Utrecht University
Department of Information and Computing Science

Applied Data Science master thesis

Unraveling Pediatric Brain Tumor Transcriptomics and Survival: A Comparative Exploration of Diverse High-Grade Tumors

Abstract

This study presents an exploration of the transcriptomic profiles of pediatric high-grade brain tumors and their correlation with survival outcomes. By leveraging RNA-sequencing data, we employed techniques including Principal Component Analysis (PCA) and differential gene expression analysis using DESeq2 to identify distinct molecular signatures and differentially expressed genes across various tumor types. Notably, our study shed light on the consistent upregulation of Melanoma Antigen Gene A3 (MAGE-A3) across different tumors, suggesting its potential role in tumorigenesis. Subsequent survival analysis, employing the Kaplan-Meier method and the log-rank test, revealed a statistically significant association between high MAGE-A3 expression and lower survival rates. This underscores the prognostic potential of MAGE-A3 in pediatric high-grade brain tumors. However, another gene of interest, COL17A1, despite being significantly downregulated, did not demonstrate a significant correlation with survival outcomes, exemplifying the complex relationships in the tumor's molecular profile. Our study contributes insights into the molecular heterogeneity of pediatric high-grade brain tumors, thereby laying the groundwork for future research to further unravel the complex genetic interactions and their implications on patient survival.

First examiner:
Dr. A. Kaznatcheev

Second examiner:
D.S. Islakoglu MSc

Candidate:
C.J. Bogaard BSc

In cooperation with:
A.H. Weldeabezgi MSc
Dr. T. Carvalheiro

July 10th, 2023

1. Introduction

1.1. Pediatric High-Grade Brain Tumors: Prevalence, Challenges, and Treatment Limitations

High-grade brain tumors are characterized by aggressive growth and often poor prognosis, representing some of the most malignant forms of brain cancers seen in children. They are hard to treat and call for fresh research to boost treatment outcomes and the quality of life for children diagnosed with these diseases. This study focuses on deciphering the immune transcriptomic profile of various high-grade brain tumors in children. Immune transcriptomic profiling is a method that uses high-throughput sequencing technologies to identify and quantify the mRNA (transcripts) within a cell or tissue sample (Conesa et al., 2016). This allows researchers to measure the activity of the immune system at the molecular level, providing a detailed look at how the immune system responds to various conditions, including disease states like cancer (Alizadeh et al., 2000; Bindea et al., 2013). Specifically, in this study, immune transcriptomic profiling is used to understand the immune response in pediatric high-grade brain tumors. High-grade brain tumors make up around 15-20% of all brain tumors in children, with the occurrence rates ranging between 0.5 and 4 cases per 100,000 kids (Ostrom et al., 2015). The 5-year survival rates for children diagnosed with high-grade brain tumors are still quite low. They range between 30–60%, which is a stark contrast to the over 90% survival rates with leukemia, another more common childhood cancer (Ward et al., 2014; Pui et al., 2015).

In recent years, the field of immunotherapy has seen substantial advancements, most notably the development of Chimeric Antigen Receptor T-cell (CAR-T) therapy. This innovative treatment modifies a patient's T cells to specifically target and destroy cancer cells, a breakthrough that has significantly enhanced outcomes for certain leukemia types. The success of CAR-T therapy highlights the power and potential of immunotherapies, and underscores the importance of understanding the immune transcriptomic profile, particularly in the context of pediatric high-grade brain tumors (Maude et al., 2018).

Factors contributing to the challenges in treating pediatric high-grade brain tumors include their relative rarity, heterogeneity, the presence of the blood-brain barrier, and the potential for severe long-term side effects from current treatment modalities (Jones & Baker, 2014; Daneman & Prat, 2015; Merchant et al., 2010).

Research into children is especially relevant, because treating pediatric high-grade brain tumors is more challenging than treating adult patients due to several factors. Firstly, pediatric tumors often exhibit distinct molecular and genetic profiles, which can affect tumor behavior, growth patterns, and response to treatments, making it difficult to extrapolate adult treatment approaches to children (Jones et al., 2012; Gupta et al., 2018). Secondly, the pediatric central nervous system is still developing, making it more vulnerable to the adverse effects of aggressive treatments, such as radiation and chemotherapy, potentially leading to long-term cognitive, motor, and endocrine impairments (Mulhern et al., 2004; Ris & Noll, 1994). By examining a range of tumors this research aims to elucidate some of the complex interplay between tumor and immune cells. Consult *Table 1* for further information on the different tumor types, incidences and prognosis.

Table 1

Tumor types, incidences, and prognosis (CNS - Central Nervous System)

Tumor Type	Incidence	Prognosis	Description	Source
AtypicalTeratoid / Rhabdoid Tumor	1-2% of pediatric CNS	15-30% 5-year survival	ATRT is a rare and fast-growing cancerous tumor of the brain and spinal cord. It commonly appears in children aged three years and younger, although it can occur in older children and adults.	Fruhwald et al. (2016)
Craniopharyngioma	Rare	Variable	Craniopharyngioma is a rare type of brain tumor derived from pituitary gland embryonic tissue. It can cause hormonal imbalances and pressure effects.	Müller (2014)
Ependymoma	8-12% of pediatric CNS	50-70% 5-year survival	Ependymomas are tumors that arise from ependymal cells lining the ventricles and central canal within the spinal cord. These tumors can occur anywhere in the central nervous system.	Pajtler et al. (2015)
Glioblastoma	3% of pediatric CNS	<20% 2-year survival	Glioblastoma is an aggressive type of cancer that can occur in the brain or spinal cord. It forms from cells called astrocytes that support nerve cells. It can occur at any age but tends to occur more often in older adults.	(Nikitovic et al., 2016); Stupp et al. (2005)
Glioma	20-30% of pediatric CNS	20-35% 5-year survival	Gliomas are a type of tumor that starts in the glial cells of the brain or the spine. They range from slow-growing (low grade) to fast-growing (high grade), with varying prognosis.	Fangusaro (2012)
Medulloblastoma	20% of pediatric CNS	60-70% 5-year survival	Medulloblastoma is a cancerous tumor—most often brain cancer. It is the most common malignant tumor of the cerebellum in children.	Ramaswamy et al. (2016)

1.2. The Tumor Microenvironment and Its Role in Immunosuppression and Immune Response

Immunotherapies are a type of treatment that use the body's immune system to find and destroy cancer cells. They have shown great potential in treating different cancers, like skin cancer (melanoma) and a type of lung cancer (non-small cell lung cancer) (Ribas & Wolchok, 2018; Reck et al., 2016), and childhood leukemia (Ward et al., 2014; Pui et al., 2015) .

The study of how the immune system responds to cancer is very important because the environment where the tumor grows plays a big role in how the cancer develops. This environment, known as the tumor microenvironment (TME), shapes how cancer cells, immune cells, and other cells interact. This complex relationship can affect how the tumor grows, how it responds to treatments, and how it can sometimes evade the immune system (Quail & Joyce, 2013). The tumor microenvironment (TME) also contains supportive cells such as cancer-associated fibroblasts and tumor-associated macrophages, which can aid in tumor growth.

Within the Tumor Microenvironment (TME), various immune cells have differing roles, playing both supportive and defensive parts in the body's response to a tumor. Some immune cells, such as a type called lymphocytes, have the capacity to recognize and eliminate cancer cells, aiding the body's fight against the tumor (Dong et al., 2017). Other types of immune cells may inadvertently assist the tumor. These cells can suppress the body's immune response, providing an opportunity for the tumor to grow and evade the immune system (Weber et al., 2017). A third group of immune cells can have a dual role depending on their state. They can either support or harm the tumor, adding another layer of complexity to the immune response in the TME (Mantovani et al., 2017). By better understanding the interplay between these immune cells and the tumor, it is possible to develop more effective strategies for cancer treatment.

When it comes to high-grade brain tumors in children, different immune responses can manifest compared to adults, making pediatric-specific research necessary for the development of effective treatments (Simon et al., 2015). High-grade brain tumors are known to suppress the immune system, making it harder for immune-based therapies to work (Jackson et al., 2019). On top of that, the blood-brain barrier (BBB), a natural defense mechanism of the brain, can prevent many treatments from reaching the brain (Daneman & Prat, 2015). By studying the immune response to these types of tumors, researchers might find new ways to overcome these challenges and improve the quality of life and outcomes for these young patients.

1.3. Decoding Immunological Signatures: The Potential for Personalized Treatments

---

Immunosignatures describe unique patterns of immune response in the context of specific diseases or conditions. An immunosignature, in simple terms, is like a unique "fingerprint" that the immune system leaves when it responds to a disease (Restifo et al., 2012). Broadly speaking, immunosignatures encompass two main components, *Immunophenotypes* and *immunomodulatory signatures*:

- *Immunophenotypes* pertain to the characterization of immune cells by identifying specific surface markers (proteins found on the cell surface). This gives us an idea of the type and state of the immune cells present in the tumor environment. It is like taking a census of the immune cells within the tumor, helping us understand what types of cells are there and in what proportion (Maecker et al., 2012).
- *Immunomodulatory signatures*, on the other hand, refer to the expression of immune-modulating factors by the tumor cells themselves. These factors, which could be proteins like cytokines or chemokines, can change the behavior of the immune cells, often leading to an environment that suppresses the immune response and allows the tumor to grow (Vinay et al., 2015). It is akin to the tumor cells 'sending signals' to trick the immune system into helping them survive and proliferate.

Immunosignatures play an integral role in shaping our understanding of disease progression and potential treatments. With this in mind, our study seeks to address two key research questions:

- **Question 1:** "Which genes are differentially expressed in pediatric high-grade brain tumors and what unique immunosignatures can we identify from these expression patterns?"
- **Question 2:** "Do these identified immunosignatures significantly influence survival outcomes in patients diagnosed with these pediatric high-grade brain tumors?"

These research questions guide our objective to not only characterize the immune transcriptomic profile of pediatric high-grade brain tumors but also to explore their association with patient survival rates. Understanding the survival outcomes associated with specific immunosignatures can provide insights into the prognosis of these aggressive cancers. While the focus on immune transcriptomic profiles can enhance our understanding of the tumor microenvironment and its role in cancer progression, linking these findings to survival outcomes brings a clinical relevance to our study.

2. Data

The utilization of patient data in this study was carried out with care and respect for patient privacy. All data were pseudonymized and analyzed collectively to ensure confidentiality. For further details, consult *Appendix A.*

2.1. Count Data: RNA-seq

RNA-seq, short for RNA sequencing, is a technology that allows for the comprehensive examination of all the RNA in a sample. The process typically starts with the conversion of RNA to cDNA, which is then broken down into short fragments (reads) that are sequenced. The sequence data is then analyzed to provide a snapshot of the cellular activity at the time the sample was collected, offering insights into the transcriptome—the complete set of RNA transcripts produced by the genome at a specific moment (Wang, Gerstein & Snyder, 2009).

The procedure to analyze RNA-seq data typically involves several key steps:

1. **RNA-seq short reads:** RNA is first converted into a library of cDNA fragments through either RNA fragmentation or cDNA synthesis. These fragments are then sequenced to produce millions of short sequence reads. The quality of these reads is paramount as the downstream analysis heavily relies on the accuracy of this data.
2. **Quality control:** This step assesses the quality of the raw sequence data to ensure its reliability for further analysis. Tools such as FastQC provide a fast and efficient way to analyze whether the raw sequence data has any problems, such as low-quality sequences, contamination, or overrepresented sequences. Another tool, Picard, provides a set of Java command-line utilities for manipulating high-throughput sequencing data and formats, providing deeper metrics for the quality control step (Andrews, 2010; Broad Institute, 2019).
3. **Mapping to reference genome:** After quality control, the sequenced reads are aligned or mapped to a reference genome. The reference genome used is GRCh38.
4. **Generating count data:** Following the mapping of reads to the reference genome, the subsequent step involves quantifying gene expression. This is typically done by counting the number of reads mapped to each gene, a process called read counting (Anders et al., 2015).

The final result is a matrix of gene expression data, with rows corresponding to genes and columns corresponding to individual samples. This data can then be used for a variety of downstream analyses, including differential gene expression analysis, clustering, and pathway analysis. As shown in *table 2*.

Table 2

Count Data

Gene	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
ENSG00000223972.5	139	54	54	43	70
ENSG00000227232.5	1081	1488	1377	778	1030
ENSG00000278267.1	13	7	7	8	29
……	…	…	…	…	…

2.2. Selected Data Exploration Results

2.2.1. Metadata

Alongside the clinical data, we also possess a metadata set. This metadata comprises valuable supplementary information that aids in our analysis and interpretation of the main data. This information can include aspects such as demographic data, relevant medical history, tumor grading, and any other data elements that provide context to our main dataset. A preview of this metadata is available in *Table 3*.

