Bioinformatic -Final Project

Report: Examination Of Gene Expression by IRIS EDA: A reproducibility Study

**Introduction and Problem Statement**

The paper chosen for this project addresses challenges in RNA-Seq experimentation by leveraging a bioinformatic approach to improve accessibility, efficiency, and reproducibility in RNA-Seq data analysis.

RNA-Seq data analysis is inherently fragmented and complex, requiring the use of multiple specialized tools createing inefficiencies and increases the likelihood of errors, particularly for researchers without advanced bioinformatics expertise. The lack of integration not only limits the potential of RNA-Seq for personalized treatments but also delays scientific discoveries and reduces the reproducibility of results. To overcome these challenges, there is a critical need for streamlined technology that makes RNA-Seq analysis more accessible, accelerates data processing, and supports the advancement of medical research and targeted treatment development.

As part of this reproducibility study, we explore IRIS-EDA (Integrated RNA-Seq Exploratory Data Analysis), a comprehensive platform designed to address these challenges. IRIS-EDA integrates all essential steps of RNA-Seq data analysis into a single, user-friendly tool. By simplifying advanced data interpretation, it democratizes access to RNA-Seq analysis for researchers across all skill levels.

IRIS-EDA disrupts conventional, fragmented analysis methods by streamlining complex workflows, accelerating the identification of genetic variables, and enhancing the discovery of disease biomarkers. This integrated platform has the potential to transform medical research by supporting the development of individualized treatments and improving patient outcomes.

**Project Aim and Objectives**

The aim of this project is to replicate and critically evaluate key Exploratory Data Analysis (EDA) techniques employed by the IRIS-EDA framework while applying them to bulk RNA-Seq data from different embryonic cell stages. The focus includes reproducing essential analyses, such as Principal Component Analysis (PCA), correlation studies, and other EDA methods central to IRIS-EDA’s approach, followed by performing Differential Gene Expression (DGE) analysis.

This project aligns with IRIS-EDA's goal of streamlining and integrating bioinformatics workflows by combining traditionally complex tasks—like quality control, visualization, and statistical analyses—into a coherent, user-friendly process. By applying these techniques to embryonic RNA-Seq datasets, the study aims to:

- Assess the reproducibility and effectiveness of IRIS-EDA’s analytical methods.

- Demonstrate how these methods enhance workflow efficiency and accessibility for bulk RNA-Seq analysis.

- Extract meaningful biological insights about gene expression patterns across embryonic cell stages through PCA, correlation, and DGE analysis.

**Materials and methods**

For the analysis carried out in this project, RNA-Seq data from the study "Tracing Pluripotency of Human Early Embryos and Embryonic Stem Cells by Single-Cell RNA-Seq," was used. The study aimed to establish a causal relationship between gene expression networks and cellular phenotypes at single-cell resolution.

The dataset includes transcriptomic profiles of 124 individual cells from human pre-implantation embryos and embryonic stem cells derived from these embryos. Single-cell RNA-Seq techniques were applied to isolate individual cells, prepare single-cell cDNAs, and sequence them using HiSeq2000. The analysis focused on the expression of known RefSeq genes.

Key highlights:

- Organism: *Homo sapiens*

- Experiment type: Expression profiling by high-throughput sequencing

- Objective: To trace the gene expression dynamics during early embryonic development and pluripotency.

- Findings: The data revealed highly dynamic expression patterns of protein-coding genes across different developmental stages.

This dataset serves as a robust resource for analyzing the transcriptional landscapes of embryonic cell stages, making it a suitable choice for applying Exploratory Data Analysis (EDA) techniques and Differential Gene Expression (DGE) analysis in the current study.

Methodology Summary

The methodology implemented in this project follows a systematic pipeline based on the outlined steps in the diagram. The process is structured as follows:

1. Loading the Dataset and quality control: Bulk RNA-Seq data from embryonic cell stages was imported and prepared for analysis. This dataset includes transcriptomic information from multiple embryonic stages, enabling comprehensive exploration of gene expression patterns. The check for null and duplicated values were perform in this step too.

2. Computing Total Counts, Variability, and Distribution: Statistical summaries and visualizations of the total read counts, variability, and gene expression distribution across samples were generated to assess data quality and identify global patterns.

3.Filtering Genes: Genes with low expression levels were filtered out based on predefined thresholds to focus on biologically significant and robustly expressed genes. This step reduced noise and enhanced the interpretability of downstream analyses.

