Re-convolving the compositional landscape of primary and recurrent glioblastoma using single nucleus RNA sequencing

Osama Al-Dalahmah[1,2], Michael Argenziano[3], Deborah Boyett[3], Akshay Save[3], Aayushi Mahrajan[3], Fatima Khan[1], John F. Tuddenham[5], Alexander R. Goldberg[1], Athanassios Dovas[1], Matei A. Banu[3], Julia Furnari[3], Tejaswi Sudhakar[3], Erin Bush[4], Andrew B. Lassman[2,5], Jeffrey N. Bruce[2,3], Peter A. Sims[2,4], Vilas Menon[5], Peter Canoll[1,2]

[1]Department of Pathology and Cell Biology, Columbia University Vagelos College of Physicians and Surgeons (VP&S), New York, NY, USA
[2]Herbert Irving Comprehensive Cancer Center, Columbia University Medical Center, New York, NY, USA
[3]Department of Neurological Surgery, Columbia University Vagelos College of Physicians and Surgeons (VP&S), New York, NY, USA
[4]Department of Systems Biology, Columbia University Vagelos College of Physicians and Surgeons (VP&S), New York, NY, USA
[5]Department of Neurology, Columbia University Vagelos College of Physicians and Surgeons (VP&S), New York, NY, USA

Running title
Re-convolving the compositional landscape of GBM

Corresponding Authors:

Osama Al-Dalahmah, Email: os2298@cumc.columbia.edu

Vilas Menon, Email: vm2545@cumc.columbia.edu

Peter Canoll, Email: pc561@cumc.columbia.edu

Abstract

Glioblastoma is an aggressive diffusely infiltrating neoplasm that spreads beyond surgical resection margins, where it intermingles with non-neoplastic brain cells. This complex tissue harboring infiltrating glioma and non-neoplastic brain cells is the origin of tumor recurrence. Thus, understanding the cellular and molecular features of the glioma margin is therapeutically and prognostically important. Here, we used single-nucleus RNA sequencing (snRNAseq) of primary and recurrent glioma to define compositional 'tissue-states' that correlate with radiographic and histopathologic features. We found that glioma cells can be clustered into proliferative, astrocyte-like/mesenchymal, and progenitor-like/proneural states in both the primary and post-treatment recurrence settings. We focused on non-neoplastic microenvironment cells including oligodendrocytes, myeloid cells, neurons, and astrocytes - the latter two are under-represented in single-cell RNAseq studies. Cell type-specific signatures of the astrocyte-like/mesenchymal glioma, and a subpopulation of non-neoplastic astrocytes correlated with poor prognosis, the latter correlated with glioma recurrence. Notably, astrocytes were enriched for metabolic and neurodegenerative gene signatures. Leveraging snRNAseq-derived compositional information, we define three tissue-states that correlate with radiographic localization of primary and recurrent glioma. Our findings define a compositional approach to the glioma microenvironment and reveal prognostically and anatomically relevant features paving the way to new mechanistic and therapeutic discoveries.

Introduction

Glioblastoma (GBM) is the most malignant glial tumor of the brain and is refractory to current treatment. Although gross surgical resection of the visible tumor is sometimes feasible, glioma cells infiltrate the brain beyond the resection margins. Most studies characterizing the molecular and cellular composition of GBM have focused on the highly cellular core of the tumor in the context of primary disease, and relatively little is known about the cellular and molecular microenvironment of the tumor margins and post-treatment recurrence. Previous studies have shown that most cells at margins are not tumor cells but rather non-neoplastic brain cells, including neurons and glia, that are reacting to the presence of the tumor and thus are not in their normal states [1]. This is the microenvironment into which tumor cells migrate and proliferate, leading to recurrence, and is also the target of adjuvant therapy. Thus, understanding the cellular milieu of the tumor-margin and recurrence, including both neoplastic and non-neoplastic cells, is vital for advancing the management of GBM.

Early studies used bulk RNA sequencing approaches to understand GBM states in MRI-localized samples from contrast-enhancing (CE) and non-contrast enhancing (NCE) margins [2-5]. More resolution is attained using single cell RNAseq (scRNAseq) approaches, which are being increasingly used to understand heterogeneity in gliomas. Several studies have employed scRNAseq from freshly resected surgical samples to explore the heterogeneity of GBM [6-12]. These studies have significantly advanced our understanding of the heterogeneity and pathology of glioma. However, application of whole cell scRNAseq is faced with practical challenges related to the limitations of acquiring and processing freshly resected glioma tissue and the technical incompatibility with banked frozen glioma tissue. Moreover, scRNAseq is limited in sampling non-neoplastic cells of the microenvironment like neurons and astrocytes, which are major constituents of the tumor-margins [6, 8-11], in part because of cell-type survivability/selection bias during tissue dissociation. Thus, while advances have been made in

defining the genetic alterations in recurrent glioma [13, 14] and features of myeloid cells [15], comprehensive analyses of cellular composition and phenotypic alterations at the margins of primary and recurrent gliomas remain a challenge.

Here, we circumvented these limitations of scRNAseq by using single nucleus RNAseq (snRNAseq), which allowed us to analyze frozen tissue, and inclusively sample cells of the microenvironment from primary and recurrent glioma. Transcriptional analysis of copy number variations (CNVs) provided a metric to distinguish neoplastic (CNVpos) and non-neoplastic (CNVneg) nuclei, and unbiased clustering revealed that primary and recurrent tumors harbor CNVpos glioma cells with similar transcriptional states. Conversely, the microenvironment of primary and recurrent glioma displayed distinct cell-type specific states and different compositional landscapes. Leveraging information from the snRNAseq-derived compositional make-up of glioma-infiltrated samples defines three generalizable "tissue states" that can be identified in more accessible bulk RNAseq samples. We applied this approach to published datasets of spatially localized bulk RNAseq samples to further define the patterns of cellular composition and cellular phenotypes in the core and margin of primary and recurrent GBM.

RESULTS

Single nucleus RNAseq reveals proliferative, astrocyte-like/mesenchymal, and progenitor-like/proneural states in both primary and recurrent GBM

Radiographically, GBM typically has a CE core surrounded by a non-enhancing margin that is highlighted by FLAIR-signal abnormality by MRI (**Figure-1A**). The histopathological features of the resected tumor can vary from highly cellular tumor with vascular proliferation to less cellular infiltrated brain. These features are shown in **Figure-S1H**, demonstrating samples with a cellular GBM core (red star in **Figure-1A, Figure-S1H PA1, PA2, PA3, and PO2_1**) and others with overlying cortex (green star in **Figure-1A, Figure-S1H PO2_2 and PO1**), which we use below.

To explore the heterogeneity of primary GBM, we analyzed several banked surgical samples using snRNAseq as shown in (**Figure-1A**). A total of 8 samples from 7 patients were selected for analysis (**Supplementary Table-1**). Neuropathological assessment of tumor cellularity ranged from cellular tumor with hallmarks of GBM, to reactive brain parenchyma with few atypical cells. This assessment was made on Hematoxylin and Eosin (H&E) stained formalin fixed paraffin embedded sections adjacent to or frozen cryosections of the frozen tissue analyzed by snRNAseq (**Figure-S1H)**. We isolated nuclei from the frozen tissue and subjected them to snRNAseq followed by downstream analyses including clustering, differential gene expression analysis, cluster marker detection, and gene set enrichment analysis (GSEA) as outlined (**Figure-1A**). 15189 nuclei passed our QC (**Supplementary Table-1**). To distinguish putative glioma cells from non-neoplastic cells, we employed an established approach that infers large scale copy number alterations/variations (CNV) from RNA expression profiles [16]. Chromosomal heat maps showing putative neoplastic nuclei are shown in **Supplementary Figure-S1A-G**. Next, we also applied a second method to label nuclei based on a "malignancy score", which we have previously shown to be a robust metric to distinguish glioma cells from non-neoplastic cells [6, 10], and the consensus nuclei designated by both methods was used for downstream analysis. Nuclei with no consensus CNV status were excluded (4.7%). Uniform

manifold approximation and projection (UMAP) plots from individual cases labeled by transformation status are shown in **Figure-1B**. We identified 7954 putative neoplastic nuclei with inferred large scale chromosomal CNV (CNVpos/glioma nuclei). Glioma nuclei showed multiple chromosomal alterations including gains of chromosome 7 and losses of chromosome 10 (**Figure-S1**). Having identified neoplastic and non-neoplastic nuclei, we aligned the datasets from multiple samples and performed clustering analyses separately on CNVpos (glioma) nuclei from all cases using shared nearest neighbor and the smart local moving algorithm [17]. A UMAP plot is shown for all primary glioma nuclei non-neoplastic nuclei color-coded by glioma state/lineage **Figure-1C**. This approach identified 6 distinct clusters: these resembled progenitors (oligodendrocyte-progenitors (gl_PN1 - proneural) and neural-progenitors (gl_PN2 - proneural), astrocytes (gl_Mes1 and gl_Mes2 - mesenchymal), and proliferative cells (gl_Pro1 and gl_Pro2).

The identity of the glioma states is akin to previously described glioma states, as demonstrated by the enrichment of several gene lists from [2, 9, 18, 19] **– (Figure-1D, supplementary Table-2)**. For example, gl_Pro1 and gl_Pro2 showed enrichment in gene sets specific for cell-cycle phases [19], with gl_Pro1 showing highest enrichment of G2/M genes (Gobin_G1) and gl_Pro2 showing highest enrichment of G1/S phase genes and DNA repair related genes (Gobin_G3). Clusters gl_PN1 showed enrichment of the Verhaak's proneural, and OPC signature genes, while gl_PN2 showed enrichment of NPC signature genes. Finally, gl_Mes1 showed enrichment of astrocyte-like signatures and Verhaak's classical signature while cluster gl_Mes2 showed enrichment of several gene sets related to reactive astrocytes, and Verhaak's mesenchymal signature [2, 7, 18]. Our clustering is consistent with that described in Neftel et al. 2019 [9] and Wang et al. 2019 [7], and the states we describe are compatible with those in Yuan et al 2018 [6]. To further clarify the cellular phenotypes represented in our glioma clusters, we measured the enrichment of the major biologic process and molecular function gene ontologies (GO) in the glioma state top gene markers (see methods)**.** GO enrichment analysis demonstrated enrichment of GO's relating to locomotion, neurogenesis, neuronal migration, and cell projection in gl_PN1 markers genes; Notch signaling, neuron development, and GABA reuptake differentiation, and synaptic signaling in gl_PN2 genes; response to organic substances, ion homeostasis, and Signaling by tyrosine kinases in gl_Mes1 genes; response to cytokines, interferon gamma, and leukocyte activation and immune response in gl_Mes2 genes; mitosis and nuclear division in gl_Pro1, and S-phase, DNA replication, and DNA repair in gl_Pro2(**Figure-1F** and **Supplementary Table-2**). The identities of the clusters can also be appreciated by examining select gene markers **Figure-1E and Supplementary Table-2.** gl_Pro1 expressed cell-cycle genes TOP2A, CENPF, and AURKB. gl_Pro2 showed highest expression of DNA damage/repair including FANCI, HELLS, and XRCC2. gl_PN2 showed high levels of CD24, MEG3, and SOX4. gl_Mes1 showed high levels of protoplasmic astrocyte genes including SLC1A3, LIFR, ATP1A2, C1orf61, and NTM, while gl_Mes2 showed highest expression levels for reactive astrocyte and immune genes including CLU, VIM, and SAT1. While our glioma states resemble those described in the literature, less is known about whether glioma cells assume similar states in the recurrent setting. Therefore, we bridged this gap by directly analyzing recurrent IDH-WT glioma samples using the same approach we used for primary GBM samples.

To define the states of IDH-WT glioma in the post-treatment recurrence setting, we analyzed 8 cases of post-recurrent IDH-WT glioma using snRNAseq (**Figure-2A**). We identified 8908 neoplastic nuclei harboring large-scale CNV (**Figure-S2**). Of the eight cases, two were

paired recurrences from the primary samples (TB5124 – recurrent of TB4916, and TB5053 – recurrent of TB4718, see supplementary results). We treated recurrent gliomas similarly to the treatment naïve primary tumors and clustered all neoplastic nuclei together. Like primary gliomas, we found that recurrent glioma clusters can be assigned two proneural, two mesenchymal, and two proliferative states (**Figure-2B**). The gene markers of the recurrent glioma states are enriched for similar ontologies to those seen for primary glioma states (**Figure-2C** and **Supplementary Table-3**), showed similar patterns of enrichment for the previously presented gene sets in **Figure-1D** (Figure-**2D**), and displayed comparable gene marker expression (**Figure-2F**). These results demonstrate that post-treatment recurrent glioma states closely resemble states observed in the primary pre-treatment setting. Indeed, Pearson correlation analysis demonstrates that corresponding states were positively correlated (**Figure-2E**). The correlation patterns reveal that gl_Mes1 and gl_Mes2 are positively correlated with each in the primary and recurrent settings. This is also seen with gl_PN1 and gl_PN2, as well as gl_Pro1 and gl_Pro2. We therefore contend that a view of primary and recurrent glioma states may benefit from simplification and embrace a viewpoint that primary and recurrent glioma states can be classified as progenitor-like/proneural (gl_PN1 and gl_PN2), astrocyte-like/mesenchymal (gl_Mes1 and gl_Mes2), and proliferative (gl_Pro2 and gl_Pro1) states. We note that astrocyte-like/mesenchymal glioma states were negatively correlated with proliferative states. Consistent with this result, our ISH findings demonstrated a significantly smaller proportion of CLU+ cells that co-expressed TOP2A (mean=31.71388837%, Standard deviation = 15.73850618, one-tailed t-test p= 0.000249641, n=5, **Figure S5A-B**).

While the transcriptional signatures of glioma are relatively well defined, the spatial distribution of these glioma states is less well understood. Given the marked difference in cellular composition between the cortex and the deeper (typically more heavily infiltrated) white matter, and the highly cellular tumor core, we asked if these different anatomic regions harbor distinct glioma states. In other words, we posited that the cellular microenvironment of glioma influences glioma states. Specifically, we hypothesized that we would find more glioma cells that resemble astrocytes (astrocyte-like/mesenchymal glioma) or neurons (progenitor-like - specifically gl_PN2) in the cortical margins. To address this question, we examined the expression of select combinations of glioma state transcripts using in situ hybridization (ISH) across the cellular tumor and the infiltrated cortical margin. We used probes to detect PTPRZ1 (high in glioma), CLU (high in astrocytes and astrocyte-like/mesenchymal glioma), SOX2 (high glioma), NOVA1 (high in progenitor-like/proneural glioma), and MEG3 (high in neurons and progenitor-like/proneural glioma - gl_PN2) in the cellular core and overlying infiltrated cortical margin in 5 cases of primary GBM (**Figure-3A, C**). We found that significantly higher proportion of PTPRZ1+ glioma cells co-expressed CLU in the cortex versus the core (**Figure-3B**). Similarly, we found that significantly higher proportion of SOX2+NOVA1+MEG3+ glioma cells in the cortex versus the core (**Figure-3D**). These findings indicate that the different glioma states have distinct distributions throughout the landscape of glioma and suggest that local tissue cellular composition and perhaps other microenvironmental influences can affect glioma states.

Transcriptional analysis of the glioma microenvironment reveals prognostically significant subpopulations of non-neoplastic astrocytes

Given the importance of glioma microenvironment in tumor progression, we decided to investigate the implications of microenvironmental states on the prognosis of GBM. To achieve that, we combined in our analysis CNVneg nuclei from primary and recurrent glioma, as well as nuclei from low-grade glioma (LGG) and epilepsy. This spectrum of disease states, all of which influence the brain microenvironment, can give us insight into the alterations in cellular and tissue states that correlate with tumor progression. Our CNVneg nuclei datasets included 16831 nuclei: 6929 from primary glioma, 6008 from post-treatment recurrent glioma, 2875 from epilepsy, and 1019 from LGG. We projected these nuclei in UMAP space and assigned cell lineages as shown in **Figure-4A**. The expression of a select number of marker genes per lineage is shown in **Figure-S5C**. We present the results on myeloid lineage nuclei in the supplementary results (**Figure-S6**), which demonstrates that monocyte-derived tumor-associated macrophages (TAMs) were enriched in recurrent glioma, while microglia-derived TAMs were enriched in primary glioma, consistent with a previous report[15].

