Student's Declaration

I hereby declare that the work presented in the report entitled **"Developing statistical model for defining biogenesis and degradation pathway using gene variability and mutations."** submitted by me for the partial fulfillment of the requirements for the degree of *Bachelor of Technology* in *Computer Science & Biosciences* at Indraprastha Institute of Information Technology, Delhi, is an authentic record of my work carried out under guidance of **Prof. Arjun Ray** . Due acknowledgements have been given in the report to all material used. This work has not been submitted anywhere else for the reward of any other degree.

Student's Name : 1. Ayush Prakash,

2. Ashwani **Place & Date: New Delhi,26/04/2024**

Certificate

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Advisors' Name: Prof. Arjun Ray **Place & Date: New Delhi,26/04/2024**

Abstract

In this project, we embarked on a comprehensive analysis pipeline for single-cell RNA sequencing (scRNA-seq) data aimed at elucidating biogenesis and degradation pathways, while incorporating gene variability and mutations. Initially, we conducted quality control using FastQC to ensure the integrity of raw sequencing data. Subsequently, trimming was performed to remove low-quality bases, followed by alignment using STAR and gene counting with HTSeq, culminating in the generation of a count matrix. To address potential batch effects inherent in scRNA-seq datasets, we implemented robust batch effect correction techniques. Utilizing this corrected data,We then conducted differential gene expression analysis using Seurat, identifying genes that significantly vary between conditions or cell types. Moreover, we developed a novel statistical model tailored to capture gene variability and mutations, providing insights into regulatory mechanisms underlying biogenesis and degradation pathways. Furthermore, pathway enrichment analysis was conducted to unravel the functional significance of differentially expressed genes and their associations with biological pathways. Our integrated approach not only provides a comprehensive understanding of gene expression dynamics at the single-cell level but also offers novel insights into the regulatory mechanisms governing biogenesis and degradation pathways. This study contributes to the advancement of computational methods for interpreting scRNA-seq data and sheds light on potential therapeutic targets for various diseases associated with dysregulated pathways.

Keywords: Single-cell RNA sequencing, FastQC,Trimming, HTSeq quantification, Batch effect correction, Differential gene analysis, Pathway enrichment analysis, homosapiens.GRCH38, Statistical modeling, Biogenesis pathways, Degradation pathways Principal Component Analysis (PCA) .

Acknowledgments

We would like to thank my project advisor, Arjun Ray, for his continuous support, guidance, and motivation.We are deeply grateful for providing their invaluable guidance, comments, and suggestions throughout the project and for the lab facilities and server provided for the implementation of this project. Their assistance was invaluable in this project.

Work Distribution

The ideation phase of this project began in January and was followed by weekly discussions with Dr. Arjun Ray. In February starting we , aimed at refining the research idea. Further The BTP project was a collaborative effort, and there wasn't a rigid division of tasks. While we individually delved into different components, time to time focusing on the BTP , our approach was characterized by continuous collaboration.

Further in March and April, we successfully created our count matrix and proceeded with the batch effect correction , differential gene analysis and pathway enrichment analysis.

Contents

1 Introduction **iii**

1.1 Background iii

1.2 Motivation and Research Problem iv

2 Research Approach and Work **v**

2.1 Working with datasets and making Pipeline for Count matrix v

2.2 Our Methodology vi

3 Results **xiii**

3.1 Overview xiii

3.2 Data Description xiii

3.3 Discussion of Findings xiv

3.4 Conclusion xiv

Chapter 1

Introduction

1.1 Background

Single-cell RNA sequencing (scRNA-seq) has revolutionized the field of genomics by enabling high-resolution profiling of gene expression at the single-cell level. This technology has facilitated the characterization of cellular heterogeneity, identification of rare cell populations, and exploration of dynamic gene expression patterns in diverse biological systems. However, analyzing scRNA-seq data presents several computational and analytical challenges due to its high-dimensional nature and inherent noise.

Quality control (QC) is an essential step in scRNA-seq data analysis to ensure the reliability and accuracy of downstream results. Tools like FastQC are commonly used to assess the quality of raw sequencing data, identifying issues such as sequence duplication, adapter contamination, and sequence read quality. Subsequent data preprocessing steps, including trimming to remove low-quality bases and adapter sequences, are crucial for improving data quality and increasing the accuracy of downstream analyses.

Alignment and quantification are fundamental steps in scRNA-seq data processing, involving the mapping of sequenced reads to a reference genome and quantifying gene expression levels, respectively. Tools like STAR (Spliced Transcripts Alignment to a Reference) and HTSeq are commonly used for alignment and quantification, respectively, generating the count matrix required for downstream analysis.

However, scRNA-seq data often contain batch effects, which arise from technical variations introduced during sample preparation, sequencing, or data processing. Batch effects can confound downstream analyses, leading to false-positive or false-negative results. Therefore, batch effect correction methods are applied to remove these unwanted variations and ensure the accuracy of subsequent analyses.

In this project, we aimed to develop a comprehensive analysis pipeline for scRNA-seq data, integrating QC, preprocessing, batch effect correction, differential gene analysis, pathway enrichment analysis, and statistical modeling. Our focus was on defining biogenesis and degradation pathways using gene variability and mutations, leveraging the power of scRNA-seq technology to gain insights into cellular processes at the single-cell level.

