Supplemental Information

Adult Human Glioblastomas Harbor Radial Glia-like Cells

Rong Wang, Roshan Sharma, Xiaojuan Shen, Ashley M. Laughney, Kosuke Funato, Philip J. Clark, Monika Shpokayte, Peter Morgenstern, Monalisa Navare, Yichi Xu, Shaghayegh Harbi, Ignas Masilionis, Gouri Nanjungud, Yanhong Yang, Gabriel Duran-Rehbein, Martin Hemberg, Dana Pe'er, and Viviane Tabar
Supplemental Methods:

Cell Division Analyses
For real time imaging assays, primary cultured CD133+ cells or tumor explants were transferred to an on-stage incubator of Vivaview (Olympus) or Zen-2 (Zeiss), and imaged using a X10 or X20 objective at 15-20 min intervals for up to 2-3 days. Maximum intensity projections of the collected stacks (50 µm at 2 µm step size) were compiled and generated into movies. The movies were analyzed with ImageJ or Imaris software. 20-30 imaged positions of *in vitro* cultures were examined. All divisions that occurred between the start of imaging and 48 hours later were analyzed. MST length was measured as the distance between the site of cytokinesis and the center of the soma right after the S phase determined by the morphology of cell body changing to round. Division angles were determined by calculating the angle between cytokinesis and the primary process (the process along which the cell body translocated). 10 imaged positions from 6 human brain tumor samples were used for analysis of division. Upon completion of time-lapse imaging, cells were fixed in 4% paraformaldehyde immediately overnight at 4ºC and then washed twice with PBS before staining with the antibodies of interest.

Determination of cleavage plane orientation
Cleavage plane orientation of mitotic cells in culture was deduced from the position of the DAPI-stained sister chromatids in anaphase or early telophase and was expressed relative to the basal-like process, with a cleavage plane to the process being defined vertical, or horizontal or oblique.

Protein Synthesis Analyses
Cells are cultured in 8 well cover slide or coverglass chamber with growth factor free N2 medium for 2 days followed by 25 ng/ml final concentration of human interleukin 1b (R&D system) stimulation for the indicated time. The protein synthesis analyses using OP-Puro are carried out using protein synthesis assay kit (Cayman) according to the manufacturer’s instruction.

Immunofluorescence staining and Microscopic analysis
Primary antibodies were rabbit anti-pH3 (ser10) (1:500, Millipore, 06-570) goat anti-FABP7 (1:500, Abcam, ab110099); rabbit anti-GLAST(1:500 LSBio, LS-A9447); rabbit anti-ARHGAP11b (1:500 Sigma, HPA064479); rabbit anti-cFOS (1:400, Santa Cruz, sc-52); mouse anti-Cy3-VIM (1:500, Sigma, C9080); rabbit anti-IL1R1 (1:500, Abcam, ab106278); mouse anti-human CRYAB (1:400, Abcam, ab13496); mouse anti-human pVIM (1:500, MBL, D076-3); rabbit anti-HOPX (1:400, Sigma, HPA030180); rabbit anti-HOPX (1:400, Santa Cruz, sc30216); mouse anti-human NESTIN (1:400, Millipore, MAB5326); mouse anti human nuclear antigen (hNA) (1:500, Molecular Probes, T-30955); mouse anti-human Ki67 (1:500, Dako Denmark, M7240); goat anti-human SOX2 (1:600; Santa Cruz, sc-17320); rabbit anti-human GFAPδ (1:1000, Chemicon, AB9598); rabbit anti-human LIFR (1:300, antibodies-online, ABIN2431519), rabbit anti-human ITGA5 (1:200, Millipore, AB1928); mouse anti-human PAX6 (1:600, Santa Cruz, sc-81649); rabbit anti-PAX6 (1:500, Biolegend, 901301). The following secondary antibodies were used: Alexa Fluor 488 donkey anti-mouse or rabbit (1:1000), Alexa Fluor 568 donkey anti-mouse or rabbit (1:1000), Alexa Fluor 647 donkey anti-goat IgG (1:1000), Alexa Fluor 647 donkey anti-rat IgG (1:1000), all from Molecular probes (Invitrogen).

