T.C.

DOKUZ EYLUL UNIVERSITY

IZMIR INTERNATIONAL BIOMEDICINE AND GENOME INSTITUTE

COMPUTATIONAL APPROACHES TO THE IDENTIFICATION OF DIAGNOSTICS AND PROGNOSTICS MARKERS IN BRAIN NEOPLASMS

MELİH ÖZBEK

MOLECULAR BIOLOGY AND GENETICS

MASTER OF SCIENCE THESIS

IZMIR-2022

THESIS CODE: DEU.IBG.MSc-2020850026

T.C.

DOKUZ EYLUL UNIVERSITY

IZMIR INTERNATIONAL BIOMEDICINE AND GENOME INSTITUTE

COMPUTATIONAL APPROACHES TO THE IDENTIFICATION OF DIAGNOSTICS AND PROGNOSTICS MARKERS IN BRAIN NEOPLAMS

MOLECULAR BIOLOGY AND GENETICS

MASTER OF SCIENCE THESIS

MELİH ÖZBEK

Supervisor

Asst. Prof. ATHANASIA PAVLOPOULOU

THESIS CODE: DEU.IBG.MSc-2020850026

Dokuz Eylül University Izmir International Biomedicine and Genome Institute Department of Genomics and Molecular Biotechnology, Molecular Biology and Genetics graduate program Master of Science student Melih Özbek has successfully completed her Master of Science thesis titled **"Computational Approaches to the Identification of Diagnostics and Prognostics Markers in Brain Neoplasms"** on the date of 04/07/2022.

HEAD OF THE COMMITTEE
Asst. Prof. Dr. Athanasia PAVLOPOULOU

MEMBER
Instructor Dr. Hasan Buğra ÇOBAN

MEMBER
Assoc. Prof. Dr. Efe SEZGİN

YÜKSEK LİSANS TEZ SAVUNMA SINAVI TUTANAĞI

Dokuz Eylül Üniversitesi İzmir Uluslararası Biyotıp ve Genom Enstitüsü Genom Bilimleri ve Moleküler Biyoteknoloji Anabilim Dalı, Moleküler Biyoloji ve Genetik Yüksek Lisans programı öğrencisi Melih Özbek **"Hesaplamalı Yöntemler Kullanılarak Beyin Neoplazmasında Prognoz ve Teşhis İşaretlerinin Belirlenmesi"** konulu Yüksek Lisans tezini 04/07/2022 tarihinde başarılı olarak tamamlamıştır.

BAŞKAN
Dr. Öğr. Üyesi Athanasia PAVLOPOULOU

JÜRİ ÜYESİ
Dr. Öğr. Gör. Hasan Buğra ÇOBAN

JÜRİ ÜYESİ
Doç. Prof. Dr. Efe SEZGİN

TABLE OF CONTENTS

INDEX OF FIGURES

LIST OF ABBREVIATIONS

BAM	Binary Alignment Map
cDNA	Complementary Deoxyribonucleic Acid
CGGA	Chinese Glioma Genome Atlas
CNS	Central Nervous System
DEA	Differential Expression Analysis
DEG(s)	Differentially Expressed Gene(s)
FPKM	Fragments per Kilobase of Transcript per Million Mapped Reads
GBM	Glioblastoma
GEO	Gene Expression Omnibus
GTEx	Genotype Tissue Expression
HGG	High(er)-grade Glioma
IDH	Isocitrate Dehydrogenase Gene
LGG	Low(er)-grade Glioma
NCBI	National Center for Biotechnology Information
NGS	Next-Generation Sequencing
PCA	Principal Component Analysis
RPKM	Reads per Kilobase of Transcript per Million Mapped Reads
RT	Radiation Therapy
SBS	Sequencing by Synthesis
TCGA	The Cancer Genome Atlas
WGCNA	Weighted Gene Co-expression Network Analysis
WHO	World Health Organization

ACKNOWLEDGEMENTS

Firstly, I would like to thank and express my deep and sincere gratitude to my supervisor Assist. Prof. Athanasia Pavlopoulou for giving me the opportunity to perform my research. I thank her for her invaluable guidance and endless patience with me. I am grateful for her help and encouragement on my PhD study. I am also thankful for her support and her sweet treats on a daily basis during my education.

Special thanks to the PhD student Halil İbrahim Toy. He was also there for me every time. I would like to thank him for his endless assistance, support and patience with me during my studies.

Their encouragement and belief in me gave me the motivation and confidence to go on, even on hard times. Therefore, I am very grateful to them.

Many thanks to the Assoc. Prof. Gökhan Karakülah and Assist. Prof. Yavuz Oktay for their support and knowledge during my studies.

I would like to thank all members of IBG for giving me opportunity to perform my study.

I am greatly indebted to my family. They were always supportive of my education and believed in me during the whole process.

I would also like to thank Gizem ÖZTÜRK for sharing her valuable time daily with lunch breaks, coffee and tea breaks and, basically, giving me motivation and believing in me.

HESAPLAMALI YÖNTEMLER KULLANILARAK BEYİN NEOPLAZMASINDA PROGNOZ VE TEŞHİS İŞARETLERİNİN BELİRLENMESİ

Melih ÖZBEK, İzmir Uluslarası Biyotıp ve Genom Merkezi, Dokuz Eylül Üniversitesi Sağlık Kampüsü, Balçova 35340 – İzmir / TÜRKİYE

ÖZET

Düşük sınıflı gliomalar (LGG) sinir sistemi 1. Sınıf tümörlerdir ve ilerledikçe en ölümcül beyin tümörlerinden biri oluyorlar. Bu gün zamanında ve doğru teşhis ve prognoza ihtiyaç vardır. Bu çalışmanın amacı düşük sınıflı gliomalarla ilişkili teşhisine ve prognozuna yönelik biyoişaretlerinin birçok farklı hesaplamalı yöntemler kullanılarak bulunmasıdır. Bu amaç için üç farklı metot kullanarak LGG ve karşılık denk gelen normal beyin dokusu arasında diferansiyel ifadeleme analizi gerçekleştirildi. Toplamda LGG ve sağlıklı normal beyin dokusu arasında dört bin dört yüz doksan altı diferansiyel ifade edilen gen tespit edildi. Daha sonra ortak diferansiyel ifade edilen genler, ağırlıklı olarak birlikte ifade edilen genlerin ağ analizinde (WGCNA), birlikte ifade edilen gen kümelerini bulmak için kullanıldı. Bu sayede klinik fenotiplerle ilişkili iki dikkate değer modül bulundu. Bu modüllerde bulunan DEGler ise birlikte ifade edilen gen ve protein etkileşim ağ analizlerinin inşasında kullanıldı. Bu analizler doğrultusunda toplam on sekiz hub gen tespit edildi: *CD74*, *CD86*, *CDC25A*, *CYBB*, *HLA-DMA*, *ITGB2*, *KIF11*, *KIFC1*, *LAPTM5*, *LMNB1*, *MKI67*, *NCKAP1L*, *NUSAP1*, *SLC7A7*, *TBXAS1*, *TOP2A*, *TYROBP*, ve *WDFY4*. Tüm tespit edilen hub genler LGG tümöründe yukarı regüle edilmiş olarak bulundu ve LGG hastlarında da olumsuz prognoz fenotiplerle ilişkili bulunmuştur. Bu sonuçlar klinik bir platformda LGG tümörünün teşhisi ve hastaların prognozu üzerine uygulanabilir.

Anahtar kelimeler: düşük sınıflı gliomalar, diferansiyel gen ifadeleme analizi, ağırlıklı gen ifade ağı analizi, transkriptomik analiz, teşshis, prognoz, biyoinformatik, system biyolojisi

COMPUTATIONAL APPROACHES TO THE IDENTIFICATION OF DIAGNOSTICS AND PROGNOSTICS IN BRAIN NEOPLASMS

Melih ÖZBEK, Izmir International Biomedicine and Genome Center, Dokuz Eylül University Health Campus, Balçova 35340 – Izmir / TURKEY

ABSTRACT

Low-grade gliomas (LGG) are central nervous system Grade I tumors, and as they progress become one of the deadliest brain tumors. Today, there is still great need for timely and accurate diagnosis and prognosis. The aim of this study, was to identify diagnostic and prognostic biomarkers associated with LGG, by employing using diverse computational approaches. For this purpose, differential gene expression analysis between LGG tissue and corresponding normal tissue was performed by using three methodologies. A total of four thousand four hundred and ninety six common differentially expressed genes (DEGs) between LGG and healthy brain tissue were detected. Next, those DEGs were used to conduct Weighted Gene Co-Expression Network Analysis (WGCNA) in order to identify clusters of co-expressed genes. In this way, two modules significantly correlated with clinical traits were detected. The DEGs comprising the modules were further used to construct gene co-expression and protein-protein interaction networks. Based on this analysis, we derived a consensus of eighteen hub genes, namely, *CD74*, *CD86*, *CDC25A*, *CYBB*, *HLA-DMA*, *ITGB2*, *KIF11*, *KIFC1*, *LAPTM5*, *LMNB1*, *MKI67*, *NCKAP1L*, *NUSAP1*, *SLC7A7*, *TBXAS1*, *TOP2A*, *TYROBP*, and *WDFY4*. All detected hub genes were up-regulated in LGG, and were also associated with unfavorable prognosis in LGG patients. These findings could be applicable in the clinical setting for diagnosing and monitoring LGG.

Keywords: low-grade gliomas, transcriptome analysis, differential gene expression analysis, weighted gene co-expression network analysis, diagnosis, prognosis, bioinformatics, systems biology

1. INTRODUCTION AND AIM

1.1. Statement and Importance of the Problem

Gliomas are cancers originating from glial cells which serve as supportive cells to the central nervous system. According to World Health Organization (WHO) general classification of gliomas, they range from Grade I to IV (the malignant form). Due to the complex etiology of the disease, nowadays, despite the availability of modern imaging technologies, a proper diagnosis and prognosis is lacking. In this study, by taking advantage of the vast amount of gene expression data stored in publicly available repositories, such as TCGA, GTEx and GEO, we aimed to detect potential biomarkers that could be useful for the diagnosis and also prognosis of low-grade gliomas. To do this, various computational approaches were employed,including differential gene expression analysis, weighted gene co-expression network analysis, and network-based methods. In this way, we detected eighteen candidate biomarkers for low- grade gliomas.

1.2. Aim of the Study

The aim of the present study was to perform differential gene expression analysis between low-grade gliomas (LGG) and corresponding normal brain tissue derived from TCGA and GTEx respectively, towards the identification of potential LGG biomarkers.

1.3. Hypothesis of the Study

It is hypothesized that biomarkers identified in this study could be used in the clinical setting for LGG diagnosis and prognosis.

2. GENERAL INFORMATION

2.1. What are Gliomas?

Cancer is a disease originating from genomic alterations that include somatic mutations and germline variants. Numerous studies on cancer tissues allowed the identification of such alterations and mutations, including, for example, the tumor suppressor genes *TP53*, *RB1* and *APC* [1]. Tumors located in the central nervous system (CNS) are defined based on their cells of origin and histopathological characteristics. Gliomas are neuroepithelial and highly heterogeneous tumors originating from neuroglial progenitor cells [2, 3]. Out of all primary brain tumors, gliomas represent 27% and account for 80% of the malignant primary brain tumors [3].

The classification of gliomas has changed throughout time, and the World Health Organization (WHO) redefined gliomas in 2016 [3]. Traditionally there are four classes of gliomas: low(er)-grade gliomas (LGGs) are classified as WHO Grade I and II tumors; high(er)-grade gliomas (HGGs) are classified as WHO Grade III and IV tumors. Contrary to previous classifications, WHO added molecular and genetic parameters that improved diagnostics and prognostics of this type of cancers. Major changes occurred in the subcategorization of gliomas because genetic mutations and prognostic factors may vary highly [4, 5]. Basically, after the discovery of mutations in the genes *isocitrate dehydrogenase gene (IDH) 1* and *IDH2*, and 1p19q deletion, when combined with their impact on short/long patient survival emerged the need of reclassification and hence new questions about optimal treatments and therapies [6]. Currently, aside from the traditional system, classification is also done based on the type of glioma cells such as astrocytoma, oligodendroglioma integrated with status of mutations in genes such as *IDH1/2*, *ATRX*, *TP53* and 1p/19q deletion [7].

2.1.2. Low-grade versus High-grade Gliomas

Although the presence of low-grade gliomas is not common in patients older than 60 years old, they generally occur in people between the ages of 20 and 40 years old. While the most common symptoms among patients is the presence of seizures, patients may have headaches too. Tumor location greatly affects the symptoms and also severity. Seizures are thought to be due to the invasion of cancer tissue into the cortex region of the brain [8]. While LGG constitutes 20% of all primary brain tumors, the survival time period of LGG patients

ranges between 4.7 and 9.8 years. Therefore, LGG patients are expected to survive longer compared to HGG patients [9]. The most malignant form of gliomas is called glioblastoma (GBM) and is recognized to be one of the deadliest brain tumor affecting adults. GBMs are classified as WHO Grade III and IV tumors. Those GBMs that transform and progress from LGGs are called secondary GBMs, while those GBMs that are formed *de novo* are called primary GBMs [10, 11]. Differences between primary and secondary GBMs arise from various genetic alterations and some clinical traits. Hence, these differences may determine the treatment options [11].

2.1.3. Treatment Options

There are different options of treatment for patients with LGG. Some LGG patients are not even aware of this condition. Therefore, timely and accurate diagnosis is critically important before tumor progression to a higher grade. Compared to the short survival rate of patients with HGGs or GBMs, patients with LGGs show higher rates of survival, which creates controversies regarding the administration of treatment in terms of dosage, side-effects and even applicability of them [12]. The conventional treatment options are surgical resection, radiotherapy and chemotherapy [13].

The aim of surgery is to achieve the maximum possible resection. Firstly, depending on the location of LGGs, resection has to be performed in a way so as to preserve the functionality of that region of the brain, for instance language functions, and is called function-based surgery. Since LGGs can be infiltrative to extend out of the foreknown borders, resection requires prior examination of imaging, which is called image-based surgery [13].

Radiotherapy or chemotherapy or combinatorial therapy are also controversial depending on the risk status of the patients. It is recommended for high risk patients to receive such treatment after surgical operations and can be followed with observational time. For low-risk patients, if they underwent a complete resection of tumor, observational time is required and sometimes radiotherapy is applied. The same holds for chemotherapy also. Radio/chemo-therapy is preferred as postoperative treatment for patients with the possibility of high recurrence of the tumor. Dosages and intensity of the treatment can vary among patients. Therefore, accurate diagnosis and prognosis is very important, since wrong treatment choice can result to decrease in the quality of life [14, 15].

2.1.4. Towards more Accurate Diagnosis and Prognosis

Generally, gliomas have poor prognosis. Despite recent advances in diagnosis and prognosis, the survival of patients with GBM ranges between 10 to 15 months. Gliomas have different subtypes, with many and complex molecular profiles due to various etiologic pathways which create a heterogeneity within the disease. While environmental factors were not found to contribute to gliomas, the genetic background constitutes a major contributing factor though [16]. Progression of LGGs to HGGs occurs with less therapeutic response and consequently poorer understanding of the disease. Therefore, better diagnosis and prognosis is an urgent need [17].

Although taking into account the molecular alterations for the classification of gliomas by WHO allowed the improvement of diagnosis and treatments, some controversies persist. Gliomas tend to have indeterminate cytological characteristics, and thus it is difficult, even for experienced clinicians to make definite conclusions. Sometimes histological features of diffuse midline gliomas, which are observed generally in young patients, resemble GBMs or low grade-diffuse astrocytomas in patients of older ages [18].

Recent findings support that molecular alterations such as 1p/19q, MGMT, and TERT are very useful for identifying tumor purity and cancer cells among tissues. Furthermore, these markers allowed defining the tumor microenvironment and hence a better prognosis and treatment [19]. For instance, mutations in *IDH1/2* cause distinct cancer types such as LGG and chondrosarcoma. Such alterations cause excessive production of an oncometabolite, D-2HG, and contributes to oncogenesis by various ways such as interfering with the cell metabolism or through epigenetic regulation [20].

Another study by Zang *et al*. (2021) points out the importance of the microenvironment of LGGs during immunotherapy. Gliomas can release inhibitory cytokines in a way to suppress and change their microenvironment and hence develop resistance to immune cells. Moreover, this can trigger immune cells to change into a pro-tumor phenotype [21]. Therefore, the identification of appropriate biomarkers and the development of rational therapeutic strategies are very important.

2.2. High-throughput RNA-Sequencing

2.2.1. Next Generation Sequencing

During a time span of 25 years, from 1977 to 2004, Sanger sequencing introduced a new era in the perspective of understanding DNA. Although it was effective and useful, it was also expensive. After the advent of next-generation sequencing (NGS) in 2005, billions of DNA templates can now be sequenced with higher efficiency and lower cost, which launched new applications of data analysis. As NGS offers major advantages, there are also disadvantages depending on the technology or technique used, such as the length of the DNA fragments (reads) to be sequenced or repetitive sequences [22].

Although currently not in use, Pyrosequencing platform by 454 Life Sciences applied the method sequencing by synthesis (SBS). Ion torrent sequencing platform also uses SBS with better accuracy than pyrosequencing. Sequencing by Illumina, on the other hand, offers fast, low in cost and also accurate service. Other sequencing platforms, such as Pacific Biosciences, Helicos fluorescent sequencing and nanopore sequencing by Oxford Nanopore Technologies, involve single-molecule sequencing [23].

2.2.2. RNA-Sequencing

According to the Central Dogma of biology, DNA molecules are transcribed into RNAs and then RNAs are translated into proteins. RNAs are the intermediate molecules between DNA and proteins. Thus, transcription of a collection of genes into RNAs could define a cell's identity and metabolism-related activities. As there are protein-coding RNAs, there are also non-coding RNAs which constitute a diverse group of RNAs with multiple regulatory functions. Transcriptome represents the information derived from RNAs. NGS technologies enabled the revolutionization of transcriptomics. By sequencing RNA molecules through their complementary DNA (cDNA), RNA-Sequencing (RNA-Seq) emerged, thereby bettering our understanding of the transcriptome such as gene expression and alternative splicing [24].

Basically, the first step in RNA-Seq analysis is the isolation of RNA molecules from samples. The overall quality of the isolated RNA molecules also affects downstream analysis, as well as the RNA molecule of interest. In the case of mRNA sequencing, the rRNA should be depleted, since rRNAs make up of at least 80% of total RNAs and therefore reduce the overall coverage. The next step is the construction of a DNA library from RNA samples, where the

extracted RNAs are reverse-transcribed into cDNA. These cDNA fragments are then amplified and sequencing adaptors are attached to the ends of these fragments. The final step is the actual sequencing of those fragments by using a variety of available platforms and methods; platform selection also can affect largely the downstream analysis [25].

2.3. Transcriptomic Analysis

Transcriptomic analysis involves two main processes, discovery and quantification. As there are no optimal pipelines to be applicable for every type of study, there are many different approaches which scientists can adopt. To carry out successfully an RNA-Seq study, a good experimental design includes the subjects of choosing optimal experiments, library type and number of replicates is very important [26].

After obtaining the RNA-Seq data, the next step is the analysis of these data. Depending on the research question, a variety of methods and tools can be applied for the downstream analysis. After getting the raw RNA-Seq reads from the sequencer, the quality of the data is checked. This step is called quality control and includes assessment of sequence quality, adaptor sequence presence, GC content or duplicated sequences; such noises could be due to contamination or PCR artifacts. In overall, quality check steps can be integrated into each step of an RNA-seq analysis and the output of each step can be quality-checked from different perspectives [26].

The next step in RNA-Seq analysis is read alignment, where the quality-checked reads are aligned/mapped to a reference genome or transcriptome. Accurate read alignment is very important and is indicated by the percentage of mapped reads, which can be quantified. Basically, raw counts are aggregated and quantified by applying different algorithms for gene/transcript-level quantification. Therefore, different outputs can be generated and stored in file formats such as gene transfer format (GTF) as raw read counts such as reads/fragments per kilobase of exon model per million mapped reads (R/FPKM) [26].

2.4. Differential Gene Expression Analysis

Downstream analysis after quantification varies depending on the research question. For instance, a user can perform differential gene expression analysis (DGEA), alternative splicing analysis or more. DGEA is carried out to discover quantitative differences between groups of

samples. For DGEA, different statistical methods have been developed that depend on different mathematical algorithms such as Poisson, negative binomial or Bayesian [27].

In this study, three different R packages were used for DGEA, namely, edgeR, DESeq2 and limma. DESeq2 and edgeR use negative binomial model for the estimation of the data and differentially expressed genes (DEGs). On the other hand, limma uses an Empirical Bayesian model. They are different methods that can be applied for RNA-Seq data analysis, and the testing strategy varies for each method. For instance, the testing strategy of edgeR includes the Exact test and the Likelihood-Ratio test. The Wald test and Likelihood-Ratio test are the testing strategies of the DESeq2 package. The testing strategy of limma on the other hand is to evaluate the posterior probability for interference [27].

2.5. Kaplan-Meier Survival Analysis

In this study Kaplan-Meier survival analysis was also performed by taking into consideration the detected hub genes. This is a method widely used to measure the mortality rate of subjects, for example groups of alive patients for a period of time after receiving treatment [28].

3. MATERIALS AND METHODOLOGY

3.1. Type of the Study

The entire study was performed by employing computational methods and resources.

3.2. Time and Place of the Study

This study was conducted in Izmir Biomedicine and Genome Institute from March 2021 to June 2022.

