Single cell RNA sequencing analysis of GFAP alternative splice isoforms in glioblastoma

Joanna Sergieva
Bioinformatics and Biocomplexity
Graduate School of Life Sciences
Utrecht University
9688307
j.n.sergieva@students.uu.nl

Supervised by
Dr. Onur Basak
Department of Translational Neuroscience
University Medical Center Utrecht

Second Reviewer
Prof.Dr. Elly Hol
Department of Translational Neuroscience
University Medical Center Utrecht

Layman’s Summary

Glioblastoma multiforme, otherwise known as grade IV glioma, is one of the most aggressive brain tumors and has no cure. Treatment aims to slow tumor progression, but patient prognosis is poor due to the highly invasive characteristics of the tumor. The cells invade and infiltrate nearby tissue, which causes the tumor to come back after its surgical removal. Another component of glioblastoma malignancy is a group of quiescent neural stem cells. These cells are not targeted by treatment since they aren’t actively dividing. However, their nondividing state is reversible and the cells are able to switch into an active state of replication to propagate the tumor. Targeting and finding these quiescent stem cells could help target tumor invasion. More insight into the cause behind glioblastoma malignancy is necessary in order to provide better treatment options for patients.

Intermediate filaments are key components of the structure of a cell and are known to play vital roles in the malignancy of certain tumors. The key intermediate filaments in glioblastoma are Vimentin, Nestin, Synemin, and Glial Fibrillary Acidic Protein or GFAP. There are several splice variants of GFAP, the most highly expressed in humans are GFAP isoforms α and δ. The GFAP gene is highly regulated in alternative splicing and differences in isoform expression can be associated with certain tumor malignancies. The ratio between isoforms GFAPδ/α was investigated to be higher in glioblastoma compared with lower grade gliomas. Therefore, GFAP isoform expression could play a role in the malignancy attributed to glioblastoma multiforme.

With the use of a bioinformatics pipeline, GFAP and other intermediate filament protein isoforms were analyzed in single cell RNA sequencing glioma datasets. The results indicated that GFAPα and GFAPδ expression is very heterogenous between tumors. However, on average GFAPα and δ expression is higher in quiescent cells compared to dividing cells. Further investigation of separate tumors and quiescent groups is necessary for the GFAP gene and its isoforms. This study generated an isoform analysis on a large single cell glioma dataset which can be reanalyzed for correlation studies in future research.

Abstract

Glioblastoma multiforme is a very aggressive brain tumor with extremely poor patient prognosis. Clinical treatment fails to fully treat glioblastoma due to its intratumoral heterogeneity and population of quiescent stem cells propagating tumor recurrence. The intermediate filament protein GFAP is highly expressed in brain matter. The ratio of two prominent splice isoforms of GFAP, α and δ, are implicated in glioma tumor grade and migration. In this study, the ratio of these two isoforms was investigated in glioblastoma using published single cell RNA SMART-seq data. Two very large datasets were investigated in this study, one with grade III and IV gliomas consisting of 14,517 cells and 41 tumors. The second dataset investigated was on pediatric ependymoma with 5,900 cells. To analyze isoform level information, the datasets were mapped with the pseudoalignment program Kallisto and then analyzed with the Seurat single cell analysis package. GFAP expression was found to be highly heterogenous between tumors and cell clusters, however on average there was no difference between ratio of GFAPδ/α between glioma grade. This led to the suggestion that cells are characterized based on GFAP isoform expression profile. When comparing dividing and nondividing cells, GFAP was found to be highly expressed in neural quiescent stem cells compared to actively dividing cells. However, invasive score was not found to be correlated with GFAPα and δ expression or ratio. Through the use of a bioinformatics pipeline a large glioma dataset was analyzed for intermediate filament expression and scored for invasion. This dataset can be further investigated for other intermediate filaments and any isoform switches between grade III to IV glioma. This analysis is highly beneficial to discern the relationship between isoform switches and the heterogeneity of glioblastoma multiforme at the single cell level.

Keywords: Glioblastoma multiforme, GFAP, intermediate filaments, single cell RNA sequencing

Introduction

Glioblastoma multiforme (GBM), also known as grade IV astrocytoma or grade IV glioma, is the most common, lethal, and aggressive tumor of the central nervous system (Uceda-Castro et al., 2022). GBM is highly prevalent with an incidence rate of 3 per 100,000 people and is 40% more common in men than women. Despite the efforts of treatment, the survival prognosis of glioblastoma is 42.4% reaching 6 months, 17.7% at 1 year, and 5% at 5 years (Delgado-López et al., 2016). Average patient survival reaches 14-15 months. While glioblastoma remains currently incurable, standard care treats GBM by conjugate use of radiation and chemotherapy following a surgical tumor resection. Adjuvant chemotherapy and radiation prolong patient survival but do not stop the invasive nature of glioblastoma cells that migrate and infiltrate into healthy tissue (Jackson et al., 2019). Treatment is able to prolong survival but does not combat invasiveness as almost all GBM cases result in tumor resurgence after resection (Uceda-Castro et al., 2022). The reason for poor survival prognostics is that glioblastoma eludes targeted therapies due to intratumoral heterogeneity and molecular plasticity (Jackson et al., 2019). Due to this heterogeneity, tumor cells are able to evade targeted therapies and continue to invade tissue (van Bodegraven et al., 2019). A better understanding of the underlying molecular mechanisms of this tumor is necessary in order to improve treatment strategies and patient prognosis.

Quiescence is a state of cell cycle growth arrest where cells enter a dormant state and stop replicating. In the context of glioblastoma, quiescent cells are a subset of tumor cells able to evade treatment. This is supported by the stem cell theory of cancer, the resistance of cancer to chemotherapy and radiotherapy due to resident tumor stem cells (Mukherjee et al., 2020). These cells evade treatment and are the cause of tumor recurrence. Tumor stem cells are able to survive treatment since current treatment is cytotoxic, relying on targeting actively replicating cells. Since these quiescent stem cells are in a dormant state, treatment only targets proliferating tumor cells and quiescent cells are able to proliferate into a recurrent tumor (Wang et al., 2018). Due to their drug resistant nature, quiescent stem cells in glioblastoma are of interest for potential therapeutic targets.

Intermediate filaments (IFs) are key components of the cytoskeletal network of a cell that are essential for cell signaling (Stassen et al., 2017). Alongside actin microfilaments and microtubules, IF's are critical for supporting cell shape, adhesions, and remain indispensable for cellular functions. The exact role of intermediate filaments in cell function and tissue integrity is still a mystery. This is due to the fact that unlike other cytoskeletal components, IF protein expression varies between cells and tissues and can represent between 0.3% - 80% of all protein expressed in the cell (van Bodegraven et al., 2021). A change in IF network composition is known to lead to a shift towards a more malignant tumor profile. This finding is exemplified in keratin 14, essential for the metastasizing invasive front of breast cancer, keratin 17, important in transforming Ewing carcinoma, and vimentin, regulating lung cancer cell adhesions in epithelial mesenchymal transformations (Stassen et al., 2017). As IF proteins have a broad involvement that is not fully understood yet, it is important to study them as a possible therapeutic target for tumor malignancy. In regards to glioblastoma multiforme, the IF proteins involved in glial cells are glial fibrillary acidic protein or GFAP, Vimentin, Synemin, and Nestin (van Bodegraven et al., 2021).

Glial fibrillary acidic protein (GFAP) is an astrocyte marker and is a protein of particular interest due to recent studies implicating relation between two of its splice variants or isoforms and an increase in tumor malignancy (Stassen et al., 2017). There are 10 known isoforms of GFAP, the most highly expressed and studied being GFAPα and GFAPδ (Uceda-Castro et al., 2022). These two isoforms have differing protein interactions and expression patterns, for instance GFAPα is prevalent in mature astrocytes and is highly upregulated in reactive gliosis. GFAPδ is enriched in human neural stem cells and subpial astrocytes of the brain (Stassen et al., 2017). Functionally, GFAPα can self-assemble into a filamentous network while only expression of GFAPδ leads to perinuclear aggregates (Uceda-Castro et al., 2022). Recent studies have discovered a higher ratio of GFAPδ/α in glioblastoma compared to grades II and III glioma (Radu et al., 2022). These GFAP expression patterns have been investigated in bulk RNA sequencing data of 528 high grade and 516 low grade glioma data from the Cancer Genome Atlas (CGA). This investigation found GFAPα levels to be significantly lower in grade IV compared with grades II and III glioma while GFAPδ levels remained the same (van Asperen et al., 2022). This shift in expression causes an observed increase in ratio of GFAPδ/α.

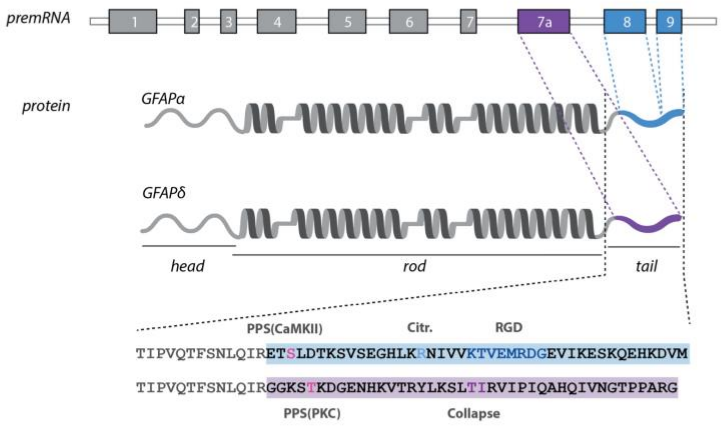
premRNA
1
2
3
4
5
6
7
7a
8
9
protein
GFAPα
GFAPδ
head
rod
tail
PPS(CaMKII)
Citr.
RGD
TIPVQTFSNLQIRETSLDTKSVSEGHLKRNIVVKTVEMRDGEVIKESKQEHKDVM
TIPVQTFSNLQIRGGKSTKDGENHKVTRYLKSLTIRVIPIQAHQIVNGTPPARG
PPS(PKC)
Collapse

Figure 1 |Depiction of molecular components of GFAPα and GFAPδ. The distinct tail regions that differ in both isoforms and cause their functional diversity. (Figure 3 of van Asperen et al. 2022)

The increase in GFAPδ/α ratio in glioblastoma, although confirmed experimentally as well as in bulk RNA sequencing, has not yet been confirmed in single cell RNA sequencing (scRNA-seq) data. Tumor heterogeneity plays a huge role in cancer progression and evasion of targeted therapies. Bulk RNA sequencing is not able to represent this heterogeneity and as a result miss vital cell subpopulations carrying important tumor information. Single cell RNA sequencing is able to overcome the limits of traditional RNA sequencing technologies by sequencing each cell of the tissue individually (Zhang et al., 2021). SMART-seq is a scRNA-seq method of interest for studying isoform level information due to its long reads, high sensitivity, and high accuracy (Picelli et al., 2014). In the case of glioblastoma, scRNA-seq can identify different tumor groups therefore unravelling the heterogeneity of the tumor. Previous papers investigating glioblastoma tumors with scRNA-seq have investigated different tumor subtypes based on molecular signature, IDH mutation, and TME composition (Venteicher et al., 2017). The paper from Couturier et al 2020 also looked at glioblastoma tumor heterogeneity using scRNA-seq to study lineage hierarchy and progenitor cells with RNA velocity. They found that progenitor cells are key targets for therapeutic opportunities and are the common originator of cell hierarchy. Studies of glioblastoma using scRNA-seq data have looked into gene level information, but have not investigated alternative splicing and isoforms of genes. Alternative splicing is known to play a significant role in development of malignancies and tumor progression (Dvinge et al., 2016). Cancer cells are able to generate advantageous splice variants at the cost of reducing efficiency of the splicing process, therefore becoming vulnerable to splicing perturbations (Bonnal et al., 2020). Investigating isoform level information using scRNA-seq would bring significant information on transcript variation between cells, cell types, tumors, and tumor types.

The purpose of this study is to analyze isoform level information of grades III and IV glioma cells for GFAP expression based on an established single cell RNA sequencing analysis pipeline. Single cell SMART-seq datasets will be analyzed with Kallisto pseudoalignment and the Seurat package in R (Bray et al. 2016 and Satija et al. 2015). Quiescence and invasion will be investigated for the intermediate filament GFAP (O'Conner et al., 2021 and Yu et al., 2020). Splice isoforms haven't been investigated in single cell glioblastoma, therefore the results of this study will be informative.

Once the pipeline is complete, the following questions will be investigated: 1) What are the expression levels of intermediate filament isoforms of Vimentin, Nestin, Synemin, and GFAP in glioblastoma? 2) Is there a difference in GFAPδ/α ratio between grade III and grade IV glioma in single cell data? 3) What are GFAPα and GFAPδ expression patterns like in quiescent compared with actively dividing cells? 4) Is there an association between GFAP expression and invasion in glioma?

Methods

Data description

For the purpose of this study, the dataset from Venteicher 2017 was selected as most appropriate due to the large number of SMART-seq cells. The paper described 14,223 cells but after acquisition the actual cell count is 21,200 cells. These cells come from 41 different patients. These tumors represent grades III and IV glioma, IDH wildtype and IDH mutant, both pediatric and adult samples. These samples were freshly collected after tumor resection from patients of the Massachusetts general hospital. The samples were FAC sorted then sequenced using SMART-seq. Further details of experimental procedures can be found in the paper Venteicher 2017 (Venteicher et al., 2017).

An additional dataset from the paper Gojo 2020 was also selected for analysis. This paper analyzed 28 primary pediatric Ependymoma's using single cell and single nuclear SMART-seq2. These tumors were both diagnostic and recurring. They were freshly collected after surgical removal. Further details about sample collection and experimental procedure can be found in the paper of Gojo 2020 (Gojo et al., 2020).

