

# **DIFFERENTIAL EXPRESSION ANALYSIS OF PANCREATIC CANCER**

Submitted in partial fulfillment of the requirements for the award of  
Master of Science in Bioinformatics and Data Science

By

**AKILAN S D (42739002)**



**DEPARTMENT OF BIOINFORMATICS  
SCHOOL OF BIO AND CHEMICAL ENGINEERING**

# **SATHYABAMA**

**INSTITUTE OF SCIENCE AND TECHNOLOGY (DEEMED TO BE UNIVERSITY)**

Accredited with Grade “A++” by NAAC | 12B Status by UGC | Approved by  
AICTE

**JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI - 600 119**

**April - 2024**



# SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY  
(DEEMED TO BE UNIVERSITY)

## DEPARTMENT OF BIOINFORMATICS

### BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **AKILAN S D (42739002)**, who carried out the final year project entitled “**DIFFERENTIAL EXPRESSION ANALYSIS OF PANCREATIC CANCER**” under my supervision from September 2023 to April 2024.

**Dr. H. JEMMY CHRISTY**

**Internal Guide**

**Dr. H. JEMMY CHRISTY**

**External Guide**

**Head of the Department**

**Dr. H. JEMMY CHRISTY**

---

**Submitted for Viva voce Examination held on 20/04/2024**

**Internal Examiner**

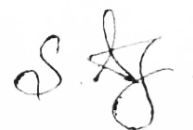
**External Examiner**

## **DECLARATION**

I **AKILAN S D (42739002)** hereby declare that the Project Report entitled “**DIFFERENTIAL EXPRESSION ANALYSIS OF PANCREATIC CANCER**” done by me under the guidance of Dr. H. JEMMY CHRISTY at Sathyabama Institute of Science and Technology, Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai –600 119, Tamil Nadu, India, and is submitted in partial fulfillment of the requirements for the award of Master of Science in Bioinformatics and Data Science.

**DATE: 20/04/2024**

**PLACE: CHENNAI**

A handwritten signature in black ink, appearing to be 'S. D.', is written above the signature line.

**SIGNATURE OF THE CANDIDATE**

## **ACKNOWLEDGEMENTS**

I am pleased to acknowledge my sincere thanks to the **Board of Management of SATHYABAMA** for their kind encouragement in doing this project for completing it successfully. I am grateful to them.

I extend my sincere gratitude to **Dr. REKHA CHAKRAVARTHY, Dean of the School of Science and Humanities**, and I am deeply grateful for her mentorship and leadership, which have been a constant source of inspiration. Her dedication to fostering a conducive academic environment has been instrumental in the successful completion of this thesis.

I convey my thanks to **DR. H. JEMMY CHRISTY, Head of the Department, Department of Bioinformatics**, for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr. H. JEMMY CHRISTY** for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the Department of Bioinformatics who were helpful in many ways for the completion of the project

## **ABSTRACT**

Pancreatic cancer is a devastating disease with a poor prognosis, often diagnosed at advanced stages when treatment options are limited. Early detection and effective therapeutic strategies are essential for improving patient outcomes. Biomarkers play a critical role in the diagnosis, prognosis, and treatment of pancreatic cancer. This project aimed to identify potential biomarkers for pancreatic cancer through differential expression analysis and pathway analysis of gene expression data obtained from The Cancer Genome Atlas (TCGA) database. Data preprocessing was conducted using the GDCprepare package in R, followed by differential expression analysis and GO enrichment analysis of biological process. The results identified a gene that were differentially expressed in pancreatic cancer samples compared to normal tissues, RN7SL1 which is identified as potential biomarkers. Additionally, and GO enrichment analysis of biological process revealed dysregulation of several critical process implicated in pancreatic cancer pathogenesis. These findings provide valuable insights into the molecular mechanisms underlying pancreatic cancer and highlight potential biomarkers and therapeutic targets for further investigation.

**KEYWORDS:** Pancreatic cancer, RNA-Seq, R programming, Differential Gene Expression Analysis, GO Enrichment Analysis

## TABLE OF CONTENTS

| CHAPTER NO | TITLE                                                         | PAGE NO |
|------------|---------------------------------------------------------------|---------|
|            | <b>ABSTRACT</b>                                               | i       |
|            | <b>LIST OF TABLES</b>                                         | iv      |
|            | <b>LIST OF FIGURES</b>                                        | iv      |
|            | <b>LIST OF ABBREVIATIONS</b>                                  | v       |
| <b>1</b>   | <b>INTRODUCTION</b>                                           | 1       |
| 1.1        | PANCREATIC CANCER                                             | 1       |
| 1.2        | R PROGRAM                                                     | 3       |
| 1.3        | DIFFERENTIAL EXPRESSION ANALYSIS                              | 4       |
| 1.4        | BIOCONDUCTOR                                                  | 6       |
| 1.5        | TCGABIOLINKS                                                  | 8       |
| 1.6        | DESEQ2                                                        | 9       |
| 1.7        | TCGA DATABASE                                                 | 10      |
| 1.8        | GO ENRICHMENT ANALYSIS OF BIOLOGICAL PROCESS                  | 11      |
| <b>2</b>   | <b>REVIEW OF LITERATURE</b>                                   | 13      |
| <b>3</b>   | <b>AIM AND SCOPE</b>                                          | 15      |
| 3.1        | AIM                                                           | 15      |
| 3.2        | OBJECTIVES                                                    | 15      |
| 3.3        | SCOPE                                                         | 15      |
| <b>4</b>   | <b>METHODOLOGY</b>                                            | 16      |
| 4.1        | DATA COLLECTION FOR THE DEA OF PANCREATIC CANCER              | 16      |
| 4.2        | DATA SETS OF THE PANCREATIC CANCER                            | 16      |
| 4.3        | DATA RETRIVAL FROM TCGA DATABASE USING R                      | 16      |
| 4.4        | DATA PRE PROCESSING OF RETRIVED DATA SETS USING R             | 16      |
| 4.5        | DIFFERENTIAL EXPRESSION ANALYSIS OF PANCREATIC CANCER USING R | 17      |

|       |                                                                              |    |
|-------|------------------------------------------------------------------------------|----|
| 4.6   | GO ENRICHMENT ANALYSIS OF BIOLOGICAL<br>PROCESS IN PANCREATIC CANCER USING R | 17 |
| 4.7   | DATA VISUALIZATION OF GENE EXPRESSION<br>LEVEL USING R                       | 17 |
| 5     | <b>RESULTS AND DISCUSSION</b>                                                | 18 |
| 5.1   | LIBRARIES USED IN THIS STUDY                                                 | 18 |
| 5.2   | SOURCE CODE FOR DATA RETRIVAL FROM<br>TCGA DATABASE                          | 20 |
| 5.3   | PANCREATIC CANCER DATA SET                                                   | 21 |
| 5.4   | SOURCE CODE TO SEPERATE CLINICAL DATA                                        | 22 |
| 5.5   | CLINICAL DATA SET                                                            | 23 |
| 5.6   | DIFFERENTIAL EXPRESSION ANALYSIS OF<br>PANCREATIC CANCER                     | 24 |
| 5.7   | RESULTS OF DESEQ2                                                            | 26 |
| 5.7.1 | TREATED VS NOT TREATED                                                       | 26 |
| 5.7.2 | TREATED VS NORMAL                                                            | 27 |
| 5.7.3 | NORMAL VS NOT TREATED                                                        | 28 |
| 5.8   | VOLCANO PLOT                                                                 | 29 |
| 5.8.1 | X-AXIS (FOLD CHANGE)                                                         | 31 |
| 5.8.2 | Y-AXIS (STATISTICAL SIGNIFICANCE)                                            | 31 |
| 5.8.3 | PLOTTING POINTS                                                              | 31 |
| 5.8.4 | HIGHLIGHTING SIGNIFICANT GENES                                               | 32 |
| 5.8.5 | INTERPRETATION                                                               | 32 |
| 5.8.6 | FILTERING                                                                    | 32 |
| 5.9   | GO ENRICHMENT ANALYSIS OF BIOLOGICAL<br>PROCESS IN FILTERED GENES            | 32 |
| 5.10  | DATA VISUALIZATION OF EXPRESSION LEVEL<br>IN FILTERED GENES                  | 37 |
| 6     | <b>SUMMARY AND CONCLUSIONS</b>                                               | 39 |
|       | <b>REFERENCES</b>                                                            | 40 |

## LIST OF TABLES

| TABLE NO  | DESCRIPTION                                        | PAGE NO |
|-----------|----------------------------------------------------|---------|
| Table 5.1 | DATA SET CONTAINS 60660 GENES OF PANCREATIC CANCER | 21      |
| Table 5.2 | CLINICAL DATA OF 178 PATIENTS                      | 23      |

## LIST OF FIGURES

| FIGURE NO | DESCRIPTION                                                  | PAGE NO |
|-----------|--------------------------------------------------------------|---------|
| 5.1       | CODE FOR THE LIBRARIES USED IN R                             | 18      |
| 5.2       | CODE TO RETRIEVE DATA FROM TCGA USING R                      | 20      |
| 5.3       | CODE TO GET CLINICAL DATA USING R                            | 22      |
| 5.4       | CODE FOR DIFFERENTIAL GENE EXPRESSION ANALYSIS IN R          | 24      |
| 5.5       | RESULT OF DEA TREATED VS NOT TREATED                         | 26      |
| 5.6       | RESULT OF DEA TREATED VS NORMAL                              | 27      |
| 5.7       | RESULT OF DEA NORMAL VS NOT TREATED                          | 28      |
| 5.8       | CODE TO CREATE VOLCANO PLOT FOR PANCREATIC CANCER USING R    | 29      |
| 5.9       | VOLCANO PLOT FOR TREATED VS NORMAL IN PANCREATIC CANCER      | 30      |
| 5.10      | VOLCANO PLOT FOR TREATED VS NOT TREATED IN PANCREATIC CANCER | 30      |
| 5.11      | VOLCANO PLOT FOR NORMAL VS NOT TREATED IN PANCREATIC CANCER  | 31      |
| 5.12      | CODE FOR GO ENRICHMENT ANALYSIS USING R                      | 32      |
| 5.13      | GO ENRICHMENT ANALYSIS BARLOT                                | 33      |
| 5.14      | RESULTS OF GENE ONTOLOGY ENRICHMENT ANALYSIS                 | 34      |

|      |                                                |    |
|------|------------------------------------------------|----|
| 5.15 | EXPRESSION BARPLOT FOR GENE<br>ENSG00000202538 | 37 |
| 5.16 | EXPRESSION BARPLOT FOR GENE<br>ENSG00000276168 | 37 |
| 5.17 | EXPRESSION BARPLOT FOR GENE<br>ENSG00000200795 | 38 |

## **LIST OF ABBREVIATIONS**

|         |                                          |
|---------|------------------------------------------|
| TCGA    | THE CANCER GENOME ATLAS                  |
| PAAD    | PANCREATIC ADENOCARCINOMA                |
| RNA-SEQ | RNA SEQUENCING                           |
| PCA     | PRINCIPAL COMPONENT ANALYSIS             |
| GO      | GENE ONTOLOGY                            |
| DEA     | DIFFERENTIAL GENE EXPRESSION<br>ANALYSIS |

# **CHAPTER – 1**

## **INTRODUCTION**

### **1.1 PANCREATIC CANCER**

Pancreatic cancer, despite being relatively less common compared to other cancers, presents a significant challenge in terms of prognosis and treatment. It's characterized by its aggressive nature, late-stage diagnosis, and limited treatment options, leading to a low 5-year survival rate. Despite advancements in medical technology, pancreatic cancer remains a major public health concern globally, with increasing incidence and mortality rates. Understanding its epidemiology, trends, and risk factors is crucial for effective prevention and management strategies. The pancreas, a vital organ located in the upper abdomen, plays a critical role in digestion and hormone regulation. Pancreatic cancer develops when abnormal mutations occur in pancreatic cells, leading to uncontrolled growth and tumour formation. The most common type, pancreatic ductal adenocarcinoma (PAAD), accounts for over 90% of cases. Unfortunately, by the time symptoms manifest, the cancer has usually progressed to an advanced stage, making early detection challenging.

Epidemiological studies provide valuable insights into the prevalence, incidence, and mortality rates of pancreatic cancer. Despite its relatively low occurrence compared to other cancers, pancreatic cancer's lethality is striking. The actual 5-year survival rates have shown minimal improvement over the past few decades, remaining below 10% in most regions globally. Surgical resection offers the best chance of survival, but the overall prognosis remains poor, particularly for non-resectable cases. In recent years, the incidence and mortality rates of pancreatic cancer have been on the rise worldwide. This upward trend poses a significant public health burden, with projections indicating further increases in the coming decades. Factors contributing to this trend include changes in lifestyle habits, environmental factors, and aging populations. Efforts to address these challenges require a multifaceted approach, including public awareness campaigns, early detection strategies, and advancements in treatment modalities.

Various risk factors have been associated with an increased likelihood of developing pancreatic cancer. These include smoking, obesity, diabetes, chronic pancreatitis, and genetic predispositions. Smoking, in particular, stands out as a significant modifiable risk factor, accounting for a substantial proportion of pancreatic cancer cases. Additionally, individuals with a family history of pancreatic cancer or certain genetic syndromes, such as hereditary pancreatitis or Lynch syndrome, are at higher risk. Early detection remains a key challenge in improving pancreatic cancer outcomes. Unlike some other cancers, pancreatic cancer often lacks specific symptoms in its early stages, leading to delayed diagnosis. By the time symptoms such as jaundice, abdominal pain, or weight loss manifest, the cancer has usually advanced to an unresectable stage. Screening methods, including imaging techniques and biomarker testing, hold promise for early detection, but their efficacy and feasibility in large-scale screening programs require further validation.

Advancements in treatment modalities, including surgery, chemotherapy, radiation therapy, and targeted therapies, have improved outcomes for some patients with pancreatic cancer. However, significant challenges remain, including treatment resistance, tumor heterogeneity, and limited therapeutic options for advanced disease. Novel approaches, such as immunotherapy and precision medicine, offer hope for improving patient outcomes, but further research is needed to optimize their effectiveness. In addition to advances in treatment, efforts to reduce the burden of pancreatic cancer must focus on prevention and early detection. Lifestyle modifications, such as smoking cessation, maintaining a healthy weight, and adopting a balanced diet, can help reduce the risk of developing pancreatic cancer. Screening programs targeting high-risk individuals, such as those with a family history or predisposing genetic factors, may facilitate early detection and intervention. Public health initiatives aimed at raising awareness about pancreatic cancer, promoting risk-reducing behaviors, and improving access to screening and treatment services are essential components of comprehensive cancer control strategies. Collaborative efforts involving healthcare professionals, researchers, policymakers, and community stakeholders are needed to address the complex challenges posed by pancreatic cancer and improve outcomes for affected individuals.

