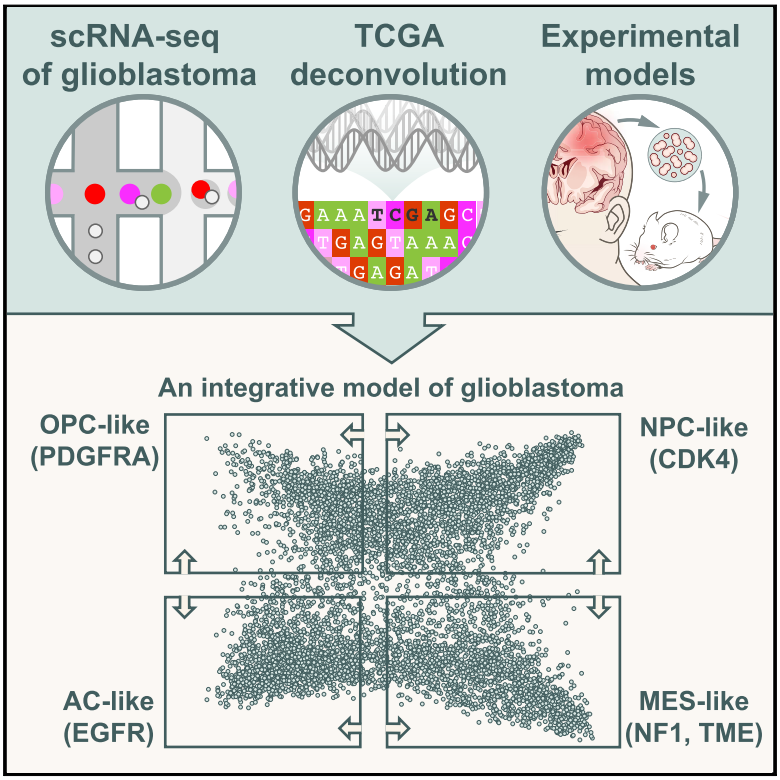
Article

Cell

An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma

Graphical Abstract



Highlights

- Four cellular states drive glioblastoma malignant cells heterogeneity
- *In vivo* single-cell lineage tracing supports plasticity between these four states
- Genetics and the microenvironment influence the frequency of cells in each state
- TCGA subtypes reflect the highest-frequency malignant states and the microenvironment

Authors

Cyril Neftel, Julie Laffy, Mariella G. Filbin, ..., Bradley E. Bernstein, Itay Tirosh, Mario L. Suvà

Correspondence

Suva.Mario@mgh.harvard.edu (M.L.S.), Itayt@weizmann.ac.il (I.T.)

In Brief

Single-cell analyses of glioblastoma samples reveal multiple cellular states, their plasticity and the genetic underpinnings of state proportions in a given tumor.

Neftel et al., 2019, Cell *178*, 1–15
August 8, 2019
https://doi.org/10.1016/j.cell.2019.06.024

CellPress

Please cite this article in press as: Neftel et al., An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma, Cell (2019), https://doi.org/10.1016/j.cell.2019.06.024

Article

Cell

An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma

Cyril Neftel,[1,2,3,4,19] Julie Laffy,[5,19] Mariella G. Filbin,[1,2,3,6,19] Toshiro Hara,[1,2,3,7,19] Marni E. Shore,[1,2,3] Gilbert J. Rahme,[1,2,3] Alyssa R. Richman,[1,2,3] Dana Silverbush,[1,2,3] McKenzie L. Shaw,[1,2,3,6] Christine M. Hebert,[1,2,3] John Dewitt,[1,2,3] Simon Gritsch,[1,2,3] Elizabeth M. Perez,[1,2,3] L. Nicolas Gonzalez Castro,[1,2,3,8] Xiaoyang Lan,[1,2,3] Nicholas Druck,[1] Christopher Rodman,[1] Danielle Dionne,[2,3] Alexander Kaplan,[8] Mia S. Bertalan,[8] Julia Small,[9] Kristine Pelton,[10] Sarah Becker,[10] Dennis Bonal,[11] Quang-De Nguyen,[11] Rachel L. Servis,[1] Jeremy M. Fung,[1] Ravindra Mylvaganam,[1] Lisa Mayr,[11] Johannes Gojo,[12] Christine Haberler,[13] Rene Geyeregger,[14] Thomas Czech,[15] Irene Slavc,[12] Brian V. Nahed,[9] William T. Curry,[9] Bob S. Carter,[9] Hiroaki Wakimoto,[9] Priscilla K. Brastianos,[8] Tracy T. Batchelor,[8] Anat Stemmer-Rachamimov,[1] Maria Martinez-Lage,[1] Matthew P. Frosch,[1] Ivan Stamenkovic,[4] Nicolo Riggi,[4] Esther Rheinbay,[1,3] Michelle Monje,[16] Orit Rozenblatt-Rosen,[2,3] Daniel P. Cahill,[9] Anoop P. Patel,[17] Tony Hunter,[7] Inder M. Verma,[7] Keith L. Ligon,[3,10] David N. Louis,[1] Aviv Regev,[2,3,18] Bradley E. Bernstein,[1,2,3] Itay Tirosh,[5,20,*] and Mario L. Suvà[1,2,3,20,21,*]

[1]Department of Pathology and Center for Cancer Research, Massachusetts General Hospital and Harvard Medical School, Boston, MA, 02114, USA
[2]Klarman Cell Observatory, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA
[3]Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA
[4]Institute of Pathology, Faculty of Biology and Medicine, Centre Hospitalier Universitaire Vaudois, Lausanne, 1011, Switzerland
[5]Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot 761001, Israel
[6]Department of Pediatric Oncology, Dana-Farber Boston Children's Cancer and Blood Disorders Center, Boston, MA 02215, USA
[7]Laboratory of Genetics, Salk Institute for Biological Studies, La Jolla, CA 92037, USA
[8]Departments of Neurology and Radiation Oncology, Division of Hematology/Oncology, Massachusetts General Hospital Cancer Center, Harvard Medical School, Boston, MA, 02114, USA
[9]Department of Neurosurgery, Massachusetts General Hospital and Harvard Medical School, Boston, MA, 02114 USA
[10]Department of Oncologic Pathology, Brigham and Women's Hospital, Boston Children's Hospital, Dana-Farber Cancer Institute, Boston, MA 02215, USA
[11]Center for Biomedical Imaging in Oncology, Lurie Family Imaging Center, Dana-Farber Cancer Institute, Boston, MA 02215, USA
[12]Department of Pediatrics and Adolescent Medicine, Medical University of Vienna, Vienna, 1090, Austria
[13]Institute of Neurology, Medical University of Vienna, Vienna, 1090, Austria
[14]Children's Cancer Research Institute (CCRI), St. Anna Kinderspital, Medical University of Vienna, Vienna, 1090, Austria
[15]Department of Neurosurgery, Medical University of Vienna, Vienna, 1090, Austria
[16]Departments of Neurology, Neurosurgery, Pediatrics and Pathology, Stanford University School of Medicine, Stanford, CA, 94305, USA
[17]Department of Neurological Surgery, University of Washington, Seattle, USA
[18]Howard Hughes Medical Institute, Koch Institute for Integrative Cancer Research, Department of Biology, MIT, Cambridge, MA 02139, USA
[19]These authors contributed equally
[20]These authors contributed equally
[21]Lead Contact
*Correspondence: Suva.Mario@mgh.harvard.edu (M.L.S.), Itayt@weizmann.ac.il (I.T.)
https://doi.org/10.1016/j.cell.2019.06.024

SUMMARY

Diverse genetic, epigenetic, and developmental programs drive glioblastoma, an incurable and poorly understood tumor, but their precise characterization remains challenging. Here, we use an integrative approach spanning single-cell RNA-sequencing of 28 tumors, bulk genetic and expression analysis of 401 specimens from the The Cancer Genome Atlas (TCGA), functional approaches, and single-cell lineage tracing to derive a unified model of cellular states and genetic diversity in glioblastoma. We find that malignant cells in glioblastoma exist in four main cellular states that recapitulate distinct neural cell types, are influenced by the tumor microenvironment, and exhibit plasticity. The relative frequency of cells in each state varies between glioblastoma samples and is influenced by copy number amplifications of the *CDK4*, *EGFR*, and *PDGFRA* loci and by mutations in the *NF1* locus, which each favor a defined state. Our work provides a blueprint for glioblastoma, integrating the malignant cell programs, their plasticity, and their modulation by genetic drivers.

INTRODUCTION

Glioblastoma (isocitrate dehydrogenase [IDH]-wild-type) is an incurable malignancy, and the main challenges underlying therapeutic failure are rooted in its heterogeneity (Louis et al., 2016). Genetic, epigenetic, and microenvironmental cues influence cellular programs and drive glioblastoma heterogeneity.

One layer of heterogeneity is reflected by previously described transcriptional subtypes. Studies of inter-tumor heterogeneity based on bulk expression profiles suggest that at least three subtypes of glioblastoma exist, namely proneural (TCGA-PN), classical (TCGA-CL), and mesenchymal (TCGA-MES) (Verhaak et al., 2010; Wang et al., 2017). These expression-based subtypes are partially enriched for selected genetic events; for example, *PDGFRA* alterations are more common in TCGA-PN glioblastoma, whereas alterations in *EGFR* are more common in TCGA-CL glioblastoma. These subtypes programs also vary within the same tumor specimen, as multi-region tumor sampling has shown that multiple subtypes can co-exist in different regions of the same tumor, longitudinal analyses demonstrated that subtypes can change over time and through therapy, and single-cell RNA-sequencing (scRNA-seq) indicated that distinct cells in the same tumor recapitulate programs from distinct subtypes (Patel et al., 2014; Sottoriva et al., 2013; Wang et al., 2017).

A second layer of heterogeneity is the developmental state of glioblastoma cells in the tumor. Glioblastoma hijacks mechanisms of neural development and contains subsets of glioblastoma stem cells (GSCs) that are thought to represent its driving force, possess tumor-propagating potential, and exhibit preferential resistance to radiotherapy and chemotherapies (Bao et al., 2006; Chen et al., 2012; Lathia et al., 2015; Parada et al., 2017). Although various markers can enrich for putative GSCs (Lathia et al., 2015), it is unknown whether different GSC markers isolate distinct or similar cellular states and whether tumors generated with different subpopulations of GSCs give rise to glioblastoma of comparable or diverse cellular composition. It also remains challenging to dissect the extent to which unidirectional hierarchies or more reversible state transitions govern glioblastoma and GSC biology (Suvà et al., 2014). Thus, a better understanding of the various sources of heterogeneity—genetic, epigenetic, developmental, and microenvironmental – in glioblastoma is a critical goal with broad implications for therapy.

scRNA-seq has emerged as a key method to comprehensively characterize the cellular states within tissues, both in health and in disease (Tanay and Regev, 2017; Tirosh and Suva, 2019). In gliomas, we have shown that one can infer the cellular architecture of tumors and relate single-cell states to genetics through inference of chromosomal copy number aberrations (CNAs) or detection of mutations in expressed transcripts (Tirosh et al., 2016b). Although these approaches have been successful in deciphering key aspects of the biology of IDH mutant and histone mutant glioma (Filbin et al., 2018; Tirosh et al., 2016b; Venteicher et al., 2017), they have proven more challenging in glioblastoma (Darmanis et al., 2017; Müller et al., 2016; Patel et al., 2014; Wang et al., 2017; Yuan et al., 2018). In particular, the relationship between genetic alterations and the diversity of epigenetic states remains unclear and poses a great challenge for the field.

Here, we used an integrative approach to understand glioblastoma transcriptional and genetic heterogeneity, combining scRNA-seq of 20 adult and 8 pediatric glioblastoma (24,131 cells in total), scRNA-seq and lineage tracing of glioblastoma models, and analysis of 401 bulk specimen from The Cancer Genome Atlas (TCGA). We find that malignant cells in glioblastoma exist in a limited set of cellular states that recapitulate (1) neural-progenitor-like (NPC-like), (2) oligodendrocyte-progenitor-like (OPC-like), (3) astrocyte-like (AC-like), and (4) mesenchymal-like (MES-like) states. Although each glioblastoma sample contains cells in multiple states, the relative frequency of each state varies between tumors. We show that such frequencies are associated with genetic alterations in *CDK4*, *PDGFRA, EGFR*, and *NF1* that each favor a particular state. Furthermore, by coupling scRNA-seq to uniquely barcoded single cells *in vivo*, we demonstrate plasticity between states and the potential for a single cell to generate all four states. Our work provides a roadmap of the cellular programs of malignant cells in glioblastoma and their plasticity and modulation by genetic drivers.

RESULTS

scRNA-Seq Charts Malignant Cells Heterogeneity in Glioblastoma

To comprehensively interrogate both inter-tumoral and intra-tumoral heterogeneity in IDH-wild-type glioblastoma, we profiled using full-length scRNA-seq (SMART-Seq2) fresh tumor samples from 28 patients spanning both adult and pediatric populations (Figures 1A and S1A; Table S1) (Picelli et al., 2014). To focus on malignant cells, we sorted cells by both viability marker and by the pan-immune marker CD45, and we profiled primarily $CD45^{-}$ cells and only to a more limited extent $CD45^{+}$ cells. In total, 7,930 cells passed our stringent quality controls; 5,730 genes were detected per cell on average, highlighting the high quality of our dataset (Figure S1B; STAR Methods). We classified cells into malignant and non-malignant cell types by combining three approaches (Figures 1A, 1B, and S1C; STAR Methods). First, we inferred CNAs on the basis of the average expression of 100 genes in each chromosomal region (Patel et al., 2014; Tirosh et al., 2016b). This analysis identified large-scale amplifications and deletions in most cells, including the glioblastoma hallmarks of chromosome 7 gain and chromosome 10 loss, which were found in most adult but not pediatric tumors (Figures 1A and S1C). Second, high expression of gene sets corresponding to markers of particular cell types classified some of the cells as macrophages, T cells, and oligodendrocytes (Figure 1B). Third, clustering (Figure 1B; STAR Methods) highlighted three small clusters of non-malignant cells, which lack CNAs and highly express markers of specific cell types. The remaining cells formed a fourth large cluster (6,864 cells) of presumed malignant cells and were associated with CNAs. The three approaches provided concordant classification of glioblastoma cells into malignant and non-malignant subsets. Malignant cells varied considerably between tumors (Figures 1C and S1D), consistent with prior studies that showed that malignant cells differ between patients to greater extent than non-malignant cells (Puram et al., 2017; Tirosh et al., 2016a; Tirosh et al., 2016b).

Malignant Cells Intra-tumoral Heterogeneity Is Dominated by a Few Expression Meta-modules

To comprehensively characterize intra-tumoral heterogeneity among the malignant cells, we identified expression programs that vary between cells within each tumor and then sought the recurrent programs (“meta-modules”) that were identified across different tumors (Filbin et al., 2018; Tirosh et al., 2016a; Tirosh et al., 2016b; Venteicher et al., 2017). First, separately

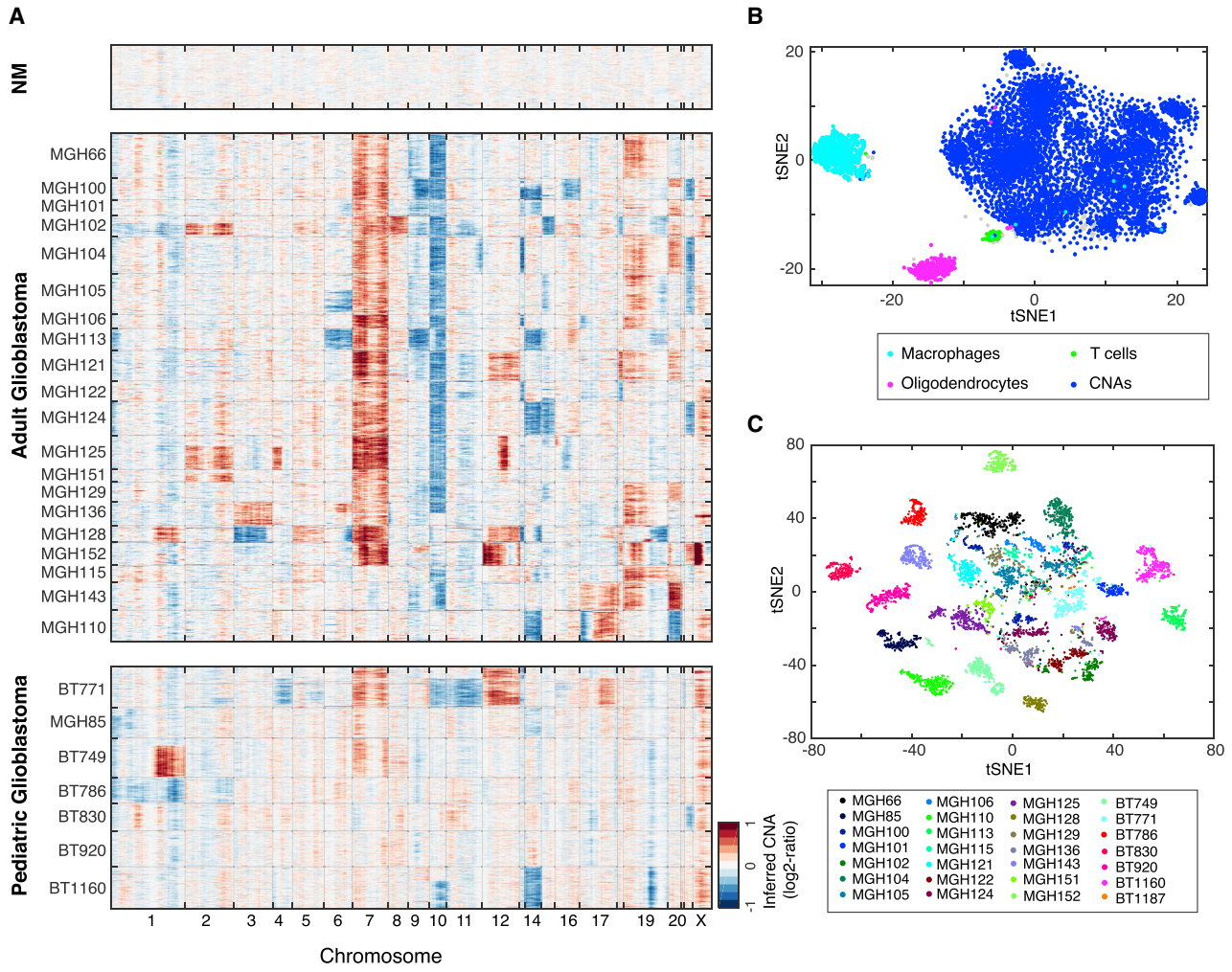


Figure 1. Classification of Single Cells from 28 Glioblastomas

(A) Inference of chromosomal CNAs on the basis of average relative expression in windows of 100 analyzed genes. Rows correspond to cells; non-malignant (NM) cells that lack CNAs are shown at the top, followed by malignant cells (with CNAs, as defined in Figure S1) ordered by tumor and within a tumor clustered by overall CNA patterns.

(B) t-distributed stochastic neighbor embedding (tSNE) plot of all single cells. Cells are colored on the basis of presence of CNAs (blue) or high expression of sets of marker genes for macrophages (cyan), oligodendrocytes (magenta), or T cells (green).

(C) tSNE plot of all malignant cells, colored by tumor.

for each tumor, we hierarchically clustered the cells on the basis of *all* genes expressed at sufficient levels (Figures 2A and S2A). Next, we conservatively retained many clusters for further analysis, including partially overlapping ones (Figure 2A, bottom), and defined for each an expression signature consisting of preferentially expressed genes. Forty-four percent of these signatures were associated almost exclusively with cell cycle genes, indicating that they correspond to the subsets of cycling cells (Figure S2B). All remaining expression signatures were subjected to further analysis in order to elucidate their biological significance. Expression signatures were highly consistent across tumors, such that on average each signature significantly overlapped (false discovery rate [FDR] < 0.01, hypergeometric test) signatures of 9 other tumors. This indicates that despite the global differences between tumors, these patterns of intra-tumoral heterogeneity reflect fundamental processes shared across tumors. We clustered the signatures, resulting in four main groups, of which two further segregated robustly into two sub-groups (Figure 2B; STAR Methods). This enabled us to define six meta-modules consisting of 39–50 genes that highly recur across overlapping signatures from multiple tumors, and each meta-module was derived from at least 6 tumors (Figures 2C and S2C; Table S2). To further demonstrate the robustness of these meta-modules, we generated scRNA-seq profiles by droplet-based scRNA-seq for 16,201 cells from 9 glioblastoma samples (of which 9,870 were inferred to be malignant) (STAR Methods), repeated the analysis, and found expression signatures highly consistent with our meta-modules (Figure S2D).

