UTRECHT UNIVERSITY

Graduate School of Natural Sciences

Universiteit Utrecht

princess máxima center pediatric oncology

**Applied Data Science master thesis**

# Differential Expression Analysis of Immune-Related Genes in Pediatric High-Grade Brain Tumors using RNA-seq Data

| | |
|---|---|
| **First examiner:** | **Candidate:** |
| Dr. A.Kaznatcheev | Amdom Weldeabezgi |
| **Second examiner:** | **In cooperation with:** |
| D.S. Islakoglu | Dr. T.F. Carvalheiro |
| | C.J Bogaard |

July 10, 2023

## Abstract

A better understanding of pediatric brain tumor immune microenvironment is crucial for developing effective immune-based treatments. The primary aim of this research was to compare the immune-related gene expressions of various high-grade pediatric brain tumors compared to Craniopharyngioma, a low-grade tumor. This is due to the absence of control group from healthy tissue at the time of this research. We also aimed to detect clusters of genes that behave similarly and their association with various tumor types.

In addressing our primary objective, we employed differential expression analysis using DESeq2 in R. We observed marked disparities in immune-related gene expression profiles among tumor types. Medulloblastoma demonstrated a striking 64% downregulation in gene expression. Contrastingly, Ependymoma and Glioma displayed 7.7% and 10% upregulation of gene expression, respectively. Specifically *MAGEA3* was top highly expressed gene across all tumor types.

Using Weighted Gene Co-expression Analysis (WGCNA), we identified eight distinct immune-related gene modules, three of which showed a strong correlation with Medulloblastoma. The gene *PIAS1* in the module with the highest positive correlation showed a notably significant association with Medulloblastoma.

**Keywords**: Pediatric brain tumors, Immunotherapy, Gene expression, Differential expression analysis, WGCNA, RNA-Seq, DESeq2.

# Acknowledgements

First and foremost, I would like to express my sincere gratitude to Utrecht University, Graduate School of Natural Sciences for providing the environment and the opportunity to carry out this research. I am deeply appreciative of the access to resources and facilities offered by Princess Máxima Center for Pediatric Oncology, which were essential to the successful completion of this project. In particular, I want to extend my heartfelt thanks to Dr. Artem Kaznatcheev, my first supervisor, whose constructive feedback and critical insights have been instrumental in ensuring the high quality of this work. I also wish to acknowledge Dr. Tiago Ferreira Carvalheiro, my supervisor at Maxima. His continuous and unwavering support, insightful suggestions, and valuable expertise was instrumental in shaping this research.

# Contents

# 1. Introduction

## 1.1 Pediatric High-Grade Brain Tumors

Brain tumors are classified by the World Health Organization (WHO) into four grades, with Grade I being the least aggressive and Grade IV being the most aggressive [1]. Pediatric high-grade brain tumors represent a particularly devastating category of cancers affecting children. Unlike lower grade tumors, these high-grade malignancies are characterized by rapid growth and a propensity for invasion, leading to significant morbidity and mortality among pediatric patients[2]. According to the Central Brain Tumor Registry of the United States (CBTRUS), brain and other CNS tumors are the most common solid cancer site in individuals aged 0-14 years, with a substantial impact on pediatric cancer mortality [2]. Children with high-grade gliomas (HGG), including diffuse midline gliomas (DMG) and glioblastoma multiforme (GBM), face a bleak 5-year overall survival rate of less than 20%[3]. The delicate and complex environment of the central nervous system (CNS) renders many conventional cancer treatments, like surgery, particularly risky [4]. In addition the diagnosis and treatment of a high-grade brain tumor can result in psychological, cognitive, and social challenges for the child and their family, there by increasing the overall burden of the disease [5].

**Table 1.1:** Pediatric high grade brain Tumor Types

| Brain Tumor Types | Characteristics |
|---|---|
| ATRT | Atypical teratoid/Rhaboid tumors are rare malignant intracranial neoplasms most commonly occurring in infants and young children. They account for only 1% to 2% of all pediatric brain tumors[6]. |
| Gliomas (including Glioblastoma) | High-grade gliomas (HGGs) occur at an incidence of 0.8 per 100,000 children per year. Approximately 20% of all childhood gliomas are HGGs[7]. |
| Craniopharyngiomas | Craniopharyngiomas are low-grade, slow-growing benign epithelial tumors and account for approximately 5% to 10% of pediatric brain tumors[8]. |
| Ependymomas | Ependymomas are the third most common brain tumor in children and account for approximately 8% to 10% of all childhood CNS tumors[9]. |
| Medulloblastoma | Medulloblastoma is an embryonal tumor of the posterior fossa and is the most common malignant brain tumor in children. It comprises up to 20% of all pediatric brain tumors[2]. |

## 1.2 Brain Tumor Immunology: the Need for Immune Microenvironment Study

The brain was traditionally thought to be an immune-privileged site, but it is now recognized that immune cells can and do infiltrate brain tumors [10]. From multiple studies over the past decade, it has become clear that the brain tumor microenvironment (TME) is a fundamental regulator of cancer progression and therapeutic efficacy in primary and metastatic brain malignancies. Insights into the biological processes within the brain TME identified potential therapeutic targets with several now under clinical evaluation such as the adoptive cell therapy with chimeric antigen receptor (CAR) T cells for patients with Glioblastoma patients that targets tumor-associated antigens[11].

Current therapeutic strategies for pediatric high-grade brain tumors are limited and often insufficient. Standard treatments typically include surgery, radiation therapy, and chemotherapy. However, these modalities can result in severe long-term side effects, including cognitive impairment and endocrine dysfunction, especially considering the developing nature of the pediatric brain[12]. Furthermore, certain high-grade brain tumors like Diffuse Intrinsic Pontine Glioma (DIPG) are notoriously difficult to treat due to their location in the brainstem, rendering surgical resection practically impossible [13]. Chemotherapy has failed to show benefit over radiotherapy, and the standard care with radiotherapy offers only temporary relief[14]. Recent advancements in cancer treatment, such as immunotherapy, present promising alternatives in light of current therapy limitations. Immunotherapy has demonstrated success in treating leukemia and complements traditional treatment methods[15]. However, applying immunotherapy to pediatric brain tumors requires generating tailored immune responses, understanding unique tumor sites, and investigating potential targets [16]. Despite the significant advancements in understanding tumor immunology in adults, the knowledge of the immune landscape in pediatric tumors, especially high-grade brain tumors, is still in its infancy[17]. Most of the existing research has been focused on adult tumors, with findings often extrapolated to pediatric cases. However, the pediatric immune system and tumor biology are distinct from those of adults, indicating that such extrapolations may not always be appropriate or accurate [18]. The current limitations in our understanding of pediatric brain tumor immunology impede the advancement of effective immunotherapies for pediatric brain tumors. Studying the immune tumor microenvironment in these patients can pave the way for the development of innovative treatment options and unveil new therapeutic avenues, ultimately enhancing clinical outcomes and improve the lives of affected children.

## 1.3 Gene Expression Analysis of RNA sequencing data

Bulk RNA sequencing, or RNA-Seq, is a technology that allows for the comprehensive quantification of the transcriptome, that is the complete set of RNA transcripts produced by the genome at a given time. RNA-Seq enables an unbiased view of the transcriptome, allowing for the discovery of novel transcripts and alternative splicing events [19]. The generation of count data begins with the preparation of an RNA sample, which is then sequenced using high-throughput technology. The sequenced reads are then mapped back to the reference genome or transcriptome. This mapping process allows to assign each read to a specific gene, resulting in raw count data. For each gene, the count is the number of reads that have been assigned to it, providing a measure of its expression level in the RNA sample [20]. When applied to cancer research, RNA-Seq can offer valuable insights into the molecular mechanisms underlying disease progression. Differential expression analysis of RNA-Seq data can identify genes that are upregulated or downregulated in tumor samples compared to normal tissues, or compared among different tumor types. These differentially expressed genes could contribute to tumorigenesis and may serve as potential targets for therapeutic intervention [21]. Differential expression analysis is an analysis method used to identify genes that exhibit significant changes in expression levels between different conditions. It provides key insights into the biological processes and pathways that are altered under different clinical conditions. By comparing gene expression profiles between different groups, we can identify genes that are specifically associated with a particular condition. These genes can serve as potential biomarkers for diagnosis, prognosis, or therapeutic targeting. Differential expression analysis can be applied to a wide variety of data types. For example, it can be used to compare gene expression profiles derived from microarray data, single-cell RNA sequencing (scRNA-seq) data, or bulk RNA sequencing data [22], [23].

## 1.4 Objectives

The primary objective of this research was to compare immune-related gene expressions in various types of high-grade pediatric brain tumors including ATRT, Ependymoma, Glioblastoma, Glioma, and Medulloblastoma, as compared to Craniopharyngioma, a low-grade and benign tumor[8]. This is due to the absence data of control group from healthy tissues at the time of the research. We addressed this objective through the application of differential gene expression analysis using DESeq2 R package.

Our secondary objective was to identify co-expressed gene modules(clusters of genes with similar co-expression) and explore their correlation with the various tumor types. In response to this objective, we utilized the Weighted Gene Correlation Network Analysis (WGCNA) method implemented in R.

By analyzing the immune-related gene expressions in these various pediatric brain tumors and understanding the gene clusters and their associations with specific tumor types, we hoped to gain insights into pediatric brain tumor immune micro-environment. The implementation of differential expression analysis in DESeq2 and WGCNA methods is detailed in methods Chapter 3.

# 2. Data

## 2.1 Sample collection and Meta data

The research cohort included 147 children patients with brain tumors. Data was collected between March 8th, 2019 and May 28th, 2022 after approval of the BioBank and Data Access Committee of the Princess Máxima Center, Utrecht, the Netherlands. The patients were diagnosed based on histopathological assessment. The age of the patients in the sample ranged from 0 to 19 years old, of which 82 of them are male and 65 are female (See Table 2.1). Tumor types included Atypical Teratoid Rhabdoid Tumor(ATRT), Craniopharyngioma, Medulloblastoma, Glioblastoma, Glioma, and Ependymoma. The tumor grades, classified as per the World Health Organization (WHO) guidelines, were included in the metadata [1], a data frame with sample rows samples( represent tumor tissues) and clinical trait columns shown in Table 2.2. The main differential expression analysis was based on count data and tumor type comparison from this data frame. We also explored the correlation between these tumor types and gene modules (gene clusters that exhibit high co-expression values across the samples).

**Table 2.1:** Distribution of Tumor types by Age Category and Gender

| Gender | Age | Tumor types | | | | | | Total |
|---|---|---|---|---|---|---|---|---|
| | | ATRT | Craniopharyngioma | Ependymoma | Glioblastoma | Glioma | Medulloblastoma | |
| Female | 0-4 | 5 | 0 | 4 | 0 | 7 | 1 | 17 |
| | 5-9 | 0 | 2 | 3 | 1 | 12 | 6 | 24 |
| | 10-14 | 0 | 1 | 0 | 0 | 8 | 4 | 13 |
| | 15-19 | 0 | 0 | 0 | 3 | 5 | 3 | 11 |
| | **Total** | **5** | **3** | **7** | **4** | **32** | **14** | **65** |
| Male | 0-4 | 5 | 2 | 12 | 1 | 8 | 6 | 34 |
| | 5-9 | 0 | 0 | 1 | 2 | 9 | 6 | 18 |
| | 10-14 | 0 | 5 | 0 | 2 | 4 | 10 | 21 |
| | 15-19 | 0 | 2 | 0 | 2 | 2 | 3 | 9 |
| | **Total** | **5** | **9** | **13** | **7** | **23** | **25** | **82** |

## 2.2 RNA Sequencing

The RNA sequencing procedure, a key part of data collection process, was undertaken by clinical and laboratory professionals at Princess Máxima Cancer Center. The process, detailed in the following paragraphs, ensured the high-quality of the gene expression data used in this study. Tumor tissue samples were acquired during standard surgical resection procedures from all the patients. Total RNA was isolated from fresh frozen tumor ma-

**Table 2.2:** Meta Data: data frame where rows are samples and columns are associated clinical traits.

| Sample | Diagnosis | Gender | Age |
|---|---|---|---|
| sample_1 | Medulloblastoma | female | 12.5 |
| sample_2 | Craniopharyngioma | male | 14.6 |
| sample_3 | Glioblastoma | female | 7.9 |
| sample_4 | Craniopharyngioma | male | 1.0 |
| sample_5 | Glioma | male | 2.0 |
| sample_6 | Glioma | female | 11.8 |
| ............... | ...................... | ...... | ..... |
| sample_147 | .. | ... | |

terial using the AllPrep DNA/RNA/Protein Mini Kit (QIAGEN) according to standard protocol on the QiaCube (Qiagen). RNA-seq libraries were generated with 300ng RNA using the KAPA RNA HyperPrep Kit with RiboErase (Roche), this libraries are complementary DNA(cDNA) which contain all the information from the tissue sample. Subsequently sequencing (reading small random sections of cDNA) was done using NovaSeq 6000 system (2x150 bp) (Illumina). The RNA sequencing data were processed as per the GATK 4.0 best practices workflow for variant calling, using a wdl and cromwell based workflow[24]. This included performing quality control with Fastqc (version 0.11.5) to calculate the number of sequencing reads and the insert size Picard (version 2.20.1) for RNA metrics output and MarkDuplicates. The raw sequencing reads were aligned using Star (version 2.7.0f) to the reference human genome called GRCh38 (complete set of human RNA) and gencode version 29 (Broad Institute. Picard. GItHub 2019). Finally, expression counts were determined at gene level using Subread Counts. These counts are mapped read counts generated from the raw read counts. These mapped counts represent the abundance of genes, which are the main focus of our analysis. The dataset contains count data for a total of 60,357 genes. Below Table2.3 is a snapshot of the count data structure.

**Table 2.3:** Count Data: Each row corresponds to a unique gene with its associated counts for each sample.

| Gene | Sample_1 | Sample_2 | Sample_3 | Sample_4 | Sample_5 |
|---|---|---|---|---|---|
| ENSG00000223972.5 | 139 | 54 | 43 | 70 | 94 |
| ENSG00000227232.5 | 1081 | 1377 | 778 | 1030 | 2218 |
| ENSG00000278267.1 | 13 | 7 | 8 | 29 | 7 |
| ENSG00000243485.5 | 134 | 69 | 47 | 39 | 20 |
| ENSG00000284332.1 | 0 | 0 | 0 | 0 | 0 |
| ENSG00000237613.2 | 31 | 27 | 1 | 20 | 1 |
| ................. | .. | .. | .. | ... | ... |
| ................. | .. | .. | .. | ... | ... |

## 2.3 Data pre-processing and Normalization

Gene counts are generally influenced by different factors that are less biologically relevant such as sequencing depth, or library composition. Each RNA sequencing experiment generates a large number of reads. However, during the quality control, alignment, and counting processes, some reads may be discarded, resulting in different total read counts(library size) for each sample. Therefore, an apparent higher gene expression in sample_-1 compared to sample_2 could be due to a larger library size of sample_1 rather than actual higher gene expression. To account for this, it is important to normalize the gene counts before performing differenial analysis. In our study normalization of count data was performed using DESeq2, which accounts sequencing depth variation among samples. The normalization processes at the back-end of the DESeq2 R package are summarized inf the following sections. The overall workflow of DESeq is depicted in Figure 2.1.

**Step 1: Estimating Pseudo-Reference Sample:** DESeq2 starts by creating a representative sample called the pseudo-reference sample. It does this by calculating the average gene expression for each gene across all samples. This helps establish a baseline for comparison.

**Step 2: Calculating Gene Expression Ratios:** DESeq2 calculates the ratios of gene expression counts for each gene in each sample relative to the pseudo-reference sample. This shows how much a gene is expressed in comparison to the average expression level.

**Table 2.4:** Normalization: size factor estimation

| Gene | Sample_1 | Sample_2 | Pseudo-reference sample |
|---|---|---|---|
| ENSG00000223972.5 | 139 | 54 | $\sqrt{139 \times 54} = 334.93$ |
| ENSG00000227232.5 | 1081 | 1377 | $\sqrt{1081 \times 1377} = 1184.49$ |
| ENSG00000278267.1 | 13 | 7 | $\sqrt{13 \times 7} = 8.37$ |
| ENSG00000243485.5 | 134 | 69 | $\sqrt{134 \times 69} = 110.63$ |
| ENSG00000284332.1 | 0 | 0 | $\sqrt{0 \times 0} = 0$ |

| Gene | Sample_1 | Sample_2 | Ratio of Sample_1/ref | Ratio of Sample_2/ref |
|---|---|---|---|---|
| . | 139 | 54 | $\sqrt{139 \times 54} = 334.93$ | $\frac{54}{334.93} = 0.161$ |
| . | 1081 | 1377 | $\sqrt{1081 \times 1377} = 1184.49$ | $\frac{1377}{1184.49} = 1.164$ |
| . | 13 | 7 | $\sqrt{13 \times 7} = 8.37$ | $\frac{7}{8.37} = 0.837$ |
| . | 134 | 69 | $\sqrt{134 \times 69} = 110.63$ | $\frac{69}{110.63} = 0.623$ |
| . | 0 | 0 | $\sqrt{0 \times 0} = 0$ | $\frac{0}{0} = \text{NaN}$ |

**Step 3:Median Ratio Calculation:** DESeq2 then finds the median (middle value) of these ratios for each sample. By using the median, which is less affected by extreme values, it ensures that rare genes or outliers don't overly influence the normalization process.The normalization_factor for sample_1 and sample_2 are calculated as median(0.415, 0.912, 1.554, 1.211, 0) = 0.924 and median(0.161, 1.164, 0.837, 0.623, 0)= 1.1469 respectively.