Pancreatic cancer remains a formidable challenge in oncology, with a low 5-year survival rate and increasing incidence and mortality rates globally. Understanding its epidemiology, trends, and risk factors is essential for developing effective prevention, early detection, and treatment strategies. Addressing the multifactorial nature of pancreatic cancer requires a coordinated effort across various sectors to reduce its burden and improve outcomes for patients and their families. (Jian-Xiong Hu C.-F. Z.-B.-C.-W.-Y., 2021)

## **1.2 R PROGRAM**

The story of the R programming language is a testament to the power of collaboration, innovation, and the pursuit of knowledge in the digital age. Born out of the necessity to democratize statistical analysis and data visualization, R has grown from its humble origins to become an indispensable tool in the fields of bioinformatics, data science, and beyond. At its core, R owes its lineage to the pioneering work of John Chambers and his team at Bell Laboratories in the 1970s. The development of the S language laid the groundwork for what would later become R, providing an interactive environment for statistical analysis. Building upon this foundation, Ross Ihaka and Robert Gentleman embarked on a journey to create a new language that would combine the functionality of S with their own innovations and ideas.

The result was R, a language named as a successor to "S" and a nod to the first initials of its creators. From its inception, R was designed to be accessible and intuitive, with a gentle learning curve that welcomed both novice and experienced programmers alike. Its built-in plotting functionality revolutionized data visualization, allowing users to generate insightful graphs with ease. As R gained traction in the early 1990s, it quickly garnered attention from statisticians, researchers, and academics worldwide. The release of the R source code under the GNU General Public License in 1995 marked a significant turning point, ensuring that R would remain freely available for all to use and contribute to. This open-source ethos laid the groundwork for a vibrant and collaborative community that continues to drive the development of R to this day.

Over the years, R has evolved from a simple statistical tool to a versatile powerhouse with applications spanning a wide range of disciplines. In bioinformatics, R serves as a linchpin for processing and analyzing complex biological data, from genomic sequences to protein structures. Its extensive ecosystem of packages caters to the diverse needs of researchers, offering solutions for everything from sequence alignment to differential expression analysis. Beyond bioinformatics, R finds utility in fields as diverse as physics, economics, and medicine. Its flexibility and extensibility make it well-suited for tackling a variety of data analysis tasks, from exploratory data analysis to predictive modeling. Moreover, R's role in promoting reproducibility and transparency in scientific research cannot be overstated. By providing a framework for creating reproducible workflows, R empowers researchers to validate and disseminate their findings with confidence.

Looking to the future, the trajectory of R appears promising, with continued advancements and innovations on the horizon. However, challenges such as scalability, performance optimization, and interoperability with emerging technologies remain relevant. Addressing these challenges will be crucial in ensuring the continued relevance and efficacy of R in an ever-evolving landscape of data science and computational biology. The story of R is one of resilience, collaboration, and innovation. From its humble beginnings as a successor to S, R has grown into a global phenomenon, empowering researchers and practitioners to unlock insights from data in ways previously unimaginable. As we reflect on its journey over the past three decades, we are reminded of the transformative power of technology to shape the future of scientific inquiry and discovery. (Federico M. Giorgi, 2022)

### **1.3 DIFFERENTIAL EXPRESSION ANALYSIS**

RNA sequencing (RNA-SEQ) has emerged as a pivotal technology in transcriptomics, offering unparalleled insights into the intricate relationship between genetic alterations and complex biological processes. Its widespread adoption stems from its ability to provide high data reproducibility and a comprehensive view of gene expression dynamics. Particularly in the field of oncology, RNA-seq holds

immense value for diagnostics, prognostics, and therapeutics of tumors, enabling the identification of aberrant transcripts, known as differentially expressed genes (DEGs), which play a crucial role in disease pathogenesis and progression. Differential analysis of RNA-SEQ data lies at the heart of unraveling the molecular mechanisms underlying diseases like cholangiocarcinoma (CHOL). It enables researchers to identify DEGs between diseased and normal tissues, shedding light on potential biomarkers and therapeutic targets. While RNA-seq analysis has been greatly facilitated by the development of open-source R/Bioconductor packages such as limma, DESeq2, and EdgeR, proficiency in the R language and the selection of appropriate methods remain challenges, especially in medical education curricula.

This protocol aims to bridge this gap by providing a detailed guide for performing differential expression analysis using limma, DESeq2, and EdgeR on CHOL RNA-seq count data extracted from The Cancer Genome Atlas (TCGA). These methods, while sharing similarities, differ in their underlying statistical frameworks and data processing steps. For instance, both EdgeR and limma require normalized RNA-seq count data, whereas DESeq2 employs its own method of library size normalization. Additionally, EdgeR is specifically tailored for RNA-seq data analysis, while limma is applicable to both microarray and RNA-seq data. The differential analysis protocols outlined in this guide offer step-by-step instructions for each method, from data preprocessing to DEG identification. Limma employs a linear model to assess DEGs, whereas EdgeR utilizes statistical methods based on the negative binomial distribution, including empirical Bayes estimation and exact tests. DESeq2, on the other hand, employs its own library size normalization method and uses a negative binomial distribution to model count data. By following this protocol, users can easily perform RNA-seq differential expression analysis and select the most appropriate method for their data. Each method has its own strengths and limitations, and the choice ultimately depends on the nature of the data and the specific research questions being addressed. For instance, EdgeR may be preferred for its robustness in handling RNA-seq data, while limma offers versatility for analyzing both microarray and RNA-seq datasets.

In summary, this protocol serves as a valuable resource for researchers and clinicians seeking to leverage RNA-seq data for differential expression analysis in the context of diseases like CHOL. By providing detailed guidance on the use of limma, DESeq2, and EdgeR, it empowers users to gain deeper insights into the molecular mechanisms underlying disease pathogenesis and identify potential targets for diagnosis and therapy. Moreover, it highlights the importance of integrating bioinformatics tools and methods into medical education curricula to equip future healthcare professionals with the necessary skills to leverage genomic data for precision medicine applications. (Shiyi Liu\*1, 2021)

#### **1.4 BIOCONDUCTOR**

Over the past decade, RNA sequencing (RNA-seq) has emerged as the gold standard for accurately profiling gene expression, offering researcher's unparalleled insights into biological systems. This surge in RNA-seq usage has been accompanied by the development of specialized software tools aimed at streamlining data analysis. Review articles by Conesa et al. and Van den Berge et al. have comprehensively covered the diverse applications and software landscape in RNA-seq analysis, highlighting the importance of robust data analysis pipelines in addressing complex biological questions. Differential expression analysis stands as one of the most commonly employed workflows in RNA-seq analysis. Researchers leverage this approach to unravel the mechanisms underlying transcriptional regulation by comparing gene expression across different conditions, genotypes, tissues, cell types, or time points. The ultimate goal is to identify robust sets of genes exhibiting changes in expression and to contextualize these findings within molecular pathways, providing valuable insights for both basic research and clinical applications. Despite the availability of established end-to-end analysis procedures, bioinformatics analyses can still pose significant challenges, particularly for users with limited programming skills. Current software implementations cover various aspects of the analysis pipeline, including quality assessment, preprocessing, alignment, quantification, and differential expression analysis. These tools, often integrated into the R/Bioconductor project, have streamlined the generation of large transcriptome profiles, laying the groundwork for downstream analyses.

In a previous study, a selection of interfaces for RNA-seq analysis was reviewed from the perspective of a life scientist, focusing on key criteria such as installation, usability, flexibility, hardware requirements, and reproducibility. Building upon these findings, a novel tool was developed to meet the diverse needs of researchers in differential expression analysis. This tool, termed "ideal," was designed as an interactive web-based application within the R/Bioconductor environment, aimed at guiding users through all stages of a complete differential expression analysis. Leveraging insights from close collaborations with wet-lab life scientists and clinicians, ideal offers an integrated platform for extracting, visualizing, interpreting, and sharing RNA-seq datasets. Similar to the Bioconductor `pcaExplorer` package, which facilitates exploratory data analysis, ideal provides a comprehensive solution tailored specifically for differential expression analysis.

To use ideal, users need to input a count matrix and experimental design information, allowing for the analysis of complex experimental designs. The tool ensures ease of reproducibility and sharing of analysis sessions, promoting effective collaboration and adherence to the FAIR Guiding Principles for scientific data. Furthermore, ideal offers a wide array of information-rich visualizations, charts, and tables for both diagnostic and downstream analysis steps, ensuring transparency and reproducibility in RNA-seq data analysis. Building upon the success of `pcaExplorer`, ideal incorporates improved documentation features and self-paced learning tours to enhance user experience. State saving and automated HTML report generation further ensure code reproducibility, aligning with the growing emphasis on reproducibility in scientific research. While there are numerous software packages available for RNA-seq analysis, ideal distinguishes itself by offering a comprehensive set of features within the Bioconductor environment. Unlike existing tools, ideal seamlessly integrates with R/Bioconductor packages, leveraging the platform's efficient core structures for flexible additional visualizations. Moreover, ideal provides a coherent user interface that combines interactivity with reproducibility, facilitating the translation of transcriptomic datasets into actionable insights.

Although existing tools accept count data and compute differentially expressed genes, they often lack the full integration and seamless combination of interactivity

and reproducibility offered by ideal. While standalone web applications are common, not all can be easily distributed as packages or deployed to private local instances. Furthermore, only a few tools provide the option to generate dynamic analysis reports using R Markdown, highlighting the unique value proposition of ideal in the RNA-seq analysis landscape. In summary, the ideal package represents a significant advancement in RNA-seq data analysis, offering a user-friendly framework within the Bioconductor environment. By integrating best practices and providing a coherent user interface, ideal empowers researchers to derive meaningful insights from transcriptomic datasets with ease and confidence. (Marini, 2020)

## **1.5 TCGABIOLINKS**

TCGAbiolinks is an R package designed to facilitate access to and analysis of data from The Cancer Genome Atlas (TCGA) project. TCGA is a comprehensive initiative led by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) that aims to characterize the genomic alterations underlying various types of cancer. TCGAbiolinks streamlines the process of retrieving TCGA data, enabling researchers to seamlessly access a wealth of genomic, transcriptomic, and clinical information for numerous cancer types. The package offers a suite of functions tailored for data retrieval, preprocessing, analysis, and visualization, empowering researchers to conduct in-depth investigations into cancer biology. One of the key features of TCGAbiolinks is its ability to download and preprocess TCGA data directly within the R environment. By leveraging the GDC Data Transfer Tool provided by the Genomic Data Commons (GDC), TCGAbiolinks simplifies the retrieval of raw sequencing data and associated clinical metadata. This functionality expedites data acquisition and ensures reproducibility, as researchers can obtain the same datasets used in published studies. Moreover, TCGAbiolinks provides robust preprocessing capabilities to normalize and filter genomic data, addressing common challenges such as batch effects and sample heterogeneity. These preprocessing steps are essential for ensuring data quality and reliability in downstream analyses. In addition to data retrieval and preprocessing, TCGAbiolinks offers a variety of analytical tools for exploring TCGA datasets. Researchers can perform differential expression

analysis, identify genetic alterations, conduct survival analyses, and visualize results using built-in functions. This comprehensive suite of analytical functionalities empowers researchers to uncover novel insights into cancer biology and identify potential biomarkers or therapeutic targets. TCGAbiolinks serves as a valuable resource for cancer researchers, providing a user-friendly interface for accessing and analyzing TCGA data within the R environment. By facilitating data integration, preprocessing, and analysis, TCGAbiolinks accelerates the pace of cancer research and contributes to our understanding of the molecular mechanisms driving tumorigenesis and disease progression.

## **1.6 DESEQ2**

DESeq2 is a powerful and widely used R package designed for the analysis of RNA sequencing (RNA-seq) data, particularly in the context of differential gene expression analysis. Developed by Love et al., DESeq2 offers robust statistical methods to identify genes that are differentially expressed across different experimental conditions or biological states. At its core, DESeq2 employs negative binomial distribution modeling to account for inherent variability in RNA-seq data, which arises due to factors such as differences in library size and gene expression levels. By modeling this variability, DESeq2 effectively normalizes the data, enabling accurate comparison of gene expression levels between samples.

One of the key features of DESeq2 is its ability to handle complex experimental designs, including multiple experimental factors and covariates. Through the use of generalized linear models (GLMs), DESeq2 can accommodate various experimental setups, such as time-course experiments or studies involving multiple treatment conditions. DESeq2 performs differential expression analysis by fitting a model to the count data for each gene, estimating parameters such as mean expression and dispersion, and testing for significant differences in expression levels between experimental groups. Importantly, DESeq2 employs rigorous statistical methods, including shrinkage estimation, to improve the stability and reliability of results, particularly when dealing with limited sample sizes. Moreover, DESeq2 provides comprehensive functionality for downstream analysis, including visualization tools for exploring differential expression results, clustering analysis,

and gene ontology enrichment analysis to uncover biological insights associated with differentially expressed genes. DESeq2 is highly regarded for its accuracy, robustness, and flexibility in analyzing RNA-seq data. Its sophisticated statistical methods and user-friendly interface make it an indispensable tool for researchers in genomics, bioinformatics, and related fields seeking to unravel the intricacies of gene expression regulation in diverse biological contexts.

