PROJECT REPORT

Project Title: Differential Gene Expression Analysis using DumbDeseq in R

Submitted by: Aprajita Gupta

HackBio Project- DumbDeseq: RNA Expression Software

Introduction

RNA sequencing (RNA-seq) has become one of the most widely used techniques in modern biology for studying how genes are expressed in different conditions. It is a powerful next-generation sequencing technique that allows for the comprehensive profiling of the transcriptome, the complete set of RNA molecules expressed in a cell at a given time. Unlike older methods such as microarrays, which required prior knowledge of gene sequences, RNA-seq can capture both known and novel transcripts with high precision. This ability to provide a digital, quantitative view of gene expression makes RNA-seq one of the most reliable tools for understanding how genes are switched on or off in different conditions (Wang, Gerstein et al. 2009). By examining gene expression patterns, researchers can gain insights into disease mechanisms, treatment effects, and even identify potential biomarkers and therapeutic targets.

One of the core applications of RNA-seq is differential gene expression analysis, which compares two or more conditions to identify genes that are expressed at significantly higher or lower levels. This approach helps in understanding how diseases disrupt normal cellular functions and how treatments may restore or alter these patterns (Love, Huber et al. 2014). Data generated from RNA-seq is often very large and complex, which is why visualization tools are essential. A commonly used method is the volcano plot, which plots the magnitude of expression change (log2 fold change) against statistical significance (-log10 p-value). This simple yet powerful representation allows researchers to quickly see which genes are most strongly and significantly affected (Cui and Churchill 2003).

In this project, I worked on a pre-processed RNA-seq dataset provided by HackBio. The dataset compared diseased cell lines with diseased cell lines treated with a compound. My goal was to apply differential expression analysis using RStudio, generate a volcano plot, to identify the most significantly upregulated and downregulated genes, and then use the GeneCards database to understand their biological roles (Stelzer, Rosen et al. 2016). This analysis provided not only hands-

on experience with RNA-seq data but also an opportunity to connect computational results with real biological meaning.

Objectives

The main goal of this project was to use RNA-seq data to understand how gene expression changes between diseased cells and diseased cells treated with a compound. The specific objectives were:

- **To analyze a pre-processed RNA-seq dataset** provided by HackBio using RStudio.
- **To generate a volcano plot** for visualizing the relationship between fold change and statistical significance of gene expression.
- **To identify significantly upregulated and downregulated genes**, based on Log2 fold change thresholds and p-values.
- **To annotate the selected genes using the GeneCards database** and explore their biological roles in cellular pathways and disease processes.

Methodology

The analysis in this project was carried out using the RNA-seq dataset provided by HackBio for Module 9. The methodology of this project was designed to analyze RNA-seq data in a structured and reproducible way. Since the dataset was already pre-processed and provided by HackBio (2025), the workflow (Figure 1) focused on downstream steps such as visualization, statistical filtering, and biological interpretation rather than raw sequence alignment and quantification. The analysis was carried out in RStudio (2024.12.0), which is one of the most widely used programming environments for bioinformatics and statistical genomics.

The first step of the analysis involved generating a volcano plot. A volcano plot is a scatterplot that combines two important features of gene expression analysis: the magnitude of change and the statistical significance of that change. On the x-axis, the log2 fold change (log2FC) values were plotted to represent how much a gene's expression increased or decreased between the treated and untreated samples. On the y-axis, the negative log10 of the p-value was plotted to reflect the statistical reliability of those changes. Genes that were both highly significant and strongly up- or downregulated appeared at the top corners of the plot. This type of visualization is commonly used in transcriptomics studies because it allows researchers to quickly identify the most relevant genes from

thousands of data point. All visualizations, including the volcano plot and histogram, were generated using base R plotting functions to visualize differential expression patterns.

