Journal of Advanced Zoology

ISSN: 0253-7214
*Volume **44** Issue **04 Year 2023** Page **105:124***

Weighted Gene Co-expression Network Analysis of Glioblastoma Gene Expression Microarray Data

G. Madhusudhan[1], P. Sujatha[2], T. Geetharathan[3]

[1]Department of Biotechnology, Sri Venkateshwara University, Tirupati -517502
[1]Department of Microbiology & Biotechnology, Bharath Institute of Higher Education and Research, Selaiyur, Chennai - 600100.

**Corresponding author's: G. Madhusudhan*

Article History

Received: 06 June 2023
Revised: 05 Sept 2023
Accepted: 02 Nov 2023

Abstract

Glioblastoma is a highly aggressive and lethal form of brain cancer characterized by its complex molecular landscape. Understanding the underlying gene expression patterns and their relationships is essential for unraveling the mechanisms driving this disease. In this study, we conducted a Weighted Gene Co-expression Network Analysis (WGCNA) on Glioblastoma gene expression microarray data to identify co-expressed gene modules and potential key regulatory genes associated with the disease. Utilizing a comprehensive dataset of Glioblastoma samples, we performed quality control and preprocessing to ensure the reliability of the data. WGCNA was employed to construct a weighted gene co-expression network, enabling the identification of modules of co-expressed genes. The correlation between these modules and clinical characteristics such as patient survival, tumor grade, and other relevant factors was assessed. Additionally, we conducted functional enrichment analysis to gain insights into the biological processes and pathways associated with the identified gene modules. Our findings revealed distinct gene modules associated with Glioblastoma progression and patient outcomes. Notably, we identified key hub genes within these modules, which may serve as potential biomarkers or therapeutic targets. Furthermore, functional enrichment analysis provided a comprehensive understanding of the biological processes and pathways influenced by these co-expressed gene modules. In conclusion, our Weighted Gene Co-expression Network analysis of Glioblastoma gene expression microarray data has shed light on the complex gene interactions and regulatory networks underlying this aggressive brain cancer. This knowledge may ultimately contribute to the development of novel diagnostic and therapeutic strategies, improving the prognosis for Glioblastoma patients..

Keywords: *Glioblastoma, Weighted Gene Co-Expression, Potential Biomarkers, Diagnostic And Therapeutic Strategies*

1. Introduction

Glioblastoma, also known as Glioblastoma Multiforme (GBM) is an aggressive type of cancer most commonly occurred in supratentorial region (frontal, temporal-parietal, and occipital lobes) and also rarely located in cerebellum and very rare in the spinal cord(1). It forms from cells called astrocytes that support nerve cells and spreads easily within months(2,3) and one of the most aggressive, invasive, and undifferentiated type of tumor which eventually results death within a short period of time. The median survival rate from the time of diagnosis was just 15 months(4,5). And 90% of glioblastoma multiforme cases develop from glial cells by multistep tumorigenesis and remaining 10% from secondary neoplasm, takes 4-5 years to develop(6). GBM, which develops as a fresh or new grows within 3 months with symptoms such as headaches, nausea, and symptoms similar to stroke and further during the progress of the disease may cause unconsciousness(7,8). The incidence of central nervous system (CNS) tumors in India ranges from 5 to 10 per 0.1 million populations with an increasing trend and accounts for 2% malignancies. This is usually from the hospitalized data of a neuro-oncology, where the majority of tumors are high-grade Glioma's accounting for 59.5% of all CNS tumors with higher urbanized male population (9). Overall, the incidence rate in males was 1.62 times higher than females across the globe(10). In the case of the United States, the incidence rate of

glioblastoma in elderly patients was 13.16 per 0.1 million population. The highest incidence of brain cancer has been observed in Australia, North America and Northern Europe in both sexes. The highest mortality rates have been reported in China, United States, India, Brazil and Russia. And in European countries, the highest incidence rate was reported in England, Ireland and Northern Europe and the lowest in Eastern Europe. And according to the present study, the incidence and mortality rates for brain cancer are higher in countries with a higher Human Development Index (HDI). The higher incidence rate in countries with a higher HDI can be explained in terms of environmental pollutants and occupational exposures of ionizing radiation and industrial radioactive sources in these countries. According to recent discoveries, GBM has been subdivided based on the mutational state of isocitrate dehydrogenase (IDH) genes. And IDH-mutant GBMs are distinct from GBMs without IDH1/2 mutation with respect to molecular and clinical features, including prognosis(11). Overall, there are two distinct subdivided types of GBMs, one is IDH-wildtype and the other is IDH-mutant glioblastoma. In addition, recent findings in pediatric GBMs regarding mutations in the histone H3F3A gene suggest that these tumors may represent a 3rd major category of GBM, separate from adult primary (IDH1/2 weighted), and secondary (IDH1/2 mutant) GBMs(12).

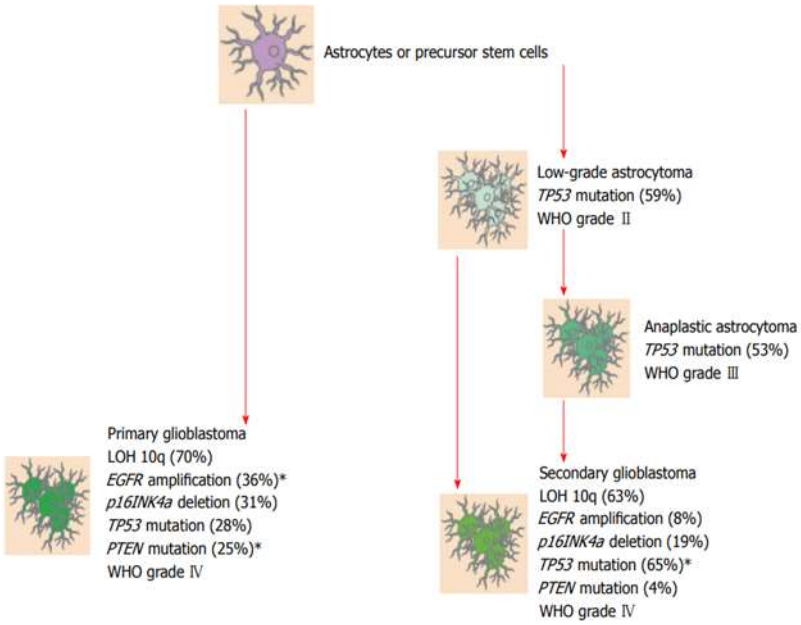


Figure 1: Genetic mutation pathways implicated in the development of malignant Gliomas

The genes IDH1 and IDH2 are molecular markers that demonstrate prognostic value in patients with glioblastomas as well as lower-grade gliomas. Isocitrate dehydrogenase (encoded by IDH1 in the cytoplasm and by IDH2 in the mitochondria) in its wild-type form produces alpha-ketoglutarate. Mutations in these genes encode an aberrant enzyme that turns alpha-ketoglutarate into an onco-metabolite, D-2 hydroxyglutarate. D-2 hydroxyglutarate controls the oncogenicity of IDH mutations. Based upon mutation status, gliomas may be classified as IDH-wild-type or IDH-mutant. IDH-wild-type gliomas include gradeIpilocytic astrocytoma and primary GBMs. Tumorigenesis in this case is independent of the IDH status and is mediated by other oncogenes. IDH-mutant gliomas include grade II and grade III gliomas as well as some secondary GBMs. And the interesting aspect is IDH mutants carry a better prognosis than IDH wild types. For example, in WHO class IV tumor, secondary GBMs (IDH mutants) carry a better prognosis than primary GBMs (IDH wildtypes)(13).

And regarding the treatment of Glioblastoma, present treatment strategies Cannot cure GBM patients completely but only extend their overall survival. Even though using chemoradiation, immunotherapy, and radiosensitizers as adjuvant therapy cannot reduce the high rates of recurrence of GBM within few months after treatment. At present lot of research activities going on regarding the development of therapies to completely cure glioblastoma. But up to now, Radiotherapy has been the backbone of the treatment(14). As of today, the treatment options available for Glioblastoma are surgery, Radiotherapy, Combined Chemo-Radiotherapy, Monoclonal Antibodies, Immunotherapy (Check-point Inhibitors, Peptide Vaccination, Adoptive cell therapy, Viral Immunology, Dendritic cells), and alternating electric field therapy(15).

Understanding the mechanisms of glioblastoma at the molecular and structural level is valuable for clinical treatment. Bioinformatics can be effectively used to analyze GBM microarray data to provide theoretical reference for further exploration of tumorigenesis mechanism and help search for potential

target genes(16). Based on bioinformatics study, some differentially expressed genes (DEGs) such as Transforming Growth Factor Beta Induced (TGFBI) and SRY-Box 4 (SOX4) were explored as the potential therapy targets for GBM. Co-expression analysis has emerged as a powerful technique for obtaining novel insights into complex mechanisms and multigene analysis of large-scale data sets, especially for identifying functional modules. As an approach of bioinformatics study, weighted gene co-expression network analysis (WGCNA) is commonly used for revealing the correlation between genes in different samples(17). Previous WGCNA shows the epigenetic events in GBM development and prognosis based on The Cancer Genome Atlas (TCGA) database. Thus, WGCNA can be used to predict genes associated with cancer development(18).Thus, the rapid development of high throughput gene expression profiling technology such as microarray and high throughput sequencing has enabled the development of many new bioinformatics data analysis methods for identifying disease related genes, characterizing disease subtypes and discovering gene signatures for disease prognosis and treatment prediction.

Different Platforms and Technologies Involved in The Project

The technologies involved in this project are microarray technology (microarray's gene expression data of glioblastoma patients been used), R language (it is a statistical analyzing language used for analyzing data), R studio (it's the IDE (Integrated Development Environment), and GUI (Graphical User Interface) of R language. And WGCNA (Weighted Gene Co-expression Network Analysis) package is been widely used, which was develop by Steve Horvath. And Cytoscape has also been used to visualize the eigengene module we got.

About DNA Microarray Technology

DNA microarrays can simultaneously measure the expression level of thousands of genes within a particular mRNA sample. This technology of expression profiling was widely used to identify diagnostic or prognostic biomarkers, to classify diseases like to differentiate tumors that are indistinguishable by microscopic examination, to monitor the response to therapy, and finally to understand the mechanisms involved in the genesis of disease processes. So, thus DNA microarrays are considered important tools for discovery in clinical medicine(19). Once the microarrays have been hybridized the resulting images are used to generate a dataset. After this the dataset is preprocessed, it been prepared for the application of data analysis methods. And further, we check the quality of data and make it ready for analysis. Here the microarray data which we used in this project belongs to the HT-HG-U133A platform.

R language

And in the case of the R language, which has been used for statistical computing and graphics. R provides a wide variety of statistical and graphical techniques and they are highly extensible. One of the key strengths of R by using it we can able to generate a wide variety of plots and also usage of packages that are developed for specific analysis. And also, one of the advantages is a large number of users work together to develop new packages and implement the latest computer technologies. And R can follow the latest trends faster than comparative software. The latest developments are being incorporated into new packages very quickly and the core code is updated on average twice a year. At the moment, the R programming environment contains the most extensive range of tools for parallel computing, machine, and deep learning, and for working with Big Data sets for genomic analysis(20).