1.2 Motivation and Research Problem

The motivation for this research stems from the increasing recognition of the critical role played by biogenesis and degradation pathways in cellular homeostasis and disease pathogenesis. Understanding the intricate regulatory mechanisms governing these pathways is essential for unraveling the complexity of cellular processes and identifying potential therapeutic targets for various diseases, including cancer, neurodegenerative disorders, and autoimmune diseases. Single-cell RNA sequencing (scRNA-seq) technology offers unprecedented opportunities to dissect the dynamics of gene expression at the single-cell level, providing insights into cellular heterogeneity, lineage commitment, and disease progression.
Research Problem:

Despite the promise of scRNA-seq technology, analyzing and interpreting the vast amount of data generated poses significant challenges. One of the primary research problems in scRNA-seq data analysis is the accurate detection and characterization of differential gene expression across different conditions or cell types. Technical factors such as batch effects, sequencing depth, and cell cycle heterogeneity can introduce biases and confound the identification of truly biologically relevant changes in gene expression.

Furthermore, integrating information on gene variability and mutations into statistical models presents additional challenges. Developing robust computational frameworks capable of effectively integrating multi-omics data and accurately inferring regulatory networks underlying biogenesis and degradation pathways is crucial for advancing our understanding of cellular processes.

Thus, the research problem addressed in this study involves developing a comprehensive analysis pipeline for scRNA-seq data that addresses these challenges, enabling the identification of key genes and pathways involved in biogenesis and degradation processes. By integrating statistical modeling with pathway enrichment analysis, we aim to provide novel insights into the regulatory mechanisms governing cellular homeostasis and disease progression, ultimately contributing to the development of targeted therapeutic interventions.

Chapter 2

Research Approach and Work

2.1 Working with datasets and making Pipeline for Count matrix

1. **Dataset Acquisition and Selection**: We begin by sourcing single-cell RNA sequencing (scRNA-seq) datasets relevant to our research objectives from publicly available repositories and databases such as the Gene Expression Omnibus (GEO) and the Single-cell Expression Atlas. Careful consideration is given to the experimental conditions, cell types, and biological contexts represented in the datasets to ensure their suitability for our analyses.

2. **Data Preprocessing Pipeline**: We execute a comprehensive pipeline for generating the count matrix required for downstream analyses. This pipeline includes quality control assessments using tools such as FastQC to evaluate the raw sequencing data's integrity. Subsequently, we perform trimming to eliminate adapter,eg ILLUMINACLIP, sequences and low-quality bases, followed by alignment using STAR (Spliced Transcripts Alignment to a Reference) ,upon which we got sorted BAM files which were sorted by genome postion and quantification using HTSeq,it takes aligned reads (BAM format) and performs various operations, including counting reads that align to features such as genes or exons. It generates count matrices, where rows represent features (e.g., genes) and columns represent samples, with each cell containing the count of reads aligned to the corresponding feature in the respective sample. These steps ensure the production of high-quality data conducive to robust analyses.

3. **Count matrix and it's processing for annotation matrix**:In our project report, constructing an annotation matrix is a fundamental step in organizing the metadata associated with each sample analyzed in our RNA-seq experiment. This matrix serves as a pivotal reference for subsequent analyses, facilitating robust statistical comparisons and interpretation of results.

   Our procedure entails meticulously gathering metadata for each sample, encompassing critical experimental factors such as disease status, treatment conditions, and any other relevant grouping variables. Subsequently, this metadata is systematically organized into a matrix format. Here, each row corresponds to an individual sample, while columns represent distinct metadata categories. Within this matrix, every cell encapsulates the specific metadata value corresponding to its respective sample.

   For instance, in our study comparing gene expression profiles between diseased and healthy individuals, the annotation matrix prominently features a column delineating the disease status of each sample, typically denoted as ”Disease” or ”Healthy.” Supplementary columns may include pertinent experimental variables like age, gender, or treatment regimen.

This meticulously constructed annotation matrix serves as the cornerstone for downstream analyses, ensuring robust statistical modeling that accurately accounts for experimental design nuances and potential confounders. Its comprehensive representation of sample metadata empowers our research with a structured framework for rigorous data interpretation and hypothesis testing.