## **1.7 TCGA DATABASE**

The Cancer Genome Atlas (TCGA) is a landmark initiative in cancer research that has revolutionized our understanding of the molecular underpinnings of cancer. Established in 2006 as a collaborative effort between the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI), TCGA represents one of the most comprehensive and ambitious projects in the field of cancer genomics. At its core, TCGA aims to systematically characterize the genomic alterations, molecular profiles, and clinical outcomes across various cancer types. By compiling a vast repository of multi-dimensional data, TCGA provides researchers with unprecedented insights into the complex landscape of cancer biology, paving the way for personalized medicine and novel therapeutic approaches. Central to the TCGA endeavor is the collection and analysis of high-throughput genomic data from thousands of cancer samples. These samples, obtained from patients diagnosed with different cancer types, encompass a diverse array of tumor tissues, including primary tumors, metastatic lesions, and adjacent normal tissues. Through state-of-the-art genomic profiling techniques such as next-generation sequencing (NGS), microarray analysis, and mass spectrometry, TCGA generates comprehensive molecular profiles that encompass genomic alterations, gene expression patterns, epigenetic modifications, and proteomic signatures. The wealth of data generated by TCGA is made freely available to the global research community through publicly accessible databases and web portals. Researchers can access and analyze this data to interrogate specific hypotheses, identify novel biomarkers, unravel disease mechanisms, and discover potential therapeutic targets. Furthermore, TCGA facilitates cross-disciplinary collaborations and data sharing initiatives, fostering a collaborative ecosystem that accelerates scientific discovery and innovation. Beyond its primary mission of characterizing cancer

genomes, TCGA also incorporates detailed clinical annotations and patient outcomes data into its datasets. This integration of clinical and molecular data enables researchers to correlate genomic alterations with clinical phenotypes, treatment responses, and disease outcomes. Such integrative analyses provide valuable insights into the molecular basis of cancer progression, metastasis, and therapeutic resistance, ultimately guiding clinical decision-making and patient care. Moreover, TCGA has catalyzed the development of advanced computational tools and bioinformatics pipelines for analyzing large-scale genomic datasets. These bioinformatics resources enable researchers to perform sophisticated data analyses, including differential gene expression analysis, pathway enrichment analysis, and integrative multi-omics profiling. By leveraging these computational resources, researchers can extract meaningful insights from TCGA data and translate them into actionable knowledge for improving cancer diagnosis, prognosis, and treatment. The Cancer Genome Atlas (TCGA) represents a groundbreaking initiative that has transformed our understanding of cancer biology and revolutionized cancer research. By providing a comprehensive molecular characterization of diverse cancer types, TCGA empowers researchers worldwide to unravel the complexities of cancer and accelerate the development of precision oncology strategies for improving patient outcomes.

## **1.8 GO ENRICHMENT ANALYSIS OF BIOLOGICAL PROCESS**

The Gene Ontology (GO) Consortium stands as a cornerstone in the field of genomics, providing a comprehensive resource for understanding the functions of genes and gene products. Over the past two years, significant advancements have been made within the consortium, notably with the introduction of the GO-Causal Activity Models (GO-CAM) annotation framework and collaborations to refine and expand the GO resource. The GO-CAM framework represents a notable enhancement in the GO Consortium's capabilities. This framework formalizes models of gene function and facilitates the representation of causal relationships between molecular activities. By linking multiple standard GO annotations, GO-CAMs provide a more comprehensive understanding of gene function, enabling researchers to delve deeper into the intricacies of molecular mechanisms. This advancement has been crucial in expanding the scope and depth of the GO

resource, empowering researchers to explore complex biological phenomena with greater precision. Moreover, efforts have been dedicated to ensuring the integrity and reliability of the GO resource. A computational schema has been implemented to check and validate the rapidly growing repository of GO-CAMs, ensuring consistency and accuracy in the annotations. Additionally, collaborations with experts in specific areas of biology have contributed to refining and expanding the GO ontology, resulting in a 10% increase in the number of GO annotations and a 25% increase in annotated gene products. Furthermore, over 9,400 new scientific articles have been annotated, enriching the resource with the latest research findings and insights.

As the project matures, the consortium continues its commitment to reviewing older annotations in light of newer discoveries and maintaining consistency with other ontologies. This iterative process ensures that the GO resource remains up-to-date and reflective of the latest advancements in the field. Furthermore, the consortium has taken steps to enhance user accessibility and support traceability and reproducibility. The redesign of the GO website facilitates quick access to documentation, downloads, and tools, streamlining the user experience and promoting widespread utilization of the resource. In addition to the GO ontology and standard annotations, the introduction of GO-CAMs has expanded the capabilities of the GO resource. GO-CAMs represent a significant step forward in capturing the complex relationships between molecular activities, enabling researchers to explore gene function in a more nuanced manner. These models offer insights into the underlying mechanisms driving biological processes, paving the way for deeper understanding and discovery in genomics and biomedical research. Overall, the GO Consortium's efforts over the past two years have significantly enhanced the comprehensiveness, accuracy, and usability of the GO resource. Through the introduction of GO-CAMs, collaborations to refine and expand annotations, and improvements in accessibility and usability, the consortium continues to uphold its mission of providing a valuable resource for the scientific community. As genomics and biomedical research continue to advance, the GO Consortium remains committed to evolving and adapting to meet the evolving needs of researchers worldwide. (Paul D Thomas, 2020)

## CHAPTER – 2

### REVIEW OF LITERATURE

Hu et al. (2021) highlight the formidable challenge posed by pancreatic cancer, characterized by its aggressive nature, late-stage diagnosis, and limited treatment options, leading to a low 5-year survival rate. Despite advancements in medical technology, pancreatic cancer remains a major public health concern globally, with increasing incidence and mortality rates. Understanding its epidemiology, trends, and risk factors is crucial for effective prevention and management strategies.

Giorgi (2022) narrates the evolution of the R programming language, tracing its roots to the collaborative efforts at Bell Laboratories in the 1970s. Initially conceived to democratize statistical analysis, R has since become indispensable in bioinformatics, data science, and beyond. Its user-friendly design and open-source nature have fostered a vibrant community that continues to drive its development. Today, R's versatility extends to diverse fields like physics and medicine, offering solutions for complex data analysis tasks and promoting reproducibility in research. While challenges such as scalability persist, R remains poised for future advancements, reflecting the transformative impact of technology on scientific inquiry.

Liu (2021) presents a comprehensive protocol for conducting differential expression analysis using RNA sequencing (RNA-seq) data in cholangiocarcinoma (CHOL) research. The protocol focuses on utilizing open-source R/Bioconductor packages, including limma, DESeq2, and EdgeR, to identify differentially expressed genes (DEGs) extracted from The Cancer Genome Atlas (TCGA). Each method has its unique statistical framework and data processing steps, with limma employing a linear model, EdgeR using negative binomial distribution-based methods, and DESeq2 employing its own normalization approach.

Marini (2020) presents "ideal," a novel web-based application within the R/Bioconductor environment, tailored for differential expression analysis in RNA sequencing (RNA-seq) data. Leveraging insights from life scientists and clinicians, ideal guides users through all stages of analysis, facilitating data extraction,

visualization, interpretation, and sharing. With features like state saving and automated HTML report generation, ideal ensures reproducibility and adherence to FAIR Guiding Principles. Unlike existing tools, ideal seamlessly integrates with R/Bioconductor packages, offering a comprehensive solution with enhanced documentation and learning features for user experience improvement. By combining interactivity with reproducibility, ideal empowers researchers to translate transcriptomic datasets into actionable insights efficiently.

Thomas (2020) highlights the Gene Ontology (GO) Consortium's recent advancements in genomics. The introduction of the GO-Causal Activity Models (GO-CAM) framework and collaborations to refine and expand the GO resource mark significant progress. GO-CAM formalizes gene function models, enabling representation of causal relationships between molecular activities. Collaborations and a computational schema ensure integrity and reliability, resulting in a 10% increase in GO annotations and 25% more annotated gene products. The consortium's commitment to reviewing older annotations and enhancing user accessibility through website redesign underscores its dedication to providing a valuable resource. Overall, these efforts have enhanced the comprehensiveness, accuracy, and usability of the GO resource, supporting continued advancements in genomics and biomedical research.

## **CHAPTER – 3**

### **AIM AND SCOPE OF THE PROPOSED STUDY**

#### **3.1 AIM**

This study aims to analyze RNA-Seq data from pancreatic cancer using R Programming language to find a gene which could be a potential biomarker.

#### **3.2 OBJECTIVES**

- Accessing the RNA-SEQ data of pancreatic cancer from TCGA database
- Pre-processing of the RNA-SEQ data
- Perform Differential gene expression analysis
- Identifying genes showing statistically significant differences in expression levels across two or more experimental or biological conditions.
- Determining expressed variants within cancer samples.
- Conducting enrichment analysis using predefined Gene Ontology terms

#### **3.3 SCOPE**

The scope of this project is to leverage the computational capabilities of the R programming language to identify potential biomarkers for pancreatic cancer. And also it focuses on the discovery and validation of biomarkers for pancreatic cancer using computational approaches supported by the R programming language.

## **CHAPTER – 4**

### **METHODOLOGY**

#### **4.1 DATA COLLECTION FOR THE DEA OF PANCREATIC CANCER**

The RNA-Seq dataset for pancreatic adenocarcinoma (PAAD), a form of brain cancer, was obtained from The Cancer Genome Atlas (TCGA) database using an R package. TCGA is a prominent initiative in cancer research dedicated to gathering and scrutinizing high quality tumor samples, providing researchers with access to associated data for analysis. The TCGA data portal provides researchers with the ability to explore, download, and analyze data for normal and tumor patients. This comprehensive database genomic sequencing data, clinical information, and other molecular characterization data obtained from thousands of cancer patients.

#### **4.2 DATA SETS OF THE PANCREATIC CANCER**

The data set used in this study contains 60660 gene identifiers of 178 people with pancreatic adenocarcinoma (PAAD) and 60660 gene identifiers of 4 normal people

#### **4.3 DATA RETRIVAL FROM TCGA DATABASE USING R**

I utilized the TCGAbiolinks package in R to download TCGA database data in the form of TSV files, specifically focusing on pancreatic adenocarcinoma (PAAD) cancer. Through its functions, users can easily access and analyze data from TCGA, download datasets as text or TSV files into their directory, perform integrative analysis, and visualize results efficiently.

#### **4.4 DATA PRE PROCESSING OF RETRIVED DATA SETS USING R**

To ensure data quality and consistency, preprocessing was conducted using the GDCprepare package in R. This involved normalization, batch correction, and quality control procedures. Proper preprocessing is crucial for accurate downstream analyses, ensuring reliable and reproducible results. By harnessing specialized tools and methodologies, we optimized the dataset for subsequent differential

expression and pathway analyses, laying the groundwork for biomarker discovery efforts.

#### **4.5 DIFFERENTIAL EXPRESSION ANALYSIS OF PANCREATIC CANCER USING R**

Differential expression analysis was performed using R programming language, employing packages such as DESeq2. The analysis identified genes that exhibited significant differences in expression levels. The result of differential expression analysis is split into three formats – Treated vs Not treated, Treated vs Normal, Normal vs not Treated

#### **4.6 GO ENRICHMENT ANALYSIS OF BIOLOGICAL PROCESS IN PANCREATIC CANCER USING R**

After conducting the differential analysis on the pancreatic adenocarcinoma (PAAD) dataset, I proceeded with Gene Ontology (GO) enrichment analysis focusing on biological processes. This method aims to elucidate the biological functions associated with the differentially expressed genes identified in the dataset. By annotating these genes based on their involvement in various biological processes, GO enrichment analysis provides valuable insights into the underlying molecular mechanisms driving PAAD progression. The process involves comparing the list of differentially expressed genes against a background set of all genes in the genome to identify overrepresented GO terms. These terms represent biological processes such as cell cycle regulation, immune response modulation, or metabolic pathways, among others. Statistical methods, often employing hypergeometric tests or Fisher's exact tests, are employed to determine the significance of enrichment.

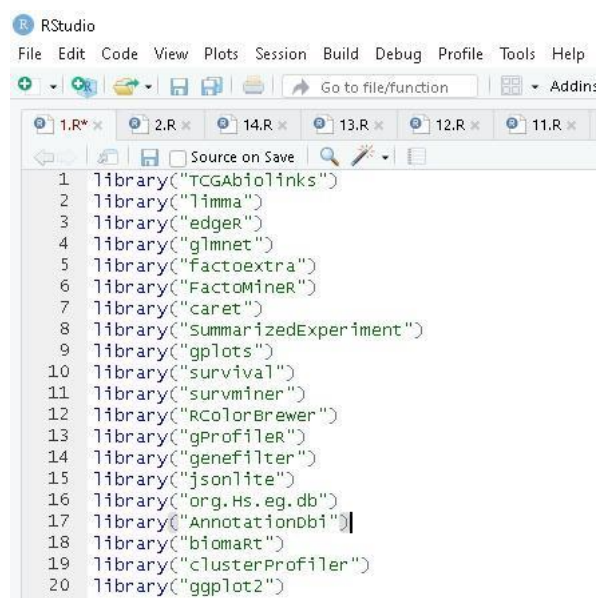
#### **4.7 DATA VISUALIZATION OF GENE EXPRESSION LEVEL USING R**

Visualizing the expression levels of the narrowed genes in untreated versus treated conditions provides a clear representation of how gene expression patterns are altered by treatment in pancreatic adenocarcinoma (PAAD). Using a bar plot, each bar represents the mean or median expression level of a gene in the two conditions, with error bars indicating variability.

## CHAPTER – 5

### RESULTS AND DISCUSSION

#### 5.1 LIBRARIES USED IN THIS STUDY

A screenshot of the RStudio interface. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, and Help. Below the menu is a toolbar with icons for saving, running, and other functions. The main editor window displays a list of R libraries being loaded, numbered 1 through 20. The libraries are: TCGAbiolinks, limma, edgeR, glmnet, factoextra, FactoMineR, caret, SummarizedExperiment, gplots, survival, survminer, RColorBrewer, gProfiler, geneFilter, jsonlite, org.Hs.eg.db, AnnotationDbi, biomaRt, clusterProfiler, and ggplot2.

```
1 library("TCGAbiolinks")
2 library("limma")
3 library("edgeR")
4 library("glmnet")
5 library("factoextra")
6 library("FactoMineR")
7 library("caret")
8 library("SummarizedExperiment")
9 library("gplots")
10 library("survival")
11 library("survminer")
12 library("RColorBrewer")
13 library("gProfiler")
14 library("geneFilter")
15 library("jsonlite")
16 library("org.Hs.eg.db")
17 library("AnnotationDbi")
18 library("biomaRt")
19 library("clusterProfiler")
20 library("ggplot2")
```

**Fig 5.1: Code for the Libraries used in R**

TCGAbiolinks: A comprehensive R package for accessing and analyzing data from The Cancer Genome Atlas (TCGA) database, facilitating genomic data retrieval, preprocessing, and downstream analysis.

glmnet: Implements Lasso and Elastic-Net regularized generalized linear models, useful for variable selection and prediction in high-dimensional datasets, commonly used in genomic data analysis.

factoextra: Provides functions to extract and visualize results from multivariate analysis techniques such as PCA, clustering, and correspondence analysis, enhancing interpretability and presentation of results.

SummarizedExperiment: Provides infrastructure for storing and manipulating assay data alongside sample and feature metadata, commonly used in bioinformatics workflows for managing high-dimensional experimental data.