Paper	**description**	**cells**	**method**	**GEO/DUOS**
Venteicher 2017	Glioblastoma and grade III glioma	21,300	SMART-seq	GSE89567 DUOS-000109 DUOS-000117
Gojo 2020	Ependymoma	7,380	SMART-seq	GSE141460 DUOS-000120

Table 1 | Number of cells per dataset and repository access codes

DUOS data access

In order to access the datasets of the paper Venteicher 2017 and Gojo 2020, we were redirected to the data sharing platform DUOS (https://duos.broadinstitute.org/) due to IRB requirements of the respective institutions. (DUOS-000109, DUOS-000117, DUOS-000120)

Account creation

The first step in the acquisition process was to create an institutional google account jointly with a library card. In order to do so, the institution must be registered in DUOS, the institutional signing official has logged into the website and has issued the permissions to the Google Identity used to apply for access. When these steps are completed, the signing official logged into DUOS, then emailed duos-support@broadinstitute.zendesk.com with the name of their institution, the signing official's name and email, and the google identity used to register in DUOS.

Application

Once the DUOS account is set up, an access request for each data set was made. The request consisted of 4 parts. The first section where the researcher identified themselves and and their collaborators with an eRA commons ID. This was followed by a data access request specifying the type of research that would be conducted with access to this dataset. A research use statement was submitted in scientific and common language to fulfill this part. The next section included questions about the extent of the project, whether it is limited to one gender, involves researching illegal behaviors, or includes the study of psychological behaviors. The final section was a research agreement following the national institute of health (NIH) genomic data sharing policy. Once these sections had been completed, the data access request was sent. After 2 months of submission, the request was approved by the committee.

Transfer of Data

After receiving the approval for the data access request, the data custodians of each retrospective dataset were contacted. They opened access for each dataset on the data repository TERRA (https://terra.bio). An institutional account was created on TERRA to access to each workspace. TERRA stores the datasets on the cloud but is also a space for running scripts and performing analysis. Due to unspecified costs of using the service, we downloaded all fastqs of each dataset on the internal servers of the University Medical Center Utrecht (UMCU) – the High-Powered Computing (HPC) network of the Utrecht bioinformatics center.

Kallisto pseudoalignment

Once the raw fastq files were downloaded onto the HPC, the process of aligning reads to a reference transcript took place. We selected the pseudoalignment program Kallisto for this due to its high accuracy and speed compared with other alignment techniques (Bray et al., 2016). For the Kallisto index we used Homo sapiens GRCh38 cDNA fasta file from Ensemble version 104 (https://www.ensembl.org/info/data/ftp/index.html) and the 'Kallisto index' function. The Ensemble index file was further manipulated for our purposed by manipulating it in Python so that for the GFAP gene only GFAP isoforms α and δ were included. In previous analyses from Alexandra de Reus, the most prominent GFAP isoform found was GFAP-201, an undocumented isoform of GFAP (de Reus, 2021). We noticed that the 3' UTR of GFAPα was longer than any other isoform, and no reads mapped to this region. We considered this an artefact, which led to the decision to exclude all other isoforms except GFAPα and GFAPδ. This filtered index improves sensitivity of the analysis at the expense of specificity.

In order to pseudoalign SMART-seq data, the function quant() was used from Kallisto. Originally applicable for bulk RNA-seq data, Kallisto quant can still be applied to SMART-seq data due to its full length reads and lack of UMI. Each cell is analyzed separately thus three output files were created for each cell. The abundance.tsv files were merged for all the cells with R and created a transcript-count file.

Making Metadata

For Venteicher 2017, there were three metadata files needed to be compiled into one. The first metadata file came from the TERRA repository, there the Venteicher data was split into two workspaces titled Regev-Suva-Tirosh and GBM2. Regev-Suva-Tirosh contained 10 tumor samples while GBM2 contained 31 tumor samples. This TERRA metadata contained information of patient data such as age, IDH type, and grade for each tumor. Once compiled for each cell, the next metadata from the authors of Venteicher was added. This metadata contained labels for malignant or non-malignant microglia/oligodendrocyte cells with tumor index scores. However, these labels were only for 5,788 out of 21,300 cells. The final metadata added was quiescence scores for 3,010 cells. These scores come from the authors of a neural G0 quiescence scoring method, which they performed on some cells of Venteicher 2017 (O'Conner et al., 2021). The Gojo 2020 dataset also had clinical metadata, however after pseudoalignment in Kallisto the cell names were lost. Therefore, this dataset only had a cluster based downstream analysis.

Single cell analysis with Seurat

The Seurat package in R was used to perform quality control metrics and analyze the single cell data (Satija et al., 2015). The initial step of analyzing Venteicher 2017 was to combine the transcript-count files of Regev-Suva-Tirosh and GBM2. Once combined, the cell count of the Venteicher 2017 dataset was 21,300 cells. With this transcript-count table a Seurat object was created using the made metadata. Quality control was performed to remove cells with high mitochondrial RNA, low feature RNA, and high-count RNA. The purpose of doing these steps was to discard low quality cells that could be lysing or apoptotic cells. After quality control the number of cells from the Venteicher dataset was 15,844 cells. After quality control, the Gojo dataset reduced in size from 7,380 cells to 5,900 cells.

The datasets were logarithmically normalized with a scale factor of 10,000 to remove nonbiological variation using the function 'NormalizeData()'. Then 2000 highly variable features used in PCA were identified with the function 'FindVariableFeatures()'. Next, the datasets were linearly transformed using the function 'ScaleData()' in order to give equal weight to genes so that highly expressed genes do not dominate downstream analysis. Dimensionality reduction was performed using Principle Component Analysis (PCA). Based on heatmaps created from PCA, 40 PC's were selected for downstream analysis for the Venteicher dataset and 35 PC's were selected for the Gojo 2020 dataset. Clustering was performed on the data using the K nearest neighbors (KNN) method using Euclidean distance and PCA with the 'FindNeighbors()' function. K was selected using the formula $\frac{\sqrt{n}}{2}$, where n is cell count. The nearest neighbors, k, chosen for Venteicher 2017 was 63 and for Gojo 2020 it was 38. The resolution selected was 1.0 for both datasets in the function 'FindClusters()'. With these parameters, 24 and 20 clusters were created for the Venteicher and Gojo datasets respectively.

Paper	**Cells after QC**	**PC's**	**k**	**resolution**	**clusters**
Venteicher 2017	15,844	40	63	1.0	24
Gojo 2020	5,900	35	38	1.0	20

Table 2 | Cell count and parameters used to make clusters per dataset

Once the Venteicher dataset had undergone quality control, PCA, and UMAP dimensionality reduction, a cell type classification based on labeled metadata could be performed. Cells without labels for malignancy and quiescence could be extrapolated based on annotated metadata. This was achieved by following the Seurat tutorial for cell type classification using an integrated reference (Hoffman, 2022). First the data was split into cells labeled for malignancy, anchor cluster, and cells unlabeled for malignancy, query cluster. Then label transfer was performed using the functions 'FindTransferAnchors()' and 'TransferData()'. The metadata was added then the query and anchor cells were merged. This process was repeated for the quiescence labels. Therefore, the labels of 5,788 cells were transferred for malignancy and 3,010 labels were transferred for quiescence. After the malignancy labels were transferred, the dataset was subset so that only malignant labeled cells were present. This reduced the cell count from 15,844 cells to 14,517 cells.

The final step of the analysis was to calculate an invasive score for both Venteicher 2017 and Gojo 2020 datasets. These scores were calculated per cell using the function 'AddModuleScore()' in Seurat with a list of invasive genes from Yu et al. 2020.

Plotting

With the Seurat object properly labeled and annotated, data visualization took place. To examine the GFAPα and GFAPδ expression across cells, the functions 'DimPlot()', 'VlnPlot()', and 'DotPlot()' from Seurat were used. The GFAPδ/α ratio was calculated using the function 'AverageExpression()', this function calculated the average expression of GFAPα and δ per group. A GFAPδ/α ratio plot was created using the ggplot package in R. This average GFAPδ/α expression was visualized between UMAP clusters, tumors, quiescent or active cells, and grade. Other intermediate filaments were investigated such as Vimentin, Nestin, and Synemin. The functions 'VlnPlot()', 'DotPlot()', and 'DimPlot()' were used to visualize their activity.

Results

Veneticher 2017 et al.

Dataset characterization

Once fully annotated and preprocessed, the Venteicher 2017 dataset was clustered using the dimensionality reduction technique UMAP (Fig. 2). Since the cell count is high, these clusters indicate that the cells cluster in tumor type instead of cell type. There are multiple separate islands per tumor, but there is also a continent in the center with overlapping tumors. The cells in the continent are more similar to each other than to the circling island around, this is clear in the UMAP cell characterization. The cells from each tumor are spread throughout the clusters, indicating validity in the clustering technique.

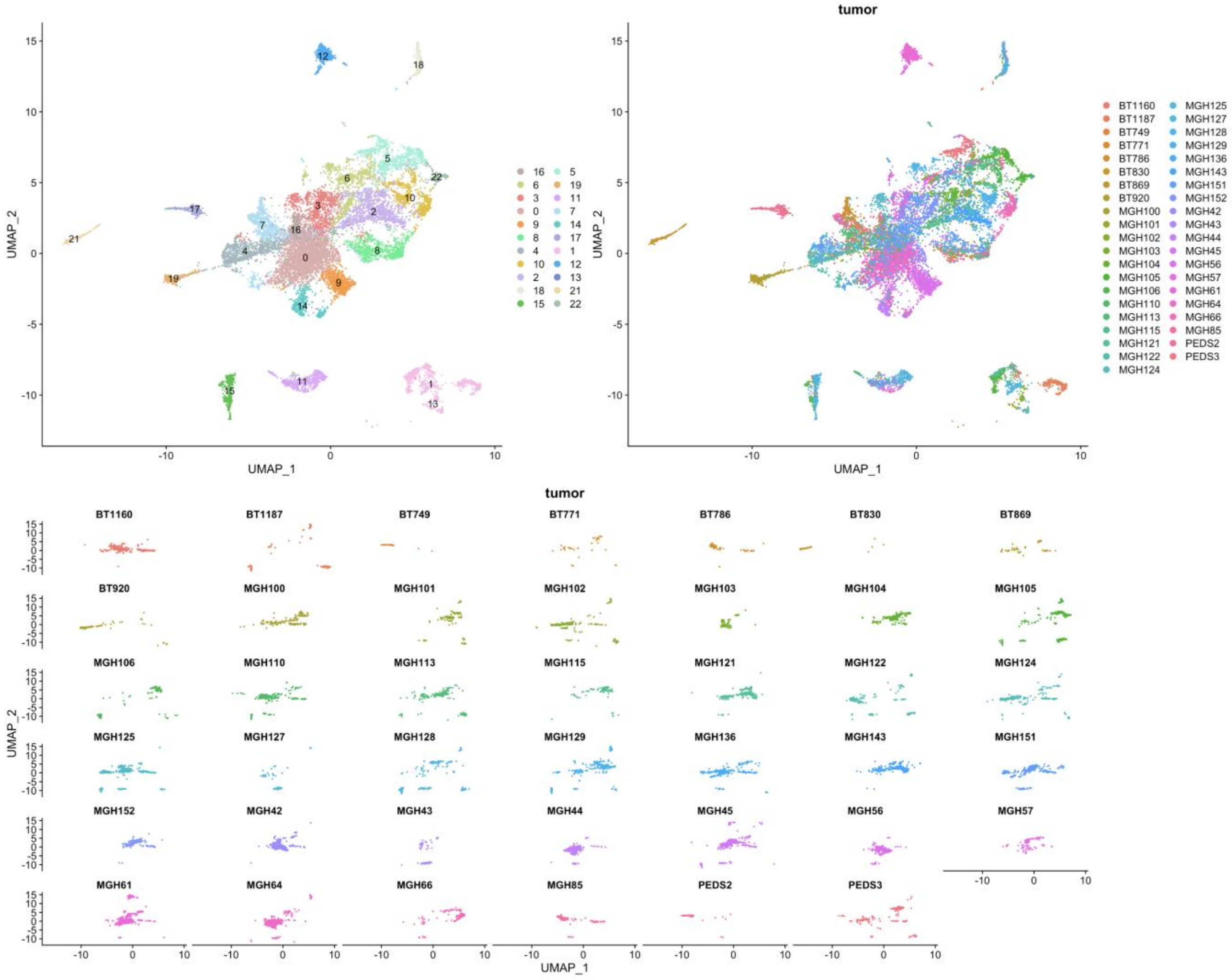

Figure 2 | UMAP dimensionality reduction of Venteicher 2017 labeled per cluster and per tumor. There are 14,517 cells encompassing 23 clusters and 41 tumors.

