nature cancer

ARTICLES

https://doi.org/10.1038/s43018-020-00154-9

Check for updates

Gradient of Developmental and Injury Response transcriptional states defines functional vulnerabilities underpinning glioblastoma heterogeneity

Laura M. Richards[1,2,18], Owen K. N. Whitley[3,4,18], Graham MacLeod[5], Florence M. G. Cavalli[6], Fiona J. Coutinho[6], Julia E. Jaramillo[3,6], Nataliia Svergun[6], Mazdak Riverin[2], Danielle C. Croucher[1,2], Michelle Kushida[6], Kenny Yu[7], Paul Guilhamon[6], Naghmeh Rastegar[6], Moloud Ahmadi[5], Jasmine K. Bhatti[5], Danielle A. Bozek[8,9], Naijin Li[3,6], Lilian Lee[6], Clare Che[6], Erika Luis[6], Nicole I. Park[2], Zhiyu Xu[2], Troy Ketela[2], Richard A. Moore[10], Marco A. Marra[10,11], Julian Spears[7,12], Michael D. Cusimano[7,12], Sunit Das[6,7,12], Mark Bernstein[12,13], Benjamin Haibe-Kains[1,2,14,15,16], Mathieu Lupien[1,2,14], H. Artee Luchman[8,9], Samuel Weiss[8,9], Stephane Angers[5,17], Peter B. Dirks[3,6,12] ✉, Gary D. Bader[3,4,15] ✉ and Trevor J. Pugh[1,2,14] ✉

Glioblastomas harbor diverse cell populations, including rare glioblastoma stem cells (GSCs) that drive tumorigenesis. To characterize functional diversity within this population, we performed single-cell RNA sequencing on >69,000 GSCs cultured from the tumors of 26 patients. We observed a high degree of inter- and intra-GSC transcriptional heterogeneity that could not be fully explained by DNA somatic alterations. Instead, we found that GSCs mapped along a transcriptional gradient spanning two cellular states reminiscent of normal neural development and inflammatory wound response. Genome-wide CRISPR–Cas9 dropout screens independently recapitulated this observation, with each state characterized by unique essential genes. Further single-cell RNA sequencing of >56,000 malignant cells from primary tumors found that the majority organize along an orthogonal astrocyte maturation gradient yet retain expression of founder GSC transcriptional programs. We propose that glioblastomas grow out of a fundamental GSC-based neural wound response transcriptional program, which is a promising target for new therapy development.

Glioblastomas (GBMs) are the most aggressive and treatment-refractory brain tumors in adults. Treatment failure is rooted in the extensive heterogeneity observed within tumors and across patients[1–4]. Molecular stratification of GBMs into transcriptional subgroups[5,6] (proneural, mesenchymal and classical) has not led to the development of successful targeted therapies[7], hindered by the inability of bulk sequencing to reflect the layered genetic, cellular and epigenetic diversity of cell states.

Single-cell RNA-sequencing (scRNA-seq) studies have highlighted the complexity of GBM biology[2,3,8–10], demonstrating that subpopulations of cells with different transcriptional subtypes and variable somatic genetic events (copy-number variations (CNVs) and mutations) coexist within a single tumor. However, the source of this functional intratumoral heterogeneity remains unclear and this has impeded the development of effective GBM treatments.

One potential source of phenotypic diversity and plasticity in GBMs lies within the rare self-renewing GSC fraction[11–14]. GSCs hijack developmental stem cell programs to drive and maintain tumor growth, as well as acquire resistance mechanisms to evade chemotherapy and radiotherapy[15–17]. However, it is still unclear how diversity within the GSC pool may affect the cellular composition and growth of GBMs.

[1]Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada. [2]Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada. [3]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. [4]The Donnelly Centre, University of Toronto, Toronto, Ontario, Canada. [5]Leslie Dan Faculty of Pharmacy, University of Toronto, Toronto, Ontario, Canada. [6]Developmental and Stem Cell Biology Program and Arthur and Sonia Labatt Brain Tumor Research Centre, The Hospital for Sick Children, Toronto, Ontario, Canada. [7]Division of Neurosurgery, St. Michael's Hospital, Toronto, Ontario, Canada. [8]Department of Cell Biology and Anatomy, University of Calgary, Calgary, Alberta, Canada. [9]Arnie Charbonneau Cancer Institute and Hotchkiss Brain Institute, University of Calgary, Calgary, Alberta, Canada. [10]Canada's Michael Smith Genome Sciences Centre, British Columbia Cancer, Vancouver, British Columbia, Canada. [11]Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia, Canada. [12]Division of Neurosurgery, Department of Surgery, University of Toronto, Toronto, Ontario, Canada. [13]Division of Neurosurgery, Toronto Western Hospital, University Health Network, Toronto, Ontario, Canada. [14]Ontario Institute for Cancer Research, Toronto, Ontario, Canada. [15]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada. [16]Vector Institute for Artificial Intelligence, Toronto, Ontario, Canada. [17]Department of Biochemistry, University of Toronto, Toronto, Ontario, Canada. [18]These authors contributed equally: Laura M. Richards, Owen K. N. Whitley. ✉e-mail: peter.dirks@sickkids.ca; gary.bader@utoronto.ca; trevor.pugh@utoronto.ca

Here, we applied scRNA-seq and genome-wide CRISPR–Cas9 screening to GSCs isolated from their in vivo primary tumor niche to study their molecular heterogeneity and function in an unbiased manner. Enriching for GSCs enabled us to observe a previously undescribed level of diversity within the cancer stem cell fraction of GBMs, a signal challenging to resolve in primary patient specimens due to the relative rarity of GSCs within the tumor bulk. We found that GSCs exist along a major transcriptional gradient between two cellular states, Developmental and Injury Response programs. Orthogonal to this GSC gradient, we identified an astrocyte maturation gradient in patient tumor cells, highlighting the transcriptional programs implicated in differentiation of GSCs into mature tumor cells that comprise the bulk. Our work provides a model that explains the source of cellular heterogeneity in GBMs and identifies a range of sensitivities of this fundamental cellular program that directly inform the development of new therapeutic strategies targeting GBMs.

Results

Transcriptional heterogeneity within GSCs. To enrich for rare stem-like cells within primary tumors, we used established serum-free culturing methods[18,19] to generate a collection of patient-derived GSCs capable of sustaining growth in vitro and initiating tumors in mice (Supplementary Table 1 and Supplementary Note 1). This method supports the growth of a diversity of clones that closely matches human GBM xenografts[12] and excludes cells of hematopoietic origin. To characterize heterogeneity in the GBM stem cell fraction, we profiled 69,393 cells from 29 early passage GSC cultures (21 adherent; 8 neurosphere) derived from 26 patients using scRNA-seq (Supplementary Table 2).

To explore GSC heterogeneity within individual patients, we clustered GSCs from each sample independently using extensive hyperparameter optimization and validation with multiple algorithms (Methods, Extended Data Fig. 1). We discovered substantial intra-GSC heterogeneity, uncovering two to six transcriptional subpopulations per GSC, totaling 86 clusters across 29 samples (Fig. 1a,b and Extended Data Fig. 1), demonstrating that in addition to the diverse cell states present in GBMs, rare GSC subpopulations within the tumor are heterogeneous themselves. For each cluster, we compared the top upregulated marker genes and across samples to identify shared subpopulations across GSCs (Supplementary Table 3). A subset of 14 clusters had increased similarity (mean Jaccard Index = 0.38 versus 0.066 for all other clusters) and shared upregulation of 358 core genes involved in cell cycling programs (Extended Data Fig. 2a–d). In addition to upregulation of canonical cell-cycle genes (*MKI67*, *TOP2A*, *AURKA*), proliferating GSC clusters overexpressed genes known to promote self-renewal and progenitor expansion in the neocortex[20] (including *ARHGAP11A* and *ARHGAP11B*). Many of these shared proliferation genes (*BRCA1*, *HMGB2*, *CDC45*) are also targets of the transcription factor TLX, part of a regulatory network governing proliferation in adult neural stem cells[21] and self-renewal in brain tumor stem cells[22]. GSCs with a larger fraction of actively cycling cells displayed increased aggressiveness and reduced survival upon implantation in an orthotopic xenograft model (Extended Data Fig. 2c). Collectively, these observations define a core GSC proliferation module, resembling aberrant neurodevelopmental programs, potentially employed by GSCs to sustain tumor growth.

Remaining intra-GSC clusters (72 of 86) had limited marker similarity (mean Jaccard Index = 0.066), suggesting a large portion of subpopulations within GSCs are specific to individual patients (Extended Data Fig. 2a). Within individual GSC samples, expression of marker genes drove divergence of transcriptionally distinct subpopulations. For example, G549_L consisted of two transcriptional states; one cluster (C1) characterized by upregulation *of EDN1* and *ADM*, both HIF-1 target genes involved in angiogenic signaling[23], while the second cluster (C2) overexpressed *ASCL1*, a transcription factor critical for neuronal differentiation that suppresses tumorigenicity in GSCs[24] (Extended Data Fig. 2e,f). These results demonstrate substantial heterogeneity both within and between the GSC pools of individual patients, with important implications for designing targeted therapies against multiple subpopulations in the tumor-initiating fraction of GBMs.

CNVs can modulate intra-GSC heterogeneity. To evaluate whether the polyclonal structures observed at the transcriptional level are a result of somatic genome alterations, we inferred CNV profiles from scRNA-seq data for each intra-GSC cluster (Fig. 1c,d; Methods). We validated CNVs inferred from scRNA-seq with matched bulk whole-genome sequencing (WGS) for a subset of 20 samples. CNV profiles from bulk WGS were more similar to averaged scRNA-seq-derived profiles from all cells versus individual clusters (Spearman's $r = 0.68$ versus $r = 0.63$, $p < 0.001$). While the aggregate data verifies our scRNA-seq CNV results, cluster-level profiles support the presence of subclonal CNVs within GSCs not detected by bulk approaches (Extended Data Fig. 3 and Supplementary Tables 4 and 5).

Amplification of chromosome 7 and deletion of chromosome 10 were common across clusters, indicating that these are likely clonal, founding events involved in the malignant transformation of neural stem cells (NSCs) to GSCs (Fig. 1c), consistent with reported frequency and evolutionary timing in GBMs[1,5,25,26]. Most GSCs harbored transcriptional clusters with unique CNV profiles ($n = 22$ of 29 samples totaling 69 clusters), indicative of extensive subclonal genomic diversification within GSCs (Extended Data Fig. 3d). For example, in G876_L all three clusters shared clonal amplification of chromosome 7, in addition to private subclonal CNVs restricted to one transcriptional cluster. Deletion of chromosome 9 was observed in 2 of the 3 clusters (C1, C2) in G876_L, while amplification of chromosome 12 was exclusive to a separate, rare cluster of cells (C3) (Fig. 1d). Furthermore, 49% of clusters ($n = 34$ of 69) had significant enrichment ($P < 0.05$, Fisher's exact test) of marker genes within altered CNV loci, highlighting the potential for subclonal CNVs to modulate transcriptional programs in GSCs (Extended Data Fig. 3e). However, not every GSC had evidence of genomic diversity. BT67_L has two transcriptional clusters presenting with identical inferred CNV profiles ($P = 0.16$, Kolmogorov–Smirnov test) (Fig. 1d). Therefore, while established GBM founder CNVs are common and clonal across GSCs, subclonal CNVs likely drive only a portion of intra-GSC heterogeneity observed between patients.

Characterizing GSC heterogeneity between patients. To map GSC transcriptional heterogeneity across patients, we used uniform manifold approximation and projection (UMAP) to visualize inter-GSC relationships (Extended Data Fig. 4a,b). Unsupervised clustering identified 61 transcriptional clusters, revealing striking patient-specific transcriptional programs, with most clusters ($n = 57$ of 61) characterized by an almost entirely unique, patient-specific GSC transcriptional profile. To ensure patient-specific clustering patterns reflect true biological signals innate to cancer cells, and not technical batch effects, we applied three batch-correction methods (Extended Data Fig. 4c–e). No batch-correction algorithm was successful in unifying clusters across all samples and were inconsistent with each other, supporting the conclusion that our samples display substantial inter-patient heterogeneity, as has been observed in tumors[2,3,27–34] and malignant cell lines[35–38] from a variety of human cancers, including GBM. Supporting this, GSCs derived from different geographical regions of the same tumor (G945-I,J,K and G946-J,K) were more similar to each other than to GSCs derived from different tumors (Extended Data Fig. 4b).

GSCs organize along a transcriptional gradient. To identify core transcriptional programs underpinning inter-GSC heterogeneity,

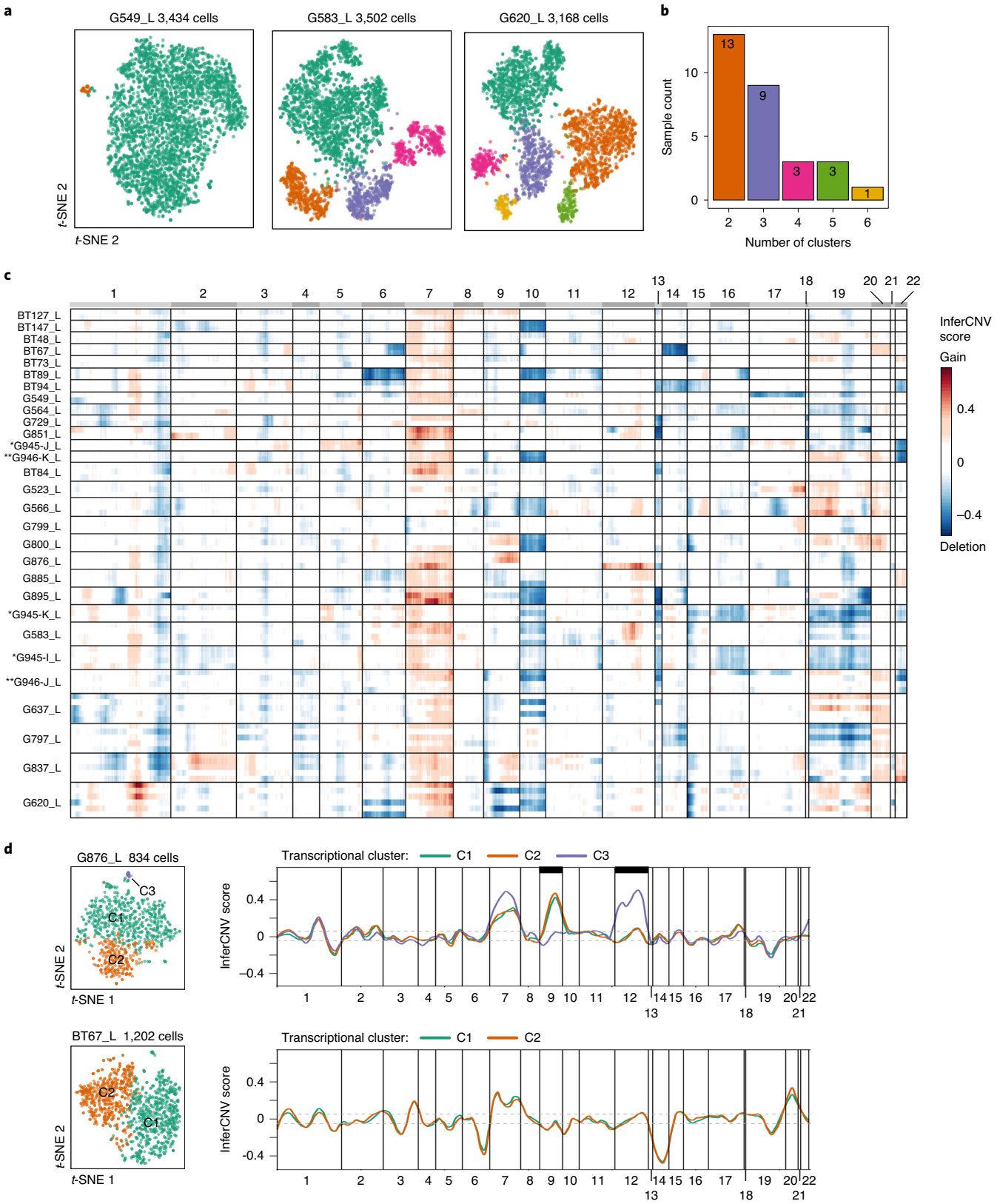


Fig. 1 | Characterizing heterogeneity within GSCs. a, *t*-distributed stochastic neighbor embedding (*t*-SNE) visualization of GSC cultures from select samples demonstrating intra-sample heterogeneity defined by the presence of multiple transcriptional clusters. Cells colored by transcriptional cluster. **b**, Breakdown of cluster number across 29 GSC cultures. **c**, Genome-wide inferred CNV profiles for 29 patient-derived GSC cultures. Columns represent genomic regions, ordered by genome position across all chromosomes. Rows represent CNVs averaged by intra-sample transcriptional cluster, with one row per cluster (Extended Data Fig. 1a). Samples ordered by increasing cluster number. **d**, Inferred CNV value (*y* axis) for select GSC cultures with (top) and without (bottom) CNV variation between transcriptional clusters. Lines are colored by intra-sample transcriptional cluster. Black bars represent regions of variable CNVs between clusters.

