

ANALYZING RNA-SEQ DATA OF BRAIN CANCER FROM TCGA USING RNALYSIS SOFTWARE AND R STUDIO

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

By

BANDLA NAGA SIVA ANJALI (42739006)



**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)

Accredited with "A" grade by NAAC | 12B Status by UGC | Approved by AICTE

Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai – 600 119

www.sathyabama.ac.in

DEPARTMENT OF BIOINFORMATICS

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **BANDLA NAGA SIVA ANJALI (42739006)**, who carried out the final year project entitled "**ANALYZING RNA-SEQ DATA OF BRAIN CANCER FROM TCGA USING RNALYSIS SOFTWARE AND R STUDIO**" under my supervision from September 2023 to April 2024.

Dr. H. JEMMY CHRISTY

Internal Guide

Dr. H. JEMMY CHRISTY

External Guide

Dr. H. JEMMY CHRISTY

Head of the Department

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

Internal Examiner

External Examiner

DECLARATION

I **BANDLA NAGA SIVA ANJALI (42739006)** hereby declare that the Project Report entitled “**ANALYZING RNA-SEQ DATA OF BRAIN CANCER FROM TCGA USING RNALYSIS SOFTWARE AND R STUDIO**” 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 that reads "B. Naga Siva Anjali". The signature is written in a cursive style with a horizontal line under the last name "Anjali".

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENTS

I am pleased to acknowledge my sincere thanks to the **Board of Management of SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY** 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

Brain tumors are relatively rare compared to other types of cancers in which Glioblastoma is the most common primary malignant brain tumor and it occurs more often in older adults and more often in men. Glioblastoma is a grade 4 brain tumor. Through the study, RNA-Seq dataset of glioblastoma multiforme extracted from the Cancer Genome Atlas Program (TCGA) using package called TCGAbiolinks in R Studio. In this project, present RNAlysis software which is a python-based software for analysis of brain cancer dataset. Intersections between gene sets can be visualized and cluster analysis provides expression patterns of the dataset. After preprocessing, 22290 gene ids are obtained out of 60660 gene ids, that dataset is used for differential gene expression analysis between experimental conditions like healthy tissue versus brain cancer. Identified 893 differential expression genes DGEs and these are up regulated and the function of 894 (DGEs), is a critical step in interpreting the results of brain cancer. In this project, used an Enrichr package in RStudio for Enrichment analysis based on the concept of enriched functional GO term enrichment analysis, were applied for the identification of key genes and pathways involved in GBM. The biological process (BP) containing 2937 enriched terms, the top enriched terms like membrane depolarization during action potential, calcium ion import, regulation of ion transmembrane transport, negative regulation of cytosolic calcium ion concentration, microtubule -based movement and providing more biologically meaningful results related to brain cancer. It was understood that studying gene changes and gene expressions in RNA-Seq data that contribute to the GBM could help understand and identify potential therapeutic targets and predictive biomarkers studies for further analysis.

KEYWORDS: Brain Cancer, RNA-Seq, Exploratory Data Analysis, Clustering Analysis, Differential Gene Expression Analysis, Enrichment Analysis

TABLE OF CONTENTS

CHAPTER NO	TITLE	PAGE NO
	ABSTRACT	i
	LIST OF TABLES	iv
	LIST OF FIGURES	iv
	LIST OF ABBREVIATIONS	v
1	INTRODUCTION	1
	1.1 BRAIN CANCER	1
	1.2 GLIOBLASTOMA MULTIFORME(GBM)	2
	1.3 RNA-SEQUENCING (RNA-Seq)	3
2	REVIEW OF LITERATURE	4
3	AIM AND OBJECTIVE OF THE PROPOSED STUDY	7
	3.1 AIM	7
	3.2 OBJECTIVES	7
4	METHODOLOGY	8
	4.1 DATA SOURCE	8
	4.2 DATASET	8
	4.3 DATA RETRIEVAL	9
	4.4 ANALYSING RNA-SEQ USING RNALysis SOFTWARE	9
	4.5 DATA PREPROCESSING AND EXPLORATORY DATA ANALYSIS	9
	4.5.1 FILTER OUT LOWLY-EXPRESSED GENES	10
	4.5.2 EXAMINE VARIANCE IN OUR DATA WITH PRINCIPAL COMPONENT ANALYSIS	11
	4.5.3 PAIR PLOT	11
	4.5.4 BOX PLOT	12
	4.6 CLUSTERING ANALYSIS	13

	4.7 DIFFERENTIAL GENE EXPRESSION ANALYSIS	14
	4.8 DATA FILTERING AND VISUALIZATION WITH CUSTOMIZED PIPELINES	15
	4.9 VISUALIZING GENE SET INTERACTIONS	16
	4.10 ENRICHMENT ANALYSIS	16
	4.10.1 ENRICHR PACKAGE FOR ENRICHMENT ANALYSIS	17 18
	4.11 WORK FLOW OF RNALYSIS PIPELINE	
5	RESULTS AND DISCUSSION	20
	5.1 SOURCE CODE FOR DATA RETRIEVAL FROM TCGA	20
	5.1.1 MERGE THE TCGA DATASET	21
	5.2 BRAIN CANCER DATASET	22
	5.3 FILTERED OUT LOW EXPRESSED GENES	23
	5.4 PRINCIPAL COMPONENT ANALYSIS	25
	5.5 PAIR PLOT	27
	5.6 BOX PLOT	28
	5.7 CLUSTERING ANALYSIS	29
	5.8 DIFFERENTIAL GENE EXPRESSION ANALYSIS	31
	5.9 VOLCANO PLOT	33
	5.10 VENN DIAGRAM	34
	5.11 ENRICHMENT ANALYSIS	35
	5.11.1 GENE LIST FOR ENRICHMENT ANALYSIS	35
	5.11.2 SOURCE CODE FOR GO ENRICHMENT ANALYSIS USING ENRICHR PACKAGE	39
	5.12 GO ENRICHMENT ANALYSIS OF BIOLOGICAL PROCESS	44
6	SUMMARY AND CONCLUSIONS	53
	REFERENCES	55

LIST OF TABLES

TABLE NO	DESCRIPTION	PAGE NO
Table 5.1	Data set of GBM contains 60662 genes	22
Table 5.2	Filtered table contains 22290 genes	23
Table 5.3	Design matrix for DGE Analysis between healthy tissue vs. brain cancer	31
Table 5.4	Differential gene expression	32
Table 5.5	Gene list data for dge analysis	35
Table 5.6	Biological process enrichment terms	42

LIST OF FIGURES

FIGURE NO	DESCRIPTION	PAGE NO
Figure 4.1	Work flow of RNALysis pipeline	18
Figure 5.1	Principal Component Analysis of GBM	25
Figure 5.1.1	Variance in PCA plot	26
Figure 5.2	Pair Plot	27
Figure 5.3	Box Plot	28
Figure 5.4	CLICOM Clustering Analysis	29
Figure 5.4.1	PCA observed in Clustering Analysis	30
Figure 5.5	Volcano Plot	33
Figure 5.6	Venn diagram	34
Figure 5.7	List of most enriched Gene Ontology (GO) terms in biological process	44
Figure 5.7(a)	Histogram of GO enrichment analysis of DGEs	49
Figure 5.7(b)	Histogram of GO enrichment analysis of DGEs	51

LIST OF ABBREVIATIONS

TCGA	The Cancer Genome Atlas
GBM	Glioblastoma Multiforme
RNA-Seq	RNA Sequencing
EDA	Exploratory Data Analysis
PCA	Principal Component Analysis
FPKM	Fragments Per Kilo base of exon per Million reads
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
CLICOM	Cliques for Combining Multiple Clusterings
IQR	Inter Quartile Range

CHAPTER 1

INTRODUCTION

1.1 BRAIN CANCER

Cancer is a disease in which some of the body cells grow uncontrollably and spread to other parts of the body. It can start almost anywhere in the human body, which is made up of trillions of cells. These cells can form tumor or invade nearby tissues and organs, causing damage and impairing their function. Cancerous tumors may also be referred to as malignant tumors.

Cancer is primarily a genetic disease, caused by changes to genes that control the way our cells function, especially how they grow and divide. Cancer occurs by a series of successive mutations in genes so that these mutations change cell functions. For children, the highest percentage types of cancer disease are blood cancer, and cancers related to the brain and lymph nodes, respectively (M Dehghani *et al.*, 2017).

Brain tumors are dangerous and some types of brain tumors are cancerous, known as malignant brain tumors these are cancerous. They typically grow rapidly and invade surrounding healthy brain structures. Brain tumors are the most primitive, invasive and malignant in humans with poor survival after diagnosis (Roda E. *et al.*, 2022). One common type of brain tumor is gliomas, can be benign and most are malignant.

About 33% of all brain tumors are gliomas, accounting for about 80% of the total malignant central nervous system (CNS) tumors in adults (Hanif F. *et al.*, 2017). Glioma is a broad category of glial brain and spinal cord tumors which originate in the glial cells or look like glial cells that surround and support neurons in the brain, including astrocytes, oligodendrocytes, and ependymal cells. Types of gliomas and related brain tumors include astrocytoma, glioblastoma, oligodendroglioma and ependymoma (Eudocia Q *et al.*, 2018).

Among these, glioblastoma (GBM) is most common type of malignant brain tumor and aggressive primary brain tumors (Birzu, *et al.*, 2021). The tumor microenvironment (TME) plays a crucial role in mediating tumor progression and invasiveness, contributing to brain tumor aggression and poor prognosis (Yekula *et al.*, 2020).

1.2 GLIOBLASTOMA MULTIFORME (GBM)

Glioblastoma multiforme (GBM) is one of the most aggressive malignancies and also the most common malignant primary tumor of the brain and central nervous system, accounting for 14.5% of all central nervous system tumors and 48.6% of malignant central nervous system tumors which is a rapidly growing type of brain tumor (Grochans *et al.*, 2022). It accounts for more than 60% of all brain tumors in adults.

Glioblastoma (GBM), also referred to as a grade IV astrocytoma, is a fast-growing and aggressive brain tumor. It invades the nearby brain tissue, but generally does not spread to distant organs. GBMs belong to a category of tumors known as astrocytomas, originating from cells called astrocytes and is the higher-grade malignant glioma (grade IV) (Grochans *et al.*, 2022). Astrocytes are endogenous native cells of the brain, phenotypically similar to glioma cells and communicate with glioma cells to play a major role in promoting tumor growth and survival (Yekula *et al.*, 2020).

These cells play a vital role in supporting nerve cells and delivering nutrients to them. GBM is the most prevalent type of primary malignant brain tumor in adults. In adults, GBM occurs most often in the cerebral hemispheres especially in the frontal and temporal lobes of the brain, with approximately 95% of tumors arising in the supratentorial region. However, a small percentage of tumors may occur in the cerebellum, brainstem, or spinal cord (Davis *et al.*, 2016).

The only identified risk factors for GBM exposure to high dose ionizing radiation is the only confirmed risk factor and certain genetic syndromes (Hanif *et al.*, 2017). GBM is

an extremely aggressive form and devastating brain cancer that can result in often leading to death within six months if left untreated. Therefore, it is crucial to seek immediate expert neuro-oncological and neurosurgical care, as prompt treatment can significantly impact overall survival.

1.3 RNA-SEQUENCING (RNA-Seq)

In recent years, RNA Sequencing (RNA-Seq) has become a functional and amazing technology for profiling gene expression. RNA-Seq is generally used to compare gene expression between conditions, and find out which genes are up- or down-regulated in each condition. RNA-Seq is a powerful tool to investigation of the cancer pathogenesis and development of new drug agents to cancer treatment.

Differential Expression (DE) analysis as the most important applications of RNA-Seq data analysis can be used for downstream systems biology analysis such as gene ontology (GO) to provide insights into cellular processes altered between biological function and Enrichment of biological pathways. This functional information is crucial for identification of molecular targets for precision therapeutics. These analyses are routinely used in a variety of research areas including type of cancer and human disease research such as brain and drug development (A. Siavoshi *et al.*, 2022).

Computational analysis of RNA-Seq data has led to many scientific advances, including novel therapeutic discoveries, detailed understanding of genetic regulatory regions, and identification of biomarkers and pathogenic mutations (Deshpande, D. *et al.*, 2023).

CHAPTER 2

REVIEW OF LITERATURE

Teichman *et al.* (2023) developed RNAlysis, a modular Python-based analysis software for RNA sequencing data. This software facilitates exploratory data analysis, data visualization, clustering analysis, and gene set enrichment analysis using RNA sequencing data from various studies. RNAlysis offers a user-friendly graphical interface, eliminating the need for researchers to write code. It is applicable to data from any organism with a reference genome.

Grochans *et al.* (2022) provided an overview of glioblastoma multiforme (GBM) pathogenesis, including epidemiological data by age, gender, and tumor location. Despite medical advancements, GBM remains a highly aggressive primary brain tumor, more common in men. The review analyzed key molecular changes triggering GBM, as well as risk factors and protective factors associated with the disease.

Davis *et al.* (2016) discussed current treatment options for GBM, including surgical resection, radiation, and chemotherapy. They highlighted advances in understanding the molecular pathology of GBM, leading to novel therapies such as tumor-treating fields, molecular receptor-targeting drugs, and immunotherapy. The importance of palliative care and symptom management throughout the continuum of care was emphasized.

Birzu *et al.* (2020) reviewed standard treatments for newly diagnosed GBM, focusing on maximal safe resection followed by radiotherapy and adjuvant chemotherapy. Despite these treatments, nearly all GBMs relapse, prompting the exploration of molecular targeted therapies and precision medicine approaches. The review also discussed ongoing clinical trials analyzing new drugs and combination regimens for recurrent GBM.

Roda *et al.* (2022) proposed novel therapeutic strategies for brain tumors, emphasizing tumor heterogeneity, the tumor microenvironment, and innovative treatments such as virotherapy and immunotherapy. They highlighted the importance of personalized pharmacological interventions targeting neuro-inflammation, the tumor microenvironment, and blood-brain barrier leakage.

Yekula *et al.* (2020) provided a comprehensive review of extracellular vesicles (EVs) in GBM, highlighting their role in tumor proliferation, angiogenesis, immune evasion, and drug resistance. They emphasized the need for further research to understand the functional implications of EV subtypes and their potential as therapeutic targets in GBM.

Silva *et al.* (2016) discussed biotechnological advances in sequencing and the analysis of genomic and epigenomic data from large international consortia such as The Cancer Genome Atlas (TCGA). They described a workflow using Bioconductor packages to extract biologically meaningful data, illustrated by analyzing brain tumors.

Maxim *et al.* (2016) presented an update to the Enrichr tool for gene set enrichment analysis, which contains a vast collection of annotated gene sets from various sources. New features include the ability to submit fuzzy sets, upload BED files, and visualize results as cluster grams. Enrichr is praised for its comprehensiveness and ease of use but lacks some flexibility compared to other tools.

McDermaid *et al.* (2019) reviewed the analysis of differential gene expression (DGE) results from RNA-Seq data. They underscored the significance of DGE analysis in understanding gene expression alterations across varied conditions. The review highlighted the complexities in interpreting DGE results due to diverse formats and information overload. Authors introduced an R/Bioconductor package named "Visualization of Differential Gene Expression Results using R." This package aims to produce comprehensive visualizations for interpreting DGE outcomes generated by popular tools like, DESeq2, and edgeR.

Siavoshi *et al.* (2022) investigated the pathogenesis of advanced epithelial ovarian cancer (EOC) using R software and Linux. They identified 333 up-regulated and 273 down-regulated genes related to EOC development. Among these genes, RPL41, ALDH3A2, ERBB2, MIEN1, RBM25, ATF4, UPF2, DDIT3, HOXB8, and IL17D, as well as pathways like Ribosome and Glycolysis/Gluconeogenesis, were highlighted as potential targets for EOC diagnosis and treatment. Further experiments and gene network analyses are needed to confirm their functions. The study emphasizes the importance of analyzing individual ovarian cancer datasets due to histologic classification variations, aiming to identify progression stages of EOC.

