Gene Expression Differences and Similarities between Bipolar Disorder and Gliomas Studied by Brain Expression Data

Miao Wang

Abstract:

Both bipolar disorder and glioma are highly prevalent diseases. Bipolar disorder is a significant mental illness, while glioma constitutes 46% of intracranial tumors. By conducting differential expression analysis, Weighted Gene Co-expression Network Analysis, and Gene Ontology enrichment analysis on brain expression data of bipolar disorder and glioma, it was found that these two conditions share three terms: protein localization to organelle, organelle localization, and protein maturation. Exploring targeted therapy and drug-targeting research on these three terms can offer novel insights for the treatment of both bipolar disorder and glioma.

Keywords: bipolar disorder, glioma, differential expression analysis, enrichment analysis

Introduction

Bipolar disorder (BP), a common mental illness, is a mood disorder characterized by episodes of mania or hypomania, as well as episodes of depression. This disorder is characterized by unstable and extreme moods, alternating between manic and depressive states. BP affects more than 1% of the global population, regardless of nationality, ethnicity, or socioeconomic level (Grande, Berk, Birmaher, & Vieta, 2016). Bipolar disease is one of the biggest causes of disability in young people, causing cognitive and functional impairment as well as increased mortality, including suicide (Grande et al., 2016).

Brain tumors, including various intracranial tumors, can be divided into two categories: benign and malignant. Malignant tumors are also called brain cancer. Glioma is a highly malignant tumor that grows very fast. Among brain tumors, the incidence of glioma accounts for approximately 46% of intracranial tumors. Depending on where the tumor grows in the brain, it will compress different nerves and affect the function of that part.

Transcriptome is the bridge connecting genetic information and biological functions. In a broad sense, it refers to the sum of all RNAs that can be transcribed in one or a group of cells under the same physiological conditions, including coding RNAs and non-coding RNAs; in a narrow sense, it refers to all mRNAs. Transcriptome sequencing analysis (RNA-seq) extracts the mRNA to be studied, reverse-transcribes it into a cDNA library, adds adapters to both ends of the DNA fragments, and uses high-throughput sequencing technology to count the number of relevant small fragments to calculate different mRNAs(Marguerat & Bähler, 2010). At the expression level, it can accurately identify alternative splicing sites and coding sequence single nucleotide polymorphisms and obtain sequence information of almost all transcripts in a specific tissue or organ of a species in a certain state. RNA-seq has been widely used in basic research, clinical diagnosis, drug development, and other fields.

The correlation between bipolar disorder and glioma has been less analyzed in past studies. By using differential expression analysis, enrichment analysis, and other analytical techniques to study the similarities and differences in brain expression between bipolar disorder and glioma, we can improve our understanding of the correlation between bipolar disorder and glioma and provide insights into the treatment of bipolar disorder.

Materials and Methods

1. Data download and collection

The main method used to analyze differential genes is transcriptome sequencing, which is used to compare the similarities and differences in brain expression between bipolar disorder and glioma. Starting with data collection, we extracted postmortem brain tissue RNA sequences from patients with bipolar disorder from the National Center for Bioinformatics Information (NCBI) database(Bowling KM, Jun 22, 2017). The data contains brain tissue RNA sequences of bipolar disorder, schizophrenia, depression, and control(Bowling KM, Jun 22, 2017). We extracted the RNA sequences of bipolar disorder and control and performed subsequent differential analyses. The glioma data were obtained from the National Cancer Institute GDC Data Portal, including RNA sequences of brain tissue from patients with and without glioma(NIH).

2. Differential expression analysis from RNA-Seq data

The high degree of data repeatability provided by lanes and flow cells, which decreases the number of technical duplicates required for the experiments, is one benefit of RNA-seq technology. Additionally, RNA-seq enables the identification and measurement of the expression of isoforms and unidentified transcripts(Agarwal et al., 2010).

The most direct way to find genes with significant expression changes between groups is to perform transcriptome sequencing to explain changes in gene expression levels on biological functions. DESeq2 mainly uses the negative binomial distribution model to perform differential analysis(Marguerat & Bähler, 2010). DESeq2 reduces technical variation among samples through normalization transformation and normalization and then estimates the dispersion of gene expression. It uses a negative binomial distribution model to identify differentially expressed genes and correct for multiple testing issues.