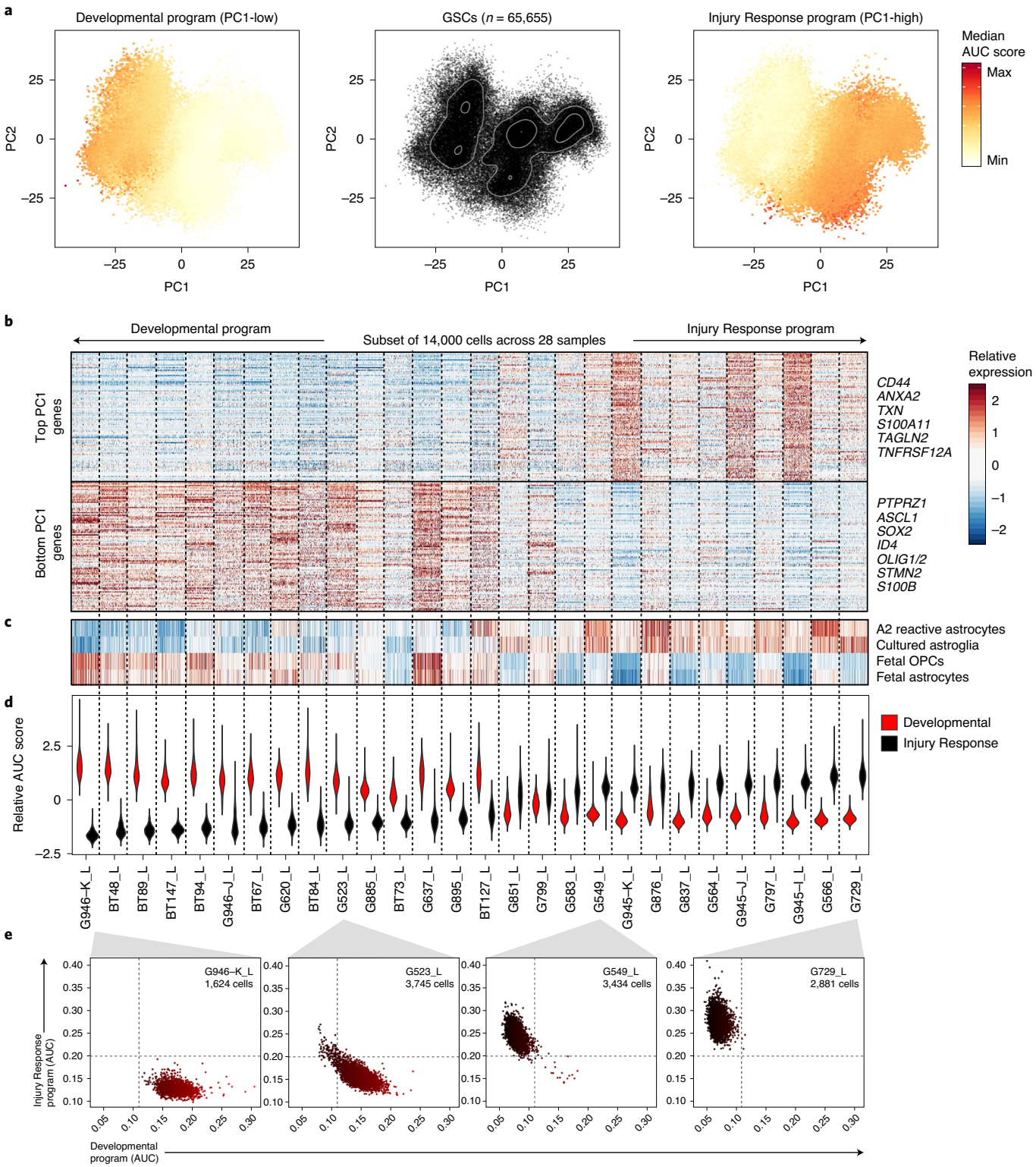


Fig. 2 | GSCs converge on a single transcriptional gradient between Developmental and Injury Response states. a, PCA of 65,655 cells from 28 GSC cultures derived from 24 patients (middle). Cells colored by expression of Developmental (PC1-low) and Injury Response (PC1-high) programs (left and right, respectively). AUC, area under the curve. **b**, Relative expression of top 100 and bottom 100 weighted genes for PC1, in a subset of 14,000 individual GSC cells (500 cells per sample, randomly selected). Select enriched genes highlighted. GSC cultures ordered by increasing median Injury Response program score as defined in Fig. 2d. **c**, Relative program score for individual cells (500 cells per sample; same cells as in Fig. 2b) for top-correlated cell-type signatures. GSC cultures ordered as in Fig. 2d. **d**, Relative signature scores of individual cells ($n = 65,655$ cells from 28 GSC cultures) evaluated for Developmental (red) and Injury Response (black) gene signatures derived from bulk RNA-seq analysis (related to Extended Data Fig. 6d,e). **e**, Single cell profiles from representative GSC cultures ($n = 4$) show that individual GSCs fall along a continuous axis between Developmental and Injury Response states. Cells are colored by relative expression of Developmental (red) and Injury Response (black) expression programs. GSC cultures with intermediate scores either contain subpopulations of both subtypes or middling scores for both states.

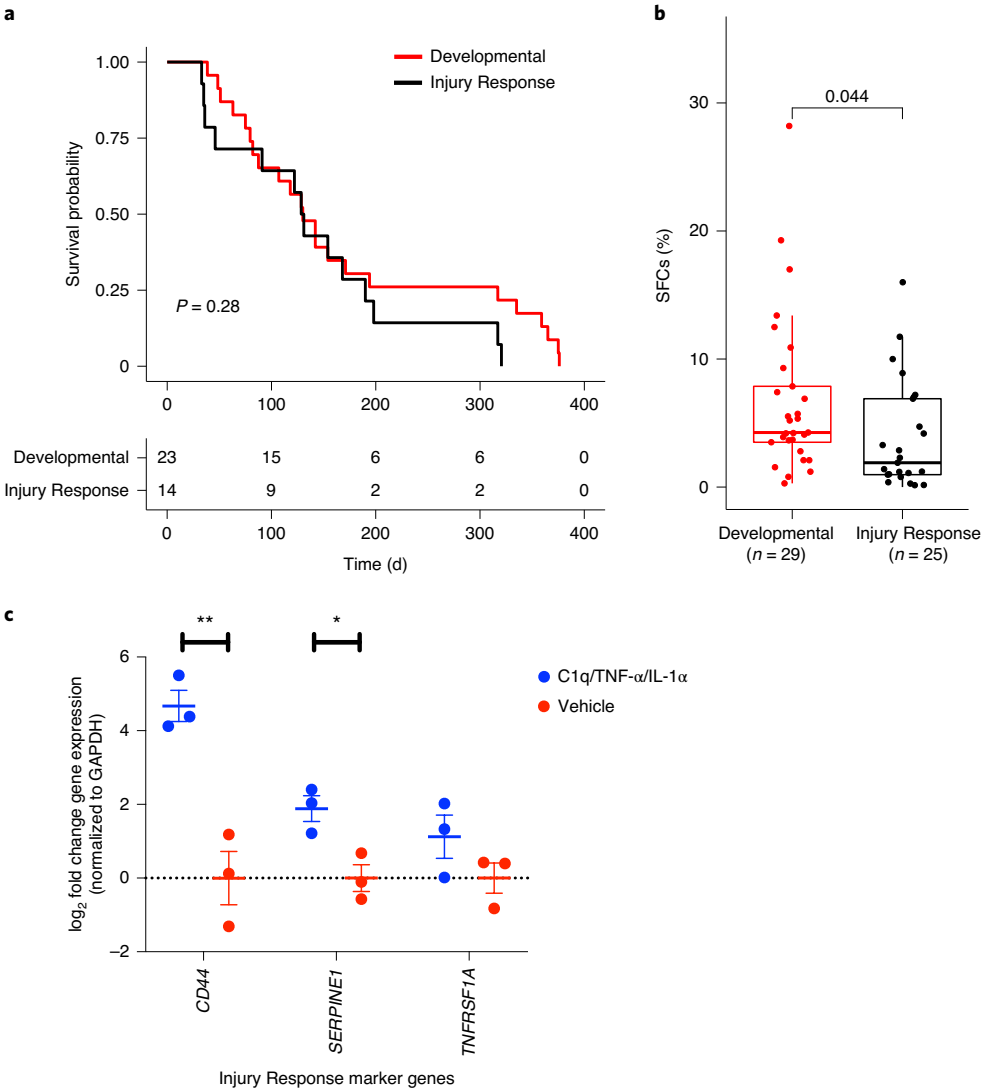


Fig. 3 | Developmental and Injury Response GSCs have functional differences and potential for plasticity. a, Kaplan–Meier curve depicting overall survival in Developmental (red; $n = 23$ GSCs) versus Injury Response (black; $n = 14$ GSCs) GSCs in an orthotopic xenograft model. *P* values determined by a two-sided log-rank test. **b**, Difference in SFCs between Developmental (red; $n = 29$ patient-derived GSCs) and Injury Response (black; $n = 25$ patient-derived GSCs) GSCs as determined by in vitro limiting dilution assays (LDAs). Box plots represent the median, first and third quartiles of the distribution and whiskers represent either 1.5-times interquartile range or the most extreme value. Each circle represents one GSC sample. A two-sided Student's *t*-test was used for statistical analysis to compare means. **c**, Cells from a Developmental GSC (G523_L) were treated for 48 h with a cytokine cocktail consisting of C1q (400 ng ml^{-1}), TNF-α (30 ng ml^{-1}) and IL-1α (3 ng ml^{-1}) or with vehicle. Gene expression was quantified by RT-qPCR and normalized to *GAPDH*. Data represent mean ± s.e.m. A two-sided Student's *t*-test was used for statistical analysis ($n = 3$ independent experiments). $^{*}P = 0.0205$; $^{**}P = 0.00506$.

we performed principal-component analysis (PCA) on the global scRNA-seq dataset of 69,393 cells. We removed one outlier GSC sample, G800_L, from downstream analysis on the basis of inflated PC2 signal, leaving 65,655 cells (Extended Data Fig. 5a). Re-running PCA without G800_L revealed a single axis of variation along PC1, separating cells into two prominent groups. (Fig. 2a).

Cells with high PC1 loadings were associated with elevated expression of mesenchymal-related genes and enrichment of pathways implicated in inflammation and immune cell activation, as well as nuclear factor (NF)-κB and STAT signaling (false discovery rate (FDR) < 0.01; Fig. 2b and Extended Data Fig. 5b,c). When compared to cell types found in developing fetal brain[39,40], mature adult brain[41–45] and malignant cell states in GBMs[2,5], these inflamed GSCs best resembled both the Cancer Genome Atlas (TCGA) Mesenchymal subtype and the mesenchymal-like cell state[2] in GBMs, as well as neuroprotective A2 reactive astrocytes (Fig. 2c, Extended Data Fig. 5b and Supplementary Table 6). Interestingly, A2 reactive astrocytes promote neuronal survival and tissue repair in response to ischemic injury[42,46], perhaps paralleling mechanisms employed by GSCs to sustain growth and self-renewal in hypoxic tumor microenvironments. Furthermore, upregulation of interferon and wound-healing programs suggests the mesenchymal-like phenotype in GSCs may be the result of microenvironment-induced transcriptional reprogramming in response to injury.

Conversely, cells with low PC1 loadings were associated with genes and pathways related to gliogenesis and neural development (for example *PTPRZ1*, *ASCL1*, *SOX2*), highlighted by the expression of oligodendrocytic (for example *OLIG1*, *OLIG2*), astrocytic (for example

CLU, *APOE*, *S100B*) and neuronal (for example *STMN3*) lineage markers (Fig. 2b and Extended Data Fig. 5b–d). Consistently, this group of GSCs strongly resembled a spectrum of developing cell types, including oligodendrocyte progenitor cells (OPCs), developing astrocytes and radial glia. Similarly, these developmental-like GSCs mirrored transcriptional profiles of multiple malignant GBM cell types, such as the Classical and Proneural subtypes reported by TCGA[5] and recently reported neural precursor (NPC), astrocyte (AC) and OPC-like cell states[2] (Fig. 2c and Extended Data Fig. 5b). This finding is indicative of a multipotent class of GSCs capable of differentiating into mature neural cell types. This result was recapitulated using Diffusion Map, an alternate dimensionality reduction method designed to identify gradients from scRNA-seq data[47] (Extended Data Fig. 6a,b).

We conclude that GSCs exist between two major transcriptional programs: one reminiscent of neural development with differentiation capacity, which we term 'Developmental' (low PC1 loadings) and the other with inflammatory and wound response signaling resembling reactive astrocytes, which we name 'Injury Response' (high PC1 loadings) (Fig. 2a).