**Step 4: Normalizing Gene Expression:** finally, count values are normalized by dividing them using the median values (size factors) as shown in Table2.5.

**Table 2.5:** Normalization of raw counts

| Gene | Sample_1 | Sample_2 |
|---|---|---|
| ENSG00000223972.5 | $\frac{139}{0.9240923} = 150.41$ | $\frac{54}{1.1469482} = 47.09$ |
| ENSG00000227232.5 | $\frac{1081}{0.9240923} = 1169.46$ | $\frac{1377}{1.1469482} = 1200.75$ |
| ENSG00000278267.1 | $\frac{13}{0.9240923} = 14.07$ | $\frac{7}{1.1469482} = 6.10$ |
| ENSG00000243485.5 | $\frac{134}{0.9240923} = 145.14$ | $\frac{69}{1.1469482} = 60.11$ |
| ENSG00000284332.1 | $\frac{0}{0.9240923} = 0.00$ | $\frac{0}{1.1469482} = 0.00$ |

The impact of normalization becomes more evident, for instance when

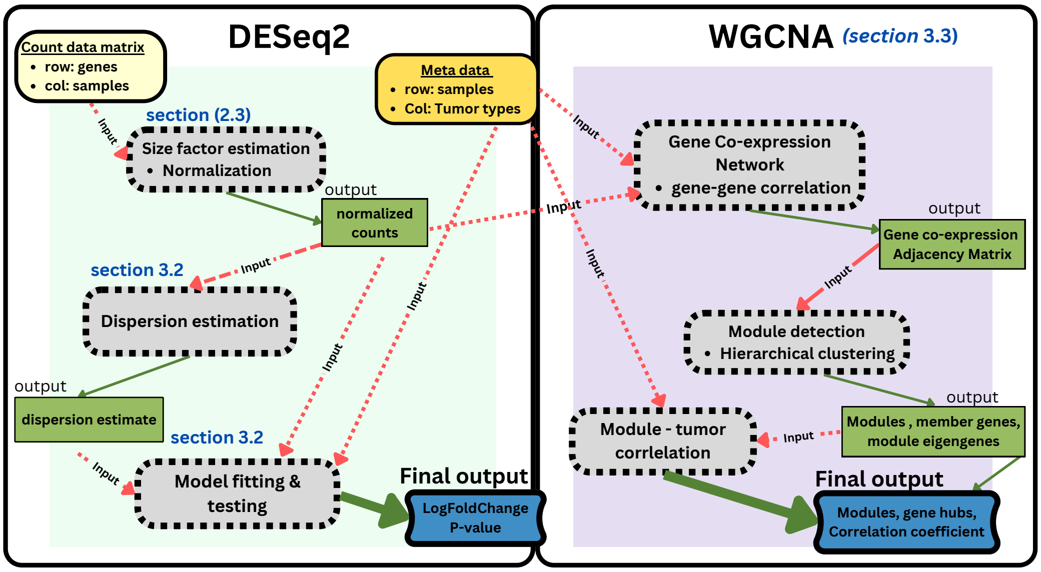


**Figure 2.1: Workflow of expression analysis using DESeq2 and WGCNA:**DESeq2 starts with the input of Count Data, which undergoes Size Factor Estimation to produce Size Factors. These are then used to normalize the Count Data, resulting in Normalized Count Data. This Normalized Count Data is used in Dispersion Estimation to generate Dispersion Estimates and in Model Fitting, along with Meta Data. Final results include LogFoldChange and P-values. The right box, labeled as WGCNA, represents the process of module detection and correlation with tumor types. It takes normalized count data and Meta Data as input and performs Network Formation to create a Network, Module Detection to identify Modules, and Correlation to establish a Correlation with tumor types. Final results include Modules, Correlation Coefficient and Hub Genes

comparing the expression levels of *MAGEA3* in glioma and medulloblastoma. Without normalization, a higher expression is observed in glioma (raw count: 1334), potentially leading to skewed interpretations. However, normalization reveals a higher *MAGEA3* expression in medulloblastoma (normalized count: 2282), illustrating the need for normalization to ensure accurate and unbiased comparisons across samples.

**Table 2.6:** Raw and normalized counts of *MAGEA3* expression in glioma and medulloblastoma samples

| | Glioma (Sample_41) | Medulloblastoma (Sample_63) |
|---|---|---|
| Raw Counts | 1334 | 975 |
| Normalized Counts | 1091 | 2282 |

## 2.4 Exploratory data analysis

To understand the underlying structure and sample relationships, we performed unsupervised analyses. Principal component analysis (PCA) was conducted using the log-transformed normalized counts of the count data. By reducing our high-dimensional data into key components, PCA captured the main trends of variation in gene expression. This gave us a visualization where we could see how samples relate to each other and spot outliers. These insights from the PCA are crucial as they provide important insights into the overall structure of the data and guide further downstream analyses[25]. The PCA plot displayed in Figure 2.2 visually demonstrates distinct clustering patterns among different tumor types. Each tumor type forms separate clusters, indicating that they contribute significantly to the observed variation in the dataset. The plot reveals a clear separation between Medulloblastoma, Craniopharyngioma, Glioma, ATRT, Ependymoma, and Glioblastoma samples, suggesting that these tumor types have distinct gene expression profiles. And there are no notable outliers observed.

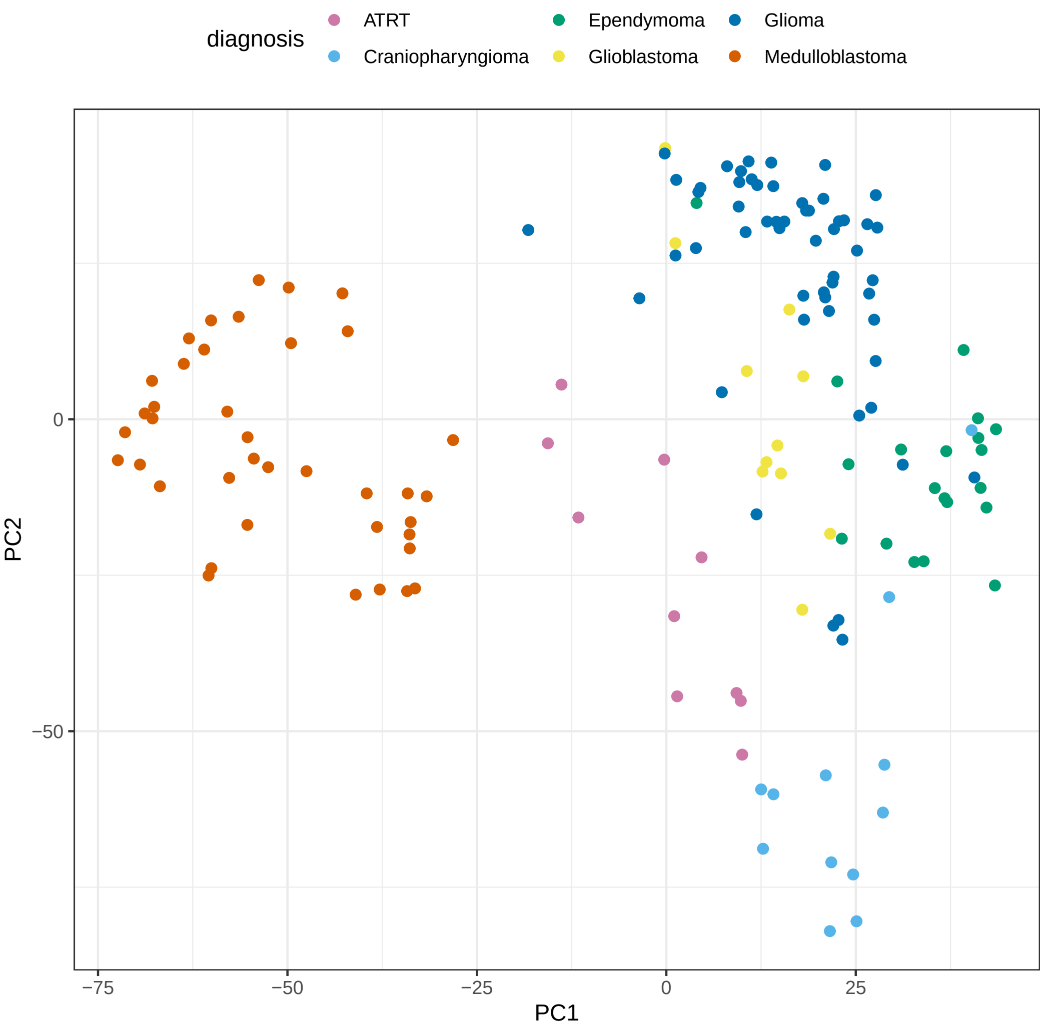


**Figure 2.2:** illustrates a Principal Component Analysis(PCA) of the gene expression data. PC1 and PC2, the axes, capture the maximum variation within the dataset. Each point represents a sample, color-coded by tumor type. The position of each sample is determined by the expression of immune-related genes in the sample. Samples that cluster together share similar gene expression profiles, highlighting potential groupings of tumors. There are no notable outlier samples.

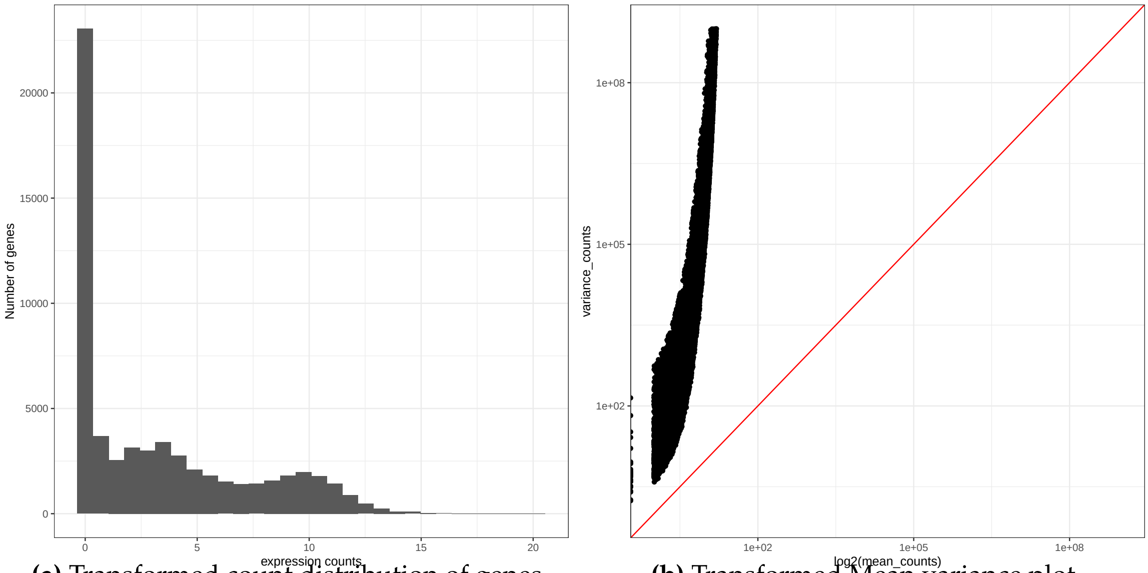
# 3. Method

This chapter discusses the details of our analysis methods, focusing on DESeq2 for differential expression analysis, and Weighted Gene Co-expression Network Analysis (WGCNA). These techniques form the core of our analysis pipeline, which was illustrated in Figure 2.1.

## 3.1 Differential Expression Analysis with DESeq2

The differential gene expression analysis was performed using the DESeq2, a commonly used R package for differential gene expression analysis of count data from high-throughput RNA sequencing. DESEq2 implements statistics such as variance estimation through a Negative Binomial Distribution, aiding in the identification of differentially expressed genes by applying negative binomial Generalized Linear Model (GLM). Additionally, DESeq2 employs statistical testing methods such as the Wald test and the Likelihood Ratio test, which are crucial for determining statistical significance. Importantly, DESeq2 also controls the false discovery rate (FDR) during multiple testing. These attributes make DESeq2 a suitable choice for RNA-Seq data analysis[26]–[29].

The count data in our research exhibits the characteristics of most RNA-seq data. A considerable number of genes with low expression levels, a long right tail, representing genes with high expression levels, and a wide dynamic range (gene expression level ranges from 0 to millions), capturing a broad span of expression levels as shown in Figure 3.1a). Moreover, the relationship between the mean and variance in the count data is not linear, with genes exhibiting higher mean expression levels tending to have greater variance across samples, as indicated by the scatter plot above the red line in Figure 3.1b). To accurately model count data, DESeq2 employs a negative-binomial (NB) distribution. The NB distribution is well-suited for RNA-seq counts due to their observed overdispersion, where counts show greater variability than expected under a Poisson distribution Figure 3.1b). The NB distribution incorporates an additional parameter, the dispersion ($\alpha$), which captures the relationship between the mean and variance of the normalized



**(a)** Transformed count distribution of genes

**(b)** Transformed Mean variance plot

**Figure 3.1:** Characteristics of gene count distribution across samples

counts. It is modeled using the formula:

$$K_{ij} \sim \text{NB}(\text{mean} = \mu_{ij}, \text{dispersion} = \alpha_i) \tag{3.1}$$

Here, $K_{ij}$ represents the observed count for gene $i$ in sample $j$, $\mu_{ij}$ denotes the mean count, and $\alpha_i$ represents the gene-specific dispersion parameter. The mean expression levels can be easily estimated using the observed normalized counts across all tumor types. In contrast, the estimation of dispersion, which will be discussed below, provides insights into the variability of gene expression.

The following sub-sections will highlight the main statistical calculations behind DESeq2 package.

### 3.1.1 Dispersion Estimation

The dispersion parameter quantifies the variability of gene counts within each tumor type, representing how the variance deviates from the mean. For instance, a dispersion of 1 indicates no deviation from the mean. DESeq2 estimates the dispersion value for each gene based on its mean expression level and observed variance across samples. The dispersion formula is given by[27]:

$$\text{Dispersion} = \frac{\text{variance of counts}}{\text{mean squared count}} \tag{3.2}$$

For instance the estimation of dispersion for Gene A and Gene B in certain tumor type is demonstrated in Table 3.1

**Table 3.1:** This table demonstrates how dispersion is estimated for each gene within one tumor Type based on the observed count data.

| Gene | Mean Expression | Variance of Counts | Dispersion |
|---|---|---|---|
| Gene A | 10 | $(10 - 7.67)^2 = 5.29$ | $5.29/100 = 0.0529$ |
| Gene B | 8 | $(8 - 5.67)^2 = 5.29$ | $5.29/64 = 0.0827$ |

In DESeq2, dispersion estimates are obtained by considering the relationship between mean expression and variation values, as illustrated in the dispersion plot figure 3.2. Moreover, to improve the reliability of dispersion estimates, DESeq2 employs a shrinkage method that shares information across genes. This approach ensures that genes with similar expression levels have similar dispersion values. The resulting shrunken dispersion values are represented by blue dots in the dispersion figure 3.2.

