Science Advances
AAAS

Supplementary Materials for

Neural network learning defines glioblastoma features to be of neural crest perivascular or radial glia lineage

Yizhou Hu *et al.*

Corresponding author: Patrik Ernfors, patrik.ernfors@ki.se

Sci. Adv. **8**, eabm6340 (2022)
DOI: 10.1126/sciadv.abm6340

The PDF file includes:

Figs. S1 to S5

Other Supplementary Material for this manuscript includes the following:

Data files S1 to S9

Supplementary Information

Table of contents:

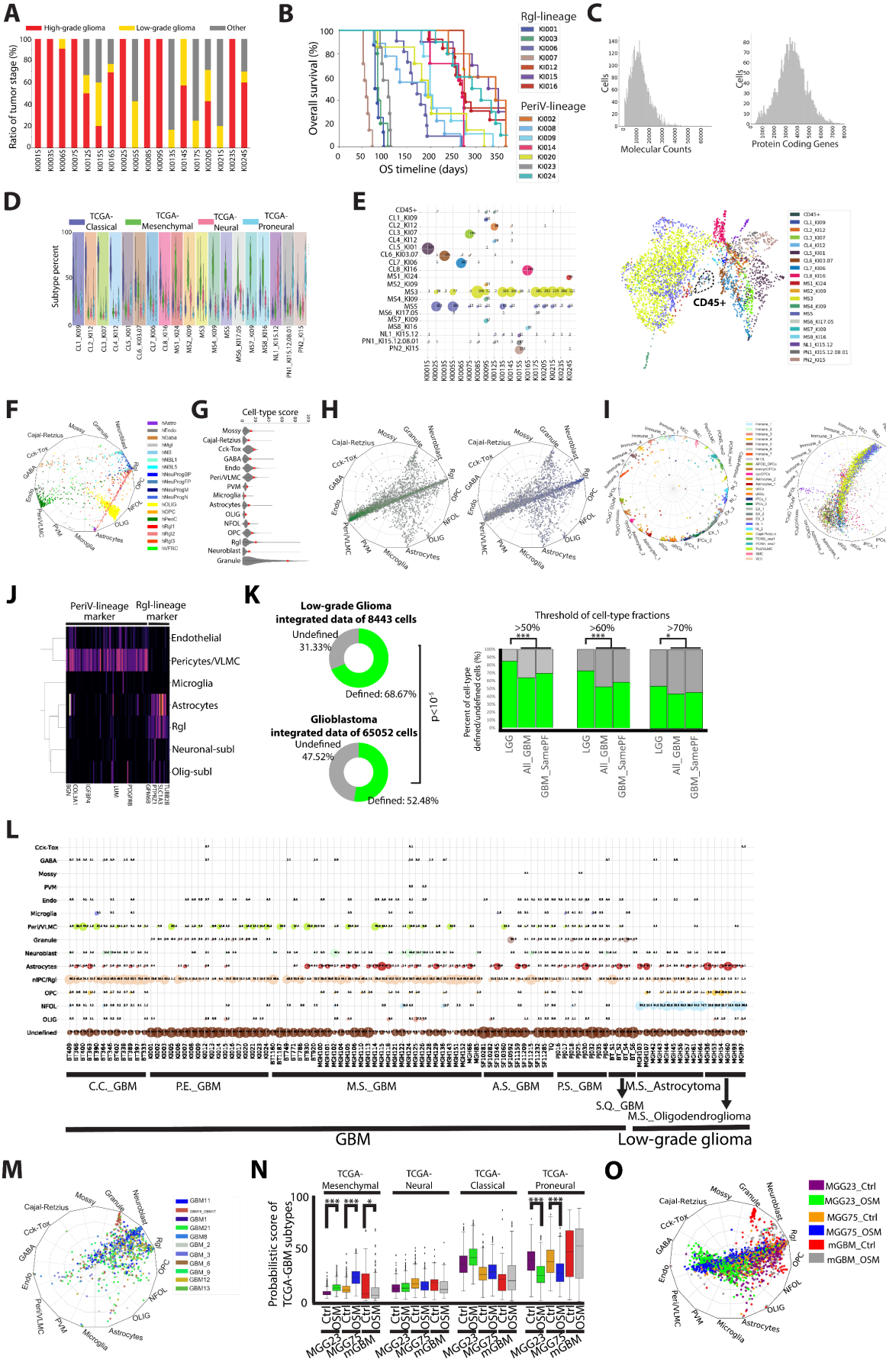

A
High-grade glioma
Low-grade glioma
Other
Ratio of tumor stage (%)
B
Overall survival (%)
OS timeline (days)
Rgl-lineage
PeriV-lineage
C
Cells
Molecular Counts
Protein Coding Genes
D
TCGA-Classical
TCGA-Mesenchymal
TCGA-Neural
TCGA-Proneural
Subtype percent
E
CD45+
F
Cajal-Retzius
Cck-Tox
GABA
Endo
Peri/VLMC
PVM
Microglia
Astrocytes
OLIG
NFOL
OPC
Rgl
Neuroblast
Granule
Mossy
G
Cell-type score
H
I
J
PeriV-lineage marker
Rgl-lineage marker
Endothelial
Pericytes/VLMC
Microglia
Astrocytes
Rgl
Neuronal-subl
Olig-subl
K
Low-grade Glioma integrated data of 8443 cells
Undefined 31.33%
Defined: 68.67%
Glioblastoma integrated data of 65052 cells
Undefined 47.52%
Defined: 52.48%
p<10-5
Threshold of cell-type fractions
>50%
>60%
>70%
Percent of cell-type defined/undefined cells (%)
LGG
All_GBM
GBM_SamePF
L
C.C._GBM
P.E._GBM
M.S._GBM
A.S._GBM
P.S._GBM
M.S._Astrocytoma
S.Q._GBM
M.S._Oligodendroglioma
GBM
Low-grade glioma
M
N
TCGA-Mesenchymal
TCGA-Neural
TCGA-Classical
TCGA-Proneural
Probabilistic score of TCGA-GBM subtypes
Ctrl
OSM
MGG23
MGG75
mGBM
O
MGG23_Ctrl
MGG23_OSM
MGG75_Ctrl
MGG75_OSM
mGBM_Ctrl
mGBM_OSM

Fig. S1 Cell type assignment of high- and low-grade glioma at single cell RNA level

(A) Histological classification of orthotopic xenografted glioma malignancy. Percent animals with high-, low-, or undefined/other glioma from cells of patients indicated along the x-axis. Color legend at the top. Tumor grades listed at the top of the chart.

(B) Survival curves of glioblastoma xenograft mice (n=6-10 animals/group). Each curve represents a group of intracranial xenograft mice injected with glioblastoma cells from one patient. Colors indicate the patients.

(C) Up, distribution of number of unique mRNA molecule counts detected by scRNA-seq in human glioblastoma cells. Down, distribution of number of protein coding genes detected in human glioblastoma cells.

(D) Percent of glioblastoma cells assigning to TCGA subtype classification for each tumor cell cluster as determined by a neural-network scoring model and visualized in violin plot. The highest median subtype scores of each glioblastoma cluster were used to define the TCGA-subtype of each glioblastoma cluster. Colors are indicated in the legend at the top.

(E) Left, quantification of the cell-type distribution of each glioblastoma cell cluster within the individual patients, indicated along the x-axis. The dot sizes represent cell numbers of each cluster found in the patient. The colors represent the defined clusters. Cell number is marked at the up-right side of each dot. Right, scRNAseq and clustering of patient-derived glioblastoma cells. UMAP visualization of glioblastoma cells from 18 patients. CD45+ cluster is outlined. Color coding based on unique cell clusters, and color legend at right.