Table 3

Metadata

Sample	Diagnosis	Gender	Age
Sample 1	Medulloblastoma	female	12.5
Sample 2	Craniopharyngioma	male	14.6
Sample 3	Glioblastoma	female	7.9
……	…	…	…

In our initial analysis, we started with a dataset comprising 182 samples, representing various types of pediatric high-grade brain tumors. However, our preliminary data cleaning process identified 33 longitudinal data points with duplicate patient identifiers, which were subsequently removed to avoid potential biases in our analysis. Our initial dataset encompassed eight pediatric high-grade brain tumors. However, for the purpose of our analysis, we chose to use simplified versions or collective names for these tumor types. The reason behind this decision was to enhance the readability and accessibility of our research, while ensuring our analysis could remain robust and significant despite the inherent variation and complexity in tumor subtypes. By consolidating subtypes under broader categories, we were able to maintain a manageable number of categories for comparison while still preserving the essential distinctions that characterize these various forms of pediatric high-grade brain tumors. This simplified approach will facilitate a more straightforward interpretation of our findings without compromising the depth and accuracy of our analysis. Additionally, we decided to exclude Astrocytoma, a cancer type represented by only a single sample, due to the imbalance this could introduce into our analysis. Following these data cleaning steps, we were left with a dataset of 147 samples, which consisted of samples from 65 female and 82 male patients, providing a gender-balanced perspective in our ensuing analyses.

Table 4 provides an overview of the distribution of tumor types within our dataset, specifically detailing the number of samples, the age range of patients, and the average age at diagnosis for each tumor type. A closer look at underscores the unique nature of each tumor type and the patient population it affects. For instance, it is evident that ATRT predominantly affects the youngest children in our study, while Glioblastoma tends to occur in older children. To further visualize the age distribution across different tumor types, we created two figures, which are available in *Appendix C*.

Table 4

Overview: distribution of tumor types

Tumor Type	Number of samples	Age range (years)	Average age at diagnosis
ATRT	10	0-4	2.2
Craniopharyngioma	12	1-17	9.6
Ependymoma	20	0 - 9	2.8
Glioblastoma	11	1 - 18	11.4
Glioma	55	0 -19	7.6
Medulloblastoma	39	2 - 19	8.9

2.2.2. Survival Data

This data provides us with information regarding the patient's outcome, allowing us to understand the lethality of each tumor type and the effectiveness of treatments *(table 5)*.

Table 5

Survival Data

Sample	Age at Diagnosis	Date Of Death
Sample 1	8	
Sample 2	0	2022-02-23
Sample 3	2	
……	…	…

With our survival data, we performed survival analysis, comparing the prognosis derived from the literature against our actual observed survival rates. This is visible in *Table 6.*

Table 6

Survival: Literature vs. our data

Tumor Type	Diagnosed	Survivors	Prognosis (literature)	Actual Survival Rate
ATRT	10	3	15-30% 5-year survival	30% 5-year survival
Craniopharyngioma	12	12	Variable	100%
Ependymoma	20	17	50-70% 5-year survival	85% 5-year survival
Glioblastoma	11	1	<20% 2-year survival	90.90% 2-year survival
Glioma	55	28	20-35% 5-year survival	83.63% 5-year survival
Medulloblastoma	39	32	60-70% 5-year survival	94.87% 5-year survival

Our analysis showcased notable discrepancies between observed survival rates and those projected in the literature for various tumor types. For Atypical Teratoid/Rhabdoid Tumor (ATRT), a survival rate of 30% at the 5-year mark is in line with the 15-30% literature prognosis, indicating consistency with established expectations for this tumor type. The Craniopharyngioma patients surprisingly demonstrated a 100% survival rate, deviating significantly from the variable prognostic expectations from the literature. This may be indicative of less aggressive disease behavior in our cohort or an exceptional efficacy of the therapeutic measures undertaken. Ependymoma and Medulloblastoma exhibited survival rates surpassing the literature's prognosis. Specifically, Ependymoma showed an 85% 5-year survival rate, exceeding the 50-70% cited in literature. Medulloblastoma showcased an impressive 94.87% survival rate, well above the 60-70% 5-year survival rate anticipated. These outcomes hint at possible selection bias towards less aggressive disease forms or exceptionally effective treatment strategies. Conversely, the observed survival rates for Glioblastoma and Glioma significantly exceeded the expectations based on existing literature. For Glioblastoma, the survival rate stood at 90.90% at the 2-year mark, far exceeding the usual prognosis of less than 20%. Glioma also exhibited a higher than anticipated 5-year survival rate at 50.91%, against the 20-35% expected survival range.

These observations underscore potential advancements in treatment strategies for these tumor types, exceeding historical expectations. Such discrepancies necessitate further investigation into the contributing factors, ranging from variations in disease biology to advancements in therapeutic techniques. A deeper understanding of these factors can help improve prognosis predictions and contribute to the development of more effective treatment methods.

2.3. Data Preparation for Analysis

While dealing with the RNA-seq data, it was necessary for us to ensure all pre-processing steps were in place before conducting our analysis. The raw RNA-seq data was carefully processed and cleaned, which was essential for making valid comparisons across samples and ensuring the accurate representation of the biological phenomena we were studying (Bahassi el & Stambrook, 2014; Schafer & Graham, 2002). In RNA-seq datasets, it is not uncommon to observe what could be considered as "missing values". This term, in this specific context, refers to genes that are not detected in some samples. These non-detected genes can occur for a variety of reasons including, but not limited to, the inherent sensitivity of RNA and its susceptibility to degradation. However, it is crucial to understand that these are not "missing" in the traditional sense, as they may indicate that certain genes are not expressed in particular samples, a phenomenon that is expected in RNA-seq data (Liu et al., 2019). In our dataset, while some genes were not detected in all samples, this does not imply missingness in the data, but rather represents a natural part of the biological variation among different samples. This absence is actually an informative part of the dataset and provides important insights into the gene expression patterns across different conditions (McCall et al., 2011). Consequently, the observed non-detection of some genes in certain samples does not present a problem for our analysis. Techniques such as DESeq2 are designed to handle this kind of data and utilize statistical models that account for the presence of non-detected genes (Love, Huber & Anders, 2014). Therefore, we can confidently proceed with our downstream analysis without worrying about potential biases introduced by these "missing values".

Our study began with 182 separate RNA-seq datasets, each providing gene count data. These individual datasets were merged into a single consolidated dataset that captured the gene counts for all the samples, facilitating further analysis. In addition to the gene count data, we also had two other types of datasets. One contained metadata, offering information about each sample like the patient's

age, diagnosis, and treatment history. The second dataset provided detailed information about the patients' survival, capturing if the children survived or if they succumbed to their disease.

Before conducting the differential expression analysis using DESeq2, an important step in our data processing was to filter and focus on immune-related genes. From the count dataset which included 60.357 genes, we only included approximately 1700 genes that were recognized as immune-related, based on information retrieved from the reputable gene database NanoString (NanoString Technologies, 2023). Focusing on these specific genes allowed us to narrow down the scope of our analysis to the most relevant factors. This is particularly useful when dealing with high-dimensional data like transcriptomic profiles, where the number of genes (variables) far exceeds the number of samples, potentially leading to statistical issues and difficulties in interpretation (Parekh et al., 2016).

Next, we focused on normalization of gene expression values. This is a crucial step because the number of sequencing reads attributed to a particular gene is not only influenced by its expression level but also by the total number of reads sequenced for a sample (Dillies et al., 2013). Therefore, to ensure a fair comparison across samples, DESeq2 was employed for normalization. DESeq2 calculates normalization factors to adjust for differences in sequencing depth across samples (Love, Huber & Anders, 2014). DESeq2 will be discussed in great detail in the *method section.*

3. Methods

3.1. Translating Biological Research into Data Science Objectives

In translating our biological research question into a data science task, we recognized our key objectives. Originally, we aimed to understand the immune transcriptomic profiles of pediatric high-grade brain tumors and their differences across tumor types. Guided by this, our investigation involved exploring the RNA-seq data to reveal underlying patterns in gene expression tied to the immune response. To translate the research question into a data science question, we focused on the following objectives:

First of all, identifying differentially expressed genes (DEGs) related to the immune response in pediatric high-grade brain tumors. This objective allowed us to narrow down the scope of our analysis to genes that were most relevant to our research question (Love, Huber & Anders, 2014). By comparing the DEGs across various tumor types, we aimed to uncover the specific immunosignatures associated with each tumor (Anders & Huber, 2010). Clustering samples based on gene expression patterns to identify unique immunosignatures.

In addition, survival analysis forms a crucial part of our study's approach, as it allows us to analyze and interpret the 'time-to-event', where the event could be any specific outcome of interest such as death, disease progression, or recovery. In the context of our research, survival analysis allows us to monitor the proportion of patients surviving over time across different tumor types. Additionally it enables us to explore how differing immune response profiles may affect patient survival times (Hosmer Jr, Lemeshow & May, 2008).

If our study finds, for instance, that certain immunosignatures are linked to poor survival rates, this could inform clinical decision-making by identifying patients who may need more aggressive or targeted treatment strategies. On the other hand, if certain immunosignatures are associated with better survival rates, this could help identify patients who might do well with less aggressive treatments, thereby reducing unnecessary side effects and improving their quality of life. This kind of data could also stimulate further research into novel therapeutic strategies, such as the development of drugs targeting the immune system in a way that mimics the beneficial immunosignatures (Galon & Bruni, 2019). Therefore, by integrating the study of immunosignatures with survival analysis, we can enhance our understanding of the interplay between the immune system and tumor biology (Therneau & Grambsch, 2000).

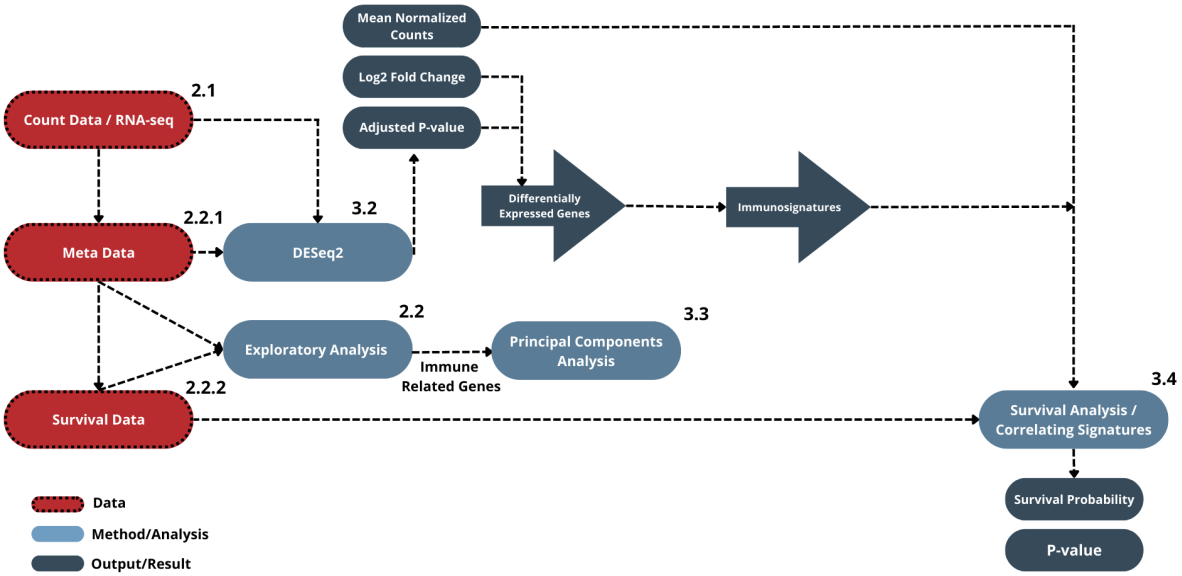
3.2. Method Selection for Analysis: DESeq2

The selection of methods for our analysis was predominantly driven by the nature of our data and the research questions we sought to answer. As our data comprised of RNA-seq counts, we required robust methods that could handle the inherent properties of this type of data, including its discrete nature, high dimensionality, and presence of technical and biological variability (Oshlack, Robinson & Young, 2010). Our entire workflow of methods is visible in *Figure 1*.

To identify differentially expressed genes (DEGs), we opted for the DESeq2 method (Love, Huber & Anders, 2014). For readers interested in a deeper mathematical exploration of these steps, we recommend the original DESeq paper by Love, Huber & Anders (2014). The article provides comprehensive mathematical underpinnings behind the model's structure and the data normalization methods employed in DESeq.

Figure 1

Workflow of methods

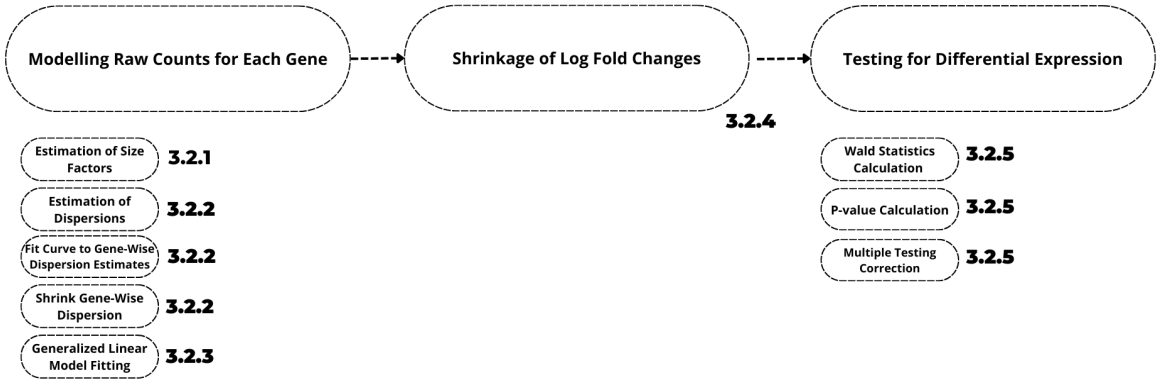


Differential gene expression analysis is a critical component in many biological studies, particularly those involving transcriptomic data. One of the robust and widely-used methods for this purpose is DESeq2, a powerful approach designed to analyze count data derived from high-throughput sequencing assays such as RNA-seq (Love, Huber & Anders, 2014). DESeq2 allows for the identification of differentially expressed genes between two or more conditions, accounting for the typical features of RNA-seq data such as overdispersion and mean-variance relationship. In the forthcoming analysis, we will be applying DESeq2 to our data to identify key genes that are differentially expressed among different samples.