RColorBrewer: Provides color palettes suitable for enhancing the visual appeal and readability of plots, offering a wide range of color schemes for categorical, sequential, and diverging data.

gProfileR: Facilitates functional enrichment analysis of gene sets, offering tools for identifying overrepresented Gene Ontology terms, pathways, and transcription factor binding sites among a list of genes.

genefilter: Implements methods for filtering and selecting genes based on various criteria such as differential expression, variance, or fold change, useful for preprocessing genomic datasets prior to downstream analysis.

jsonlite: Provides functions for reading, writing, and manipulating JSON (JavaScript Object Notation) data, commonly used for storing and exchanging structured data in web applications and APIs.

org.Hs.eg.db: An annotation package providing mappings between gene identifiers and corresponding annotations for the human genome, facilitating the interpretation of genomic data.

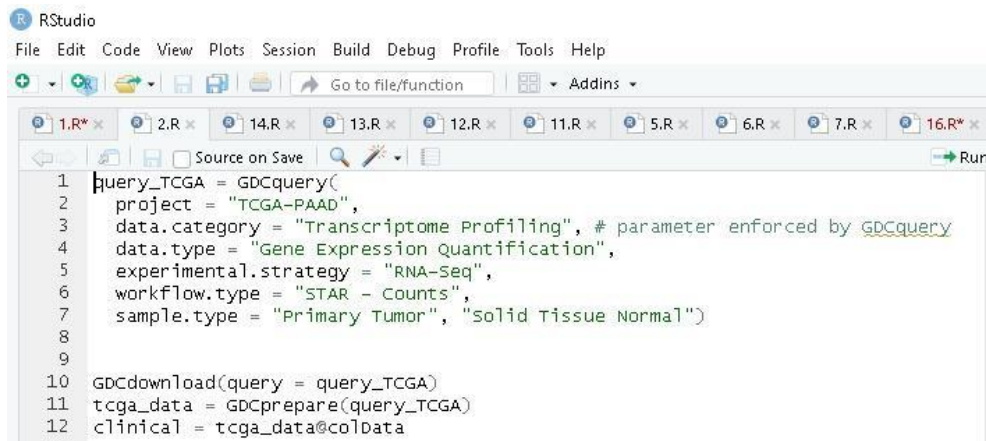
AnnotationDbi: Offers infrastructure for managing and querying annotation databases, supporting the integration of various annotation resources with genomic datasets.

biomaRt: Provides access to biological data resources such as Ensembl through the BioMart web service, allowing for programmatic retrieval of gene annotations, sequences, and functional annotations.

clusterProfiler: Performs functional enrichment analysis of gene clusters or sets, offering

ggplot2: Offers additional plotting functionalities beyond base R graphics, including enhanced customization options for creating high-quality plots for exploratory data analysis and visualization of results.

## 5.2 SOURCE CODE FOR DATA RETRIEVAL FROM TCGA

A screenshot of the RStudio interface. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, and Help. Below the menu is a toolbar with icons for file operations and a search bar. The main editor window displays R code for querying and downloading TCGA data. The code is as follows:

```
1 query_TCGA = GDCquery(  
2   project = "TCGA-PAAD",  
3   data.category = "Transcriptome Profiling", # parameter enforced by GDCquery  
4   data.type = "Gene Expression Quantification",  
5   experimental.strategy = "RNA-Seq",  
6   workflow.type = "STAR - Counts",  
7   sample.type = "Primary Tumor", "Solid Tissue Normal")  
8  
9  
10 GDCdownload(query = query_TCGA)  
11 tcga_data = GDCprepare(query_TCGA)  
12 clinical = tcga_data@colData
```

**Fig 5.2: Code to retrieve data from TCGA using R**

This R code segment initiates the process of obtaining and preparing transcriptomic data from The Cancer Genome Atlas (TCGA) database specifically for pancreatic adenocarcinoma (PAAD) investigation. It begins by defining a query object, specifying parameters such as the TCGA project (designated as "TCGA-PAAD" for PAAD), data category (transcriptome profiling), data type (gene expression quantification), and experimental strategy (RNA-Seq). It also includes criteria for selecting sample types, targeting both primary tumor and solid tissue normal samples. Subsequently, the code downloads the data using the specified query parameters via the `GDCdownload` function, facilitating interaction with the Genomic Data Commons (GDC) API. Next, the downloaded data undergoes preparation for analysis through the `GDCprepare` function. This step organizes the data into a `SummarizedExperiment` object, effectively integrating gene expression counts with associated sample metadata. Finally, the code extracts clinical metadata from the prepared TCGA data object, enabling the retrieval of pertinent patient information such as demographics and clinical characteristics.

### 5.3 PANCREATIC CANCER DATA SET

**Table 5.1: Data set contains 60660 genes of pancreatic cancer**

|                    |       |       |       |       |       |
|--------------------|-------|-------|-------|-------|-------|
| ENSG00000000003.15 | 2684  | 3654  | 880   | 1090  | 1361  |
| ENSG00000000005.6  | 0     | 1     | 0     | 4     | 0     |
| ENSG00000000419.13 | 1139  | 1620  | 941   | 974   | 1148  |
| ENSG00000000457.14 | 609   | 916   | 372   | 438   | 711   |
| ENSG00000000460.17 | 179   | 314   | 174   | 171   | 216   |
| ENSG00000000938.13 | 321   | 433   | 69    | 300   | 496   |
| ENSG00000000971.16 | 2473  | 10156 | 1659  | 6457  | 7855  |
| ENSG00000001036.14 | 2051  | 1710  | 1075  | 2185  | 473   |
| ENSG00000001084.13 | 1230  | 1097  | 732   | 1544  | 494   |
| ENSG00000001167.14 | 1078  | 1476  | 773   | 830   | 1139  |
| ENSG00000001460.18 | 438   | 943   | 466   | 508   | 280   |
| ENSG00000001461.17 | 2002  | 3087  | 1477  | 1564  | 2074  |
| ENSG00000001497.18 | 2816  | 2186  | 958   | 1484  | 934   |
| ENSG00000001561.7  | 790   | 2388  | 447   | 860   | 389   |
| ENSG00000001617.12 | 2666  | 1731  | 3066  | 2216  | 2874  |
| ENSG00000001626.16 | 29541 | 9813  | 3359  | 7358  | 1583  |
| ENSG00000001629.10 | 2432  | 2727  | 2658  | 2614  | 1277  |
| ENSG00000001630.17 | 211   | 213   | 56    | 91    | 20    |
| ENSG00000001631.16 | 227   | 332   | 189   | 270   | 153   |
| ENSG00000002016.18 | 388   | 393   | 366   | 378   | 461   |
| ENSG00000002079.14 | 49    | 150   | 133   | 79    | 92    |
| .....              | ..... | ..... | ..... | ..... | ..... |
| ENSG00000002330.14 | 521   | 583   | 319   | 546   | 335   |

This data from TCGA, is derived from RNA sequencing experiments. Each row corresponds to a specific gene, identified by its Ensembl gene ID, while the columns represent different patient. It contains 60660 gene data of 178 people. The values in the table represent the counts of RNA transcripts for each gene in each sample. This data is crucial for differential expression analysis, where changes in gene

expression levels between different conditions, such as cancer versus normal tissues, are assessed. It provides insights into the transcriptional activity of genes and aids in understanding the molecular mechanisms underlying diseases like cancer.

## 5.4 SOURCE CODE TO SEPARATE CLINICAL DATA

A screenshot of the RStudio interface. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, and Help. Below the menu is a toolbar with icons for file operations and a search bar. The main editor window shows R code with line numbers 1 through 13. The code filters a dataframe named 'clinical' based on the condition 'clinical\$definition == "Primary solid Tumor"', selecting columns 'patient', 'vital\_status', 'days\_to\_death', 'days\_to\_last\_follow\_up', 'treatments', and 'gender'. It then assigns the result to 'clin\_df', sets rownames, and prints the dataframe.

```
1 clin_df = clinical[clinical$definition == "Primary solid Tumor",  
2                  c("patient",  
3                    "vital_status",  
4                    "days_to_death",  
5                    "days_to_last_follow_up",  
6                    "treatments",  
7                    "gender")]  
8 |  
9  
10 clin_df$listData$patient <- rownames(clin_df)  
11  
12 # Viewing the updated DataFrame  
13 print(clin_df)
```

**Fig 5.3: Code to get clinical data using r**

This R code filters and subsets clinical data from a dataframe called `clinical`, focusing on patients diagnosed with primary solid tumors. It selects specific columns including patient identifiers, vital status (alive or deceased), days to death for deceased patients, days to last follow-up for living patients, treatments received, and gender. By applying the condition `clinical$definition == "primary solid tumor"`, only rows corresponding to patients with primary solid tumors are retained. This subset, stored in a new dataframe named `clin_df`, is likely intended for further analysis or visualization aimed at understanding the clinical characteristics and outcomes of patients with primary solid tumors.

## 5.5 CLINICAL DATA SET

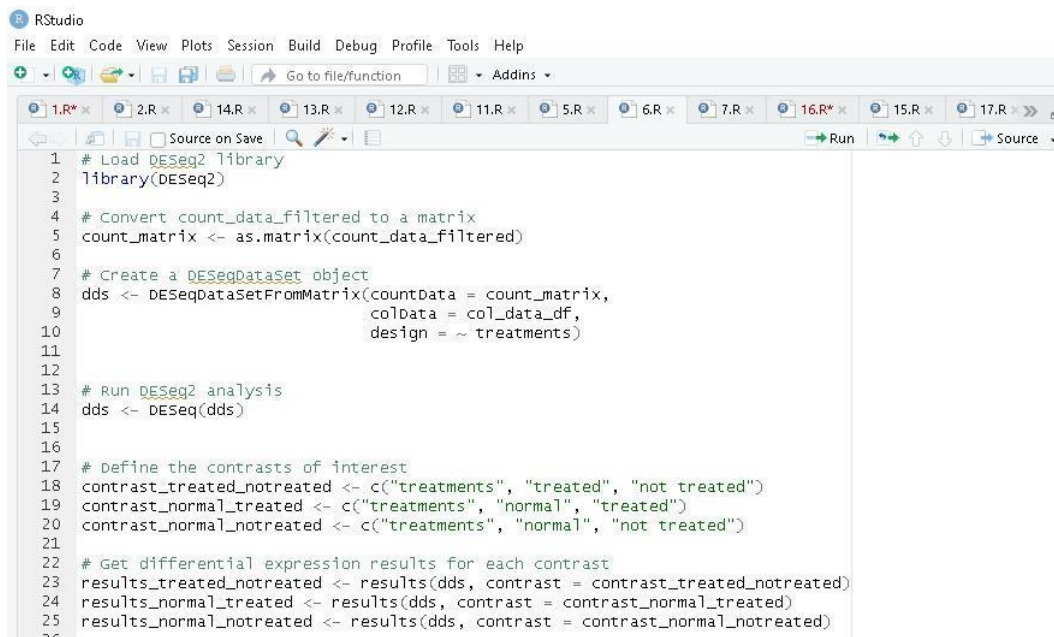
**Table 5.2: Clinical data of 178 patients**

|                              | vital_status | days_to_death | days_to_last_follow_up | treatments  | gender |
|------------------------------|--------------|---------------|------------------------|-------------|--------|
| TCGA-IB-A6UG-01A-32R-A33R-07 | Dead         | 41            | NA                     | not treated | male   |
| TCGA-3A-A9IU-01A-11R-A39D-07 | Dead         | 458           | NA                     | treated     | male   |
| TCGA-FB-A545-01A-11R-A26U-07 | Dead         | 732           | 385                    | treated     | female |
| TCGA-IB-AAUQ-01A-22R-A41I-07 | Dead         | 183           | NA                     | treated     | male   |
| TCGA-HV-A5A3-01A-11R-A26U-07 | Dead         | 128           | NA                     | treated     | male   |
| TCGA-IB-7897-01A-21R-2204-07 | Dead         | 486           | NA                     | not treated | female |
| TCGA-HV-AA8V-01A-11R-A41B-07 | Alive        | NA            | 920                    | treated     | male   |
| TCGA-HV-A5A5-01A-11R-A26U-07 | Alive        | NA            | 289                    | treated     | female |
| TCGA-HZ-8638-01A-11R-2404-07 | Dead         | 151           | 91                     | not treated | female |
| TCGA-F2-7276-01A-11R-2156-07 | Dead         | 216           | NA                     | treated     | male   |
| TCGA-F2-6880-01A-11R-2156-07 | Alive        | NA            | 295                    | treated     | male   |
| TCGA-M8-A5N4-01A-11R-A26U-07 | Alive        | NA            | 584                    | treated     | female |
| TCGA-3A-A9IJ-01A-11R-A39D-07 | Alive        | NA            | 1854                   | not treated | male   |

|                              |      |     |   |         |        |
|------------------------------|------|-----|---|---------|--------|
| TCGA-HZ-8637-01A-11R-2404-07 | Dead | 517 | 0 | treated | female |
|------------------------------|------|-----|---|---------|--------|

This clinical data includes information on vital status, days to death, days to last follow-up, treatment received, and gender for each patient sample. Vital status indicates whether the patient is alive or deceased, while days to death and days to last follow-up represent the time elapsed from diagnosis to death or last follow-up, respectively. Treatment details specify whether the patient received treatment or not, and gender identifies the patient's sex. This clinical data is crucial for correlating patient outcomes with various factors such as treatment response, survival times, and demographic characteristics, aiding in clinical research and treatment decision-making in cancer care.

## 5.6 DIFFERENTIAL EXPRESSION ANALYSIS OF PANCREATIC CANCER



```

1 # Load DESeq2 library
2 library(DESeq2)
3
4 # Convert count_data_filtered to a matrix
5 count_matrix <- as.matrix(count_data_filtered)
6
7 # Create a DESeqDataSet object
8 dds <- DESeqDataSetFromMatrix(countData = count_matrix,
9                               colData = col_data_df,
10                              design = ~ treatments)
11
12
13 # Run DESeq2 analysis
14 dds <- DESeq(dds)
15
16
17 # Define the contrasts of interest
18 contrast_treated_notreated <- c("treatments", "treated", "not treated")
19 contrast_normal_treated <- c("treatments", "normal", "treated")
20 contrast_normal_notreated <- c("treatments", "normal", "not treated")
21
22 # Get differential expression results for each contrast
23 results_treated_notreated <- results(dds, contrast = contrast_treated_notreated)
24 results_normal_treated <- results(dds, contrast = contrast_normal_treated)
25 results_normal_notreated <- results(dds, contrast = contrast_normal_notreated)
26

```

**Fig 5.4: Code for differential gene expression analysis in R**

This R code utilizes the DESeq2 package, a popular tool for analyzing RNA-Seq data, to perform differential expression analysis between different treatment conditions in a dataset.