The top scoring genes and functional enrichments of the meta-modules (Figure 2C and S2E; Tables S2 and S3) highlight

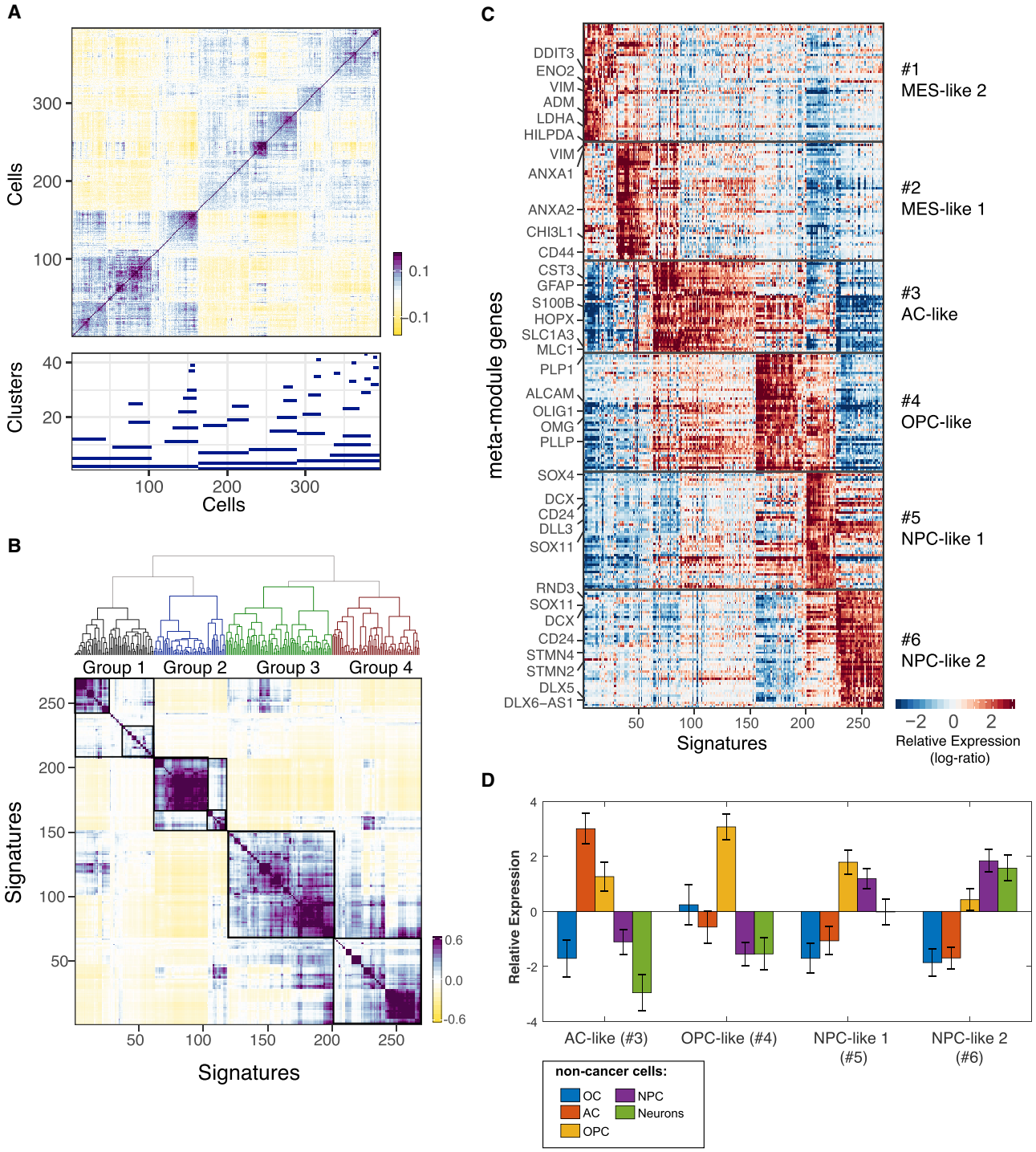


Figure 2. Expression Signatures of Intra-tumoral Heterogeneity among Malignant Cells

(A) Top: cell-to-cell correlation matrix of malignant cells from MGH105, with cells ordered by hierarchical clustering. Shown on the bottom is the assignment of cells to potential overlapping clusters.

(B) Hierarchical clustering of signatures for 269 potential clusters defined from 27 tumors. Groups of potential clusters are highlighted at the top and were used to define meta-modules.

(C) Meta-modules composed of genes consistently upregulated in potential clusters of the same group. Selected genes are indicated (see Table S2 for a full list).

(D) Relative expression of meta-modules across neurodevelopment-related cell types as measured by scRNA-seq (Darmanis et al., 2017; Darmanis et al., 2015; Tirosh et al., 2016b). Error bars correspond to standard error.

two meta-modules that were associated with high expression of mesenchymal-related genes (e.g., *VIM*) and gene sets ($p < 10^{-9}$, hypergeometric test). One of these meta-modules was strongly associated with hypoxia-response genes (e.g., *HILPDA*), stress (e.g., *DDIT3*), and glycolytic (e.g., *ENO2* and *LDHA*) genes, suggesting that in some tumors the mesenchymal state is linked to hypoxia and increased glycolysis. We defined these as mesenchymal-like (MES-like) meta-modules: hypoxia-independent (MES1) and -dependent (MES2) signatures.

The other four meta-modules were associated with neuro-developmental genes, characteristic of neuronal/glial lineages or progenitor cells. These included astrocytic markers in meta-module #3 (*S100B*, *GFAP*, *SLC1A3*, *GLAST*, and *MLC1*), oligodendroglial lineage markers in meta-module #4 (*OLIG1*, *OMG*, *PLP1*, *PLLP*, *TNR*, and *ALCAM*), stem and progenitor cell signatures in meta-modules #5 and #6, including NPC markers (*SOX4*, *SOX11*, and *DCX*) (Tirosh et al., 2016b; Venteicher et al., 2017). Consistently, comparing the meta-modules to neural cell type signatures from scRNA-seq of fetal brains, adult brains, and non-malignant cells from gliomas, meta-modules #3, #4, and #6 were most highly expressed in astrocytes, oligodendrocytic precursor cells (OPCs), and neural progenitor cells (NPCs), respectively (Figures 2D and S2F) (Darmanis et al., 2017; Darmanis et al., 2015; Nowakowski et al., 2017; Tirosh et al., 2016b).

Therefore, the meta-modules mimic developmental cell types but with important distortions from normal programs (Table S4) and were named accordingly as AC-like, OPC-like, and NPC-like. NPC-like was further subdivided into two subprograms (NPC1 and NPC2) (STAR Methods; Table S2) that were distinguished by inclusion of OPC-related genes in NPC1 (e.g., OLIG1 and TNR) versus neuronal lineage genes in NPC2 (e.g., *STMN1*, *STMN2*, *STMN4*, *DLX5-AS1*, and *DLX6-AS1*) (Figure S2E; Table S2), likely reflecting the potential of NPCs to differentiate toward either OPCs or neurons. Each of the meta-modules had additional features beyond the corresponding cell types, which might reflect their distortion compared with normal cell type programs. Thus, although the AC-like meta-module was most highly expressed by astrocytes, it was also expressed in radial glia (RG) and contained RG markers such as *HOPX* (Pollen et al., 2015). Overall, intra-tumoral heterogeneity in glioblastoma largely corresponds to cellular states resembling NPCs, OPCs, astrocytes, and mesenchymal cells. These states were largely consistent between adult and pediatric tumors and were also observed in pediatric samples when analyzed independently (Figure S2G -I).

Cycling Cells and Hybrid Cellular States in Glioblastoma

Next, we classified the cells from all tumors by expression of the meta-modules and cell cycle programs (Figures 3A, 3B, and S3A). Between 3% and 51% of the cells in each tumor were identified as cycling on the basis of the expression of cell cycle signatures (Figure S3B). Cycling cells were enriched in the OPC-like and NPC-like states (Figure 3C), particularly in pediatric tumors (Figure S3C). This is consistent with proliferation of normal OPC and neural precursors, and with our previous observations in IDH mutant and H3K27M mutant glioma, which are driven by proliferating NPC-like and OPC-like cells, respectively (Filbin et al., 2018; Tirosh et al., 2016b; Venteicher et al., 2017). However, in glioblastoma, unlike in other classes of gliomas, the other cellular states—AC-like and MES-like—also contain considerable subsets of proliferating cells, possibly reflecting its very aggressive nature (Figure 3C).

Interestingly, although most glioblastoma cells corresponded primarily to one of the four states, 15% of the cells highly expressed two distinct meta-modules and hence were defined as "hybrid" states (Figures 3A and 3D). Some combinations of meta-modules were rarely observed, whereas others (AC-like/MES-like, NPC-like/OPC-like, and AC-like/OPC-like) were as common as expected by a simple model of independence between expression of the different meta-modules (Figure 3D; STAR Methods). Thus, our data supports a model whereby glioblastoma cells span four main cellular states and their intermediate hybrids, each with proliferative potential, but with higher proliferation of NPC-like and OPC-like states. The meta-modules, hybrids states, and proliferation patterns were confirmed by RNA *in situ* hybridization (RNA-ISH) in ten glioblastoma specimens (Figures 3E, S3E, and S3F). Finally, we developed a "cell-state plot" to summarize the distribution of cells across these states and their intermediates (Figure 3F; STAR Methods), which demonstrates the diversity of proliferating cells and is used below for further analysis.

Limited Relationship between Genetic Subclones and Intra-tumoral State Diversity

Next, we asked whether intra-tumoral cell state diversity could directly reflect genetic subclones within the tumor. Detection of genetic mutations within individual cells from scRNA-seq data is limited by the partial transcriptome coverage of the transcriptome. However, large-scale CNAs, such as full-chromosome or chromosome-arm events, might be robustly detected on the basis of the average upregulation or downregulation of large sets of genes within each chromosomal region, as previously demonstrated (Filbin et al., 2018; Patel et al., 2014; Puram et al., 2017; Tirosh et al., 2016b). Inferred CNAs (STAR Methods) enabled robust detection of 37 genetic subclones in 12 of the tumors, with 2–5 distinct subclones in each tumor (Figures 4A, 4B, and S4A).

Notably, each of the 37 subclones contain cells in multiple cellular states, as defined by the four quadrants of the cell-state plot (Figure 4C and S4B). Thus, cell state is not strictly determined by any of the genetic subclones, although some of the subclones are biased toward specific states. To quantify this bias, we compared the cellular states between pairs of cells from the same and from different subclones. On average, 49% of all pairs of cells from the same tumor had the same state. The overall fraction of same-state pairs was comparable among pairs of cells from the same and from different subclones (51% versus 46%), as only 8 of the 37 subclones had an increased fraction of same-state pairs (Figure S4C) (63% on average).

We also assessed the number of differentially expressed (DE) genes between subclones in the same tumor (Figure S4D). Subclones had a median of 20 DE genes, most of which were not associated or correlated with the meta-modules and rather were often located within the CNA loci that distinguished the subclones. Although we can only detect some genetic events, the limited relation between CNA-defined subclones and expression states corresponding to the meta-modules suggests

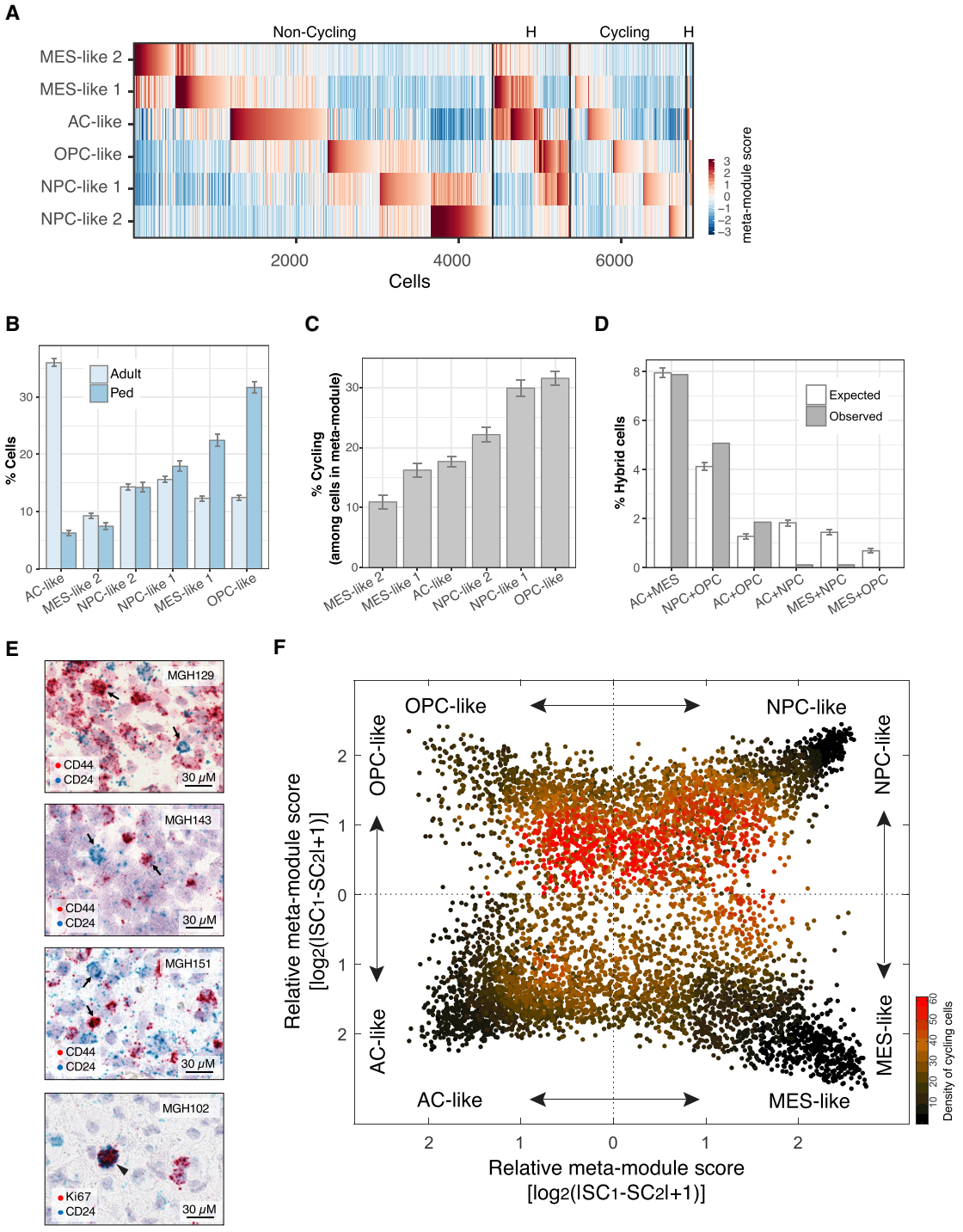


Figure 3. Assignment of Malignant Cells to Cellular States and Their Hybrids

(A) Heatmap showing the meta-module scores of all non-cycling cells (left) and cycling cells (right). Within each group, the cells are ordered by their maximal score, for cells mapping to one meta-module, followed by cells mapping to two meta-modules (hybrid states, denoted as "H").

(B) Bar plot showing the percentage of cells with the highest score for each meta-module. Adult and pediatric tumors are separated in order to demonstrate their distinct distributions. Error bars correspond to standard error, calculated by bootstrapping.

(C) Bar plot showing the percentage of cycling cells among cells with highest score for each of the meta-modules. Error bars correspond to standard error, calculated by bootstrapping.

(D) Bar plot showing the observed and expected percentages of hybrid cells (co-expressing two distinct meta-modules) out of all malignant cells. Expected percentages and their standard errors were calculated by shuffling the cell scores (STAR Methods).

(legend continued on next page)

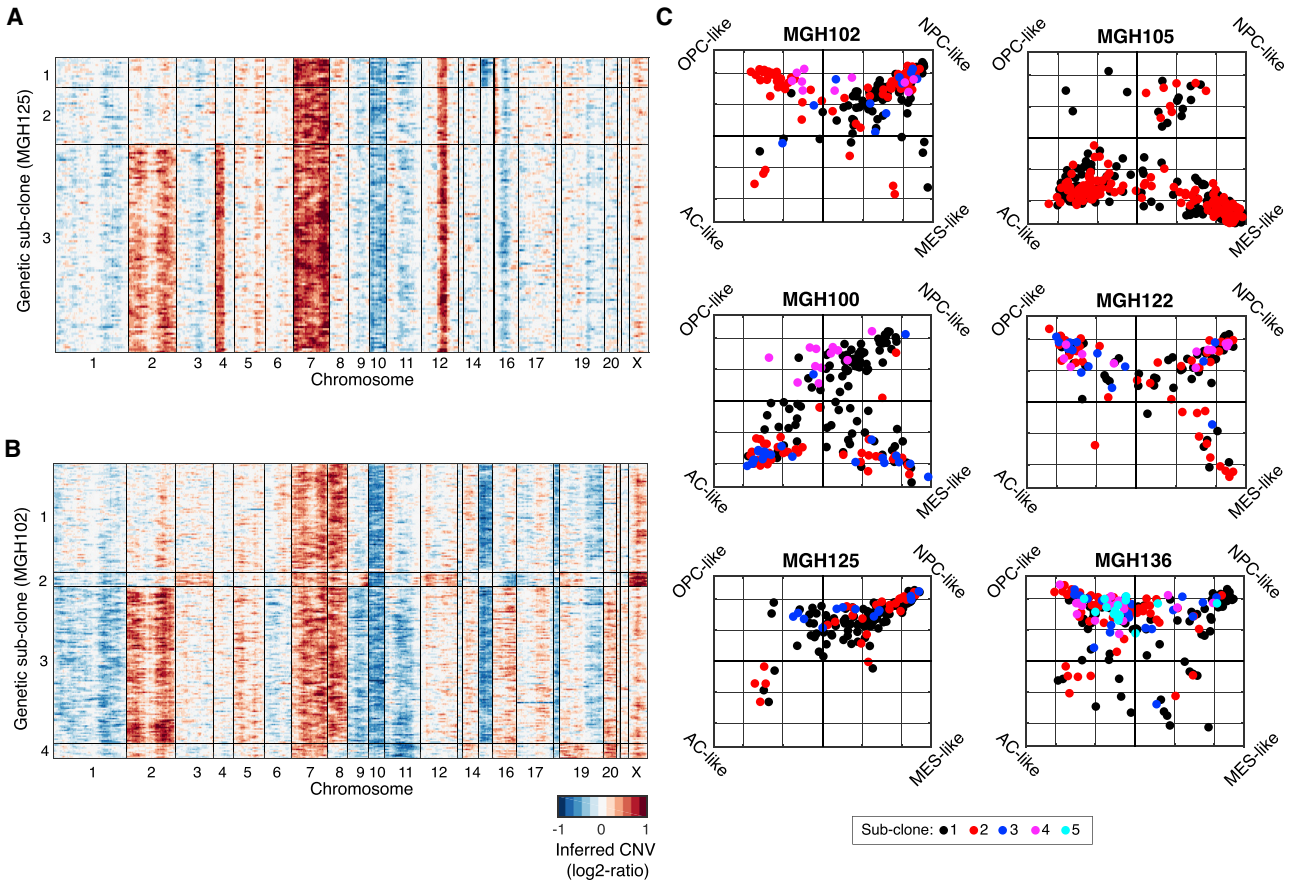


Figure 4. Intra-tumoral Heterogeneity at the Genetic and Expression Levels

(A and B) Identification of genetic subclones by CNAs. Shown are the inferred CNAs of malignant cells in MGH125 (A) and MGH102 (B), separated into genetic subclones on the basis of amplifications or deletions of specific chromosomes (STAR Methods).

(C) Cell-state plots (as in Figure 3F) for six tumors with CNA-based subclones. Cells are colored by their subclone.

that much of the intra-tumoral diversity in expression states is not driven by genetic subclones. This is consistent with our prior observations in IDH mutant and H3K27M mutant gliomas (Filbin et al., 2018; Tirosh et al., 2016b; Venteicher et al., 2017).