For immunostaining cells, fixed cells or explant cultures or tissue sections were permeabilized and blocked in PBS buffer containing 0.3% (for explant) or 0.1% (for cell) Triton-X100 and 4% FBS for 1h at RT. Primary antibodies were diluted in blocking PBS buffer containing 0.1% Triton-X100 and 4% FBS, then incubated with cultures overnight at 4 ºC. Cultures incubated with primary antibodies were washed 3 times with blocking PBS buffer, 6 minutes per wash, followed by incubation in secondary antibodies diluted in blocking buffer with 1:1000 for 1h at RT.

Stained sections or whole explants were examined with confocal laser scanning SP8 microscopy (Leica Microsystems, Bannockburn, IL; Carl Zeiss MicroImaging GmbH, Germany). The data was analyzed with Fiji, Velocity (Mountain View, CA) or LSM5 (Carl Zeiss MicroImaging GmbH, Germany) software and quantified in pixels using MetaMorph (Molecular Devices, Downingtown, PA) or Imaris image analysis software. 3D images were constructed with Imaris software.
RNA-FISH

Monolayer cultures are fixed in 4% PFA for 10 min at room temperature on slides and stored in PBS till further process. Cryosections were baked at 60 °C for 30 min - 1 hr and fixed for 15 min at 4°C in 4% paraformaldehyde in 1X PBS to decrease sample detachment, followed by a PBS wash and dehydration in an ethanol series (50% ethanol 5min-70% ethanol 5min- 100% ethanol 5min -100% ethanol 5min-ari dray 5 min, all at RT) before being loaded on the instrument for staining. Slides were stained on Leica Bond RX using RNAscope LS Multiplex Fluorescent Reagent Kit (ACD Diagnostics cat#3228000). The cell culture pre-treatment consisted of a 5-minute incubation with Protease III (proprietary enzyme formulation) at room temperature. Tissue pre-treatment was ER2 buffer for 15min at 95C followed by 15 min of Protease III at 40C. The C1 probe was not diluted as it was ready to use. The 50X C2 probe was added to C1 probe at 1:50 dilution. The TSA Fluor Alexa 488 and Alexa Fluor 594 reagents were prepared and used according to manufacturer instructions, at 1:200 dilution (Thermo Fisher Life Technologies). Slides were counterstained with DAPI and mounted with Mowiol mounting media. *EGFR* C1 and *CRYAB* C2 probes (ACD Diagnostics) were used for double RNA FISH. The hybridization was performed for 3 hours at 40C using ACD Diagnostics hybridization buffer.

DNA-FISH

FISH analysis was performed on frozen sections in three GBM patients to validate copy number variation (CNV) detected by sc-RNA seq. Based on CNV result, a custom probe-set was designed for each patient: a four-color probe to detect chromosome 10, 15, 11 and *EGFR* (GBM27); a three-color probe to detect chromosome 10, 9 and *EGFR* (GBM28); and a three-color probe to detect chromosome 10, 11 and *EGFR* (GBM29). The bacterial artificial chromosome (BAC) or plasmid clones used in the probe mix were as follows: *EGFR* (RP5-1019E12 and RP11-339F13; labeled with Cy5- or Orange-dUTP), chromosome 9 (RP11-747P3 and RP11-574G7; labeled with Green-dUTP), chromosome 10 (RP11-54P13 and RP11-718J13; labeled with Red-dUTP), chromosome 11 (RP11-300I16 and RP11-804L21; labeled with Green-dUTP), and chromosome 15 (RP11-14709 and RP11-342L10; labeled with orange-
dUTP). Probe labeling, tissue processing, hybridization, post-hybridization washing and fluorescence detection were performed according to standard laboratory procedures established at the Molecular Cytogenetics Core Facility. Slides were scanned using a Zeiss Axioplan 2i epifluorescence microscope equipped with a megapixel charge-coupled device camera (CV-M4+CL, JAI) controlled by Isis 5.5.9 imaging software (MetaSystems Group Inc, Waltham, MA). The entire section was first scanned under 63× objective, possible intratumoral heterogeneity assessed, and representative regions imaged through the depth of the tissue (compressed or merged stack of 12 z-section images were taken at 0.5 µ intervals). For each patient, a minimum of 50 discrete nuclei were scored. In all three patients, the presence of EGFR amplification or copy number gain assisted in marking/confirming tumor cell origin. Amplification was defined as ≥6 copies or presence of at least one small signal-cluster (≥4 signals resulting from tandem duplications). Clusters of high-level amplification (≥20 copies) cannot be accurately counted and therefore given a score of 20. Loss was defined as <2 copies or lower copy number relative to control. Chromosome 9 and 11 served as control in respective probe-mix.

