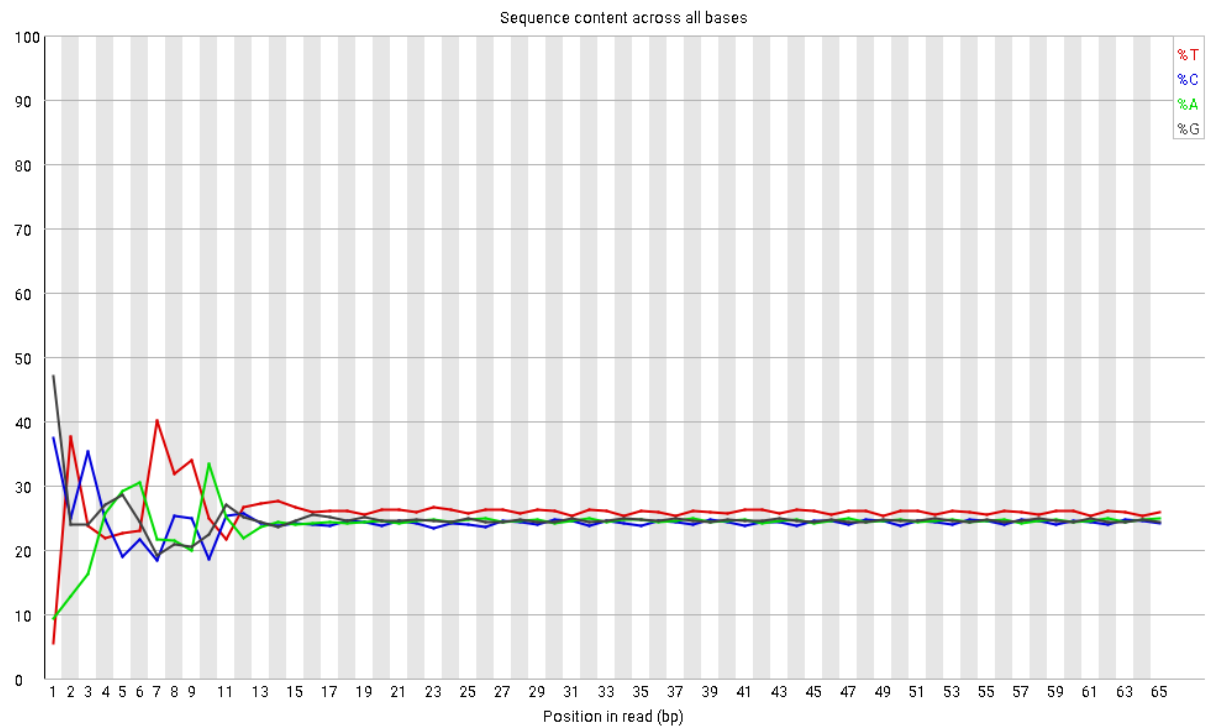
**Fig 1b: Base per Sequence content of SRR3414629**



**Fig 1c: Base per Sequence Quality of SRR3414635**



**Fig 1d: Base per Sequence Content of SRR3414635**



The FastQC report indicated by critical parameters such as sequence length distribution, per-sequence GC content, and adapter content pass the quality assessment, signifying their high quality (see attached files for more). Consequently, read alignment was initiated without requiring trimming using HISAT2.

## READS ALIGNMENT

### HISAT2

An index genome was created using the GRCm39 primary reference genome obtained from the Gencode database. Indexing represent different kmers of the genome and also allows us to align reads even if they are not exact match with reference genome. This index genome facilitates speedy analysis by pre-processing the reference sequences and organizing them in a format optimized for alignment algorithms like HISAT2. The alignment using HISAT2 was done with the following command:

```
hisat2 -p 4 -x indexes/GRCm39 SRR3414635.fastq.gz -S SRR3414635.sam 2> SRR3414635.hisat
```

This command utilized the HISAT2 aligner with the specified parameters (-p for number of threads, -x for the index, and input and output files). The resulting output was a SAM file (SRR3414635.sam) containing the aligned data, and a summary report of the HISAT2 alignment was redirected to a file named SRR3414635.hisat. Subsequently, further processing was conducted to create uniquely mapped reads, sort them, and generate a machine-readable BAM. This .bam file was used in creating the .counts file for subsequent analyses using htseq-count.

This comprehensive RNA-seq pipeline (Fig 2a) streamlined the entire process, ensuring efficient processing of multiple samples, and facilitating further investigations.

**Fig 2a: HISAT2 RNA-seq Pipeline**

```

judeaneke@JudeAneke: ~/DN x + v
j:/bin/bash
# Set timer
SECONDS=0

# pmd = HISAT2

# Build HISAT2 index
hisat2-build /home/Students/M06-310/JudeAneke/HISAT2_indexes/data/GRCm39.genome.fa HISAT2_indexes/GRCm39_idx

# List of input fastq files
FASTQ_FILES=( "SRR3414629_1.fastq" "SRR3414631_1.fastq" "SRR3414635_1.fastq" "SRR3414636_1.fastq" "SRR3414637_1.fastq" )

# Path to Bowtie2 index
HISAT2_INDEX="/home/Students/M06-310/JudeAneke/HISAT2_indexes/data/HISAT2/HISAT2_indexes/GRCm39_idx"

# Loop through each fastq file
for file in "${FASTQ_FILES[@]}; do
    # Extract the sample ID from the filename
    sample_id=$(echo "$file" | cut -d '_' -f 1)

    # Run HISAT2 alignment for each sample
    hisat2 -p 4 -x "$HISAT2_INDEX" -U "/home/Students/M06-310/data/$file" -S "${sample_id}.sam" 2> "${sample_id}.hisat"
    echo "HISAT2 finished running for ${sample_id}!"

    # Filter the alignment file to include only reads with NM:i:1
    grep -P "NM:i:1" "${sample_id}.sam" > "${sample_id}.uniq.sam"
    echo "Uniquely mapped reads created successfully for ${sample_id}!"

    # Run samtools to create sorted bam files
    samtools view -b "${sample_id}.uniq.sam" | samtools sort -o "${sample_id}.uniq.sorted.bam"
    echo "Bam file created and sorted successfully for ${sample_id}!"

    # Create index bam file
    samtools index "${sample_id}.uniq.sorted.bam"
    echo "Bam indexed successfully for ${sample_id}!"

    # Run htseq-count for each sample
    htseq-count --strandedno "${sample_id}.uniq.sorted.bam" /home/Students/M06-310/JudeAneke/HISAT2_indexes/data/gencode.vM34.primary_assembly.annotation.gtf > "${sample_id}.counts"
    echo "Count files created successfully for ${sample_id}!"
done

# Calculate and display the elapsed time
duration=$SECONDS
time_summary="Alignment, filtering, and counting completed in $((duration / 60)) minutes and $((duration % 60)) seconds."

# Echo the time summary to a file
echo "$time_summary" > hisat2_time_summary.txt

# Display the time summary
echo "$time_summary"

# Exit the script
exit
1,1 Top

```

## Reads Alignment Summary of Reprogrammed Samples using HISAT2

### SRR3414629.hisat Report

- 21106089 reads; of these:
- 21106089 (100.00%) were unpaired; of these:
- 236348 (1.12%) aligned 0 times
- 18581373 (88.04%) aligned exactly 1 time
- 2288368 (10.84%) aligned >1 times
- 98.88% overall alignment rate

### SRR3414630.hisat Report

- 15244711 reads; of these:
- 15244711 (100.00%) were unpaired; of these:
- 165838 (1.09%) aligned 0 times
- 13325584 (87.41%) aligned exactly 1 time
- 1753289 (11.50%) aligned >1 times
- 98.91% overall alignment rate

#### **SRR3414631.hisat Report**

- 24244069 reads; of these:
- 24244069 (100.00%) were unpaired; of these:
- 274384 (1.13%) aligned 0 times
- 21169297 (87.32%) aligned exactly 1 time
- 2800388 (11.55%) aligned >1 times
- 98.87% overall alignment rate

#### **Reads Alignment Summary of Control Samples using HISAT2**

#### **SRR3414635.hisat Report**

- 20956475 reads; of these:
- 20956475 (100.00%) were unpaired; of these:
- 236243 (1.13%) aligned 0 times
- 18644700 (88.97%) aligned exactly 1 time
- 2075532 (9.90%) aligned >1 times
- 98.87% overall alignment rate

#### **SRR3414636.hisat Report**

- 20307147 reads; of these:
- 20307147 (100.00%) were unpaired; of these:
- 228180 (1.12%) aligned 0 times
- 18039904 (88.84%) aligned exactly 1 time
- 2039063 (10.04%) aligned >1 times
- 98.88% overall alignment rate

#### **SRR3414637.hisat Report**

- 20385570 reads; of these:
- 20385570 (100.00%) were unpaired; of these:
- 232763 (1.14%) aligned 0 times
- 18049280 (88.54%) aligned exactly 1 time
- 2103527 (10.32%) aligned >1 times
- 98.86% overall alignment rate

**Table 1: Alignment Summary**

<b>Reads Alignment</b>	<b>SRR3414629</b>	<b>SRR3414630</b>	<b>SRR3414631</b>	<b>SRR3414635</b>	<b>SRR3414636</b>	<b>SRR3414637</b>
Number of mapped reads:	18,581,373	13,325,584	21,169,297	18,644,700	18,039,904	18,049,280
Total number of alignments:	18,581,373	13,325,584	21,169,297	18,644,700	18,039,904	18,049,280
Number of non-unique alignments:	0	0	0	0	0	0
Aligned to genes:	16,152,993	11,528,784	18,536,741	16,399,422	15,875,862	15,855,506
Ambiguous alignments:	513,717	341,465	581,235	547,646	528,975	516,666
No feature assigned:	1,911,210	1,452,629	2,047,186	1,693,947	1,631,404	1,672,722
Missing chromosome in annotation:	3,453	2,706	4,135	3,685	3,663	4,386
Not aligned:	0	0	0	0	0	0
<b>Reads Genomic Origin</b>						
Exonic:	16,152,993 / 89.42%	11,528,784 / 88.81%	18,536,741 / 90.05%	16,399,422 / 90.64%	15,875,862 / 90.68%	15,855,506 / 90.46%
Intronic:	1,666,146 / 9.22%	1,258,690 / 9.7%	1,763,365 / 8.57%	1,484,108 / 8.2%	1,425,995 / 8.15%	1,451,329 / 8.28%
Intergenic:	245,064 / 1.36%	193,939 / 1.49%	283,821 / 1.38%	209,839 / 1.16%	205,409 / 1.17%	221,393 / 1.26%
Intronic/intergenic overlapping exon:	420,357 / 2.33%	289,718 / 2.23%	474,950 / 2.31%	421,627 / 2.33%	413,691 / 2.36%	390,537 / 2.23%