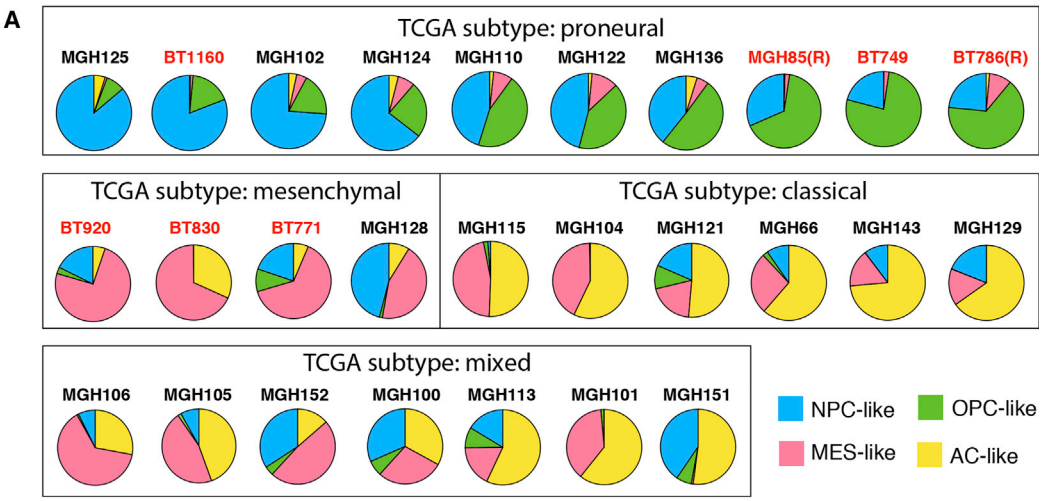
Defined Genetic Drivers Influence the Distribution of Cellular States

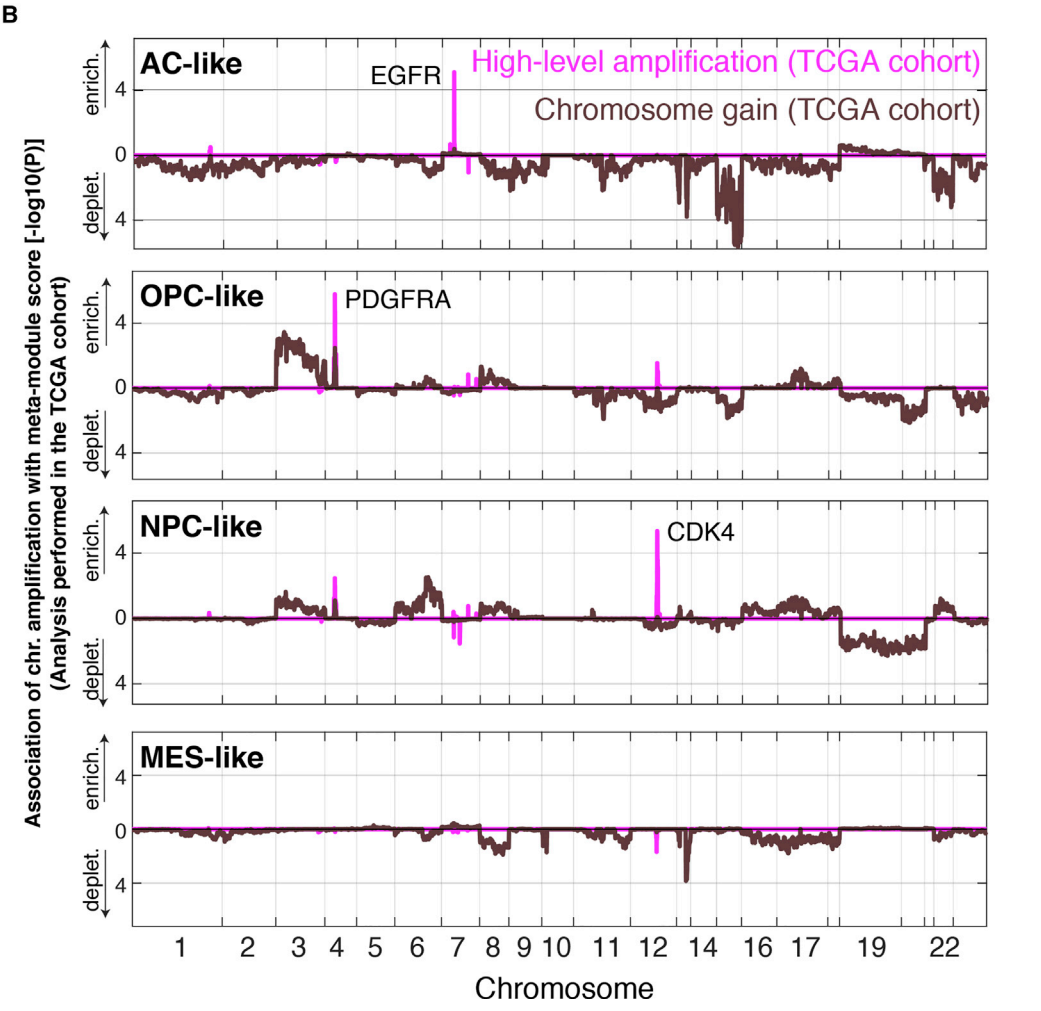
Each of the tumors contained cells in at least two of the four cellular states, with most tumors containing all four states (Figure S5A), but the frequencies of states varied between tumors (Figure 5A) and even to some extent between different regions of the same tumor (Figure S5B). Most tumors consisted primarily of NPC-like plus OPC-like cells, or of AC-like plus MES-like cells, although some tumors had other patterns (Figure 5A). Furthermore, for each of the four, states there were some tumors in which that state was the most frequent. Notably, adult and pediatric glioblastomas appeared to have similar patterns, although AC-like cells were depleted in pediatric versus adult glioblastoma (Figures 5A and 3B).

The preponderance of a particular state (or combination of two states) in each tumor is highly consistent with three bulk subtypes previously defined by TCGA (Figures 5A and S5C). Whereas the TCGA-CL and TCGA-MES subtypes correspond to tumors enriched for the AC-like and MES-like states, respectively, the TCGA-PN subtype corresponds to the combination of two distinct cellular states, OPC-like and NPC-like (Figures 5A and S5C), reflecting the typical co-occurrence of these two states in glioblastoma (Figure 5A), which hinders the ability to distinguish their contributions in bulk RNA-seq. Similarly, the TCGA-MES subtype corresponds to a combination of the MES-like malignant state defined above and a preponderance of microglia and macrophages (Figure S5C), supporting

(E) *In situ* RNA hybridization of glioblastoma for NPC-like (CD24), MES-like (CD44), and proliferation (Ki67) markers. Arrows highlight representative cells positive for CD24 (blue) or CD44 (red). Arrowhead highlights a cell co-expressing CD24 and Ki67.

(F) Two-dimensional representation of cellular states. Each quadrant corresponds to one cellular state, the exact position of malignant cells (dots) reflect their relative scores for the meta-modules, and their colors reflect the density of cycling cells (STAR Methods).





(legend on next page)

their potential role in sustaining the MES-like state of malignant cells. A fourth subtype that was proposed previously (TCGA-Neural) appears to primarily reflect the preponderance of non-malignant oligodendrocytes and neurons (Figure S5C), consistent with recent observations (Wang et al., 2017).

We hypothesized that the fact that each tumor contains multiple cellular states, but that specific states are enriched in subsets of tumors (i.e., tumor subtypes) can be explained by the genetics and/or microenvironment of individual tumors favoring particular cellular states over others. This could be due, for example, to facilitation or inhibition of certain cellular transitions. To identify such effects, we first used the single-cell profiles to search for genes that correlate with high frequency of each state but are not themselves part of the expression program of that state. For example, we searched for differentially expressed genes between the tumors with high frequency of AC-like cells (AC-high tumors) and those with low frequency of AC-like cells (AC-low) within our 28-tumor cohort. To control for differences in the proportions of states between those cohorts, we separately compared cells in each of the four states (Figure S5D). Thus, although AC-high tumors contain primarily AC-like cells they also contain sufficient numbers of MES-like, NPC-like, and OPC-like cells for comparison to the same states in AC-low tumors. This analysis identified 22 genes that were consistently higher in AC-high tumors than in AC-low tumors (Figure S5D), and 16–41 genes that were associated with an abundance of each of the other three states (Figure S5E).

The gene with highest upregulation in AC-high tumors was *EGFR*, which was markedly higher (> 7-fold) in AC-high than in AC-low tumors, among cells from each of the four states (Figure S5E). These results suggest that tumors with *EGFR* aberrations, and therefore high levels of *EGFR* across all cellular states, might favor a high frequency of AC-like cells, consistent with previous reports of *EGFR* as a regulator of astrocytic differentiation (Sun et al., 2005).

To systematically examine the association between cellular states and genetics, we next turned to 401 bulk samples from the TCGA glioblastoma dataset. Bulk expression profiles reflect an average of the diverse tumor constituents, and therefore, the expression of each meta-module defines a crude estimate for the abundance of the corresponding cellular state in bulk samples. We scored each bulk sample for expression of each of the four meta-modules and examined the association between expression scores and genetic features (Figure 5B). As expected from the above analysis, TCGA tumors with high-level genetic amplifications of *EGFR* are significantly associated with higher AC-like bulk scores ($p < 10^{-5}$). Similarly, high-level amplifications of *PDGFRA* and *CDK4* were associated with OPC-like and NPC-like scores, respectively, consistent with the known roles of these genes as OPC and NPC regulators in normal development (Lim and Kaldis, 2012; Zhu et al., 2014). Thus, although OPC-like and NPC-like abundances are largely coupled and together define the TCGA-PN subtype, they are distinct enough to detect differential associations with amplifications of relevant regulators. Several point mutations were also correlated with particular cellular states, such as *NF1* alterations in MES-high tumors (Figure S5F), but CNAs had stronger effects (Figure 5B); each cellular state was significantly associated with specific CNAs. We also observed that deletion of chromosome arm 5q and the MES-like state were negatively related across the TCGA dataset (Figure S5G), suggesting that deletion of genes on this chromosome arm could limit the number of MES-like cells. This chromosomal region encodes potential regulators of mesenchymal expression programs (*SMAD5* and *TGFBI*), as well as multiple cytokines and chemokines (*CSF2*, *IL3*, *IL4*, *IL5*, *IL13*, and *CXCL14*) that could be involved in communication with microglia/macrophages and other immune cells (Wang et al., 2017).

EGFR Drives an AC-Like Program and *CDK4* an NPC-Like Program in Mouse Neural Cells

We hypothesized that some of these genetic alterations might favor specific cellular states by increasing their growth and/or by inducing state transitions. To test this hypothesis, we overexpressed *CDK4*, *EGFR*, and control GFP in primary mouse neural progenitor cells derived from embryonic stem cells (STAR Methods) and performed phenotypic characterization and scRNA-seq. Supporting our model, NPCs overexpressing *EGFR* induced an AC-like program, as assessed by both GFAP staining and scRNA-seq analysis (Figures 6A–6C, S6A, and S6B; STAR Methods). Conversely, cells overexpressing *CDK4* induced an NPC-like program (Figures 6D and S6B; STAR Methods). Thus, these oncogenes favor transition of non-cancer progenitor cells toward cellular states that they also correlate with in the tumor context. Because *EGFR* and *CDK4* have established roles in driving cellular proliferation, we additionally tested the effect of these oncogenes on proliferation of the respective cell types. We observed that mouse neural progenitor cells proliferated more upon overexpression of *CDK4* than of *EGFR* or GFP control (Figure 6E), whereas mouse astrocytes proliferated more upon overexpression of *EGFR* than of *CDK4* or GFP control (Figure 6F). This suggests that distinct neural cell types respond differently to these glioblastoma oncogenes and mirrors the associations between genetics and cell states that we observed across the TCGA dataset. Altogether, our results support a model in which oncogenes such as *EGFR* and *CDK4* play a key role in regulating both transitions and growth of specific

Figure 5. The Distribution of Glioblastoma Cellular States Is Associated with Chromosomal Amplifications across the TCGA Glioblastoma Cohort

(A) Pie charts displaying the fraction of cells in four cellular states in each glioblastoma from our cohort. Tumor indices are above each pie chart; pediatric tumors are indicated in red and recurrent tumors with "R." Tumors are grouped by bulk TCGA subtype as labeled.

(B) Analysis of the TCGA glioblastoma cohort shows that high-level amplifications of *EGFR*, *PDGFRA*, and *CDK4* are associated with high bulk scores for the AC-like, OPC-like, and NPC-like cellular states, respectively. Shown are significance values, $-\log_{10}$(P value), for the association of chromosomal amplifications with high bulk scores (shown above the zero line, indicating enrichment of the cellular state) or with low bulk scores (shown below the zero line, indicating depletion of the cellular state). Single chromosome gains are distinguished from high-level amplifications (Brennan et al., 2013), which are found to have significant associations with three cellular states (AC-like, OPC-like, and NPC-like), whereas no associations of chromosomal amplifications were found with the MES-like state.

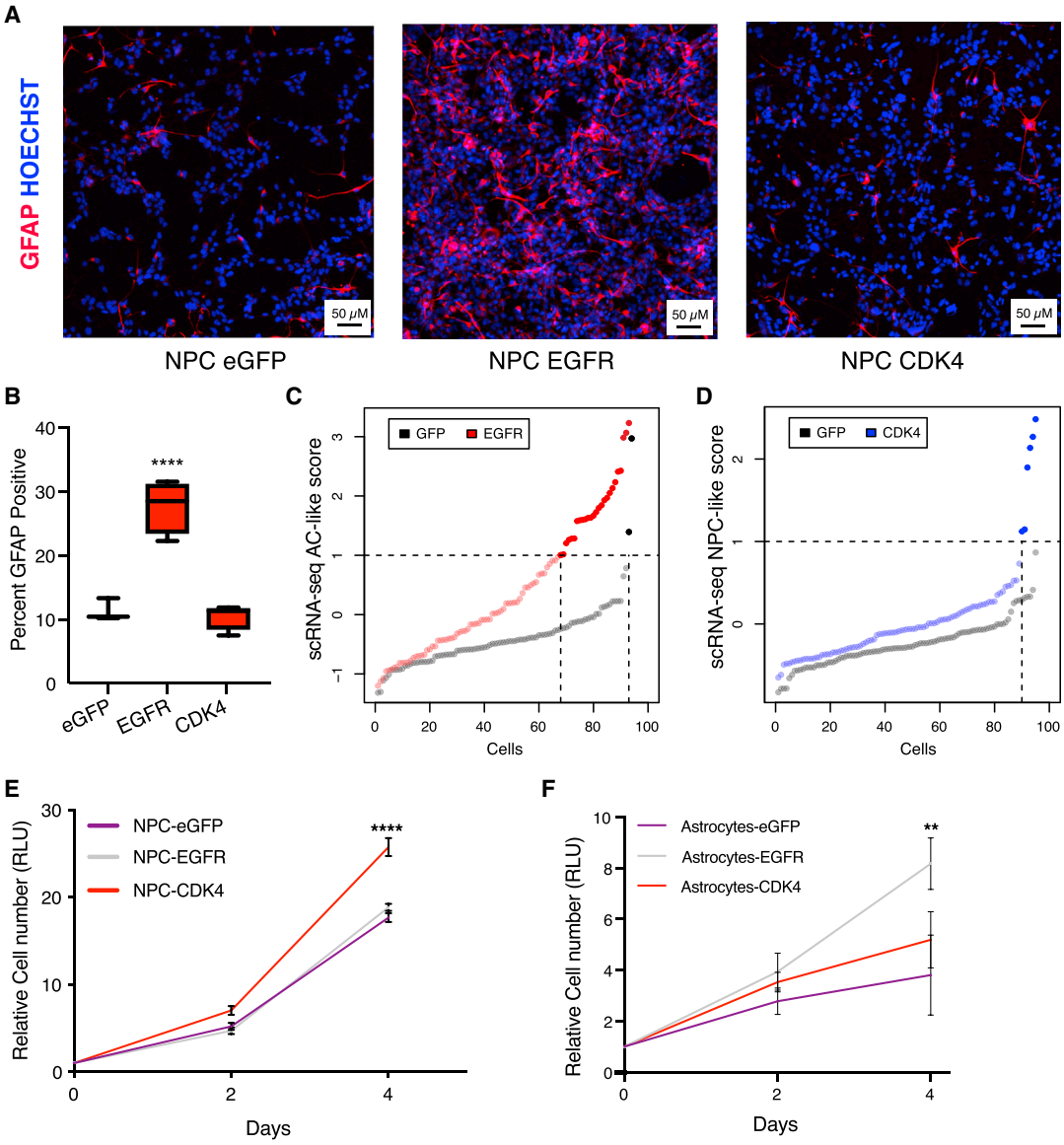


Figure 6. Glioblastoma Oncogenes Drive Defined Cellular States

(A) Micrographs of immunofluorescence of mouse NPCs overexpressing *EGFR*, *CDK4*, or eGFP immunostained for the astrocytic marker GFAP (red).

(B) Quantification of GFAP$^+$ cells shown in (A) (STAR Methods).

(C) scRNA-seq scores for the AC-like signature (y axis) of ranked cells (x axis) overexpressing EGFR (red) or GFP (black) (STAR Methods).

(D) scRNA-seq scores for the NPC-like signature (y axis) of ranked cells (x axis) overexpressing CDK4 (blue) or GFP (black).

(E) Growth curve using NPCs overexpressing eGFP, *EGFR*, or *CDK4* shows increased proliferation ($p < 0.0001$) in *CDK4*-expressing cells. Abbreviation is as follows: RLU, Relative Light Units (arbitrary value).

(F) Growth curve of astrocytes derived from the engineered NPCs (STAR Methods) shows significant ($p < 0.002$, ANOVA) increase in growth of astrocytes overexpressing EGFR.

neurodevelopmental cell types, and therefore when they occur in tumors, they might not only drive tumor progression but also shape the distribution of cellular states within the tumor.

Demonstration of Cellular Plasticity by Combined scRNA-Seq and Cellular Barcoding

Whereas defined genetic events appear to drive the identity of the most common cellular states, we speculated that genetics might incompletely skew toward specific cellular states, such that a diversity of states is maintained through cellular plasticity. To experimentally test the capacity of cells to transition between states, we sought to isolate cells in a specific state, use them to initiate tumors in a patient-derived xenograft (PDX) model, and determine the state distribution in the resulting tumor.

First, to isolate cells of a specific state, we searched for cell-surface markers from the meta-module genes and identified

CD24 and CD44 among the 4 top-scoring genes for the NPC-like and MES-like states, respectively. Next, we isolated CD24-high cells, CD44-high cells, and unselected malignant cells ($CD45^-$) from a fresh tumor sample (MGH143) (Figure 7A). We selected a tumor with the *EGFRvIII* genetic alteration (Figure S7A), a constitutively active mutant of EGFR and accordingly a high proportion of AC-like cells and smaller proportions of NPC-like and MES-like cells. However, these latter states were efficiently enriched in the CD24-high and CD44-high fractions, as demonstrated by scRNA-seq of the sorted populations (Figure 7A, 7B, and S7B). We then tested the three fractions (CD24-high, CD44-high, and $CD45^-$) for tumor-initiation potential by orthotopic xenografts in immunocompromised mice (Figure 7A and 7B). Each of the populations robustly initiated glioblastoma in multiple mice, indicating their tumor-initiating potential (Figure S7C and S7D). Upon tumor development, we analyzed the PDXs by scRNA-seq to determine the spectrum of cellular states in these models and compare them to the injected patient sample (Figure 7B).

Regardless of the population used to initiate the PDX—$CD45^-$ (containing mostly AC-like cells), NPC-like, or MES-like—the derived tumor contained all three states in a similar distribution (Figure 7B). Indeed, in almost all cases, the derived tumor recapitulated the distribution of cellular states found in the original patient sample. The only exception was one PDX derived from the CD24-high fraction; but even this PDX had a decreased proportion of the sorted cellular state and an increased proportion of AC-like cells, the most common cellular state in the patient sample and the one associated with *EGFR* amplification. These results show that the baseline distribution of cellular states can be recapitulated in the mouse brain microenvironment and, moreover, that cellular states transition from the sorted state to other states, from a single sorted population. These results were also recapitulated when the analysis was repeated for distinct PDX CNA subclones (Figure S7E-G).

To further demonstrate cellular plasticity in glioblastoma at single-cell resolution, we combined scRNA-seq with cellular barcoding in both a genetic mouse model and in PDX models. First, we modified a mouse model of glioblastoma in which lentiviruses harboring H-Ras and shP53 are stereotactically injected into the hippocampus of GFAP-cre animals, such that each transformed cell would additionally harbor a unique and heritable genetic tag (Figure 7C; STAR Methods) (Friedmann-Morvinski et al., 2012). scRNA-seq analysis of the resulting mouse tumors showed that each tumor contained three of the four cellular states identified in human glioblastoma (Figures 7D, S6C, and S6D). Because of cell proliferation, many of the heritable barcodes were identified in multiple cells, which were also identified by scRNA-seq. Importantly, 39% of those barcodes were seen among cells in different states, unambiguously demonstrating common plasticity among states.

Second, to assess whether plasticity is also observed in human glioblastoma, we derived two primary human cell cultures from patient samples (MGH143 and MGG23), infected them with lentiviruses harboring unique barcodes (Figure 7E; STAR Methods), and orthotopically transplanted them into recipient immunocompromised mice. scRNA-seq and barcode analysis of the resulting mouse tumors identified human glioblastoma cells that share the same genetic barcode but correspond to different states of glioblastoma. Notably, there were several instances of a single barcode found in cells of four different states, demonstrating that a single cell can give rise to all four states of glioblastoma observed in patients (Figures 7F and S6E; STAR Methods). Overall, these results are consistent with glioblastoma cells displaying plasticity of states, and with a baseline distribution reflecting a steady state that emerges by cellular transitions and the tumor's genotype.