**ImmunofISH**

Immuno-FISH (pVIM-EGFR) was performed on frozen section in four GBM patients to confirm tumor cell origin of RG-like cells. Briefly, fixed frozen sections were permeabilized with 0.3% Triton X100/2%FBS/1XPBS for 1 hour at RT and incubated with mouse anti-human pVim (1:100, MBL D076-3) overnight at 4°C. The following day, sections were washed with 1xPBS, fixed in 4% PFA for 20 minutes at RT, washed in 1XPBS, co-denatured at 80°C for 4 minutes with 5-10uL of EGFR (Red) DNA-FISH probe, and hybridized for 48-72 hours at 37°C. Following hybridization, sections were washed with wash buffer (0.3% Tween 20 in 0.4 XSSC), incubated with goat anti-mouse Alexa 647 (1:100, Invitrogen) for 1 hour at 37°C, washed with wash buffer, rinsed with water and stained with DAPI. Slides were scanned, imaged and analyzed as described in DNA-FISH section.

**Single cell-based cDNA and library preparation**

Sorted tumor cells from human GBM specimens were washed once and resuspended in PBS containing 0.05% BSA. 21ul of a cellular suspension at 500 cells/ul with 95% viability, were
loaded onto the 10X Genomics Chromium platform to generate barcoded single-cell Gel Bead-In-Emulsions (GEMs). Single-cell RNA sequencing (scRNA-seq) libraries were prepared according to 10X Genomics specifications (Single Cell 3’ V2 Reagent Kits User Guide PN-120233, 10x Genomics, Pleasanton, CA, USA). GEM-Reverse Transcription (RT) (55 °C for 2 h, 85 °C for 5 min; held at 4 °C) was performed in a C1000 Touch Thermal cycler with 96-Deep Well Reaction Module (Bio-Rad, Hercules). After RT, GEMs were broken, and the single-strand cDNA was cleaned up with DynaBeads MyOne Silane Beads (Thermo Fisher Scientific, Waltham, MA) and SPRIselct Reagent Kit (0.6 × SPRI; Beckman Coulter). cDNA was amplified for 14 cycles using the C1000 Touch Thermal cycler with 96-Deep Well Reaction Module (98 °C for 3 min; 98 °C for 15 s, 67 °C for 20 s, and 72 °C for 1 min x 14 cycles; 72 °C for 1 min; held at 4 °C). cDNA quality was analyzed using an Agilent Bioanalyzer 2100 (Santa Clara, CA). The resulting cDNA was sheared to ~200bp using a Covaris S220 instrument (Covaris, Woburn, MA) and cleaned using 0.6 x SPRI beads. The products were end-repaired, ‘A’-tailed and ligated to adaptors provided in the kit. A unique sample index for each library was introduced through 10 cycles of PCR amplification using the recommendations provided by in the kit (98 °C for 45 s; 98 °C for 20 s, 60 °C for 30 s, and 72 °C for 20 s x 14 cycles; 72 °C for 1 min; held at 4 °C). After two SPRI cleanups, libraries were quantified using Qubit fluorometric quantification (Thermo Fisher Scientific, Waltham, MA) and the quality assessed on an Agilent Bioanalyzer 2100. Four libraries were pooled and clustered on a HiSeq2500 rapid mode at 10pM on a pair end read flow cell and sequenced for 98 cycles R1, followed by 14bp I7 Index (10X Barcode), 8bp I5 Index (sample Index) and 10bp on R2 (UMI).