First, the count data, which typically represents the number of sequencing reads mapped to each gene, is converted into a matrix format using the `as.matrix()` function. This matrix, `count.matrix`, serves as the input data for DESeq2 analysis.

Next, a `DESeqDataSet` object, `dds`, is created using `DeseqDatasetMatrix()` function. This function requires the count data matrix `countdata`, sample metadata `coldata`, and a design formula `design` specifying the experimental design and the variables of interest, in this case, the treatment conditions.

The `DeESeq` function is then used to run the DESeq2 analysis on the `dds` object, estimating size factors and dispersion parameters, and fitting a negative binomial generalized linear model to the count data.

After the DESeq2 analysis is performed, contrasts of interest are defined using the `contrast` argument. Three specific contrasts are defined: between treated and untreated samples `contrast_treated_notreated`, between normal and treated samples `contrast_normal_treated`, and between normal and untreated samples `contrast_normal_notreated`.

The `result()` function is then applied to extract the differential expression results for each defined contrast. This function calculates fold changes, log2 fold changes, p-values, and adjusted p-values (typically using the Benjamini-Hochberg procedure) for each gene, indicating the extent of differential expression between the specified conditions.

In summary, this code segment enables the user to conduct differential expression analysis between different treatment conditions in an RNA-Seq dataset using DESeq2, providing valuable insights into the genes and pathways that are modulated by the treatments under investigation.

## 5.7 RESULTS OF DESEQ2

### 5.7.1 Treated vs Not treated

|                     | baseMean    | log2FoldChange | lfcSE       | stat         | pvalue      | padj        |
|---------------------|-------------|----------------|-------------|--------------|-------------|-------------|
| ENSG00000000003.15  | 1804.081313 | 0.080463147    | 0.125705452 | 0.640092739  | 0.52211231  | 0.923968679 |
| ENSG00000000005.6   | 13.02124896 | 0.175562004    | 0.507745228 | 0.345767906  | 0.729517147 | 0.963379493 |
| ENSG000000000419.13 | 1420.733241 | 0.044753923    | 0.069187981 | 0.646845342  | 0.517732044 | 0.923258766 |
| ENSG000000000457.14 | 714.3607259 | -0.037064973   | 0.084606761 | -0.438085233 | 0.661324499 | 0.95364347  |
| ENSG000000000460.17 | 216.0645361 | 0.175468805    | 0.098939075 | 1.77350359   | 0.076145298 | 0.642535442 |
| ENSG000000000938.13 | 897.6750832 | -0.032652034   | 0.195325972 | -0.167166886 | 0.867238739 | 0.981501044 |
| ENSG000000000971.16 | 6077.562164 | -0.19185284    | 0.168032699 | -1.141758959 | 0.253554226 | 0.833793221 |
| ENSG000000001036.14 | 2279.02762  | -0.093557583   | 0.093903702 | -0.996314104 | 0.319097553 | 0.858519271 |
| ENSG000000001084.13 | 1431.150204 | 0.032545846    | 0.101872792 | 0.319475351  | 0.74936608  | 0.964985419 |
| ENSG000000001167.14 | 1230.250061 | 0.006756289    | 0.085543767 | 0.078980491  | 0.93704814  | 0.993330805 |
| ENSG000000001460.18 | 547.8525767 | 0.05009834     | 0.089516332 | 0.559655861  | 0.575714194 | 0.935406267 |
| ENSG000000001461.17 | 1906.034633 | 0.045313516    | 0.08656687  | 0.523451018  | 0.600660429 | 0.94036285  |
| ENSG000000001497.18 | 1891.025498 | 0.060406616    | 0.066091758 | 0.913981077  | 0.360726789 | 0.874186506 |
| ENSG000000001561.7  | 1323.931253 | -0.154474511   | 0.117448744 | -1.315250426 | 0.188425753 | 0.791601992 |
| ENSG000000001617.12 | 1751.673571 | 0.283104538    | 0.141268774 | 2.004013558  | 0.045068608 | 0.56550954  |
| ENSG000000001626.16 | 11963.27278 | -0.123661622   | 0.338121766 | -0.365731031 | 0.714565789 | 0.960033547 |
| ENSG000000001629.10 | 2569.709708 | 0.083403368    | 0.087836301 | 0.94953188   | 0.342350166 | 0.865416072 |
| ENSG000000001630.17 | 143.4654058 | -0.040892475   | 0.186641766 | -0.219096056 | 0.826575223 | 0.974781227 |
| ENSG000000001631.16 | 289.5709694 | -0.048554021   | 0.069756892 | -0.696046217 | 0.486399887 | 0.914944178 |
| ENSG000000002016.18 | 428.3115002 | -0.020950042   | 0.114117238 | -0.1835835   | 0.854340208 | 0.980373784 |
| ENSG000000002079.14 | 112.42899   | 0.558163636    | 0.30526045  | 1.828483306  | 0.067477049 | 0.627412043 |
| ENSG000000002330.14 | 504.0320588 | 0.040434787    | 0.169698061 | 0.238274889  | 0.811667897 | 0.972484804 |
| ENSG000000002549.12 | 4585.634859 | 0.0189525      | 0.112781062 | 0.168046833  | 0.866546435 | 0.98148853  |
| ENSG000000002586.20 | 12966.55333 | -0.107693421   | 0.111326909 | -0.967362    | 0.3333631   | 0.861852833 |
| ENSG000000002587.10 | 2449.16346  | 0.077554196    | 0.22036169  | 0.351940465  | 0.724882911 | 0.962254794 |
| ENSG000000002726.21 | 4883.372526 | -0.397102188   | 0.284854217 | -1.394054092 | 0.163301271 | 0.769104902 |
| ENSG000000002745.13 | 20.646992   | -0.095831535   | 0.256866833 | -0.373078663 | 0.709089898 | 0.959381803 |
| ENSG000000002746.15 | 109.3043072 | -0.505167823   | 0.259049822 | -1.950079794 | 0.051166609 | 0.583582379 |
| ENSG000000002822.15 | 39.38688283 | 0.176869554    | 0.198800185 | 0.889685061  | 0.373635018 | 0.878574476 |
| ENSG000000002834.18 | 12288.58863 | 0.048227368    | 0.077430581 | 0.622846518  | 0.533385386 | 0.927346465 |
| ENSG000000002919.15 | 1282.435397 | -0.007219656   | 0.069538899 | -0.103821832 | 0.917310741 | 0.991096228 |
| ENSG000000002933.9  | 6713.284093 | -0.169488193   | 0.171225209 | -0.989855373 | 0.322244815 | 0.860170302 |
| ENSG000000003056.8  | 2773.999079 | 0.116771263    | 0.104080291 | 1.121934442  | 0.261890315 | 0.838346903 |
| ENSG000000003096.14 | 167.6002272 | -0.518238733   | 0.182327689 | -2.842347944 | 0.004478259 | 0.240133751 |
| ENSG000000003137.8  | 165.8299816 | -0.10524388    | 0.193405185 | -0.544162659 | 0.586329545 | 0.93763709  |
| ENSG000000003147.19 | 2296.588972 | -0.076395352   | 0.125951625 | -0.606545193 | 0.544152782 | 0.928847942 |
| ENSG000000003249.15 | 929.7809919 | 0.158336366    | 0.171593583 | 0.922740601  | 0.356142402 | 0.871848899 |
| ENSG000000003393.15 | 1323.696821 | -0.132480057   | 0.070269298 | -1.885319218 | 0.059386753 | 0.603812862 |
| ENSG000000003400.15 | 1427.367599 | -0.075514367   | 0.11523502  | -0.655307447 | 0.512269827 | 0.921689413 |
| ENSG000000003402.20 | 6239.843505 | -0.044515773   | 0.082707222 | -0.538233198 | 0.590416064 | 0.938648288 |
| ENSG000000003436.16 | 4162.591376 | 0.077372382    | 0.155975449 | 0.496054876  | 0.619855695 | 0.94373369  |
| ENSG000000003509.16 | 543.8444172 | 0.028559184    | 0.057583101 | 0.495964678  | 0.619919332 | 0.94373369  |
| ENSG000000003756.17 | 3255.489683 | -0.044907409   | 0.088785588 | -0.50579615  | 0.612999762 | 0.942338837 |
| ENSG000000003987.14 | 115.8488823 | -0.509186288   | 0.231752567 | -2.19711175  | 0.028012466 | 0.482204613 |
| ENSG000000003989.18 | 2925.136177 | 0.091264994    | 0.278008073 | 0.328281813  | 0.742698594 | 0.964576795 |
| ENSG000000004059.11 | 4463.548201 | -0.024854718   | 0.079472876 | -0.312744663 | 0.75447466  | 0.966152619 |
| ENSG000000004139.14 | 320.8215145 | -0.049216172   | 0.124898153 | -0.39405044  | 0.6935438   | 0.957691279 |
| ENSG000000004142.12 | 4419.842925 | 0.125665349    | 0.064356232 | 1.952652385  | 0.050860799 | 0.582475355 |
| ENSG000000004399.13 | 6266.090234 | 0.059707729    | 0.102277519 | 0.583781551  | 0.559367292 | 0.932627608 |

**Fig 5.5: Result of DEA Treated vs Not treated**

## 5.7.2 Treated VS Normal

|                     | baseMean    | log2FoldChange | lfcSE       | stat         | pvalue      | padj        |
|---------------------|-------------|----------------|-------------|--------------|-------------|-------------|
| ENSG00000000003.15  | 1804.081313 | 0.318442459    | 0.352890346 | 0.902383596  | 0.366853132 | 0.999996589 |
| ENSG00000000005.6   | 13.02124896 | -1.728165488   | 1.46521466  | -1.179462324 | 0.23821413  | 0.999996589 |
| ENSG000000000419.13 | 1420.733241 | -0.251985181   | 0.194647661 | -1.294570815 | 0.195468326 | 0.999996589 |
| ENSG000000000457.14 | 714.3607259 | -0.003909962   | 0.237904103 | -0.016435034 | 0.986887331 | 0.999996589 |
| ENSG000000000460.17 | 216.0645361 | 0.006800851    | 0.278201567 | 0.024445769  | 0.980497041 | 0.999996589 |
| ENSG000000000938.13 | 897.6750832 | 0.017279112    | 0.548502796 | 0.031502322  | 0.974868941 | 0.999996589 |
| ENSG000000000971.16 | 6077.562164 | 0.475738595    | 0.471756747 | 1.00844047   | 0.313243053 | 0.999996589 |
| ENSG00000001036.14  | 2279.02762  | 0.17900053     | 0.263691593 | 0.67882532   | 0.497248549 | 0.999996589 |
| ENSG00000001084.13  | 1431.150204 | -0.067922346   | 0.286174468 | -0.237345932 | 0.812388432 | 0.999996589 |
| ENSG00000001167.14  | 1230.250061 | 0.02946667     | 0.24033042  | 0.122608989  | 0.902416736 | 0.999996589 |
| ENSG00000001460.18  | 547.8525767 | 0.191814082    | 0.251372989 | 0.763065603  | 0.445424272 | 0.999996589 |
| ENSG00000001461.17  | 1906.034633 | -0.033676925   | 0.243161842 | -0.138495928 | 0.889848487 | 0.999996589 |
| ENSG00000001497.18  | 1891.025498 | -0.199208156   | 0.185823111 | -1.072031106 | 0.283706061 | 0.999996589 |
| ENSG00000001561.7   | 1323.931253 | -0.130738741   | 0.329997964 | -0.396180447 | 0.691971915 | 0.999996589 |
| ENSG00000001617.12  | 1751.673571 | -0.015073719   | 0.396636331 | -0.038003878 | 0.96968459  | 0.999996589 |
| ENSG00000001626.16  | 11963.27278 | 0.759987323    | 0.949286717 | 0.800587756  | 0.423370341 | 0.999996589 |
| ENSG00000001629.10  | 2569.709708 | 0.024481873    | 0.246662941 | 0.099252334  | 0.920937923 | 0.999996589 |
| ENSG00000001630.17  | 143.4654058 | -1.058708667   | 0.529480244 | -1.999524399 | 0.045551645 | 0.999996589 |
| ENSG00000001631.16  | 289.5709694 | -0.203973403   | 0.197815255 | -1.031130804 | 0.302479484 | 0.999996589 |
| ENSG00000002016.18  | 428.3115002 | -0.148812001   | 0.321076653 | -0.463478113 | 0.6430217   | 0.999996589 |
| ENSG00000002079.14  | 112.42899   | -0.783515293   | 0.858845764 | -0.912288709 | 0.361616757 | 0.999996589 |
| ENSG00000002330.14  | 504.0320588 | -0.43711419    | 0.477117164 | -0.916156916 | 0.359584599 | 0.999996589 |
| ENSG00000002549.12  | 4585.634859 | -0.165767456   | 0.316702209 | -0.52341743  | 0.600683798 | 0.999996589 |
| ENSG00000002586.20  | 12966.55333 | 0.070495756    | 0.312570011 | 0.22553589   | 0.821562408 | 0.999996589 |
| ENSG00000002587.10  | 2449.16346  | -0.850735055   | 0.618875064 | -1.374647493 | 0.169240757 | 0.999996589 |
| ENSG00000002726.21  | 4883.372526 | -0.884934365   | 0.799846798 | -1.10637983  | 0.268562152 | 0.999996589 |
| ENSG00000002745.13  | 20.646992   | -0.105591901   | 0.727509872 | -0.145141537 | 0.884599124 | 0.999996589 |
| ENSG00000002746.15  | 109.3043072 | 0.014600477    | 0.728888932 | 0.02003114   | 0.984018531 | 0.999996589 |
| ENSG00000002822.15  | 39.38688283 | 0.200161448    | 0.557490013 | 0.359040419  | 0.719564853 | 0.999996589 |
| ENSG00000002834.18  | 12288.58863 | 0.057216226    | 0.217403591 | 0.263179765  | 0.792412027 | 0.999996589 |
| ENSG00000002919.15  | 1282.435397 | -0.288730837   | 0.19574817  | -1.47501168  | 0.140209403 | 0.999996589 |
| ENSG00000002933.9   | 6713.284093 | -0.511644367   | 0.480792988 | -1.064167697 | 0.287252738 | 0.999996589 |
| ENSG00000003056.8   | 2773.999079 | 0.184877688    | 0.29220821  | 0.632691629  | 0.526935037 | 0.999996589 |
| ENSG00000003096.14  | 167.6002272 | 0.31611019     | 0.512612234 | 0.616665325  | 0.537455497 | 0.999996589 |
| ENSG00000003137.8   | 165.8299816 | -0.712917565   | 0.546121584 | -1.305419134 | 0.191750167 | 0.999996589 |
| ENSG00000003147.19  | 2296.588972 | -0.400577392   | 0.353828779 | -1.132122132 | 0.257583092 | 0.999996589 |
| ENSG00000003249.15  | 929.7809919 | -0.320873422   | 0.48200157  | -0.665710325 | 0.505596273 | 0.999996589 |
| ENSG00000003393.15  | 1323.696821 | 0.054762615    | 0.197527965 | 0.277239806  | 0.781595976 | 0.999996589 |
| ENSG00000003400.15  | 1427.367599 | -0.217424283   | 0.323790359 | -0.671497086 | 0.501903915 | 0.999996589 |
| ENSG00000003402.20  | 6239.843505 | -0.196022551   | 0.232281049 | -0.843902468 | 0.398723909 | 0.999996589 |
| ENSG00000003436.16  | 4162.591376 | 0.446861257    | 0.437885062 | 1.020498976  | 0.307491875 | 0.999996589 |
| ENSG00000003509.16  | 543.8444172 | -0.102202569   | 0.162514067 | -0.628884444 | 0.529424711 | 0.999996589 |
| ENSG00000003756.17  | 3255.489683 | -0.216843481   | 0.249412993 | -0.869415336 | 0.384619998 | 0.999996589 |
| ENSG00000003987.14  | 115.8488823 | -0.741915451   | 0.655539431 | -1.131763271 | 0.257733973 | 0.999996589 |
| ENSG00000003989.18  | 2925.136177 | -0.581728229   | 0.780608785 | -0.745223779 | 0.45613645  | 0.999996589 |
| ENSG00000004059.11  | 4463.548201 | -0.111586391   | 0.223210456 | -0.499915608 | 0.617134502 | 0.999996589 |
| ENSG00000004139.14  | 320.8215145 | -0.495511317   | 0.352406501 | -1.406078817 | 0.159700722 | 0.999996589 |
| ENSG00000004142.12  | 4419.842925 | -0.113540476   | 0.180760551 | -0.628126414 | 0.529921131 | 0.999996589 |
| ENSG00000004399.13  | 6266.090234 | 0.481962418    | 0.287123373 | 1.678589985  | 0.09323198  | 0.999996589 |