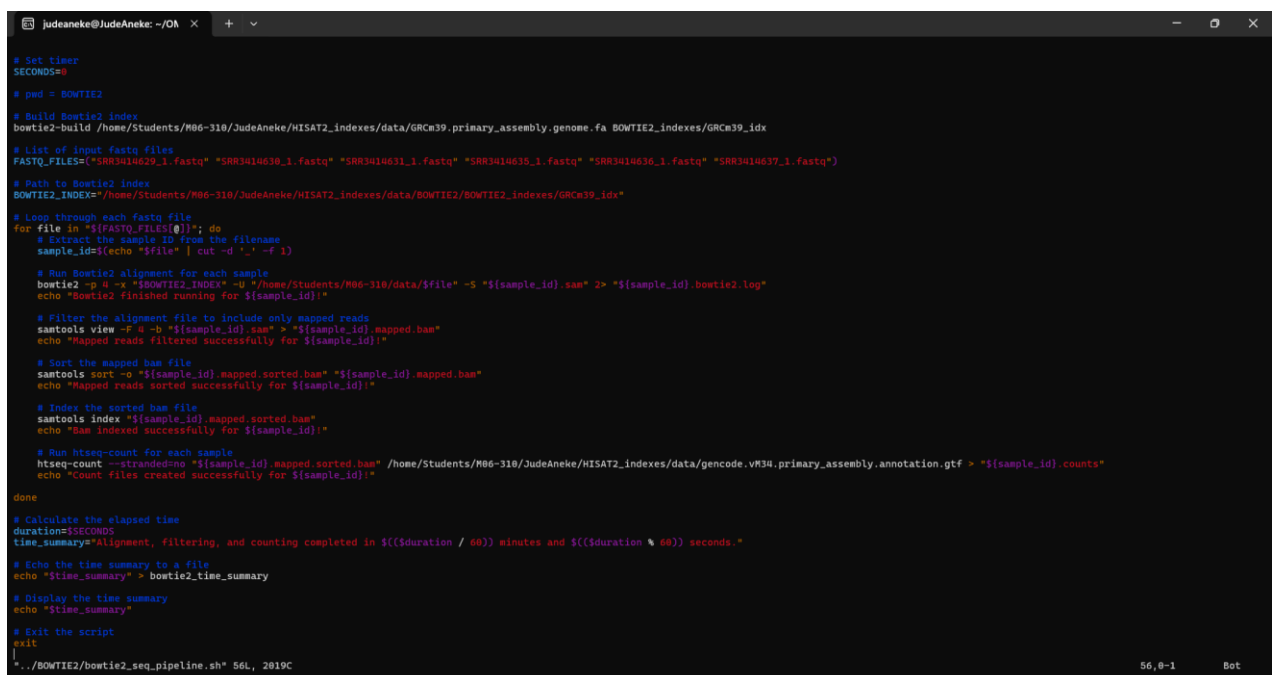
## BOWTIE2

A similar process was conducted utilizing the BOWTIE2 alignment algorithm to align sequencing reads against the GRCm39 primary reference genome from the Gencode database. Bowtie2 was used to build index file using the same reference genome. The alignment was executed with the BOWTIE2 command:

```
bowtie2 -p 4 -x indexes/GRCm39_idx -U /home/Students/M06-310/data/SRR3414635 -S SRR3414635.sam 2> SRR3414635.bowtie2.log
```

Specific parameters (-x for index reference, input and output files) were used to ensure precise alignment. The resulting SAM file (SRR3414635.sam) contained the aligned data, accompanied by a summary report redirected to SRR3414635.bowtie2.log. Subsequent steps included filtering for uniquely mapped reads, sorting, creating .bam and .counts files (Fig 2b).

**Fig 2b: BOWTIE2 RNA-Seq Pipeline**



```
judeaneke@JudeAneke: ~/OH
# Set timer
SECONDS=0

# pwd = BOWTIE2

# Build Bowtie2 index
bowtie2-build /home/Students/M06-310/JudeAneke/HISAT2_indexes/data/GRCm39.primary_assembly.genome.fa BOWTIE2_indexes/GRCm39_idx

# List of input fastq files
FASTQ_FILES=("SRR3414629_1.fastq" "SRR3414630_1.fastq" "SRR3414631_1.fastq" "SRR3414635_1.fastq" "SRR3414636_1.fastq" "SRR3414637_1.fastq")

# Path to Bowtie2 index
BOWTIE2_INDEX="/home/Students/M06-310/JudeAneke/HISAT2_indexes/data/BOWTIE2/BOWTIE2_indexes/GRCm39_idx"

# Loop through each fastq file
for file in "${FASTQ_FILES[@]}; do
    # Extract the sample ID from the filename
    sample_id=$(echo "$file" | cut -d '_' -f 1)

    # Run Bowtie2 alignment for each sample
    bowtie2 -p 4 -x "$BOWTIE2_INDEX" -U "/home/Students/M06-310/data/$file" -S "${sample_id}.sam" 2> "${sample_id}.bowtie2.log"
    echo "Bowtie2 finished running for ${sample_id}"

    # Filter the alignment file to include only mapped reads
    samtools view -F 4 -b "${sample_id}.sam" > "${sample_id}.mapped.bam"
    echo "Mapped reads filtered successfully for ${sample_id}"

    # Sort the mapped bam file
    samtools sort -o "${sample_id}.mapped.sorted.bam" "${sample_id}.mapped.bam"
    echo "Mapped reads sorted successfully for ${sample_id}"

    # Index the sorted bam file
    samtools index "${sample_id}.mapped.sorted.bam"
    echo "Bam indexed successfully for ${sample_id}"

    # Run htseq-count for each sample
    htseq-count --stranded=no "${sample_id}.mapped.sorted.bam" /home/Students/M06-310/JudeAneke/HISAT2_indexes/data/gencode.vR34.primary_assembly.annotation.gtf > "${sample_id}.counts"
    echo "Count files created successfully for ${sample_id}"
done

# Calculate the elapsed time
duration=$SECONDS
time_summary="Alignment, filtering, and counting completed in $((duration / 60)) minutes and $((duration % 60)) seconds."

# Echo the time summary to a file
echo "$time_summary" > bowtie2_time_summary

# Display the time summary
echo "$time_summary"

# Exit the script
exit

~/BOWTIE2/bowtie2_seq_pipeline.sh 56L, 2019C 56,0-1 Bot
```

## Reads Alignment Summary of Reprogrammed Samples using BOWTIE2

### SRR3414629.bowtie2.log Report

- 21106089 reads; of these:
- 21106089 (100.00%) were unpaired; of these:
- 2194257 (10.40%) aligned 0 times
- 13916350 (65.94%) aligned exactly 1 time



- 4995482 (23.67%) aligned >1 times
- 89.60% overall alignment rate

#### **SRR3414630.bowtie2.log Report**

- 15244711 reads; of these:
- 15244711 (100.00%) were unpaired; of these:
- 1492956 (9.79%) aligned 0 times
- 9958511 (65.32%) aligned exactly 1 time
- 3793244 (24.88%) aligned >1 times
- 90.21% overall alignment rate

#### **SRR3414631.bowtie2.log Report**

- 24244069 reads; of these:
- 24244069 (100.00%) were unpaired; of these:
- 2454317 (10.12%) aligned 0 times
- 15713965 (64.82%) aligned exactly 1 time
- 6075787 (25.06%) aligned >1 times
- 89.88% overall alignment rate

### **Reads Alignment Summary of Control Samples using BOWTIE2**

#### **SRR3414635.bowtie2.log Report**

- 20956475 reads; of these:
- 20956475 (100.00%) were unpaired; of these:
- 2229346 (10.64%) aligned 0 times
- 14146972 (67.51%) aligned exactly 1 time
- 4580157 (21.86%) aligned >1 times
- 89.36% overall alignment rate

#### **SRR3414636.bowtie2.log Report**

- 20307147 reads; of these:
- 20307147 (100.00%) were unpaired; of these:
- 2138601 (10.53%) aligned 0 times
- 13669490 (67.31%) aligned exactly 1 time
- 4499056 (22.16%) aligned >1 times
- 89.47% overall alignment rate

#### **SRR3414637.bowtie2.log Report**

- 20385570 reads; of these:
- 20385570 (100.00%) were unpaired; of these:
- 2030584 (9.96%) aligned 0 times
- 13747478 (67.44%) aligned exactly 1 time
- 4607508 (22.60%) aligned >1 times
- 90.04% overall alignment rate

## **ALIGNMENT QUALITY CONTROL**

Before proceeding with our Alignment QC, we filtered out reads with multiple alignments leaving only uniquely mapped reads for further analysis. I used this code to filter:

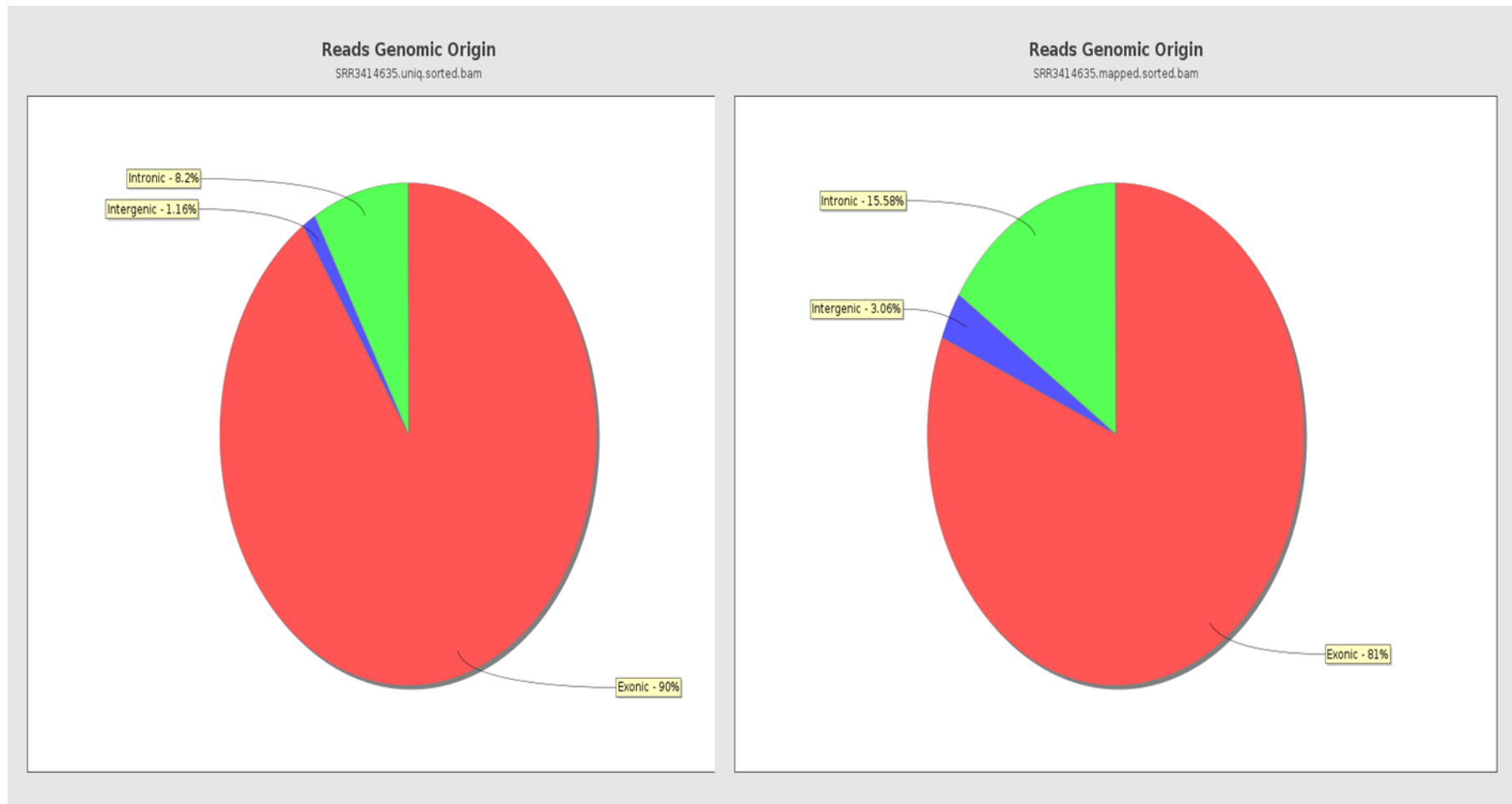
```
grep -P '^@/NH:i:1$' SRR3414636.sam > SRR3414636.uniq.sam
```