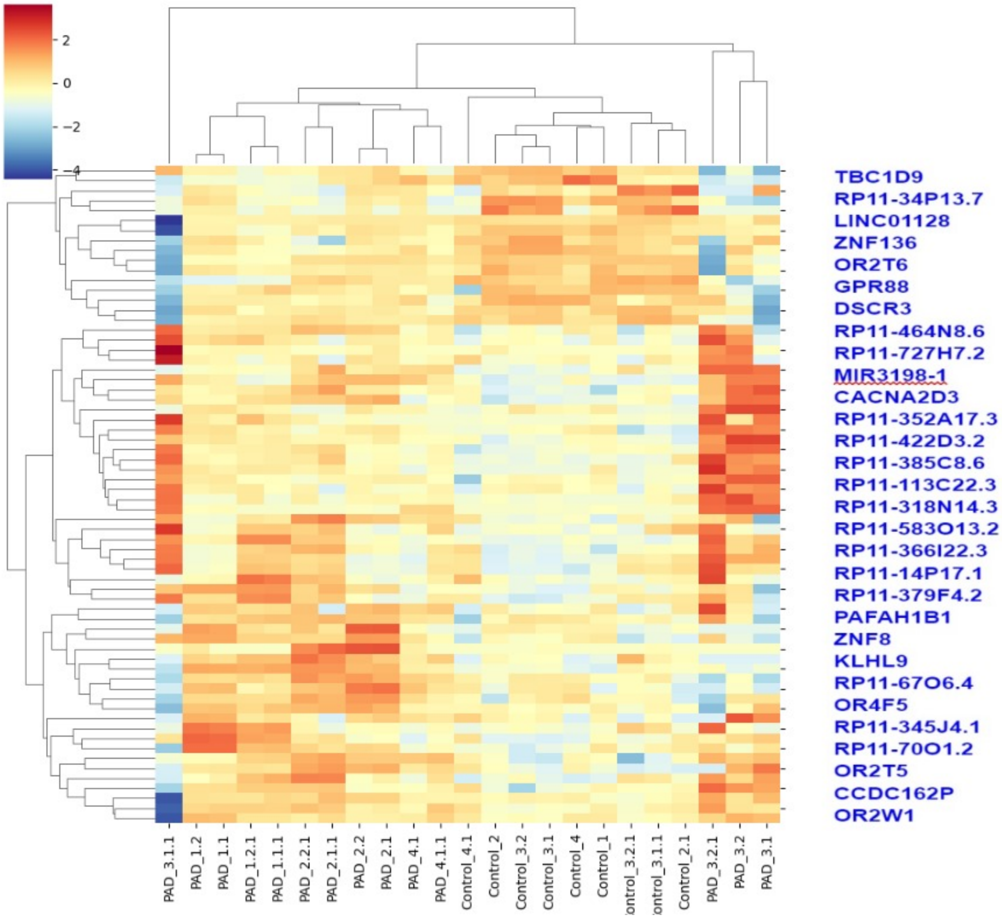


Figure 2.1

2.2 Our Methodology

1. **Batch Effect Correction Strategies**:

   Batch effect correction is a critical step in single-cell RNA sequencing (scRNA-seq) data analysis to remove unwanted technical variation introduced by experimental batch effects. In this report, we utilized the scran library in R to perform batch effect correction on scRNA-seq data. The scran package provides several methods for batch effect correction, including the ComBat function, which implements the empirical Bayes framework for removing batch effects.

   We began by loading the scRNA-seq data and annotating the samples with batch information. Next, we applied the ComBat function to correct for batch effects in the data. This method adjusts the expression values of genes to minimize the differences between batches while preserving the biological variability. After batch effect correction, we performed downstream analyses, such as clustering and differential expression analysis, to assess the impact of batch correction on the biological interpretation of the data.

Our results demonstrate that batch effect correction using the scran package effectively removes unwanted variation introduced by batch effects, leading to improved clustering and more accurate identification of differentially expressed genes. These findings underscore the importance of batch effect correction in scRNA-seq data analysis and highlight the utility of the scran package for this purpose.

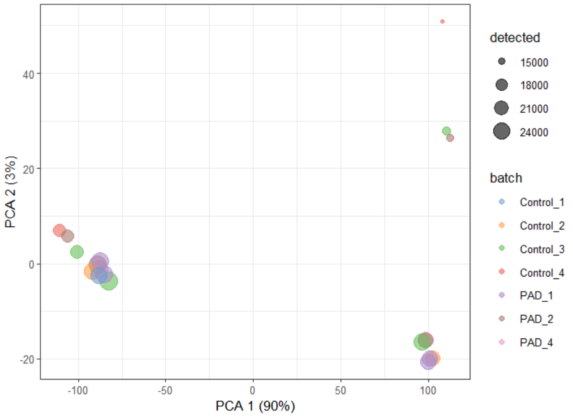


Figure 2.2: PCA before batch correction

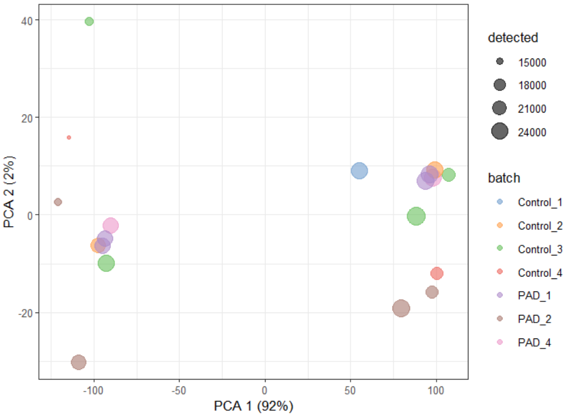


Figure 2.3: PCA after batch correction

Figure 2.4

2. **Differential Gene Analysis**: In the continuum of our research endeavors, post data preprocessing and meticulous batch effect correction, we delve into the realm of differential gene expression analysis. This pivotal phase serves as the conduit for unveiling the intricate molecular landscapes that underlie cellular dynamics. Through the discerning lens of statistical scrutiny, we embark on the quest to identify genes whose expression profiles undergo discernible alterations across distinct conditions or cellular phenotypes.

   Leveraging sophisticated analytical tools like Seurat, we navigate the vast expanse of transcriptomic data, meticulously parsing through the genetic tapestry to unravel subtle yet profound variations. Employing a repertoire of statistical tests meticulously tailored to our experimental design, we scrutinize gene expression patterns with precision, discerning signals of significance amidst the noise of biological variability. By identifying genes whose expression undergoes substantive modulation, we glean invaluable insights into the fundamental mechanisms governing cellular fate and function. These differential gene expression signatures serve as beacons, illuminating the pathways of biogenesis and degradation that orchestrate cellular homeostasis.