5. Correlation Analysis: Pairwise correlations were computed to investigate relationships between samples and identify clusters or groupings based on similarity in gene expression profiles.

6. Principal Component Analysis (PCA): PCA was conducted to reduce dimensionality and visualize the overall variance in the dataset, highlighting key trends and differences among embryonic stages.

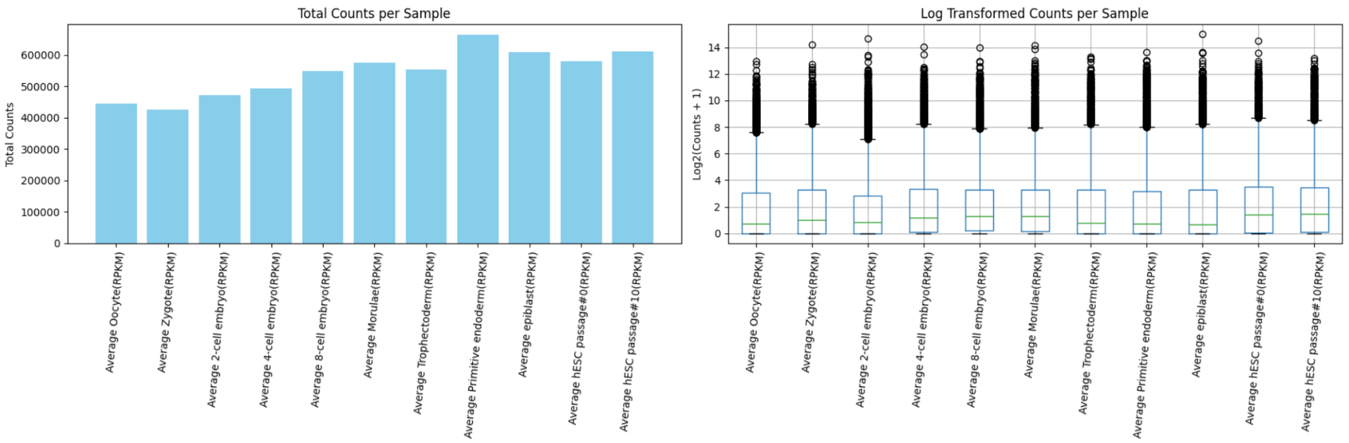
7. Dendrogram Construction: Hierarchical clustering was performed to create a dendrogram, which provided insights into sample similarities and the hierarchical organization of gene expression patterns.

8. Differential Gene Expression (DGE) Analysis: Finally, a DGE analysis was carried out to identify genes that show significant expression changes across different embryonic stages. This step highlighted potential biomarkers or key drivers of developmental processes.

This structured approach allowed for a thorough exploration of the bulk RNA-Seq dataset, integrating multiple analytical techniques to derive meaningful biological insights while ensuring data quality and reproducibility.

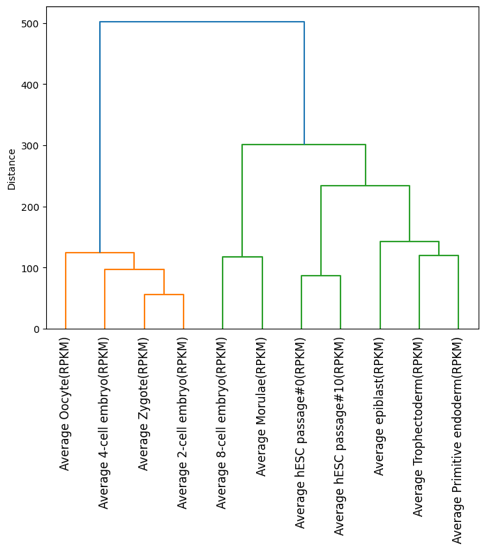
Results

1. Total counts per sample and log-transformed counts

The left bar plot illustrates the total counts per sample (in RPKM), demonstrating a consistent distribution of reads across embryonic cell stages, with higher counts observed in later stages, such as the primitive endoderm. The right box plot depicts the log-transformed counts per sample, showcasing the variability and distribution of gene expression values, where most genes cluster at lower expression levels, while a few outliers represent highly expressed genes. These visualizations confirm the quality and comparability of the dataset across samples, validating its reliability for subsequent analyses.