In the following sections, the implementation of DESeq2 will be explained in more detail, detailing each step of the process, from normalization of raw counts and modeling of count data, to the shrinkage of log2 fold changes and the testing for differential expression. Consult *Figure 2* for the structure of the entire DESeq2 analysis.

Figure 2

DESeq2 Analysis



We will consider the following subset of our data for a more detailed walk-through of the DESeq2 analysis steps *(copy of Table 2)*:

Gene	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
ENSG00000223972.5	139	54	54	43	70
ENSG00000227232.5	1081	1488	1377	778	1030
ENSG00000278267.1	13	7	7	8	29

3.2.1. Size Factor Estimation

Before any analysis, DESeq normalizes the raw count data to account for sequencing depth and library composition. DESeq2 expects a matrix of raw counts, which reflect gene abundance, but can also depend on other factors such as gene length, sequencing depth, and library composition. DESeq2 estimates size factors to account for these "uninteresting" factors, making the expression levels more comparable between samples. For instance, in the provided data, the count for ENSG00000227232.5 is considerably higher than the counts for the other two genes. This difference could be due to a higher sequencing depth rather than an actual increased expression of ENSG00000227232.5.

DESeq2 calculates size factors to scale the counts for each gene and normalize these discrepancies across all samples. It does so by calculating a pseudo-reference sample by taking the geometric mean for each gene across all samples. The size factor for each sample is then calculated as the median of the ratios of the counts in each sample to the pseudo-reference sample (Love, Huber & Anders, 2014). The normalization method DESeq employs is called the median of ratios. Here is the simplified version of how it works:

1. A pseudo-reference sample is created by calculating the geometric mean for each gene across all samples. This geometric mean serves as a kind of "average" gene expression level. The geometric mean is used instead of the arithmetic mean because it is less sensitive to extreme values, providing a more stable reference point.
2. For each gene in each sample, the ratio of the gene's count to the corresponding count in the pseudo-reference sample is calculated.
3. The median of these ratios in each sample is calculated and taken as the "size factor" for that sample.
4. Each raw count is then divided by the respective sample's size factor to get normalized counts.

In the context of our study, normalization plays an essential role in accurate representation and interpretation of the gene expression data. For instance, let us consider the expression of MAGE-A3 in two samples, Sample 41 and Sample 63, as depicted in *Table 7.* Before normalization, the raw count of MAGE-A3 is higher in Sample 41 (1334 counts) compared to Sample 63 (975 counts). If we were to interpret this data without normalization, we might conclude that the expression of MAGE-A3 is higher in Sample 41. However, this could be misleading as it doesn not account for differences in sequencing depth or other potential biases between the two samples. Upon applying normalization, the counts for MAGE-A3 drastically change. The normalized count for Sample 41 is 1091, while for Sample 63, it is 2282. After normalization, we observe that MAGE-A3 expression is actually higher in Sample 63, a conclusion opposite to what we would have drawn from the raw counts.

Table 7

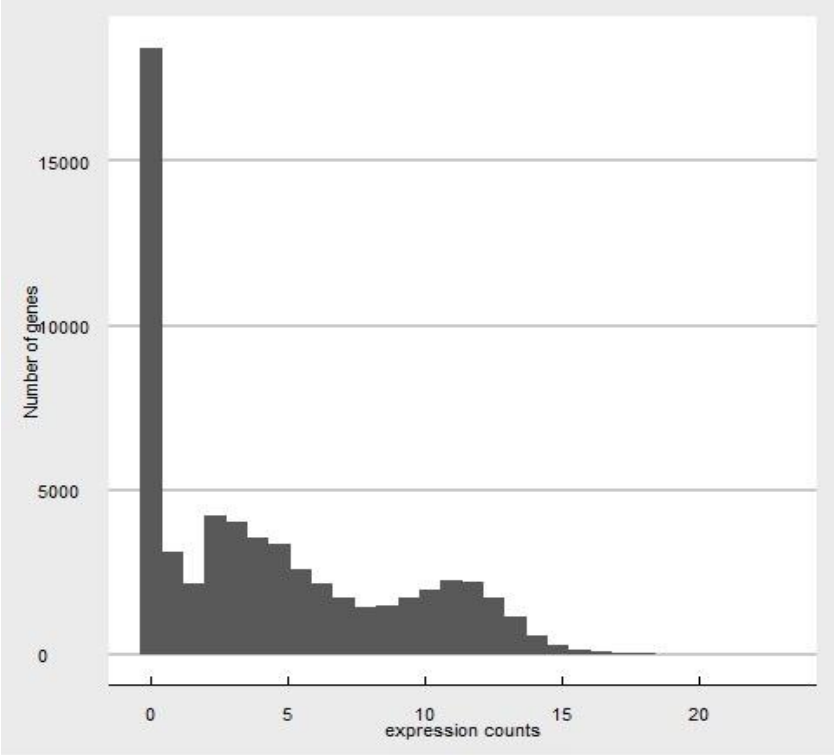
Example of Impact of Normalization on MAGE-A3

MAGE-A3	Sample 41	Sample 63
Raw Counts	1334	975
Normalized Counts	1091	2282

DESeq models count data using the negative binomial distribution. This distribution is particularly suited to RNA-seq count data, which is discrete, over-dispersed (the variance is greater than the mean), and non-negative. The distribution in our RNA-seq count data is not normally distributed, as shown in *figure 3*. This is essential because the count data is modeled using a generalized linear model (GLM) that assumes a negative binomial distribution (Anders and Huber, 2010). The negative binomial distribution has two parameters: the mean (μ) and the dispersion (α). The mean is directly estimated from the normalized counts, but the dispersion is more complex.

Figure 3

Distribution RNA-seq



Note: In Figure 3, the x-axis represents the expression counts, signifying the frequency of gene expression, whereas the y-axis denotes the number of genes expressed at that frequency. The graph represents a negative binomial distribution, which is typical in RNA-seq data.

3.2.2. Dispersion Estimation

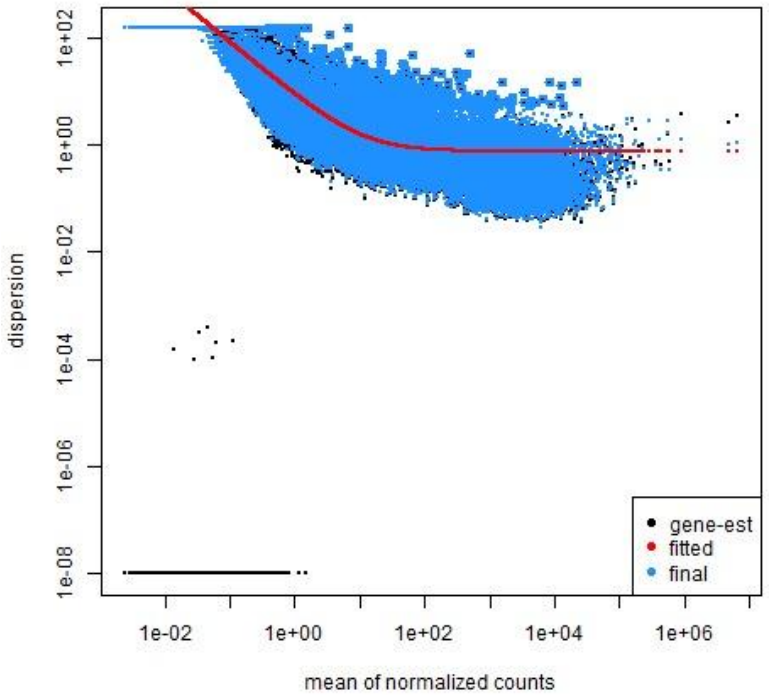
Dispersion is a measure of variability in the data. For example, ENSG00000278267.1 in the sample data shows lower counts and less variation across samples compared to ENSG00000227232.5, which would result in a lower dispersion estimate. Dispersion is a statistical metric that gauges variability in the data, computed as the square of the coefficient of variation. If a gene has a dispersion value of 0.04, it suggests a 20% fluctuation around the anticipated mean (Love, Huber & Anders, 2014). DESeq provides a method for estimating the dispersion parameter (α) for each gene. Here are the steps involved:

1. An initial estimate of dispersion for each gene is calculated using maximum likelihood estimation, creating a "dispersion plot".
2. A trend line (red line in the dispersion plot) is fitted to these initial dispersion estimates.
3. The dispersion estimates are then "shrunk" towards the fitted curve using an empirical Bayes approach to obtain final dispersion estimates. This process of shrinkage helps to stabilize the estimates, reducing the influence of random variability.

Based on the description, it seems that our data aligns well with what is expected in this kind of analysis. The dispersion estimates and the process of shrinking them towards the trend line appear to operate as expected, which is crucial for the reliability of the downstream analysis, like identifying differentially expressed genes. Consult *Figure 4* for our Dispersion Curve.

Figure 4

Gene-Wise Dispersion Curve



Note. The x-axis represents the mean normalized counts, indicating the average gene expression levels, while the y-axis shows dispersion, reflecting the variability of gene expression across different samples. Each point on the graph corresponds to a gene, with its position representing the relationship between the gene's average expression level and its dispersion.

3.2.3. Generalized Linear Model

DESeq employs the normalized counts, size factors, and dispersion estimates to construct a generalized linear model (GLM) for each gene. This GLM, grounded in the negative binomial distribution, helps determine the expected true count for each gene in every sample, which is adjusted for the size factor. The final model DESeq utilizes can be written as:

$$log2(count_{ij}) = \beta_0 + \beta_i \bullet x_j + \epsilon$$

Where:

- β_0 is the log2 expression level in the control samples (reference).
- β_i is the log2 Fold Change in gene i.
- x_j is 1 for treated and 0 for control.
- ϵ is the random error.

The hypothesis that DESeq tests for each gene is whether the log2FC (β_i) is equal to 0, meaning there is no difference in expression between the two conditions. If β_i is significantly different from zero, then the gene is considered differentially expressed.

3.2.4. Shrinkage of Log Fold Changes

In the next step, DESeq2 calculates the log2 fold changes for each gene. Fold changes measure the relative difference in gene expression between conditions. The log2 fold changes are then shrunk to stabilize the estimates and reduce the effect of outliers or genes with low counts (Love, Huber & Anders, 2014).

3.2.5. Testing for Differential Expression and Generating P-Values

1. A Wald test is performed for each gene to identify those that are differentially expressed between conditions (Love, Huber & Anders, 2014). The Wald test generates a p-value for each gene that reflects the strength of evidence against the null hypothesis (no difference in expression between conditions).
2. Finally, to control the rate of false positives due to multiple hypothesis testing, DESeq2 applies the Benjamini-Hochberg procedure to adjust the p-values, effectively controlling the false discovery rate (FDR) at a certain level, set to 5% in our study (Benjamini & Hochberg, 1995).

3.3. Method Selection for Analysis: Principal Component Analysis (PCA)

To explore our RNA-seq data thoroughly and gain insights into the structure within our samples, we leveraged Principal Component Analysis (PCA), a powerful technique for dealing with high-dimensional datasets. PCA is frequently used in genomics to unravel the internal structure of gene expression data and account for the variance within the data (Jolliffe & Cadima, 2016).
The principal utility of PCA lies in its ability to transform an original set of variables into a smaller set of uncorrelated variables, termed as principal components, without losing much of the original data's variance (Ringnér, 2008). This reduction in dimensionality eases the interpretation and visualization of complex datasets.When applying PCA to our RNA-seq data, our primary aim was to explore inherent patterns or clusters within the data and highlight any potential outliers requiring further exploration (Abdi & Williams, 2010).

3.4. Method Selection for Analysis: Survival Analysis

Survival analysis is a branch of statistics focused on analyzing the time until an event of interest occurs. It is often employed in medical research to study the time from treatment to an event such as death or relapse (Rich, 2010). In the context of high-grade pediatric brain tumors, survival analysis allows us to explore the duration from diagnosis or the start of treatment until an event like death or progression of the disease.