### 3.1.2 Model Fitting

The differential expression analysis in DESeq2 uses a generalized linear model (GLM)[27]. In our case the GLM captures the relationship between gene expression of tumor types as compared to the reference tumor type. For our analysis we used, DESeq2 that applies a simplified form of the GLM where we only consider a single predictor, namely the tumor type. The simplified model can be expressed as:

$$\log_2(count_{ij}) = \beta_0 + \beta_i \cdot x_j + \epsilon \tag{3.3}$$

Here, $\beta_0$ represents the baseline log2 expression level for the reference tumor type, and $\beta_i$ denotes the log2 fold change for each gene between the ref-

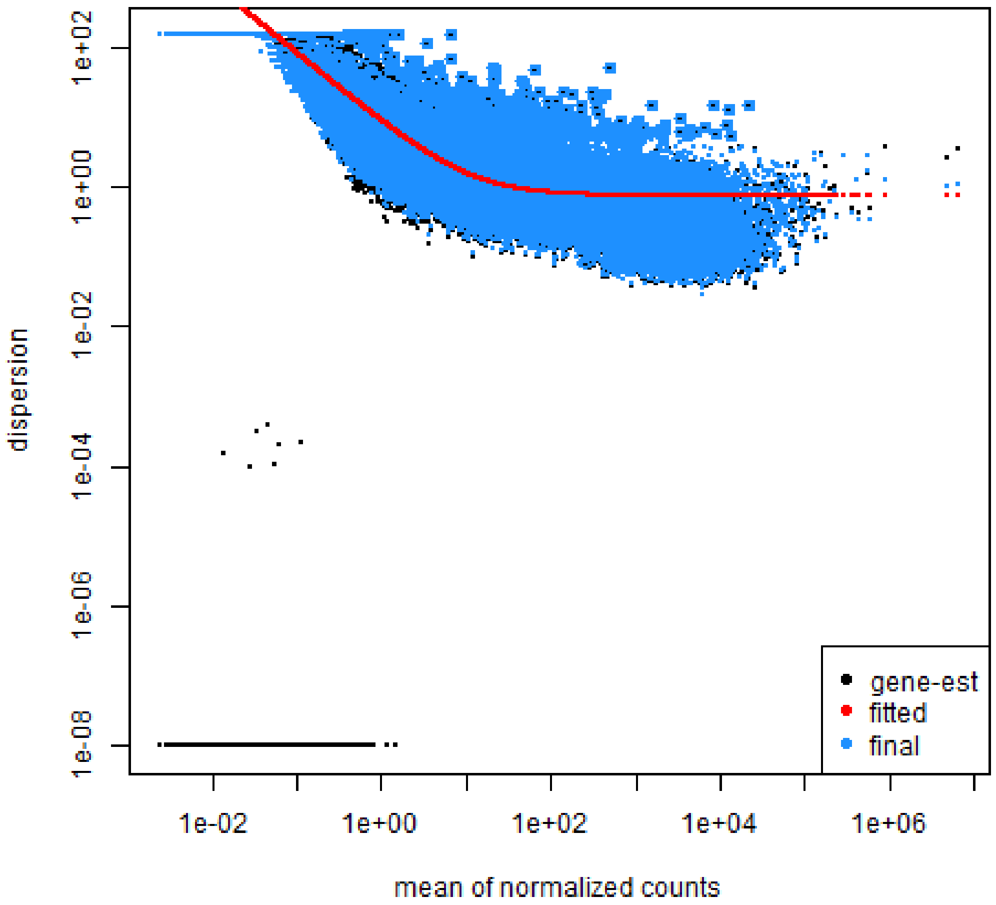


**Figure 3.2:** The plot shows each gene represented by a black dot, where the x-axis represents the mean expression and the y-axis represents the dispersion.

erence tumor type and the tumor type of interest. The variable $x_j$ indicates the tumor type for each sample, taking the value 1 for samples from tumor type of interest, 0 otherwise. For instance, if *Gene A* in Medulloblastoma shows a log2 fold change of -1.96, this means *Gene A*'s expression is approximately halved (since $2^{-1.96} \approx 0.5$ or a 50% decrease) in Medulloblastoma compared to Craniopharyngioma. In this example, the negative log2 fold change indicates that Gene A is downregulated in Medulloblastoma relative to Craniopharyngioma, and the magnitude of the fold change (-1.96) suggests that its expression level in Medulloblastoma is about half of that in Craniopharyngioma.

### 3.1.3 Hypothesis Testing

DESeq2 implements hypothesis testing to assess the significance of differential expression each gene, in specifc tumor types compared to Craniopharyngioma. The goal of hypothesis testing is to determine whether the observed difference in read counts for a given gene is greater than what would be expected due to natural random variation. The null hypothesis $H_0$ assumes no differential expression, which translates to a log2 fold change ($log2FC$) of zero for a given gene. To test this hypothesis, DESeq2 utilizes the Wald test, a statistical test that evaluates whether the data provides sufficient evidence to reject the null hypothesis. For a detailed mathematical explanation of how dispersion estimates and shrinkage, model fitting are performed in DESeq2, readers are encouraged to refer to the original article and accompanying methodological chapter by Love, Huber, and Anders [27].

### 3.1.4 Implementation of DESeq2 in R

DESeq2 package was implemented in R *version 4.2.2*. Due to absence of samples from health tissue at the time of this research, Craniopharyngioma, a benign and low-grade type of tumor was used as a baseline comparison tumor type during the differential expression analysis. The implementation of DESeq2 in R together with R code is presented under the appendix Section 7.1.

## 3.2 Weighted Gene Co-expression Network Analysis (WGCNA)

For the Weighted gene co-expression network analysis(WGCNA), the WGCNA R package was used. This was used to detect the presence of gene module (cluster) structure after gene co-expression network construction. Afterward, we explored the correlation between modules using eigengene[1] and their correlation with the tumor type[30]. The steps detailing the implementation of WGCNA are provided below.

Step 1: Construction of a similarity matrix

The construction of our weighted gene co-expression network started by defining a measure of similarity between each pair of genes based on their co-expression (based on normalized counts) across all samples. We utilized the absolute value of Pearson correlation as similarity measure ($s_{ij}$), which is defined between gene $i$ and gene $j$, across all samples as [30]:

$$s_{ij} = |cor(i,j)| \tag{3.4}$$

We then generated a similarity matrix dataset, $S = [s_{ij}]$, with a pairwise similarity measure for all genes.

**Table 3.2:** Gene co-expression matrix, with each entry representing the pairwise correlation coefficient between the expression profiles of two genes, rounded to one decimal place. This table only shows the first five genes.

| **Gene ID**[2] | **ENSG001** | **ENSG002** | **ENSG003** | **ENSG004** | **ENSG005** |
|---|---|---|---|---|---|
| ENSG001 | 1.0 | 0.5 | 0.1 | 0.2 | 0.2 |
| ENSG002 | 0.5 | 1.0 | 0.2 | 0.2 | -0.0 |
| ENSG003 | 0.1 | 0.2 | 1.0 | 0.4 | 0.0 |
| ENSG004 | 0.2 | 0.2 | 0.4 | 1.0 | 0.2 |
| ENSG005 | 0.2 | -0.0 | 0.0 | 0.2 | 1.0 |
| ... | ... | ... | ... | ... | ... |

[1] Eigengene is defined as the first principal component of a given module. It can be considered a representative of the gene expression profiles in a module.

Step 2: Transformation of the similarity matrix into an adjacency matrix

The transformation of the similarity matrix $S$ into an adjacency matrix $A$ emphasizes significant gene-gene relationships and minimize the influence of weaker, potentially spurious associations. Our transformation was based on a power adjacency function, ensuring the construction of a weighted network that retains information about the strength of gene interactions (co-expression values were transformed to [0, 1]).

For a given pair of genes $i$ and $j$, the adjacency $a_{ij}$ is computed as:

$$a_{ij} = s_{ij}^{\beta} \quad (i \neq j) \tag{3.5}$$

Here, $\beta$ denotes the soft-thresholding parameter that fine-tunes the sensitivity of the network. In this analysis, based on the scale-free topology criterion, a $\beta$ value of 4 was chosen. This specific $\beta$ value ensured that the network adhered to the scale-free topology (with $R^2 > 0.8$). Choosing a parameter is a trade-off between a high scale-free topology fit (R2) and the mean number of connections for the network. For instance, a parameter value that leads to an R2̂ value close to 1 may lead to networks with very few connections. The authors of the methodology, Bin Zhang and colleagues suggest, choosing a parameter value that leads to satisfying scale-free topology, such as $R^2 > .8$. Our choice of $\beta$ was in line with this recommendation. Based on this, the adjacency matrix was generated as displayed in table3.3.

Step 3: Hierarchical Clustering and Module Detection

After creating the gene network, in this step, we applied average linkage hierarchical clustering for module detection. Modules are clusters of densely interconnected genes. This process resulted in a dendrogram where each branch represents a module of highly co-expressed genes. The definition of modules was achieved by Dynamic Branch Cut methods of the

[2] Gene ID*s are made up for demonstration.

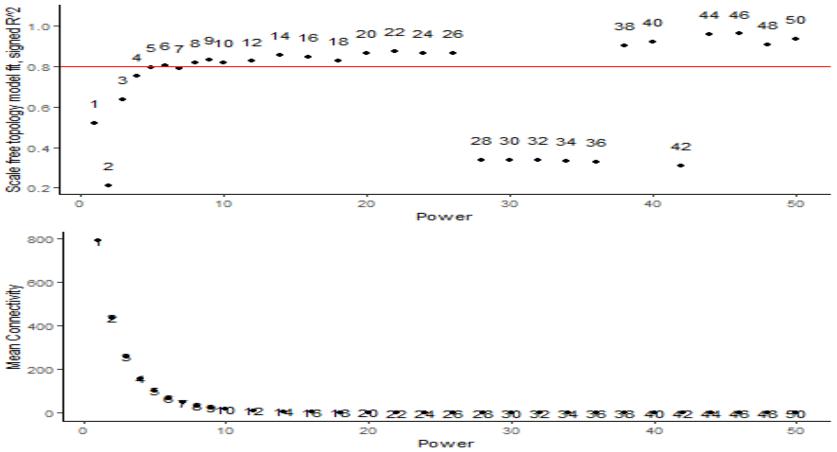


**Figure 3.3:** Network properties for different hard and soft thresholds, hard thresholds (top row) and soft thresholds (bottom row): the top plot visualizes scale-free topology with regression fitting index $R^2$(top plot), and the bottom plot is mean connectivity. Points are labeled by the corresponding adjacency function parameter.

**Table 3.3:** Snapshot of the adjacency matrix, derived from the gene co-expression matrix using a power function (power = 4). Each cell represents the adjacency between two genes, calculated as the absolute value of the co-expression raised to the power of 4.

| | **ENSG001** | **ENSG002** | **ENSG003** | **ENSG004** | **ENSG005** |
|---|---|---|---|---|---|
| **ENSG001** | 1.00 | 0.06 | 0.00 | 0.00 | 0.00 |
| **ENSG002** | 0.06 | 1.00 | 0.00 | 0.00 | 0.00 |
| **ENSG003** | 0.00 | 0.00 | 1.00 | 0.04 | 0.00 |
| **ENSG004** | 0.00 | 0.00 | 0.04 | 1.00 | 0.00 |
| **ENSG005** | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| **...** | ... | ... | ... | ... | ... |

dendrogram applied in WGCNA R package. This was indeed one of the limitations as it was difficult to determine the presence of clusters[31].

Step 4: Module Clinical traits correlation

Finally, the relationship between the identified modules of gene expression and the tumor types were assessed using correlation analysis. To assess the relationships, the module eigengenes[3] for each module, calculated

[3] **ModuleEigengenes** summarizes the gene expression of entire co-expression modules. This is done by performing singular value decomposition (SVD) on a subset of the scaled expression matrix containing only genes that are assigned to each module. The module eigengene (ME), defined as the first dimension of the SVD matrix, retains the most variation, and we use this vector as a summary of gene expression for the whole module.[30]

in the previous step as the first principal component of the gene expression data for the module, were correlated with the tumor types. This correlation essentially measures how much the major pattern of gene expression in a module corresponds with each type of tumor. Each module eigengene represents the major gene expression profile for a module. Therefore, by correlating these with tumor types, we aimed to identify modules whose gene expression patterns are strongly associated with particular tumor types. These modules, in turn, provide insights into the groups of genes that act might be critical to characterize pediatric brain tumors. The correlation coefficients range from -1 to 1, with values close to -1 indicating a strong negative association (as the module eigengene value increases, the likelihood of the particular tumor type decreases), values close to 1 indicating a strong positive association (as the module eigengene value increases, the likelihood of the particular tumor type also increases), and values close to zero indicating no or weak association. Here, we conducted this correlation analysis for each module against each tumor type, providing a comprehensive overview of which modules are relevant to each type of tumor and the nature of their relationship, whether they are positively or negatively associated. This step in our analysis helps in identifying gene networks that may play significant roles in specific types of pediatric brain tumors that require further analysis.

# 4. Results

The findings from differential expression and WGCNA analysis are presented in the sections that follow.

## 4.1 Differential Expression Analysis

For the differential analysis, the gene expressions in ATRT, Ependymoma, Glioblastoma, Glioma, and Medulloblastoma are computed in comparison to Craniopharyngioma. This allowed us to identify key differences in gene expression profiles among these tumor types. The differential expression analysis yields a DESeq object, where each row is a unique gene Ensembl identifier, and columns represent various statistical metrics. While all these metrics contribute to the overall understanding of gene expression variability, our analysis primarily focuses on the log2FoldChange and padj values. These two metrics, respectively, allow us to identify the magnitude and direction of gene expression changes, and to account for multiple testing to minimize false discovery rate. The details of these metrics, along with others, are provided in Table 4.1. We defined a widely used cut-off point for the *significantly differentially expressed* genes if thier adjusted P-value is less than 0.05. Moreover, upregulated or downregulated genes are defined by *adjusted P-value* less than 0.05 and greater than 2 absolute *logfoldchange*.

**Table 4.1:** Description of Key Columns in Differential Expression Analysis DESeq Result

| Statistic | Description |
|---|---|
| baseMean | The mean of normalized count values, averaged over all samples. It provides a measure of the overall expression level of a gene across all samples. |
| **log2FoldChange** | **Represents the log2 fold change in gene expression between Medulloblastoma and Craniopharyngioma (the baseline). A negative value indicates downregulation in Medulloblastoma compared to Craniopharyngioma, while a positive value indicates upregulation.** |
| lfcSE | The standard error of the log2 fold change estimate. It provides a measure of the uncertainty associated with the log2 fold change estimate. |
| stat | It is the Wald statistic, which is the log2 fold change estimate divided by its standard error. It is used to test the null hypothesis that the log2 fold change is zero (i.e., there is no difference in gene expression between the two conditions). |
| pvalue | The p-value associated with the Wald test. A small p-value (typically, less than 0.05) suggests that we can reject the null hypothesis of no difference in gene expression between the two conditions. |
| **padj** | **The p-value adjusted for multiple testing using the Benjamini-Hochberg procedure. It controls the false discovery rate, which is the expected proportion of false positives among all genes declared differentially expressed.** |

From the differential expression analysis identified significant differences in the expression of immune-related genes across different tumor types when compared to Craniopharyngioma. Medulloblastoma, for example, exhibited the highest percentage of downregulated genes (64%), suggesting a potential suppression of certain immune-related functions in this tumor type, although further functional annotation is required. On the other hand, Ependymoma had 7.7% of genes upregulated. Similarly, Glioma showed 10% upregulated and 45% downregulated genes, while Glioblastoma also demonstrated significant differential gene expression. These findings provide baseline insights into the immune landscapes of these tumors.

**Table 4.2:** Differential Expression Analysis of Immune-Related Genes: Each row represents the type of brain tumor. The columns provide the number and percentage of genes that are upregulated and downregulated in each tumor type compared to Craniopharyngioma.

| DEGs compared to Craniopharyngioma | Upregulated Genes (%) | Downregulated Genes (%) |
|---|---|---|
| ATRT | 121 (7.1%) | 722 (43%) |
| Ependymoma | 131 (7.7%) | 823 (48%) |
| Glioblastoma | 98 (5.8%) | 548 (32%) |
| Glioma | 173 (10%) | 763 (45%) |
| Medulloblastoma | 201 (12%) | 1092 (64%) |

The number of Differentially Expressed Genes (DEGs) varied between 343 in Glioblastoma and 863 in Medulloblastoma. Remarkably, each tumor type displayed a distinct set of DEGs not identified in the other tumor types. Specifically, ATRT had 30, Ependymoma had 37, Glioblastoma had 8, Glioma had 14, and Medulloblastoma had 271. The barplot 4.1 below shows the number of differentially expressed genes (DEGs) across five tumor types. Each blue dot represents DEGs intersection and its corresponding bar denotes the number of genes in the specific tumor type. For instance, the first bar with a single dot under Medulloblastoma reveals 271 exclusively differentially expressed number of genes in that type, while the subsequent bar implies 179 genes that are differentially expressed shared with at least one other tumor type.

The Volcano plot presented in Figure 4.2 showcases the differentially expressed genes in ATRT compared to Cranipharyngioma. The dots are individula gene. Significantly expressed genes are in red color. The X-axis shows the logFoldChange. Upregulated genes are at the right side of zero, while downregulated genes are ploted left to zero. The absolute logfoldchanges increase as we move either dierction from zero. Beside the MA plot provides interesting insight on the relationship between the fold change and the average expression of genes. the blue dots are genes that are differentially expressed, and, genes deviating from the center line indicate the absolute value change. Genes below the central line being downregulated and genes above the central are upregulated. Further illustrations, including additional Volcano and MA plots per each tumor type, can be found in the Appendix (Section 8.1) to support the presented results.