1. Dendrogram

The dendrogram visualizes the hierarchical clustering of samples based on their RPKM values, grouping embryonic cell stages by their similarity in gene expression profiles. Early stages, such as oocyte, zygote, and 2-cell embryo, form a distinct cluster, reflecting their shared transcriptional characteristics. Later stages, including morula, trophectoderm, and primitive endoderm, cluster separately, highlighting their divergence in gene expression as development progresses. This analysis demonstrates the progressive differentiation of embryonic stages and validates the dataset's ability to reveal biologically meaningful relationships among samples.



Correlation analysis

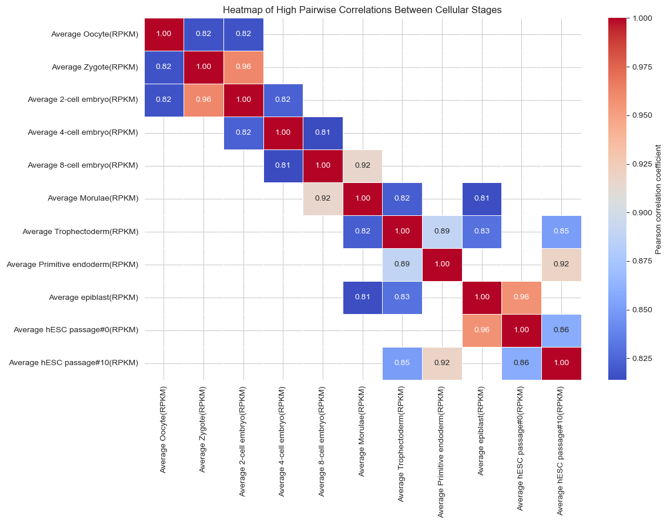
The heatmap displays the pairwise Pearson correlation coefficients between average RPKM values across different embryonic stages. From this plot it can be observed that:

1. High Correlations in Adjacent Stages: Early stages (e.g., zygote and 2-cell embryo) and later stages (e.g., primitive endoderm and epiblast) exhibit strong correlations (>0.9), reflecting similar gene expression patterns during closely related developmental phases.

2. Divergence Between Early and Late Stages: Lower correlations are observed between early stages (e.g., oocyte) and later stages (e.g., hESC passage#10), indicating significant transcriptional divergence as development progresses.

3. Clustered Relationships: Similar patterns of gene expression are evident within distinct clusters of stages, supporting the biological progression of embryonic development.

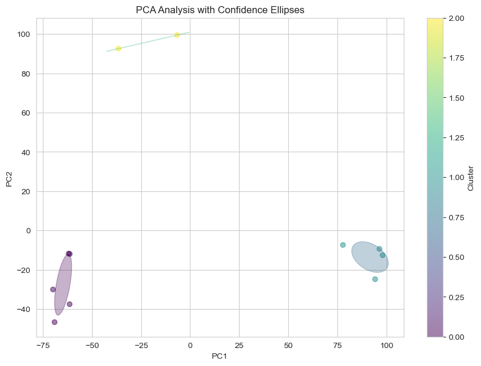
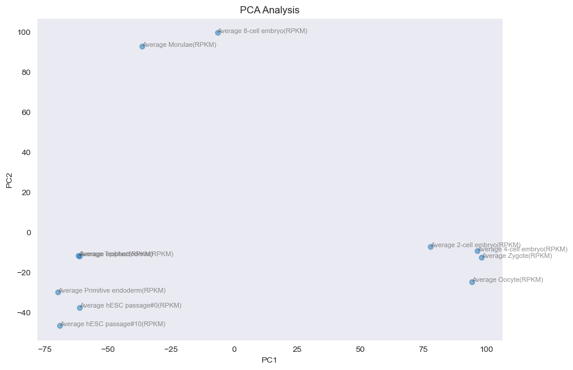
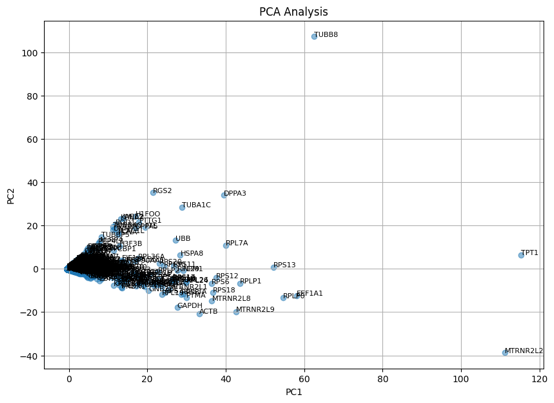
This correlation analysis highlights stage-specific transcriptional similarities and differences, providing insight into the transitions and regulatory mechanisms underlying embryonic development.