DISCUSSION

A better understanding of the multiple sources of heterogeneity in glioblastoma and of their inter-relationships is a critical goal for neuro-oncology, with broad implications for therapy. Here, we started by performing the most comprehensive scRNA-seq analysis of glioblastoma to date, analyzing extensively 28 tumors from both adults and children. Each tumor is unique and the diversity within a tumor is driven by a combination of factors: genetic, epigenetic, and microenvironmental. Yet, we found that the diversity of malignant cells in glioblastoma converges to few recurrent expression signatures, given that almost all signatures of intra-tumoral heterogeneity were mapped to either the cell cycle or one of four general cellular states. Although heterogeneity is often viewed as a major barrier, the convergence of cellular signatures on a few common patterns of heterogeneity might help identify dependencies in glioblastoma shared by many tumors.

Each of those recurrent cellular states is associated with cycling cells, and the highest fractions are observed in NPC-like and OPC-like states, particularly in pediatric tumors. Analysis of intermediate states in patients, as well as PDX and lineage-tracing experiments, indicates plasticity between those four states, with multiple possible transitions. This suggests that each tumor is composed of cells in multiple cellular states that might proliferate or transition to other states. Such dynamic behavior implies that rates of proliferation and transitions would ultimately define a steady-state distribution, and that one might expect a similar distribution across different tumors, reflecting the intrinsic propensity of each cell to proliferate or transition to the other states. Yet, we observe widely different distributions between tumors, such that each state is most common in some tumors and least common in others, indicating that additional factors might influence the proliferation and transition rates. We propose that certain genetic factors dictate certain transition rates that in turn define a steady-state distribution. One such relevant genetic factor appears to be *EGFR* aberrations, which is associated with a relative abundance of AC-like cells both in our cohort as well as in the larger TCGA dataset. Similarly, amplifications of *CDK4* and *PDGFRA* are associated with abundance of the NPC-like and OPC-like states, respectively, whereas Chr5q deletions and *NF1* alterations affect the frequency of MES-like states (Figure 7G). These genetic associations might also provide an explanation for the differential distribution of cellular states between adult and pediatric glioblastoma: in pediatric tumors *EGFR* alterations are less common than in adult tumors, mirroring their decreased frequency of

Please cite this article in press as: Neftel et al., An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma, Cell (2019), https://doi.org/10.1016/j.cell.2019.06.024

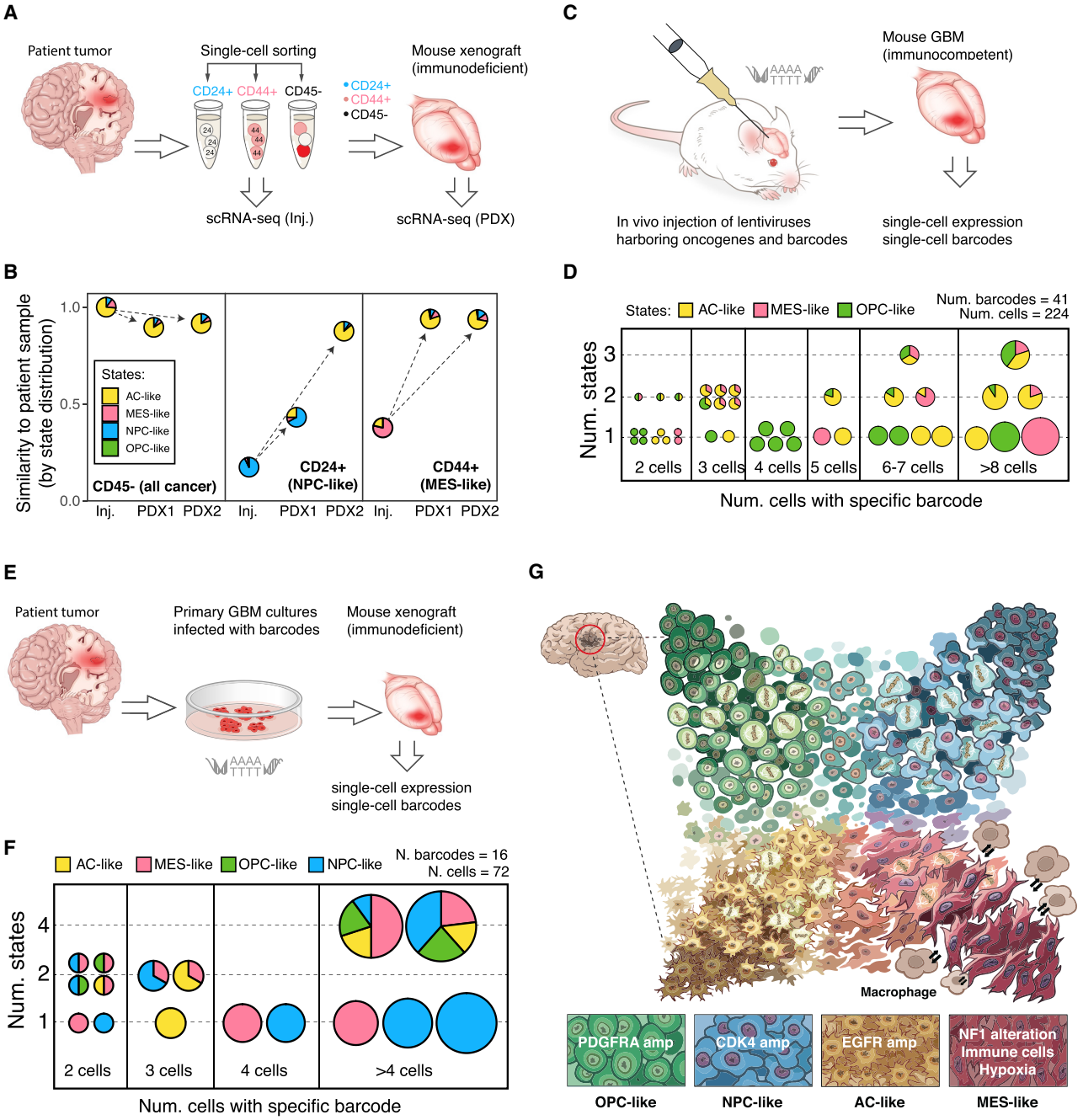


Figure 7. Cellular Transitions in Glioblastoma

(A) Experimental workflow. Different fractions of cells were sorted from patient sample MGH143 and injected orthotopically into immunocompromised mice to generate PDXs. The patient sample and the PDX subpopulations were subjected to scRNA-seq.

(B) Samples described in (A) are each represented by a pie chart depicting the fraction of cells in four states. Pie charts are positioned on the x axis on the basis of their sorted fraction and whether they represent injected or PDX sample, and on the y axis on the basis of their compositional similarity to the original patient sample (one minus the Manhattan distance over the fractions of four states).

(C) Experimental workflow. Lentiviruses harboring oncogenes and unique barcodes were injected into the mouse hippocampus (STAR Methods) and the resulting tumors were analyzed by scRNA-seq.

(D) Barcodes which were identified in multiple cells are each represented by a pie chart depicting the fraction of cells in each state. Pie charts are positioned on the basis of the number of cells with the respective barcode (x axis), and the number of cellular states observed among these cells (y axis). Pie chart sizes are proportional to $\log_2$ of the number of cells.

(legend continued on next page)

AC-like cells (Figure 3B). The association between genetics and cellular states could form the biological basis for the TCGA bulk expression subtypes, as tumor genetics might determine the more frequent cellular state and hence the average (i.e., bulk) expression profile (Verhaak et al., 2010; Wang et al., 2017).

Experimentally, this model is supported by overexpression experiments in neural progenitor cells, linking genetic drivers to specific cellular states (Figure 6), and by previous studies in which overexpression of *EGFR* in nestin$^+$ neural progenitors drove the formation of an astrocytoma-like tumor, whereas overexpression of *PDGFRA* in the same cells drove an oligodendroglioma-like tumor (Holland et al., 1998). Thus, signaling pathways relevant to certain cellular states during normal development can be selected for during tumorigenesis and play a role in stabilizing specific malignant cellular states. More generally, this model is consistent with the view that oncogenesis involves the generation of a self-renewing population with defects in differentiation capacity. Each of these genetic drivers might skew a particular cellular state toward self-renewal and therefore might promote the generation of tumors driven primarily by that specific state. This model would explain why PDXs derived from different cellular states remarkably converged toward the same distribution of states as observed in the patient sample. Thus, certain genetic drivers in glioblastoma might dictate certain transition probabilities and define the steady-state distribution. It is tempting to speculate that such capacity to define state transitions is being selected for and that *EGFR*, *PDGFRA*, and *CDK4* would be selected not only to promote glioblastoma growth, but to expand and stabilize a certain state within the glioblastoma ecosystem. Targeting such genetic drivers might modulate the distribution of states and could possibly lead to an alternative distribution driven primarily by the self-renewal of a different state. Such a scenario might explain the limited efficacy of targeting a single signaling pathway in glioblastoma.

In conclusion, we have elucidated the spectrum of expression states of glioblastoma cells and their plasticity, identifying cellular programs that recapitulate neural development, cell cycle, and influences of the microenvironment. By showing that specific genetic drivers of glioblastoma influence the frequency of those states, we provide a cellular correlate to glioblastoma genetic heterogeneity and a model that explains why different bulk expression programs, such as the TCGA subtypes, are enriched for defined genetic alterations. Further studies will be needed to assess translational opportunities and to evaluate the effect of existing therapeutic approaches on the spectrum of cellular states that drive glioblastoma.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cell.2019.06.024.

ACKNOWLEDGMENTS

We thank Ania Hupaloska and Sébastien Perroud for help with graphics. This work was supported by grants to M.L. Suvà from the Sontag Foundation, the

(E) Experimental workflow. Primary cultures are established from glioblastoma samples (MGH143 and MGG23) and infected with lentiviruses harboring unique barcodes, xenografted into mouse brains and tumors formed are analyzed by scRNA-seq.

(F) Unique barcodes from (E) are displayed as shown in (D).

(G) Model for the cellular states of glioblastoma and their genetic and micro-environmental determinants. Mitotic spindles indicate cycling cells. Lighter or darker tones indicate strength of each program. Intermediate states are shown in between the four states and indicate transitions.

Howard Goodman Fellowship, the Merkin Institute Fellowship, the Wang Family Fund, the Smith Family Foundation, the Chan-Zuckerberg Initiative, the Alex's Lemonade Stand Foundation, the V Foundation for Cancer Research, and The Swiss National Science Foundation Sinergia grant to M.L.Suvà. and I.S. I.T. is the incumbent of the Dr. Celia Zwillenberg-Fridman and Dr. Lutz Zwillenberg Career Development Chair, and was supported by the Zuckerman STEM Leadership Program, the Human Frontiers Science Program, the Mexican friends new generation, and the Benoziyo Endowment Fund. M.G.F. was supported by a Career Award for Medical Scientist from Burroughs Wellcome Fund and K12 Paul Calabresi Career Award for Clinical Oncology (K12CA090354). C.N. was supported by the Placide Nicod Foundation. T.Hara was supported by Grant-in-Aid for JSPS Fellows from the Japan Society for the Promotion of Science. A.R. was supported by funds from the Howard Hughes Medical Institute, the Klarman Cell Observatory, STARR cancer consortium, NCI grants 1U24CA180922 and R33CA202820, the Koch Institute Support (core) (P30CA14051) from the National Cancer Institute, the Ludwig Center and the Broad Institute. A.R. is a scientific advisory board member for ThermoFisher Scientific, Syros Pharmaceuticals and Driver Group. A.P.P. was supported by a Career Award for Medical Scientist from Burroughs Wellcome Fund. D.P.C. is supported by the BWF Career Award in the Medical Sciences, and Tawingo Chair Fund and has received consulting fees from Lilly and Merck (unrelated to this work). B.E.B. was supported by the NIH Common Fund and National Cancer Institute (DP1CA216873), the American Cancer Society, the Ludwig Center at Harvard Medical School, and the Bernard and Mildred Kayden MGH Research Institute Chair. B.E.B. is an advisor and equity holder for Fulcrum Therapeutics, 1CellBio, HiFiBio, and Arsenal Biosciences; is an advisor for Cell Signaling Technologies; and has equity in Nohla Therapeutics. Flow-cytometry and sorting services at MGH were supported by NIH shared instrumentation grants 1S10RR023440-01A1 and P30CA014195. Flow-cytometry and sorting services at the Salk Institute were supported by NIH grant S10-OD023689. I.M.V. and T.Hunter were supported by NIH grant R01CA195613 and the Helmsley Center for Genomic Medicine.

AUTHOR CONTRIBUTIONS

C.N., J.L., M.G.F., T.H., I.T., and M.L.Suvà conceived the project, designed the study, and interpreted results. C.N., M.G.F., M.E.S., A.R.R., C.M.H., and M.L.Shaw collected glioblastoma single cells and generated sequencing data. I.T. and J.L. performed computational analyses. T.Hara. designed lentiviral barcodes and performed lineage-tracing experiments. G.R. developed mouse neural cell cultures used in the study and performed overexpression experiments. D.S., L.X., E.P., and E.R. provided support for single-cell genetic analyses. J.M.F., R.L.S., and R.M. provided flow-cytometry expertise. J.S., A.K., M.S.B., N.G.C., T.T.B., and P.K.B. identified and consented patients for the study. N.D., S.G., J.D., M.E.S., and C.N. performed ISH and IHC experiments. K.P., D.B., and Q.D.N. performed mouse PDX and animal follow-up. C.R., D.D., S.B., L.M., J.G., C.H., R.G., T.C., B.V.N., W.T.C., B.S.C., I.M.V., A.S.R., M.M.L., M.P.F, I.S., T.Hunter., H.W., N.R., M.M., O.R.R., K.L.L, M.M., D.P.C., A.P.P., D.N.L., A.R. and B.E.B. provided experimental and analytical support. I.T. (computational part) and M.L.S. (experimental part) jointly supervised this work and interpreted results. I.T., A.R., and M.L.S. wrote the manuscript with feedback from all authors.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: September 13, 2018
Revised: March 27, 2019
Accepted: June 13, 2019
Published: July 18, 2019

REFERENCES

Bao, S., Wu, Q., McLendon, R.E., Hao, Y., Shi, Q., Hjelmeland, A.B., Dewhirst, M.W., Bigner, D.D., and Rich, J.N. (2006). Glioma stem cells promote radioresistance by preferential activation of the DNA damage response. Nature *444*, 756–760.

Brennan, C.W., Verhaak, R.G., McKenna, A., Campos, B., Noushmehr, H., Salama, S.R., Zheng, S., Chakravarty, D., Sanborn, J.Z., Berman, S.H., et al.; TCGA Research Network (2013). The somatic genomic landscape of glioblastoma. Cell *155*, 462–477.

Chen, J., Li, Y., Yu, T.S., McKay, R.M., Burns, D.K., Kernie, S.G., and Parada, L.F. (2012). A restricted cell population propagates glioblastoma growth after chemotherapy. Nature *488*, 522–526.

Darmanis, S., Sloan, S.A., Zhang, Y., Enge, M., Caneda, C., Shuer, L.M., Hayden Gephart, M.G., Barres, B.A., and Quake, S.R. (2015). A survey of human brain transcriptome diversity at the single cell level. In Proceedings of the National Academy of Sciences of the United States of America.

Darmanis, S., Sloan, S.A., Croote, D., Mignardi, M., Chernikova, S., Samghababi, P., Zhang, Y., Neff, N., Kowarsky, M., Caneda, C., et al. (2017). Single-cell RNA-seq analysis of infiltrating neoplastic cells at the migrating front of human glioblastoma. Cell Rep. *21*, 1399–1410.

Filbin, M.G., Tirosh, I., Hovestadt, V., Shaw, M.L., Escalante, L.E., Mathewson, N.D., Neftel, C., Frank, N., Pelton, K., Hebert, C.M., et al. (2018). Developmental and oncogenic programs in H3K27M gliomas dissected by single-cell RNA-seq. Science *360*, 331–335.

Friedmann-Morvinski, D., Bushong, E.A., Ke, E., Soda, Y., Marumoto, T., Singer, O., Ellisman, M.H., and Verma, I.M. (2012). Dedifferentiation of neurons and astrocytes by oncogenes can induce gliomas in mice. Science *338*, 1080–1084.

Holland, E.C., Hively, W.P., DePinho, R.A., and Varmus, H.E. (1998). A constitutively active epidermal growth factor receptor cooperates with disruption of G1 cell-cycle arrest pathways to induce glioma-like lesions in mice. Genes Dev. *12*, 3675–3685.

Kerman, B.E., Kim, H.J., Padmanabhan, K., Mei, A., Georges, S., Joens, M.S., Fitzpatrick, J.A., Jappelli, R., Chandross, K.J., August, P., and Gage, F.H. (2015). In vitro myelin formation using embryonic stem cells. Development *142*, 2213–2225.

Lathia, J.D., Mack, S.C., Mulkearns-Hubert, E.E., Valentim, C.L., and Rich, J.N. (2015). Cancer stem cells in glioblastoma. In Genes & development (United States), pp. 1203–1217.

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. B.M.C. Bioinformatics, ed. (England), pp. 323.

Lim, S., and Kaldis, P. (2012). Loss of Cdk2 and Cdk4 induces a switch from proliferation to differentiation in neural stem cells. Stem Cells *30*, 1509–1520.

Louis, D.N., Ohgaki, H., Wiestler, O.D., and Cavenee, W.K. (2016). WHO classification of tumors of the central nervous system, Revised 4th edition (Lyon: IARC).

Müller, S., Liu, S.J., Di Lullo, E., Malatesta, M., Pollen, A.A., Nowakowski, T.J., Kohanbash, G., Aghi, M., Kriegstein, A.R., Lim, D.A., and Diaz, A. (2016). Single-cell sequencing maps gene expression to mutational phylogenies in PDGF- and EGF-driven gliomas. Mol. Syst. Biol. *12*, 889.

Nowakowski, T.J., Bhaduri, A., Pollen, A.A., Alvarado, B., Mostajo-Radji, M.A., Di Lullo, E., Haeussler, M., Sandoval-Espinosa, C., Liu, S.J., Velmeshev, D., et al. (2017). Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. Science *358*, 1318–1323.

Parada, L.F., Dirks, P.B., and Wechsler-Reya, R.J. (2017). Brain tumor stem cells remain in play. J. Clin. Oncol. *35*, 2428–2431.

Patel, A.P., Tirosh, I., Trombetta, J.J., Shalek, A.K., Gillespie, S.M., Wakimoto, H., Cahill, D.P., Nahed, B.V., Curry, W.T., Martuza, R.L., et al. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science *344*, 1396–1401.

Picelli, S., Faridani, O.R., Bjorklund, A.K., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. In Nat Protoc (England), pp. 171–181.

Pollen, A.A., Nowakowski, T.J., Chen, J., Retallack, H., Sandoval-Espinosa, C., Nicholas, C.R., Shuga, J., Liu, S.J., Oldham, M.C., Diaz, A., et al. (2015).

Molecular identity of human outer radial glia during cortical development. Cell *163*, 55–67.

Puram, S.V., Tirosh, I., Parikh, A.S., Patel, A.P., Yizhak, K., Gillespie, S., Rodman, C., Luo, C.L., Mroz, E.A., Emerick, K.S., et al. (2017). Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. Cell *171*, 1611–1624.

Sottoriva, A., Spiteri, I., Piccirillo, S.G., Touloumis, A., Collins, V.P., Marioni, J.C., Curtis, C., Watts, C., and Tavaré, S. (2013). Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. Proc. Natl. Acad. Sci. USA *110*, 4009–4014.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550.

Sun, Y., Goderie, S.K., and Temple, S. (2005). Asymmetric distribution of EGFR receptor during mitosis generates diverse CNS progenitor cells. Neuron *45*, 873–886.

Suvà, M.L., Rheinbay, E., Gillespie, S.M., Patel, A.P., Wakimoto, H., Rabkin, S.D., Riggi, N., Chi, A.S., Cahill, D.P., Nahed, B.V., et al. (2014). Reconstructing and reprogramming the tumor-propagating potential of glioblastoma stem-like cells. Cell *157*, 580–594.

Tanay, A., and Regev, A. (2017). Scaling single-cell genomics from phenomenology to mechanism. Nature *541*, 331–338.

Tirosh, I., and Suva, M.L. (2019). Deciphering Human Tumor Biology by Single-Cell Expression Profiling. Annual Review of Cancer Biology *3*.

Tirosh, I., Izar, B., Prakadan, S.M., Wadsworth, M.H., 2nd, Treacy, D., Trombetta, J.J., Rotem, A., Rodman, C., Lian, C., Murphy, G., et al. (2016a). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. In Science (United States), pp. 189–196.

Tirosh, I., Venteicher, A.S., Hebert, C., Escalante, L.E., Patel, A.P., Yizhak, K., Fisher, J.M., Rodman, C., Mount, C., Filbin, M.G., et al. (2016b). Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. In Nature (England), pp. 309–313.