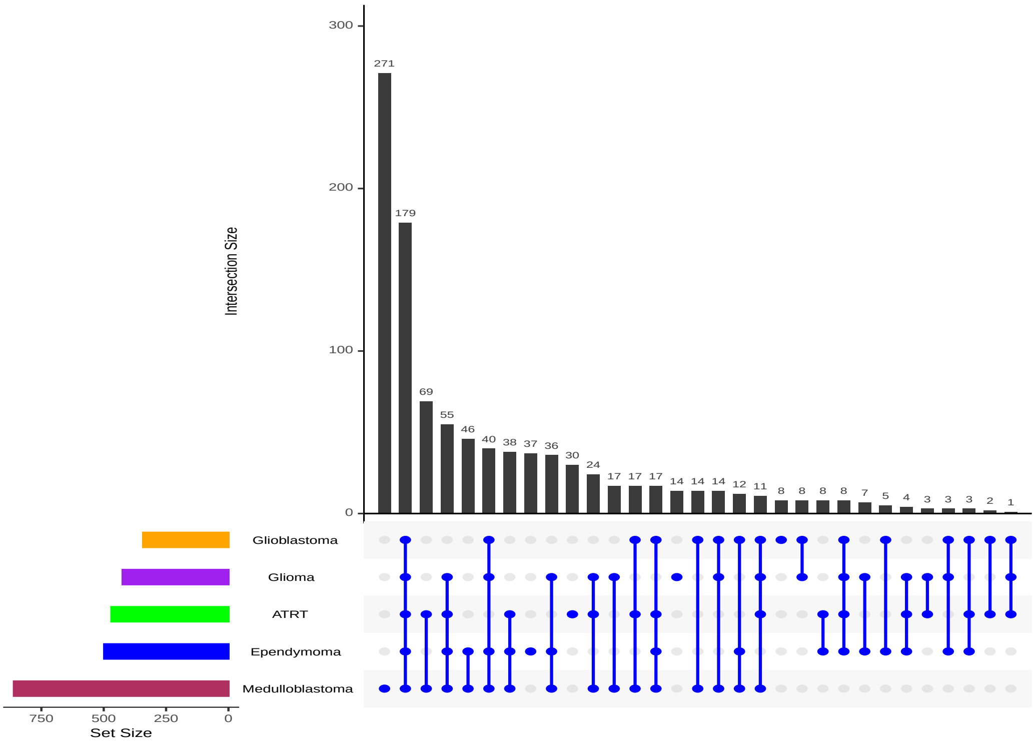


**Figure 4.1:** UpSetR shows overlapping differentially expressed genes among five tumor types (ATRT, Ependymoma, Glioblastoma, Glioma, and Medulloblastoma).

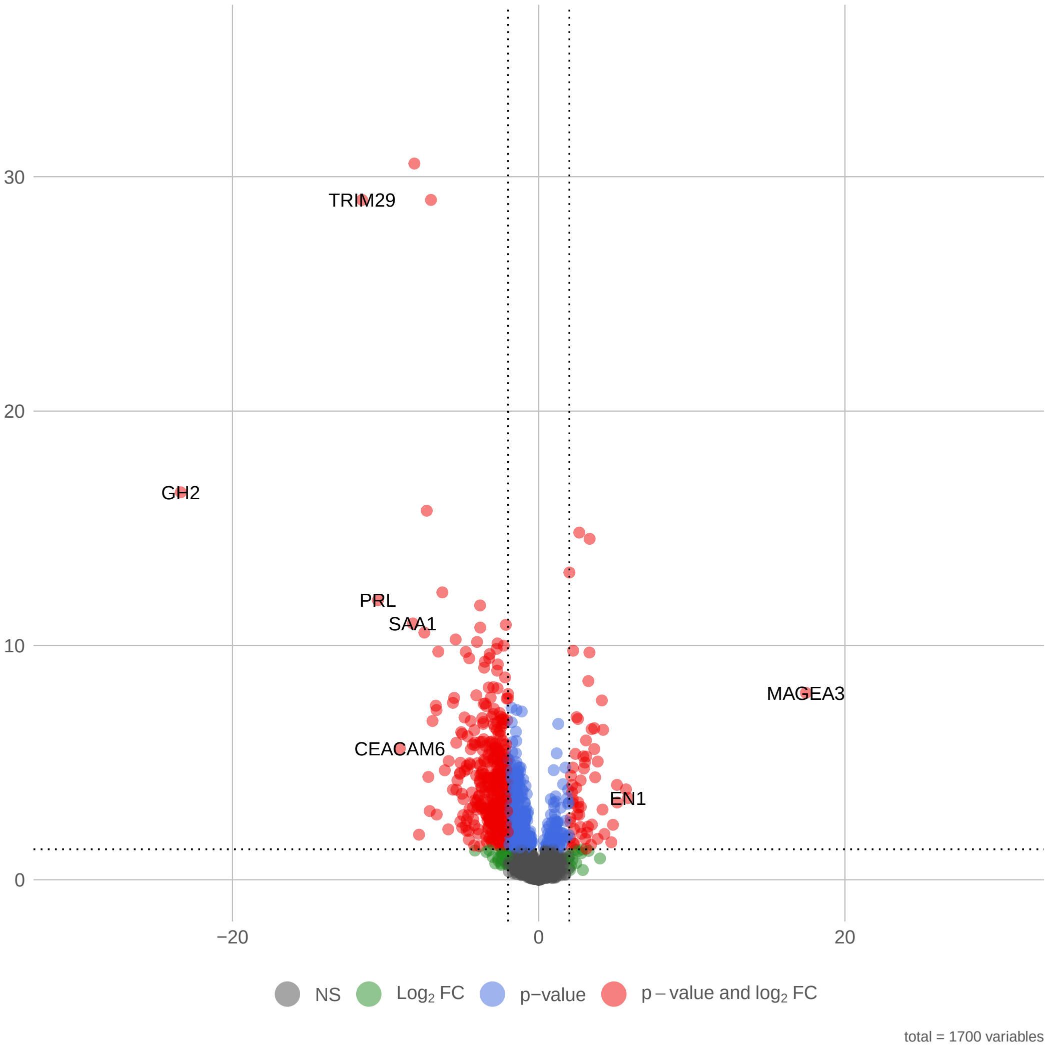


**Figure 4.2:** The x-axis represents the log2 fold change values, and the y-axis represents the -log10 transformed p-values. Each point on the plot corresponds to a gene. Genes with significant differential expression are highlighted in red and lie beyond the absolute Log2FoldChange threshold, which is 2. For instance, the gene MAGEA3, which is indicated by a point at approximately 14 on the x-axis and 6 on the y-axis, is expressed at a level that is approximately 16384 times upregulated in ATRT (2̂(14) = 16384) compared to Craniopharyngioma.

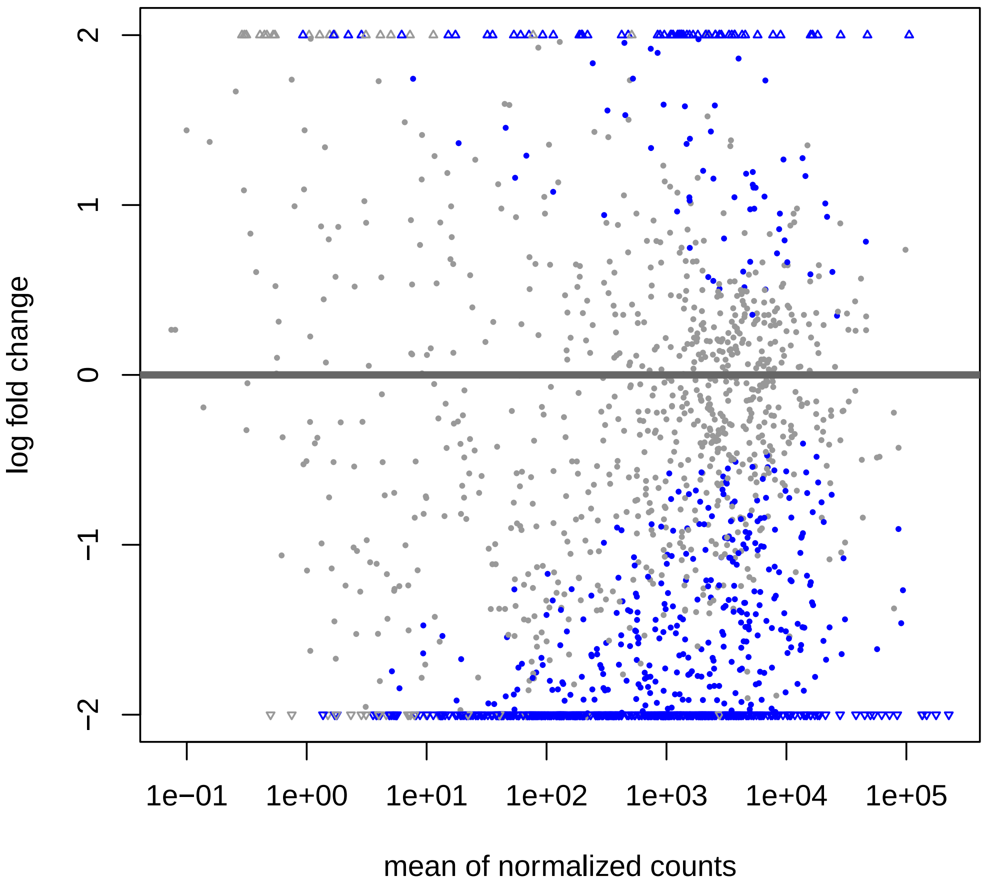


**Figure 4.3:** The x-axis represents the average normalized count, while the y-axis represents the log2 fold change. Each blue point on the plot corresponds to genes that are significantly differentially expressed (beyond the absolute value of log2 fold change threshold, P-value <0.05). For instance, a gene with a log2 fold change of -2 is expressed at a level that is one-fourth ($2^{-2} = 0.25$) of its expression level in Craniopharyngioma, indicating that it is four times downregulated in ATRT tumors compared to Craniopharyngioma.

We ranked the top five upregulated and downregulated genes based on a p-value <0.05 and an absolute logFoldChange value greater than 2. The gene *MAGEA3* appeared as the top consistently upregulated gene across all tumor types with the highest fold change in Medulloblastoma (approximately 25,794,709-fold[1]), and the lowest in Glioma (approximately 514,825-

[1] LogFoldChange is calculated from the log2 fold change using the formula: $FC = 2^{log2FC}$. For instance, for 'SYCP1' in Medulloblastoma, where $log2FC = 12.477781$, $FC =$

fold). Moreover, gene *SYCP1* was the second top upregulated gene across all tumor types except in ATRT (see Table 4.3). The downregulation pattern of gene expression is less consistent across tumor types. For instance, *COL17A1* is downregulated accross all tumor types except in ATRT. It is worth noting this is not the gene with highest LogFoldChange except in Medulloblastoma tumor types. Moreover, *MMP12* is most underegulated gene in Ependymoma and Glioma tumor types. There are no similar studies that use craniopharyngioma as refernece tumor establish this down regulation comparison of gene expressions. The top 5 most downregulated genes are summarized in Table4.3 with their respective LogFoldChange.

**Table 4.3:** Top 5 Upregulated and Downregulated genes with their respective LogFoldChange values across different tumor types

| **Upregulated Genes(*Highlighted in red are genes that are upregulated across all tumor*)** | | | | |
|---|---|---|---|---|
| **ATRT** | **Ependymoma** | **Glioblastoma** | **Glioma** | **Medulloblastoma** |
| MAGEA3 (14.91) | MAGEA3 (11.70) | MAGEA3 (21.87) | MAGEA3 (18.95) | MAGEA3 (24.66) |
| EN1 (5.83) | SYCP1 (8.01) | SYCP1 (5.84) | SYCP1 (6.68) | SYCP1 (12.47) |
| SYCP1 (5.81) | RELN (5.69) | DLL3 (5.55) | MAGEA12 (7.56) | MAGEA12 (7.56) |
| PRAME (5.69) | FBP2 (5.57) | EN1 (5.16) | MAGEC2 (4.99) | GNGT1 (6.84) |
| PMCH (5.12) | HOXD4 (5.50) | MAGEA12 (5.03) | KLRC2 (4.98) | RELN (6.73) |
| **Downregulated Genes(*genes highlighted in green are genes downregulated frequently*)** | | | | |
| GH2 (-25.12) | MMP12 (-11.84) | PRL (-11.71) | MMP12 (-12.03) | COL17A1 (-10.87) |
| CCL11 (-23.97) | PITX2 (-11.23) | COL17A1 (-11.25) | PRL (-10.57) | TRIM29 (-10.76) |
| TRIM29 (-11.54) | SERPINB7 (-10.97) | FGF19 (-9.67) | CCL11 (-10.19) | PLA2G2A (-9.54) |
| PRL (-10.50) | CEACAM6 (-9.62) | GH2 (-9.47) | COL17A1 (-9.46) | LAMB3 (-9.28) |
| CEACAM6 (-9.07) | COL17A1 (-9.45) | ALDH3B2 (-8.60) | WNT3A (-9.46) | SAA1 (-9.08) |

$2^{12.477781} \approx 5,684$. This means the expression of 'SYCP1' is about 5,684 times higher in Medulloblastoma compared to the Craniopharyngioma.

## 4.2 Weighted Correlation Network Analysis

In the application of the Weighted Gene Co-expression Network Analysis (WGCNA), eight distinct gene modules emerged. These modules are clusters of genes with similar expression patterns, consisted of different numbers of genes, ranging from 35 to 408 and the top 5 most interconnected (hub) genes presented in table 4.4.

**Table 4.4:** Gene Modules and Their Top 5 highly connected genes in their respective module

| Modules | Number of Genes | Top 5 Hub Genes |
|---|---|---|
| Module1 | 35 | *ZNF205, MBD3, DVL1, TELO2, MAP2K2* |
| Module2 | 408 | *TRIM33, MAPK8, CCNT2, ATF2, TXNDC16* |
| Module3 | 40 | *CDC42, PSMA1, PRDX3, PSMA6, PSMC2* |
| Module4 | 73 | *RAD51, CCNB2, CHEK1, BIRC5, CDC25A* |
| Module5 | 59 | *ARHGEF6, MASP1, PRKCA, FEZ1, CX3CR1* |
| Module6 | 356 | *CD53, C3AR1, LAPTM5, ITGB2, PTPRC* |
| Module7 | 358 | *IL7R, IL2RG, SLAMF1, SLAMF6, IL1RN* |
| Module8 | 38 | *COL4A1, COL4A2, HSPG2, CDH5, ITGA1* |

### 4.2.1 Module -trait correlation

Following the module detection steps from the previous subsection, we explored the potential of module to clinical trait correlation, particularly their relationships with various types of tumors. To this end, we computed the correlations of these modules' eigengenes with different tumor types. The eigengenes, serving as representative gene expression profiles within the respective modules, demonstrated varied degrees of correlation across the tumor types. The relationship between the module eigengenes and the tumor types is captured in the correlation table 4.4 shown below. Notably, three modules exhibited a high correlation with Medulloblastoma: Module 2, Module 5, and Module 6, with correlation coefficients of 0.82, -0.74, and -0.69, respectively.

To understand the gene's significance in correlation with the tumor types, we looked at Module 2, which showed a strong positive correlation with Medulloblastoma. The correlation analysis based normalized mean count of each sample and tumor type of Medulloblastoma (1 if "yes" and 0 if "no"), allowed us to detect individual genes that exhibited a high relationship with Medulloblastoma. Based on this analysis, gene *PIAS1* was the most significantly correlated gene with Medulloblastoma. *PIAS1*, short for Protein Inhibitor of Activated *STAT 1* as part of the PIAS family, *PIAS1* is known to play a regulatory role in the JAK-STAT signaling pathway, a critical pathway in immune responses and cellular processes like proliferation and differentiation [32]. The role of *PIAS1* has been illustrated as a key player in the process of Epithelial-mesenchymal-transition (EMT), a process linked to tumor metastasis. Additionally, *PIAS1*, acting as a SUMO E3 ligase, is downregulated by TGF$beta$, a potent inducer of EMT. This suggests that *PIAS1* serves as a negative regulator of EMT and its downregulation may, therefore, facilitate EMT and possibly tumor progression[33]. Although we couldn't find concrete evidence on the role of PIAS1 in Medulloblastoma tumor types, this initial finding could be used as a baseline for further research.