The Limma program utilizes linear model modeling as its underlying methodology. Originally developed for analyzing microarray data, it has since been expanded to include RNA-seq data analysis. The limma package is commonly used for gene expression chip data analysis, and many functionalities of the edgeR package depend on the limma package. Limma employs the empirical Bayes model to enhance result reliability, making it a widely used tool for differential analysis. Limma-voom, originally designed for microarray data analysis, has also been adapted for transcriptome data analysis. In the initial stage of RNA-Seq data analysis, the raw read count is transformed into log2-counts-per-million (logCPM). Then, the mean-variance relationship is estimated and modeled. Two distinct modeling methods are used in this field: Precision Weight (voom) and empirical Bayes prior trend (limma-trend) (Ritchie et al., 2015).

edgeR is a widely used tool for conducting differential expression analysis of RNA-seq expression profiles with biological replicates. It employs various statistical test methods based on the negative binomial distribution, such as empirical Bayes estimation, exact test, generalized linear model, and quasi-likelihood test(Robinson, McCarthy, & Smyth, 2010). Additionally, edgeR utilizes the negative binomial distribution for performing statistical tests. Before conducting the test, it is important to standardize the read count expression matrix to eliminate any group differences that may arise due to library size and composition.

3. WGCNA (weighted gene co-expression network analysis)

After conducting differential analysis, we performed a Weighted Gene Co-expression Network Analysis (WGCNA)(Langfelder & Horvath, 2008). WGCNA is a method used to identify gene clusters (modules) exhibiting high correlations in their expression patterns. These modules can be characterized by their module feature genes (e.g., eigengenes) or hub genes, which connect modules to sample traits. This analysis can help identify potential biomarker genes or therapeutic targets. WGCNA focuses on assigning weights to gene co-expression relationships and identifying sets of genes that change collaboratively. For instance, if a certain type of gene increases or decreases together in different samples, the expression pattern remains consistent. The weighting aspect involves assigning weights to gene pairs based on their correlation. Higher correlation results in higher weights, while lower correlation leads to lower weights.

The methodology of Weighted Gene Co-expression Network Analysis (WGCNA) consists of two main components: expression cluster analysis and phenotype correlation(Zhang & Horvath, 2005). These components involve four key steps: calculating correlation coefficients between genes, identifying gene modules, constructing a co-expression network, and establishing associations between modules and traits(Zhang & Horvath, 2005). The initial step involves computing the correlation coefficient, specifically the Person Coefficient, between two given genes. A screening threshold is typically set to assess the similarity of expression patterns. Genes that surpass the threshold are considered similar. However, when the threshold is set at 0.8, it becomes difficult to demonstrate a statistically significant distinction between values of 0.8 and 0.79. Therefore, WGCNA analysis uses a weighted value derived from the correlation coefficient, where the gene correlation coefficient is exponentiated to the power of N. This transformation ensures that the relationships between genes in the network adhere to the distribution of scale-free networks, which is biologically significant. The next step involves constructing a hierarchical clustering tree using the correlation coefficients among genes. The tree branches represent distinct gene modules, each represented by a different color. Genes are categorized into modules based on expression patterns, using the weighted correlation coefficient to measure association. Genes with similar patterns are clustered together within these modules. Analyzing gene expression patterns makes it possible to categorize tens of thousands of genes into concise and informative modules.

Results

1. Bipolar results:

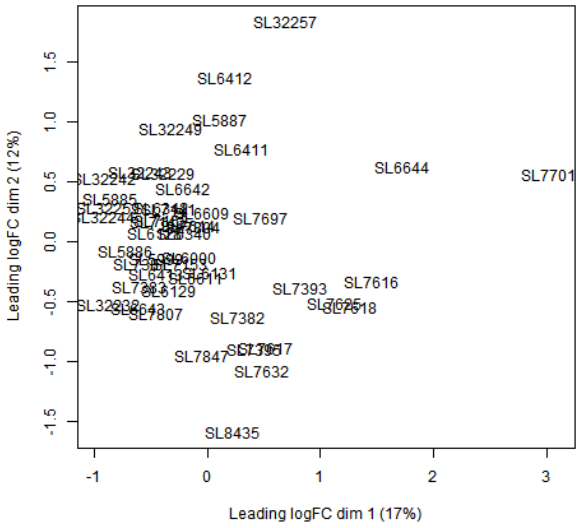

Figure 1. plotMDS can use logFC to view the grouping of samples. The distance between BP samples corresponds to the main biological variation coefficient between these samples.