To validate the existence of two GSC states, we profiled a larger cohort of 72 GSCs (38 adherent, 34 neurosphere) with bulk RNA-seq, a subset of which ($n=23$ of 72) overlap with those profiled by scRNA-seq. Using a resampling procedure, bulk GSC profiles separated into two stable clusters (Extended Data Fig. 6c,d). Consistent with our scRNA-seq data, differential gene expression and pathway enrichment analysis identified one GSC cluster enriched for pathways involved in neuro- and gliogenic signaling and development (consistent with the Developmental subtype) and another enriched for inflammatory response programs (consistent with the Injury Response subtype) (Extended Data Fig. 6e,f and Supplementary Table 7).

At the population level using bulk RNA-seq profiling, GSCs were categorized discretely as Developmental or Injury Response (Extended Data Fig. 6d). However, at the single-cell level, we observed a transcriptional gradient between the two states (Extended Data Fig. 7). For each patient, GSCs occupied a discrete range within the Developmental and Injury Response spectrum. (Fig. 2d,e and Extended Data Fig. 7c). Patient localization to a range of the gradient is not the result of technical artifacts, as the same gradient existed after correcting the expression matrix for batch by matching mutual nearest neighbors[48] across samples (Extended Data Fig. 7d,e). Furthermore, cells from multiple patients mapped to overlapping regions of the Injury Response–Developmental gradient, supporting common cellular phenotypes across patients (Methods; Extended Data Fig. 7f). Thus, profiling GSCs from many samples is necessary to characterize the full spectrum of possible transcriptional states giving rise to bulk GBM.

Developmental and Injury Response GSC states have functional differences and exhibit plasticity. We functionally validated the presence of the two GSC transcriptomic states using core cancer stem cell assays. Using in vitro limiting dilution assays as a readout of self-renewal, we found that Developmental GSCs had higher rates of sphere-forming cells (SFCs) compared to Injury Response GSCs ($P=0.044$, Student's t-test) (Fig. 3b). Furthermore, Developmental gene signature scores were correlated with the proportion of SFCs (Spearman's $r=0.30$, $P=0.027$), whereas Injury Response gene signature scores were negatively correlated (Spearman's $r=-0.32$, $P=0.018$), demonstrating that GSC functional properties vary along the transcriptional gradient.

To assess disease aggressiveness and tumorigenic potential between the two GSC states, we engrafted 37 GSC lines intracranially into immunocompromised mice. In line with stratification of patients with GBMs into transcriptional subgroups[5], we did not observe a difference in survival between Developmental and Injury Response GSCs in an orthotopic xenograft model ($P=0.28$, log-rank test), suggesting that both GSC states give rise to equally aggressive tumors (Fig. 3a). However, we did observe a difference in tumorigenicity. Developmental GSCs ($n=23$ of 23) had significantly higher rates of tumor formation compared to Injury Response GSCs ($n=11$ of 14; $P=0.047$, Fisher's exact test), perhaps highlighting the requirement of the tumor microenvironment to perpetuate the Injury Response GSC phenotype. Collectively, these assays demonstrate that functional properties governing GSC phenotype are associated with the gradient of transcriptional states.

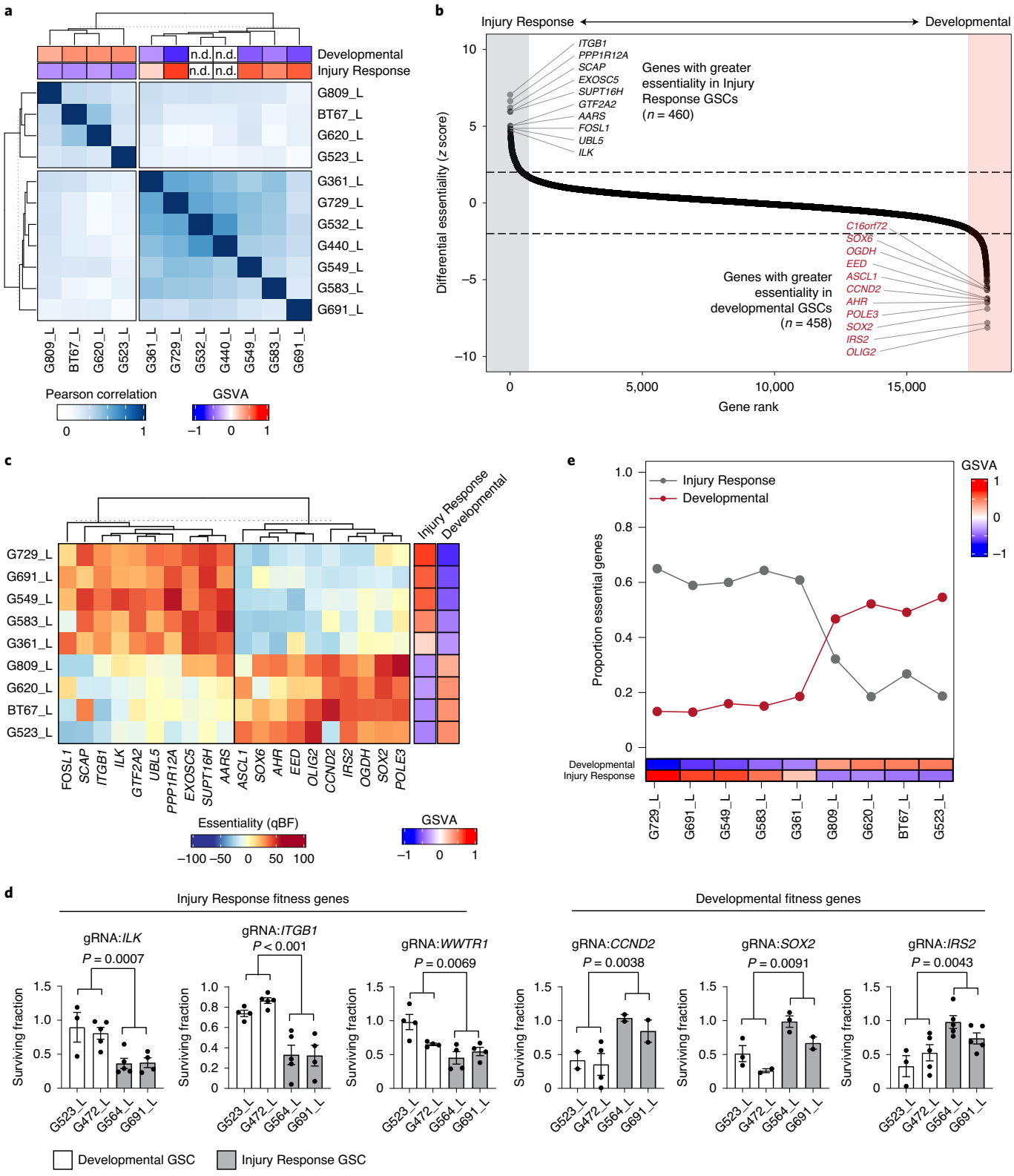
Given the continuous nature of GSC phenotypes along the transcriptional gradient, we investigated the possibility of plasticity between Developmental and Injury Response states. We treated a Developmental GSC (G523_L) with an inflammatory cytokine cocktail (C1q, tumor necrosis factor (TNF)-α and IL-1α) and assessed the expression of Injury Response gene markers (*CD44*, *SERPINE1* and *TNFRSF1A*) by quantitative PCR with reverse transcription (RT–qPCR) (Fig. 3c). The cytokine cocktail induced expression of Injury Response genes after 48 h, demonstrating the potential for microenvironment-induced conversion of GSCs from a Developmental to Injury Response state. These assays mimic conditions in the tumor microenvironment to inform the potential of plasticity between GSC states and the origins of inflammatory signals we observed in vitro. These results suggest that inflammatory cytokines previously found to be secreted by microglia to induce the formation of reactive astrocytes[42], may also induce the expression of Injury Response genes in Developmental subgroup GSCs.

Functional dependencies identified by genome-wide CRISPR screens reflect Developmental–Injury Response gradient position. To identify functional dependencies and potential therapeutic targets underpinning the Developmental–Injury Response gradient, we performed genome-wide CRISPR–Cas9 dropout screens using the 70-k TKOv3 library[49] (70,948 guides targeting 18,053 protein-coding genes) in 11 GSCs, a subset of which overlapped

Fig. 4 | Genome-wide CRISPR screens identify essential regulators of the transcriptional gradient in GSCs. a, Pearson correlation between CRISPR screens ($n=11$ GSC cultures), ordered by hierarchical clustering. Columns annotated with gene set variation analysis (GSVA) gene signature scores from matched bulk RNA-seq. n.d. denotes no bulk RNA-seq data available for sample. **b**, Rank order plot depicting differential fitness scores between Developmental ($n=4$) and Injury Response ($n=5$) GSC screens. Rank is according to differential fitness z scores (average qBF for Injury Response GSC screens, average qBF for Developmental GSC screens). Top ten hits per group are labeled. **c**, Heat map of quantile normalized gene fitness qBF scores for the top ten differentially essential genes between Developmental and Injury Response GSCs. Rows ordered by position on the transcriptional gradient (related to Fig. 2d). Rows are annotated with GSVA gene signature scores from matched bulk RNA-seq. **d**, Validation of state-specific fitness genes identified in CRISPR–Cas9 screens. Cas9-expressing Developmental (G523_L and G472_L; white) and Injury Response (G564_L and G691_L; gray) GSCs were transduced with lentivirally expressed gRNAs targeting indicated genes. gRNA-infected cells were grown in competitive proliferation assays against control cells expressing AAVS1 targeting gRNAs for 14 d, at which point relative cell number was assessed by flow cytometry. P values were calculated using Welch's t-test (two-sided) comparing pooled Injury Response and Developmental replicates. Bars represents mean ± s.e.m. Data points represent independent biological replicates from $n=2$–5 independent experiments per gRNA. **e**, Line plot depicting the proportion of Injury Response (gray line) and Developmental (red line) fitness genes (as defined in Fig. 4b) that are essential in each GSC. Samples are ordered by position on the transcriptional gradient (related to Fig. 2d) and annotated with GSVA gene signature scores from matched bulk RNA-seq.

those profiled by bulk ($n = 9$ of 11) and scRNA-seq ($n = 6$ of 11). We used the BAGEL algorithm[50,51] to normalize gRNA reads for sample sequencing depth, calculate fold change for each guide RNA from the T0 baseline and compute a quantile normalized Bayes factor (qBF) for each gene, representing a confidence measure that knockout of a specific gene reduced fitness (Supplementary Table 8). Notably, unsupervised clustering of variable essential genes (1,345 genes; qBF > 10 in 3–9 of 11 screens) recapitulated Developmental and Injury Response groups, consistent with observations from bulk and scRNA-seq (Fig. 4a and Extended Data Fig. 8). These data emphasize the fundamental role of the GSC gradient in governing essential cellular phenotypes.

Next, we calculated the difference in qBF scores between Developmental and Injury Response GSCs to identify differentially



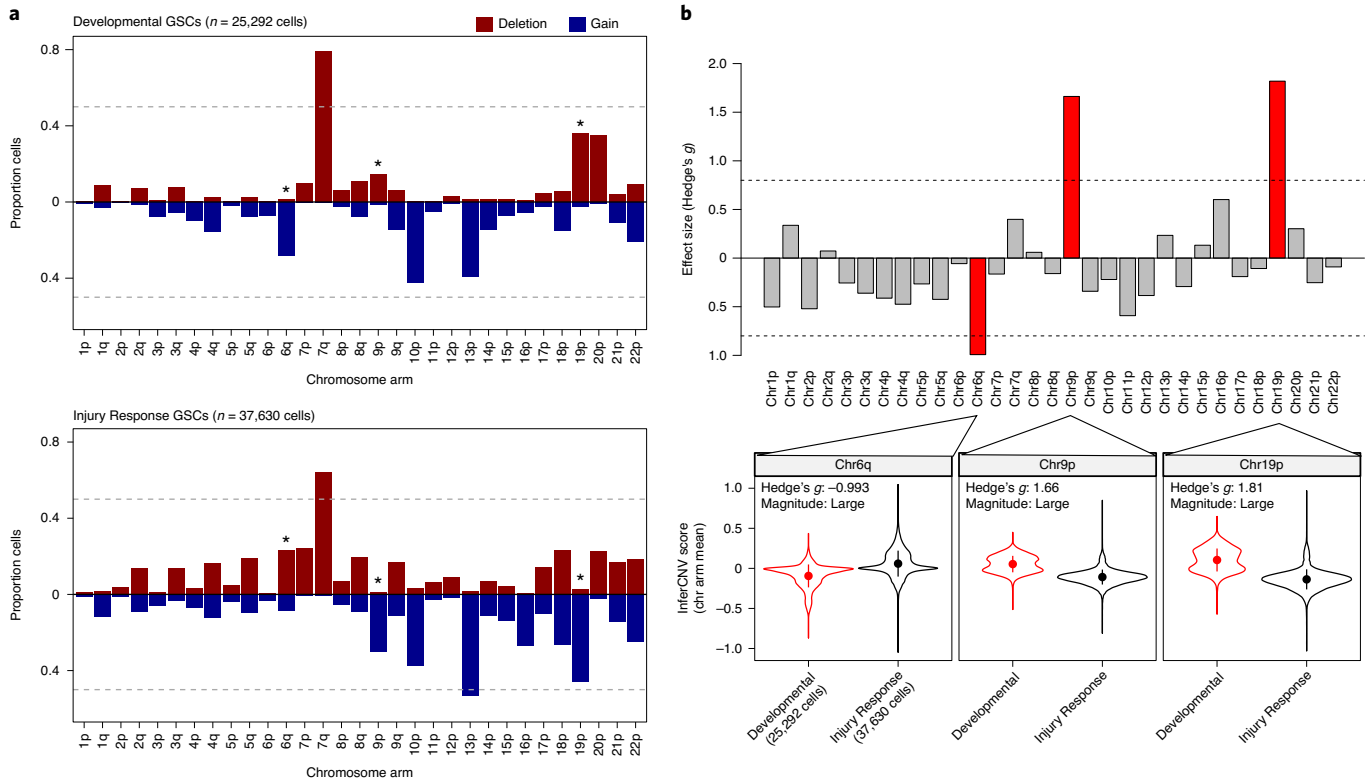


Fig. 5 | Genetic alterations influence GSC state. a, Frequency of amplifications (red) and deletions (blue) across chromosomal arms within cells classified as being Developmental (left) or Injury Response-like (right). Regions variably altered between states are denoted by asterisks. **b**, Comparison of InferCNV scores between Developmental ($n = 25,292$ cells) and Injury Response ($n = 37,630$ cells) GSCs across chromosome arms. Bar plot of effect size calculated with Hedge's g. Chromosome arms with a 'large' effect size (defined as >0.8, red bars) were determined to be variably altered between groups. The central dot in the violin plot represents the mean and whiskers represent s.d. Tips of the violin plot extend to the minimum and maximum values of the distribution. A two-sided Wilcoxon test was used for statistical analysis to compare means.

essential genes (Supplementary Table 9). Examination of top differential fitness genes (z score cutoff of >2 or <−2) in each respective GSC state identified dependencies resembling gene expression markers and biological processes identified in the transcriptomics data (Fig. 4b). Injury Response GSCs were dependent on genes related to inflammation and integrin signaling (for example *ITGB1*, *ILK*) for their proliferation, whereas Developmental GSCs were dependent on genes implicated in neurodevelopment (for example *OLIG2*, *SOX2*, *ASCL1*) (Fig. 4c).