To further compare the performance of HISAT2 and BOWTIE2, we analyzed their alignment results using various metrics. For example, we examined the alignment rate, the distribution of reads across genomic features (such as exons, introns, and intergenic regions), and the quality of the alignment.

### **Comparison between HISAT2 and BOWTIE2 Alignment**

Alignment quality control (QC) involved a comparative analysis between the alignment rates achieved by the HISAT2 and BOWTIE2 algorithms. QualiMap, a tool commonly used for assessing sequencing alignment quality, was employed to evaluate and compare the sequencing maps generated by HISAT2 and BOWTIE2. This comparative analysis aimed to provide insights into the performance and accuracy of both alignment algorithms.

**Fig 7: Percentage of Mapped Exon**



>>>>>> Reads genomic origin (HISAT2 left)

exonic = 16,399,422 (90.64%)

intronic = 1,484,108 (8.2%)

intergenic = 209,839 (1.16%)

overlapping exon = 421,627 (2.33%)

>>>>>> Reads genomic origin (BOWTIE2 right)

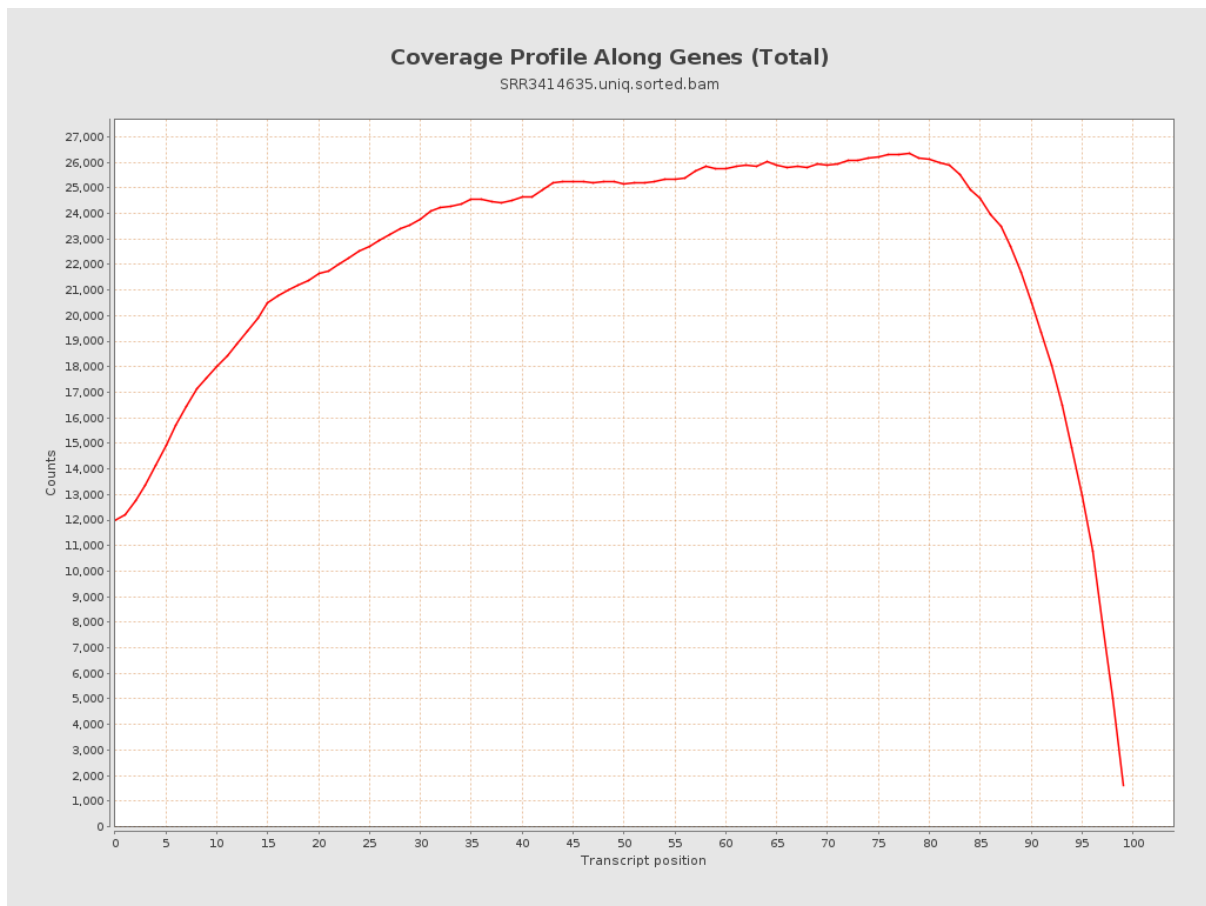
exonic = 14,753,937 (81.37%)

intronic = 2,824,659 (15.58%)

intergenic = 554,347 (3.06%)

overlapping exon = 1,409,371 (7.77%)

**Fig 3a: SRR3414635 Total Coverage Profile Along Genes (HISAT2)**



>>>>>> Reads alignment (HISAT2)

reads aligned = 18,644,700

total alignments = 18,644,700

secondary alignments = 0

non-unique alignments = 0

aligned to genes = 16,399,422

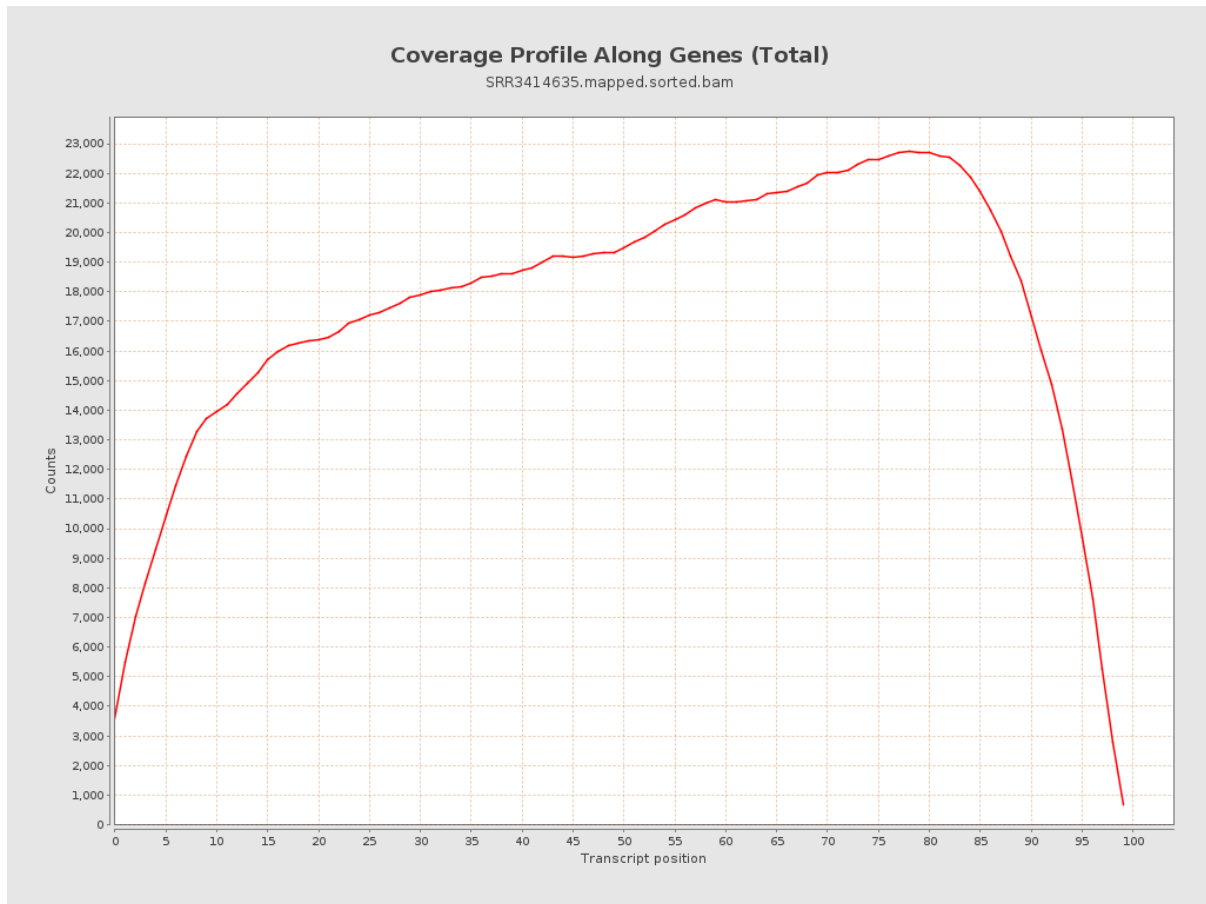
ambiguous alignments = 547,646

no feature assigned = 1,693,947

not aligned = 0

SSP estimation (fwd/rev) = 0.02 / 0.98

**Fig 3b: SRR3414635 Total Coverage Profile Along Genes (BOWTIE2)**



>>>>>> Reads alignment (BOWTIE2)

reads aligned = 18,727,129

total alignments = 18,727,129

secondary alignments = 0

non-unique alignments = 0

aligned to genes = 14,753,937

ambiguous alignments = 586,837

no feature assigned = 3,379,006

not aligned = 0

There are notable differences in the alignment results obtained using BOWTIE2 and HISAT2 as showed in Fig 3a, 3b & 3c for the considered sample SRR3414635. For BOWTIE2, out of 18,727,129 reads aligned, 14,753,937 (81.37%) were aligned to genes, with 586,837 reads having ambiguous alignments. In terms of genomic origin, the majority of aligned reads (81.37%) were exonic, followed by intronic (15.58%), overlapping exon (7.77%), and intergenic (3.06%) regions (Fig 3a).