3.3. Materials and Methods of the Study

In this study, raw RNA-Seq data from TCGA, GTEx and GEO were used.

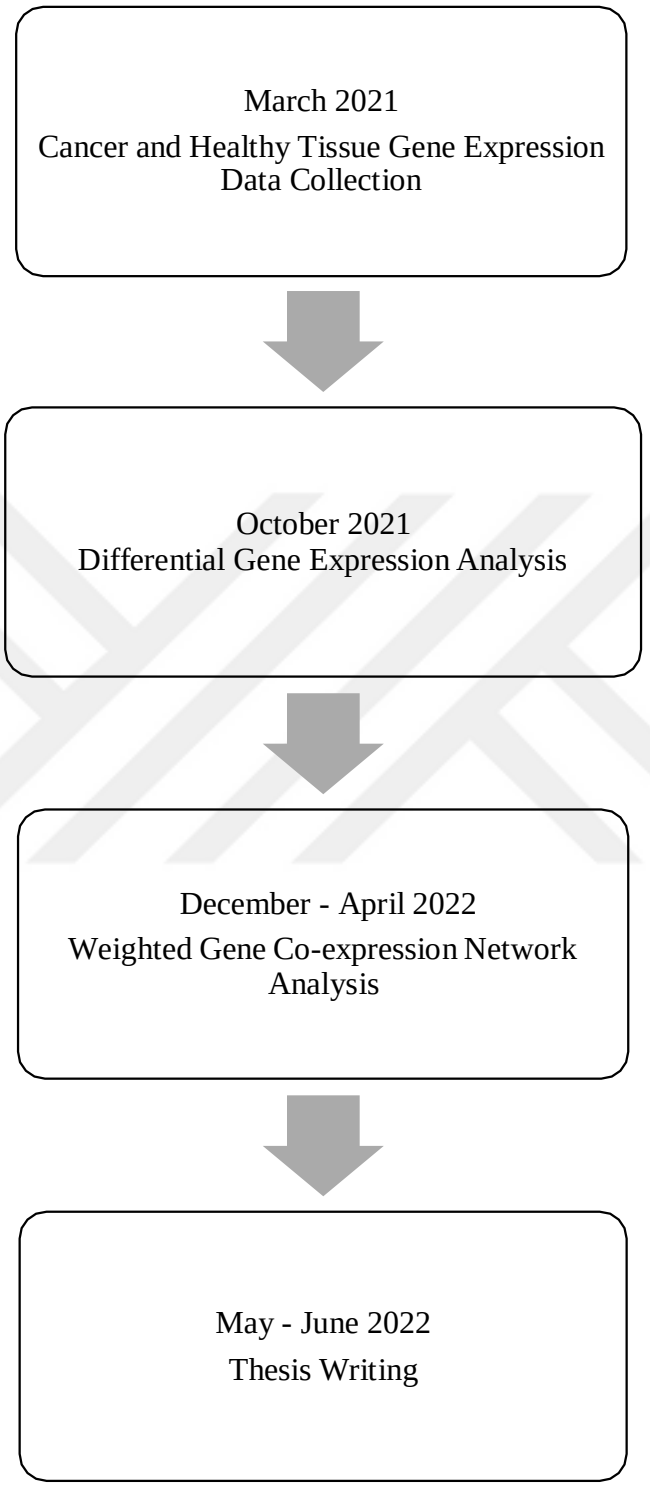
3.4. Variables of the Study

The variables of the study are the samples used for the analysis, that is, two tumor sample groups obtained from the TCGA and GEO databases, and one set of corresponding normal brain tissue data obtained from GTEx.

3.5. Tools for Data Collection

The R programming software platform (http://www.r-project.org/) (version 4.1.1) and Linux (GNU P. Free Software Foundation. Bash (version 3.2. 48) [Unix shell program], 2007) platform were used to conduct all analyses.

3.6. Study Plan and Calendar



3.7. Data Evaluation

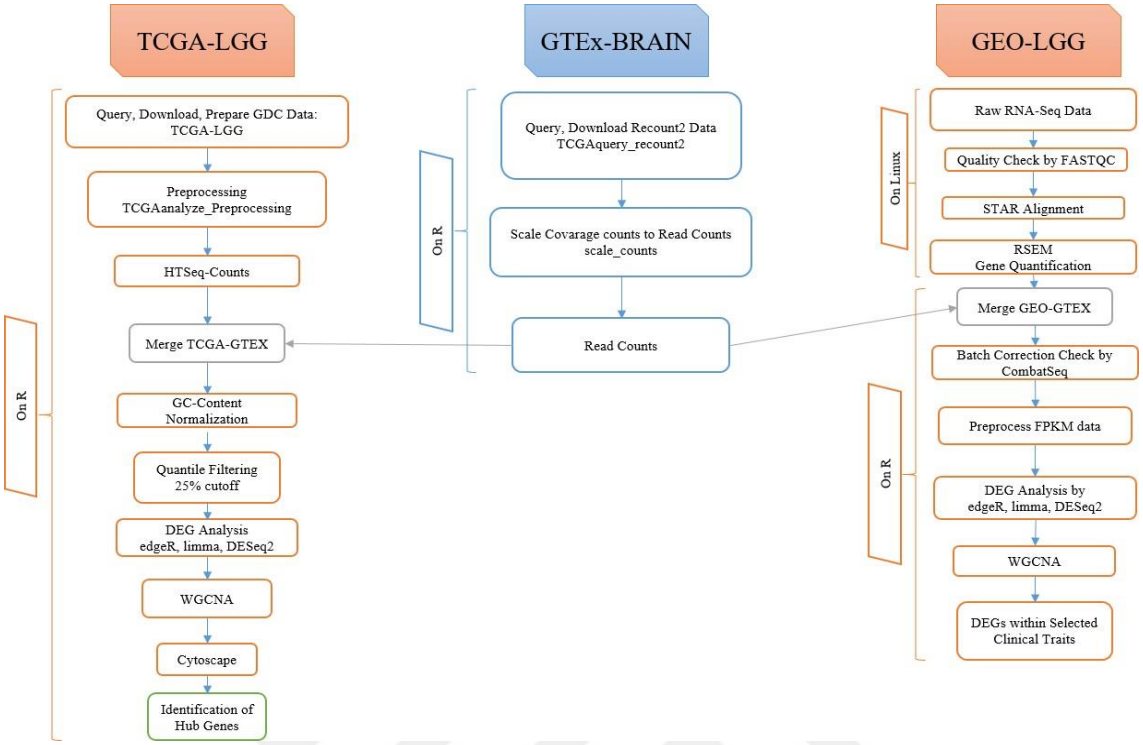


Figure 1. Overall pipeline of the procedure followed in this study.

3.7.1. Data Acquisition and Pre-processing

3.7.1.1. Acquisition of Low-Grade Glioma Data from TCGA

On the R programming environment (https://www.R-project.org/), TCGA (https://www.cancer.gov/tcga) derived LGG data were downloaded by using the 'TCGAbiolinks' package [29].

Target query was created by using the project, data.category, data.type and workflow.type attributes via the *GDCquery* function. Using information within the target query, harmonized RNA-Seq data was downloaded with the *GDCprepare* function.

A total of 529 samples were downloaded and processed via the GDC portal (https://portal.gdc.cancer.gov/) from the TCGA database.

Then the downloaded data were preprocessed by using the cor.cut attribute setting as 0.6 of the *TCGAanalyze_Preprocessing* function of the 'TCGAbiolinks' package, which was

carried out to detect low correlated samples, termed outliers, which resulted with a count data matrix to be used in further steps.

3.7.1.2. Acquisition of Brain Transcriptomic Data from GTEx

On the R environment, corresponding normal brain data from the Recount2 project [30] of the GTEx (31, https://gtexportal.org) database were downloaded by using the *TCGAquery_recount2* function of the 'TCGAbiolinks' package. The downloaded file was a Ranged Summarized Experiment (RSE) object. By using the *scale_counts* function of the 'recount' package, count data were scaled. A total of 120 samples were downloaded and scaled from the Recount2 platform of the GTEx.

3.7.1.3. Acquisition and Generation of LGG data from GEO

The public repository of NCBI GEO DataSets [32] was thoroughly searched for LGG gene expression data. In this way, the eligible dataset Series GSE184941 was retrieved, which contains mRNA profiles of human low- and high-grade glioma samples. A total of 79 LGG out of 180 samples in the GSE184941 were analyzed in this study.

For this analysis, the GTEx RNA-Seq pipeline was followed. FASTQ files were downloaded by using the 'SRA Tool Kit' v.2.11.1 [33] with the following attributes:

--skip-technical
--readids
--dumpbase
--clip
--split-3

For the read alignment in the FASTQ files, firstly index files of GENCODE Human Reference Genome Release 39 [34] were generated by 'STAR' v.2.6 [35].

To generate 'STAR' v.2.6 index files, primary GENCODE Release 39 was provided as the reference genome by using the *genomeFastaFiles* attribute. Main annotation file was given with GTF file extension by using --*sjdbGTFfile*. Another attribute of STAR v.2.6 was also used, *sjdbOverhang 'number'*; the number was 151 and was calculated by the (Average Read Length sequenced)-1.

For the alignment of the raw RNA-Seq reads, the aligner 'STAR' v.2.6 tool was used with the following attributes:

--*runMode alignReads* to align FASTQ files to reference genome

--*readFilesCommand zcat* because FASTQ files were compressed

--*readFilesIn* to tell the path of paired-end FASTQ files

--*genomeDir* to give path of index files

--*outFileNamePrefix* to name output files

--*outSAMtype BAM SortedByCoordinate* is used to generate sorted and aligned BAM files

--*quantMode TranscriptomeSAM* is needed for further steps of the analysis.

'STAR' v.2.6 provides alignment results of each sample with many informative sections about the started/finished time, input read length, and mapped reads. Information aboutuniquely mapped reads is given in number and percentages. This also indicates the quality of each sample mapped to the reference genome. Therefore, we excluded the samples of which the uniquely mapped percentage is lower than 70%. Out of the 79 GEO-derived LGG samples,we retained 34 samples for subsequent analysis.

For gene expression levels estimation (i.e., quantification), index files created by 'RSEM' v.1.3.3 [36] were used.

To generate 'RSEM' v.1.3.3 [36] index files, the primary GENCODE Release 39 was given as the reference genome.

The *rsem-prepare-reference* function of the 'RSEM' v.1.3.3 tool was used to generate index files. The main annotation file was provided with the GTF file extension by using the *gtf* attribute.

In the quantification step by 'RSEM' tool, the following attributes of *rsem-calculate-expression* were used:

--*alignments* and --*paired-end* to specify that RNA-Seq files are aligned and paired-end

--*bam* to tell the tool that files are bam formatted

--*no-bam-output* to specify path of output files generated from STAR as input here

-*tag* to generate same-named files.

The next step was to collect all count data and combine them into a matrix file with corresponding sample ID and gene ID. To this end, the *rsem-generate-data-matrix* function of

'RSEM' was used without providing any extra options or attributes. In this way, the count data matrix of 34 LGG samples derived from GEO database were obtained.

3.7.1.4. Merging TCGA-GTEx and GEO-GTEx data

After preprocessing TCGA-derived LGG count data and obtaining the count data of GTEx, these data were merged based on the matching gene IDs. Merged TCGA-GTEx count data were further normalized. By using the function *TCGAanalyze_Normalization* of 'TCGABiolinks' by using the EDASeq protocol [37], merged TCGA-GTEx data count data were normalized for *gcContent*. Then quantile filtering was done with 25% cutoff. These steps were also done to further enrich our merged TCGA-GTEx count data.

Upon acquiring GEO-derived LGG data from 'RSEM', count data of GEO-LGG and count data derived from GTEx were merged based on matching gene IDs. Then, batchcorrection was performed by 'ComBat-seq' [38], an adjustment tool that uses negative binomialregression to detect batch effects. It accepts count data as an input and checks the data by comparing an expected distribution of the data (https://github.com/zhangyuqing/ComBat-seq).Results from 'ComBat-seq' showed that the merged GEO-GTEx count data required no batch correction.

3.7.2 Principle Component Analysis of Transcriptome Profiles

The data that are used in this study are high-dimensional. There are hundreds of columns with sample IDs and tens of thousands of rows with gene IDs. Principal Component Analysis (PCA) is applied for dimensionality reduction and allows to detect variances among different groups. PCA is a widely used method, applied also to RNA-Seq data analysis, since these are considered 'big data' and normally require great computational power to detect variances [39]. In our study, PCA was employed to examine for clustering of the input transcriptomic data and for the presence of any outliers.

Before performing PCA, firstly raw read counts of both merged TCGA-GTEx and GEO-GTEx data were normalized by converting them to FPKM values by using the *count2fpkm* function of 'RNAAgeCalc' [40].

Before performing PCA, a prefiltering process was applied on the merged FPKM TCGA-GTEx and GEO-GTEx data. The FPKM data were filtered so the expression value of

the genes was greater than ‘1’ in at least half of the samples in each group (TCGA-GTEx and GEO-GTEx). Next the merged and filtered FPKM TCGA-GTEx, as well as merged FPKM GEO-GTEx data were log2 transformed in the normalization step.

Then a base R function, *prcomp*, was used to generate PCA objects of the filtered/merged log2 transformed FPKM TCGA-GTEx and the filtered/merged log2 transformed FPKM GEO-GTEx, respectively.

To visualize the PCA plots, the *fviz_pca_ind* function of the ‘factoextra’ package (https://CRAN.R-project.org/package=factoextra) was used.

3.8. Differential Expression Analysis

After prefiltering step, outlier and clustering check by PCA, for DGEA only raw count data was provided as input to the tools. Count data with sample ids after PCA and genes after prefiltering were used in DGEA tools.

For DGES, three software tools were used, namely, ‘edgeR’ v3.34.1 [41], ‘limma’ v3.48.3 [42] and ‘DESeq2’ v1.32.0 [43]. Output differentially expressed genes (DEGs) were selected based on an absolute log2 fold-change ($|\log2FC| \geq 2$; the threshold for the False Discovery Rate (FDR)-adjusted p-value was set less than 0.05.

To visualize the results of DEGs, heatmap plots of DEGs generated by each of the three methods were created by using the ‘pheatmap’ package v1.0.12 (https://cran.r-project.org/package=pheatmap). DEGs were filtered from the filtered/merged FPKM TCGA-GTEx and the filtered/merged FPKM GEO-GTEx, and these data were used to generate the heatmap plot.

3.8.1. edgeR

The method in ‘edgeR’ uses a negative binomial distribution to analyze DEGs and accepts raw read count data as input. Herein, filtered and merged raw count data were used.

Firstly, by providing the count data as input to DGEList, we created a *DGEList* object for edgeR. Then the *calcNormFactors* function was used to normalize library sizes. Next, a model matrix was created since there were two groups (cancer vs normal). Then dispersion

estimation was made by using the *estimateDisp* function. DGEA then was performed by applying the likelihood ratio tests (LRT) by using the *glmFit* and *glmLRT* functions of 'edgeR'.

The 'edgeR' method resulted to a total of 17337 DEGs between TCGA and GTEx. After exclusion of the non-significant DEGs, 804 up-regulated and 3972 down-regulated genes remained. There were 19769 genes detected between GEO-GTEx and after exclusion there were 4117 up-regulated and 4382 down-regulated genes.

3.8.2. limma

Another package used for DGEA in this study was 'limma', which is widely used for analyzing gene expression data such as RNA-Seq data. It takes advantage of linear models to perform DEA.

For DGEA by 'limma', raw read count data were provided as input as in the case of 'edgeR'. Firstly, a *DGElist* object was created of count data. Then the *calcNormFactors* function of edgeR was used. Then, the *voom* function of limma was used, which is based on an approach called precision weights. Next, the *lmfit* function was applied with designed/model matrix, followed by the *eBayes* function.

A total of 17341 DEGs between TCGA and GTEx were detected by 'limma'. After exclusion of the non-significant ones, there were 709 up-regulated and 4475 down-regulated genes. On the other hand, between GEO and GTEx, we had a total of 19568 DEGs. After exclusion there were 4861 up-regulated and 4090 down-regulated genes.

3.8.3. DESeq2

The third package, 'DESeq2', uses a negative binomial liner model to perform DGEA. It also accepts only normalized count data as input for the analysis.

Firstly, the *DESeqDataSetFromMatrix* function was used with the model/design matrix. Then *DESeq* function was used and later DEGs were extracted.

DESeq2 detected a total of 17171 DEGs between TCGA and GTEx. After exclusion of the not significant DEGs, there were 580 up-regulated and 4332 down-regulated DEGs. On the

other hand, 19812 DEGs were detected between GEO and GTEx. After filtration, there were 4203 up-regulated and 4350 down-regulated DEGs.

The next step of this study was to perform Weighted Gene Co-Expression Network Analysis [44]. For this analysis, only the common DEGs detected by 'edgeR', 'limma' and 'DESeq2' and the samples derived from TCGA and GEO were used.

3.9. Weighted Correlation Network Analysis

Weighted Correlation Network Analysis (WCNA) is commonly used for gene expression correlation network analysis. For instance, it is used to identify highly correlated gene expression patterns, modules or hub genes among genes or clusters. Furthermore, the detected correlated genes are usually considered of biological significance and can be used to identify candidate biomarkers or therapeutic targets.

3.9.1. Weighted Gene Co-Expression Network Analysis (WGCNA)

In this study, the 'WGCNA' R package v1.70-2 [44] and its guide/workflow was used to find highly correlated genes, clusters, modules and modules-LGG clinical trait relationships.

3.9.1.1 WGCNA Data Input

For 'WGCNA', the FPKM matrix data of the detected DEGs from TCGA and GEO were used.

To normalize the FPKM data, the FPKM data matrices of TCGA and GEO were log2(data+1) transformed. To convert the Ensembl IDs into the official genes symbols, the *getBM* function in the 'biomaRt' package and the 'org.HS.eg.db' package for genome annotation for human were used. Those Ensembl IDs that had no corresponding gene symbol were discarded.

Then, the normalized FPKM data with gene symbols were checked with the function *goodSamplesGenes* of the 'WGCNA' package for possible outliers. No outliers were detected for both TCGA and GEO samples.

3.9.1.2. Clinical Trait Data Acquisiton

Clinical data of samples derived from TCGA were acquired using the *GDCprepare* function of 'TCGABiolinks'. The clinical traits of interest from TCGA were Primary Diagnosis, Age at Diagnosis, Vital Status, Sample Type, Site of Resection Biopsy, Prior Treatment, Gender, Race, and Tissue Organ Origin.

Clinical data of samples derived from GEO were downloaded by using the website of GEO database (https://www.ncbi.nlm.nih.gov/geo/). Selected clinical traits for GEO derived samples were Age, Progressed Tumor Pathology, IDH Mutant Status, Vital Status, Gender, Initial Tumor Pathology, and Patient Initial Grade.

3.9.2. Network Construction and Module Detection

3.9.2.1. Analysis of Network Topology and Soft-Thresholding Power

To construct a weighted co-expression network and calculate adjacency, a soft thresholding power is needed. It is actually a mathematical value to power the correlation of genes to the chosen threshold. To do this, as suggested in the guide/workflow, the *pickSoftThreshold* function of 'WGCNA' was used by providing FPKM data with DEGs to the function. This function conducts network topology analysis and enable the selection of the correct soft-thresholding power. This function returns a Scale Independence plot and a Mean Connectivity plot. The soft-thresholding power of TCGA and GEO was six and four, respectively.

After determining the correct soft-thresholding power, the next step was the calculation of adjacencies by using that power. For this purpose, the *adjacency* function of the 'WGCNA' package was used by also providing the soft-thresholding power to the function.

3.9.2.2. Topological Overlap Matrix (TOM)

To eliminate any noise, the adjacency matrix was transformed to Topological Overlap Matrix (TOM) and then dissimilarity was calculated by using the 'WGCNA' functions, '*TOMsimilarity(adjacency)*' and dissTOM = 1 – TOM. In this way, only highly correlated genes are grouped together.

Next, by using hierarchical clustering, trees of genes (i.e., dendograms) were plotted.

3.9.2.3. Module Identification

The next step was the detection of modules in the dendrograms. For this purpose, the *cutreeDynamic* function of 'WGCNA' was used. Since highly correlated genes were detected in the previous step, those genes were then assigned into color-coded modules.

3.9.2.4. Merging Modules

After the identification of highly correlated genes within modules, with their assigned colors, modules with similar expression profiles were detected. Since highly co-expressed genes display similar module profiles, they were merged, as suggested by the workflow of 'WGCNA'.

To this end, the module eigenenes of modules were calculated to estimate co-expression similarity and then were clustered again. For this purpose, the functions '*moduleEigengenes*', '*cor*', again '*hclust*' and '*mergeCloseModules*' of the 'WGCNA' package were used. The plot of these modules was generated by using the *plotDendroAndColors* function of 'WGCNA'.

3.9.2.5. Association of Modules with LGG Clinical Traits

The relationship between the gene modules and LGG clinical traits were investigated. By using eigengene information of each module, external traits were correlated and the most significant associations were found. To this end, traits without numbers were converted to their mathematical abbreviations by using the *cor* and *corPvalueStudent* functions. Heatmaps of this analysis were generated by using the *labeledHeatmap* function, which include the Spearman correlation coefficient of the modules with clinical traits and the corresponding p-value. Those modules significantly correlated with clinical traits based on a correlation coefficient above 0.3 and p-value less than 0.05 were selected. The genes within these selected modules were used for gene co-expression and protein-protein interaction network analysis.