CHAPTER 3

AIM AND OBJECTIVE OF THE PROPOSED STUDY

3.1 AIM

This study aims to analyze RNA-Seq data from brain cancer using RNALysis software for preprocessing and identifying differentially expressed genes. It will conduct clustering analysis, visualize gene expression profiles. Perform enrichment analysis with functional Gene Ontology enriched terms using the Enrichr package in R Studio.

3.2 OBJECTIVES

- Assessing the quality and integrity of RNA-Seq data pertaining to brain cancer.
- Identifying over- or under-represented classes of genes, biological pathways, and processes enriched based on gene expression patterns.
- 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.

CHAPTER 4

METHODOLOGY

4.1 DATA SOURCE

The RNA-Seq dataset for glioblastoma multiforme (GBM), 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 spanning approximately 30 different tumor types. This comprehensive database genomic sequencing data, clinical information, and other molecular characterization data obtained from thousands of cancer patients.

4.2 DATA SET

The dataset analyzed in this study focuses on Glioblastoma multiforme (GBM), a rapidly proliferating type of brain tumor extensively utilized in cancer research. This brain dataset consists of 60,660 gene identifiers, where rows and columns represent samples of Glioblastoma Multiforme (GBM). A total of 13 cases have been observed, encompassing both males and females.

The dataset includes parameters such as experimental strategy (RNA-Seq), data type (Gene expression quantification), data category (Transcriptome profiling). The sample type is recurrent tumor, and the primary site is the brain. These quantification metrics enable the analysis of differential gene expression, identification of biomarkers, exploration of biological pathways, and correlation of gene expression with clinical outcomes in brain cancer samples.

4.3 DATA RETRIEVAL

The TCGA brain cancer dataset, specifically Glioblastoma Multiforme (GBM), was obtained using RStudio with the assistance of the "TCGAbiolinks" package. This package facilitates direct retrieval of the dataset from the TCGA database by providing sample IDs and relevant parameters. 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. Utilizing TCGAbiolinks streamlines the data retrieval process and enhances the accessibility and usability of TCGA datasets for research purposes (Silva T. C *et al.*, 2016).

4.4 ANALYSING RNA-SEQ USING RNALysis SOFTWARE

RNALysis is a modular Python-based analysis software designed specifically for RNA sequencing data. Its main window provides a user-friendly interface for filtering, processing, analyzing, normalizing, and summarizing data tables, with a focus on RNA sequencing data. RNALysis offers the flexibility to build customized analysis pipelines and conduct exploratory data analysis, data visualization, clustering analysis, and gene-set enrichment analysis.

RNALysis enables researchers to analyze and visualize the data using tools and functions available within the software. By providing relevant parameters, researchers can tailor the analysis to their specific research objectives and gain insights into the molecular landscape of brain cancer. RNALysis facilitates efficient and comprehensive analysis of RNA-Seq data, empowering researchers to uncover novel findings and contribute to our understanding of brain cancer biology (Teichman G. *et al.*, 2023).

4.5 DATA PREPROCESSING AND EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is a fundamental approach in analyzing RNA-Seq datasets, particularly in cancer research, aimed at comprehensively understanding the

characteristics, patterns, and quality of the data. In RNA-Seq data analysis, EDA is crucial for gaining insights into gene expression profiles, identifying potential outliers or anomalies, and assessing data quality. Removing or correcting missing values, filtering out low-quality data. Creating plots such as histograms, boxplots and pair plots, etc., to explore gene expression distributions, identify patterns, and detect outliers.

Additionally, dimensionality reduction is a technique which is used to simplify complex high-dimensional data into a lower-dimensional representation while retaining as much information as possible and maintaining or improving the model's performance (Al-khassaweneh M *et al.*, 2023). Dimensionality Reduction Techniques like principal component analysis (PCA) can help visualize high-dimensional data and identify clusters or patterns.

4.5.1 FILTER OUT LOWLY-EXPRESSED GENES

Filtering out lowly expressed genes is a crucial preprocessing step in RNA-Seq data analysis, particularly in the context of TCGA-GBM data. This step helps to reduce noise and improve the accuracy of downstream analyses, such as differential gene expression analysis or clustering.

In RNA analysis, users can preprocess their data by normalizing it through various methods and eliminating rows with missing data from their tables. Additionally, we can filter out lowly expressed genes by setting a threshold parameter, ensuring that genes with expression levels below this threshold in the table are removed. One commonly used metric for this purpose is FPKM (Fragments Per Kilo base of transcript per Million mapped reads), which adjusts for factors like sequencing depth and gene length, making it suitable for comparing gene counts within a sample (Teichman G. *et al.*, 2023).

By applying this filtering step, genes with low expression levels across all samples are excluded, resulting in a more refined dataset for subsequent analyses. This enhances

the reliability and interpretability of downstream analyses performed on the TCGA-GBM RNA-Seq dataset.

4.5.2 EXAMINE VARIANCE IN OUR DATA WITH PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis (PCA) is a robust dimensionality reduction technique utilized in Exploratory Data Analysis (EDA). It is a statistical method that transforms high-dimensional data into a lower-dimensional space while preserving the most crucial patterns or variations in the data. PCA accomplishes this by identifying principal components, which are linear combinations of the original variables that capture the maximum variance in the data.

By converting the original high-dimensional gene expression data into a new set of variables known as principal components, PCA unveils the underlying structure of the data. The initial principal component captures the highest variance, followed by subsequent components capturing decreasing amounts of variance. These principal components, sorted by their contribution to the variance, reveal the axes of maximal variance (Yasser Perera *et al.*, 2021).

In the context of RNA-Seq data analysis of TCGA-GBM, PCA serves to visualize the variance among samples, identify potential clusters of samples with similar gene expression profiles, and explore the relationships between samples. This exploratory analysis yields valuable insights into the fundamental biology of GBM and facilitates the interpretation of subsequent analyses (Teichman G. *et al.*, 2023).

4.5.3 PAIR PLOT

Pair plots, also known as scatterplot matrices, are invaluable visualizations utilized in data analysis to concurrently explore relationships between multiple variables. In the context of RNA-Seq data analysis in brain cancer, pair plots serve as effective tools for

visualizing relationships among gene expression levels, clinical variables, and other relevant features.

Pair plots provide a concise visualization of high-dimensional data, enabling the identification of patterns and clusters within the dataset. By presenting scatterplots of each variable pair in the dataset matrix, pair plots empower researchers to assess correlations, detect outliers, and recognize potential associations between variables.

In RNA-Seq data analysis of brain cancer, pair plots play a crucial role in elucidating relationships between gene expression profiles and clinical characteristics, such as tumor subtype, patient outcome, or treatment response. They contribute to a deeper understanding of the complex interplay between genetic factors and clinical outcomes (Teichman G. *et al.*, 2023).

4.5.4 BOX PLOT

Exploratory data analysis utilizes statistical techniques to unveil patterns hidden within a dataset. One such technique is the "Box plot," which visually summarizes and compares groups of data. The box plot employs the median, quartiles, and extreme values to convey the distribution's level, spread, and symmetry. It aids in pattern identification, outlier detection, and distribution comparison across different sample groups or conditions (Liao C. *et al.*, 2023).

In a box plot, the central box represents the interquartile range (IQR), containing the middle 50% of the data, with the line within denoting the median. Whiskers extend from the box to depict variability beyond the upper and lower quartiles. The length of the whiskers is often determined by a specified range, such as 1.5 times the IQR. Data points outside the whiskers are considered outliers and may be plotted individually. The lower whisker boundary corresponds to the minimum dataset value, while the upper boundary corresponds to the maximum value (Teichman G. *et al.*, 2023).

4.6 CLUSTERING ANALYSIS

Clustering analysis is a fundamental technique utilized in bioinformatics and gene expression analysis to partition datasets into groups, or clusters, based on similarities among samples or features. RNAlysis offers support for various clustering methods, enabling researchers to group genes based on their expression patterns across samples. Additionally, the Gap Statistic method is employed to determine optimal options for the number of clusters in the analysis.

Methods:

K-means Clustering Method is used to identify patterns and group similar samples or genes together based on their expression profiles.

Parameters are like number of clusters determined by Gap Statistic method, Power transform = true. K-means is a partitioning clustering algorithm that assigns each sample to the nearest centroid. It aims to minimize the within-cluster variance and is suitable for datasets with predefined cluster numbers.

K- Medoids clustering Method also known as partitioning around medoids , is a variation of the K-means clustering algorithm that uses actual data points (medoids) as cluster centers instead of centroids.

Parameters are like number of clusters = 14, Metric = Spearman. K-medoids is a distance- based clustering algorithm similar to K-means but utilizes representative data points called "medoids" as cluster centers. It is robust to outliers and minimizes total dissimilarity within clusters.

Hierarchical clustering method defined as each data point is considered a separate cluster. Then, iteratively merge the closest pair of clusters based on the calculated distances to form a hierarchical tree or dendrogram.

Parameters are like number of clusters = 14, Metric = Euclidean, Linkage = Ward. Hierarchical clustering builds a tree-like hierarchy of clusters by recursively merging or splitting clusters based on similarity measures.

HDBSCAN identifies clusters based on the density of data points in the feature space. It groups together closely packed points as clusters and identifies points that are isolated as noise.

Parameters applied in this are minimum cluster size = 150, Metric = Jackknife, Method = Eom. HDBSCAN clusters data points based on density, identifying regions of high density as clusters while labeling outliers as noise. It automatically determines the number of clusters and is robust to varying densities and shapes of clusters.

CLICOM

CLICOM (Cliques for combining multiple clusterings) is an ensemble-based clustering algorithm that combines multiple clusterings to obtain a final clustering with better quality. It scales efficiently and produces output quickly (Teichman G. *et al.*, 2023). Parameters applied in Clicom are evidence threshold = 05, min_cluster_size = 75.

4.7 DIFFERENTIAL GENE EXPRESSION ANALYSIS

Differential gene expression (DGE) analysis is a computational technique employed in bioinformatics, especially in RNA-Seq data analysis, to pinpoint genes that exhibit significant expression differences across distinct experimental conditions or biological states. Following preprocessing, the gene identifiers in the GBM cancer dataset are subjected to DGE analysis using the DESeq2 function within the RNaLysis software. This method enables the identification of genes that play crucial roles in distinguishing between different experimental conditions or biological states, providing insights into the underlying molecular mechanisms of the studied phenomenon (Teichman G. *et al.*, 2023).

The DESeq2 function offers insights into up regulated and down regulated genes, a fundamental aspect in comprehending gene regulation and its implications in various biological processes, including development and disease. The output of differential gene expression analysis usually includes a list of DEGs ranked by their level of significance, along with statistical measures such as fold change, p-values, and

adjusted p-values (to account for multiple testing). These DEGs provide valuable insights into the molecular mechanisms underlying the studied biological phenomena and serve as candidates for further functional validation and exploration (A. Siavoshi *et al.*, 2022).

4.8 DATA FILTERING AND VISUALIZATION WITH CUSTOMIZED PIPELINES

The customized pipeline was developed within the RNALysis software to facilitate data filtering and visualization of differential expression analysis results. The pipeline included a series of functions, with a focus on visualizing results using volcano plots. Volcano plots are scatter plots serve as a comprehensive visualization tool, depicting significance versus fold-change on the y and x axes, respectively.

The Volcano plot displays each gene as a dot, with significantly up regulated genes ($\log_2$ fold change ≥ 1) colored in red, and significantly down regulated genes ($\log_2$ fold change ≤ -1) colored in blue. The x-axis represents $\log_2$ (Fold Change), while the y-axis represents the p-value.

Within the RNALysis software, the pipeline was configured to filter genes by statistical significance and split them by $\log_2$ fold-change direction using predefined parameters. This enabled the identification of significantly up- and down-regulated genes in the differential expression tables.

Upon applying the pipeline to the differential expression tables, the output consisted of two sets of volcano plots (positive and negative), along with new tables containing the significantly up- and down-regulated genes. These outputs provided valuable insights into gene expression changes associated with the studied conditions or biological states (Teichman G. *et al.*, 2023).

4.9 VISUALIZING GENE SET INTERACTIONS

Venn diagrams serve as powerful tools for visualizing the intersections between different sets of genes, particularly in the context of differential gene expression analysis. Each circle in the diagram represents a gene set, while the overlapping regions depict genes shared between the sets. In this gene sets represents DGE analysis between healthy tissues versus brain cancer and up and down regulated genes.

Analyzing the Venn diagram allows for the identification of genes common to multiple gene sets (intersection) and those unique to specific sets (non-intersection). This analysis offers valuable insights into the relationships between different biological processes, pathways, or experimental conditions represented by the gene sets. By examining the overlapping regions of the Venn diagram, we can identify genes shared between two or more conditions, indicating potential common regulatory mechanisms or biological pathways (Teichman G. *et al.*, 2023).

4.10 ENRICHMENT ANALYSIS

Enrichment analysis is a computational method used to identify biological pathways, molecular functions, or cellular components that are significantly overrepresented within a set of genes of interest. This analysis helps researchers gain insights into the underlying biological processes associated with their experimental data.

The process typically involves comparing the input gene list with a Gene Ontology (GO) terms or pathway databases. Statistical methods, such as the hypergeometric test or Fisher's exact test, are then applied to determine if certain categories are overrepresented in the input list compared to what would be expected by chance.

Enrichment analysis results in a list of enriched terms ranked by their statistical significance, often accompanied by additional metrics like p-values and adjusted p-

values. Enriched terms with low p-values are deemed statistically significant and biologically relevant. The top enriched terms offer valuable insights into the biological processes, functions, or pathways associated with the input gene list (Maxim V *et al.*, 2016).

4.10.1 ENRICHR PACKAGE FOR ENRICHMENT ANALYSIS

The Enrichr package serves as a valuable tool for conducting enrichment analysis directly within the R environment. Through interaction with the Enrichr web service or by accessing data from Bioconductor packages, we can leverage Enrichr for Gene Ontology (GO) enrichment analysis and functional annotation of genes. Enrichr allows users to input gene lists like DEGs and conduct enrichment analysis, revealing crucial biological insights about the genes and identifies overrepresented terms within the input gene list, understanding the functional significance of their gene sets.

Enrichr provides interactive visualizations like bar charts and heatmaps, facilitating the interpretation of enrichment results and enhancing the understanding of underlying biological mechanisms (Maxim V *et al.*, 2016).

- **Enrichr:** This function allows users to input gene lists and perform enrichment analysis. Users can specify the gene list, organism, and library to be used for enrichment analysis.
- **Enrich:** This function performs enrichment analysis using Enrichr and returns the enrichment results as a data frame. Users can specify parameters such as the gene list, library, and significance threshold.
- **Enrichr_barplot:** This function generates a bar plot to visualize the top enriched terms or pathways identified in the enrichment analysis.