**Single cell RNA sequencing data processing**

The Sequence Quality Control (SEQC) (Azizi et al., 2018) package was used to process raw scRNA-seq reads to a transcript count matrix, including de-multiplexing, alignment, barcode and UMI error correction, and generation of a digital expression matrix. Alignment was performed to the hg38 annotation restricted to transcribed, polyadenylated RNA of length > 200 nucleotides to increase mapping specificity. Viable single cells were identified based on total mRNA abundance, number of unique transcripts, and mitochondria fraction. Briefly, viable cells were distinguished from captured ambient mRNA based on library size. In practice, this was achieved
by constructing an empirical cumulative density function of cell sizes and finding the minimum of its second derivative. Additionally, cells with low complexity libraries were filtered by regressing the number of genes detected per cell against the number of molecules contained in that cell. Genes detected in less than ten cells or genes with low expression levels, identified as those with count values < 6 standard deviations from the second mode on the distribution of counts/gene were also excluded from downstream analysis (retained 16,377 unique genes). After filtering, a total of 12,367 cells from 3 patients (GBM 27: 4570, GBM 28: 4341, and GBM 29: 3456 cells) with a median library size of 6,152 molecules per cell were kept for downstream analysis. The filtered count matrix was then normalized by dividing the expression level of each gene in a cell by the cell’s total library size and then scaling by the median library size of all cells.

While single cell technologies offer an unprecedented opportunity to obtain the abundance of all expressed mRNA in each single cell, the resulting data is heavily impaired by noise. The most noticeable of all is the low sampling of mRNA transcripts, which results in missing transcripts for majority of the genes in majority of the cells, also known as “drop-outs”. Several computational tools have been proposed to get around this problem, and while many of them make parametric assumptions on the distribution of gene expression, MAGIC (Markov Affinity Graph Based Imputation of Cells) is a non-parametric graph-based non-linear imputation and denoising method that has been used to robustly denoise and impute scRNA-seq data in various settings (Bakhoum et al., 2018; Takahashi et al., 2019; Vuong et al., 2018). Briefly, MAGIC begins with the construction of a $k$-nearest neighbor graph of the cells using Euclidean distance. The distance graph is then symmetrized $D = D + D^T$, where $D$ denotes the distance graph and $D^T$ is the transpose of $D$. The distance graph is then converted into an affinity matrix ($A$). If $D(i,j)$ represents the distance between cells $i$ and $j$ then the affinity between them is given by $e^{(-\frac{D(i,j)}{\sigma^2})}$ where $\sigma$ is the width of the kernel. We set $\sigma$ to be distance to the $k/3^{th}$ neighbor for each cell to ensure that it adapts to data density. The affinity matrix is then normalized into a Markov matrix ($M$). Let $Z$ be the diagonal matrix whose non-zero entries are the sum of rows of $A$ that is $Z(j,j) = \sum_k A(j,k)$. Then $M = Z^{-1}A$. The resulting Markov matrix mimics a random walk along the graph of the data. The $(i,j)^{th}$ entry of $M$ provides the probability that a random walker at the $i^{th}$ state will jump to the $j^{th}$ state in one step. MAGIC then uses the resulting Markov matrix to run a diffusion process along the graph by raising $M$ to a power $t$ and right
multiplying the normalized data to obtain an imputed and denoised data; that is, 
\[ \text{Imputed data} = M^t \times \text{Data}. \]
We applied MAGIC based imputation to the median-normalized count matrix to de-noise and recover missing gene values using conservative parameters \((t = 3, k = 21)\). Imputation was performed using the first 40 principal components of the normalized count matrix (accounting for > 95% variance in data).

**Single cell visualization and clustering**

All cells were visualized using the Barnes-hut approximate version of t-Distributed Stochastic Neighbor Embedding (t-SNE) (Amir el et al., 2013) computed on the first 20 principal components of the imputed count matrix. To obtain a gross-level understanding of the various subpopulations present in the data, we used PhenoGraph based clustering (van Dijk et al., 2017). PhenoGraph is a graph based unsupervised clustering algorithm that has been shown to successfully and robustly identify biologically sound cellular subpopulations from single-cell data in various settings (Jaitin et al., 2016; Lavin et al., 2017). We used PhenoGraph to compute clusters directly on the imputed count matrix (using parameter \(k = 100\)) and 16 phenotypic cell types were identified. The CD45+ clusters showing up-regulation of immune cell markers were subsequently excluded from further analysis of tumor cells only.