3.9.2.6. Visualization of Gene Networks

To visualize the weighted gene co-expression networks, heatmap plots were generated based on these networks. By each row and column corresponding to a gene and sample, respectively, heatmaps of the gene networks were generated by using the *TOMplot* function of 'WGCNA'.

3.9.3. Protein-Protein Interaction Network and Detection of Hub Genes

Two types of networks were generated for this study, gene co-expression and protein-protein interaction networks. Cytoscape v3.9.1 [45] platform, an open source program, was used for network analysis and visualization.

To this end, the adjacency matrix (co-expression data matrix) of genes of the selected modules were filtered by setting the gene co-expression value of 0.5 as threshold, so as to increase the robustness of the study.

Protein-Protein Interaction (PPI) network analysis was performed by providing as input the genes of the co-expression data matrix to the database Search Tool for the Retrieval of Interacting Genes (STRING) v1.7.0 [46]. The interactions of the corresponding gene products were investigated and a protein-protein interaction network was generated.

The gene-gene and protein-protein association data were uploaded to the Cytoscape and the Cytohubba [47] plugin of Cytoscape which allows network investigation by twelve different ranking algorithms: Degree method (Deg), Maximum Neighborhood Component (MNC), Density of Maximum Neighborhood Component (DMNC), Maximum Clique Centrality (MCC), Closeness (Clo), EcCentricity (EC), Radiality (Rad), BottleNeck (BN), Stress (Str), Betweenness (BC), Edge Percolated Component (EPC). In this way, the most highly connected/correlated genes/proteins in the gene co-expression and PPI networks were selected by using the node-degree filter.

3.10. Overall Survival Analysis

The prognostic value of the detected ‘hub’ genes was explored (see 3.9.3). To assess whether the expression levels of these genes are associated with the overall survival (i.e., a person is either alive or dead) of LGG patients, the web-based tool GEPIA (Gene Expression Profiling Interactive Analysis) [48, 49] version 2 (http://gepia2.cancer-pku.cn/) was applied. GEPIA2 retrieves and analyzes survival data from the integrated TCGA pan-cancer clinical data resource [50]. The LGG patient cohorts were divided into ‘high-risk’ and ‘low-risk’, by setting the cutoff values for high and low gene expression at 50%.

3.11. Limitations of the Study

There were hundreds of samples and tens of thousands of genes. Therefore, powerful computational processes and hence powerful hardware were required.

3.12. Ethics Committee Approval

Izmir Biomedicine and Genome Center Noninvasive Researches Ethics Committee Approval with the protocol number “2021-038” was granted for this thesis on the 21st November 2021.

4. RESULTS

4.1. Transcriptomics Profies

In the PCA plots in Figure 2, there are two distinct clusters which represent transcriptomic data of the LGG samples derived from TCGA database and normal brain data derived from GTEx database. In the generated plot, the principle components PC1 and PC2 represent the 36.4% and 8.4% of the entire data, respectively.

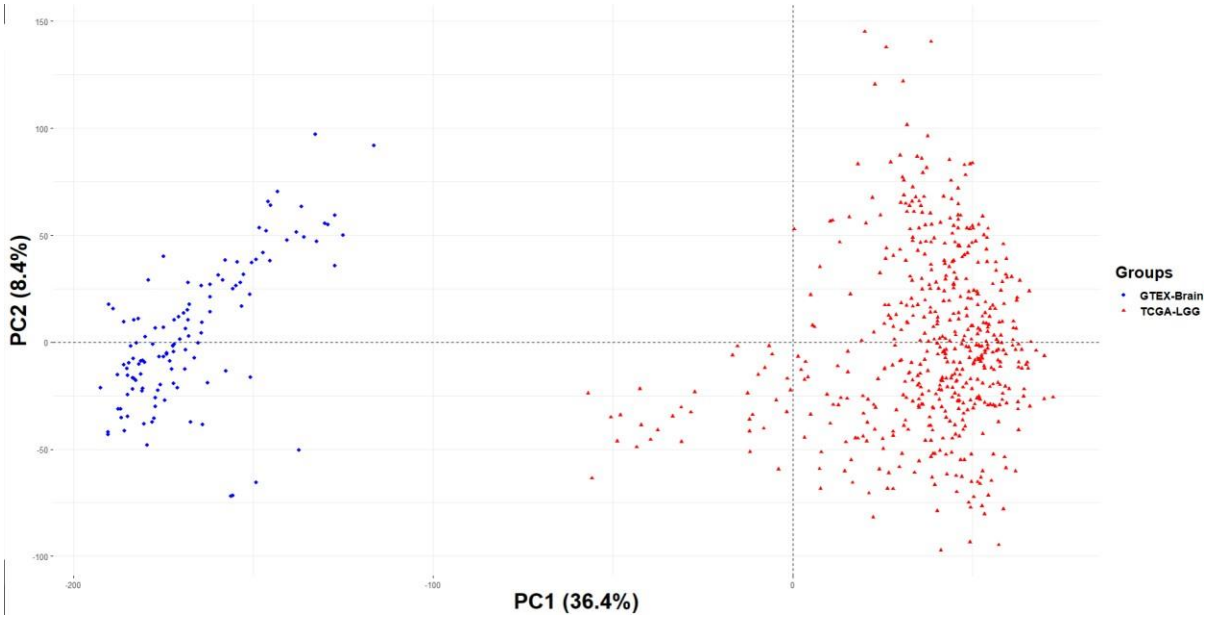


Figure 2. Principal Component Analysis of TCGA-GTEx gene expression data. The x-axis represents PC1 and the y-axis represents PC2. Red dots indicate the TCGA samples and blue dots indicate the GTEx samples.

In the PCA plot in Figure 3, there are two discernible clusters which correspond to gene expression data of the LGG samples derived from the GEO database and normal brain tissue samples derived from the GTEx database. In the plot, PC1 and PC2 represent the 45.6% and 14.8% of the whole data, respectively.

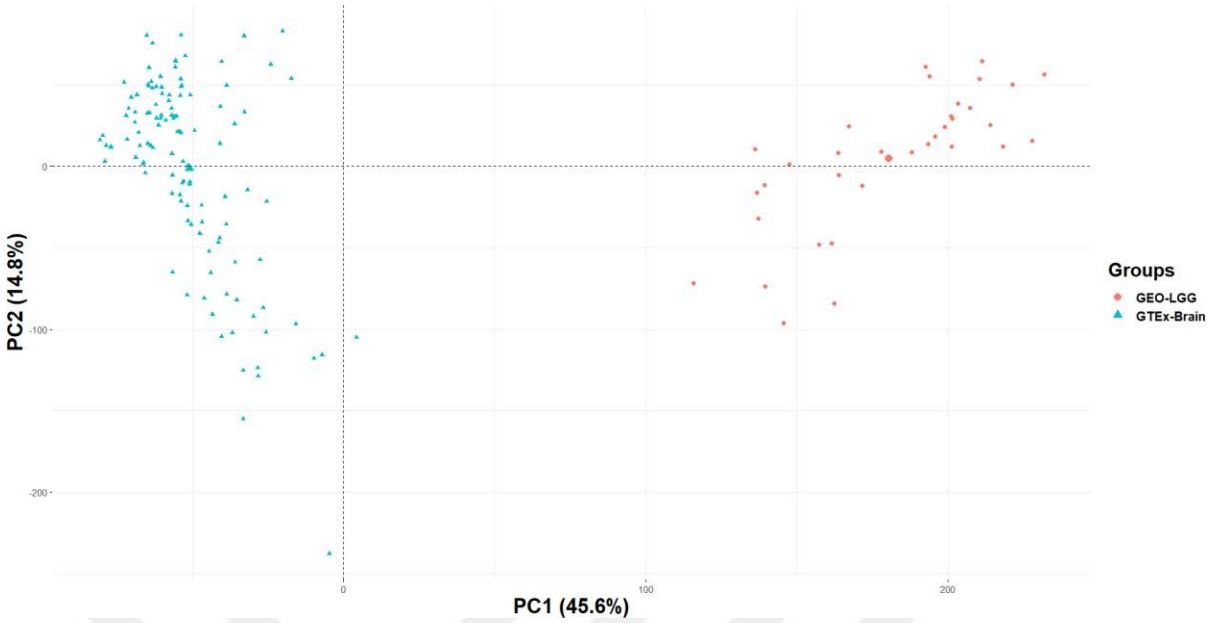


Figure 3. Principal Component Analysis of GEO-GTEx transcriptomic data. The x-axis represents PC1 and the y-axis represents PC2. Red dots indicate the GEO samples and blue dots indicate the GTEx samples.

4.2. Differential Gene Expression Analysis

Analysis of differential gene expression between TCGA-GTEx and GEO-GTEx was performed by applying three different methods implemented in the software tools ‘edgeR’, ‘DESeq2’ and ‘limma’. The volcano plots and heatmaps of DEGs were generated. Thestatistical results and generated plots are described below.

4.2.1. DGEA between TCGA and GTEx

4.2.1.1. edgeR

DGEA by ‘edgeR’ between TCGA and GTEx samples resulted to a total of 17337 genes. When logFC and FDR were taken into consideration, 804 up-regulated and 3972 down-regulated genes were detected.

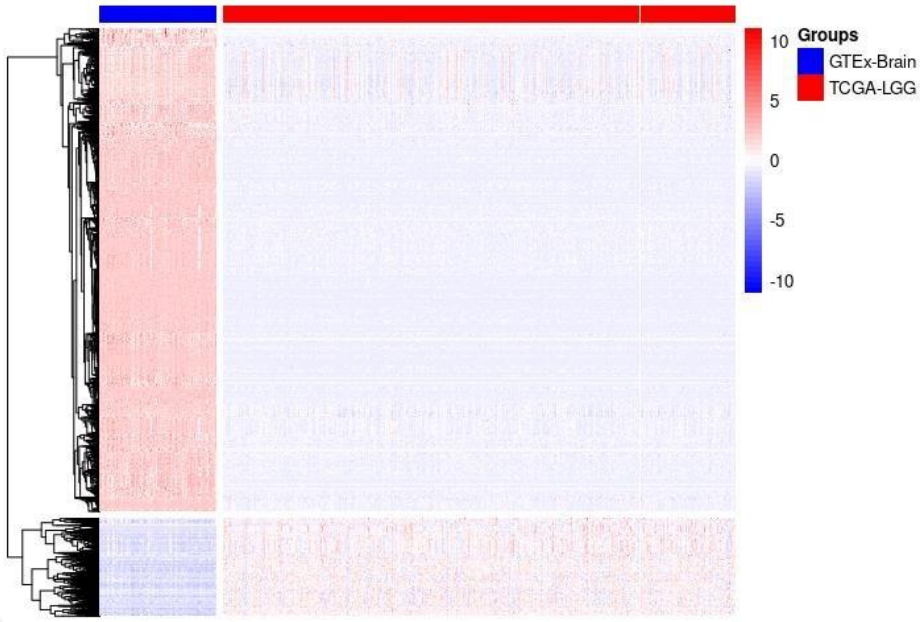


Figure 4. Heatmap of DEGs detected by edgeR between TCGA-LGG and GTEx-Brainsamples. Each row represents a DEG. Up-regulated and down-regulated genes are indicated byred and blue color, respectively. A tree depicting the clustering of the genes is shown on the left. GTEx-Brain and TCGA-LGG samples are shown in red and blue color, respectively.

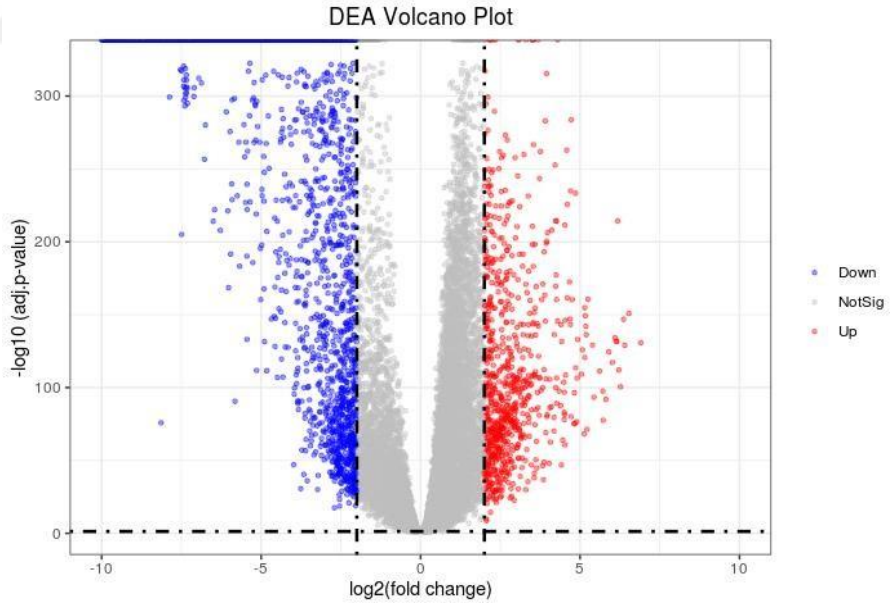


Figure 5. Volcano Plot of DEGs detected by edgeR between TCGA-LGG and GTEx-Brain samples. The x-axis represents log2(Fold Change) and the y-axis represents –log10(adjusted p-value). A dotted line parallel to the x-axis indicates the number 1.301 or –log10(0.05). A dotted line parallel to the y-axis indicates +2 and -2 fold-change thresholds. Up-regulated and down-regulated genes are denoted by red and blue points, respectively; the not significantly DEGs are denoted by gray points.

4.2.1.2. DESeq2

DGEA by 'DESeq2' between TCGA and GTEx samples resulted to a total of 17171 genes. When logFC and FDR were considered, 580 up-regulated and 4332 down-regulated genes detected.

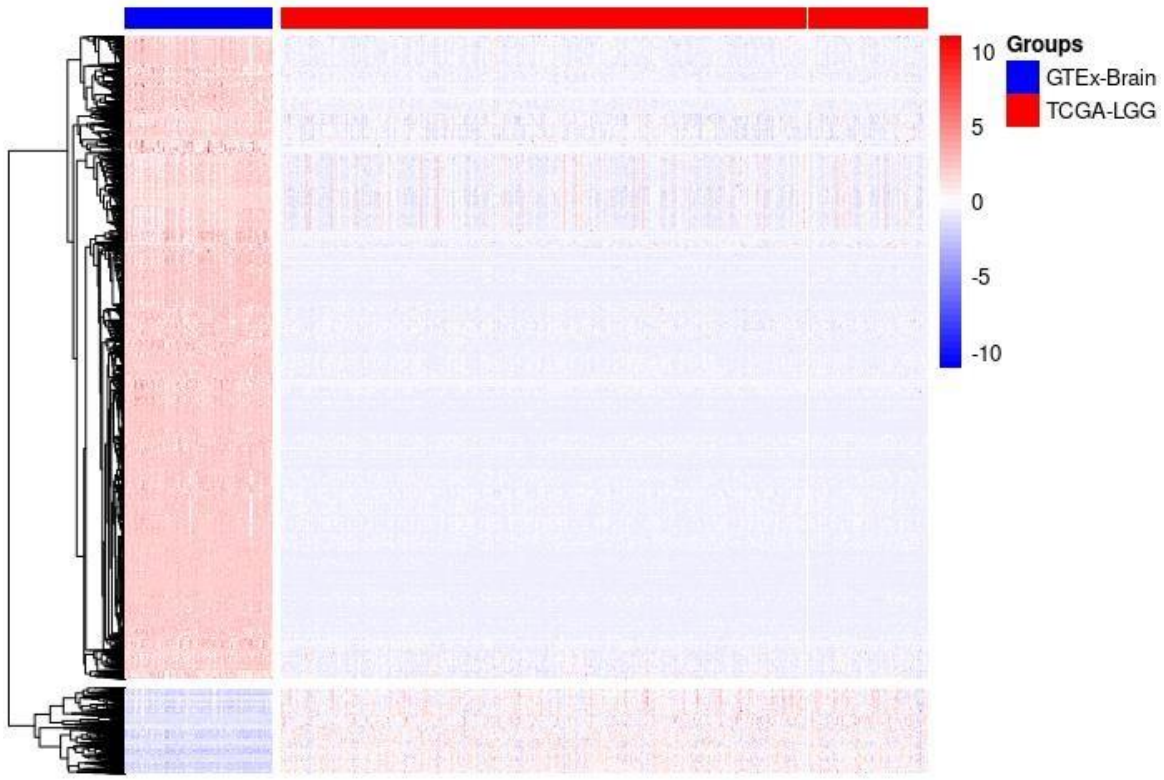


Figure 6. Heatmap of DEGs detected by DESeq2 between TCGA-LGG and GTEx-Brain samples. Each row represents a DEG. Up-regulated and down-regulated genes are indicated by red and blue color, respectively. A tree depicting the clustering of the genes is shown on the left. GTEx-Brain and TCGA-LGG samples are shown in red and blue color, respectively.

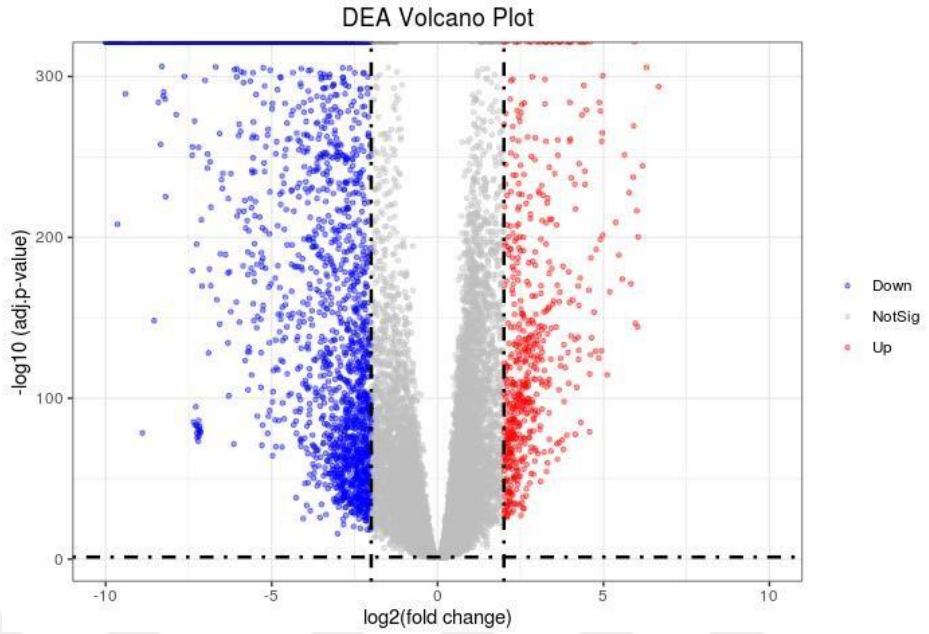


Figure 7. Volcano Plot of DEGs detected by DESeq2 between TCGA-LGG and GTEx-Brain samples. The x-axis represents log2(Fold Change) and the y-axis represents –log10(adjusted p-value). A dotted line parallel to the x-axis indicates the number 1.301 or –log10(0.05). A dotted line parallel to y-axis indicates +2 and -2 fold-change thresholds. Up-regulated and down-regulated genes are denoted by red and blue points, respectively; the not significantly DEGs are denoted by gray points.

4.2.1.3. limma

DGEA by 'limma' between TCGA and GTEx samples resulted with a total of 17341 genes. When logFC and FDR were taken into consideration, 709 up-regulated and 4475 down-regulated genes detected.

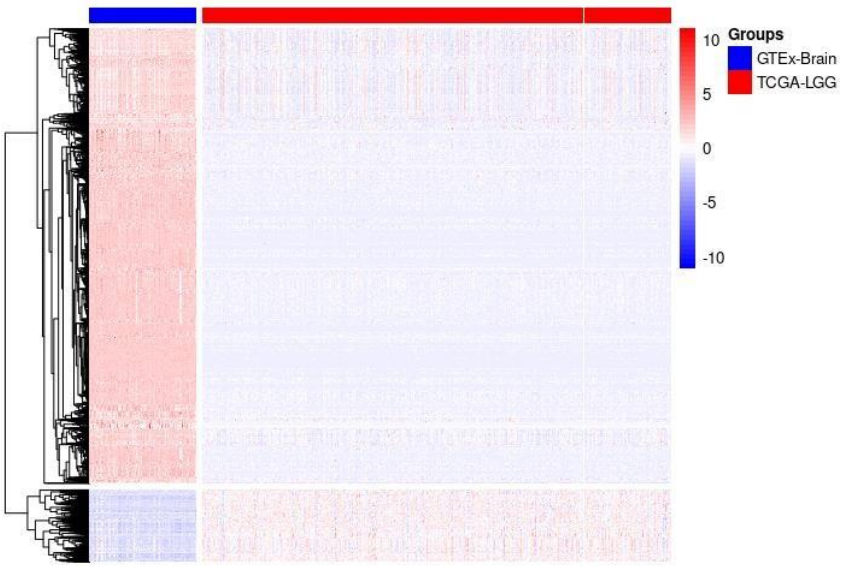


Figure 8. Heatmap of DEGs detected by limma between TCGA-LGG and GTEx-Brain samples. Each row represents a DEG. Up-regulated and down-regulated genes are indicated by red and blue color, respectively. A tree depicting the clustering of the genes is shown on the left. GTEx-Brain and TCGA-LGG samples are shown in red and blue color, respectively.