We focused on astrocytes, which are key elements of the glioma microenvironment and are not well represented in glioma single-cell RNAseq datasets [6-12, 20]. A recent landmark paper implicates GBM-associated astrocytes in promoting an immunosuppressive microenvironment [21]. Moreover, the distinction between tumor-astrocytes and reactive-astrocytes is of major diagnostic importance in neuropathology. Thus, we first analyzed astrocytes (707 nuclei – 284 from primary glioma, 254 from recurrent glioma, 45 from LGG, and 121 from epilepsy) in isolation from other cell types, reduced their dimensions, and clustered them using the shared nearest neighbor algorithm into 9 clusters **Figure-4B.** Clusters 1 and 8 showed expression of many of the protoplasmic genes such as SLC1A2 and GLUL, whereas clusters 6 and 7 showed high expression of reactive genes including CLU, LGALS3, and VIM. One cluster was characterized by expression of ribosomal genes (Cluster 3), while another had a hypoxic signature characterized by expression of VEGFA and HILPDA (Cluster 5). While an over-representation of ribosomal genes may be a sign of low-coverage, we filtered cells with less than 10,000 reads and 400 detected genes and thus interpret this ribosomal gene signature as of potential biologic relevance - future studies will investigate that further. Clusters 2, 4, and 5 were each characterized by expression of genes associated with other lineages including MEG3 and PDFGRA (Cluster 5), GRIA1, TNR (Cluster 4), and PLP1 as well as SNAP25 (cluster 2). We previously showed that reactive astrocytes can mis-express non-lineage genes in neurodegenerative setting [22]. Cluster 0 did not show many differentially expressed genes and is thus considered a baseline cluster (**Figure-4D)**. The astrocyte cluster markers are provided in **Supplementary Table-4**. These astrocyte clusters were distributed between primary glioma, recurrent glioma, LGG, and epilepsy (**Figure-4C**), but not equally so. Specifically, clusters 2, 3, and 5 were mostly represented in primary glioma, clusters 6 and 7 in recurrent glioma, and cluster 1 in epilepsy. Moreover, cluster 0 was represented in all glioma conditions, mostly in primary and recurrent gliomas and least in epilepsy, while cluster 8 was represented in recurrent glioma and epilepsy (**Figure-4E**). Because cluster 6 shows enrichment for several genes previously associated with poor glioma outcomes, including CD44, CHI3L1, LGALS3, CLU and APOE ([23-25]), we asked if the gene signature for cluster 6 (**Supplementary Table-4**) has prognostic relevance in survival data associated with publicly available datasets. We performed GSEA of the top gene markers for astrocyte clusters, using the TCGA-GBM [26] and the Chinese Glioma Genome Atlas (CGGA) databases [27, 28]. 318 IDH-WT samples (TCGA - 139, CGGA - 179) were used for survival analysis. Enrichment of the gene signature for astrocyte cluster 6 is significantly associated with poor survival (p=0.007 negative log rank test - **Figure-**

4F) on univariate analysis by the Kaplan-Meier method. To investigate the relationship between survival and enrichment of each astrocyte cluster top gene markers, we determined the hazard ratio of death given enrichment of each gene set using the cox proportional hazard regression model, accounting for age as a covariate (**Figure-4G**). Only positive enrichment of cluster 6 genes was significantly associated with increased hazard ratio of death by more than 1.5 fold (confidence interval 1.08-2.20, p.value 0.015 **Figure-4G**), an effect independent of age. Of note, astrocyte-like/mesenchymal glioma-states signatures were also significantly associated with poor survival (**SFigure-S7**). Since astrocytes and glioma shared gene signatures (for example, CLU expression in gl_Mes2 and Astrocyte cluster 6), we performed differential gene expression analysis between primary and recurrent glioma non-neoplastic astrocytes and all CNVpos glioma nuclei and identified 1456 and 1620 genes were higher in astrocytes compared to glioma and 3380 were higher glioma compared to astrocytes. Examples of genes higher in non-neoplastic astrocyte include genes associated with Alzheimer's disease (CLU, APOE), metallothionein genes (MT1H, MT1G, MT1M, MT1F, MT1E, MT1X, MT2A, and MT3 – increased in reactive astrocytes), Synuclein genes (SNCA, SNCB, and SNCG), WIF1, CHI3L2 (associated with poor prognosis in glioma[29]), ALDOC, ALDOA, AQP4, carbonic anhydrases CA2 and CA11, and CXCL14, a cytokine implicated in promoting glioma invasion [30] (**Supplementary Table-4)**. Conversely, genes higher in glioma include EGFR, PTPRZ1, NOVA1, CD24, Nestin (NES), SOX5, and SOX4. We used KEGG pathway enrichment analysis to query the function of these genes (**Figure-4H**). Further analysis of the differentially expressed genes showed that several KEGG pathways were enriched in genes higher in non-neoplastic astrocytes, with some relating to neurodegeneration such as Parkinson disease, prion disease. Notably, these signatures are highly enriched in oxidative phosphorylation genes (**supplementary Table-4**), which is dysregulated in neurodegenerative diseases [31]. Moreover, other metabolic pathways enriched in astrocyte DGE included metabolism of fatty acids, glycolysis, TCA cycle, and ferroptosis. Conversely, KEGG pathways increased in tumor-astrocytes were largely related to DNA replication, cancer-related pathways including ERBB and MAPK signaling, as well as cell-cycle **Figure-3H** and **Supplementary Table-4**.

Re-convolution of snRNAseq identifies three tissue states based on cellular composition of glioma and its microenvironment

Given the heterogeneity of cellular states of glioma and non-neoplastic cells in the glioma microenvironment, we asked whether specific myeloid and astrocyte states are correlated with glioma states. We first asked if specific glioma, or brain microenvironment lineages were differentially abundant or deplete in recurrent glioma using a regression model [32] to test for differential abundance (**Figure-5A**). The results showed that gl_Mes2 were significantly more abundant in recurrent glioma, while gl_PN1 were more abundant in primary glioma, (Benjamini-Hochberg adjusted p values (q-value) 3.99e-2 and 1.318e-5 respectively). Of the microenvironment lineages, OPCs were significantly more abundant in primary glioma (q-value 1.085e-03).

To uncover patterns of 'tissue states" with correlated cell states/lineages, we took advantage of the relatively unbiased sampling of cellular composition in the brain tumor microenvironment provided by snRNAseq. We approximated the cellular composition of each surgical sample by recombining the cells from all the distinct cell populations, as identified by snRNA-seq, to create a compositional matrix containing the abundance of all cell types across

all samples (**Supplementary Table-1**). For the cellular composition matrix (**Supplementary Table-1**), we used three major astrocyte clusters (Ast1-3) and five immune states (Myel1, moTAM, mgTAM, prTAM, and T cells) as described in the supplementary results and methods. We then used principal component analysis of the resulting cellular composition matrix and identified the compositional features that account for the variance across the samples (**Figure-5B**). The results showed that the relative abundance of CNVpos glioma cells versus CNVneg non-neoplastic cells (neurons, oligodendrocytes, OPCs) is the major feature of the first principal component, and the abundance of reactive astrocytes (Ast3), macrophage-like myeloid cells (moTAM), and T cells is the major feature of the second principal component. Notably, the abundance of a specific subpopulation of mesenchymal glioma cells (gl_Mes2) was also highly correlated with the second principal component (PC2) These finding indicate that specific subpopulations of neoplastic and non-neoplastic cells tend to co-inhabit glioma samples. Additionally, PC2 was positively correlated with the recurrent glioma condition (p.value= 0.027). To further explore this prognostic relevance in IDH-WT GBM survival TCGA and CGGA dataset, we performed a log-rank test on samples with positive versus negative enrichment of the PC2 signature and found that positive enrichment is significantly associated with poor survival (**Figure-5C**). We also determined the hazard ratio of death given PC2 signature using the cox proportional hazard regression model, considering age and MGMT methylation as a covariate and found PC2 signature is associated with increased hazard of death (**Figure-S7A,** hazard ratio 1.4 (CI 1.07-1.9), p value 0.015), an effect independent of age and MGMT methylation status. This data show that examining “tissue-states” as defined by cellular composition, reveals prognostically-relevant gene signatures that span across both neoplastic and non-neoplastic cell states that co-inhibit tissue samples.

Driven by the above findings, we clustered the snRNAseq samples into 3 distinct “tissue-states” based on these approximated cellular compositions; **tissue-state A** samples are predominantly composed of non-neoplastic brain cells, including neurons oligodendrocytes, and OPCs**, tissue-state B** samples are enriched in reactive astrocytes, myeloid/macrophages, and T-cells, and **tissue-state C** samples are predominantly composed of CNVpos glioma cells (**Figure 6A-B and Figure S10**). To generate a gene signature for each tissue state, we combined the snRNAseq for all nuclei in each sample and performed differential gene expression analysis between tissue-state clusters, using the re-convolved expression profile of each sample as a biological replicate. This analysis identified the top-differentially expressed genes unique to each tissue state (**Supplementary Table-7**). To assess the generalizability of the three tissue-states, we performed single sample GSEA analysis for the tissue state gene signatures using a dataset of bulk RNAseq analysis performed on 91 primary and recurrent MRI-localized samples from 39 patients. We found that these samples separated into 3 compositional clusters based on their enrichment score for snRNAseq-defined “tissue-states” (**Figure 6C**). We refer to the compositional clusters and tissue-states interchangeably henceforth. These three tissue-states are further demonstrated by projecting the RNA-expression levels for canonical markers of the predominant cell types for each tissue-state in **Figure 6E** showing RBFOX3 (neuronal marker) in tissue-state A, CD68 (myeloid marker) in tissue-state B, and MKI67 (proliferation marker) in tissue-state C. Sox2 (pan-glioma marker) was widely distributed across the samples, indicating variable degrees of tumor infiltration across samples in all three tissue-states. . Further analysis revealed that these tissue-state gene signatures are enriched for specific biologically relevant functional ontologies. For example, Tissue-state A is enriched of genes involved in synaptic transmission, tissue-state B is

enriched for genes associated with oxidative stress and inflammation, and tissue state C is associated with cell proliferation (**Figure 6D**). To further validate these findings, we quantified total cellularity and the IHC labeling indices SOX2, NeuN, CD68, and Ki67 in 45 recurrent and primary glioma samples and found that total cellularity was highest in cluster C, which also had the highest abundance of SOX2+ and Ki67+ cells, while cluster A had the highest abundance of NeuN+ cells, and Cluster B had the highest abundance of CD68+ cells. While Clusters A and B resemble normal and reactive brain tissue, the SOX2 and Ki67 labeling indices indicate that these clusters comprise samples with variable levels of glioma infiltration. Together, these findings show that samples can be classified into 3 tissue-states/clusters, each characterized by their predominant cellular constituents, that correspond to infiltrated brain (tissue-state A), and reactive/inflammatory tissue (tissue-state B) and cellular/proliferative tumor (tissue-state C).

Glioma-associated tissue states are associated with distinct metabolic states driven by lineage composition

Unbiased analysis of enrichment of KEGG pathways in genes differentially expressed between compositional clusters/tissue-states reveals they exhibit enrichment of multiple unique and specific pathways, supporting the existence of tissue state-specific signaling pathways (**Figure S8A**). Interestingly, several of the tissue state-enriched pathways are metabolic pathways. For example, tissue-state A showed highest enrichment for oxidative phosphorylation and beta-alanine metabolism, whereas tissue-state C was most enriched for pyrimidine, folate, and branched chain amino acid metabolism. Conversely, tissue-state B showed highest enrichment of fatty acid oxidation and glycolysis (**Figure S8B**). We reasoned this could be driven by the expression of pathway genes in specific cell types that compose the tissue-state. To explore this possibility, we ranked the averaged then scaled normalized gene-expression per lineage and generated a modified average rank lineage-distributed pathway enrichment score (**Figure S8C**). The results show that cell-cycle and other glioma-related pathways are driven by genes highest in glioma, Taurine metabolism and calcium signaling are driven by genes highest in neurons, and PPAR and glutathione metabolism are driven by genes highest in astrocytes and myeloid cells. As an example, the normalized expression of genes in the glutathione pathway is shown in **Figure S8D.** Notably, while the enrichment of the glutathione pathway was highest in astrocytes and myeloid cells, the expression of the pathway genes was distributed across multiple cell types, suggesting that the metabolic status of a tissue can have distinct, but functionally related effects on different cell types in that tissue. Overall these results demonstrate that tissue-states exhibit lineage-driven enrichment of metabolic pathways.

The spatial landscape of glioma associated tissue-states in primary and recurrent GBM

To understand the spatial landscape of primary and recurrent glioma, we mapped the distribution of our “tissue-state” signatures in space in primary and recurrent GBM. First, we tested one of our cases that we utilized for snRNAseq (PO2) and took 48 localized biopsies that we analyzed using plate-seq [33]. Immunofluorescence of frozen sections taken prior to analysis revealed a cellular DAPI-dense glioma core and a NeuN rich cortical margin (**Figure S9A**). We conducted GSEA analysis of our tissue-state signatures in the RNAseq data from the localized biopsies and mapped that against the location of the biopsies (**Figure S9B**). Tissue-state C signature was highest in the core, compared to tissue-state A signature, which was highest in the cortical margin. Tissue-state B signature showed a more patchy distribution with foci of

enrichment in both the core and margin. Interestingly, the intermediate region between the core and cortex, showed mixed enrichment across all three tissue-states. This data highlights the anatomic localization of tissue-state signatures and underscores the heterogeneous patterns in the intermediate non-cortical "margin" region. Next, to assess the generalizability of these results, we conducted the same enrichment analysis on a dataset of MRI-localized primary and recurrent GBM samples [2] (**Figure S9C**). As expected, control brain samples showed high enrichment of tissue-state A signature while the contrast-enhancing samples in primary GBM showed high enrichment of tissue-state C signature. Notably, the contrast enhancing samples from post-treatment recurrent GBM showed highest enrichment of tissue states B and C, indicating a mixture of recurrent tumor and treatment effect. The FLAIR+ primary GBM samples showed a wide distribution of enrichment of all three tissue types while the FLAIR+ recurrent GBM samples showed enrichment of tissue-states A and B, with only a few samples showing enrichment of tissue state C. This data indicates that the non-enhancing margins of recurrent GBM samples predominantly represent reactive/gliotic brain tissue with relatively low levels of tumor infiltration, whereas the non-enhancing margins of primary GBM can contain a wider range of pathological features, including regions of abundant glioma infiltration.

DISCUSSION

In this work, we investigated the transcriptional landscape of glioma and its microenvironment in primary and post-treatment recurrent IDH-WT glioma using snRNAseq. Understanding heterogeneity of GBM is important for guiding treatment and meeting the challenge of recurrence. Furthermore, understanding cellular heterogeneity at the margins of glioma is especially important given that surgical resection of the contrast enhancing tumor core inevitably leaves behind infiltrated cortex and white matter in the resection bed which are largely composed of non-neoplastic cells. Recent studies revealed a diversity of glioma states that resemble cell lineages found during development and adulthood [6, 9-12, 20]. Our study provides novel insight into glioma states in post-treatment recurrent GBM, which can be classified, as with primary GBM, into astrocyte-like/mesenchymal, progenitor-like/proneural, and proliferative states that correlate with their primary GBM counterparts. The comprehensive and relatively unbiased sampling of different cell types also provide novel insight into the cellular and transcriptional landscape of primary and recurrent GBM microenvironment, including astrocytes, which are poorly represented in single cell RNAseq datasets [11]. Importantly, we discovered that select gene signatures of both neoplastic and non-neoplastic cells were prognostically relevant and detail a prognostically relevant astrocyte signature that is correlated with myeloid and mesenchymal glioma states. We demonstrate that, leveraging insight into correlated cellular states and lineages that co-inhabit tissue samples identifies signatures we use to classify primary and recurrent GBM tissue into three tissue states: infiltrated brain, reactive/inflammatory tissue, and cellular/proliferative tumor. We find that these tissue states can also be identified in more accessible bulk RNAseq samples and correlate with immunohistochemical profiles and metabolic signatures. We used this approach to classify radiographically localized primary and recurrent GBM samples and highlight margin-specific tissue states. This is significant given that transcriptional signatures in the glioma-infiltrated tissue correspond to compositionally defined tissue-states with distinct biological / metabolic features that likely correspond to specific therapeutic vulnerabilities. The therapeutic implications of our findings help expand the target of

therapy from targeting one gene or one cell type, to targeting to tissue states comprising cell populations that co-inhabit the tissue under defined metabolic constraints.

Our analysis of glioma states revealed they can be grouped into three main categories based on commonalities with previously described glioma-state signatures: astrocyte-like/mesenchymal (gl_Mes1 and gl_Mes2), progenitor (gl_PN1 and gl_PN2), and proliferative (gl_Pro1 and gl_Pro2). We found that the astrocyte-like/mesenchymal signatures (gl_Mes1 in primary GBM and gl_Mes2 in recurrent GBM) portended poor survival in the IDH-WT GBM survival series (TCGA and CGGA), consistent with previous studies [5]. Moreover, GBM recurrence was correlated with increased gl_Mes2 and reduced gl_PN1 abundance in our dataset. This is consistent with results from bulk RNAseq on paired primary-recurrent glioma pairs, where the authors found enrichment of the mesenchymal subtype, and depletion of the classical subtype [5]. The classical subtype resembles our gl_PN1, which showed positive enrichment scores of the Verhaak's classical gene set. Our gl_PN1 was indeed depleted in the recurrent setting. Of the non-neoplastic cell types, OPCs were depleted in recurrence. This may be explained by the fact the OPCs are the proliferative cell type in the brain and glioma treatment with chemotherapy and radiotherapy depletes proliferative cells, as have been previously demonstrated[34]. That being stated, our dataset is based on largely unpaired snRNAseq from primary and recurrent GBM and is of relatively smaller size, which may preclude us from further interpretation.

Our analysis of the cellular phenotypes in the glioma microenvironment revealed that subpopulations of non-neoplastic astrocytes show enrichment for abnormal transcriptional signatures that are also seen in the context of neurodegenerative diseases. In contrast to CNVpos neoplastic astrocytes, which express high levels of proliferation and glioma genes, a subpopulation of non-neoplastic astrocytes (Ast3) displayed a reactive signature reminiscent of astrocytes described in neurodegenerative diseases Huntington disease, Parkinson disease and Alzheimer's disease cortex [22, 35]. It is notable that the non-neoplastic glia in the microenvironment of glioma exhibited neurodegenerative signatures even though most of the cases we profiled were less than 65 years old - the age after which neurodegenerative diseases become more prevalent. Our findings thus uncover commonalities in the astrocytic response to glioma and neurodegeneration. Given that GBM is a disease of the elderly population, one can envisage that future astrocyte modifying therapies developed for neurodegenerative disease may have a role in therapeutically modifying the microenvironment of GBM.

One of the main results highlighted by our analysis of cellular composition is that specific cell types are correlated with each other, indicating that they co-inhabit the same tissue-states. For example, Ast3, gl_Mes2, moTAM, and T-Cells were all correlated with the second principal component that account for the variation in sample composition. This axis was positively correlated with the post-treatment recurrence condition. Moreover, the gene signatures for the aforementioned cell types (Ast3, gl_Mes2, moTAM, and T-Cells) are significantly enriched in tissue-state B (**Figure-S10**). Together these findings suggest that the transcriptional phenotypes of both neoplastic and non-neoplastic cells are subject to common microenvironmental influences that may define the tissue state they co-inhabit. Furthermore, our analysis identified therapeutically relevant metabolic pathways that are enriched in specific tissue-states. For example, the glutathione pathway, which determines a cell's sensitivity to ferroptosis-inducing drugs [36] is most enriched in tissue-state B, and is driven by gene expression in astrocytes,

myeloid cells and glioma cells. These finding highlight how combining a tissue-state approach with single cell resolution can reveal new cell type-specific therapeutic targets in GBM.