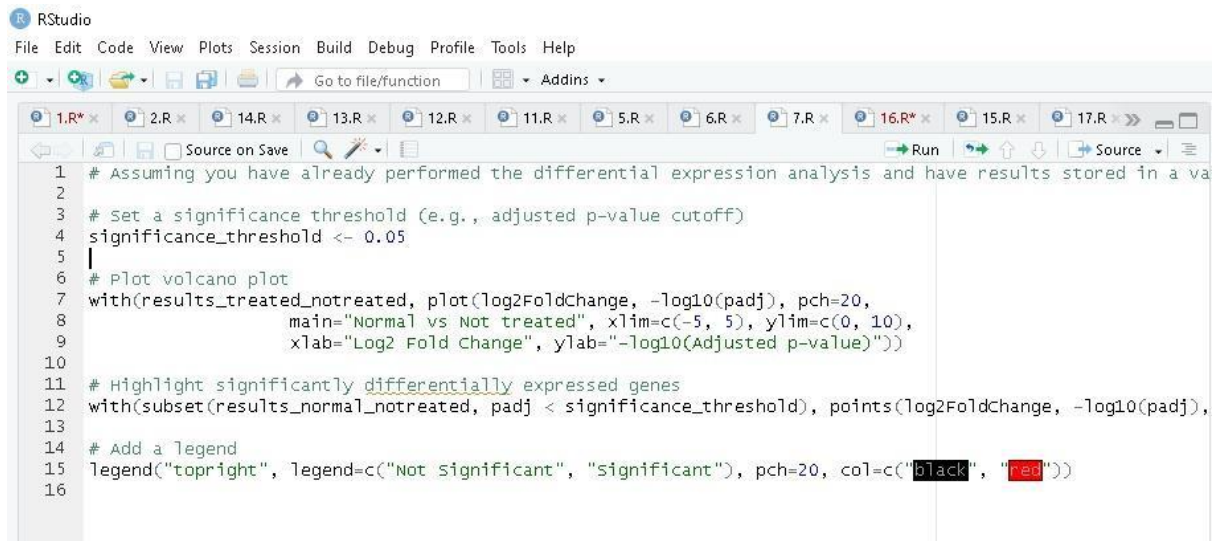
**FIG 5.6: Result of DEA Treated vs Normal**

### 5.7.3 Normal vs Not treated

|                     | baseMean    | log2FoldChange | lfcSE       | stat         | pvalue      | padj        |
|---------------------|-------------|----------------|-------------|--------------|-------------|-------------|
| ENSG00000000003.15  | 1804.081313 | 0.398905606    | 0.364551914 | 1.094235392  | 0.273851753 | 0.999444956 |
| ENSG00000000005.6   | 13.02124896 | -1.552603484   | 1.511152904 | -1.027429772 | 0.304218138 | 0.999444956 |
| ENSG000000000419.13 | 1420.733241 | -0.207231258   | 0.20105364  | -1.030726217 | 0.302669227 | 0.999444956 |
| ENSG000000000457.14 | 714.3607259 | -0.040974934   | 0.245737991 | -0.166742368 | 0.867572767 | 0.999444956 |
| ENSG000000000460.17 | 216.0645361 | 0.182269656    | 0.287390884 | 0.634222121  | 0.525935873 | 0.999444956 |
| ENSG000000000938.13 | 897.6750832 | -0.015372923   | 0.566615298 | -0.027131147 | 0.978355132 | 0.999444956 |
| ENSG000000000971.16 | 6077.562164 | 0.283885756    | 0.487341565 | 0.582519071  | 0.560217099 | 0.999444956 |
| ENSG00000001036.14  | 2279.02762  | 0.085442947    | 0.272398309 | 0.313669153  | 0.753772332 | 0.999444956 |
| ENSG00000001084.13  | 1431.150204 | -0.0353765     | 0.295618854 | -0.119669295 | 0.904745125 | 0.999444956 |
| ENSG00000001167.14  | 1230.250061 | 0.036222959    | 0.248259768 | 0.145907486  | 0.883994422 | 0.999444956 |
| ENSG00000001460.18  | 547.8525767 | 0.241912422    | 0.259677354 | 0.931588441  | 0.351549259 | 0.999444956 |
| ENSG00000001461.17  | 1906.034633 | 0.011636591    | 0.251188036 | 0.046326217  | 0.963050244 | 0.999444956 |
| ENSG00000001497.18  | 1891.025498 | -0.13880154    | 0.191946197 | -0.723127324 | 0.469601664 | 0.999444956 |
| ENSG00000001561.7   | 1323.931253 | -0.285213251   | 0.340880412 | -0.836695925 | 0.40276351  | 0.999444956 |
| ENSG00000001617.12  | 1751.673571 | 0.268030819    | 0.409742392 | 0.654144713  | 0.513018577 | 0.999444956 |
| ENSG00000001626.16  | 11963.27278 | 0.636325701    | 0.980648039 | 0.648882857  | 0.516414096 | 0.999444956 |
| ENSG00000001629.10  | 2569.709708 | 0.10788524     | 0.254809187 | 0.423396195  | 0.67200622  | 0.999444956 |
| ENSG00000001630.17  | 143.4654058 | -1.099601142   | 0.546613757 | -2.011660204 | 0.044255768 | 0.999444956 |
| ENSG00000001631.16  | 289.5709694 | -0.252527424   | 0.204216209 | -1.236568954 | 0.216247152 | 0.999444956 |
| ENSG00000002016.18  | 428.3115002 | -0.169762043   | 0.331638028 | -0.511889557 | 0.608728307 | 0.999444956 |
| ENSG00000002079.14  | 112.42899   | -0.225351657   | 0.887157037 | -0.254015521 | 0.799483567 | 0.999444956 |
| ENSG00000002330.14  | 504.0320588 | -0.396679403   | 0.492836607 | -0.804890297 | 0.420882988 | 0.999444956 |
| ENSG00000002549.12  | 4585.634859 | -0.146814956   | 0.327160893 | -0.4487546   | 0.653608692 | 0.999444956 |
| ENSG00000002586.20  | 12966.55333 | -0.037197665   | 0.322895067 | -0.115200476 | 0.908286221 | 0.999444956 |
| ENSG00000002587.10  | 2449.16346  | -0.773180859   | 0.639308005 | -1.20940275  | 0.226508153 | 0.999444956 |
| ENSG00000002726.21  | 4883.372526 | -1.282036552   | 0.826263236 | -1.551607885 | 0.120756075 | 0.999444956 |
| ENSG00000002745.13  | 20.646992   | -0.201423436   | 0.751081769 | -0.268177772 | 0.788562484 | 0.999444956 |
| ENSG00000002746.15  | 109.3043072 | -0.490567347   | 0.752814852 | -0.651644087 | 0.5146308   | 0.999444956 |
| ENSG00000002822.15  | 39.38688283 | 0.377031002    | 0.57602235  | 0.654542314  | 0.512762475 | 0.999444956 |
| ENSG00000002834.18  | 12288.58863 | 0.105443593    | 0.224585122 | 0.469503912  | 0.63870949  | 0.999444956 |
| ENSG00000002919.15  | 1282.435397 | -0.295950493   | 0.202181384 | -1.463787057 | 0.14325214  | 0.999444956 |
| ENSG00000002933.9   | 6713.284093 | -0.681132559   | 0.496671704 | -1.371393929 | 0.170252189 | 0.999444956 |
| ENSG00000003056.8   | 2773.999079 | 0.301648951    | 0.301863164 | 0.999290364  | 0.317654052 | 0.999444956 |
| ENSG00000003096.14  | 167.6002272 | -0.202128543   | 0.529452362 | -0.381769084 | 0.702632653 | 0.999444956 |
| ENSG00000003137.8   | 165.8299816 | -0.818161445   | 0.563951081 | -1.450766693 | 0.146844837 | 0.999444956 |
| ENSG00000003147.19  | 2296.588972 | -0.476972744   | 0.365503365 | -1.304974975 | 0.191901369 | 0.999444956 |
| ENSG00000003249.15  | 929.7809919 | -0.162537056   | 0.497912453 | -0.326437017 | 0.744093735 | 0.999444956 |
| ENSG00000003393.15  | 1323.696821 | -0.077717442   | 0.204033535 | -0.380905239 | 0.703273565 | 0.999444956 |
| ENSG00000003400.15  | 1427.367599 | -0.29293865    | 0.334468815 | -0.875832473 | 0.381121114 | 0.999444956 |
| ENSG00000003402.20  | 6239.843505 | -0.240538324   | 0.239949563 | -1.002453683 | 0.316124526 | 0.999444956 |
| ENSG00000003436.16  | 4162.591376 | 0.52423364     | 0.452353145 | 1.158903493  | 0.246495525 | 0.999444956 |
| ENSG00000003509.16  | 543.8444172 | -0.073643385   | 0.167831047 | -0.438794767 | 0.660810252 | 0.999444956 |
| ENSG00000003756.17  | 3255.489683 | -0.26175089    | 0.257642948 | -1.015944322 | 0.309655901 | 0.999444956 |
| ENSG00000003987.14  | 115.8488823 | -1.251101739   | 0.67682694  | -1.848480999 | 0.064532791 | 0.999444956 |
| ENSG00000003989.18  | 2925.136177 | -0.490463235   | 0.806391948 | -0.608219409 | 0.543041963 | 0.999444956 |
| ENSG00000004059.11  | 4463.548201 | -0.136441109   | 0.23057871  | -0.591733333 | 0.554029172 | 0.999444956 |
| ENSG00000004139.14  | 320.8215145 | -0.544727489   | 0.363931715 | -1.496784879 | 0.134449242 | 0.999444956 |
| ENSG00000004142.12  | 4419.842925 | 0.012124873    | 0.186728611 | 0.064933128  | 0.948227244 | 0.999444956 |
| ENSG00000004399.13  | 6266.090234 | 0.541670146    | 0.296610883 | 1.826197813  | 0.067820471 | 0.999444956 |

**Fig 5.7: Result of DEA Normal vs Not treated**

## 5.8 VOLCANO PLOT

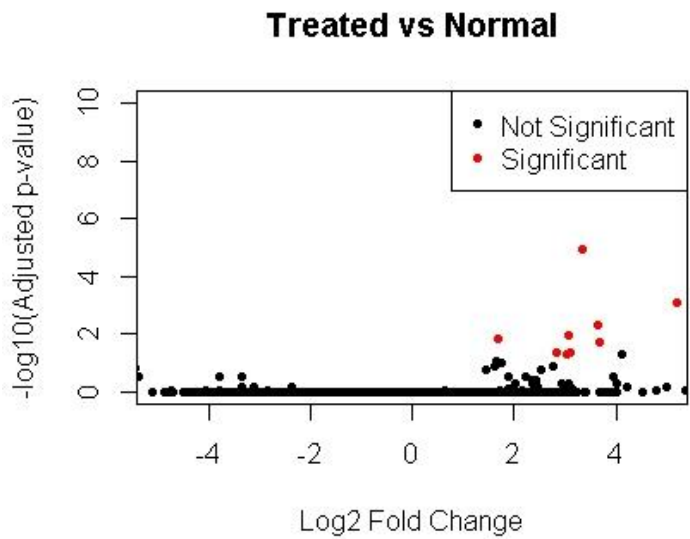
The image shows the RStudio interface with a script editor containing R code for creating a volcano plot. The code is as follows:

```
1 # Assuming you have already performed the differential expression analysis and have results stored in a variable named 'results'
2
3 # Set a significance threshold (e.g., adjusted p-value cutoff)
4 significance_threshold <- 0.05
5
6 # Plot volcano plot
7 with(results_treated_notreated, plot(log2FoldChange, -log10(padj), pch=20,
8   main="Normal vs Not treated", xlim=c(-5, 5), ylim=c(0, 10),
9   xlab="Log2 Fold Change", ylab="-log10(Adjusted p-value)"))
10
11 # Highlight significantly differentially expressed genes
12 with(subset(results_normal_notreated, padj < significance_threshold), points(log2FoldChange, -log10(padj),
13   pch=20, col="red"))
14
15 # Add a legend
16 legend("topright", legend=c("Not significant", "significant"), pch=20, col=c("black", "red"))
```