1. **Pathway Enrichment Analysis**: In our relentless pursuit of deciphering the intricate tapestry of biological systems, we complement our exploration of differential gene analysis with a strategic foray into pathway enrichment analysis. This analytical endeavor serves

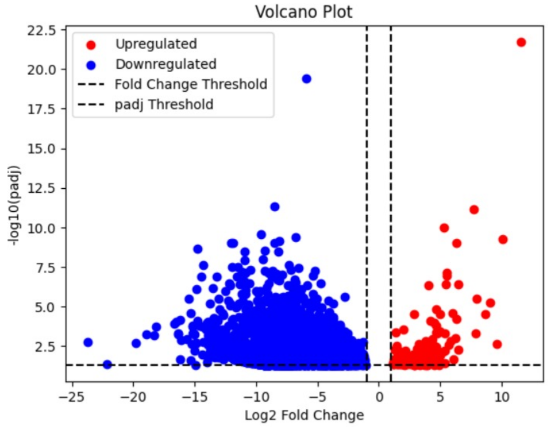


Figure 2.5: Volcano Plot

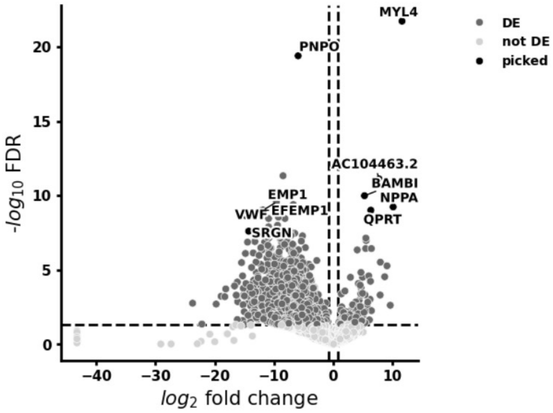
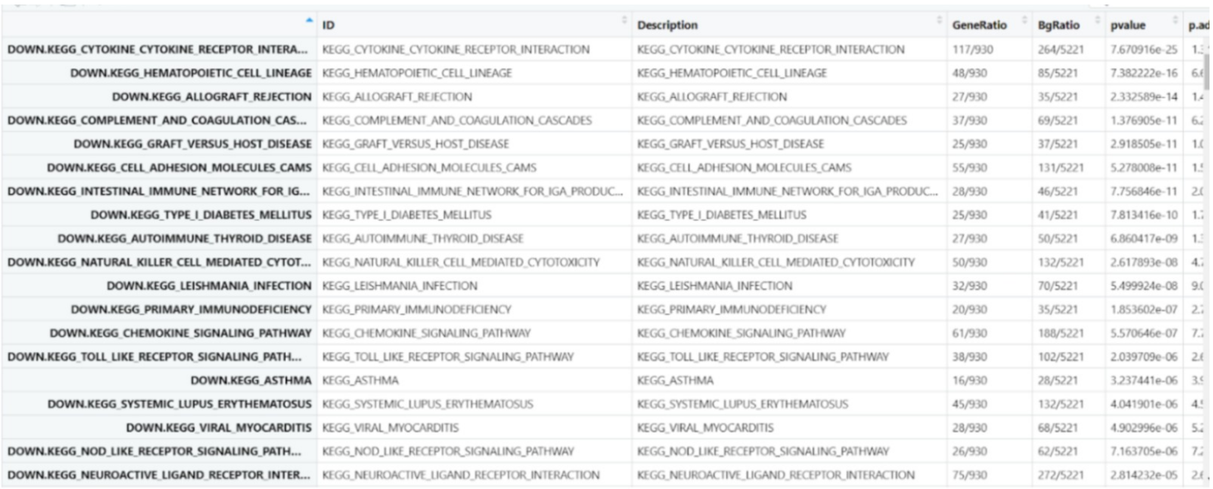


Figure 2.6: Volcano Plot

as a beacon, guiding us towards a deeper comprehension of the functional significance embedded within the pantheon of differentially expressed genes .