Methods

Human subjects and glioma tissue

Frozen primary untreated GBM tissue was acquired from the Bartoli brain tumor bank at Columbia University Medical Center. All diagnoses were rendered by specialized neuropathologists. Study protocols were approved by Columbia University Medical Center Institutional Review Board. All clinical samples were de-identified prior to analysis. Analyses were carried out in alignment with the principles outlined in the WMA Declaration of Helsinki and the Department of Health and Human services Belmont Report. Informed written consent was provided by all patients. The demographics of the cases used are provided in **Supplementary Table 1**.

Extraction of nuclei and snRNAseq procedure

Nuclei were isolated from frozen surgical resection specimen as described in Al-Dalahmah O et al. 2020. Briefly, frozen tissue samples measuring from 5 x 4 x 3 mm to 10 x 6 x 3 mm were homogenized using a dounce homogenizer in ice-cold 30% sucrose 0.1% Triton-X 100 based homogenization buffer. 10-15 strokes of the loose dounce pestle were followed by 10-15 strokes of the tight dounce pestle on ice. Mixing using a P1000 pipette followed before filtration through a BD Falcon 40um filters. Filtration was repeated after a 10-minute spin at 1000g at 4c. A cleanup step followed using a density gradient step as described in [38]. The nuclear pellet was suspended in 1% BSA in PBS resuspension buffer containing RNAse inhibitors. A final filtration step using 20um Flowmi ™ filters followed before dilution to 700-1200 nuclei per ul in resuspension buffer. The nuclear suspensions were processed by the Chromium Controller (10x Genomics) using single Cell 3' Reagent Kit v2 or v3 (Chromium Single Cell 3' Library & Gel Bead Kit v2, catalog number: 120237; Chromium Single Cell A Chip Kit, 48 runs, catalog number: 120236; 10x Genomics).

Sequencing and raw data analysis

Sequencing of the resultant libraries was done on Illumina NOVAseq 6000 platformV4 150bp paired end reads. Alignment was done using the CellRanger pipeline (10X Genomics) to GRCh38.p12 (refdata-cellranger-GRCh38-1.2.0 file provided by 10x genomics). Count matrices were generated from BAM files using default parameters of the DropEst pipeline (Petukhov V et al. 2018). Filtering and QC was done using the scater package (3). Nuclei with percent exonic reads from all reads in the range of 25-75% were included. Nuclei with percent mitochondrial reads aligning to mitochondria genes of more than 19% were excluded. Genes were filtered by keeping features with >10 counts per row in at least in 31 cells. Further filtering of low quality cells was done to include cells with at least 400 detected genes and 10,000 reads.

Single Nuclei RNAseq analysis

Sequencing and analysis of raw data

Sequencing of the resultant libraries was done on Illumina NOVAseq 6000 platformV4 150bp paired end reads. We used 10X chromium v2 chemistry for samples PO1 and PO2, and v3 chemistry for samples PA1, PA2, and P3. Read alignment was done using the CellRanger pipeline (v3.1 - 10X genomics) to reference GRCh38.p12 (refdata-cellranger-GRCh38-1.2.0 file provided by 10x genomics). Count matrices were generated from BAM files using default parameters of the DropEst pipeline [39].

Data-cleanup

Filtering and QC was done using the scater package [40, 41]. Nuclei with percent exonic reads from all reads in the range of 25-73% were included. Nuclei with percent mitochondrial reads aligning to mitochondria genes of more than 15% were excluded. Genes were filtered by keeping features with >10 counts per row in at least in 31 cells. The count matrix of each sample was normalized by first running the quickcluster function, then estimating size-factors by calling scran::computeSumFactors() function with default options and clusters set to clusters identified by calling quickcluster function. scater::normalize() function was then used to generated normalized counts. Doublet identification was done using scran::doubletCells function with default options, and cells with doublet score of NMADs > 3 were excluded as we described previously [22].

Combining multiple datasets from different sequencing batches

To control sequencing and technical batches, we utilized canonical correlation analysis in Seurat [42] accounting for batch and mitochondrial read percentage for CNVneg nuclei. For CNVpos nuclei, we accounted for case and mitochondrial read percentage.

Pre-Clustering and clustering of nuclei

Pre-clustering of nuclei was done in Seurat using the shared nearest neighbor smart local moving algorithm. PCA reduction was used as the reduction in the FindNeighbors() step. Pre-cluster identity determination was done using geneset enrichment analysis of lineage markers [22] and by inspecting cluster markers generated by scran::findmarkers(direction="up") function. Microglia +/- oligodendrocytes were used as negative control cell for InferCNV pipeline (below). Once CNVneg cells were verified, cells from all cases were aligned using Seurat and clustered. Clusters with mixed identities based on enrichment of multiple lineage genes were sub-clustered iteratively until all "pre-clusters" showed pure identities. Only then do we combine the pre-clusters of the same lineage into lineages (Astrocytes, neurons, oligodendrocytes, myeloid, endothelial). For subclustering of astrocytes and myeloid cells, we analyzed the nuclei in isolation of other lineages, and re-aligned them in Seurat, and reduced the dimensions before subclustering. For CNVpos nuclei, unbiased clusters were combined into glioma states/lineages based on similarity in marker expression and enrichment for known gene sets described in Figures 1D and 2C.

Count normalization

Raw counts were normalized in Seurat using the sctransform function SCT() function with default settings and controlling for percent mitochondrial gene expression [43].

Copy number variation analysis of snRNAseq

To detect putative neoplastic tumor cells, we used combination of marker expression and large scale copy number variation inference as per the InferCNV R package [16]. We used the default parameter as described in the package documentation. As a control population, we used microglia and oligodendrocytes from case PO2_1. Iteratively, CNVneg clusters including Oligodendrocytes and Neurons were identified and added as control cells. Different gene window sizes were tested (50, 100, 200) and yield similar results. We then applied an orthogonal approach to label putative neoplastic cells based on previous approaches described in [6, 44]. Briefly, Log2+1 counts were averaged across chromosomes for each nucleus. A principal component analysis (PCA) was performed on autosomal chromosomes in factominer R package [45]. Chromosomes with high correlation with PC2 were the same as the ones

shown in the detected by inferCNV() with the exception is PO1, where no neoplastic cells were detected by InferCNV or CONICS. A malignancy score was calculated by dividing the log2 gained chromosome counts over the sum of those that are lost (selection was limited to three chromosomes or less). The scores were then z-neoplastic per sample. To identify putative neoplastic nuclei in this method, we next performed K means clustering of the scaled malignancy scores in R using the kmeans function and centers argument set to 2. PO1 does not show bimodal malignancy score distribution and the results of kmeans clustering were not considered. For case PA3, only a minority of nuclei had malignancy scores > 2 standard deviations above the mean. Therefore, these cells were identified using outlier detection in a normal distribution as done in the getOutliers(, method = "I", rho=c(0.1,3))$iRight) in R. getOutliers is part of extremevalues R package https://github.com/markvanderloo/extremevalues. Only the consensus nuclei that were identified as CNV positive in both approaches were considered for analysis. Less than 7.0% of the nuclei were called alternately by the two methods and were excluded from the analysis. Identification of CNVpos nuclei in recurrent glioma samples was conducted through a combination of InferCNV and identification of clusters with high expression of tumor markers SOX2 and PTPRZ1.

Geneset enrichment analysis and Gene Ontology Analysis

The average normalized counts per gene per cluster was calculated. The resultant cluster-wise count matrix was used as input to the GSVA pipeline [46]. Gene sets used for various tests are provided in the supplementary material (**Supplementary Table-1**). The options used for performing the GSVA pipeline are as follows: method= ssgsea, kcdf="Gaussian", mx.diff=TRUE. Heat maps were generated using the heatmap.2 in R function from the package gplots (R Package) and scores z-scaled were indicated. Ontology enrichment analysis in gProfiler with default settings [47]. For GSEA of the combined TCGA and CGGA dataset, the enrichment was performed using method = “ssgsea” option on normalized counts, which normalizes the enrichment scores for each gene set per sample. GSEA in **Figure S10** was conducted using pre-ranked GSEA and was performed as described in Subramanian et al 2005 with 1000 permutations [48]. Log-2-fold-change between cluster B and the remaining clusters was used to rank the genes for the analysis, and marker genes from each sub-lineage was used for the gene sets.

Survival Analysis

Survival analysis was performed using the survfit() function in the survival package in R [49, 50], using the binarized enrichment of each of the gene sets as the covariate in the formula. For cox proportional hazards, the function coxph() was used in R, and the covariates are indicated in the main figures.

Correlation Analysis

Correlation between deconvolution of values of microglial as well as astrocytic clusters, and glioma proportions was done using Pearson correlation (function cor() in R). The heatmap was generated using the corrplot package in R.

Identification of glioma state and lineage top gene markers

The lineage specific genes were determined using scater::findmarkers(…, direction =”up”) function on the top-level lineages (Neurons, astrocytes, microglia, undetermined, oligodendrocytes, OPC, and endothelial cells). The glioma-state specific genes were determined using scater::findmarkers(…, direction =”up”) function on the neoplastic glioma

states only. To select specific lineage/glioma state markers, we further filtered the top markers generated above by selecting the genes with positive log-fold change values in 90% or more of the cluster-to-cluster comparisons. The top 150 genes were selected and are provided in **Supplementary Tables 2-4** for primary glioma, recurrent glioma, and non-neoplastic lineages, respectively.

Differential gene expression analysis

EdgeR glmQLFTest was used and the top 3000 differentially expressed genes with an FDR cutoff of 25% [51] were extracted. Only datapoints with adjusted p-values less than 0.05 were used in downstream analysis.

Principal component analysis

PCA analysis was done in factorminer R package[45]. A matrix of snRNAseq sample x cell type/cluster was used as input. The composition of each snRNAseq sample astrocyte and myeloid subclusters was consolidated as follows: Astrocytes subclusters 0, 1, and 8 were combined as Ast1 (baseline/protoplasmic); subclusters 2, 3, 4, and 5 as Ast2, and subclusters 6 and 7 as Ast3. Myeloid subclusters 0 and 9 were combined as Myel1 (baseline), subclusters 1 and 3 – moTAM (monocyte derived TAMs); subclusters 2, 4, 6, and 8 as mgTAM (microglia-derived TAMs); subcluster 5 was kept as prTAM (proliferative TAMs); and subcluster 7 was kept as T-cells. The all CNV positive cells were consolidated as CNVpos, and the glioma states were included as supplementary quantitative variables. Condition (primary versus recurrent) was used as a supplementary qualitative variable.

Bulk RNAseq using Plate-Seq

RNA extraction was done using the RNeasy Mini Kit from Qiagen (Catalog No 74106). RNAseq was on spatially localized biopsies was performed using Plate-seq as described [33]. 75bp paired end sequencing was performed on Illumina NextSeq platform and read alignment was done using STAR [52] with analysis done as previously described [2]. FPKM values were used in GSEA analysis. The count matrix for the TCGA GBM dataset was downloaded using the GDCquery tool in R. The Chinese Glioma Genome Atlas (CGGA) RNAseq datasets [27, 28] was downloaded from (http://www.cgga.org.cn/download.jsp). The counts were normalized using the vst() function in deseq2 R package [53]. IDH-WT only samples were kept from both datasets (TCGA: 139 samples, CGGA: 179 samples) and used for downstream analysis.

Immunohistochemistry, histology, and in situ hybridization

Standard chromogenic Immunohistochemistry was done as described previously [22]. Paraffin-embedded formalin-fixed tissue sections or fresh frozen sections briefly fixed in 4% PFA, for 10 min (40 C) in 4% PFA in PBS. Paraffin sections after deparaffinization were treated with antigen unmasking solution according manufacture recommendations (Vector Laboratories, Burlingame, CA). The following antibodies and dilutions were used SOX2 (1:200, Mouse monoclonal, Abcam, Ab218520), KI67 (1:500, rat monoclonal polyclonal, Thermo Scientific, 14-5698-80), CD68 (1:200, mouse monoclonal, Abcam cat# ab955), NeuN (1:1000, mouse monoclonal, Millipore, MAB377). For fluorescent IHC, secondary antibody conjugated to fluorophores: anti-mouse Alexa Fluor 488 and 594 and anti-rabbit Alexa Fluor 488 and 594; goat or donkey (1:300, ThermoFisher Scientific, Eugene, OR) were applied for 1 hr at room temperature. In situ hybridization was done using RNAscope™ multiplex fluorescent v2 (ACDbio cat no 323100) per the manufacturer's protocol in 5-micron paraffin-embedded, formalin-fixed tissue sections. We used predesigned probes for PTPRZ1, CLU, TOP2A, NOVA1, MEG3, and SOX2 from ACDbio; cat# 584781, 584771, 470321, 400871, 584801, and 400871, respectively. Fluorescent images

were taken on a Zeiss 810 Axio confocal microscope at 40X. Brightfield fluorescent images were taken on an Aperio LSM™ slide scanner at 20X and 40X.

Quantification of ISH

For quantification of in situ hybridization images we used the positive cell detection function in Qupath v0.2.3 [54]. We only quantified signal contained in DAPI-positive nuclei. First, DAPI positive nuclei were detected using the cell detection tool. Next, subcellular detection function was employed to segment puncta per each of the three probe channels. A random forest classifier was used to classify nuclei to be positive or negative, with a minimum of two puncta per channel to classify a nucleus as positive for the probe. Infiltrated cortex and cellular tumor core were annotated by a neuropathologist.

Statistical testing

Statistical comparisons were done using one-way ANOVA (or Kruskal Wallis test) and Tuckey post-hoc comparison in R. Statistical testing for RNAseq application is reported in the main text or respective methods section. Differential abundance analysis was done employing a moderated regression model in ANCOMBC with default parameters, assigned Condition (primary vs post-treatment recurrence) and CNVpos proportions in the design formula, and as described by the authors ([32]). One tailed paired t-tests were done to compare the core and margin percentages of the same case (Figure 3). A one sample t-test was conducted to determine if the percentage of TOP2A+ that were CLU+ was less than 50%.

Data availability

Data for spatial transcriptomic can be queried using an interactive web app: https://vmenon.shinyapps.io/gbm_expression/. All other data will be deposited as in GEO repositories prior to publication.

Acknowledgment

The results published here are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. This research was funded by NIH/NINDS R01NS103473 (PC, PAS, JNB), The William Rhodes and Louise Tilzer-Rhodes Center for Glioblastoma at New York-Presbyterian Hospital (OA, VM, PC), Herbert Irving Comprehensive Cancer Center pilot research grant (OA), and the NIH/NCI Cancer Center Support Grant P30CA013696 (OA, ABL, JNB, PAS, PC).

This research was funded by NIH/NINDS R01NS103473 (PC, PAS, JNB), William Rhodes and Louise Tilzer-Rhodes Center for Glioblastoma (RCG) Collaborative Research Initiative grant (OA, VM, PC), HICCC pilot research grant (OA), and in part through the NIH/NCI Cancer Center Support Grant P30CA013696. This research was supported by the Genomics and High Throughput Screening Shared Resource and the Digital Computational Pathology Laboratory in the Department of Pathology and Cell Biology at Columbia University Irving Medical Center. We thank Dr. Claudia Deoge for help with RNAscope. We thank the immunohistochemistry core in the Department of Pathology and Cell Biology at Columbia University Irving Medical Center for help with CD14 and GFAP-SOX2 immunostains. We thank Al-Fardthakh© for support with aspects of bulk RNAseq data analysis.

Author contribution

OS, JNB, PAS, VM, and PC designed the study; OA, DB, AS, AM, FK, JFT, ARG, AT, MAB, JF, TS, EB, JNB and PC conducted the experiments, including tissue procurement, tissue

processing, and tissue analysis. OA, MA, PAS, VM, and PC analyzed the data. OA, and PC wrote the manuscript and all authors edited, read, and approved the final manuscript.

Supplementary Results

Comparison of CNV profiles between primary and recurrent glioma pairs

Not surprisingly, the recurrent tumors did not show identical chromosomal CNVs with their primary counterparts. While TB5014 retained the CNV of the TB4916 (gain of 7, loss of 10 and 14) and acquired additional alterations including gains in chromosomes 19 and 20 (**Figures S1F and S2A)**, TB5053 showed a complex gains and losses across multiple chromosomes (**Figures S1G and S2B)**.

Survival analysis of glioblastoma state signatures

To determine if enrichment of our glioma states has prognostic relevance we performed GSEA of the top gene markers for each glioma state, both primary and recurrent gene sets (**Supplementary Table 2 and 3**), using the TCGA-GBM [26] and the Chinese Glioma Genome Atlas (CGGA) databases [27, 28]. 318 IDH-WT samples (TCGA - 139, CGGA - 179) were used for survival analysis. Kaplan-Meyer survival curves are shown on **Figure-S7B-C** for primary and recurrent glioma gene sets, respectively. The astrocyte-like/mesenchymal signatures Mes1 in primary and Mes2 in recurrent glioma significantly predicted poor survival. These two gene sets share 46 genes including CD44, and CHI3L1, which may explain the effect of primary gl_Mes1 and recurrent gl_Mes2 with survival.

