Differential Gene Expression Analysis Report

Introduction

The main objective is to identify genes that were differentially expressed between two experimental conditions called "Condition A" and "Condition B," using RNA-seq. Each condition is represented by three biological replicates, totaling six RNA-seq samples. RNA-seq is a high-throughput sequencing technique that allows for precise quantification of transcript levels and provides comprehensive information about gene expression under different conditions.

The present study will be done to make use of a systematic workflow comprising quality control of raw data, adapter trimming, read alignment to the reference genome, gene-level quantification, and statistical analysis that identifies DEGs.

The design of this analytical pipeline was to ensure accuracy, reproducibility, and in-depth insight into transcriptomic changes, culminating in the identification of key genes upregulated and downregulated that might implicate important biological processes or pathways. Such information is very instrumental in the provision of insights into the molecular mechanisms behind experimental conditions.

Materials and Methods

1. Quality Control-FastQC:

FastQC was utilized to evaluate the quality of the raw sequencing data. This step identified potential issues that could impact downstream analyses. Key metrics assessed included:

a) **Per Base Quality Scores:** Allowed the identification of low-quality regions that might require trimming.

b) **GC Content Distribution:** Checked for anomalies indicating possible contamination or bias.

c) **Adapter Content:** Highlighted residual adapter sequences requiring removal.

d) **Sequence Duplication Levels** Provided insights into library complexity and sequencing depth.

The generated FastQC reports served as an initial checkpoint to ensure the data met quality standards for subsequent steps.

2. Adapter Trimming(Cutadapt):

Adapter trimming was performed using Cutadapt to remove residual adapter sequences and low-quality bases, which could interfere with read alignment and quantification. The steps involved were:

a) Identifying adapter sequences based on predefined adapter files.

b) Removing adapters and trimming low-quality bases at read ends to retain only high-quality reads.

c) Reassessing trimmed reads with FastQC to confirm improved quality and successful adapter removal.

This step ensured that only high-quality, adapter-free reads were passed to the alignment stage.

3. Read Alignment(STAR):

The trimmed reads were aligned for the reference genome using STAR (Spliced Transcripts Alignment to a Reference), it is a high-performance aligner optimized for RNA-seq data. Key procedures includes:

a) Indexing the reference genome with a corresponding annotation file.

b) Aligning reads to the genome to produce BAM files containing spliced alignments.

c) Assessing alignment statistics, including uniquely mapped reads, multi-mapped reads, and overall alignment rates.

The high alignment rates confirmed the efficacy of the trimming and alignment process.

4. Feature Counting(featureCounts)

Gene expression quantification was performed using featureCounts, a tool from the Subread package. Steps included:

a) Assigning aligned reads to genomic features based on the annotation file.

b) Generating a raw count matrix for all genes across the six samples.

The resulting count matrix formed the foundation for downstream statistical analysis.

5. Differential ExpressionAnalysis(DESeq2)

Differential expression analysis was conducted using DESeq2, a robust statistical tool tailored for RNA-seq data. The process involved:

a) Condition Definition: Grouping samples into "Condition A" and "Condition B."

b) Normalization: Accounting for library size and sequencing depth to ensure fair comparisons.

c) Statistical Modeling and Significance Testing: Employing a negative binomial distribution to model count data and identifying DEGs using adjusted p-values (FDR < 0.05) and log2 fold-change thresholds.

```
library(DESeq2)

counts <- read.table("featureCounts_output.txt", header=TRUE,
comment.char="#")

rownames(counts) <- counts[,1]
counts <- counts[,-c(1:6)]  # Remove annotation columns

colnames(counts) <- paste0("sample_", 0:5)
```

```
# Define conditions
condition <- factor(c("A", "A", "A", "B", "B", "B"))
colData <- data.frame(condition)
rownames(colData) <- colnames(counts)

# Filtering
keep <- rowSums(counts) >= 10
counts <- counts[keep,]

dds <- DESeqDataSetFromMatrix(countData=counts,
                              colData=colData,
                              design=~condition)

dds <- DESeq(dds)
res <- results(dds)

print("Result summary:")
summary(res)

sig_res <- res[!is.na(res$padj) & abs(res$log2FoldChange) >= 2 &
res$padj < 0.05, ]

res_up <- sig_res[sig_res$log2FoldChange >= 2, ]
res_down <- sig_res[sig_res$log2FoldChange <= -2, ]

write.table(res_up, file="deseq2_up.txt", sep="\t", quote=FALSE)
write.table(res_down, file="deseq2_down.txt", sep="\t", quote=FALSE)

q("no")
```

6. Result Filtering and export

DEGs were categorized based on their expression changes:

a) **Upregulated Genes:** Initially, a log2 fold-change >= 2 was used, but no genes met this stringent threshold. Relaxing the criterion to log2 fold-change >= 1 yielded additional upregulated gene, saved in deseq2_up_manual.txt.

b) **Downregulated Genes:** Genes with log2 fold-change <= -2 and FDR < 0.05 were saved in deseq2_down.txt.

7. To run the workflow

The entire analysis was executed using Snakemake:

snakemake --cores 4

And then pack the results

tar -cvf results.tar featureCounts_output.txt deseq2_up.txt deseq2_down.txt workflow.smk protocol.pdf

Results

The analysis identified DEGs between the experimental conditions. Key outcomes included:

a) **deseq2_up.txt:** No genes satisfied the initial stringent upregulation threshold (log2 fold-change >= 2).

b) **deseq2_down.txt:** Several genes demonstrated significant downregulation, providing insights into potential repressed biological pathways.

c) **deseq2_up_manual.txt:** Relaxing the threshold to log2 fold-change >= 1 revealed additional upregulated gene, broadening the scope of discovery.

These findings underscore the importance of balancing statistical stringency with biological relevance, especially in exploratory studies.

Discussion

This study successfully implemented a robust RNA-seq pipeline to identify DEGs. Several key observations emerged:

a) **Stringency in Thresholds:** The absence of upregulated genes under stringent criteria highlights the challenge of balancing sensitivity and specificity. Adjusting thresholds dynamically can help uncover biologically relevant genes without inflating false positives.

b) **Pipeline Robustness:** The integration of tools like FastQC, Cutadapt, STAR, featureCounts, and DESeq2 ensured high-quality data processing, enabling accurate and reproducible results.

c) **Exploratory Adjustments:** Relaxing thresholds revealed genes with moderate upregulation, showcasing the value of exploratory adjustments in gene expression studies.

Conclusion

This RNA-seq analysis successfully identified DEGs between experimental conditions using a structured computational workflow.The findings provide insights into gene expression changes and serve as a basis for further biological interpretation.