AUTOMATING RNA-SEQ ANALYSIS FOR GENE ABUNDANCE

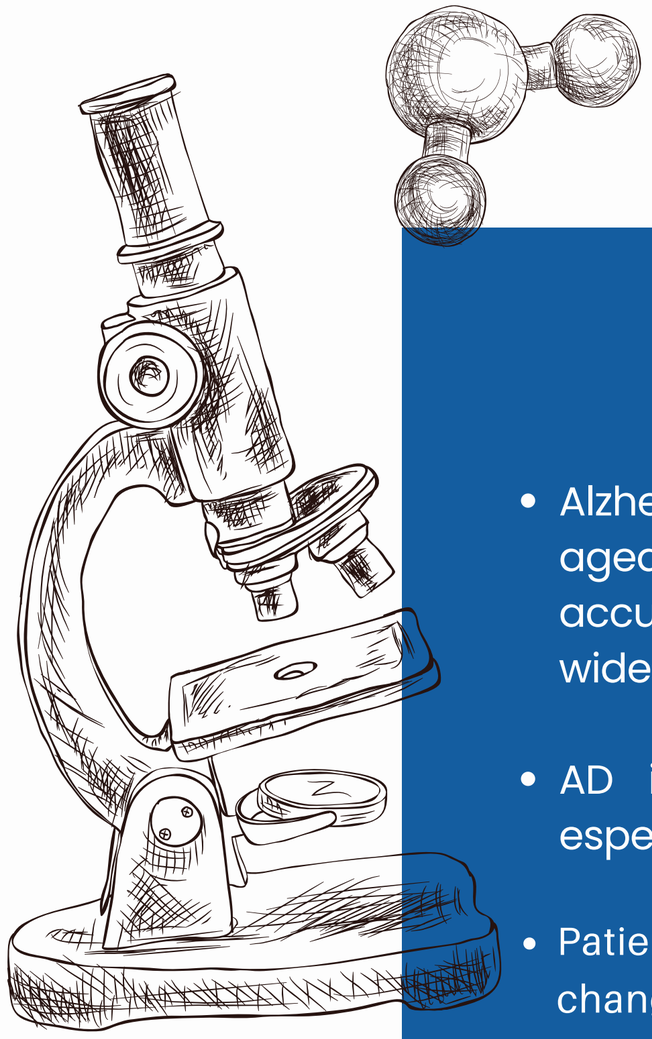
GUIDED BY :

SRIDHAR SRINIVASAN ,DR. RAMACHANDRA PRASAD, SIDDHI KURTADIKAR,MEENAKSHI

PRESENTED BY:

HAREESH T

TABLE OF CONTENTS



Introduction

- Alzheimer's disease (AD) is the most prevalent dementia that affects the aged globally. It is typified by the pathology of amyloid-β (Aβ) accumulation, development of neurofibrillary tangles (NFTs), and widespread neurodegeneration in the brain
- AD is primarily characterized by progressive neurological decline, especially selectively targeted memory loss and cognitive dysfunction
- Patients at risk of sporadic late-onset AD may carry unique genetic changes, including clusterin, TREM2, and APOE variants. However, the interaction between specific LOAD risk alleles and disease pathogenesis remains elusive.

Introduction to RNA-seq

1 RNA-seq is an experimental protocol that produces cDNA sequences from whole RNA molecules. It is followed by the creation of libraries and massively parallel deep sequencing

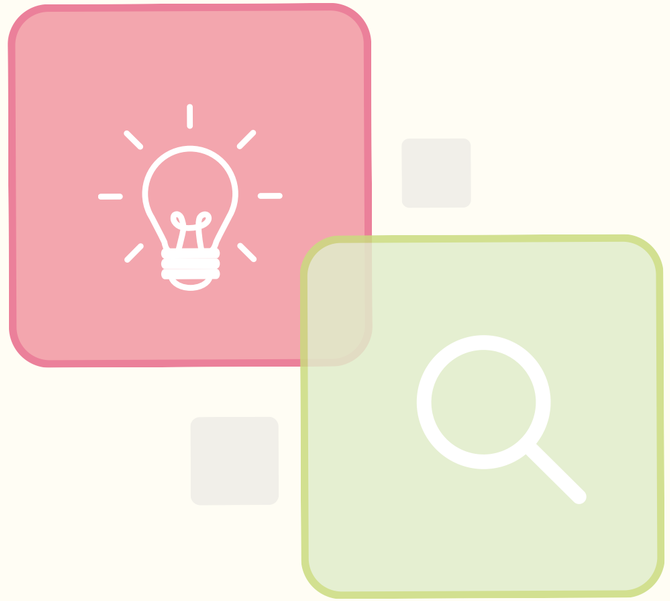
2 Comprehensive Analysis: RNA-seq captures both coding and non-coding RNAs, providing a complete view of the transcriptome, including mRNAs, long non-coding RNAs (lncRNAs), and small RNAs like microRNAs.

3 Quantitative Accuracy: RNA-seq provides quantitative measurements of gene expression levels, enabling precise comparisons across different conditions and treatments.

4 Clinical Applications: RNA-seq is increasingly being used in clinical settings for diagnostic purposes, such as identifying gene expression signatures for disease classification, prognosis, and treatment response.

ADVANTAGES

The capacity to characterize transcriptome dynamics at single-nucleotide resolution is a key benefit of RNA-Seq. As a result, the sequenced transcript reads can reflect transcription from both the paternal and maternal alleles, providing coverage across heterozygous locations



Hybridization-based microarray technologies were initially utilized in transcriptomics research due to their high-throughput and low cost. However, these techniques have drawbacks such as requiring prior sequence knowledge, causing cross-hybridization artifacts, and limited quantification of highly and lowly expressed genes

APPLICATIONS

COMPARATIVE TRANSCRIPTOMICS

A significant use of RNA sequencing is comparing transcriptomes across distinct developmental phases, between a disease state and normal cells, or between certain experimental stimuli and physiological settings

CANCER RESEARCH:

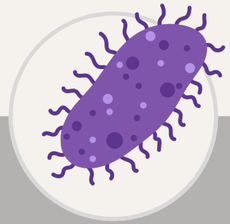
RNA-seq is extensively used to identify differentially expressed genes, novel transcripts, and gene fusions in various types of cancers. It helps in understanding the molecular mechanisms driving tumor development and progression

NEUROBIOLOGY:

RNA-seq is utilized to study the transcriptome of the brain and other components of the nervous system, shedding light on the molecular basis of neurological disorders such as Alzheimer's disease, Parkinson's disease, and autism.

PERSONALIZED MEDICINE

RNA-seq can be used to profile gene expression in individual patients, enabling the development of personalized treatment plans based on the specific molecular characteristics of a patient's disease.



MICROBIOME STUDIES

RNA-seq, specifically metatranscriptomics, allows researchers to study the functional activity of microbial communities in various environments, including the human gut, soil, and ocean.

