SHAHD KHALID 221001066

Abstract:

The paper illustrates a full RNA-seq analysis pipeline that covers all steps, starting from raw data acquisition and culminating in the derivation of functional insights. It is implemented in Bash and R, covering data preprocessing, differential gene expression studies, and enrichment analyses, including GSEA. Solid integration among HISAT2, DESeq2, and visualization tools underlines automation in genomics as central to actionable biological insights.

Introduction:

RNA-seq has become a cornerstone in transcriptomics, enabling the investigation of gene expression profiles across different conditions. The following report describes a pipeline using state-of-the-art alignment tools such as HISAT2 and differential expression analyses like DESeq2, with further enhancement through GSEA for biological interpretation. This pipeline was designed in such a way that raw sequencing data would be analyzed efficiently to produce reproducible results that offer insight into cellular mechanisms and pathways.

Results:

Pipeline Execution

Raw data from SRA repositories were downloaded and pre-processed. Quality control indicated high-quality reads post-trimming.

Alignments had high mapping rates against the reference genome GRCh38 using HISAT2.

With adjusted p-value < 0.05, a total of X genes were highly upregulated, and Y genes were downregulated.

Enrichment Analysis

Among top pathways identified by GSEA are [Pathway Name 1], emphasizing their role in [biological context].

A graph showing a line of blue lines

Description automatically generated with medium confidence

Description: fastaqc

Mean Quality Scores (Blue Line): Represents the average quality score at each base position.

Interquartile Range (Shaded Area): Indicates the variability in quality scores.

Thresholds:

20 (Orange Line): Minimum acceptable quality.

30 (Green Line): High-quality threshold.

A graph with red and blue dots

Description automatically generated

The red dots: represent the genes with significant differential expression (Significant Genes).

The blue dots: represent the genes that were not statistically significant (Non-significant Genes).

The x-axis: represents the mean gene expression (Mean Expression).

The y-axis: represents the log2 fold change value (Log2 Fold Change).

This plot helps identify genes that showed clear expression changes between the studied samples.

A graph of a volcano plot

Description automatically generated

Axis: Log2 fold change of gene expression (left: downregulated, right: upregulated).

Y-Axis: -Log10 p-value (higher values mean greater statistical significance).

Red Points: Genes whose changes are significant (adjusted p-value < 0.05).

Blue Points: Genes with non-significant changes.

Dashed Lines:

Vertical line at 0: No expression change.

Horizontal line: Significance threshold.

A graph showing a number of green lines

Description automatically generated with medium confidence

X-Axis: Gene rank (negative = downregulated, positive = upregulated).

Y-Axis: Enrichment score (ES) profile.

Green Line: Running ES, with peaks marking the highest enrichment.

Blue Line: Peak enrichment location in the gene rank.

Text Box: Normalized Enrichment Score (NES = 2.05) and p-value (0.001), indicating significance.

Conclusion:

This integration of robust tools for RNA-seq analysis enables efficient and reproducible workflows. The results provide a valuable resource for understanding gene expression dynamics and biological pathways under studied conditions. Future enhancements may include additional enrichment analyses and validation experiments.

.