Using competitive cell proliferation assays, we validated three hits each from Developmental (*CCND2*, *SOX2*, *IRS2*) and Injury Response (*ILK*, *ITGB1*, *WWTR1*) GSC states by testing individual gene knockouts (two gRNAs per gene) in a panel of four GSC lines (two Developmental and two Injury Response) (Fig. 4d). GSCs were preferentially sensitive to knockdown of gene hits from their respective transcriptional state. Injury Response GSCs were sensitive to knockdown of Injury Response gene hits, but not Developmental hits and vice versa, demonstrating that GSC states have unique and specific functional dependencies underpinning cellular growth.

Pathway analysis on differentially essential genes revealed Injury Response GSCs were more sensitive to perturbations in basic cellular functions such as cell cycle, splicing and DNA repair, as well as immune related signaling pathways (Extended Data Fig. 8e). Interestingly, Developmental GSCs relied on aerobic respiration, whereas Injury Response GSCs were more dependent on glycolysis. Under hypoxic conditions, tumor-initiating cells in GBMs upregulate glycolysis to promote drug resistance and stemness[52], suggesting that GSC fitness is influenced by their microenvironmental niche. This is consistent with our expression data showing upregulation of transcriptional programs related to hypoxia and angiogenesis in Injury Response GSCs (Extended Data Figs. 5d and 6e) and demonstrates GSC functional dependencies are reflective of their transcriptional programming.

Furthermore, we observed that GSCs organize along an essentiality gradient, mirroring the transcriptional gradient (Fig. 4e). The most Developmental GSCs, as defined by expression data (G523_L), were dependent on the greatest fraction of Developmental fitness genes. The same observation was true in Injury Response GSCs. GSCs located at the center of the gradient (for example, G809_L and G361_L), potentially representing mixed Developmental/Injury Response phenotypes, were the most reliant on fitness genes from both GSC states. Regardless of position on the gradient, all GSCs possessed essential genes from both ends of the spectrum, suggesting that combinatorial targeting of essential genes implicated in core Developmental and Injury Response processes could have general therapeutic benefit across patients.

Position on GSC gradient is associated with specific copy-number variants. Next, we hypothesized that specific CNVs may be preferentially enriched within Developmental and Injury Response GSC subtypes. Using gene signature scoring, we categorized cells into Developmental or Injury Response subtypes and compared the frequency and signal of CNVs across chromosome arms within these two groups (Fig. 5). To obtain a pure view of genetic heterogeneity within states, we excluded hybrid or unknown cells (2,733 of 65,655 cells; 4%) from the analysis, defined as cells classified into both

subtypes and neither subtype, respectively. Generally, Developmental and Injury Response GSCs shared similar CNV profiles (Fig. 5a). Full or partial gain of chromosome 7 (79% Developmental, 64% Injury Response) and loss of chromosome 10 (42% Developmental, 38% Injury Response) occurred at similar, high frequencies in both GSC subtypes, consistent with reports that place these CNVs at the apex of GBM somatic evolution[25].

In contrast, to established founder CNVs, we identified three chromosome arms, 6q, 9p and 19p, as being differentially altered between Developmental and Injury Response GSCs. Chromosome arm 6q was frequently amplified in Injury Response cells (23% versus 1%) and deleted in Developmental cells (28% versus 8%) (effect size = 0.99) (Fig. 5b). This chromosomal region encodes potential regulators of the Injury Response phenotype, including *TNFAIP3*, involved in TNF signaling and cytokine-mediated inflammatory responses. Chromosome arm 19p was more frequently deleted in Injury Response cells (46% versus 2%) and amplified in Developmental cells (36% versus 3%) (effect size = 1.81). Deletion of chromosome arm 9p, encompassing the *CDKN2A/B* locus, was exclusive to the Injury Response state (30% versus 1%) (effect size = 1.66) and is implicated in GBM initiation[25]. Both Developmental and Injury Response marker genes were enriched in state-specific altered regions of the genome ($P < 0.0001$, chi-squared test), suggesting that somatic CNVs can affect position in the GSC gradient.

Heterogeneity in GBMs is defined by two transcriptional axes. To determine where the Developmental–Injury Response GSC gradient lies within the cellular architecture of GBMs, we profiled 44,712 cells from seven GBM tumors using scRNA-seq. Using a combination of unbiased clustering, cell-type marker expression and CNV inference, we determined that 14,207 of 30,505 cells were malignant tumor cells (Fig. 6). We performed PCA on the combined 79,862 cancer cell dataset (65,655 GSCs and 14,207 tumor cells) to identify shared transcriptional programs between GSCs and GBM tumor cells. The first two principal components defined two core axes of variation explaining the genesis of heterogeneity in GBMs (Fig. 7a). The first, a differentiation trajectory between stem-like GSCs and differentiated tumor cells and the second recapitulating the Developmental–Injury Response gradient that we observed in GSCs alone (Fig. 7a). To investigate transitional dynamics between GSCs and differentiated tumor cells, we ran RNA velocity in combination with Diffusion Map on a subset of cells ($n = 20{,}343$ cells; Methods; Fig. 7b). In general, the vector field points from the root of GSCs (DM1-high) to the tail of tumor cells (DM1-low), indicating directional flow from a stem-like phenotype to differentiated tumor cell and further supporting the gradients that we identified by PCA.

Separation between GSCs and tumor cells along the differentiation trajectory underscores the presence of distinct transcriptional programs involved in the transition from stem-like initiating cells to mature differentiated tumor cells in GBMs (Fig. 7a). Tumor cells most distant from GSCs, at the end of the differentiation trajectory, resemble mature nonproliferative astrocytes[41,53], expressing canonical markers such as *GFAP*, *AQP4* and *APOE* (Fig. 7c,d and Extended Data Fig. 9a). Conversely, the GSC pool was enriched for gene signatures related to progenitor cells, such as NPCs and young astrocytes, as well as elevated expression of *H2FAZ*, a gene involved in regulating gliogenesis in neural precursor cells[54]. The second transcriptional gradient was correlated with the Developmental–Injury Response gradient that we observed in GSCs. Both tumor cells and GSCs expressed markers of Developmental (for example *OLIG1*, *OLIG2*) and Injury Response (for example *CD44*) states (Fig. 7c,d).

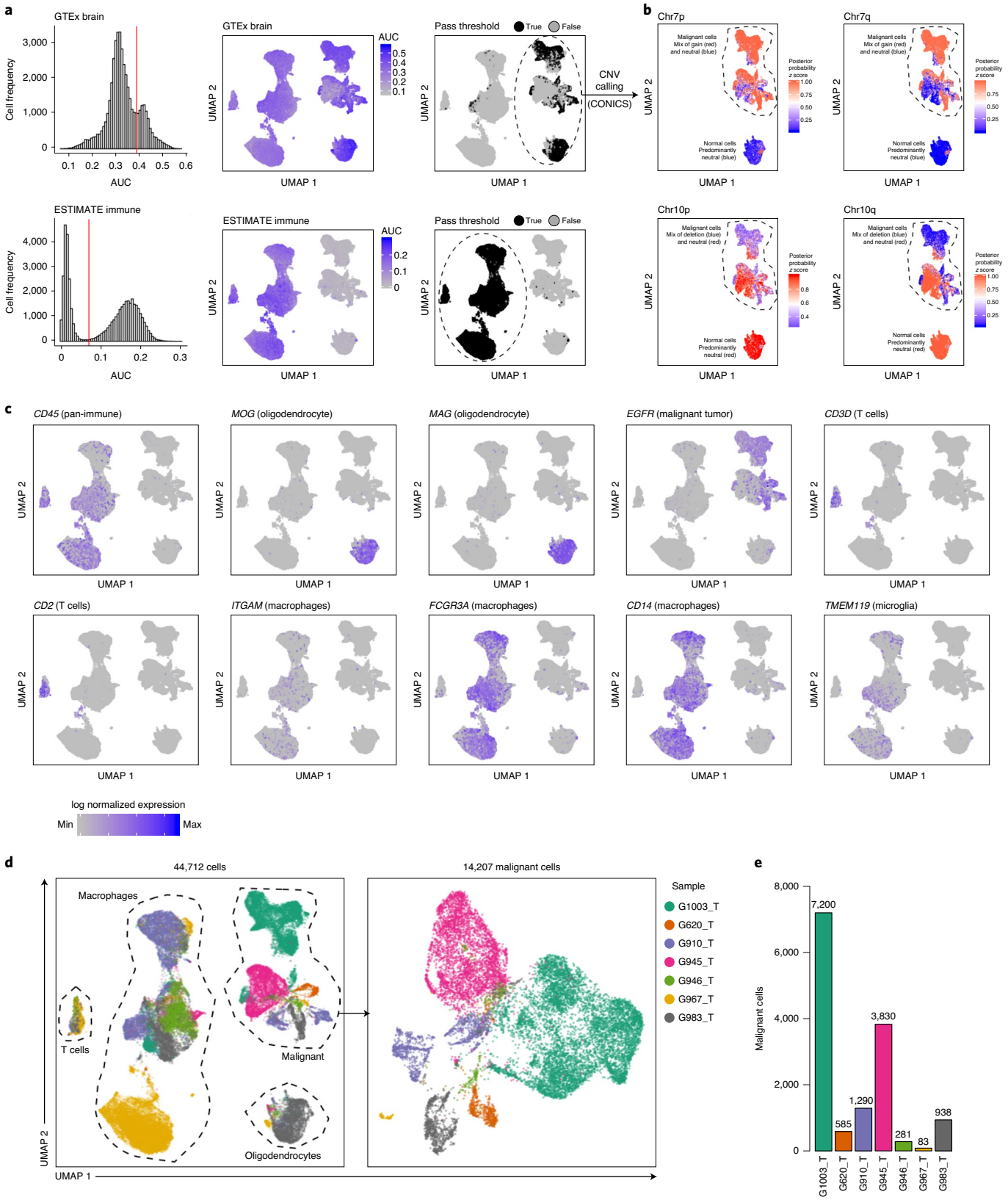
We further interpreted our two gradients in the context of previously described cell types in adult and pediatric GBM[2]. We projected GSCs and tumor cells onto a cellular state map consisting of NPC, OPC, astrocyte-like and mesenchymal-like quadrants (Extended Data Fig. 9b). GSCs were capable of recapitulating all four cell states found in patient tumors. Developmental GSCs commonly mapped to astrocyte-like/OPC/NPC cell states, whereas Injury Response GSCs mapped predominantly to a mesenchymal-like state. Patient tumor cells were predominantly astrocyte-like, confirming the phenotypes observed in our differentiation trajectory (Extended Data Fig. 9c). Together these findings demonstrate that, despite culture conditions and lack of microenvironment, GSCs mirror cell types found in primary tumors and represent a major transcriptional axis underpinning GBMs.

GSC gradient between Developmental and Injury Response is recapitulated in primary tumors. Although discovered in GSCs, primary tumor cells also organize along the transcriptional gradient (Fig. 8a). Tumor cells resembled the Developmental state more often, however Injury Response-like tumor cells were visible in every tumor (Fig. 8b–d). To validate the presence of rare Injury Response GSCs in a larger cohort, we profiled an additional ten patient tumors (42,334 of 53,853 nuclei were malignant) using single-nuclei RNA-seq (snRNA-seq) (Fig. 8a and Extended Data Fig. 9d–f) and analyzed four public GBM sc/snRNA-seq datasets[2,8–10] (52 tumors; 49,018 malignant cells per nuclei) (Supplementary Table 10; Methods). Across all datasets, Developmental and Injury Response programs were anti-correlated (mean Pearson's $r = -0.70$; Fig. 8e), mirroring patterns observed in our original discovery cohort. Tumor cells spanned the complete range of phenotypes discovered in our GSCs, including rare Injury Response-like tumor cells (Fig. 8a,e). The presence of fewer Injury Response-like cells relative to Developmental-like cells in primary tumors could be the result of hindered differentiation capacity, limiting contribution of cells to the tumor bulk[24]. Thus, our panel of GSC lines successfully acts as a model to help explain global expression patterns in GBMs, including rare tumor-initiating cell types.

To determine whether tumor cells harbor CNVs of their matched GSC states, we categorized tumor cells as Developmental or Injury Response-like based on the upper quartile of respective transcriptional program scores (Extended Data Fig. 10a). Next, we identified tumor cells harboring at least one Developmental ($chr6q^-$, $chr9p^+$, $chr19p^+$) or one Injury Response ($chr6q^+$, $chr9p^-$, $chr19p^-$) CNV. Developmental and Injury Response-like tumor cells were significantly enriched for their corresponding state-specific

Fig. 6 | Classification of malignant cells in GBM tumors. a, Gene signature scoring and classification of cells into broad brain and immune lineages. Distribution of AUCell scores across cells. Vertical red line represents the classification threshold. Cells with an AUC value greater than the threshold were determined to be active for a given gene signature (left). UMAP visualization of cells colored by AUC (middle) and whether they are active (black) for a given gene signature (right) ($n = 44{,}412$ cells from seven tumors). **b**, UMAP subsetted by cells classified as being of brain origin. Cells colored by scaled posterior probability from CONICS single-cell CNV inference tool for select chromosome arms. Higher probability (red) represents a cell likely belonging to the Gaussian mixture model component with a higher expression mean ($n = 44{,}412$ cells from seven tumors; same as in **a**). **c**, Expression of pan-immune (*PTPRC/CD45*), macrophage (*ITGAM/CD11B*, *FCGR3A/CD16A*, *CD14*), microglia (*TMEM119*), T-cell (*CD2*, *CD3D*), oligodendrocyte (*MOG*, *MAG*) and putative tumor cell (*EGFR*) markers ($n = 44{,}412$ cells from seven tumors; same as in **a**). **d**, Clustering of 44,712 cells from patient GBM tumors. Cells are colored by patient and annotated by cell type (left). Re-clustering of malignant cells only (right; $n = 14{,}207$ cells), colored by patient. **e**, Quantification of malignant cells across patients, totaling 14,207 cells.

CNVs compared to tumor cells with lower transcriptional scores (Developmental $P<0.0001$, Injury Response $P<0.0001$; chi-squared test). Individual tumor cells rarely harbored CNVs from both Developmental and Injury Response states ($n=658$ of 14,207 cells; 4.6%), suggesting that these may be mutually exclusive events and that, in addition to transcriptional programs, tumor cells inherit genetic alterations of their founder GSCs (Extended Data Fig. 10b). These results further support the potential for CNVs to



influence GSC and subsequent tumor cell transcriptional state, although further validation is needed beyond the seven patients' tumors available in this cohort.