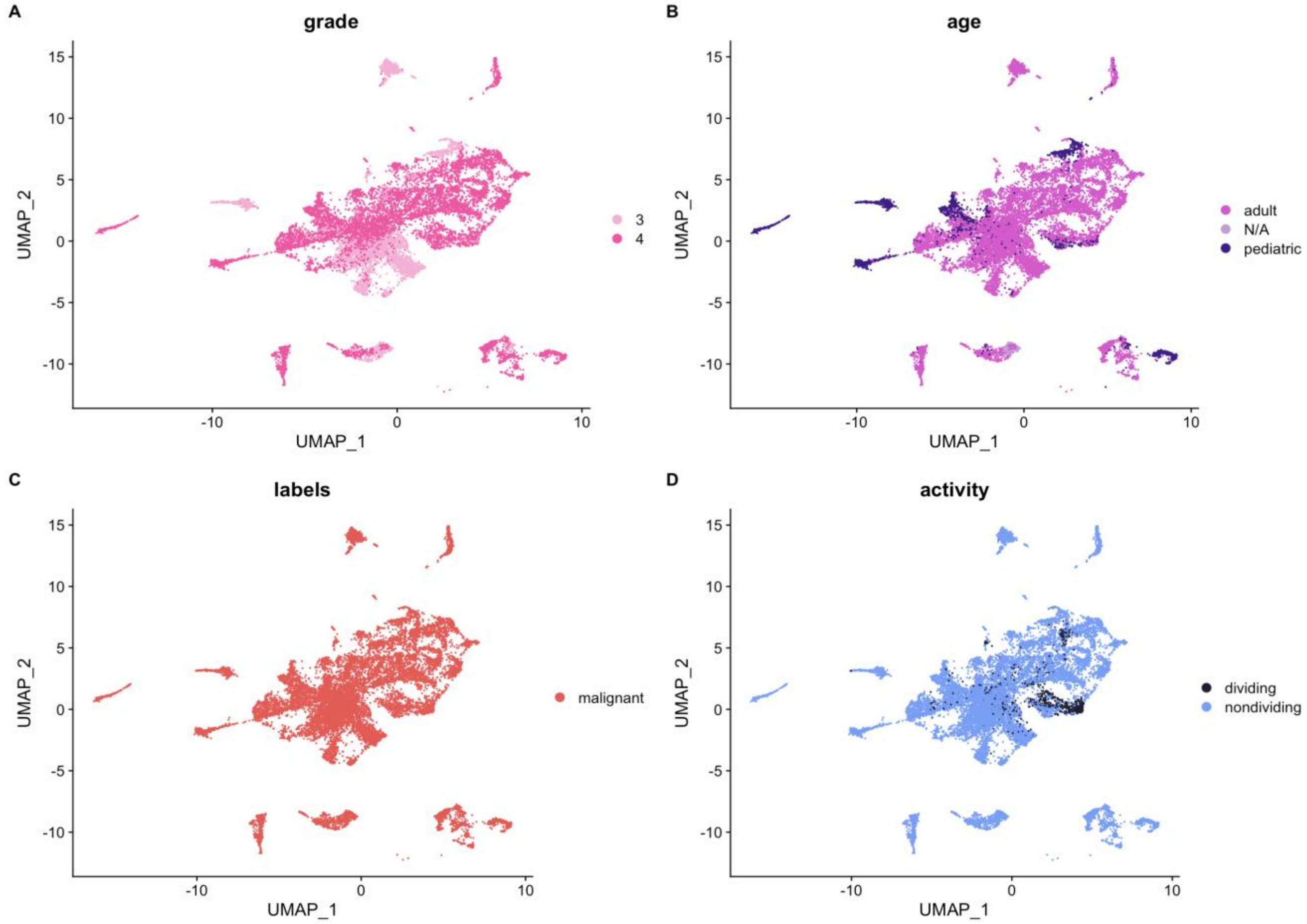

Figure 3 | UMAP of Venteicher 2017 with 14,517 cells labeled for grade, and patient age, malignancy, and activity.

The dataset from Venteicher 2017 et al. was characterized with dimensionality reduction technique UMAP and labeled with clinical metadata and computed metadata in Figure 3. This dataset consisted of grades 3 and grade 4 glioma, otherwise known as glioblastoma. The tumors come from both pediatric and adult samples. Every cell was labeled as a malignant tumor cell. These labels were created from a label transfer of annotated metadata from the authors of the paper. After selecting for malignancy, the cell count decreased from 15,844 cells to 14,517 cells. Activity level information was acquired from a label transfer of metadata from O'Connor et al. 2020. Based on these labels, most tumor cells are labeled as nondividing quiescent cells in the G_0 state. There are dividing cells in most clusters, but there is a concentration of dividing cells in cluster 8. The cells were also labeled for IDH type, but since the labels were not completely annotated, IDH analysis was inconclusive (Fig. 19).

Expression levels of intermediate filament proteins

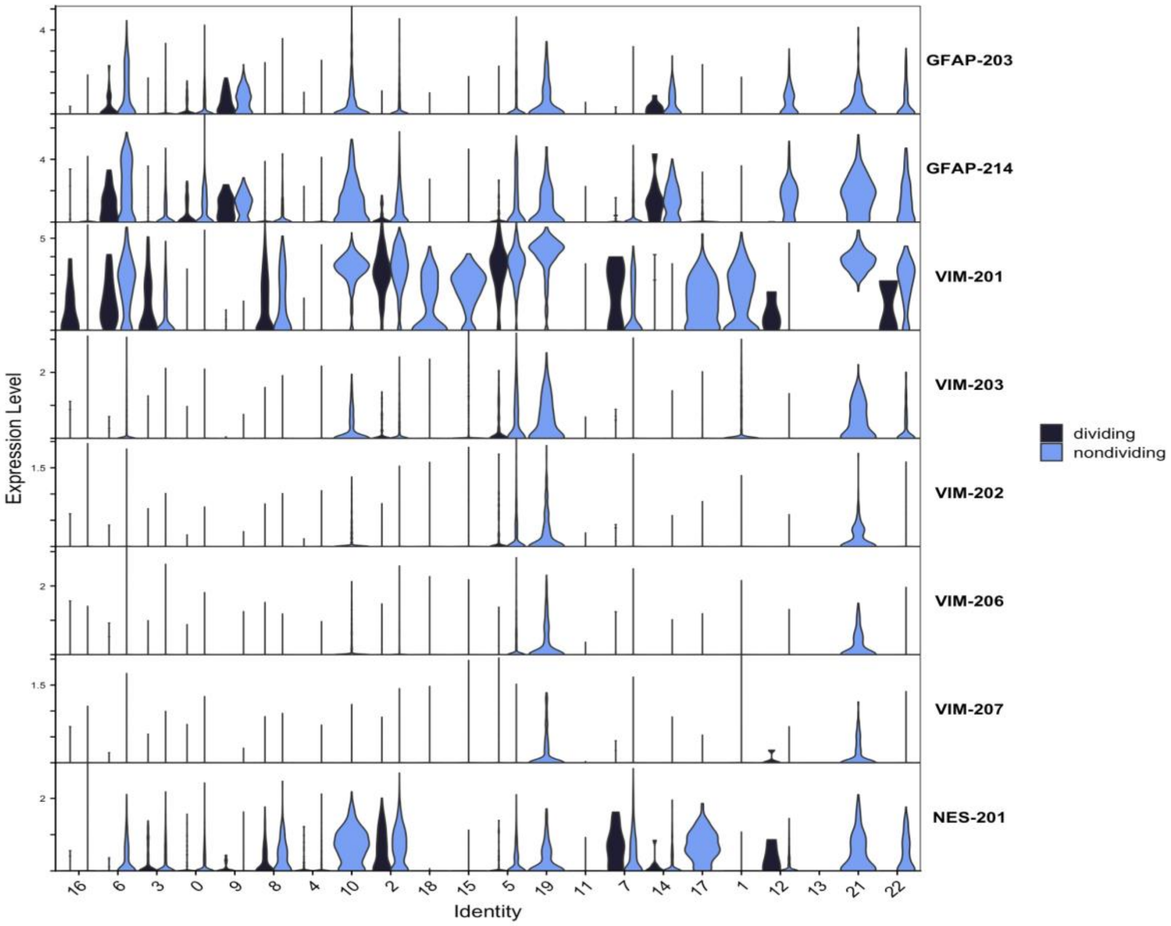

Figure 4 | Violin plot of intermediate filament expression per UMAP cluster. Each cluster is split by activity, with dividing and nondividing groups. The intermediate filaments investigated were GFAP, Vimentin, and Netsin.

Intermediate filaments GFAP, Vimentin, and Nestin were found in detectable levels in this dataset (Fig. 4). The final intermediate filament present in astrocytes, Synemin, was not detected and therefore was excluded in the analysis. There were two GFAP isoforms detected, GFAP-214 which is GFAPα and GFAP-203 which is GFAPδ. This was due to the use of a filtered index in the Kallisto mapping stage of the analysis. GFAP-214 (GFAPα) has higher expression than GFAP-203 (GFAPδ). There were five isoforms of Vimentin detected. Of these isoforms, VIM-201 has the highest expression. In regards to Nestin, only one isoform was detected in this dataset, NES-201. The expression levels of these intermediate filament isoforms were investigated based on dividing and nondividing groups per Seurat cluster in figure 4. There are more nondividing cells than dividing cells in the dataset, therefore some clusters such as cluster 19 only have IF expression in nondividing cells. In most groups, there is higher expression of IF proteins in quiescent cells compared to dividing cells. However, there are certain clusters where similar expression levels are observed between nondividing and dividing cells. This pattern is visible in cluster 2 for Nes-201, clusters 2, 5, 8, and 6 for VIM-201, and clusters 14, 9, and 6 for GFAPα and GFAPδ. In the Nestin and Vimentin plots, but not in GFAP, some clusters even had higher expression in dividing cells than quiescent cells. These clusters were clusters 7 and 12 for NES-201 and clusters 7, 12, and 16 for VIM-201. The levels of intermediate filament expression were also analyzed per tumor in figures 20 and 21.

GFAPα and GFAPδ expression

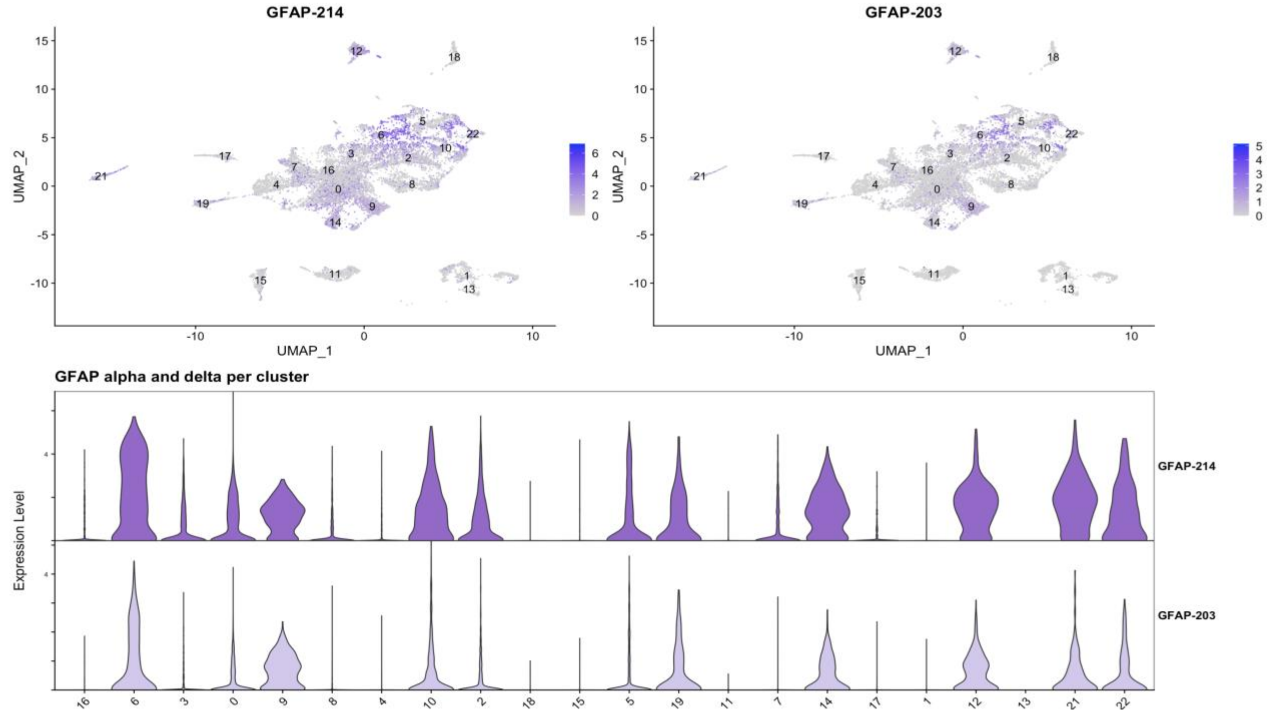

Figure 5 | GFAP-214 (GFAPα) and GFAP-203 (GFAPδ) expression visualized in a UMAP dimensionality reduction. Violin plot visualizing the distribution of GFAPα and GFAPδ expression in clusters generated by UMAP in Seurat analysis

The next part of the analysis was a closer look at intermediate filament GFAP. The expression of GFAPα and δ isoforms was analyzed per UMAP clusters and tumors looking at glioma grade and cell activity in Figure 5. There are UMAP clusters with little to no GFAP expression and clusters with high expression of GFAPα and δ. For the most part, GFAPα is more highly expressed in all clusters, coinciding with the literature. Important to note is that analysis was conducted using a filtered index for only GFAPα and GFAPδ. Therefore, only these two isoforms were investigated. When examining GFAP expression levels per cluster, there are certain clusters with low of no GFAP at all (1,4,7,8,11,15,16,17,18). In the clusters where GFAP is expressed, GFAPα levels are always higher than those of GFAPδ. A notable difference is in clusters 9 and 6, where GFAPδ levels are close to those of GFAPα. Therefore, these two clusters exhibit a high GFAPδ/α ratio.