(F) Radar plot visualization of cell-type scores of human brain cells in relation to the trained reference cell types. Hochgerner et al. data was used for training (*20*) and testing was performed by an integrated dataset of human embryonic midbrain (*21*) and cortex (*22*). Each dot represents one cell. Color coding is based on cell types. The position of each dot indicates the cell-type scores between that cell and the training cell types- which are indicated outside of each bend in the radar wheel. Abbreviations: Cajal-Retzius, Cajal-Retizus cells; Endo, Endothelial cells; GABA, GABAergic neurons; Granule, granule neurons; Mossy, mossy cells; NFOL, newly formed oligodendrocytes; OPC, oligodendrocyte precursor cells; Peri, pericytes; PVM, perivascular macrophage; Rgl, Rgl cells and neuronal intermediate progenitor cells; VLMC, vascular and leptomeningeal cells.

(G) Violin plot of cell-type scores obtained from the same model in (F) after random permutation of the features in the training dataset. Red X represent the 95% confident value. All reference cell types in this model have low baseline (<20) except granule cells, which is around 60.

(H) Same radar plot as Fig. 1B, with different color coding. Each cell was classified based on similarity to Peri/VLMC cells (green) or Rgl cells (blue) and color coded thereafter. Gray color represents cells with low similarity to either of the two cell types.

(I) Left, radar-plot showing cell-type scores of the reference dataset after training by the neural-network scoring model and testing the same dataset. This reference dataset contains cells of the human fetal brain (gestational week 9 to 28)(*23*). Right, Radar plot visualization of cell-type score of glioblastoma cells in relation to the trained cell types. Color coding based on previously annotated unique cell clusters. The position of each dot indicates the cell-type score of the testing cells with names of trained cell types outside each bend in the radar wheel. Abbreviations: EX, cortical excitatory neuron; IN, inhibitory neuron; Cajal-Retzius, Cajal-Retzius cell; Pons-neu, projection neuron in pons; cycOPCs, cycling oligodendrocyte precursor cells; noncycOPCs, non-cycling

oligodendrocyte precursor cells; APOD_OPCs, APOD oligodendrocyte precursor cells; NFOL, newly formed oligodendrocytes; Astrocytes, astrocytes; vRGs, ventricular radial glial cells; oRGs, outer sub-radial glial cells; IPCs, intermediate progenitor cells; Peri/VLMC, pericytes/vascular and leptomeningeal cells; SMC, smooth muscle cell; VEC, vascular endothelial cell; Immune, immune cells.
(J) Heatmap of differential gene expression among the lineage types of brain cells from the mouse dentate gyrus (*20*). Hierarchical clustering was performed via the linkage method of "average" on the correlation distance between the observations. Each row represents one cell lineage type, as shown at the right side. Each columns represent the differentially expressed genes same to Figure 1C, and the labeled lineages were listed at the top. Selected differential expressed gene names listed at the bottom. The colors from dark-blue to light-yellow represents the expression levels from the minimal to maximal.
(K) Left, donut charts showing the quantitative distribution of low-grade glioma cells (*9, 25*) (left) or glioblastoma cells (*5, 7, 10, 24, 26*) merged from our data and available published studies (right) of Fig. 1 D and E, respectively. Green color represents cells that could be defined by a high machine learning cell-type score and grey color represents undefined cells, thus with low cell-type score. Right, percentage bar chart of the quantitative distribution of low-grade glioma cells, LGG (*9, 25*) or glioblastoma cells, All_GBM (*5, 7, 10, 24, 26*), or glioblastoma cells processed from the same platform, GBM_SamePF (*5*). Green color represents the defined cells (high cell-type score), and grey color represents the undefined cells (low cell-type score).
(L) The percent of low-grade glioma or glioblastoma cells from 100 patients assigning with high cell-type score by machine learning to endogenous cell types of the brain. Dot colors represent the defined normal brain cell types, and the dot size represents the cell percentage in each patient. Endogenous cell types of the brain were indicated on the Y-axis. Integrated analysis of available data from scRNA sequenced glioma studies, as indicated on the x-axis. P.E. GBM, data from this study; C.C._GBM, data from(*7*), M.S. GBM, data from (*5*), A.S._GBM, data from (*10*), P.S._GBM, data from (*26*), S.Q._GBM, data from (*24*), M.S._Astrocytoma, data from (*25*), M.S._Oligodendroglioma, data from (*9*).
(M) Radar plot visualization of cell-type scores of brain tumor cells carrying GFAP-Cre induced genetic alterations (*27*) in relation to the trained cell types as described in (F). Each dot represents one cell. Color coding is based on sample ID. The position of each dot indicates the cell-type scores between that cell and the training cell types- which are indicated outside of each bend in the radar wheel.
(N) Box chart represents the significant TCGA-subtype score of control and OSM treated GBM cells (OSM). The box colors indicated in (O). *, $p<0.05$; ***, $p<0.001$.
(O) Radar plot visualization of cell-type scores of controls or OSM treated GBM cells (*27*) in relation to the trained cell types as described in (F). Each dot represents one cell. Color coding is based on sample ID. The position of each dot indicates the cell-type scores between that cell and the training cell types- which are indicated outside of each bend in the radar wheel. The dot colors indicated at the right side.