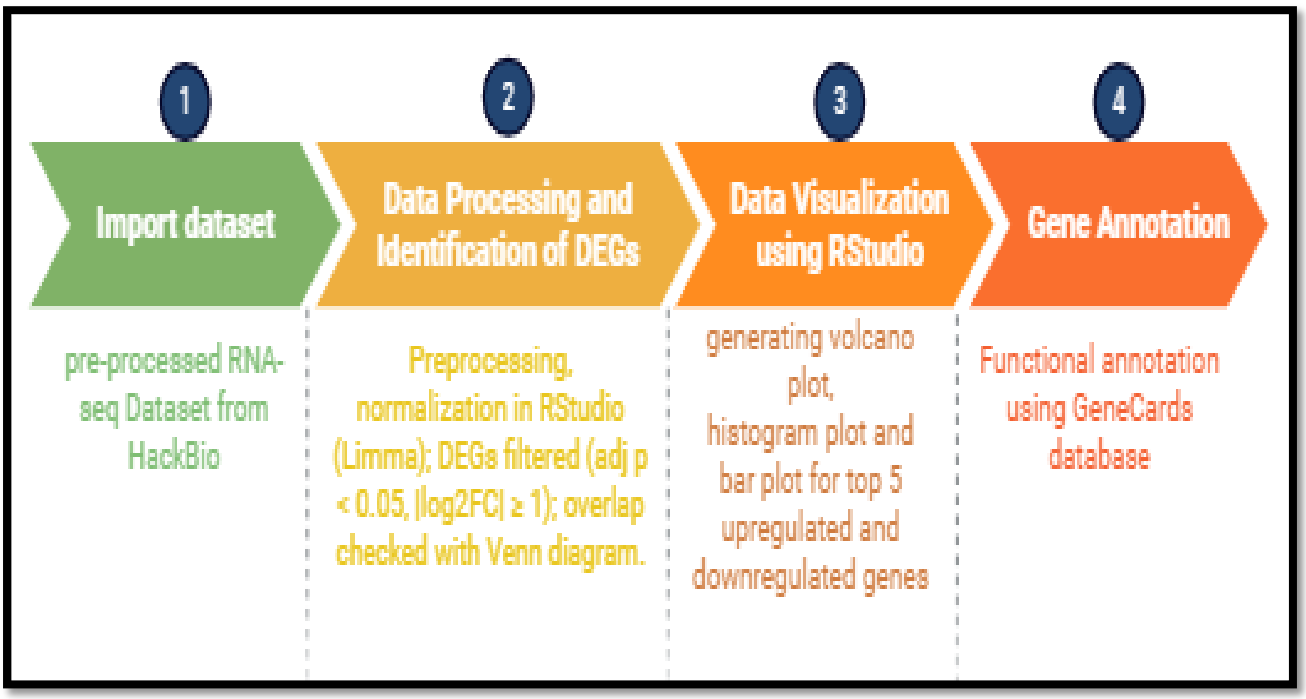

Figure 1: Workflow of the methodology used in the project.

In the second step, genes were classified into upregulated and downregulated categories using strict thresholds. Genes with a log2FC greater than 1 and p-value less than 0.01 were classified as upregulated, while those with a log2FC less than -1 and p-value less than 0.01 were classified as downregulated. Applying these filters ensured that only genes with biologically meaningful fold changes and strong statistical evidence were included in the analysis.

The third step focused on functional interpretation. For this purpose, the GeneCards database was used. GeneCards is an integrative resource that compiles information on gene functions, pathways, interactions, and disease relevance. Each of the top upregulated and downregulated genes was individually searched in GeneCards, and their biological roles were noted. This step added critical biological meaning to the results, helping to move from raw statistical values to functional insights that explain how the treatment may influence cellular processes.

Results

The RNA-seq dataset contained a total of 16,406 genes, with log2 fold change (log2FC) values ranging from –2.13 to +1.54. Out of 16,406 genes, 19 were significantly upregulated and 91 were significantly downregulated. A statistical overview revealed that the distribution of fold changes was

centered around 0, which means that most genes showed little or no differential expression between the two conditions. However, a subset of genes showed significant positive or negative changes in expression. The results were explored using three main visualizations: a volcano plot, a histogram of fold changes, and a bar plot of the top differentially expressed genes.