**Single cell data RNASeq analysis**

Data from individual patients were analyzed as follows: We first denoised and imputed each sample using MAGIC (40 PCA components, \(k = 30, t = 3\)) as previously described. This was followed by PhenoGraph clustering (\(k = 50\)) on the imputed data. Subsequently we used EMD analysis to annotate each cluster and identify differentially expressed genes in each cluster. In order to visualize the underlying structure of the data, we considered the diffusion components for each patient separately. Diffusion maps (Coifman and Lafon, 2006) are graph based non-linear dimensionality reduction technique. They map high dimensional single cell data to a much lower dimensional Euclidean space while maintaining the neighborhood structure in the vicinity of each cell. This approach has been successfully utilized to analyze single-cell data and primarily to extract continuous structure from data (Farrell et al., 2018; Haghverdi et al., 2016). The diffusion maps or diffusion components are defined as the eigenvectors of the Markov
transition matrix, which we have defined above in the discussion of MAGIC. The resulting eigenvectors comprise the new low-dimensional coordinates of each cell.

**Identification of differentially expressed gene sets among tumor clusters**

To identify differentially expressed genes per tumor cluster, we used Earth Mover’s Distance (EMD) to compute distances between the distribution of expression of a given gene in each cluster against the rest of the data \((45)(44)\). EMD is particularly appealing for single cell data because unlike traditional measures such as cluster mean/median that collapse single cell level information into a single number, EMD retains gene information at single cell level in the form of distributions. Thus, it allows us to harness the entire dynamic range and distribution of gene expression in a cluster and compare it against the expression of the same gene in rest of the data. Notably, EMD has been previously used successfully in single cell data analysis to obtain differentially expressed genes.

Mathematically, in one dimension EMD can be computed as the L1-norm of the difference in the cumulative distribution function of the two densities. If \(g_1\) is the probability density of gene \(g\) in cluster A and \(g_2\) is the probability density of gene \(g\) in rest of the data then \(\text{EMD}(g) = |\text{cdf}(g_1) - \text{cdf}(g_2)|\). To approximate the probability density, we use histograms with 30 bins and the edges of histogram are set to be the minimum and maximum expression level of the gene in the entire data. In this way, we provide an EMD score to each gene, then we identify genes that have EMD score more than 2.5 standard deviations from the mean (or less than mean minus 2.5 standard deviation) across all the genes as differentially expressed. This ensures that the selected genes are statistically deviant from average.

In addition to identifying differentially expressed genes, we were interested in assessing the expression level (high or low) of genes belonging to certain genesets in each cluster. For this, given a geneset, we computed EMD for each gene in the geneset and obtained the mean EMD score across all genes for that geneset. We identified RG-like clusters of each individual patient based on EMD score of oRG specific gene sets (\(\geq\)average of oRG EMD score of all clusters +standard deviation of EMD score of all clusters /2). Cycling clusters for GBM28 and 29 are defined by EMD score of either of 4 types of cycling gene sets (\(\geq\)average of cycling EMD score
of all clusters +standard deviation of cycling EMD score of all clusters /2). In GBM27, the cycling EMD scores (earlyG1, G1/S) of non-RG-like clusters are significantly higher than those of RG-like clusters, resulting in only 1 cycling RG-like cluster when using the calculation mentioned above. Thus, Cycling clusters for GBM27 are defined by EMD score of either of 4 types of cycling gene sets (>=average of cycling EMD score of RG-like clusters +standard deviation of cycling EMD score of RG-like clusters /2). 1.1, 1.2, 1, and 0.95 of oRG EMD score are separately used as the cutoff value to define RG-like clusters in combined data, GBM27, GBM28, and GBM29. Cycling gene sets and all other gene sets used in this study are from both KEGG pathways and gene ontology (Dessimoz and Skunca, 2017; Kanehisa and Goto, 2000).