3.4.1. Application of the Kaplan-Meier Method in Our Research

In our survival analysis, we'll primarily use the Kaplan-Meier method, which provides a non-parametric estimation of the survival function from lifetime data (Kaplan & Meier, 1958). In essence, this method generates a survival curve that depicts the proportion of patients surviving over time. In practice, we will create separate Kaplan-Meier survival curves for different tumor types. This approach will enable us to visualize and compare the survival probabilities for each tumor type over time. We will use a similar approach to compare survival probabilities associated with cytotoxic and immunosuppressive gene signatures. By overlaying these survival curves on the same graph, we will be able to visually compare survival experiences across different groups. Any significant divergence between the curves could suggest differences in survival probabilities. In essence, Kaplan-Meier survival curves allow us to observe and compare the survival outcomes of different patient groups over time. The vertical drops in the curve signify the occurrence of the event (death or disease progression), while the flat parts represent periods with no events. The differences in these curves can reveal the influence of tumor type and gene signature on patient survival.

Even though our dataset presents complete information, with only 'event' data (deaths) and no censored (dropout) data, the use of the Kaplan-Meier method provides consistency and reproducibility, two critical aspects of any scientific endeavor. Being a standard method in survival analysis, the Kaplan-Meier method ensures that our research can be reliably compared with others (Bland & Altman, 1998). By employing the Kaplan-Meier method, our research remains in line with the standard survival analysis methodologies, enhancing its validity and reliability.

The essential working of the method in steps:

1. For each point in time where at least one event (like death or disease progression) occurred, we note down how many people were still in the study and had not had the event yet. This is our 'at risk' group.
2. Then, we look at how many people in the 'at risk' group had the event at this specific time. This gives us a ratio of the number of events to the number of people 'at risk'.
3. We subtract this ratio from 1 to get the survival probability at that specific time point.
4. Finally, we multiply all these survival probabilities together up to the time of interest. This gives us the overall probability of survival up to that time.

3.4.2. Statistical Significance Evaluation using Log-Rank Test

To statistically test the observed differences in survival curves, we will use the log-rank test, which can validate whether the observed differences in survival probabilities are statistically significant or due to chance (Bland & Altman, 2004). If the log-rank test returns a significant p-value, we can reject the null hypothesis of no difference in survival between the groups being compared.

3.4.3. Immunosignatures and Survival Analysis

The term "immunosignature" is traditionally used to depict a snapshot of the host immune response through gene expression patterns associated with immune function (Stafford et al., 2014). However, in the context of our study, we have broadened this term to represent the top altered genes in each tumor type, both up and downregulated, irrespective of their direct association with immune function. This approach mirrors that taken by other studies that have also examined top altered genes to identify signatures correlating with disease outcome (Tsai et al., 2015; Zhang et al., 2019). For instance, Tsai et al. (2015) selected the top altered genes in hepatocellular carcinoma to form a signature related to patient prognosis, while Zhang et al. (2019) used the most differentially expressed genes to predict survival in lung adenocarcinoma. Due to time constraints, this focused approach allows us to gain insights into potential relationships between these key gene expression alterations and patient survival outcomes, without necessitating a more exhaustive gene-by-gene analysis. Such targeted approaches have been used in similar high-throughput studies to efficiently identify the most impactful genetic alterations (Tarazona et al., 2011).

Building upon our revised understanding of "immunosignatures," our approach is closely aligned with the two main components we have identified: immunophenotypes and immunomodulatory signatures. When we focus on the most significantly altered genes, we indirectly gain insight into the different types of immune cells in the tumor microenvironment, thereby decoding immunophenotypes. These genes could serve as potential indicators of immune cell states and proportions within the tumors.Regarding immunomodulatory signatures, our study of key gene alterations might uncover factors produced by the tumor cells that influence the immune response. Upregulated genes could highlight elements produced excessively by tumor cells, potentially skewing the immune response to favor tumor growth. So, by broadening the scope of 'immunosignatures', we are capturing a snapshot of both the immune cell characteristics and the tumor's immune-modulating factors, providing insights into the complex tumor-immune dynamics.

In order to establish potential links between the immunosignatures (upregulated and downregulated genes) of the various tumor types and survival, we performed a series of analyses, using the Kaplan-Meier estimator and log-rank (Kleinbaum & Klein, 2010).

1. **Gene Expression Categorization:** For each gene of interest, the samples were classified into two distinct groups: those exhibiting high expression (defined as expression levels above the median) and those with low expression. This categorization allowed us to compare the survival rates of patients with high gene expression to those with lower expression.
2. **Survival Analysis:** Survival times were evaluated for each group, and the aim was to identify any significant disparities in survival between patients with high gene expression and those with low expression. We leveraged the Kaplan-Meier method to estimate these survival probabilities. The Kaplan-Meier survival curves were plotted separately for the high-expression and low-expression groups.
3. **Significance Testing:** The log-rank test was utilized to determine the statistical significance of any differences observed in the survival analysis.

4. Results

4.1. Principal Component Analysis: Delineation of Tumor Types

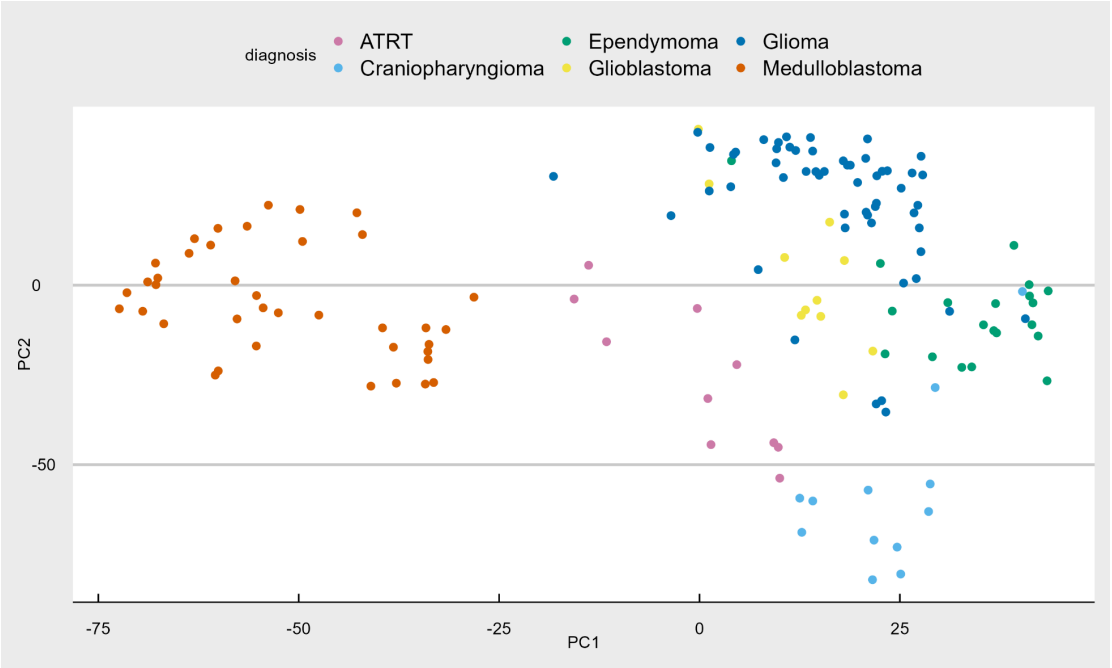
To gain insights into the overall gene expression patterns within different types of pediatric high-grade brain tumors, we focused our analysis on our 1700 immune-related genes. Having established this focus, we employed Principal Component Analysis (PCA). This technique reduces the dimensionality of our complex gene expression dataset, enabling the visualization of the primary sources of variation (Abdi & Williams, 2010).

Our PCA plot revealed clear and distinct clusters corresponding to each tumor type (*Figure 5*). This separation indicates that each tumor type possesses a unique gene expression profile, further validating the inherent molecular differences between them. These distinct gene expression patterns could be reflective of the diverse pathophysiological processes underpinning each tumor type, which can potentially be exploited for differential diagnosis and targeted therapies.

Moreover, the PCA also displayed certain outliers within these clusters. These outliers represent samples that have different gene expression profiles compared to the majority of the other samples within their tumor type. These outliers are not merely data anomalies but can serve as starting points for further analysis. They may represent unique subtypes of the tumors that are not yet recognized or may point to individual variations in the tumor's genetic makeup.

Figure 5

Principal Component Analysis color-coded according to tumor type



4.2. Differential Gene Expression Analysis

We performed a differential gene expression analysis comparing each tumor type to the reference point, Craniopharyngioma. Craniopharyngioma was chosen as the reference point for comparison due to its unique molecular profile and its often benign nature, providing an interesting contrast to the aggressive malignant characteristics typically associated with other pediatric high-grade brain tumors. The term "upregulated" refers to the process where a gene has an increased rate of expression (Alberts et al., 2002). In contrast, "downregulation" is the process where a gene has a decreased rate of expression (Alberts et al., 2002). Analyzing the results of our differential expression study, it is clear that specific genes show significantly altered expression in different pediatric high-grade brain tumors compared to the reference point, Craniopharyngioma. *Table 8* provides an overview of the number and percentage of significantly differentially expressed genes in each tumor type.

Table 8

Significantly differentially expressed genes

Tumor Type	Upregulated	Downregulated
Atypical Teratoid/Rhabdoid Tumor	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%)

As shown in the table, a varying percentage of genes showed significant differential expression in each tumor type, indicating the unique molecular signatures of each. The subsequent sections will delve into the specific genes of interest, visually represented through a volcano plot (*Figure 6*), and further discuss the top upregulated and downregulated genes for each tumor type

To provide a overview of our gene expression data, we utilized a volcano plot, a type of scatter-plot that displays the statistical significance (p-values) versus fold-change of differential expression on the y and x axes, respectively. In our volcano plot (*Figure 6*), each point represents an individual gene. The y-axis displays the negative logarithm (base 10) of the p-value, thus amplifying the distinction between statistically significant and non-significant results. The x-axis represents the log2 fold change of the gene expression between Craniopharyngioma and Glioma tumor types. Genes with a higher absolute log2 fold change (either positive or negative) are located farther from the y-axis, indicating a greater difference in expression between the two tumor types.

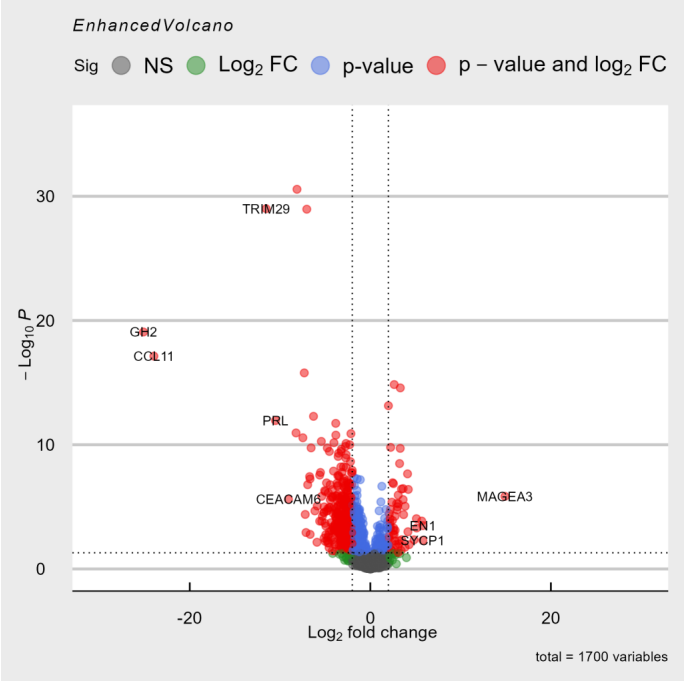
The genes upregulated in Medulloblastoma compared to Craniopharyngioma are represented by points to the right of the plot, while the downregulated genes are shown to the left. The genes with statistically significant differential expression (e.g., $p < 0.05$) are higher on the plot. We have chosen to show these results as an example because these result differ the most in comparison to craniopharyngioma in terms of upregulated and downregulated genes.

The volcano plot aids in visually emphasizing the genes that are significantly differentially expressed and exhibit a substantial fold-change. These genes, marked in *Figure 6*, provide interesting

leads for further investigation into their potential role in the pathogenesis of pediatric high-grade brain tumors.

Figure 6

Volcano plot differentially expressed genes (Craniopharyngioma vs. Medulloblastoma)



The results in *Tables 9 and 10* highlight the top differentially expressed genes across various types of pediatric high-grade brain tumors, providing insights into the molecular changes associated with these diseases (Bustin & Nolan, 2017).

Table 9

Top upregulated genes

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)

The MAGEA3 gene was consistently upregulated across all analyzed tumor types. MAGEA3 belongs to the melanoma-associated antigen gene family and has been linked with poor prognosis and survival in different cancer types (Jin et al., 2021). Overexpression of this gene suggests its possible role in the aggressive behavior of these pediatric tumors. SYCP1 also showed consistent upregulation across various tumor types. SYCP1 is typically involved in meiosis, and its dysregulation could suggest cell cycle abnormalities in these tumor types (Crichton et al., 2022). EN1 is a homeobox transcription factor previously associated with tumorigenesis and cancer progression in different tissues (Peluffo et al., 2019).