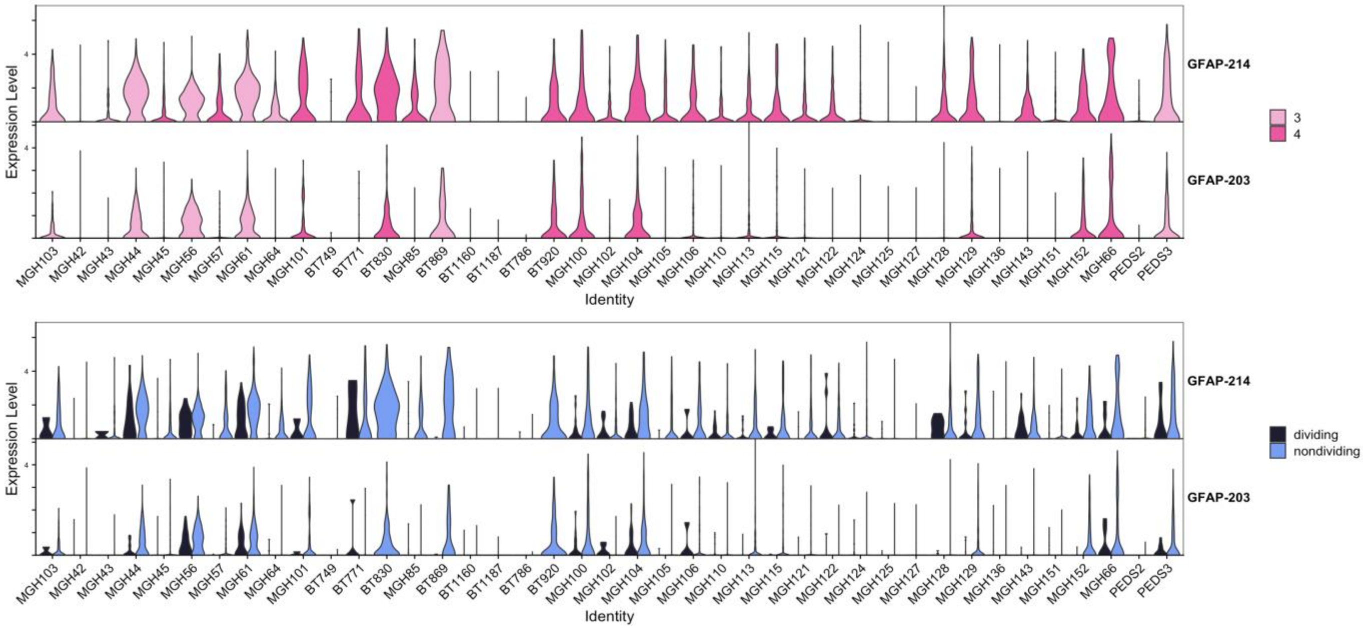

Figure 6 | Violin Plot visualizing distribution of GFAPα and δ isoform expression levels per separate tumor. The tumors are colored by glioma grade on the top plot and are split between quiescent and dividing cells in the bottom violin plot.

To assess if similar GFAP expression patterns are also observed per tumor, violin plots of GFAPα and δ expression per tumor were created (Fig. 6). The tumors starting with MGH are adult tumors while those starting with BT are pediatric. In the violin plots of GFAPα and δ expression levels per grade, there is a large amount of heterogeneity between tumors. In all cases if GFAP is expressed, GFAPα levels are higher than those of GFAP δ. This is noticeable again by tumor as by cluster in figure 5. Certain tumors do not express any GFAP, for example: MGH125, MGH127, MGH42, BT1160, BT1187, and BT786. There are other tumors which only express GFAPα but not GFAPδ, these seem to be primarily glioblastoma. Finally, there are also certain tumors in which GFAPα and GFAPδ are expressed, tumor MGH56 even has similar GFAPα and GFAPδ levels. The GFAP isoforms were also investigated by activity in the bottom plot of figure 6. As a whole, GFAP seems to be more highly expressed in quiescent cells per tumor, the same finding as investigation by cluster. There are certain tumors with higher GFAPα expression levels in dividing than quiescent cells such as MGH44, MGH56, BT771, and MGH128.

GFAPδ/α ratio in Venteicher 2017

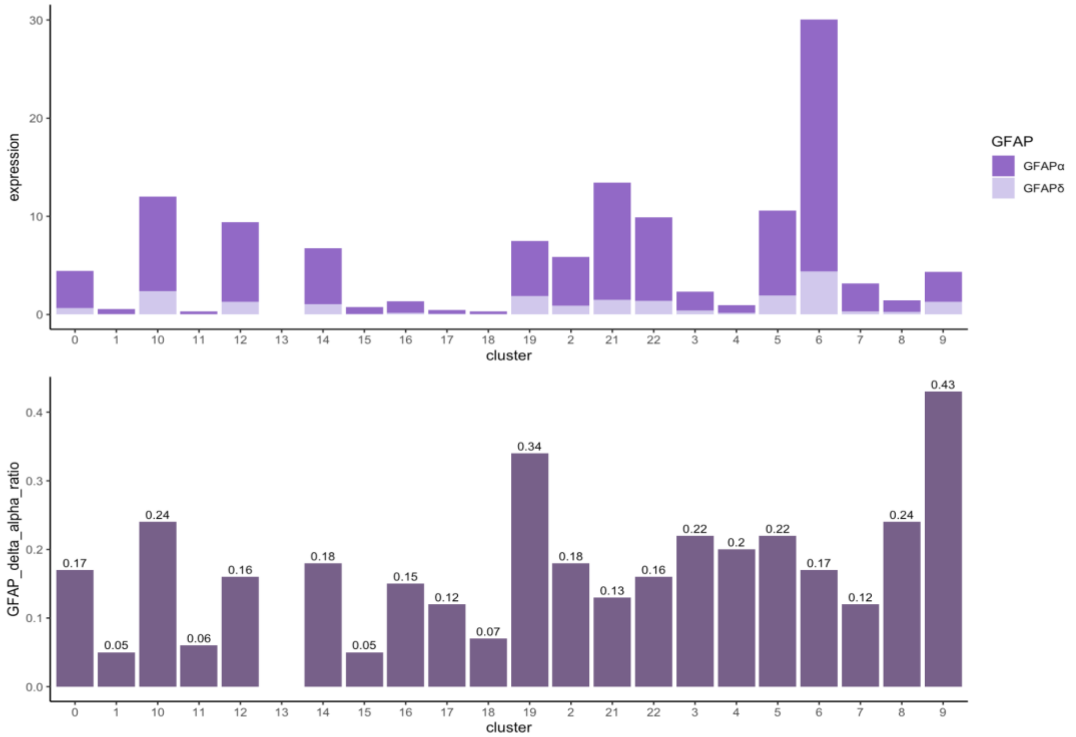

Figure 7 | Average GFAP expression and ratio per UMAP cluster

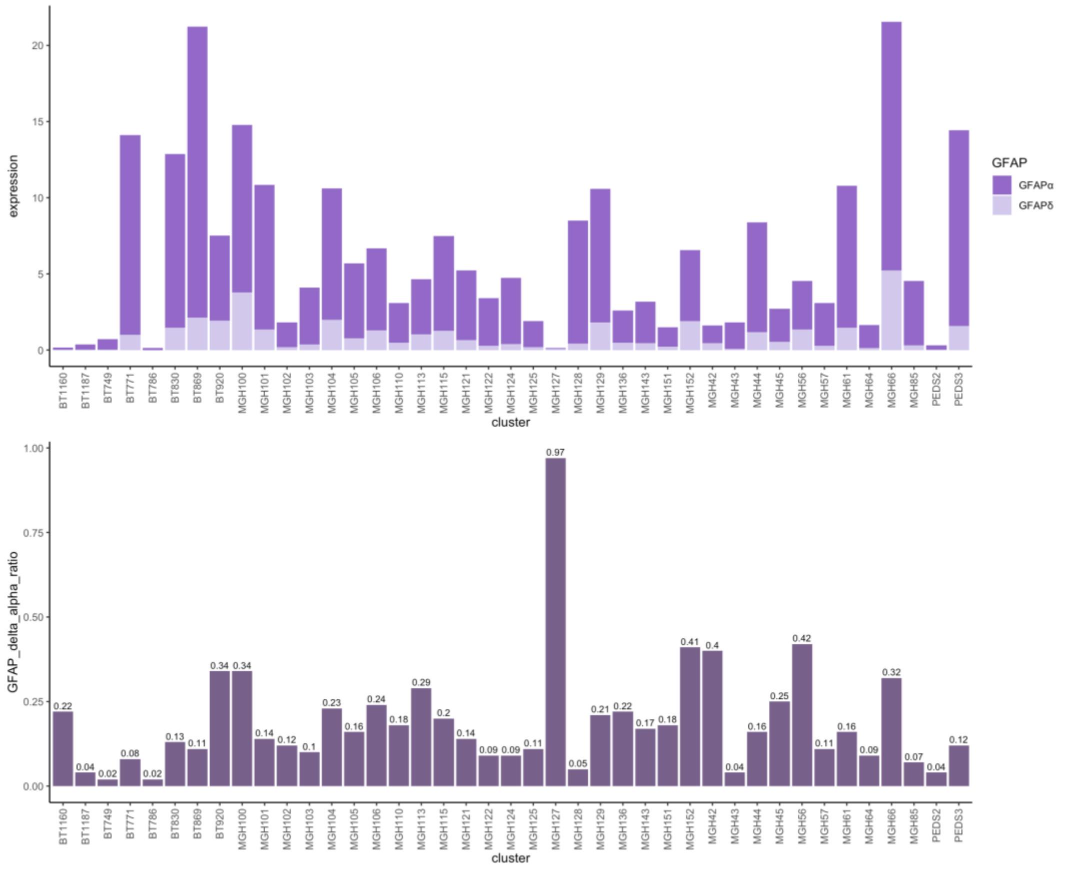

Figure 8 | Average GFAP expression and ratio per tumor sample

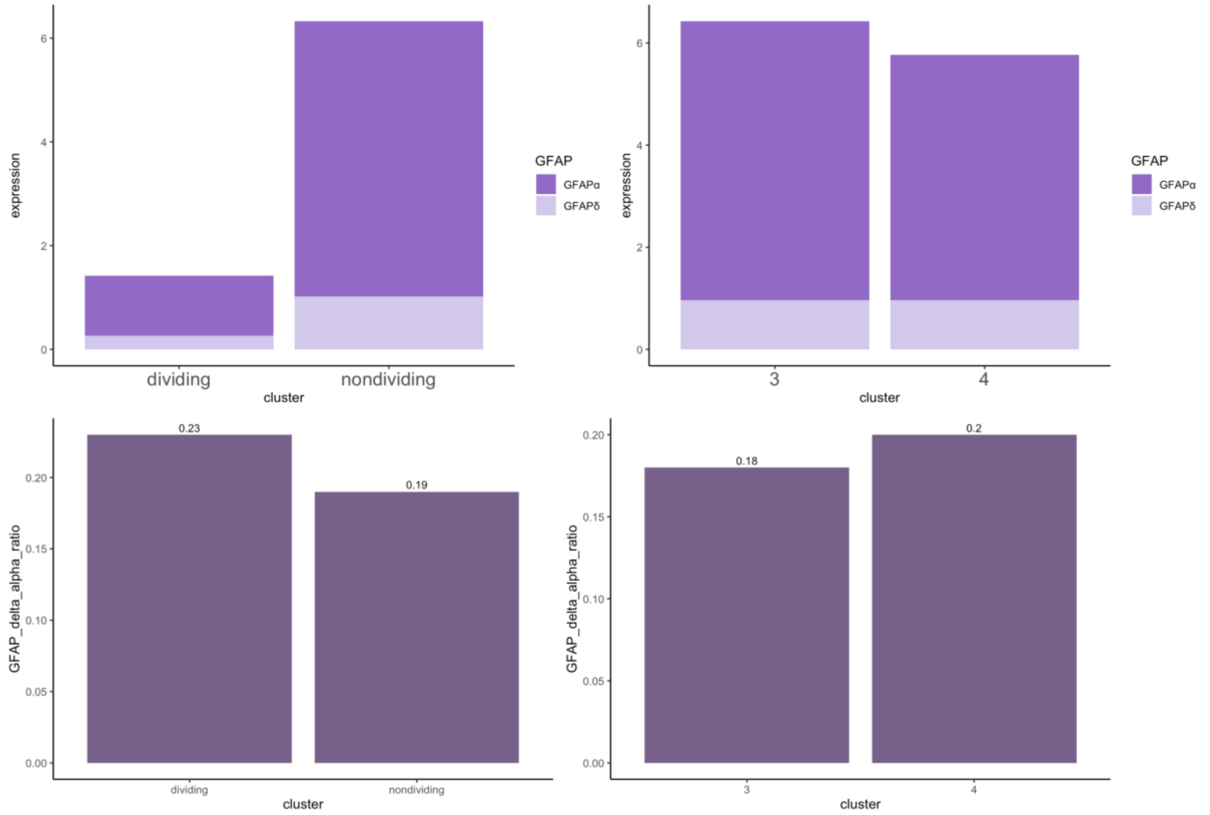

Figure 9 | Average GFAP expression and ratio per activity and tumor grade.

The average expression and ratio plots of GFAP isoforms α and δ were grouped according to Seurat cluster, tumor, activity, and glioma grade. GFAP expression between clusters is very heterogenous, but a notable cluster with the highest GFAP overall expression is cluster 6 with a GFAPδ/α ratio of 0.17 (Fig. 7). The cluster with the highest GFAPδ/α ratio is cluster 9 with a ratio of 0.43, almost as much GFAPδ expression as GFAPα. When studying average GFAP expression among tumors, there is also a lot of heterogeneity (Fig. 8). Certain tumors do not express any GFAP while others express very high levels. Tumors MGH66 and BT869 have the highest GFAP expression, but the ratio of GFAPδ/α is higher in MGH66, 0.32, compared to BT869, 0.11. It's important to consider both GFAP expression and ratio, since the tumor with the highest ratio of 0.93 is MGH127, but the expression of this tumor is almost 0. Considering GFAP levels in dividing and neural quiescent cells, there is explicit higher expression of GFAP in nondividing quiescent cells (Fig. 9). However, there does not seem to be a real change in isoform δ/α ratio between the two groups. Finally, investigation of average GFAP expression levels and δ/α ratio between grades III and IV glioma did not find a difference between grade (Fig. 9). GFAPδ/α ratio seems similar based on this single cell data which conflicts with previous studies. However, there are high amounts of heterogeneity between tumors and these values are averaged. It would be of interest to subdivide tumors into class based on GFAPδ/α ratio and expression then observe changes between grade and GFAP class.