R Studio

RStudio is an Integrated Development Environment (IDE) for the R language. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, debugging, and workspace management. It provides a graphical user interface to interact with the R language.

Weighted Gene co-expression Network (WGCNA)

Weighted gene co-expression network analysis (WGCNA) package was originally developed by Steve Horvath for analyzing gene co-expression from microarray data. The main aim of WGCNA is to find biologically more interesting modules. Gene co-expression networks are increasingly used to explore the system-level functionality of genes. In the network, the nodes represent genes, and nodes are connected if the corresponding genes are significantly co-expressed across appropriately chosen tissue samples(21). The purpose of analyzing gene expression data with WGCNA package with R is to identify biologically meaningful modules. And to find those prominent hub gens in those modules.

2. Materials And Methods

This bibliographic review was carried out based on the analysis of scientific articles, collected from high-impact databases such as: Doaj, Pubmed, Medline, Science Direct, Researchgate, Scielo, in a systematic way with a focus on the study variables that are dental surgical treatments (dependent

variable) in patients with chronic renal failure (independent variable), from the last 6 years from the date, so the study period was defined from 2016 to 2021 inclusive.

Gathered all this information, a large study was conducted, pointing out and extracting the highlights of all the authors on the subject of research, analysis and conclusions about dental management focused on patients with chronic renal failure.

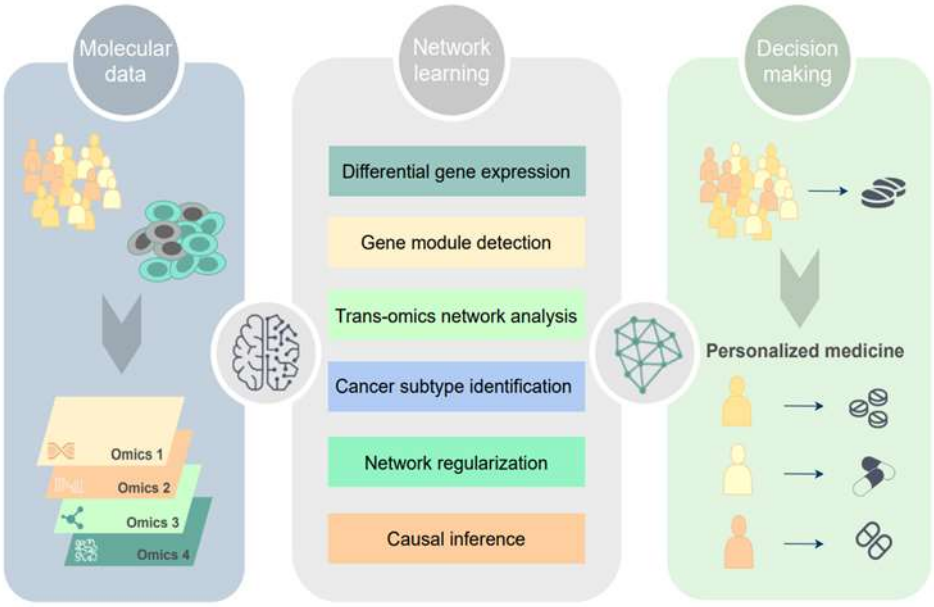


Figure 6: Overview of current strategies in precision medicine for network learning in Glioblastoma

Thus, the usage of Network science has been well-established and prominent across biological domains. By using biological data identifying the crucial metabolic pathways, subtypes and many more is a crucial part of research and it will progress the mode of understanding the disease. Whereas in case of glioblastoma, the progress of discovering curing methods was slow, thus by combing the network science to analyze the huge expression datasets will help in the mode of understanding the depts of the disease. GBM networks are essential to understand the information flow and better informing drug development and pre-clinical studies. Overall, all this analysis will help to design personalized medicine. These are all first step of Precision Oncology(22).

Cytoscape

And also, the Cytoscape tool is used to analyze the module eigengenes and to detect hub genes in those modules and their interactions. It's an open-source bioinformatics software platform for visualizing molecular interaction networks. By this software, we analyze interactions between modules we identified and further detect hub genes(23).

3. Results and Discussion

About Data and Microarray Technology

The data which we took to analyze with R package WGCNA (Weighted Gene Co-expression Network Analysis) was the Microarray's gene expression data of patients suffering with glioblastoma. Which contains of 215 patient's data with 12042 gene expression levels. The Microarray data which we used in the analysis belongs to HT-HG-U133A platform (that's one of the technologies of microarray's chip). And the database of the data was TCGA (The Cancer Genome Atlas), whereas it is the collection of expression data from different cancers. The TCGA was a project, begun in 2005 to catalogue genetic mutations responsible for cancer. Whereas TCGA was one of the high throughput genome analysis techniques to improve the ability to diagnose, treat and prevent cancer through a better understanding of the genetic basis of the disease. Microarray is a multiplex lab-on-a-chip. It is a two-dimensional array on a solid substrate usually a glass slide or silicon thin-film cell that assays large amounts of biological material using high throughput screening miniatured, multiplexed and parallel processing and detection methods. Whereas with the help of this microarray gene expression data we further analyze with the help of R studio, which is a part of R-program (it's a statistical language used worldwide to analyze many things.

Installing R and R Studio

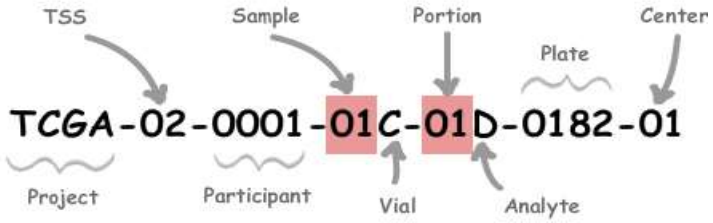
At first install R to deal with the data, by downloading R and R studio (which acts as IDE (interactive development environment) for R program), whereas R studio consists of better graphical user interface to interact with R language and process our work. So, after successfully installing R and R studio we process our gene expression data with the help of R packages.

Installing Required Packages

The packages where we installed for this project are WGCNA (Weighted Gene Co-expression Network Analysis), Which is required to analyze the data completely and make a cluster and further to visualize the data. The WGCNA was a crucial package for this project, whereas the whole project runs on the basis of this Package. And further we installed packages like SNF Tool (Similarity Network Fusion Tool), RcolorBrewer and dhga. Whereas we use SNF Tool for the purpose of standard normalization of our gene expression data. And RcolorBrewer for the purpose of visualize heatmaps in more colorful way. And dhga package to identify hub genes in the gene expression data. So those where the packages we used for analysis of our project.

Loading Data and Further Analysis Using R Studio

At first, we load the gene expression data using delim function, which the loaded data consists of 215 columns with of 12042 entries, here the columns are patient ids and rows where the expression data of genes. Usually, the data we loaded consists of TCGA barcodes, because it's a TCGA data from that respective project by NIH (national institute of health). So, we going to cut the long barcode to some extent up to vial portion for better visualization purposes.



TCGA barcode example By using R code in R studio, we cut the portion, analyte, plate and center part of the TCGA barcode.

Next, we check out whether the data we loaded was normalized or not by plotting a boxplot. Whereas the boxplot was one of the ways to graphically depict the loaded gene expression data to check whether the data is normalized or not. After checking we got to know that the data is not normalized. So, we go for doing standard normalization and quantile normalization. Whereas by doing standard normalization process the data seems somewhat normalized, better than before. But after doing quantile normalization the data been completely normalized and it is ready for further steps. And also, at every stage of data processing we plotted a heatmap to check out the variance exhibited by data. Whereas we plotted heatmap for unnormalized data and after standard normalization and also after quantile normalization of data. And observed the level of noise in the heatmaps been reduced after quantile normalization. Where the data processing been completed and ready for WGCNA analysis.

Whereas from this point the WGCNA (weighted gene co-expression network analysis) package been used to analyze the data. And at first, we choose soft-threshold power to check mean connectivity of genes. Whereas we check by plotting a connectivity plot for choosing better soft-threshold power. This a was the crucial step which the whole results will alter and biologically meaningful modules/clusters will not come if we chose wrong soft-threshold power. And also, in the connectivity plot we check the R^2 value, which should be greater than 0.8 (which was the criteria to eligible scale-free Topology). Whereas the scale-free Topology entails that the presence of hub nodes/genes that are connected to a large number of other nodes.

At next step we transform the gene expression data to an adjacency matrix, whereas the adjacency matrix was a symmetric square matrix used to represent the data. And further we transform the adjacency matrix to a topological overlap matrix. And also, typically we prefer signed network because at last we end up with biologically meaningful modules. So, after obtaining the topological overlap matrix, we calculate dissTOM (dissimilarity of Topological Overlap Matrix). Then further we cluster the Gene expression data with hclust (hierarchical clustering) function. In hierarchical clustering we see the connectivity between genes in the Gene expression data. And after hierarchical clustering we constructed the modules/clusters of gene expression data, based on the expression

levels. And further we plotted the modules we detected, whereas based on the size of the module the colors been organized automatically.

Next, we visualize the network we constructed by some modes of visualizing methods. At first, we cut the cluster tree, but the real challenge is at what height we have to cut-off to result biologically meaningful things. As we know large height values lead to big modules and small values lead to tight modules. Whereas in reality we should use different thresholds to see how robust the findings are. And next we plot the TOMplot, which was the one of important thing used for visualizing the networks we created. And further we plotted classical multi-dimensional plot (MDS plot). Whereas the multi-dimensional scaling plot was one of the visualizing tools of network, where it takes the input data of dissimilarity measure and translates in to Euclidean distances.

And further we do heatmaps of prominent modules which we detected and biologically meaningful. Here we done heatmaps for Turquoise, blue, brown, pink, black, purple, red, yellow, and grey modules we detected and believed as biologically important. And visualized the mode of expression of genes in each module based on color intensities. And next we visualize a cluster tree based on module eigengenes of modules. Whereas eigengene is not a real gene, it's just a weighted average of all module genes. These allow one to relate modules to each other and also to determine whether modules should be merged or to define eigengene networks. And further the relation between module eigengenes been studied, which depicts how modules been related with each other through eigengene values. And then we exported those module eigengenes which has most significant value to cytoscape format. And from cytoscape software we analyzed those networks to detect the hub genes.

After loading data (Gene expression data consists of 215 columns with of 12042 entries) by using delim function, at first, we check the quality of the data. So, to check the quality of the data we plotted boxplots and heatmaps after numerous Normalization processes. Boxplots are which used for the purpose of visualizing quality of the data. It displays how the data is present in the dataset and amount of noise in it.