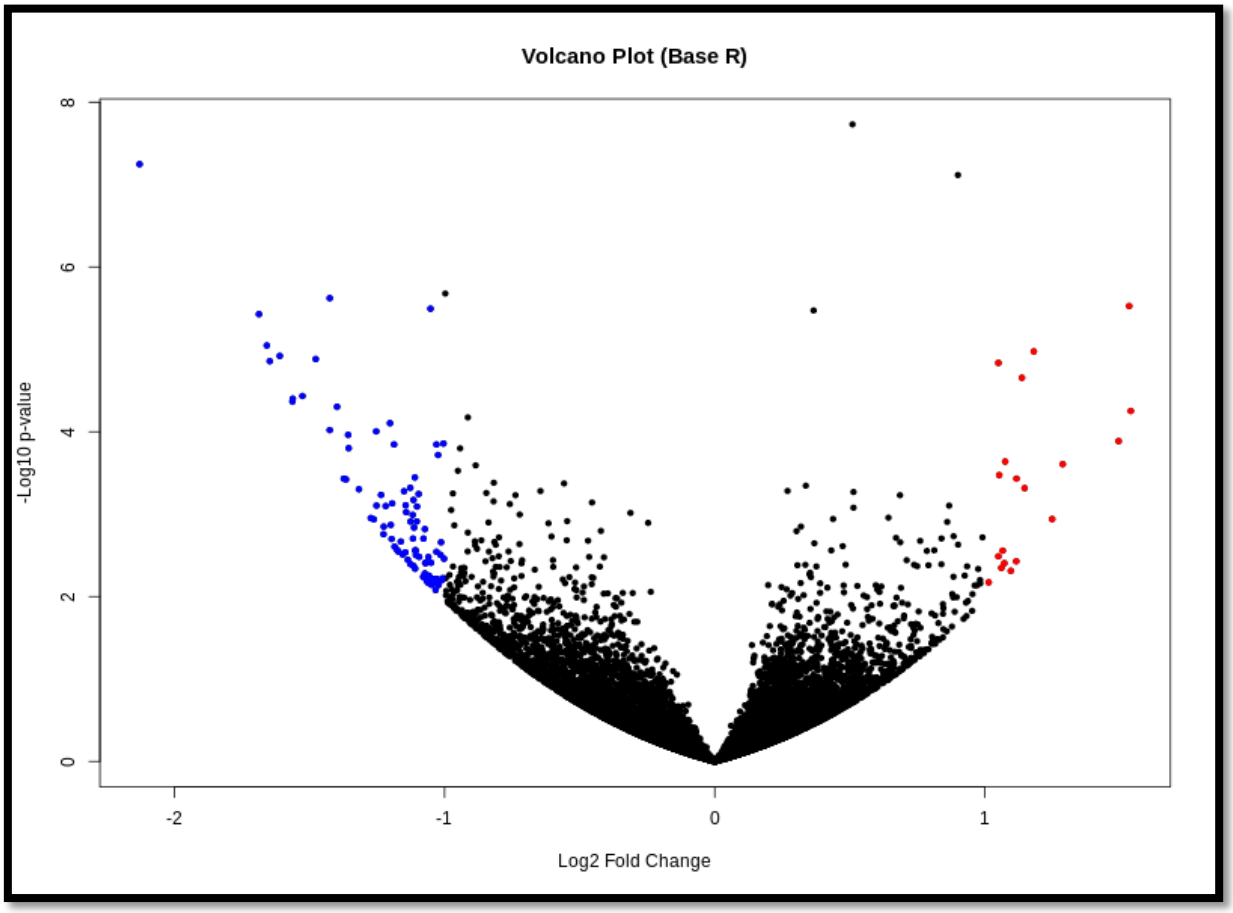

Figure 2: Volcano plot showing differentially expressed genes (generated using base R).

The volcano plot (Figure 2) combines both the magnitude of change (log2FC) and the statistical significance (–log10 p-value) for all genes. In this figure, three categories of genes are visible:

- **Upregulated genes (red dots):** Genes with log2FC > 1 and p-value < 0.01. These genes are located on the right side of the plot, far from zero, showing that they are expressed at much higher levels in the treatment condition compared to control.
- **Downregulated genes (blue dots):** Genes with log2FC < –1 and p-value < 0.01. These appear on the left side of the plot and represent those whose expression has significantly decreased under treatment.

- **Non-significant genes (black dots):** Most genes fall in the central cluster around log2FC = 0, meaning they show no meaningful difference in expression between conditions.

The shape of the plot is a characteristic "volcano," where only a relatively small proportion of genes reach both strong fold change and high statistical significance. This confirms that the biological response is likely driven by a focused set of genes rather than global transcriptome-wide changes.