Density plots of GFAP alpha percent

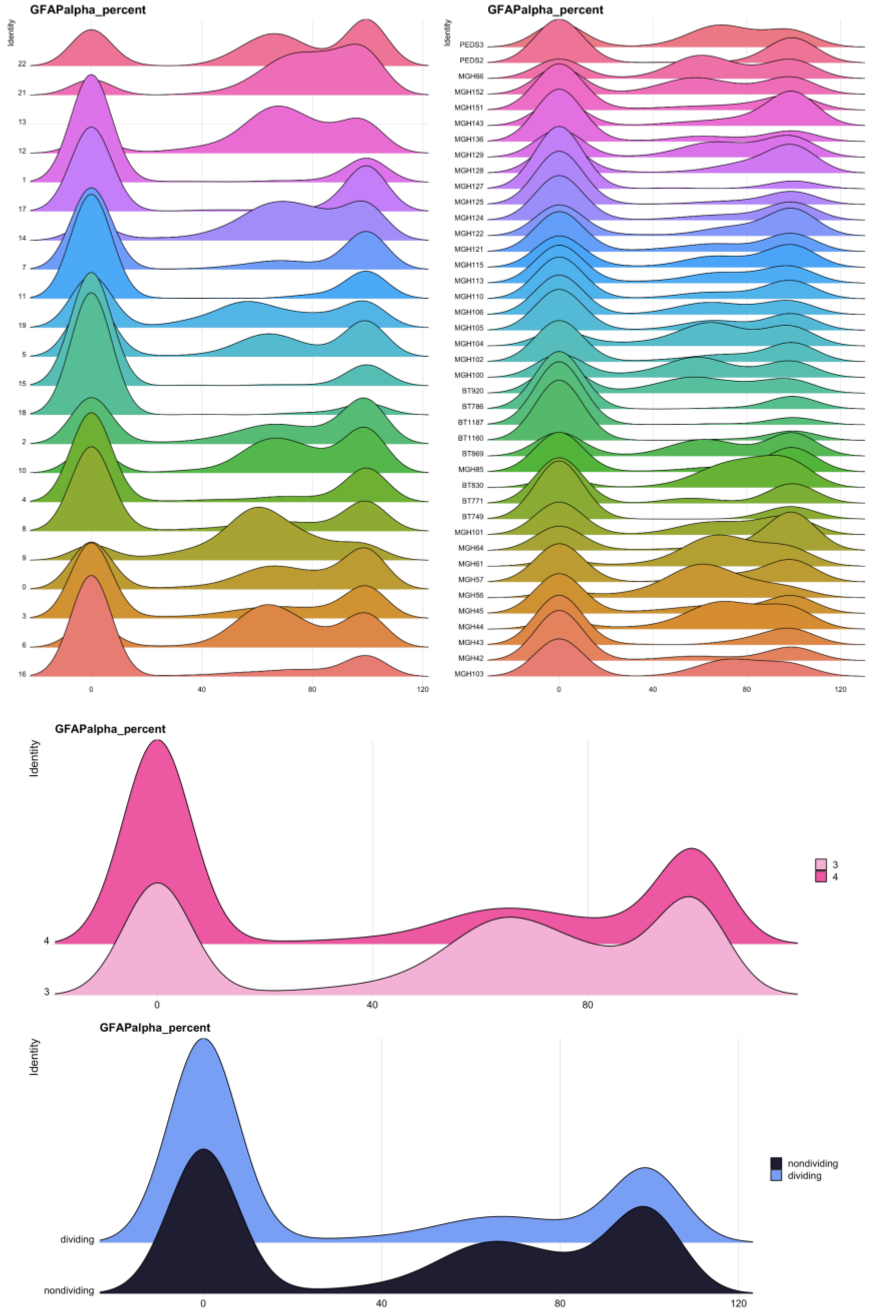

Figure 10 | Density plots of GFAP alpha percentage in clusters, tumors, grade, and activity.

In order to investigate GFAP δ/α ratio per cell, the percentage of GFAPα out of total GFAP expression was plotted in a ridge plot (Fig. 10). GFAPα percent was plotted by cluster, tumor, grade, and activity. In each plot, there are 3 ridges noticeable per group. A prominent ridge at 0, indicative of GFAP negative cells. A second ridge at 100, indicating cells only expressing GFAPα. A third usually smaller ridge between a GFAPα percent of 40 to 90, indicative of the cells that express GFAPα and GFAPδ. Looking into these patterns of GFAP isoform expression shows an informative expression pattern per group. The cells in cluster 9 are mostly expressing GFAPα and δ. Tumor MGH56 also shows a similar pattern. These GFAP cell subpopulations are able to point out interesting groups for sub analysis.

Invasion scoring

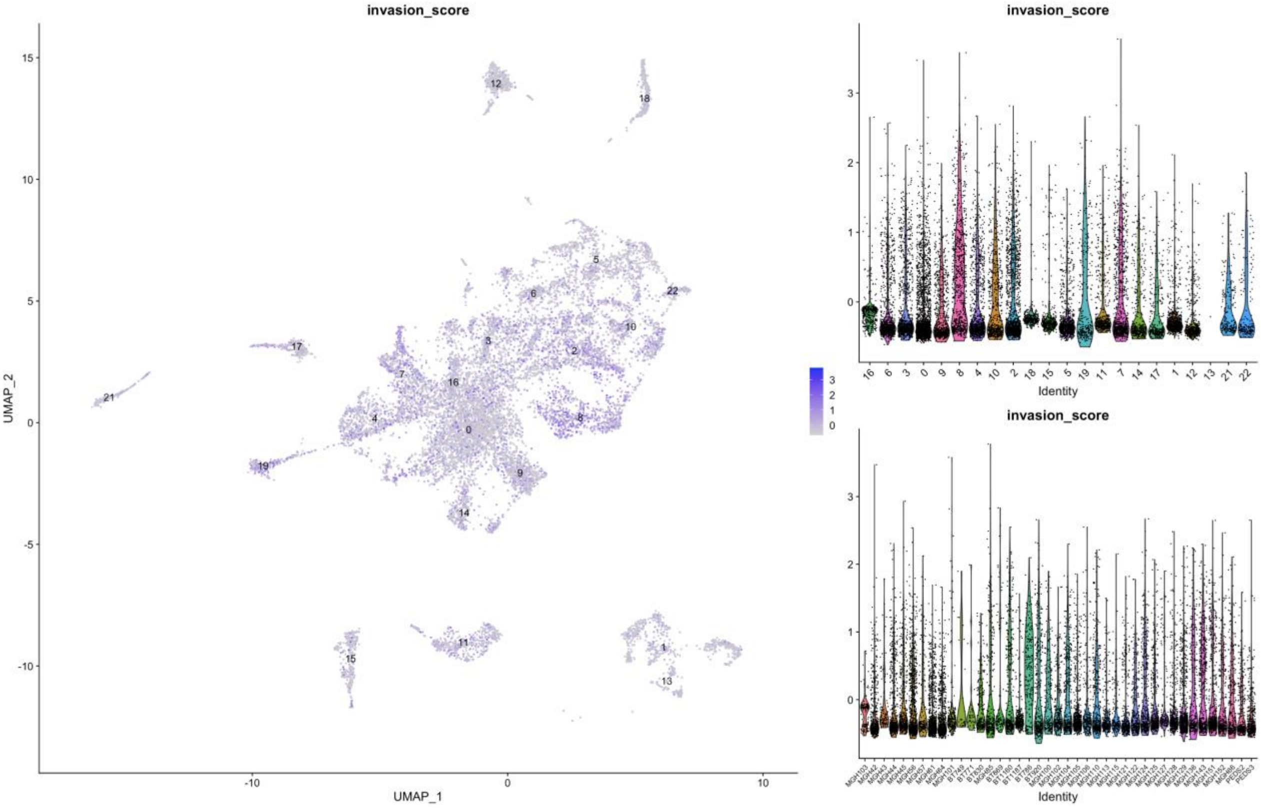

Figure 11 | UMAP and violin plot of invasion score per cluster and tumor.

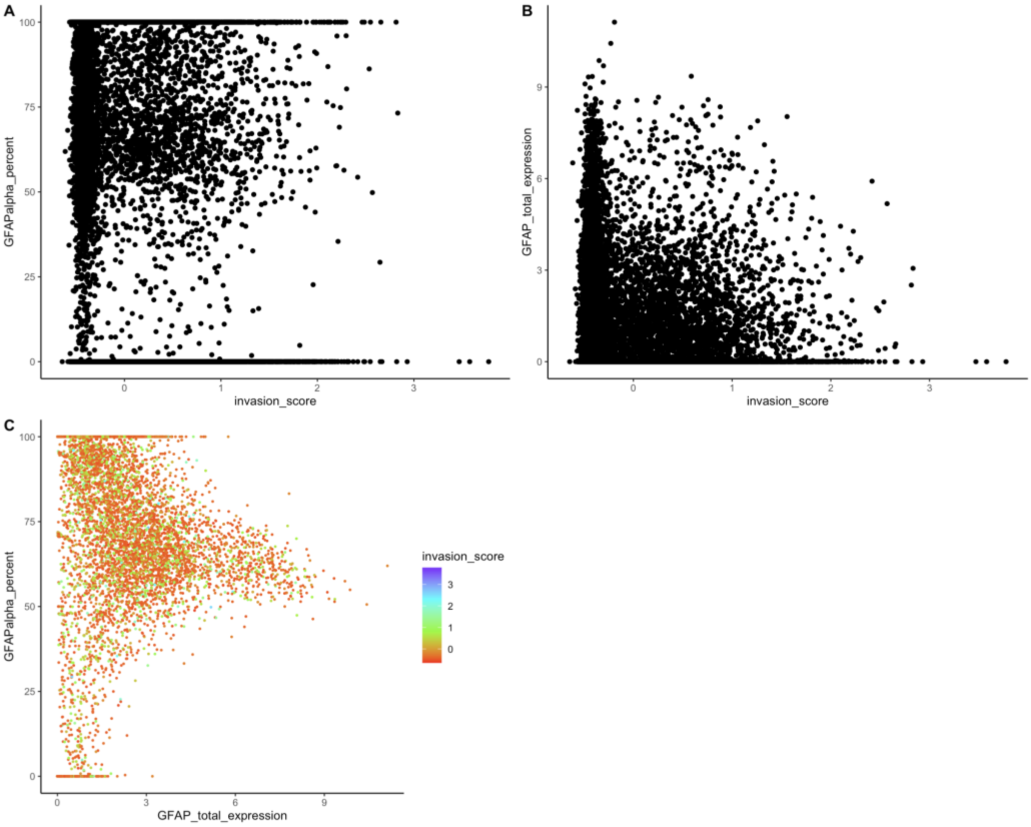

Figure 12 | Invasion score as a function of GFAPα percentage and GFAP total expression

Invasion scores were calculated by cell based on the list of 271 invasive genes (Fig. 11). These scores are mostly below 0 with few cells having high invasive scores per cluster or tumor. The average invasive score per tumor is below 0, but cluster 8 and tumor BT876 has a bit of a higher average invasive score. Cluster 8 was previously noted to have the most dividing cells in its cluster. When analyzing invasion score as a function of GFAP expression of GFAP α percent, there does not seem to be much of a pattern in terms of invasion (Fig. 12). The same conclusion is visible when graphing total GFAP expression as a function of GFAPα percent colored by invasion score. While there is no clear association with invasion scores, as GFAP total expression increases, GFAPα percent seems decrease.

Gojo 2020 et al.

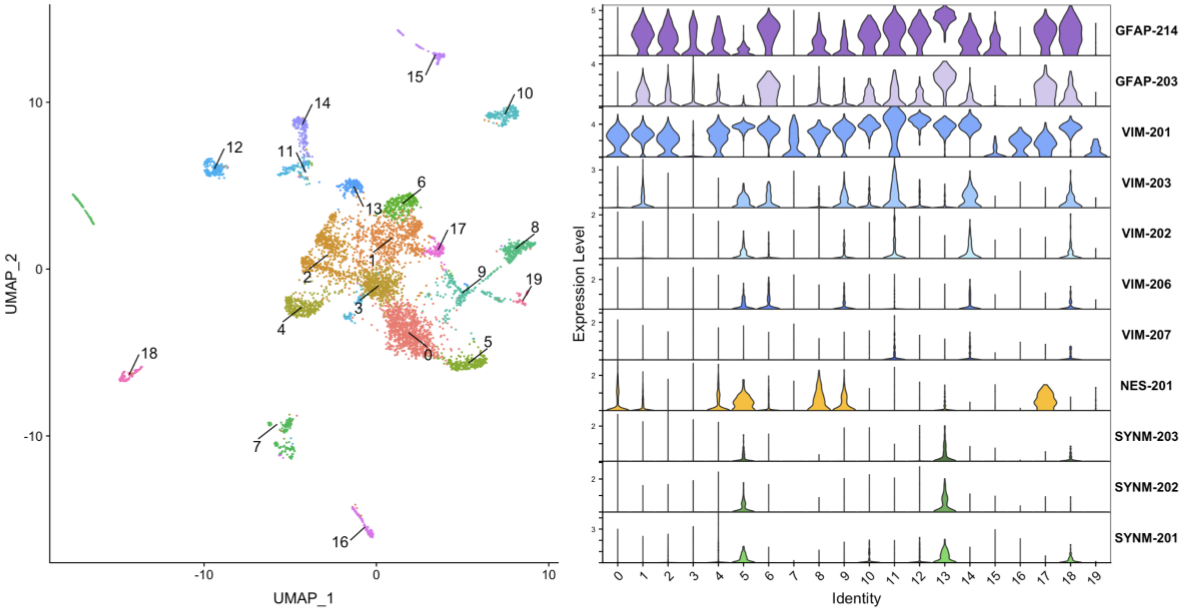

Figure 13 | UMAP dimensionality reduction and a violin plot of intermediate filament isoforms present in Gojo datasets 5,900 cells.