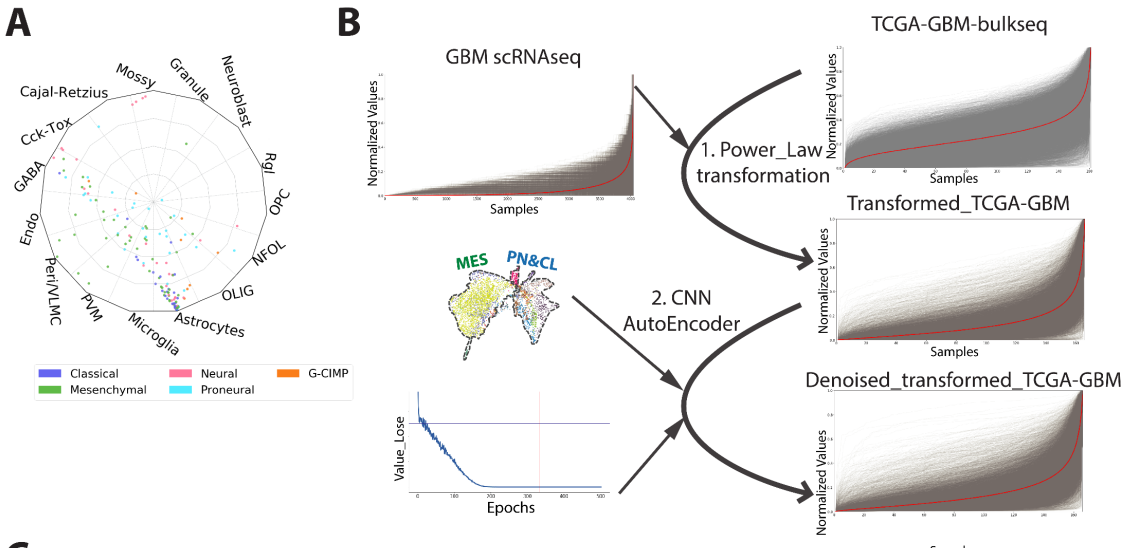

A
Cajal-Retzius
Mossy
Granule
Neuroblast
Rgl
OPC
NFOL
OLIG
Astrocytes
Microglia
PVM
Peri/VLMC
Endo
GABA
Cck-Tox
Classical
Mesenchymal
Neural
Proneural
G-CIMP
B
GBM scRNAseq
Normalized Values
Samples
1. Power_Law transformation
TCGA-GBM-bulkseq
Transformed_TCGA-GBM
MES
PN&CL
2. CNN AutoEncoder
Value_Lose
Epochs
Denoised_transformed_TCGA-GBM
C
G-CIMP
Olig-subl 28.6%
Astro-subl 14.3%
Rgl-like 57.1%
Proneural
Olig-subl 25.0%
Astro-subl 20.0%
Rgl-like 55.0%
Classical
Rgl-like 12.9%
PeriV-lineage 3.2%
Astro-subl 83.9%
Neural
Astro-subl 52.6%
Rgl-like 36.8%
PeriV-lineage 10.5%
D
Cck-Tox
GABA
Mossy
PVM
Endo
Microglia
Peri/VLMC
Granule
Neuroblast
Astrocytes
Rgl
OPC
NFOL
OLIG
Classical
Mesenchymal
Neural
Proneural
E
Overall survival
OS Timeline (Months)
Rgl-like (N=27)
PeriV-lineage (N=19)
Astro-subl (N=51)
Olig-subl (N=9)
p values:
PeriV _ Rgl: 0.021
PeriV _ Astro: 0.048
PeriV _ Olig: 0.026
F
Rgl-lineage
PeriV-lineage
Astro-subl
Olig-subl
Rgl-like
TP53
TTN
PTEN
AHNAK2
FLG
EGFR
PKHD1
PCLO
HYDIN
MUC17
OBSCN
PIK3CA
ATRX
ADGRV1
NF1
Cell_Type
Gender
TCGA_Subtypes
G
Log10_pVal
Log2 Fold_Change
PeriV-lineage GBM
Rgl-lineage GBM
H
Training_Reference:
Prediction_Data:
Probabilistic Score of Rgl-type Tumor
Probabilistic Score of PeriV-type Tumor
84.6%
61.1%
Rgl-lineage
PeriV-lineage
Undefined
I
Overall survival
OS Timeline (Months)
Rgl-lineage (N=18)
PeriV-lineage (N=4)
Undefined (N=7)
IDHmut (N=10)

Fig. S2 Cell-type assignment, methylation status and survival of deconvoluted bulk tumor data

(A) Cell-type scores of glioblastoma bulk RNA sequenced tumors generated from the original sequencing data of the TCGA (*2*) from the UCSC Cancer Browser, and visualized in relation to brain cell types as a radar plot. Each dot represents one bulk tumor.

(B) A schematic view of the workflow for deconvolution of bulk sequencing data. Step 1, Power-Law transformation, by comparing the γ values of both GBM bulk tissue data from TCGA and our reference glioblastoma single cell data, the expression matrix of bulk sequencing was fit to the same distribution of single cell sequencing via Power – Law transformations. Step 2, CNN-Autoencoder, after the training of the reference glioblastoma scRNAseq data with the defined clusters (indicated as butterfly shape UMAP plot), the CNN model was performed for the deconvolution of the transformed dataset of glioblastoma bulk tissue. For the curve plot in grey color, the scaled expression of each gene was visualized by a curved line plot, the x axis represents the cell/sample which was sorted by the expression value of the gene. Thus, we obtained the distribution of gene expression of each dataset and visualized. Red curve in each plot represents the mean values of all grey curves. For the learning rate plot in blue color at the left bottom, the mean squared error between each element in the input (MSELoss) were evaluated against the epoch. The MSELoss values and epoch steps were visualized as a curve plot.

(C) Pie charts showing the quantitative distribution of assignment of glioblastoma cells to the different endogenous cell types of the Rgl-lineage (Rgl-like, radial glia cells like; Olig-subl, oligodendrocytes sublineage; Astro-subl, Astrocytes sublineage) and PeriV-lineage (perivascular lineage) among TCGA-subtypes of glioblastoma. Colors represent the corresponding normal brain cells.

(D) The percent of cell-type defined glioma cells assigning to TCGA-subtypes by machine learning. TCGA-subtypes of glioblastoma indicated on the x-axis. Endogenous cell types of the brain indicated on the Y-axis. Dot colors represent the defined normal brain cell types, and the dot size represents the cell percentage in each TCGA-subtype of glioblastoma.

(E) Patient survival of IDH1 wt glioblastoma from TCGA stratified to the Rgl-lineage (Rgl-like, radial glia cells like; Olig-subl, oligodendrocytes sublineage; Astro-subl, Astrocytes sublineage) and PeriV-lineage (perivascular lineage). Color legend is listed at the right-top of the plot.

(F) Analysis of the mutational status of glioblastoma from TCGA after stratification into Rgl-lineage type and PeriV-lineage type glioblastoma as well as sublineages as in (E). Spike line plot of the significant mutational status. Defined cell type (color legend same as Fig. 2B), gender (red, female; blue, male), and TCGA-subtypes (color legend same as Fig. 2A) is indicated at the bottom.

(G) Volcano plot represents the differential methylated sites between Rgl-lineage and PeriV-lineage type glioblastoma tumors. The lineage-type specific markers are highlighted with color.

(H) Binary classification of glioblastoma using a neural-network scoring model based on methylation. Violin plots represent the probabilistic score of Rgl-lineage or PeriV-lineage tumors. Left, according to FDR significance at both sides (red dash line, <0.05), the reference tumors were assigned to PeriV-lineage or Rgl-lineage. Cell-type defined tumors at transcriptional level were colored with blue (Rgl-lineage) and green (PeriV-lineage). Right, the model was thereafter applied for scoring the prediction datasets,

and all tumors were classified into 3 groups: Rgl-lineage or PeriV-lineage according to FDR p <0.05, or undefined-type that did not show significant FDR value.
(I) Patient survival of IDH1 wt CCGA-glioblastomas assigned to Rgl-lineage, PeriV-lineage or undefined glioblastoma based on methylation. Color legend is listed at the right-top of the plot.