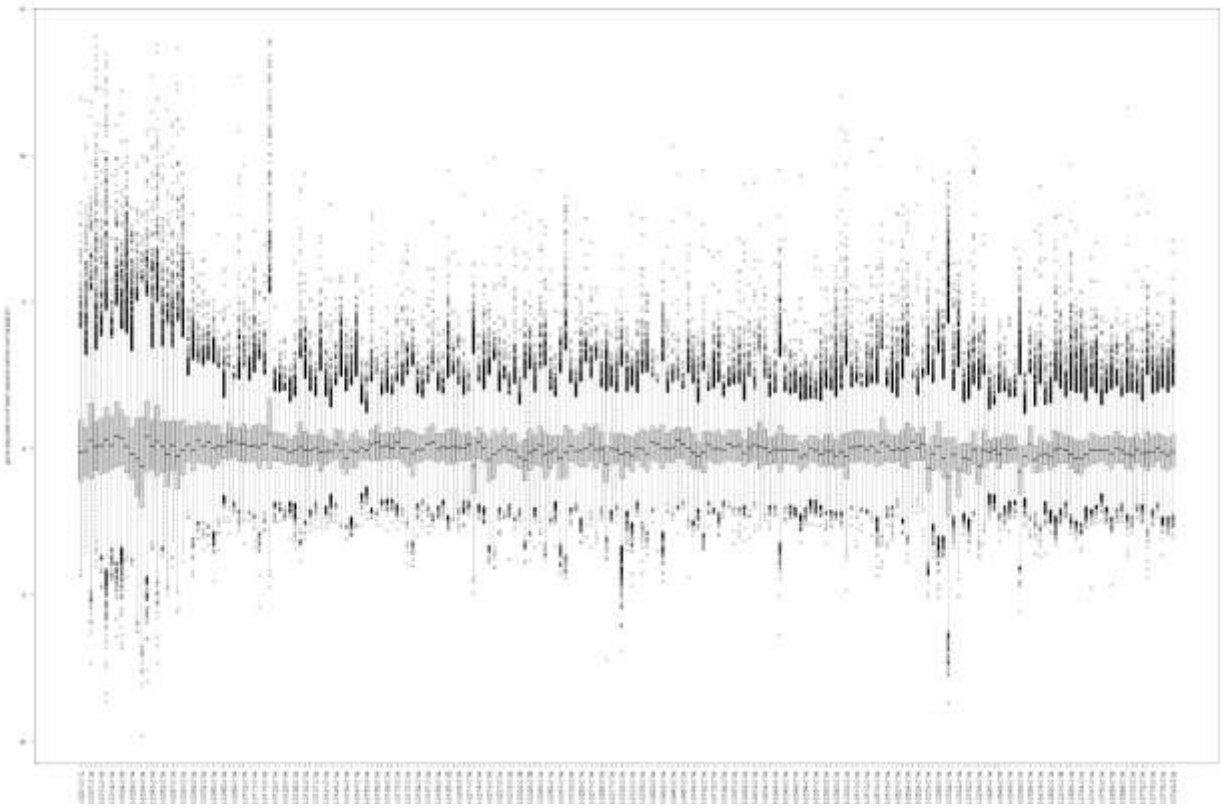


Figure 1: Boxplot before Normalization of gene expression data of glioblastoma multiforme

From the above figure the boxplot depicts that there is a huge noise in the data, which is clearly visualized from that boxplot. So, next we do both Standard Normalization and Quantile Normalization to reduce the noise of the data.

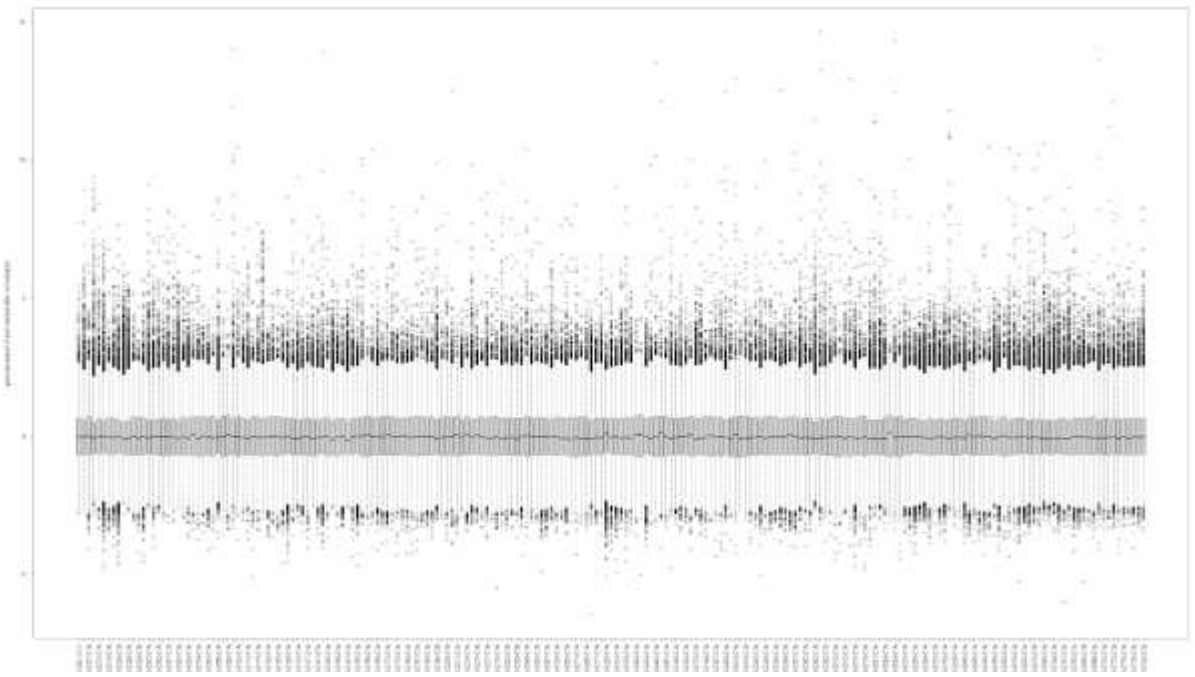


Figure 2: Boxplot after Standard Normalization of gene expression data of glioblastoma multiforme

From the above figure, still we can observe some noise in the data, but it is better regarding unnormalized dataset boxplot. So, further we go for Quantile Normalization of the Standard Normalized dataset. Quantile Normalization will disable all the noise in the dataset.

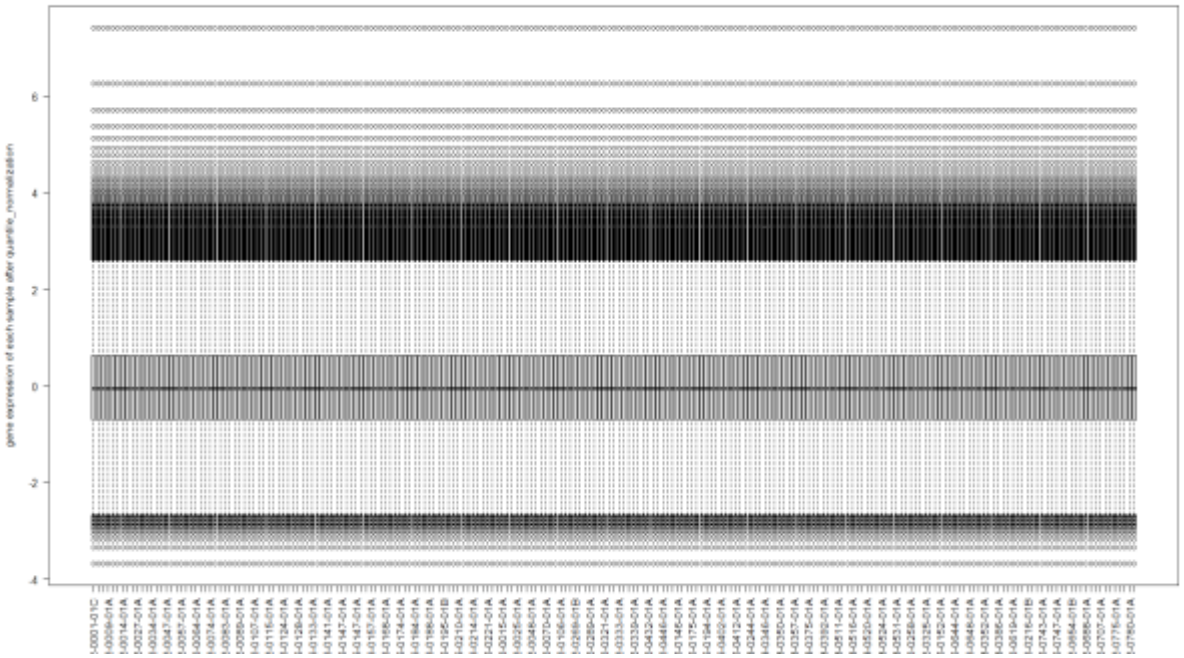


Figure 3: Boxplot after Quantile Normalization of gene expression data of glioblastoma multiforme .From the above boxplot of quantile normalized data, it is clear that the data is completely normalized. Hence, we can visualize flat lines in the boxplot. Now let's plot heatmaps of the gene expression data of glioblastoma at different levels of cleaning

.

4.4 heatmap of unnormalized gene expression data 4.5. heatmap of standard normalized

Gene expression data

Figure 4: heatmap of quantile normalized gene expression data of glioblastoma multiforme

Whereas in the above three heatmaps, we can clearly observe the amount of noise in the unnormalized data was more. And in the heatmap of quantile normalized one has noise least. And also, in these heatmaps we clearly observe the connections between genes been clustered and some areas of heatmaps been over expressed with much color intense. The heatmap is which depict adjacencies or topological overlaps with light colors denoting low adjacency(overlap) and darker color denotes higher adjacency(overlap).

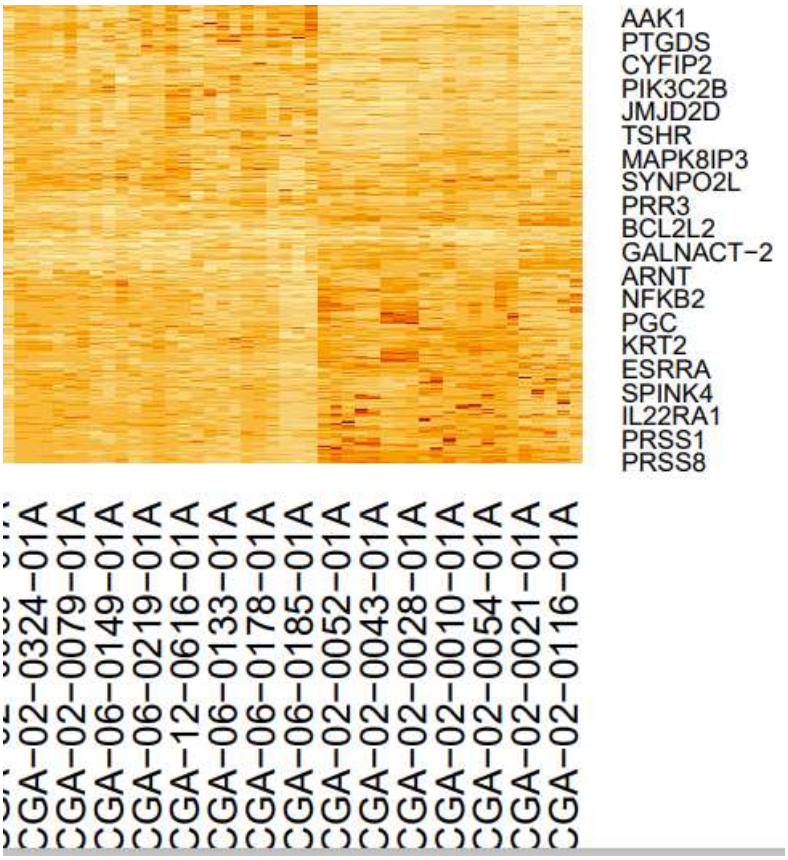


Figure 5: heatmap showing expression of some genes with dark red color The right side of the heatmap consists genes expressed in patients and below the TCGA ids where the patient ids.