Once quality control and preprocessing steps were completed, the Gojo dataset on primary pediatric ependymoma consisted of 5,900 cells and 20 clusters (Fig. 13). All four glial intermediate filament proteins were found in detectable levels. GFAP isoforms α and δ, five isoforms of Vimentin, one isoform of Nestin, and three isoforms of Synemin were detected. GFAPα and GFAPδ are expressed in most clusters of the ependymoma dataset. Specific clusters with high GFAPδ/α ratio are clusters 6 and 13. VIM-201 is has the highest expression out of this group of isoforms and is expressed in each cluster. Clusters 18 and 14 have cells that express all vimentin isoforms. The one isoform of Nestin, NES-201, is expressed in cluster 17, 8, and 5. The three Synemin isoforms detected, SYNM-201, SYNM-202, and SYNM-203, all overlap in terms of expression. Synemin is expressed in clusters 5, 13, and 18.

GFAPα and GFAPδ expression per cluster

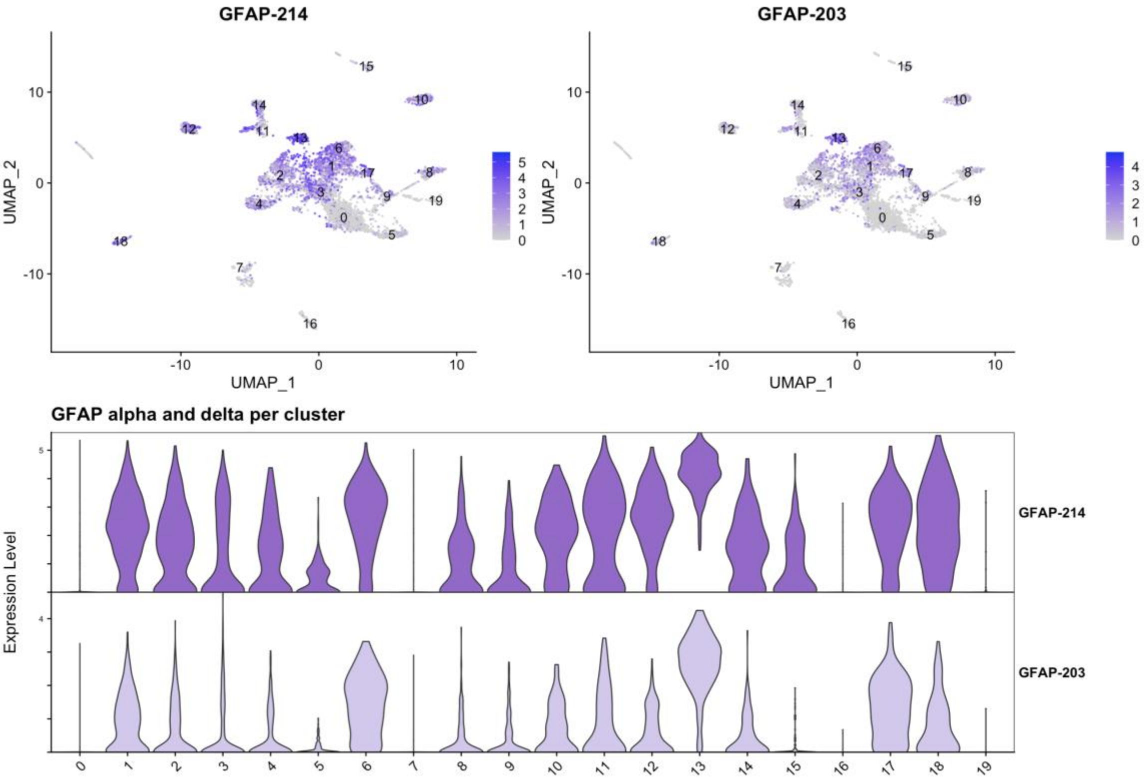

Figure 14 | UMAP with GFAP expression of Gojo 2020

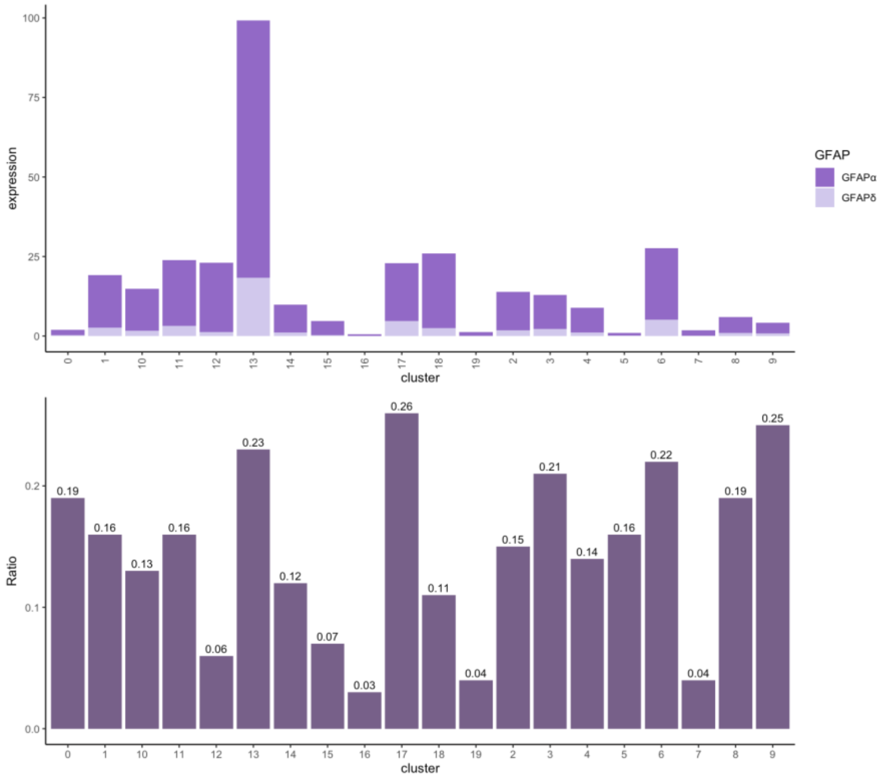

Figure 15 | GFAP average expression and ratio per Seurat cluster

GFAP isoforms α and δ were investigated more closely in the Gojo 2020 dataset (Fig. 14). The average expression and ratio of GFAPα and GFAPδ were also calculated (Fig. 15). There is a lot of heterogeneity between clusters, similar in pattern to the Veneticher dataset. Cluster 13 has the highest average GFAP expression and the second highest GFAPδ/α ratio. Clusters 6 and 17 also have high GFAP ratios and good GFAP expression levels. There clusters are interesting to investigate further.

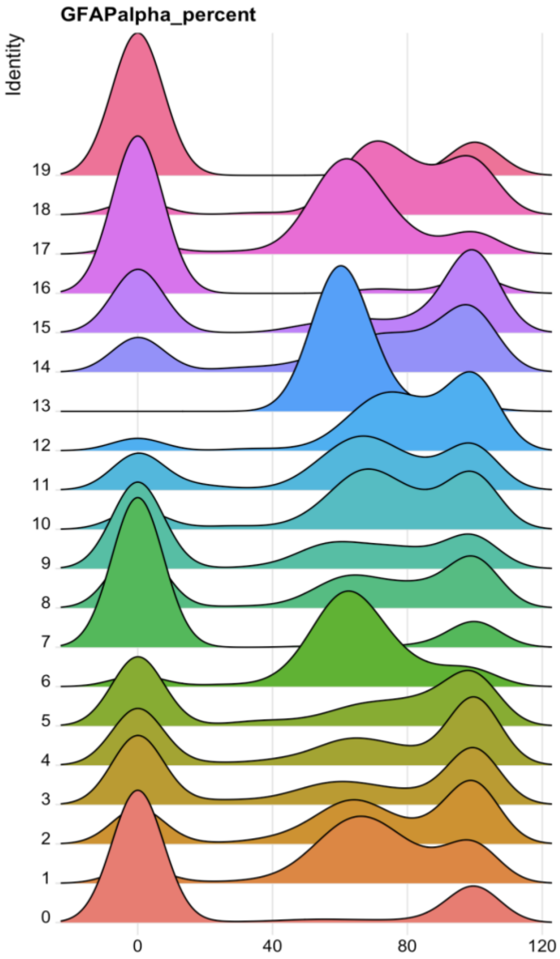

Figure 16 | Density plot of GFAP alpha percentage per cell in each cluster

GFAP cell population was investigated by analyzing the percent of total GFAP expression that is GFAPα (Fig. 16). Similar results as the Veneticher dataset occurred in the Gojo pediatric ependymoma dataset. There were 3 peaks on the density plot, each indicated a specific GFAP expressing cell subtype. A GFAPα percent of 0 indicated that the cell doesn't express GFAP, a GFAPα percent of 100 indicated that the cell only expresses GFAPα, and finally a GFAPα percentage between 40 to 80 indicated that the cell expressed both GFAPα and GFAPδ.

Invasion scoring

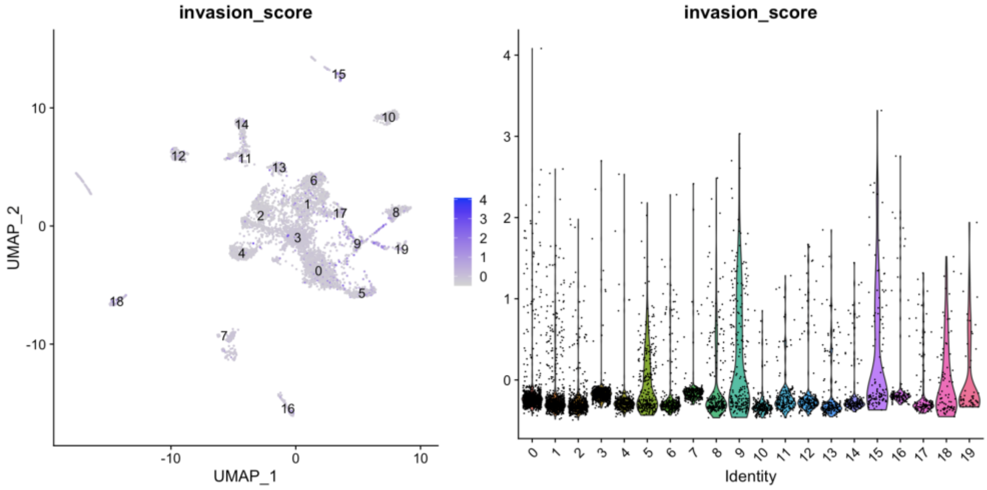

Figure 17 | UMAP and violin plot of invasion score per cluster

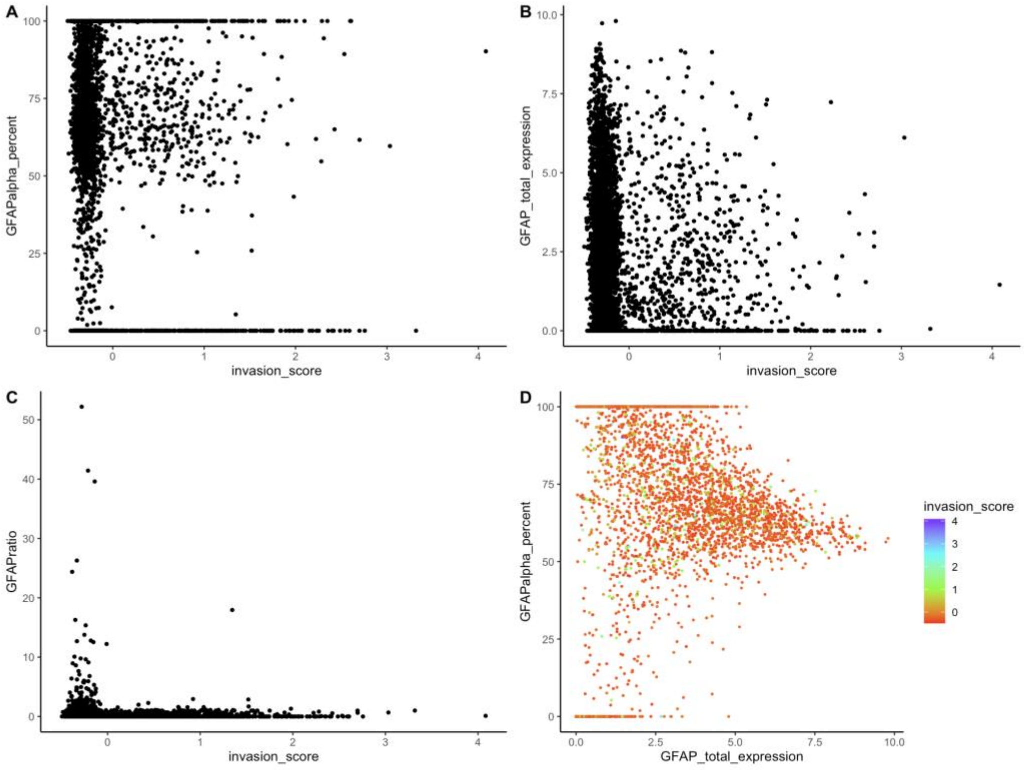

Figure 18 | Invasion score as a function of GFAP alpha percentage, total GFAP expression, and GFAP ratio.

Invasion scoring was performed per cell for the ependymoma dataset Gojo et al 2020 based on the list of 271 invasive genes (Fig. 17). Invasion in relation to GFAP was analyzed plotted in scatterplots of invasion score as a function of GFAPα percentage, total GFAP expression, and GFAP ratio (Fig. 18). There seems to not be a clear correlation between invasion score and GFAP. The same pattern is seen when plotting GFAP total expression by GFAPα percentage colored by invasion score. However, in the ependymoma dataset, there is also a pattern of decreasing GFAPα percentage with increasing total GFAP expression.