On the other hand, for HISAT2, out of 18,644,700 reads aligned, 16,399,422 (90.64%) were aligned to genes, with 547,646 reads showing ambiguous alignments. The distribution of aligned reads based on genomic origin showed that the majority (90.64%) were exonic, followed by intronic (8.2%), overlapping exon (2.33%), and intergenic (1.16%) regions (Fig 3a).

The Bowtie2 alignment, filtering, and counting process took approximately 351 minutes and 20 seconds to complete, which is faster compared to the approximately 421 minutes it took for the HISAT2 process.

These differences suggest variations in the alignment speed, efficiency and specificity between the two algorithms, potentially influencing downstream analyses and interpretations of the sequencing data. So, BOWTIE2 is faster while HISAT2 is more efficient in dealing with introns.

VISUALIZATION

Fig 4a: IGV Visualization showing Reads and coverage track

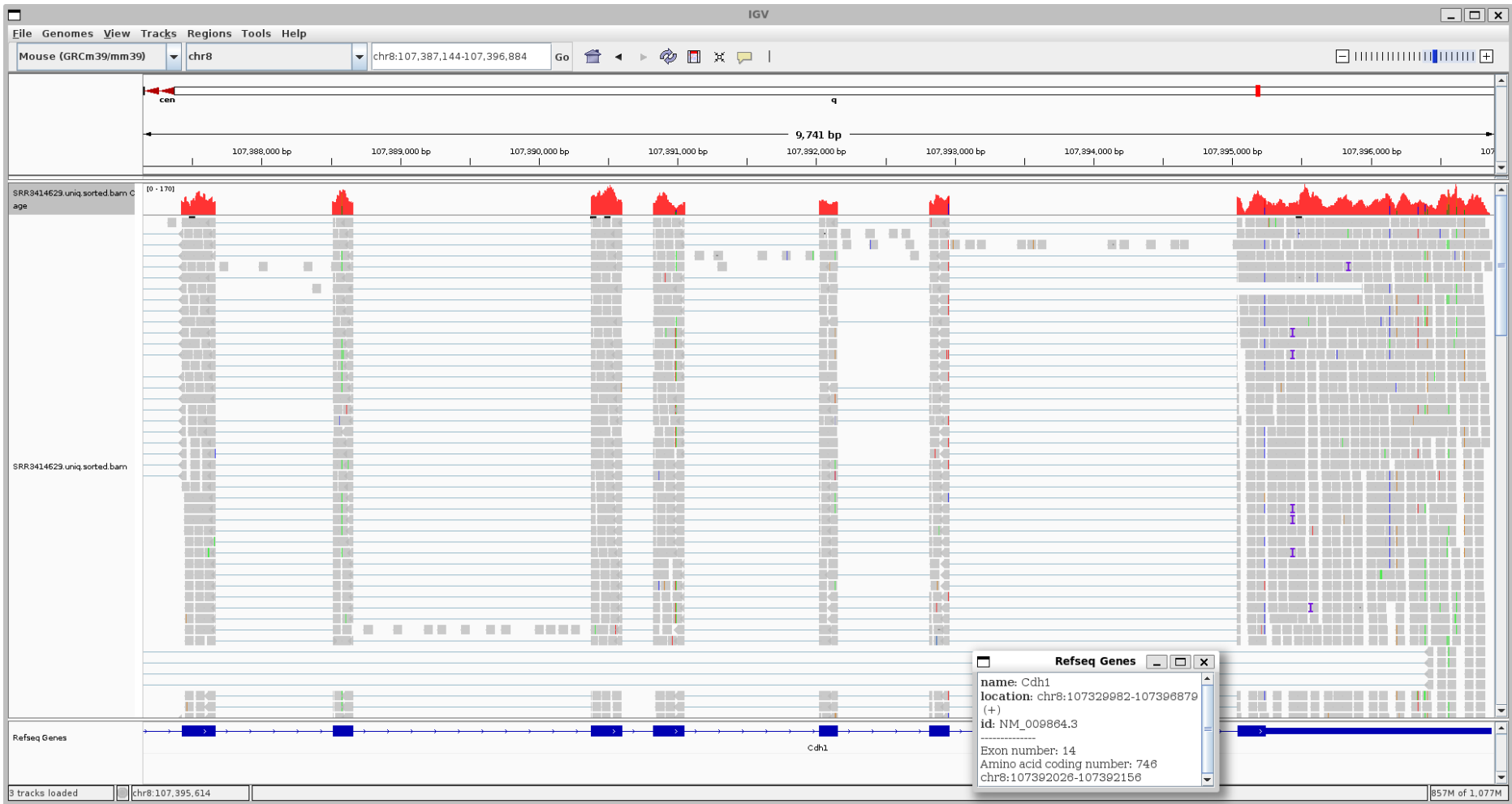


Fig 4b: IGV Visualization showing Line Plots of Reads using Bedgraph.tdf

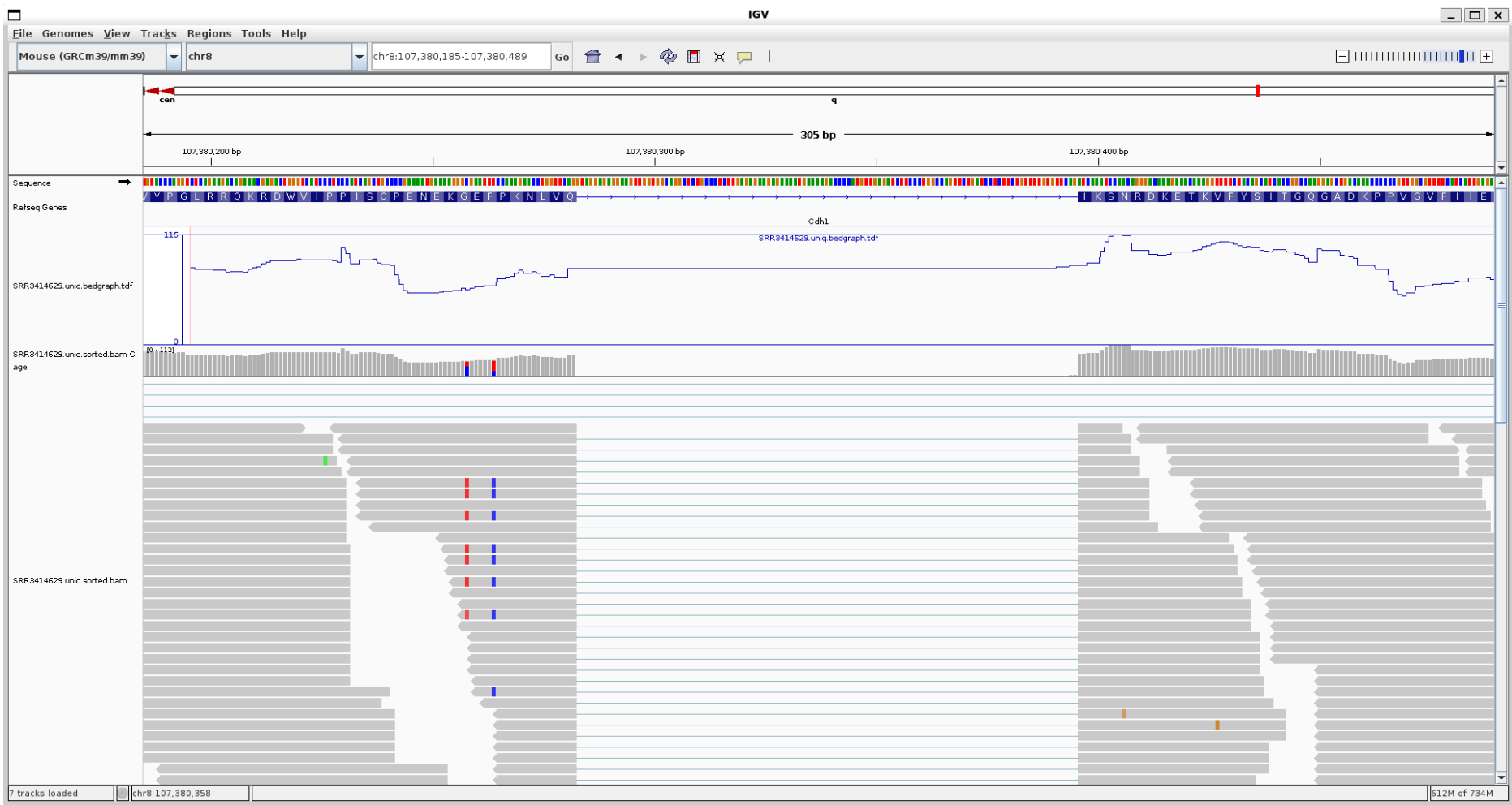
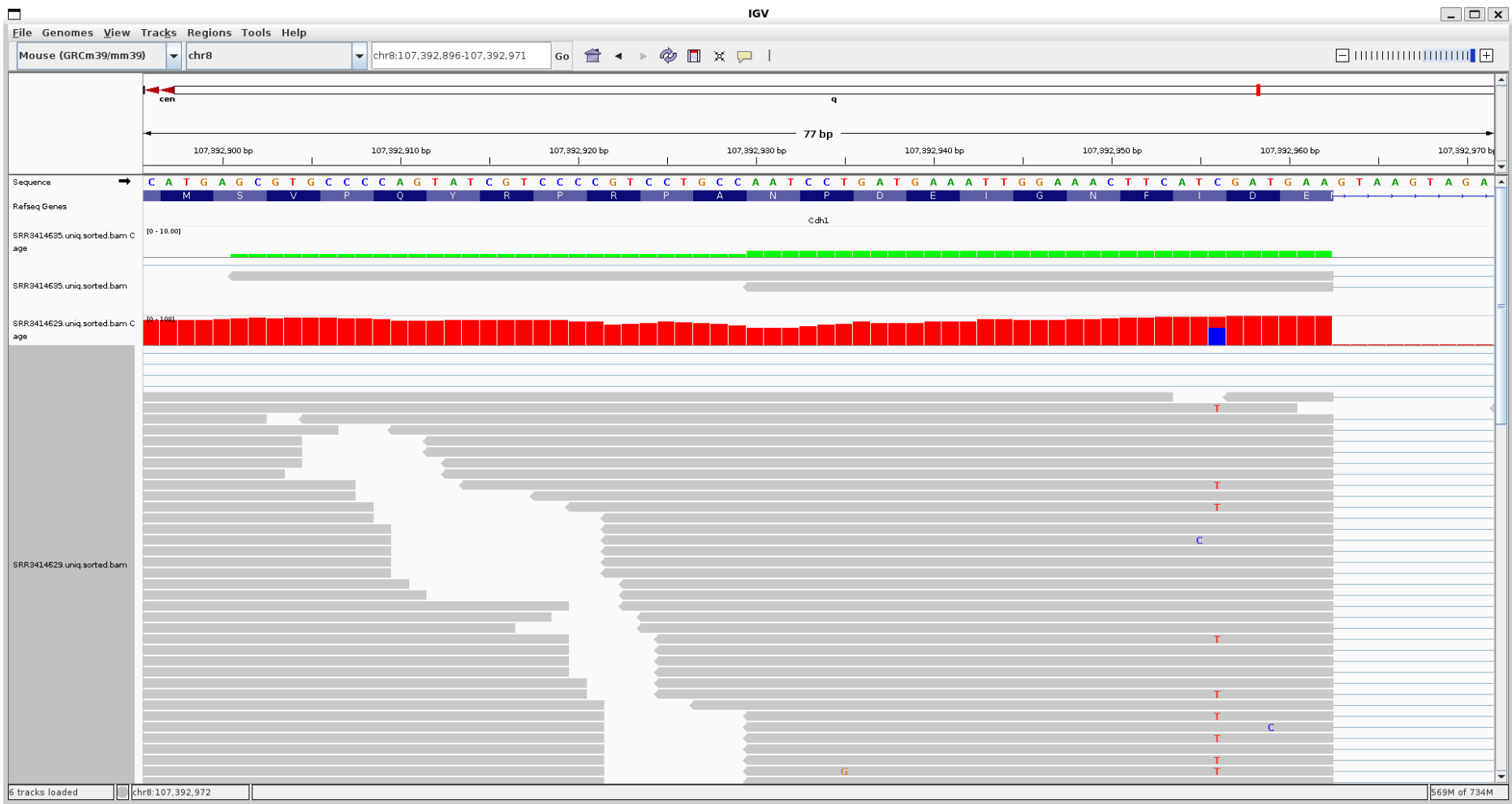