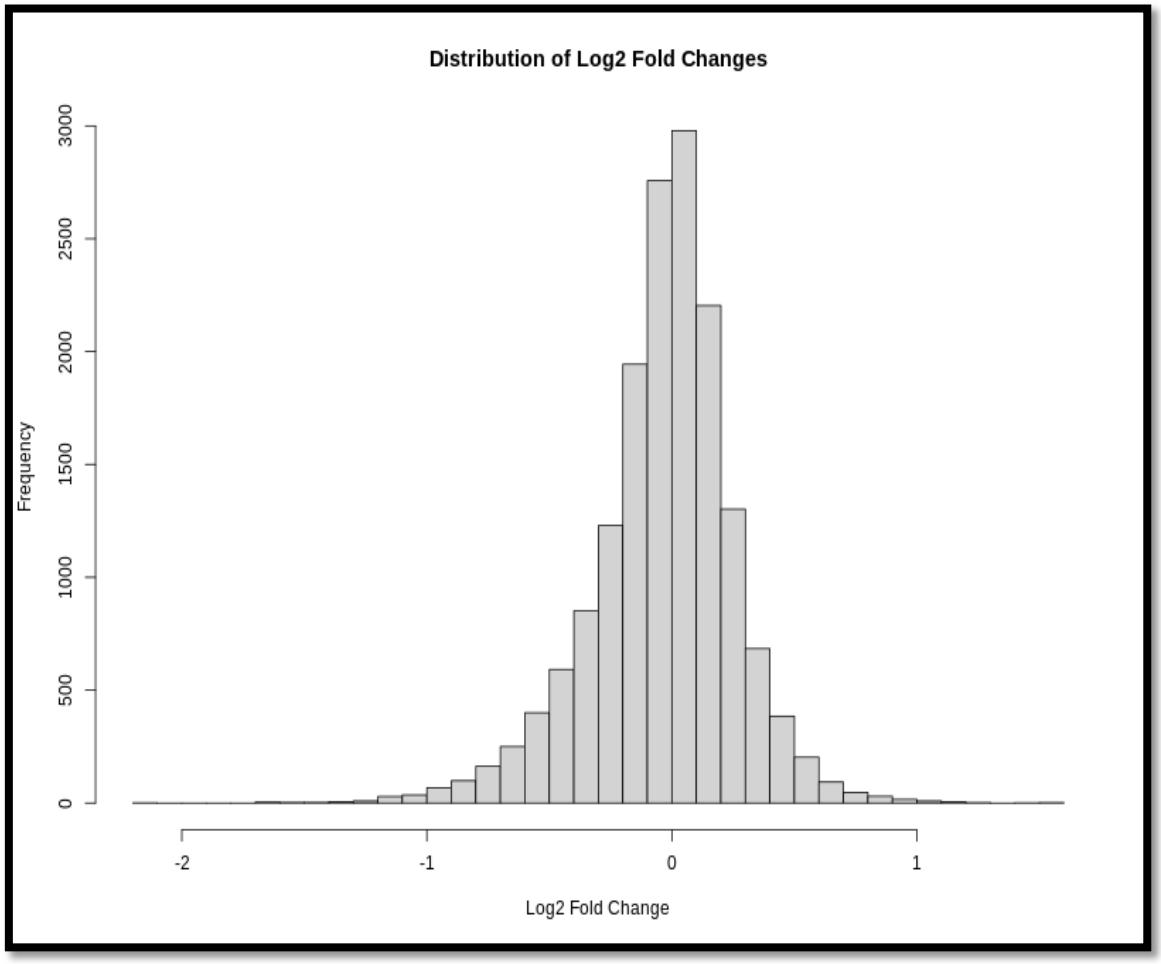

Figure 3: Histogram showing distribution of log2 fold changes (base R visualization).

The histogram (Figure 3) shows the distribution of log2 fold changes across all genes. The plot is approximately bell-shaped and centered around zero, which means most genes have minimal expression change between the two conditions.

- The **peak near zero** indicates that the majority of genes remain stable.
- The **tails extending on either side** represent genes with large positive or negative fold changes. Although these are fewer in number, they are biologically important because they represent the strongest candidates for functional analysis.

This histogram provides a complementary view to the volcano plot, confirming that while most of the genome remains unaffected, a small subset of genes stands out with strong regulation.

To highlight the most affected genes, a bar plot was generated (Figure 4). This plot compares the top five upregulated and downregulated genes based on log2FC values.

- **Upregulated genes (Blue bars):**
  - *DTHD1* (+1.54), *EMILIN2* (+1.53), and *PI16* (+1.50) showed the strongest increases in expression (Table 1).
  - *C4orf45* (+1.29) and *FAM180B* (+1.25) also ranked among the top upregulated genes, though with slightly smaller fold changes.