Discussion

The aim of this study was to perform a single cell intermediate filament isoform analysis of glioblastoma multiforme. Previous work on this project established the bioinformatics analysis pipeline used in this study on the datasets of Venteicher et al. 2017 and Gojo et al 2020. The pipeline psuedoaligned and mapped reads using the program Kallisto, then analyzed the output in the single cell package Seurat. The annotated single cell isoform level datasets were investigated for intermediate filament expression, specifically focusing on the role of GFAP in quiescence and invasion in gliomas.

Intermediate filaments that are known to be highly expressed in glial cells were analyzed in both Venteicher and Gojo datasets. Nestin, GFAP, and Vimentin were found in detectable levels in both datasets while Synemin was only found to be expressed in Gojo 2020. Expression of these isoforms differs per cluster, but the most highly expressed intermediate filament was an isoform of Vimentin, VIM-201. Vimentin was also listed as one of the 271 invasive genes from the list of Yu et al. 2020; therefore, this gene and its isoforms would be interesting to analyze in glioblastoma for further research.

The expression levels of isoform GFAPα were found as the highest, corresponding with previous literature about GFAP isoforms (Stassen et al., 2017). However, when comparing GFAPδ/α ratio in grade III glioma to glioblastoma, there is no difference in ratio or expression levels which does not align with current studies (van Asperen et al., 2022). This finding is due to the fact that GFAPδ/α ratio was calculated as an average in this analysis. This variability is supported by calculating average GFAPδ/α ratio and expression by tumor. In Figure 8, the heterogeneity between tumors was well characterized, as there were tumors that do not express GFAP in both grades III and IV gliomas. This makes it clear that there are high levels of intertumoral variation in regards to GFAPδ/α ratio and GFAP expression as a whole. It was further established that this heterogeneity is also intratumoral, there are cells that do not express GFAP, those that only express GFAPα and those that express both GFAPα and GFAPδ. The distributions of these cell subpopulations differ between tumors, but there are certain tumors where a high amount of GFAPδ/α expressing cells can be observed. Therefore, if GFAP is a gene of interest in glioblastoma malignancy, it is for specific tumors and not all encompassing. It would be of interested to classify tumors as GFAP positive or negative and analyze the characteristics of these subgroups with a differentially expressed genes analysis. It is also interesting to note that as total GFAP expression increases, the percentage of GFAPα decreases. This pattern can be observed in both glioma and ependymoma datasets. The ependymoma dataset from Gojo et al. 2020 could function as a malignancy benchmark of sorts for glioblastoma. Ependymoma tumors are less deadly that glioblastoma, with a survival rate higher than 80% after 5 years. However, even though these tumors are more likely to be cured, they also have high recurrence (Celano et al., 2016). Certain patients in the Gojo dataset had four tumor recurrences, three of which were sequenced in the dataset. Of course, in order to perform any significant comparison and analysis of the Gojo data with glioblastoma in Venteicher et al., clinical metadata annotations must also be made for Gojo et al. 2020.

Quiescent neural stem cells present in glioblastoma function as a reservoir that harboring tumor recurrence after cytotoxic treatment (Sadcheva et al., 2019). The high GFAP expression levels in quiescent cells compared to actively dividing cells could be indicative of GFAP's role in tumor malignancy. The ratio of GFAPδ/α itself is not differing significantly, but the expression of GFAPα and GFAPδ is significantly higher in quiescent compared to dividing cells. It could be possible that cells overexpress GFAP while quiescent state of arrested cell growth to prepare themselves for division and invasion later on. This analysis is correlative and it is difficult to find out whether this difference in expression levels of GFAP causes quiescence or is a product of quiescence. However, the relationships investigated in this report may help guide further experimentation. To confirm any correlations, in vitro experiments should be performed through the use of a glioblastoma organoid model. GFAP isoform knockout lines could be developed to test the effect of GFAPα or GFAPδ loss on the invasive characteristics of the glioblastoma tumor.

Invasion score was not found to be correlated with GFAP expression or ratio in Venteicher and Gojo datasets. However, the cluster with the highest average invasive score was also the cluster with the most dividing cells, cluster 8 in Venteicher 2017. This is indicative that the genes used for invasive scoring may be correlated with those for cell division. The genes used for scoring invasion originate from the paper Yu et al. 2020, it is possible that the genes represent actively replicating cells, not necessarily ones that resist treatment. Since Vimentin was used to generate invasive score, it makes sense that VIM-201 is a more interesting gene than GFAP when looking at invasion (Fig. 22). It would be interesting to further investigate GFAP using a different list or invasive genes, ones that correlate with quiescent neural stem cells. However, more studies identifying genes as relevant for invasiveness must be performed. A limitation of investigating markers for invasive cells is that the tumor samples collected in the datasets come from surgical resections from clinicians. The clinician's goal is to remove the bulk of the tumor, but they do not remove excess surrounding tissue which likely contains invading cells propagating tumor recurrence. Therefore, the datasets do not include the true invasive cells of interest.

A limitation of the current analysis is the parameters used for clustering. Clusters represent tumor classes, but the clusters are fluid and the number of clusters changes based on resolution, PC's, and k. The current parameters were selected for optimal coverage in Venteicher 2017 and Gojo 2020 based on PC plots and ease of biological interpretation of clusters. Changing any of these parameters will result in a different number of clusters calculated. It would be beneficial to have a larger number of clusters when looking for rare tumor varieties or rare cell types. Further analysis could use the python library schist which uses a nested stochastic block model opposed to Leiden clustering in to find the best parameters for the model (Morelli et al., 2021). This program was not used in the current analysis due to time constraints.

The use of a filtered GFAPα and GFAPδ index is also a limitation in this study. In the previous work on this project by Alexandra de Reus, it was found that after mapping with Kallisto, high levels of transcript GFAP-201 were discovered. This is a protein coding isoform according to the Ensemble database, but this GFAP-201 is not mentioned as a known isoform in literature. She concluded that this isoform was an artifact from Kallisto, since Kallisto relies on pseudoalignment instead of alignment to increase computational speed (de Reus, 2021). Mapping using a filtered index solved this problem, but there is a chance some of the discarded isoforms are incorrectly pseudoaligned GFAPα and GFAPδ isoforms. Further research could remap the datasets analyzed in this study using the mapping tool STAR which has improved accuracy compared to Kallisto (Dobin et al., 2012).

Another limitation of this analysis is that a large label transfer was used for neural quiescence labels, from 3,010 cells to 11,507 unlabeled cells. This extrapolation likely has an impact on the accuracy of the labels. In order to improve this, the neural G_0 neural network classification model could be used (O'Connor et al., 2020). This is the same technique that labeled the 3,010 cells from Venteicher as nondividing or dividing used in this analysis.

Further analysis should be performed on differentially expressed genes in GFAP expressing populations of cells. For example, finding marker genes for cells not expressing any GFAP, cells expressing both GFAPα and δ isoforms, and for those expressing only GFAPα. This investigation can be followed up with a gene regulatory network and splice factor analysis. That way we can understand which splice factors are responsible for the regulation of GFAP isoforms in glioblastoma.

Conclusion

In conclusion, glioblastoma tumors are very heterogenous expressing different levels and isoform ratio of GFAPα and δ. The expression of GFAPα and δ isoforms is higher in nondividing quiescent cells compared to actively dividing cells. The datasets annotated in this analysis serve as a great starting off point for deeper analysis into the intricacies of the intermediate filament network in glioblastoma. Further research on the single cell datasets mapped here would bring valuable insight into the molecular characteristics of glioblastoma multiforme.

Acknowledgements

I would like to express my gratitude to my supervisor Dr. Onur Basak for guiding me through the world of neuroscience these past months. It has been a pleasure working with you and thank you so much for sharing your knowledge. In addition, I would like to thank Dr. Emma van Bodegraven for her ideas on glioblastoma, intermediate filaments, and scoring methods for invasion. I am also grateful to Prof. Dr. Elly Hol for her expertise on the protein GFAP and glial cells which made this project feel important. I would like to sincerely thank Alexandra de Reus for establishing this bioinformatics pipeline in her internship and helping me understand the analysis. Finally, I want to thank all the lab members of the Hol and Basak lab groups for their insightful questions, presentations, and fun lab bonding activities. I really enjoyed working with you all.

References

Bonnal, S. C., López-Oreja, I., & Valcárcel, J. (2020). Roles and mechanisms of alternative splicing in cancer — implications for care. *Nature Reviews Clinical Oncology*, *17*(8), 457–474. https://doi.org/10.1038/s41571-020-0350-x

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, *34*(5), 525–527. https://doi.org/10.1038/nbt.3519

Broad Data Use Oversight System. (n.d.). Retrieved December 30, 2022, from https://duos.broadinstitute.org/

Celano, E., Salehani, A., Malcolm, J. G., Reinertsen, E., & Hadjipanayis, C. G. (2016). Spinal cord ependymoma: A review of the literature and case series of Ten patients. *Journal of Neuro-Oncology*, *128*(3), 377–386. https://doi.org/10.1007/s11060-016-2135-8

Chen, R., Smith-Cohn, M., Cohen, A. L., & Colman, H. (2017). Glioma subclassifications and their clinical significance. *Neurotherapeutics*, *14*(2), 284–297. https://doi.org/10.1007/s13311-017-0519-x

Comba, A., Faisal, S. M., Dunn, P. J., Argento, A. E., Hollon, T. C., Al-Holou, W. N., Varela, M. L., Zamler, D. B., Quass, G. L., Apostolides, P. F., Abel, C., Brown, C. E., Kish, P. E., Kahana, A., Kleer, C. G., Motsch, S., Castro, M. G., & Lowenstein, P. R. (2022). Spatiotemporal analysis of glioma heterogeneity reveals COL1A1 as an actionable target to disrupt tumor progression. *Nature Communications*, *13*(1). https://doi.org/10.1038/s41467-022-31340-1

Couturier, C. P., Ayyadhury, S., Le, P. U., Nadaf, J., Monlong, J., Riva, G., Allache, R., Baig, S., Yan, X., Bourgey, M., Lee, C., Wang, Y. C., Wee Yong, V., Guiot, M.-C., Najafabadi, H., Misic, B., Antel, J., Bourque, G., Ragoussis, J., & Petrecca, K. (2020). Single-cell RNA-seq reveals that glioblastoma recapitulates a normal neurodevelopmental hierarchy. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-17186-5

de Reus, A. (2021). Gfap isoform expression in glioblastoma single cell Rna sequencing datasets (thesis).

Delgado-López, P. D., & Corrales-García, E. M. (2016). Survival in glioblastoma: A review on the impact of treatment modalities. *Clinical and Translational Oncology*, *18*(11), 1062–1071. https://doi.org/10.1007/s12094-016-1497-x

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2012). Star: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Dvinge, H., Kim, E., Abdel-Wahab, O., & Bradley, R. K. (2016). RNA splicing factors as oncoproteins and tumour suppressors. *Nature Reviews Cancer*, *16*(7), 413–430. https://doi.org/10.1038/nrc.2016.51

Gojo, J., Englinger, B., Jiang, L., Hübner, J. M., Shaw, M. K. L., Hack, O. A., Madlener, S., Kirchhofer, D., Liu, I., Pyrdol, J., Hovestadt, V., Mazzola, E., Mathewson, N. D., Trissal, M., Lötsch, D., Dorfer, C., Haberler, C., Halfmann, A., Mayr, L., … Filbin, M. G. (2020). Single-cell RNA-seq reveals cellular hierarchies and impaired developmental trajectories in pediatric ependymoma. *Cancer Cell*, *38*(1). https://doi.org/10.1016/j.ccell.2020.06.004

Hoffman, P. (2022, January 11). *Mapping and annotating query datasets*. Seurat. Retrieved March 5, 2023, from https://satijalab.org/seurat/articles/integration_mapping.html

Jackson, C. M., Choi, J., & Lim, M. (2019). Mechanisms of immunotherapy resistance: Lessons from glioblastoma. *Nature Immunology*, *20*(9), 1100–1109. https://doi.org/10.1038/s41590-019-0433-y

Morelli, L., Giansanti, V., & Cittaro, D. (2021). Nested stochastic block models applied to the analysis of Single Cell Data. *BMC Bioinformatics*, *22*(1). https://doi.org/10.1186/s12859-021-04489-7

O'Connor, S. A., Feldman, H. M., Arora, S., Hoellerbauer, P., Toledo, C. M., Corrin, P., Carter, L., Kufeld, M., Bolouri, H., Basom, R., Delrow, J., McFaline-Figueroa, J. L., Trapnell, C., Pollard, S. M., Patel, A., Paddison, P. J., & Plaisier, C. L. (2021). Neural G0: A quiescent-like state found in neuroepithelial-derived cells and glioma. *Molecular Systems Biology*, *17*(6). https://doi.org/10.15252/msb.20209522

Picelli, S., Faridani, O. R., Björklund, Å. K., Winberg, G., Sagasser, S., & Sandberg, R. (2014). Full-length RNA-seq from single cells using SMART-SEQ2. *Nature Protocols*, *9*(1), 171–181. https://doi.org/10.1038/nprot.2014.006

Radu, R., Petrescu, G. E., Gorgan, R. M., & Brehar, F. M. (2022). GFAPδ: A promising biomarker and therapeutic target in glioblastoma. *Frontiers in Oncology*, *12*. https://doi.org/10.3389/fonc.2022.859247

Sachdeva, R., Wu, M., Johnson, K., Kim, H., Celebre, A., Shahzad, U., Graham, M. S., Kessler, J. A., Chuang, J. H., Karamchandani, J., Bredel, M., Verhaak, R., & Das, S. (2019). BMP signaling mediates glioma stem cell quiescence and confers treatment resistance in glioblastoma. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-51270-1

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., & Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, *33*(5), 495–502. https://doi.org/10.1038/nbt.3192

Stassen, O. M. J. A., van Bodegraven, E. J., Giuliani, F., Moeton, M., Kanski, R., Sluijs, J. A., van Strien, M. E., Kamphuis, W., Robe, P. A. J., & Hol, E. M. (2017). GFAPδ/gfapα ratio directs astrocytoma gene expression towards a more malignant profile. *Oncotarget*, *8*(50), 88104–88121. https://doi.org/10.18632/oncotarget.21540