4.11 WORK FLOW OF RNALYSIS PIPELINE

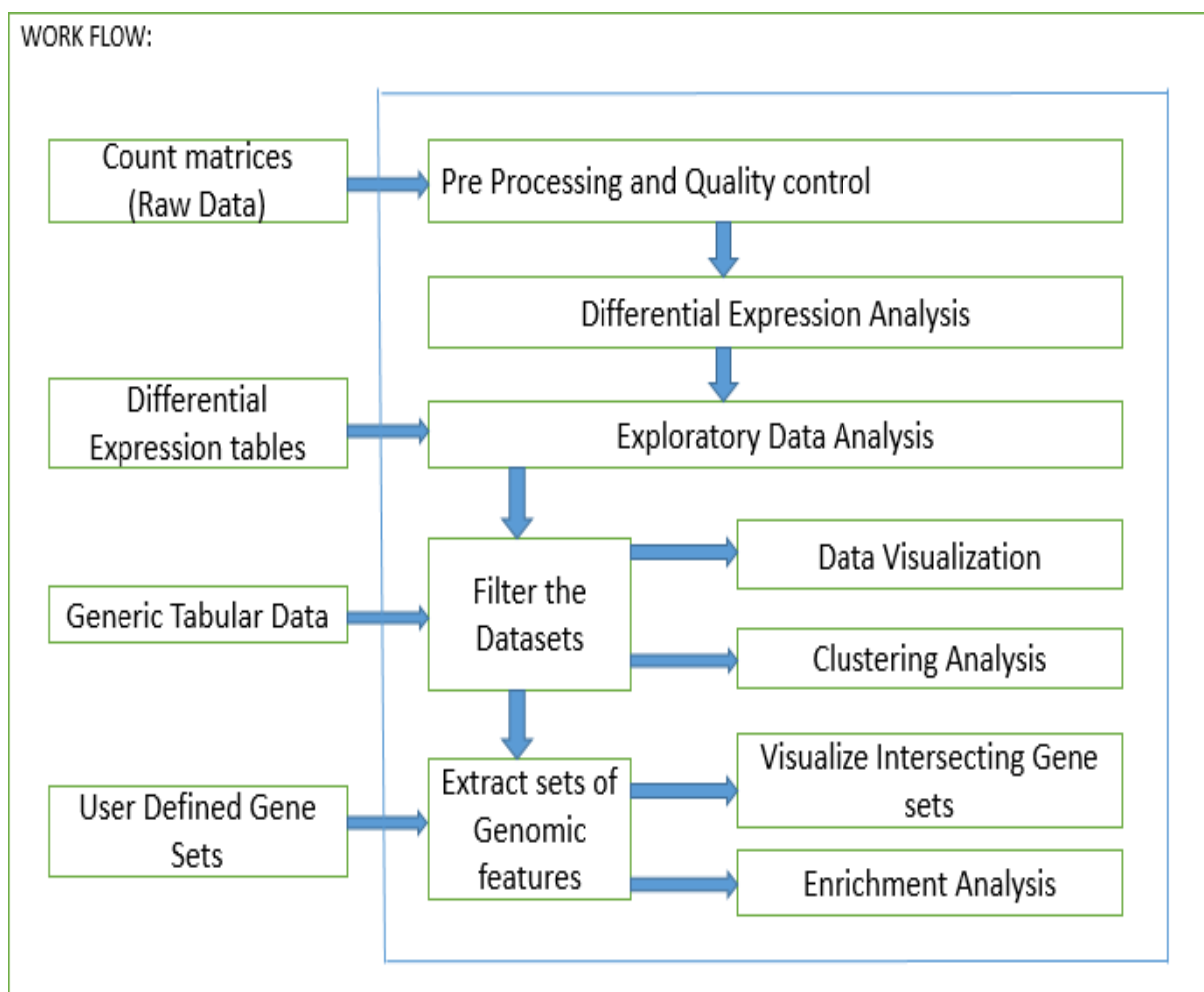


Figure 4.1: work flow of the RNALysis

Figure 4.1: The workflow of the RNALysis platform encompasses three main sections. In the top section, the analysis can commence at various stages, ranging from raw files to more processed data tables such as count matrices or differential expression tables. Users have the flexibility to input their data at any point in the analysis pipeline.

In the middle section, the data tables undergo a series of transformations, including filtering, normalization, and other customizable functions. These processes enable users to clean and refine their data, tailor their analysis to specific biological questions,

and prepare the data for downstream analysis. RNAlysis also provides users with a broad assortment of customizable clustering methods to help recognize genes with similar expression patterns, and visualization methods to aid in data exploration.

In the bottom section, users can focus on gene sets of interest by visualizing intersections between different gene sets, extracting gene lists resulting from set operations applied to their gene sets and data tables, and performing enrichment analysis (Teichman G. *et al.*, 2023).

CHAPTER 5

RESULTS AND DISCUSSION

5.1 SOURCE CODE FOR DATA RETRIEVAL FROM TCGA

```
1
2 if (!require("BiocManager", quietly = TRUE))
3   install.packages("BiocManager")
4 BiocManager::install(version = "3.18")
5
6 BiocManager::install("TCGAbiolinks")
7 library(TCGAbiolinks)
8
9 listSamples <- c(
10   "TCGA-19-0957-10B-01D-2366-09", "TCGA-14-1034-01A-01D-0518-04",
11   "TCGA-06-0152-02A-01R-2005-01", "TCGA-06-0125-01A-01T-0213-07",
12   "TCGA-14-0736-02A-01D-2280-08", "TCGA-06-0211-01A-01D-0242-05",
13   "TCGA-06-0210-01A-01R-0233-01", "TCGA-06-0190-01A-01D-0234-02",
14   "TCGA-19-4065-02A-11R-2005-01", "TCGA-06-0171-02A-11D-2366-09",
15   "TCGA-06-0221-02A-11R-2005-01", "TCGA-19-1389-10C-01D-0591-01",
16   "TCGA-14-1402-01A-01D-0521-05"
17 )
18 query <- GDCquery(
19   project = "TCGA-GBM",
20   data.category = "Transcriptome Profiling",
21   data.type = "Gene Expression Quantification",
22   experimental.strategy = "RNA-Seq",
23   barcode = listSamples
24 )
25 GDCdownload(query)
26
27
```

The provided code written in R for downloading RNA-Seq data related to glioblastoma multiforme (GBM) from The Cancer Genome Atlas (TCGA) using the TCGAbiolinks package. The script loads the TCGAbiolinks libraries into the R environment using the library function. The listSamples vector contains a list of TCGA sample identifiers (barcode) for GBM samples. These identifiers uniquely identify each sample in the TCGA dataset.

5.1.1 MERGE THE TCGA DATA

```
1
2
3 # merge the files into one tsv file
4 install.packages(c("readr", "dplyr"))
5 library(readr)
6 library(dplyr)
7
8 file1 <- read_tsv("file.path/GDCdata/TCGA-GBM/Transcriptome_Profiling/
9                   Gene_Expression_Quantification/89dad61a-6e26-4fd2-b277-1e438092af86/
10                   664dc12a-5d76-4efb-bb33-f0f58c937062.rna_seq.augmented_star_gene_counts.tsv")
11 file2 <- read_tsv("file.path/GDCdata/TCGA-GBM/Transcriptome_Profiling/
12                   Gene_Expression_Quantification/d8b840c7-e6cd-4cf9-824f-131fb18bf3b4/
13                   08dce278-ba8d-4ba0-90a1-e40820fd1740.rna_seq.augmented_star_gene_counts.tsv")
14 file3 <- read_tsv("filepath/GDCdata/TCGA-GBM/Transcriptome_Profiling/
15                   Gene_Expression_Quantification/ef068398-be98-409e-86aa-927f83bc5441/
16                   f5e230a7-b044-4190-8b61-36260d8cd54c.rna_seq.augmented_star_gene_counts.tsv")
17
18 merged_data <- inner_join(file1, file2, by = "ID") %>%
19   inner_join(file3, by = "ID")
20 merged_data <- bind_rows(file1, file2, file3)
21 write_tsv(merged_data, "merged_data.tsv")
22
23
24
```

R script merges three TSV files containing data of TCGA GBM into one combined TSV file. The script installs and loads the required packages: readr for reading TSV files and dplyr for data manipulation. The inner_join function from the dplyr package is used to merge the data frames file1, file2, and file3 based on the common column "ID". The merged_data data frame is written to a new TSV file named merged_data.tsv using the write_tsv function from the readr package.

5.2 BRAIN CANCER DATA SET

gene_id	gene_name	unstranded	stranded_first	stranded_second	tpm_unstranded	fpm_unstranded	fpm_unstranded
ENSG00000000003.15	TSPAN6	7285	3558	3731	131	36	36
ENSG00000000005.6	TNMD	8	2	6	0	0	0
ENSG000000000419.13	DPM1	3422	1707	1716	232	64	63
ENSG000000000457.14	SCYL3	1486	1179	1243	18	5	5
ENSG000000000460.17	C1orf112	1070	1050	1018	15	4	4
ENSG000000000938.13	FGR	1791	899	897	43	12	11
ENSG000000000971.16	CFH	7783	3887	3898	78	22	24
ENSG00000001036.14	FUCA2	6199	4547	4480	173	49	51
ENSG00000001084.13	GCLC	4104	2463	2483	40	10	11
ENSG00000001167.14	NFYA	3674	2051	2127	81	21	21
ENSG00000001460.18	STPG1	1139	636	666	11	3	3
ENSG00000001461.17	NIPAL3	8855	4573	4432	81	22	21
ENSG00000001497.18	LAS1L	8224	4107	4148	55	15	15
ENSG00000001561.7	ENPP4	2740	1364	1377	50	14	13
ENSG00000001617.12	SEMA3F	2564	1301	1278	42	12	11
ENSG00000001626.16	CFTR	271	143	142	2	0	0
ENSG00000001629.10	ANKIB1	16198	8319	7988	185	49	46
ENSG00000001630.17	CYP51A1	596	308	300	14	4	3
ENSG00000001631.16	KRIT1	1010	536	573	13	4	3
ENSG00000002016.18	RAD52	2005	1018	990	36	10	9

ENSG00000288675.1	AP0066216	194	188	7	2	2	0

Table 5.1: Data set of GBM contains 60660 genes

Table 5.1: The Glioblastoma Multiforme dataset, comprising 13 cases retrieved from The Cancer Genome Atlas (TCGA) database through the R package called TCGAbiolinks and applied parameters such as experimental strategy (RNA-Seq), data type (Gene expression Quantification), and data category (transcriptome profiling). The samples were specifically selected based on sample type (recurrent tumor), primary

site (brain), and project (TCGA-GBM). Upon retrieval, three open files were acquired from the 13 cases of both male and female, which were subsequently merged using the R package like readr, dplyr. The resulting dataset encompasses 60,660 gene ids facilitating comprehensive analysis of gene expression patterns within the context of Glioblastoma Multiforme.

5.3 FILTERED OUT LOW EXPRESSED GENES

gene_id	gene_name	Unstranded	stranded_first	stranded_second	tpm_unstranded	fpkm_unstranded	fpkm_unstranded
ENSG00000000003.15	TSPAN6	7285	3558	3731	131	36	36
ENSG000000000419.13	DPM1	3422	1707	1716	232	64	63
ENSG000000000457.14	SCYL3	1486	1179	1243	18	5	5
ENSG000000000460.17	C1orf112	1070	1050	1018	15	4	4
ENSG000000000938.13	FGR	1791	899	897	43	12	11
ENSG000000000971.16	CFH	7783	3887	3898	78	22	24
ENSG00000001036.14	FUCA2	6199	4547	4480	173	49	51
ENSG00000001084.13	GCLC	4104	2463	2483	40	10	11
ENSG00000001167.14	NFYA	3674	2051	2127	81	21	21
ENSG00000001460.18	STPG1	1139	636	666	11	3	3
ENSG00000001461.17	NIPAL3	8855	4573	4432	81	22	21
ENSG00000001497.18	LAS1L	8224	4107	4148	55	15	15
ENSG00000001561.7	ENPP4	2740	1364	1377	50	14	13
ENSG00000001617.12	SEMA3F	2564	1301	1278	42	12	11
ENSG00000001626.16	CFTR	271	143	142	2	0	0
ENSG00000001629.10	ANKIB1	16198	8319	7988	185	49	46
ENSG00000001630.17	CYP51A1	596	308	300	14	4	3
ENSG00000001631.16	KRIT1	1010	536	573	13	4	3
ENSG00000002016.18	RAD52	2005	1018	990	36	10	9
ENSG00000002330.14	BAD	1524	3035	3127	74	20	20

ENSG00000288675.1	AP006621.6	150	194	188	7	2	2

Table 5.2: Filtered table contains 22290 genes

Table 5.2: RNAlysis facilitates this preprocessing by offering various normalization methods and the option to eliminate rows with missing data or reduce noise in the dataset. For this analysis, the "Filter genes with low expression in all columns" parameter was selected to effectively remove genes with low expression levels across all samples. The parameter threshold was set to 50 FPKM (fragments per kilo base of transcript per million mapped reads). Consequently, genes with expression levels below this threshold in all columns of the dataset were filtered out.

Initially, the dataset contained 60,660 genes. However, after applying the filtering process, the dataset was refined to include 22,290 genes. This filtering step ensures that only genes with sufficiently high expression levels across all samples are retained, thereby improving the quality of the dataset for downstream analyses.

5.4 PRINCIPAL COMPONENT ANALYSIS

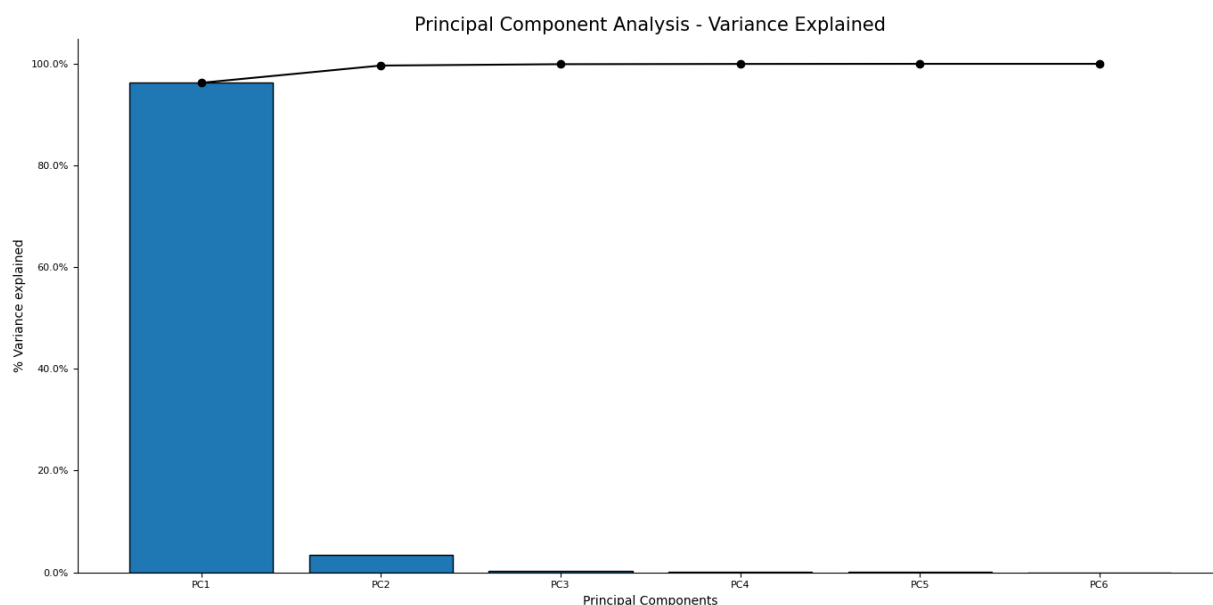


Figure 5.1: PCA Plot of GBM

Figure 5.1: In the GBM dataset, comprising 22,290 genes after applying a filtering function, a power transform is applied to the data before standardization and PCA. This preprocessing step helps mitigate undesirable characteristics of count data, such as mean-variance dependence and extreme values. During PCA, the first principal component (PC1) shows maximum variance i.e. 99% in the data and explains the most variation in the dataset, followed by subsequent components in decreasing order of variance.