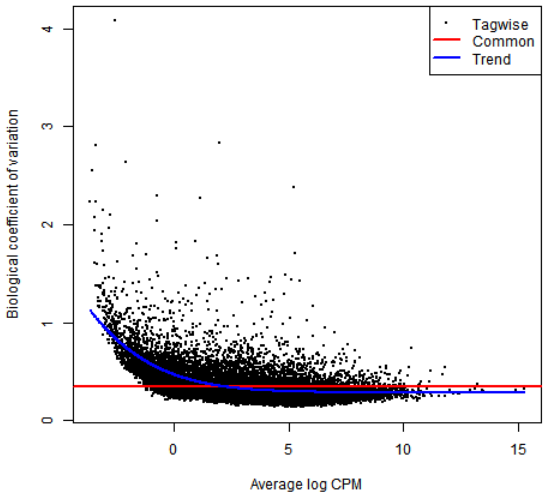

Figure 2. plotBCV reflects the fitting of BP genes with different expression levels to the model.

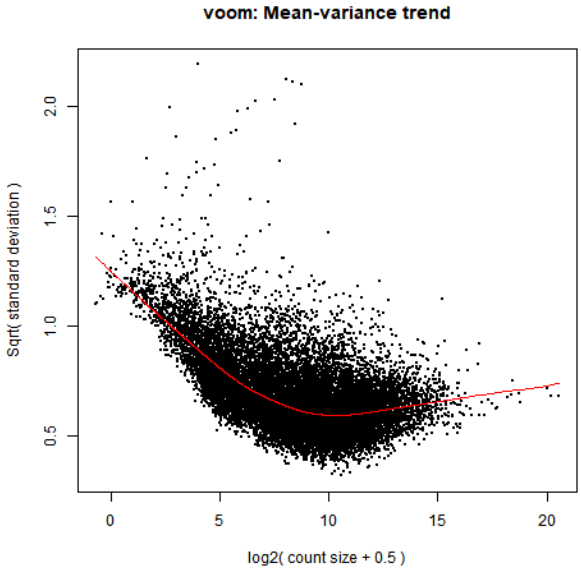

Figure 3. Voom transformation converts the read count data into log2-counts per million (logCPM) by estimating the mean-variance relationship and using it to calculate the appropriate observation-level weight. Finally, the data is linearly modeled.

	limma	logFC	DESeq2
limma	1.0000000	0.7946691	0.8714990
logFC	0.7946691	1.0000000	0.7989693
DESeq2	0.8714990	0.7989693	1.0000000

Figure 4. The three major differential expression analysis packages, limma, edgeR, and DESeq2, each have their benefits. This figure shows the result of comparing these three differential expression analysis packages.

Table 1. This table compiles the top 15 genes from the three differential expression analysis package analysis results. These genes have a higher probability of being differential genes.

	limma_logFC	adj. p	edgeR_logFC	FDR	DESeq2_logFC	padj
ENSG00000225972	-0.317025933	0.764723	-4.579606885	3.32E-06	0.435237303	0.293836
ENSG00000229344	0.57201484	0.462224	3.310005167	2.21E-05	0.300992832	0.420718
ENSG00000173702	-1.14487117	0.40319	-3.003257833	0.000131	-0.925958848	0.110307
ENSG00000230916	-0.771728124	0.458732	-3.413970532	0.000131	-0.581656502	0.449236
ENSG00000134668	0.45277879	0.537915	2.795070257	0.000131	0.358071435	0.43111
ENSG00000115386	-1.06170259	0.513752	-9.061935588	0.000245	-4.397275018	0.119342
ENSG00000015520	0.901573521	0.40319	2.645472663	0.000291	0.632647164	0.215697
ENSG00000102287	0.889185989	0.40319	1.55315816	0.000316	1.672818166	0.000279
ENSG00000127954	0.997932525	0.40319	2.001437486	0.000373	2.119968476	0.000282
ENSG00000198868	0.548597766	0.557744	3.147086396	0.000834	0.286986876	0.720211
ENSG00000163638	1.121776809	0.40319	2.433164875	0.001082	2.535730748	0.001562
ENSG00000137558	0.974169525	0.40319	2.135034039	0.002406	2.145450503	0.000229
ENSG00000240409	-0.137104675	0.901613	-3.304358852	0.002406	0.186648973	0.822344
ENSG00000188425	0.492651532	0.598424	3.407884522	0.002794	0.103523839	0.930844