PCA

The PCA analysis for genes (first plot) highlights the contribution of specific genes to the variance in the dataset. Genes like TUBB8 and PT1 stand out as drivers of variation, with clusters representing patterns of co-expression or functional similarity.

The PCA for samples (second plot) reveals distinct clustering of embryonic stages, reflecting transcriptional differences across developmental progression. Early stages, such as oocytes and zygotes, cluster separately from later stages like morulae and hESCs, emphasizing their unique gene expression profiles. Also, the plot with confidence ellipses reinforces these patterns by grouping samples with similar variance, indicating robust stage-specific transcriptional signatures. Collectively, these analyses validate the dataset's ability to capture biologically meaningful differences across genes and samples.



DGE analysis

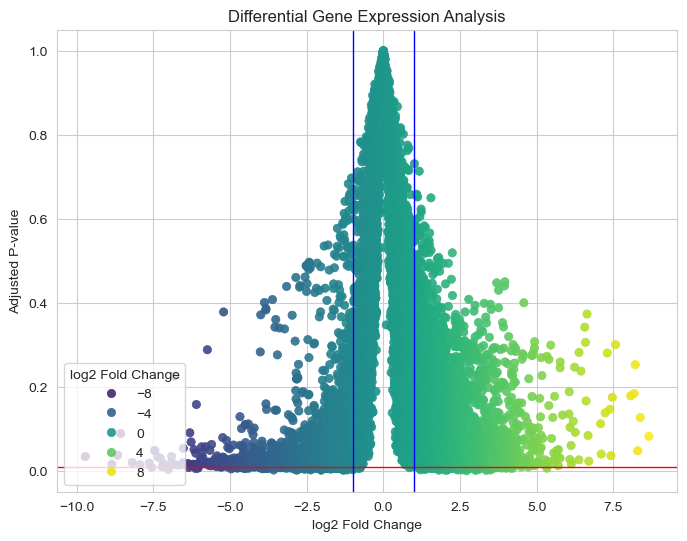
The volcano plot illustrates the differential gene expression (DGE) analysis comparing early embryonic stages (control group) and later developmental stages (treated group). Genes with significant adjusted p-values (<0.01) and absolute log2 fold changes (>1) are highlighted as differentially expressed. Our notable findings include:

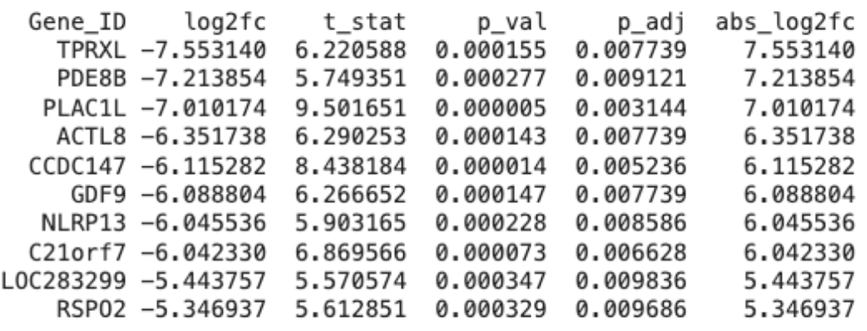
1. Highly Downregulated Genes: Genes like TPRXL, PDE8B, and PLAC1L show strong downregulation (negative log2 fold changes) in treated samples, indicating decreased expression in later stages compared to earlier stages.

2. Threshold Filtering: The blue lines mark the log2 fold change thresholds of ±1, while the red line represents the adjusted p-value cutoff of 0.01, effectively separating significant DGE genes.

3. Biological Insights: The identified DGE genes (e.g., ACTL8, CCDC147, and GDF9) suggest specific pathways or cellular functions being modulated during later developmental stages.

This analysis identifies genes that may play key roles in the transition between early and late embryonic stages, contributing to the understanding of developmental processes. Further functional enrichment of these genes could uncover pathways critical to cell differentiation and pluripotency.