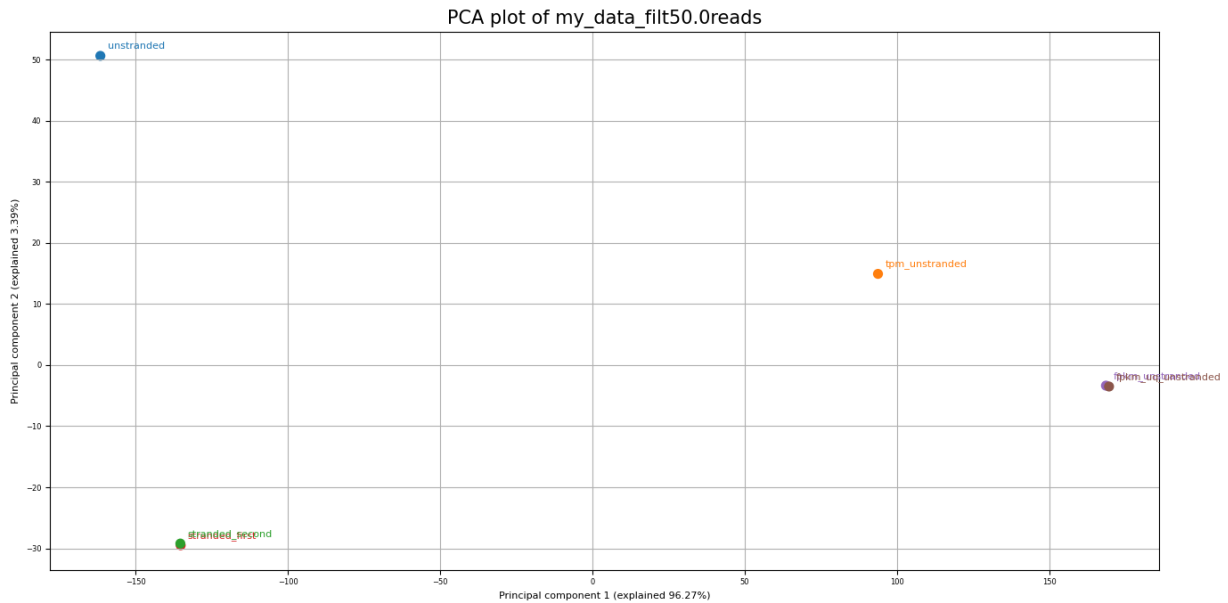


Figure 5.1.1: Variance in PCA plot

Figure 5.1.1: In the PCA analysis, the first two principal components account for over 99% of the data's variance, indicating a strong representation of the dataset's variability within these components. Notably, the first principal component segregates the samples based on their expression levels of "fpkm_unstranded" and "fpkm_uq_unstranded" on one end, while positioning "unstranded" on the opposite end, with "stranded_first" and "stranded_second" falling below.

On the other hand, the second principal component positions the samples featuring "fpkm_unstranded" and "fpkm_uq_unstranded" toward the center of the graph. This configuration offers valuable insights into the diverse expression patterns evident among the samples, thereby enabling enhanced interpretation of the dataset.

5.5 PAIR PLOT

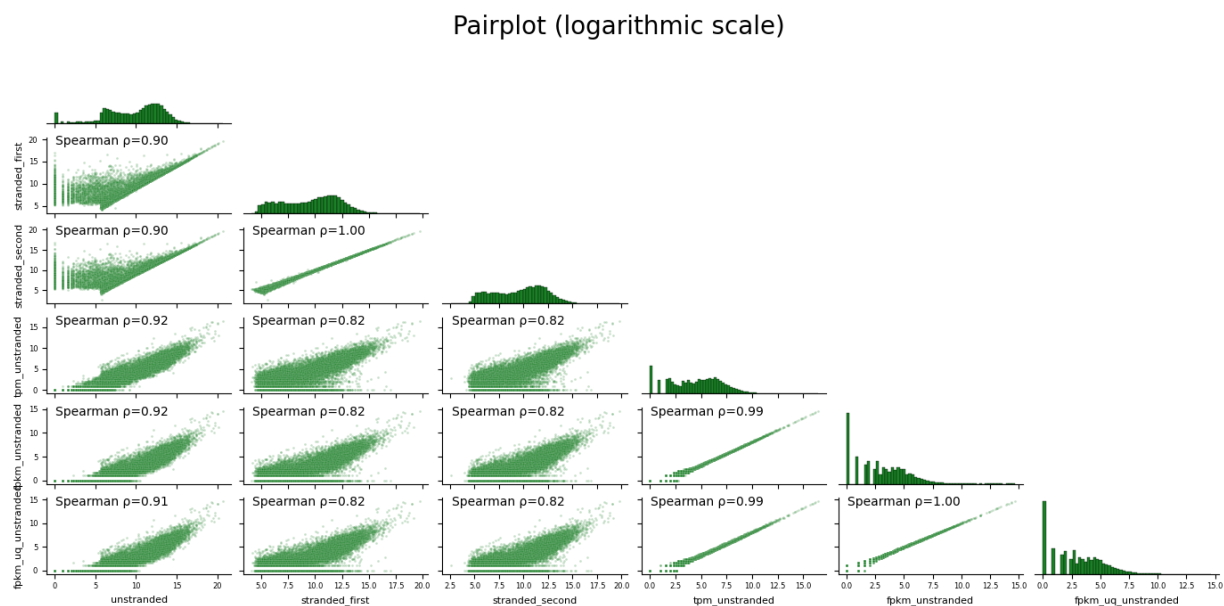


Figure 5.2: Pair plot

Figure 5.2: Pair plots provide a visual representation of the pairwise relationships between samples in our dataset, accompanied by histograms depicting gene expression values within each sample. Within the scatter plots, each dot represents the log of normalized expression for a single gene.

As a data visualization method in exploratory data analysis, pair plots offer valuable insights into patterns and correlations between gene expression levels across samples. After preprocessing, dataset comprising 22,290 gene IDs from the GBM cancer dataset, pair plots serve as a comprehensive tool for visualizing the gene expression landscape, aiding in the identification of potential associations and patterns within the data.

5.6 BOX PLOT

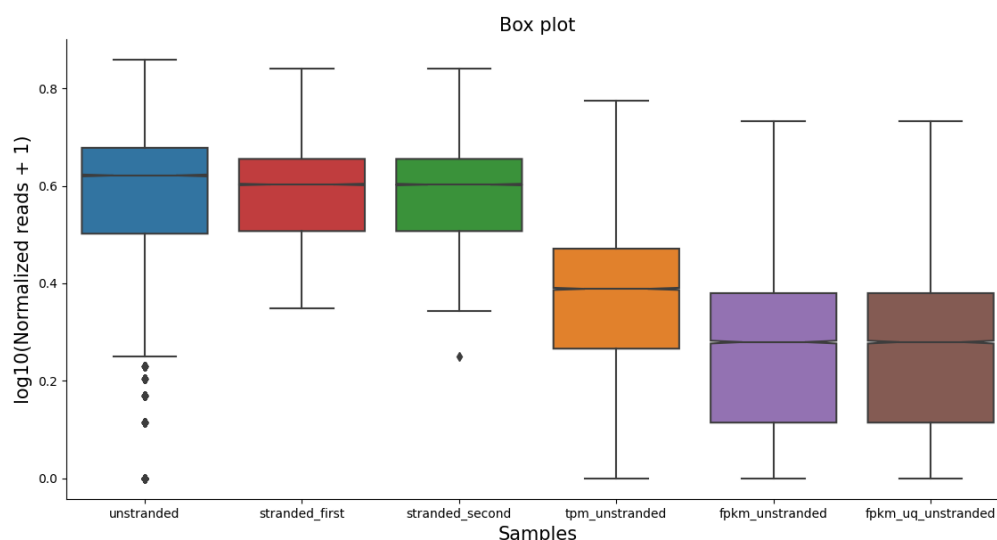


Figure 5.3: Box plot

Figure 5.3: The box plot visually illustrates the distribution of $\log_{10}$ normalized reads + 1 scores along the y-axis and sample names along the x-axis within the cancer dataset, which encompasses 22,290 gene IDs. Serving as a data visualization tool in exploratory data analysis, this plot offers insights into the variability and distribution of gene expression levels across the dataset.

The central box of the plot represents the middle 50% of the data, with a horizontal line indicating the median. Whisker lines extend from the box to the highest and lowest data points. Any data points lying beyond the whiskers are identified as outliers and are depicted as rhombic points, providing visibility into potential anomalies in gene expression levels.

5.7 CLUSTERING ANALYSIS

Results of Emsemble Clustering for 5 clustering solutions,
evidence_threshold=0.5, and min_cluster_size=75

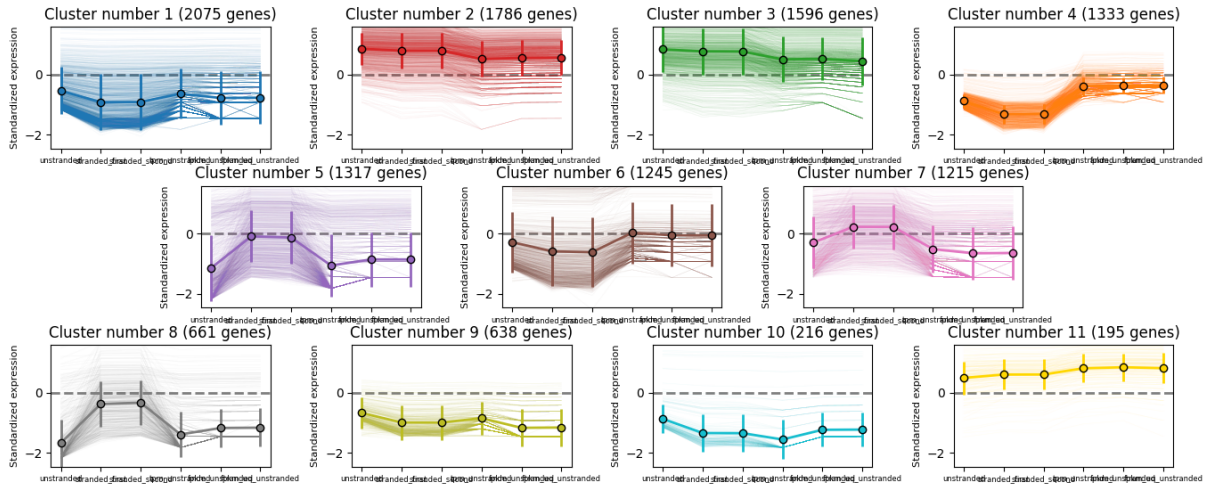


Figure 5.4: CLICOM clustering analysis

Figure 5.4: In RNA sequencing clustering analysis, various algorithms are employed to group genes based on their expression patterns. These include distance-based (K-Medoids), density-based (HDBSCAN), hierarchical (Hierarchical), and ensemble-based (CLICOM) algorithms, which combines five multiple clustering setups with an evidence threshold of 0.50 and a minimal cluster size of 75.

The input dataset for the clustering analysis consists of 22,290 genes after applying a filtering function. Using RNALysis and selecting the CLICOM tool, the analysis generates gene clusters. Each cluster's expression profile is visually depicted, with center lines representing the cluster means and standard deviations across samples. This comprehensive clustering approach enables the exploration of gene expression patterns and facilitates the identification of functionally related gene groups within the dataset.

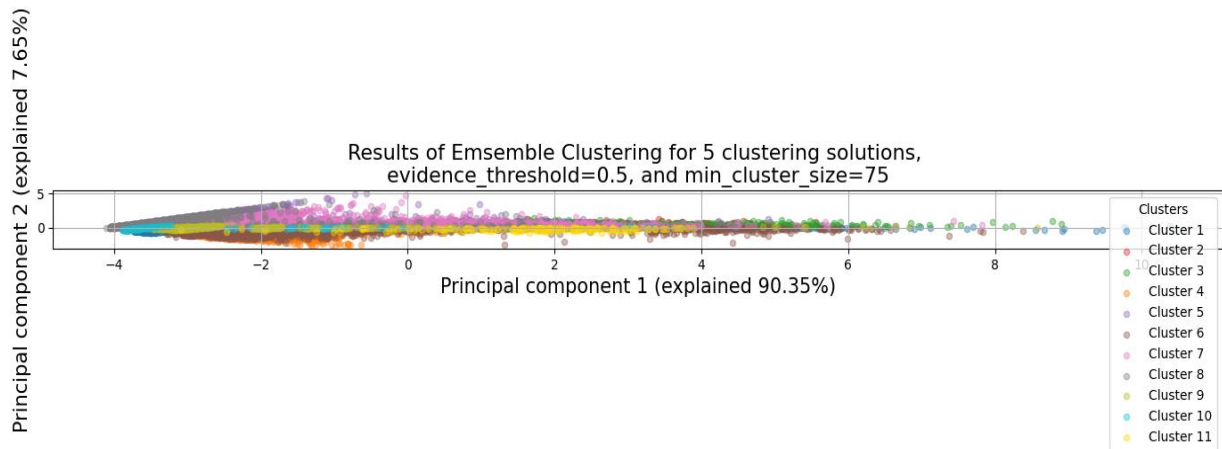


Figure 5.4.1: PCA observed in clustering analysis

Figure 5.4.1: The Principal Component Analysis (PCA) projection visualizes the power-transformed and standardized gene expression data, with each dot representing an individual gene. To enhance interpretability, the points are color-coded based on the cluster to which they belong, as determined by the CLICOM clustering algorithm. This visualization enables the examination of gene expression patterns and their grouping into distinct clusters.

The PCA projection captures 98% of the variance, as depicted in Figure 5.4.1 of the CLICOM clustering projection. This visualization offers valuable insights into the relationships between genes and their clustering behavior, facilitating the interpretation of gene expression data and the identification of biologically relevant gene groups.

5.8 DIFFERENTIAL GENE EXPRESSION ANALYSIS

Samples	Condition	Batch
unstranded	Healthy Tissue	A
stranded_first	Healthy Tissue	B
stranded_second	Healthy Tissue	C
tpm_unstranded	Brain Cancer	A
fpkm_unstranded	Brain Cancer	B
fpkm_uq_unstranded	Brain Cancer	C

Table 5.3: Design Matrix for dge analysis between healthy tissue vs.
brain cancer

Table 5.3: The design matrix serves as a crucial component in differential gene expression analysis, delineating the various sample conditions and experimental factors under investigation. The matrix distinguishes between different sample types within each condition for Cancer such as "Healthy Tissue" and "Brain Cancer" are defined, highlighting the distinct experimental groups being compared.

Additionally, batch information is integrated into the design matrix through labels such as "A," "B," and "C." These labels signify different experimental batches or sequencing runs, and accounting for batch effects is imperative in downstream analysis to ensure the robustness and reliability of the results.