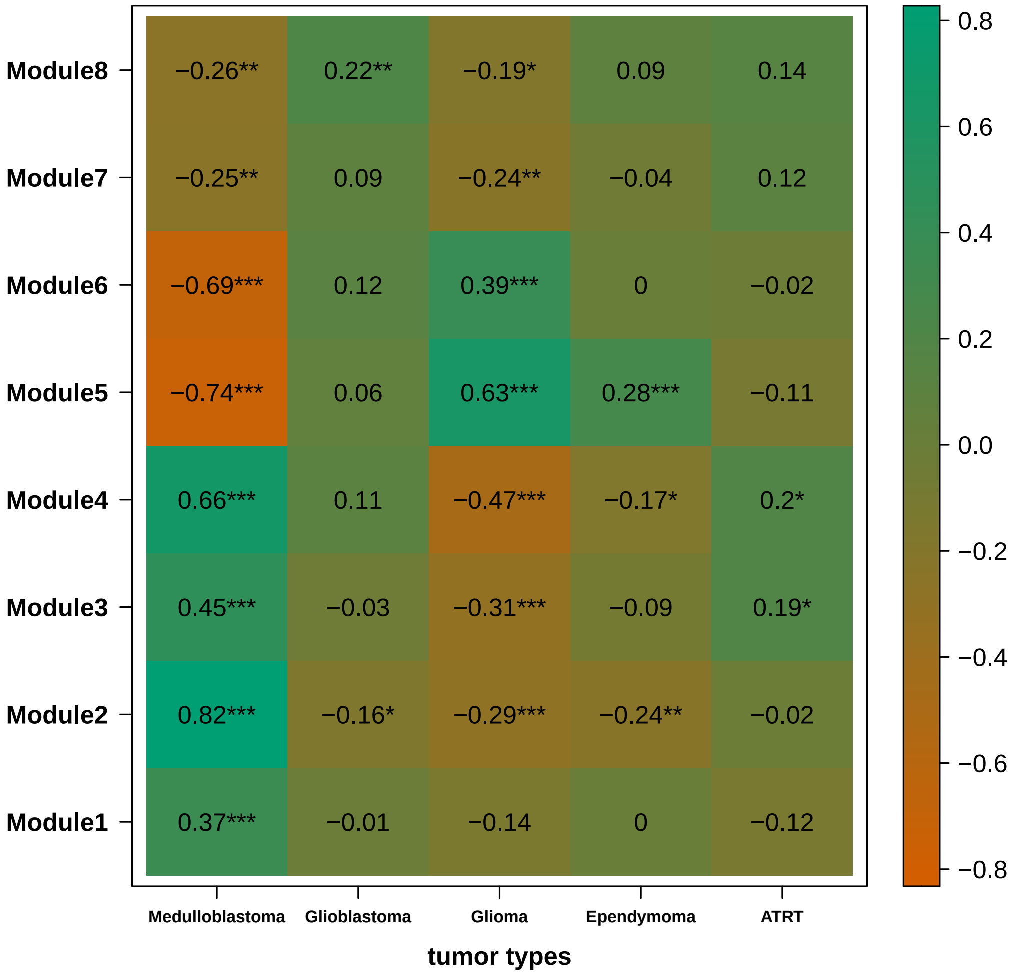


**Figure 4.4:** *The heatmap show the correlation between gene modules (Y-axis) and tumor types (X-axis). Each cell in the heatmap represents the correlation between a specific gene module and a tumor type, with the color indicating the strength and direction of the correlation. For instance, gene module 2 shows a strong positive correlation (0.82) with Medulloblastoma, suggesting that the expression of genes in this module is associated with the presence of this tumor type. Conversely, Module 5 has a strong negative correlation (-0.74) with Medulloblastoma, indicating that the genes in this module are less likely to be expressed in this type of tumor.*

# 5. Discussion

Our study was designed with two key objectives in mind. First, we aimed to identify the differences in immune-related gene expression across different pediatric brain tumors. Second, we sought to detect clusters of genes that behave similarly and their relationship with different tumor types.

## 5.1 Differential Expression Analysis

In response to our first objective, we employed differential expression analysis using DESeq2 R package[27]. We used Craniopharyngioma, a benign and low-grade tumor as a reference tumor. This was due to the absence of data from healthy tissue samples at the time of this research. In this analysis, We identified substantial disparities in the immune-related gene expression among the tumors types as compared to the reference tumor. Strikingly, Medulloblastoma exhibited the highest proportion of downregulated genes (64%). Medulloblastoma is an embryonal tumor of the posterior fossa and is the most common malignant brain tumor in children. It comprises up to 20% of all pediatric brain tumors[2]. On the other hand, Ependymoma, Glioma, and Glioblastoma displayed notable proportions of upregulated genes, suggesting a possible distinctive immune landscape within these tumors. To get a better understanding of the differential expression analysis, we have applied a commonly used threhsold of P-value less than 0.05, and absolute LogFoldChange greater than 2. This enabled us to rank top 5 upregulated and downregulated gene list. To our knowledge, there are no similar researches available to draw a comparison, however, most of the top upregulated genes are well studied in different contexts in previous research. For instance *MAGEA3*, is a tumor-specific shared antigen that is often expressed in lung cancer and melanoma. In addition, other tumor types express *MAGEA3* less frequently. Its expression is absent in normal tissues, with the exception of the testis and the placenta, but its level is linked to the severity of the illness as well as the patient's prognosis[34]. Moreover, in line with the findings of our study, *MAGEA3*, which was consistently highly expressed across all tumor types, it has been reported that, it is one of the most frequently expressed cancer-testis (CT) antigens in pediatric brain tumors, with expression noted in 56% of cases [35], [36]. This antigen expression can be leveraged for immunotherapy, considering the potential of CT antigens as targets for cancer treatment, one study suggested[37]. Importantly, previous studies have successfully used demethylating agents such as decitabine (DAC) to upregulate the expression of *MAGEA3*, thereby increasing the visibility of tumor cells to the immune system and enhancing their susceptibility to cytotoxic T lymphocytes [38]. *MAGEA3* was significantly upregulated when a protein called fibronectin was silenced in human thyroid carcinoma cells. This upregulation led to increased cell migration and invasion, contributing to the progression of human thyroid cancer [39]. These findings further highlights the potential relevance of *MAGEA3* in the behavior and progression of various tumor types.

In contrast one of the downregulated gene in our research, *COL17A1* also known as (Collagen XVII) is a transmembrane protein involved in maintaining the link between intra- and extracellular structural elements, essential for epidermal adhesion. *COL17A1* is encoded by the *COL17A1* gene and consists of three $\alpha 1$ chains. Previous research has connected chromosomal translocations, especially those resulting in oncogenic fusion proteins, with the initiation of human cancer [40]. It has been observed that this gene is usually upregulated in GBM, contrasting with our result where it is downregulated compared to Craniopharyngioma[41]. Additional studies found a new *PTEN-COL17A1* fusion gene that could increase *COL17A1* expression, leading to enhanced tumor invasiveness through upregulated matrix metalloproteinase (MMPs) expression [42]. Other researches indicated that *COL17A1* might serve as a prognostic biomarker and therapeutic target for GBM. In GBM samples, *COL17A1* gene expression has been identified in less than 1% of cases [42]. Moreover, *COL17A1* expression was found in malignant, but not benign, melanocytic tumors. There is a correlation between increased *COL17A1* expression and poor outcomes in colorectal carcinoma, suggesting a role in tumor progression [43], [44]. *MMP12*, Matrix Metallopeptidase 12, is another gene we identified as being significantly downregulated in Ependymoma and Glioma tumor types. Matrix Metallopeptidase 12 (*MMP12*), also known as Human Macrophage Metalloelastase (HME), is an enzyme that belongs to the matrix metalloproteinase family. These enzymes are known for their role in the degradation of extracellular matrix, contributing to physiological processes like embryogenesis and tissue remodeling, and pathological processes like inflammation and tumor invasion [45]. Although the significance of MMP12 in human brain tumors is not well established, one research demonstrated the role of Matrix Metalloproteinas(MMPs), including MMP12, in the progression of human brain tumors [46]. In another research it has been shown to play a role in the development of atherosclerosis and pathogenesis of aortic aneurysm [47], [48]. *MMP12* has also been detected in glioma cell line [49]. In a comprehensive study by Zarin et al. that examined *MMP2*, *MMP9*, and *MMP12* expression across sixty human brain tumors, the expression levels of *MMP12* was found to be higher in some tumors compared to other MMPs tested, however it has lower expression level among grade III tumor types when compared to the expression level of *MMP2*.

## 5.2 Gene Module Detection and Correlation

In addressing our second objective, the weighted gene co-expression network analysis (WGCNA), and WGCNA R package was employed[30]. The WGCNA allowed us to identify eight distinct modules, each encompassing a different set of genes whcih bahve similarly based on their expression. The high connectivity of certain genes within their respective modules suggests these genes could hold a crucial role in these co-expression networks. They might be the key players in the biological processes or pathways related to each tumor type. Interestingly, Modules 2 which includes genes such as *TRIM33, MAPK8, CCNT2, ATF2, TXNDC16*, and module 5 (*ARHGEF6, MASP1, PRKCA, FEZ1, CX3CR1*), and module 6 (*CD53, C3AR1, LAPTM5, ITGB2, PTPRC*) demonstrated a high correlation with Medulloblastoma. These modules might contain genes significantly contributing to the molecular characteristics of Medulloblastoma. For instance The finding of TRIM33 as a highly connected gene in Module2 aligns with the studies, such as that by Koso et al., who employed transposon-based insertional mutagenesis to model medulloblastoma in mice. In their investigation, TRIM33 was one of the Common Insertion Sites (CIS) genes in both their screen and in the medulloblastoma model developed in Ptch1 and Trp53 mutant backgrounds. This overlap underlines TRIM33's possible role as a driver gene in medulloblastoma pathogenesis[50]. Moreover, the gene PIAS1 from Module 2, in particular, demonstrated a strong positive correlation with Medulloblastoma. Given PIAS1's regulatory role in the JAK-STAT signaling pathway and the epithelial-mesenchymal-transition (EMT), understanding its role might offer new insights into the pathophysiology of Medulloblastoma and potentially uncover new therapeutic targets[33].

## 5.3 Limitations

Our results could have been impacted by the lack of comparison(refernce) group of healthy tissue in the gene expression analysis. While this method has enabled us on how the gene expression of various tumor types differs, it does not provide a comprehensive picture of how these gene expressions differ from those of normal, healthy tissue. This omission may have an impact on our findings because the identified genes may not be specifically dysregulated in the tumor types under investigation but rather may represent a general response to any pathological condition in the brain tissue. Future studies could overcome this by using comparisons from existing healthy tissue banks or reference gene expression databases.

# 6. Appedix A

## 6.1 Ethical and legal consideration of the data

The study was conducted with strict adherence to ethical and EU General Data Protection Regulation(GDPR) guidelines. Prior to the collection of clinical samples and RNA sequencing data, informed consent was obtained from parents or legal guardians after providing a detailed explanation of the study's purpose. To ensure privacy and confidentiality, all patient samples were pseudo-anonymized. This pseudonymization process involved replacing personal identifiers in the dataset with artificial identifiers. Thus, the connection between patients' identities and their corresponding data was protected while still permitting individual-level data analysis. Access to the data is granted only to authorized individuals or entities, and this access is controlled by the BioBank and Data Access Committee of the Princess Máxima Center, Utrecht, the Netherlands. This process of 'controlled access' ensures that the data is protected and only accessible to those with the necessary permissions.

Before granting access to the data, a data sharing agreement (DSA) is signed. This DSA, composed by the BioBank and Data Access Committee and the legal department of the Princess Máxima Center, aligns with the specific requirements of the Princess Máxima Center. The DSA ensures that the data is used responsibly and ethically and that the privacy and rights of the individuals whose data is included in the dataset are upheld. In the analysis phase, samples are recorded with their research identifiers, further ensuring the privacy of the individuals involved.

# 7. Appendix B

## 7.1 Running DESeq2 in R

DESeq2 package was implemented in*R version 4.2.2*. Craniopharyngioma, a benign and low-grade type of tumor was used as a baseline comparison tumor type during the differential expression analysis. This approach allowed us to identify key differences in gene expression profiles among these tumor types. The steps present the detailed implementation of DESeq2 in our research.

**Creation of DESeqDataSet:** from the matrix of count data, a DESeqDataSet object was created. The DESeqDataSet is an R object that encapsulates the count data along with meta-data for the samples and genes. In the process of creating this object, raw count data was inputted, alongside a table of sample information of the "meta data" that describes variables such as patient tumor type, gender and age.

**Design Formula:** the design formula used in the DESeq2 analysis describes the variables that will be used for normalization and the variables that will be tested for differential expression. In this study, the design formula was specified as

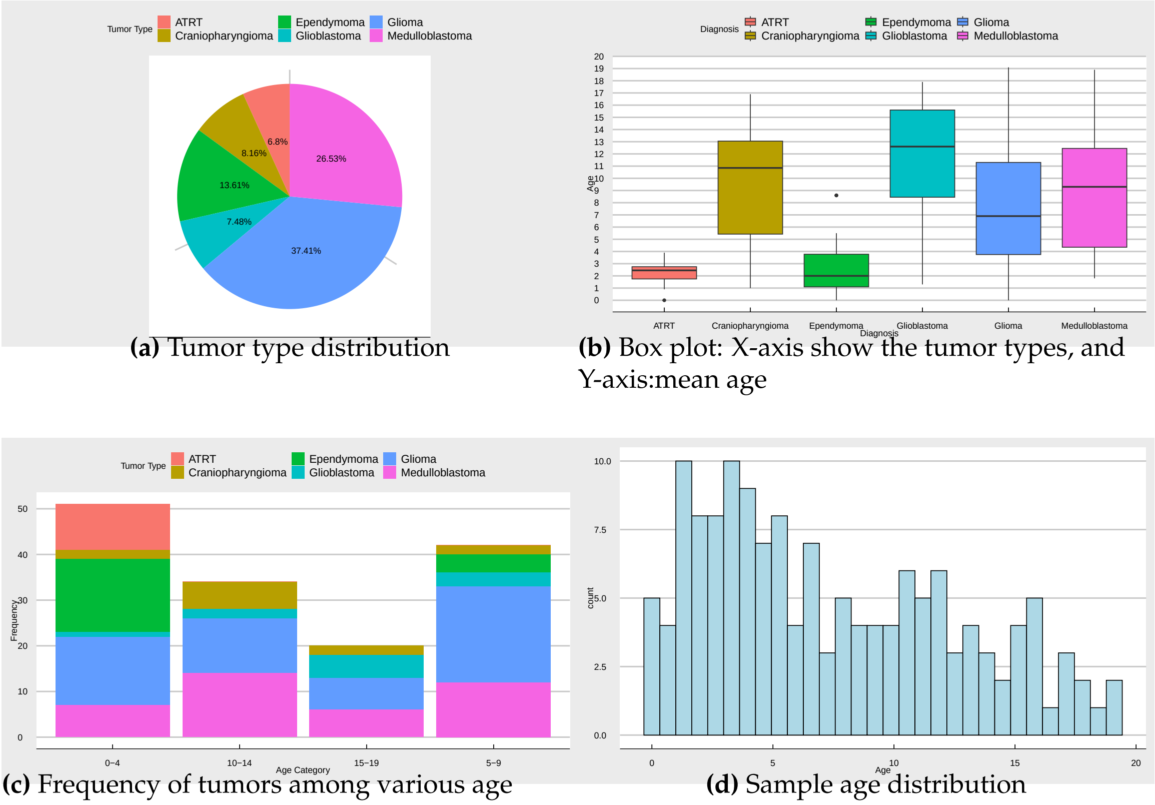
```
design ~ tumor_type
```

**Adjustment for Multiple Testing:** the p-values obtained from the DESeq2 analysis were adjusted for multiple testing using the Benjamini-Hochberg procedure. This procedure controls the false discovery rate (FDR), which is the expected proportion of false positives among all rejected hypotheses. By controlling the FDR, we limit the probability of making Type I errors when performing multiple comparisons. After running full gene expression analysis, we have filtred 1773 immune related genes for further downnstream analysis.

Genes are identified as 'significantly differentially expressed' if they met two stringent criteria: an adjusted P value of less than 0.05 and an absolute log2 fold change of at least 2. This dual-threshold approach ensures that only genes with both statistically significant changes ($P < 0.05$) and substantial expression differences (log2 fold change > 2) were selected for further downstream biological pathway analysis. This rigorous selection process enhances the reliability of the subsequent pathway analysis by focusing on the most biologically relevant genes.