Venteicher, A.S., Tirosh, I., Hebert, C., Yizhak, K., Neftel, C., Filbin, M.G., Hovestadt, V., Escalante, L.E., Shaw, M.L., Rodman, C., et al. (2017). Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. Science *355*, eaai8478.

Verhaak, R.G., Hoadley, K.A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M.D., Miller, C.R., Ding, L., Golub, T., Mesirov, J.P., et al.; Cancer Genome Atlas Research Network (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell *17*, 98–110.

Wakimoto, H., Mohapatra, G., Kanai, R., Curry, W.T., Jr., Yip, S., Nitta, M., Patel, A.P., Barnard, Z.R., Stemmer-Rachamimov, A.O., Louis, D.N., et al. (2012). Maintenance of primary tumor phenotype and genotype in glioblastoma stem cells. Neuro-oncol. *14*, 132–144.

Wang, Q., Hu, B., Hu, X., Kim, H., Squatrito, M., Scarpace, L., deCarvalho, A.C., Lyu, S., Li, P., Li, Y., et al. (2017). Tumor evolution of glioma-intrinsic gene expression subtypes associates with immunological changes in the microenvironment. Cancer cell *32*, 42–56.

Yuan, J., Levitin, H.M., Frattini, V., Bush, E.C., Boyett, D.M., Samanamud, J., Ceccarelli, M., Dovas, A., Zanazzi, G., Canoll, P., et al. (2018). Single-cell transcriptome analysis of lineage diversity in high-grade glioma. Genome Med. *10*, 57.

Zhu, Q., Zhao, X., Zheng, K., Li, H., Huang, H., Zhang, Z., Mastracci, T., Wegner, M., Chen, Y., Sussel, L., and Qiu, M. (2014). Genetic evidence that Nkx2.2 and Pdgfra are major determinants of the timing of oligodendrocyte differentiation in the developing CNS. Development *141*, 548–555.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Anti-human CD45-VioBlue, (clone REA747)	Miltenyi Biotec	Cat#130-110-775; RRID: AB_2658242
Anti-human CD24-APC, (clone REA832)	Miltenyi Biotec	Cat#130-112-846; RRID: AB_2656558
Anti-human CD44-VioBlue (clone REA690)	Miltenyi Biotec	Cat#130-113-906; RRID: AB_2726396
Anti-mouse CD45-VioBlue, mouse (clone 30F11)	Miltenyi Biotec	Cat#130-119-130; RRID: AB_2751631
Anti-mouse CD16/32	BD Biosciences	Cat#553142; RRID: AB_39465
PerCP anti-mouse CD45 (30-F11)	BD Biosciences	Cat#550994; RRID: AB_394003
Anti-human GFAP	Agilent	Cat#Z033429-2
Bacterial and Virus Strains		
Endura Electrocompetent cells	Lucigen	Cat#60242-2
Biological Samples		
C57BL/6 Mouse Embryonic Fibroblasts, irradiated	GIBCO	Cat#A34961
Chemicals, Peptides, and Recombinant Proteins		
Buffer TCL	QIAGEN	Cat#1031576
Bovine Serum Albumin	Sigma Aldrich	Cat#A3059
Calcein, AM, cell-permeant dye	Invitrogen	Cat#C3100MP
TO-PRO-3 Iodide	Invitrogen	Cat#T3605
Maxima H Minus Reverse Transcriptase	Thermo Scientific	Cat#EP0753
KAPA HiFi HotStart ReadyMix	Roche	Cat#KK2602
dNTP Mix (10 mM each)	Thermo Scientific	Cat#R0192
Recombinant RNase Inhibitor	Takara	Cat#2313B
Betaine solution	Sigma-Aldrich	Cat#B0300
Magnesium chloride 1M	Invitrogen	Cat#AM9530G
RNAscope Hydrogen Peroxide	ACDbio	Cat#322335
RNAscope Target Retrieval Reagent	ACDbio	Cat#322000
RNAscope Protease Plus	ACDbio	Cat#322331
Neurobasal Medium	GIBCO	Cat#21103049
N-2 Supplement (100X)	GIBCO	Cat#17502048
B-27 Supplement (50X), serum free	GIBCO	Cat#17504044
GlutaMAX Supplement	GIBCO	Cat#35050061
Penicillin-Streptomycin	GIBCO	Cat#15140122
Recombinant Human EGF Protein	R&D Systems	Cat#236-EG-200
Recombinant Human FGF basic/FGF2	R&D Systems	Cat#4114-TC-01M
TrypLE Express Enzyme (1X), phenol red	GIBCO	Cat#12605010
Knockout DMEM	GIBCO	Cat#10829018
Embryonic stem-cell FBS	GIBCO	Cat#16141061
MEM Non-Essential Amino Acids Solution (100X)	GIBCO	Cat#11140050
L-Glutamine (200 mM)	GIBCO	Cat#25030081
Penicillin-Streptomycin	GIBCO	Cat#15140122

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
2-Mercaptoethanol	Sigma-Aldrich	Cat#M6250
ESGRO Recombinant Mouse LIF Protein	Sigma-Aldrich	Cat#ESG1107
Q5 Hot Start High-Fidelity 2X Master Mix	New England Biolabs	Cat#M0494
Gibson Assembly Master Mix	New England Biolabs	Cat#E2611
EcoRI-HF	New England Biolabs	Cat#R3101
Lipofectamine 2000 Transfection Reagent	Invitrogen	Cat#11668019
Penicillin-Streptomycin	GIBCO	Cat#15140122
Human EGF	PeproTech	Cat#AF-100-15
Human FGF-basic	PeproTech	Cat#AF-100-18B
TrypLE Express Enzyme (1X), no phenol red	GIBCO	Cat#12604013
0.1% typeI collagenase	Thermo Fisher Scientific	Cat#17100-017
Calcein Blue AM	BD Biosciences	Cat#564060
Zombie NIR (BioLegend)	BioLegend	Cat#423106
Hoechst 33342, Trihydrochloride, Trihydrate - 10 mg/mL Solution in Water	Fisher	Cat#H3570
VECTASHIELD Antifade Mounting Medium	Vectorlabs	Cat#H-1000; RRID: AB_2336789
Critical Commercial Assays		
Brain Tumor Dissociation Kit (P)	Miltenyi	Cat#130-095-942
Dead Cell Removal Kit	Miltenyi Biotec	Cat#130-090-101
CD45 MicroBeads, human	Miltenyi Biotec	Cat#130-045-801; RRID: AB_2783001
CD24 MicroBead Kit, human	Miltenyi Biotec	Cat#130-095-951
CD44 MicroBeads, human	Miltenyi Biotec	Cat#130-095-194
Agencourt RNAClean XP	Beckman Coulter	Cat#A66514
Agencourt AMPure XP	Beckman Coulter	Cat#A63882
Nextera XT DNA Library Preparation Kit (96 samples)	Illumina	Cat#FC-131-1096
NextSeq 500/550 High Output Kit v2.5 (75 Cycles)	Illumina	Cat#20024906
Bioanalyzer High Sensitivity DNA Analysis	Agilent	Cat#5067-4626
Qubit dsDNA HS Assay kit	Invitrogen	Cat#Q32854
RNAscope 2.5 HD Duplex Detection Kit	ACDbio	Cat#322430
QIAquick Gel Extraction Kit	QIAGEN	Cat#28706
QIAquick PCR Purification Kit	QIAGEN	Cat#28106
PureLink HiPure Maxiprep kit	Thermo Fisher Scientific	Cat#K2100-17
Deposited Data		
Single-cell RNA-sequencing data	Gene Expression Omnibus	GEO: GSE131928
Single-cell RNA-sequencing data	Broad Institute Single-Cell Portal	SCP393
Experimental Models: Cell Lines		
MGH143	This paper (Table S1)	N/A
MGG23	Wakimoto et al., 2012	N/A
Experimental Models: Organisms/Strains		
NOD.Cg-Prkdcscid Il2rgtm1Wjl/SzJ (NSG)	Jackson laboratory	Cat#5557
FVB-Tg(GFAP-cre)25Mes/J (backcrossed to C57BL/6J)	Jackson laboratory	Cat#4600
Oligonucleotides		
Custom LNA Oligonucleotide (1 μmol synthesis)	QIAGEN	Cat#339413
Oligo(dT) Primer (5′–AAGCAGTGGTATCAACGCAGA GTACTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN-3′)	IDT	N/A
ISPCR Primer (5′-AAGCAGTGGTATCAACGCAGAGT-3′)	IDT	N/A

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Hs-CD24	ACDbio	Cat#313021, Cat#313021-C2
Hs-CD44	ACDbio	Cat#311271-C2
Hs-PDGFRA	ACDbio	Cat#604481-C2
Hs-MKI67	ACDbio	Cat#591771, Cat#591771-C2
Hs-S100B	ACDbio	Cat#430891
Recombinant DNA		
DNA fragment with 16xN mixed bases	Integrated DNA Technologies	N/A
pLV-CMV-GFP	This paper	N/A
pTomo-HrasV12-IRES-GFP-shp53	Dinorah Friedmann-Morvinski et al.,2012	N/A
Barcode amplify primer Fw (5′-ccctttgggccg cctccccgcctggGCAACTGACTGAAATGCCTC-3′)	Integrated DNA Technologies	N/A
Barcode amplify primer Rv (5′-attggtcttaaaggtaccg agctcgTCAAGCCTCAGACAGTGGTTC-3′)	Integrated DNA Technologies	N/A
Other		
Falcon 4-well Culture Slide	Corning	Cat#354114

LEAD CONTACT AND MATERIALS AVAILABILITY

Further information and requests for resources and reagents should be directed to and will be fulfilled by Mario L. Suvà (Suva.Mario@mgh.harvard.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human Subjects

Adult patients at Massachusetts General Hospital (MGH) and pediatric patients and their parents at Boston Children's hospital provided preoperative informed consent to take part in the study in all cases after the Institutional Review Board Protocols DF/HCC 10-417 and DF/HCC 15-370B. Patients were males and females. Clinical characteristics are summarized in (Table S1).

Cell lines

Patient derived primary cultures (MGH143, MGG23) were grown in Neurobasal Medium (GIBCO 21103-049) supplemented with 1X N2/B27 (GIBCO), 1% Penicillin/Streptomycin (GIBCO), 1X Glutamax (GIBCO), 20 ng/mL EGF and 20 ng/mL bFGF (FGF2). The details of MGH143 are summarized in (Table S1). MGG23 is detailed in (Wakimoto et al., 2012).

Mouse NPCs were established from mouse ES cells (V6.5) using previously published protocols (Kerman et al., 2015). NPCs were propagated with DMEM:F12 (GIBCO 11320-033 with L-glutamine/Sodium Bicarbonate) supplemented with 1X N2/B27, 1% Penicillin/Streptomycin, 1 μg/mL laminin (from EHS tumor), 20 ng/mL EGF and 20 ng/mL bFGF (FGF2). NPCs were cultured on poly-l-ornithine coated tissue culture treated dishes. Astrocytes were derived and propagated from NPCs using NPC media supplemented with 4% Fetal Bovine Serum (FBS).

METHOD DETAILS

Tumor acquisition and single-cell sorting

Fresh tumors were collected directly from the operating room at the time of surgery and presence of glioblastoma was confirmed by frozen section. Tumors were mechanically and enzymatically dissociated using a papain-based brain tumor dissociation kit (Miltenyi Biotec) as previously reported (Filbin et al., 2018; Patel et al., 2014; Tirosh et al., 2016b; Venteicher et al., 2017). Tumor cells were blocked in 1% bovine serum albumin in Hanks buffered saline solution (BSA / HBSS). Tumors were first stained first with CD45-Vioblue direct antibody conjugate (clone REA747, Miltenyi Biotec) for 30 min at 4°C. Cells were washed with cold PBS, and then re-suspended in 1 mL of BSA / HBSS containing 1 μM calcein AM (Life Technologies) and 0.33 μM TO-PRO-3 iodide (Life Technologies) to co-stain for 30 min before sorting. For MGH143, CD45 MicroBeads (Miltenyi Biotec) were used to remove immune cells. CD45 negative cells were stained with calcein AM (Life Technologies) for viability and CD24-APC (human antibody, clone REA832, Miltenyi Biotec) and CD44-VioBlue (human antibody, clone REA690, Miltenyi Biotec) to sort subpopulations of viable non-immune cells. Sorting was performed with the FACS Aria Fusion Special Order System (Becton Dickinson) using 488 nm (calcein AM, 530/30 filter),

640nm (TO-PRO-3 or CD24-APC, 670/14 filter), and 405 nm (CD45-VioBlue or CD44-VioBlue, 450/50 filter) lasers. Standard, strict forward scatter height versus area criteria were used to discriminate doublets and gate only singleton cells. Viable single cells were identified as calcein AM positive and TO-PRO-3 negative. We sorted individual, viable, immune, and non-immune single cells into 96-well plates containing TCL buffer (QIAGEN) with 1% beta-mercaptoethanol. Plates were frozen on dry ice immediately after sorting and stored at −80°C prior to whole transcriptome amplification, library preparation and sequencing. For samples processed on the 10x genomics platform, dead cells were removed from single-cell suspensions using Dead Cell Removal Kit (Miltenyi Biotec).

RNA *in situ* hybridization

Paraffin-embedded tissue sections from tumors from Massachusetts General Hospital and Boston Children's Hospital were obtained according to Institutional Review Board-approved protocols. Sections were mounted on glass slides and stored at −80°C. Slides were stained using the RNAscope 2.5 HD Duplex Detection Kit (Advanced Cell Technologies, Cat. No. 322430), as previously described (Filbin et al., 2018; Tirosh et al., 2016b; Venteicher et al., 2017). Briefly, slides were baked for 1 h at 60°C, deparaffinized and dehydrated with xylene and ethanol. The tissue was pretreated with RNAscope Hydrogen Peroxide (Cat. No. 322335) for 10 min at room temperature and RNAscope Target Retrieval Reagent (Cat. No. 322000) for 15 min at 98°C. RNAscope Protease Plus (Cat. No. 322331) was then applied to the tissue for 30 min at 40°C. Hybridization probes were prepared by diluting the C2 probe (red) 1:50 into the C1 probe (green). Advanced Cell Technologies RNAscope Target Probes used included Hs-CD24 (Cat. No. 313021; Cat. No. 313021-C2), Hs-CD44 (Cat. No. 311271-C2), Hs-PDGFRA (Cat. No. 604481-C2), Hs-S100B (Cat. No. 430891), Hs-MKI67 (Cat. No. 591771; 591771-C2). Probes were added to the tissue and hybridized for 2 h at 40°C. A series of 10 amplification steps were performed using instructions and reagents provided in the RNAscope 2.5 HD Duplex Detection Kit. Tissue was counterstained with Gill's hematoxylin for 25 s at room temperature followed by mounting with VectaMount mounting media (Vector Laboratories). For ISH quantification, at least 1,000 cells were counted in representative areas of the tumors.

Intracranial patient-derived xenografts

Fresh tumor cells isolated directly from human glioblastoma at the time of surgery were stereotactically injected into the right striatum of 5- to 12-week-old female NSG mice (NOD.Cg-Prkdcscid Il2rgtm1Wjl/SzJ, The Jackson Laboratory, Bar Harbor, ME). Briefly, mice were anesthetized with 2% isoflurane mixed with medical air and placed on a stereotactic frame. The skull of the mouse was exposed through a small skin incision, and a small burr hole was made using a 25-gauge needle at the selected stereotactic coordinates. The cells suspended in 6uL PBS were loaded on a 33- gauge Hamilton syringe, and injected slowly using the following coordinates: 2.0 mm lateral of the bregma, and 2 mm deep to the cortical surface. Upon completing injection, the needle was left in place for another minute, then withdrawn slowly to help reduce cell reflux. After closing the scalp with suture and staple, mice were returned to their cages placed on a warming pad and visually monitored until full recovery. Mice were then checked daily for signs of distress, including seizures, ataxia, weight loss, and tremors. Mice were also monitored by imaging using small animal MRI, first 8 weeks after injection and again when they started to display neurological symptoms, including head tilt, seizures, sudden weight loss, loss of balance, and ataxia. Mice were sacrificed as soon as they became symptomatic, brains were collected directly after euthanasia and patient derived xenograft tumors (PDX) were processed the same day for single cell sorting, using the same protocol as for primary human tumors. All animal studies were performed according to Dana-Farber/Harvard Cancer Center Institutional and the Salk Institute Animal Care and Use Committee (IACUC)-approved protocols.

Small animal MRI

MRI experiments were performed on a Bruker BioSpec 7T/30 cm USR horizontal bore Superconducting Magnet System (Bruker Corp., Billerica, MA) equipped with the B-GA12S2 gradient and integrated with up to 2nd order room temperature shim system, which provides a maximum gradient amplitude of 440 mT/m and slew rate of 3440 T/m/s. The Bruker-made 23 mm ID birdcage volume radiofrequency (RF) coil was used for both RF excitation and receiving. The Bruker AutoPac with laser positioning was used for accurate definition of the region of interest. Animals were anesthetized with 1.5% isoflurane mixed in medical air at a flow rate of 2 L/min. Body temperature was maintained at 37°C using a warm air fan. A pressure- transducer for respiratory gating was placed on the abdomen. Animal respiration and temperature were monitored and regulated by the SAII (Sa Instruments Inc., Stony Brook, NY) monitoring and gating system model 1025T. Bruker Paravision 6.0.1 was used for MRI data acquisition. T2-weighted images were obtained by a fast spin echo (RARE) sequence with fat suppression using the following parameters: TR = 6,000 ms, TE = 36 ms, FOV = 19.2 × 19.2 mm2, matrix size = 256 × 192, spatial resolution = 75 × 100 μm2, slice thickness = 0.5 mm, number of slices = 29, rare factor = 15, number of averages = 8, acquisition time 7 min. Images were analyzed and tumor volumes extracted using the semi-automatic segmentation analysis software ClinicalVolumes (ClinicalVolumes, London, UK).

Barcoded lentiviral vector design, construction, and production

A 393 base pair DNA fragment incorporated with 16 × N mixed bases (estimated diversity of $> 4 \times 10^9$) and well-used primer sequences for efficient amplification (CAT-R Fw / LucN Rv / pBABE5′ / EF1a Fw) was synthesized by Integrated DNA Technologies. This DNA fragment was amplified by polymerase chain reaction (PCR), limited for 20 cycles to reduce potential biases introduced during amplification, using Q5 high fidelity polymerase (New England Biolabs) and a set of primers containing either CAT-R Fw or EF1a Fw sequence and 25bp overlapped sequence with a lentiviral vector for Gibson assembly reaction. The amplified PCR product

was confirmed and extracted by agarose gel electrophoresis using QIAquick Gel Extraction Kit (QIAGEN), then purified again with QIAquick PCR Purification Kit (QIAGEN) and cloned by Gibson assembly method (New England Biolabs) into either HrasV12-IRES-GFP-shp53 (Friedmann-Morvinski et al., 2012) or GFP alone digested with EcoRI (New England Biolabs). EcoRI site locates after GFP code region and Woodchuck hepatitis virus Posttranscriptional Regulatory Element (WPRE), and before 3′ Long terminal repeat (LTR) containing viral polyA motif. The assembled vector was amplified using Endura Electrocompetent cells (Lucigen) on LB agar plates limited for 12–14 h to reduce potential biases introduced by the competition between colonies and purified with PureLink HiPure Maxiprep kit (Thermo Fisher Scientific). The lentiviral vector with 16 mixed bases was verified by Sanger sequencing of the corresponding barcode region. VSV-pseudotyped 3rd generation lentiviruses were produced by Lipofectamine 2000 (Thermo Fisher Scientific)-based transfection of 293T cells (5 × 10^{6}cells / 15-cm plate, 20 plates) with a transfer plasmid, packaging plasmids (GAG/POL, RSV-Rev) and an envelope plasmid (VSV-G). 2 μM sodium butyrate was added into medium on the occasion of transfection and medium change to increase the viral production. Transfection efficiency was evaluated based on fluorescence expression. Supernatant was collected 48 and 72 h post-transfection, and lentiviruses were concentrated by the ultracentrifugation. Biological titer of lentiviruses was evaluated with 293T-based on fluorescence expression.

INTRACRANIAL INJECTION OF BARCODED LENTIVIRUS

Lentiviruses were stereotactically injected into the hippocampus of 6- to 16-week-old hGFAP-cre mice (The Jackson Laboratory, Bar Harbor, ME). All mice were maintained under pathogen-free conditions at the Salk Institute, and all procedures performed were approved by the Institutional Animal Care and Use Committee. Lentiviruses (1 × 10^5 IU) suspended in 1uL PBS were loaded on a 33-gauge Hamilton syringe, and injected slowly (0.1uL/30sec-1min) using the following coordinates: 2.0 mm posterior, 1.5 mm lateral, and 2.3 mm dorsal to the bregma. Upon completing injection, the needle was left in place for 3 min, then withdrawn slowly to help reduce virus reflux in 2 min.