TRANSCRIPTOMIC INSIGHTS INTO AD

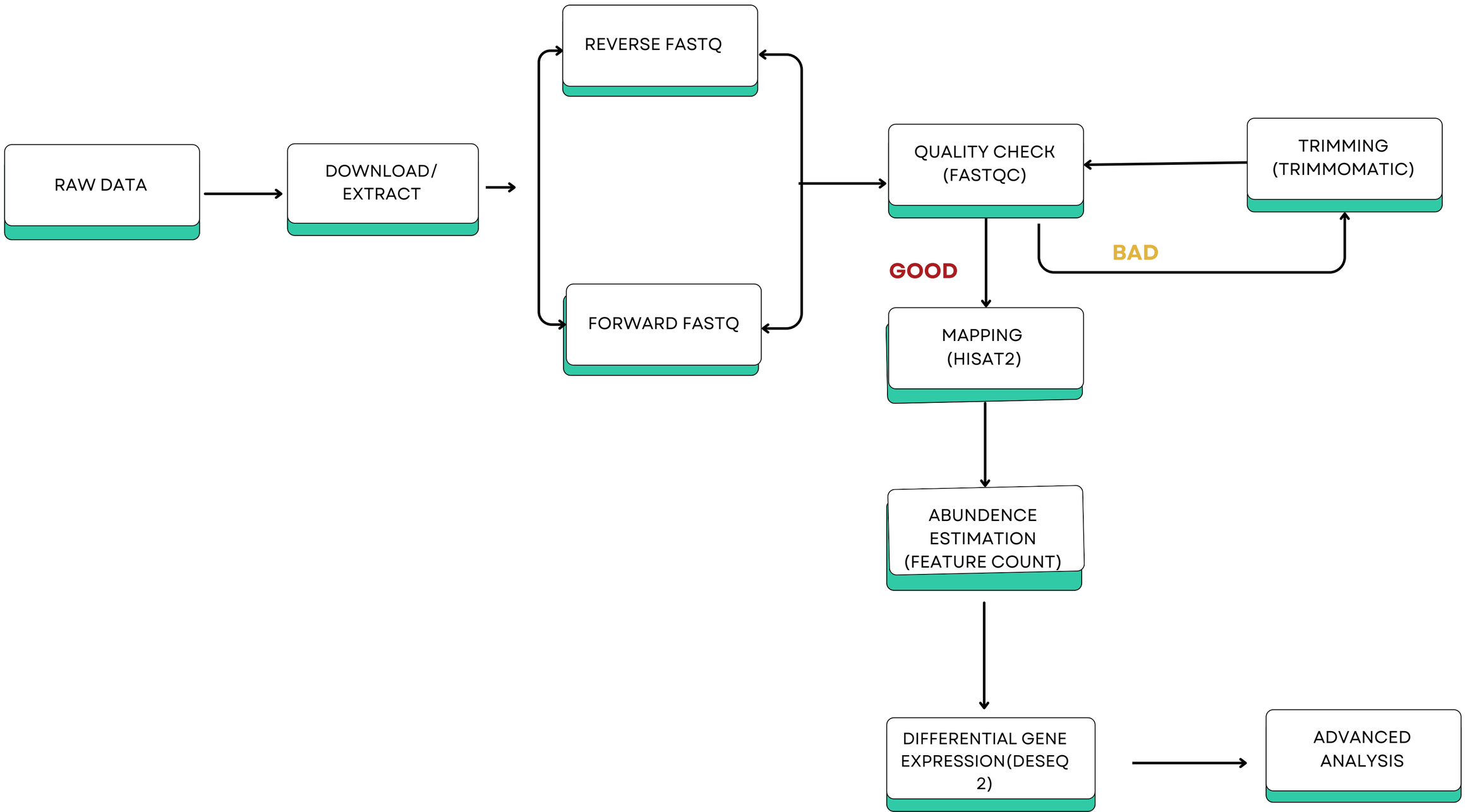
Studying mRNA expression in Alzheimer's disease (AD) patients is crucial because it can provide insights into the underlying molecular mechanisms of the disease.

Understanding the gene expression changes in AD can help identify potential therapeutic targets and biomarkers for early diagnosis.

The primary factor affecting RNA quality is brain pH, which is strongly connected with the agonal condition.Although there is evidence of mRNA's varying susceptibility to degradation throughout the agonal phase, RNA degrades from the 5′ to 3′ end

By analyzing mRNA expression, researchers can uncover novel transcriptomic events that may be critical in the pathogenesis of AD, leading to advancements in treatment and prevention strategies

RNA SEQ PIPELINE



Command	Tool Version	Input	Output	Time Taken (Approx
fasterq-dump SRR123456	fasterq-dump 2.10.9	SRA accession number	FASTQ file(s)	20mins
seqtk sample -s100 SRR.fastq 1087458 > read.fastq	seqtk 1.3-r116	FASTQ file(s)	Subsampled FASTQ file(s)	10 mins
fastqc file1.fastq file2.fastq file3.fastq	FastQC 0.11.9	Subsampled FASTQ file(s)	Quality control report(s)	15 mins
java -jar Trimmomatic-0.39.jar PE -threads $threads $read1 $read2 \$OutputForwardPaired $C	Trimmomatic 0.39	Subsampled FASTQ file(s)	Trimmed FASTQ files	5 mins
hisat2-build genome.fa genome_index	HISAT2 2.2.1	reference_genome.fa	HISAT index files (.ht2)	36 mins
hisat2-build -p 8 reference.fa index_basename	HISAT2 2.2.1	HISAT index files (.ht2), Trimmed FASTQ files	SAM alignment file (.sam)	36 mins
samtools view [options] <input.bam/sam> [region]	SAMtools 1.13	SAM alignment file (.sam)	BAM alignment file (.bam)	5 minns
samtools sort [options] <input.bam> -o <output.bam>	SAMtools 1.13	BAM alignment file (.bam)	Sorted BAM file (.bam)	5 mins
samtools index sorted.bam	SAMtools 1.13	Sorted BAM file (.bam)	BAM index file (.bai)	10 minutes
featureCounts -a annotation.gtf -o counts.txt -O -p -t exon -g gene_id sample1.bam sample2.bam	Subread 2.0.1	GTF annotation file, BAM alignment files	Gene expression counts file (.txt)	10 mins
dds <- DESeq(your_count_data, design = ~ group)	DEseq2.34.0	count data obtained from high-throughput sequencing experiments	data structures and statistical summaries generated during the differential expression analysis	1 min