Fig 4c: IGV Visualization showing Coverage Track of two Samples

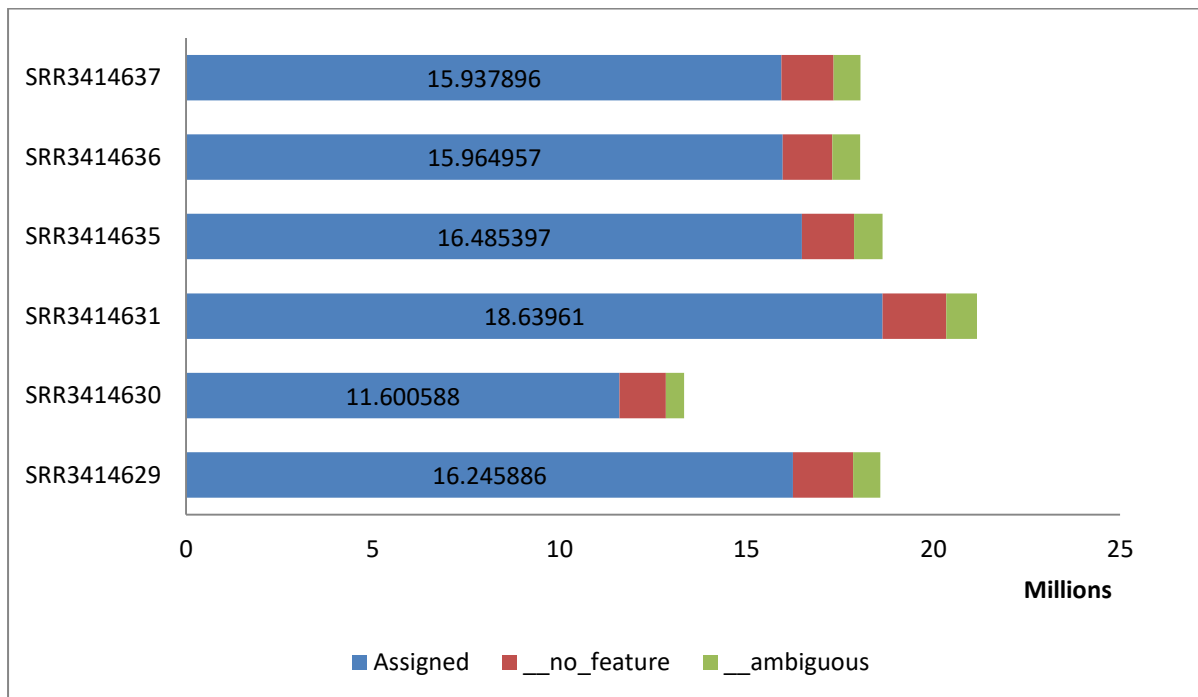


In Fig 4a, a screenshot from IGV showcases the *Cdh1* gene (chr8:107,387,144-107,396,884) of the reprogrammed sample (SRR3414629), revealing the reads from the alignment and highlighting various exons and introns. The reads are well aligned, as indicated by minimal alignment reads within the introns. The read coverage, depicted in red, illustrates the alignment counts or coverage of each nucleotide, represented as a bar chart. Further zooming into Exon number 14 and the sequence track displays the first amino acid coded for in the exon. Optionally, a 3-band track can be revealed by clicking on exon showing a 3-frame translation of the amino acid sequence corresponding to the nucleotide sequence.

Conversion of the SRR3414629.bedgraph file to SRR3414629.bedgraph.tdf using igvtools facilitates visualization in IGV, with the TDF file serving as a compressed binary format similar to .bigWig files used in genome browsers (Fig 4b). Moreover, IGV allows visualization of reads counts from two different samples simultaneously, revealing the expression counts. This was compared between SRR3414629 and the control SRR3414635 showing a higher expression level of the *Cdh1* gene in the reprogrammed sample compared to the control (Fig 4c).

## COUNTING THE NUMBER OF READS

**Fig 5: HTSeq: Counts Assignments**



### HTSeq: SRR3414629.counts

- \_\_no\_feature 1607490
- \_\_ambiguous 727997
- \_\_too\_low\_aQual 0
- \_\_not\_aligned 0
- \_\_alignment\_not\_unique 0

### HTSeq: SRR3414630.counts

- \_\_no\_feature 1241225
- \_\_ambiguous 483771
- \_\_too\_low\_aQual 0
- \_\_not\_aligned 0
- \_\_alignment\_not\_unique 0

### HTSeq: SRR3414631.counts

- \_\_no\_feature 1703434
- \_\_ambiguous 826253
- \_\_too\_low\_aQual 0
- \_\_not\_aligned 0
- \_\_alignment\_not\_unique 0

**HTSeq: SRR3414635.counts**

- \_\_no\_feature 1392968
- \_\_ambiguous 766335
- \_\_too\_low\_aQual 0
- \_\_not\_aligned 0
- \_\_alignment\_not\_unique 0

**HTSeq: SRR3414636.counts**

- \_\_no\_feature 1333663
- \_\_ambiguous 741284
- \_\_too\_low\_aQual 0
- \_\_not\_aligned 0
- \_\_alignment\_not\_unique 0

**HTSeq: SRR3414637.counts**

- \_\_no\_feature 1395438
- \_\_ambiguous 715946
- \_\_too\_low\_aQual 0
- \_\_not\_aligned 0
- \_\_alignment\_not\_unique 0

Fig 5 shows the total number of counts (aligned reads) of each sample. This value is calculated by:

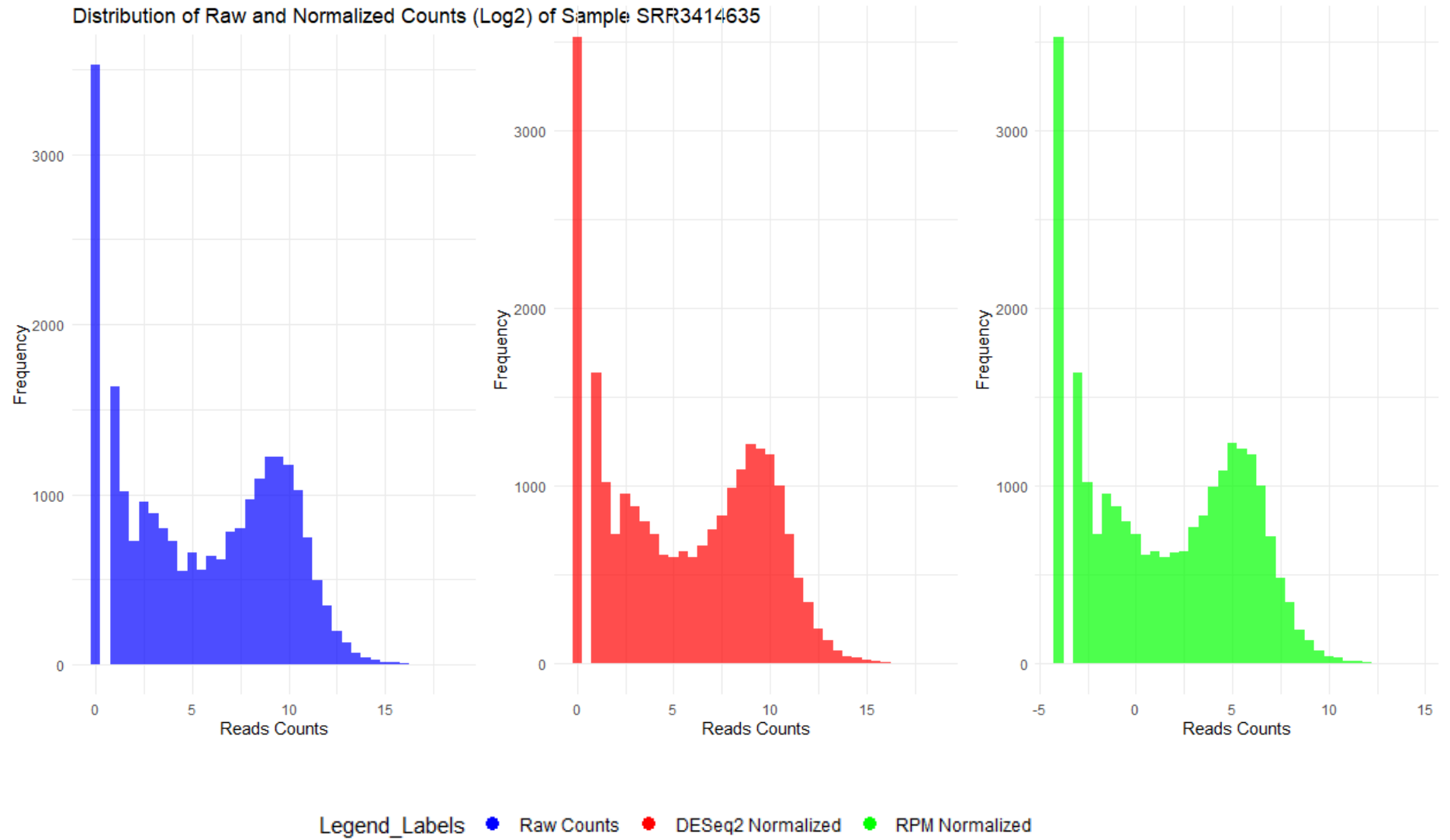
Assigned reads (counts) = Total uniquely mapped reads – (\_\_no\_feature reads) – (\_\_ambiguous)

\*Assigned reads and counts are used interchangeably.