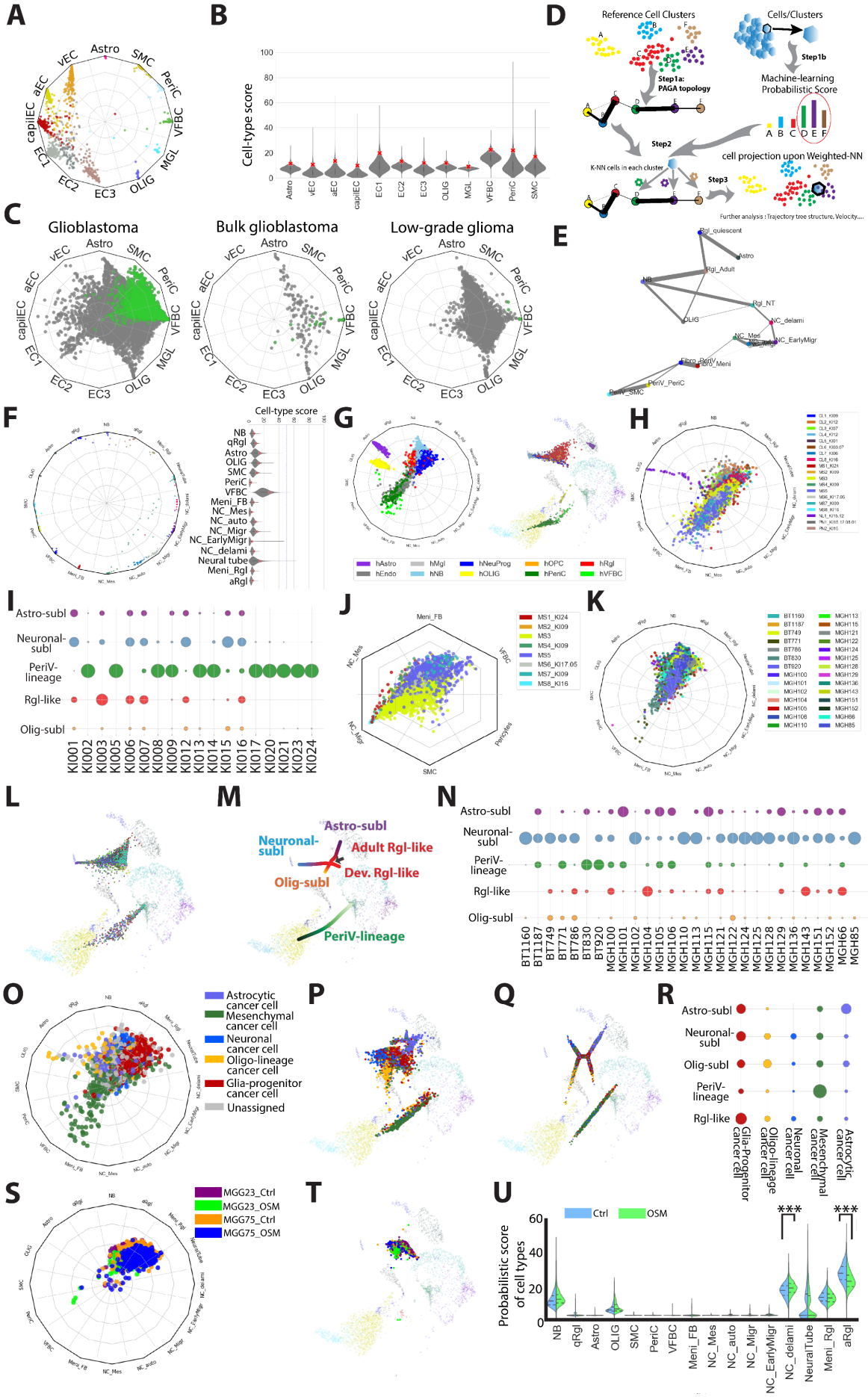

A
B
Cell-type score
D
Reference Cell Clusters
Cells/Clusters
Step1a: PAGA topology
Step1b
Machine-learning Probabilistic Score
Step2
K-NN cells in each cluster
Step3
cell projection upon Weighted-NN
Further analysis : Trajectory tree structure, Velocity....
C
Glioblastoma
Bulk glioblastoma
Low-grade glioma
E
F
G
H
I
Astro-subl
Neuronal-subl
PeriV-lineage
Rgl-like
Olig-subl
J
K
L
M
Astro-subl
Neuronal-subl
Adult Rgl-like
Dev. Rgl-like
Olig-subl
PeriV-lineage
N
O
Astrocytic cancer cell
Mesenchymal cancer cell
Neuronal cancer cell
Oligo-lineage cancer cell
Glia-progenitor cancer cell
Unassigned
P
Q
R
S
MGG23_Ctrl
MGG23_OSM
MGG75_Ctrl
MGG75_OSM
T
U
Ctrl
OSM
Probabilistic score of cell types
***
***

Fig. S3 Assignment of glioblastoma cells to the developing central nervous system and neural crest using SWAPLINE

(A) Radar plot showing cell-type scores of the reference normal brain and blood vessel cell types after training to build learning model and thereafter testing the reference dataset. Abbreviations: Astro, astrocytes; aEC, arterial and arteriolar endothelial cells; capilEC, capillary endothelial cells; EC1-3, endothelial cells1-3; MGL, microglia; OLIG, oligodendrocytes; PeriC, pericytes; SMC, smooth muscle cells; vEC, venous endothelial cells; vFB, vascular fibroblast-like cells.

(B) Violin plot represents the baseline cell-type scores obtained from the same model as used in (A) after random permutation of the features in the training dataset. Red X represent the 95% confident value. All reference cell types in this model has low baseline.

(C) Radar plot showing the cell-type scores of integrated datasets of glioblastomas (left), deconvoluted glioblastomas from TCGA (middle), and low-grade glioma (right) in relation to the reference cell-types. Green, the previous defined perivascular--lineage cells and tumors; grey, all cells/tumors not previously defined as of the perivascular-lineage.

(D) Workflow of SWAPLINE method, which includes three PCR-like steps: Step1, denaturation, step 1a, deconstructing the topology (PAGA) of the reference developmental cell types and confirming the relatively flatten topology of the reference plot. Step1b, calculate the probabilistic score of each cell to each reference cell type. Step2, annealing, select the top topologically nearest cell types with the combined highest probabilistic score of that predicting cell as anchor cell type, and then extract the coordinates of top N nearest cells of the predicting cell from each anchor cell type. Step3, extension, the projected coordinate of that predicting cell was calculated via the linear combination of probabilistic score weighted distance of all nearest cells from all anchor cell types. All predicting cells were assigned one by one via these steps.

(E) Topology-preserving map of single cells extrapolated from Partition-based graph abstraction (PAGA). Each cluster node represents the central position of the cluster in the UMAP. From the start point of neural crest cell delamination (NC-delami in the plot), maximum top 3 linked lines of each cluster were selected following the lineage trace, indicating the maximum top 3 NN based closest clusters of that cluster. The line width represents the relation intensity between cluster node, the wider line, the higher relation intensity. Abbreviations: Astro, astrocytes; NB, neuroblasts; OLIG, oligodendrocytes; Rgl-quiescent, quiescent adult Rgl cells; Rgl_Adult, active adult Rgl cells; Rgl_NT, developmental Rgl cells and neural tube cells; Fibro_Meni, meningeal fibroblasts; Fibro_PeriV, perivascular fibroblasts; NC_Auto, neural crest autonomic progenitors; NC_EarlyMigr, early migratory neural crest progenitors; NC_LateMigr, late migratory neural crest progenitors; NC_Mes, neural crest mesenchymal progenitors; NC_Sensory, neural crest sensory neuron progenitors; PeriV_PeriC, pericytes; PeriV_SMC, perivascular smooth muscle cells.

(F) Left, radar-plot showing cell-type scores of the reference dataset after training by the neural-network scoring model. Abbreviations: Astro, astrocytes; aRgl, active adult radial glial cells; Meni_FB, meningeal fibroblasts; NB, neuroblasts; NC_auto, neural crest autonomic progenitors; NC_delami, neural crest delaminating progenitors; NC_EarlyMigr, neural crest early migratory progenitors; NC_Migr, neural crest late migratory progenitors; NC_Mes, neural crest mesenchymal progenitors; NeuralTube, developmental radial glial cells and neural tube cells; OLIG, oligodendrocytes; PeriC, pericytes; qRgl, quiescent adult radial glial cells; SMC, perivascular smooth muscle cells; vFB, perivascular fibroblasts. Right, violin plot showing baseline cell-type scores