A fraction of primary tumor cells resembling GSCs were evident at the intersection of the Developmental–Injury Response and differentiation gradients. To characterize candidate stem-like cells within patient tumors more precisely, we trained a logistic regression classifier to find GSC-like tumor cells (Extended Data Fig. 10c; Methods). In agreement with the PCA, 2,062 GSC-like tumor cells were found in the overlapping region between GSCs and tumor cells. Every tumor contained a fraction of cells resembling GSCs (median 14%) (Extended Data Fig. 10d). Notably, the tumor with the highest proportion of GSC-like cells was the only IDH1 mutant (p.R100Q, G620_T) in the cohort (Fig. 8d). IDH1 mutations promote convergence toward a proneural phenotype[55], similar to what we term 'Developmental', potentially explaining the increased overlap with Developmental GSCs. Compared to the differentiated tumor bulk, GSC-like tumor cells have upregulated expression of stemness genes (for example *SOX4*, *SOX11*, *STMN1)* that overlap with markers of our GSC gradient (Extended Data Fig. 10e–j). These data demonstrate that substantial overlap exists between GSCs cultured from patient tumors and GSCs found directly within surgical GBM samples.

Discussion

Single-cell profiling of adult and pediatric GBMs has characterized the diverse landscape of cellular states and genetic abnormalities present across and within individual tumors[2,3,8–10]. However, the fundamental source of this heterogeneity remains unclear. In this study, we comprehensively characterized cellular phenotypes of purified GSCs at the root of gliomagenesis using a combination of scRNA-seq and genome-wide CRISPR screening. We verified these phenotypes using sc/snRNA-seq of primary tumors and defined the relationship between GSCs and bulk progeny tumor cells.

While GSCs from each patient were composed of multiple transcriptionally and genetically distinct subpopulations, all GSCs converged on a single biological axis, spanning two recurrent cell states defined by neurodevelopmental and inflammatory programs. Previously, GSC subtypes have been interpreted using the proneural and mesenchymal classifications derived from bulk RNA-seq of GBM tumors[5,6,56,57] or based on similarity to neural subtypes found in normal or fetal brain development[10]. In contrast, our analyses suggest that both neural developmental and wound response programs account for a large portion of heterogeneity in GSCs and that plasticity could be mediated, in part, through cytokine signaling. Our results support a model centered around brain tumor stem cell development where transcriptional heterogeneity in GBMs can be explained by a combination of phenotypic gradients; a GSC gradient between regenerative and wound response programs and a bulk GBM gradient between stem-like and astrocyte-like differentiated cells.

In response to invasive brain injuries, such as stab wounding or ischemia, astrocytes are known to increase proliferation and reactivate stem cell potential as a part of reactive astrogliosis[58,59]. The strong correlation between reactive astrocyte expression signatures and the Injury Response phenotype suggests that these GSCs may arise under similar conditions as reactive astrogliosis, such as hypoxia or neuroinflammation, both common features of the tumor microenvironment in GBMs. We demonstrated that Developmental GSCs can be converted to a more Injury Response-like phenotype following exposure to inflammatory cytokines. Although initially discovered in our in vitro model of GSCs, the Injury Response state was also observed in primary tumors, suggesting that this state could arise via interactions with activated microglia[42] and act as a neurodevelopmental driver via growth factor based cell–cell communication. We cannot, at this stage, exclude whether Injury Response programs could arise autonomously in cells and further understanding of deviation from a Developmental state requires additional experiments.

The presence of GSC state-specific CNVs suggests that the position on the Developmental–Injury Response gradient may be influenced by early somatic alterations. Established founder somatic copy-number alterations (chromosomes 7 and 10) may be responsible for the malignant transformation of astrocyte-like NSCs to GSCs[25,26] with less-prevalent CNVs (19p, 6q, 9p) influencing the Developmental–Injury Response gradient position at which each GSC begins generation of bulk tumor. This creates a framework to further explore the influence of somatic variants and mutations on cellular states in the stem-like compartment of GBM and resultant heterogeneity in patient tumors. One model could be the acquisition of somatic alterations in pre-GSC development cells that lie dormant until subject to injury, thereby triggering differentiation toward an Injury Response state that is redirected toward generation of abnormal, bulk cancer cells.

In conclusion, our observations have two important consequences. First, we may be able to explain GBMs across patients by a single biological model that involves combined mixtures of inflammatory wound-healing cells and NPC/OPC-like cells that cause aberrant neural growth. We hypothesize that GBM forms as a response to neural tissue wounding in the context of a mutated genomic background and that the output of this process is the dual generation of a brain growth and repair response that is derived from genetically abnormal brain precursor cells. This tissue regeneration-oriented interpretation contrasts with previous[2,10] studies and the traditional cancer stem cell discourse that emphasizes cancer stem cell roots solely in a developmental stem cell paradigm. Second, the heterogeneity we have discovered at the GSC level suggests that therapies must be developed to simultaneously target both developmental and inflammatory processes observed in GBMs and GSCs. Further, our CRISPR screens directly identify a range of targetable sensitivities within this GBM-generating biological program. This paradigm may help identify new approaches to treating GBMs.

Methods

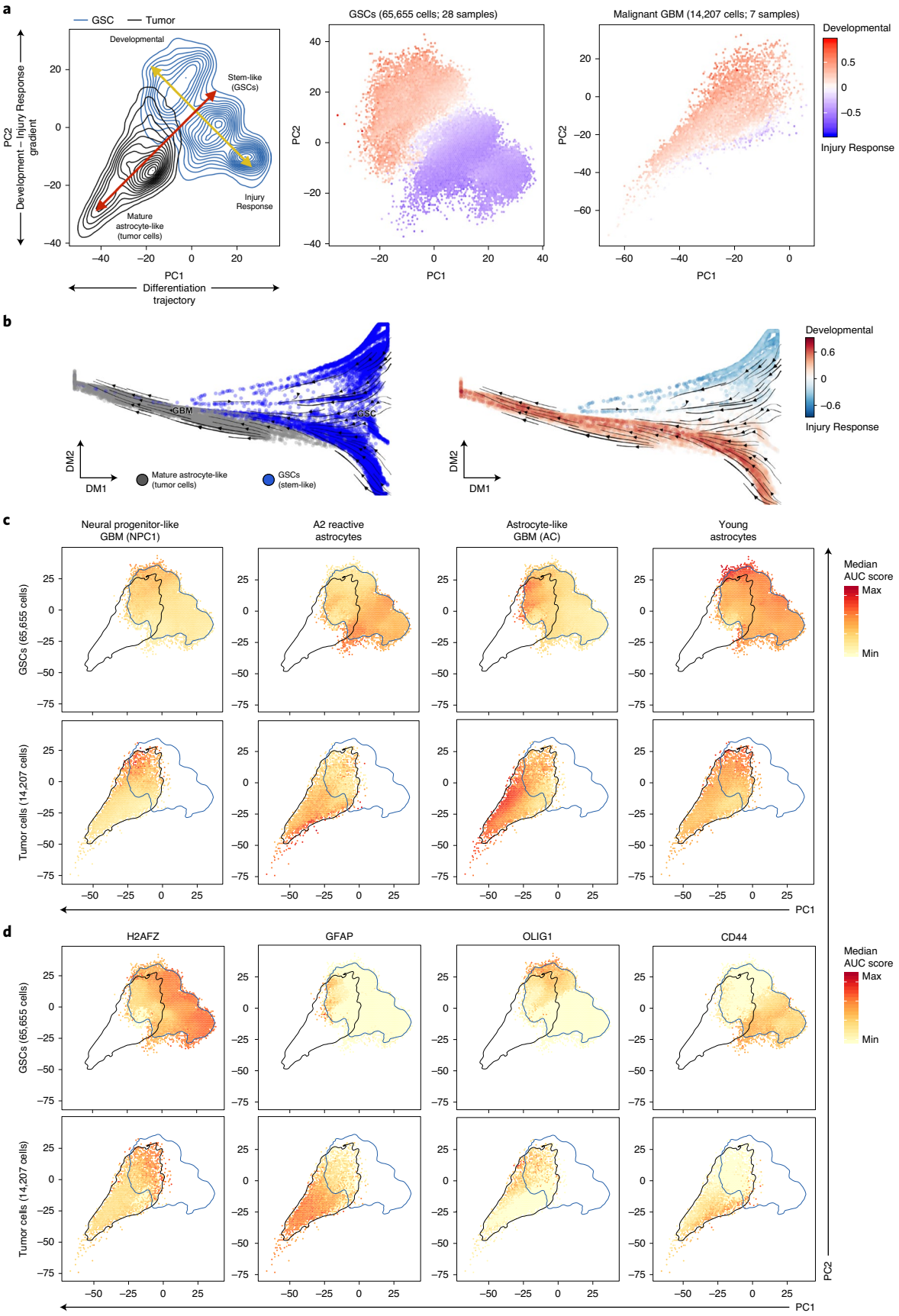
Patient samples and derivation of GSC cultures. All samples were obtained following informed consent from patients. All experimental procedures were performed in accordance with the Research Ethics Board at The Hospital for Sick Children (REB1000025582, REB0020010404), the University Health Network, the University of Calgary Ethics Review Board and the Health Research Ethics Board of Alberta, Cancer Committee and Arnie Charbonneau Cancer Institute Research Ethics Board (REB HREBA-CC-160762).

Patient-derived GSC primary cell lines were derived as either adherent (denoted 'G###_L') or free-floating sphere (denoted 'BT###_L') cultures from

Fig. 7 | Heterogeneity in GBMs is defined by two transcriptional axes. a, PCA of 79,862 cells highlights overlap between GSCs (blue; $n = 65{,}655$ cells) and malignant GBM tumor cells (black; $n = 14{,}207$ cells) (left). GSCs (middle) and tumor cells (right) are colored by expression of Developmental (red) and Injury Response programs (blue). The GSC transcriptional gradient is represented by a yellow arrow and the astrocyte maturation gradient is represented by a red arrow. **b**, Velocity field superimposed on Diffusion Map embeddings of a subset of 20,343 cells from Fig. 7a (maximum 500 cells per sample, randomly selected). Cells are colored by cell type (left) and difference in Developmental and Injury Response scores (right). **c**, Visualization of top-scoring cell-type signatures that are most descriptive of GSC or tumor cell populations. PCA plots binned into hexagons (hexbins). Hexbins represent median AUC score of all overlapping cells within a given coordinate. Contour lines represent an outline of GSC (blue) and tumor cell (black) data points on the PCA plot. **d**, Visualization of select top- and bottom-loading PC1 and PC2 genes. Hexbins represent median normalized gene expression of all overlapping cells within a given coordinate. Contour lines represent an outline of GSC (blue) and tumor cell (black) data points on the PCA plot.

tumor suspensions. GSCs were cultured in serum-free self-renewal medium. Detailed culture conditions are described in Supplementary Note 1. All assays were completed with cultures between passage P8–P12.

Proliferation assays. Cells were plated in equal numbers in a 24-well plate: triplicate wells for technical replicates and in four biological replicates of each technical triplicate. Each set of technical triplicates was lifted and absolute cell



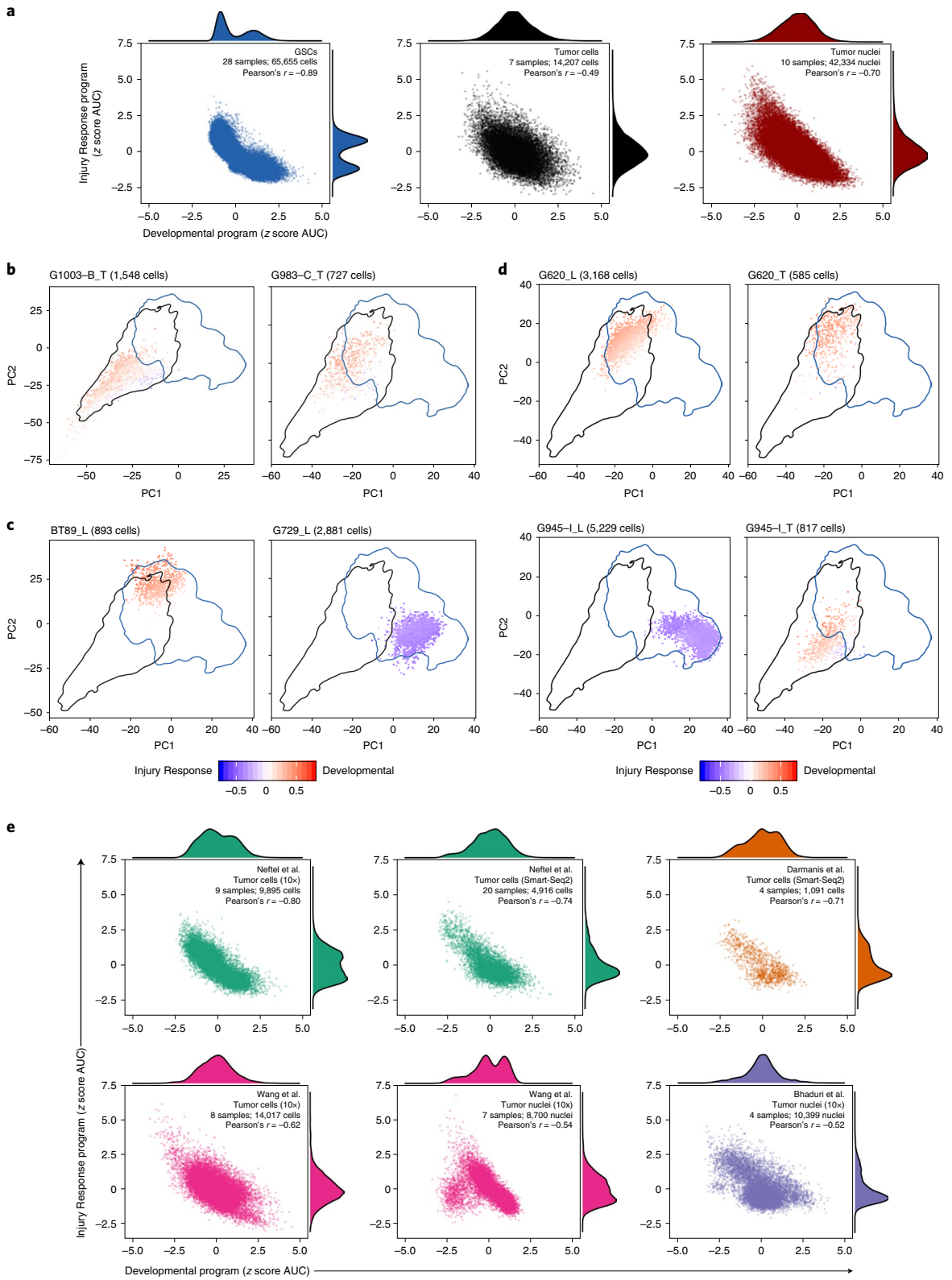


Fig. 8 | GSC transcriptional states are reflected in patient tumors. a, Scoring of individual GSCs (blue; $n = 65{,}655$ cells from 28 GSC cultures) and tumor cells profiled by scRNA-seq (black; $n = 14{,}207$ from seven tumors) or snRNA-seq (dark red; $n = 42{,}334$ cells from ten tumors) for Developmental (*x* axis) and Injury Response (*y* axis) transcriptional programs. **b**–**d**, Distribution of cells from select tumors (**b**), GSCs (**c**) and matched GSC–tumor pairs (**d**) on a PCA plot (related to Fig. 7a). Cells are colored by median expression of Developmental (red) and Injury Response programs (blue) and grouped into hexbins. Contour lines represent outline of GSC (blue) and tumor cell (black) data points on PCA plot. **e**, Projecting malignant cells from four public GBM sc/snRNA-seq datasets recapitulates (cumulative $n = 49{,}018$ cells or nuclei from 52 tumors) the Developmental to Injury Response gradient. Visualization of scaled Developmental (*x* axis) and Injury Response (*y* axis) program scores across malignant cells from multiple public datasets.

number was quantified at several discrete time points over culture. Population doubling time was calculated over exponential phase of growth using the calculation: $(t2-t1)/3.32\times(\log n2-\log n1)$, where t = time and n = number of cells.