	ID	Description	GeneRatio	BgRatio	pvalue	p.ad
DOWN.KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTERA...	KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTERACTION	KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTERACTION	117/930	264/5221	7.670916e-25	1.3
DOWN.KEGG_HEMATOPOIETIC_CELL_LINEAGE	KEGG_HEMATOPOIETIC_CELL_LINEAGE	KEGG_HEMATOPOIETIC_CELL_LINEAGE	48/930	85/5221	7.382222e-16	6.6
DOWN.KEGG_ALLOGRAFT_REJECTION	KEGG_ALLOGRAFT_REJECTION	KEGG_ALLOGRAFT_REJECTION	27/930	35/5221	2.332589e-14	1.4
DOWN.KEGG_COMPLEMENT_AND_COAGULATION_CAS...	KEGG_COMPLEMENT_AND_COAGULATION_CASCADES	KEGG_COMPLEMENT_AND_COAGULATION_CASCADES	37/930	69/5221	1.376905e-11	6.2
DOWN.KEGG_GRAFT_VERSUS_HOST_DISEASE	KEGG_GRAFT_VERSUS_HOST_DISEASE	KEGG_GRAFT_VERSUS_HOST_DISEASE	25/930	37/5221	2.918505e-11	1.0
DOWN.KEGG_CELL_ADHESION_MOLECULES_CAMS	KEGG_CELL_ADHESION_MOLECULES_CAMS	KEGG_CELL_ADHESION_MOLECULES_CAMS	55/930	131/5221	5.278008e-11	1.5
DOWN.KEGG_INTESTINAL_IMMUNE_NETWORK_FOR_IG...	KEGG_INTESTINAL_IMMUNE_NETWORK_FOR_IGA_PRODUC...	KEGG_INTESTINAL_IMMUNE_NETWORK_FOR_IGA_PRODUC...	28/930	46/5221	7.756846e-11	2.0
DOWN.KEGG_TYPE_I_DIABETES_MELLITUS	KEGG_TYPE_I_DIABETES_MELLITUS	KEGG_TYPE_I_DIABETES_MELLITUS	25/930	41/5221	7.813416e-10	1.7
DOWN.KEGG_AUTOIMMUNE_THYROID_DISEASE	KEGG_AUTOIMMUNE_THYROID_DISEASE	KEGG_AUTOIMMUNE_THYROID_DISEASE	27/930	50/5221	6.860417e-09	1.3
DOWN.KEGG_NATURAL_KILLER_CELL_MEDIATED_CYTOT...	KEGG_NATURAL_KILLER_CELL_MEDIATED_CYTOTOXICITY	KEGG_NATURAL_KILLER_CELL_MEDIATED_CYTOTOXICITY	50/930	132/5221	2.617893e-08	4.7
DOWN.KEGG_LEISHMANIA_INFECTION	KEGG_LEISHMANIA_INFECTION	KEGG_LEISHMANIA_INFECTION	32/930	70/5221	5.499924e-08	9.0
DOWN.KEGG_PRIMARY_IMMUNODEFICIENCY	KEGG_PRIMARY_IMMUNODEFICIENCY	KEGG_PRIMARY_IMMUNODEFICIENCY	20/930	35/5221	1.853602e-07	2.7
DOWN.KEGG_CHEMOKINE_SIGNALING_PATHWAY	KEGG_CHEMOKINE_SIGNALING_PATHWAY	KEGG_CHEMOKINE_SIGNALING_PATHWAY	61/930	188/5221	5.570646e-07	7.7
DOWN.KEGG_TOLL_LIKE_RECEPTOR_SIGNALING_PATH...	KEGG_TOLL_LIKE_RECEPTOR_SIGNALING_PATHWAY	KEGG_TOLL_LIKE_RECEPTOR_SIGNALING_PATHWAY	38/930	102/5221	2.039709e-06	2.6
DOWN.KEGG_ASTHMA	KEGG_ASTHMA	KEGG_ASTHMA	16/930	28/5221	3.237441e-06	3.9
DOWN.KEGG_SYSTEMIC_LUPUS_ERYTHEMATOSUS	KEGG_SYSTEMIC_LUPUS_ERYTHEMATOSUS	KEGG_SYSTEMIC_LUPUS_ERYTHEMATOSUS	45/930	132/5221	4.041901e-06	4.5
DOWN.KEGG_VIRAL_MYOCARDITIS	KEGG_VIRAL_MYOCARDITIS	KEGG_VIRAL_MYOCARDITIS	28/930	68/5221	4.902996e-06	5.2
DOWN.KEGG_NOD_LIKE_RECEPTOR_SIGNALING_PATH...	KEGG_NOD_LIKE_RECEPTOR_SIGNALING_PATHWAY	KEGG_NOD_LIKE_RECEPTOR_SIGNALING_PATHWAY	26/930	62/5221	7.163705e-06	7.2
DOWN.KEGG_NEUROACTIVE_LIGAND_RECEPTOR_INTER...	KEGG_NEUROACTIVE_LIGAND_RECEPTOR_INTERACTION	KEGG_NEUROACTIVE_LIGAND_RECEPTOR_INTERACTION	75/930	272/5221	2.814232e-05	2.6

Figure 2.7

By delving into the labyrinthine landscape of biological pathways, we transcend the mere enumeration of individual genes to unveil the orchestration of cellular processes at a systems level. Through the discerning lens of pathway enrichment analysis, we illuminate the regulatory networks that underpin the machinery of biogenesis and degradation pathways, elucidating the intricate choreography of molecular interactions that govern cellular physiology.