OUTLINE OF WORK

TRIMMOMATIC

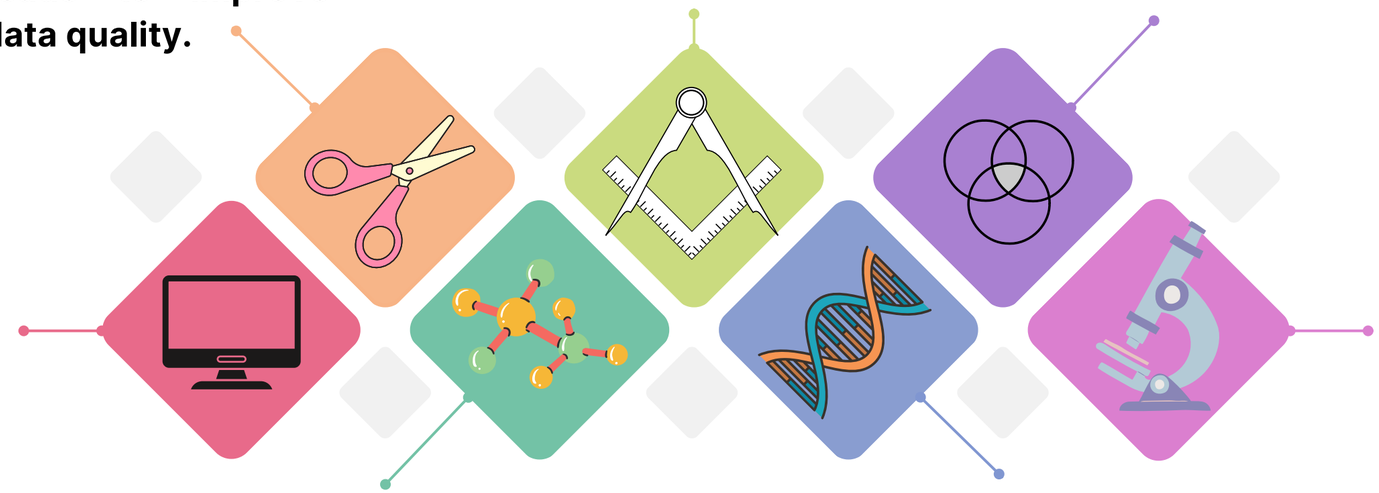
trim adapters and filter low-quality reads to improve data quality.

HISAT2:

Align trimmed reads to a reference genome or transcriptome.

DESeq2

Identify differentially expressed genes between conditions.



FASTQC

Helps assess read quality and provide summary reports to identify potential issues in sequencing data.

Post FastQC

Perform another round of quality control to ensure that trimming has improved read quality if trimming was performed.

FeatureCounts

Quantify gene expression levels from aligned reads.

Advanced Analysis

PCA, GO analysis with GO Shiny, and clustering with ClustVis for interpreting biological significance and regulatory patterns.

ANALYSIS PLAN

Analysis plan	Upregulated	Downregulated
AD vs WT	437	753

FILTERED GENES

MOST UP REGULATED

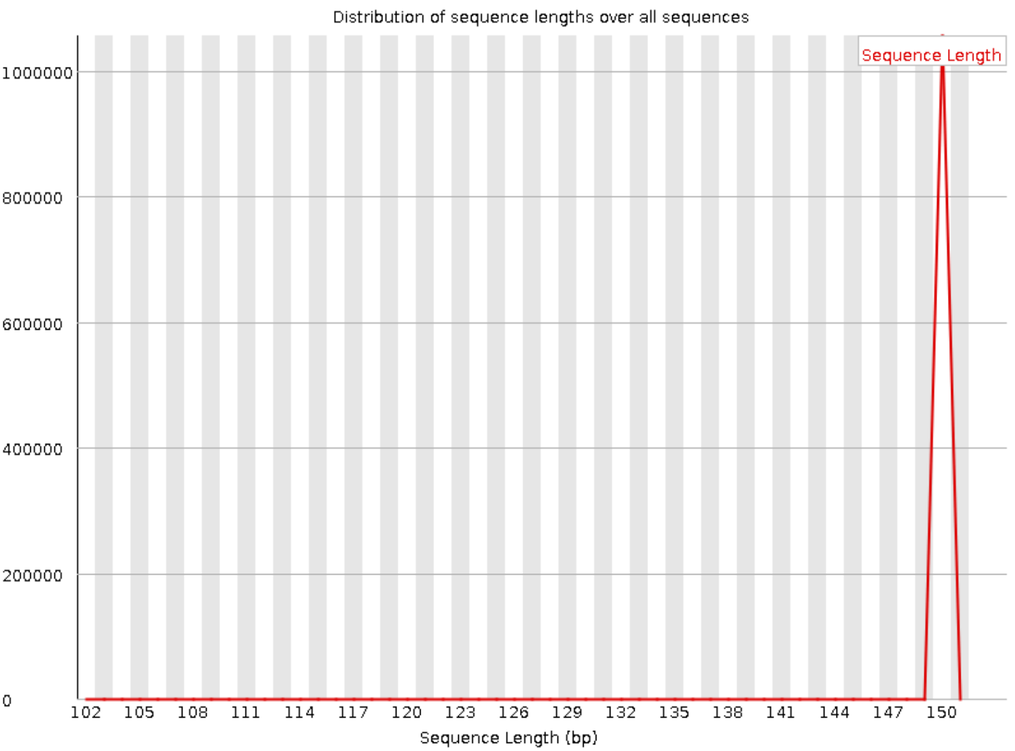
SL.NO	GENE ID	LOG2FOLDCHANGE	PVALUE
1	ENSG00000144452	26.0802720858051	5.0182311476509E-08
2	ENSG00000225792	25.219216606338	1.35936223323593E-07
3	ENSG00000268713	10.7972174473106	0.0000292368195871955
4	ENSG00000287243	10.6476755872563	0.0000292368195871955