In vitro labeling of patient derived cells with barcoded lentiviruses

Patient derived cells (MGH143, MGG23) were incubated with serially diluted barcoded lentiviruses for 12 h in Neurobasal Medium (GIBCO 21103-049) supplemented with 1X N2/B27 (GIBCO), 1% Penicillin/Streptomycin (GIBCO), 1X Glutamax (GIBCO), 20 ng/mL EGF and 20 ng/mL bFGF (FGF2). Barcoded cells were washed three times with PBS, and then dissociated with pre-warmed TrypLE Express (GIBCO) to prepare single-cell suspension (a range of 2x10^4-1x10^5 cells per mouse) for intracranial injection. Remaining cells were further cultured for 48 h to evaluate GFP expression and lentiviral infection efficiency based on flowcytometric analysis.

Fluorescence-activated cell sorting of GFP positive mouse and human GBM cells

All mice were perfused with ice-cold PBS after euthanasia. The collected brains were mechanically and enzymatically dissociated using a papain-based brain tumor dissociation kit (Miltenyi Biotec) supplemented with 0.1% typel collagenase (Thermo Fisher Scientific) / PBS. Cells were first stained with calcein Blue AM (Life Technologies) and Zombie NIR (BioLegend) for 30 min at 4°C, and with anti-mouse CD16/32 (BD Biosciences) for 5min. After washing cells with ice-cold 2% FBS / PBS, cells were then stained with anti-mouse CD45-PerCP (clone 30-F11, BD Biosciences) for 30 min at 4°C. Sorting was performed with The Becton Dickinson Influx cytometer (Becton Dickinson) using 640 nm (Zombie NIR, 750LP filter), 355 nm (Calcein Blue AM, 460/50 filter), 488 nm (CD45-PerCP, 692/40 filter) and 488 nm (GFP, 530/40 filter) lasers. Side scatter (SSC) width versus forward scatter (FSC) area, and Trigger Pulse Width versus FSC criteria were used to discriminate doublets and gate only singleton cells. Viable single cells were identified as calcein blue AM positive and Zombine NIR negative. We sorted viable CD45 negative/GFP positive single-cells into 96-well plates containing 5uL of TCL buffer (QIAGEN) with 1% beta-mercaptoethanol. Plates were frozen immediately after sorting and stored at −80°C prior to whole transcriptome amplification, library preparation and sequencing.

In vitro overexpression experiments

NPCs were established from mouse ES cells (V6.5) using previously published protocols (Kerman et al., 2015). NPCs were propagated with DMEM:F12 (GIBCO 11320-033 with L-glutamine/Sodium Bicarbonate) supplemented with 1X N2/B27, 1% Penicillin/Streptomycin, 1 μg/mL laminin (from EHS tumor), 20 ng/mL EGF and 20 ng/mL bFGF (FGF2). NPCs were cultured on poly-l-ornithine coated tissue culture treated dishes. Astrocytes were derived and propagated from NPCs using NPC media supplemented with 4% Fetal Bovine Serum (FBS). Constructs utilized: piggyback plasmid with CMV promoter driving the expression of T2A-eGFP (empty vector), EGFR (human)-T2A-eGFP, or CDK4(mouse)-T2A-eGFP. All constructs were validated by Sanger sequencing in addition to whole plasmid next-generation sequencing. Cell proliferation was measured in 96 well plates (2,000 cells per well) using ATPlite as per the manufacturer's instructions. Cells were lysed on day 0 (1 h post-plating), day 2, and day 4. Quantification was normalized to day 0 data.

Imaging of GFAP positive cells in mouse NPC cultures

NPCs engineered to express eGFP, EGFR, or CDK4 were plated in 8-chamber glass slides (BD biosciences) at a density of 80,000 cells per well. Cells were fixed using 4% formaldehyde, permeated with 0.5% Triton X-100 for 10 min at 4°C, blocked using PBS

supplemented with 10% normal goat serum, and then stained overnight with a GFAP antibody (Dako, Z0334) at a dilution of 1:2500. Cells were then washed with PBS and stained with a secondary goat anti-rabbit conjugated with Alexa-555 (1:500 dilution in blocking buffer). Nuclei were stained with Hoechst 33342 (1:10,000 in PBS). Slides were then mounted with Vectashield mounting media and imaged in a Zeiss LSM 800 confocal microscope. Maximum intensity images were then assembled on imageJ. The Find Maxima tool on imageJ was then used to count GFAP positive cells with noise tolerance set to eliminate background positive signal.

QUANTIFICATION AND STATISTICAL ANALYSIS

Single-cell RNA-seq data generation and processing

Smart-seq2 whole transcriptome amplification, library construction, and sequencing were performed as previously published (Filbin et al., 2018; Picelli et al., 2014; Tirosh et al., 2016b; Venteicher et al., 2017). As quality control, we examined the number of genes detected in each cell (Figure S1B). We observed a bimodal distribution and conservatively excluded 28% of the sequenced cells with fewer than 3,000 detected genes. Among the remaining cells, we detected on average 5,730 genes per cell, highlighting the high quality of our scRNA-seq dataset. Expression levels were quantified as $E_{i,j} = \log_2(TPM_{i,j}/10+1)$, where $TPM_{i,j}$ refers to transcript-per-million for gene *i* in sample *j*, as calculated by RSEM (Li and Dewey, 2011). TPM values were divided by 10 since we estimate the complexity of single cell libraries in the order of 100,000 transcripts and would like to avoid counting each transcript ~10 times, as would be the case with TPM, which may inflate the difference between the expression level of a gene in cells in which the gene is detected and those in which it is not detected. For the remaining cells, we calculated the aggregate expression of each gene as $E_a(i) = \log_2(\text{average}(TPM_{i,1...n})+1)$, and defined the set of analyzed genes as those with $E_a > 4$. We then defined relative expression over the remaining cells and the analyzed genes, by centering the expression levels per gene, $Er_{i,j} = E_{i,j}$-average$[E_{i,1...n}]$. For a subset of samples, single cells were processed through the 10X Chromium 3′ Single Cell Platform using the Chromium Single Cell 3′ Library, Gel Bead and Chip Kits (10X Genomics, Pleasanton, CA), following the manufacturer's protocol. Briefly, 7,000 cells were added to each channel of a chip to be partitioned into Gel Beads in Emulsion (GEMs) in the Chromium instrument, followed by cell lysis and barcoded reverse transcription of RNA in the droplets. Breaking of the emulsion was followed by amplification, fragmentation, and addition of adaptor and sample index.

tSNE analysis and identification of non-malignant cell types

Relative expression values were used to classify all cells passing quality control by tSNE, using the MATLAB implementation tsne, with default parameters (Figure 1B). Three small clusters were apparent, which were associated with high expression of markers for three non-malignant cell types. We thus defined sets of marker genes for each of those cell types and scored each cell by their average expression. For macrophages: *CD14*, *AIF1*, *FCER1G*, *FCGR3A*, *TYROBP, CSF1R. For T cells: CD2*, *CD3D*, *CD3E*, *CD3G*. For oligodendrocytes: *MBP*, *TF*, *PLP1*, *MAG*, *MOG*, *CLDN11*. Cells were classified to each of these cell types by scores above 4. A second tSNE analysis was performed only for malignant cells and with "NumPCAComponents" equal to 30 (Figure 1C).

Definition of single-cell gene signature scores

Given a set of genes (G_j) reflecting an expression signature of a specific cell type or biological function, we calculate for each cell *i*, a score, $SC_j(i)$, quantifying the relative expression of G_j in cell *i*, as the average relative expression (*Er*) of the genes in G_j, compared to the average relative expression of a control gene-set (G_j^{cont}): $SC_j(i) = \text{average}[Er(G_j,i)] - \text{average}[Er(G_j^{cont},i)]$. The control gene-set is defined by first binning all analyzed genes into 30 bins of aggregate expression levels (*Ea*) and then, for each gene in the gene-set G_j, randomly selecting 100 genes from the same expression bin. In this way, the control gene-set has a comparable distribution of expression levels to that of G_j, and the control gene set is 100-fold larger, such that its average expression is analogous to averaging over 100 randomly selected gene-sets of the same size as the considered gene-set.

CNA inference from single-cell data

CNAs were estimated by sorting the analyzed genes by their chromosomal location and applying a moving average to the relative expression values, with a sliding window of 100 genes within each chromosome, as we have previously described (Patel et al., 2014; Tirosh et al., 2016b). Cells classified to each of the non-malignant cell types were used to define a baseline of normal karyotype, such that their average CNA value was subtracted from all cells. We then scored each cell for two CNA-based measures. "CNA signal" reflects the overall extent of CNAs, defined as the mean of the squares of CNA values across the genome. "CNA correlation" refers to the correlation between the CNA profile of each cell and the average CNA profile of all cells from the corresponding tumor, except for those classified by gene expression as non-malignant. Cells were then classified as malignant by CNA analysis if they had CNA signal above 0.02 and CNA correlation above 0.4 (Figure S1C).

Integrated definition of malignant cells

We then combined the CNA classification with the tSNE-based and the gene-set based classifications, such that the final list of malignant cells included those which were classified as malignant by CNA, were part of the malignant tSNE cluster, and were not classified to any of the non-malignant cell types based on the marker gene-sets. Similarly, cells were classified to each non-malignant cell type only if these assignments were concordant among the three analyses.

Identification of intra-tumor variability programs using hierarchical clustering

First, we used average linkage hierarchical clustering of the single malignant cells from each of the tumors separately, using one minus the Pearson correlation (across all analyzed genes) as the distance metric. In order to select clusters without a pre-defined strict threshold on the number of clusters or their level in the hierarchical tree, we first recovered all potential clusters, and then excluded them by size, by their signal for differential expression, and by their redundancy with other clusters, in the following ways: (1) We excluded clusters that consist of less than 5 cells or more than 80% of the malignant cells in the respective tumor. (2) For each cluster, we estimated the number of preferentially expressed genes: we identified all genes with 3-fold higher average expression in the cluster than in all other malignant cells from the same tumor, and corresponding p values below 0.05 (using a t test and corrected for False Discovery Rate with the Benjamini-Hochberg method). We then count the number of significant genes with adjusted p values below 0.05 (N_{sig1}) and below 0.005 (N_{sig2}), respectively. All clusters with both $N_{sig1} > 50$ and $N_{sig2} > 10$ were defined as having sufficient signal of differential expression and retained for further analysis. (3) For each pair of clusters with Jaccard index above 75%, we excluded the cluster with lower N_{sig1}. Applying this approach to 27 tumors revealed 479 clusters, including (as desired) many cases of both a large cluster and its smaller sub-clusters. Finally, we used the differentially expressed genes (N_{sig1}) as each cluster's signature, yielding 479 signatures.

Integration of individual signatures into meta-modules

Jaccard indices, reflecting the overlap between pairs of signatures, were used for hierarchical clustering of the signatures by average linkage. Four groups of signatures were identified, of which two had robust separation into two subgroups (Figure 3), resulting in six groups of signatures which were used as the basis for defining six meta-modules. For each group of signatures, we defined the meta-module based on the average expression $\log_2$-ratios, across the corresponding signatures: for each signature, an expression log-ratio was defined by comparing all the cells in the corresponding potential cluster to all other malignant cells in the same tumor. These log-ratios were then averaged across all signatures that constitute a group (or subgroup), which in each of the cases included at least six different tumors. Each meta-module was then defined as all genes whose average log-ratio was above 2 and was restricted to the 50 genes with highest log-ratios for that group of programs.

Identification of cycling cells

Meta-modules were also defined for the G1/S and G2/M phases of the cell cycle, from analysis of the cell-cycle related signatures. Next, the cell scores for these meta-modules were used to classify cells as cycling or non-cycling (Figures S3A and S3B). For each of the two cell cycle scores, the distribution of scores for all malignant cells was fitted to a normal distribution and a threshold of $p < 0.001$ was used for distinguishing cycling cells.

Assignment of cells to meta-modules and their hybrids

Malignant cells were first assigned to the meta-module with the highest score, including the six meta-modules (MES1-like, MES2-like, NPC1-like, NPC2-like, AC-like, OPC-like) but excluding the cell cycle meta-modules. For most analyses, we collapsed the MES1 and MES2 groups of cells into one group of MES-like cells, and similarly, the NPC1 and NPC2 cells into one group of NPC-like cells. Next, we defined hybrids as those that also had a high score for a second meta-module (not including the MES1/MES2 or NPC1/NPC2 distinction) by three criteria: (1) the score for the second meta-module was higher than 1. (2) The score for the second meta-module was higher than that of 10% of the cells that map to this meta-module (as their top-scoring meta-module). (3) The difference in score between the second meta-module and the third meta-module was at least 0.3.

The percentages and patterns of hybrids were largely unchanged when we used different criteria. An ''expected number'' of hybrids for each pair of meta-modules (Figure 3D) was defined by shuffling the meta-module scores of cells in each tumor. Each meta-module was shuffled independently such that any relationship between the meta-modules was eliminated while the distribution of scores was unchanged as were the differences in distribution between tumors. This shuffling was performed 100 times and in each case we used the criteria defined above to count the number of hybrids. The mean and standard deviations of these counts were then used as a control for the expected number of hybrids.

Identification of genetic subclones by inferred CNAs

In each cell, we defined the average inferred CNA values for each chromosome, or chromosome arm. Next, we examined for each tumor if one or more of those chromosome arms have a bimodal distribution of CNA values among malignant cells. We fitted the CNA values to a bimodal Gaussian distribution using MATLAB's fitgmdist function and examined the posterior probabilities of cells to the two modes. In most tumors and for most chromosome arms, the two modes were highly similar and cells were not confidently assigned to distinct modes. However, in particular cases (specific chromosome arms in specific tumors) the two modes were highly distinct such that most cells could be confidently assigned to only one mode. In those cases, we defined subclones whenever > 80% of the malignant cells in a tumor had a posterior probability higher than 0.95 for one of the two modes, and at least 10 cells were assigned to each of the two modes. The few cells which were not confidently assigned to any mode were excluded from the subclonal analysis. If a tumor had only one chromosome (or chromosome arm) with bimodal distribution, we defined two clones corresponding to the two modes. If a tumor had multiple such chromosomes then we considered all combinations of modes with at least three cells as subclones.

Characterization of meta-modules by comparison to external data

We characterized the meta-modules by four complementary approaches. (1) First, by the relative expression scores of non-malignant cell types for each of the meta-modules (Figure 2). To this end, we obtained scRNA-seq data for non-malignant brain cells from multiple sources (Darmanis et al., 2015; Nowakowski et al., 2017; Pollen et al., 2015; Tirosh et al., 2016b). For each source, we aggregated cells by their cell type classification, defined the average expression profile of each cell type (or used the respective data generated by the original study) and finally defined relative expression for cell types by subtracting the mean expression of all cell types. (2) Second, by the global expression similarities (Pearson correlations) between 47 non-malignant cell types in the developing human cortex (Nowakowski et al., 2017) and the average profiles of 250 randomly sampled glioblastoma cells mapping to each of the meta-modules (Figure S2F). (3) Third, by enrichment of meta-module genes in non-malignant cell types (Figure S2F) (Nowakowski et al., 2017). Enrichment was defined as the significance level for the fraction of meta-module genes most highly expressed in a given cell type (-$\log_{10}$(P value)) and calculated using a hypergeometric test. (4) We similarly tested for enrichment of the meta-modules in C2 and C5 gene-sets (N = 10,679) from MSigDB (Subramanian et al., 2005) using the clusterProfiler::enricher function in R (Figure S2E).

Two-dimensional representation of malignant cellular states

Cells were first separated into OPC/NPC versus AC/MES by the sign of $D = max(SC_{opc}, SC_{npc}) - max(SC_{ac}, SC_{mes})$, and D defined the y axis of all cells. Next, for OPC/NPC cells (i.e., $D > 0$), the x axis value was defined as $\log_2(|SC_{opc} - SC_{npc}|+1)$ and for AC/MES cells (i.e., $D < 0$), the x axis was defined as $log_2(|SC_{ac} - SC_{mes}|)$. To visualize the enrichment of subsets of cells (e.g., cycling cells, Figure 3F) across the two-dimensional representation, we calculated for each cell the fraction of cells that belong to the respective subset among its 100 nearest neighbors, as defined by Euclidean distance, and these fractions were displayed by colors.

Bulk scores defined for TCGA samples

Expression data from TCGA samples was based on the agilent microarray platform, as these had the highest number of profiled samples. Expression scoring of bulk samples for meta-modules were done as described above for single cells, with two exceptions. (1) The expression of genes in bulk samples reflects the combined effect from multiple expressing cell types and therefore many genes, which are good markers for a particular cellular state in single cell data may not be good markers in bulk data. To exclude such genes, we first defined initial bulk scores by the average expression of meta-modules. Next, we calculated the correlation of each meta-module gene with the initial scores. Genes were excluded if their correlation was below 0.4 or if the correlation was higher for a different meta-module. The remaining genes were then used to define refined bulk scores. (2) Genes which are not part of the meta-modules but were found to be associated with high frequency of the meta-module (Figures S5D and S5E) were included in this analysis.

Association of bulk scores with CNAs

Chromosomal copy-number losses, gains, as well as high-level amplifications, were obtained from TCGA (Brennan et al., 2013). For each gene with at least 10 tumors that have a specific CNA pattern (gain, loss or high-level amplification) t test was used to compare the bulk scores between all tumors with and those without that. significance values corresponding to $-\log_{10}$(P), where P is the t test p value, are shown in Figures 5B and S5G and were above 5 for all genetic events that are noted in the text (*EGFR*, *PDGFRA*, *CDK4* amplifications and chr5q deletion). We further examined the fraction of tumors with each of those events (as well as with downregulation of NF1) in the subset of tumors with highest bulk scores for each meta-module (excluding tumors in which all scores are below 1) and found a significant enrichment ($p < 0.001$, hypergeometric test) for each of the positive associations (Figure S5F).

Assignment of TCGA subtypes to tumors profiled by scRNA-seq

We simulated bulk expression levels of each tumor as $E_{i,J} = \log_2(TPM_{i,J}+1)$, where J refers to all malignant cells in that tumor. The resulting bulk profiles were subsequently scored for three TCGA subtypes (Verhaak et al., 2010) and assigned to their highest scoring subtype or to a "mixed" category if the difference in score between the first and second subtypes was less than 0.05.

Integration of the 10X Genomics data

Processing and analysis of a second dataset generated by the 10x Genomics platform was performed as stated above for the SMART-Seq2 data, with the following exceptions: Preprocessing: (i) Due to larger variability in the number of detected genes between tumors compared to the SMART-Seq2 data, we excluded all sequenced cells whose number of detected genes was less than half or more than twice the mean number of genes detected across cells coming from the same tumor (26% of all sequenced cells). (ii) Identification of non-malignant cells: Non-malignant cells were identified by either one or both of two methods. (1) We defined sets of marker genes for normal cell types (see earlier STAR Methods) and scored all cells for their average expression of each signature. Bimodal distributions were observed for macrophage and oligodendrocyte scores. Accordingly, cutoffs of > = 0.5 and > = 3 were chosen in order to define non-malignant cells. (2) All cells were hierarchically clustered with average linkage and pairwise Pearson correlations as a distance metric. Three large clusters were identified, and of these one was associated with high expression of markers for non-malignant cells, specifically macrophages. 100% of the cells in this cluster were captured by the method described in #1. In total, 40% of cells passing quality control were non-cancer cells as defined by both methods.

(iii) Generation of gene signatures: We analyzed 9,870 cancer cells from 9 tumors to identify coherent signatures of differential gene expression in our data. 577 signatures were defined and compared to one another by pairwise Jaccard overlaps, as before. Hierarchical clustering of the overlaps revealed a strong meta-module for cell-cycle signatures. The remaining 78% of non-cycling signatures were compared to those generated by the SMART-Seq2 platform (and discussed in the main text).

Comparison of 10x and SMART-Seq2 results

448 non-cycling signatures from the 10x data were hierarchically clustered with average linkage and using the signatures' pairwise Jaccard overlaps as distance metric. The clustered signatures were subsequently scored for their correspondence to the six meta-modules in the main text (NPC1, NPC2, OPC, AC, MES1, and MES2), based on the expression of the cells in the corresponding cluster compared to all other malignant cells in the same tumor, with scores defined as described above (see Definition of single-cell gene signature scores).

Analysis of barcoded cells

Barcodes used in this study include unique (i.e., variable) 16-nucleotide sequences that were identifiable by two common flanking 29-nucleotide sequences (see STAR Methods). To assign cells to barcodes, we first searched for the flanking sequences in scRNA-seq reads to locate the barcode sequences. Barcodes were then assigned to cells if counted at least three times in the cell, or at least three times more than any other barcode. Unassigned cells were excluded from downstream analyses. Next, cells with detected barcodes were assigned to the meta-module with the highest score if the score was at least 1 and the difference to the next highest score was at least 0.5 (see above for definition of single-cell scores). Cells not assigned to meta-modules by these criteria were also excluded. Of the cells retained, 53% and 78% belonged to barcoded populations with at least two cells (Figures 7C and 7D and Figures 7E and 7F, respectively).