# 8. Appendix C

## 8.1 Additional Differential Expression Anlysis Results



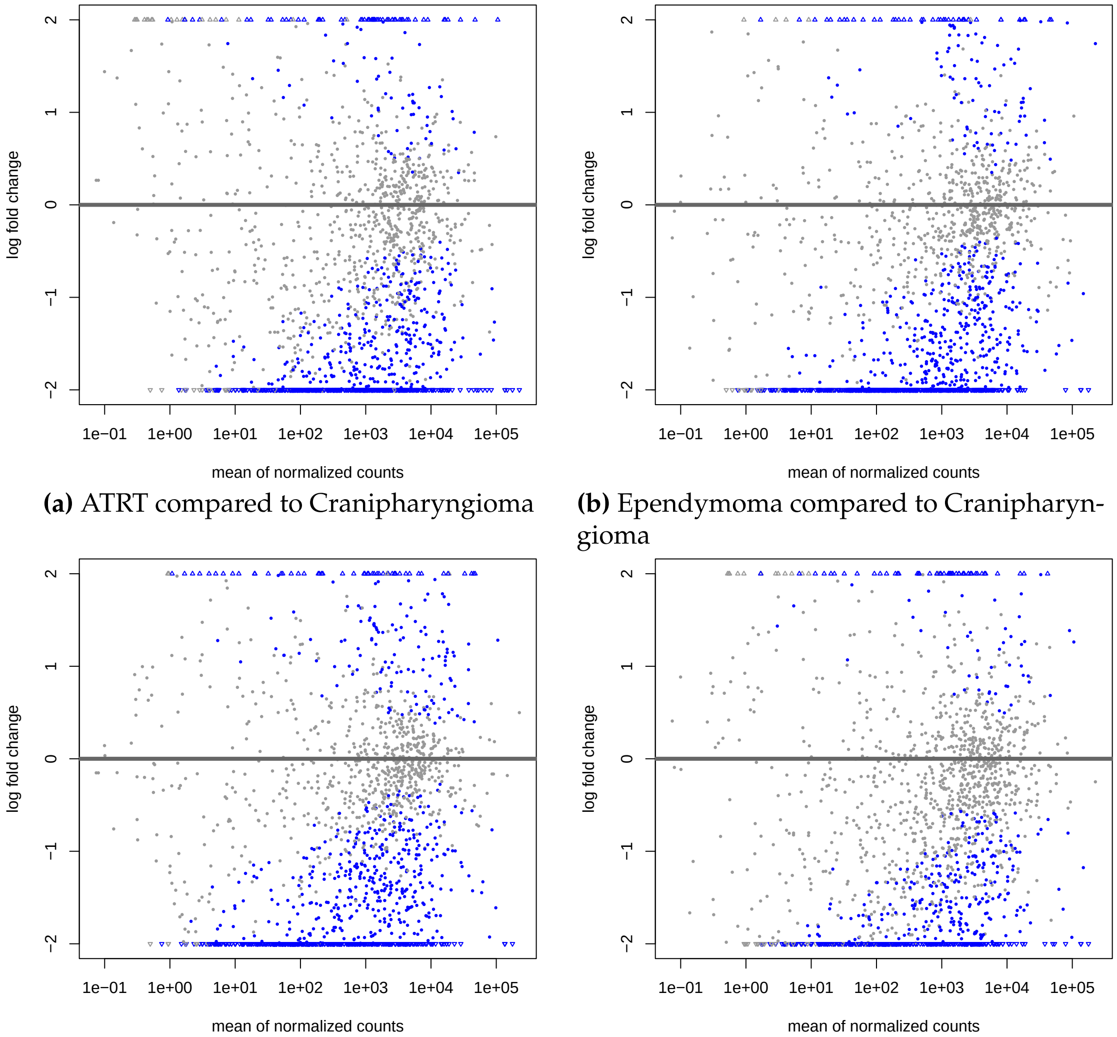
**(a)** Tumor type distribution

**(b)** Box plot: X-axis show the tumor types, and Y-axis:mean age

**(c)** Frequency of tumors among various age groups

**(d)** Sample age distribution

**Figure 8.1: Sample characteristics:** *ATRT exclusively occurs in the 0-4 age group. Medulloblastoma and Glioma are most prevalent in the 10-14 and 0-4 age groups respectively, while Glioblastoma peaks in the 15-19 age group.*

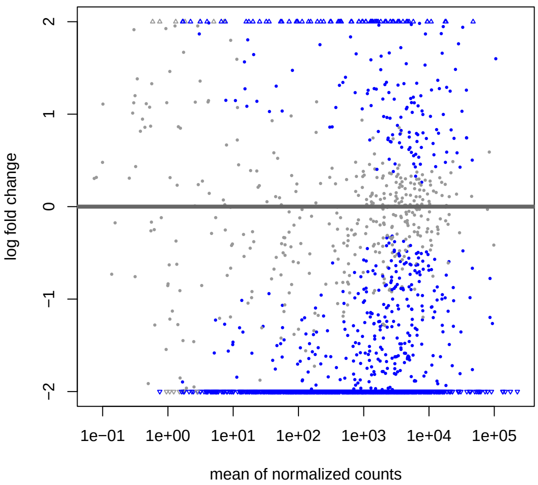


**(a)** ATRT compared to Cranipharyngioma

**(b)** Ependymoma compared to Cranipharyngioma

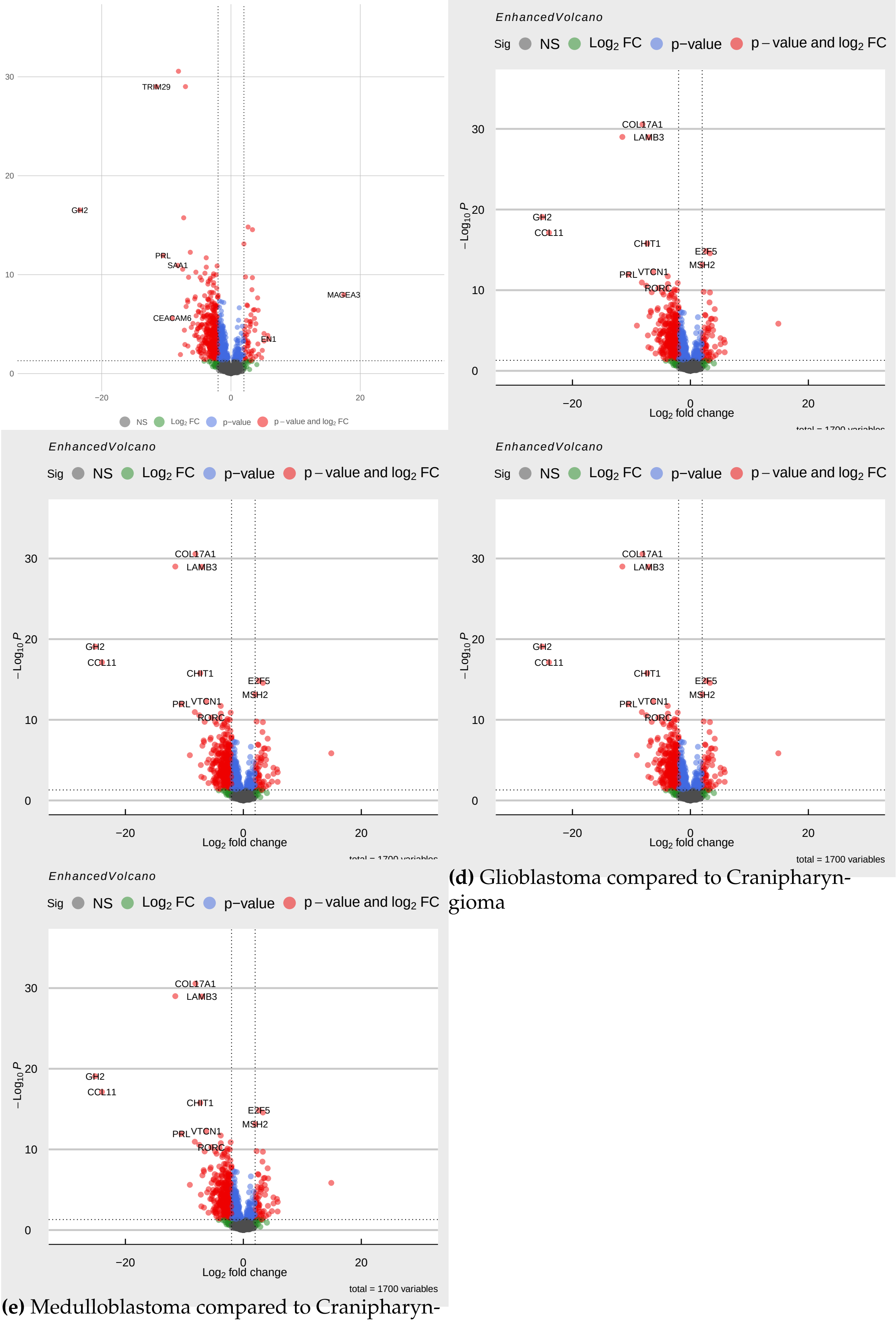
**(c)** Glioma compared to Cranipharyngioma

**(d)** Glioblastoma compared to Cranipharyngioma



**(e)** Medulloblastoma compared to Cranipharyngioma

**Figure 8.2:** *The x-axis represents the average normalized counts, a measure of gene expression level, while the y-axis represents the log2 fold change, indicating the degree of differential expression between the two tumor types. Each point on the plot corresponds to a gene. Genes that are significantly differentially expressed (beyond the absolute value of log2 fold change threshold). For instance, a gene with a log2 fold change of -2 is expressed at a level that is one-fourth (2ˆ(-2) = 0.25) of its expression level in Craniopharyngioma, indicating that it is four times downregulated in ATRT tumors compared to Craniopharyngioma.*



**(d)** Glioblastoma compared to Cranipharyngioma

**(e)** Medulloblastoma compared to Cranipharyngioma

**Figure 8.3:** *The x-axis represents the log2 fold change values, and the y-axis represents the -log10 transformed p-values. Each point on the plot corresponds to a gene. Genes with significant differential expression are highlighted and lie beyond the absolute log2 fold change threshold and above the -log10 p-value threshold. For instance, the gene LAMB3, which is indicated by a point at approximately -10 on the x-axis and 90 on the y-axis, is expressed at a level that is approximately 1/1024th (2ˆ-10) = 0.0009765625) of its expression level in Craniopharyngioma, indicating that it is*

## 8.2 Weighted Gene Co-expression Analyis

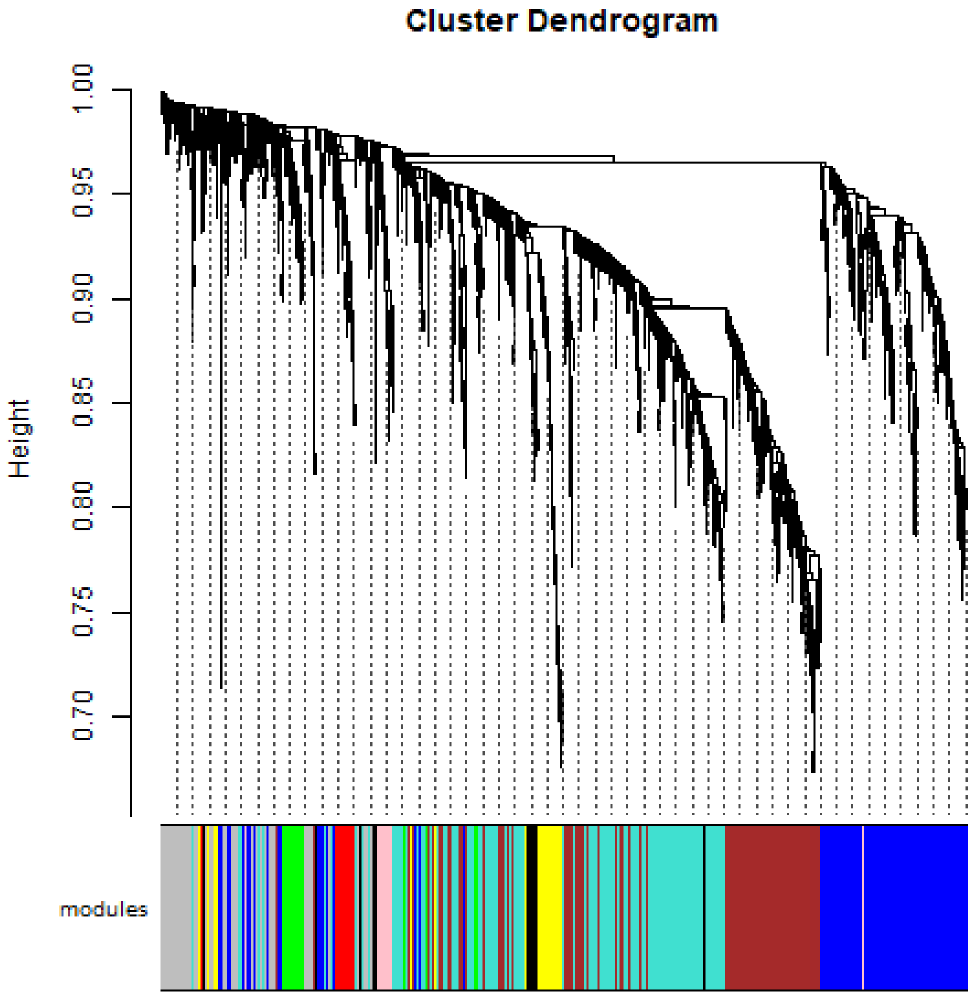
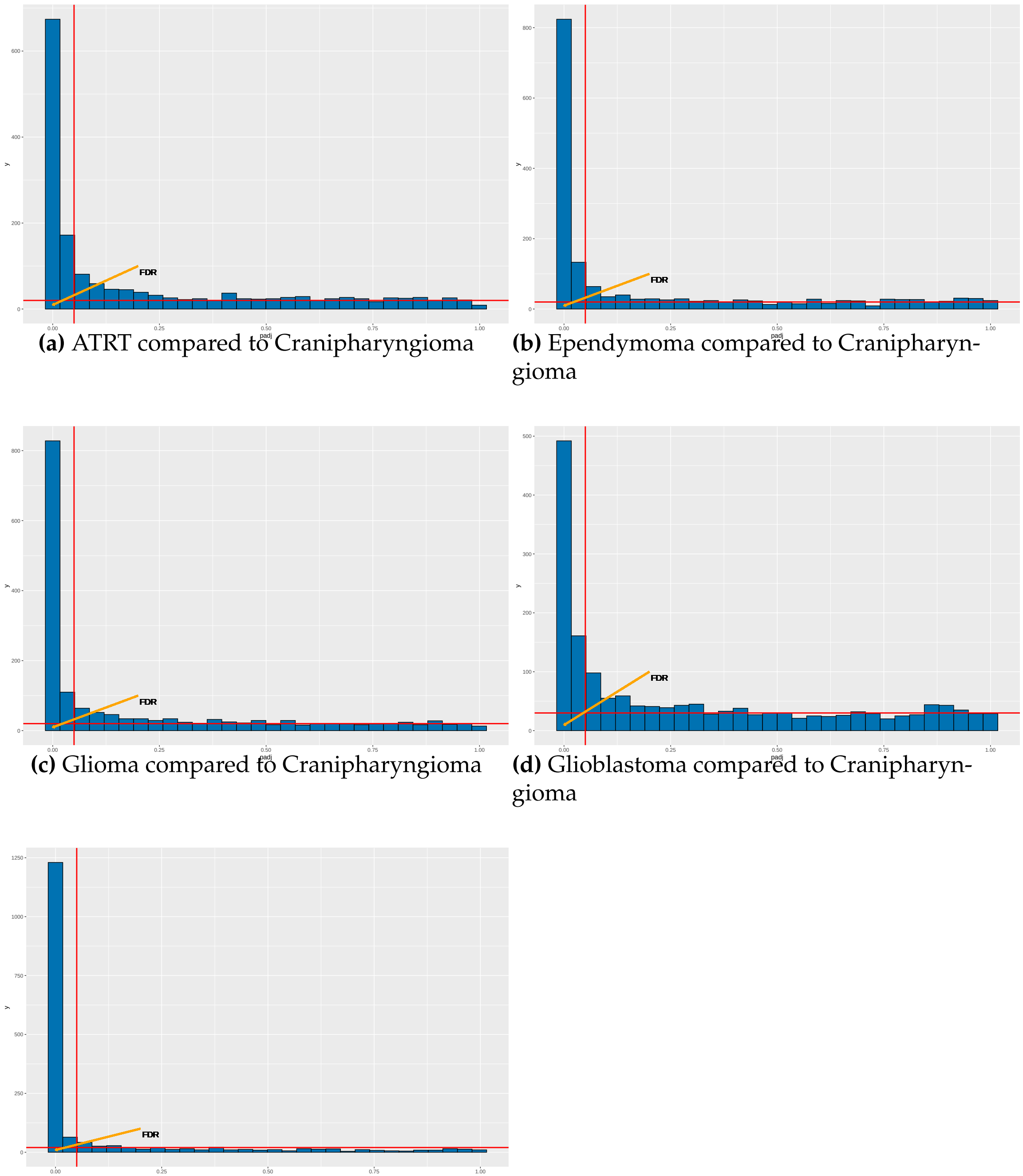