Table 1: Top 5 up-regulated genes

S. No.	Gene	log2FoldChange	pvalue	padj
1	EMILIN2	1.534	2.980e-06	0.006809
2	POU3F4	1.181	1.060e-05	0.015840
3	LOC285954	1.050	1.460e-05	0.015920
4	VEPH1	1.137	2.210e-05	0.022670
5	DTHD1	1.540	5.590e-05	0.043710

- **Downregulated genes (Red bars):**
  - *TBX5* (–2.13) was the most strongly downregulated gene, followed by *IFITM1* (–1.69), *TNN* (–1.66), *COL13A1* (–1.65), and *IFITM3* (–1.61) (Table 2).

Table 2: Top 5 down-regulated genes

S. No.	Gene	log2FoldChange	pvalue	padj
1	TBX5	-2.129	5.660e-08	0.0004191
2	IFITM1	-1.687	3.740e-06	0.0068090
3	LAMA2	-1.425	2.390e-06	0.0068090
4	CAV2	-1.052	3.210e-06	0.0068090
5	TNN	-1.658	8.970e-06	0.0147200

The magnitude of these changes, particularly the sharp downregulation of *TBX5*, suggests that these genes may play central roles in the biological pathways influenced by the treatment.

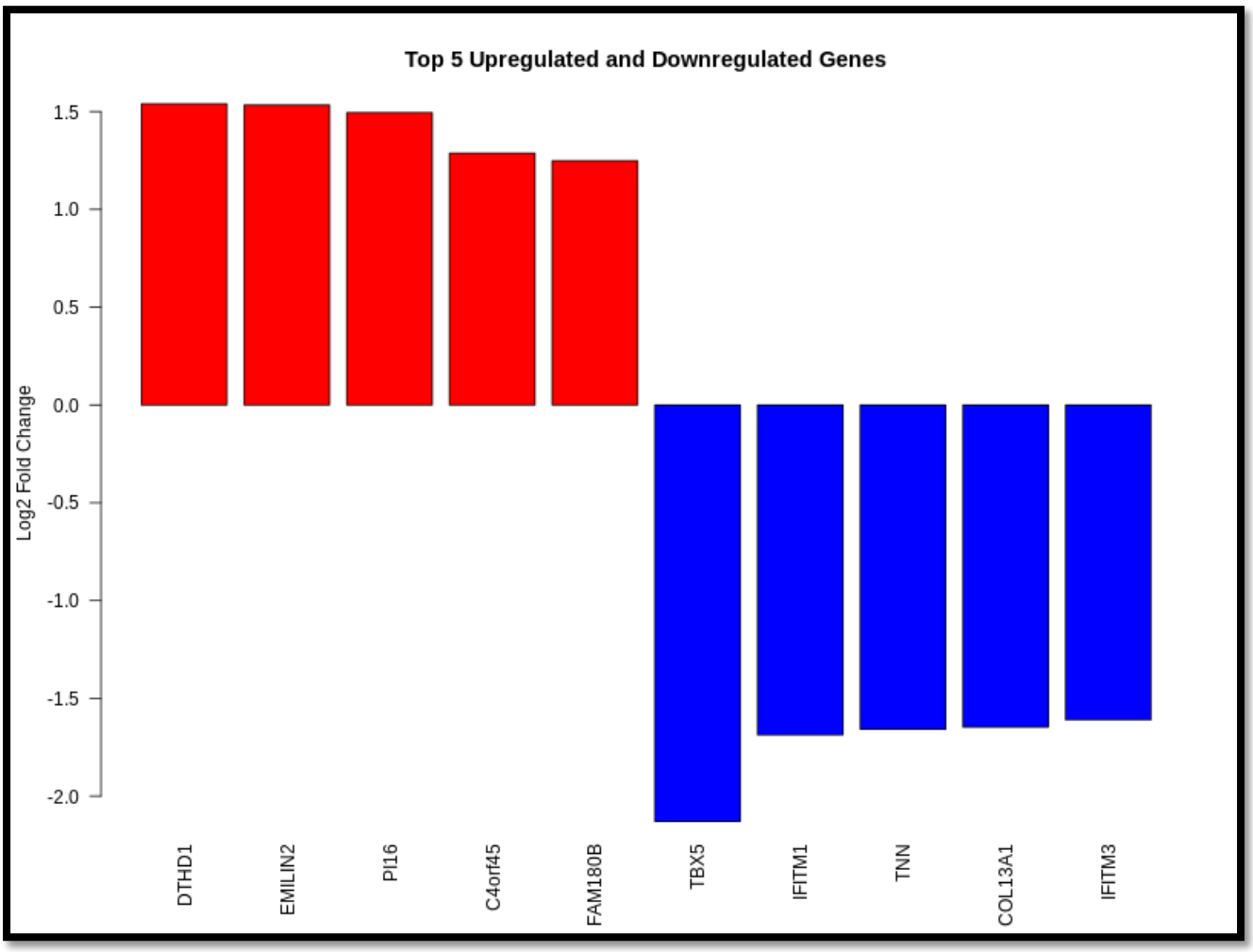

Figure 4: Top 5 Upregulated and Downregulated Genes

Functions of the Top 5 Upregulated Genes (Based on GeneCards):