For conducting differential gene expression analysis, it's essential to load the design matrix and specify the R installation path within the DESeq tool in the RNaLysis software. Loading the design matrix allows for the incorporation of experimental design factors into the analysis, enabling the comparison of gene expression levels between different conditions or groups.

gene_id	gene_name	Base Mean	log2Fold Change	lfcSE	stat	pvalue	padj
ENSG00000269968.1	AC006064.4	5359.73	11.0929	3.9067	2.839	0.0045	0.010639
ENSG00000258017.2	AC011603.2	4278.17	10.630	2.8953	3.6717	0.00024	0.0007
ENSG00000234961.1	AL133415.1	3850.11	10.616	3.9067	2.7175	0.00657	0.015005
ENSG00000273149.1	AL138963.3	1821.79	9.5369	3.9067	2.441	0.01464	0.030781
ENSG00000265401.1	AC093484.3	1703.55	9.4405	3.9067	2.416	0.01567	0.032653
ENSG00000266844.1	AC093330.1	1232.31	8.9802	3.8765	2.3165	0.02052	0.041478
ENSG00000203865.11	ATP1A1-AS1	1333.99	8.9076	2.653	3.3573	0.00078	0.002127
ENSG00000210151.2	MT-TS1	911.522	8.537	3.9067	2.1852	0.02887	0.056097
ENSG00000227766.1	AL671277.1	822.911	8.3904	3.9067	2.1476	0.03174	0.06088
ENSG00000262831.1	AC145207.3	761.005	8.2780	3.9067	2.1188	0.03409	0.06487
ENSG00000225339.4	AL354740.1	706.692	8.2055	2.9075	2.8221	0.00477	0.011174
ENSG00000176868.2	AL358781.1	700.529	8.1586	3.9067	2.0883	0.03676	0.069327
ENSG00000288597.1	AC234782.2	780.604	8.0625	2.3887	3.3751	0.00073	0.002002
ENSG00000279569.1	AC020763.4	718.282	8.0360	2.7641	2.9072	0.00364	0.008749
ENSG00000271581.1	AL671883.2	706.118	8.0219	2.8540	2.8106	0.00494	0.01154
ENSG00000238045.9	AC009133.1	659.924	7.9012	2.7078	2.9178	0.00352	0.00849
ENSG00000262413.1	AC145207.2	573.124	7.8685	3.9068	2.0140	0.04400	0.081047
ENSG00000267469.1	AC005944.1	553.350	7.8179	3.9068	2.0011	0.04538	0.083173
ENSG00000272325.2	NUDT3	620.671	7.7746	2.5466	3.0529	0.00226	0.005652
ENSG00000269292.1	AC093503.1	485.982	7.6310	3.9068	1.9532	0.05079	0.09173

ENSG00000112096.19	SOD2	6128.17	2.0008	0.0776	25.753	2.98E-146	3.77E-144

Table 5.4: Differential gene expression analysis

Table 5.4: The dataset, consisting of 22,290 gene IDs, undergoes differential gene expression analysis using the RNAlysis software. The software offers default tools for analysis, including the "Run DESeq2 differential expression" function.

To perform the analysis, follow these steps: Select the "Run DESeq2 differential expression" function. Provide the design matrix containing sample conditions (e.g., healthy tissue vs. brain cancer). Specify the R installation folder for proper execution. Initiate DESeq2 to commence the analysis.

DESeq2 will identify differentially expressed genes, including both up and down regulated genes, accompanied by p-values and log2FoldChange values. These results provide insights into gene expression alterations associated with the transition from healthy tissue to brain cancer.

5.9 VOLCANO PLOT

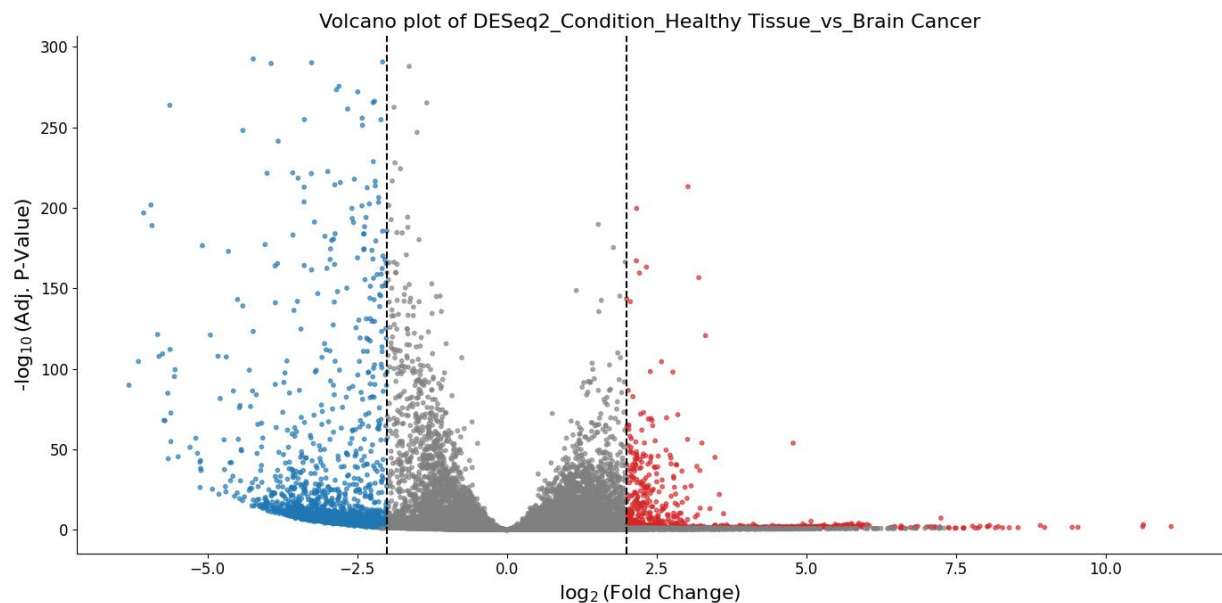


Figure 5.5: Volcano plot of GBM

Figure 5.5: The differential expression analysis comparing healthy tissue versus brain cancer was visualized using Volcano plots, which were generated and extracted from each respective differential expression table. This process was automated through the creation and application of a Pipeline within RNALysis. The Pipeline included filtering parameters such as "Filter by statistical significance" and "Split by log2 fold-change direction", which were saved and applied for subsequent analysis.

Upon analysis, it was found that out of the 22,290 genes examined, 6,202 genes were up regulated, while 16,088 genes were down regulated. Notably, the maximum log2 fold change observed for differentially expressed genes was 11.09, associated with the gene ID ENSG00000269968.1 and gene name AC006064.4, while the minimum log2

fold change was -6.30, linked to the gene ID ENSG00000265917.1 and gene name MIR3685. These findings provide insights into the differential expression patterns between healthy tissue and brain cancer samples.

5.10 VENN DIAGRAM

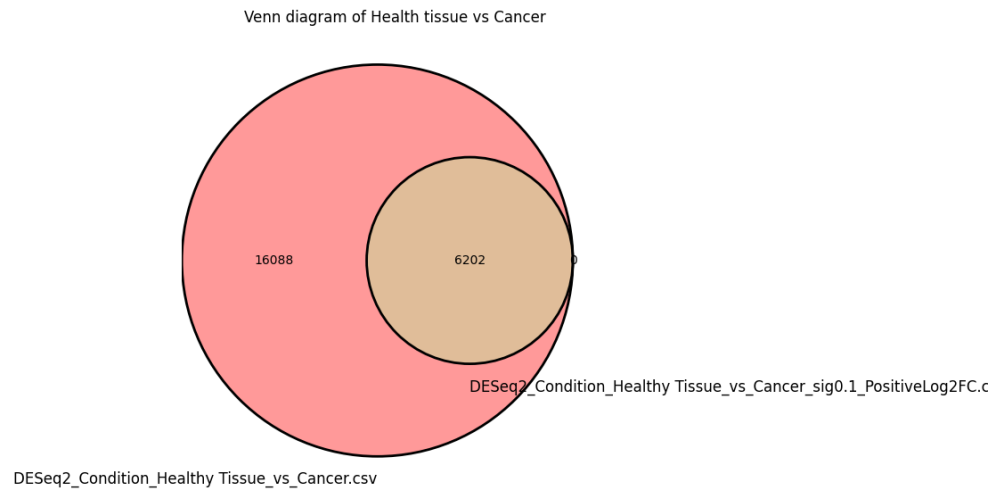


Figure 5.6: Venn diagram shows intersections between gene sets

Figure 5.6: A Venn diagram was generated to visually represent the intersections between the gene sets derived from the original dataset, the significantly up regulated genes, and the significantly down regulated genes. It illustrates the overlap between these gene sets, providing insights into the shared and unique genes among them. Each circle in the diagram represents a gene set, and the overlapping regions indicate genes that are shared between the sets.

Analyzed 22,290 gene ids in differential gene expression analysis between the conditions and interpreted 16,088 genes as down regulated. 6,202 genes were identified as up regulated.

5.11 ENRICHMENT ANALYSIS

5.11.1 GENE LIST FOR ENRICHMENT ANALYSIS

AC006064.4	STX18-AS1	AC008771.1	RUSC1-AS1	AC010320.2	PAX6
AC011603.2	DARS-AS1	AC023509.2	COG8	AC005726.2	CIITA
AL133415.1	DNAH10	LAMTOR5-AS1	ELP1	CAPN12	MIR100HG
AL138963.3	KCNMA1	KDM4A-AS1	USP34	MIF-AS1	SCFD1
AC093484.3	ACSM3	AC084125.2	ATM	SLC12A5-AS1	ELF3-AS1
AC093330.1	TTTY14	USP50	NDUFB9	AC006064.2	AL157392.3
ATP1A1-AS1	AC023355.1	AL356804.1	AC010642.2	SRI-AS1	PCBP1
MT-TS1	AGAP2-AS1	GSN-AS1	HNRNPU	AL354813.1	CIT
AL671277.1	B3GNT4	NAPA-AS1	H1-10-AS1	NDUFB2-AS1	FAM27C
AC145207.3	GRM4	AC109587.1	PCLO	AC044849.1	TCF4
AL354740.1	MMADHC-DT	GAS6-AS1	KALRN	AC011511.5	CCDC171
AL358781.1	BHLHE40-AS1	AL450998.2	CFLAR	AC083798.2	GPSM2
AC234782.2	DYNC1H1	AL162258.2	COL19A1	AC012459.1	AC025165.1
AC020763.4	EIF3E	EAFF1-AS1	KAT6A	PCBP1-AS1	AFAP1
AL671883.2	ELP4	CARS1-AS1	KLF8	PRAL	MMACHC
AC009133.1	ADPGK-AS1	AC135050.1	MBP	AC093849.4	LRRC75A
AC145207.2	LDLRAD4	AL161668.3	BMP2K	AL356481.3	TUBGCP4
AC005944.1	NRG1	HELLPAR	ST3GAL3	DLGAP4-AS1	TPTEP1
NUDT3	LACTB2-AS1	AC025164.1	EHMT1	H2AZ1-DT	SCN9A
AC093503.1	SPEF2	RCCD1-AS1	KLC1	CD44-AS1	CCNK
THY1-AS1	PRKG1	AL121581.3	ARIH1	MGAT2	PLCG1-AS1
MIR1282	MEG8	IFNAR2	NEAT1	C6orf201	B3GALT5
AL137857.1	AL022311.1	CCDC152	PPP3CB-AS1	AC125494.1	EPB41L4A
AC092069.1	HYPK	AC008676.3	KANSL1	AC211433.1	NKAIN3
AC131944.1	MTMR3	DPP9-AS1	SLC25A26	AC005329.1	GIT2
ZACN	SMAD2	MORC2-AS1	AC006504.5	ZNRD1ASP	STXBP5-AS1
AC003681.1	AP4B1-AS1	AL031846.2	PCNX4	AC005306.1	CNNM2
ILK	AC010198.2	PDXP-DT	RBMS3	DNAH11	LCOR
AC015909.1	LRRC7	PITRM1-AS1	MYLK	MIF	AC087289.5
JSRP1	LINC00299	LINC01572	GDAP1	CXCR5	MYO15A
ATP6V1C2	MIR29B2CHG	SNRK-AS1	C2orf16	KCNQ1OT1	NF1
CADM3-AS1	MYT1L	AC074212.1	PTCD1	AP003068.1	STK24
GPS2	LAMA1	CLN5	BCL11A	AC015971.1	CSPP1
AP002490.1	SLC19A1	CFAP73	NAV1	TMEM51-AS1	PCSK6
AC097382.2	SPG7	CACTIN-AS1	SLC8A1	AC107081.1	CDH23
AFAP1-AS1	TMEM161B-AS1	ARHGEF38	PLEKHA1	SLC31A2	INTS6
PPT2-EGFL8	AL035530.2	AL669831.3	TANC2	AC004882.3	PAX8-AS1

CLMAT3	RNF157-AS1	TLDC2	C12orf60	HCG25	DPP4
LDLRAD2	NRDE2	AC004893.2	SCN2A	AL589935.1	PDE4DIP
AC022400.7	AC004877.1	AC011484.1	HIC1	CCDC73	DEXI
PCOLCE-AS1	ATG9B	AC007750.1	GABRG2	AC018362.1	ANKRD34C-AS1
AP001267.1	AL033523.1	AC114271.1	CFAP46	LINC01504	AL589740.1
AL662797.1	GARS1-DT	AC244517.5	CD302	SNHG14	ZFH3
AC010883.3	SLC46A1	SEPTIN4-AS1	WDFY2	AC025043.1	LPP
AC018761.2	CASC2	AC024909.1	CACNA1D	AL158151.1	NTRK3
AP003352.1	NBEAL1	ATG10	USP8	TOP3B	LINC00511
AL020996.1	CCDC88A	AL133227.1	IGF2R	ACTA2-AS1	ANK3
VASH1-AS1	CASC15	ELOA-AS1	GNAI1	AC025682.1	BASP1-AS1
VIM-AS1	ZBTB20	TTN-AS1	KCNC4	AL109628.2	ABCB9
AC006115.2	CCDC85C	AC019257.9	AC107959.1	AP000781.1	TRABD2A
AC009630.1	CSMD3	GRM3-AS1	ARHGEF28	TMEM262	AC012063.1
AC067930.5	SHPRH	DNAJC9-AS1	HDAC8	EML5	REL
COX20	NEDD4L	AC020913.2	MSI2	AC009119.1	B9D1
AL354733.3	CACNB4	FLJ40194	HSPG2	AC134772.1	ALG11
AC120057.3	S100A1	VIL1	ATRX	AC002401.1	PCDHA10
SRRM5	OPRL1	AC063952.1	HECW2	AC090197.1	PEAK1
CFAP92	CHD8	ENTPD1-AS1	MED13L	XIST	MCF2L
HOXA-AS3	VPS53	PXN-AS1	FILIP1	AC019197.1	C21orf62-AS1
AC087164.1	NUTM2A-AS1	SLC5A2	MAGI2	AC007342.6	FAN1
AC010761.1	MMS22L	AC009690.2	MCF2L2	AC087276.2	MICAL3
MFF-DT	NFASC	AC245884.1	P2RX5-TAX1BP3	AC005606.2	GBF1
MFF-DT	TSC2	AC068580.3	ZBED3-AS1	MEIOC	SPTB
MGAT4C	PHF3	JPX	NBPF20	AP000757.1	NCKAP1
ANGPTL6	PPP1R12B	AL031714.1	SZT2	A2M-AS1	BRWD1
RDM1P5	CSMD1	PRMT5-AS1	ADGRV1	SPDYE3	RYR3
AC010226.1	VPS13D	BEST1	AC027290.2	TMEM184A	IRF1-AS1
FLNB-AS1	CD226	AC106886.4	FLRT2	CPB2-AS1	DNAJC10
RAPH1	SLC7A6	AC008522.1	ZC2HC1C	EIF1B-AS1	LINC00847
AC123768.4	AC005670.3	SLC3A1	HMCN2	AC138028.4	TTC7B
NUTM2B-AS1	ATP7B	AC002116.2	SYNE2	AL358334.4	SSPOP
RNF217-AS1	MAPT-AS1	CCDC144NL-AS1	FEM1A	AC011405.1	AKAP13
AC107027.1	PBX1	AL138689.2	LINC00662	PTRH1	XKR4
ITGA9-AS1	ACSS3	AC008894.2	TRIM59	DAAM2-AS1	KCNH1
GPR39	CNKS3	TREH	ITPR1	TCTE1	SLC1A2
KCNMB3	CYTOR	AC012640.4	FOXP1	AL035106.1	MARK3
SLC18A2	CARMN	AC073896.2	KIF5C	AC073254.1	SH3D19
NEB	CASK	AC007938.2	TAF10	LINC00334	NCAPD3
AC023024.1	PAQR3	C5orf66	TMEM30A-DT	EPHA1	MYH11