Table 10

Top downregulated genes

ATRT	Ependymoma	Glioblastoma	Glioma	Medulloblastoma
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 (-10.57)	PLA2G2A (-9.54)

In terms of downregulated genes, GH2 and CCL11 were found to be downregulated in ATRT. GH2 is associated with growth hormone activity, and its downregulation may affect tumor growth. CCL11 is a small cytokine associated with the inflammatory response. Downregulation might imply an altered immune response within the tumor environment (Miller et al., 2010). Furthermore, the PITX2 and PRL genes, involved in pituitary development and hormone regulation respectively, were found downregulated in Ependymoma (Kelberman et al., 2009; Suh et al., 2002). This downregulation could be linked to alterations in hormone-related pathways, which may play a role in tumor progression.

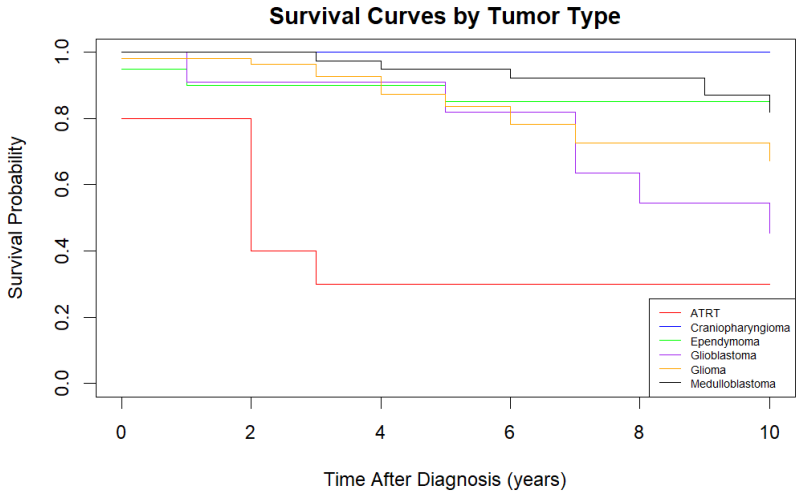
It is important to interpret these results with caution as changes in gene expression do not necessarily mean the gene itself is a cause or effect of cancer development. Many factors, including alterations in gene regulation, gene mutations, and interactions with other genes or proteins, can contribute to changes in gene expression. While this study provides insight into the potential role of these genes in pediatric high-grade brain tumors, more comprehensive functional studies are needed to understand their specific roles in tumorigenesis, progression, and potential as therapeutic targets.

4.3. Survival Analysis

We conducted a survival analysis for six types of pediatric high-grade brain tumors: Atypical Teratoid/Rhabdoid Tumor (ATRT), Craniopharyngioma, Ependymoma, Glioblastoma, Glioma, and Medulloblastoma. Our findings revealed significant differences between the observed and expected survival outcomes for each tumor type ($p < 0.00001$), implying that the actual survival rates in our sample deviated markedly from the rates reported in the literature. In essence, our Kaplan-Meier survival analysis, in combination with chi-square tests, substantiates that the actual survival rates for these pediatric high-grade brain tumors in our dataset significantly differ. Consult *Table 11 and Figure 7* for the complete results

Figure 7

Survival Curve Per Tumor Type



Note. Figure 7 illustrates the survival curves per tumor type. Each line represents a different tumor type, showing the survival probability over time since diagnosis. The p-value (p = 1e-08) suggests that the differences in survival rates among the tumor types are statistically significant.

Table 11

Survival Analysis results

Tumor Type	Diagnosed	Survivors	Expected Deaths	Chi-square
Atypical Teratoid/Rhabdoid Tumor	10	3	1.82	15.73
Craniopharyngioma	12	12	5.41	6.16
Ependymoma	20	17	7.90	3.65
Glioblastoma	11	1	3.39	14.14
Glioma	55	28	19.45	4.71
Medulloblastoma	39	32	16.04	7.43

4.4. Immunosignatures and Survival Analysis

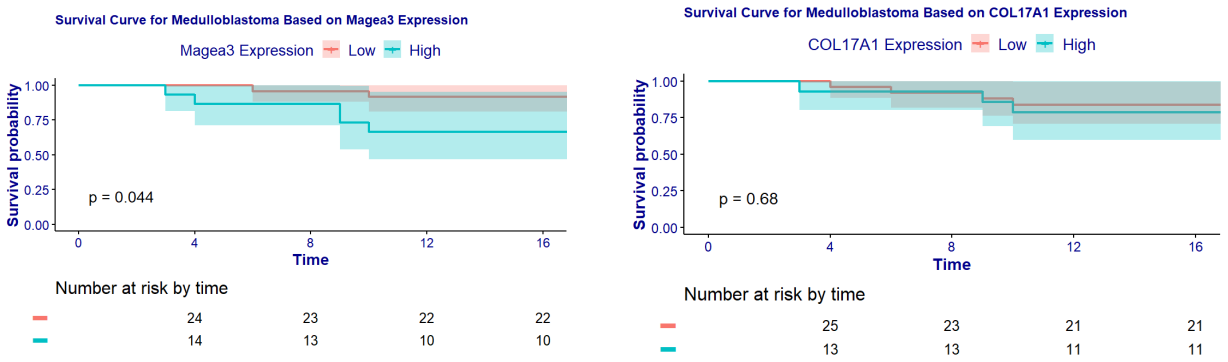
Pediatric high-grade brain tumors demonstrate distinct molecular profiles, as evidenced by our differential gene expression analysis. To further explore the prognostic implications of these molecular changes, we focused on the most significantly upregulated and downregulated genes from each tumor type to generate an "immunosignature" for each tumor type.

Using the Kaplan-Meier survival analysis method and the log-rank test, we aimed to correlate these immunosignatures with survival outcomes. Due to time constraints we will only use the top regulated genes for Medulloblastoma. MAGE-A3 emerged as the most significantly upregulated gene, and its expression demonstrated a statistically significant association with survival outcomes, with a p-value of 0.044. Particularly, we observed that samples with high expression of MAGE-A3 have a significantly lower survival expectancy than those with low expression. This finding emphasizes the potential role of MAGE-A3 as a marker of disease progression and suggests its value as a prognostic factor in Medulloblastoma, indicating poorer prognosis with higher gene expression. This points to the critical need to consider the level of MAGE-A3 expression in therapeutic planning and management of the disease.

In contrast, despite being the most significantly downregulated gene, COL17A1 did not show a significant correlation with survival outcomes, with a p-value of 0.68. This observation suggests a complex scenario where the influence of COL17A1 on survival may be indirect, dependent on other genetic or clinical factors. Consult *Figure 8* for the results of the curves.

Figure 8

Survival Curve Medulloblastoma Based on Magea3 and COL17A1 Expression



Note. Figure 8 represents the survival curves for Medulloblastoma patients based on the expression of Magea3 and COL17A1. Each line demonstrates survival probability over time for patients with high or low expressions of these two genes.

5. Conclusion

5.1. Summary of Findings

Our data analysis of pediatric high-grade brain tumors has delivered significant insights into the molecular landscape and survival rates of these aggressive malignancies. The Principal Component Analysis (PCA) effectively delineated distinct clusters corresponding to each tumor type, indicating the presence of unique gene expression profiles for each.

We then performed differential expression analysis, which led to the identification of several differentially expressed genes across the various tumor types. In particular, key genes such as MAGEA3 and SYCP1 showed consistent upregulation across different tumor types. In the context of Medulloblastoma, the expression of MAGEA3 presented a notable correlation with survival outcomes, with a statistically significant p-value of 0.044. This suggests MAGEA3's potential role not only in disease progression but also as a prognostic indicator.

On the other hand, genes like GH2, CCL11, PITX2, and PRL were found to be downregulated in certain tumors. Interestingly, COL17A1 emerged as the most significantly downregulated gene in Medulloblastoma. However, its expression did not demonstrate a significant association with survival outcomes, returning a p-value of 0.68. This points towards the complex interactions and dependencies between gene expression and survival, warranting further investigation.

These findings not only highlight the molecular diversity of pediatric high-grade brain tumors but also offer potential targets for future research and potential therapeutic interventions. They further emphasize the significance of the immunosignature and its correlation with survival outcomes in these tumors.

Our survival analysis also presented a contrast between the anticipated survival rates based on the literature and the actual survival rates observed in our dataset. Despite the generally poor prognosis associated with these tumors, certain types, such as Ependymoma and Medulloblastoma, exhibited higher survival rates than anticipated. This data underscores the importance of continued study and data collection to improve prognostic accuracy.

5.2. Answer to the Research Question

In response to our first research question: "Which genes are differentially expressed in pediatric high-grade brain tumors and what unique immunosignatures can we identify from these expression patterns?" - our study has successfully accomplished the goal of characterizing the immune transcriptomic profile of pediatric high-grade brain tumors. DESeq2 analysis led us to distinguish unique gene expression patterns and differentially expressed genes tied to these tumors that may influence their development. A noteworthy finding was the upregulation of the gene MAGEA3 across various tumor types, which adds depth to our understanding of the complex tumor microenvironment and its pivotal role in the progression of these malignancies.

Addressing our second research question: "Do these identified immunosignatures significantly influence survival outcomes in patients diagnosed with these pediatric high-grade brain tumors?" - we investigated the correlation of identified immunosignatures with patient survival rates. Our results indicate the significant role of MAGEA3, specifically in Medulloblastoma, where its expression was significantly associated with survival outcomes (p-value 0.044). This suggests its promising role as a prognostic marker. However, the case of the gene COL17A1 highlighted the complex interactions at play. Despite being the most significantly downregulated gene in Medulloblastoma, it did not show a significant association with survival outcomes (p-value 0.68).

Furthermore, our comparison of expected versus observed survival rates revealed potential discrepancies, indicating a possible underestimation of survival chances in certain tumor types. This finding emphasizes the importance of ongoing data collection and analysis to enhance the accuracy of prognostic predictions.

By connecting the dots between our findings on immune transcriptomics and survival outcomes, we offer a clinically relevant viewpoint to the study of pediatric high-grade brain tumors. Our findings underscore the potential impact of molecular heterogeneity on survival, laying the groundwork for future research.

6. Discussion

6.1. Implications of Findings

The consistent upregulation of MAGEA3 across different tumor types, as identified in our research, underscores its potential significance in the tumorigenesis and progression of these aggressive tumors. While the specific role of MAGEA3 in these tumors requires further elucidation, its presence in multiple tumor types suggests a broad spectrum of influence that warrants further investigation. Similarly, the downregulation of genes like COL17A1, despite no significant correlation with survival outcomes in our study, adds to our understanding of the molecular complexities of these tumors.

Our research also brings the importance of understanding the nuanced relationships between gene expression and survival outcomes into focus. Survival rates for certain tumor types were higher than anticipated, highlighting the potential need to revise prognostic models. This observation underscores the need for continuous data collection and meticulous analysis to refine our understanding of these aggressive diseases and enhance prognostic accuracy.

Taken together, our findings serve to broaden the existing body of knowledge in pediatric high-grade brain tumor research. The correlations identified here can help guide future research, which can potentially translate into improved diagnostic methods, treatment planning, and therapeutic interventions. This study lays a robust groundwork for future exploration into the relationships between gene expression, tumorigenesis, and patient outcomes, promising potential advancements in the field of pediatric oncology.

6.2. Comparison with Previous Studies

Our findings offer both unique insights and confirming evidence for the known heterogeneity of pediatric high-grade brain tumors. The distinct gene expression profiles for each tumor type found in our study is in line with previous genomic analyses that have revealed fundamental molecular differences between various pediatric brain tumors (Sturm et al., 2014; Ramaswamy et al., 2016). Our study builds on this by linking specific gene expression profiles with survival outcomes, extending the understanding of how these molecular differences could potentially translate to clinical prognosis.

For instance, we found that the MAGEA3 gene was consistently upregulated across all analyzed tumor types. This is in line with the work of Jin et al. (2021), who linked the MAGEA3 gene to poor prognosis and survival in different cancer types. Our findings suggest that overexpression of this gene may contribute to the aggressive behavior of these pediatric tumors.

Moreover, we observed distinct survival rates across the various tumor types, which differed significantly from the expected survival rates from the literature. While previous studies have highlighted the poor prognosis associated with Glioblastoma (Ostrom et al., 2015) and Medulloblastoma (Ramaswamy et al., 2016), our study observed considerably higher survival rates. This divergence could be due to multiple factors, including advancements in treatments or potential selection bias within our study population.

Our study also reaffirms the role of certain genes, like EN1, associated with tumorigenesis and cancer progression in different tissues (Peluffo et al., 2019), demonstrating the interconnected nature of cancer biology across different cancer types.

In contrast, we also observed some genes which have not been frequently reported in literature, providing new directions for future research. One such gene is GH2, associated with growth hormone activity, which we found to be downregulated in Atypical Teratoid/Rhabdoid Tumor

(ATRT). This novel finding could stimulate further research into its potential role in tumor growth and the possibility of targeting such genes for treatment.

6.3. Limitations

Even though our study has provided significant insights, it is not without its limitations. The absence of a healthy control group in our dataset limited our ability to fully contextualize the differential gene expression observed in pediatric high-grade brain tumors. In future studies, inclusion of a healthy control group could offer a more robust comparison for the characterization of the transcriptomic landscape of these tumors.

Additionally, our dataset, despite being comprehensive, only represents a subset of all pediatric high-grade brain tumors. This could potentially introduce selection bias, which should be considered when interpreting the findings. Future research should aim to utilize larger and more diverse cohorts, with integrated molecular and clinical data to offer more comprehensive insights.