**Figure 8.5:** *The dendrogram represents the hierarchical clustering of genes, resulting in the identification of eight distinct modules. Below the dendrogram, each color box represents a module*



**(a)** ATRT compared to Cranipharyngioma

**(b)** Ependymoma compared to Cranipharyngioma

**(c)** Glioma compared to Cranipharyngioma

**(d)** Glioblastoma compared to Cranipharyngioma

**(e)** Medulloblastoma compared to Cranipharyngioma

**Figure 8.4: Pvalue and False Discovery Rate(FDR):** *The plot illustrates the relationship between the adjusted p-values (padj) and the number of genes associated with each padj value. The upper right corner represents genes correctly identified as differentially expressed. Conversely, the lower left corner indicated by the yellow arrow are genes that are false positives, i.e., they are incorrectly identified as differentially expressed.*

# 9. Appendix D

## 9.1 R code

The code used for the analysis in this paper is available at the following GitHub repository: https://github.com/AmdomH/DEA-Immune-Related-Genes

```

#Load library
library(tidyverse)
library(DESeq2)
library(readxl)
library(pheatmap)
library(psych) # for descriptive statistics
library(EnhancedVolcano) # for enhanced volcano plot
library(ggthemes)
if(!requireNamespace("WGCNA", quietly = TRUE)) install.packages("WGCNA")
if (!requireNamespace("BiocManager", quietly = TRUE)) install.packages("BiocManager")
if(!requireNamespace("GO.db", quietly = TRUE))  BiocManager::install("GO.db", update = TRUE, ask = FALSE)
library(WGCNA)
library(GEOquery)
library(ComplexHeatmap) # required for ComplexHeatmap
library(EnhancedVolcano)# for nicer volcano plots
library(UpSetR) #bar plots
library(gridExtra)
library(CorLevelPlot)

#Set working directory
setwd("C:/Users/aweldeab/surfdrive3/Documents/Data source files")
#Load count data and Meta data
# Load count_data from .Rdata file
load("count_data.Rdata")
load("meta.Rdata")
load("gene.Rdata")
load("immune_genes.Rdata")

###################################
# Chapter 01: Descriptive Analysis
####################################
ggplot(meta, aes(x=diagnosis, fill=gender)) +
  geom_bar(position="dodge") +
  labs(x="Diagnosis", y="Count", fill="Gender") +
  theme_solarized()

meta%>%
  group_by(gender)%>%
  summarise(perc= n()/nrow(meta) * 100)
```

```

#create table
table(meta$diagnosis, meta$age_cat)

#Box plot
ggplot(meta, aes(x=diagnosis, y=age, fill=diagnosis)) +
  geom_boxplot() +
  geom_jitter(width = 0.2, size = 1, color = "black")+
  ylab("Age in Years")+
  labs( x="Diagnosis", y="Age", fill="Diagnosis") +
  scale_y_continuous(breaks= c(0:20))+
  theme_economist_white()

################################################
## Chapter 02 Exploratory Analysis
###############################################
dds<- DESeqDataSetFromMatrix(countData = count_data,
                             colData = meta,
                             design = ~ diagnosis)
dds_normal<- estimateSizeFactors(dds)


sizeFactors(dds_normal)[1:5] #check top 5 samples
count_normal<- counts(dds_normal, normalized = TRUE) #Extrating Normalized counts
count_normal[1:5,1:5]


vsd <- vst(dds_normal, blind = TRUE)#scaling the data
vsd_mat<- assay(vsd)#Extract the vst matrix
vsd_mat[1:3, 1:3]

#subset the immune related genes
by <- join_by(gene_name==GeneName)
immune_genes<- left_join(immune_genes, gene,by)
immune_genes%>%head()

vsd_immune_mat <- vsd_mat[rownames(vsd_mat) %in% immune_genes$ID.x, ]
dim(vsd_immune_mat)
vsd_immune_cor<- cor(vsd_immune_mat) #compute correlation
vsd_immune_cor[1:5,1:5]

set.seed(123)
# Create the Heatmap object
p<-Heatmap(vsd_immune_cor,
             name = "cor",
             show_column_names = FALSE,  # hide column labels
             show_row_names = FALSE,  # hide row labels
             cluster_rows = TRUE,  # cluster rows
             cluster_columns = TRUE,  # cluster columns
             top_annotation = HeatmapAnnotation(diagnosis = meta$diagnosis),
             column_title_rot = 90
)

draw(p, heatmap_legend_side = "right")# Draw the heatmap
#Plot PCA
pcaData <- plotPCA(vsd, intgroup = "diagnosis", returnData = TRUE)
cbbPalette <- c("#CC79A7", "#56B4E9", "#009E73", "#F0E442", "#0072B2", "#
    D55E00")
pcaData%>%
  ggplot(aes(PC1, PC2, color=diagnosis))+
  geom_point(size = 2, alpha=1)+
  scale_color_manual(values =cbbPalette)+
  theme_economist_white()

################################################
# Chapter 03 Differential analysis
#################################################
dds <- DESeq(dds)# Perform differential expression analysis
norm_counts <- counts(dds, normalized = TRUE)
norm_counts[1:5, 1:5]

plotDispEsts(dds)
#Checking the distributin count data
ggplot(count_data) +
  geom_histogram(aes(x = log2(sample_63+1)), stat = "bin", bins = 30) +
  xlab("expression counts") +
  ylab("Number of genes") +
  theme_economist_white()

#Plot mean-variance
mean_counts <- apply(count_data[,1:147], 1, mean)
variance_counts <- apply(count_data[,1:147], 1, var)
df <- data.frame(mean_counts, variance_counts)
p<-ggplot(df) +
  geom_point(aes(x=log2(mean_counts), y=variance_counts)) +
  scale_y_log10(limits = c(1,1e9)) +
  scale_x_log10(limits = c(1,1e9)) +
  geom_abline(intercept = 0, slope = 1, color="red")+
  theme_economist_white()
p

#create contrasts
diagnosis <- c('Craniopharyngioma', 'ATRT', 'Ependymoma', 'Glioblastoma', '
    Glioma', 'Medulloblastoma')
contrasts <- lapply(setdiff(diagnosis, "Craniopharyngioma"), function(x) c("
    diagnosis", x, "Craniopharyngioma"))
contrasts
results.list <- lapply(contrasts, function(con) results(dds, contrast = con,
    alpha = 0.05))
```

```

#### ATRT vs Craniopharyngioma
ATRT_res<-results.list[[1]]
ATRT_res
summary(ATRT_res)
#### Ependymoma vs Craniopharyngioma
Ependymoma_res<- results.list[[2]]
Ependymoma_res
summary(Ependymoma_res)
#### Glioblastoma vs craniopharungioma
Glioblastoma_res<- results.list[[3]]
Glioblastoma_res
summary(Glioblastoma_res)
#### Glioma vs Craniopharyngioma
Glioma_res<- results.list[[4]]
Glioma_res
summary(Glioma_res)
#### Medulloblastoma vs Craniopharyngioma
Medulloblastoma_res<- results.list[[5]]
Medulloblastoma_res
summary(Medulloblastoma_res)

## Add gene names
#idx <- match( rownames(res), gene$ID )
ATRT_res$geneName<- gene$GeneName[match(rownames(ATRT_res), gene$ID )]
Ependymoma_res$geneName<- gene$GeneName[match(rownames(Ependymoma_res), gene$ID )]
Glioblastoma_res$geneName<- gene$GeneName[match(rownames(Glioblastoma_res), gene$ID )]
Glioma_res $geneName<- gene$GeneName[match(rownames(Glioma_res ), gene$ID )]
Medulloblastoma_res$geneName<- gene$GeneName[match(rownames(Medulloblastoma_res), gene$ID )]
head(ATRT_res)
#Filter immune related genes
idx <- which(ATRT_res$geneName %in% immune_genes$gene_name)#create index
ATRT_immune_res <- ATRT_res[idx,]
Ependymoma_immune_res<- Ependymoma_res[idx,]
Glioblastoma_immune_res<- Glioblastoma_res[idx,]
Glioma_immune_res<- Glioma_res[idx,]
Medulloblastoma_immune_res<- Medulloblastoma_res[idx,]
head(Medulloblastoma_immune_res)


## Summary results
summary(ATRT_immune_res );
head(ATRT_immune_res)
summary(Ependymoma_immune_res);
head(Ependymoma_immune_res)
summary(Glioblastoma_immune_res)
head(Glioblastoma_immune_res)
```

```r
summary(Glioma_immune_res);
head(Glioma_immune_res)
summary(Medulloblastoma_immune_res);
Medulloblastoma_immune_res

#################################################
# Chapter 04 Filtering DE genes
#################################################
# ATRT
top_upregulated_ATRT <- ATRT_immune_res %>%as.data.frame()%>%
  arrange(desc(log2FoldChange)) %>%
  head(5)
top_downregulated_ATRT <- ATRT_immune_res %>%as.data.frame()%>%
  arrange(log2FoldChange) %>%
  head(5)
# Ependymoma
top_upregulated_Ependymoma <- Ependymoma_immune_res %>%as.data.frame()%>%
  arrange(desc(log2FoldChange)) %>%
  head(5)
top_downregulated_Ependymoma <- Ependymoma_immune_res %>%as.data.frame()%>%
  arrange(log2FoldChange) %>%
  head(15)
# Glioblastoma
top_upregulated_Glioblastoma <- Glioblastoma_immune_res %>%as.data.frame()%>%
  arrange(desc(log2FoldChange)) %>%
  head(5)
top_downregulated_Glioblastoma <- Glioblastoma_immune_res %>%as.data.frame()%>%
  arrange(log2FoldChange) %>%
  head(5)
# Glioma
top_upregulated_Glioma <- Glioma_immune_res %>%as.data.frame()%>%
  arrange(desc(log2FoldChange)) %>%
  head(5)
top_upregulated_Glioma
top_downregulated_Glioma <- Glioma_immune_res %>%as.data.frame()%>%
  arrange(log2FoldChange) %>%
  head(5)
top_downregulated_Glioma
# Medulloblastoma
top_upregulated_Medulloblastoma <- Medulloblastoma_immune_res %>%as.data.frame()%>%
  arrange(desc(log2FoldChange)) %>%
  head(5)
top_downregulated_Medulloblastoma <- Medulloblastoma_immune_res %>%as.data.frame()%>%
  arrange(log2FoldChange) %>%
  head(5)
#################################################
# Chapter 05 Visualizations
```

```r
################################################
#Volcano plot
p1<- EnhancedVolcano(ATRT_immune_res,
                     lab = ATRT_immune_res$geneName,
                     x = 'log2FoldChange',
                     y = 'padj',
                     pCutoff = 0.05,
                     FCcutoff = 2,
                     cutoffLineType = 'dotted',
                     labSize = 3,
                     title=NULL,
                     xlim = c(-30, 30),
                     selectLab = c(top_upregulated_ATRT$geneName,top_downregulated_ATRT$geneName))
p1 + theme_economist_white()
#MA Plot
plotMA(ATRT_immune_res, ylim=c(-2,2))
#Adjusted P-values plots
ggplot(as.data.frame(ATRT_immune_res), aes(padj)) +
  geom_histogram(color = "black", fill = "#0072B2", bins = 30) +
  geom_vline(aes(xintercept = 0.05), color = "red",size = 1) +
  geom_hline(aes(yintercept = 20), color = "red",size = 1) +
  geom_segment(aes(x = 0.2, y = 100, xend = 0.001, yend = 10), arrow = arrow(length = unit(0.1, "cm")), size=1.5, color="orange") +
  geom_text(aes(x = 0.2, y = 100, label = "FDR"), hjust = -0.1, vjust = 1.5, size = 5, color = "black")
labs(x = "padj", y = "Frequency") +
  scale_x_continuous(breaks = seq(0, 1, by = 0.2)) +
  theme_minimal()

# Get DEGs for each result
# Order results based on padj
ATRT_immune_res_filter <- subset(ATRT_immune_res, padj < 0.05 & abs(log2FoldChange) > 2)
Ependymoma_immune_res_filter <- subset(Ependymoma_immune_res, padj < 0.05 & abs(log2FoldChange) > 2)
Glioblastoma_immune_res_filter <- subset(Glioblastoma_immune_res, padj < 0.05 & abs(log2FoldChange) > 2)
Glioma_immune_res_filter <- subset(Glioma_immune_res, padj < 0.05 & abs(log2FoldChange) > 2)
Medulloblastoma_immune_res_filter <- subset(Medulloblastoma_immune_res, padj < 0.05 & abs(log2FoldChange) > 2)
ATRT_immune_res_filter <- ATRT_immune_res_filter[order(ATRT_immune_res_filter$padj),]
Ependymoma_immune_res_filter <- Ependymoma_immune_res_filter[order(Ependymoma_immune_res_filter$padj),]
Glioblastoma_immune_res_filter <- Glioblastoma_immune_res_filter[order(Glioblastoma_immune_res_filter$padj),]
Glioma_immune_res_filter <- Glioma_immune_res_filter[order(Glioma_immune_res_filter$padj),]
```

```
Medulloblastoma_immune_res_filter <- Medulloblastoma_immune_res_filter[order(Medulloblastoma_immune_res_filter$padj),]
ATRT_immune_res_filter
ATRT_DEGs <- ATRT_immune_res_filter$geneName
Ependymoma_DEGs <- Ependymoma_immune_res_filter$geneName
Glioblastoma_DEGs <-Glioblastoma_immune_res_filter$geneName
Glioma_DEGs <- Glioma_immune_res_filter$geneName
Medulloblastoma_DEGs <- Medulloblastoma_immune_res_filter$geneName

# Create a list of DEGs
degs <- list(ATRT = ATRT_DEGs, Ependymoma = Ependymoma_DEGs, Glioblastoma = Glioblastoma_DEGs,
             Glioma = Glioma_DEGs, Medulloblastoma = Medulloblastoma_DEGs)
# Make the UpSet plot
upset(fromList(degs), order.by = "freq", set_size.show=FALSE, matrix.color = "blue", text.scale = 1,
      sets.bar.color =c("maroon", "blue", "green", "purple", "orange")
)

#################################################
# Chapter 06 WGCNA
################################################
count_data_filter <- count_data %>%
  rownames_to_column("ENSEMBL") %>%
  filter(ENSEMBL %in% rownames(ATRT_immune_res)) %>%
  column_to_rownames("ENSEMBL")
count_data_filter[1:3,1:3]

gsg <- goodSamplesGenes(t(count_data_filter)) #Explore good samples
summary(gsg)
gsg$allOK

table(gsg$goodGenes);table(gsg$goodSamples)

data <- count_data_filter[gsg$goodGenes == TRUE,]# remove genes that are detectd as outliers
#PCA
pca <- prcomp(t(data))
pca.dat <- pca$x
pca.var <- pca$sdev^2
pca.var.percent <- round(pca.var/sum(pca.var)*100, digits = 2)
#create this as data.frame
pca.dat <- as.data.frame(pca.dat)
ggplot(pca.dat, aes(PC1, PC2, color= )) +
  geom_point() +
  geom_text(label = rownames(pca.dat)) +
  labs(x = paste0('PC1: ', pca.var.percent[1], ' %'),
       y = paste0('PC2: ', pca.var.percent[2], ' %'))

samples.to.be.excluded <- c("sample_81")#Outliers
```

```
data.subset <- data[,!(colnames(data) %in% samples.to.be.excluded)]
colData <- meta %>%
  filter(!row.names(.) %in% samples.to.be.excluded)

# create dds
dds2 <- DESeqDataSetFromMatrix(countData = data.subset,
                               colData = colData,
                               design = ~ 1) # not specifying model
## remove all genes with counts < 15 in more than 75% of samples (31*0.75=23.25)
## suggested by WGCNA on RNAseq FAQ
dds75 <- dds2[rowSums(counts(dds2) >= 15) >= 24,]
nrow(dds75) #

# perform variance stabilization
dds_norm <- vst(dds75)

# get normalized counts
norm.counts <- assay(dds_norm) %>%
  t()


## 4 Network Construction: The main network
# Choose a set of soft-thresholding powers
power <- c(c(1:10), seq(from = 12, to = 50, by = 2))
# Call the network topology analysis function
sft <- pickSoftThreshold(norm.counts,
                         powerVector = power,
                         networkType = "signed",
                         verbose = 5)
#### Visualize
sft.data <- sft$fitIndices

# visualization to pick power
a1 <- ggplot(sft.data, aes(Power, SFT.R.sq, label = Power)) +
  geom_point() +
  geom_text(nudge_y = 0.1) +
  geom_hline(yintercept = 0.8, color = 'red') +
  labs(x = 'Power', y = 'Scale free topology model fit, signed R^2') +
  theme_classic()
a2 <- ggplot(sft.data, aes(Power, mean.k., label = Power)) +
  geom_point() +
  geom_text(nudge_y = 0.1) +
  labs(x = 'Power', y = 'Mean Connectivity') +
  theme_classic()
grid.arrange(a1, a2, nrow = 2)


##4A convert matrix to numeric
norm.counts[] <- sapply(norm.counts, as.numeric)
```

```
soft_power <- 4
temp_cor <- cor
cor <- WGCNA::cor
## Co-expression
# Calculate the co-expression (correlation) matrix
correlation_matrix <- cor(norm.counts)
# View the correlation matrix
correlation_matrix[1:5, 1:5]

### Adjacancy matrix
# Calculate the adjacency matrix based on the chosen power (beta)
A = adjacency(norm.counts, power = soft_power)

##4B memory estimate w.r.t blocksize
bwnet <- blockwiseModules(norm.counts,
                          maxBlockSize = 2000,
                          TOMType = "signed",
                          power = soft_power,
                          mergeCutHeight = 0.25,
                          numericLabels = FALSE,
                          randomSeed = 1234,
                          verbose = 3)
cor <- temp_cor

#Module Eigengenes
module_eigengenes <- bwnet$MEs
# get number of genes for each module
table(bwnet$colors)

#Plot dendogram and modules
plotDendroAndColors(bwnet$dendrograms[[1]], bwnet$colors,
                    "Modules",
                    dendroLabels = FALSE,
                    addGuide = TRUE,
                    hang= 0.03,
                    guideHang = 0.05)

#module trait associations
# create traits file - binarize categorical variables
trait<-colData %>%
  mutate(gender_bin = ifelse(grepl('female', gender), 1, 0)) %>%
  select(gender_bin)

# binarize categorical variables
colData$type <- factor(colData$diagnosis, levels = c("Craniopharyngioma", "
    Medulloblastoma",   "Glioblastoma","Glioma", "Ependymoma","ATRT" ))

type.out <- binarizeCategoricalColumns(colData$type,
                                       includePairwise = FALSE,
                                       includeLevelVsAll = TRUE,
                                       minCount = 1)
colnames(type.out)

traits <- cbind(trait, type.out)
# Define numbers of genes and samples
nSamples <- nrow(norm.counts)
nGenes <- ncol(norm.counts)
nSamples;nGenes

module.trait.corr <- cor(module_eigengenes, traits, use = 'p')
module.trait.corr

module.trait.corr.pvals <- corPvalueStudent(module.trait.corr, nSamples)
module.trait.corr.pvals

# visualize module-trait association as a heatmap
heatmap.data <- merge(module_eigengenes, traits, by = 'row.names')
head(heatmap.data)

heatmap.data <- heatmap.data %>% ### renames the columns for visisbility
  rename_all(~str_replace_all(., "data.", "")) %>%
  rename_all(~str_replace_all(., ".vs.all", "")) %>%
  rename_all(~str_replace_all(., "_bin", ""))


colnames <- names(heatmap.data)
# Rename the selected ones
colnames[1:8] <- c("Module1", "Module2", "Module3", "Module4", "Module5", "Module6", "Module7", "Module8")
# Assign the new names back to the dataframe
names(heatmap.data) <- colnames
cbbPalette <- c("#D55E00","#009E73")
par(cex.axis=0.8, cex.main=1.5)
CorLevelPlot(heatmap.data,
             x = names(heatmap.data)[11:15],
             y = names(heatmap.data)[1:8],
             cexLabX = .7,
             titleX = "tumor types",
             signifCutpoints = c(0, 0.001, 0.01, 0.05, 1),
             cexMain = 1,
             colFrame = "white",
             col = cbbPalette
)

#### Module mapping
module.gene.mapping <- as.data.frame(bwnet$colors)
module.gene.mapping%>%head()

##### ADD Gene Names
#Intramodular analysis: Identifying driver genes
```

```r
#Calculate the module membership and the associated p-values
#intramodular connectivity is calculated as the correlation of the eigengene and the gene expression profile.
module.membership.measure <- cor(module_eigengenes, norm.counts, use = ’p’)
module.membership.measure.pvals <- corPvalueStudent(module.membership.measure, nSamples)
module.membership.measure.pvals[1:9,1:10]

turquoise<-module.membership.measure.pvals["MEturquoise",]
hub_t<-as.data.frame(turquoise)%>%
  arrange(turquoise)%>%
  head()
ATRT_immune_res[rownames(hub_t),]

###Using the gene significance you can identify genes that have a high significance for trait of interest
#Using the module membership measures you can identify genes with high module membership with module
#who has high correlation with Meduloblastoma.

# Calculate the gene significance and associated p-values
gene.signf.corr <- cor(norm.counts, traits$data.Medulloblastoma.vs.all, use = ’p’)
gene.signf.corr.pvals <- corPvalueStudent(gene.signf.corr, nSamples)
gene.signf.corr.pvals %>%
  as.data.frame() %>%
  arrange(V1) %>%
  head(5)

top2_MB<- Medulloblastoma_immune_res[c("ENSG00000033800.13", "ENSG00000115738.10"), ]
top2_MB

#Identifying top 5 module hubs
hub_genes <- function(module_color) {
  ME <- module.membership.measure.pvals[paste0("ME", module_color), ]
  module_genes <- names(bwnet$colors)[bwnet$colors == module_color]
  module_pvals <- ME[module_genes]
  hub_genes <- sort(module_pvals, decreasing = FALSE)
  top_5 <- hub_genes[1:5]
  top_5_df <- data.frame(ID = names(top_5), pvalue = top_5)
  top_5_df <- dplyr::inner_join(top_5_df, gene, by = "ID")
  return(top_5_df)
}
colors<-unique(bwnet$colors)[1:8]
for (i in 1:length(colors)) {
  print(bwnet$colors[i])
  print(hub_genes(colors[i]))
}