EMILIN2 (Elastin Microfibril Interface Located Protein 2) is a extracellular glycoprotein involved in anchoring smooth muscle cells to elastic fibers and possibly in elastic fiber formation and vascular integrity. It also downregulates Wnt signaling, suppressing cell growth and migration in cancers like breast cancer. POU3F4 (POU Class 3 Homeobox 4) is the member of the POU-III class of neural transcription factors. It plays a crucial role in inner ear development, particularly the mesenchyme of the periotic bone. It also mediates neural differentiation and cell fate specification. Mutations are linked to X-linked non-syndromic deafness (DFN3). LOC285954 is annotated as a long non-coding RNA (lncRNA). VEPH1 (Ventricular Zone Expressed PH Domain–Containing 1) is a PH-domain protein predicted to bind phosphatidylinositol-5-phosphate. It involved in negative regulation of

SMAD-mediated TGF-β receptor signaling and likely acts as an intracellular adaptor modulating pathways such as TGF-β signaling. DTHD1 (Death Domain–Containing Protein 1) encodes a death domain–containing protein involved in formation of signaling complexes, notably within apoptotic pathways. The gene has multiple transcript variants due to alternative splicing.

Table 3: Functional summary of the top 5 upregulated and top 5 downregulated genes identified in the RNA-seq analysis based on GeneCards database annotations (Stelzer, Rosen et al. 2016).

Regulation Status	Gene	Key Function (GeneCards-derived)
Upregulated	**EMILIN2**	**ECM glycoprotein anchoring elastic fibers; suppresses Wnt signaling**
	POU3F4	**Neural TF for inner ear development and cell fate specification**
	LOC285954	**Long non-coding RNA; function unknown**
	VEPH1	**PH-domain adaptor; inhibits TGF-β/SMAD signaling**
	DTHD1	**Death domain protein; involved in apoptotic signaling**
Downregulated	**TBX5**	**T-box TF essential for heart and limb development**
	IFITM1	**Interferon-inducible membrane protein; antiviral, growth-regulating**
	LAMA2	**Laminin subunit for ECM structure; linked to muscular dystrophy**
	CAV2	**Caveolar scaffold protein in membrane signaling**
	TNN	**ECM glycoprotein for neurite and bone tissue dynamics**

Functions of the Top 5 Downregulated Genes (Based on GeneCards):

TBX5 (T-Box Transcription Factor 5) is a T-box family transcription factor characterized by a DNA-binding domain. Central to heart and limb development, including regulation of gene expression during cardiogenesis (e.g., binding NPPA promoter). It interacts with transcription partners like

NKX2-5 and GATA4. Mutations are associated with Holt-Oram syndrome, atrial septal defect, and related congenital defects. IFITM1 (Interferon-Induced Transmembrane Protein 1) is a part of the IFITM family this transmembrane protein is induced by interferons. It plays roles in osteoblast differentiation, restrains viral infections (e.g., HCV) by directing virions to lysosomal degradation, and contributes to interferon-gamma's antiproliferative effects by inhibiting ERK activation and inducing G1 arrest (p53-dependent). LAMA2 (Laminin Subunit Alpha-2) encodes the alpha-2 chain of laminin, a core component of the basement membrane ECM. It mediates cell attachment, migration, and tissue organization during embryonic development. Mutations in LAMA2 are known to cause congenital merosin-deficient muscular dystrophy (MDC1A). CAV2 (Caveolin-2) is a scaffolding protein of caveolae (membrane microdomains). It forms hetero-oligomeric complexes with CAV1, facilitating caveolae formation, recruiting CAVIN proteins, and regulating signal transduction via interaction with G-protein alpha subunits. TNN (Tenascin-N) is a member of the tenascin family, believed to be an extracellular matrix glycoprotein involved in neurite outgrowth and cell migration, particularly in hippocampal development and bone formation.