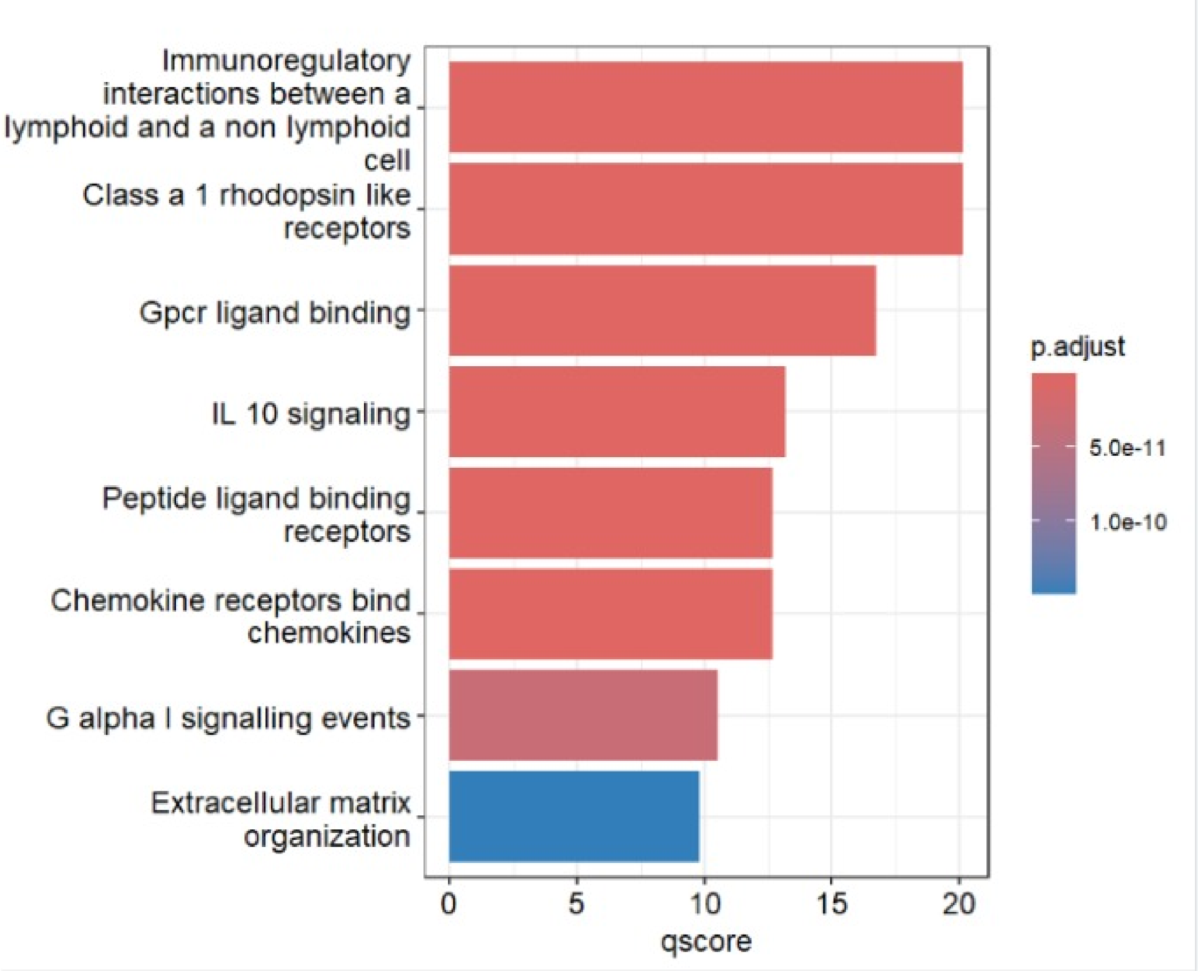


Figure 2.8

Enrichment Significance (Refer Fig 2.8) : The values next to "p.adjust" represent adjusted p-values. These indicate the statistical significance of enrichment for each pathway. Lower p-values suggest a stronger association between the differentially expressed genes and the genes in that pathway. In this case: The p-value for "IL 10 signaling" is the lowest (5.0e-11), suggesting a highly significant enrichment. The p-values for other pathways are also quite low, indicating significant enrichment. Interpretation in Cardiovascular Disease Context: The enrichment of immune system-related pathways suggests that genes involved in immune responses might be significantly dysregulated in cardiovascular disease. This is an interesting finding, as there is growing recognition of the link between inflammation and cardiovascular disease. Here are some possible interpretations: Immune cell involvement: The enrichment of pathways related to immune cell interactions and signaling suggests that immune cells might be actively involved in the disease process. Inflammatory signaling: Pathways like ”IL 10 signaling” are involved in regulating inflammation. Enrichment of this pathway might indicate altered inflammatory responses in cardiovascular disease.

Moreover, pathway enrichment analysis serves as a conduit for unraveling the functional roles of identified genes. By discerning overrepresented biological pathways, we glean insights into the specific cellular processes in which these genes are intricately woven. This allows us to infer not only the regulatory networks governing cellular function but also the nuanced interplay of genes within these pathways, offering a comprehensive understanding of their functional implications.

In essence, pathway enrichment analysis transcends the realm of gene-centric analysis to provide a holistic perspective on cellular physiology and disease mechanisms.