Intracranial GSC xenografts. Six- to 16-week-old female NOD/scid gamma or CB17/SCID mice (Charles River Laboratories) were orthotopically transplanted with GSCs for survival studies. A total of 100,000 cells dissociated to a single-cell suspension were transplanted into the right striatum or at the following coordinates: 1 mm anterior of bregma, 2 mm to the right of the midline and 3 mm deep. Mice were housed in groups of three to five and maintained on a 12 h light–dark schedule with a temperature of 22 ± 1 °C and relative humidity of 50 ± 5%. Food and water were available ad libitum. All attempts were made to minimize handling time during surgery and treatment so as not to unduly stress the animals. Animals were observed daily after surgery to ensure there were no unexpected complications. All animal protocols described in this study were approved by the Animal Care Committee at the Hospital for Sick Children and the University of Calgary, operating under the Guidelines of the Canadian Council on Animal Care. All animal work procedures were in accordance with the Guide to the Care and Use of Experimental Animals published by the Canadian Council on Animal Care and the Guide for the Care and Use of Laboratory Animals issued by the National Institutes of Health.

Limiting dilution assays. GSCs grown adherently were plated as serial dilutions on nonadherent 96-well plates with the highest density at 2,000 cells per well and the lowest at 2 cells per well. Each cell dose was plated in six technical replicates. GSCs grown as neurospheres were seeded in 100 µl of medium into the inner 60 wells of a 96-well plate at ten cell densities, as serial dilutions from 512 cells to 1 cell per well, with six replicate wells per cell density. Each LDA plate was counted as one technical replicate. After plating, LDA plates were incubated at 37 °C and 5% CO_2 for 14 or 21 d when all wells were scored for the presence or absence of spheres. The SFC was calculated using Extreme Limiting Dilution Analysis software[61]. Three biological replicates from each GSC culture were plated.

Cytokine treatment and RT–qPCR. GSCs were seeded at a density of 350,000 cells per well into six-well plates coated with poly-L-ornithine and laminin. After 24 h in NS medium, fresh medium containing vehicle or cytokines was added, with final concentrations as follows: TNF-α (30 ng µl^{-1}), C1q (400 ng µl^{-1}) and IL-1α (3 ng µl^{-1}). Cell pellets were collected after 48 h of treatment and stored at −80 °C until RNA extraction. RNA was extracted from cells using RNeasy Mini kit (QIAGEN). The Transcriptor First Strand cDNA Synthesis kit (Roche) was used to reverse transcribe 1 µg of RNA. Quantitative PCR was performed using SsoFast EvaGreen Supermix (BioRad) and the CFX Connect Real-Time PCR detection system (BioRad). Primers are listed in Supplementary Table 11.

Single-cell and single-nuclei RNA-seq. *Generation of single cell and nuclei suspensions.* We generated single-cell suspensions from viably cryopreserved, dissociated GSC lines by thawing and resuspending in a solution of PBS and BSA. For patient GBM tumors, high-quality single-cell suspensions were generated by dissociating biopsied tissues in accutase and DNase. Post-dissociation red blood cells (RBC lysis solution, Miltenyi) and cellular debris from damaged cells (Miltenyi) were removed. We generated single-nuclei suspensions from snap-frozen tumors. Tissues were minced on dry ice and dissolved in lysis buffer (0.32 M sucrose, 5 mM $CaCl_2$, 3 mM $Mg(Ac)_2$, 20 mM Tris-HCl (pH 7.5), 0.1% Triton-X-100, 0.1 mM EDTA (pH 8.0)), followed by homogenization with a pellet pestle. Nuclei integrity and quantity was assessed with SYBR Green II RNA Gel stain (Thermo Fisher Scientific). Nuclei were filtered through a 40-µm cell strainer and sorted for intact nuclei using DAPI (Sigma-Aldrich) on a BD Influx FACS sorter. Using a hemocytometer, nuclei or cells were re-suspended according to 10X Genomics concentration guidelines to obtain a target of 2,000–6,000 nuclei per sample. Cells had a minimum final viability of 70%.

Library preparation and sequencing. Library preparation was carried out as per the 10X Genomics Chromium single-cell protocol using the v2 chemistry reagent kit. Cell or nuclei suspensions were loaded onto individual channels of a Chromium Single-Cell Chip along with reverse transcription master mix and single cell 3′ gel beads. Complementary DNA underwent a two-stage purification process with Dynal MyONE Silane beads (Thermo Fisher Scientific), followed by SPRISelect beads (Beckman Coulter). Libraries were sequenced on an Illumina 2500 in High Output mode using the 10X Genomics recommended sequencing parameters. Samples were quantified by KAPA Library Quantification kit (Roche) and normalized to achieve the desired median read depth per cell (target mean 60,000 reads per cell).

Single-cell and single-nuclei RNA-seq data pre-processing. We used the 10X Genomics CellRanger software pipeline (v.2) to demultiplex cell barcodes and map reads to the GRCh38 human reference transcriptome using STAR aligner. snRNA-seq data were aligned to a custom GRCh38 pre-mRNA reference transcriptome that included intron sequences to accurately quantify nuclear unspliced messenger RNA. We calculated the number of reads per cell barcode using the BamTagHistogram function in the Drop-seq Alignment Cookbook[62]. We determined the number of cells per sample using the cumulative fraction of reads corresponding to cell barcode in a library. Cell barcodes were sorted in decreasing order and the inflection point was identified using the R package Dropbead[63] (v.0.3.1) to distinguish between empty droplets and droplets containing a cell. The raw matrix of gene counts versus cells from CellRanger (v.2) output was filtered by the list of true cell barcodes from Dropbead. We processed the resultant unique molecular identifier (UMI) count matrix using the R package Seurat[64,65] (v.2.3.4) as described below and defined detected genes as those with >0 UMIs.

Data filtration. We discarded cells with >4 median absolute deviations, up to a maximum of 40%, of UMI counts belonging to expressed mitochondrial genome genes, potentially indicative of damaged cells with compromised cellular membranes. Probable cell multiplets were removed if log-library size or log-genes detected were more than 3 median absolute deviations above the median. Low-quality cells with fewer than 350 genes detected were also removed. We removed lowly expressed genes detected in fewer than 1% of cells in a sample. Quality control metrics are outlined in Supplementary Table 2.

Data normalization. Expression normalization was performed using the LogNormalize() function in Seurat. To adjust for differences in library size and cell cycle, we regressed on the number of UMIs, mitochondrial content and cell-cycle difference (described below) using a linear model during gene scaling and centering. Expression values were scaled across all samples and cells in a given dataset. Scaled z score residuals ('relative expression') were used for dimensionality reduction and clustering. For visualizations, we clipped relative expression to the range (−2.5, 2.5) to prevent outliers from dominating the scale.

Adjusting for cell-cycle signal. To preserve biological signal separating cycling and noncycling cells, while removing uninteresting differences in cell cycle, we used the 'Alternate Workflow' in Seurat. (https://satijalab.org/seurat/v2.4/cell_cycle_vignette.html; Supplementary Note 1). First, we assigned cell-cycle scores to individual cells on the basis of expression of previously published G2/M and S-phase gene signatures[33], using the CellCycleScoring() function. Cells expressing neither G2/M nor S-phase marker genes were assigned to G1. Next, we calculated the difference between S-phase and G2M-phase scores for each cell to give a 'Cell Cycle Difference Score' and regressed the difference in phases with a linear regression model as described above.

Dimensionality reduction. PCA was conducted on all expressed genes, excluding ribosomal transcripts. Significant principal components, as determined by the inflection point in a scree plot, were used as inputs for nonlinear dimensionality reduction techniques (t-SNE and UMAP), as well as cell clustering. Diffusion Map[66] was performed on the same subset of genes as PCA using the RunDiffusion() implementation in Seurat. Due to memory constraints, Diffusion Map was run on a subset of cells by randomly downsampling each sample to a maximum of 500 cells.

Clustering and visualization. To identify intra-GSC and inter-GSC clusters, we performed iterative SNN-Cliq-inspired clustering on significant principal components using a smart local moving algorithm as implemented in Seurat with a range of resolutions from 0.1 to 1. The R package scClustViz[67] (v.1.2.1) was used to perform differential expression testing (Wilcoxon rank-sum test, FDR < 0.05) between clusters for all resolutions to assess the biological relevance of each cluster solution. Genes with a detection rate difference between clusters of 0.15 or greater were included in differential testing. To select the optimal resolution, we selected the clustering solution with the greatest silhouette value from all solutions with a median of >20 DE genes per cluster. We performed benchmarking against four other clustering algorithms to assess reproducibility of our clustering solutions (details in Supplementary Note 1). Clusters were visualized using t-SNE and UMAP.

RNA velocity. Briefly, sorted BAM files were run through Velocyto[68] (v.0.17.13) to generate spliced and unspliced counts. Counts for cells with Diffusion Map coordinates were filtered, normalized and log-transformed using scvelo[69] (v.0.2.2), and visualizations were generated in scvelo. Further details are provided in Supplementary Note 1.

Single-cell gene signature scoring and pathway analysis. Gene signature activity in single cells, with the exception of cell-cycle stage, was quantified using AUCell[70] (v.1.4.1). Gene signatures were curated from the literature and derived from differentially expressed genes between the Injury Response and Developmental clusters identified by bulk RNA-seq (Supplementary Table 7 and Supplementary Note 1). When directly comparing the difference of AUCell scores between two gene signatures, such as in Fig. 7a, AUCell scores were normalized between (0,1) by subtracting the minimum and dividing by the range.

Marker genes and principal component top/bottom-loading gene lists were annotated using over-representation analysis in clusterProfiler[71] (v.3.10.1), with a

q-value cutoff of 0.01 after multiple comparison adjustment with the Benjamini–Hochberg procedure (Supplementary Note 1). The enrichment map was generated using genes ranked by the negative of PC1 loadings as described in Supplementary Note 1 (Extended Data Fig. 6d).

Single-cell CNV analysis. CNVs were called from scRNA-seq data using inferCNV (v.0.3, https://github.com/broadinstitute/infercnv). CNVs were estimated by sorting expressed genes by genomic location and averaging relative expression of genetically adjacent genes using a sliding window of 100 genes. Resultant expression levels were compared to a reference panel of 600 normal, diploid oligodendrocyte cells from six primary tumors. Individual CNV scores were averaged across intra-GSC clusters to visualize transcriptional clusters with unique CNV profiles in Fig. 2b. To validate the accuracy of our single-cell CNV calls, we compared inferCNV scores and WGS CNV $\log_2$ ratios at the gene level for a cohort of 20 GSCs profiled with both technologies. Discrete inferCNV cutoffs that define single copy gain (0.17) or loss (−0.15) were determined using the median inferCNV score of genes deleted or gained by GISTIC[72] (v.2.0.23) on matched WGS data (Extended Data Fig. 3a–c).

CNV enrichment analysis. To assess how inferred CNVs may influence marker gene profiles of GSC clusters, we first identified GSCs with variable CNVs across chromosome arms by binning loci into deletion, neutral and gain bins using inferCNV score cutoffs as described in Extended Data Fig. 3d. We then assessed the proportion and enrichment of cluster marker genes that reside within CNV loci that are variable between clusters using a Fisher's exact test (Extended Data Fig. 3e).

To identify CNVs specific to Developmental versus Injury Response-like GSCs, we averaged CNV signals of all genes across chromosome arms for each cell. Chromosome arms with <50 expressed genes were excluded. Next, we classified cells as being either Developmental-like or Injury Response-like using gene signature scoring. We excluded hybrid cells, defined as cells scoring as positive or negative for both states. We then compared the intensity of CNV signal, represented by inferCNV scores, between Developmental-like or Injury Response-like cells across chromosome arms. Variably altered regions between GSC subtypes were identified using effect size (large magnitude, Hedge's $g \geq 0.8$; Fig. 5b).

Identification of malignant cells in patient GBMs. To discern tumor cells from normal cells, we used a three-step approach involving unbiased clustering, CNVs and expression of cell-type specific marker genes. First, we used UMAP to visualize all cells in the same transcriptional space. Second, we classified cells as being of 'brain' or 'immune' origin using gene signature scoring (Supplementary Note 1). We identified malignant tumor cells within the brain fraction using single-cell CNV inference (Supplementary Note 1). Finally, we validated our cell-type annotations with expression of canonical cell-type marker genes for immune, macrophage, microglial, T-cell, oligodendrocyte and putative tumor cells (Fig. 6).

Re-analyzing public sc/snRNA-seq datasets. Whenever possible, we used cell annotations provided in the publications to label cells (for example, tumor, immune, oligodendrocytes). In the absence of annotations, we re-processed the data using our clustering pipeline as described above. Malignant cells were then identified using a combination of unbiased clustering, marker gene expression and scaled expression of genes on chromosome 7 and 10 as a proxy for CNVs. Normalized gene expression matrices were used for gene signature scoring.

Projection onto GBM cell-state map. The two-dimensional cell-state representation map was created as described by Neftel et al.[2] (Extended Data Fig. 9b). Cells were scored for cell-state gene signatures using the AddModuleScore() function in Seurat. NPC1/2 and MES1/2 scores were averaged to represent one score each for NPC and mesenchymal states. Cells were then separated into OPC/NPC and astrocyte/mesenchymal lineages by the sign of $D = \max(SC_{OPC}, SC_{NPC}) - \max(SC_{AC}, SC_{MES})$; where SC represents the transcriptional program score, D represents the y axis value, AC represents astrocytes and MES represents mesenchymal cells. Next, for OPC/NPC cells ($D > 0$), the x axis value was computed as $\log_2(SC_{OPC} - SC_{NPC} + 1)$ and for astrocyte/mesenchymal cells ($D < 0$), the x axis was computed as $\log_2(SC_{AC} - SC_{MES} + 1)$.

Identification of GSC-like tumor cells with a logistic regression classifier. The scRNA-seq dataset consisting of all tumor and cultured GSC cells minus all G800_L cells was split into an 80% training set and a 20% test set, with the split stratified by the two classes (tumor and GSC). The first two principal components (Fig. 7a) were used as inputs to be mapped to labels. Hyperparameter optimization, model selection and final class predictions were performed as described in Supplementary Note 1.