FILTERED GENES

MOST DOWN REGULATED

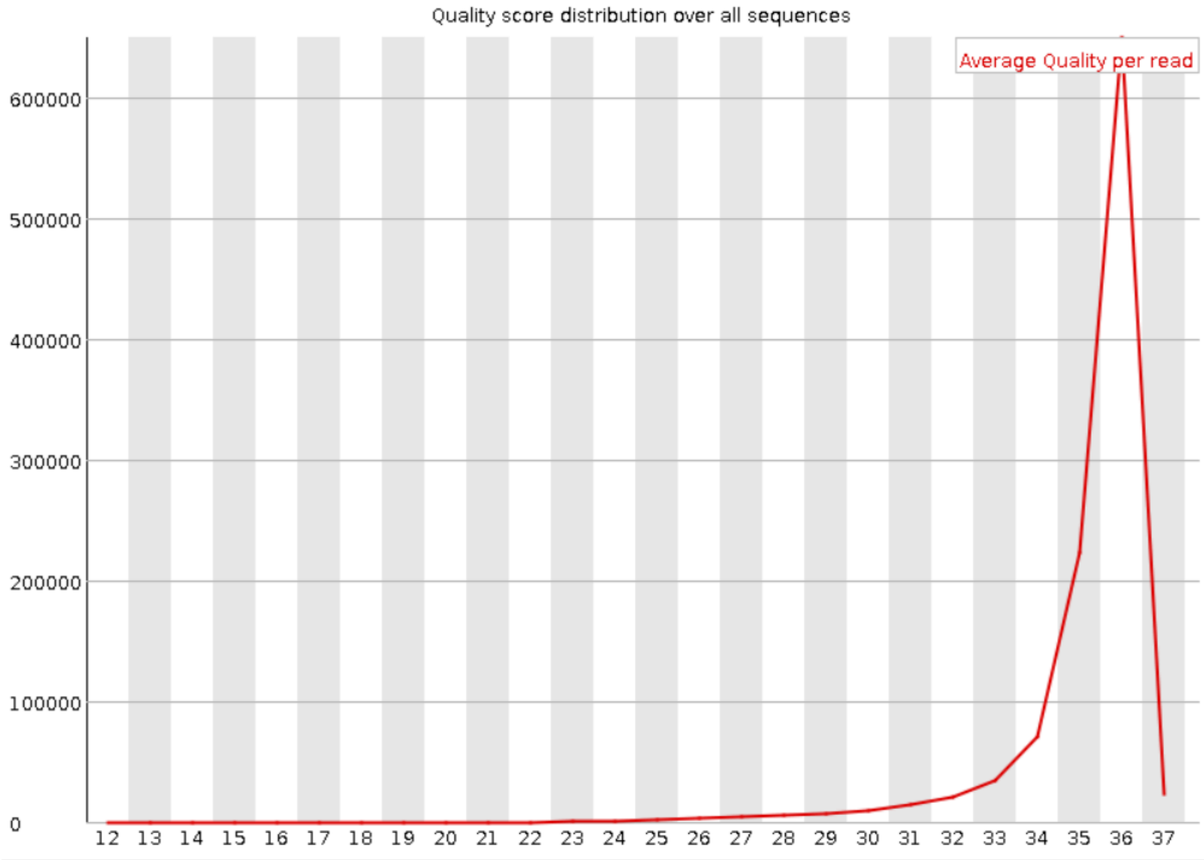
SL.NO	GENE ID	LOG2FOLDCHANGE	PVALUE
1	ENSG00000138119	-1.000684375	0.049947 2311476509E-08
2	ENSG00000116396	-1.000684375	0.0452374793882364
3	ENSG00000152558	-1.000748679	0.000059797517647401 8
4	ENSG00000020406 5	-1.003444576	0.001851046879320 26

RESULTS



Basic Statistics

Measure	Value
Filename	AD_01_1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1087458
Sequences flagged as poor quality	0
Sequence length	150
%GC	49



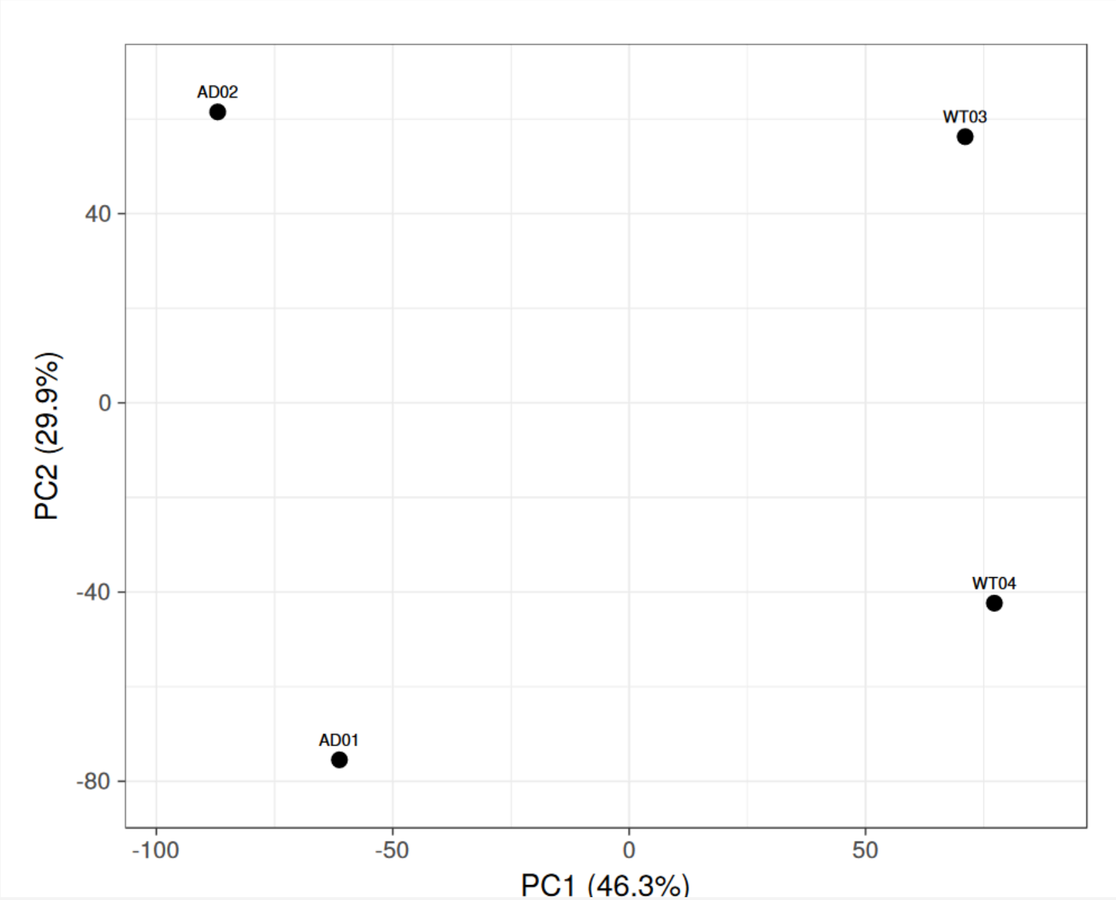
Preprocessing, using Trimmomatic, has likely improved the quality and consistency of the data, as indicated by FastQC. The reduction in variability and concentration around a single value suggests removal of outliers or artifacts, resulting in more reliable data for analysis.