Wgcna Analysis Results

From here by using WGCNA package to analyze the glioblastoma multiforme gene expression data

Choosing Soft-Threshold Power And Checking Mean Connectivity of Genes

Here we choose soft-thresholding power as 9 (We chosen power as 9 because as per fig 4.8 (connectivity plot for choosing better soft-thresholding power) R^2 should be greater than 0.8 (which indicates Scale-free Topology). Thus, for the power 9, R^2 value is satisfied and also mean connectivity plot been checked (low power the connectivity is high).

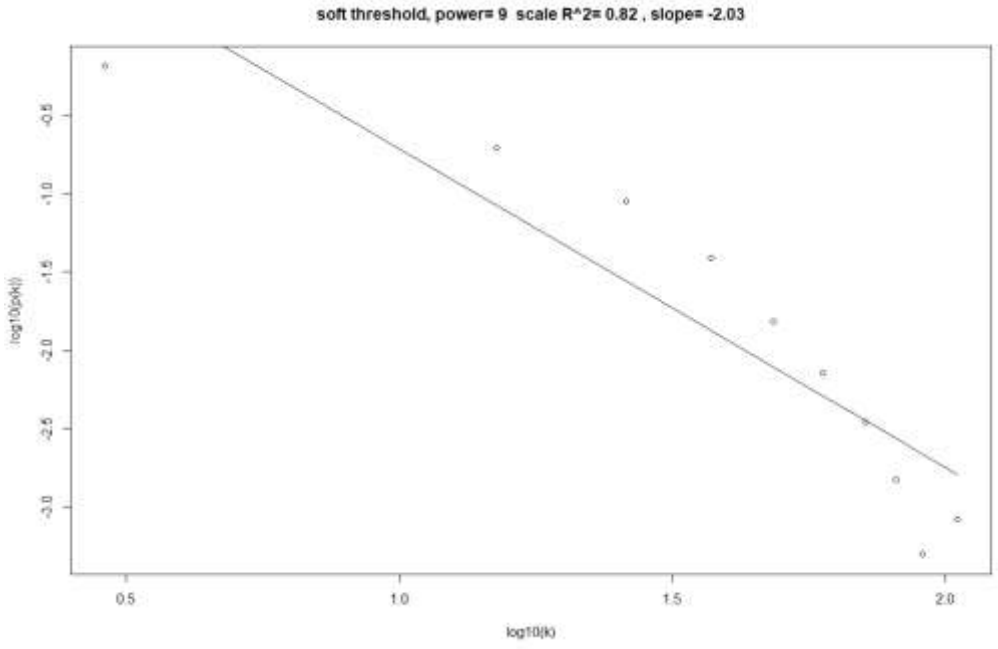


Figure 6: Connectivity plot for choosing better soft-threshold power

In the above connectivity plot the R^2 value was greater than 0.8, which signifies that it qualified Scale-free Topology criteria. Thus, the soft-threshold power been chosen as 9. Whereas choosing a best soft-threshold power was very crucial for detection of best biologically significant clusters/modules. There are two types of weighted correlational networks. unsigned network (note: default □□□power) =12, doesn't threat the same way it threats negative correlations. signed network (note: default □□(power) = 6, we generally prefer signed networks, we transform the correlation coefficient lies between 0 and 1. We prefer signed networks because we need modules which are biologically meaningful. Thus, if we want to go from a signed network to unsigned network we have.

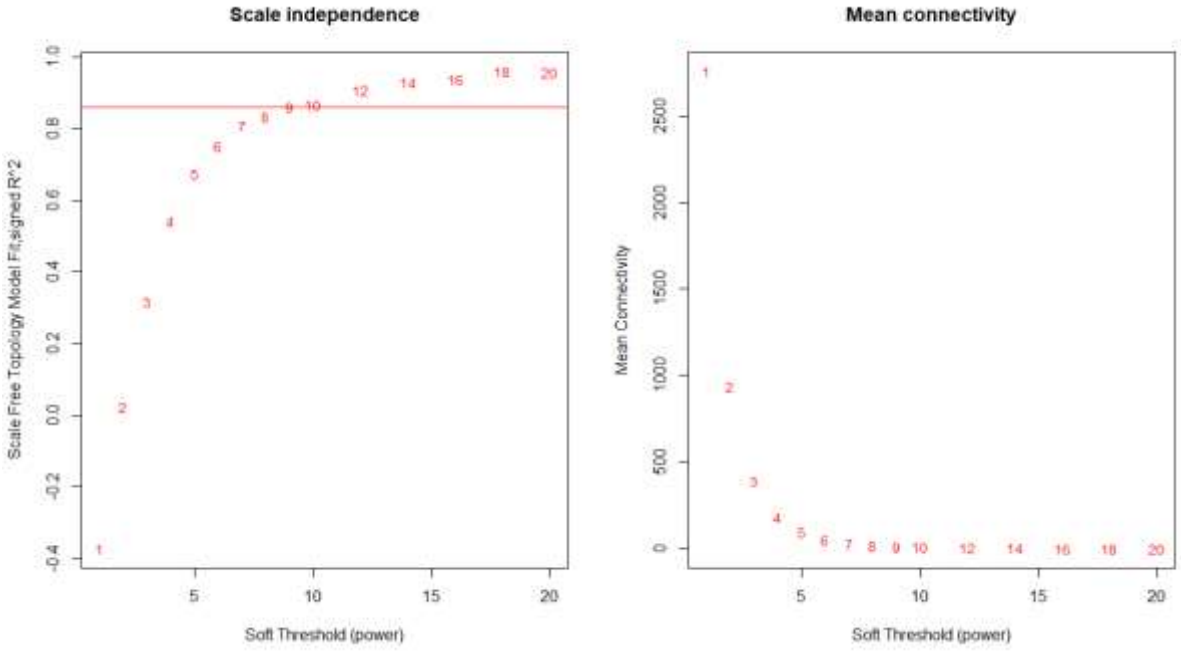


Figure 7: Plots of soft-threshold power and mean connectivity of gene expression data

Usually, the choice of type of network was depends on data analyst requirements, where we chose the signed network, because the correlation coefficient values lies between 0 and 1, and also we will get modules which are biologically meaningful, which is the requirement of our analysis. So signed network been chosen and power =9 been selected. Choosing a soft-threshold power was an important step, which determines the modules we get and the quality of modules/clusters we get. So, for better outcome of results, we have to choose correct threshold power, so we get better modules which are useful in the coming steps.

Converting To Adjacency Matrix And Topological Overlap Matrix And Further Disstom

In this step, we transform the dataset in to adjacency matrix, and further we transform adjacency matrix in to Topological matrix. And further to measure the dissimilarity in the network we transform TOM (Topological overlap matrix) in to dissTOM (dissimilarity Topological overlap matrix).

Clustering The Gene Expression Data With Hclust (Hierarchical Clustering) Function

In this step, after calculating the dissimilarity topological overlap matrix, we further cluster the genes through hierarchical clustering to construct modules further.

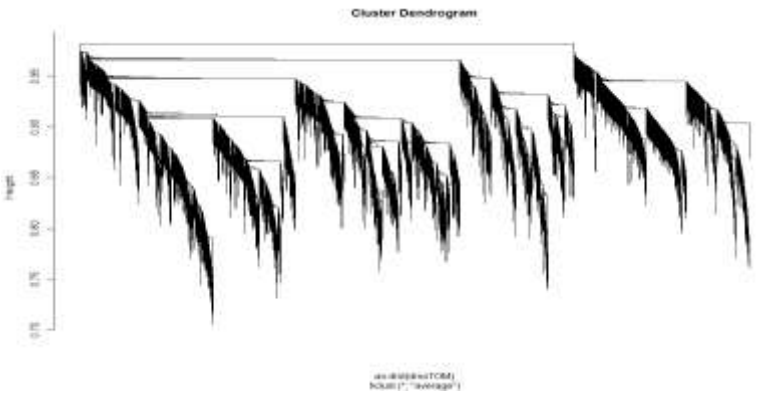


Figure 8: Hierarchical clustering of gene expression data of glioblastoma multiforme based on dissTOM (dissimilarity of topological overlap matrix)

In case of Hierarchical clustering, we use an algorithm that groups similar genes based on dissimilarity topological overlap matrix. Whereas each cluster is distinct from each other, but the objects within each cluster are broadly similar to each other. The data we used was a dissTOM which is generated from Topological overlap matrix. Hierarchical clustering starts by treating each observation as a separate cluster. And further it identify the two clusters that are closest together and merge the two most similar clusters. This process continues until all the clusters are merged together. In the above figure on the Y-axis the heights where the branch heights of clusters and in x-axis indicates that clustering based on dissTOM matrix.

Construction And Plotting The Modules of Gene Expression Data

After clustering the genes based on dissTOM matrix, we further construct those in to modules and plot them to visualize them clearly. A total of 27 modules been detected in our analysis.

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
456	2123	1841	1193	906	812	526	493	481	417	409	406	374	338	230	198	149	109	94
19	20	21	22	23	24	25	26	27										
75	74	67	60	52	49	44	35	31										

Those where the 27 modules detected, with minimum module size of 31 and maximum 2123. But for the further analysis we have to check which modules which efficient for our biological understanding. And further we plot these modules in a better visualized manner with differentiating colors based on module size. In each module the genes which are closely related will aggregate, but between each modules there will be no relation. It's almost like a clustering procedure we done before.

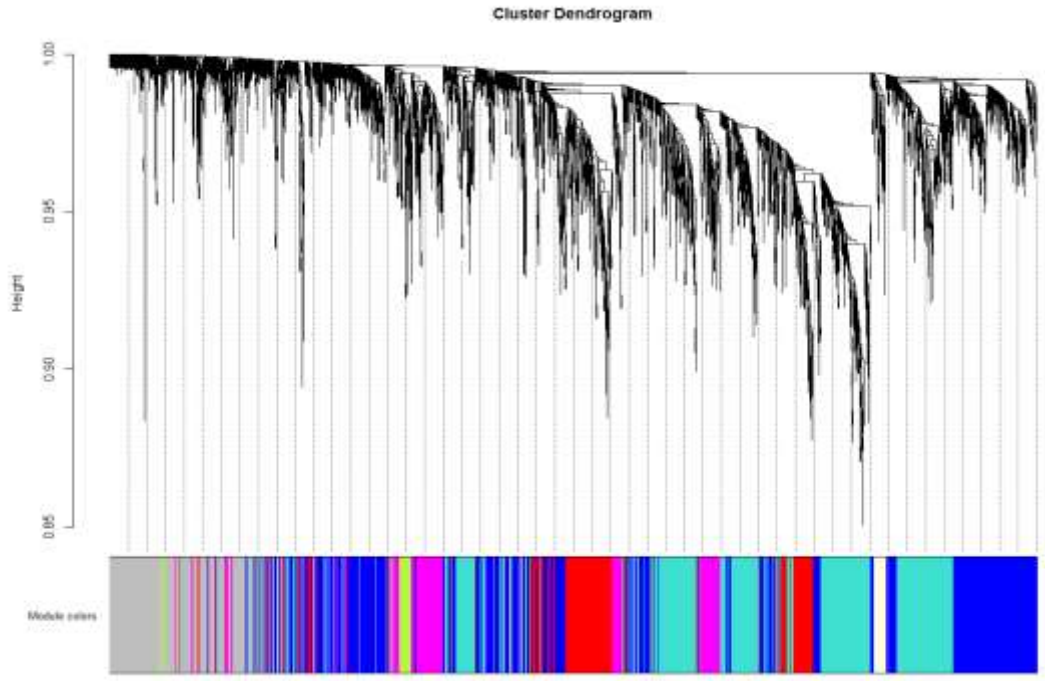


Figure 9: Cluster Dendrogram of detected modules of gene expression data

Next We Cut The Cluster Dendrogram At Branch Cut Height At 0.94 To Obtain Biological Meaningful Modules