**NORMALIZATION**

Normalization in RNA-seq refers to the process of adjusting gene expression measurements to account for differences in library size, sequencing depth, and other technical factors across samples, enabling meaningful comparison between them. To show the effect of normalization we will consider the assigned reads of a gene (Cdh1) before and after normalization using the RPM and RPKM normalization.

**Fig 6a: Histogram of the Distribution of Raw and Normalized Log2 Counts of SRR3414635**



## Counts of selected gene Normalized with RPM and RPKM

Total counts = No. of uniquely mapped reads – \_\_no\_feature - \_\_ambiguous

For example the Total counts of SRR3414629 is given by:

Total mapped Reads = 18581373 – 1607490 – 727997 = 16245886

RPM and RPKM Normalization of ENSMUSG00000000303.14 (Cdh1)

Gene Length = 66895 (chr 8: 107,329,983 - 107,396,878)

RPM = counts/Total Reads(million)

RPKM (Reads Per Kilobase Million) = RPM/Gene Length(kb)

Gene	Number of Counts per samples						Gene Length (kb)
	SRR3414629	SRR3414630	SRR3414631	SRR3414635	SRR3414636	SRR3414637	
ENSMUSG00000000303.14 (Cdh1)	5181	4414	7692	94	99	99	66.895
Total Counts (million)	16.245886	11.600588	18.639610	16.485397	15.964957	15.937896	

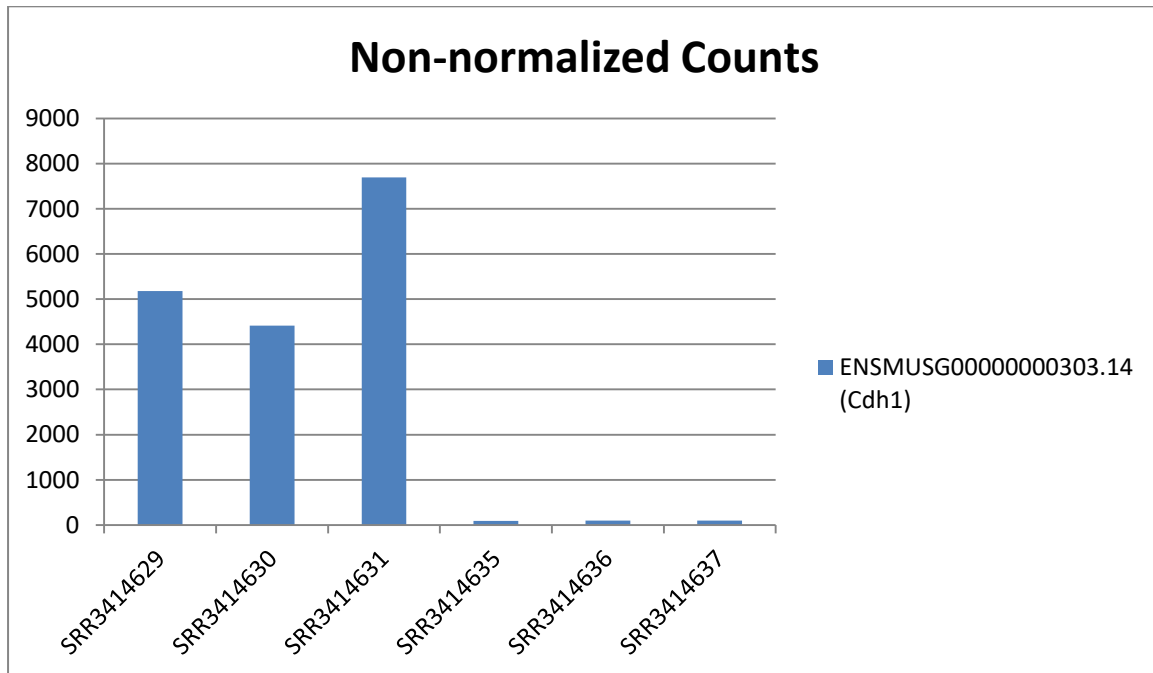
We divide the read counts per gene by the total mapped to each sample.

Gene	RPM (using the per million scale factor) Normalize for read depth					
ENSMUSG00000000303.14 (Cdh1)	318.9115078	380.498	412.66958	5.7020162	6.201082	6.21161

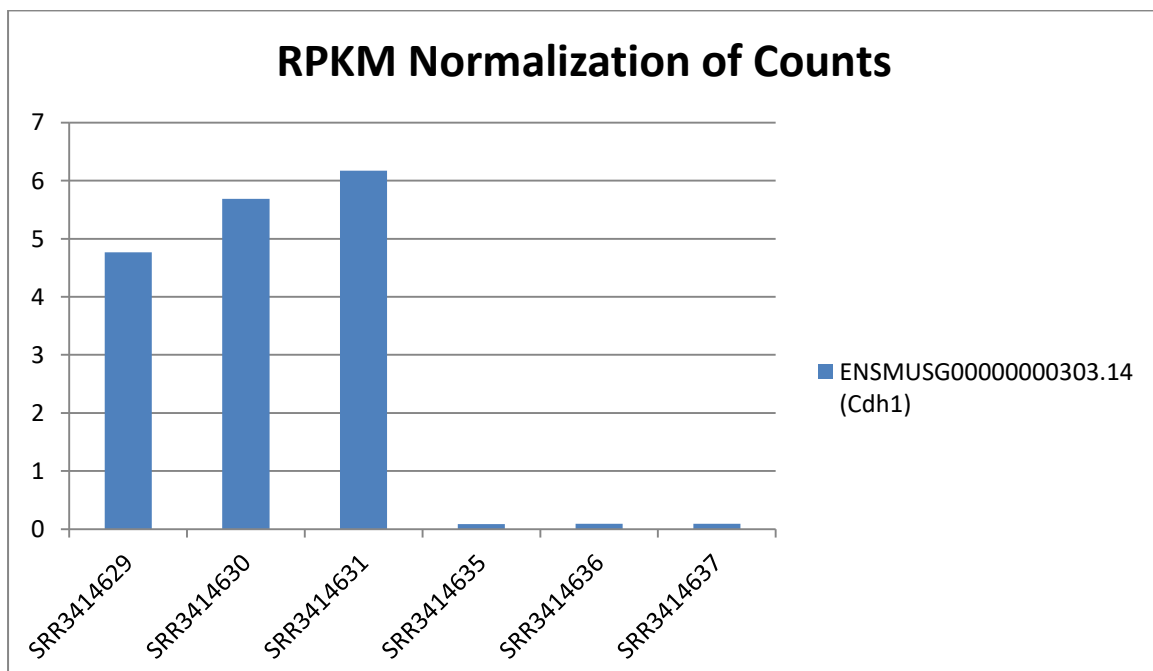
RPKM further normalizes for gene length.

Gene	RPKM (using per kb scale factor) Normalize for gene length					
ENSMUSG00000000303.14 (Cdh1)	4.7673444	5.68798	6.16891	0.08523	0.09269	0.09285
	62	8	51	83	9	6

**Fig 6b: Non-normalized Counts of Gene**



**Fig 6c: RPKM Normalized Counts of Gene**



**Fig 6d: DESeq2 Normalized Counts of Gene**

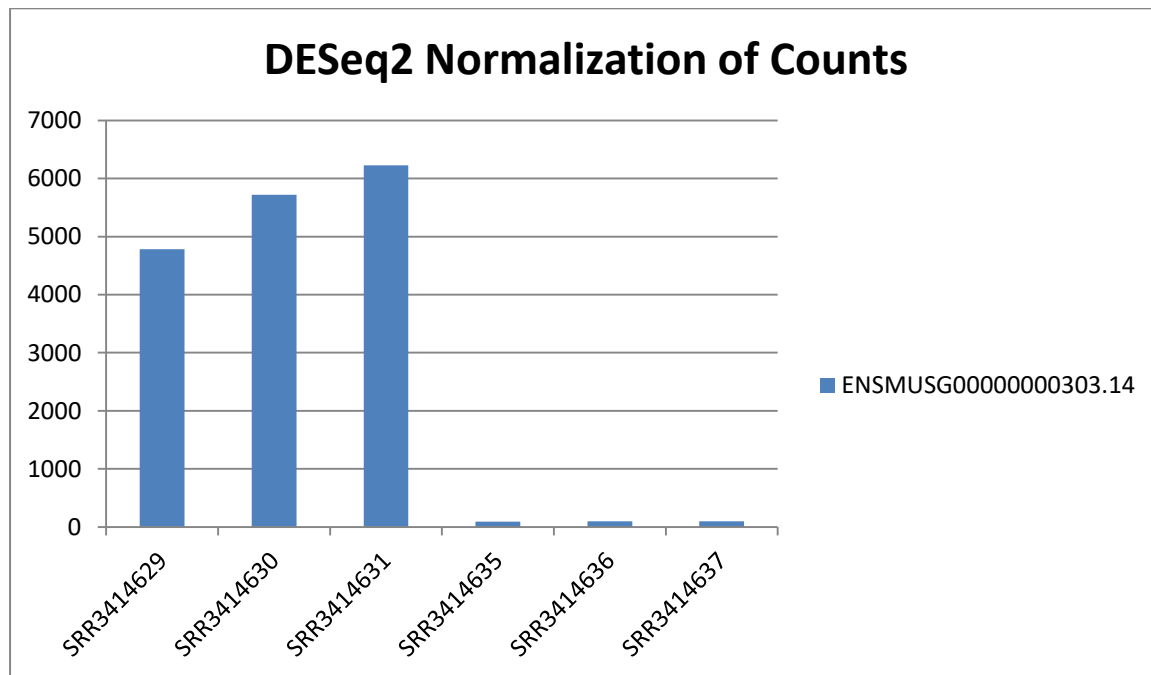


Fig 6a shows the distribution of raw counts, DESeq2 normalization and RPM normalization in Log2. The difference can be seen clearly on the range of the counts in x-axis. To further illustrate this difference we manually calculated the RPKM value of Chd1 gene. The counts y-axis shows the effect of various normalization methods on the counts of chd1.