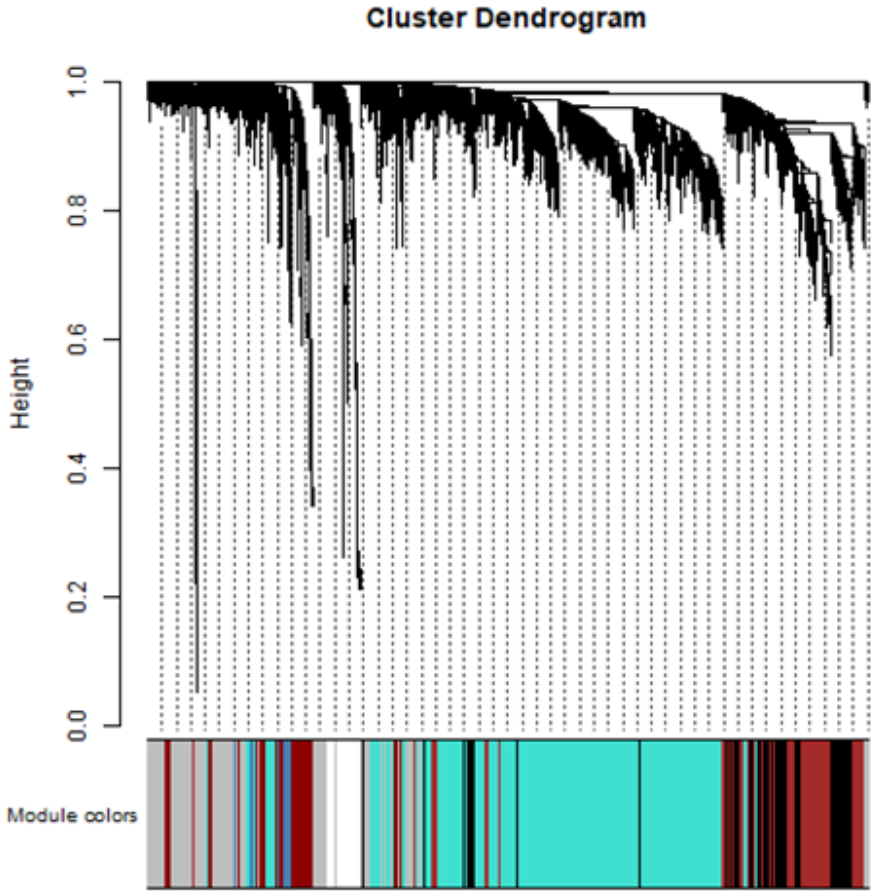

Figure 5. WGCNA; Hierarchical cluster analysis detection of the co-expression clusters determined by WGCNA.