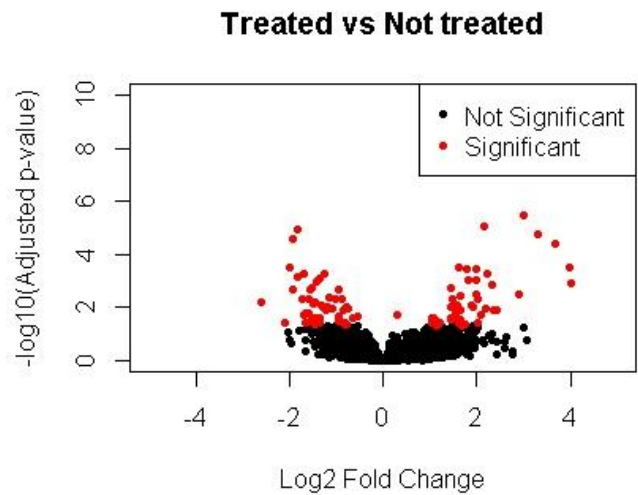
**Fig 5.8: Code to create volcano plot for pancreatic cancer using R**

This R code snippet is used to create a volcano plot, a common visualization method for displaying results from differential expression analysis. Assuming that the differential expression analysis results are stored in a variable named 'results', the first step sets a significance threshold, typically a corrected p-value cutoff, which determines the genes considered significantly differentially expressed. The volcano plot is then generated using the 'plot' function. The x-axis represents the log2 fold change, which quantifies the magnitude of gene expression differences between the two conditions (in this case, normal vs. not treated), while the y-axis represents the negative log10-transformed adjusted p-values. Adjusted p-values account for multiple testing and are commonly used to assess the statistical significance of differential expression. In the plot, each data point represents a gene, with its position determined by its log2 fold change and adjusted p-value. The 'pch=20' argument specifies the point character as a solid circle. Significantly differentially expressed genes are highlighted in red using the 'points' function, which overlays red circles on top of the corresponding data points with adjusted p-values below the significance threshold. Finally, a legend is added to the plot using the 'legend' function, indicating the colors used to represent significant and non-significant genes. This visualization allows for easy identification of genes with substantial changes in expression levels and high statistical significance, aiding in the

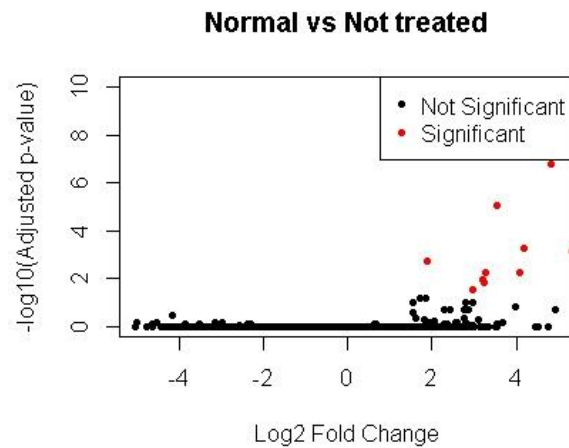
prioritization of candidate genes for further investigation. The volcano plot effectively conveys both the magnitude and statistical confidence of gene expression differences between conditions, providing valuable insights into the underlying biology of the studied phenomenon.



**Fig 5.9: Volcano plot for Treated vs Normal in pancreatic cancer**



**Fig 5.10: Volcano plot for Treated vs Not treated in pancreatic cancer**



**Fig 5.11: Volcano plot for Normal vs Not treated in pancreatic cancer**

### **5.8.1 X-axis(fold change)**

The X-axis typically represents the fold change in gene expression between two conditions. Each point on the plot represents a gene, with its position on the X-axis indicating how much its expression level changes between the two conditions. For example, genes that are upregulated in condition A compared to condition B will be located on the right side of the plot, while genes that are downregulated will be on the left side.

### **5.8.2 Y-axis(statistical significance)**

The Y-axis represents the statistical significance of the observed changes in gene expression. It is usually represented as the negative logarithm of the p-value obtained from statistical tests such as t-tests or ANOVA. The higher the point on the Y-axis, the more statistically significant the change in gene expression.

### **5.8.3 Plotting points**

Each gene is represented by a point on the plot, with its position determined by both its fold change and statistical significance. Genes with a large fold change and high statistical significance are typically considered the most interesting, as they represent the most pronounced and reliable changes in expression between the two conditions.

#### 5.8.4 Highlighting significant genes

I have used different colors to highlight genes that meet predefined criteria for statistical significance, such as an adjusted p-value below a certain threshold (e.g.,  $p < 0.05$ ) or a fold change greater than a specified value.

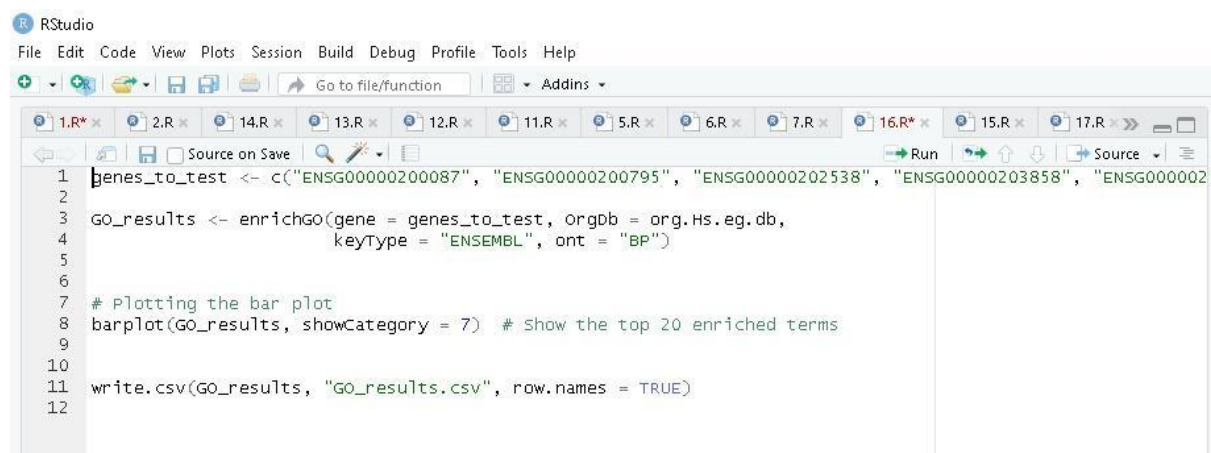
### 5.8.5 Interpretation

By examining the volcano plot, we can see that the upregulated genes are more likely to be involved with the pancreatic cancer

### 5.8.6 Filtering

Then the Highly expressed genes are filtered and compared with other results which led to 15 genes that are most likely involved

## 5.9 GO ENRICHMENT ANALYSIS OF BIOLOGICAL PROCESS IN FILTERED GENES

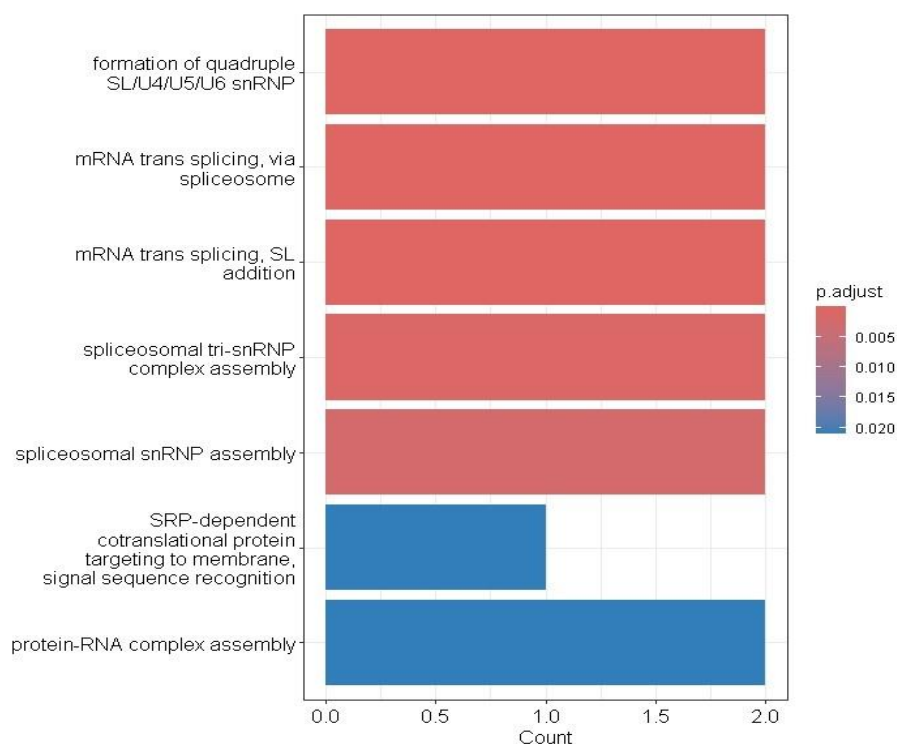


**Fig 5.12: Code for GO enrichment analysis using R**

This R code snippet performs Gene Ontology (GO) enrichment analysis for a set of genes of interest, represented by their Ensembl gene IDs. Firstly, the `genes_to_test` vector contains a list of Ensembl gene IDs for the genes under investigation. These genes may have been identified through prior analysis, such as differential expression analysis, and are now being further explored for their functional

relevance. The `enrichGO` function from the `clusterProfiler` package is then used to conduct the GO enrichment analysis. This function takes several parameters, including the input gene list (`gene`), the organism database `OrgDB`, here specified as `org.hs.eg.db` for *Homo sapiens*), the type of key identifier used (`keytype`, set to "ENSEMBL" to match the input gene IDs), and the ontology aspect of interest (`onf`, specified as "BP" for biological process).

The results of the GO enrichment analysis are stored in the `GO_results` object, which contains information about enriched GO terms along with statistical significance measures such as p-values and adjusted p-values. To visualize the enriched GO terms, a bar plot is created using the `barplot` function. The `showcategory` parameter specifies the number of top enriched terms to display in the plot. Here, `showcategory = 7` indicates that the top 7 enriched terms will be shown in the bar plot. Finally, the enriched GO results are saved as a CSV file named "GO\_results.csv" using the `write.csv` function. Setting `rownames = true` ensures that row names (corresponding to the GO terms) are included in the CSV file.



**Fig 5.13: GO enrichment analysis barplot**

| ID         | Description                                                                              | GeneRatio | BgRatio   | pvalue      | p.adjust    | qvalue      | geneID                            |
|------------|------------------------------------------------------------------------------------------|-----------|-----------|-------------|-------------|-------------|-----------------------------------|
| GO:0000353 | formation of quadruple SL/U4/U5/U6 snRNP                                                 | 02-Nov    | 10/21261  | 1.09E-05    | 7.28E-05    | 1.53E-05    | ENSG000000200795/ENSG000000202538 |
| GO:0000365 | mRNA trans splicing, via spliceosome                                                     | 02-Nov    | 10/21261  | 1.09E-05    | 7.28E-05    | 1.53E-05    | ENSG000000200795/ENSG000000202538 |
| GO:0045291 | mRNA trans splicing, SL addition                                                         | 02-Nov    | 10/21261  | 1.09E-05    | 7.28E-05    | 1.53E-05    | ENSG000000200795/ENSG000000202538 |
| GO:0000244 | spliceosomal tri-snRNP complex assembly                                                  | 02-Nov    | 41/21261  | 0.000197369 | 0.000986844 | 0.000207757 | ENSG000000200795/ENSG000000202538 |
| GO:0000387 | spliceosomal snRNP assembly                                                              | 02-Nov    | 69/21261  | 0.000560222 | 0.002240889 | 0.000471766 | ENSG000000200795/ENSG000000202538 |
| GO:0006617 | SRP-dependent cotranslational protein targeting to membrane, signal sequence recognition | 01-Nov    | 13/21261  | 0.006706978 | 0.021073021 | 0.004436426 | ENSG000000276168                  |
| GO:0022618 | protein-RNA complex assembly                                                             | 02-Nov    | 266/21261 | 0.007961257 | 0.021073021 | 0.004436426 | ENSG000000200795/ENSG000000202538 |
| GO:0071826 | protein-RNA complex organization                                                         | 02-Nov    | 274/21261 | 0.008429209 | 0.021073021 | 0.004436426 | ENSG000000200795/ENSG000000202538 |
| GO:0006614 | SRP-dependent cotranslational protein targeting to membrane                              | 01-Nov    | 26/21261  | 0.013373036 | 0.024595974 | 0.0051781   | ENSG000000276168                  |
| GO:0000377 | RNA splicing, via transesterification reactions with bulged adenosine as nucleophile     | 02-Nov    | 379/21261 | 0.015672433 | 0.024595974 | 0.0051781   | ENSG000000200795/ENSG000000202538 |
| GO:0000398 | mRNA splicing, via spliceosome                                                           | 02-Nov    | 379/21261 | 0.015672433 | 0.024595974 | 0.0051781   | ENSG000000200795/ENSG000000202538 |
| GO:0006613 | cotranslational protein targeting to membrane                                            | 01-Nov    | 31/21261  | 0.015926057 | 0.024595974 | 0.0051781   | ENSG000000276168                  |
| GO:0000375 | RNA splicing, via transesterification reactions                                          | 02-Nov    | 383/21261 | 0.015987383 | 0.024595974 | 0.0051781   | ENSG000000200795/ENSG000000202538 |
| GO:0045047 | protein targeting to ER                                                                  | 01-Nov    | 53/21261  | 0.027088154 | 0.038697363 | 0.008146813 | ENSG000000276168                  |
| GO:0072599 | establishment of protein localization to endoplasmic reticulum                           | 01-Nov    | 57/21261  | 0.029105216 | 0.038806955 | 0.008169885 | ENSG000000276168                  |

**Fig 5.14: Results of Gene ontology enrichment analysis**

Formation of quadruple SL/U4/U5/U6 snRNP: This process involves the assembly of the spliceosomal complex, a crucial machinery responsible for pre-mRNA splicing. The formation of quadruple small nuclear ribonucleoprotein particles (snRNPs) comprising U4, U5, and U6 snRNAs is essential for spliceosome function and proper RNA splicing.

mRNA trans splicing, via spliceosome: Trans-splicing is a mechanism where exons from different pre-mRNAs are joined together to produce a mature mRNA. This process, facilitated by the spliceosome, allows for the diversification of transcripts and can contribute to gene regulation and protein diversity.

mRNA trans splicing, SL addition: This process involves the addition of a spliced leader (SL) sequence to mRNA transcripts during trans-splicing. SL addition is prevalent in organisms like trypanosomes and nematodes and plays a role in mRNA processing and translation initiation.

Spliceosomal tri-snRNP complex assembly: Tri-snRNP complexes are essential components of the spliceosome involved in pre-mRNA splicing. Assembly of these complexes is critical for spliceosome function and proper mRNA processing.