ALIGNING READS TO REFERENCE GENOME

	SAMPLE NAME	OVERALL ALIGNMENT RATE	
01	AD_01	91.82%	01
02	AD_02	92.42%	02
03	WT_03	93.82%	03
04	WT_04	92.44%	04

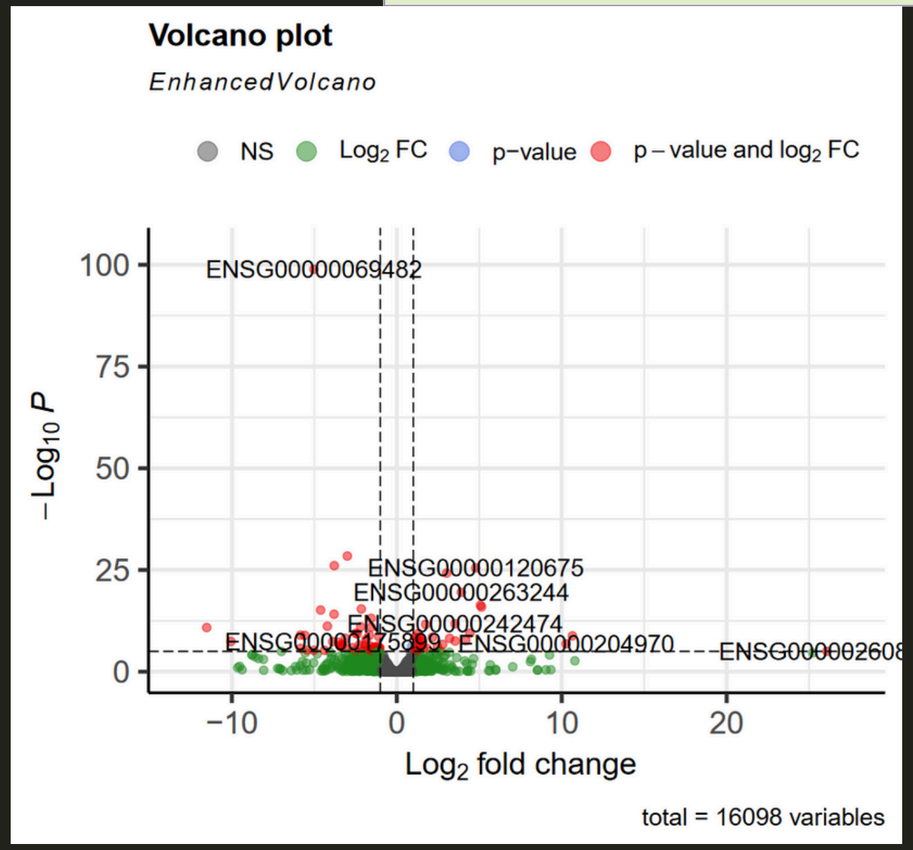
Primary Analysis

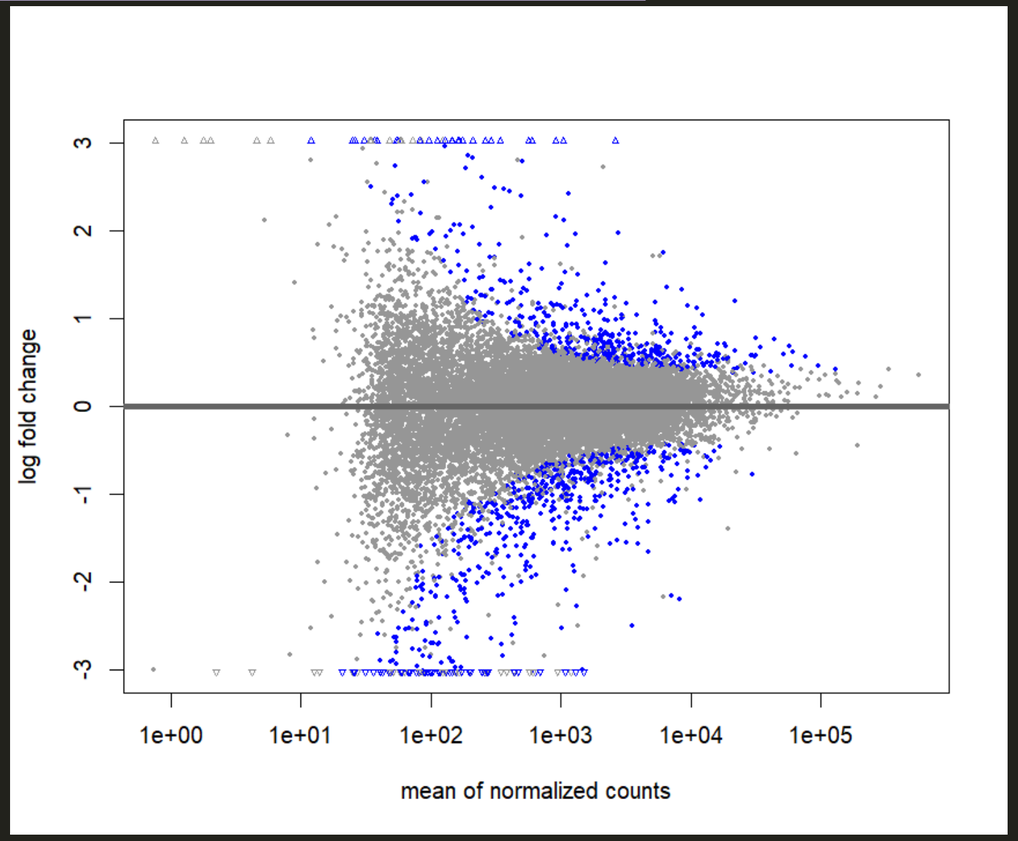
CLUSTVIS PCA



ClustVis' PCA graph simplifies complex data visualization by reducing dimensions to principal components. It shows axes representing maximum variance, revealing clusters and separations.