Analysis of low-grade glioma and epilepsy cases

To sample states of myeloid cells and astrocytes across different disease states, we chose to analyze the microenvironment of low-grade glioma (LGG) and temporal lobe epilepsy. We conducted snRNAseq on 6 cases: two IDH-mutant oligodendroglioma (TB3652 & TB3926), one IDH-mutant astrocytoma (TB4100), and three temporal lobe epilepsies (TB4189, TB4437, & TB4957). We identified 970, 1154, 1036 nuclei for LGG cases TB3652, TB3926, and TB4100, respectively. We identified CNVpos nuclei using a combination of chromosomal CNV, clustering, and tumor marker expression as shown in **Figure S3**. Cases TB3652 and TB3926 had typical chromosome 1p and 19q codeletions (**Figure S3A, D**), and harbored 817 and 942 CNVpos nuclei, respectively (**Figure S3B, E**). The tumor nuclei expressed tumor markers SOX2, EGFR, and PTPRZ1, and/or OPC markers DSCAM and TNR; myeloid cells expressed a CD74, C3, ITGAX (CD11c), ITM2B, and/or HLA-B; while oligodendrocytes expressed MBP and MOG (**Figure S3C, F**). 382 CNVpos nuclei were found in case TB4100, which did not harbor CNVs across most cells, and CNVpos nuclei were identified by clustering and marker expression as noted above. Of the epilepsy cases, we identified 2558, 179, and 138 nuclei in cases TB4189, TB4437, and TB4957, respectively. **Figure S4A-C** show marker expression in cases TB4437, TB4189, and TB4957, where markers of astrocytes (GFAP, AQP4, SLC1A2, SLC1A3), neurons (RBFOX3, MEG3, GAD1, and SLC17A6), myeloid cells (CD74, ITGAX, C3, ITM2B), oligodendrocytes (MBP, MOG, OPALIN, and CNP), and OPCs (DSCAM, TNR, SOX2, and PDGFRA). The CNVneg nuclei from all LGG and epilepsy cases were combined with those from primary and recurrent IDH-WT GBM and were analyzed as presented in the section below (myeloid cells) and main text (astrocytes).

Astrocyte signatures and classifications

Based on unbiased clustering of astrocytes, we identified multiple clusters that share in gene expression. For example, clusters 1 and 8 show expression protoplasmic astrocytes gene

including SLC1A2, NTM, GLUL, FGFR3, and GPC5. Moreover, clusters 6 and 7 showed expression multiple reactive genes including LGALS3, LGALS1, CLU, and CD44. Overall, all clusters showed expression of canonical astrocyte gene GFAP, as well as variable levels of expression of other astrocyte genes including AQP4, S100B, GPC5, SLC1A2, and ATP1B2 (**Figure-S5C-D**). Furthermore, expression of non-astrocyte genes including PLP1, MBP, CCK, and MEG3 was seen in clusters 2 and 4, and ribosomal genes in cluster 3 (**Figure-3D**). Based on the overlap of gene expression between astrocyte clusters, and the resemblance to known astrocyte phenotypes we derived three gene sets (**Supplementary Table-5**), which represent three major astrocyte states (protoplasmic, reactive-1, and reactive-2), and then re-clustered astrocyte nuclei using Ward D2 hierarchical clustering on the Manhattan distance of the enrichment scores (**Figure-S5E**), into a protoplasmic cluster (Ast1), and two reactive clusters (Ast2 and Ast3). These astrocyte states correspond closely to those used in the compositional PCA (**figure-5B**). A select number of gene markers for each cluster is shown in **Figure-S5F**. The scores of each gene set are projected on astrocyte nuclei in 3D tSNE (**Figure-S5G-I**).

Analysis of myeloid cell states

Myeloid cells have been implicated in modulating glioma migration, infiltration, and progression [37]. We identified 5925 nuclei we classified as myeloid cells. Unbiased clustering revealed 8 subclusters which we then used to assign the specific myeloid lineages. We merged clusters with similar enrichment scores of gene sets representing microglia-derived tumor-associated macrophages (mgTAM), monocyte-derived TAMs (moTAM), proliferative TAMs (prTAMs), and T-cells as described in [15] - see methods (section on PCA analysis). The enrichment of these gene sets in the final myeloid states is provided in

Figure S6E. A subset of myeloid cells showed mixed enrichment scores across mgTAM, moTAM, and dendritic cells, and were considered baseline (referred to as Myel1). Overall, we classified 2678, 1346, 1364, 360, and 177 nuclei as Myel1, moTAM, mgTAM, prTAM, and T-cells, respectively, and these are shown in 3D tSNE space in **Figure S6A**. Myel1 state showed higher expression of SAT1, CEBPD, and GLUL (**Figure S6C**-top row). moTAM showed highest expression of CD163, MS4A4E, NHSL1, FMN1, and MSR1 (**Figure S6C**-2nd row). mgTAM showed highest expression of SORL1, RIN3, ITGAX, HS3ST4, and FRMD4A (**Figure S6C**-3rd row). prTAMs showed highest expression of CST3, MEF2A, DBI, PLXDC2, and DOCK4 (**Figure S6C**-4th row). Finally, T-cells showed highest expression of CD2, CD247, CD96, FYN, and SKAP1 (**Figure S6C**-5th row). Different myeloid states were accounted for different conditions (**Figure S6B).** While Myel1 was present in Epilepsy, primary and recurrent GBM, mgTAM was the main state found in LGG, but was also in primary and recurrent GBM. moTAM, T-cells, and prTAM were found in primarily in recurrent GBM (**Figure S6D)**. The gene-wise DGE between myeloid states and the myeloid state markers are provided in **Supplementary Table-6**.

References

1. Quail, D.F. and J.A. Joyce, *The Microenvironmental Landscape of Brain Tumors.* Cancer Cell, 2017. **31**(3): p. 326-341.
2. Gill, B.J., et al., *MRI-localized biopsies reveal subtype-specific differences in molecular and cellular composition at the margins of glioblastoma.* Proc Natl Acad Sci U S A, 2014. **111**(34): p. 12550-5.
3. Puchalski, R.B., et al., *An anatomic transcriptional atlas of human glioblastoma.* Science, 2018. **360**(6389): p. 660-663.
4. Jin, X., et al., *Targeting glioma stem cells through combined BMI1 and EZH2 inhibition.* Nat Med, 2017. **23**(11): p. 1352-1361.
5. Wang, Q., et al., *Tumor Evolution of Glioma-Intrinsic Gene Expression Subtypes Associates with Immunological Changes in the Microenvironment.* Cancer Cell, 2017. **32**(1): p. 42-56 e6.
6. Yuan, J., et al., *Single-cell transcriptome analysis of lineage diversity in high-grade glioma.* Genome Med, 2018. **10**(1): p. 57.
7. Wang, L., et al., *The Phenotypes of Proliferating Glioblastoma Cells Reside on a Single Axis of Variation.* Cancer Discov, 2019. **9**(12): p. 1708-1719.
8. Patel, A.P., et al., *Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma.* Science, 2014. **344**(6190): p. 1396-401.
9. Neftel, C., et al., *An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma.* Cell, 2019. **178**(4): p. 835-849 e21.
10. Levitin, H.M., et al., *De novo gene signature identification from single-cell RNA-seq with hierarchical Poisson factorization.* Mol Syst Biol, 2019. **15**(2): p. e8557.
11. Darmanis, S., et al., *Single-Cell RNA-Seq Analysis of Infiltrating Neoplastic Cells at the Migrating Front of Human Glioblastoma.* Cell Rep, 2017. **21**(5): p. 1399-1410.
12. Bhaduri, A., et al., *Outer Radial Glia-like Cancer Stem Cells Contribute to Heterogeneity of Glioblastoma.* Cell Stem Cell, 2020. **26**(1): p. 48-63.e6.
13. Lee, J.K., et al., *Spatiotemporal genomic architecture informs precision oncology in glioblastoma.* Nat Genet, 2017. **49**(4): p. 594-599.
14. Barthel, F.P., et al., *Longitudinal molecular trajectories of diffuse glioma in adults.* Nature, 2019. **576**(7785): p. 112-120.
15. Pombo Antunes, A.R., et al., *Single-cell profiling of myeloid cells in glioblastoma across species and disease stage reveals macrophage competition and specialization.* Nat Neurosci, 2021. **24**(4): p. 595-610.
16. Tickle, T., et al., *inferCNV of the Trinity CTAT Project.* 2019.
17. Butler, A., et al., *Integrating single-cell transcriptomic data across different conditions, technologies, and species.* Nat Biotechnol, 2018. **36**(5): p. 411-420.
18. Verhaak, R.G., et al., *Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1.* Cancer Cell, 2010. **17**(1): p. 98-110.
19. Gobin, M., et al., *A DNA Repair and Cell-Cycle Gene Expression Signature in Primary and Recurrent Glioblastoma: Prognostic Value and Clinical Implications.* Cancer Res, 2019. **79**(6): p. 1226-1238.
20. Weng, Q., et al., *Single-Cell Transcriptomics Uncovers Glial Progenitor Diversity and Cell Fate Determinants during Development and Gliomagenesis.* Cell Stem Cell, 2019. **24**(5): p. 707-723 e8.

21. Henrik Heiland, D., et al., *Tumor-associated reactive astrocytes aid the evolution of immunosuppressive environment in glioblastoma.* Nat Commun, 2019. **10**(1): p. 2541.
22. Al-Dalahmah, O., et al., *Single-nucleus RNA-seq identifies Huntington disease astrocyte states.* Acta Neuropathol Commun, 2020. **8**(1): p. 19.
23. Pelloski, C.E., et al., *YKL-40 expression is associated with poorer response to radiation and shorter overall survival in glioblastoma.* Clin Cancer Res, 2005. **11**(9): p. 3326-34.
24. Wu, G., et al., *Expression of CD44 and the survival in glioma: a meta-analysis.* Biosci Rep, 2020. **40**(4).
25. He, X., et al., *Increased LGALS3 expression independently predicts shorter overall survival in patients with the proneural subtype of glioblastoma.* Cancer Med, 2019. **8**(5): p. 2031-2040.
26. Brennan, C.W., et al., *The somatic genomic landscape of glioblastoma.* Cell, 2013. **155**(2): p. 462-77.
27. Zhao, Z., et al., *Comprehensive RNA-seq transcriptomic profiling in the malignant progression of gliomas.* Sci Data, 2017. **4**: p. 170024.
28. Liu, X., et al., *A radiomic signature as a non-invasive predictor of progression-free survival in patients with lower-grade gliomas.* Neuroimage Clin, 2018. **20**: p. 1070-1077.
29. Liu, L., et al., *CHI3L2 Is a Novel Prognostic Biomarker and Correlated With Immune Infiltrates in Gliomas.* Front Oncol, 2021. **11**: p. 611038.
30. Fazi, B., et al., *The Expression of the Chemokine CXCL14 Correlates with Several Aggressive Aspects of Glioblastoma and Promotes Key Properties of Glioblastoma Cells.* Int J Mol Sci, 2019. **20**(10).
31. Cenini, G., A. Lloret, and R. Cascella, *Oxidative Stress in Neurodegenerative Diseases: From a Mitochondrial Point of View.* Oxid Med Cell Longev, 2019. **2019**: p. 2105607.
32. Lin, H. and S.D. Peddada, *Analysis of compositions of microbiomes with bias correction.* Nat Commun, 2020. **11**(1): p. 3514.
33. Bush, E.C., et al., *PLATE-Seq for genome-wide regulatory network analysis of high-throughput screens.* Nat Commun, 2017. **8**(1): p. 105.
34. Lopez, K.A., et al., *Convection-enhanced delivery of topotecan into a PDGF-driven model of glioblastoma prolongs survival and ablates both tumor-initiating cells and recruited glial progenitors.* Cancer Res, 2011. **71**(11): p. 3963-71.
35. Mathys, H., et al., *Single-cell transcriptomic analysis of Alzheimer's disease.* Nature, 2019. **570**(7761): p. 332-337.
36. Stockwell, B.R., et al., *Ferroptosis: A Regulated Cell Death Nexus Linking Metabolism, Redox Biology, and Disease.* Cell, 2017. **171**(2): p. 273-285.
37. Markovic, D.S., et al., *Microglia stimulate the invasiveness of glioma cells by increasing the activity of metalloprotease-2.* J Neuropathol Exp Neurol, 2005. **64**(9): p. 754-62.
38. Krishnaswami, S.R., et al., *Using single nuclei for RNA-seq to capture the transcriptome of postmortem neurons.* Nat Protoc, 2016. **11**(3): p. 499-524.
39. Petukhov, V., et al., *dropEst: pipeline for accurate estimation of molecular counts in droplet-based single-cell RNA-seq experiments.* Genome Biol, 2018. **19**(1): p. 78.
40. McCarthy, D.J., et al., *Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R.* Bioinformatics, 2017. **33**(8): p. 1179-1186.
41. Lun, A.T., D.J. McCarthy, and J.C. Marioni, *A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor.* F1000Res, 2016. **5**: p. 2122.
42. Stuart, T., et al., *Comprehensive Integration of Single-Cell Data.* Cell, 2019. **177**(7): p. 1888-1902 e21.
43. Hafemeister, C. and R. Satija, *Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression.* Genome Biol, 2019. **20**(1): p. 296.

44. Levitin, H.M., J. Yuan, and P.A. Sims, *Single-Cell Transcriptomic Analysis of Tumor Heterogeneity.* Trends Cancer, 2018. **4**(4): p. 264-268.
45. Lê, S., J. Josse, and F. Husson, *FactoMineR: A Package for Multivariate Analysis.* Journal of Statistical Software, 2008. **25**(1): p. 1-18.
46. Hanzelmann, S., R. Castelo, and J. Guinney, *GSVA: gene set variation analysis for microarray and RNA-seq data.* BMC Bioinformatics, 2013. **14**: p. 7.
47. Reimand, J., et al., *g:Profiler-a web server for functional interpretation of gene lists (2016 update).* Nucleic Acids Res, 2016. **44**(W1): p. W83-9.
48. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15545-50.
49. Therneau, T.M. and P.M. Grambsch, *Modeling Survival Data: Extending the Cox Model*. 2000, New York: Springer.
50. Therneau, T.M., *A Package for Survival Analysis in R*. 2020.
51. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-40.
52. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.
53. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biol, 2014. **15**(12): p. 550.
54. Bankhead, P., et al., *QuPath: Open source software for digital pathology image analysis.* Sci Rep, 2017. **7**(1): p. 16878.

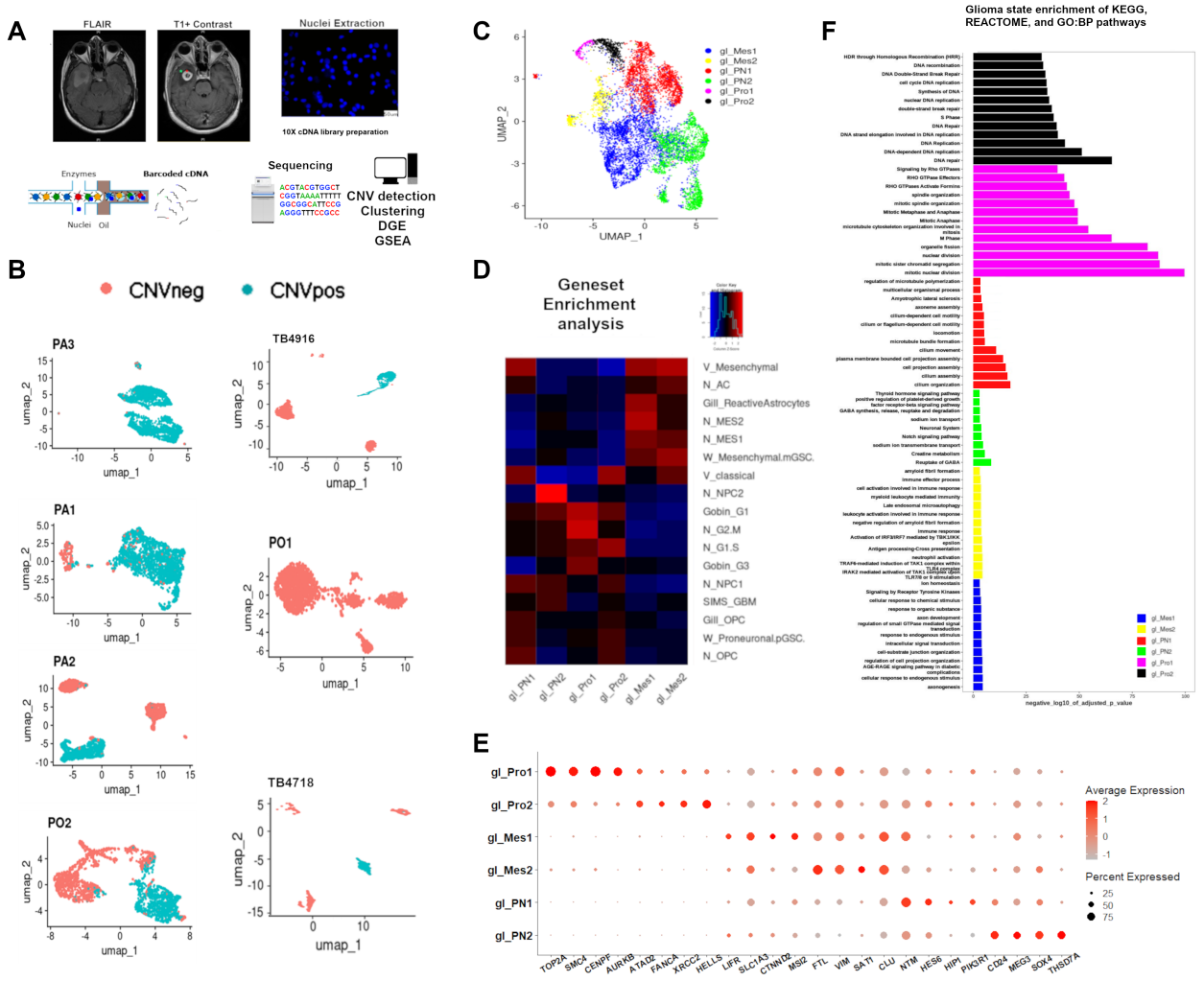

Figure 1: snRNAseq-derived transcriptional states of putative neoplastic nuclei from primary IDH-wildtype GBM samples

A) Outline of Analytic Design: T 2/FLAIR and post-contrast T1 MRI sequences of a glioblastoma showing the classic radiological appearance of a glioblastoma (Case PO2); with a ring enhancing mass (red star) with surrounding increased FLAIR signal (green star). The tumor was resected and banked (frozen). Nuclei are extracted from frozen tissue and are subjected to droplet based single nuclei RNA sequencing using the 10X chromium platform. The resultant barcoded cDNA is then sequenced and analyzed. Analyses performed include identification of putative neoplastic cells by identifying cells with inferred copy number variations (CNV), clustering, differential gene expression (DGE), and gene set enrichment analysis (GSEA). Scale bars = 50 um. **B)** Uniform-manifold approximation and projection (UMAP) graphs showing putative neoplastic (CNVpos) and non-neoplastic (CNVneg) nuclei from the seven primary IDH-wildtype glioma cases selected for analysis indicated by subpanels b1-b7. **C**) UMAP plot showing all putative CNVpos (**C**) nuclei from the seven primary glioma cases aligned and projected in shared UMAP spaces. The nuclei are color-coded by glioma state: Oligodendrocyte-progenitor-like (proneural - gl_PN1), Neural-progenitor-like (proneural - gl_PN2), Mesenchymal/astrocyte like (gl_Mes1 and gl_Mes2), and proliferative (gl_Pro1 & gl_Pro2). **D**) Geneset enrichment analysis (GSEA) of selected gene sets from Verhaak et al. 2009 (v), Gobin M et al 2019, Gill et al 2014, Wang et al. 2019 (W), and Neftel et al. 2019 (N) showing enrichment of genes specific for states described in the literature in our described glioma states. **E**) Gene ontology (GO) term enrichment analysis (KEGG and REACTOME pathways and biological process GO) of the major terms enriched in glioma state top gene markers. The bars represent the negative log10 of the false discovery rate adjusted p.value, and are color-coded as in **C**.