obtained from the same model as in left panel after random permutation of the features in the training dataset. Red X representing the 95% confident value. All reference cell types in this model have significant low baseline.
(G) Left, radar plot visualization of cell-type scores of human brain cells in relation to the trained mouse brain cell types as described in Fig. S3, F. Each dot represents one cell. Color code is based on cell types defined in (*21, 22*). The position of each dot indicates the cell-type score between that cell and the training (endogenous) cell types- which are indicated outside of each bend in the radar wheel. Abbreviations: hAstro, human astrocytes; hEndo, human endothelial cells; hMgl, human microglia; hNeuProg, human neuronal progenitors; hNB, human neuroblasts; hOLIG, human oligodendrocytes; hOPC, human oligodendrocyte progenitor cells; hRgl, radial glial cells; hPeriC, human pericytes; hvFB, human perivascular fibroblasts. Right, integrative projection of human brain cells to the reference cell types in Figure 3A. Background dots "X" represents the reference cells, corresponding to Figure 3A. Solid dots "." indicate individual human brain cells and their distance to each reference cell types, showing that the model with high accuracy places previously annotated cells into the expected position of the developing nervous system.
(H) Radar plot visualization of cell-type score of glioblastoma cells in relation to the trained cell types. Each dot represents one cell. Color coding based on defined cell clusters. The position of each dot indicates the cell-type score between that cell and the training (endogenous) cell types- which are shown outside of each bend in the radar wheel. Abbreviations: Astro, astrocytes; aRgl, active adult radial glial cells; Meni_FB, meningeal fibroblasts; NB, neuroblasts; NC_auto, neural crest autonomic progenitors; NC_delami, neural crest delaminating progenitors; NC_EarlyMigr, neural crest early migratory progenitors; NC_Migr, neural crest late migratory progenitors; NC_Mes, neural crest mesenchymal progenitors; NeuralTube, developmental radial glial cells and neural tube cells; OLIG, oligodendrocytes; PeriC, pericytes; qRgl, quiescent adult radial glial cells; SMC, perivascular smooth muscle cells; vFB, perivascular fibroblasts.
(I) Distribution of relative cell number of each patient in lineage branches. Dot colors represent the lineage branches, and the dot size represents the cell percentage. Abbreviations: Rgl-like, radial glia cells like; Olig-subl, oligodendrocytes sublineage; Neuronal-subl, neuronal sublineage; Astro-subl, Astrocytes sublineage; PeriV-lineage, perivascular lineage.
(J) Radar plot visualization of cell-type score of glioblastoma cells in relation to the trained cell types of cranial neural crest cell reference dataset via integrating the migrating cranial neural crest cells, neural crest mesenchymal progenitor cells (*33*), meningeal cells (*34*), and brain perivascular cells (*17*), Each dot represents one cell. Color coding based on defined PeriV-lineage cell clusters, The position of each dot indicates the cell-type score between that cell and the training (endogenous) cell types- which are shown outside of each bend in the radar wheel. Abbreviations: Meni_FB, meningeal fibroblasts; NC_Migr, neural crest migratory progenitors; NC_Mes, neural crest mesenchymal progenitors; Pericytes, pericytes; SMC, perivascular smooth muscle cells; VFBC, perivascular fibroblasts.
(K, O, S) Radar plot visualization of cell-type scores of glioblastoma cells from public GBM scRNAseq dataset (*5*) (K), or (*7*) (O), or (*27*) (S) in relation to the trained mouse brain cell types as described in (F). Each dot represents one cell. Color code is based on patient IDs(*5*) (K) or defined states (*7*) (O) or treatment condition (*27*) (S). The position of each dot indicates the cell-type score between that cell and the training (endogenous) cell types- which are indicated outside of each bend in the radar wheel.

Abbreviations: Astro, astrocytes; aRgl, active adult radial glial cells; Meni_FB, meningeal fibroblasts; NB, neuroblasts; NC_auto, neural crest autonomic progenitors; NC_delami, neural crest delaminating progenitors; NC_EarlyMigr, neural crest early migratory progenitors; NC_Migr, neural crest late migratory progenitors; NC_Mes, neural crest mesenchymal progenitors; NeuralTube, developmental radial glial cells and neural tube cells; OLIG, oligodendrocytes; PeriC, pericytes; qRgl, quiescent adult radial glial cells; SMC, perivascular smooth muscle cells; vFB, perivascular fibroblasts.
(L, P, T) Projection of GBM cells from public GBM scRNAseq dataset (*5*) (L), or dataset (*7*) (P), or dataset (*27*) (T) to the reference plot. “X” Markers represent the reference cells. Marker dots “.” represent the projected developmental position of individual glioblastoma cells to native cell types.
(M, Q) the plot of principal branching tree summarizes the differentiational status of each glioblastoma cell from dataset (*5*) (M) or (*7*) (Q). Each branch was composed of glioblastoma cells that belong to that developmental lineage, and cell positions represent their current transcriptional status along that lineage. Colors represent the distance of each cell to the hub point and thus indicating the cell status along that lineage from red to branch specific color (M), or represents the defined states (Q). Abbreviations: Adult radial glial cells-like, Adult Rgl-like; Developmental radial glial cells-like, Dev. Rgl-like; oligodendrocytes-sublineage, Olig-subl; Astrocytes-sublineage, Astro-subl; perivascular lineage, PeriV-subl; Neuronal-sublineage, neuronal -subl.
(N, R) The distribution of relative cell number of each patient from dataset (*5*) (N) or each defined state from dataset (*7*) (R) in the defined lineage branches. Dot colors represent the lineage branches (N) or the originally defined type (R), and the dot size represents the cell percentage in each patient/state. Abbreviations: Rgl-like, radial glia cells like; Olig-subl, oligodendrocytes sublineage; Neuronal-subl, neuronal sublineage; Astro-subl, Astrocytes sublineage; PeriV-lineage, perivascular lineage.
(U) Violin chart represents the change of the significant cell type score between the control and the OSM treated GBM cells (OSM) in relation to the trained mouse brain cell types, as described in (F). The group colors indicated at the top. ***, $p<0.001$.