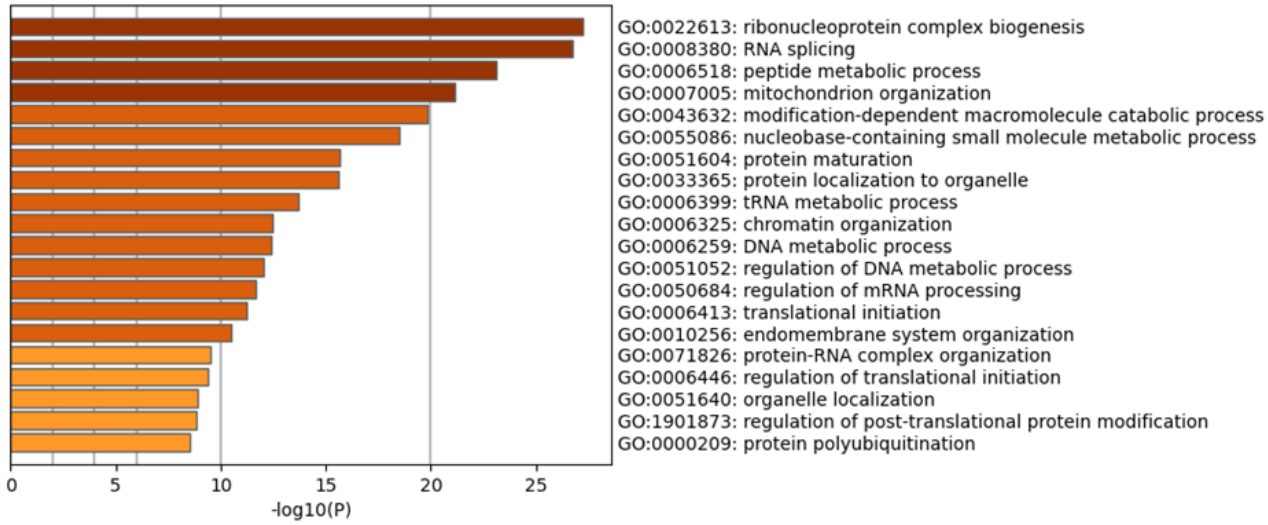

Figure 6. Gene Ontology Enrichment Analysis of the red module.

The output of MDS demonstrates unsupervised clustering of samples, revealing the similarities and dissimilarities between them. Figure 1 illustrates distinct groups within the bipolar disorder data. Additionally, the figure indicates that samples clustered closely together exhibit higher similarity, whereas far-apart samples have little or no similarity. PlotBCV (Figure 2) depicts the fitting of genes with varying expression levels to the model. BCV, an abbreviation for biological coefficients of variance, represents the actual variation between samples. When the model fits accurately, the distribution of tagwise (represented by black dots) will merge with the trend curve (shown in blue). Voom, an abbreviation for Variance modeling at the observation level, tackles the issue of representing discreteness and variability in RNA-Seq data, as shown in Figure 3. It transforms discrete data into continuous data, enabling statistical and differential expression analyses. By exploiting the diversity within gene count data, voom estimates the relationship between mean and variance for each gene, reshaping the count data to fit a normal distribution(Law, Chen, Shi, & Smyth, 2014). This alteration facilitates the employment of linear models for examining differential expression and conducting hypothesis testing. Figure 4 presents the comparison results of different analysis packages. It reveals variations among these analysis methods, albeit not substantial. Table 1 presents the findings of the differential expression analysis. The results obtained from the three differential expression analysis packages demonstrate a similar trend. The table displays the top 15 genes that are the most likely differential genes. The provided images exhibit the outcomes of WGCNA and GO ontology enrichment analysis. Figure 5 illustrates the results of gene cluster analysis through a dendrogram depicting clusters. Initially, the co-expression correlation coefficient is computed based on the measured levels of gene expression. Subsequently, genes are organized into clusters, and a gene tree is formulated utilizing Euclidean distance. The gene tree undergoes pruning using dynamic shearing, ultimately yielding gene modules. This strategy facilitates the analysis of a reduced number of gene modules rather than a vast number of individual genes. The cluster dendrogram's lower segment signifies each module's designated color. Following this, GO enrichment analysis is performed on each module to attain further insights into gene functionalities, as shown in Figure 6.

2. GBM results:

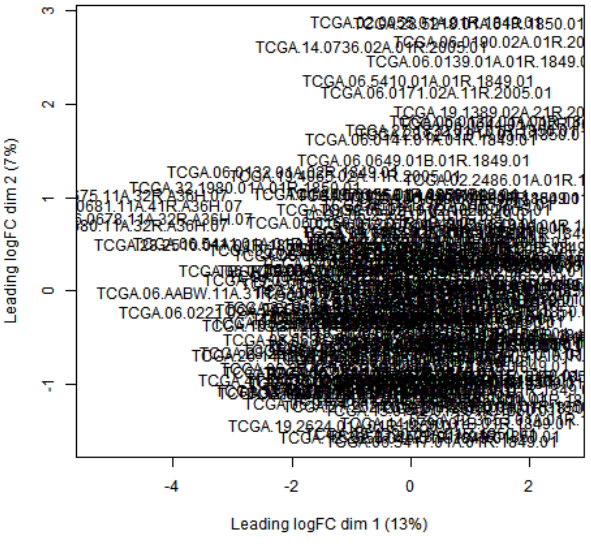

Figure 7. Plot MDS presents the grouping of GBM samples.