DESEQ2 RESULTS

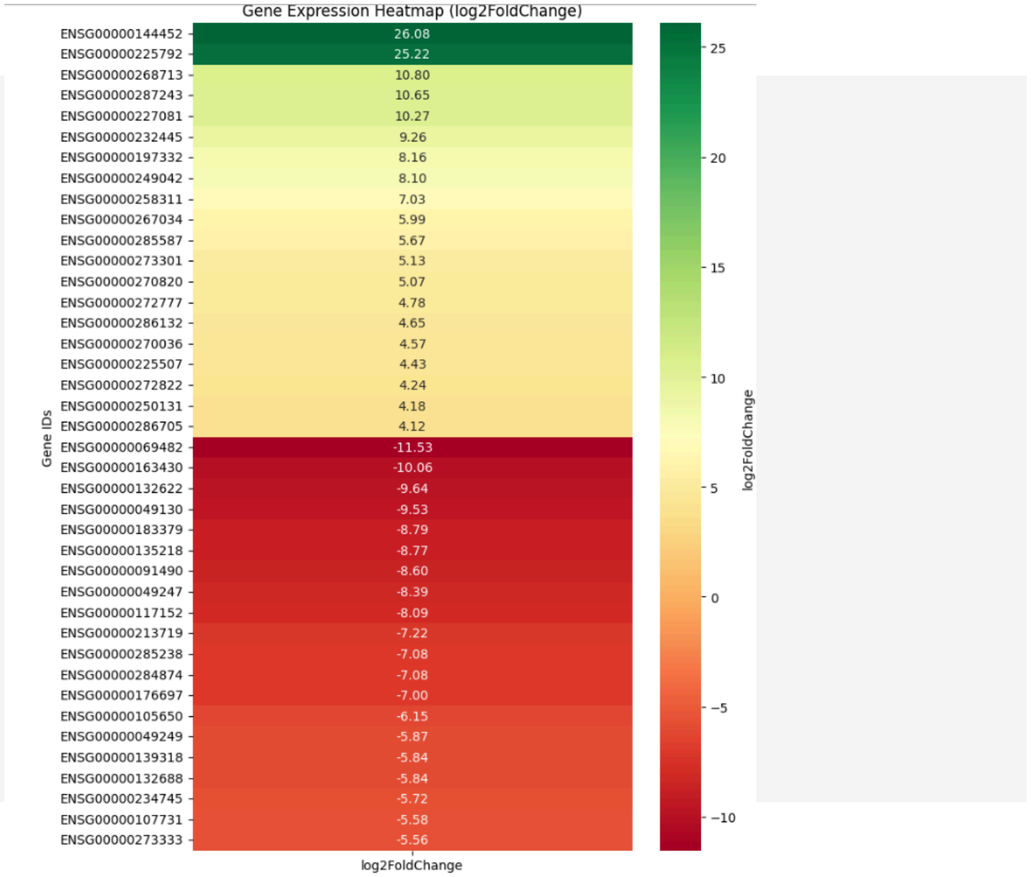




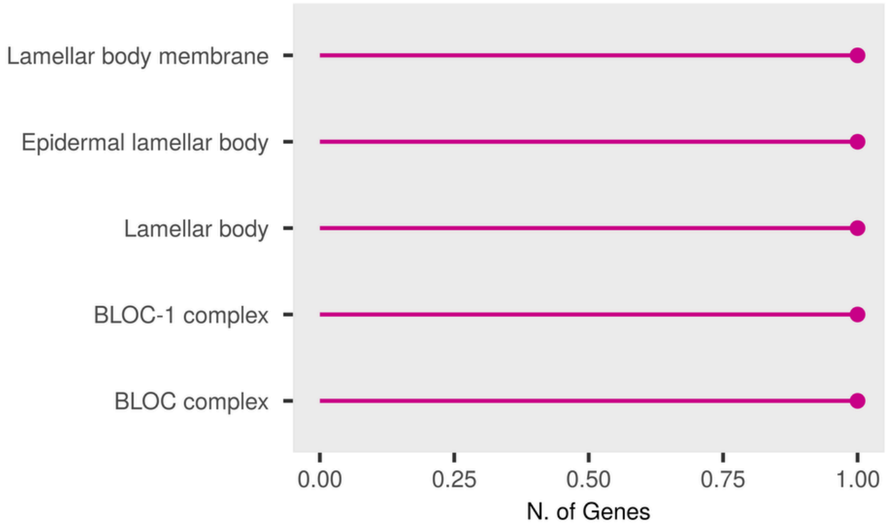
Gene Id	Log2 fold change	pvalue	Regulation
ENSG00000260836	26.08027209	5.02E-08	Up regulated
ENSG00000130707	-1.000684375	0.045237479	Down regulated

Here ENSG00000260836 is approximately 26.062 times upregulated compared to the control sample

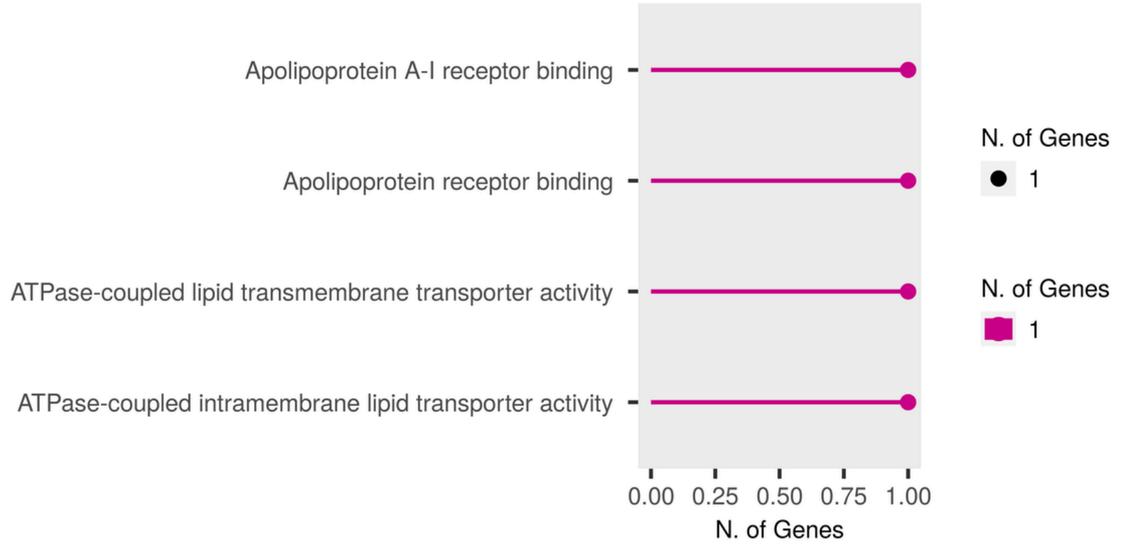
HEATMAP OF UP REGULATED AND DOWN REGULATED GENES