Uceda-Castro, R., van Asperen, J. V., Vennin, C., Sluijs, J. A., van Bodegraven, E. J., Margarido, A. S., Robe, P. A., van Rheenen, J., & Hol, E. M. (2022). GFAP splice variants fine-tune glioma cell invasion and tumour dynamics by modulating migration persistence. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-021-04127-5 van Asperen, J. V., Robe, P. A. J. T., & Hol, E. M. (2022). GFAP alternative splicing and the relevance for disease – a focus on diffuse gliomas. *ASN Neuro*, *14*, 175909142211020. https://doi.org/10.1177/17590914221102065

van Bodegraven, E. J., & Etienne-Manneville, S. (2021). Intermediate filaments from tissue integrity to single molecule mechanics. *Cells*, *10*(8), 1905. https://doi.org/10.3390/cells10081905

van Bodegraven, E. J., Asperen, J. V., Robe, P. A. J., & Hol, E. M. (2019). Importance of GFAP isoform-specific analyses in astrocytoma. *Glia*, *67*(8), 1417–1433. https://doi.org/10.1002/glia.23594

van Bodegraven, E. J., Asperen, J. V., Sluijs, J. A., Deursen, C. B., Strien, M. E., Stassen, O. M., Robe, P. A., & Hol, E. M. (2019). GFAP alternative splicing regulates glioma cell–ECM interaction in a dusp4-dependent manner. *The FASEB Journal*, *33*(11), 12941–12959. https://doi.org/10.1096/fj.201900916r

Venteicher, A. S., Tirosh, I., Hebert, C., Yizhak, K., Neftel, C., Filbin, M. G., Hovestadt, V., Escalante, L. E., Shaw, M. K. L., Rodman, C., Gillespie, S. M., Dionne, D., Luo, C. C., Ravichandran, H., Mylvaganam, R., Mount, C., Onozato, M. L., Nahed, B. V., Wakimoto, H., … Suvà, M. L. (2017). Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. *Science*, *355*(6332). https://doi.org/10.1126/science.aai8478

Wang, L., Shang, Z., Zhou, Y., Hu, X., Chen, Y., Fan, Y., Wei, X., Wu, L., Liang, Q., Zhang, J., & Gao, Z. (2018). Autophagy mediates glucose starvation-induced glioblastoma cell quiescence and chemoresistance through coordinating cell metabolism, cell cycle, and survival. *Cell Death & Disease*, *9*(2). https://doi.org/10.1038/s41419-017-0242-x

Yu, K., Hu, Y., Wu, F., Guo, Q., Qian, Z., Hu, W., Chen, J., Wang, K., Fan, X., Wu, X., Rasko, J. E. J., Fan, X., Iavarone, A., Jiang, T., Tang, F., & Su, X.-D. (2020). Surveying brain tumor heterogeneity by single-cell RNA-sequencing of multi-sector biopsies. *National Science Review*, *7*(8), 1306–1318. https://doi.org/10.1093/nsr/nwaa099

Zhang, Y., Wang, D., Peng, M., Tang, L., Ouyang, J., Xiong, F., Guo, C., Tang, Y., Zhou, Y., Liao, Q., Wu, X., Wang, H., Yu, J., Li, Y., Li, X., Li, G., Zeng, Z., Tan, Y., & Xiong, W. (2021). Single-cell RNA sequencing in cancer research. *Journal of Experimental & Clinical Cancer Research*, *40*(1). https://doi.org/10.1186/s13046-021-01874-1

Supplementary figures

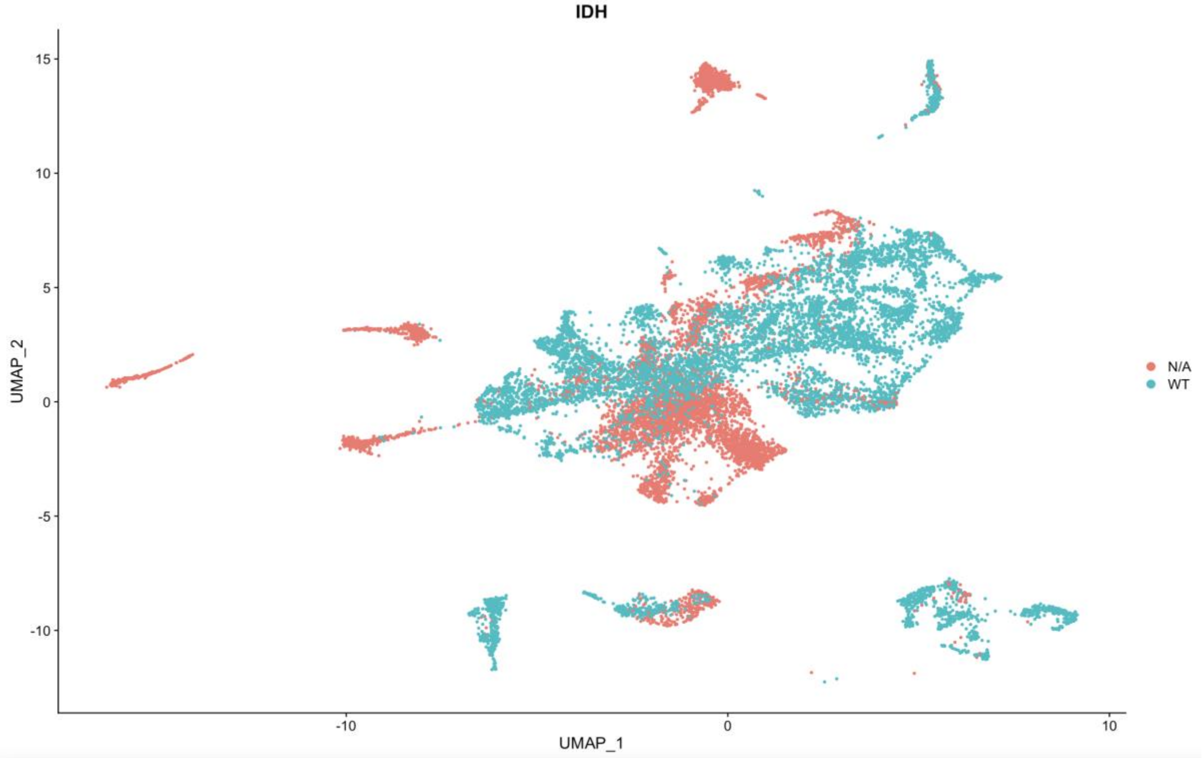

Figure 19 | UMAP of cells labeled for IDH mutation type in Venteicher 2017. The metadata provided by the authors was incomplete and only provided IDH information for half of the dataset, all of which were WT.

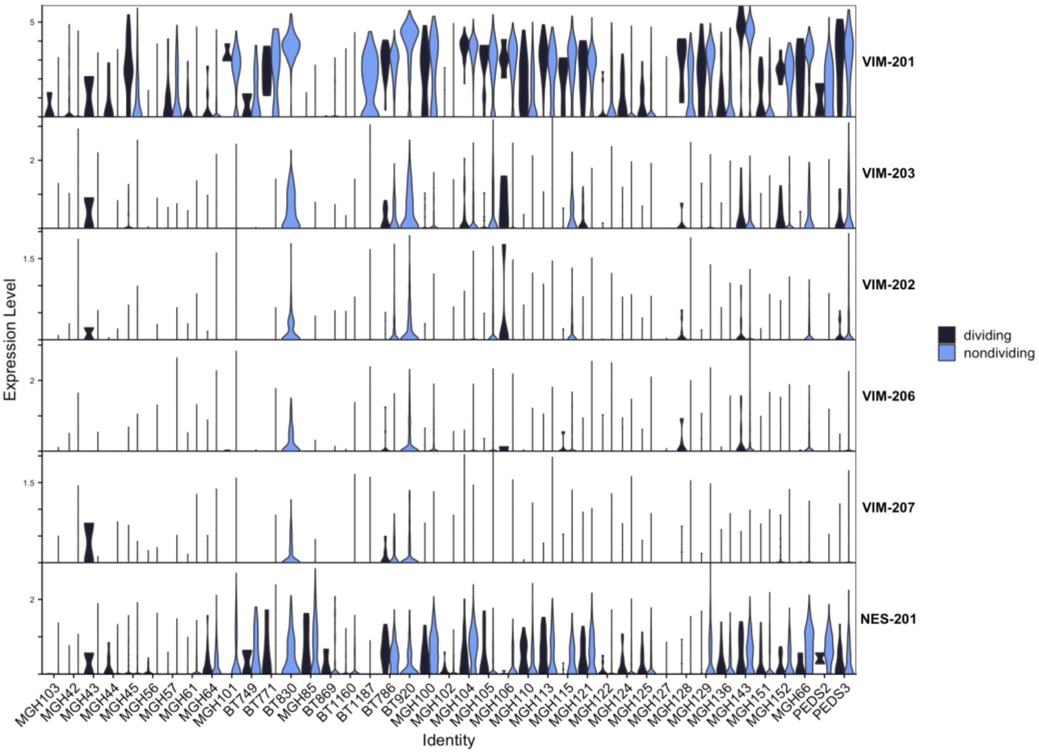

Figure 20| Distribution of expression levels of intermediate filament isoforms of Vimentin and Nestin for Venteicher 2017.

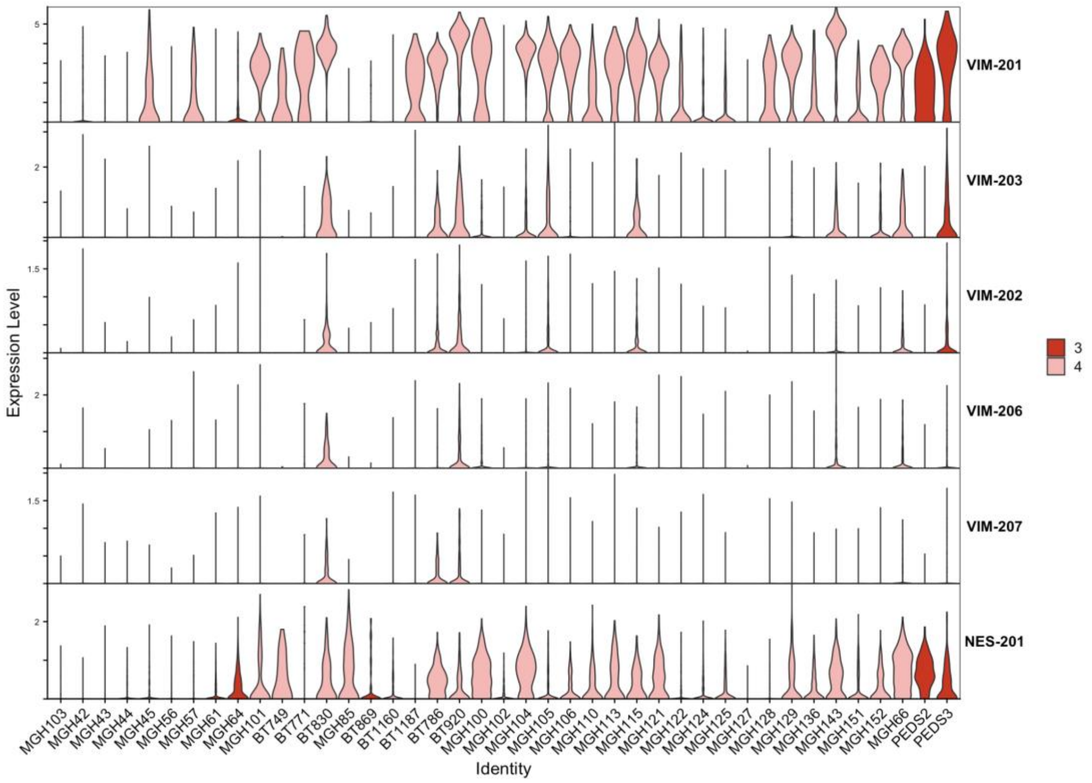

Figure 21 | Distribution of expression levels of intermediate filament isoforms of Vimentin and Nestin for Venteicher 2017. The expression is measured per tumor which is classified as either grade 3 or 4 glioma.

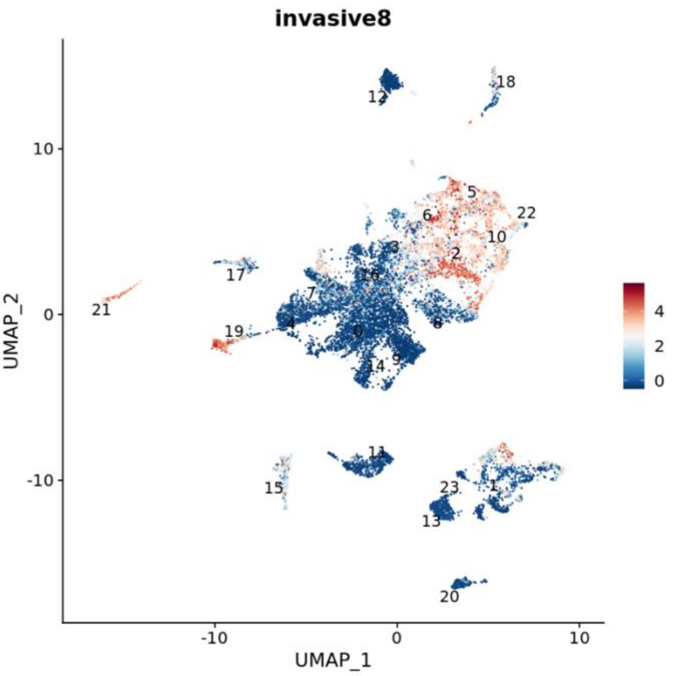

Figure 22 | Invasion score of Vimentin isoform VIM-201 in Venteicher 2017 dataset. One of the top 10 highest invasion isoforms based on expression levels.