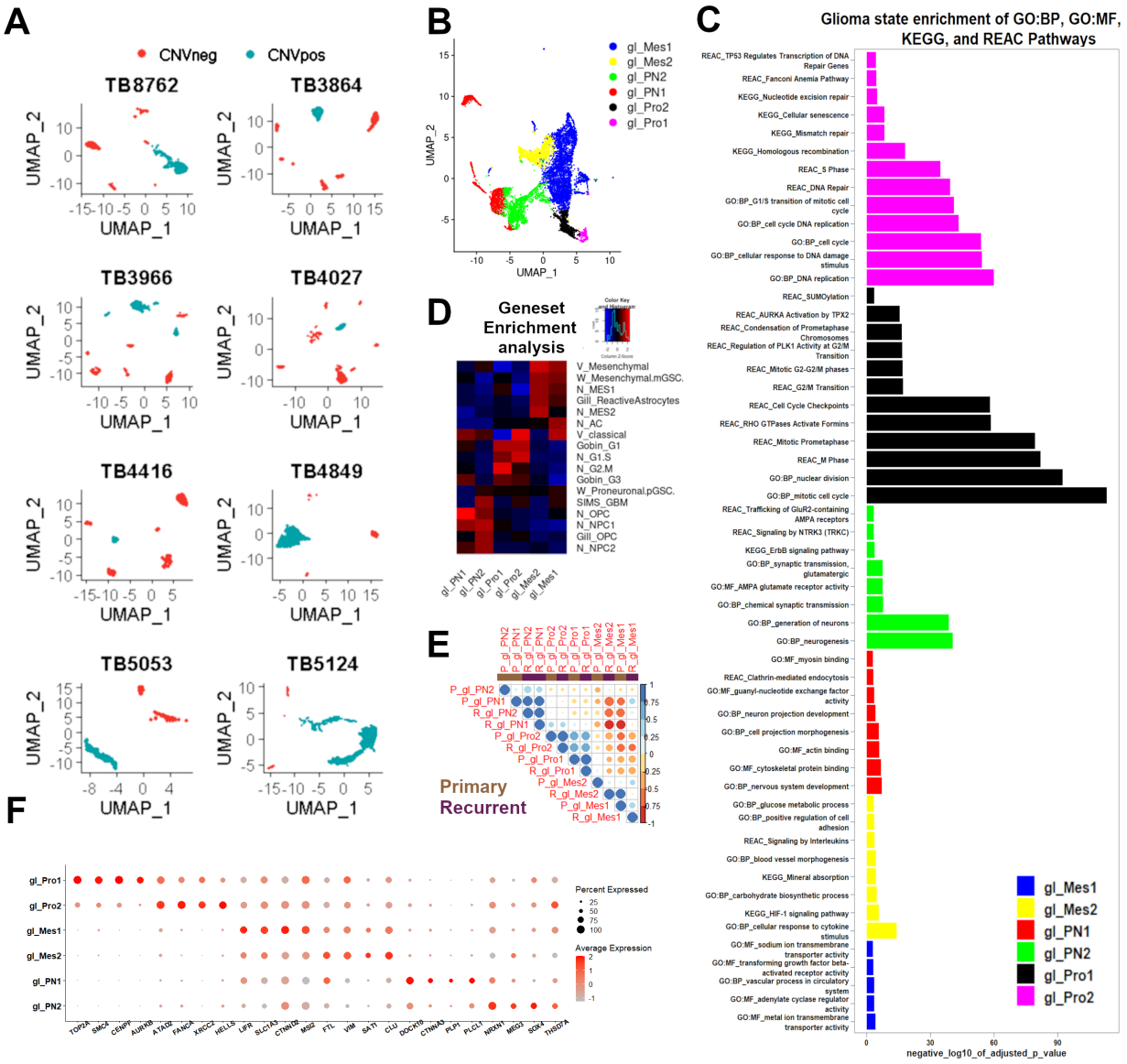

Figure 2 snRNAseq-derived transcriptional states of putative neoplastic nuclei from post-treatment recurrent IDH-wildtype GBM samples

A) Uniform-manifold approximation and projection (UMAP) graphs showing putative neoplastic (CNVpos) and non-neoplastic (CNVneg) nuclei from the eight post-treatment recurrent IDH-wildtype glioblastoma cases. **B**) UMAP plot showing all putative CNVpos nuclei from the eight recurrent glioma cases aligned and projected in shared UMAP spaces. The nuclei are color-coded by glioma state: Oligodendrocyte-progenitor-like (proneural - gl_PN1), Neural-progenitor-like (proneural - gl_PN2), Mesenchymal/astrocyte like (gl_Mes1 and gl_Mes2), and proliferative (gl_Pro1 & gl_Pro2). **C**) Gene ontology (GO) term enrichment analysis (KEGG and REACTOME pathways and biological process GO) of the major terms enriched in glioma state top gene markers. The bars represent the negative log10 of the false discovery rate adjusted p.value and are color-coded as in B. **D**) Geneset enrichment analysis (GSEA) of selected gene sets from Verhaak et al. 2009, Gobin M et al 2019, Gill et al 2014, and Neftel et al. 2019 showing enrichment of genes specific for states described in the literature in our described glioma states. **E**) Correlation heatmap between glioma states in primary and post-treatment recurrent GBM based on expression on glioma state marker genes. The size and color of the circles denote the strength of correlation. **F**) Gene expression dot plots showing select gene marker expression in glioma states.

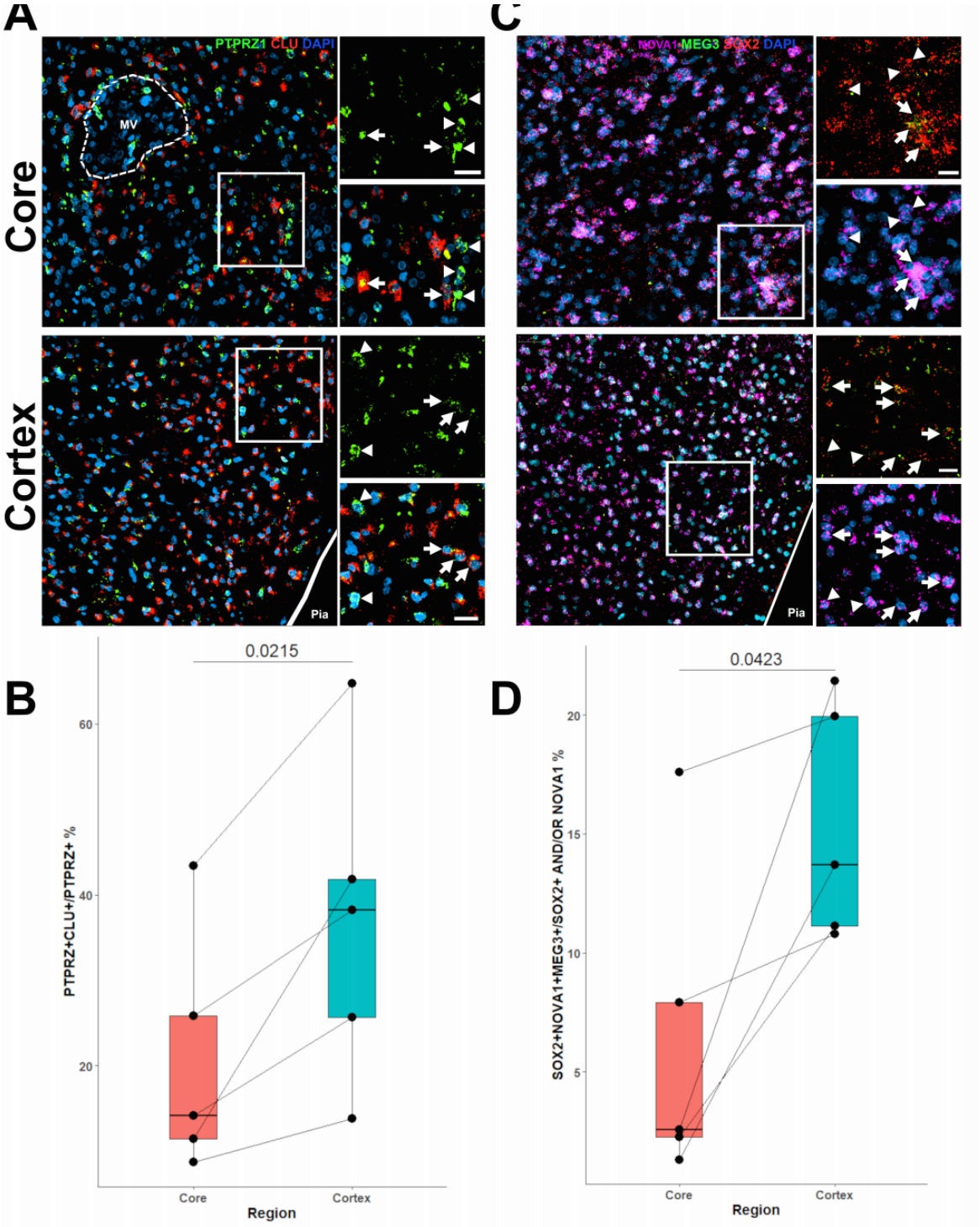

Figure 3: The spatial landscape of glioma states across the cellular tumor core and infiltrated cortical margin

A) Confocal images showing optical sections of in situ hybridization for PTPRZ1 and CLU in the core (upper row) and cortex (lower row). The pial surface is outlined (lower row). High-power images of the insets show that PTPRZ1+ CLU+ cells (arrows) are more abundant in the cortex, while PTPRZ1+CLU- (arrowheads) are more numerous in the core. scale bars = 20 µm. M.V: Microvascular proliferation **B)** Quantification of PTPRZ1 and CLU expression across the core (orange boxplot) and cortex (green boxplot). The data is shown as boxplots, with the bar indicating the median. Paired t-test, N=5. The p value is indicated. **C**) Confocal images showing optical sections of in situ hybridization for NOVA, SOX2, and MEG3 in the core (upper row) and cortex (lower row). The pial surface is outlined (lower row). High-power images of the insets show that NOVA1+SOX2+MEG3+ cells (arrows) are more abundant in the cortex, while MEG3- cells (arrowheads) are more numerous in the core. scale bars = 20 µm. **C**) Quantification of MEG3+NOVA1+SOX2+ cells as a proportion of all tumor cells (SOX2+ and/or NOVA1+) across the core (orange boxplot) and cortex (green boxplot). The data is shown as boxplots, with the bar indicating the median. Paired t-test, N=5. The p value is indicated.

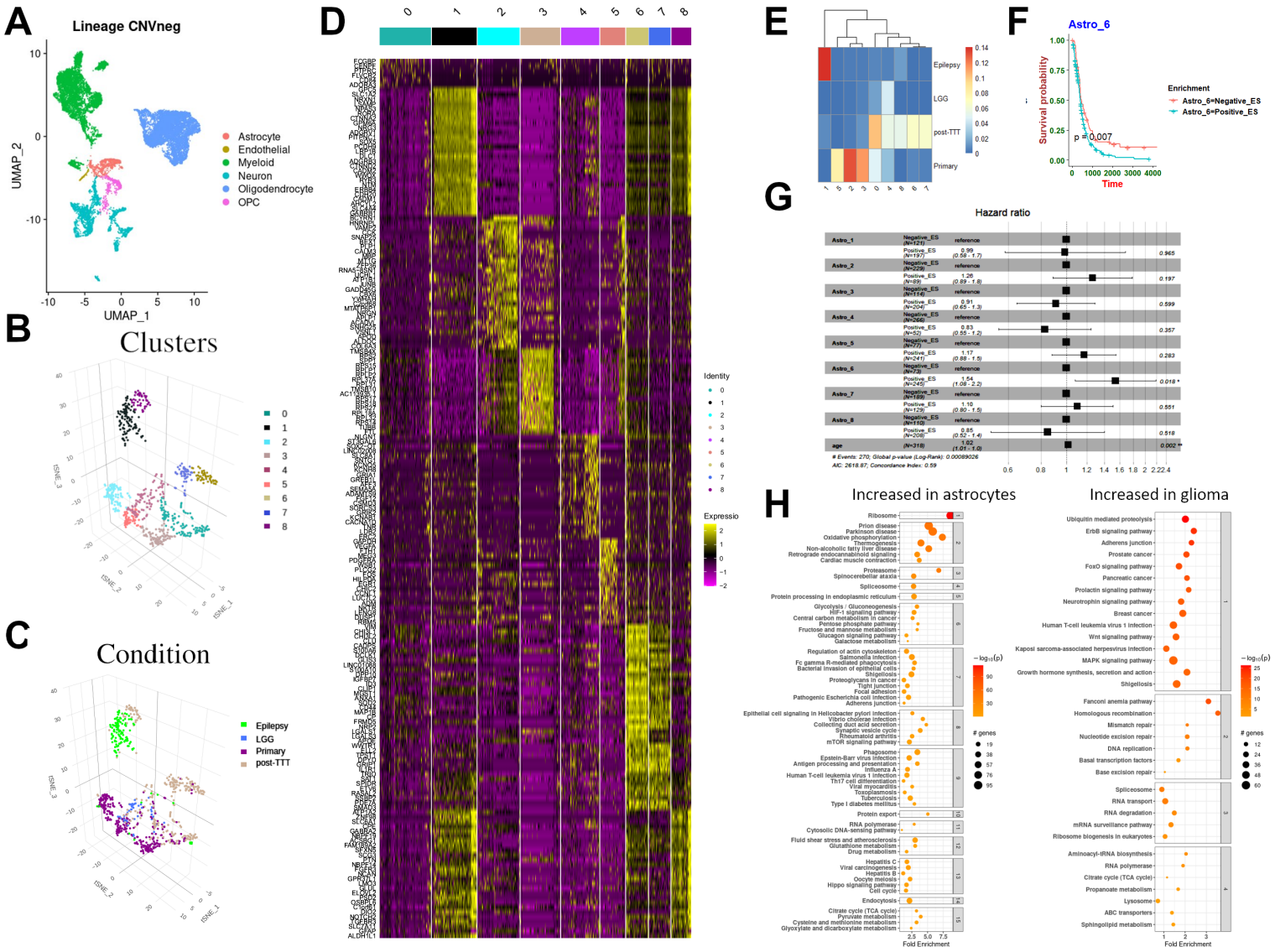

Figure 4: Astrocytes in the glioma microenvironment are heterogeneous and prognostically relevant

A) Uniform-manifold approximation and projection (UMAP) graphs showing putative non-neoplastic (CNVneg) from primary glioma, recurrent glioma, low grade glioma (LGG) - and epilepsy (see supplementary data for the analysis of LGG and epilepsy cases). The nuclei are color-coded by lineage (Oligodendrocytes, oligodendrocyte-precursor cells (OPC), neurons, astrocytes, myeloid cells, and endothelial cells). Three-dimensional UMAP plots showing all astrocyte nuclei from color-coded by unbiased cluster (**B**) and condition (**C**). Scaled gene expression heatmap showing select gene marker expression for the unbiased astrocyte clusters. Clusters 1 and 8 show high expression of protoplasmic astrocyte genes, while clusters 6 and 7 show high expression of reactive genes. Cluster 0 has very few genes differentially expressed and is considered baseline. **E**) Heatmap showing the proportion of nuclei in each cluster (columns) contributed by condition (rows). **F**) Kaplan-Meier survival curves for the combined TCGA and CGGA RNAseq datasets. The samples were classified based on enrichment of Astrocyte cluster 6 signature genes into positive or negative enrichment. Statistical significance was computed using the log rank test. **G**) Cox proportional hazard ratio of survival in the combined TCGA and CGGA dataset given enrichment of each of the astrocyte cluster specific signature. Age is also included in the model. The p values are shown on the left, bars indicate confidence intervals (also noted on the right). Enrichment of each gene set was categorized as negative or positive. **H**) Active subnetwork enrichment analysis of KEGG pathways in genes differentially expressed in glioma-associated astrocytes compared to primary and recurrent IDH-WT glioma. Fold enrichment is represented on the x-axis and the pathways in the y-axis. The pathways are clustered to denote shared genes driving enrichment. The size of the circle per pathway denotes the number of enriched genes, and the negative log10 of the adjusted p value is represented by color. Pathways enriched in genes significantly increased in astrocytes are shown on the right and include neurodegenerative diseases and oxidative phosphorylation. Pathways enriched in genes significantly decreased in astrocytes are shown on the left and include DNA replication and Wnt signaling.