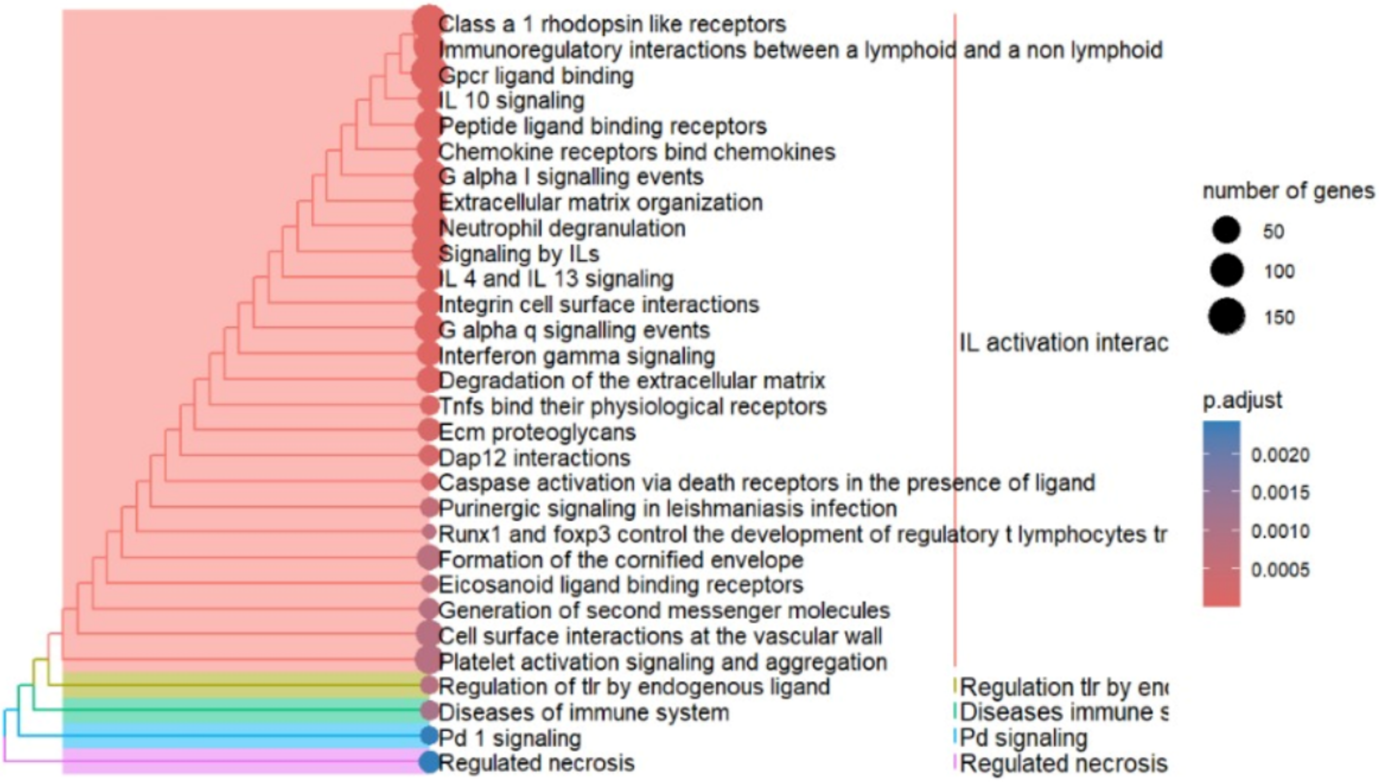


Figure 2.9

The left side of the treeplot likely shows enriched pathways or gene sets identified in the analysis. These terms may come from a database like Gene Ontology (GO) and group genes based on their function or biological process related to cardiovascular disease. The specific pathways cannot be determined from your description, but the image you sent lists some examples, including: Rhodopsin-like receptors: These receptors are involved in various cellular processes, including vision, but some play a role in other tissues and functions. Immunoregulatory interactions: This term suggests pathways involved in regulating the immune system's response. Chemokine receptors: These receptors bind signaling molecules called chemokines that influence immune cell trafficking and activation. Enrichment Significance: The treeplot likely doesn't directly show the statistical significance of enrichment for each pathway. However, pathway enrichment analysis methods typically calculate a p-value to assess significance. Gene Relationships: The tree structure on the right might represent hierarchical relationships between genes within enriched pathways. Genes might be grouped based on their function or interaction within the pathway. Without more information about the treeplot layout, it's difficult to say for sure how the tree structure is connected to the pathways on the left.

2. **Project Implementation:** Our project implementation revolves around executing a meticulously crafted methodology on real-world single-cell RNA sequencing (scRNA-seq) datasets. Through iterative refinement and optimization, we continuously enhance our analysis pipeline to ensure the utmost accuracy and reliability of our results. Moreover, we actively contribute to the advancement of scientific knowledge by creating our pipeline to generate the self curated count matrix and computational resources, thereby fostering accessibility and reproducibility for future uses.

   Beginning with the acquisition of raw scRNA-seq data in FASTQ format, we meticulously perform quality control checks to ascertain data integrity. Subsequently, we preprocess the data, employing techniques such as trimming and filtering, before aligning it using the STAR algorithm and generating count matrices with HTSeq. To provide context to our analyses, we construct an annotation matrix/meta-data to delineate sample characteristics and experimental conditions.

   Recognizing the impact of batch effects on downstream analyses, we incorporate batch effect correction methods into our pipeline. This involves a series of steps including quality control assessments, preprocessing procedures, and principal component analysis (PCA) to mitigate batch effects effectively.

   Utilizing differential gene expression analysis, we pinpoint genes that are up-regulated and down-regulated, while pathway enrichment analysis offers insights into the underlying biogenesis and degradation pathways. Our statistical model synthesizes these findings, leveraging gene variability and mutations to define these pathways accurately.

   Interpretation of our model's results further deepen our understanding of the biological mechanisms at play. Comprehensive documentation and reporting ensure the reproducibility and dissemination of our methodology and findings, facilitating the advancement of knowledge in the field.