Bulk RNA-seq. *Library preparation and sequencing.* RNA was extracted from frozen cell pellets using the AllPrep DNA/RNA Universal kit (QIAGEN). Strand-specific sequencing libraries were prepared from 500 ng total RNA using poly(A) capture of transcripts with the NEBNext Poly(A) mRNA Magnetic Isolation Module (E7490L, NEB). Libraries were quantified with the Qubit dsDNA HS Assay kit (Thermo Fisher Scientific). Clusters were generated on the Illumina cluster station and sequence was run on the Illumina HiSeq2500 (indexed lane using V4 chemistry) platform following the manufacturer's instructions.

Data pre-processing and clustering. Strand-specific 75-bp paired-end reads were aligned to hg38 reference using STAR[73] (v.2.4.2a) and annotated with University of California Santa Cruz (UCSC) source from the Illumina iGenome reference. The 'ReadsPerGene' raw counts from STAR were used for downstream analysis. Genes were filtered for those with at least five counts across all samples. DESeq2 (ref. [74]) (v.1.22.2) was used to calculate size factors for each sample and perform variance stabilizing transformation. Batch correction was performed to incorporate technical and biologically relevant features into the model (Supplementary Note 1).

Variance stabilizing-transformed bulk RNA-seq data for 72 GSC lines were used as inputs for clustering. We assessed the similarity of clusters obtained on random subsamples of data to a full data clustering solution across a range of clusters ($k = 2$–4 clusters) using an adjusted Rand index and spectral clustering (Supplementary Note 1).

Differential gene expression and pathway analysis. Differential gene expression analysis was carried out on count data using DESeq2 (ref. [74]) incorporating batch status as a covariate in the expression model. Developmental and Injury Response signatures were defined as upregulated genes (FDR < 0.05, two-sided Wald test) in the corresponding clusters identified above (Supplementary Table 7 and Extended Data Fig. 5c,d). GSVA[75] (v.1.30.0) was used to assess the activity of gene signatures across samples. Gene set enrichment analysis (GSEA)[76] (v.3.0) was performed on genes ranked by differential expression. Full details on GSEA analysis and data visualization with EnrichmentMap are provided in Supplementary Note 1.

Whole-genome sequencing. DNA was extracted from frozen cell pellets using the AllPrep DNA/RNA Universal kit (QIAGEN) and whole blood samples using the QiaAmp DNA Blood Midi kit (QIAGEN). Illumina-compatible sequencing libraries were constructed from 500 ng gDNA using TruSeq DNA PCR-free kits (New England Biolabs) and sequenced with paired-end 150-base reads on the Illumina HiSeqX platform to a median depth of 60× for GSCs and blood normals. Sequence data quality checks were performed with FastQC (v.0.11.5) and aligned to the human reference genome hg38 with bwa[77] (v.0.7.15). A detailed description of additional pre-processing, quality control and copy-number calling is provided in Supplementary Note 1.

Genome-wide CRISPR–Cas9 screens. We performed CRISPR–Cas9 screens using the 70-k TKOv3 library[49] (Addgene, 90294) using previously established protocols[51] with cells cultured as described above. A minimum of 8×10^7 cells were transduced with gRNA library-expressing lentivirus in the presence of 0.8 µg polybrene at a multiplicity of infection of 0.3. At 24 h after transduction, lentiviral medium was removed and cells were cultured with 2 µg ml^{-1} puromycin for 48–72 h to select for integration of lentiviral cassette. After selection, surviving cells were pooled and T0 samples of a minimum of 1.5×10^7 cells were collected and frozen at −80 °C for gDNA extraction. The remaining cells were then divided into 2–3 replicates of 1.5×10^7 cells and cultured for 14-cell doublings under standard culture conditions, maintaining a minimum of 1.5×10^7 cells per replicate at all times (~200-fold library coverage). At time points of approximately 10- and 14-cell doublings, we collected cell pellets of 1.5×10^7 cells and stored them at −80 °C for gDNA extraction. A detailed description of gDNA extraction, library preparation and sequencing is provided in Supplementary Note 1.

Analysis of genome-wide screen data. DNA sequencing reads for each CRISPR screen were mapped to TKOv3 library gRNAs and normalized for sample sequencing depth. We calculated the fold change for each gRNA from the T0 baseline and computed a qBF for each gene representing a confidence score that gene knockout produces a fitness defect (Supplementary Table 8). Additional details regarding quality control and gRNA complexity are provided in Supplementary Note 1.

To identify differentially essential genes, the difference in average qBF scores between Injury Response and Developmental GSC screens was calculated for each gene. The resulting differences were transformed into z scores and a cutoff of $>|2|$ was used to identify essential genes in each respective GSC state (Supplementary Table 9). Pathway analysis was performed on the ranked gene list generated by calculating the difference in average qBF scores between Injury Response and Developmental GSC screens and visualized with EnrichmentMap as described in the Supplementary Note 1.

Competitive proliferation assays. For validation of gene knockouts producing fitness defects, Cas9-expressing GSCs were first engineered via lentiviral transduction as previously described[51]. Cas9-expressing GSCs were then transduced with either Lentiguide-gRNA-NLS-eGFP-2A-PURO targeting specific genes of interest or Lentiguide-gRNA-NLS-mCherry-2A-PURO constructs targeting the *AAVS1* locus. Each gene was targeted with two unique gRNAs. At 24 h after transduction, cells were selected with 2 µg ml^{-1} puromycin for 48–72 h. Co-culture competitive proliferation assays were set up by mixing approximately 50,000 red cells (nls-mCherry gRNA-AAVS1) and 50,000 green cells (nls-eGFP

gRNA-gene of interest). One half of this mixture was seeded in a six-well plate and the other half was subjected to flow cytometry using a CytoFlex S (Beckman Coulter) to assess the relative proportion of red and green cells at the start of the experiment. Cells were cultured for 14 d at which point they were collected and subjected to flow cytometry as above to assess the relative proportion of red and green cells. Relative cell fitness was calculated as the percentage of green cells at T14 divided by the percentage of green cells at T0, with normalization to an *AAVS1* versus *AAVS1* competition assay. The following gRNAs were used for gene knockout in competition assays (Supplementary Table 11).

Statistics and reproducibility. No statistical method was used to predetermine sample size. Cells with insufficient library complexity were excluded from the analyses as described in the methods. G800_L was removed as an outlier based on PCA (Extended Data Fig. 5a). Investigators were not blinded to the study of human sequencing data. Plotting and statistical analysis was performed in the R statistical environment (v.3.5.0 and v.3.6.1) and GraphPad Prism (v.8).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Bulk RNA-seq (EGAS00001003070 and EGAS00001004395), WGS (EGAS00001004395), sc and snRNA-seq (EGAS00001004656) datasets generated and analyzed in this study are available through the European Genome-Phenome Archive repository in the form of FASTQ or BAM files. Processed sc and snRNA-seq data are publicly available through the Broad Institute Single-Cell Portal (https://singlecell.broadinstitute.org/single_cell/study/SCP503) and CReSCENT[60] (https://crescent.cloud; study ID CRES-P23). All other data supporting the findings of this study are available from the corresponding author on reasonable request. Original CSV files for Supplementary Tables 2–9 are available in the Supplementary Information. Previously published scRNA-seq data that were re-analyzed in this study are available from the following sources: Wang et al.[9] (GSE138794), Bhaduri et al.[10] (http://cells.ucsc.edu/?ds=gbm), Neftel et al.[2] (https://singlecell.broadinstitute.org/single_cell/study/SCP393/) and Darmanis et al.[8] (http://gbmseq.org/). Source data are provided with this paper.

Code availability

Code necessary to reproduce the core analyses presented in this study are available without restrictions at https://github.com/pughlab/su2c-gsc-scrna.

Received: 3 April 2020; Accepted: 16 November 2020;
Published online: 04 January 2021

References

1. Brennan, C. W. et al. The somatic genomic landscape of glioblastoma. *Cell* **155**, 462–477 (2013).
2. Neftel, C. et al. An integrative model of cellular states, plasticity, and genetics for glioblastoma. *Cell* https://doi.org/10.1016/j.cell.2019.06.024 (2019).
3. Patel, A. P. et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396–1401 (2014).
4. Meyer, M. et al. Single cell-derived clonal analysis of human glioblastoma links functional and genomic heterogeneity. *Proc. Natl Acad. Sci. USA* **112**, 851–856 (2015).
5. Verhaak, R. G. W. et al. An integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in *PDGFRA*, *IDH1*, *EGFR* and *NF1*. *Cancer Cell* **17**, 98 (2010).
6. Wang, Q. et al. Tumor evolution of glioma-intrinsic gene expression subtypes associates with immunological changes in the microenvironment. *Cancer Cell* **32**, 42–56 (2017).
7. Carlsson, S. K., Brothers, S. P. & Wahlestedt, C. Emerging treatment strategies for glioblastoma multiforme. *EMBO Mol. Med.* **6**, 1359–1370 (2014).
8. Darmanis, S. et al. Single-cell RNA-seq analysis of infiltrating neoplastic cells at the migrating front of human glioblastoma. *Cell Rep.* **21**, 1399–1410 (2017).
9. Wang, L. et al. The phenotypes of proliferating glioblastoma cells reside on a single axis of variation. *Cancer Discov.* https://doi.org/10.1158/2159-8290.CD-19-0329 (2019).
10. Bhaduri, A. et al. Outer radial glia-like cancer stem cells contribute to heterogeneity of glioblastoma. *Cell Stem Cell* **26**, 48–63 (2020).
11. Berezovsky, A. D. et al. Sox2 promotes malignancy in glioblastoma by regulating plasticity and astrocytic differentiation. *Neoplasia* **16**, 193–206 (2014).
12. Lan, X. et al. Fate mapping of human glioblastoma reveals an invariant stem cell hierarchy. *Nature* **549**, 227–232 (2017).
13. Natsume, A. et al. Chromatin regulator PRC2 is a key regulator of epigenetic plasticity in glioblastoma. *Cancer Res.* **73**, 4559–4570 (2013).
14. Singh, S. K. et al. Identification of human brain tumour initiating cells. *Nature* **432**, 396–401 (2004).
15. Bao, S. et al. Glioma stem cells promote radioresistance by preferential activation of the DNA damage response. *Nature* **444**, 756–760 (2006).
16. Chen, J. et al. A restricted cell population propagates glioblastoma growth after chemotherapy. *Nature* **488**, 522–526 (2012).
17. Liu, G. et al. Analysis of gene expression and chemoresistance of CD133$^+$ cancer stem cells in glioblastoma. *Mol. Cancer* **5**, 67 (2006).
18. Pollard, S. M. et al. Glioma stem cell lines expanded in adherent culture have tumor-specific phenotypes and are suitable for chemical and genetic screens. *Cell Stem Cell* **4**, 568–580 (2009).
19. Kelly, J. J. P. et al. Proliferation of human glioblastoma stem cells occurs independently of exogenous mitogens. *Stem Cells* **27**, 1722–1733 (2009).
20. Florio, M. et al. Human-specific gene *ARHGAP11B* promotes basal progenitor amplification and neocortex expansion. *Science* **347**, 1465–1470 (2015).
21. Zhang, C.-L., Zou, Y., He, W., Gage, F. H. & Evans, R. M. A role for adult TLX-positive neural stem cells in learning and behaviour. *Nature* **451**, 1004–1007 (2008).
22. Zhu, Z. et al. Targeting self-renewal in high-grade brain tumors leads to loss of brain tumor stem cells and prolonged survival. *Cell Stem Cell* **15**, 185–198 (2014).
23. Ouafik, L. et al. Neutralization of adrenomedullin inhibits the growth of human glioblastoma cell lines in vitro and suppresses tumor xenograft growth in vivo. *Am. J. Pathol.* **160**, 1279–1292 (2002).
24. Park, N. I. et al. ASCL1 reorganizes chromatin to direct neuronal fate and suppress tumorigenicity of glioblastoma stem cells. *Cell Stem Cell* **21**, 209–224.e7 (2017).
25. Körber, V. et al. Evolutionary trajectories of IDH^{WT} glioblastomas reveal a common path of early tumorigenesis instigated years ahead of initial diagnosis. *Cancer Cell* https://doi.org/10.1016/j.ccell.2019.02.007 (2019).
26. Lee, J. H. et al. Human glioblastoma arises from subventricular zone cells with low-level driver mutations. *Nature* **560**, 243–247 (2018).
27. Filbin, M. G. et al. Developmental and oncogenic programs in H3K27M gliomas dissected by single-cell RNA-seq. *Science* **360**, 331–335 (2018).
28. Gojo, J. et al. Single-cell RNA-seq reveals cellular hierarchies and impaired developmental trajectories in pediatric ependymoma. *Cancer Cell* **38**, 44–59 (2020).
29. Hovestadt, V. et al. Resolving medulloblastoma cellular architecture by single-cell genomics. *Nature* **572**, 74–79 (2019).
30. Izar, B. et al. A single-cell landscape of high-grade serous ovarian cancer. *Nat. Med.* https://doi.org/10.1038/s41591-020-0926-0 (2020).
31. Ledergor, G. et al. Single cell dissection of plasma cell heterogeneity in symptomatic and asymptomatic myeloma. *Nat. Med.* **24**, 1867 (2018).
32. Puram, S. V. et al. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell* **171**, 1611–1624 (2017).
33. Tirosh, I. et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).
34. Tirosh, I. et al. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. *Nature* **539**, 309–313 (2016).
35. Ben-David, U. et al. Genetic and transcriptional evolution alters cancer cell line drug response. *Nature* **560**, 325–330 (2018).
36. Kinker, G. S. et al. Pan-cancer single-cell RNA-seq identifies recurring programs of cellular heterogeneity. *Nat. Genet.* **52**, 1208–1218 (2020).
37. Krieger, T. G. et al. Single-cell analysis of patient-derived PDAC organoids reveals cell state heterogeneity and a conserved developmental hierarchy. Preprint at *bioRxiv* https://doi.org/10.1101/2020.08.23.263160 (2020).
38. McFarland, J. M. et al. Multiplexed single-cell transcriptional response profiling to define cancer vulnerabilities and therapeutic mechanism of action. *Nat. Commun.* **11**, 4296 (2020).
39. Nowakowski, T. J. et al. Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science* **358**, 1318–1323 (2017).
40. Zhong, S. et al. A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex. *Nature* **555**, 524–528 (2018).
41. Cahoy, J. D. et al. A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. *J. Neurosci.* **28**, 264–278 (2008).
42. Liddelow, S. A. et al. Neurotoxic reactive astrocytes are induced by activated microglia. *Nature* **541**, 481–487 (2017).
43. John Lin, C.-C. et al. Identification of diverse astrocyte populations and their malignant analogs. *Nat. Neurosci.* **20**, 396–405 (2017).
44. Chai, H. et al. Neural circuit-specialized astrocytes: transcriptomic, proteomic, morphological, and functional evidence. *Neuron* **95**, 531–549 (2017).
45. Morel, L. et al. Molecular and functional properties of regional astrocytes in the adult brain. *J. Neurosci.* **37**, 8706–8717 (2017).
46. Miller, S. J. Astrocyte heterogeneity in the adult central nervous system. *Front. Cell. Neurosci.* **12**, 401 (2018).
47. Haghverdi, L., Buettner, F. & Theis, F. J. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* **31**, 2989–2998 (2015).
48. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).