While our "immunosignature" approach proved insightful, it may not capture the full complexity of the tumor's molecular profile, which could involve subtle changes in several genes rather than drastic changes in a few. Especially in our survival analysis, which examined the associations between survival outcomes and the expression of only two genes (MAGEA3 and COL17A1), was inherently limited in scope. While the analysis of these genes did reveal promising potential, it is necessary to acknowledge that it represents just a fraction of the possible gene-survival associations within our dataset. Future studies should aim to expand this analysis to include more genes and explore the potential relationships therein.

Finally, it is worth noting that our project was conducted within a limited time frame of just eight weeks. This constraint inevitably meant that many potential avenues of exploration and analysis fell beyond our reach. Many aspects of this complex topic, from extending the survival analysis to exploring additional tumor types, were not feasible within our limited timeframe. This, however, does not diminish the value of our findings but rather highlights the potential for further research.

6.4. Future Directions

Our study paves the way for several exciting avenues for future research. Primarily, the roles of the identified genes, especially MAGEA3 and SYCP1, warrant further investigation. Detailed exploration into their precise roles in tumorigenesis and progression of pediatric high-grade brain tumors could provide potential targets for therapeutic interventions.

Moreover, we limited our survival analysis to two genes (MAGEA3 and COL17A1) and to one tumor type (Medulloblastoma) due to constraints of time and resources. Therefore, expanding survival analysis to include more genes and diverse tumor types will likely offer a more holistic understanding of how gene expression correlates with survival outcomes. Such an expanded investigation could serve to enhance the prognostic accuracy and inform the development of personalized therapeutic strategies.

In addition, the "immunosignatures" identified in this study, based on the expression of around 1700 immune-related genes, should be considered a starting point, open to refinement or expansion. The selection of these genes was driven by the current state of knowledge and available resources, and hence is subject to potential biases and limitations. Future research should seek to confirm and possibly broaden these immunosignatures, incorporating new discoveries and emerging data.

Lastly, to provide a more comprehensive view of differential gene expression in pediatric high-grade brain tumors, future studies could benefit from integrating data from healthy controls. This

would allow for a more robust identification of gene expression patterns uniquely associated with these tumors.

While our findings provide valuable insights into the complex molecular landscape of pediatric high-grade brain tumors, they also highlight the importance of continuous research efforts to refine and expand our understanding, with the ultimate goal of improving patient prognosis and therapeutic strategies.

7. Literature

Abdi, H., & Williams, L. J. (2010). Principal component analysis. Wiley Interdisciplinary Reviews: *Computational Statistics*, 2(4), 433-459.

Adiconis, X., et al. (2013). Comparative analysis of RNA sequencing methods for degraded or low-input samples. *Nature Methods*, 10(7), 623–629. https://doi.org/10.1038/nmeth.2483

Alberts, B., Johnson, A., Lewis, J., et al. (2002). Molecular Biology of the Cell (4th ed.). New York, NY: *Garland Science*.

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., ... & Levy, R. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769), 503-511.

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, 11(10), R106. https://doi.org/10.1186/gb-2010-11-10-r106

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Astivia, O. L. O., & Zumbo, B. D. (2019). Heteroskedasticity in Multiple Regression Analysis: What it is, How to Detect it and How to Solve it with Applications in R and SPSS. *Practical Assessment, Research and Evaluation*, 24(1), 1. https://doi.org/10.7275/q5xr-fr95

Bahassi el, M., and Stambrook, P.J. (2014). Next-generation sequencing technologies: breaking the sound barrier of human genetics. *Mutagenesis* 29, 303–310.

Bhatia, A., Kumar, Y., & Cancer, C. (2017). MAGE-A3: an attractive target molecule for cancer immunotherapy. *Cellular immunology*, 318, 1-6.

Bland JM, Altman DG. The logrank test. *BMJ*. 2004;328(7447):1073.

Bourgon, R., Gentleman, R., & Huber, W. (2010). Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences*, 107(21), 9546-9551.

Broad Institute. (2019). Picard Toolkit. Broad Institute, GitHub repository. http://broadinstitute.github.io/picard/

Bustin, S., & Nolan, T. (2017). Talking the talk, but not walking the walk: RT-qPCR as a paradigm for the lack of reproducibility in molecular research. *European Journal of Clinical Investigation*, 47(10), 756-774. https://doi.org/10.1111/eci.12801

Chomczynski, P. (1993). A reagent for the single-step simultaneous isolation of RNA, DNA and proteins from cell and tissue samples. *BioTechniques*, 15(3), 532–537.

Clark, T. G., Bradburn, M. J., Love, S. B., & Altman, D. G. (2003). Survival analysis part IV: further concepts and methods in survival analysis. *British Journal of Cancer*, 89(5), 781–786. https://doi.org/10.1038/sj.bjc.6601119

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., ... & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome biology*, 17(1), 13.

Crichton, J. H., Dunce, J. M., Dunne, O. M., Salmon, L. J., Devenney, P. S., Lawson, J. A., Adams, I. R., & Davies, O. (2022). Structural maturation of SYCP1-mediated meiotic chromosome synapsis through conformational remodelling by molecular adapter SYCE3. *bioRxiv (Cold Spring Harbor Laboratory*). https://doi.org/10.1101/2022.03.06.483192

Daneman, R., & Prat, A. (2015). The blood–brain barrier. Cold Spring Harbor *Perspectives in Biology*, 7(1), a020412.

Deng, J., Wang, E. W., Jenkins, R. W., Li, S., Dries, R., Yates, K. A., Chhabra, S., Huang, W., Liu, H., Aref, A. R., Ivanova, E. P., Paweletz, C. P., Bowden, M., Zhou, C. W., Herter-Sprie, G. S., Weiss, J., Bisi, J. E., Lizotte, P. H., Merlino, A. A., . . . Wong, K. (2017). CDK4/6 Inhibition Augments Antitumor Immunity by Enhancing T-cell Activation. *Cancer Discovery*, 8(2), 216–233. https://doi.org/10.1158/2159-8290.cd-17-0915

Bindea, G., Mlecnik, B., Tosolini, M., Kirilovsky, A., Waldner, M., Obenauf, A. C., ... & Galon, J. (2013). Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. *Immunity*, 39(4), 782-795.

Dobin, A., et al. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 29(1), 15–21. https://doi.org/10.1093/*bioinformatics*/bts635

Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences,* 95(25), 14863–14868. https://doi.org/10.1073/pnas.95.25.14863

Ellison, D. W., Dalton, J., Kocak, M., Nicholson, S. L., Fraga, C., Neale, G., ... & Onar-Thomas, A. (2011). Medulloblastoma: clinicopathological correlates of SHH, WNT, and non-SHH/WNT molecular subgroups. *Acta neuropathologica*, 121(3), 381-396.

Fangusaro, J. (2012). Pediatric high grade glioma: a review and update on tumor clinical characteristics and biology. *Frontiers in Oncology*, 2, 105.

Frühwald, M. C., Biegel, J. A., Bourdeaut, F., Roberts, C. T., & Chi, S. N. (2016). Atypical teratoid/rhabdoid tumors—current concepts, advances in biology, and potential future therapies. *Neuro-oncology,* 18(6), 764–778. https://doi.org/10.1093/neuonc/nov264

Gupta, N., Goumnerova, L. C., Manley, P., Chi, S. N., & Neuberg, D. (2018). Prospective feasibility and safety assessment of surgical biopsy for patients with newly diagnosed diffuse intrinsic pontine glioma. *Neuro-oncology*, 20(11), 1547-1555.

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: The next generation. *Cell*, 144(5), 646-674.

Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., & Xu, W. (2018). Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genomics-Proteomics*, 15(1), 41-51.

Hunger, S. P., & Mullighan, C. G. (2015). Acute lymphoblastic leukemia in children. New England *Journal of Medicin*e, 373(16), 1541-1552.

Jackson, C. M., Choi, J., & Lim, M. (2019). Mechanisms of immunotherapy resistance: Lessons from glioblastoma. *Nature Immunology*, 20(9), 1100-1109.

Jin, J., Tu, J., Ren, J., Cai, Y., Chen, W., Zhang, L., Zhang, Q., & Zhu, G. (2021). Comprehensive Analysis to Identify MAGEA3 Expression Correlated With Immune Infiltrates and Lymph Node Metastasis in Gastric Cancer. *Frontiers in Oncology*, *11*. https://doi.org/10.3389/fonc.2021.784925

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Jones, C., & Baker, S. J. (2014). Unique genetic and epigenetic mechanisms driving pediatric diffuse high-grade glioma. *Nature Reviews Cancer*, 14(10), 651-661.

Jones, D. T., Jäger, N., Kool, M., Zichner, T., Hutter, B., Sultan, M., ... & Lichter, P. (2012). Dissecting the genomic complexity underlying medulloblastoma. *Nature*, 488(7409), 100-105.

Kassambara A, Kosinski M, Biecek P, Fabian S. survminer: Drawing Survival Curves using 'ggplot2'. R package version 0.4.7. 2020.

Kaplan EL, Meier P. Nonparametric Estimation from Incomplete Observations. *J American Statistical Association*. 1958;53(282):457–481.

Kelberman, D., Rizzoti, K., Lovell-Badge, R., Robinson, I. C. a. F., & Dattani, M. T. (2009). Genetic Regulation of Pituitary Gland Development in Human and Mouse. *Endocrine Reviews*, 30(7), 790–829. https://doi.org/10.1210/er.2009-0008

Kleinbaum DG, Klein M. Survival Analysis. *Springer New York.* 2012.

Kline, C., Joseph, N. M., Grenert, J. P., Van Ziffle, J., Talevich, E., Onodera, C., Aboian, M., Cha, S. I., Solomon, D. A., Braunstein, S., Torkildson, J., Samuel, D., Bloomer, M., De Alba Campomanes, A. G., Banerjee, A., Butowski, N., Raffel, C., Tihan, T., Bollen, A. W., . . . Nicolaides, T. (2016). Targeted next-generation sequencing of pediatric neuro-oncology patients improves diagnosis, identifies pathogenic germline mutations, and directs targeted therapy. *Neuro-oncology*, now254. https://doi.org/10.1093/neuonc/now254

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2014). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13, 8-17.

Liao, Y., Smyth, G. K., & Shi, W. (2019). The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Research*, 47(8), e47. https://doi.org/10.1093/nar/gkz114

Liu, Y., Zhou, J., & White, K. P. (2019). RNA-seq differential expression studies: more sequence or more replication?. *Bioinformatics*, 30(3), 301-304.

Louis, D. N., Perry, A., Reifenberger, G., Von Deimling, A., Figarella-Branger, D., Cavenee, W. K., Ohgaki, H., Wiestler, O. D., Kleihues, P., & Ellison, D. W. (2016). The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathologica*, 131(6), 803–820. https://doi.org/10.1007/s00401-016-1545-1

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Maecker, H. T., McCoy, J. P., & Nussenblatt, R. (2012). Standardizing immunophenotyping for the Human Immunology Project. *Nature Reviews Immunology*, 12(3), 191-200.

Mantovani, A., Marchesi, F., Malesci, A., Laghi, L., & Allavena, P. (2017). Tumour-associated macrophages as treatment targets in oncology. *Nature Reviews Clinical Oncology*, 14(7), 399-416.

Maude, S. L., Laetsch, T. W., Buechner, J., Rives, S., Boyer, M., Bittencourt, H., ... & Rheingold, S. R. (2018). Tisagenlecleucel in children and young adults with B-cell lymphoblastic leukemia. *New England Journal of Medicine*, 378(5), 439-448.

McCall, M. N., Uppal, K., Jaffee, H. A., Zilliox, M. J., & Irizarry, R. A. (2010). The Gene Expression Barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes. *Nucleic Acids Research*, 39(suppl_1), D1011–D1015. https://doi.org/10.1093/nar/gkq1259

Merchant, T. E., Pollack, I. F., & Loeffler, J. S. (2010). Brain tumors across the age spectrum: Biology, therapy, and late effects. *Seminars in Radiation Oncology*, 19(1), 58-66.

Miller, A. M., Asquith, D. L., Pfeil, A., Anderson, L. A., Holmes, W. M., McKenzie, A. N. J., Xu, D., Sattar, N., McInnes, I. B., & Liew, F. Y. (2010). Interleukin-33 Induces Protective Effects in Adipose Tissue Inflammation During Obesity in Mice. *Circulation Research*, 107(5), 650–658. https://doi.org/10.1161/circresaha.110.218867

Mulhern, R. K., Merchant, T. E., Gajjar, A., Reddick, W. E., & Kun, L. E. (2004). Late neurocognitive sequelae in survivors of brain tumours in childhood. *The Lancet Oncology,* 5(7), 399-408.

Müller, H. L. (2014). Craniopharyngioma. *Endocrine reviews*, 35(3), 513-543.