DATA AND CODE AVAILABILITY

Data generated for this study are available through the Broad Institute Single-Cell Portal. (https://portals.broadinstitute.org/single_cell/study/SCP393/single-cell-rna-seq-of-adult-and-pediatric-glioblastoma) and the Gene Expression Omnibus (GEO: GSE131928). The Code supporting the current study is available from the corresponding author on request.

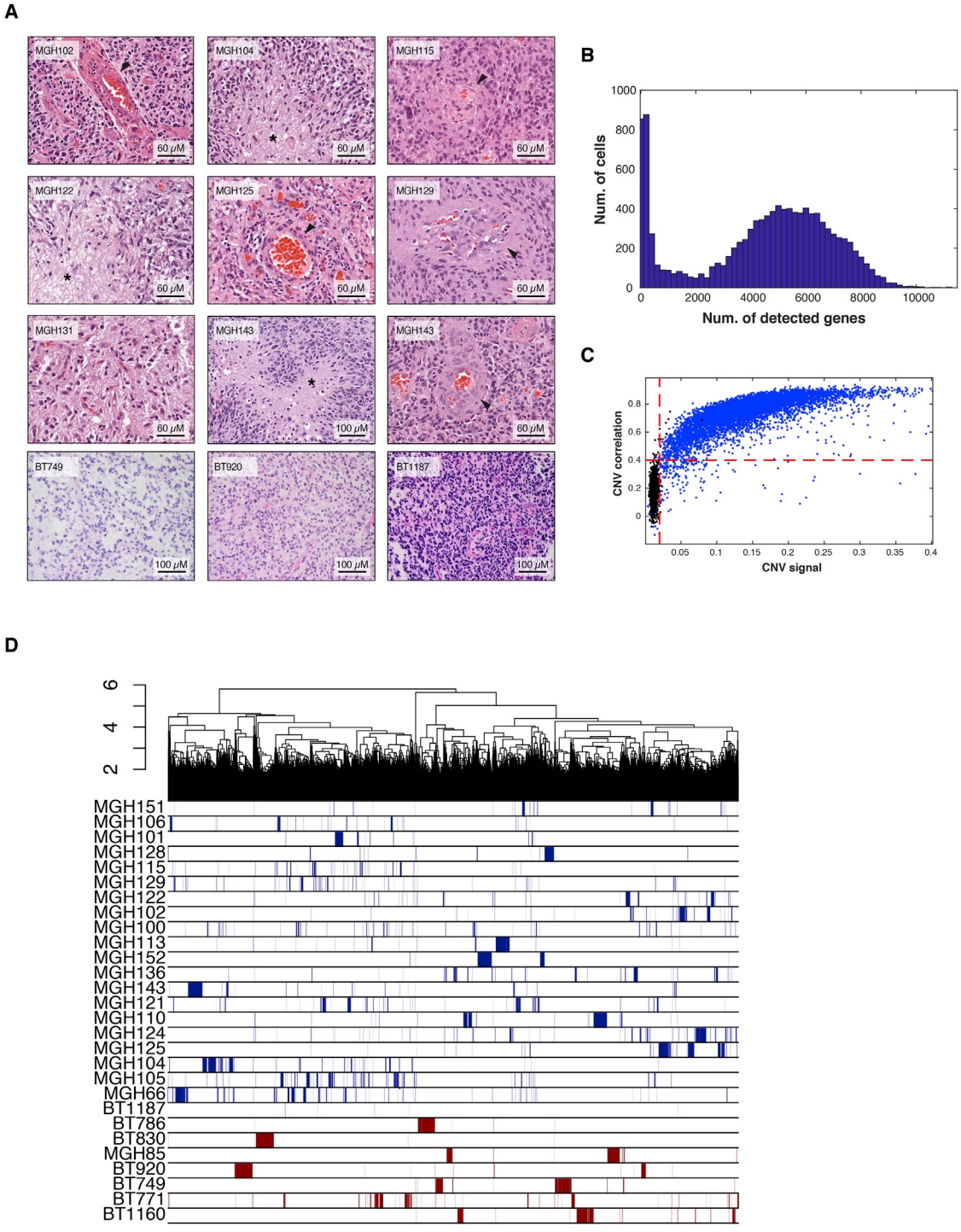


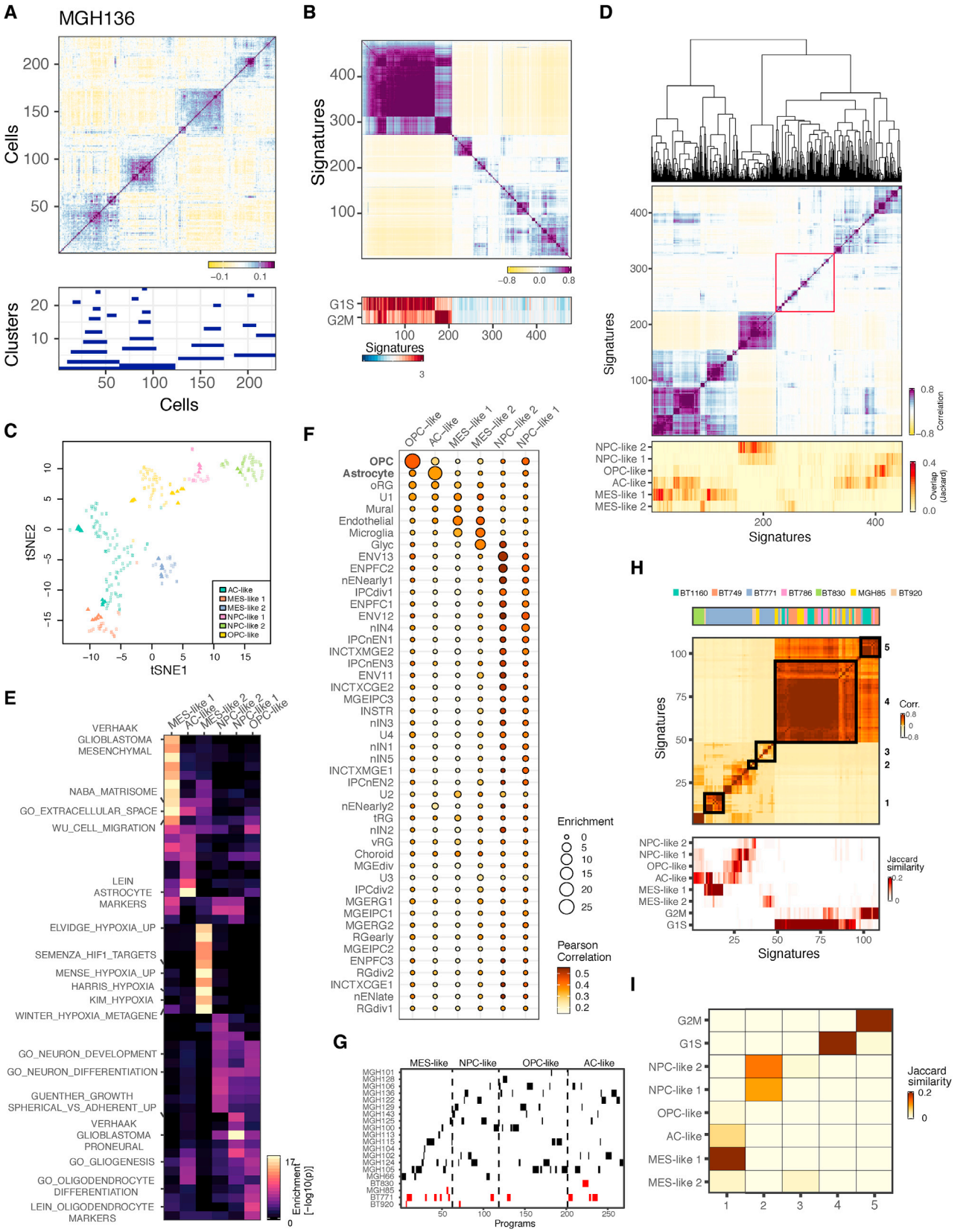
Figure S1. Related to Figure 1

(A) H&E stain of a representative subset of glioblastomas in our cohort. Images show features of high-grade glioma with important pleomorphism and cyto-nuclear atypia. Arrowheads highlight vascular proliferation, stars highlight areas of necrosis, both hallmark features of glioblastoma.

(B) Distribution of the number of genes detected in each of the sequenced cells.

(C) Classification of cells (dots) to malignant and non-malignant based on CNA signal (x axis) and CNA correlation (y axis), based on thresholds indicated by red dashed lines. CNA signal reflect the extent of CNAs while CNA correlations reflect the similarity between the CNA pattern of a single cell and that of other malignant cells from the same tumor (see Method Details). Cells mapping to non-malignant cell types are shown in black while the rest are shown in blue.

(D) Top: Hierarchical clustering of all malignant cells by their expression profiles. Bottom: Assignment of cells to tumors. Adult and pediatric tumors are highlighted in blue and red, respectively.



(legend on next page)

Figure S2. Related to Figure 2

(A) Top: cell-to-cell correlation matrix of malignant cells from MGH136 ordered by hierarchical clustering. Bottom: assignment of cells to potential clusters.

(B) Top: Hierarchical clustering of 479 potential clusters defined from analysis of 27 tumors. Bottom: expression scores of programs for the G1/S and G2/M cell cycle signatures.

(C) tSNE plot of all non-cycling signatures clustered by their cell scores and colored by assignment into meta-modules. Circles and triangles signify signatures derived from adult and pediatric tumors, respectively.

(D) 10x-derived expression signatures are consistent with the meta-modules derived from Smart-Seq2 analysis. Non-cycling signatures derived from the 10x dataset were clustered hierarchically according to their pairwise correlations (shown at the top), and their similarity to the six meta-modules are evaluated by Jaccard indices (shown in the lower) reflecting the fraction of overlapping genes. This analysis demonstrates that most 10x signatures are part of apparent clusters which are consistent with the meta-modules defined by the Smart-Seq2 analysis. An exception is a subset of signatures highlighted by a red square. These additional signatures are characterized by weak correlations with one another and with all other signatures (except for those derived from overlapping clusters of cells) and therefore do not constitute a recurrent module. Furthermore, they are primarily associated with high expression of either ribosomal protein genes or hemoglobin and are otherwise not associated with coherent functional annotations. We therefore considered that these signatures primarily reflect technical confounders.

(E) Functional enrichment analysis of meta-modules across C2 and C5 gene-sets in MSigDB (Subramanian et al., 2005). The top ten gene-sets for each meta-module are shown in the heatmap (see also Table S3) and are ordered by hierarchical clustering.

(F) Meta-module similarities to neurodevelopmental cell types, profiled by scRNA-seq (Nowakowski et al., 2017), are shown by two complementary measures: Colors indicate the correlations of cell types and meta-modules by their global expression values; Circle sizes indicate the significance levels for the enrichment of meta-module genes among those most highly expressed in a given cell type ($-\log_{10}$(P value)). Shown in bold are cell types that were uniquely ascribed to a meta-module as defined by an enrichment level at least two-fold higher than remaining cell types and a correlation value among the top three.

(G) Assignment of signatures to tumors (adult in black; pediatric in red). Signatures were ordered by the hierarchical clustering pattern shown in Figure 2B into the four meta-modules, which are separated by dashed lines.

(H and I) Meta-modules identified when restricting the analysis shown in Figure 2B to signatures from pediatric tumors. (H) Middle: Hierarchical clustering of 109 signatures defined by analysis of 7 pediatric tumors. Black squares denote five potential groups of pediatric-only signatures that are derived from multiple tumors. Top: Assignments of signatures to tumors. Bottom: signature similarities to the six meta-modules. (I) Jaccard similarities of the six meta-modules from the main analysis (y axis; as shown in Figure 2B) with the five pediatric-only meta-modules (x axis) which were derived from the groups of signatures defined in (H), and numbered by their position from left to right.

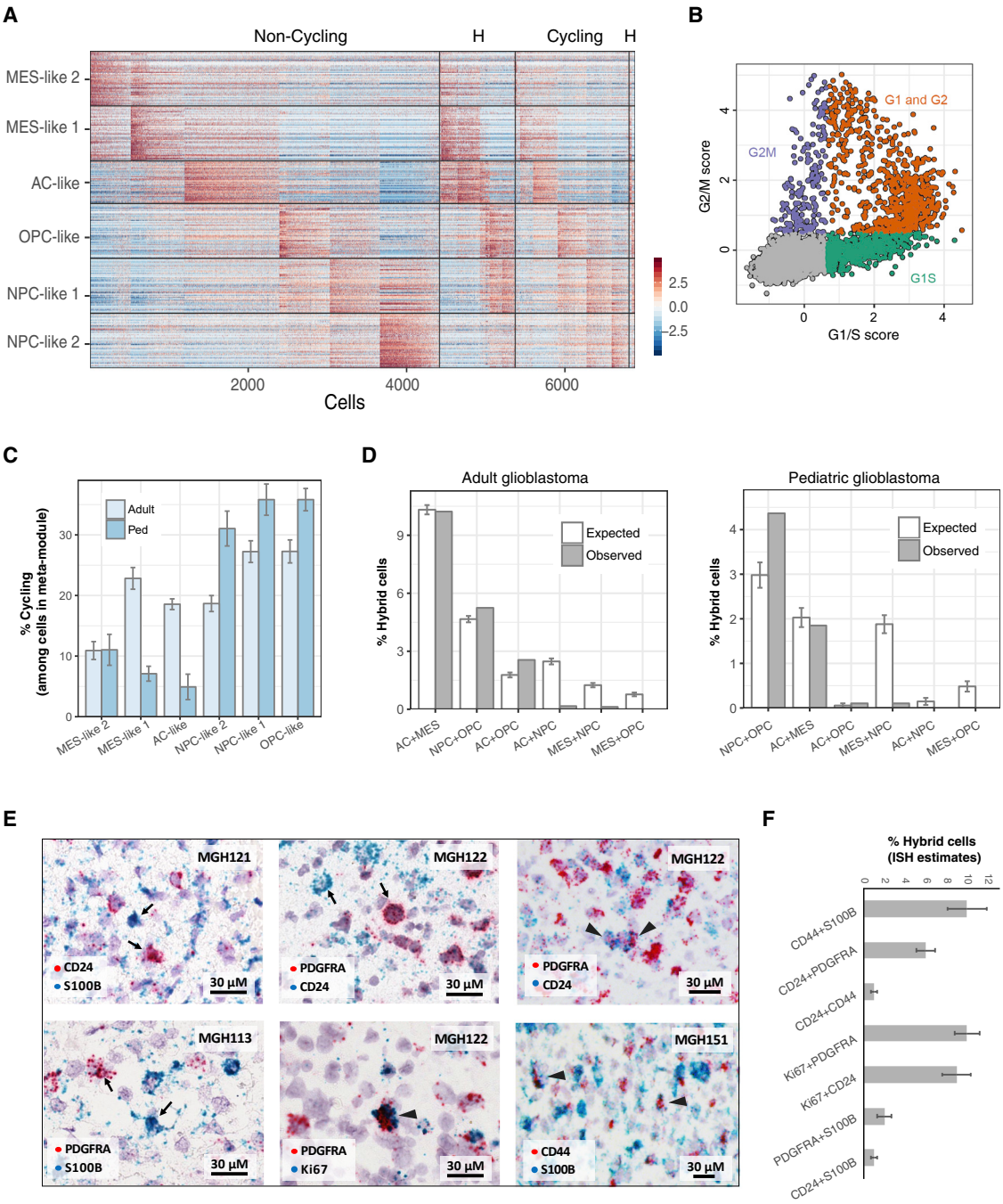


Figure S3. Related to Figure 3

(A) Heatmap showing meta-module expression, with rows corresponding to all genes in the meta-modules, and columns corresponding to all malignant cells, separated to non-cycling (left) and cycling cells (right). Within each group, the cells are ordered by their maximal score, for cells mapping to one meta-module, followed by cells mapping to two meta-modules (H, hybrid states).

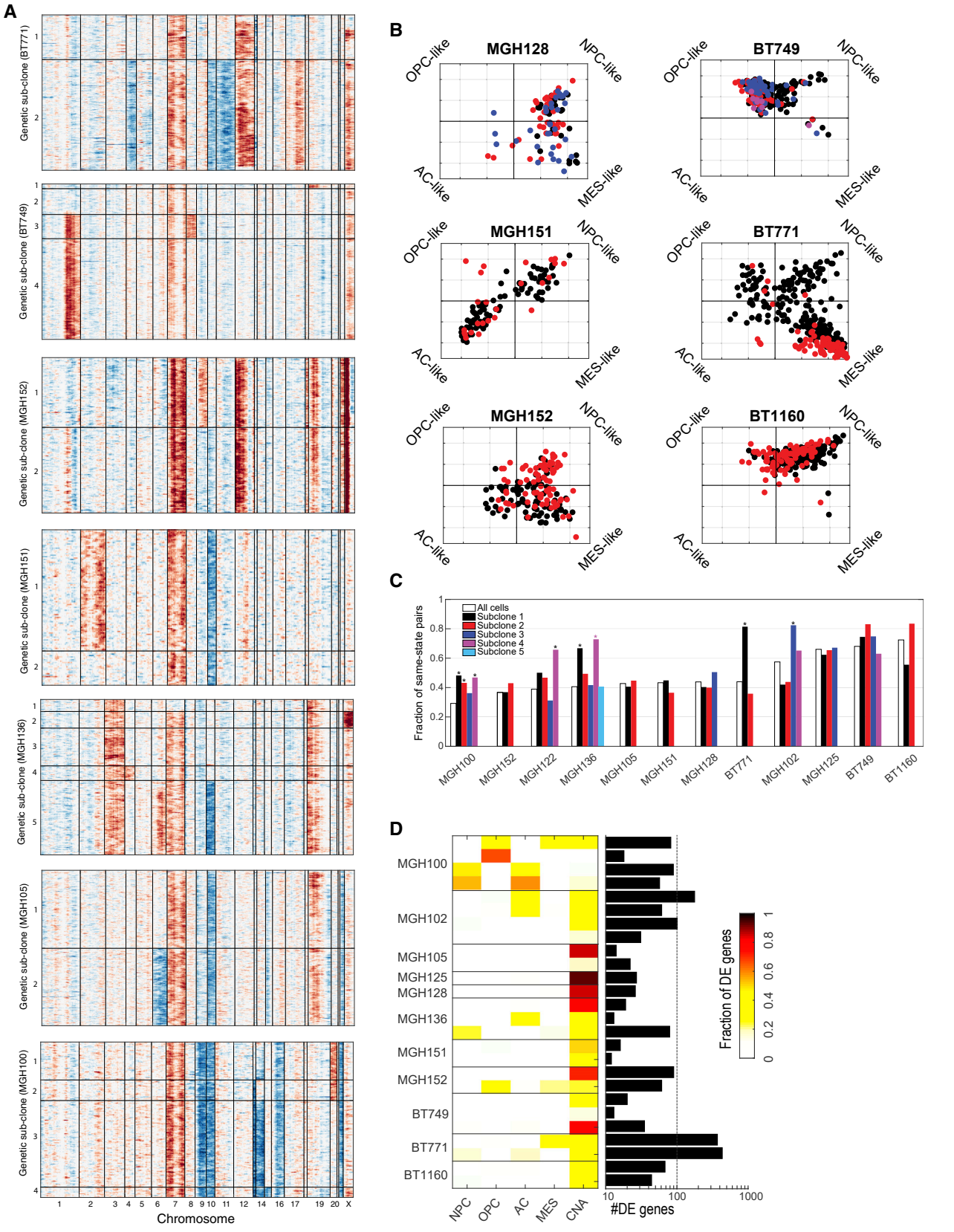
(B) Identification of cycling cells. Cell scores for the G1/S and G2/M signatures are shown for all malignant cells. Cells defined as cycling are colored as green (G1/S), purple (G2/M) or red (both).

(C) Bar plot showing the percentage of cycling cells among cells with highest score for each of the meta-modules. Adult and pediatric tumors are separated in order to demonstrate their distinct distributions. Error bars correspond to standard error, calculated by bootstrapping.

(D) Bar plots showing the percentage of hybrid cells (co-expressing two distinct meta-modules) out of all malignant cells in adult (left) and pediatric (right) glioblastoma samples. Expected percentages of hybrid cells (assuming the scores for meta-modules are independent of one another) were defined as those found after independently shuffling the cell scores for each meta-module; error bars correspond to standard error, calculated by shuffling the cell scores 100 times.

(legend continued on next page)

(E) *In situ* RNA hybridization of glioblastoma for NPC-like (*CD24*), OPC-like (*PDGFRA*), AC-like (S100B) and proliferation (*Ki67*) markers. Arrows in each panel highlight representative positive cells for respective markers. Arrowheads highlight co-expression of *PDGFRA* and *Ki67*.
(F) Quantification of the percentage of cells co-expressing markers (hybrid states, cycling cells) by RNA ISH across ten glioblastoma specimen. Error bars correspond to standard deviation across tumors.