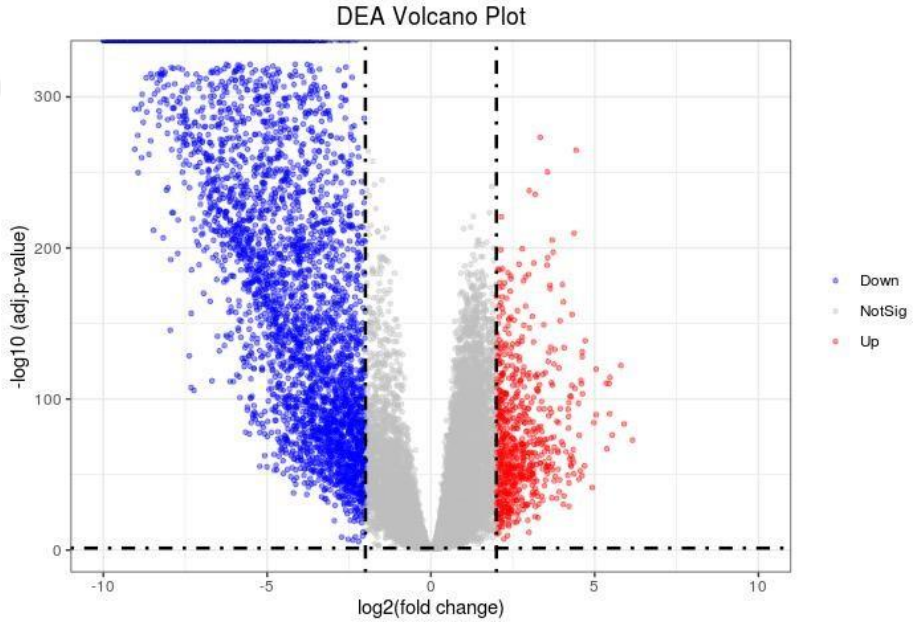


Figure 9. Volcano Plot of DEGs detected by limma between TCGA-LGG and GTEx-Brain samples. The x-axis represents log2(Fold Change) and the y-axis represents –log10(adjusted p-value). A dotted line parallel to the x-axis indicates the number 1.301 or –log10(0.05). A dotted line parallel to the y-axis indicates +2 and -2 fold-change thresholds. Up-regulated and down-regulated genes are denoted by red and blue points, respectively; the not significantly DEGs are denoted by gray points.

4.2.2. DEA between GEO and GTEx

4.2.2.1. edgeR

DEA by edgeR between GEO and GTEx samples resulted to a total of 19789 genes. When logFC and FDR were considered, 4117 up-regulated and 4382 down-regulated genes were detected.

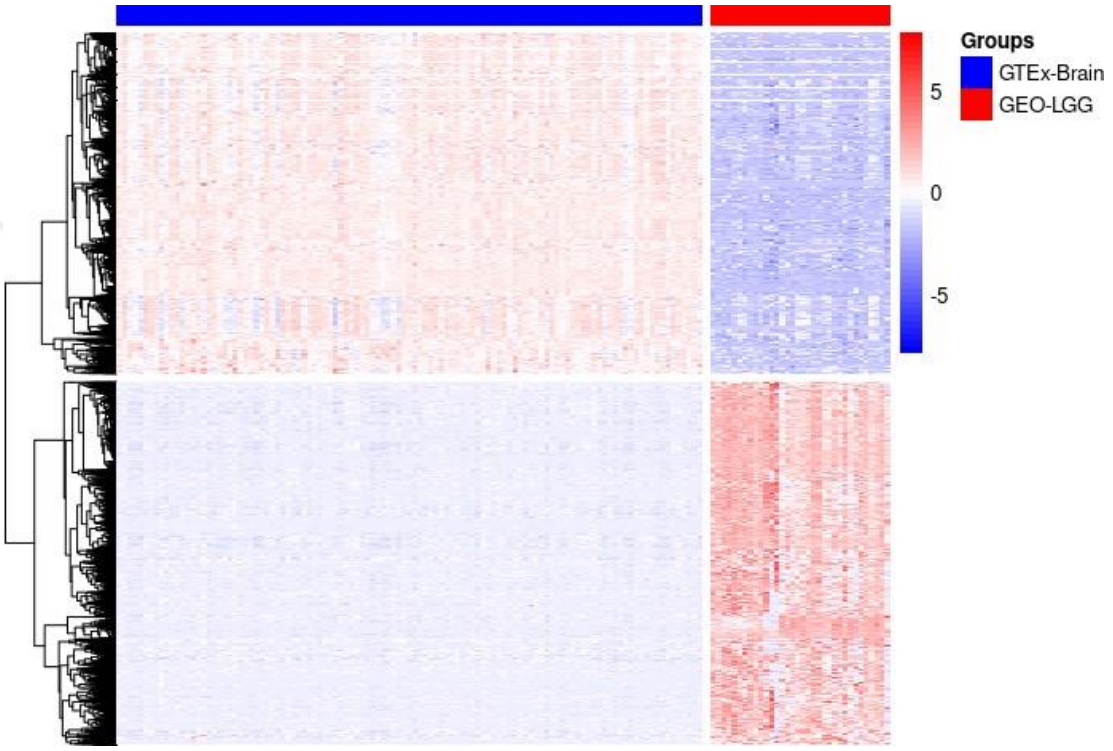


Figure 10. Heatmap of DEGs detected by edgeR between GEO-LGG and GTEx-Brain samples. Each row represents a DEG. Up-regulated and down-regulated genes are indicated by red and blue color, respectively. A tree depicting the clustering of the genes is shown on the left. GTEx-Brain and TCGA-LGG samples are shown in red and blue color, respectively.

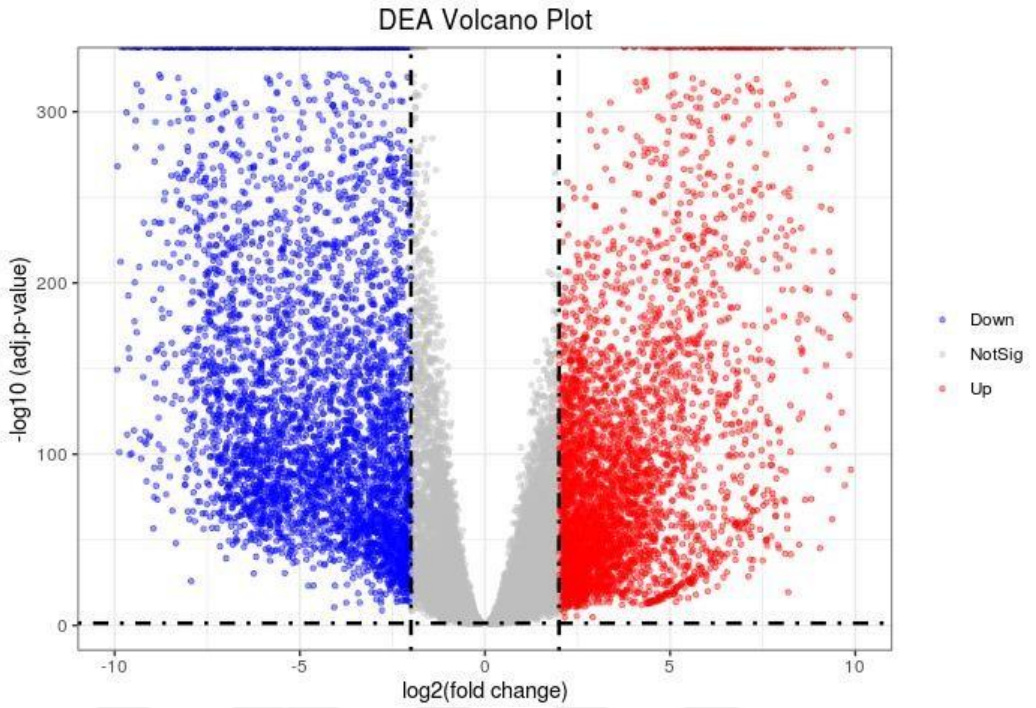


Figure 11. Volcano Plot of DEGs detected by edgeR between GEO-LGG and GTEx-Brain samples. The x-axis represents log2(Fold Change) and the y-axis represents –log10(adjusted p-value). A dotted line parallel to the x-axis indicates the number 1.301 or –log10(0.05). A dotted line parallel to the y-axis indicates +2 and -2 fold-change thresholds. Up-regulated and down-regulated genes are denoted by red and blue points, respectively; the not significantly DEGs are denoted by gray points.

4.2.2.2. DESeq2

DGEA by DESeq2 between GEO and GTEx samples resulted to a total of 19812 genes. When logFC and FDR were taken into account, there were 4203 up-regulated and 4350 down-regulated genes.

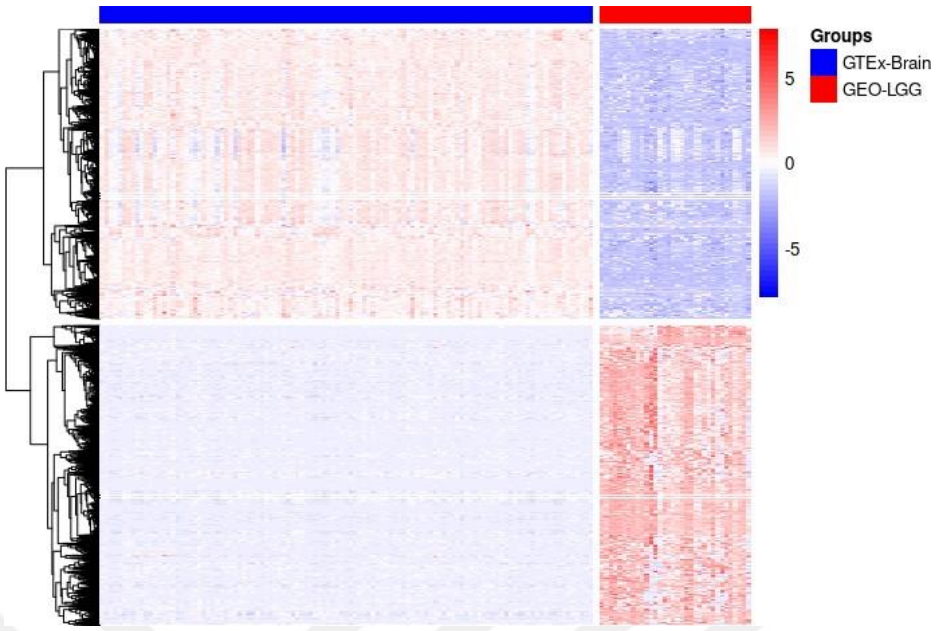


Figure 12. Heatmap of DEGs detected by DESeq2 between GEO-LGG and GTEx-Brain samples. Each row represents a DEG. Up-regulated and down-regulated genes are indicated by red and blue color, respectively. A tree depicting the clustering of the genes is shown on the left. GTEx-Brain and TCGA-LGG samples are shown in red and blue color, respectively.

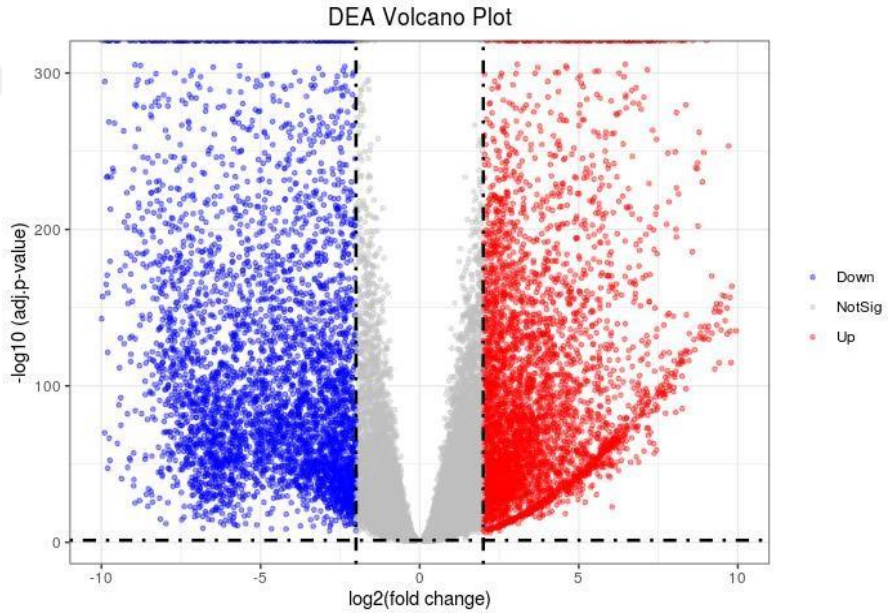


Figure 13. Volcano Plot of DEGs detected by DESeq2 between GEO-LGG and GTEx-Brain samples. The x-axis represents log2(Fold Change) and the y-axis represents –log10(adjusted p-value). A dotted line parallel to the x-axis indicates the number 1.301 or –log10(0.05). A dotted line parallel to the y-axis indicates +2 and -2 fold-change thresholds. Up-regulated and down-regulated genes are denoted by red and blue points, respectively; the not significantly DEGs are denoted by gray points.

4.2.2.3. limma

DGEA by 'limma' between GEO and GTEx samples resulted to a total of 19568 genes. When logFC and FDR were taken into consideration, 4862 up-regulated and 4090 down-regulated genes detected.

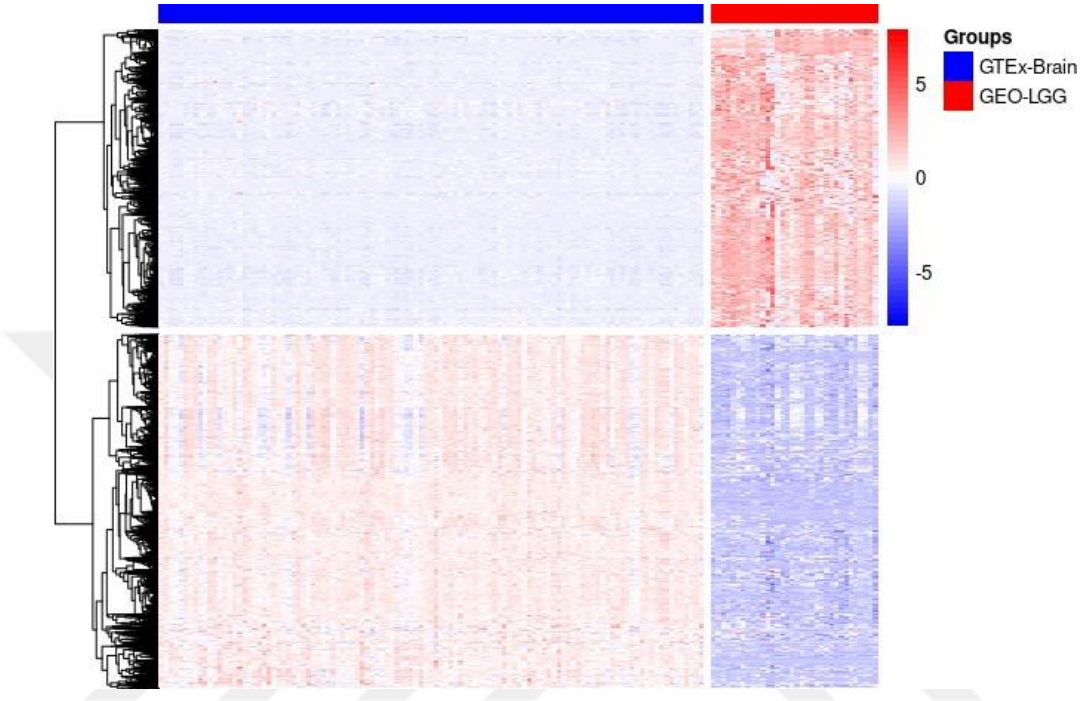


Figure 14. Heatmap of DEGs detected by limma between GEO-LGG and GTEx-Brain samples. Each row represents a DEG. Up-regulated and down-regulated genes are indicated by red and blue color, respectively. A tree depicting the clustering of the genes is shown on the left. GTEx-Brain and TCGA-LGG samples are shown in red and blue color, respectively.

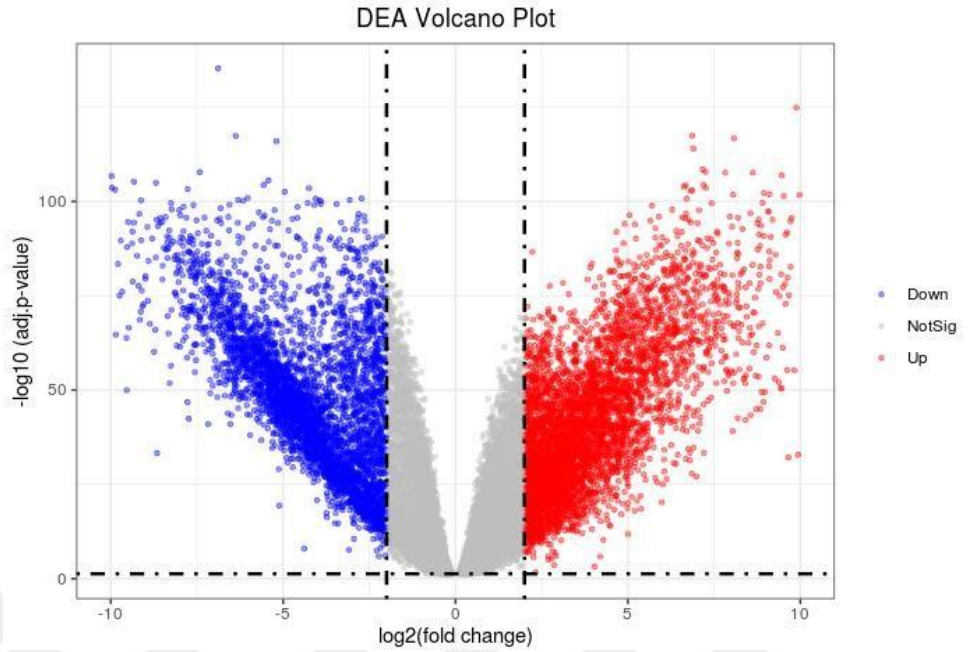


Figure 15. Volcano Plot of DEGs detected by limma between GEO-LGG and GTEx-Brain samples. The x-axis represents log2(Fold Change) and the y-axis represents –log10(adjusted p-value). A dotted line parallel to the x-axis indicates the number 1.301 or –log10(0.05). A dotted line parallel to the y-axis indicates +2 and -2 fold-change thresholds. Up-regulated and down-regulated genes are denoted by red and blue points, respectively; the not significantly DEGs are denoted by gray points.

4.2.3. Common DEGs betwenn TCGA and GTEx

DGEA between TCGA and GTEx by ‘edgeR’, ‘DESeq2’ and ‘limma’ resulted to a total of common 4496 genes, where 531 of these genes were up-regulated and 3965 of these genes were down-regulated.

DSEA between GEO and GTEx by ‘edgeR’, ‘DESeq2’ and ‘limma’ resulted to a total of common 8049 genes, where 4000 of these genes were up-regulated and 4049 of these genes were down-regulated.

Of particular note, the common DEGs between TCGA and GTEx represent 60% of the TCGA DEGs.

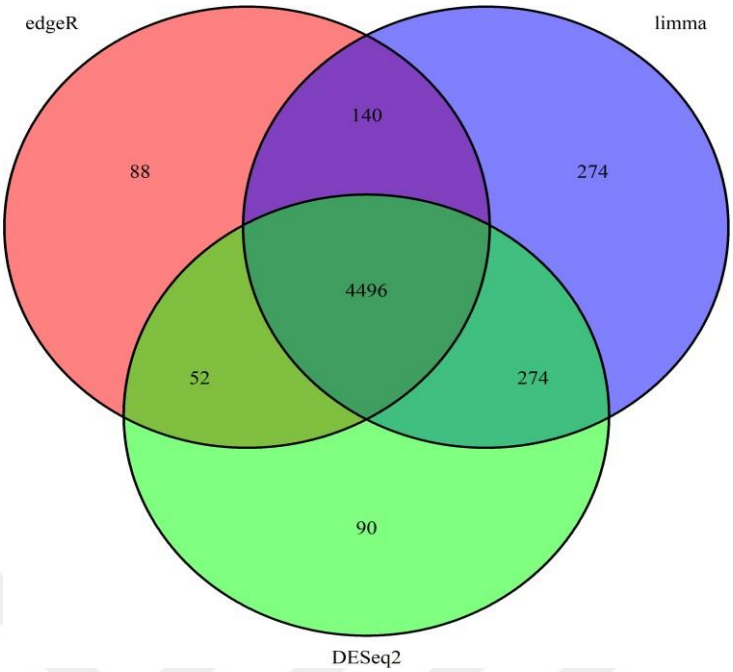


Figure 16. Venn diagram of the number of DEGs between TCGA and GTEx detected by edgeR, DESeq2 and limma. Common, shared and unique DEGs are indicated; edgeR, DESeq2and limma are represented by red, green and blue color, respectively.