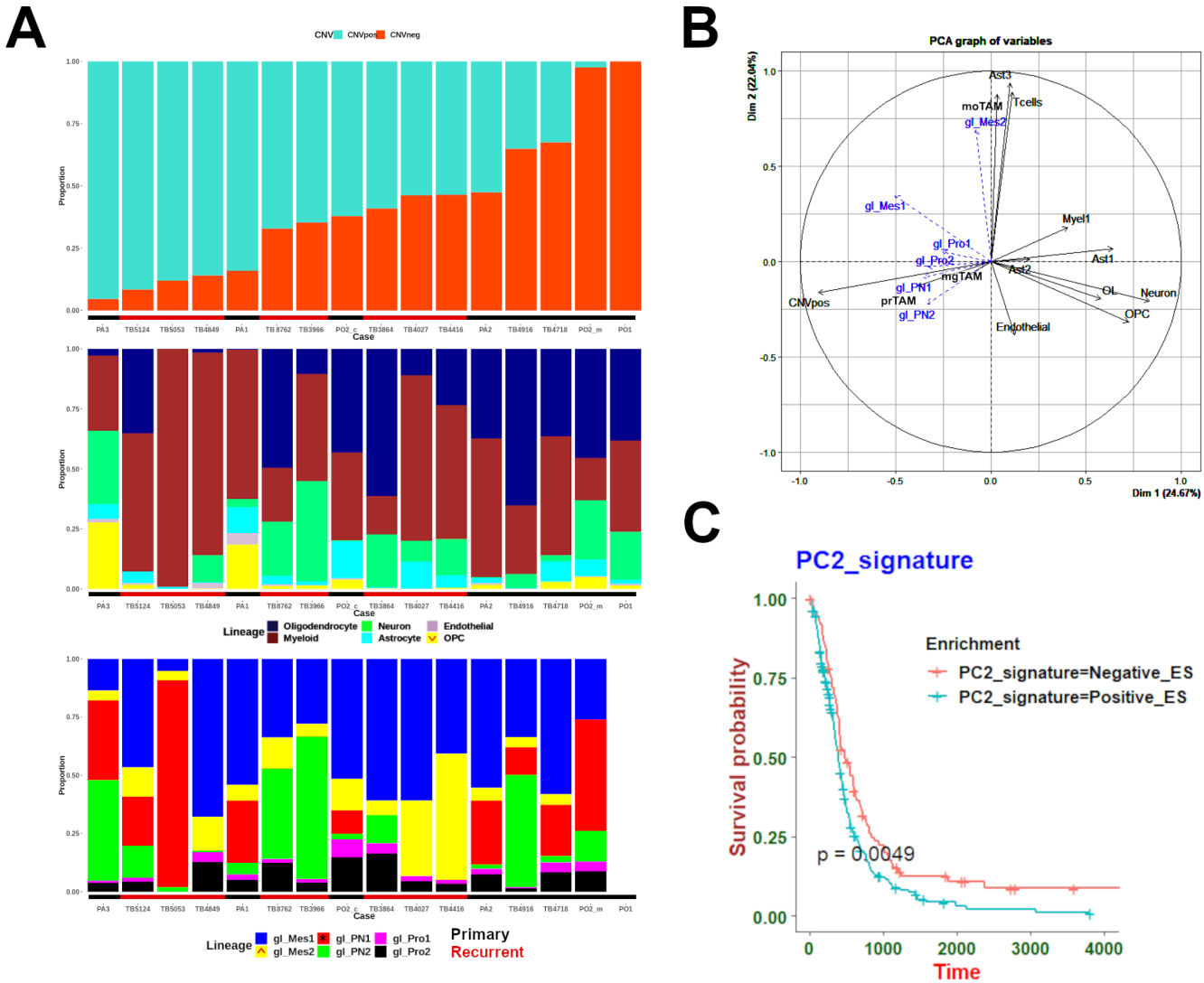

Figure 5: Correlated glioma and microenvironmental states better predict prognosis

A) Bar plots demonstrating the fractional composition of each one of 16 samples analyzed by snRNAseq (8 primary IDH-WT glioma from 7 patients, one case was divided to core and overlying cortex, and 8 recurrent IDH-WT glioblastoma) The first row of bar plots represent the fraction of neoplastic (CNVpos) and non-neoplastic (CNVneg) nuclei. The middle row represents the fraction of the non-neoplastic nuclei contributed by neurons, oligodendrocytes, OPCs, astrocytes, myeloid cells, endothelial cells, and astrocytes. The red arrowhead in the OPC legend box indicate that OPCs were significantly reduced in recurrent GBM as determined by differential abundance analysis. The bottom row represents the fraction of the neoplastic nuclei contributed by gl_PN1, gl_PN2, gl_Mes1, gl_Mes2, gl_Pro1, and gl_Pro2. The red arrowhead in the gl_Mes2 legend box and the black star in gl_PN1 indicate that gl_Mes2 and gl_PN1 were significantly increased and reduced in recurrent GBM, respectively, as determined by differential abundance analysis. **B**) Principal component analysis of the fractional composition matrix of 19 samples encompassing eight primary and eight recurrent gliomas plus three epilepsy samples. The tissue composition matrix consists of the percentage of nuclei per each tissue state. Immune cell states are: mgTAMs (microglia-derived Tumor-associated macrophages), moTAM (monocyte-derived TAMs), prTAM (proliferative TAM), Myel1 (baseline myeloid cells), and T cells. Astrocyte states include baseline (protoplasmic) astrocytes (Ast1), reactive CD44+ astrocytes (Ast3), and astrocytes with mixed identity (Ast2) – see text and supplementary results for additional description of these cell states. CNVpos represents the total percentage of all tumor states per sample. Individual tumor states were not used in PCA calculation, rather their coordinates were predicted including (Supplementary quantitative variables). **C**) Kaplan-Meier survival plot graphing survival in the combined TCGA and CGGA RNAseq datasets. The samples were classified based on enrichment of gene signatures of microenvironment states correlated with PC2 into positive or negative enrichment. Statistical significance was computed using the log rank test. Note the p value is lower here compared with Figure 3E.

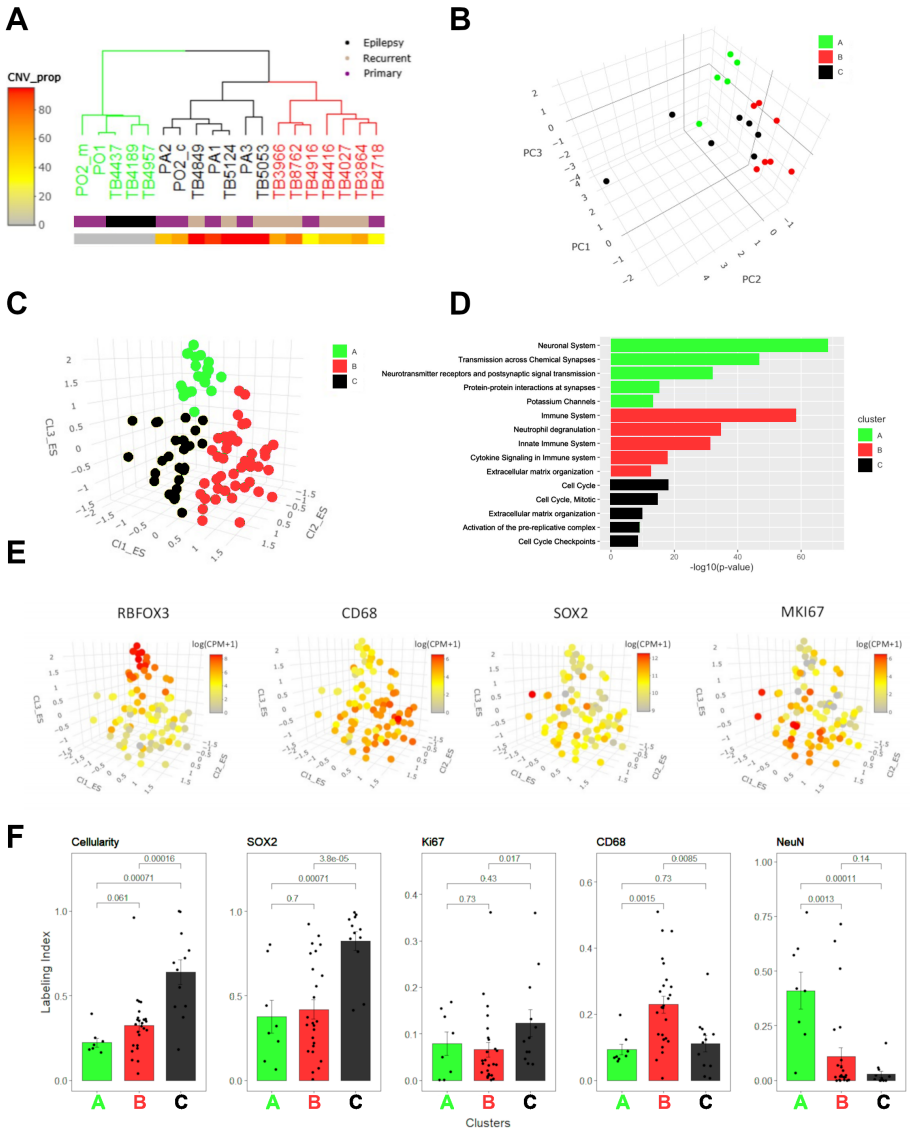

Figure 6: Tissue composition analysis defines "tissue states" that are recapitulated in bulk RNAseq data and validated by quantitative immunohistochemical analysis

A) Dendrogram of hierarchically clustered glioma and epilepsy samples based on Minkowski sample distance analysis drawn from the fractional composition matrix (see Figure 4A). Three clusters were identified and are color-coded on the dendrogram in black (Tissue-state C), red (Tissue-state B), and green (Tissue-state A). The condition (primary, recurrent and epilepsy) is indicated in the top bar below. The proportion of neoplastic nuclei (CNVprop) is indicated in the bottom bar. **B**) Three-dimensional scatter plot showing the samples in **A** projected in the first three principal component loadings – see figure 4B for PCA analysis. The samples are color-coded by cluster designation as in A. **C**) Bulk RNAseq samples from 91 primary and recurrent IDH-WT glioblastoma samples projected in the space represented by the enrichment of the three gene signatures characteristic of the tissue states in B. The samples were clustered on the Minkowski distance of the enrichment scores into three clusters A-C and are color-coded as such. **D**) Gene ontology term analysis of the genes uniquely and differentially expressed in each of the clusters in **C**. The bar plots are color coded as per the clusters in C. KEGG, REACTOME, or Biological Process GO pathways are shown in the y-axis. Negative log10 of the adjusted p-value is shown on the x-axis. **E**) Normalized expression of select genes characteristic of each of the clusters projected onto the compositional-signature enrichment score space shown in C. Red denotes high expression, and grey denotes low expression. NeuN (RBFOX3) is highest in the samples of Cluster A. CD68 is highest in the samples of cluster B. SOX2 is highest in the samples of clusters B and C. MKI67 is highest in the samples of cluster C. **F**) Quantification of histological cellularity analysis (far left) as well as immunohistochemistry labeling indices of (from left to right) SOX2, KI67, CD68, and NeuN. The labeling index is shown on the y-axis. Note that the y-axis for the cellularity graph is total cellularity normalized to the most cellular sample. The sample clusters are labeled (A-C) as in **C**. p values were calculated using Kruskal-Wallis test and are indicated on the graphs. N=8 for cluster A, 25 for cluster B, and 12 for cluster C.

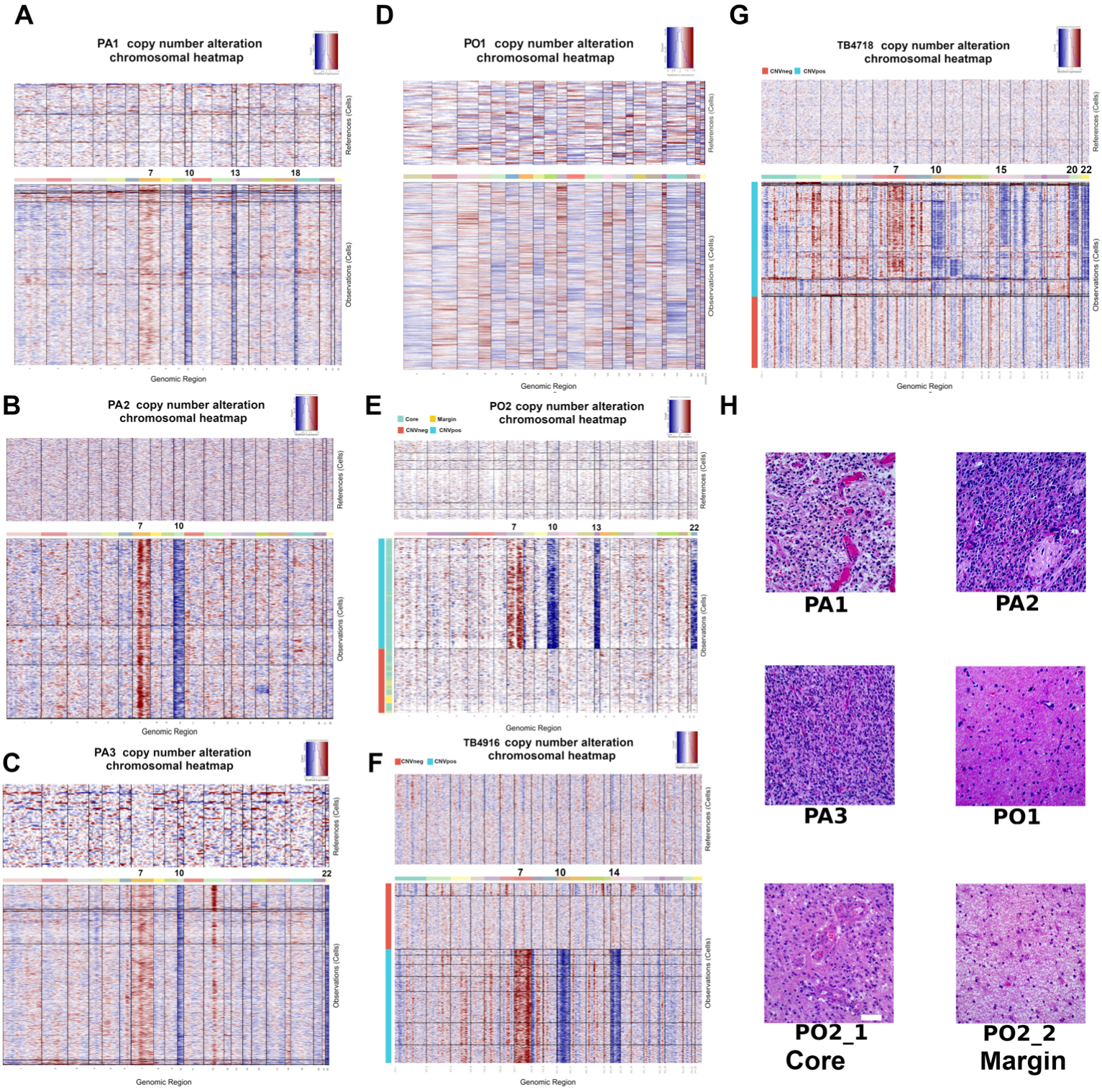

Figure S1: Identifying neoplastic nuclei in primary GBM by copy number alterations inferred from snRNAseq, and histopathology of corresponding tissue samples

Large-scale chromosomal copy number alterations were inferred from RNA expression using InferCNV R package (see methods for details). The heat maps show gains (red) and losses (blue) in case PA1 (**A**), PA2 (**B**), PA3 (**C**), PO1 (**D**), PO2 (two samples – core and margin) (**E**), TB4916 (**F**), and TB4718 (**G**). Representative Hematoxylin and Eosin-stained section of the brain tissue used for single nuclei RNAseq of the first five cases (**H**). Some cases showed clear infiltration with glioma cells PA1, PA2, PA3, and PO2_c, PO2_2. Cases PO1 and PO2_m showed no clear evidence of cellular tumor.

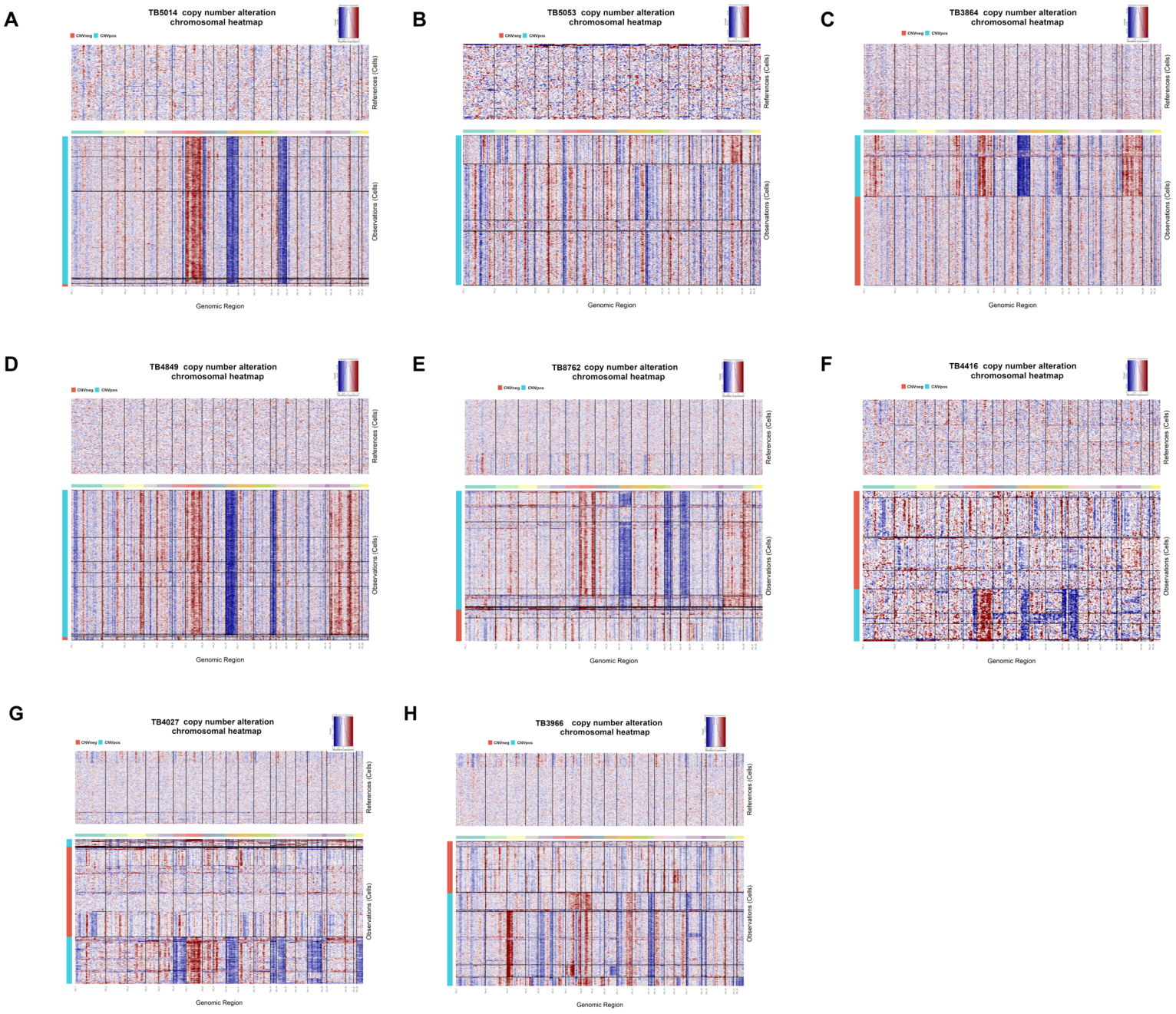

Figure S2: Identifying neoplastic nuclei in post-treatment recurrent GBM by copy number alterations inferred from snRNAseq

Large-scale chromosomal copy number alterations were inferred from RNA expression using InferCNV R package. The heat maps show gains (red) and losses (blue) in case TB5014 (**A**), TB5053 (**B**), TB3864 (**C**), TB4898 (**D**), TB8762 (**E**), TB4416 (**F**), and TB4027 (**G**), and TB3966 (**H**).