Looking at the non-normalize bar chart, in Fig 6b, one would conclude that the difference between the gene expression of the gene in the control is about 1 count for every 8000 counts in the reprogrammed sample. While the gene expression of the reprogrammed sample is obviously high than that of the control the level of difference is misleading. This difference is not due to biology but to sequencing depth. This effect is similar to the effect of scaling. The RPKM normalizes for depth of reads and length of genes (Fig 6c). This corrects for biases in sequencing length and gene length.

There are other tools that can be used to normalize expression counts more effectively. DESeq2 normalizes and checks for genes that are differentially expressed based on a given threshold. From the above plot, we can see normalization using DESeq2 (Fig 6d).



## DIFFERENTIAL EXPRESSION OF GENES

I used Deseq2 in r and a threshold of  $\text{padj} < 0.001$  for this experiment. We started our DESeq analysis by visualizing the raw counts and this showed great disparity even within replicates of same condition.

**Fig 7a: Non-Normalized and Normalize Counts of Samples**

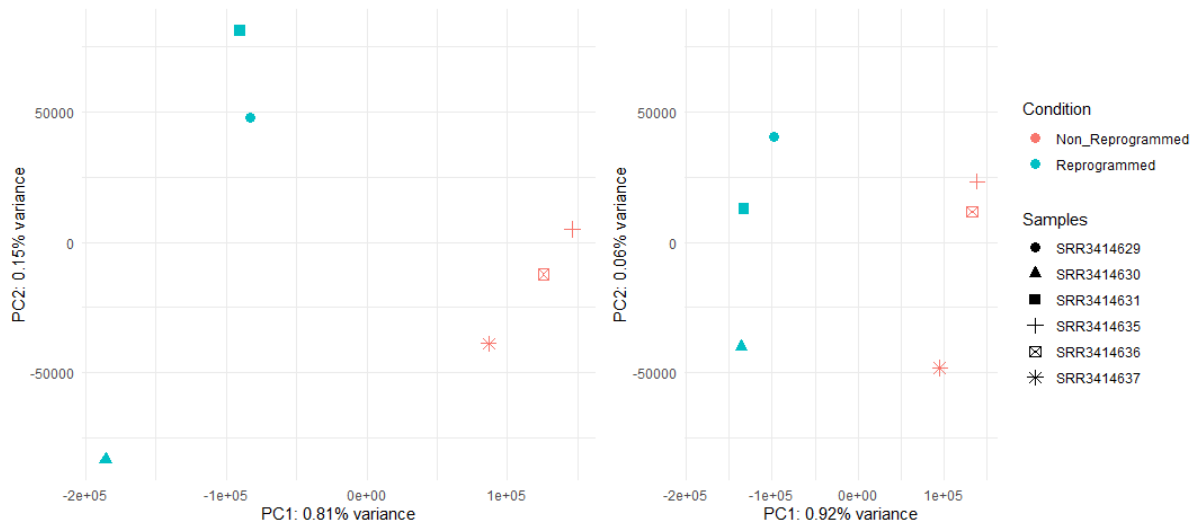
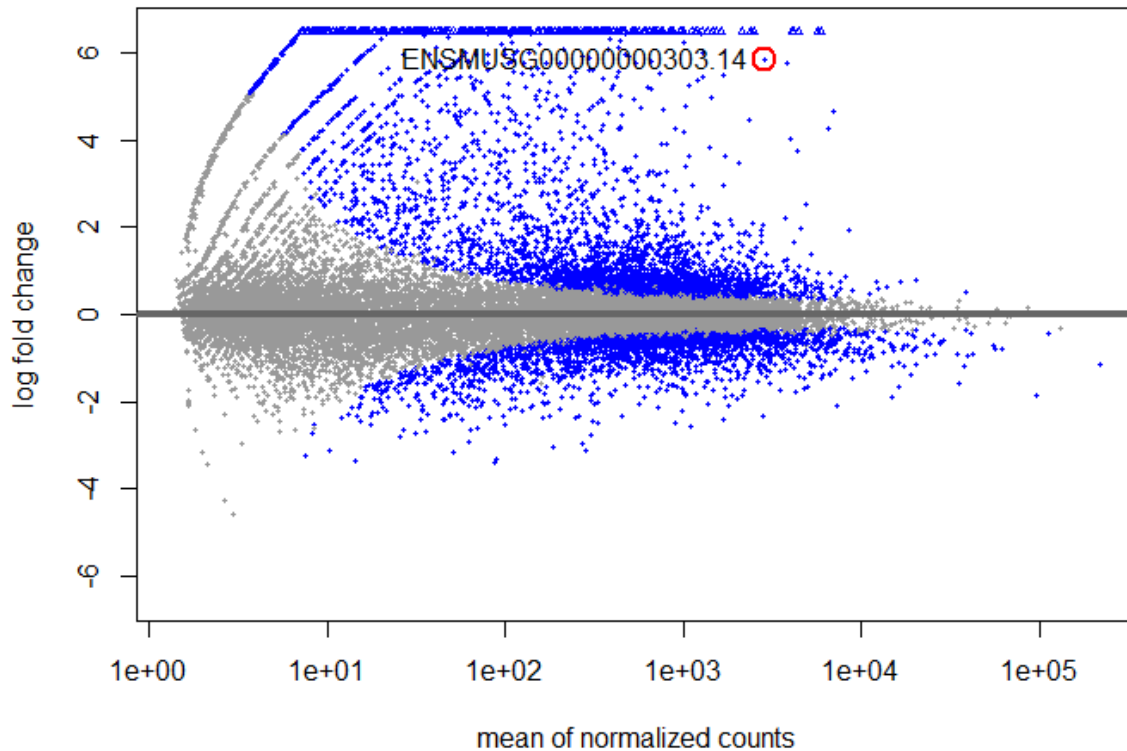


Fig 7a shows the visualization of the counts of each sample before and after DESeq2 normalization. The normalized samples are clearly seen to be within closer range compared to the raw counts.

**Fig 7b: MA Plot showing LogFC of DE Genes**



In DESeq2, the function *plotMA* shows the log2 fold changes attributable to a given variable over the mean of normalized counts for all the samples in the *DESeqDataSet*. Points will be colored blue if the adjusted *p* value is less than 0.001. Points which fall out of the window are plotted as open triangles pointing either up or down. You can see our gene of interest location indicated by the colour red shows it is highly differentially expressed.

#### Differential Analysis Summary

*out of 20933 with nonzero total read count*

*adjusted p-value < 0.001*

*LFC > 0 (up) : 3191, 15%*

*LFC < 0 (down) : 2219, 11%*

*outliers [1] : 3, 0.014%*

low counts [2] : 1624, 7.8%

(mean count < 3)

[1] see 'cooksCutoff' argument of ?results

[2] see 'independentFiltering' argument of ?results

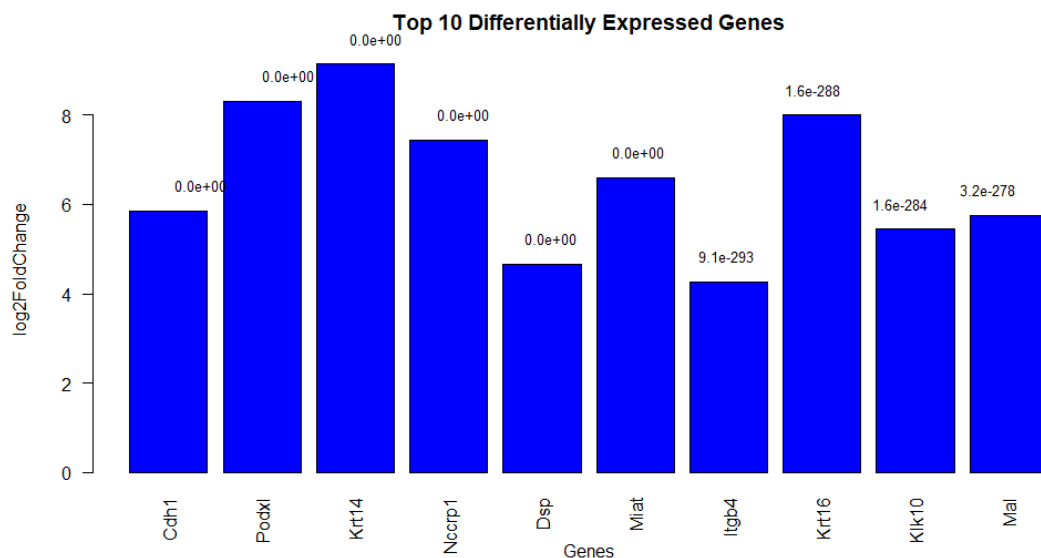
**Table 2: List of Differentially Expressed Genes (padj < 0.001)**

Gene_ID	log2FoldChange	Gene_Annotation	padj (< 0.001)
ENSMUSG00000000303.14	5.847092164	Cdh1	0.0E+00
ENSMUSG000000025608.10	8.3058837	Podxl	0.0E+00
ENSMUSG000000045545.9	9.13790358	Krt14	0.0E+00
ENSMUSG000000047586.5	7.430686036	Nccrp1	0.0E+00
ENSMUSG000000054889.11	4.654156309	Dsp	0.0E+00
ENSMUSG000000097767.10	6.595314866	Miat	0.0E+00
ENSMUSG000000020758.16	4.270336649	Itgb4	9.1E-293
ENSMUSG000000053797.11	7.995814586	Krt16	1.6E-288
ENSMUSG000000030693.11	5.438888122	Klk10	1.6E-284
ENSMUSG000000027375.15	5.756475119	Mal	3.2E-278
ENSMUSG000000024406.17	4.482568991	Pou5f1	5.5E-278
ENSMUSG000000037820.16	4.034731856	Tgm2	2.0E-275
ENSMUSG000000056602.14	5.374733383	Fry	6.0E-259
ENSMUSG000000037185.10	5.000475971	Krt80	5.0E-258
ENSMUSG000000026413.13	5.51518737	Pkp1	6.7E-255
ENSMUSG000000047281.4	5.261909517	Sfn	1.3E-254
ENSMUSG000000031995.10	6.011086535	St14	5.1E-243
ENSMUSG000000034282.4	6.82596039	Evpl	1.1E-217
ENSMUSG000000019102.11	3.762064531	Aldh3a1	2.5E-217