Discussion

The transcriptomic analysis carried out in this study highlights a distinctive set of genes with significant expression changes, offering insights into the underlying molecular adjustments triggered under the experimental condition.

Upregulated Genes: Signatures of Remodeling and Stress Adaptation

The upregulated gene set is dominated by EMILIN2, VEPH1, POU3F4, LOC285954, and DTHD1, each contributing unique but complementary roles to cellular dynamics. EMILIN2 encodes an extracellular matrix glycoprotein that contributes to vascular stability and elastic fibre assembly. Its role in modulating pericyte recruitment and ensuring vessel integrity has been strongly emphasized in recent literature (Fejza, Camicia et al. 2023). Elevated expression of EMILIN2 in the present dataset may signify a protective or reparative response, where the tissue attempts to counteract stress-induced structural instability by reinforcing its extracellular framework.

VEPH1 (Ventricular Zone Expressed PH Domain Containing 1) represents another upregulated gene with important implications. Functioning as an adaptor protein, VEPH1 modulates a variety of signalling pathways, including AKT, TGF-β, and Hippo. Its ability to suppress tumour angiogenesis and reduce vascularisation by impairing AKT activation has been reported in ovarian cancer models (Shathasivam, Kollara et al. 2017). Within the context of the present data, the increased expression

of VEPH1 may indicate a broader role in controlling cell proliferation and signalling fidelity during stress. POU3F4, a transcription factor, has a recognised function in developmental processes, particularly in the auditory system, with mutations linked to congenital deafness. Although its role in the experimental setting remains less direct, its upregulation could signify compensatory transcriptional adjustments in response to altered signalling environments (Stelzer, Rosen et al. 2016).

The gene LOC285954, annotated as a long non-coding RNA, is comparatively less characterized. Non-coding RNAs have increasingly been recognized as pivotal regulators of gene expression and chromatin states, and its induction here may reflect epigenetic fine-tuning in response to environmental cues.

Finally, DTHD1, encoding a death-domain-containing protein, may be tied to apoptotic signalling. While still under-characterized, its presence among the upregulated set suggests that programmed cell death or stress-induced survival checkpoints are being actively engaged. Together, the upregulated genes project a scenario of tissue remodeling, signalling regulation, and apoptotic control, hallmarks of cells attempting to maintain equilibrium under potentially destabilising conditions.

Downregulated Genes: Developmental and Structural Suppression

In contrast, the downregulated group, comprising TBX5, IFITM1, LAMA2, CAV2, and TNN, presents a picture of suppressed developmental, structural, and immune-associated functions. TBX5, a key transcription factor in the T-box family, is indispensable for cardiac and limb development. Its known association with Holt-Oram syndrome underscores its centrality in organogenesis and cardiac conduction. The sharp downregulation observed here may represent an attenuation of developmental or differentiation-linked programs, possibly diverting cellular energy towards maintenance rather than growth (Stelzer, Rosen et al. 2016).

IFITM1 (Interferon-Induced Transmembrane Protein 1) plays an essential role in the innate immune response, especially in restricting viral entry and modulating cell proliferation. Reduced expression of IFITM1 could signal a dampened antiviral state, leaving cells more vulnerable to external stressors or infections (Gomez-Herranz, Taylor et al. 2023). Moreover, downregulation of IFITM1 may indicate a compromised ability to control unchecked proliferation, consistent with immune-evasion strategies often seen in stressed or diseased tissues.

LAMA2, encoding the laminin α2 chain, is a fundamental structural component of the basement membrane. It contributes not only to the architecture of tissues but also to signalling events influencing cell adhesion, migration, and differentiation. Its suppression may weaken extracellular matrix stability, potentially altering the mechanical and biochemical cues received by cells. CAV2, a structural protein of caveolae, coordinates membrane trafficking and signalling. Its downregulation could disrupt caveolae-dependent signalling pathways, such as those linked to lipid regulation, receptor internalisation, and endocytosis. Given its role in maintaining membrane integrity, suppressed CAV2 expression suggests compromised communication between extracellular cues and intracellular responses (Parton and del Pozo 2013). Finally, TNN (Tenascin-N), part of the tenascin family of extracellular glycoproteins, is involved in neural development, cell adhesion, and tissue repair. Reduced expression of TNN may reflect diminished neural plasticity or a general decrease in tissue remodeling capacity (Joester and Faissner 2001). Together, these downregulated genes suggest a scenario where cells are deprioritizing developmental and structural programs, possibly as a trade-off to channel resources into survival and remodeling.