Conclusions

The analyses conducted provide valuable insights into the transcriptional dynamics during human embryonic development. The PCA and dendrogram analyses revealed distinct clustering of embryonic stages, highlighting that gene expression profiles are tightly regulated and specific to each developmental phase. Early stages, such as oocytes and zygotes, display unique transcriptional signatures that differ significantly from later stages like trophectoderm and epiblast. This progression underscores the dynamic changes in gene expression that occur as cells transition through developmental milestones.

The differential gene expression (DGE) analysis identified key genes with significant changes in expression between early and late stages, such as TPRXL and PDE8B, which are strongly downregulated in later stages. These findings suggest that certain genes active in early development are repressed as cells differentiate and acquire specialized functions. The correlation heatmap further supports this observation, with strong correlations between adjacent stages like zygote and 2-cell embryo, reflecting smooth transcriptional transitions. Conversely, lower correlations between early stages (e.g., oocyte) and late stages (e.g., hESC passage#10) highlight the divergence in gene expression as pluripotent cells progress toward lineage specification.

Quality control metrics, including total counts and log-transformed distributions, confirmed the reliability and consistency of the dataset, validating its suitability for downstream analyses. Additionally, several significant genes identified in the DGE analysis, such as GDF9, CCDC147, and PLAC1L, may play critical roles in embryonic development and serve as potential markers for specific stages or targets for further functional studies.

Overall, this study provides a comprehensive view of the molecular mechanisms driving cell differentiation and highlights the transcriptional complexity underlying human embryonic development. These findings not only enhance our understanding of early development but also offer valuable implications for stem cell research, regenerative medicine, and reproductive biology. The analytical workflow demonstrated here establishes a robust framework for exploring transcriptomic datasets, ensuring reproducibility and meaningful biological interpretations.

* Why do they believe the paper was written in the first place?

Today, there is an increasing need for analysis and interpretation of RNA-seq data, and many researchers and users have limited experience with computational tools to proceed. Furthermore, the existence of barriers and bottlenecks in RNA-seq data has made this type of tool (IRIS-EDA) a need for the scientific community.

* Identification of the original researcher's hypothesis.

There is no hypothesis statement per se in this paper. The development and application of a web tool called IRIS-EDA for the analysis of gene expression data is described, but without any hypothesis testing.

* Identification of any substantiating or corroborating information the original authors use to support their findings.

The authors use several sources of information to support their claims about the need for a tool like IRIS-EDA. They refer to previous studies and cite several scientific articles that emphasize the importance of next-generation sequencing and RNA-Seq data analysis. These references support their assertion regarding the increasing availability of large-scale gene expression data and the need for efficient analysis tools. They also mention existing differential gene expression (DGE) analysis tools such as DESeq, DESeq2, edgeR, limma, Cuffdiff, Cuffdiff2 and sleuth, demonstrating their familiarity with the current landscape of available tools. This understanding allows them to identify the limitations of existing solutions and justify the need for a new tool such as IRIS-EDA.

* A statement of the validity of the methods and results.

The authors employ a combination of established methods, flexible experimental design options, discovery-based analysis, interactive visualizations, a real-world application example, FAIR compliance, and open-source availability to support the validity of IRIS-EDA’s methods and results. These approaches build confidence in the tool's reliability and its capacity to produce meaningful, reproducible outcomes.

* Identification of potential drawbacks or oversights the original authors "missed."

IRIS-EDA is a valuable tool for the analysis of gene expression data, but it has some limitations and potential oversights. It is based on R, which may present a learning curve for users unfamiliar with programming, even with its GUI. While it supports large scRNA-Seq datasets, details on performance optimisation for such data are limited to the supplementary materials, leaving questions about scalability and processing. The tool focuses primarily on differential gene expression (DGE) analysis, with fewer options for other analyses such as isoform or gene fusion analysis. In addition, planned features such as functional enrichment and network analysis are not yet included, and the authors do not detail long-term maintenance or update plans. These aspects highlight areas for improvement to enhance its usability and sustainability.

Bibliography

* Monier, B., McDermaid, A., Wang, C., Zhao, J., Miller, A., Fennell, A., & Ma, Q. (2019). IRIS-EDA: An integrated RNA-Seq interpretation system for gene expression data analysis. *PLoS computational biology*, *15*(2), e1006792.
* Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., ... & Tang, F. (2013). Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nature structural & molecular biology*, *20*(9), 1131-1139.