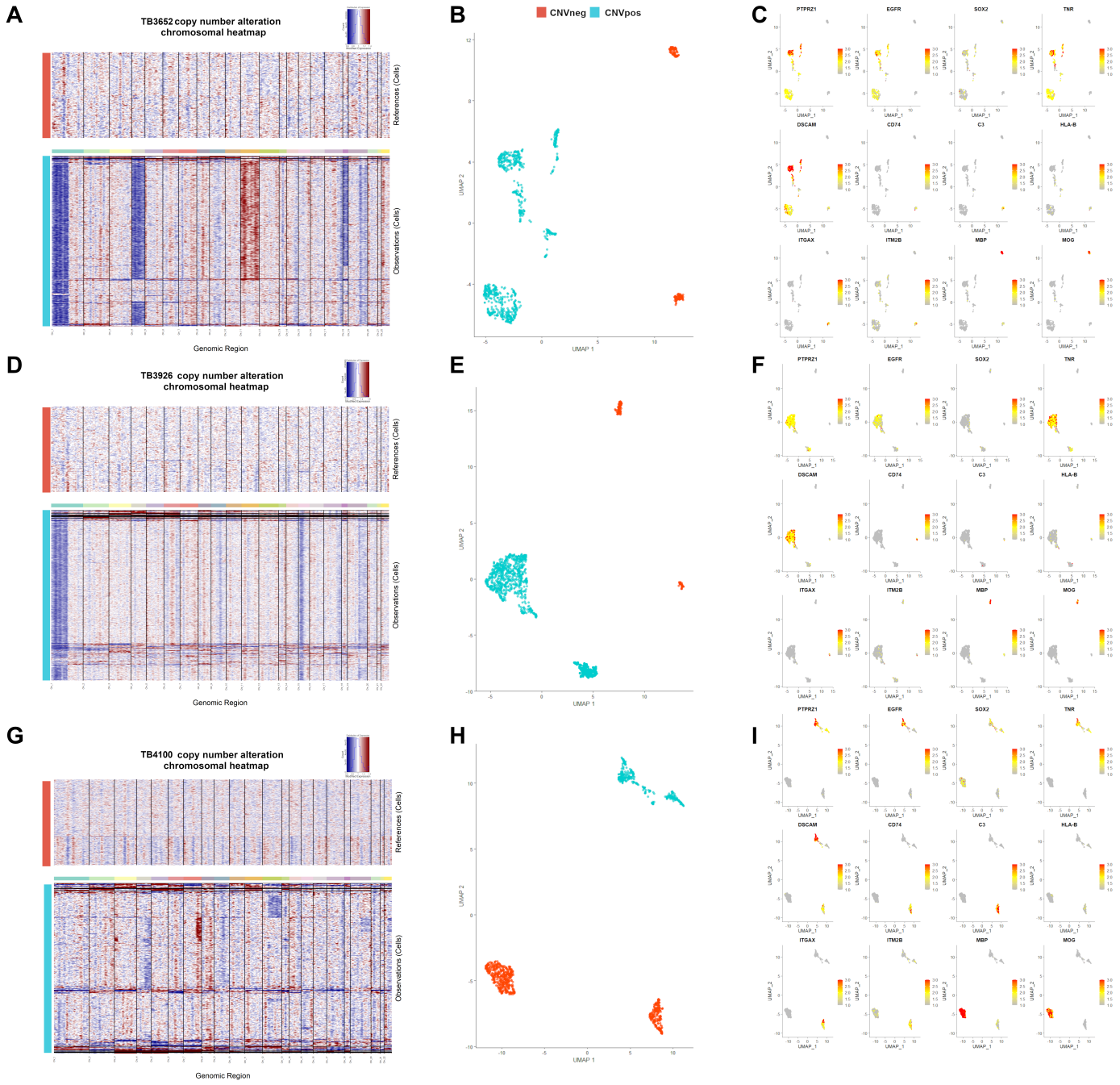

Figure S3: Analysis of Low-grade glioma samples using single nucleus RNAseq

Large-scale chromosomal copy number alterations were inferred from RNA expression of cases TB3652 (**A**), TB3926 (**D**) – both IDH1-mutant oligodendrogliomas, and TB4100 (**G**) – IDH-mutant astrocytoma. Uniform manifold approximation and projection (UMAP) plots of the three cases are shown in panels **B**, **E**, and **H**, color-coded by copy number alteration status. Gene expression UMAPs showing markers of tumor cells (PTPRZ1, EGFR, SOX2, TNR, and DSCAM), immune cells (CD74, C3, HLA-B, ITGAX, ITM2B), and oligodendrocytes (MBP, MOG) of cases TB3652, TB3926, and TB4100 in panels **C, F,** and **I**, respectively.

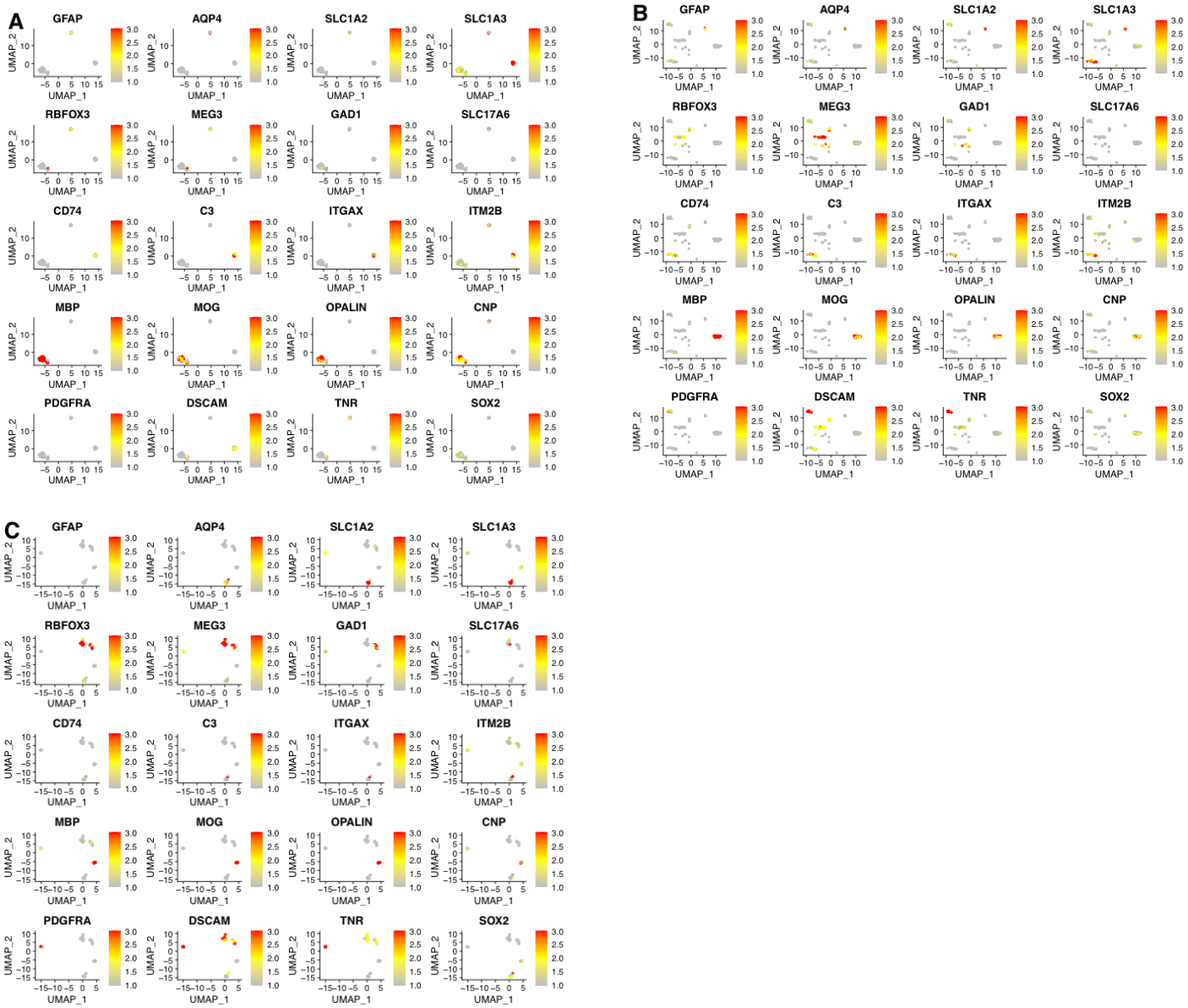

Figure S4: Analysis of Epilepsy samples using single nucleus RNAseq

A-C) Uniform-manifold approximation and projection (UMAP) graphs plots showing normalized gene expression of select lineage markers for cases TB4437 (**A**), TB4189 (**B**), TB4957 (**C**). The markers include astrocyte markers (GFAP, AQP4, SLC1A2, and SLC1A3), neuron makers (RBFOX3, MEG3, GAD1, SLC17A6), myeloid markers (CD74, C3, ITGAX, ITM2B), oligodendrocyte markers (MBP, MOG, OPALIN, CNP), and OPC markers (PDGFRA, DSCAM, TNR, and SOX2).

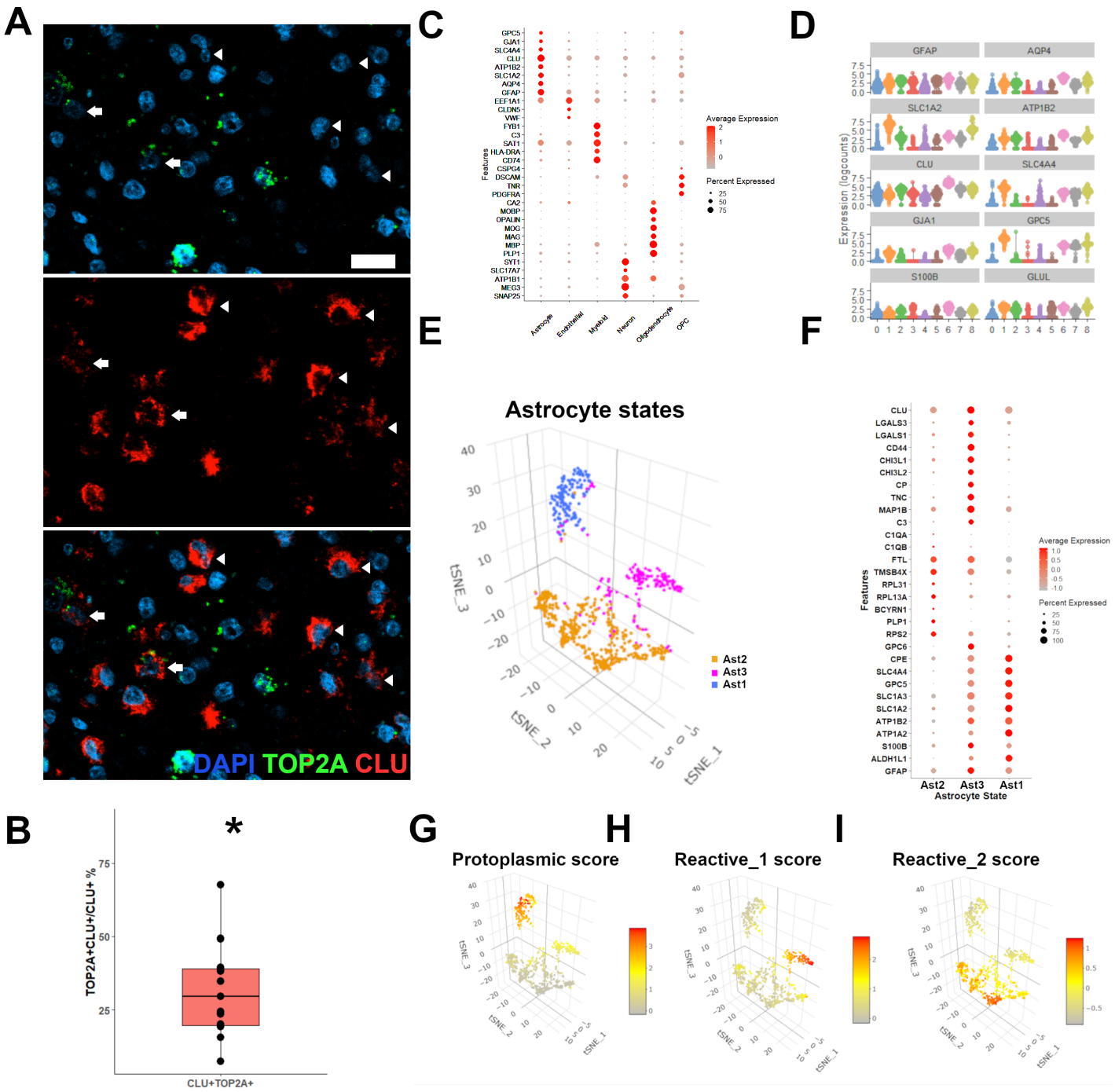

Figure S5: Related to Figure 3. Astrocytes states in glioma

A) Confocal images showing optical sections of in situ hybridization for TOP2A and CLU in the GBM infiltrated tissue. Arrows indicate CLU+TOP2A+ cells, and arrowheads indicate CLU+TOP2A- cells. scale bar = 20 µm. **B)** Quantification of TOP2A and CLU expression. The percentage of TOP2A+CLU+/CLU+ cells is shown as a boxplot. One-sample t-test, N=5. *=p value < 0.001**. C**). Gene expression dot plots showing select gene marker expression for the non-neoplastic lineages. **D**) Violin plots showing expression of select astrocyte markers in the unbiased clusters. Note all clusters show expression of GFAP. **E**) Uniform-manifold approximation and projection (UMAP) graphs plots showing all astrocyte nuclei clustered as per the genesets in supplementary table 5. The astrocyte nuclei were clustered in protoplasmic astrocytes (Ast1), reactive astrocytes with misexpression of non-astrocyte genes (Ast3), and reactive astrocytes with expression of inflammatory genes (Ast3). **F**). Gene expression dot plots showing select gene marker expression for the astrocyte clusters. **G-I**) Gene-set scores for each of the genesets used for clustering in A shown as a color-gradient on the UMAP.