AL139353.2	MACROH2A1	SPATA21	AC020916.1	AL606760.2	CLK3
PCSK7	BAZ2B	AMN	INTS6-AS1	UXT-AS1	AGO3
AC008840.1	G3BP1	AC008264.2	KAT6B	AP001610.2	LRRRC8C-DT
CBR3-AS1	ATG7	AC024560.5	PCDH9	SERHL	FRAS1
C10orf55	ZNF26	SPSB3	KMT2A	SH3TC2	GVQW3
HULC	ANK2	AC015802.6	IQCK	RHOF	LCORL
NDP-AS1	SLITRK5	PRC1-AS1	C1RL-AS1	HSD17B1	PHF20L1
MCM3AP-AS1	SELENOP	PPM1F-AS1	RPS6KA5	AC079601.1	SILC1
ADD3-AS1	CMC1	AP005329.1	SPATA6L	AL355388.2	SGCE
SPAG17	SLC7A6OS	AL592166.1	SOX2-OT	KMO	CPLANE1
AC024588.1	PIGT	NUCB1-AS1	CCNI2	AL512770.1	RAB11FIP1
AP001972.3	PIGN	COL4A2-AS1	LINC01748	AC091045.1	LUCAT1
TBCE	RTTN	AP001107.1	LINC00632	PTPRG-AS1	LRP8
BX322234.1	VPS13C	AL159978.1	FIG4	SLC5A10	HSPA12A
PVT1	SH3BP5-AS1	AC010973.2	ATR	MCMD2C2	ITSN1
AC005520.5	TMED3	AL450306.1	MIPOL1	AP001001.1	TRRAP
AL080317.2	SBF2	ABALON	DNASE1	HYDIN	AL161772.1
MYLK-AS1	TNRC6B	CERS6-AS1	AC092279.1	AC079921.1	MICAL2
IGF1	PSMG4	SPAG5-AS1	RNF213	PKP4-AS1	MYO9A
AP002748.6	LIMD1	AC139530.1	GRIN2B	CTC-338M12.4	ASPH
FARP1	PRPF40B	AP000866.5	MBNL1-AS1	PCA3	CACNA1C
CCDC39	PTPRO	AC023593.1	CSNK2A1	HDHD2	FTO
RGL4	POLR3H	TTN	SNX22	AC068733.3	LRP1
IGSF10	UBE3A	AC009060.1	CNKSR2	SLC26A10	LARS1
JAKMIP2-AS1	USP15	SCAND2P	PSMA2	AC110611.1	NUMA1
AC020931.1	PIK3CA	U47924.2	XPC-AS1	MUC12	COL4A2
SLC8A1-AS1	BACE1-AS	FAM166A	AC129492.6	Z92544.1	AP001062.1
AD000864.1	FKTN	EHMT2-AS1	C12orf73	AF131216.1	HECTD4
NREP	KCNJ6	AC036108.2	ALG1L9P	PKD2L2	GPATCH2L
ABCC11	OBI1-AS1	AC080038.3	SRGAP1	WDR38	N4BP2L2
STARD4-AS1	KIAA1109	AC245052.4	CFAP251	AC027601.1	NCAM1-AS1
FAM198B-AS1	PEX26	AP001029.2	WDR73	AP003108.2	KMT2D
KCP	SNED1	AC009107.2	OPRM1	AL356124.1	PKD1
ZNF337-AS1	LINC01184	CTBP1-AS	AHNAK	AL590627.1	TOR1AIP2
NSUN5P2	ABI2	NFIA	ABRAXAS1	MACF1	HEXA
AL160270.1	SMG1	GEMIN7-AS1	NSD2	ARL2BP	CPT2
LIX1L-AS1	LINC01234	GMDS-DT	FRMD4A	AL117329.1	HMCN1
AC012358.2	NBPF10	AL133230.1	CENPP	CCDC30	LRRK1
GTF2IRD1P1	PDE10A	AL139423.1	MUC20-OT1	AL137139.2	ADGRD1
EGFR-AS1	ENOSF1	PMFBP1	HEMK1	MIR99AHG	ZNF891
TSPOAP1-AS1	CTSC	PINK1-AS	EML6	DCHS2	ADGRD1

CEP164P1	AC046134.2	SAP30L-AS1	HERC2	AC021092.1
AL122035.2	LAMA3	MKNK1-AS1	WDR83	AC103740.1
AC110015.1	B3GAT2	SYCP3	NTRK2	AL135999.1
AC107081.3	CEP290	FTX	HTT	SLC24A5
AC073525.1	BRAF	AL022322.2	SCN1A	GRM5-AS1
AP000812.4	ZEB2	AL121987.2	ANKRD11	SOD2
AC020659.1	CR1	GLS2	SPECC1	KDM5B
MIR497HG	FAT4	L3MBTL2-AS1	DLEU2	ZNF831
AL645768.1	MATR3	AC087521.2	NOTCH2	AUTS2
AC005329.3	DOCK7	SPATA1	NALCN	UBR4
AC135050.7	NR2F1-AS1	CLN3	SKP2	MBD5
AC008079.1	XPO1	DNAAF1	ITGA1	NABP1
LINC00649	SPTBN5	DST	EPB41	FAM153CP
CNTD1	DSE	AC006460.1	VPS13A	MIR4435-2HG
LHFPL5	SARM1	BSCL2	NPHP1	MED17
CATSPER2P1	GDF9	AC131571.1	C12orf45	AC125611.4
FOXH1	TF	ENO3	STARD9	AC068888.1
AP000688.1	ARHGEF12	SYNE1	NDUFA10	LINC00882
AC005703.5	SEMA3F-AS1	STARD5	TRPV6	CAPZB
AL161452.1	SYNC	SNX32	FGF13	FAT3
IPO9-AS1	PHLDB1	CFLAR-AS1	ATP6V0A2	HMGA2
AC245140.2	RNASEH2B	LINC00472	ARHGAP29	LINC00461
Z99774.1	KIF6	MRNIP	COPB2	DNAH9
IGFBP7-AS1	CDK13	DNAH7	CAPN3	AC037459.2
AP006621.1	MEG3	CACNA1A	LYST	SLC35E3
PABPC4-AS1	AC012443.2	IQCH-AS1	LAMB1	C4orf50
AP003465.1	NRXN1	ARID1B	UNC80	KDM6A
AC010285.3	ADGRL2	AC009950.1	SORL1	AAK1
CERKL	EEF1D	KYNU	AHNAK2	CLVS1
USP3-AS1	EFHC1	SLC4A5	RUBCN	RUNDC3A-AS1
AL109811.1	DNAH17	SCN1A-AS1	SCARB2	NDUFS7
OBSCN	ZC3H14	VAC14-AS1	ZNF827	GTF3C2-AS1
EMSLR	TMED2-DT	AC069209.2	AC093866.1	BTD
AC006059.5	AL355312.2	MAP3K15	DLEC1	ATXN1

Table 5.5: Gene list for enrichment analysis

Table 5.5: The gene list extracted from the differential gene expression analysis comprises 893 genes, selected based on their p-values exceeding a threshold of 1. These genes will be subjected to enrichment analysis to identify overrepresented biological pathways, functions, or processes.

5.11.2 SOURCE CODE FOR ENRICHMENT ANALYSIS USING ENRICHR PACKAGE:

1. SOURCE CODE

```
1 library(enrichR)
2
3 websiteLive <- getOption("enrichR.live")
4 if (websiteLive) {
5   listEnrichrSites()
6   setEnrichrSite("Enrichr") # Human genes
7 }
8 if (websiteLive) dbs <- listEnrichrDbs()
9 if (websiteLive) head(dbs)
10 dbs <- c("GO_Molecular_Function_2015", "GO_Cellular_Component_2015", "GO_Biological_Process_2015")
11
12 if (websiteLive) {
13   enriched <- enrichr(c("AC006064.4", "AC011603.2", "AL133415.1", "AL138963.3",
14     "AC093484.3", "AC093330.1", "ATP1A1-AS1", "MT-TS1",.....), dbs)
15 }
16
17
18
```

The provided R code utilizes the enrichR package to perform enrichment analysis on a list of genes. First, it sets up the Enrichr website and retrieves available databases if the website is live. Then, it specifies the databases to be used for enrichment analysis, focusing on gene ontology (GO) terms related to molecular function, cellular component, and biological process. This process evaluates whether these genes are significantly associated with any GO terms from the selected databases

2. CODE FOR BIOLOGICAL PROCESS

```
19 #BIOLOGICAL PROCESS
20 if (websiteLive) enriched[["GO_Biological_Process_2015"]]
21 Data <- enriched[["GO_Biological_Process_2015"]]
22 if (websiteLive) {
23   plotEnrich(enriched[[3]], showTerms = 15, numChar = 80, y = "Count", orderBy = "P.value")
24 }
25 top_rows <- Data[16:30, ]
26 if (websiteLive) {
27   plotEnrich(top_rows, showTerms = 15, numChar = 100, y = "Count", orderBy = "P.value")
28 }
29
30 |
```

The provided R code continues the enrichment analysis by focusing on the biological process aspect of gene ontology (GO) terms. First, it extracts the results related to biological processes from the enriched object and assigns it to the Data variable. Then, it plots the top enriched biological process terms using the PlotEnrich function, displaying the top 15 terms ordered by p-value.

Additionally, it selects a subset of the top enriched biological process terms (rows 16 to 30) from the Data object and assigns them to the top_rows variable. Finally, it plots these selected terms using the plotEnrich function, again displaying the top 15 terms ordered by p-value.

3. CODE FOR CELLULAR COMPONENT

```
31 #CELLULAR COMPONENT
32 Data1 <- enriched[["GO_Cellular_Component_2015"]]
33 if (websiteLive) {
34   plotEnrich(Data1, showTerms = 15, numChar = 80, y = "Count", orderBy = "P.value")
35 }
36 top_rows1 <- Data1[16:30, ]
37 if (websiteLive) {
38   plotEnrich(top_rows1, showTerms = 15, numChar = 80, y = "Count", orderBy = "P.value")
39 }
40
41 |
```

This section of the code focuses on the cellular component aspect of gene ontology (GO) terms in the enrichment analysis. It first extracts cellular component results from the enriched object and plots the top 15 enriched terms by p-value. Then, it selects a subset of top terms (rows 16 to 30) and plots them similarly. These plots provide insights into cellular component associations with the input genes.

4. CODE FOR MOLECULAR FUNCTION:

```
42 #MOLECULAR FUNCTION
43 Data2 <- enriched[["GO_Molecular_Function_2015"]]
44 if (websiteLive) {
45   plotEnrich(Data2, showTerms = 15, numChar = 80, y = "Count", orderBy = "P.value")
46 }
47 top_rows2 <- Data2[16:30, ]
48 if (websiteLive) {
49   plotEnrich(top_rows2, showTerms = 15, numChar = 80, y = "Count", orderBy = "P.value")
50 }
51
52
```

This section of the code focuses on the molecular function aspect of gene ontology (GO) terms in the enrichment analysis. It first extracts molecular function results and plots the top 15 enriched terms by p-value. Then, it selects a subset of top terms (rows 16 to 30) and plots them similarly. These plots provide insights into molecular function associations with the input genes. .

Term	Overlap	P. value	Adjusted .p.value	Genes
Membrane depolarization during action potential (GO:0086010)	9/38	3.40E-05	0.09984	SCN9A;CACNA1A;CACNA1D;ANK2;CACNA1C;SCN2A;NALCN;SLC8A1;SCN1A
Post embryonic development (GO:0009791)	12/84	3.37E-04	0.49560	SMAD2;SZT2;PLEKHA1;SCN9A;TBCE;ITPR1;SOD2;ATG7;SLC8A1;SLC18A2;GABRG2;IGF2R
Membrane depolarization (GO:0051899)	13/106	8.70E-04	0.79360	NRXN1;CACNA1A;ANK2;CACNA1D;CACNA1C;NALCN;GRIN2B;SLC8A1;CACNB4;ATXN1;SCN9A;SCN2A;SCN1A
Positive regulation of transporter activity (GO:0032411)	8/52	2.00E-03	0.79360	ASPH;KMT2A;S100A1;CNKSR3;HTT;NEDD4L;ANK2;ANK3
Calcium ion import (GO:0070509)	6/32	2.59E-03	0.79360	CACNA1A;CASK;CACNA1D;CACNA1C;NALCN;SLC8A1
Regulation of transmembrane transport (GO:0034762)	26/327	3.29E-03	0.79360	AHNAK;NRXN1;KCNC4;NEDD4L;HTT;CACNA1A;CACNA1D;CACNA1C;NALCN;OAZ3;SCN9A;SCN1A;KCNH1;KCNJ6;SLC31A2;S100A1;CASK;ANK2;ANK3;OPRM1;CACNB4;JSRP1;ASPH;CNKSR3;KCNMA1;SCN2A
Positive regulation of ion transmembrane transporter activity (GO:0032414)	7/45	3.83E-03	0.79360	ASPH;S100A1;CNKSR3;HTT;NEDD4L;ANK2;ANK3
Regulation of ion transmembrane transport (GO:0034765)	25/314	3.83E-03	0.79360	AHNAK;NRXN1;KCNC4;NEDD4L;HTT;CACNA1A;CACNA1D;CACNA1C;NALCN;SCN9A;SCN1A;KCNH1;KCNJ6;SLC31A2;S100A1;CASK;ANK2;ANK3;OPRM1;CACNB4;JSRP1;ASPH;CNKSR3;KCNMA1;SCN2A
Negative regulation of cytosolic calcium ion concentration (GO:0051481)	3/8	4.18E-03	0.79360	OPRM1;RYR3;SLC8A1

Actin filament depolymerization (GO:0030042)	3/8	4.18E-03	0.79360	VIL1;MICAL3;MICAL2
Microtubule-based movement (GO:0007018)	17/187	4.32E-03	0.79360	CCDC39;DYNC1H1;DNAH11;DNAH10;DNAH17;DST;DNAH7;DNAAF1;STARD9;HTT;DNAH9;SPG7;KLC1;KIF6;CLN3;KIF5C;HYDIN
Behavioral response to pain (GO:0048266)	4/16	4.65E-03	0.79360	GIT2;SCN9A;CACNA1A;GRIN2B
Calcium ion transmembrane transport (GO:0070588)	13/128	4.73E-03	0.79360	ITPR1;CACNA1A;CACNA1D;PKD2L2;CACNA1C;OPRM1;NALCN;GRIN2B;PKD1;RYR3;CACNB4;TRPV6;ASPH
Amino acid transport (GO:0006865)	13/128	4.73E-03	0.79360	NTRK2;SLC46A1;GLS2;SLC1A2;CACNA1A;SLC3A1;HTT;CLN3;SLC25A26;SLC7A6;CACNB4;CPT2;SLC19A1
Regulation of calcium ion transmembrane transport (GO:1903169)	7/48	5.11E-03	0.79360	JSRP1;CACNB4;AHNAK;ASPH;S100A1;HTT;ANK2

Table 5.6: Biological process enriched terms

Table 5.6: Through GO enrichment analysis of 894 differentially expressed genes, it was found that genes were mainly enriched in biological processes such as regulation of ion transmembrane transport, calcium ion import, amino acid transport, actin filament depolymerization and transmembrane transport.