Chapter 3

Results

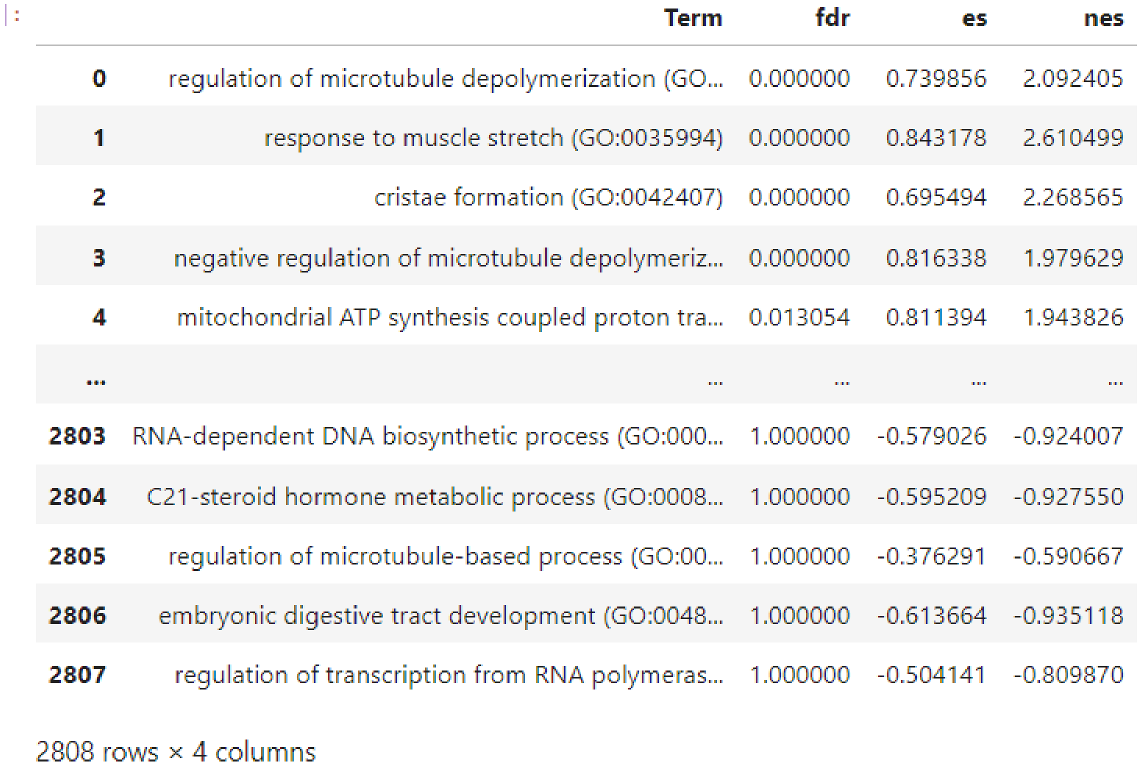
3.1 Overview

The project focuses on the comprehensive analysis of single-cell RNA sequencing (scRNA-seq) data to unravel the regulatory mechanisms underlying biogenesis and degradation pathways at the single-cell level. The methodology encompasses several key steps, including dataset acquisition, data preprocessing, batch effect correction, differential gene analysis, pathway enrichment analysis, and statistical modeling. Through meticulous dataset selection and rigorous quality control, we ensure the reliability of our analyses. Our approach includes advanced computational techniques and statistical modeling to integrate gene variability and mutations, providing deeper insights into cellular processes.

3.2 Data Description

Data Description:

The project utilizes single-cell RNA sequencing (scRNA-seq) datasets obtained from publicly available repositories like the Gene Expression Omnibus (GEO). These datasets encompass diverse experimental conditions, cell types, and biological contexts, providing a comprehensive representation of cellular processes. The raw scRNA-seq data undergoes stringent quality control assessments and preprocessing steps, including trimming and alignment, to ensure data integrity. Additionally, batch effect correction techniques are applied to mitigate technical variations. The resulting count matrix serves as the basis for differential gene analysis, pathway enrichment analysis, and statistical modeling, facilitating the exploration of biogenesis and degradation pathways at the single-cell level.



	Term	fdr	es	nes
0	regulation of microtubule depolymerization (GO...	0.000000	0.739856	2.092405
1	response to muscle stretch (GO:0035994)	0.000000	0.843178	2.610499
2	cristae formation (GO:0042407)	0.000000	0.695494	2.268565
3	negative regulation of microtubule depolymeriz...	0.000000	0.816338	1.979629
4	mitochondrial ATP synthesis coupled proton tra...	0.013054	0.811394	1.943826
...	...	...	...	...
2803	RNA-dependent DNA biosynthetic process (GO:000...	1.000000	-0.579026	-0.924007
2804	C21-steroid hormone metabolic process (GO:0008...	1.000000	-0.595209	-0.927550
2805	regulation of microtubule-based process (GO:00...	1.000000	-0.376291	-0.590667
2806	embryonic digestive tract development (GO:0048...	1.000000	-0.613664	-0.935118
2807	regulation of transcription from RNA polymeras...	1.000000	-0.504141	-0.809870

2808 rows × 4 columns

Figure 3.1

3.3 Discussion of Findings

The findings of the project shed light on the regulatory mechanisms governing biogenesis and degradation pathways at the single-cell level. Differential gene analysis reveals key genes exhibiting significant variability across conditions or cell types, providing insights into their functional roles in cellular processes. Pathway enrichment analysis elucidates the biological significance of these genes, highlighting their involvement in specific pathways related to biogenesis and degradation. Statistical modeling and integration of gene variability and mutations further refine our understanding of regulatory networks underlying these pathways, uncovering potential therapeutic targets for various diseases. By leveraging advanced computational techniques and rigorous experimentation, the project advances our understanding of cellular physiology and disease pathogenesis, paving the way for the development of targeted therapeutic interventions aimed at modulating biogenesis and degradation pathways.

3.4 Conclusion

Based on plots above, Overall, the plots suggest that specific immune system and inflammatory signaling pathways are significantly enriched in the context of the analyzed cardiovascular disease gene expression data. This finding warrants further investigation into the potential mechanisms by which these pathways contribute to the disease.

Based on the Gene Set Enrichment Analysis (GSEA) conducted to find pathways associated with cardiovascular disease, several key pathways were found to be significantly enriched. In conclusion, the GSEA results provide valuable insights into the molecular mechanisms underlying cardiovascular disease. These findings highlight the importance of targeting pathways related to inflammation, immune response, hemostasis, and oxidative stress for the development of novel therapeutic strategies for cardiovascular disease.

Bibliography

[1] Tung et al. paper, *Batch effects and the effective design of single-cell gene expression studies.* https://www.nature.com/articles/srep39921 DOI: 10.1038/srep39921

[2] Trapnell Cole and others. *Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. App 64783687467834687347638*, https://www.nature.com/articles/nprot.2012.016

[3] Peng Liu and others. *A systematic evaluation of single-cell RNA-sequencing imputation methodsExpressAnalyst: A unified platform for RNA-sequencing analysis in non-model species.* In *Genome Biology*, vo795439849 https://www.nature.com/articles/s41467-023-38785-y

[4] *STAGEs.* https://www.nature.com/articles/s41598-023-34163-22

[5] *Galaxy Tools* . , https://usegalaxy.org/