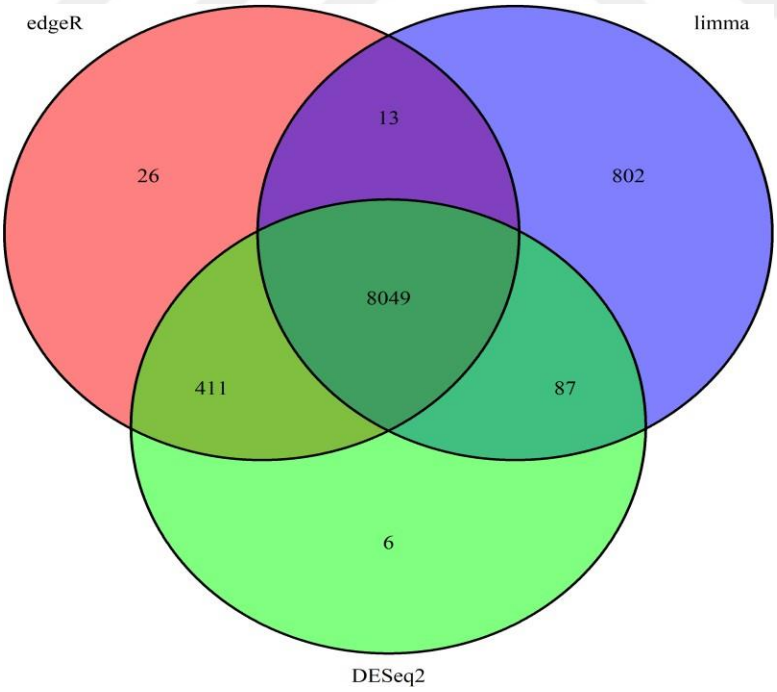


Figure 17. Venn diagram of the number of DEGs between GEO and GTEx detected by edgeR, DESeq2 and limma. Common, shared and unique DEGs are indicated; edgeR, DESeq2 and limma are represented by red, green and blue color, respectively.

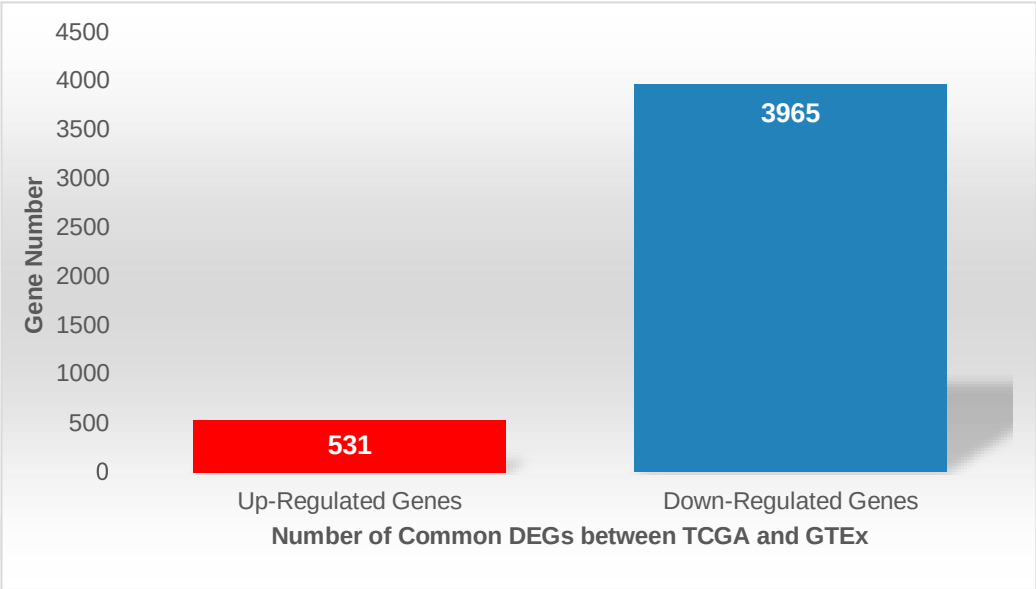


Figure 18. Bar plot of the number of common DEGs detected by edgeR, DESeq2 and limma between TCGA and GTEx samples. There are 531 up-regulated and 3965 down-regulated genes. Up-regulated genes are represented by red color and down-regulated genes are represented by blue color.

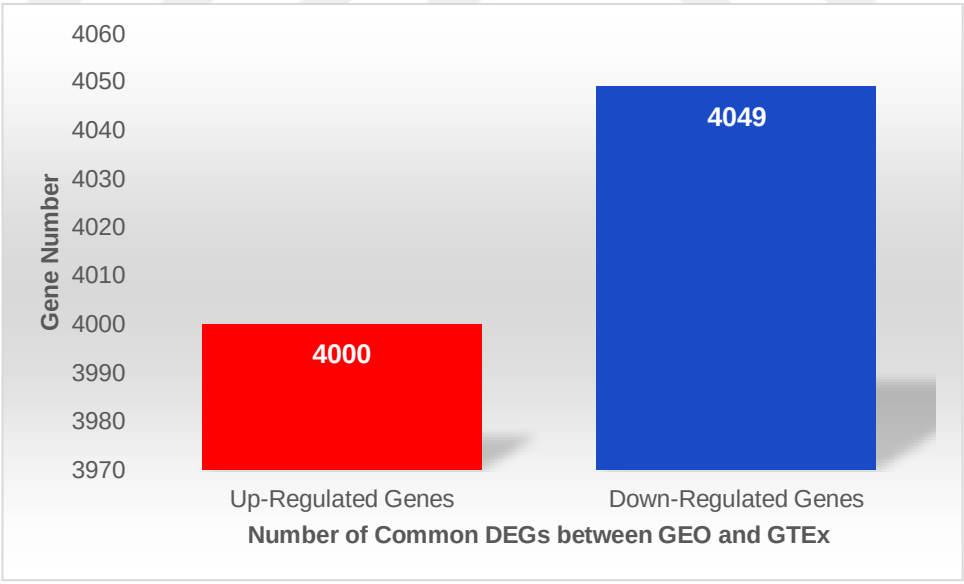


Figure 19. Bar plot of the number of common DEGs detected by edgeR, DESeq2 and limma between GEO and GTEx samples. There are 4000 up-regulated and 4049 down-regulated genes. Up-regulated genes are represented by red color and down-regulated genes are represented by blue color.

4.3. WGCNA

4.3.1. Selection of Soft-Thresholding Power

The scale independence and mean connectivity plots and values were generated by the given TCGA and GEO sample data. Power values were determined by the number below the set 0.8 limit, as shown in the Scale Independence plots (Figure 20 and 22). The power for TCGA was selected as 6 and for GEO as 4 by WGCNA.

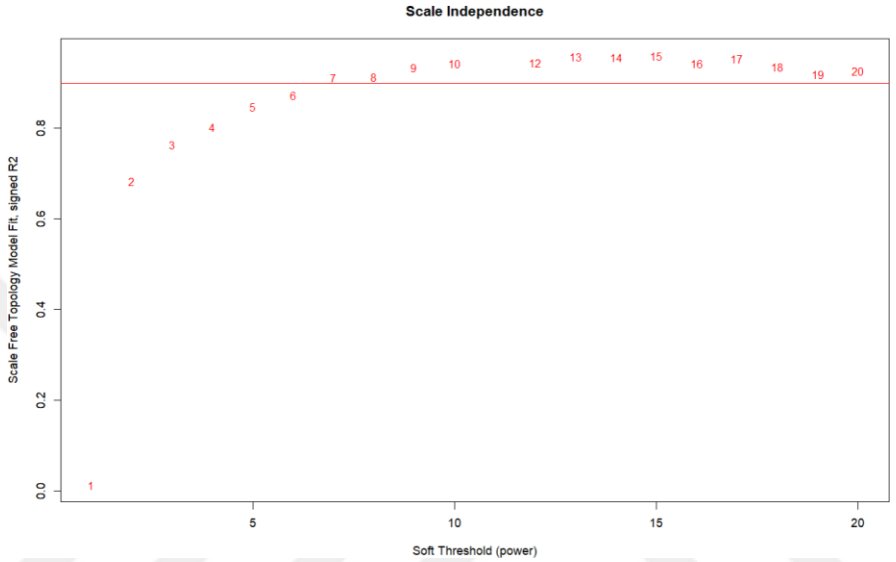


Figure 20. Scale Independence plot of TCGA. Based on TCGA-derived common DEGs, the power value was determined as 6.

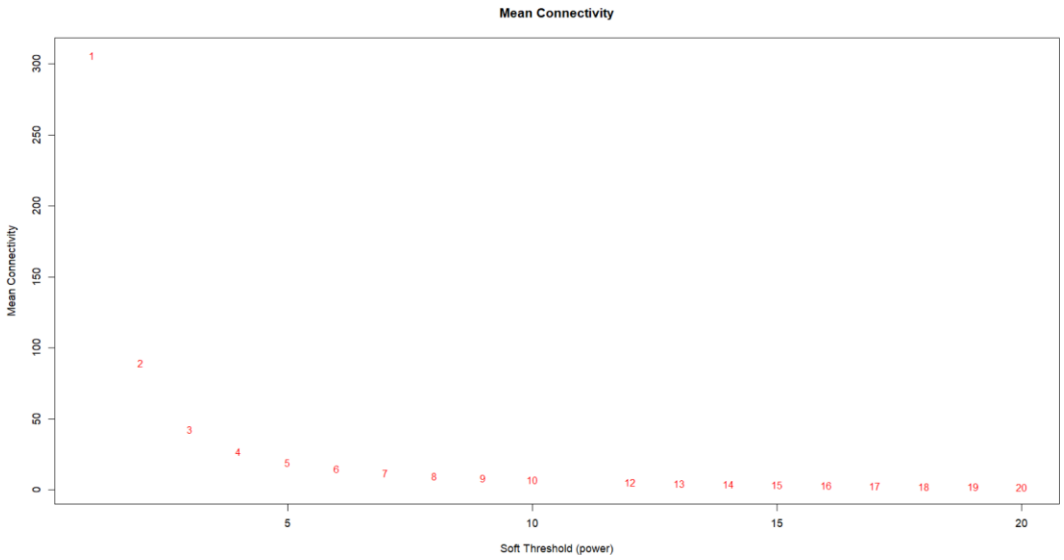


Figure 21. Mean Connectivity plot of TCGA. Various soft-thresholding powers were tested for mean connectivity.

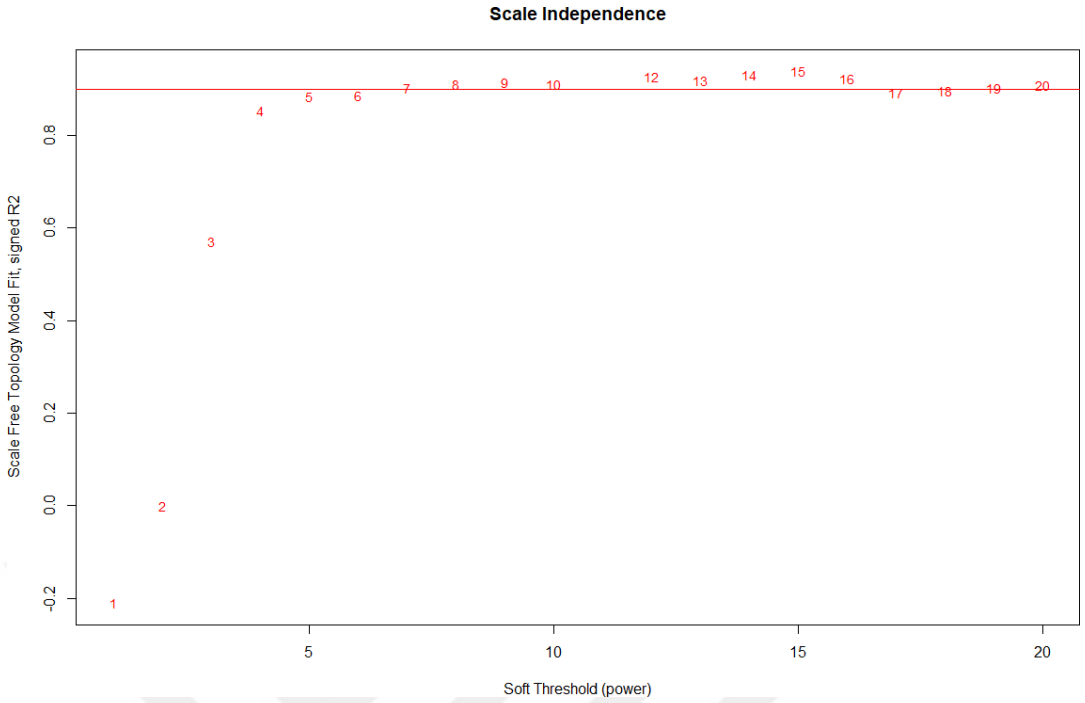


Figure 22. Scale Independence plot of GEO. Based on GEO-derived common DEGs, the power value was determined as 4.

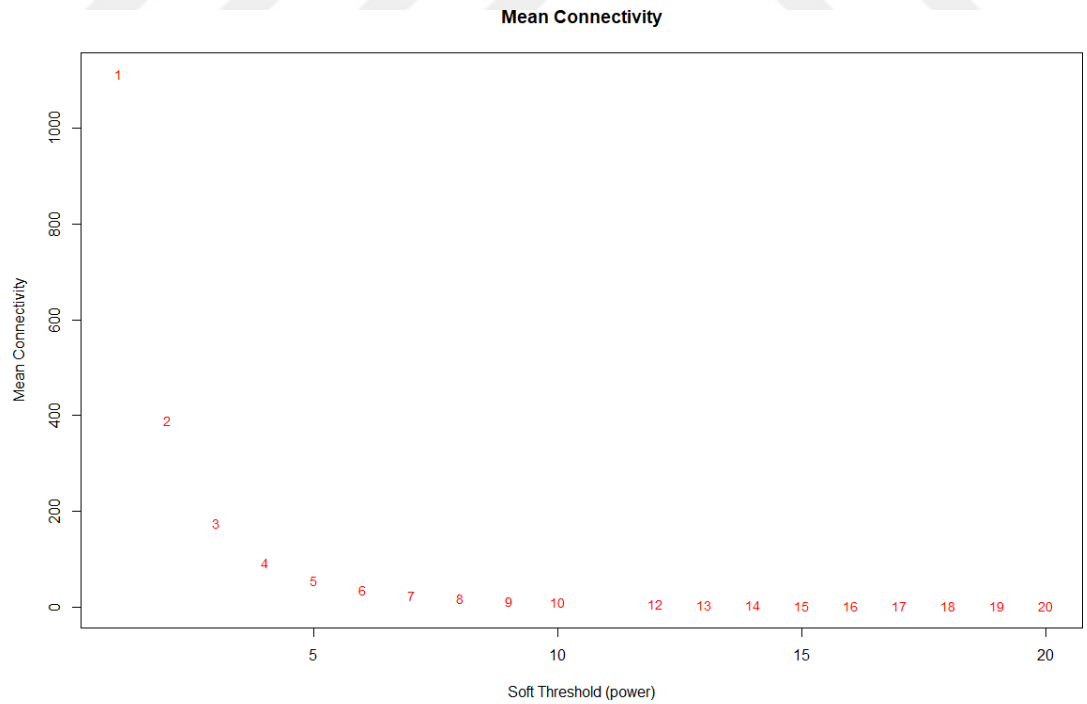


Figure 23. Mean Connectivity plot of GEO. Various soft-thresholding powers were tested for mean connectivity.

4.3.2. Identification of Modules

Genes were clustered and modules were detected. A color was assigned to each module. Genes with higher similarities or co-expressed genes have a higher probability to be represented by the same module. Gray module contains the unassigned genes.

For TCGA data, WGCNA detected seven modules. For GEO data, WGCNA detected 21 modules.

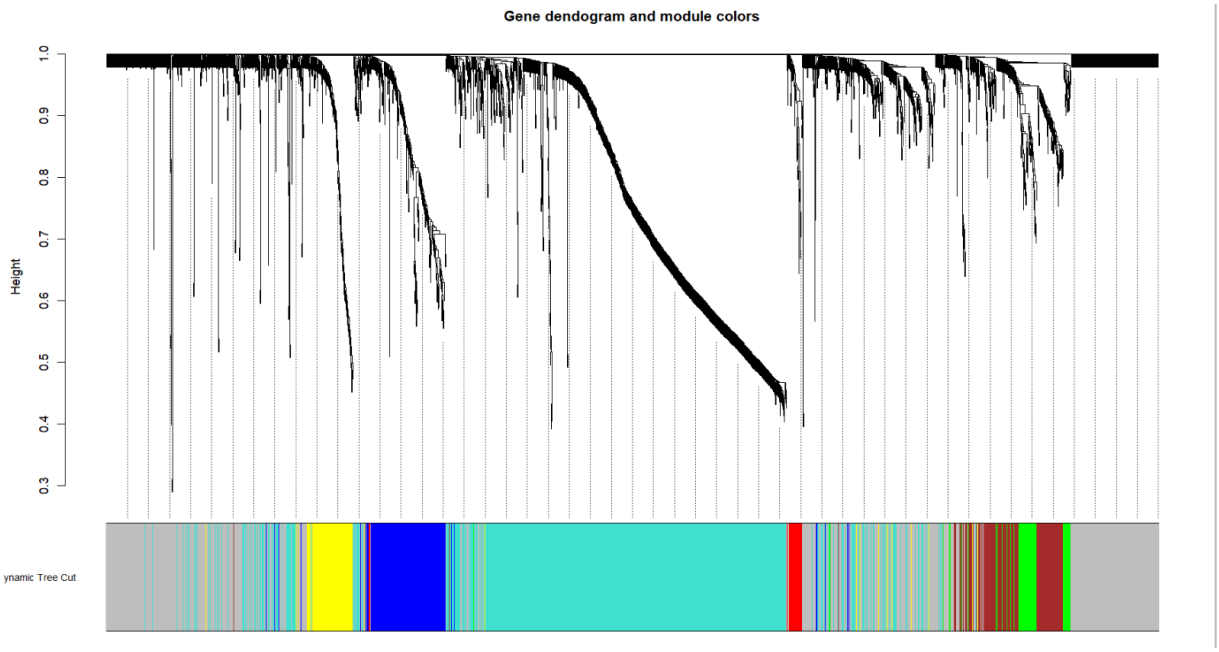


Figure 24. Gene Dendogram and Module Colors of TCGA DEGs. The upper part represents the gene dendogram tree that was generated with dissimilarity based on topological overlap. The lower part depicts the modules and assigned module colors. There are overall seven modules.

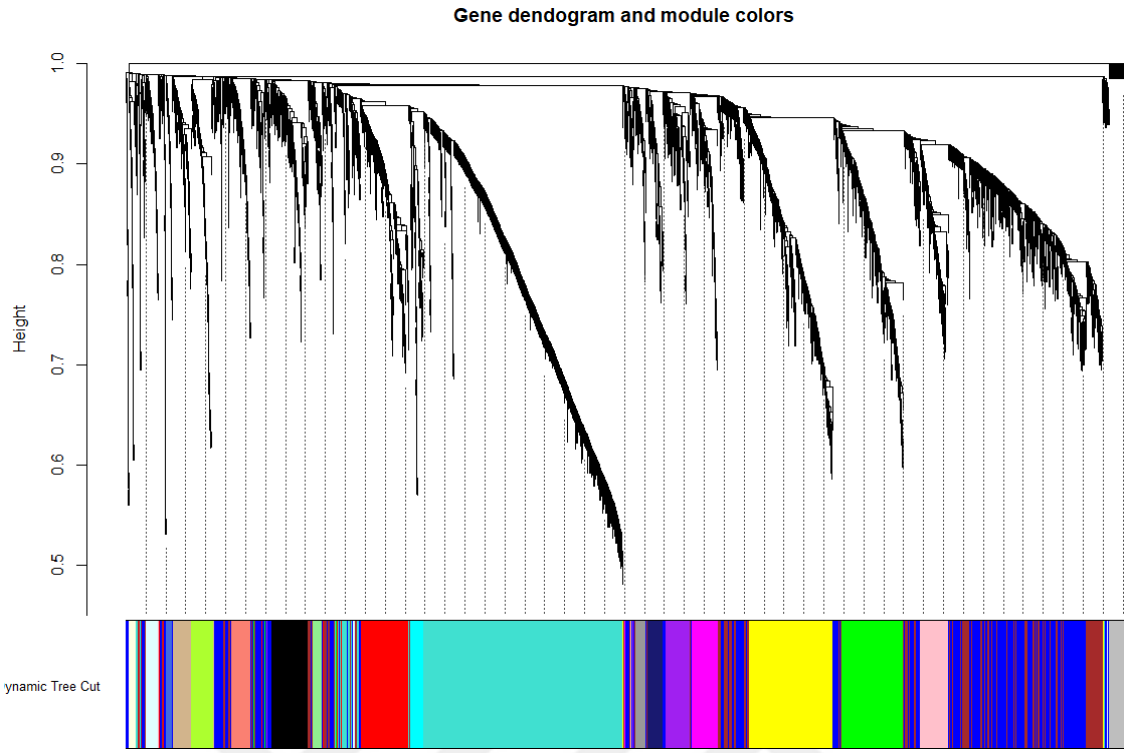


Figure 25. Gene Dendogram and Module Colors of GEO DEGs. The upper part represents the gene dendogram tree that was generated with dissimilarity based on topological overlap. The lower part depicts the modules and assigned module colors. There are overall 21 modules.

4.3.3. Module Eigengene Identification and Clustering

A dendogram tree of modules was generated. Any modules out of the tree height threshold, were merged with the upper modules within the tree. The dissimilarity threshold was 0.25.