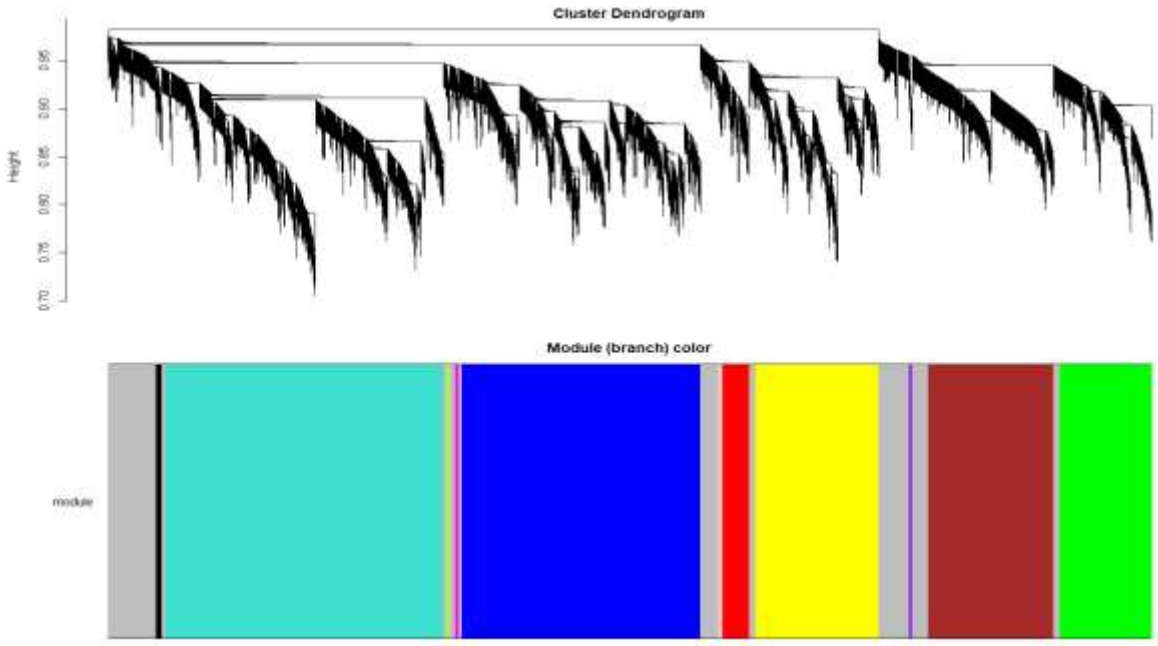


Figure 10: Cluster dendrogram with module colors after branch height cut at 0.94

In the above figure we used the function cutreeStaticColor, which generally cuts the cluster dendrogram at particular height (we assigned to cut at 0.94) and assigns colors according to size of the modules detected. GREY IS RESERVED to colour genes that are not part of any module. We only consider modules that contain at least 30 genes. We known that modules correspond to branches of the tree, but we cut-off at particular height because to obtain biologically meaningful modules. Because we dealing with a huge dataset of 12042 genes, in that we need those genes clustered and which are biologically important. The colours are assigned based on module size. Turquoise (325 genes) (others refer to it as cyan) colours the largest module, next comes blue (161 genes), next brown (146 genes), etc. Let's look out which module colour and size of the modules.

```
> table(colorh1)
colorh1
     black       blue      brown      green greenyellow       grey    magenta       pink
        68       2749       1439       1058         34       1642         36         53
    purple        red  turquoise     yellow
        35        304       3209       1415
> |
```

As we known that grey is reserved to the genes which are not part of any module, Thus the largest module is turquoise module with 3209 genes, next comes blue with 2749 genes. etc. Here we can notice that grey module with 1642 genes which are not part of any modules, which are considered as background genes.

TOMplot

Whereas TOMplot was the one of the important things used for visualizing the networks we created, in this pot the hierarchial clustering based on dissimilarity topological overlap matrix values and the module colors and further there overlapping is been clearly seen in the plot.

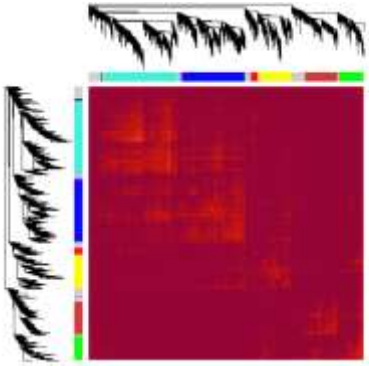


Figure 11: TOMplot

Generally, rows and columns of this tom plot sorted according to hierarchical clustering and now for a statistician only matters a hierarchical clustering from which the modules been cut. But to visualize underlying connectivity pattern we use TOM, which is more virtual to visualize at biology point. In case of branches those genes, which are at tip of the branch, they are also part of that cluster which they have high module membership and with more connectivity, in case of genes at top of the branch have weaker connections towards the cluster.

MDS Plot

Multi-dimensional scaling (MDS) was one of the visualizing tools of networks, whereas it takes the input data of dissimilarity measure and translates in to Euclidean distances. Here we visualize the same network where each dot represents a gene, and the modules roughly represent a finger of a hand in multi-dimensional scaling plot. The genes which are at the tip of fingers are the intra-modular hubs, where as some genes which are grey at the centre of multi-dimensional scaling plot where usually background genes, which are not most taking part of any process. There will be thousands of genes which are not part of any network or process, which they are not able to decide which way they should go. So, they situated at the centre of the plot with grey color (remember that we used to denote genes which are not part of any module, which are usually grey in color) which were called as background genes.

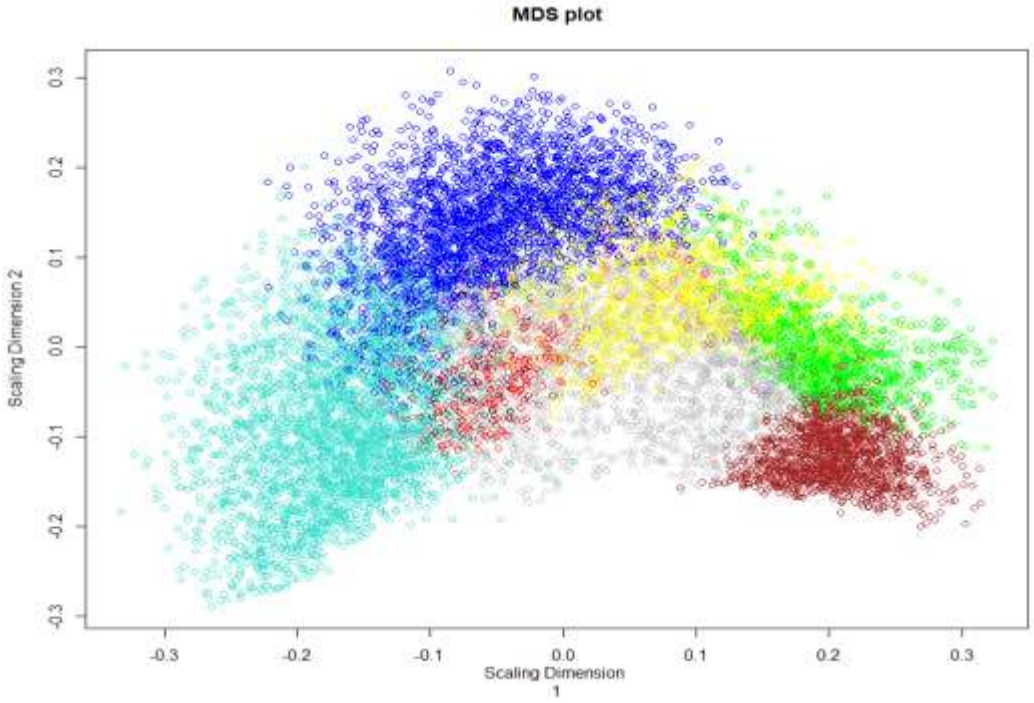


Figure 12: classical multi-dimensional scaling plot for visualizing the network based On dissimilarity topological overlap matrix (dissTOM).

Heatmaps of Prominent Modules Detected

In case of heatmap view of a module, where columns = tissue samples, and the rows = genes and color band indicates module membership. which is useful to measure the co-expression of genes, where if there is high expression the overlap color intensifies and if low at expression it remains light in color. Generally, we will notice red, black and green bands. Whereas red and black indicates high expression, but the green indicates low expression of that specific genes in that person. where the color indicates the level of expression of genes.

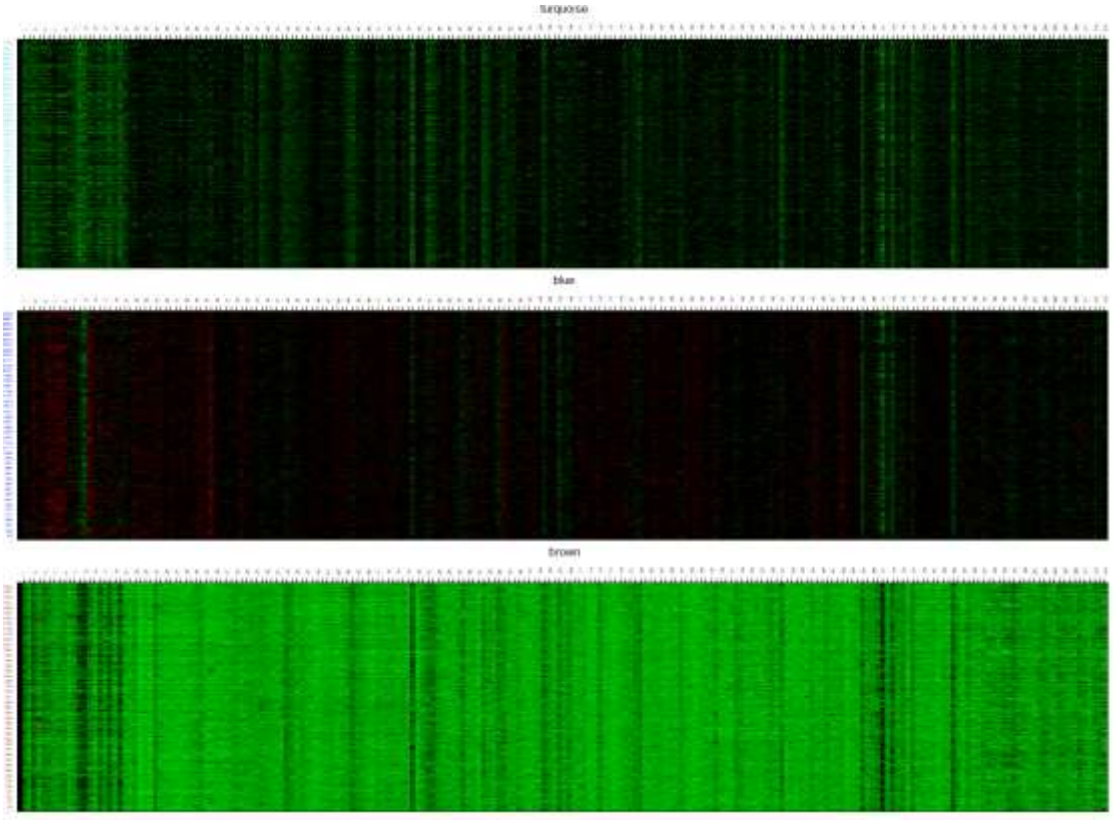


Figure 13: heatmap of turquoise, blue and brown modules

In the above figure, we can clearly observe that the turquoise color have dark green and black bands, which signifies that the gene co-expression was nominal, and in case of blue module we can observe red (which means over-expression), dark green and black which signifies most of the genes been expressed and the blue module is biologically significant. And further in case of brown module we can observe light green (low co-expression), and very little black, dark green and red bands, where totally the brown module co-expression is not so good, where that module is considered as biologically less significant.