NanoString Technologies, Inc. (2023, June 6). *NanoString: Experience the Power of Spatial Biology*. NanoString. https://nanostring.com/

Nikitovic, M., Stanić, D., Pekmezovic, T., Gazibara, M. S., Bokun, J., Paripovic, L., Grujičić, D., Sarić, M. M., & Mišković, I. (2016). Pediatric glioblastoma: a single institution experience. *Childs Nervous System*, 32(1), 97–103. https://doi.org/10.1007/s00381-015-2945-6

Oshlack, A., Robinson, M. D., & Young, M. D. (2010). From RNA-seq reads to differential expression results. *Genome Biology*, 11(12), 220. https://doi.org/10.1186/gb-2010-11-12-220

Ostrom QT, de Blank PM, Kruchko C, Petersen CM, Liao P, Finlay JL, Stearns DS, Wolff JE, Wolinsky Y, Letterio JJ, Barnholtz-Sloan JS. Alex's Lemonade Stand Foundation Infant and Childhood Primary Brain and Central Nervous System Tumors Diagnosed in the United States in 2007–2011. *Neuro-Oncology*. 2015;16 Suppl 10:x1-x36. doi:10.1093/neuonc/nou327.

Ostrom, Q.T., Gittleman, H., & Kruchko, C. (2015). CBTRUS statistical report: Primary brain and central nervous system tumors diagnosed in the United States in 2008-2012. *Neuro-oncology*, iv1-iv62.

Pajtler, K. W., Witt, H., Sill, M., Jones, D. R., Hovestadt, V., Kratochwil, F., Wani, K., Tatevossian, R. G., Punchihewa, C., Johann, P., Reimand, J., Warnatz, H., Ryzhova, M., Mack, S., Ramaswamy, V., Capper, D., Schweizer, L., Sieber, L., Wittmann, A., . . . Pfister, S. M. (2015). Molecular Classification of Ependymal Tumors across All CNS Compartments, Histopathological Grades, and Age Groups. *Cancer Cell*, 27(5), 728–743. https://doi.org/10.1016/j.ccell.2015.04.002

Parekh, S., Ziegenhain, C., Vieth, B., Enard, W., & Hellmann, I. (2016). zUMIs - A fast and flexible pipeline to process RNA sequencing data with UMIs. *GigaScience*, 7(6). https://doi.org/10.1093/gigascience/giy059

Peluffo, G., Subedee, A., Harper, N., Kingston, N. L., Jovanović, B. S., Flores, F., Stevens, L. M., Beca, F., Trinh, A., Chilamakuri, C. S. R., Papachristou, E. K., Murphy, K. A., Su, Y., Marusyk, A., D'Santos, C., Rueda, O. M., Beck, A. H., Caldas, C., Carroll, J. S., & Polyak, K. (2019). EN1 Is a Transcriptional Dependency in Triple-Negative Breast Cancer Associated with Brain Metastasis. *Cancer Research*, 79(16), 4173–4183. https://doi.org/10.1158/0008-5472.can-18-3264

Princess Máxima Center. (n.d.). Princess Máxima Center for Pediatric Oncology. Retrieved from https://www.prinsesmaximacentrum.nl/en/

Princess Máxima Center (2023). Sequencing Data Processing Guidelines. Utrecht, Netherlands.

Pocock SJ, Clayton TC, Altman DG. Survival plots of time-to-event outcomes in clinical trials: good practice and pitfalls. *The Lancet*. 2002;359(9318):1686-1689.

Pui, C. H., Yang, J. J., Hunger, S. P., Pieters, R., Schrappe, M., Biondi, A., ... & Horibe, K. (2015). Childhood acute lymphoblastic leukemia: progress through collaboration. *Journal of Clinical Oncology*, 33(27), 2938-2948.

Quail, D. F., & Joyce, J. A. (2013). Microenvironmental regulation of tumor progression and metastasis. *Nature Medicine*, 19(11), 1423-1437.

Ramaswamy, V., Remke, M., Bouffet, E., Faria, C. C., Perreault, S., Cho, Y. J., ... & Northcott, P. A. (2016). Risk stratification of childhood medulloblastoma in the molecular era: the current consensus. *Acta neuropathologica*, 131(6), 821-831.

Reck, M., Rodríguez-Abreu, D., & Robinson, A. G. (2016). Pembrolizumab versus chemotherapy for PD-L1-positive non-small-cell lung cancer. *New England Journal of Medicine*, 375(19), 1823-1833.

Restifo, N. P., Dudley, M. E., & Rosenberg, S. A. (2012). Adoptive immunotherapy for cancer: harnessing the T cell response. *Nature Reviews Immunology*, 12(4), 269-281.

Ribas, A., & Wolchok, J. D. (2018). Cancer immunotherapy using checkpoint blockade. *Science*, 359(6382), 1350-1355.

Ringnér, M. (2008). What is principal component analysis?. *Nature biotechnology*, 26(3), 303-304.

Ris, M. D., & Noll, R. B. (1994). Long-term neurobehavioral outcome in pediatric brain-tumor patients: review and methodological critique. *Journal of Clinical and Experimental Neuropsychology,* 16(1), 21-42.

Ross, J. A., Vega, J. A., Plant, A., MacDonald, T. J., Becher, O. J., & Hambardzumyan, D. (2021). Tumour immune landscape of paediatric high-grade gliomas. *Brain*, 144(9), 2594–2609. https://doi.org/10.1093/brain/awab155

Schafer, J.L., and Graham, J.W. (2002). Missing data: our view of the state of the art. *Psychological Methods* 7, 147–177.

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., & Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetic*s, 15(2), 121–132. https://doi.org/10.1038/nrg3642

Soneson, C., Love, M. I., & Robinson, M. D. (2015). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, 4.

Stafford, P., Halperin, R., Legutki, J. B., Magee, D. M., Galgiani, J., & Johnston, S. A. (2014). Physical characterization of the "immunosignaturing effect". *Molecular & Cellular Proteomics*, 13(4), 1119–1125. https://doi.org/10.1074/mcp.M113.032821.

Stupp, R., Mason, W. P., van den Bent, M. J., Weller, M., Fisher, B., Taphoorn, M. J., ... & Mirimanoff, R. O. (2005). Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *New England Journal of Medicine,* 352(10), 987-996.

Sturm, D., Witt, H., Hovestadt, V., Khuong-Quang, D. A., Jones, D. T., Konermann, C., ... & Sahm, F. (2014). Hotspot mutations in H3F3A and IDH1 define distinct epigenetic and biological subgroups of glioblastoma. *Cancer Cell*, 22(4), 425-437.

Suh, H., Gage, P. J., Drouin, J., & Camper, S. A. (2002). Pitx2is required at multiple stages of pituitary organogenesis: pituitary primordium formation and cell specification. *Development*, 129(2), 329–337. https://doi.org/10.1242/dev.129.2.329

Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A., Conesa, A. (2011). Differential expression in RNA-seq: A matter of depth. *Genome Research*, 21(12), 2213-2223.

Therneau TM, Grambsch PM. Modeling Survival Data: Extending the Cox Model. *Springer New York*. 2000.

Tsai, C.A., Chen, J.J. (2015). Multivariate analysis of variance test for gene set analysis. *Bioinformatic*s, 29(8), 1067-1073.

Vinay, D. S., Ryan, E. P., Pawelec, G., Talib, W. H., Stagg, J., Elkord, E., ... & Lichtor, T. (2015). Immune evasion in cancer: Mechanistic basis and therapeutic strategies. *Seminars in cancer biology*, 35, S185-S198.

Ward, E., DeSantis, C., Robbins, A., Kohler, B., & Jemal, A. (2014). Childhood and adolescent cancer statistics, 2014. CA: *A Cancer Journal for Clinicians*, 64(2), 83-103.

Weber, R., Fleming, V., Hu, X., Nagibin, V., Groth, C., Altevogt, P., Utikal, J., & Umansky, V. (2017). Myeloid-derived suppressor cells hinder the anti-cancer activity of immune checkpoint inhibitors. *Frontiers in Immunology*, 9, 1310.

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3), 37-52.

Zhang, H., Meltzer, P., Davis, S. (2019). RCircos: an R package for Circos 2D track plots. *BMC Bioinformatic*s, 14, 244.

8. Appendix

Appendix A. Ethical and Legal Considerations

Our study adhered strictly to the basic criteria for processing sequencing data as defined by the Princess Máxima Center. The following criteria were maintained: The Princess Máxima Center was the joint data controller for the biological data used in this study. All included subjects were patients of the center, fulfilling the first criteria (Princess Máxima Center, 2023). Informed consent was obtained from all patients or their guardians, fulfilling the second criteria. All metadata were collected from patients who provided informed consent and were accessed via the Central Subject Registry (CSR) (Princess Máxima Center, 2023). To maintain patient confidentiality and adhere to ethical standards, all directly identifiable data in the dataset were either removed or pseudonymized. This is in compliance with the third basic criteria outlined by the center (Princess Máxima Center, 2023).

Requesting sequencing data or generating new sequencing data was carried out as per the guidelines of the Princess Máxima Center. When sequencing data was not available for selected cohorts, new sequencing data was generated by the Laboratory of Pediatric Oncology (LPO). All requests were initiated by contacting the Sequence Facility of the LPO and the procedure included the delivery of a completed institutional metadata sheet before processing. This was done to improve the FAIRness (Findability, Accessibility, Interoperability, and Reusability) of the data (Princess Máxima Center, 2023). Following the generation of new sequencing data, the process involved subsequent data analyses and registration of the sequenced data with the delivered metadata on the Máxima storage. The corresponding metadata was registered on the Trecode platform (Princess Máxima Center, 2023).