ADVANCED ANALYSIS ON SHINY GO 0.80
UP REGULATED

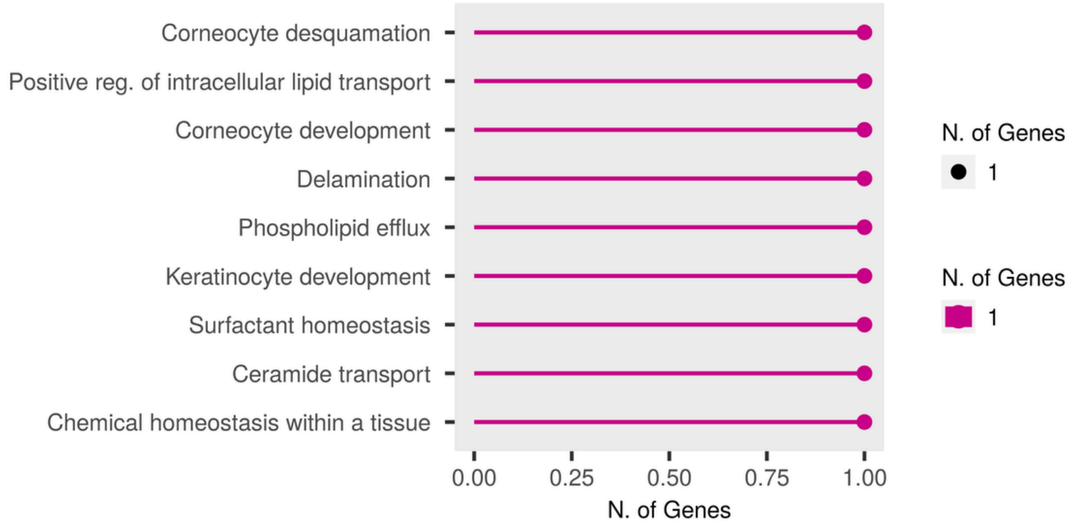


CELLULAR COMPONENT

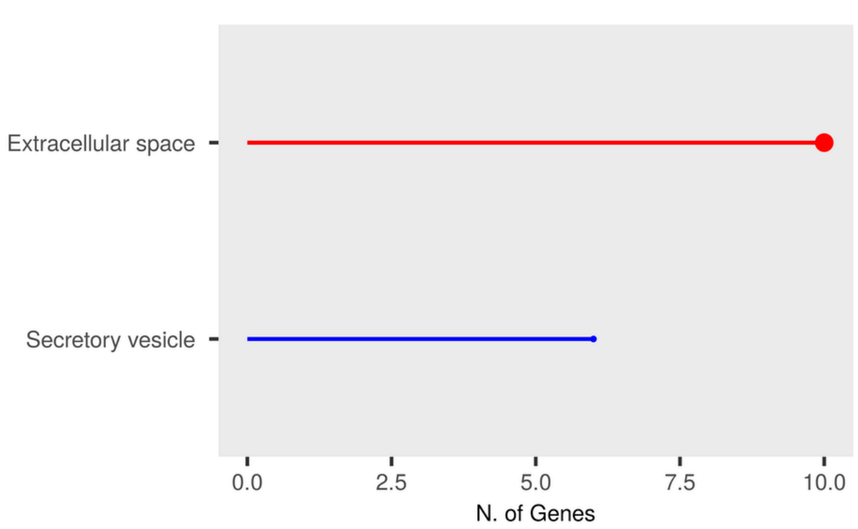


MOLECULAR FUNCTION

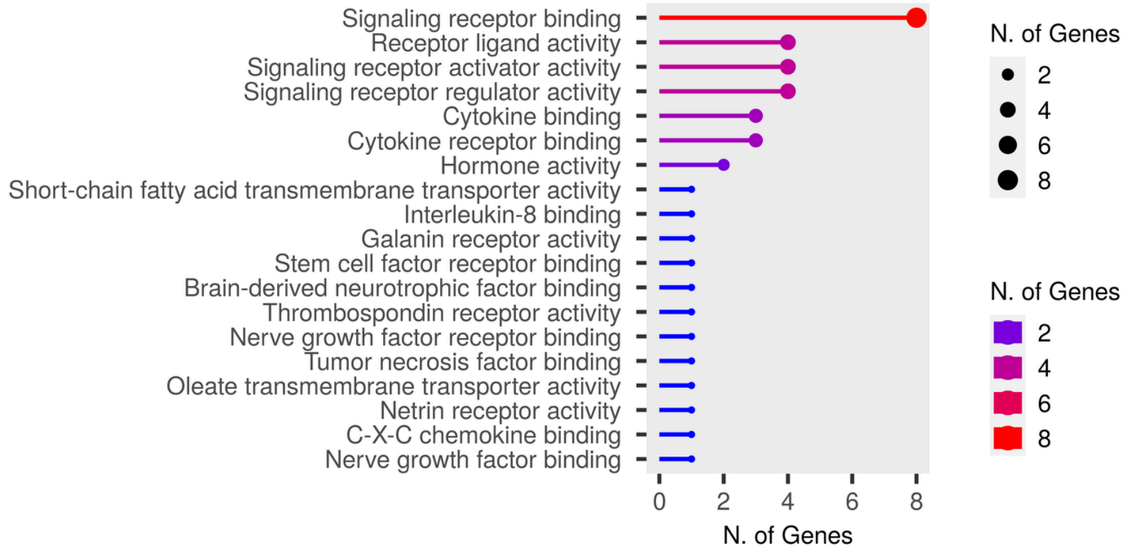
BIOLOGICAL PROCESS



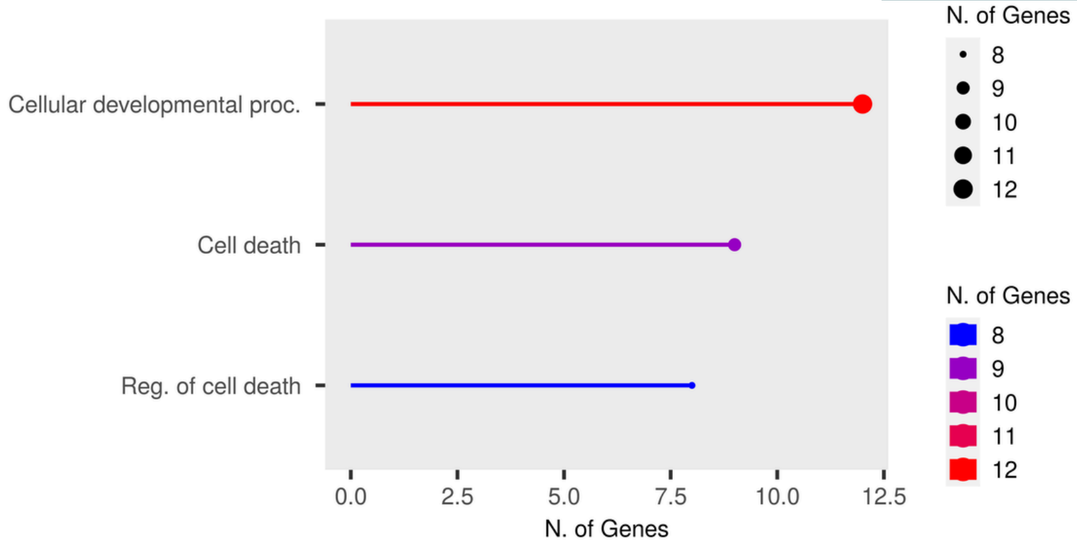
ADVANCED ANALYSIS ON SHINY GO 0.80
DOWN REGULATED