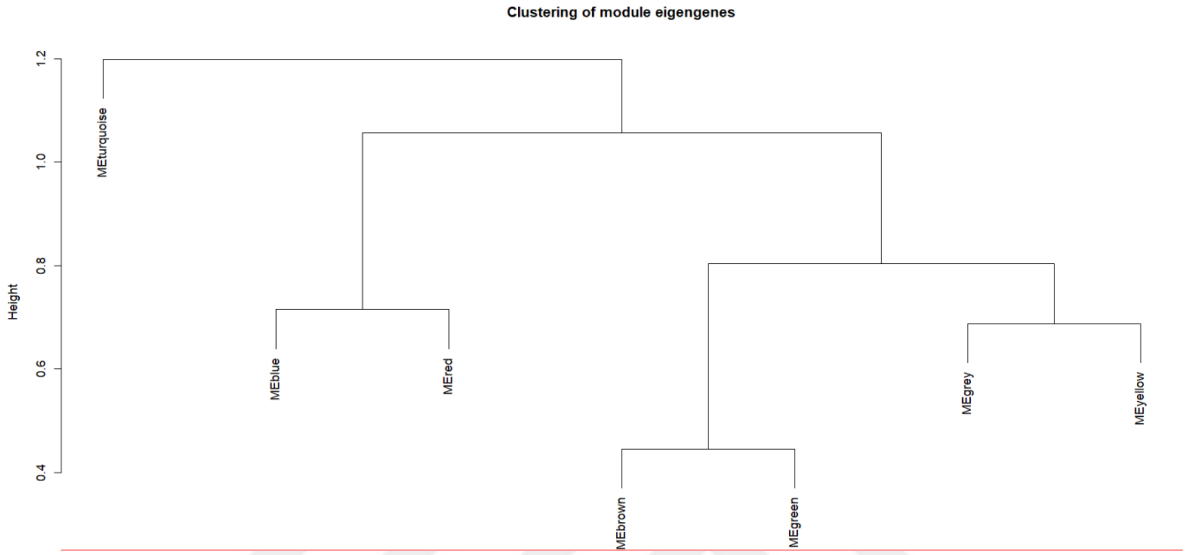


Figure 26. Dendogram tree of TCGA module eigengenes. Modules are shown in branches and the red line indicates the dissimilarity threshold which is 0.25.

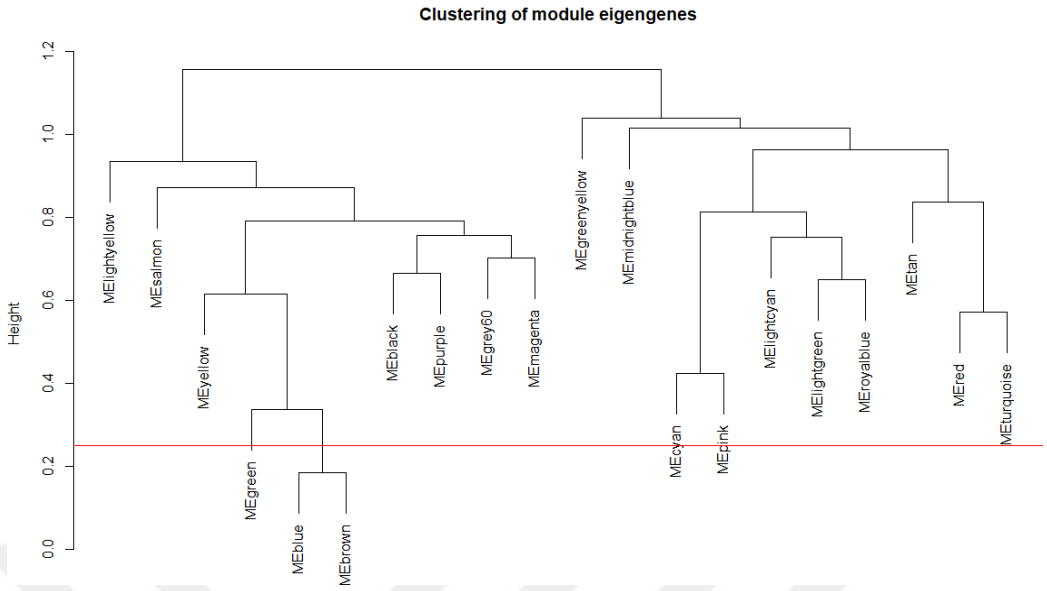


Figure 27. Dendogram tree of GEO module eigengenes. Modules are shown in branches and the red line indicates the dissimilarity threshold which is 0.25.

As mentioned previously, based on the dissimilarity threshold, modules were merged. Merged module plots including the old and merged modules with assigned colors are shown in Figure 28 and 29.

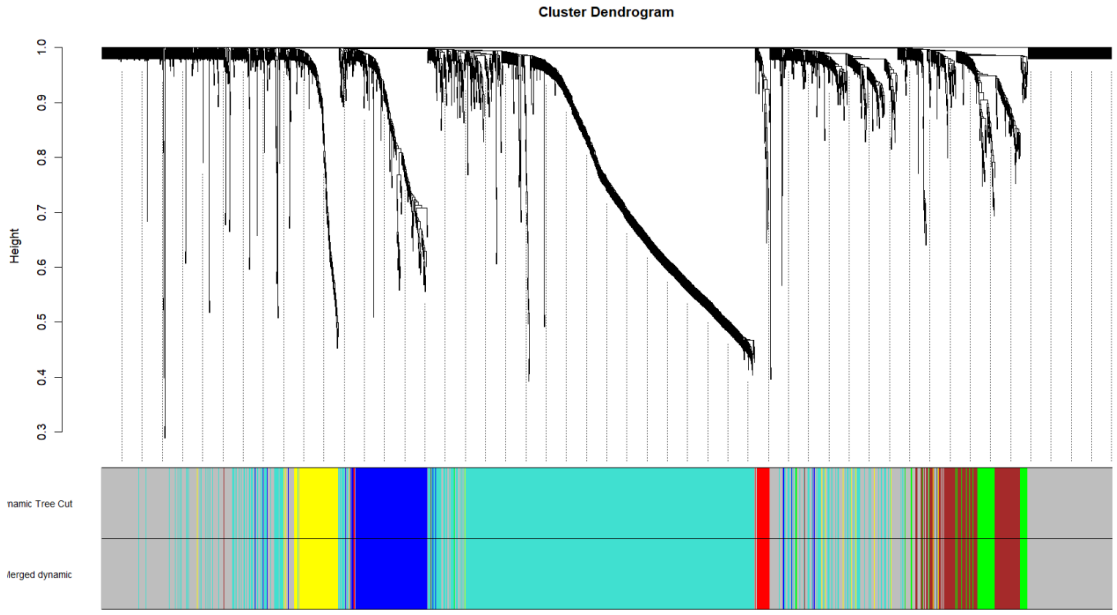


Figure 28. Gene Dendogram and old and new Module Colors of TCGA DEGs. The upper part represents the gene dendogram tree that was generated with dissimilarity based on topological overlap. The lower part indicates the old and new modules and their assigned module colors.

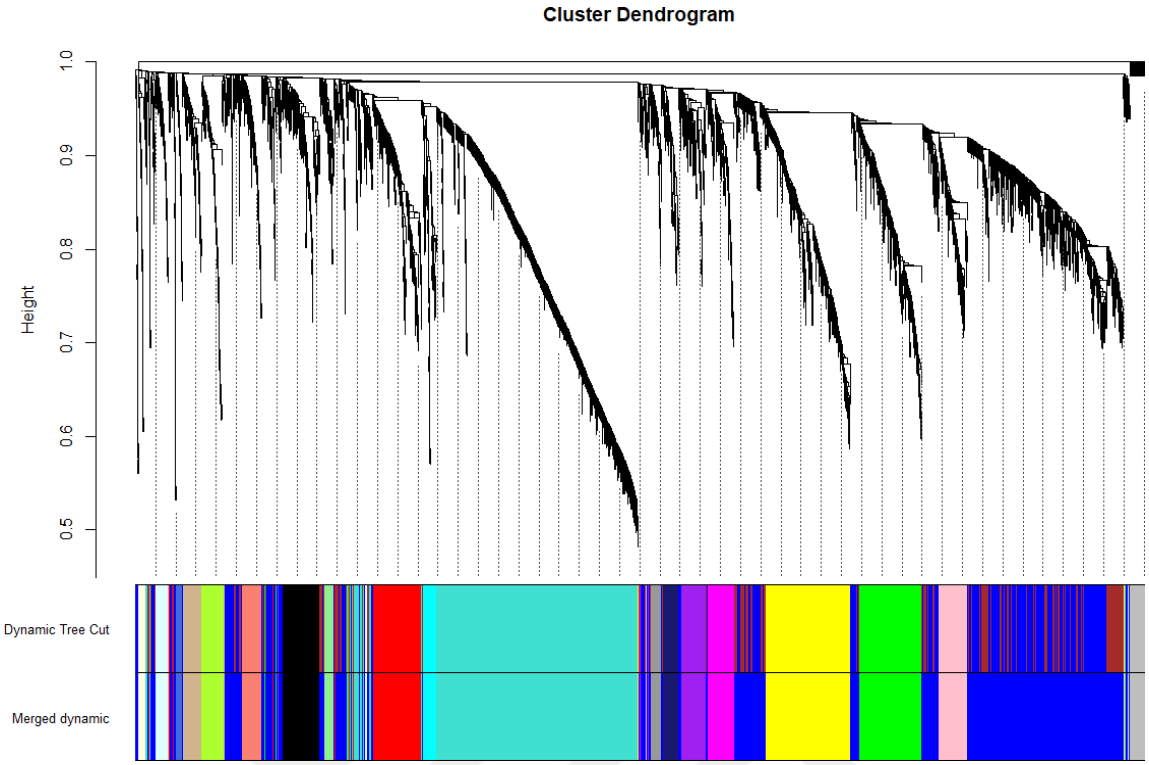


Figure 29. Gene Dendogram and old and new Module Colors of GEO DEGs. The upper part represents the gene dendogram tree that was generated with dissimilarity based on topological overlap. Lower part represents the old and new modules and their assigned module colors.

4.3.4. Relationship between Modules and Traits

In this step of the study, modules highly associated with clinical traits were selected. The relationship between modules and clinical traits were examined by the corresponding correlation coefficient greater than 0.3 and p-value < 0.05. Then following clinical traits were selected.

For TCGA, nine clinical traits were used:

- Primary Diagnosis: Date of diagnosis based on the number of days.
- Age at Diagnosis: Age the patient was diagnosed based on the number of years.
- Vital Status: Current vital status of the patient (alive or dead).
- Sample Type: Type of material obtained from regions of the brain.
- Site of Resection Biopsy: Part of the brain that the resection was performed.

- Prior Treatment: Information about early treatments after diagnosis.
- Gender: Sex of the patient (male or female).
- Race: Race of the patients.
- Tissue Organ Origin: Origin of the organ tissue where tumor originated.

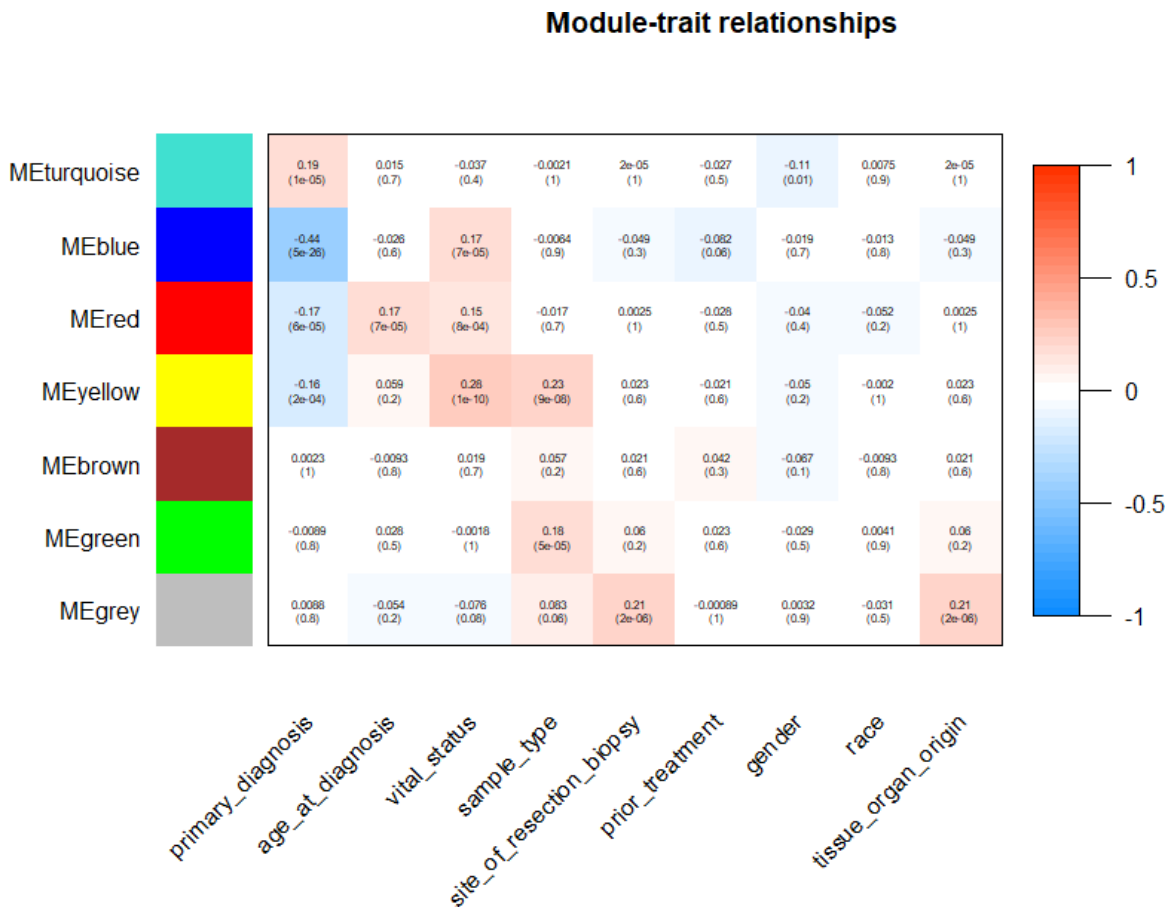


Figure 30. Heatmap of the relationship between modules and clinical traits of TCGA. Rows represent modules and columns represent clinical traits. Blue indicates a negative correlation and red indicates a positive correlation. Numbers in indices are the correlation coefficient values and corresponding p-values.

For GEO, seven clinical traits were considered:

- Age: Age of patients.
- Progressed Tumor Pathology: Current stage of tumor pathology.
- IDH Mutant Status: If patients carry a mutation in the *IDH* gene.

- Vital Status: If patients are dead or alive.
- Gender: Sex of patients.
- Initial Tumor Pathology: Initial state of the tumor pathology.
- Patient Initial Grade: Glioma grade that the patient was diagnosed.

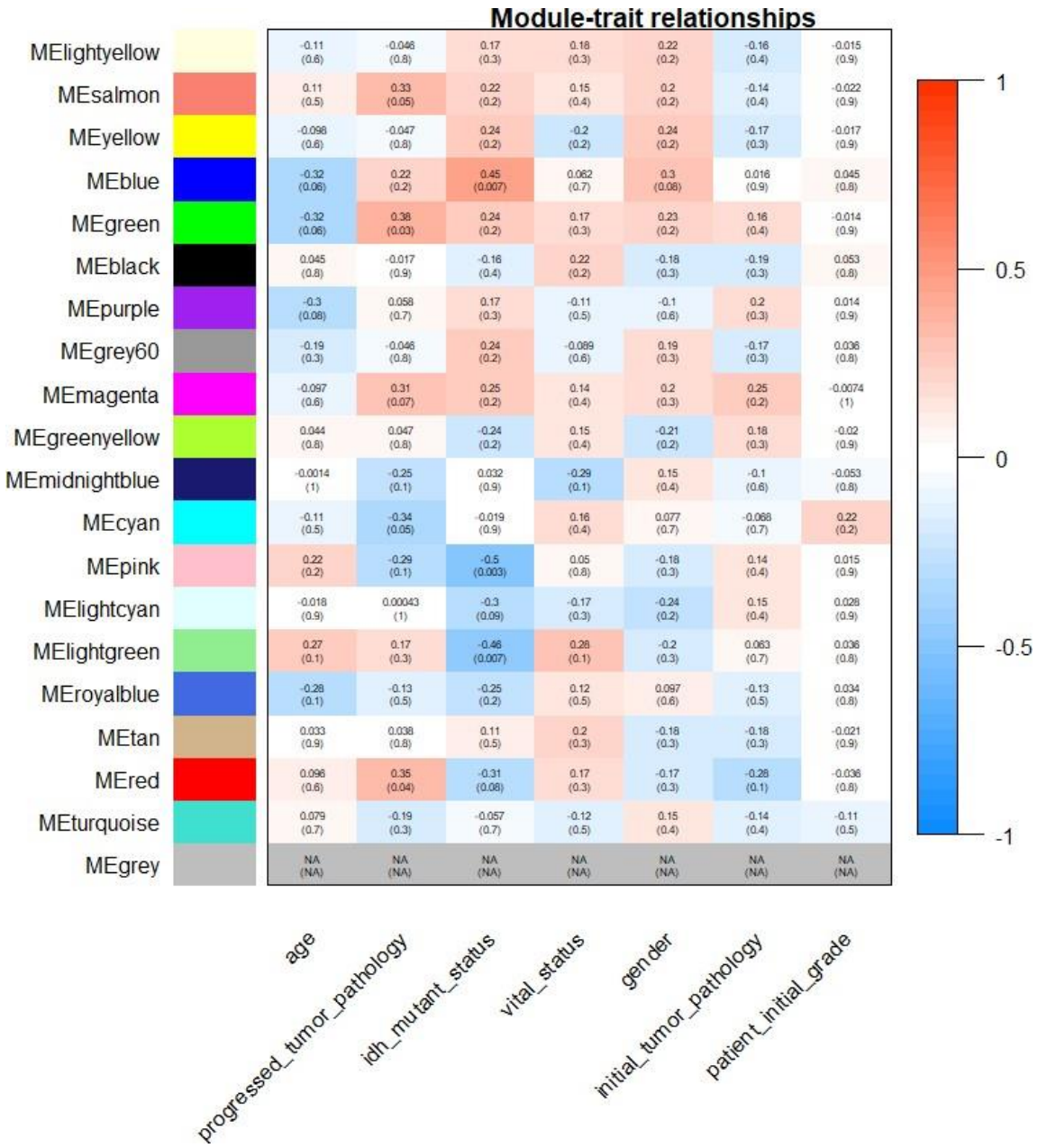


Figure 31. Heatmap of the relationship between modules and clinical traits of GEO. The modules are represented by rows modules and the clinical traits by columns. Blue indicates a negative correlation and red indicates a positive correlation. Numbers in indices are the correlation coefficient values and corresponding p-values.

For TCGA, two out of seven modules were selected. Blue module contains two hundred and twenty five (225) genes; yellow module contained one hundred and forty two (142) genes. For GEO, seven out of twenty modules were selected. Salmon module contained one hundred and three (103); blue module contained one thousand six hundred and twenty one (1621); green module contained three hundred and thirty nine (339); cyan module contained eighty one (81); pink module contained one hundred sixty four (164); lightgreen module contained fifty one (51); red module contained two hundred and seventy two (272). The commonly detected DEGs related with clinical traits were found to represent 17% of TCGA DEGs.

4.3.5. Gene Network Visualization

TOM plots with dendogram tree and also modules of TCGA and GEO gene co-expression networks were generated.

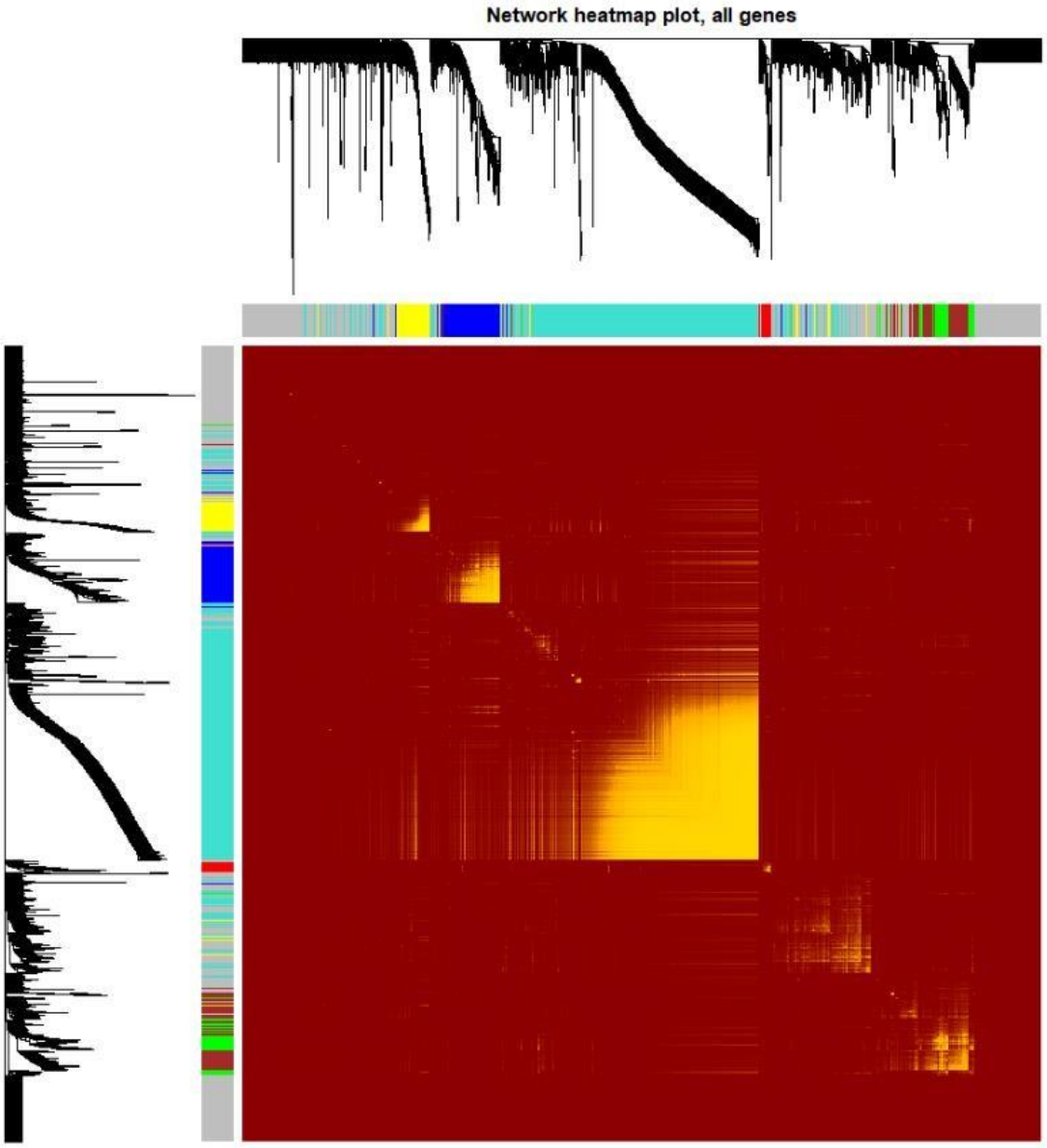


Figure 32. TOM with gene network heatmap of TCGA. Outer layer is the dendogram tree of each gene. Middle layer corresponds to the merged modules and their assigned colors. Inner layer is the heatmap. Lighter color indicates higher topological overlaps and darker color indicates lower topological overlaps.