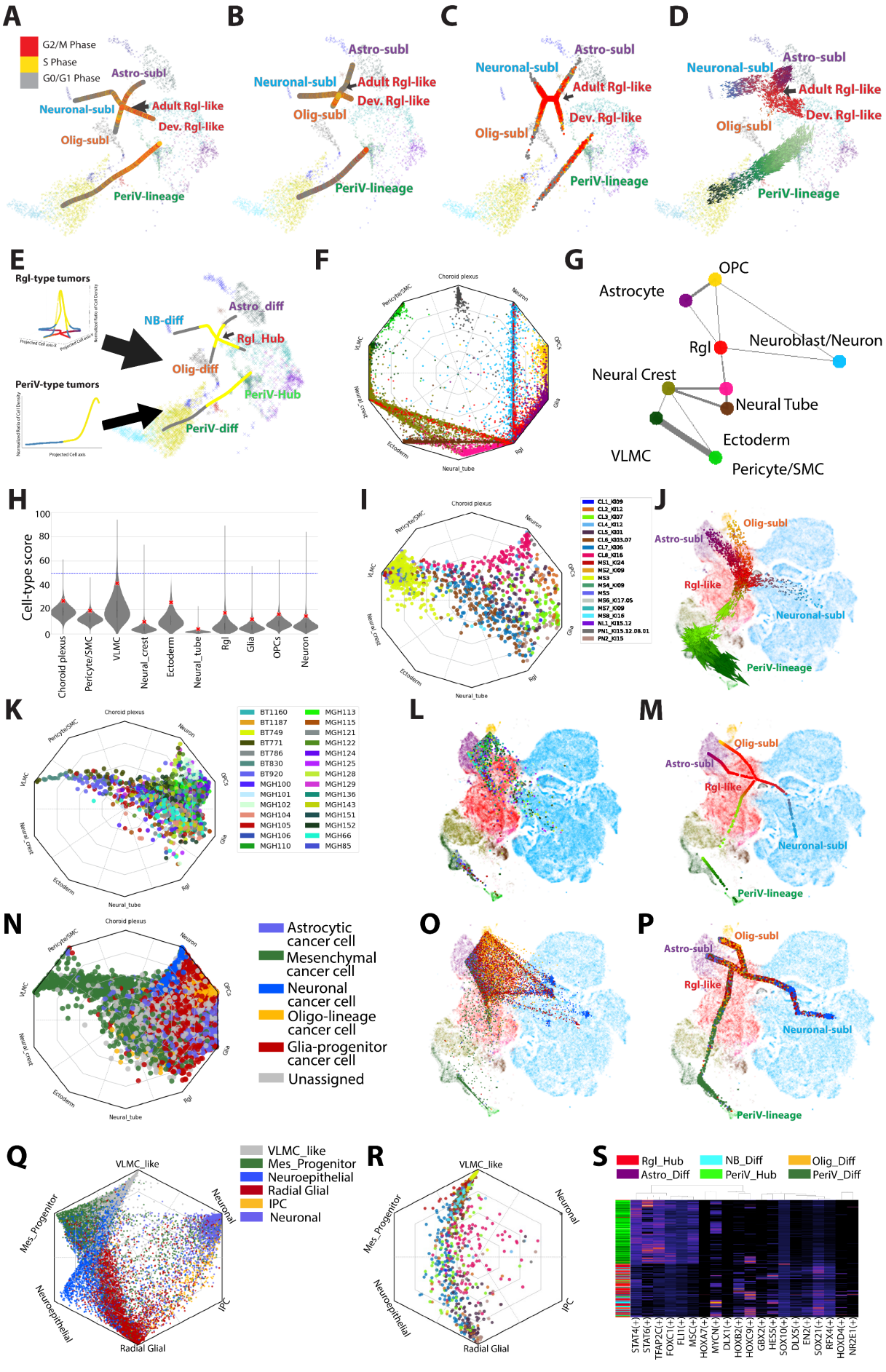

A
G2/M Phase
S Phase
G0/G1 Phase
Astro-subl
Neuronal-subl
Adult Rgl-like
Dev. Rgl-like
Olig-subl
PeriV-lineage
B
C
D
E
Rgl-type tumors
PeriV-type tumors
NB-diff
Astro_diff
Rgl_Hub
Olig-diff
PeriV-Hub
PeriV-diff
F
Choroid plexus
Neuron
OPCs
Glia
Rgl
Neural_tube
Ectoderm
Neural_crest
VLMC
Pericyte/SMC
G
OPC
Astrocyte
Rgl
Neuroblast/Neuron
Neural Crest
Neural Tube
Ectoderm
VLMC
Pericyte/SMC
H
Cell-type score
Choroid plexus
Pericyte/SMC
VLMC
Neural_crest
Ectoderm
Neural_tube
Rgl
Glia
OPCs
Neuron
I
J
K
L
M
N
Astrocytic cancer cell
Mesenchymal cancer cell
Neuronal cancer cell
Oligo-lineage cancer cell
Glia-progenitor cancer cell
Unassigned
O
P
Q
VLMC_like
Mes_Progenitor
Neuroepithelial
Radial Glial
IPC
Neuronal
R
S
Rgl_Hub
Astro_Diff
NB_Diff
PeriV_Hub
Olig_Diff
PeriV_Diff
STAT4(+)
STAT6(+)
TFAP2C(+)
FOXC1(+)
FLI1(+)
MSC(+)
HOXA7(+)
MYCN(+)
DLX1(+)
HOXB2(+)
HOXC9(+)
GBX2(+)
HES5(+)
SOX10(+)
DLX5(+)
EN2(+)
SOX21(+)
RFX4(+)
HOXD4(+)
NR2E1(+)

Fig. S4 Assignment of glioblastoma cells to the developing central nervous system and neural crest, RNA velocity and cell cycle

(A-C) Cell cycle estimations of glioblastoma cells from our dataset (A), from dataset (*5*) (B), and from dataset (*7*) (C), were projected to the principal branching tree and colored by cell cycle status: G2/M phase cycling cells(red), S phase(yellow), and G0/G1 noncycling cells (gray). Abbreviations: Adult radial glial cells-sublineage, Adult Abbreviations: Adult radial glial cells-like, Adult Rgl-like; Developmental radial glial cells-like, Dev. Rgl-like; oligodendrocytes-sublineage, Olig-subl; Astrocytes-sublineage, Astro-subl; perivascular lineage, PeriV-subl; Neuronal-sublineage, neuronal -subl.

(D) Quiver visualization of RNA velocity of glioblastoma cells in branching tree plot. Arrow of each glioblastoma cell pointing to the direction of future status, extrapolated from RNA velocity estimates. Abbreviations: Adult radial glial cells-like, Adult Rgl-like; Developmental radial glial cells-like, Dev. Rgl-like; oligodendrocytes-sublineage, Olig-subl; Astrocytes-sublineage, Astro-subl; perivascular lineage, PeriV-subl; Neuronal-sublineage, neuronal -subl.

(E) Schematic illustration of the cut-off for the extraction of progenitor glioblastoma cells. Left, relative density score of each cell. For Rgl-lineage, projected cells were folked from Figure 3C as a 2-dimension plot at the bottom. The normalized ratio of cell density was measured via kernel-density estimation and visualized at the z axis. For PeriV lineage, projected cells were visualized as a line plot at the x axis, which represents the central line of the original line plot at the bottom. The normalized ratio of cell density was measured via kernel-density estimation and visualized at the y axis. yellow color indicating the top 50% cells at high ratio of cell density as progenitor glioblastoma cells, blue color indicating the rest 50% cells as differentiated glioblastoma cells. Right, visualization of progenitor and differentiated glioblastoma cells according to the quantification from left plot. The plot was folked from Figure 3C, yellow representing progenitor glioblastoma cells and grey represent differentiated glioblastoma cells. Rgl-sublineage cells (Rgl_Hub), differentiated neuroblast-sublineage cells (NB_diff), differentiated oligodendrocyte- sublineage cells (Olig_diff), differentiated astrocyte- sublineage cells (Astro_diff), neural crest progenitor-sublineage cells (PeriV_Hub), differentiated neural crest perivascular- sublineage cells (PeriV_diff).

(F) Radar-plot showing cell-type scores of the reference dataset after training by the neural-network scoring model and testing the same dataset. Abbreviations: Glia, glioblast/astrocytes; Rgl, radial glial cells; VLMC, vascular leptomeningeal cells; Pericyte/SMCs, pericytes and perivascular smooth muscle cells; Neuron, neuroblasts and maturing neurons; OPCs, Oligodendrocyte precursor cells.