\* check attached file for all differentially expressed genes at padj < 0.001

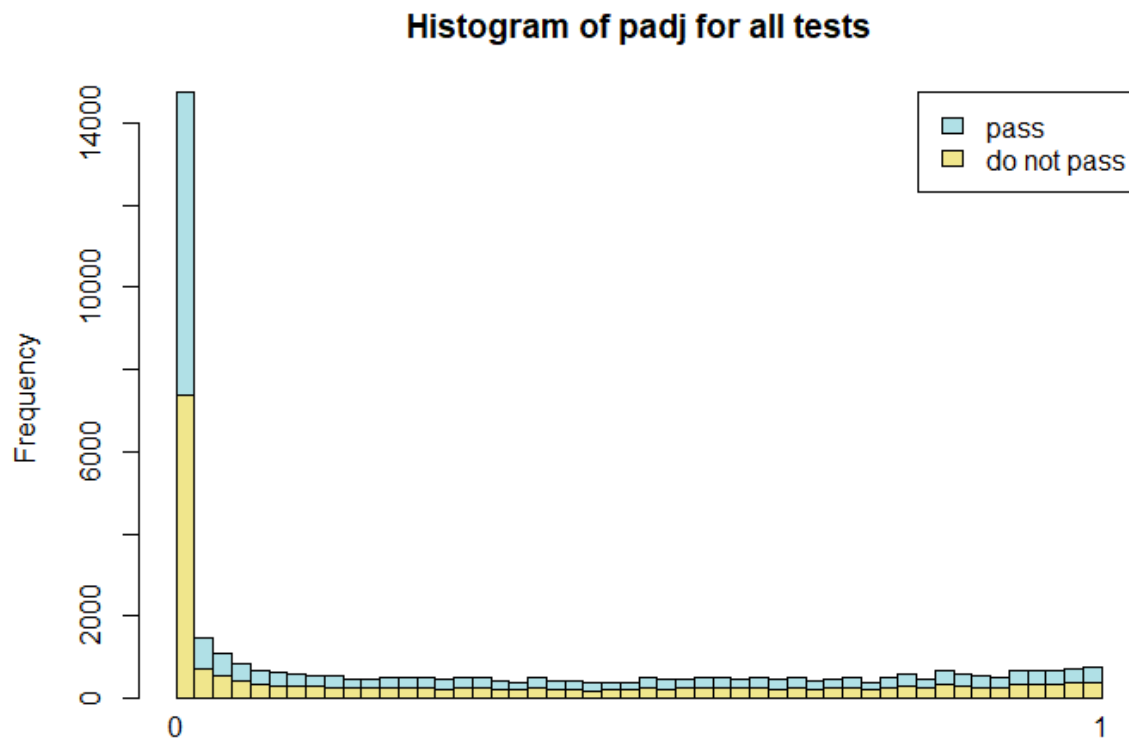
\*There are 5377 genes with adjusted p-value < 0.001

**Fig 7c: Top 10 Differentially Expressed Genes**



Histogram of p values for all tests. The area shaded in blue indicates the subset of those that pass the filtering, the area in khaki those that do not pass:

**Fig 7d: Histogram of P-adjusted for all tests**



The x-axis is divided into 50bins comprising of all padj values. The y-axis indicates the number of test within a particular padj value.

### Gene Set Enrichment Analysis (GSEA)

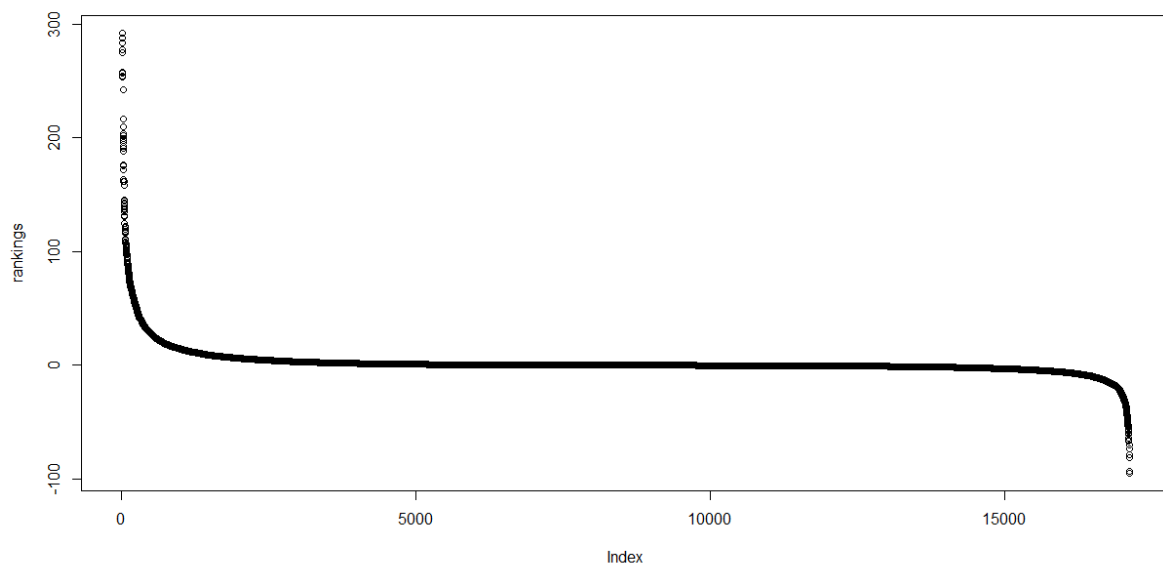
GSEA is a computational method used to identify groups of genes (pathways) that are significantly more active or inactive in one biological condition compared to another. The procedure involves ranking genes based on a metric that considers both the fold change in expression and the statistical significance of that change. In our analysis, we used log-transformed adjusted p-values from differential expression analysis for this ranking (Fig 8a).

We first prepared gene sets from a .gmt file using the `prepare_gmt` function. Then, we applied the `fgsea` function to perform GSEA, considering these ranked genes and the predefined gene

sets. This analysis generated a list of significantly enriched pathways based on their normalized enrichment scores (NES) and adjusted p-values (padj).

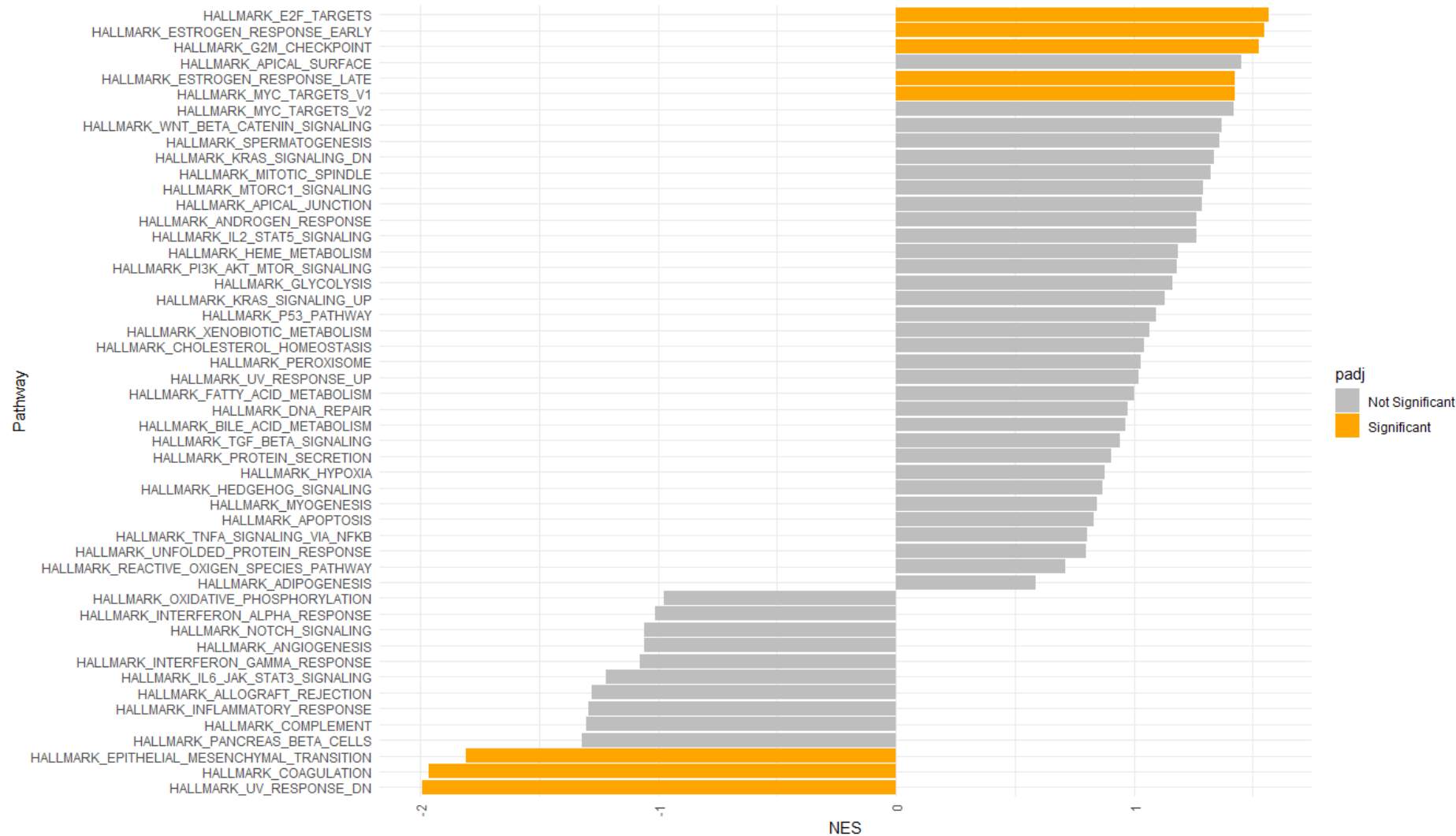
Finally, we created a vertical bar chart to visualize pathway enrichment. The chart shows the NES for each pathway, with bars colored orange to highlight pathways that are significantly enriched based on a padj cutoff of 0.05 (Fig 8b).

**Fig 8a: Ranked Index of Genes**



\* Ranking using  $\text{sign}(\text{df\$log2fc}) * (-\log_{10}(\text{df\$pval}))$  as metric.

Fig 8b: Hallmark Pathways from using Normalized Enrichment Score form GSEA



## **CONCLUSION**

In conclusion, this OMICS RNA-Seq bioinformatics experiment was conducted on sequencing data from six samples. This report highlights the importance of quality control measures, the utilization of alignment algorithms like HISAT2 and BOWTIE2 and the significance of normalization techniques in ensuring accurate gene expression analysis. It also demonstrated the manual calculation of RPKM normalization using Chd1. Through detailed analyses and visualizations, this report demonstrates the identification of differentially expressed genes between conditions, emphasizing the statistical significance of these differences. This experiment also shows potential pathways and biological processes of differentially expressed genes.