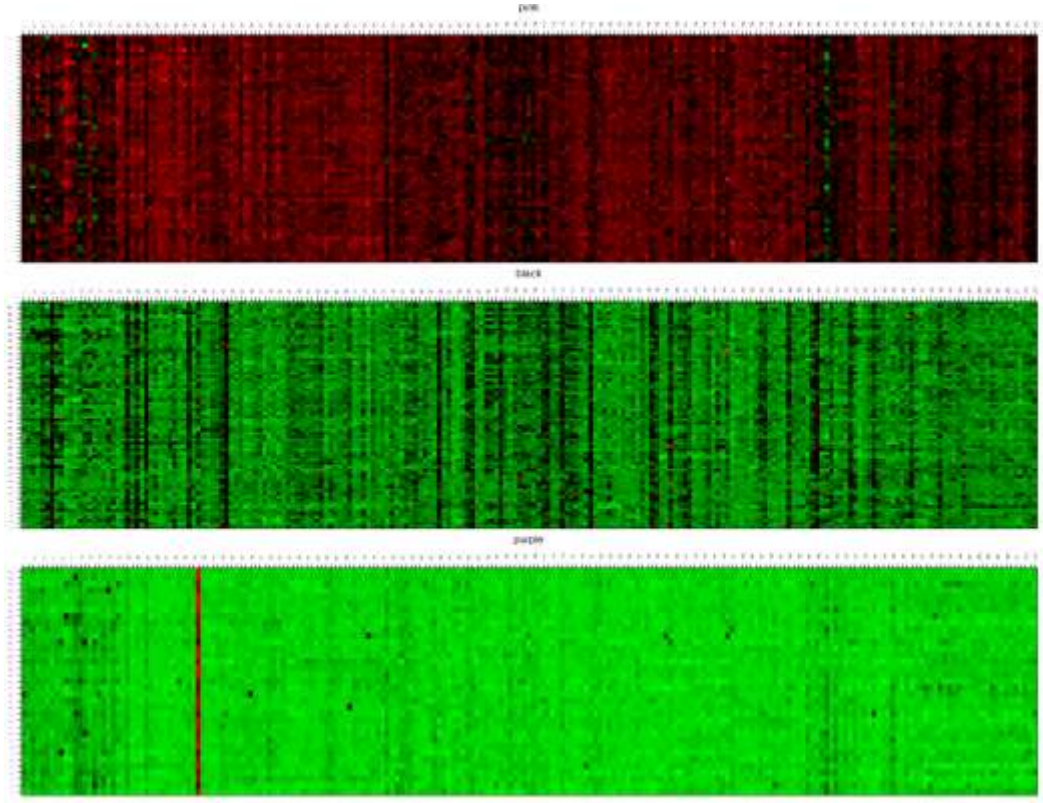


Figure 14: Heatmap of pink, black and purple modules

Where as in case of figure 4.16, where pink module consists of almost full of red color bands (which signifies that it is most significant module due to most co-expression). And in case of black module the gene co-expression is nominal. And in case of purple module the gene co-expression was too low, but at sample id 38, where all genes of the purple module been expressed too high.

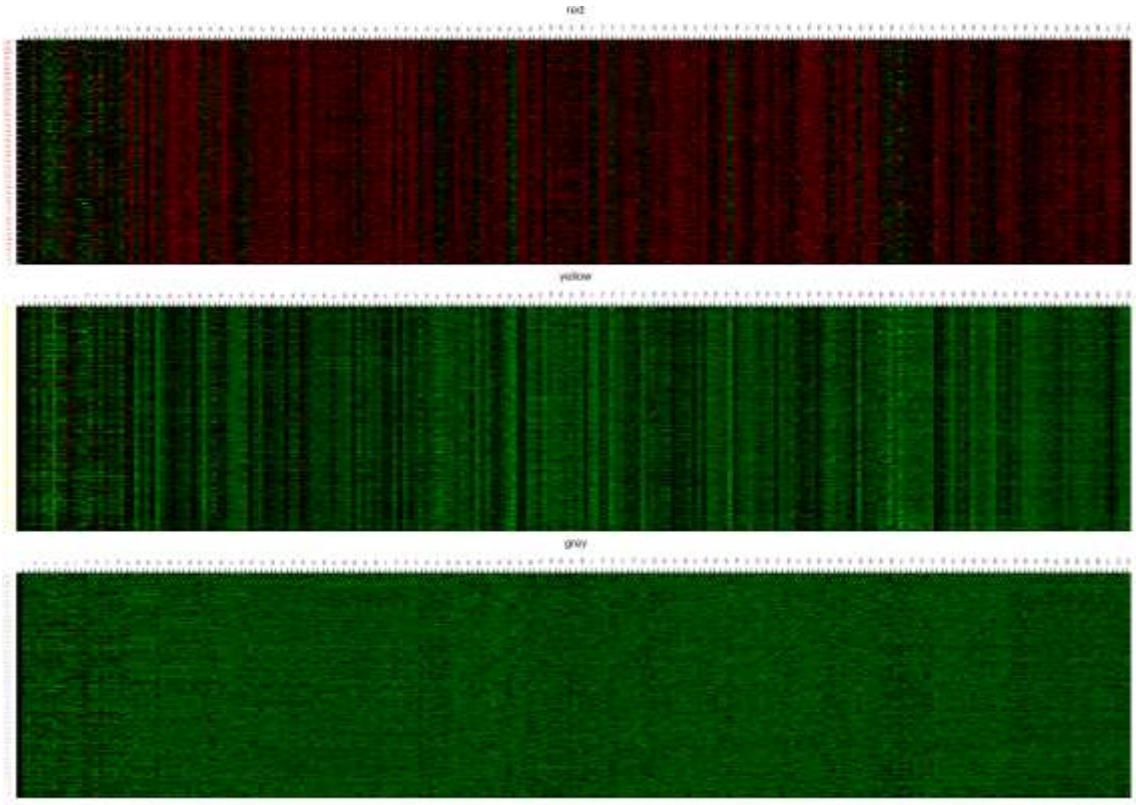


Figure 15: Heatmap of red, yellow and grey modules

In case of figure 4.17, whereas the red color module was a significant one where as the bands where almost red in color (which signifies the gene co-expression is high in that module and it is biologically significant module). Where as in in case of yellow and grey module the gene expression is nominal, which is not that high and not much low.

Module Eigengene

eigengene is not a real gene, it's just a weighted average of all module genes. These allow one to relate modules to each other, and also allows one to determine whether modules should be merged or to define eigengene networks. These allow one to relate modules to clinical traits and SNPs (which avoids multiple comparison problem). And also these allow one to define a measure of module membership: KME=cor(x,ME). Thus, eigengene been used for the purpose of visualization of networks.

Clustering Tree Based on The Module Eigengenes of Modules

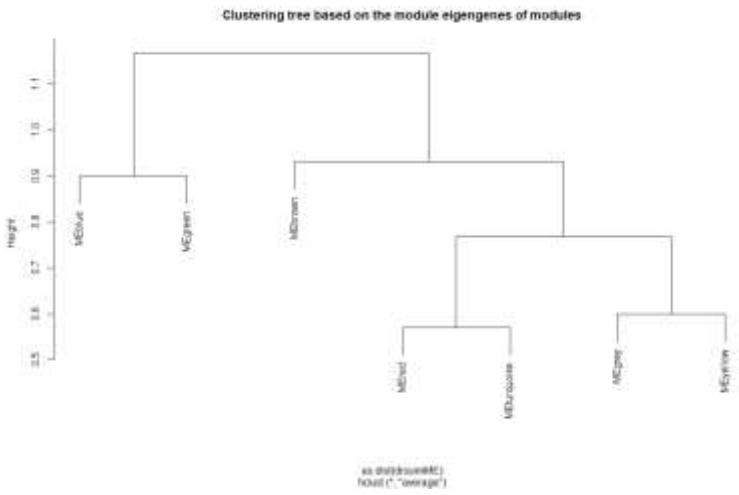


Figure 16: Clustering tree based on module eigengenes of module

In case of figure 4.18, we used to calculate the relation between the modules based on module eigengene values, whereas eigengene values was nothing but a weighted average of all module genes. So, usually based on the module eigengene values, we plotted a cluster tree diagram, which signifies how modules been related with each other. And in case of above clustering tree of modules eigengenes we can clearly observe close relation between red and turquoise modules, which is clearly seen in figure 4.19 (Relations between module eigengenes plot).