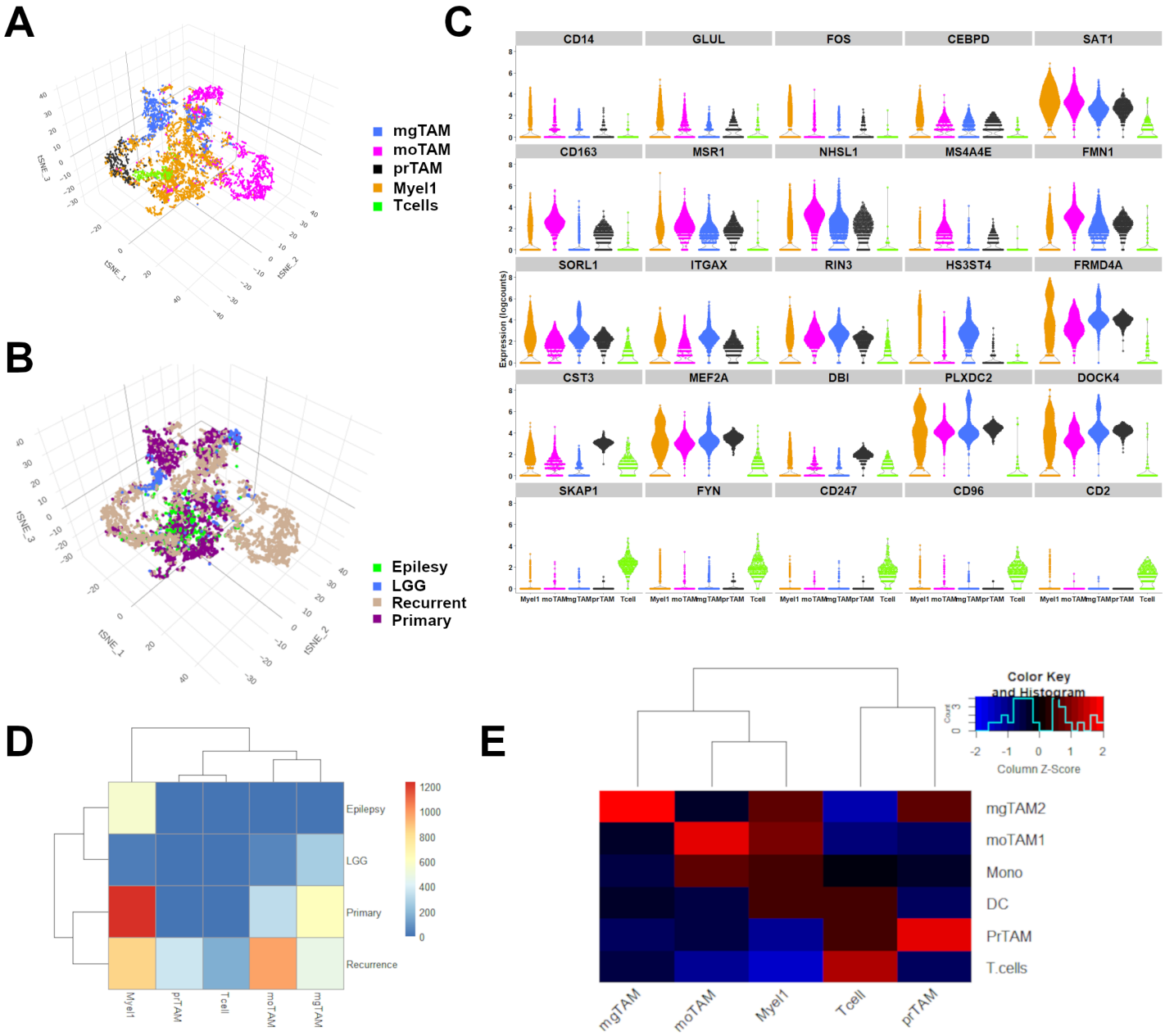

Figure S6: The transcriptional landscape of microglia/myeloid cells in glioma

A) Uniform-manifold approximation and projection (UMAP) graphs plots showing all myeloid nuclei from color-coded by cluster (**B**) and condition (primary glioma, recurrent glioma, low grade glioma (LGG), and epilepsy (**C**). Gene expression violin plots showing select gene marker expression for the immune cell clusters from top to bottom; Myel1, mgTAM, moTAM, prTAM, and T cells. **D**) Heatmap showing the proportion of nuclei in each cluster (columns) contributed by condition (rows). **E**) Heatmap showing the scaled enrichment scores of gene sets derived from Movahedi et al 2021 in the nuclei pooled from each myeloid cluster.

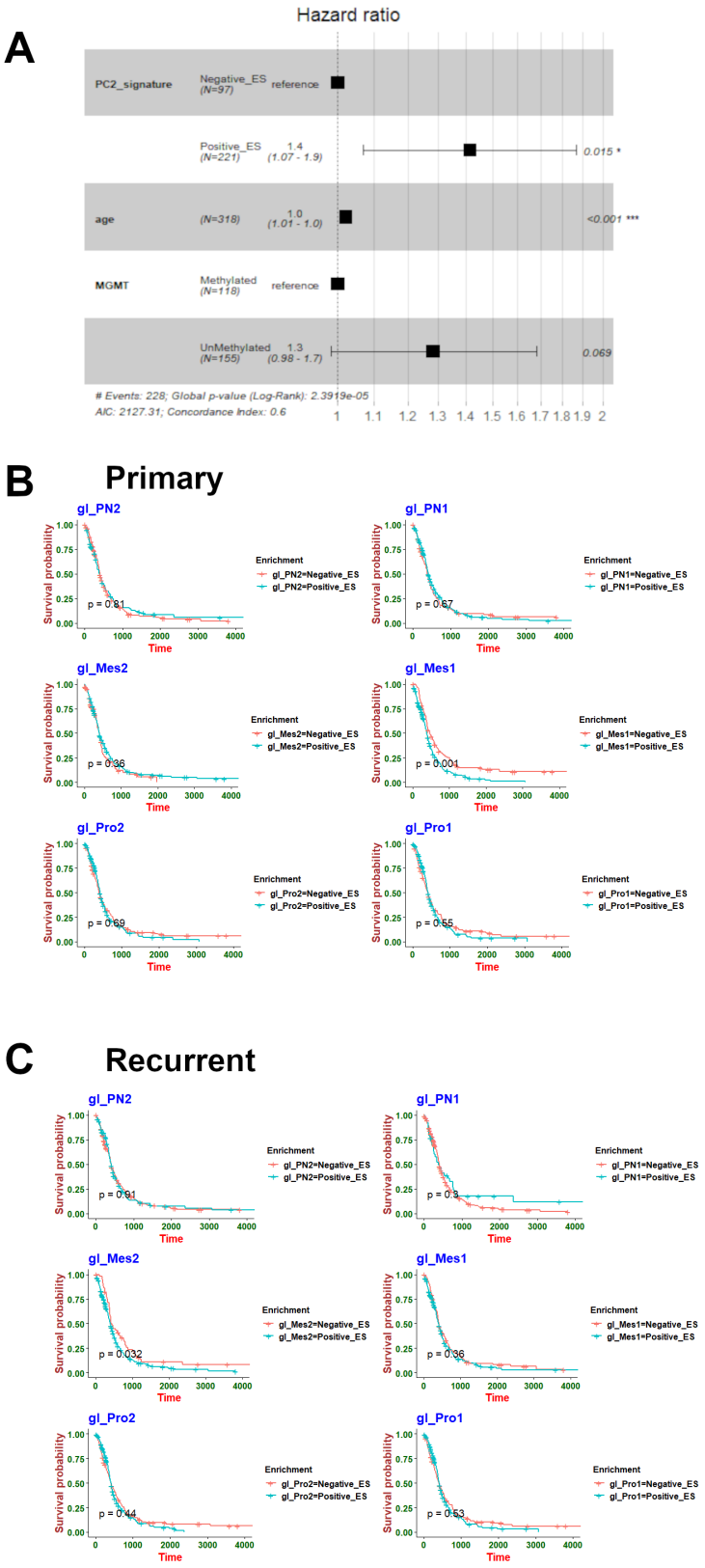

Figure S7: Related to Figure 4, astrocyte-like/mesenchymal glioma-states signatures are associated with poor survival

Cox proportional hazard ratio of survival in the combined TCGA and CGGA dataset given enrichment of each of the PC2 gene signature (Figure 4B). Age and MGMT-methylation status are also included in the model. The p values are shown on the left, bars indicate confidence intervals (also noted on the right). Enrichment of the PC2 geneset was categorized as negative or positive.

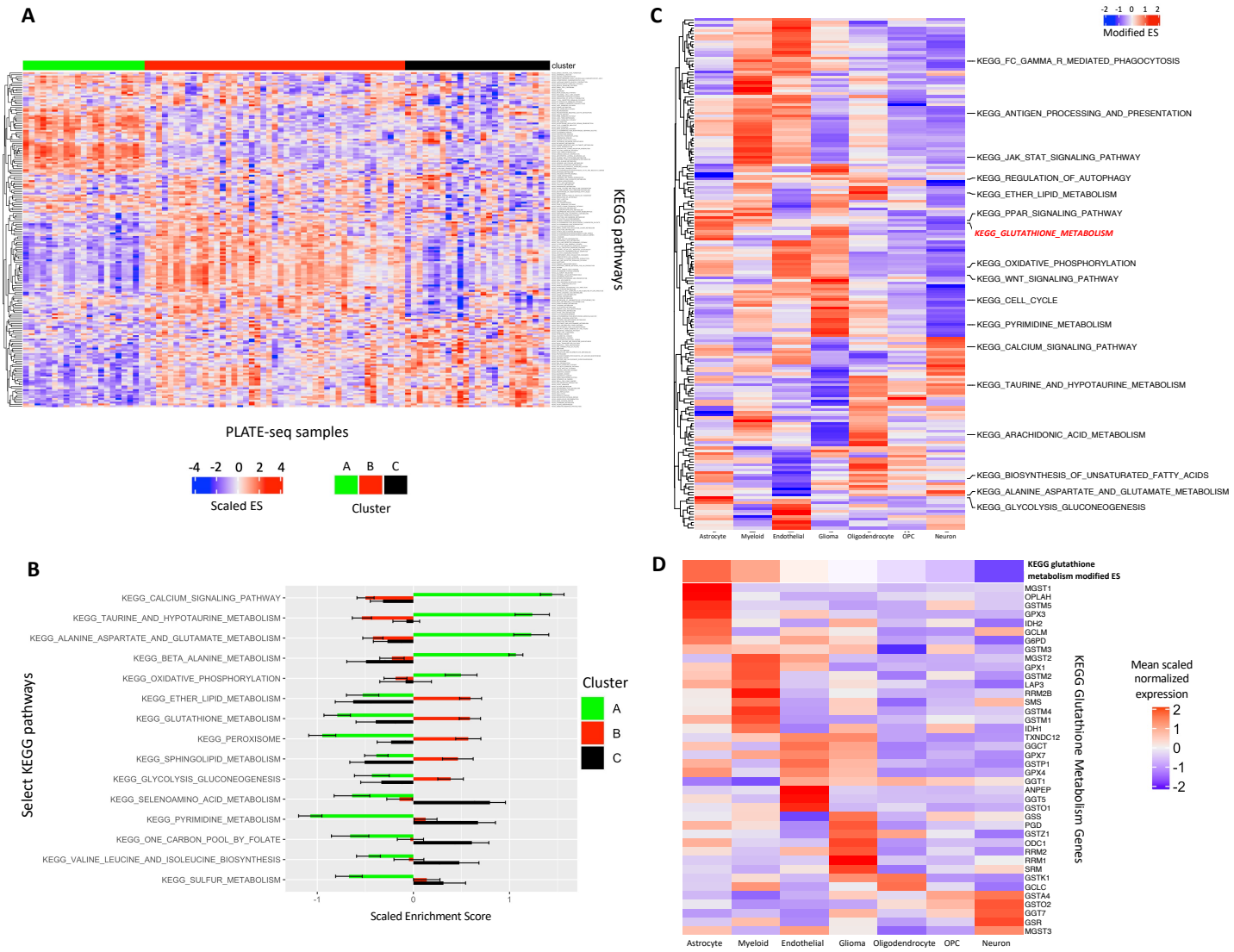

Figure S8: Metabolic pathways in tissue states

A) Heatmap displaying scaled enrichment scores for all KEGG pathways across all PLATE-seq samples. Heatmap is grouped by cluster, annotated by the horizontal bar at the top. Hierarchical clustering was performed on the rows (pathways), demonstrating cluster-specific metabolic programs. **B**) Bar plot displaying scaled ssGSEA scores for select KEGG metabolic programs from A. Bar plots represent mean scaled ssGSEA score ± standard error for each of the three clusters for a given pathway. **C**) Heatmap of lineage-specific modified enrichment score for all KEGG pathways. Select pathways shown in B are identified on the right of the heatmap. **D**) Representative example showing a heatmap displaying mean lineage-specific scaled normalized expression of genes in the KEGG Glutathione Metabolism gene set – which was most enriched in astrocytes – the modified normalized enrichment score per lineage is shown on the top.

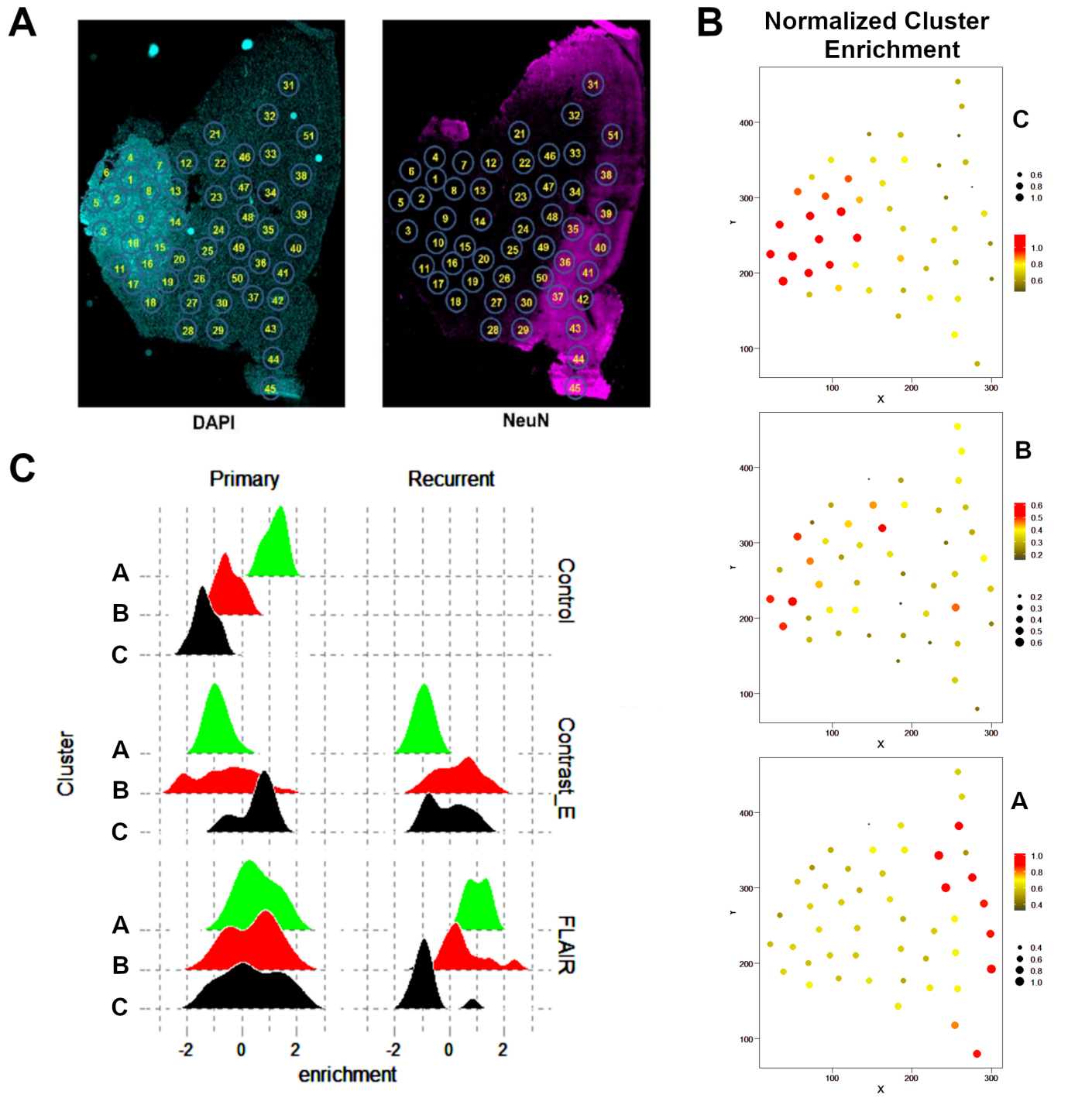

Figure S9: The spatial landscape of tissue states in GBM

A) Outline of spatial transcriptomic analysis of infiltrating margin of GBM. DAPI (left) and NeuN (right) immunostains of frozen sections from case PO2, for which snRNAseq was done. Each circle represents a biopsy on which bulk RNAseq was done. After the biopsies were taken, the specimen was bisected along the dashed white line (y-axis) and subjected to snRNASeq. **B**) Enrichment analysis of each of the spatially mapped biopsies using the genesets of the three compositional clusters (see text for details) displaying normalized single sample GSEA enrichment scores for the tumor cluster (C - upper panel), the tumor-reactive cluster (B– middle panel), and the normal brain cluster (A – lower panel). The enrichment scores are coded by color and size. The normalized RNA data for the spatial biopsy map is available in an interactive web interface at https://vmenon.shinyapps.io/gbm_expression/. **C**) Enrichment of the compositional cluster gene sets as in B, applied to the Gill et al. 2014 MRI localized biopsy dataset. The normalized enrichment scores for the shown in histograms for the control brain, contrast enhancing (Contrast_E) samples, and FLAIR-abnormal margin samples from primary and recurrent samples.

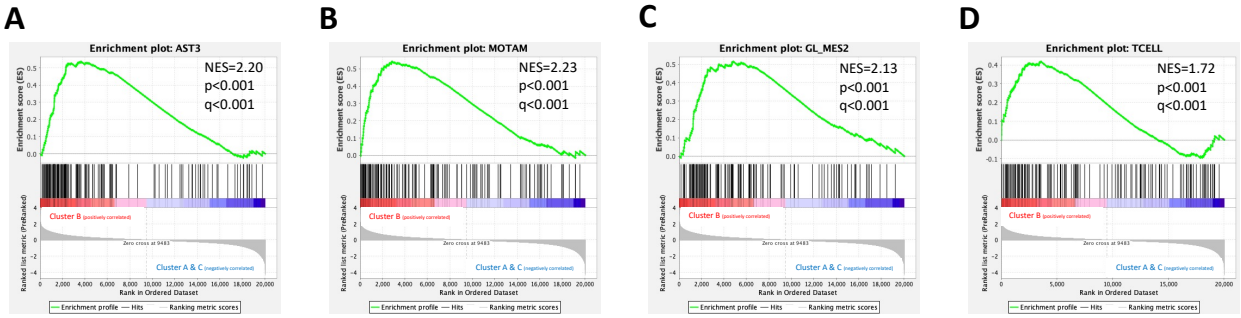

Figure S10: Enrichment of select cell type-specific gene signatures in Cluster B

Pre-ranked Gene Set Enrichment Analysis (GSEA) comparing Cluster B samples with Cluster A & C samples for 4 sub-lineages: Ast3 (**A**), MoTAM (**B**), gl_Mes2 (**C**), and T-cells (**D**). Marker genes for each sub-lineage were used as the gene set for each analysis. Normalized Enrichment Score (NES) is displayed, along with p-value and FDR-adjusted q-value.