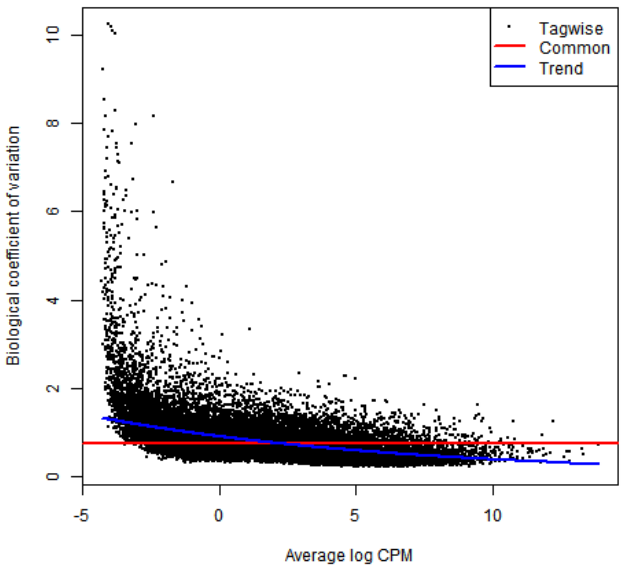

Figure 8. The plotBCV function is utilized to visualize fitting a gradient boosting machine (GBM) sample to the model, considering various expression levels.

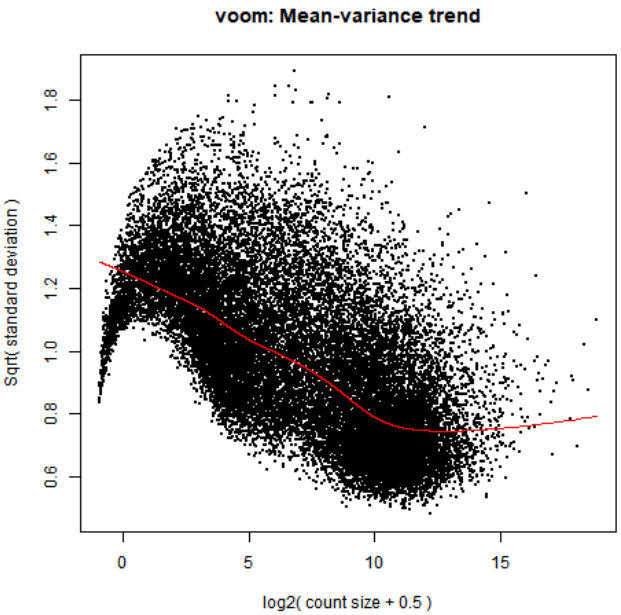

Figure 9. Variance modeling at the observational level.

Table 2. Top 15 genes from the analysis results of the three differential expression analysis packages for GBM samples.

	limma_logFC	adj. p	edgeR_logFC	FDR	DESeq2_logFC	padj
ENSG00000266489.3	-6.561120177	3.04E-37	-6.34851552	5.87E-16	-6.619032444	0.000119
ENSG00000285883.1	-5.857001039	3.25E-34	-6.02868803	3.13E-24	-6.354909923	3.86E-07
ENSG00000198883.12	-6.829918548	3.99E-33	-5.75780147	9.02E-38	-5.882468697	4.53E-11
ENSG00000233123.1	-7.607911803	3.99E-33	-6.75819341	9.92E-20	-6.880423218	2.97E-05
ENSG00000225110.3	-7.618416071	1.25E-31	-5.78979128	8.01E-23	-5.912617423	6.52E-07
ENSG00000204933.3	-5.369981578	1.25E-31	-5.51866413	9.27E-16	-5.842939527	7.41E-05
ENSG00000160396.9	-5.874017289	1.86E-31	-5.15703605	5.17E-38	-5.279541246	2.58E-12
ENSG00000177570.15	-4.519480933	2.24E-31	-4.21829844	4.43E-47	-4.341222509	1.67E-18
ENSG00000107864.15	-3.81344377	4.56E-31	-3.70115119	3.02E-47	-3.828007686	3.16E-21
ENSG00000196972.9	-4.208800211	1.80E-30	-4.05257957	7.09E-45	-4.179960574	2.09E-18
ENSG00000113319.13	-4.139011536	2.87E-30	-3.8787623	1.12E-43	-4.000206055	2.61E-18
ENSG00000033122.21	-5.050851471	5.06E-30	-4.51434709	1.12E-38	-4.639514817	3.23E-14
ENSG00000198785.7	-4.894709344	8.49E-30	-4.5603169	1.08E-38	-4.677548486	4.95E-14
ENSG00000144550.13	-4.563484946	1.43E-29	-4.39112197	8.86E-44	-4.519993664	9.53E-17