CELLULAR COMPONENT



MOLECULAR FUNCTION

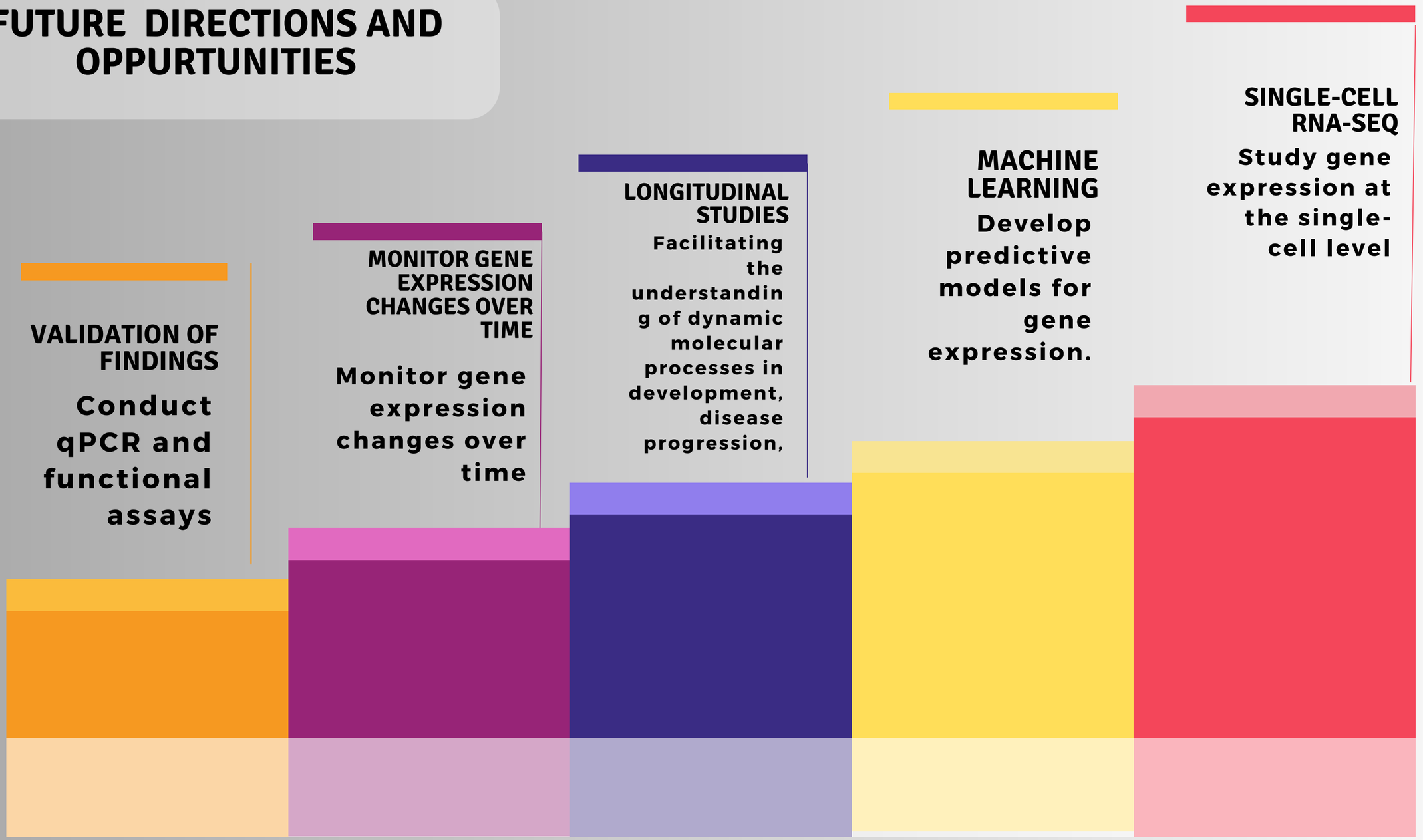


BIOLOGICAL PROCESS

AUTOMATION SAMPLE REVIEW

```
(base) bioinfo@MSI:~/rna_seq/code$ ./rna_seq_pipeline.sh
```

FUTURE DIRECTIONS AND OPPURTUNITIES



KEY LEARNINGS

1
RNA-seq Pipeline building

2
Database integration

3
Linux commands

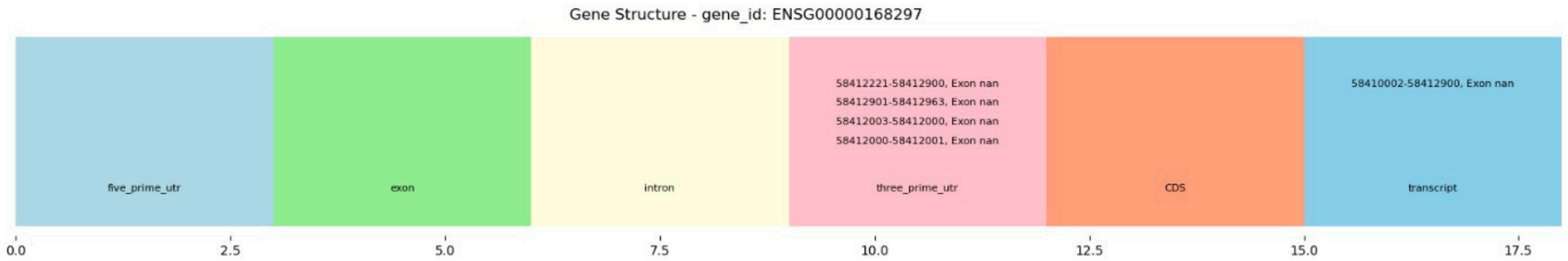
4
Gathering data from leading companies focused on precision medicine

5
Pipeline Automation

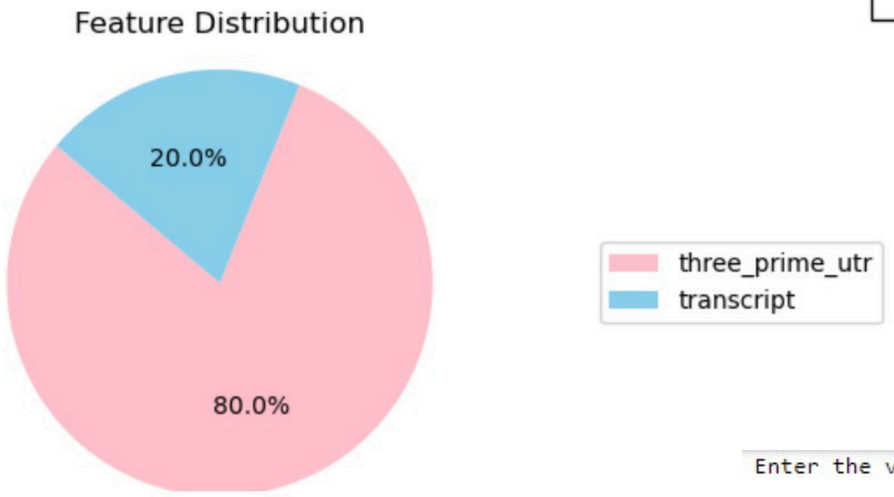
6
GTF visualizer

7
Exposure to certain NGS based tools

GTF Visualizer



Feature	Counts
three_prime_utr	4
transcript	1




```
Enter the value you want to search for (gene_id, gene_name, transcript_id, transcript_name): PXK

Results:
Chromosome         Feature    Start      End         gene_id gene_name   transcript_id transcript_name  exon_number Strand
         3 three_prime_utr 58412221 58412900 ENSG00000168297       PXK ENST00000468776         PXK-207          NaN      +
         3 three_prime_utr 58412901 58412963 ENSG00000168297       PXK ENST00000468776         PXK-207          NaN      +
         4 three_prime_utr 58412003 58412000 ENSG00000168000       PXK ENST00000468776         PXK-207          NaN      +
         5 three_prime_utr 58412000 58412001 ENSG00000168000       PXK ENST00000468776         PXK-207          NaN      +
         7      transcript 58410002 58412900 ENSG00000168000       PXK ENST00000468776         PXK-207          NaN      +
       NaN             NaN 58412901 58412963 ENSG00000168297       PXK ENST00000468776         PXK-207          NaN      +
```

THANK YOU

Grateful for the Opportunity!

As my internship concludes, I extend heartfelt thanks to 3BIGS for the invaluable experience and guidance provided by my mentors ., Sridhar, Siddhi , Meenakshi .

Looking forward to carrying these lessons forward!

HAREESH T