Relation between module eigengenes (which depicts how modules related with each other Through eigengene values

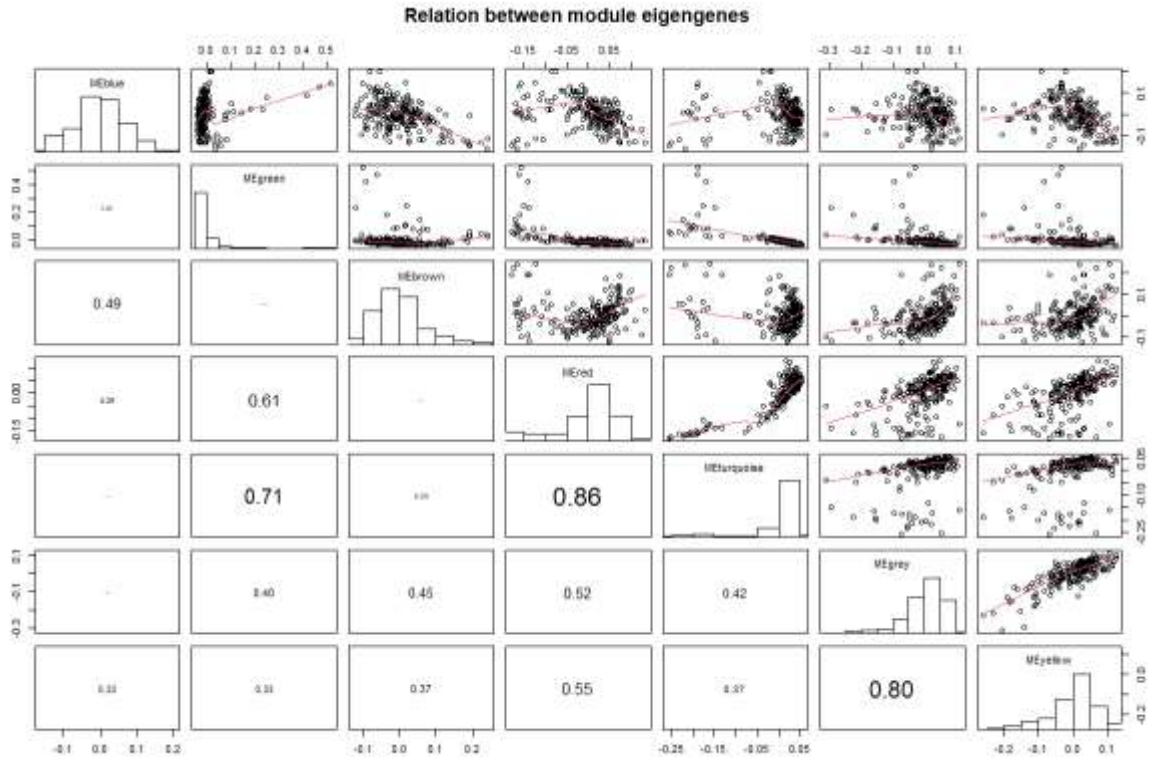


Figure 17: Relation between module eigengenes

In the above figure 4.19, which we generalize the relation between modules using module eigengenes, where as we can observe some high correlation among the modules such as, in case of red and turquoise the relation is high (around 0.86 as depicted in the figure). And also, in case of module grey and yellow the relation is high (around 0.80 as in the figure). And we can also note that the relation between green and turquoise was around 0.71, which is also good. And in case of less significantly related module the values where low like around 0.10 and 0.20. And in case of no relation between two eigengene modules there will be no value in the box, which significantly empty box. In case of above figure, there is no relation between blue and turquoise modules and also there is no relation between brown and red module eigengenes.

Further Prominent Modules Relationships From Module Eigengenes Been Preserved To Analyze With Cytoscape

From the above module eigengenes plot, we can observe the relation between red-turquoise modules been high around 0.86. And also, in case of module grey and yellow modules the relation is 0.80. And also, in case of green-turquoise modules the relation is about 0.71. So, these modules been preserved in cytoscape format, further to analyze the network and to detect prominent hub genes with the help of cytoscape tool.

Analysis of Prominent Highly Correlated Eigengene Modules With Cytoscape Tool

With the help of a cytoscape software, we know analyze those networks and find out top 5 hub genes in each highly correlated eigengene module.

Red-Turquoise eigengene module

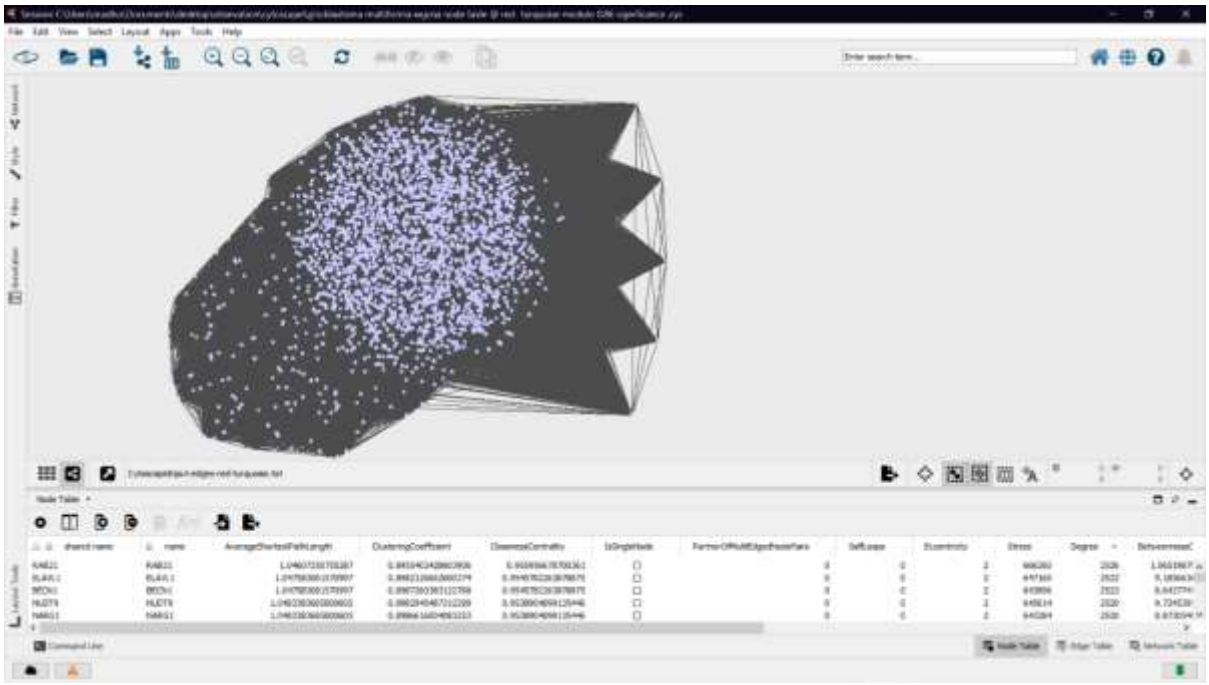


Figure 18: Top 5 hub genes of Red-Turquoise module with 0.86 significance value

In the above Red-Turquoise module the top 5 hub genes are RAB21, ELAVL1, BECN1, NUDT9, and NARG1. These hub genes been identified based on degree of connectivity of each gene to many neighborhood Genes.

-RAB21 GENE

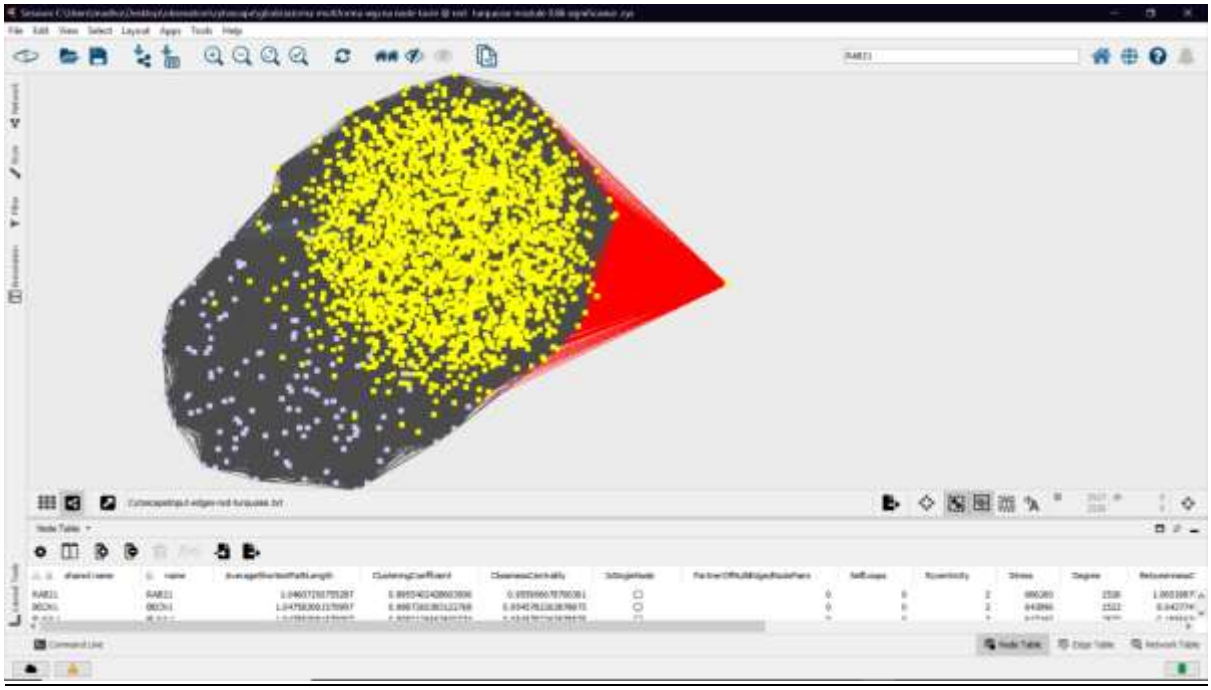


Figure 20: RAB21 gene connectivity across the Network

RAB21 Gene was related to RAS oncogene family, There where so many researches going on regarding the expression of RAB21 gene in Glioma's and other type of cancer's too. There was a study that Knockdown of RAB21 Gene inhibits proliferation and induces apoptosis in Human Glioma cells. ELAVL1 gene is a part of ELVL family, and these are RNA-binding protein coding genes (RBPs). And recent studies found that aberrant expression RBPs could effect cellular functions, leading to the occurrence and progression of various cancers, including Gliomas. And this gene been identified as a crucial oncogenic driver and promote malignant peripheral nerve sheet tumor growth and metastasis. Thus, it can be considered as a new therapeutic target.

BECN1 (beclin 1) gene acts as tumor suppressor and is an essential mediator of autophagy. Beclin 1 also interacts with Bcl-2 and can induce apoptosis by activating the mitochondrian permeabilizing function of proapopotic multidomain protein from Bcl-2 family. And also it is observed that Beclin 1 expression decreases with tumor progression. And NUDT9 gene belongs to NUDT family of genes, And these been exhibited associations with overall survival of Glioblastoma. And finally, in case of NARG1 gene there is no sufficient research work on it.

Grey-Yellow Eigengene Module

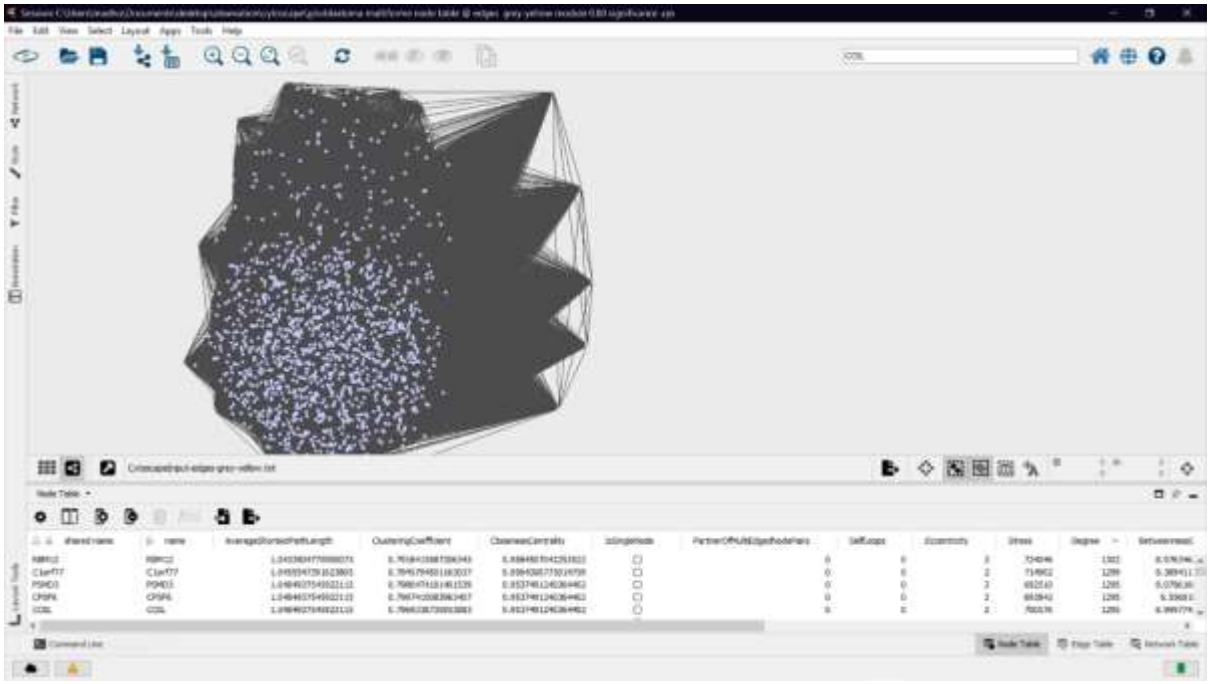


Figure 21: Top 5 Hub genes of Grey-Yellow module with 0.80 significance value

In the above Grey-Yellow module the Top 5 Hub genes are RBM12, C1orf77, PSMD3, CPSF6, and COIL. These hub genes been identified based on degree of connectivity of each gene to many neighborhood Genes.