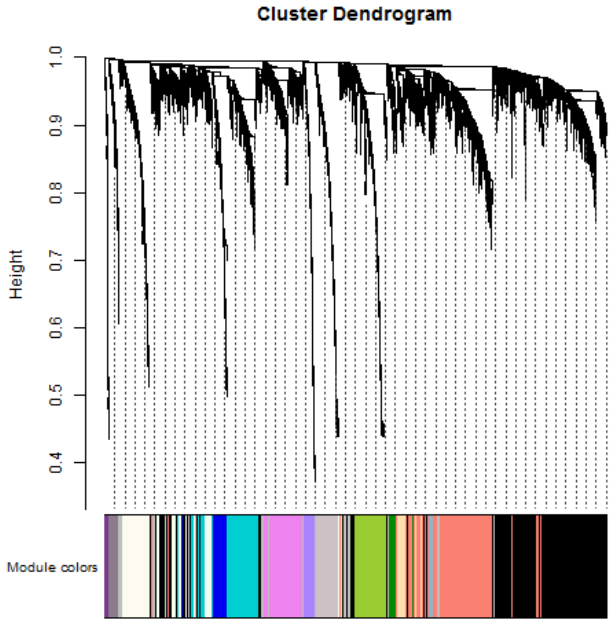

Figure 10. Cluster dendrogram of GBM brain expression presents the relationship between different genes. Close-related genes are clustered into one group where each color represents a group.

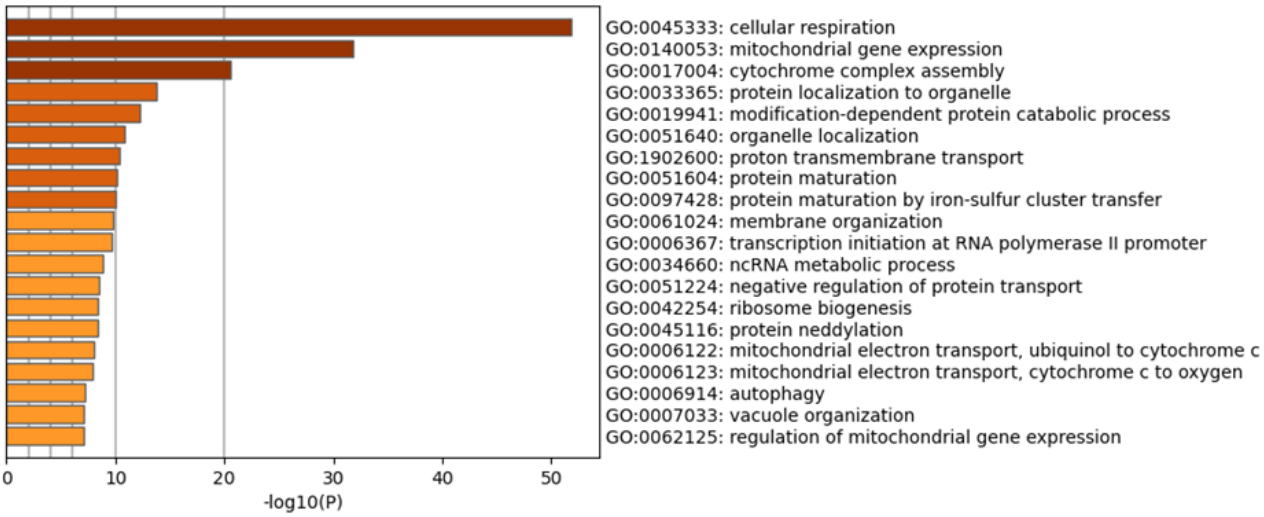

Figure 11. Gene Ontology Enrichment Analysis of the pink module.

The output presents the results of the differential expression analysis, WGCNA, and GO enrichment analysis of GBM data. Figures 7, 8, and 9 display MDS, plotBCV, and voom plots, respectively. These figures indicate that the GBM data exhibits grouping tendencies. Figure 7 illustrates that the majority of samples are clustered together. In Figure 9, most of the data is concentrated around the red line, while the rest is scattered. Table 2 lists the top 15 genes likely to be differential genes. Figure 10 presents the outcome of the Weighted Gene Co-expression Network Analysis (WGCNA). The clustering dendrogram illustrates the relationship between genes and assigns genes with high correlation and weight to a specific module. Each module is represented by a different color, indicating that genes within a module have a stronger association with each other. Then, by performing a go enrichment analysis on the pink module, we can know what the function of each gene is (see Figure 11).