5.12 GO ENRICHMENT ANALYSIS OF BIOLOGICAL PROCESS

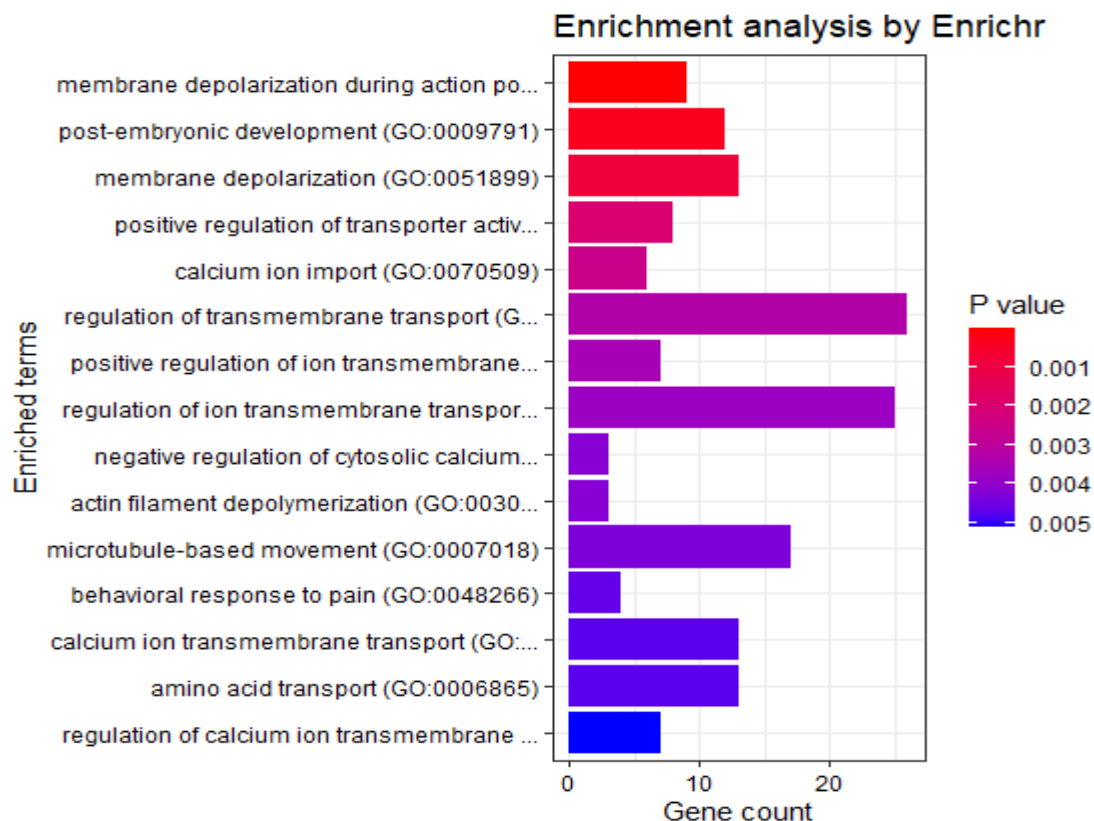


Figure 5.7: GO enriched terms in biological process pathway

Figure 5.7: The GO enrichment analysis of 894 differentially expressed genes in brain cancer revealed a diverse landscape of biological processes, comprising a total of 2937 enriched terms. Among the top 15 enriched terms were crucial processes. This comprehensive analysis underscores the intricate molecular mechanisms underlying brain cancer pathogenesis, offering valuable insights into potential therapeutic targets and diagnostic biomarkers for further investigation.

The GO term "membrane depolarization during action potential (GO:0086010)" was enriched with genes such as SCN9A, CACNA1A, CACNA1D, ANK2, CACNA1C, SCN2A, NALCN, SLC8A1, and SCN1A. This enrichment suggests their involvement in processes related to depolarization of the cell membrane during action potential generation. Interestingly, a study by Ozkucur et al. (2015) demonstrated that depolarization of the membrane potential in mature neurons led to significant changes

in glia cell density and morphology. These findings are particularly relevant in the context of postnatal brain development and neurological diseases, where alterations in neuron-glia interactions are commonly observed, including in glioblastoma multiforme (GBM).

The GO term "post-embryonic development (GO:0009791)" was enriched with genes like SMAD2, SZT2, PLEKHA1, SCN9A, TBCE, ITPR1, SOD2, ATG7, SLC8A1, SLC18A2, GABRG2, and IGF2R. Curry et al. (2021) implicated altered expression profiles of growth factors and their receptors, including epidermal growth factor (EGF), platelet-derived growth factor (PDGF), and vascular endothelial growth factor (VEGF), in the development and progression of primary brain tumors. These factors play crucial roles in post-embryonic development and are also implicated in the pathogenesis of GBM.

The GO term "membrane depolarization (GO:0051899)" was enriched with genes like NRXN1, CACNA1A, ANK2, CACNA1D, CACNA1C, NALCN, GRIN2B, SLC8A1, CACNB4, ATXN1, SCN9A, SCN2A, and SCN1A. This enrichment indicates their involvement in regulating membrane depolarization processes, particularly in excitable cells like neurons. Yang et al. (2013) highlighted the significance of voltage-gated ion channels and transporters, such as sodium and potassium channels, in fine-tuning action potential waveforms, which is crucial for normal neuronal function and may also have implications in diseases like GBM.

The Gene Ontology (GO) term "Positive regulation of transporter activity (GO:0032411)" revealed enrichment of genes such as ASPH, KMT2A, S100A1, CNKSR3, HTT, NEDD4L, ANK2, and ANK3. Notably, major neurological diseases, including glioblastoma multiforme (GBM), are associated with structural and functional changes in the blood-brain barrier (BBB). The impact of these neurodegenerative diseases on transporter regulation, particularly the ATP-binding cassette (ABC) transporter P-glycoprotein (P-gp), has been extensively characterized (Taggi, V. et al., 2022).

The GO term "Calcium ion import (GO:0070509)" showed enrichment of genes such as CACNA1A, CASK, CACNA1D, CACNA1C, NALCN, and SLC8A1. In addition to their diverse roles in normal brain function, proteins involved in calcium transport are implicated in various brain cancers, including GBM. These calcium-transporting proteins play crucial roles in transmitting signals from the brain tumor microenvironment to cancer cells, highlighting an area of research that necessitates further investigation (Maklad A. et al., 2010).

The GO term "Regulation of transmembrane transport (GO:0034762)" exhibited enrichment of genes including AHNAK, NRXN1, KCNC4, NEDD4L, HTT, CACNA1A, CACNA1D, CACNA1C, NALCN, OAZ3, SCN9A, SCN1A, KCNH1, KCNJ6, SLC31A2, S100A1, CASK, ANK2, ANK3, OPRM1, CACNB4, JSRP1, ASPH, CNKSR3, KCNMA1, and SCN2A. Understanding the mechanistic pathways and connectivity involved in regulating transporter expression and activity, particularly in critical barriers such as the BBB, holds promise for identifying novel therapeutic targets for GBM and other neurological disorders (Taggi, V. et al., 2023).

The GO term "Positive regulation of ion transmembrane transporter activity (GO:0032414)" highlighted enrichment of genes such as ASPH, S100A1, CNKSR3, HTT, NEDD4L, ANK2, and ANK3. Ion transporters play pivotal roles in maintaining ionic homeostasis and cellular signal transduction under physiological conditions. Notably, ion transporters such as NKCC-1 and NHE-1 have emerged as significant players in cancer progression, including GBM (Cong D et al., 2015).

The GO term "Regulation of ion transmembrane transport (GO:0034765)" demonstrated enrichment of genes including AHNAK, NRXN1, KCNC4, NEDD4L, HTT, CACNA1A, CACNA1D, CACNA1C, NALCN, SCN9A, SCN1A, KCNH1, KCNJ6, SLC31A2, S100A1, CASK, ANK2, ANK3, OPRM1, CACNB4, JSRP1, ASPH, CNKSR3, KCNMA1, and SCN2A. Understanding the roles of ion transporters, particularly sodium-potassium-chloride cotransporter isoform 1 (NKCC-1) and sodium-hydrogen exchanger isoform 1 (NHE-1), in GBM and other malignant tumors is critical for

identifying novel therapeutic strategies (Cong D et al., 2015).

The GO term "Negative regulation of cytosolic calcium ion concentration (GO:0051481)" indicated enrichment of genes such as OPRM1, RYR3, and SLC8A1. Dysregulation of calcium signaling pathways is implicated in various neurological disorders, including GBM. Understanding the intricate mechanisms governing calcium ion concentration may offer insights into novel therapeutic interventions (Maklad A. et al., 2010).

The GO term "Actin filament depolymerization (GO:0030042)" revealed enrichment of genes including VIL1, MICAL3, and MICAL2. Actin dynamics play crucial roles in cytoskeletal regulation, and dysregulation of actin depolymerization is implicated in cancer progression, including GBM. Understanding the role of actin-regulatory proteins in tumor biology may unveil potential therapeutic targets (Hotulainen P et al., 2009).

The GO term "Microtubule-based movement (GO:0007018)" demonstrated enrichment of genes including CCDC39, DYNC1H1, DNAH11, DNAH10, DNAH17, DST, DNAH7, DNAAF1, STARD9, HTT, DNAH9, SPG7, KLC1, KIF6, CLN3, KIF5C, and HYDIN. Microtubules play essential roles in cellular processes such as cell division and intracellular transport. Dysregulated microtubule-based movement may contribute to the pathogenesis of GBM, underscoring the importance of investigating these mechanisms further (Dehmelt L et al., 2006).

The GO term "Behavioral response to pain (GO:0048266)" revealed enrichment of genes including GIT2, SCN9A, CACNA1A, and GRIN2B. Pain perception and management are critical aspects of patient care in GBM. Understanding the molecular mechanisms underlying pain response may facilitate the development of targeted therapies to alleviate pain and improve patient outcomes (Cho Y et al., 2021).

The GO term "Calcium ion transmembrane transport (GO:0070588)" exhibited enrichment of genes including ITPR1, CACNA1A, CACNA1D, PKD2L2, CACNA1C,

OPRM1, NALCN, GRIN2B, PKD1, RYR3, CACNB4, and TRPV6. Dysregulated calcium ion transport is implicated in various aspects of tumor biology, including GBM. Investigating the roles of calcium-transporting proteins may offer insights into novel therapeutic strategies for targeting tumor-specific characteristics (Cong D et al., 2015).

The GO term "Amino acid transport (GO:0006865)" revealed enrichment of genes including NTRK2, SLC46A1, GLS2, SLC1A2, CACNA1A, SLC3A1, HTT, CLN3, SLC25A26, SLC7A6, CACNB4, CPT2, and SLC19A1. Amino acid transporters play crucial roles in brain physiology and are implicated in various neurological disorders, including GBM. Investigating the dysregulation of amino acid transport may provide insights into novel therapeutic avenues for GBM treatment (Zaragoza R et al., 2020).

The GO term "Regulation of calcium ion transmembrane transport (GO:1903169)" demonstrated enrichment of genes including JSRP1, CACNB4, AHNAK, ASPH, S100A1, HTT, and ANK2. Calcium ion transport is tightly regulated and is essential for various physiological processes in the brain. Dysregulation of calcium ion transport is implicated in neurological disorders, including GBM. Investigating the regulatory mechanisms governing calcium ion transport may offer insights into novel therapeutic strategies for GBM treatment (Maklad A. et al., 2010).

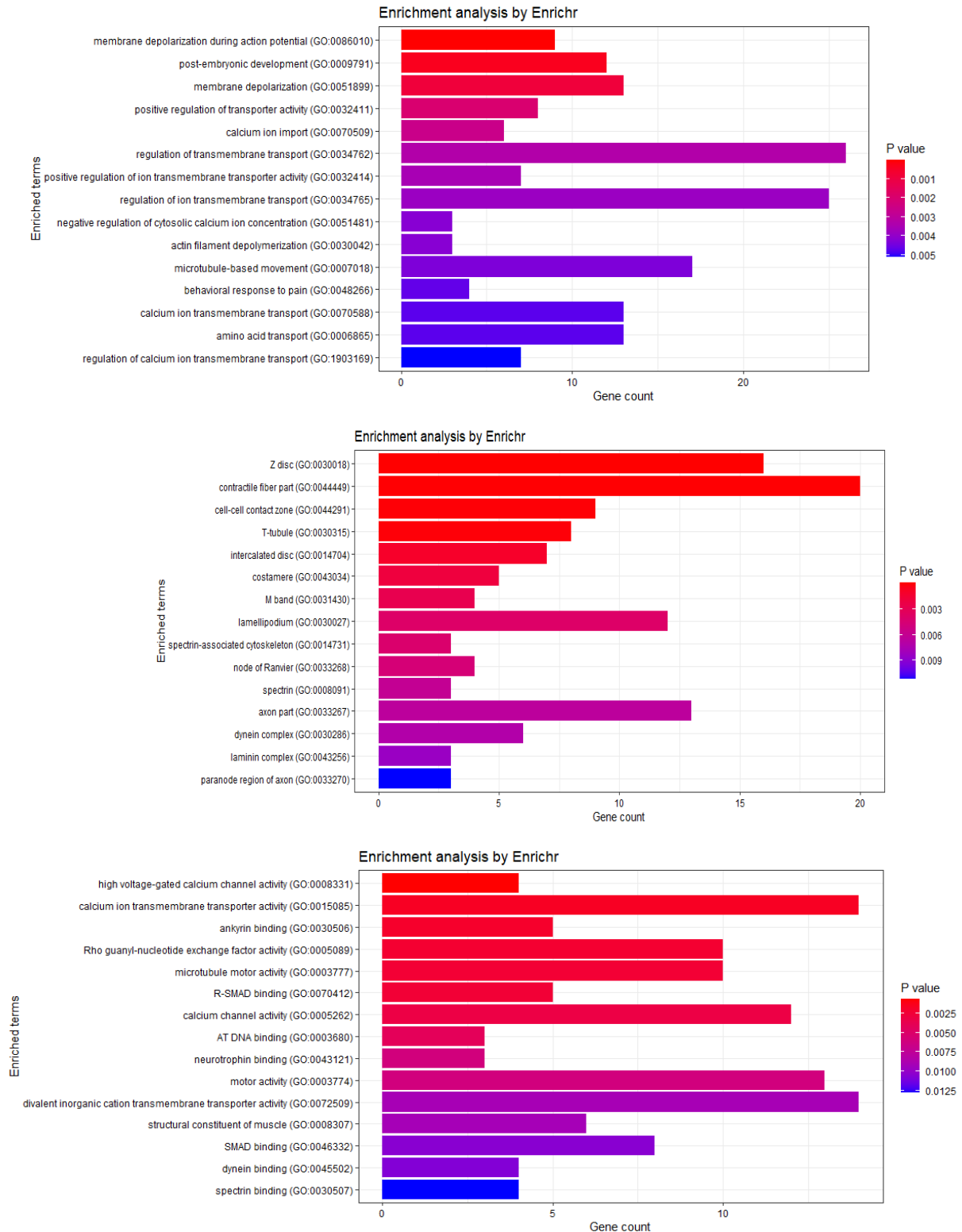


Figure 5.7(a): GO enriched in biological processes (BP, MF, CC)

Figure 5.7(a): The histograms depicting GO enrichment analysis reveals the distribution of enriched terms across pathways like biological processes (BP), cellular components (CC), and molecular functions (MF). The X-axis represents the number of genes enriched, while the Y-axis represents the modules of GO enriched. The color shade reflects the significance of P values. This analysis focuses on the top 15 terms within each pathway, although there are a total of 2937 terms for biological processes, 392 terms for cellular components, and 622 terms for molecular functions.