Appendix B. R Code

```
# Load Packages
library(tidyverse)
library(DESeq2)
library(readxl)
library(pheatmap)
library(psych)
library(EnhancedVolcano)
library(ggthemes)
library(survival)
library(purrr)
library(survminer)

# Set Working Environment
setwd("C:/Users/cbogaar3/surfdrive/Shared/Data source files")

# Load all data
count_data <- readRDS("count_data.Rdata")
meta <- readRDS("meta.Rdata") # Includes Survival Data
gene <- readRDS("gene.Rdata")
immune_genes <- readRDS("immune_genes.Rdata")

##################################
# 1. Exploratory analysis (meta)
##################################

# View the first few rows of the data
head(meta)

# Summary of the data
summary(meta)

# Checking structure
str(meta)

# Count of Patients by Gender
gender_counts <- meta %>%
group_by(gender) %>%
summarise(count = n())
print(gender_counts)

# Plot Age distribution
ggplot(meta, aes(x = age_at_diag)) +
geom_histogram(binwidth = 1, fill = "blue", color = "black") +
labs(x = "Age at Diagnosis", y = "Count", title = "Age at Diagnosis Distribution")

# Count of Patients by Diagnosis_simple (Tumor type)
diagnosis_simple_counts <- meta %>%
group_by(Diagnosis_simple) %>%
summarise(count = n())
print(diagnosis_simple_counts)

# Boxplot of Age at Diagnosis by Diagnosis_simple
ggplot(meta, aes(x = Diagnosis_simple, y = age_at_diag)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = 90)) +
  labs(x = "Tumor Type (Diagnosis Simple)", y = "Age at Diagnosis", title = "Age at
Diagnosis by Tumor Type")

# Count of Patients by Gender and Diagnosis_simple
gender_diagnosis_counts <- meta %>%
  group_by(gender, Diagnosis_simple) %>%
  summarise(count = n())
print(gender_diagnosis_counts)
```

```
# Boxplot of Age at Diagnosis by Tumor Type with Individual Data Points
ggplot(meta, aes(x = Diagnosis_simple, y = age_at_diag, fill = Diagnosis_simple)) +
  geom_boxplot(alpha = 0.7) +
  geom_jitter(width = 0.3, size = 1, alpha = 0.5) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(x = "Tumor Type", y = "Age at Diagnosis", fill = "Tumor Type") +
  ggtitle("Age at Diagnosis by Tumor Type")

# Boxplot of Age at Diagnosis by Tumor Type
ggplot(meta, aes(x = Diagnosis_simple, y = age_at_diag, fill = Diagnosis_simple)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(x = "Tumor Type", y = "Age at Diagnosis", fill = "Tumor Type") +
  ggtitle("Age at Diagnosis by Tumor Type")

####################################
# 2. Exploratory analysis (survival)
####################################

# Death Percentage
death_percentage <- sum(meta$event_occur == 1) / nrow(meta) * 100
print(paste0("Percentage of children who have died: ", round(death_percentage, 2), "%"))

# Death Probability per tumor type
death_probability_by_tumor_type <- meta %>%
  group_by(Diagnosis_simple) %>%
  summarise(death_probability = sum(event_occur == 1) / n() * 100)
print(death_probability_by_tumor_type)

# Visualisation
ggplot(death_probability_by_tumor_type, aes(x = reorder(Diagnosis_simple,
-death_probability), y = death_probability, fill = Diagnosis_simple)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  xlab("Tumor Type") +
  ylab("Death Probability (%)") +
  labs(fill = "Tumor Type") +
  ggtitle("Probability of Death by Tumor Type")

# Mean Survival Time by tumor type
mean_survival_by_tumor_type <- meta %>%
  group_by(Diagnosis_simple) %>%
  summarise(mean_survival = mean(Time_after_diagnosis))
print(mean_survival_by_tumor_type)

# Bar chart for Mean Survival Time by Tumor Type
ggplot(mean_survival_by_tumor_type, aes(x = Diagnosis_simple, y = mean_survival, fill =
Diagnosis_simple)) +
  geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(x = "Tumor Type", y = "Mean Survival Time", fill = "Tumor Type") +
  ggtitle("Mean Survival Time by Tumor Type")

# Kaplan Meier for simple Diagnosis
meta$Diagnosis_simple <- as.factor(meta$Diagnosis_simple)
fit <- survfit(surv_object ~ meta$Diagnosis_simple)
color_palette <- c("red", "blue", "green", "purple", "orange", "black")
plot(fit, col = color_palette, lty = 1,
     xlab = "Time After Diagnosis (years)", ylab = "Survival Probability",
     main = "Survival Curves by Tumor Type", xlim = c(0, 10))
legend("bottomright", legend = levels(meta$Diagnosis_simple),
       col = color_palette, lty = 1, cex = 0.6)

# Calculate log-rank
logrank_test <- survdiff(surv_object ~ meta$Diagnosis_simple)
```

```
print(logrank_test)

##########################################
# 3. DESeq2: Differentially Expressed Genes
##########################################

# DESeq2: DESeqDataSetFromMatrix
dds <- DESeqDataSetFromMatrix(
  countData = count_data,
  colData = meta,
  design = ~ diagnosis
)

# Normalization
dds_normal <- estimateSizeFactors(dds)

# Check top 5 samples
sizeFactors(dds_normal)[1:5]

# Extracting Normalized counts
count_normal <- counts(dds_normal, normalized = TRUE)
count_normal[1:5, 1:5]

# Unsupervised Learning - Clustering using PCA
vsd <- vst(dds_normal, blind = TRUE)
vsd_mat <- assay(vsd)
vsd_immune_mat <- vsd_mat[rownames(vsd_mat) %in% immune_genes$ID, ]
vsd_immune_cor <- cor(vsd_immune_mat)

# Plot correlation heatmap
library(ComplexHeatmap)
library(circlize)
png("figures/cor.png")
p <- Heatmap(
  vsd_immune_cor,
  name = "cor",
  show_column_names = FALSE,
  show_row_names = FALSE,
  cluster_rows = TRUE,
  cluster_columns = TRUE,
  top_annotation = HeatmapAnnotation(diagnosis = meta$diagnosis),
  column_title_rot = 90
)
draw(p, heatmap_legend_side = "right")
dev.off()

# Set reference level - Craniopharyngioma
dds$diagnosis <- relevel(dds$diagnosis, ref = "Craniopharyngioma")

# PCA Analysis - Plot PCA
plotPCA(vsd, intgroup = "diagnosis")

# DESeq Core Analysis
dds <- DESeq(dds)

# Getting the normalized counts
norm_counts <- counts(dds, normalized = TRUE)
norm_counts[1:5, 1:5]

# Checking the distribution of count data
png("figures/count_disribution.png")
p <- ggplot(count_data) +
  geom_histogram(aes(x = log2(sample_63 + 1)), stat = "bin", bins = 30) +
  xlab("expression counts") +
  ylab("Number of genes")
p
dev.off()
```

```
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)
png("figures/mean_var.png")
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")
p
dev.off()

# Filter the data for "Craniopharyngioma" and "Medulloblastoma" tumor types
filtered_data <- plotCounts(dds, gene = "ENSG00000188290.10", intgroup = "diagnosis",
returnData = TRUE) %>%
  filter(diagnosis %in% c("Craniopharyngioma", "Medulloblastoma"))

# Create a density plot
ggplot(filtered_data, aes(count)) +
  geom_density(aes(fill = diagnosis), alpha = 0.5) +
  ggtitle("Gene HES4") +
  labs(fill = "Diagnosis")

# Define contrasts
diagnosis <- c("Craniopharyngioma", "ATRT", "Ependymoma", "Glioblastoma", "Glioma",
"Medulloblastoma")
contrasts <- lapply(setdiff(diagnosis, "Craniopharyngioma"), function(x) c("diagnosis", x,
"Craniopharyngioma"))

# Perform differential expression analysis for each contrast
results.list <- lapply(contrasts, function(con) results(dds, contrast = con, alpha = 0.05))

# Add gene names to the results
ATRT_res <- results.list[[1]]
ATRT_res$geneName <- gene$GeneName[match(rownames(ATRT_res), gene$ID)]
Ependymoma_res <- results.list[[2]]
Ependymoma_res$geneName <- gene$GeneName[match(rownames(Ependymoma_res), gene$ID)]
Glioblastoma_res <- results.list[[3]]
Glioblastoma_res$geneName <- gene$GeneName[match(rownames(Glioblastoma_res), gene$ID)]
Glioma_res$geneName <- gene$GeneName[match(rownames(Glioma_res), gene$ID)]
Medulloblastoma_res <- results.list[[5]]
Medulloblastoma_res$geneName <- gene$GeneName[match(rownames(Medulloblastoma_res),
gene$ID)]

# Filter immune related genes for each contrast
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,]

# Summary results
summary(ATRT_immune_res)
summary(Ependymoma_immune_res)
summary(Glioblastoma_immune_res)
summary(Glioma_immune_res)
summary(Medulloblastoma_immune_res)

# Filter based on padj < 0.05 and abs(log2FoldChange) > 2
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)

# Order results based on padj
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 <- MedulThere was an interruption in the code. Here's the
continuation:

```R
# loblastoma_immune_res_filter[order(Medulloblastoma_immune_res_filter$padj),]

ATRT_immune_res_filter
Ependymoma_immune_res_filter
Glioblastoma_immune_res_filter
Glioma_immune_res_filter
Medulloblastoma_immune_res_filter

##########################################
# 4. Immunosignatures and Survival Analysis
##########################################

# Get the top upregulated gene for Medulloblastoma
top_upregulated_med <-
Medulloblastoma_immune_res_filter[which.max(Medulloblastoma_immune_res_filter$log2FoldChang
e), ]
# Get the top downregulated gene for Medulloblastoma
top_downregulated_med <-
Medulloblastoma_immune_res_filter[which.min(Medulloblastoma_immune_res_filter$log2FoldChang
e), ]

# Retrieve expression levels for specified genes
magea3_exp_samples <- norm_counts["ENSG00000221867.9", ]
col17a1_exp_samples <- norm_counts["ENSG00000065618.21", ]

# Convert row names to a separate column
meta$Sample_ID <- rownames(meta)

# Add this information to your metadata dataframe.
meta$magea3_exp <- magea3_exp_samples[meta$Sample_ID]
meta$col17a1_exp <- col17a1_exp_samples[meta$Sample_ID]

# Create new binary variables based on the median expression of each gene
meta$magea3_high <- ifelse(meta$magea3_exp > median(meta$magea3_exp, na.rm = TRUE), 1, 0)
meta$col17a1_high <- ifelse(meta$col17a1_exp > median(meta$col17a1_exp, na.rm = TRUE), 1,
0)

# Replace NA values with the maximum value of the column
meta$Time_after_diagnosis[is.na(meta$Time_after_diagnosis)] <-
max(meta$Time_after_diagnosis, na.rm = TRUE)
print(meta$Time_after_diagnosis)

# Load the required libraries
library(survival)
library(survminer)

# Assume that meta$status is the status variable: 1 for event/death, 0 for censored
# And Time_after_diagnosis is the survival time
fit_magea3 <- survfit(Surv(Time_after_diagnosis, status) ~ magea3_high, data = meta)
fit_col17a1 <- survfit(Surv(Time_after_diagnosis, status) ~ col17a1_high, data = meta)

# Plot the survival curves
ggsurvplot(fit_magea3, data = meta, risk.table = TRUE, pval = TRUE, conf.int = TRUE,
```
```

```
           legend.title = "Magea3 Expression", legend.labs = c("Low", "High"),
           title = "Kaplan-Meier Survival Curve Based on Magea3 Expression")

ggsurvplot(fit_col17a1, data = meta, risk.table = TRUE, pval = TRUE, conf.int = TRUE,
           legend.title = "COL17A1 Expression", legend.labs = c("Low", "High"),
           title = "Kaplan-Meier Survival Curve Based on COL17A1 Expression")

# Plot the survival curves without risk table
ggsurvplot(fit_magea3, data = meta, risk.table = FALSE, pval = TRUE, conf.int = TRUE,
           legend.title = "Magea3 Expression", legend.labs = c("Low", "High"),
           title = "Kaplan-Meier Survival Curve Based on Magea3 Expression")

ggsurvplot(fit_col17a1, data = meta, risk.table = FALSE, pval = TRUE, conf.int = TRUE,
           legend.title = "COL17A1 Expression", legend.labs = c("Low", "High"),
           title = "Kaplan-Meier Survival Curve Based on COL17A1 Expression")

# Subset data for only Medulloblastoma patients
medulloblastoma_meta <- meta[meta$diagnosis == "Medulloblastoma",]

# Assume that meta$status is the status variable: 1 for event/death, 0 for censored
# And Time_after_diagnosis is the survival time
fit_magea3 <- survfit(Surv(Time_after_diagnosis, status) ~ magea3_high, data =
medulloblastoma_meta)
fit_col17a1 <- survfit(Surv(Time_after_diagnosis, status) ~ col17a1_high, data =
medulloblastoma_meta)

# Plot the survival curves with risk table at the top
ggsurvplot(fit_magea3, data = medulloblastoma_meta, risk.table = TRUE, risk.table.y.text =
FALSE,
           risk.table.title = "Number at risk by time",
           pval = TRUE, conf.int = TRUE, legend.title = "Magea3 Expression",
           legend.labs = c("Low", "High"), risk.table.height = 0.25,
           title = "Survival Curve for Medulloblastoma Based on Magea3 Expression",
           tables.theme = theme_cleantable(), font.main = c(12, "bold", "darkblue"),
           font.submain = c(15, "bold", "darkblue"),
           font.caption = c(14, "plain", "darkblue"),
           font.xaxis = c(14, "bold", "darkblue"),
           font.yaxis = c(14, "bold", "darkblue"),
           font.tickslab = c(12, "plain", "darkblue"),
           font.legend = c(14, "plain", "darkblue"))

ggsurvplot(fit_col17a1, data = medulloblastoma_meta, risk.table = TRUE, risk.table.y.text =
FALSE,
           risk.table.title = "Number at risk by time",
           pval = TRUE, conf.int = TRUE, legend.title = "COL17A1 Expression",
           legend.labs = c("Low", "High"), risk.table.height = 0.25,
           title = "Survival Curve for Medulloblastoma Based on COL17A1 Expression",
           tables.theme = theme_cleantable(), font.main = c(12, "bold", "darkblue"),
           font.submain = c(15, "bold", "darkblue"),
           font.caption = c(14, "plain", "darkblue"),
           font.xaxis = c(14, "bold", "darkblue"),
           font.yaxis = c(14, "bold", "darkblue"),
           font.tickslab = c(12, "plain", "darkblue"),
           font.legend = c(14, "plain", "darkblue"))

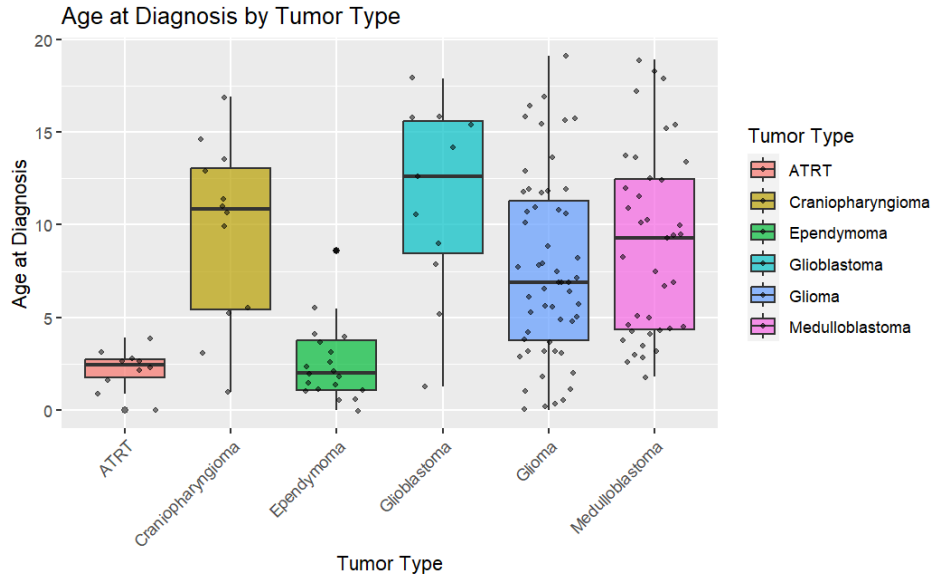
##########################################
# 5. The END
##########################################
```

Appendix C. Additional Information

Figure 9 is a boxplot mapping the age at diagnosis for each tumor type. This boxplot offers a visual representation of the minimum, first quartile, median, third quartile, and maximum age at diagnosis for each tumor type, effectively illustrating the range and distribution of ages at diagnosis.

Figure 9

Age at Diagnosis by Tumor Type



In addition to this, we constructed *Figure 10*, which is a graph showing the distribution of different tumor types within each age category. This second figure provides a clear picture of how the prevalence of each tumor type changes across different age groups. Together, these figures offer comprehensive insights into the relationship between age and tumor type.

Figure 10

Distribution of Tumor Types per Age Category