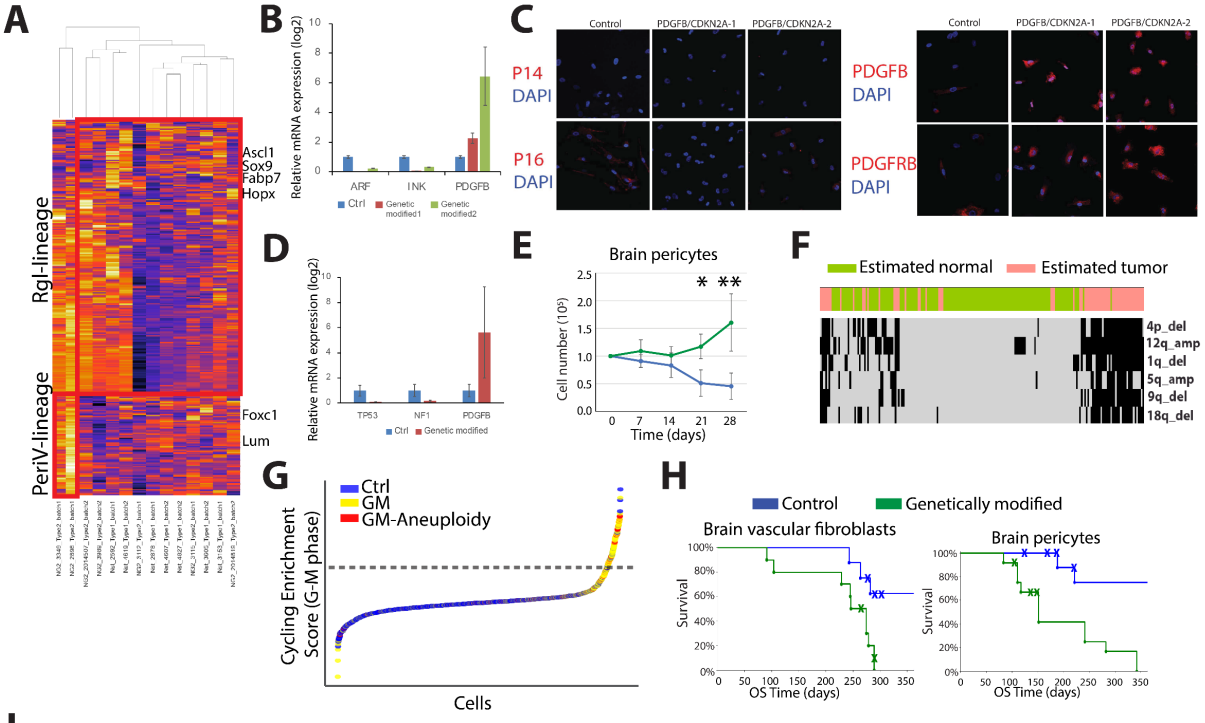
(G) Topology-preserving map of single cells extrapolated from Partition-based graph abstraction (PAGA). Each cluster nodes represent the central position of the cluster in the UMAP. From the start point of Ectoderm, maximum top 3 linked lines of each cluster were selected following the lineage trace, indicating the maximum top 3 NN based closest clusters of that cluster. The Line width representing the relation intensity between cluster node, the wider line, the higher relation intensity.

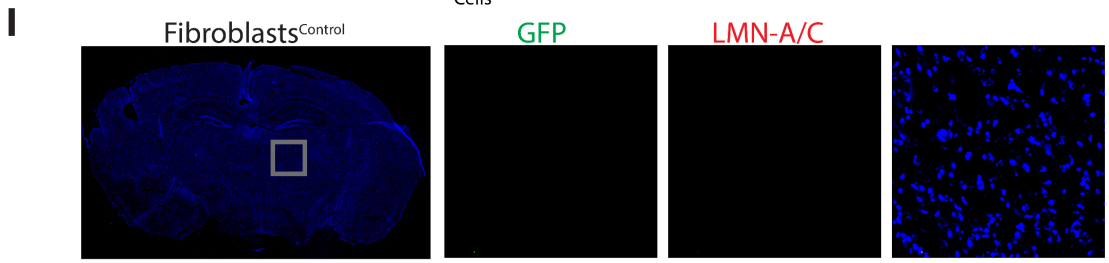
(H) Violin plot showing baseline cell-type scores obtained from the same model as in (F) panel after random permutation of the features in the training dataset. Red X representing the 95% confident value. All reference cell types in this model have significant low baseline.

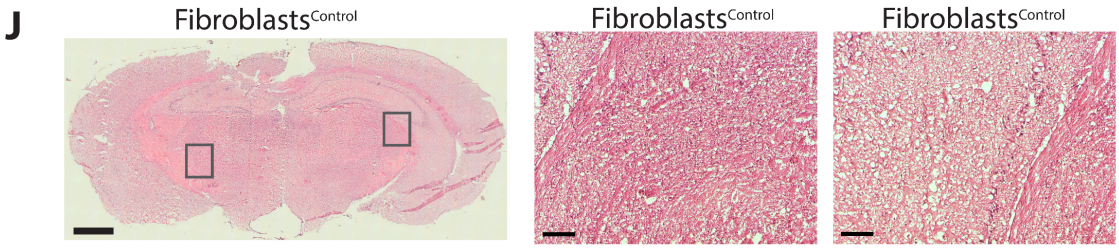
(I) Radar plot visualization of cell-type score of glioblastoma cells in relation to the trained reference cell types of developmental mouse brain. Each dot represents one

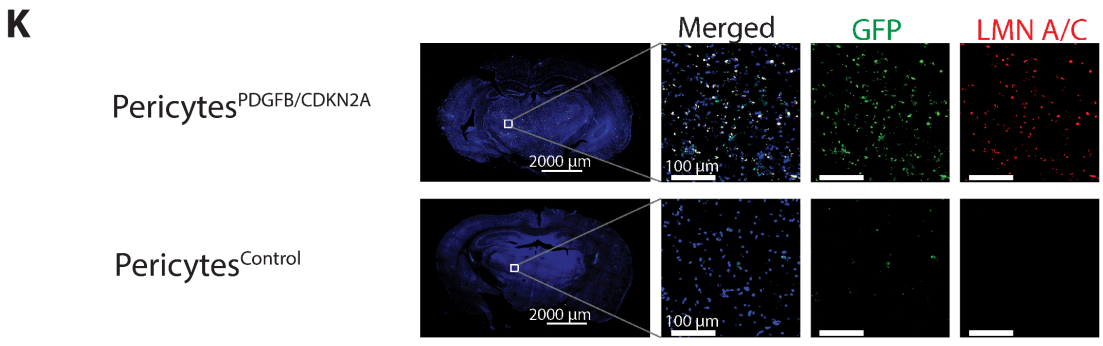
cell. Color coding based on defined cell clusters. The position of each dot indicates the cell-type score between that cell and the training (endogenous) cell types- which are shown outside of each bend in the radar wheel.
(J) Quiver visualization of RNA velocity of glioblastoma cells in branching tree plot. Arrow of each glioblastoma cells pointing to the direction of future status, extrapolated from RNA velocity estimates. Abbreviations: Rgl-like, radial glia cells like; Olig-subl, oligodendrocytes sublineage; Neuronal-subl, neuronal sublineage; Astro-subl, Astrocytes sublineage; PeriV-lineage, perivascular lineage.
(K, N) Radar plot visualization of cell-type score of GBM cells from public GBM scRNAseq dataset (13) (K), and from dataset (60) (N), in relation to the trained cell types of developmental mouse brain. Each dot represents one cell. Color coding based on GBM patients. The position of each dot indicates the cell-type score between that cell and the training (endogenous) cell types- which are shown outside of each bend in the radar wheel.
(L, O) SWAPLINE projection of GBM cells from public GBM scRNAseq public GBM scRNAseq dataset (13) (L) and from dataset (60) (O) to the reference plot of developmental mouse brain. Marker “X” represents the reference cells. Marker “.” represents the projected developmental position of individual glioblastoma cells to native cell types.
(M, P) The plot of principal branching tree summarizes the differentiational status of each glioblastoma cell from public GBM scRNAseq dataset (13) (M) and from dataset (60) (P). Each branch was composed of glioblastoma cells that belong to that developmental lineage, and cell positions represent their current transcriptional status along that lineage. Colors from red to branch specific color represent the distance of each cell to the hub point (M), or represents the defined states (P), and thus indicating the cell status along that lineage. Abbreviations: Rgl-like, radial glia cells like; Olig-subl, oligodendrocytes sublineage; Neuronal-subl, neuronal sublineage; Astro-subl, Astrocytes sublineage; PeriV-lineage, perivascular lineage.
(Q) Radar-plot showing cell-type scores of the reference dataset after training by the neural-network scoring model and testing the same dataset. Dataset is the human fetal brain (gestational week 6 to 10). Abbreviations: Neuronal, neuroblasts & neurons; Radial Glial, radial glial cells; IPC, intermediate progenitor cell; Neuroepithelial, neuroepithelial cells; VLMC-like, vascular and leptomeningeal like cells.
(R) Radar plot visualization of cell-type scores of glioblastoma cells in relation to the trained reference cell types of (Q). color coding based on unique cell clusters. The position of each dot indicates the cell-type score between that cell and the trained cell types- which are indicated outside each bend in the radar wheel. Abbreviations: Neuronal, neuroblasts & neurons; Radial Glial, radial glial cells; IPC, intermediate progenitor cell; Neuroepithelial, neuroepithelial cells; VLMC-like, vascular and leptomeningeal like cells.
(S) Heatmap showing normalized SCENIC regulon activity of the glioblastoma sublineage types. Each row represents a cell, and its sublineage type is indicated as color on the left. The Identified common TF signatures between SCENIC and TF enrichments (Fig 4A-B) are listed at the bottom. The color legend of the sublineage type is listed at the top including differentiated cells of each sublineage (Diff) and the most undifferentiated progenitor like cells (Hub). The colors from dark-purple to yellow represents the expression levels from the minimal to maximal. Rgl-sublineage cells (Rgl_Hub), differentiated neuroblast- sublineage cells (NB_diff), differentiated oligodendrocyte- sublineage cells (Olig_diff), differentiated astrocyte- sublineage cells