Spliceosomal snRNP assembly: This process refers to the assembly of small nuclear ribonucleoprotein particles (snRNPs), which are key components of the spliceosome. Proper assembly of snRNPs is necessary for spliceosome activity and accurate splicing of pre-mRNA transcripts.

SRP-dependent cotranslational protein targeting to membrane, signal sequence recognition: Signal recognition particle (SRP)-dependent cotranslational targeting is a mechanism for directing nascent polypeptides to the endoplasmic reticulum (ER) membrane for membrane insertion or secretion. This process involves the recognition of signal sequences in the nascent polypeptide chain by the SRP complex. Protein-RNA complex assembly: This process encompasses the formation of complexes between proteins and RNA molecules. Such complexes play diverse roles in cellular processes, including RNA processing, regulation of gene expression, and post-transcriptional modifications. Protein-RNA complex organization: Once formed, protein-RNA complexes undergo organization and

structural rearrangements to perform specific functions. This process ensures proper coordination and functionality of the complex in various cellular processes.

Protein targeting to ER: This process involves the delivery of newly synthesized proteins to the endoplasmic reticulum (ER) for proper folding, modification, and trafficking. Protein targeting to the ER is essential for maintaining cellular homeostasis and is critical for the biogenesis of secretory and membrane proteins.

Establishment of protein localization to endoplasmic reticulum: This process involves the initial steps in targeting and anchoring proteins to the endoplasmic reticulum membrane, ensuring their proper localization and function within the ER compartment. From the enriched biological processes these three potential genes that could serve as biomarkers:

ENSG00000200795: This gene is associated with several enriched biological processes, including mRNA Trans splicing, spliceosomal complex assembly, and protein-RNA complex organization. Given its involvement in these critical cellular processes, dysregulation of ENSG00000200795 could potentially serve as a biomarker for aberrant splicing events or disruptions in RNA processing pathways, which are commonly observed in various diseases, including cancer.

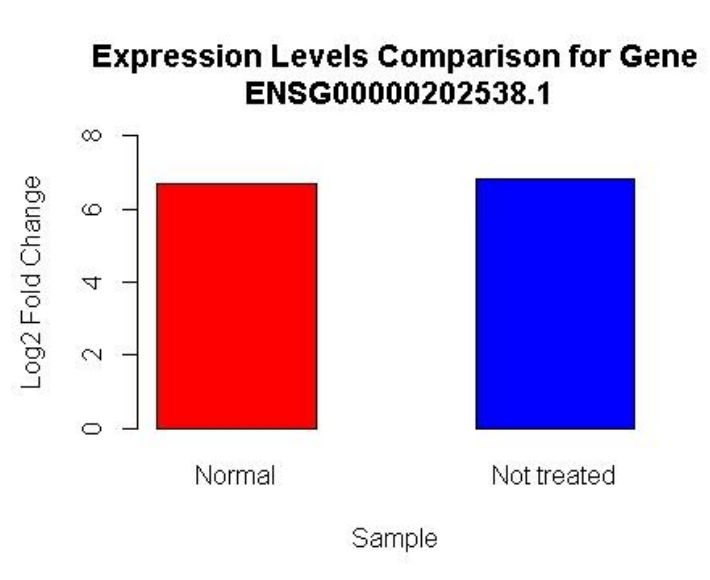
ENSG00000202538: Similar to ENSG00000200795, this gene is also implicated in mRNA Trans splicing, spliceosomal complex assembly, and protein-RNA complex organization. Dysregulation of ENSG00000202538 may reflect abnormalities in RNA processing machinery and spliceosome function, making it a candidate biomarker for conditions associated with splicing defects, such as certain types of cancer or genetic disorders.

ENSG00000276168: This gene is involved in SRP-dependent cotranslational protein targeting to the membrane and protein targeting to the endoplasmic reticulum (ER). Dysregulation of ENSG00000276168 could indicate disruptions in protein localization and trafficking processes, which are crucial for cellular homeostasis and proper function of the ER. Aberrant expression or activity of this gene may thus serve as a biomarker for ER-related stress conditions, such as ER stress in cancer cells or neurodegenerative diseases.

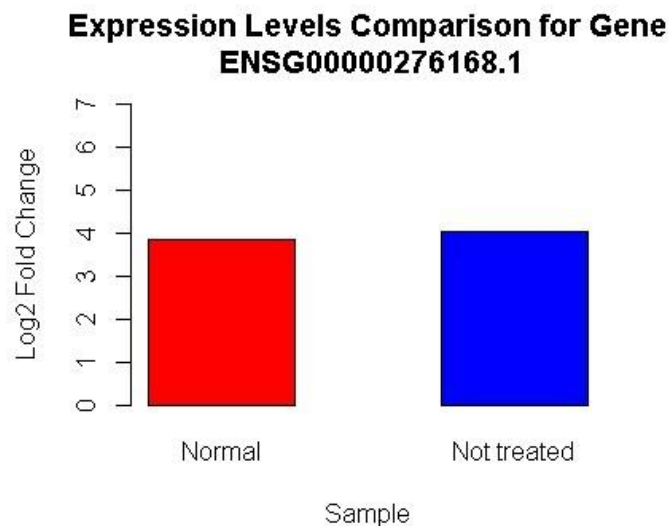
## 5.10 DATA VISUALIZATION OF EXPRESSION LEVEL IN FILTERED GENES

Data visualization has been done via bar plots as it is efficient for comparing expression level of single gene

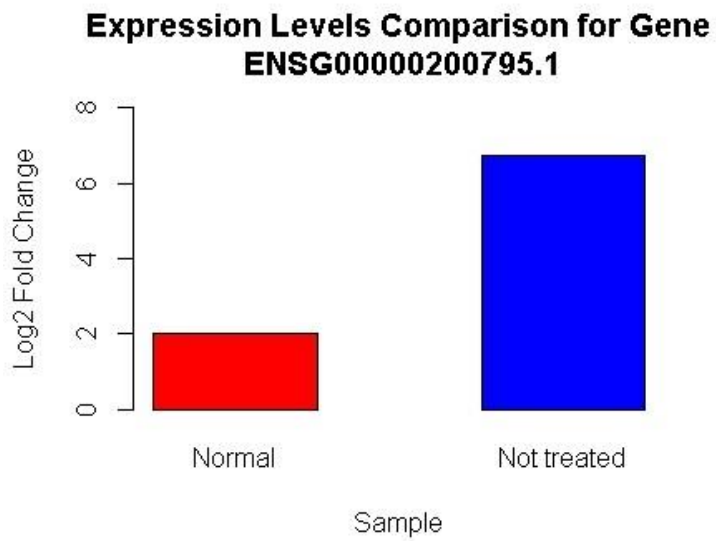
By comparing the expression levels of each candidate gene with normally expressed genes, identified a single gene which could be a potential biomarker.



**Fig 5.15: Expression barplot for gene ENSG00000202538**



**Fig 5.16: Expression barplot for gene ENSG00000276168**



***Fig 5.17: Expression barplot for gene ENSG00000200795***

From these barplots I have identified that gene ENSG00000200795 - RN7SL1 could be an potential biomarker for pancreatic cancer

## **CHAPTER – 6**

### **SUMMARY AND CONCLUSIONS**

The findings of this study underscore the complexity of pancreatic cancer biology and the importance of integrated analyses in biomarker discovery. The identified gene offer valuable insights into the molecular mechanisms underlying pancreatic cancer pathogenesis and provide potential targets for therapeutic intervention.

In conclusion, this project employed differential expression analysis techniques to identify potential biomarkers for pancreatic cancer. The integration of genomic data, pathway analysis, and expression outcomes enhances our understanding of pancreatic cancer biology and facilitates the discovery of novel therapeutic targets.

Future research endeavors should focus on validating the identified biomarkers and elucidating their functional significance in pancreatic cancer biology. Additionally, prospective clinical studies are needed to evaluate the clinical utility of the identified biomarker in real-world settings.

## REFERENCES

1. Al-khassaweneh M, Bronakowski M, Al-Sharoa E. Multivariate and Dimensionality-Reduction-Based Machine Learning Techniques for Tumor Classification of RNA-Seq Data. *Applied Sciences*. 2023; 13(23):12801. <https://doi.org/10.3390/app132312801>.
2. Armanious, H., Adam, B., Meunier, D., Formenti, K., & Izevbaye, I. (2020). Digital gene expression analysis might aid in the diagnosis of thyroid cancer. *Current oncology (Toronto, Ont.)*, 27(2), e93–e99. <https://doi.org/10.3747/co.27.5533>
3. Bai, R., Rebelo, A., Kleeff, J., & Sunami, Y. (2021). Identification of prognostic lipid droplet-associated genes in pancreatic cancer patients via bioinformatics analysis. *Lipids in health and disease*, 20(1), 58. <https://doi.org/10.1186/s12944-021-01476-y>
4. Deshpande, D., Chhugani, K., Chang, Y., Karlsberg, A., Loeffler, C., Zhang, J., & Mangul, S. (2023). RNA-Seq data science: From raw data to effective interpretation. *Frontiers in Genetics*, 14, 997383.
5. Ge SX, Son EW, Yao R. iDEP: an integrated web application for differential expression and pathway analysis of RNA-Seq data. *BMC Bioinformatics*. 2018 Dec 19; 19(1):534. doi: 10.1186/s12859-018-2486-6. PMID: 30567491; PMCID: PMC6299935.
6. Jamail I and Moussa A (2020) Current State-of-the-Art of Clustering Methods for Gene Expression Data with RNA-Seq. *Pattern Recognition [Working Title]*. IntechOpen.
7. Klein A. P. (2021). Pancreatic cancer epidemiology: understanding the role of lifestyle and inherited risk factors. *Nature reviews. Gastroenterology & hepatology*, 18(7), 493–502. <https://doi.org/10.1038/s41575-021-00457-x>

8. Koch CM, Chiu SF, Akbarpour M, Bharat A, Ridge KM, Bartom ET, Winter DR. A Beginner's Guide to Analysis of RNA Sequencing Data. *Am J Respir Cell Mol Biol*. 2018 Aug; 59(2):145-157. doi: 10.1165/rcmb.2017-0430TR. PMID: 29624415; PMCID: PMC6096346.
  
9. Khozyainova, A. A., Valyaeva, A. A., Arbatsky, M. S., Isaev, S. V., Iamshchikov, P. S., Volchkov, E. V., Sabirov, M. S., Zainullina, V. R., Chechekhin, V. I., Vorobev, R. S., Menyailo, M. E., Tyurin-Kuzmin, P. A., & Denisov, E. V. (2023). Complex Analysis of Single-Cell RNA Sequencing Data. *Biochemistry. Biokhimiia*, 88(2), 231–252. <https://doi.org/10.1134/S0006297923020074>
  
10. Liao, C., Wang, X. *TCGApilot*: an R package for integrative pan-cancer analysis and visualization of TCGA multi-omics data. *BMC Bioinformatics* 24, 483 (2023). <https://doi.org/10.1186/s12859-023-05615-3>.
  
11. Manipur, I., Granata, I., Maddalena, L. et al. Clustering analysis of tumor metabolic networks. *BMC Bioinformatics* 21 (Suppl 10), 349 (2020).
  
12. McDermaid, A., Monier, B., Zhao, J., Liu, B. and Ma, Q., 2019. Interpretation of differential gene expression results of RNA-seq data: review and integration. *Briefings in bioinformatics*, 20(6), pp.2044-2054.
  
13. Paczkowska, M., Barenboim, J., Sintupisut, N. et al. Integrative pathway enrichment analysis of multivariate omics data. *Nat Commun* 11, 735 (2020).
  
14. Qin, C., Yang, G., Yang, J., Ren, B., Wang, H., Chen, G., Zhao, F., You, L., Wang, W., & Zhao, Y. (2020). Metabolism of pancreatic cancer: paving the way to better anticancer strategies. *Molecular cancer*, 19(1), 50. <https://doi.org/10.1186/s12943-020-01169-7>
  
15. Seyed Hossein Hassanpour, Mohammadamin Dehghani, Review of cancer from perspective of molecular, *Journal of Cancer Research and Practice*,

16. Siavoshi, M. Taghizadeh, E. Dookhe, M. Piran, Gene expression profiles and pathway enrichment analysis to identification of differentially expressed gene and signaling pathways in epithelial ovarian cancer based on high throughput RNA-Seq data, *Genomics*, Volume 114, Issue 1, 2022, Pages 161-170, ISSN 0888-7543, <https://doi.org/10.1016/j.ygeno.2021.11.031>.
17. Silva TC, Colaprico A, Olsen C, D'Angelo F, Bontempi G, Ceccarelli M, Noushmehr H. *TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages*. *F1000Res*. 2016 Jun 29;5:1542. doi: 10.12688/f1000research.8923.2. PMID: 28232861; PMCID: PMC5302158
18. Teichman, G., Cohen, D., Ganon, O. et al. RNAlysis: analyze your RNA sequencing data without writing a single line of code. *BMC Biol* 21, 74 (2023).
19. Vidman L, Källberg D, Rydén P (2019) Cluster analysis on high dimensional RNA-seq data with applications to cancer research - An evaluation study. *PLoS ONE* 14(12): e0219102.
20. Vincent, A., Herman, J., Schulick, R., Hruban, R. H., & Goggins, M. (2011). Pancreatic cancer. *Lancet (London, England)*, 378(9791), 607–620. [https://doi.org/10.1016/S0140-6736\(10\)62307-0](https://doi.org/10.1016/S0140-6736(10)62307-0)
21. Wang, T., Li, B., Nelson, C.E. and Nabavi, S., 2019. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC bioinformatics*, 20(1), pp.1-16.
22. Yu, W., Diao, Y., Zhang, Y., Shi, Y., Lv, X., Zhang, C., Zhang, K., Yao, W., Huang, D., & Zhang, J. (2022). Bioinformatic analysis of FOXN3 expression and prognostic value in pancreatic cancer. *Frontiers in oncology*, 12, 1008100. <https://doi.org/10.3389/fonc.2022.1008100>

23. Zhao, Z., & Liu, W. (2020). Pancreatic Cancer: A Review of Risk Factors, Diagnosis, and Treatment. *Technology in cancer research & treatment*, 19, 1533033820962117. <https://doi.org/10.1177/1533033820962117>
24. Zhu, Q., Sun, Y., Zhou, Q., He, Q., Qian, H. "Identification of key genes and pathways by bioinformatics analysis with TCGA RNA sequencing data in hepatocellular carcinoma". *Molecular and Clinical Oncology* 9, no. 6 (2018): 597-606. <https://doi.org/10.3892/mco.2018.1728>.