(legend on next page)

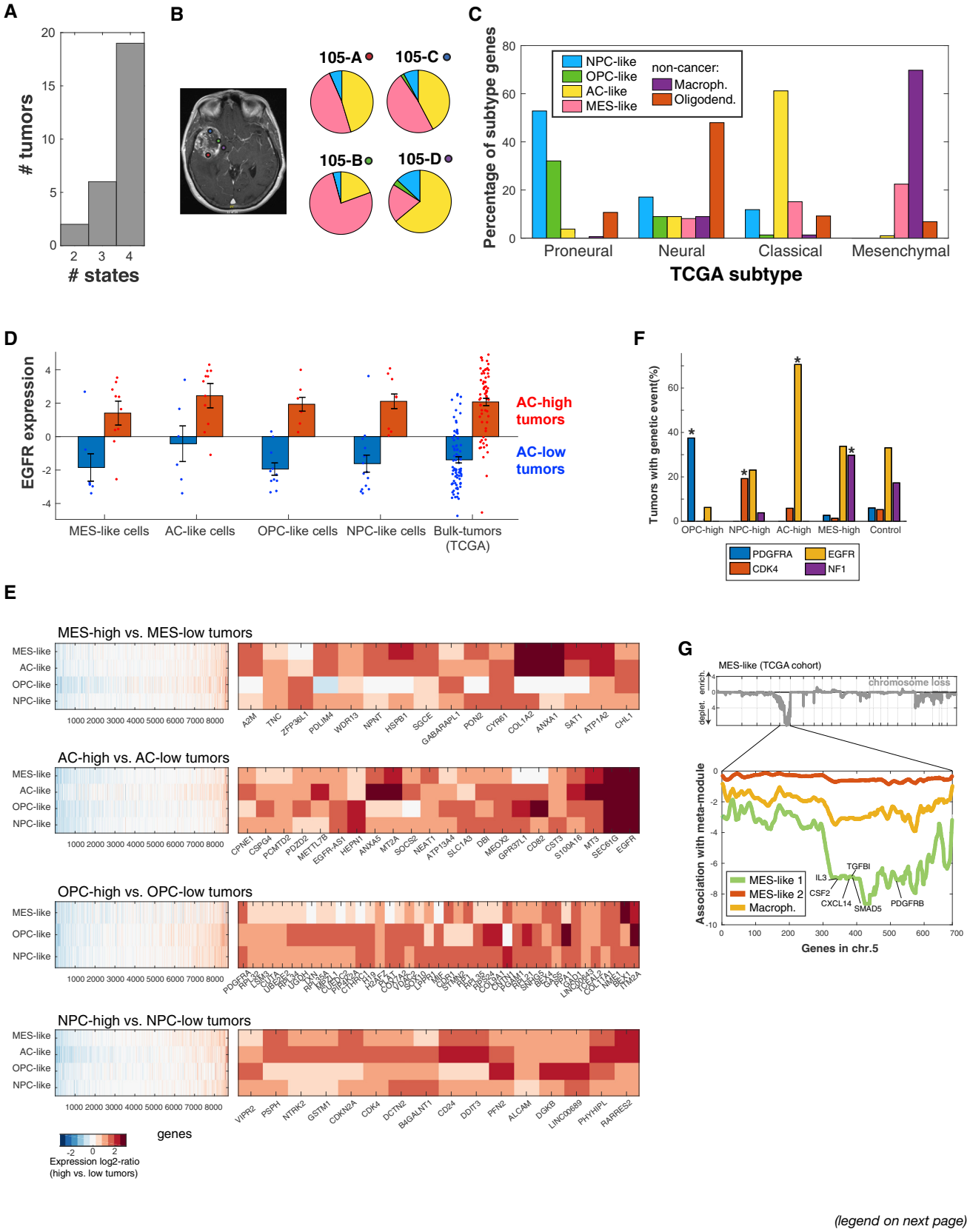
Figure S4. Related to Figure 4

(A) Identification of genetic subclones by CNAs. Shown are the inferred CNAs of malignant cells in (from top to bottom) BT771, BT749, MGH152, MGH151, MGH136, MGH105, and MGH100, separated into genetic subclones based on the amplifications/deletions of specific chromosomes.

(B) Cell-state plots (as in Figure 3F) for six tumors with CNA-based subclones. Cells are colored by their subclone assignments (see color legend in C).

(C) Fraction of cell-pairs that map to the same state (out of four states represented by quadrants of B), among all cells from each tumor (white), and among cells from individual subclones (as defined by color legend). Eight out of 37 subclones had a significantly high rate of same-state cell-pairs (defined as $p < 0.05$ by a permutation test) and these are highlighted by asterisks.

(D) Analysis of differentially expressed (DE) genes between subclones, comparing each individual subclone to the other subclones in the same tumor. Heatmap shows the fraction of DE genes that correlate (Pearson $R > 0.3$) with each of the meta-modules or that are located within CNA loci that distinguished the subclones. Black bar to the right indicates the number of DE genes.



(legend on next page)

Figure S5. Related to Figure 5

(A) For each tumor we counted the number of distinct cellular states (out of four) which were detected. Shown are the number of tumors with two to four states (none of the tumors had less than two states).

(B) Left: MRI image of MGH 105 with colored dots for areas sampled. Right: Pie charts (as in figure 5A) for the four spatial regions of MGH105.

(C) Linking TCGA subtypes to cellular subpopulations. Marker genes for each of the four TCGA subtypes were classified to one of six cellular programs based on the subpopulation of cells in which it had the highest expression level in our scRNA-seq data. The six programs correspond to the four malignant states (as defined by the meta-modules) and two non-malignant cell types: macrophages and oligodendrocytes (see Figure 1). Shown are the percentage of genes for each TCGA subtype that are classified to each of the six programs.

(D) Identifying genes associated with enrichment of particular cellular states. For each of the four malignant cellular states we defined tumors with high and low fractions of cells and examined the differential expression between them. This was done separately for cells in each cellular state (rows in each panel) to control for differences in tumor composition. Shown are differential expression (log_2-ratio for high versus low tumors) for all genes (left) and for genes significant in at least two cellular states (right), with genes ranked by average differential expression.

(E) EGFR expression is higher in AC-high than AC-low tumors, both in single cells and in bulk tumors. For each of the four cellular states, the average relative EGFR expression in all cells in that state is shown for AC-low tumors (blue dots) and AC-high tumors (red dots). Bars and error-bars show the mean and standard error across the two subsets of tumors. The rightmost pair of bars show EGFR expression in bulk tumors of TCGA divided by bulk AC-like scores into AC-high and AC-low tumors (see STAR Methods).

(F) Enrichment of genetic events in TCGA tumors with high bulk scores for particular meta-modules. For each of the four meta-module signatures we defined a subset of TCGA tumors with high bulk scores and examined the fraction of tumors with three high-level amplifications (PDGFRA, CDK4 and EGFR) and with downregulation of NF1. As a control, we compared these proportions to all TCGA glioblastomas. Asterisks indicate significant enrichment ($p < 0.01$, hyper-geometric test).

(G) Top panel: Association of chromosomal losses with TCGA bulk scores for the MES-like state, as shown for chromosomal gains in Figure 5B. Bottom: zoom in on chromosome 5, including significance values ($-log_{10}$(P value)) for the association of losses with bulk scores of the MES1 and MES2 meta-modules and the bulk scores for a macrophage signature, demonstrating that the effect shown at the top panel is largely restricted to the MES1 meta-module.

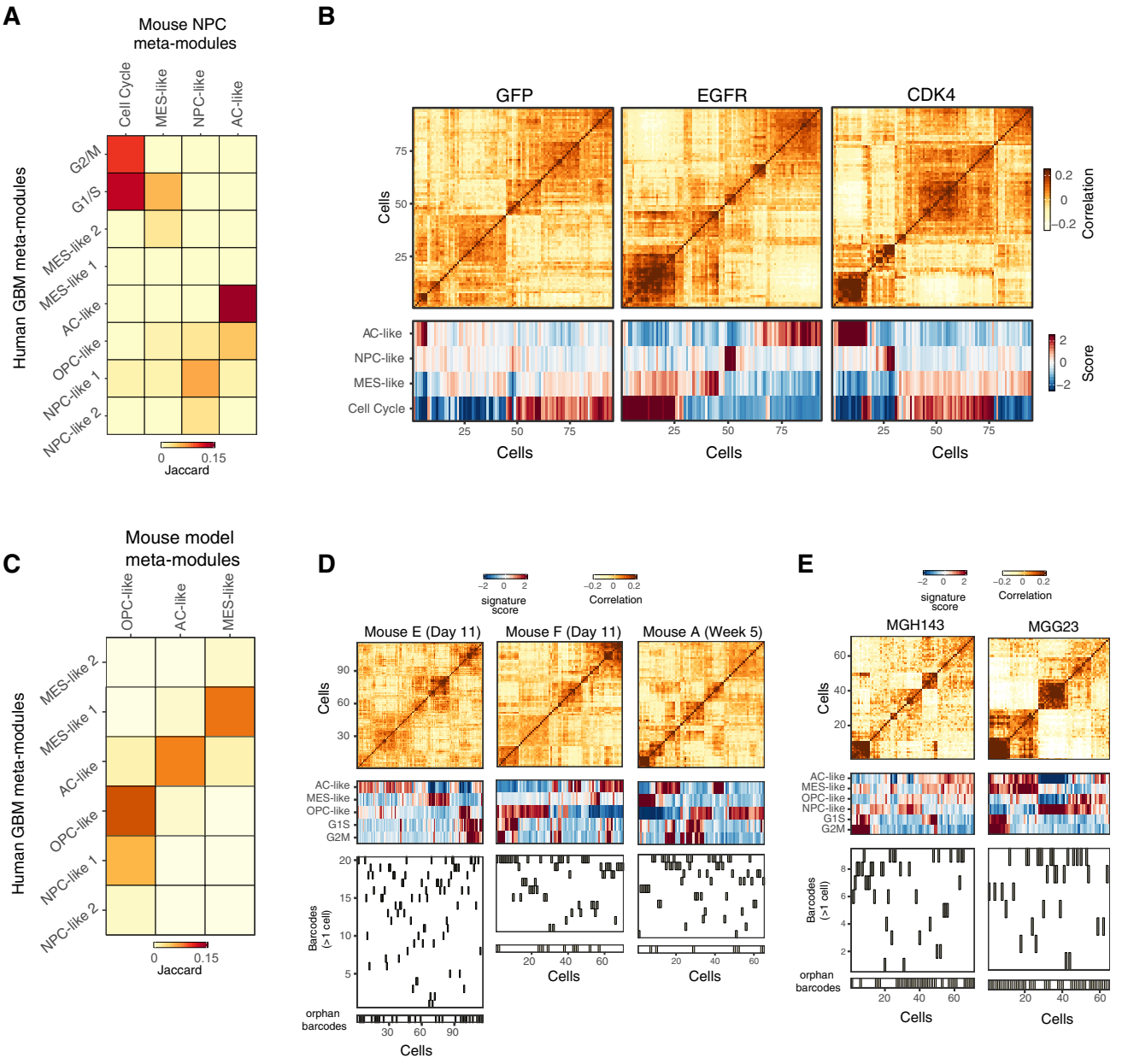


Figure S6. Related to Figures 6 and 7

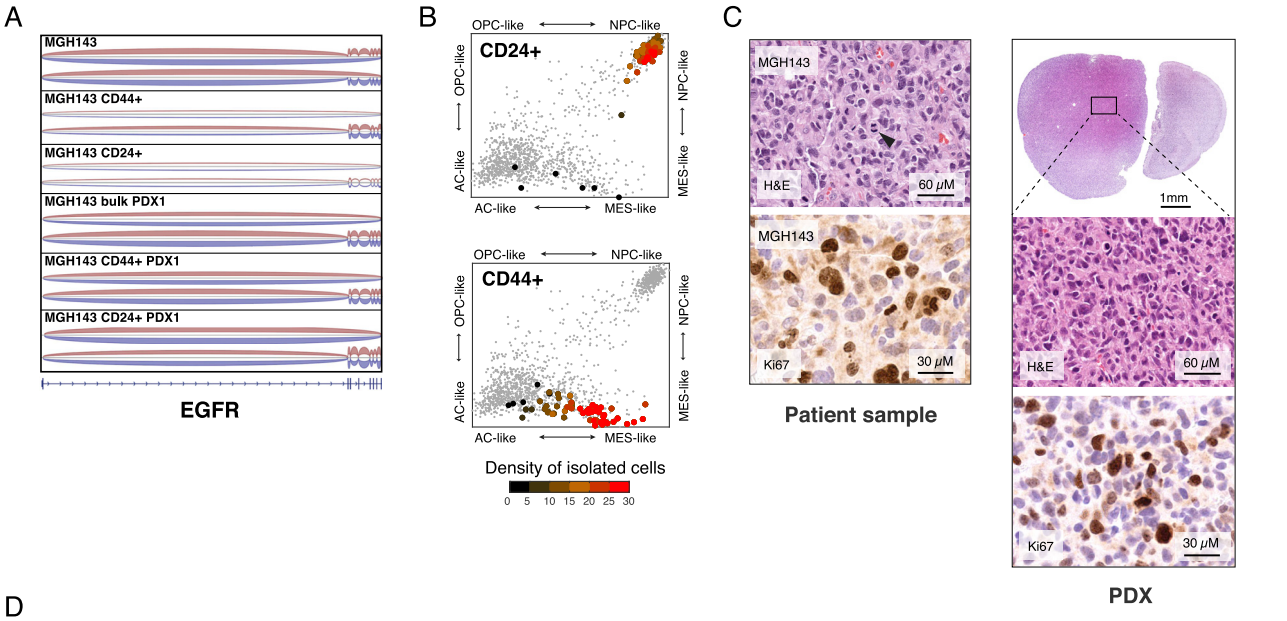
(A) Jaccard similarities of the six meta-modules from the main human glioblastoma analysis (x axis) with four meta-modules derived using the same approach from analysis of mouse NPC scRNA-seq data (y axis). Comparisons were based on human-mouse ortholog genes, and mouse meta-modules were named according to the most similar human meta-module.

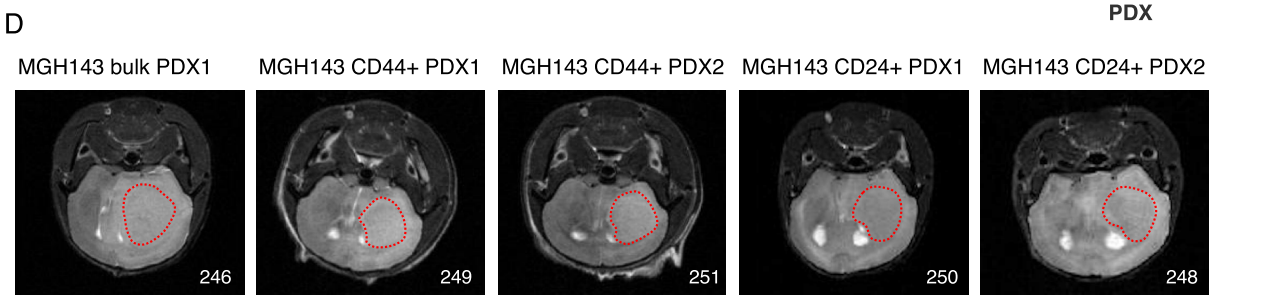
(B) Top: cell-to-cell correlation matrices of mouse NPCs overexpressing GFP, EGFR, or CDK4, ordered by hierarchical clustering. Bottom: expression scores for mouse meta-modules defined in (A).

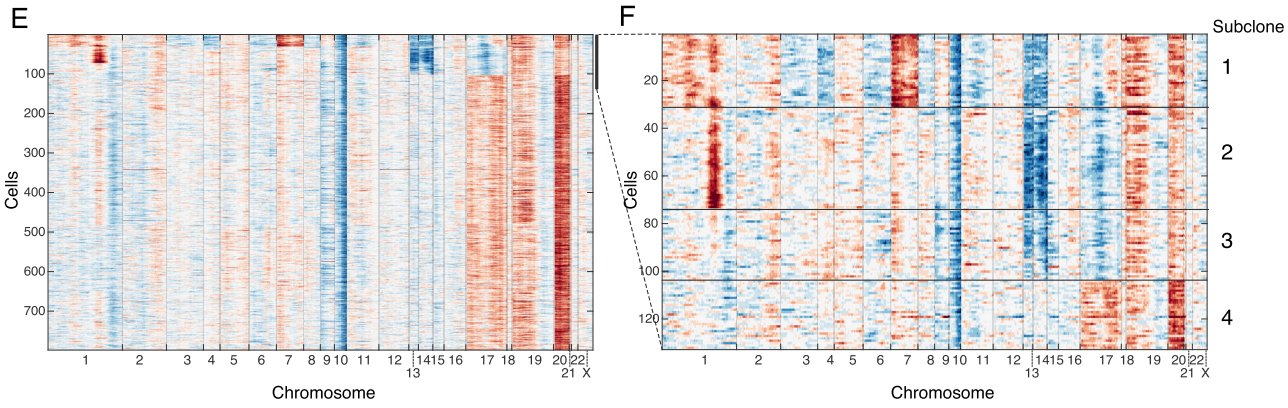
(C) Jaccard similarities of the six meta-modules from the main human glioblastoma analysis (x axis) with four meta-modules derived using the same approach from analysis of the genetic mouse model scRNA-seq data (y axis). Comparisons were based on human-mouse ortholog genes, and mouse meta-modules were named according to the most similar human meta-module.

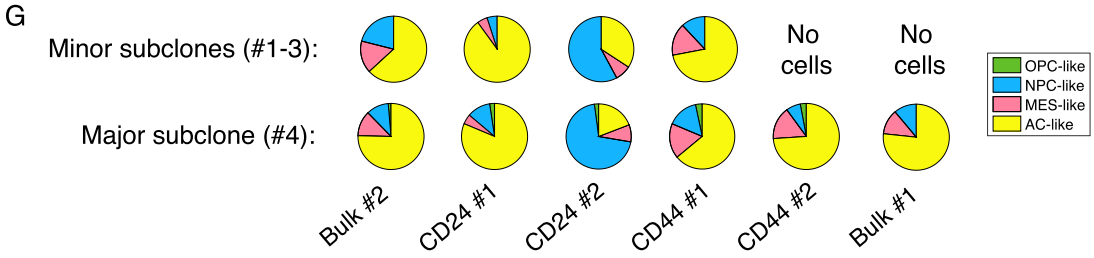
(D) Top: cell-to-cell correlation matrices of malignant cells from three barcoded mouse glioblastoma models at day 11 (two left panels) or week 5 (right panel) post transformation (STAR Methods), ordered by hierarchical clustering. Middle: expression scores for mouse meta-modules defined in (C). Bottom: assignment of cells (ordered as in top and middle panels) to barcodes; barcodes were ordered by the number of cells that map to them (highest at the top), and all orphan barcodes (each seen only in one cell) were combined in the lowest panel to enable compact visualization.

(E) Same as (B and D) for the patient-derived barcoded cells which were injected into immunocompromised mice.









(legend on next page)

Figure S7. Related to Figure 7

(A) EGFR locus, showing only exons 1–8. Exons junction-spanning reads (from +strand in red, from -strand in blue) mapped in different individual cells in MGH143 patient and PDX for unsorted, $CD44^+$ or $CD24^+$ sorted fractions showing that each fraction has both EGFR wild-type reads and reads that skip exons 2–7 (EGFR vIII).

(B) Two-dimensional representation of the cells in each sample based on their cellular states (as in Figure 3F). Small gray dots reflect all cells from MGH143 and its associated PDXs, while larger black-to-red dots reflect cells in the corresponding sample, colored by the density of cells that are from the same sample (STAR Methods).

(C) H&E stain and Ki67 immunohistochemistry of MGH143 patient sample (left) and bulk PDX (right). Tumor patient and PDX show similar morphology (features of high-grade glioma with important pleomorphism and cytonuclear atypia) and proliferation.

(D) Small animal MRI performed at the time of initial neurological symptoms (2–4 months after injection), for the PDX of MGH143 bulk, $CD24^+$ and $CD44^+$ subsets.

(E) Inference of chromosomal CNAs in PDX samples based on average relative expression in windows of 100 analyzed genes. Rows correspond to cells, which were ordered by hierarchical clustering across all loci in which only a subset of the cells had CNAs.

(F) Zoomed in view highlights the 4 subclones detected among the PDX cells, with three minor subclones (#1–3, each consisting of 3%–6% of the cell) and one major clone (#4, consisting of 87% of the cells, including all cells below the zoomed-in section).

(G) Pie charts displaying the fraction of cells in the four cellular states for each PDX, when separated between cells of the major (bottom) and minor (top) subclones; minor subclones were combined in this analysis because each of them is too small to be analyzed independently. This analysis demonstrates that the distributions of cellular states are largely decoupled from the segregation into genetic subclones.