When viewed collectively, the differential expression pattern suggests a biological system caught between adaptive remodeling and functional suppression. On the one hand, upregulated genes such as EMILIN2 and VEPH1 appear to be orchestrating protective structural and signalling responses to preserve cellular homeostasis. On the other, downregulation of TBX5, LAMA2, and IFITM1 points toward reduced developmental signalling, weakened extracellular stability, and dampened immune readiness. This balance reflects a shift in cellular priorities, from growth and development toward short-term survival and compensatory repair.

Future Directions

While the current analysis provides valuable insights, functional validation through in vitro assays, animal models, or patient-derived data will be essential to confirm these transcriptomic predictions. Specifically, it will be important to test whether EMILIN2 and VEPH1 upregulation indeed supports tissue resilience, and whether suppression of TBX5 and LAMA2 leads to functional deficits in development or structure. Further integration of these findings with proteomics and pathway-level analysis could provide a more holistic picture of how these genes interact within larger regulatory networks. The analysis and visualization were re-implemented using base R plotting functions as per the HackBio feedback to ensure reproducibility and alignment with course methods.

Conclusion

This project explored differential gene expression using a pre-processed RNA-seq dataset, comparing diseased cell lines with those treated by a compound X. Through R-based analysis and visualization with a volcano plot, identified subsets of genes showing significant changes and annotated their functions using the GeneCards database.

The analysis highlighted that most genes remained unchanged, but a defined group displayed meaningful alterations. The top upregulated genes, **EMILIN2, VEPH1, POU3F4, LOC285954, and DTHD1** were linked to extracellular matrix stability, signaling regulation, neural development, and apoptotic pathways, suggesting an adaptive response aimed at maintaining cellular balance under stress. In contrast, the downregulated genes, **TBX5, IFITM1, LAMA2, CAV2, and TNN** were primarily involved in developmental regulation, immune defense, and structural organization. Their suppression indicates a cellular shift away from growth and differentiation toward conserving resources and remodeling processes. This shows that treatment did not broadly disrupt the transcriptome but selectively modulated key biological pathways. This targeted regulation reflects a balance between protective adaptation and functional suppression. While further experimental validation is essential, the study demonstrates how RNA-seq data combined with statistical filtering and functional annotation can yield meaningful biological insights.

References:

Cui, X. and G. A. Churchill (2003). "Statistical tests for differential expression in cDNA microarray experiments." Genome Biol **4**(4): 210.

Fejza, A., L. Camicia, G. Carobolante, E. Poletto, A. Paulitti, G. Schinello, E. Di Siena, R. Cannizzaro, R. V. Iozzo, G. Baldassarre, E. Andreuzzi, P. Spessotto and M. Mongiat (2023). "Emilin2 fosters vascular stability by promoting pericyte recruitment." Matrix Biol **122**: 18-32.

Gomez-Herranz, M., J. Taylor and R. D. Sloan (2023). "IFITM proteins: Understanding their diverse roles in viral infection, cancer, and immunity." J Biol Chem **299**(1): 102741.

Joester, A. and A. Faissner (2001). "The structure and function of tenascins in the nervous system." Matrix Biol **20**(1): 13-22.

Love, M. I., W. Huber and S. Anders (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." Genome Biol **15**(12): 550.

Parton, R. G. and M. A. del Pozo (2013). "Caveolae as plasma membrane sensors, protectors and organizers." Nat Rev Mol Cell Biol **14**(2): 98-112.

Shathasivam, P., A. Kollara, T. Spybey, S. Park, B. Clarke, M. J. Ringuette and T. J. Brown (2017). "VEPH1 expression decreases vascularisation in ovarian cancer xenografts and inhibits VEGFA and IL8 expression through inhibition of AKT activation." Br J Cancer **116**(8): 1065-1076.

Stelzer, G., N. Rosen, I. Plaschkes, S. Zimmerman, M. Twik, S. Fishilevich, T. I. Stein, R. Nudel, I. Lieder, Y. Mazor, S. Kaplan, D. Dahary, D. Warshawsky, Y. Guan-Golan, A. Kohn, N. Rappaport, M. Safran and D. Lancet (2016). "The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses." Curr Protoc Bioinformatics **54**: 1 30 31-31 30 33.

Wang, Z., M. Gerstein and M. Snyder (2009). "RNA-Seq: a revolutionary tool for transcriptomics." Nat Rev Genet **10**(1): 57-63.