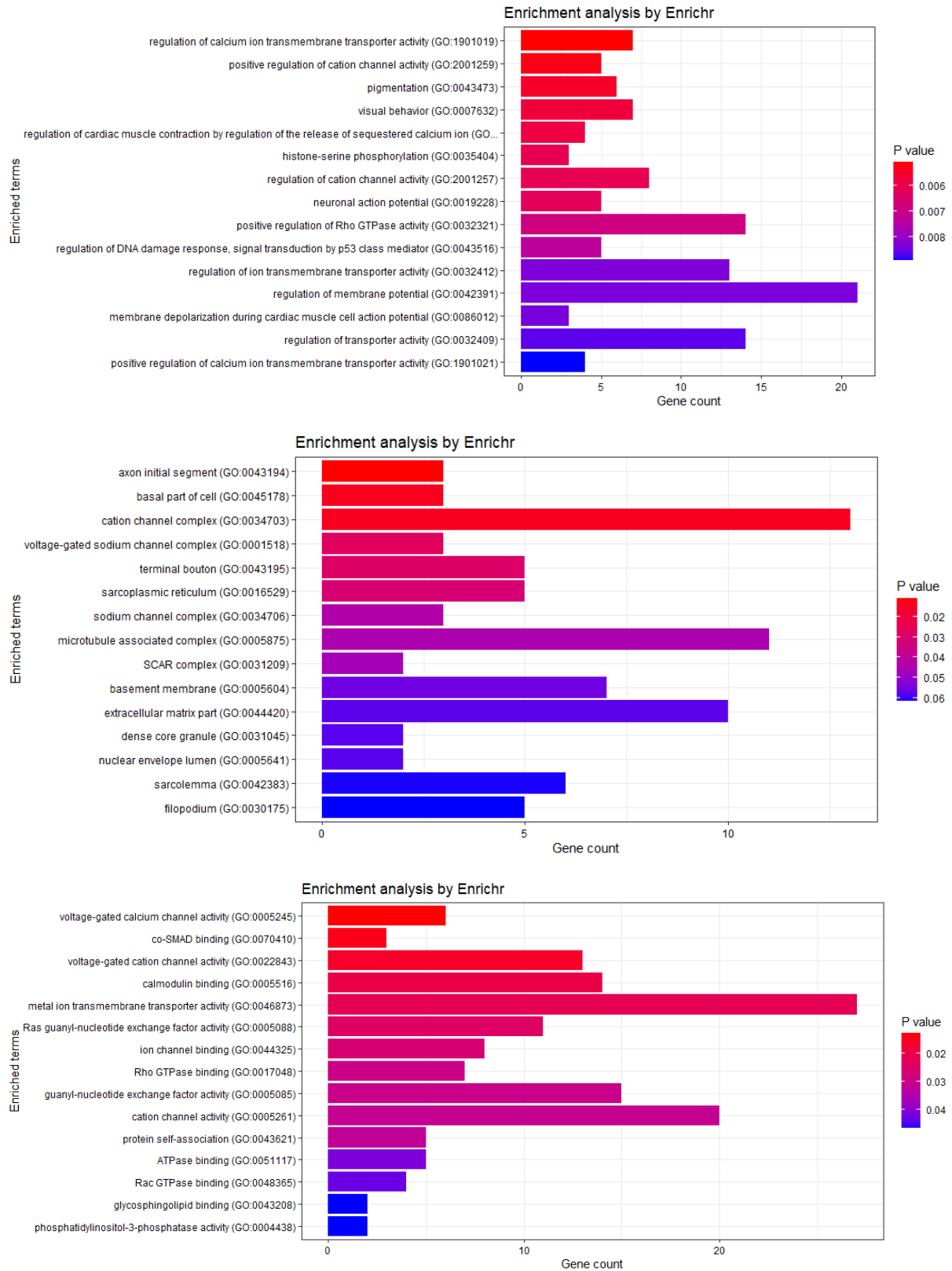


Figure 5.7(b): GO enriched in biological processes (BP, MF, CC)

Figure 5.7(b): The histograms depicting GO enrichment analysis reveals the distribution of next 15 enriched terms across pathways like biological processes (BP), cellular components (CC), and molecular functions (MF). The X-axis represents the number of genes enriched, while the Y-axis represents the modules of GO enriched. The color shade reflects the significance of P values. This analysis focuses on the next top 15 terms within each biological pathway, although there are a total of 2937 terms for biological processes, 392 terms for cellular components, and 622 terms for molecular functions.

CHAPTER 6

SUMMARY AND CONCLUSION

Malignant primary brain tumors remain among the most difficult cancers to treat, with a 5 year overall survival no greater than 35%. The most common malignant primary brain tumors in adults are glioblastoma multiforme.

The study was done by using large dataset which contains 60660 gene ids in RNA-Seq of glioblastoma multiforme which is extracted from the cancer genome atlas (TCGA) database and analyzing this dataset using RNALysis software and using packages like TCGAbiolinks and Enrichr packages in R studio to get a wholistic view on the expression of genes and GO enriched terms and pathways.

This field involves analyzing Rna-Seq data of Glioblastoma multiforme by performing data preprocessing, exploratory data analysis, clustering, differential gene expression analysis and enrichment analysis and how these enriched terms functions, pathways and genes contribute to the brain cancer or tumor by using the RNALysis software and R studio.

The main goal of this project is to present RNALysis software to analyze the large RNA-Seq data of brain cancer, and understand that, which is also used for the TCGA datasets and also analyze different cancers datasets like raw count or expression data. The scope of this project is to perform enrichment analysis based on the concept of enriched functional GO term enrichment and pathway analysis of differential expressed genes. Perform different functional categories according to the GO term enrichment analysis for Biological process (BP), molecular function (MF) and cellular component (CC).

In conclusion, the present study sheds light on the intricate pathogenesis of advanced glioblastoma multiforme (GBM) and identifies several key genes and pathways warranting further investigation as potential prognostic markers. Utilizing RNALysis

software and R Studio, pinpointed a comprehensive set of significantly regulated genes, including notable candidates like SCN1A, CACNA1C, ANK2, CASK, and JSRP1, which may play pivotal roles in GBM development.

Gene Ontology enrichment analysis of 843 identified genes revealed their involvement in crucial biological processes relevant to brain cancer. Particularly, terms such as membrane depolarization during action potential, regulation of transporter activity, calcium ion import, ion transmembrane transport, and actin filament depolymerization emerged as top enriched categories, highlighting their significance in GBM pathogenesis.

Nevertheless, further validation through molecular biological experiments, computational analyses, and gene co-expression network studies is imperative to elucidate the precise functions of these identified genes in GBM. Recognizing the importance of investigating gene changes and expressions in RNA-Seq data can offer valuable insights into potential therapeutic targets and predictive biomarkers for GBM, paving the way for future research endeavors in this field.

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. Birzu, C., French, P., Caccese, M., Cerretti, G., Idbaih, A., Zagonel, V., & Lombardi, G. (2020). Recurrent Glioblastoma: From Molecular Landscape to New Treatment Perspectives. *Cancers*, 13(1), 47. <https://doi.org/10.3390/cancers13010047>
3. Cho, Y., & Hong, G. S. (2021). Behavioral and Physiological Pain Responses in Brain-Injured Patients Who Are Unable to Communicate in the Intensive Care Unit. *Pain management nursing : official journal of the American Society of Pain Management Nurses*, 22(1), 80–85. <https://doi.org/10.1016/j.pmn.2020.08.006>
4. Cong, D., Zhu, W., Kuo, J. S., Hu, S., & Sun, D. (2015). Ion transporters in brain tumors. *Current medicinal chemistry*, 22(10), 1171–1181. <https://doi.org/10.2174/0929867322666150114151946>
5. Curry, R. N., & Glasgow, S. M. (2021). The Role of Neurodevelopmental Pathways in Brain Tumors. *Frontiers in cell and developmental biology*, 9, 659055. <https://doi.org/10.3389/fcell.2021.659055>
6. Davis M. E. (2016). Glioblastoma: Overview of Disease and Treatment. *Clinical journal of oncology nursing*, 20(5 Suppl), S2–S8. <https://doi.org/10.1188/16.CJON.S1.2-8>
7. Dehmelt, L., Nalbant, P., Steffen, W., & Halpain, S. (2006). A microtubule-based, dynein-dependent force induces local cell protrusions: Implications for neurite

- initiation. *Brain cell biology*, 35(1), 39–56. <https://doi.org/10.1007/s11068-006-9001-0>
8. Deshpande, D., Chhugani, K., Chang, Y., Karlsberg, A., Loeffler, C., Zhang, J., Muszyńska, A., Munteanu, V., Yang, H., Rotman, J., Tao, L., Balliu, B., Tseng, E., Eskin, E., Zhao, F., Mohammadi, P., P Łabaj, P., & Mangul, S. (2023). RNA-seq data science: From raw data to effective interpretation. *Frontiers in genetics*, 14, 997383. <https://doi.org/10.3389/fgene.2023.997383>
 9. Ge, S.X., Son, E.W. & Yao, R. iDEP: an integrated web application for differential expression and pathway analysis of RNA-Seq data. *BMC Bioinformatics* **19**, 534 (2018). <https://doi.org/10.1186/s12859-018-2486-6>
 10. Grochans, S., Cybulska, A. M., Simińska, D., Korbecki, J., Kojder, K., Chlubek, D., & Baranowska-Bosiacka, I. (2022). Epidemiology of Glioblastoma Multiforme-Literature Review. *Cancers*, 14(10), 2412. <https://doi.org/10.3390/cancers14102412>
 11. Hanif, F., Muzaffar, K., Perveen, K., Malhi, S. M., & Simjee, S.hU. (2017). Glioblastoma Multiforme: A Review of its Epidemiology and Pathogenesis through Clinical Presentation and Treatment. *Asian Pacific journal of cancer prevention: APJCP*, 18(1), 3–9. <https://doi.org/10.22034/APJCP.2017.18.1.3>
 12. Hotulainen, P., Llano, O., Smirnov, S., Tanhuanpää, K., Faix, J., Rivera Baeza, C., & Lappalainen, P. (2009). Defining mechanisms of actin polymerization and depolymerization during dendritic spine morphogenesis. *Journal of Cell Biology*, 185(2), 323-339. <https://doi.org/10.1083/jcb.200809046>
 13. Ishiuchi, S., Yoshida, Y., Sugawara, K., Aihara, M., Ohtani, T., Watanabe, T., Saito, N., Tsuzuki, K., Okado, H., Miwa, A., Nakazato, Y., & Ozawa, S. (2007). Ca²⁺-permeable AMPA receptors regulate growth of human glioblastoma via

Akt activation. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 27(30), 7987–8001.
<https://doi.org/10.1523/JNEUROSCI.2180-07.2007>

14. 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.
15. Jamal, A., Handayani, A., Septiandri, A.A., Ripmiatin, E. and Effendi, Y., 2018. Dimensionality reduction using pca and k-means clustering for breast cancer prediction. *Lontar Komput. J. Ilm. Teknol. Inf*, 9(3), pp.192-201
16. Jayakumar, A. R., & Norenberg, M. D. (2010). The Na-K-Cl Co-transporter in astrocyte swelling. *Metabolic brain disease*, 25(1), 31–38.
<https://doi.org/10.1007/s11011-010-9180-3>
17. J. Ricardo McFaline-Figueroa, Eudocia Q. Lee, Brain Tumors, *The American Journal of Medicine*, Volume 131, Issue 8, 2018, Pages 874-882, ISSN 0002-9343, <https://doi.org/10.1016/j.amjmed.2017.12.039>
18. Koch, C. M., Chiu, S. F., Akbarpour, M., Bharat, A., Ridge, K. M., Bartom, E. T., & Winter, D. R. (2018). A Beginner's Guide to Analysis of RNA Sequencing Data. *American journal of respiratory cell and molecular biology*, 59(2), 145–157. <https://doi.org/10.1165/rcmb.2017-0430TR>
19. Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S. L., Jagodnik, K. M., Lachmann, A., McDermott, M. G., Monteiro, C. D., Gundersen, G. W., & Ma'ayan, A. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research*, 44(W1), W90–W97. <https://doi.org/10.1093/nar/gkw377>

20. 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>
21. Maklad A, Sharma A, Azimi I. Calcium Signaling in Brain Cancers: Roles and Therapeutic Targeting. *Cancers*. 2019; 11(2):145. <https://doi.org/10.3390/cancers1102014>
22. Manipur, I., Granata, I., Maddalena, L. et al. Clustering analysis of tumor metabolic networks. *BMC Bioinformatics* 21 (Suppl 10), 349 (2020)
23. Martin J van den Bent, Marjolein Geurts, Pim J French, Marion Smits, David Capper, Jacqueline E C Bromberg, Susan M Chang, Primary brain tumors in adults, *The Lancet*, Volume 402, Issue 10412, 2023, Pages 1564-1579, ISSN 0140-6736, [https://doi.org/10.1016/S0140-6736\(23\)01054-1](https://doi.org/10.1016/S0140-6736(23)01054-1)
24. 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
25. Özkucur, N., Quinn, K. P., Pang, J. C., Du, C., Georgakoudi, I., Miller, E., Levin, M., & Kaplan, D. L. (2015). Membrane potential depolarization causes alterations in neuron arrangement and connectivity in cocultures. *Brain and behavior*, 5(1), 24–38. <https://doi.org/10.1002/brb3.295>
26. Paczkowska, M., Barenboim, J., Sintupisut, N. et al. Integrative pathway enrichment analysis of multivariate omics data. *Nat Commun* 11, 735 (2020)
27. 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

28. Roda, E., & Bottone, M. G. (2022). Editorial: Brain Cancers: New Perspectives and Therapies. *Frontiers in neuroscience*, 16, 857408. <https://doi.org/10.3389/fnins.2022.857408>

29. Seyed Hossein Hassanpour, Mohammadamin Dehghani, Review of cancer from perspective of molecular, *Journal of Cancer Research and Practice*, Volume 4, Issue 4, 2017, Pages 127-129, ISSN 23113006, <https://doi.org/10.1016/j.jcrpr.2017.07.001>

30. Silva, T. C., Colaprico, A., Olsen, C., D'Angelo, F., Bontempi, G., Ceccarelli, M., & Noushmehr, H. (2016). *TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages*. *F1000Research*, 5, 1542. <https://doi.org/10.12688/f1000research.8923.2>

31. 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>

32. Stewart, Geoffrey, and Mahmood Al-Khassaweneh. 2022. "An Implementation of the HDBSCAN* Clustering Algorithm" *Applied Sciences* 12, no. 5: 2405.

33. Taggi, V., Riera Romo, M., Piquette-Miller, M., Meyer Zu Schwabedissen, H. E., & Neuheff, S. (2022). Transporter Regulation in Critical Protective Barriers: Focus on Brain and Placenta. *Pharmaceutics*, 14(7), 1376. <https://doi.org/10.3390/pharmaceutics14071376>

34. Teichman, G., Cohen, D., Ganon, O., Dunskey, N., Shani, S., Gingold, H., & Rechavi, O. (2023). RNAlysis: analyze your RNA sequencing data without

writing a single line of code. *BMC biology*, 21(1), 74.
<https://doi.org/10.1186/s12915-023-01574-6>

35. 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
36. Yang, M., & Brackenbury, W. J. (2013). Membrane potential and cancer progression. *Frontiers in physiology*, 4, 185.
<https://doi.org/10.3389/fphys.2013.00185>
37. Yekula, A., Yekula, A., Muralidharan, K., Kang, K., Carter, B. S., & Balaj, L. (2020). Extracellular Vesicles in Glioblastoma Tumor Microenvironment. *Frontiers in immunology*, 10, 3137.
<https://doi.org/10.3389/fimmu.2019.03137>
38. Zaragozá R. (2020). Transport of Amino Acids Across the Blood-Brain Barrier. *Frontiers in physiology*, 11, 973.
<https://doi.org/10.3389/fphys.2020.00973>
39. Zhu, Q., Sun, Y., Zhou, Q., He, Q., & Qian, H. (2018). Identification of key genes and pathways by bioinformatics analysis with TCGA RNA sequencing data in hepatocellular carcinoma. *Molecular and clinical oncology*, 9(6), 597–606.
<https://doi.org/10.3892/mco.2018.1728>