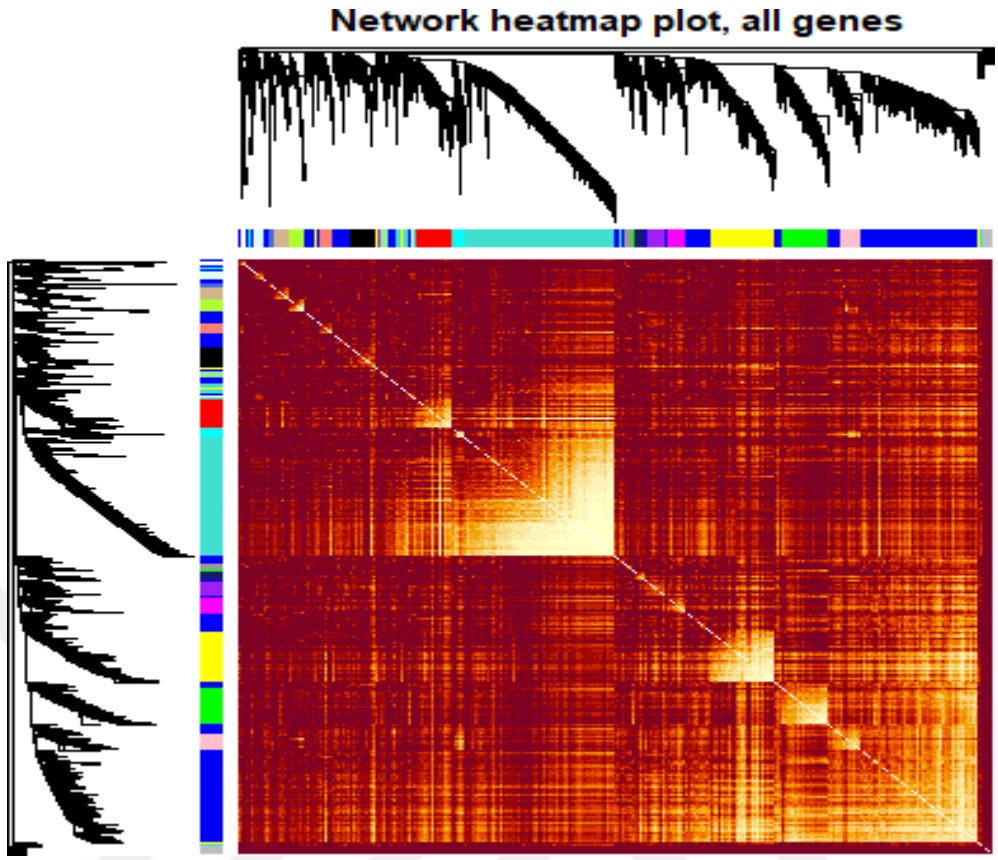


Figure 33. TOM with gene network heatmap of GEO. Outer layer is the dendogram tree of each gene. Middle layer is the merged modules and their assigned colors. Inner layer corresponds to the heatmap. Lighter color indicates higher topological overlaps and darker color indicates lower topological overlaps.

4.4. Co-Expression and PPI Network Analysis and Identification of Hub Genes

4.4.1. Gene Co-Expression Network Analysis

Hub gene identification was carried out by using node degree filtration and 12 algorithms for network topology analysis using the Cyohubba plugin of the Cytoscape software. The gene co-expression network had 76 nodes/genes and 522 edges/interactions. There was a total of 48 genes with a degree ≥ 10. Analysis by Cytohubba resulted to a total of 35 hub genes.

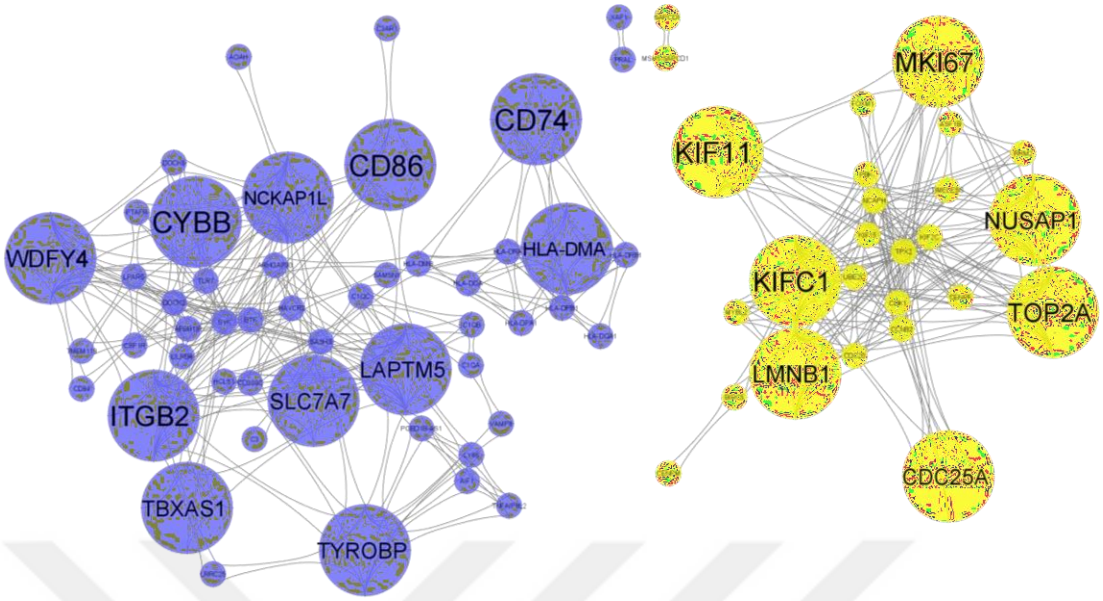


Figure 34. Co-expression network of genes from selected modules of TCGA. The hub genes are highlighted. The node color correspond to the blue and yellow modules. Lines indicate the edges or interactions between nodes/genes. Modules are grouped into two clusters.

4.4.2. Protein-Protein Interaction Network

The *STRING* protein search option of Cytoscape resulted to a total of identified 73 proteins/nodes and 558 interactions/edges. Network analysis resulted to 38 proteins with node degree ≥ 10. PPI network analysis by Cytohubba also resulted to 38 proteins.

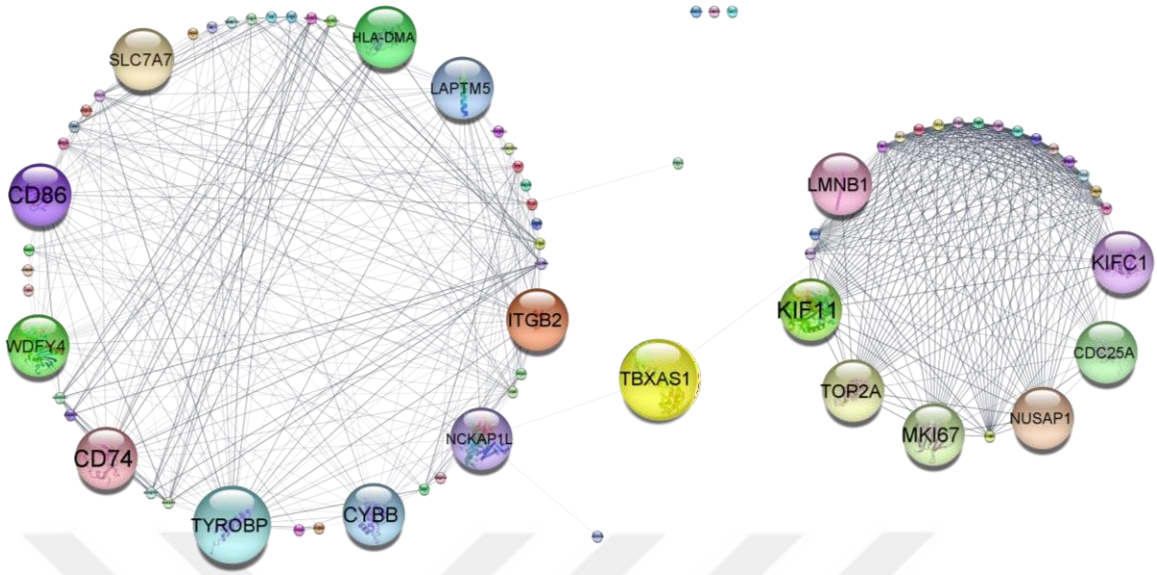


Figure 35. Protein-Protein Network Analysis. Bigger nodes indicate hub genes. The proteins are clustered into two distinct groups. The gene product *TBXAS1* connects the two clusters. Each node represents a protein. Lines indicate the interactions among gene products.

4.4.3. Determination of Hub Genes

By combining the results from the gene co-expression network analysis and Cytohubba, we detected a total of 18 hub genes: *CD74*, *CD86*, *CDC25A*, *CYBB*, *HLA-DMA*, *ITGB2*, *KIF11*, *KIFC1*, *LAPTM5*, *LMNB1*, *MKI67*, *NCKAP1L*, *NUSAP1*, *SLC7A7*, *TBXAS1*, *TOP2A*, *TYROBP*, and *WDFY4*. All detected hub genes were found to be up-regulated in LGG.

4.5. Prognostic Potential of LGG Hub Genes

The elevated expression of the *CD74*, *CD86*, *CDC25A*, *HLA-DRA*, *ITGB2*, *KIF11*, *KIFC1*, *LAPTM5*, *LMNB1*, *MKI67*, *NCKAP1L*, *NUSAP1*, *SLC7A7*, *TBXAS1*, *TOP2A*, *TYROBP* and *WDFY4* hub genes was found to be singificantly associated with unfavorable overall survival in LGG patients, as indicated by hazard ratio (HR) values above 1 and p-values < 0.05 (Figure 36). For *CYBB*, a HR value of 1.3 was not statistically significant (p = 0.14). Therefore, LGG patients with enhanced expression of these pivotal genes have an increased mortality risk, and may die at a higher rate per unit time as compared to patients with decreased expression of the corresponding genes.

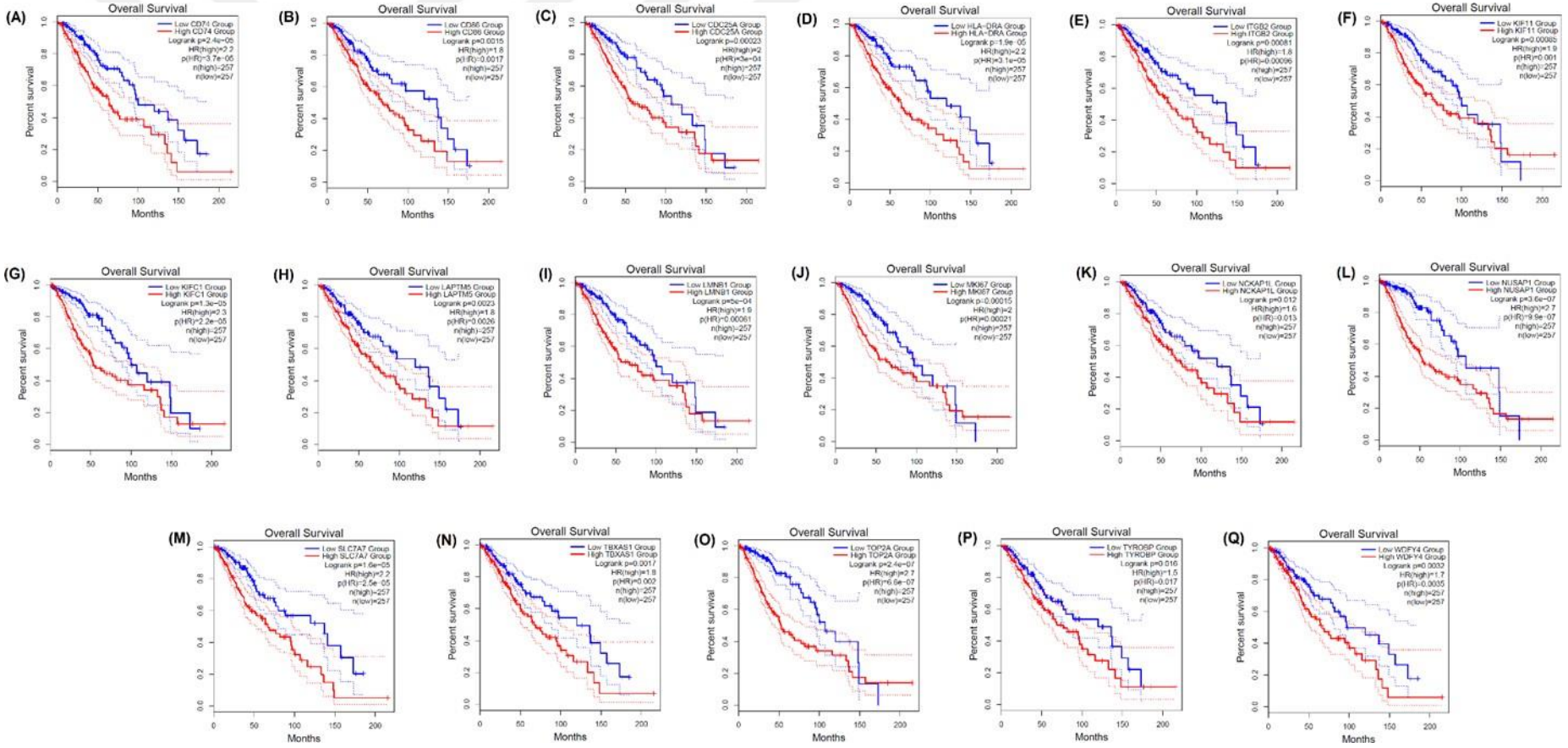


Figure 36. Kaplan–Meier curves depicting the prognostic significance of seventeen signature genes (A) *CD74*, (B) *CD86*, (C) *CDC25A*, (D) *HLA-DRA*, (E*) ITGB2*, (F) *KIF11*, (G*) KIFC1*, (H) *LAPTM5*, (I) *LMNB1*, (J) *MKI67*, (K) *NCKAP1L*, (L) *NUSAP1*, (M) *SLC7A7*, (N) *TBXAS1*, (O) *TOP2A*, (P) *TYROBP* and (Q) *WDFY4* for overall survival in LGG patients. The HR "HR (high)" and the corresponding p-values "p(HR)" are indicated. The 95% confidence intervals (CI) are denoted by dotted lines. The number of high-risk and low-risk LGG patient groups are represented by "n(high)" and "n(low)," respectively.

5. DISCUSSION

Gliomas are tumors that originate from glial cells which support the central nervous system. Based on the general classification of gliomas, there are four major grades of gliomas ranging from low-grade to high-grade or glioblastoma, the malignant form. Due to cancer heterogeneity, diagnosis is difficult, even for experienced medical doctors. Sometimes even scanned images of brain can be confusing because some types of gliomas resemble each other. Therefore, glioma treatment options become highly debatable. In the case of wrong diagnosis, treatments may result to decreased life standards due to after-treatment effects or even reduced life span. In this study, it was aimed to detect potential diagnostic and prognostic biomarkers between TCGA-derived and GTEx-derived gene expression data.

Due to the advancement of high-throughput technologies, researchers can access NGS generated data easily. In this study, LGG gene expression (RNA-Seq) data were acquired from TCGA, which is a collective database of sequenced genomes from samples across many cancer types. The corresponding control data were obtained from the corresponding normal brain tissue from GTEx. Differential gene expression analysis was performed between these two RNA-Seq datasets, followed by weighted gene co-expression network analysis. The detected DEGs between TCGA and GTEx samples were used to identify hub genes based on twelve different algorithms for network topology.

As a cross-platform validation option, LGG RNA-Seq data were extracted from the GEO database and transcriptomic profiling was performed by following the same procedure as in the case of TCGA data. DGEA between GEO-LGG and GTEx normal brain samples was performed, followed by WGCNA.

Collectively, 4496 genes were detected between TCGA and GTEx samples, of which 531 were up-regulated and 3965 were down-regulated. On the other hand, there were detected 8049 genes between GEO and GTEx samples, of which 4000 of were up-regulated and 4049 were down-regulated. In addition, there was a rather significant 60% overlap between the TCGA-GTEx and GEO-GTEx DEGs with respect to TCGA DEGs.

WGCNA of those TCGA-derived and GEO-derived LGG DEGs associated with clinicopathological traits, resulted in two significant modules comprised of 363 genes and seven

modules consisting of 2617 genes, respectively. There was also a 17% overlap between the clinically related DEGs of TCGA and GEO with respect to TCGA DEGs.

By using the genes within selected modules correlated with clinical traits for network topology analysis, a total of eighteen 'hub' genes were identified: *CD74, CD86, CDC25A, CYBB, HLA-DMA, ITGB2, KIF11, KIFC1, LAPTM5, LMNB1, MKI67, NCKAP1L, NUSAP1, SLC7A7, TBXAS1, TOP2A, TYROBP,* and *WDFY4*.

CD74 codes for a protein that regulates antigen presentation to immune cells. In general, *CD74* was demonstrated to play a role in the development of many types of cancers. In a study by Xu *et al*. (2021) revealed that *CD74* expression was higher in glioma cells compared to normal brain cells. Moreover, the expression of *CD74* was higher in HGG compared to LGG. It was suggested that *CD74* is a biomarker of LGG for poor prognosis and also could be a therapeutic target for glioma [51]. In another study, expression of *CD74* was shown tocontribute to resistance to treatments (themozolomide) in GBM [52]. In our study, *CD74* was found to be up-regulated in TCGA-LGG samples.

CD86 is another protein-coding gene and expressed by antigen presenting cells. A recent study by Ahmed *et al*. (2022) explored tumor microenvironment. Their study suggests that *CD86* may actually be a potential biomarker for prognosis of GBM and found that mRNA expression of *CD86* was high in GBM patients [53]. Also, Qiu and coworkers (2021) found that *CD86* overexpression is an unfavorable marker for LGG prognosis, as in our study [54].

CYBB gene encodes the beta chain of Cytochrome b. This protein has a crucial function in ion channels and beta chain. In a study where screening for biomarkers of LGG was performed, *CYBB* was shown to be one of the key DEGs [55].

Although *HLA-DMA* was previously associated with breast cancer, it is also known to have a role in cancer progression and resistance to drugs. Yin *et al*. (2022) found *HLA-DMA* to have prognostic value in GBM [56]. We also found *HLA-DMA* to be up-regulated in our TCGA-LGG samples and represent a potential prognostic marker of LGG.

ITGB2 codes for the beta chain of an integrin which forms integrin heterodimers. Integrins are basically involved in cell-surface related functions such as cell adhesion or signaling. A recent study by Xu and coworkers (2022) examined *ITGB2* gene in terms of prognosis in glioma patients. They found that the glioma grade increases relatively to the

expression level of *ITGB2*, and they suggest that *ITGB2* could be a novel predictor for glioma [57]. Another study that aimed to detect hub genes by using data derived from TCGA and CGGA in LGG found *ITGB2* to be a key gene. Including *ITGB2*, several key genes of this study are also similar to the one identified in our study [58].

KIF11, shown to be up-regulated in LGG, encodes a motor protein which plays a role in the spindle dynamics of cells. A study about the inhibitory effects of *KIF11* on mice found that *KIF11* is up-regulated in GBM, specifically in proliferating and migrating cells [59]. One important reason that *KIF11* represents a good target is that its mostly explored inhibitors are not neurotoxic [60]. Another study found that *KIF11* was up-regulated in gliomas and also was negatively associated with the survival of patients, based on TCGA-derived data. Moreover, it is suggested that *KIF11* is necessary for the stemness of glioma cells and thereby cell proliferation [61].

KIFC1, a kinesin motor thought to have a role in the clustering of the centrosomes of cancer cells, is involved in several types of cancers. Wu and coworkers (2021) found that there is elevated expression of *KIFC1* in GBM and its inhibition suppressed proliferation and resistance to temozolomide [62]. Another study showed that *KIFC1* is up-regulated in Grade III gliomas [63], like in our study.