Green-Turquoise Eigengene Module

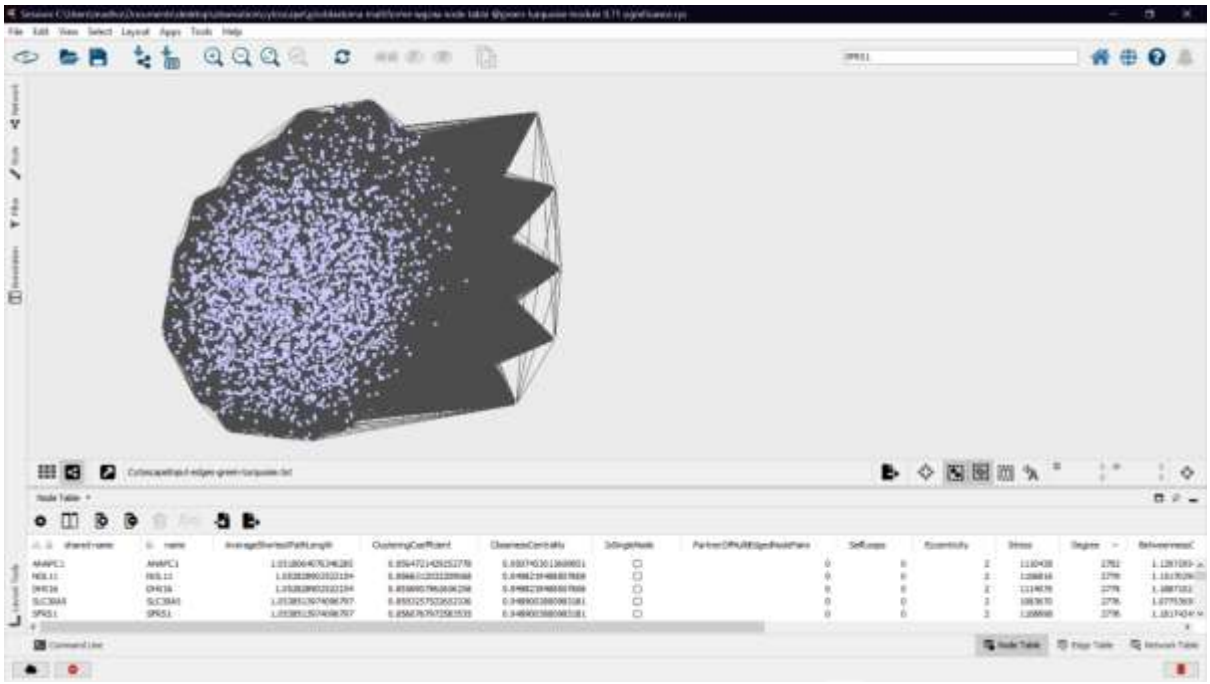


Figure 22: Top 5 Hub genes of Green-Turquoise eigengene module with 0.71 significance value

In the above Green-Turquoise eigengene module Top 5 Hub genes are ANAPC1, NOL11, DHX16, SLC30A5, and SFRS1**.** These hub genes been identified based on degree of connectivity of each gene

to many neighborhood Genes. The principal objective of this dissertation work is to find out crucial hub genes which involved in advancing the disease prognosis. Finding out the crucial hub genes will enhance the mode of treatment, which currently not so well advanced. So, for the purpose of enhancing precision medicine for Glioblastoma this was one of the steps. Generally, GBM was one of the most aggressive cancers which resulting few weeks of life after diagnosis. Incidence been rising based on advancement of technological factors. Developed countries had highest incidence rate. Thus, it is necessary to find out new methods to approach to treat Glioblastoma by using New Technologies and Artificial Intelligence.

The methodology we used to analyze the Genomic Expression Data was by WGCNA (Weighted Gene Co-expression Network Analysis) Package which is present in R studio develop by Steve Horvath. This WGCNA package been widely used for the purpose of analyzing Genomic Expression Data. Where initially we load the data and further, we visualize it. And later we export the biologically meaningful modules to cytoscape and visualize them at Gene level to find out the Hub Genes. Those Hub Genes where the crucial factors in those 12042 Genes which we analyzed for 215 patients. Those helpful for further analysis like designing precision medicine for this aggressive cancer. Thus, we analyze the molecular data with network learning processes for differential gene expression, Gene module detection, Trans-omics network analysis, cancer subtype identification, network regularization, causal inference for decision making to design personalized medicine. Thus, Finding the Hub genes was crucial thing which we done by analyzing the Gene Expression Data, which further useful for apoptosis of glioma cells and saving lives.

4. Conclusion

By using WGCNA package in Rstudio and Cytoscape Software we found 15 Hub Genes in 3 Eigengene Modules. In Red-Turquoise module we find out RAB21 Gene which was related to RAS oncogene family. And by Knockdown of RAB21 Gene inhibits proliferation and induces apoptosis in Human Glioma cells. Thus, those hub genes where crucial and they will help for better cure of this aggressive cancer. And also ELAVL1 been identified as a crucial carcinogenic driver which can be used as a new therapeutic target. And also BECN1 gene acts as tumor suppressor and it is an essential mediator of autophagy. NUDT9 gene also been exhibited associations with overall survival of glioblastoma. And also in case of green-yellow eigengene module and green-turquoise eigengene module several can genes led new pathways for discovery new therapeutics.

References:

1. Chakrabarti, I., Cockburn, M., Cozen, W., Wang, Y. P., & Preston-Martin, S. (2005). A population-based description of glioblastoma multiforme in Los Angeles County, 1974-1999. Cancer, 104(12), 2798–2806.
2. Urbanska, K., Sokolowska, J., Szmidt, M., & Sysa, P. (2014). Glioblastoma multiforme - An overview. Wspolczesna Onkologia, 18, 307–312.
3. Ray, S. K. (2010). Glioblastoma: Molecular mechanisms of pathogenesis and current therapeutic strategies. Glioblastoma Mol Mech Pathog Curr Ther Strateg, 1–431.
4. Hanif, F., Muzaffar, K., Perveen, K., Malhi, S. M., & Simjee, S. U. (2017). Glioblastoma multiforme: A review of its epidemiology and pathogenesis through clinical presentation and treatment. Asian Pacific Journal of Cancer Prevention, 18(1), 3–9.
5. Llaguno, S. R. A., & Parada, L. F. (2016). Cell of origin of glioma: Biological and clinical implications. British Journal of Cancer, 115(12), 1445–1450.
6. Tso, C. L., Freije, W. A., Day, A., Chen, Z., Merriman, B., Perlina, A., et al. (2006). Distinct transcription profiles of primary and secondary glioblastoma subgroups. Cancer Research, 66(1), 159–167.
7. Kleihues, P., & Ohgaki, H. (1999). Primary and secondary glioblastomas: From concept to clinical diagnosis. Neuro-Oncology, 1(1), 44–51.
8. Karcher, S., Steiner, H. H., Ahmadi, R., Zoubaa, S., Vasvari, G., Bauer, H., et al. (2006). Different angiogenic phenotypes in primary and secondary glioblastomas. International Journal of Cancer, 118(9), 2182–2189.
9. Munshi, A. (2016). Central nervous system tumors: Spotlight on India. South Asian Journal of Cancer, 5(3), 146–147.
10. Chen, B., Chen, C., Zhang, Y., & Xu, J. (2021). Recent incidence trend of elderly patients with glioblastoma in the United States, 2000–2017. BMC Cancer, 21(1), 1–10.
11. D'Alessio, A., Proietti, G., Sica, G., & Scicchitano, B. M. (2019). Pathological and molecular features of glioblastoma and its peritumoral tissue. Cancers, 11(4).
12. Aldape, K., Zadeh, G., Mansouri, S., Reifenberger, G., & von Deimling, A. (2015). Glioblastoma: pathology, molecular mechanisms, and markers. Acta Neuropathologica, 129(6), 829–848.
13. Joshi, S. K. (2015). Molecular pathogenesis of glioblastoma multiforme: Nuances, obstacles, and implications for treatment. World Journal of Neurology, 5(3), 88.

14. Batash, R., Asna, N., Schaffer, P., Francis, N., & Schaffer, M. (2017). Glioblastoma Multiforme, Diagnosis and Treatment; Recent Literature Review. Current Medicinal Chemistry, 24(27), 3002–3009.
15. Raizer, J., & Parsa, A. (2015). Current Understandings and Treatment of Gliomas. Cancer Treatment and Research, 163.
16. Bo, L. J., Wei, B., Li, Z. H., Wang, Z. F., Gao, Z., & Miao, Z. (2015). Bioinformatics analysis of miRNA expression profile between primary and recurrent glioblastoma. European Review for Medical and Pharmacological Sciences, 19(19), 3579–3586.
17. Langfelder, P., & Horvath, S. (2008). WGCNA: An R package for weighted correlation network analysis. BMC Bioinformatics, 9.
18. Xiang, Y., Zhang, C. Q., & Huang, K. (2012). Predicting glioblastoma prognosis networks using weighted gene co-expression network analysis on TCGA data. BMC Bioinformatics, 13(Suppl 2), S12.
19. Tarca, A. L., Romero, R., & Draghici, S. (2006). Analysis of microarray experiments of gene expression profiling. American Journal of Obstetrics and Gynecology, 195(2), 373–388.
20. Hackenberger, B. K. (2020). R software: Unfriendly but probably the best. Croatian Medical Journal, 61(1), 66–68.
21. Zhang, B., & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. Statistical Applications in Genetics and Molecular Biology, 4(1).
22. Lopes, M. B., Martins, E. P., Vinga, S., & Costa, B. M. (2021). The role of network science in glioblastoma. Cancers, 13(5), 1–22.
23. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: A Software Environment for Integrated Models. Genome Research, 13(11), 426.