```

```
### END ################
```

# Bibliography

[1] D. N. Louis, A. Perry, P. Wesseling, *et al.*, "The 2021 WHO Classification of Tumors of the Central Nervous System: A summary," *Neuro-Oncology*, vol. 23, no. 8, Jun. 2021. DOI: 10.1093/neuonc/noab106. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8328013/ (visited on 05/02/2023).

[2] Q. T. Ostrom, N. Patil, G. Cioffi, K. Waite, C. Kruchko, and J. S. Barnholtz-Sloan, "Cbtrus statistical report: Primary brain and other central nervous system tumors diagnosed in the united states in 2013-2017," *Neuro Oncol*, vol. 23, pp. ii1–ii105, Suppl 2 2021.

[3] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2018," *CA: a cancer journal for clinicians*, vol. 68, no. 1, pp. 7–30, 2018.

[4] C. Jones, M. A. Karajannis, D. T. W. Jones, *et al.*, "Pediatric high-grade glioma: Biologically and clinically in need of new thinking," *Neuro Oncol*, vol. 19, pp. 153–161, 2 2017.

[5] M. Pogorzala, E. Steliarova-Foucher, G. Gatta, R. S. Arora, and L. A. G. Ries, "Survival of children with central nervous system malignant tumours in europe," *Eur J Cancer*, vol. 49, pp. 2214–2224, 9 2013.

[6] J. A. Biegel, "Molecular genetics of atypical teratoid/rhabdoid tumors," *Neurosurgical focus*, vol. 20, no. 1, pp. 1–7, 2006.

[7] K.-W. Liu, K. W. Pajtler, B. C. Worst, S. M. Pfister, and R. J. Wechsler-Reya, "Molecular mechanisms and therapeutic targets in pediatric brain tumors," *Science signaling*, vol. 10, no. 470, eaaf7593, 2017.

[8] M. L. Garrè and A. Cama, "Craniopharyngioma: Modern concepts in pathogenesis and treatment," *Current opinion in pediatrics*, vol. 19, no. 4, pp. 471–479, 2007.

[9] R. J. Packer, "Chemotherapy: Low-grade gliomas of the hypothalamus and thalamus," *Pediatric neurosurgery*, vol. 32, no. 5, pp. 259–263, 2000.

[10] A. Louveau, I. Smirnov, T. J. Keyes, *et al.*, "Structural and functional features of central nervous system lymphatic vessels," *Nature*, vol. 523, no. 7560, pp. 337–341, 2015.

[11] L. Feldman, C. Brown, and B. Badie, "Chimeric antigen receptor (car) t cell therapy for glioblastoma," *NeuroMolecular Medicine*, pp. 1–6, 2022.

[12] D. S. Bitterman, S. M. MacDonald, T. I. Yock, *et al.*, "Revisiting the role of radiation therapy for pediatric low-grade glioma," *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*, vol. 37, no. 35, pp. 3335–3339, 2019.

[13] A. Mackay, A. Burford, D. Carvalho, and et al., "Integrated molecular meta-analysis of 1,000 pediatric high-grade and diffuse intrinsic pontine glioma," *Cancer Cell*, vol. 34, 520–537.e5, 4 2018.

[14] C. Jones, M. A. Karajannis, D. T. Jones, Kieran, *et al.*, "Pediatric high-grade glioma: Biologically and clinically in need of new thinking," *Neuro-oncology*, vol. 19, no. 2, pp. 153–161, 2017.

[15] S. L. Maude, T. W. Laetsch, Buechner, *et al.*, "Tisagenlecleucel in children and young adults with b-cell lymphoblastic leukemia," *New England Journal of Medicine*, vol. 378, no. 5, pp. 439–448, 2018.

[16] S. S. Wang, P. Bandopadhayay, and M. R. Jenkins, "Towards immunotherapy for pediatric brain tumors," *Trends in immunology*, vol. 40, no. 8, pp. 748–761, 2019.

[17] S. Sherif, J. Roelands, Mifsud, *et al.*, "The immune landscape of solid pediatric tumors," *Journal of Experimental & Clinical Cancer Research*, vol. 41, no. 1, p. 199, 2022.

[18] T. A. McEachron and L. J. Helman, "Recent advances in pediatric cancer research," *Cancer research*, vol. 81, no. 23, pp. 5783–5799, 2021.

[19] Z. Wang, M. Gerstein, and M. Snyder, "Rna-seq: A revolutionary tool for transcriptomics," *Nature reviews genetics*, vol. 10, no. 1, pp. 57–63, 2009.

[20] A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, McPherson, *et al.*, "A survey of best practices for rna-seq data analysis," *Genome biology*, vol. 17, no. 1, pp. 1–19, 2016.

[21] C. Trapnell, B. A. Williams, G. Pertea, *et al.*, "Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation," *Nature biotechnology*, vol. 28, no. 5, pp. 511–515, 2010.

[22] J. Chen, E. E. Bardes, B. J. Aronow, and A. G. Jegga, "Toppgene suite for gene list enrichment analysis and candidate gene prioritization," *Nucleic acids research*, vol. 37, no. suppl_2, W305–W311, 2009.

[23] A. P. Patel, I. Tirosh, Trombetta, *et al.*, "Single-cell rna-seq highlights intratumoral heterogeneity in primary glioblastoma," *Science*, vol. 344, no. 6190, pp. 1396–1401, 2014.

[24] J. Bathke and G. Lühken, "Ovarflow: A resource optimized gatk 4 based open source variant calling workflow," *BMC bioinformatics*, vol. 22, pp. 1–18, 2021.

[25] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[26] M.-H. Tran. "The basics of deseq2 – a powerful tool in differential expression analysis for single-cell rna-seq." (2022).

[27] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for rna-seq data with deseq2," *Genome biology*, vol. 15, no. 12, pp. 1–21, 2014.

[28] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "Edger: A bioconductor package for differential expression analysis of digital gene expression data," *bioinformatics*, vol. 26, no. 1, pp. 139–140, 2010.

[29] M. D. Luecken and F. J. Theis, "Current best practices in single-cell rna-seq analysis: A tutorial," *Molecular systems biology*, vol. 15, no. 6, e8746, 2019.

[30] P. Langfelder and S. Horvath, "Wgcna: An r package for weighted correlation network analysis," *BMC Bioinformatics*, vol. 9, no. 1, p. 559, 2008.

[31] P. Langfelder, B. Zhang, and S. Horvath, "Defining clusters from a hierarchical cluster tree: The dynamic tree cut package for r," *Bioinformatics*, vol. 24, no. 5, pp. 719–720, 2008.

[32] K. Shuai and B. Liu, "The jak-stat pathway," *Cell research*, vol. 10, no. 2, pp. 87–92, 2000.

[33] S. J. Netherton and S. Bonni, "Suppression of tgf$\beta$-induced epithelial-mesenchymal transition like phenotype by a pias1 regulated sumoylation pathway in nmumg epithelial cells," *PloS one*, vol. 5, no. 11, e13971, 2010.

[34] N. Peled, A. B. Oton, F. R. Hirsch, and P. Bunn, "Mage a3 antigen-specific cancer immunotherapeutic," 2009.

[35] J. F. Jacobs, O. M. Grauer, F. Brasseur, *et al.*, "Selective cancer-germline gene expression in pediatric brain tumors," *Journal of neuro-oncology*, vol. 88, pp. 273–280, 2008.

[36] S. M. Oba-Shinjo, O. L. Caballero, A. A. Jungbluth, *et al.*, "Cancer-testis (ct) antigen expression in medulloblastoma," *Cancer immunity*, vol. 8, no. 1, 2008.

[37] D. K. Krishnadas, F. Bai, and K. G. Lucas, "Targeting cancer-testis antigens in recurrent pediatric brain tumors," *Journal of Neuro-Oncology*, vol. 123, pp. 193–195, 2015.

[38] L. Bao, K. Dunham, and K. Lucas, "Mage-a1, mage-a3, and ny-eso-1 can be upregulated on neuroblastoma cells to facilitate cytotoxic t lymphocyte-mediated tumor cell killing," *Cancer Immunology, Immunotherapy*, vol. 60, pp. 1299–1307, 2011.

[39] W. Liu, S. Cheng, S. L. Asa, and S. Ezzat, "The melanoma-associated antigen a3 mediates fibronectin-controlled cancer progression and metastasis," *Cancer research*, vol. 68, no. 19, pp. 8104–8112, 2008.

[40] A. V. Chernov, S. Baranovskaya, V. S. Golubkov, *et al.*, "Microarray-based transcriptional and epigenetic profiling of matrix metalloproteinases, collagens, and related genes in cancer," *Journal of Biological Chemistry*, vol. 285, no. 25, pp. 19 647–19 659, 2010.

[41] V. Senner, S. Ratzinger, S. Mertsch, S. Grässel, and W. Paulus, "Collagen xvi expression is upregulated in glioblastomas and promotes tumor cell adhesion," *FEBS letters*, vol. 582, no. 23-24, pp. 3293–3300, 2008.

[42] X. Yan, C. Zhang, T. Liang, *et al.*, "A pten-col17a1 fusion gene and its novel regulatory role in collagen xvii expression and gbm malignance," *Oncotarget*, vol. 8, no. 49, p. 85 794, 2017.

[43] T. Krenács, Kiszner, *et al.*, "Collagen xvii is expressed in malignant but not in benign melanocytic tumors and it can mediate antibody induced melanoma apoptosis," *Histochemistry and cell biology*, vol. 138, pp. 653–667, 2012.

[44] J. M. Moilanen, N. Kokkonen, S. Löffek, J. P. Väyrynen, *et al.*, "Collagen xvii expression correlates with the invasion and metastasis of colorectal cancer," *Human pathology*, vol. 46, no. 3, pp. 434–442, 2015.

[45] H. Nagase, R. Visse, and G. Murphy, "Structure and function of matrix metalloproteinases and timps," *Cardiovascular research*, vol. 69, no. 3, pp. 562–573, 2006.

[46] C. Colton, J. Keri, W.-T. Chen, and W. Monsky, "Protease production by cultured microglia: Substrate gel analysis and immobilized matrix degradation," *Journal of neuroscience research*, vol. 35, no. 3, pp. 297–304, 1993.

[47] S.-i. Matsumoto, T. Kobayashi, M. Katoh, *et al.*, "Expression and localization of matrix metalloproteinase-12 in the aorta of cholesterol-fed rabbits: Relationship to lesion development," *The American journal of pathology*, vol. 153, no. 1, pp. 109–119, 1998.

[48] J. A. Curci, S. Liao, M. D. Huffman, S. D. Shapiro, R. W. Thompson, *et al.*, "Expression and localization of macrophage elastase (matrix metalloproteinase-12) in abdominal aortic aneurysms.," *The Journal of clinical investigation*, vol. 102, no. 11, pp. 1900–1910, 1998.

[49] S. Wagner, C. Stegen, H. Bouterfa, *et al.*, "Expression of matrix metalloproteinases in human glioma cell lines in the presence of il-10," *Journal of neuro-oncology*, vol. 40, pp. 113–122, 1998.

[50] H. Koso, A. Tsuhako, E. Lyons, *et al.*, "Identification of foxr2 as an oncogene in medulloblastoma," *Cancer research*, vol. 74, no. 8, pp. 2351–2361, 2014.