LAPTM5 encodes protein E3 which is a transmembrane receptor. There are controversial results regarding the role of *LAPTM5* in gliomas. A study found *LAPTM5* to be a biomarker of poor prognosis in GBMs [64]. On the other hand, another study found *LAPTM5* to be a tumor suppressor and its inhibition actually promoted invasiveness of GBMs based on *in vitro* and *in vivo* experiments [65]. Nevertheless, in this study we found *LAPTM5* to be up-regulated in LGG and also to be a poor prognostic marker for the overall survival of LGG patients.

The product of *LMNB1* is one of the three component proteins of nuclear lamina. In a recent study based on microarray data, it was shown that *LMNB1* expression was higher in gliomas, and was also related with poor survival rates. Moreover, inhibition of *LMNB1* suppressed glioma proliferation [66]. Another study showed that *LMNB1* is up-regulated in glioma cells. They have also shown that overexpression of *lamin* genes causes abnormalities in human astrocytes which are actually glial cells that support the central nervous system [67].

NCKAP1L, found to be up-regulated in our study, was also found to be a hub gene in LGG in another study, using TCGA and CGGA data. In the same study, the genes *LAMC1*, *CD74*, *HLA-DMA*, *ITGB2*, *TYROBP*, *LAPTM5*, and *CYBB* [68], which were also identified in our study, were shown to constitute hub genes.

SLC7A7 belongs to the family of *solute carrier 7* genes, which have a role in nutrition intake since they are transporters. In a study, where the genes of this family were examined in LGG by using TCGA-derived data, it was shown that *SLCA7A* was significantly up-regulated in patients with LGG and was also a poor prognostic marker [69], similarly to the findings of our study. In another study, they examined a Chinese population of 736 glioma patients and 793 normal subjects and found that polymorphisms in *SLC7A7* actually contribute to the heterogeneity of gliomas [70].

TBXAS1 codes for a member of the cytochrome P450 superfamily. These proteins have a role in drug metabolism and also the synthesis of fat molecules. A study found high expression of *TBXAS1* in diffuse LGG patients, which was also associated with poor overall survival [71], like in our study.

The product of *TOP2A* is a DNA topoisomerase with a critical role in transcription. A novel variant of *TOP2A* was identified in GBM patients, which was associated with altered transcriptional regulation and decreased survival [72]. Another study also supports the relationship of *TOP2A* upregulation of with poor prognosis in gliomas, and also suggests that *MKI67*, another hub gene found to be upregulated in our study, is correlated with *TOP2A* overexpression [73]. *TYROBP* encodes a transmembrane signaling protein. In a study by Lu and colleagues (2021), where glioma datasets derived from Oncomine, GEPIA2 and CGGA were examined, it was found that *TYROBP* was one of the hub genes, upregulated in LGG. Their results revealed also that TYROBP was implicated in pivotal signaling pathways such as JAK-STAT, suggesting that TYROBP could actually contribute largely to LGG, through its involvement in signaling networks [74].

Symptoms of brain tumors, with headache being the most common one, are generally considered to be non-significant or possibly benign tumors, thereby leading to poor diagnosis and prognosis. The hub genes detected in this study could represent a rapid way to possibly test their significance in brain neoplasms through blood tests. In a study by Butler and colleagues (2019) tested a methodology, called ATR-FTIR, on blood tests, and successfully differentiated

brain cancer and normal cells with a 93.2% sensitivity and 92.8% specificity. Another study by Podnar and colleagues (2019) used machine learning on blood tests of patients with brain tumors in order to investigate any differences with 96% sensitivity and 74% specificity [75, 76]. In a similar manner, the eighteen signature genes identified in this study could be tested for their diagnostic potential for patients with gliomas.

6. CONCLUSION AND FUTURE ASPECT

In this study differential gene expression analysis, weighted gene co-expression network analysis and network-based analyses were performed in order to detect signature genes of particular importance in LGG. We detected eighteen hub genes, all up-regulated in LGG, and also significantly associated with unfavorable prognosis in LGG patients. Their significance in LGG was also examined on the basis of their relationship with LGG clinical traits. Therefore, these genes could be taken into consideration in the clinical setting as candidate diagnostic or prognostic biomarkers, for the accurate, timely and cost-effective diagnosis of LGG and for monitoring LGG patients' progression. The findings of this study could lay the foundation for further in vivo and in vitro experimental studies towards the elucidation of the underlying mechanisms of low-grade glioma genesis.

7. REFERENCES

1. Nakagawa H, Fujita M. Whole genome sequencing analysis for cancer genomics and precision medicine. Cancer Science. 2018;109(3):513-522.
2. Forst D, Nahed B, Loeffler J, Batchelor T. Low-Grade Gliomas. The Oncologist. 2014;19(4):403-413.
3. Wirsching H, Weller M. The Role of Molecular Diagnostics in the Management of Patients with Gliomas. Current Treatment Options in Oncology. 2016;17(10).
4. Delgado-López P, Corrales-García E, Martino J, Lastra-Aras E, Dueñas-Polo M. Diffuse low-grade glioma: a review on the new molecular classification, natural history and current management strategies. Clinical and Translational Oncology. 2017;19(8):931-944.
5. Ferris S, Hofmann J, Solomon D, Perry A. Characterization of gliomas: from morphology to molecules. Virchows Archiv. 2017;471(2):257-269.
6. Brown T, Bota D, van Den Bent M, Brown P, Maher E, Aregawi D et al. Management of low-grade glioma: a systematic review and meta-analysis. Neuro-Oncology Practice. 2018;6(4):249-258.
7. Dono A, Ballester L, Primdahl D, Esquenazi Y, Bhatia A. IDH-Mutant Low-grade Glioma: Advances in Molecular Diagnosis, Management, and Future Directions. Current Oncology Reports. 2021;23(2).
8. Schiff D. Low-grade Gliomas. CONTINUUM: Lifelong Learning in Neurology. 2017;23(6):1564-1579.
9. Kumthekar P, Raizer J, Singh S. Low-Grade Glioma. Cancer Treatment and Research. 2014;:75-87.
10. Chen R, Cohen A, Colman H. Targeted Therapeutics in Patients With High-Grade Gliomas: Past, Present, and Future. Current Treatment Options in Oncology. 2016;17(8).
11. Claus E, Walsh K, Wiencke J, Molinaro A, Wiemels J, Schildkraut J et al. Survival and low-grade glioma: the emergence of genetic information. Neurosurgical Focus. 2015;38(1):E6.

12. Wang T, Mehta M. Low-Grade Glioma Radiotherapy Treatment and Trials. Neurosurgery Clinics of North America. 2019;30(1):111-118.

13. Rossi M, Sani S, Nibali M, Fornia L, Bello L, Byrne R. Mapping in Low-Grade Glioma Surgery. Neurosurgery Clinics of North America. 2019;30(1):55-63.

14. Carabenciov I, Buckner J. Controversies in the Therapy of Low-Grade Gliomas. Current Treatment Options in Oncology. 2019;20(4).

15. Tom M, Cahill D, Buckner J, Dietrich J, Parsons M, Yu J. Management for Different Glioma Subtypes: Are All Low-Grade Gliomas Created Equal?. American Society of Clinical Oncology Educational Book. 2019;(39):133-145.

16. Atkins I, Kinnersley B, Ostrom Q, Labreche K, Il'yasova D, Armstrong G et al. Transcriptome-Wide Association Study Identifies New Candidate Susceptibility Genes for Glioma. Cancer Research. 2019;79(8):2065-2071.

17. Xu S, Wang Z, Ye J, Mei S, Zhang J. Identification of Iron Metabolism-Related Genes as Prognostic Indicators for Lower-Grade Glioma. Frontiers in Oncology. 2021;11.

18. Perry A, Wesseling P. Histologic classification of gliomas. Handbook of Clinical Neurology. 2016;:71-95.

19. Liu Y, Xu Z, Jin T, Xu K, Liu M, Xu H. Ferroptosis in Low-Grade Glioma: A New Marker for Diagnosis and Prognosis. Medical Science Monitor. 2020;26.

20. de Botton S, Mondesir J, Willekens C, Touat M. IDH1 and IDH2 mutations as novel therapeutic targets: current perspectives. Journal of Blood Medicine. 2016;Volume 7:171-180.

21. Zhang Y, Yang X, Zhu X, Bai H, Wang Z, Zhang J et al. S100A gene family: immune-related prognostic biomarkers and therapeutic targets for low-grade glioma. Aging. 2021;13(11):15459-15478.

22. McCombie W, McPherson J, Mardis E. Next-Generation Sequencing Technologies. Cold Spring Harbor Perspectives in Medicine. 2018;9(11):a036798.

23. Hong M, Tao S, Zhang L, Diao L, Huang X, Huang S et al. RNA sequencing: new technologies and applications in cancer research. Journal of Hematology; Oncology. 2020;13(1).

24. Kukurba K, Montgomery S. RNA Sequencing and Analysis. Cold Spring Harbor Protocols. 2015;2015(11):pdb.top084970.

25. Hrdlickova R, Toloue M, Tian B. RNA-Seq methods for transcriptome analysis. WIREs RNA. 2016;8(1).

26. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A et al. A survey of best practices for RNA-seq data analysis. Genome Biology. 2016;17(1).

27. Huang H, Niu Y, Qin L. Differential Expression Analysis for RNA-Seq: An Overview of Statistical Methods and Computational Software. Cancer Informatics. 2015;14s1:CIN.S21631.

28. Kishore J, Goel M, Khanna P. Understanding survival analysis: Kaplan-Meier estimate. International Journal of Ayurveda Research. 2010;1(4):274.

29. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, Sabedot T, Malta TM, Pagnotta SM, Castiglioni I, Ceccarelli M, Bontempi G, Noushmehr H (2015). " TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data." Nucleic Acids Research. doi: 10.1093/nar/gkv1507, http://doi.org/10.1093/nar/gkv1507.

30. Collado-Torres L, Nellore A, Kammers K, Ellis S, Taub M, Hansen K et al. Reproducible RNA-seq analysis using recount2. Nature Biotechnology. 2017;35(4):319-321.

31. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S et al. The Genotype-Tissue Expression (GTEx) project. Nature Genetics. 2013;45(6):580-585.

32. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository Nucleic Acids Res. 2002 Jan 1;30(1):207-10

33. Leinonen R, Sugawara H, Shumway M. The Sequence Read Archive. Nucleic Acids Research. 2010;39(Database):D19-D21.

34. Frankish A, Diekhans M, Jungreis I, Lagarde J, Loveland J, Mudge J et al. GENCODE 2021. Nucleic Acids Research. 2020;49(D1):D916-D923.

35. Dobin A, Davis C, Schlesinger F, Drenkow J, Zaleski C, Jha S et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2012;29(1):15-21.

36. Li B, Dewey C. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12(1).

37. Risso D, Schwartz K, Sherlock G, Dudoit S (2011). "GC-Content Normalization for RNA-Seq Data." BMC Bioinformatics, 12(1), 480.

38. Zhang Y, Parmigiani G, Johnson W. ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genomics and Bioinformatics. 2020;2(3).

39. Jolliffe I, Cadima J. Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2016;374(2065):20150202.

40. Ren X, Kuan P. RNAAgeCalc: A multi-tissue transcriptional age calculator. PLOS ONE. 2020;15(8):e0237006.

41. Robinson M, McCarthy D, Smyth G. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2009;26(1):139-140.

42. Ritchie M, Phipson B, Wu D, Hu Y, Law C, Shi W et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research. 2015.

43. Love, M.I., Huber, W., Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 Genome Biology 15(12):550 (2014)

44. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9(1).

45. Shannon P, Markiel A, Ozier O, Baliga N, Wang J, Ramage D et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Research. 2003;13(11):2498-2504.

46. Szklarczyk D, Gable A, Lyon D, Junge A, Wyder S, Huerta-Cepas J et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Research. 2018;47(D1):D607-D613.

47. Chin C, Chen S, Wu H, Ho C, Ko M, Lin C. cytoHubba: identifying hub objects and sub-networks from complex interactome. BMC Systems Biology. 2014;8(S4).

48. 4. Li C, Tang Z, Zhang W, Ye Z, Liu F. GEPIA2021: integrating multiple deconvolution-based analysis into GEPIA. Nucleic Acids Research. 2021;49(W1):W242-W246.

49. 5. Tang Z, Li C, Kang B, Gao G, Li C, Zhang Z. GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. Nucleic Acids Research. 2017;45(W1):W98-W102.

50. 6. Liu J, Lichtenberg T, Hoadley K, Poisson L, Lazar A, Cherniack A et al. An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. Cell. 2018;173(2):400-416.e11.

51. Xu S, Li X, Tang L, Liu Z, Yang K, Cheng Q. CD74 Correlated With Malignancies and Immune Microenvironment in Gliomas. Frontiers in Molecular Biosciences. 2021;8.

52. Presti M, Mazzon E, Basile M, Petralia M, Bramanti A, Colletti G et al. Overexpression of macrophage migration inhibitory factor and functionally-related genes, D-DT, CD74, CD44, CXCR2 and CXCR4, in glioblastoma. Oncology Letters. 2018;.

53. Ahmed M, Hernández-Verdin I, Bielle F, Verreault M, Lerond J, Alentorn A et al. Expression and Prognostic Value of CD80 and CD86 in the Tumor Microenvironment of Newly Diagnosed Glioblastoma. Canadian Journal of Neurological Sciences / Journal Canadien des Sciences Neurologiques. 2022;:1-9.

54. Qiu H, Tian W, He Y, Li J, He C, Li Y et al. Integrated Analysis Reveals Prognostic Value and Immune Correlates of CD86 Expression in Lower Grade Glioma. Frontiers in Oncology. 2021;11.

55. Guo F, Yan J, Ling G, Chen H, Huang Q, Mu J et al. Screening and Identification of Key Biomarkers in Lower Grade Glioma via Bioinformatical Analysis. Applied Bionics and Biomechanics. 2022;2022:1-12.

56. Yin X, Wu Q, Hao Z, Chen L. Identification of novel prognostic targets in glioblastoma using bioinformatics analysis. BioMedical Engineering OnLine. 2022;21(1).

57. Xu H, Zhang A, Han X, Li Y, Zhang Z, Song L et al. ITGB2 as a prognostic indicator and a predictive marker for immunotherapy in gliomas. Cancer Immunology, Immunotherapy. 2021;71(3):645-660.

58. Li G, Wu Z, Gu J, Zhu Y, Zhang T, Wang F et al. Metabolic Signature-Based Subtypes May Pave Novel Ways for Low-Grade Glioma Prognosis and Therapy. Frontiers in Cell and Developmental Biology. 2021;9.

59. Venere M, Horbinski C, Crish J, Jin X, Vasanji A, Major J et al. The mitotic kinesin KIF11 is a driver of invasion, proliferation, and self-renewal in glioblastoma. Science Translational Medicine. 2015;7(304).

60. Gampa G, Kenchappa R, Mohammad A, Parrish K, Kim M, Crish J et al. Enhancing Brain Retention of a KIF11 Inhibitor Significantly Improves its Efficacy in a Mouse Model of Glioblastoma. Scientific Reports. 2020;10(1).

61. Liu B, Zhang G, Cui S, Du G. Upregulation of KIF11 in TP53 Mutant Glioma Promotes Tumor Stemness and Drug Resistance. Cellular and Molecular Neurobiology. 2021;42(5):1477-1485.

62. Wu J, Wang X, Yuan X, Shan Q, Wang Z, Wu Y et al. Kinesin Family Member C1 Increases Temozolomide Resistance of Glioblastoma Through Promoting DNA Damage Repair. Cell Transplantation. 2021;30:096368972199146.

63. Dai J, Bing Z, Zhang Y, Li Q, Niu L, Liang W et al. Integrated mRNAseq and microRNAseq data analysis for grade III gliomas. Molecular Medicine Reports. 2017;16(5):7468-7478.

64. M. Hajj G, Silva F, Bellis B, Lupinacci F, Bellato H, Cruz J et al. Aberrant expression of RSK1 characterizes high-grade gliomas with immune infiltration. Molecular Oncology. 2019;14(1):159-179.

65. Berberich A, Bartels F, Tang Z, Knoll M, Pusch S, Hucke N et al. LAPTM5–CD40 Crosstalk in Glioblastoma Invasion and Temozolomide Resistance. Frontiers in Oncology. 2020;10.

66. Zhou D, Wang M, Zhang Y, Wang K, Zhao M, Wang Y et al. Screening and identification of LMNB1 and DLGAP5, two key biomarkers in gliomas. Bioscience Reports. 2021;41(5).

67. Pei S, Wang X, Wang X, Huang H, Tao H, Xie B et al. Aberrant nuclear lamina contributes to the malignancy of human gliomas. Journal of Genetics and Genomics. 2022;49(2):132-144.

68. Su J, Long W, Ma Q, Xiao K, Li Y, Xiao Q et al. Identification of a Tumor Microenvironment-Related Eight-Gene Signature for Predicting Prognosis in Lower-Grade Gliomas. Frontiers in Genetics. 2019;10.

69. Liu W, Ji Y, Ren R, Zhang G. The Solute Carrier Family 7 Genes Are Potential Diagnostic and Prognostic Biomarkers in Lower Grade Glioma. 2021;.

70. Fan S, Zhao Y, Li X, Du Y, Wang J, Song X et al. Genetic variants in SLC7A7 are associated with risk of glioma in a Chinese population. Experimental Biology and Medicine. 2013;238(9):1075-1081.

71. Wang S, Jin F, Fan W, Liu F, Zou Y, Hu X et al. Gene expression meta-analysis in diffuse low-grade glioma and the corresponding histological subtypes. Scientific Reports. 2017;7(1).

72. Gielniewski B, Poleszak K, Roura A, Szadkowska P, Krol S, Guzik R et al. The novel, recurrent mutation in the TOP2A gene results in the enhanced topoisomerase activity and transcription deregulation in glioblastoma. 2020;.

73. Zhou T, Wang Y, Qian D, Liang Q, Wang B. Over-expression of TOP2A as a prognostic biomarker in patients with glioma. Int J Clin Exp Pathol. 2018 Mar 1;11(3):1228-1237. PMID: 31938217; PMCID: PMC6958105.

74. Lu J, Peng Y, Huang R, Feng Z, Fan Y, Wang H et al. Elevated TYROBP expression predicts poor prognosis and high tumor immune infiltration in patients with low-grade glioma. BMC Cancer. 2021;21(1).

75. Butler H, Brennan P, Cameron J, Finlayson D, Hegarty M, Jenkinson M et al. Development of high-throughput ATR-FTIR technology for rapid triage of brain cancer. Nature Communications. 2019;10(1).

76. Podnar S, Kukar M, Gunčar G, Notar M, Gošnjak N, Notar M. Diagnosing brain tumours by routine blood tests using machine learning. Scientific Reports. 2019;9(1).

8. APPENDIX

8.1. Curriculum Vitae

MELIH ÖZBEK

Year of Birth	
Address:	İzmir Uluslararası Biyotıp ve Genom Enstitüsü (İBG-İzmir) Balçova 35340 İzmir
Telephone:	+
Fax:	
e-mail:	

EDUCATION

Country	University	Faculty/Institute	Field	Degree	Year of Graduation
TR	Dokuz Eylul University	İzmir International Biomedicine and Genome Institute	Molecular Biology and Genetics	M.Sc.	2022
TR	Izmir Institute of Technology	Faculty of Science	Molecular Biology and Genetics	B.Sc.	2019

ACADEMIC EXPERTISE

Institution	Country	Laboratory	Department	Title
Karolinska Institute	Sweden	Assist. Prof. Karolinska Tammimies	Department of Women's and Children's Health	Intern
Izmir Institute of Technology	Turkey	Prof. Dr. Sami DOĞANLAR	Molecular Biology and Genetics	Intern

RESEARCH FIELDS

Cancer Bioinformatics

Computational Biology

Caner, Neurodegenerative Diseases

8.2. Ethics Committee Approval

İZMİR BİYOTIP VE GENOM MERKEZİ
GİRİŞİMSEL OLMAYAN ARAŞTIRMALAR ETİK KURULU (İBG-GOEK)
KARARI

Toplantı Tarihi	: 25/11/2021	**Toplantı Günü**	: Perşembe
Toplantı Sayısı	: 13	**Toplantı Saati**	: 10:30

Sayın Doktora Öğretim Üyesi Athanasia PAVLOPOULOU,

2021-038 Protokol No'lu; sorumlusu olduğunuz "Hesaplamalı Yöntemler Kullanılarak Beyin Neoplazmasında Prognoz ve Teşhis İşaretlerin Belirlenmesi" başlıklı araştırmanın uygulanmasında etik açıdan sakınca olmadığına oy birliği ile karar verilmiştir.

Bilgilerinizi ve gereğini rica ederiz.

Prof. Dr. H. Alper BAĞRIYANIK **Başkan**	**Dr. Serap ERKEK ÖZHAN** **Başkan Yardımcısı**
Prof. Dr. Sedef AKGÜNGÖR **Üye** (Katılmadı)	**Prof. Dr. İnci ALAGAOGLU** **Üye**
Prof. Dr. Gülgün OKTAY **Üye** (Katılmadı)	**Prof. Dr. Hilmi ORHAN** **Üye**
Prof. Dr. Şermin GENÇ **Üye**	**Doç. Dr. Canan Aslı YILDIRIM** **Üye** (Katılmadı)
Doç. Dr. Can KÜÇÜK **Üye** (Katılmadı)	**Dr. Öğr. Üyesi Yavuz OKTAY** **Üye**