Discussion

By analyzing the results of the differential expression

analysis of BP and GBM, we can observe that both the BP and GBM data are available. Figure 1 demonstrates clear groupings in the BP data, distinguishing between controls and BP patients. Furthermore, it reveals that the clustered data exhibit high similarity. Figure 2 (plotBCV) is primarily used to assess the gene expression stability and the data quality in these samples. The plot indicates that the BCV value is small, indicating relatively stable gene expression among these samples and ensuring high-quality and reliable data. Similarly, the same principles apply to GBM data samples. These samples are grouped and have a higher quality compared to other data. They exhibit greater consistency and contain reliable rows.

In the outcomes of both GO enrichment analyses for BP and GBM, we notice the existence of GO:0033365. This specific term denotes the process of protein localization to organelle, which mainly regulates the arrangement of proteins within cells. The subcellular localization of proteins holds immense importance as it dictates their functionality and provides a physiological framework(Hung & Link, 2011). Faulty protein subcellular localization is associated with various diseases. Genetic mutations and abnormal expression of cargo proteins or transport receptors may cause deviations in protein localization, eventually resulting in human disorders(Hung & Link, 2011).

Another feature that BP and GBM share is their organelle localization (GO: 0051640), which is the same as saying that their proteins are localized in organelles. It is necessary for organelles' proper functioning and continued existence that proteins are positioned correctly within them. When proteins enter organelles to which they do not belong, they have the potential to interfere with the function of other proteins, which could ultimately result in the death of the cell. Additionally, improper localization of proteins can affect signaling pathways, leading to inconsistencies in signaling and disrupting the communication that occurs within cells.

The maturation of proteins is the last common genetic trait, referred to as GO:0051604 in the gene ontology. Protein maturation is extremely important to proteins' functioning, regulation, and overall health. It is the process by which a protein moves from a synthetic state to a functional one, where it can exert its biological functions. It is referred to as post-translational modification. This process involves folding, modifying, and activating the protein to ensure that it can properly carry out the function for which it was designed(Saraogi & Shan, 2014).

All of the genes that we have discussed above are connected to the malfunctioning of protein function. I hypothesize that these genes could be singled out for treatment in the case of BP or GBM. It will be vital in the creation of new drugs and research into diseases to pay attention to these genes in the future. This will allow for the prevention of protein function failure or mutation, which has the potential to lead to the treatment of a wide variety of human disorders.

Reference

[1] Agarwal, A., Koppstein, D., Rozowsky, J., Sboner, A., Habegger, L., Hillier, L. W., . . . Gerstein, M. (2010). Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays. BMC Genomics, 11(1), 1-16.

[2] Bowling KM, R. R., Lasseigne BN, Cooper SJ, Myers RM. (Jun 22, 2017). RNA-sequencing of human post-mortem brain tissues. Retrieved from: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE80655

[3] Grande, I., Berk, M., Birmaher, B., & Vieta, E. (2016). Bipolar disorder. The Lancet, 387(10027), 1561-1572. doi:10.1016/S0140-6736(15)00241-X

[4] Hung, M.-C., & Link, W. (2011). Protein localization in disease and therapy. Journal of cell science, 124(20), 3381-3392.

[5] Langfelder, P., & Horvath, S. (2008). WGCNA is an R package for weighted correlation network analysis. BMC Bioinformatics, 9(1), 1-13.

[6] Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome biology, 15(2), 1-17.

[7] Marguerat, S., & Bähler, J. (2010). RNA-seq: from technology to biology. Cellular and molecular life sciences, 67, 569-579.

[8] NIH. TCGA-GBM [Gene Level Copy Number]. Retrieved from: https://portal.gdc.cancer.gov/repository

[9] Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. bioinformatics, 26(1), 139-140.

[10] Saraogi, I., & Shan, S.-o. (2014). Co-translational protein targeting the bacterial membrane. Biochimica et Biophysica Acta (BBA)-Molecular Cell Research, 1843(8), 1433-1441.

[11] Zhang, B., & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. Statistical applications in genetics and molecular biology, 4(1).