(Astro_diff), neural crest progenitor- sublineage cells (PeriV_Hub), differentiated neural crest perivascular- sublineage cells (PeriV_diff).


A
RgI-lineage
PeriV-lineage
Ascl1
Sox9
Fabp7
Hopx
Foxc1
Lum
B
Relative mRNA expression (log2)
ARF
INK
PDGFB
Ctrl
Genetic modified1
Genetic modified2
C
Control
PDGFB/CDKN2A-1
PDGFB/CDKN2A-2
P14
DAPI
P16
DAPI
PDGFB
DAPI
PDGFRB
DAPI
D
Relative mRNA expression (log2)
TP53
NF1
PDGFB
Ctrl
Genetic modified
E
Brain pericytes
Cell number (10⁵)
Time (days)
F
Estimated normal
Estimated tumor
4p_del
12q_amp
1q_del
5q_amp
9q_del
18q_del
G
Ctrl
GM
GM-Aneuploidy
Cycling Enrichment Score (G-M phase)
Cells
H
Control
Genetically modified
Brain vascular fibroblasts
Brain pericytes
Survival
OS Time (days)


I
Fibroblasts^Control
GFP
LMN-A/C


J
Fibroblasts^Control
Fibroblasts^Control
Fibroblasts^Control


K
Pericytes^PDGFB/CDKN2A
Pericytes^Control
Merged
GFP
LMN A/C
2000 µm
100 µm

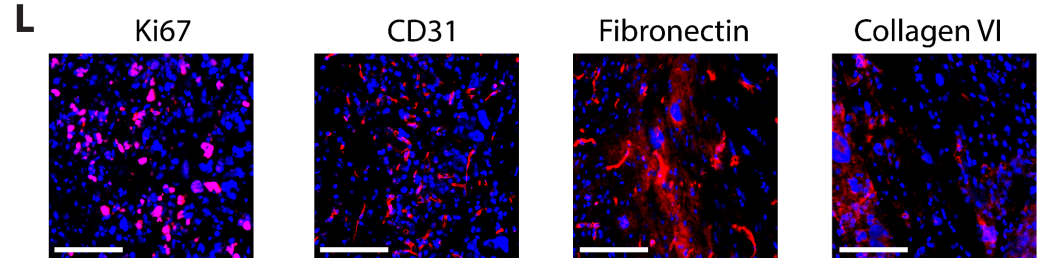

L
Ki67
CD31
Fibronectin
Collagen VI

Fig. S5 *In vivo* initiation of tumors from perivascular cells

(A) Heatmap representing the hierarchical clustering of the transcriptional profiles of mouse glioblastoma induced by knocking-out Pten/Tp53/Nf1 in NG2 expressing cells or in Nestin (Nst) expressing cells of the mouse brain. Each column is one glioblastoma tumor, rows are differentially expressed genes contributing to clustering. Selected marker gene names are indicated on right. Note that two of seven mouse tumors arising from NG2 expressing cells cluster together and display a unique perivascular signature (left). Rgl-type and PeriV-type tumors were highlighted in the red frames.

(B, D) Bar chart showing the relative expression as determined by qPCR quantification of mRNA of indicated genes in control or genetically altered human brain pericytes (B), or genetic modified human brain fibroblasts (D). Y axis is log2 transformed relative expression, gene names are listed at the x-axis. Color coding is indicated for control and genetically modified cells. Three independent measurements.

(C) Immunohistochemical analysis of expression of indicated proteins in cultured pericytes before or after genetic changes. Control cells and two independent experiments of genetic modification is shown. Genetically modified cells highly expressed PDGFB (right) but low levels of p16(INK4a) and p14(ARF) (left). Three independent measurements.

(E) *In vitro* proliferation of brain pericytes with/without carrying genetic alterations of patient-derived glioblastoma (green). means ± SD, 3 independent measurements. **, $p<0.05$; **, $p<0.01$.

(F) Probabilities of estimated CNV change of control and genetic modified human brain fibroblasts. Black indicates high probability and grey indicates low probability. The CNV events were listed at the right side. Color at the top indicates the estimated normal/tumor status.

(G) Enrichment Score of cycling event (G2-M phase) of all cells, dashed line indicates the cut off of 0. X axis represents cells, and Y axis represents the enrichment score of cycling event. Each dot represents one cell, color indicates the status of cells: Ctrl, control fibroblast; GM, genetic modified; GM-Aneuploidy, genetic modified with aneuploidy feature estimated by CNV.

(H) Tumor-related survival curve of mice xenografted with genetically modified pericytes or fibroblasts, as indicated. X represents mice that died of or was complicated by lymphoma and removed. Median survival in pericytes (142: 188 days; log-rank p value: 0.0012) a nd in fibroblasts (245: 282 days; log-rank p value: 0.0402), of genetically modified vs. control (days).

(I) Representative fluorescence image of coronal section from mouse xenograft of control fibroblasts. Tumor region or comparable region in control section is outlined and magnified. Green, GFP; red, human LAM A/C; blue, DAPI.

(J) Representative haematoxylin and eosin staining of coronal section from mouse xenograft of control fibroblasts. Two regions outlined and magnified. Scale bar: 1000 μm, whole section; 100 μm, magnified figures.

(K) Immunohistochemical analysis of GFP that was forced to be expressed in both control and genetically modified cells and of the human LAM A/C antigen in the xenografted mouse brains. Images are from normal pericyte (control) or genetically modified pericyte (genetic modified) xenografted mice. Grey box enlarged at right. Green, GFP; red, human LAM A/C; blue, DAPI. Scale bar: 1000 μm for the whole section, 100 μm for magnified images.

(L) From left to right, representative fluorescence staining (red color) of Ki67, CD31, Fibronectin, Collagen VI in the mouse xenograft of genetically modified fibroblasts. Blue, DAPI. Scale bar 100 μm.