49. Hart, T. et al. Evaluation and design of genome-wide CRISPR/SpCas9 knockout screens. *G3 Genes Genom. Genet.* **7**, 2719–2727 (2017).
50. Hart, T. & Moffat, J. BAGEL: a computational framework for identifying essential genes from pooled library screens. *BMC Bioinf.* **17**, 164 (2016).
51. MacLeod, G. et al. Genome-wide CRISPR-Cas9 screens expose genetic vulnerabilities and mechanisms of temozolomide sensitivity in glioblastoma stem cells. *Cell Rep.* **27**, 971–986 (2019).
52. Zhou, Y. et al. Metabolic alterations in highly tumorigenic glioblastoma cells preference for hypoxia and high dependency on glycolysis. *J. Biol. Chem.* **286**, 32843–32853 (2011).
53. Lein, E. S. et al. Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445**, 168–176 (2007).
54. Su, L. et al. H2A.Z.1 crosstalk with H3K56-acetylation controls gliogenesis through the transcription of folate receptor. *Nucleic Acids Res.* **46**, 8817–8831 (2018).
55. Philip, B. et al. Mutant *IDH1* promotes glioma formation in vivo. *Cell Rep.* **23**, 1553–1564 (2018).
56. Bhat, K. P. L. et al. Mesenchymal differentiation mediated by NF-κB promotes radiation resistance in glioblastoma. *Cancer Cell* **24**, 331–346 (2013).
57. Xie, Y. et al. The human glioblastoma cell culture resource: validated cell models representing all molecular subtypes. *EBioMedicine* **2**, 1351–1363 (2015).
58. Sirko, S. et al. Reactive glia in the injured brain acquire stem cell properties in response to sonic hedgehog. *Cell Stem Cell* **12**, 426–439 (2013); erratum **12**, 629 (2013).
59. Robel, S., Berninger, B. & Götz, M. The stem cell potential of glia: lessons from reactive gliosis. *Nat. Rev. Neurosci.* **12**, 88–104 (2011).
60. Mohanraj, S. et al. CReSCENT: CanceR single cell ExpressioN toolkit. *Nucleic Acids Res.* **48**, W372–W379 (2020).
61. Hu, Y. & Smyth, G. K. ELDA: extreme limiting dilution analysis for comparing depleted and enriched populations in stem cell and other assays. *J. Immunol. Methods* **347**, 70–78 (2009).
62. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
63. Alles, J. et al. Cell fixation and preservation for droplet-based single-cell transcriptomics. *BMC Biol.* **15**, 44 (2017).
64. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
65. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).
66. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13**, 845–848 (2016).
67. Innes, B. T. & Bader, G. D. scClustViz – single-cell RNAseq cluster assessment and visualization. *F1000Research* **7**, 1522 (2019).
68. La Manno, G. et al. RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
69. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-020-0591-3 (2020).
70. Aibar, S. et al. SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).
71. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS J. Integr. Biol.* **16**, 284–287 (2012).
72. Mermel, C. H. et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).
73. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
74. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
75. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinf.* **14**, 7 (2013).
76. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
77. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

Acknowledgements

Research was supported by Stand Up To Cancer (SU2C) Canada Cancer Stem Cell Dream Team Research Funding (SU2C-AACR-DT-19-15) provided by the Government of Canada through Genome Canada and the Canadian Institute of Health Research, with supplemental support from the Ontario Institute for Cancer Research, through funding provided by the Government of Ontario. SU2C Canada is a Canadian Registered Charity (reg. no. 80550 6730 RR0001). Research Funding is administered by the American Association for Cancer Research International – Canada, the Scientific Partner of SU2C Canada. L.M.R. was supported by an Ontario Graduate Scholarship and the Frank Fletcher Memorial Fund from the University of Toronto. O.K.N.W. was supported by funding from the Canadian Institute for Health Research, the Cecil Yip Doctoral Research Award and the David Stephen Cant Graduate Scholarship in Stem Cell Research from the University of Toronto. Funding to H.A.L. and S.W. was from Canadian Institute for Health Research. T.J.P. holds the Canada Research Chair in Translational Genomics and is additionally supported by the Princess Margaret Cancer Foundation, a Senior Investigator Award from the Ontario Institute for Cancer Research and the Government of Canada through Genome Canada and the Ontario Genomics Institute (OGI-167). Additional infrastructure support came from the Canada Foundation for Innovation, Leaders Opportunity Fund (CFI no. 32383); Ontario Ministry of Research and Innovation, Ontario Research Fund Small Infrastructure Program; Ontario Institute for Cancer Research; the Chan Zuckerberg Initiative; and the Princess Margaret Cancer Foundation. P.B.D. is additionally supported by the Canadian Institutes for Health Research, the Ontario Institute for Cancer Research, the Terry Fox Research Institute, the Hospital for Sick Children Foundation, the Bresler family, Jessica's Footprint Foundation, the Hopeful Minds Foundation and B.R.A.I.N. Child. P.B.D. holds a Garron Family Chair in Childhood Cancer Research at The Hospital for Sick Children. G.D.B. was supported by NRNB (US National Institutes of Health, National Center for Research Resources grant no. P41 GM103504). T.J.P., P.B.D. and G.D.B. are supported by a Canadian Cancer Society Impact Grant. We thank R. Hassam, O. Cseh and I. Restall (University of Calgary) for technical assistance and the Calgary Brain Tumor and Tissue Bank for providing patient samples for the establishment of the BT cell line. We thank the staff of the Princess Margaret Genomics Centre (www.pmgenomics.ca), Bioinformatics and High-Performance Computing Core and the BC Cancer Agency Genome Sciences Centre (www.bcgsc.ca) for their expertise in generating the sequencing data used in this study.

Author contributions

L.M.R., O.K.N.W., P.B.D., G.D.B. and T.J.P. conceived the project, designed the study and interpreted results. N.S., M.R., T.K., Z.X. and L.M.R. generated sc and snRNA-seq data. L.M.R. and O.K.N.W. performed scRNA-seq analysis. F.J.C., F.M.G.C. and P.G. generated and pre-processed bulk RNA-seq or WGS data. O.K.N.W. and F.M.G.C. performed bulk RNA-seq analysis. L.M.R. and F.M.G.C. performed WGS analysis. F.J.C., M.K., N.R., L.L., C.C., H.A.L. and J.E.J. derived GSC cultures used in the study and performed LDAs, xenografts and cytokine assays. G.M., M.A., D.A.B., J.E.J., N.L., E.L., N.I.P., J.K.B. and M.K. performed genome-wide CRISPR–Cas9 screens. G.M. and L.M.R. analyzed screen data. K.Y., J.S., S.D., M.B. and M.D.C. provided tumor tissue. F.J.C., D.C.C., M.K., M.L., B.H.K., H.A.L., S.W., M.A.M., R.A.M. and S.A. provided experimental and analytical support. L.M.R., G.D.B., P.B.D. and T.J.P. wrote the manuscript with feedback from all authors.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s43018-020-00154-9.

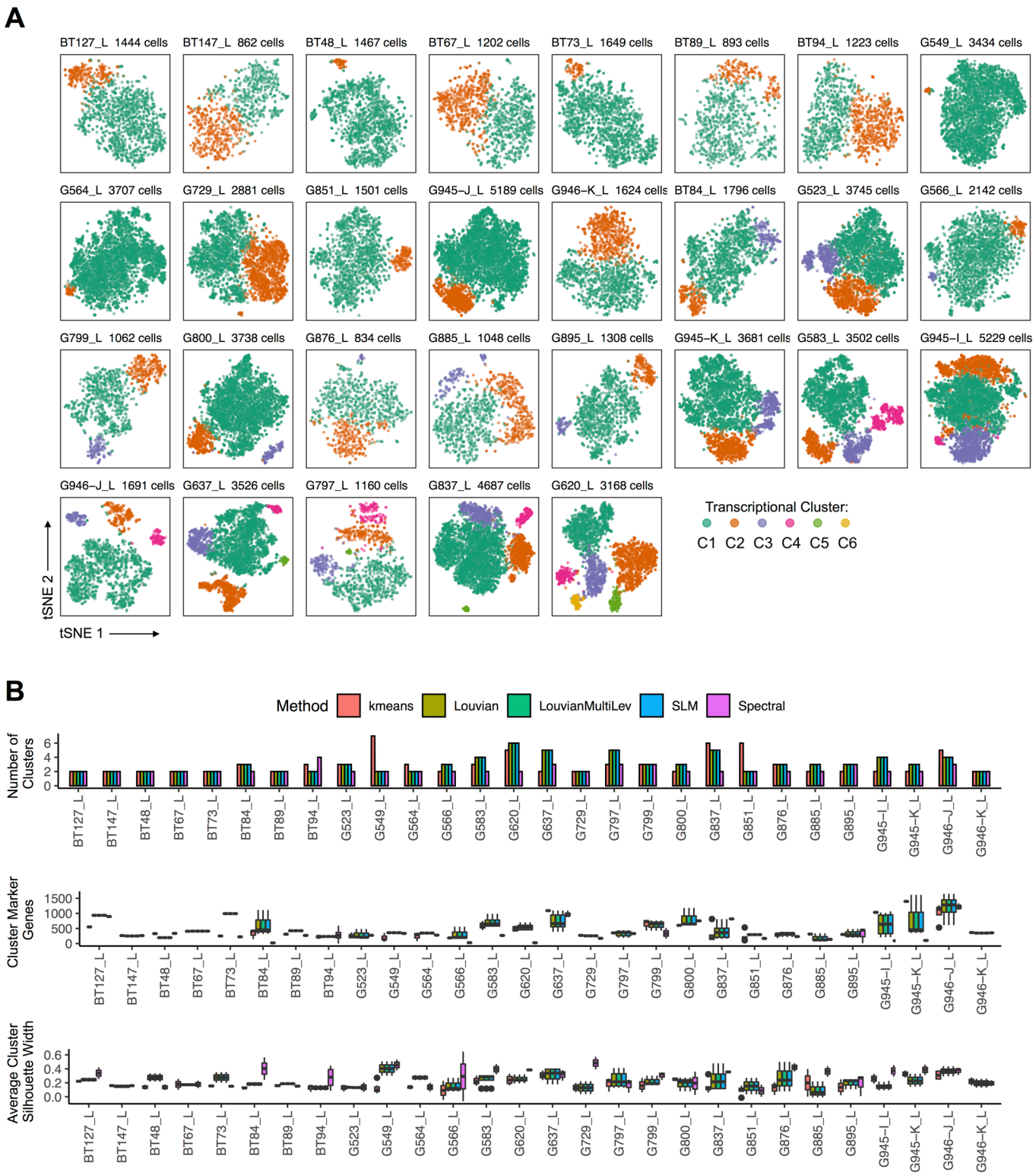
Supplementary information is available for this paper at https://doi.org/10.1038/s43018-020-00154-9.

Correspondence and requests for materials should be addressed to P.B.D., G.D.B. or T.J.P.

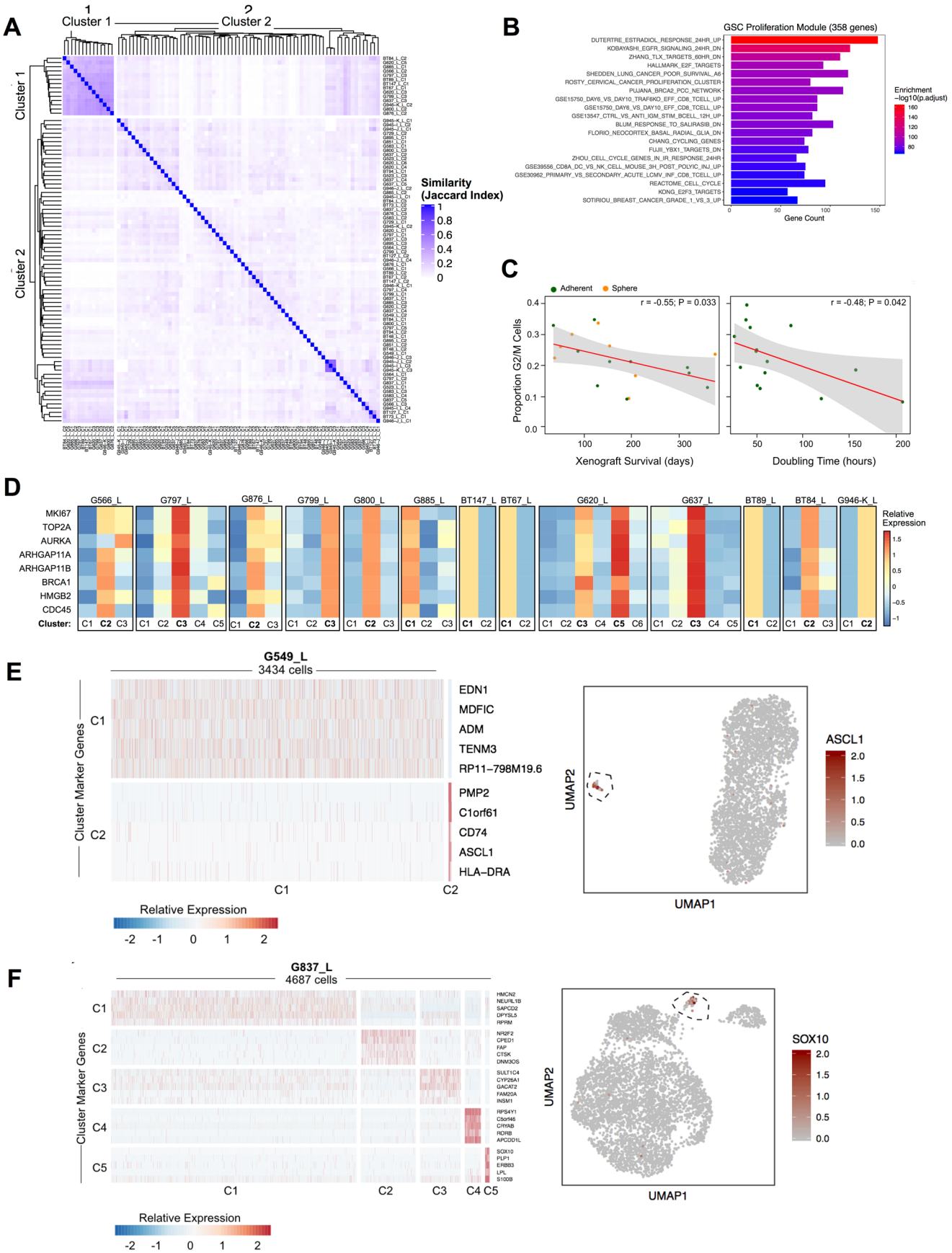
Peer review information *Nature Cancer* thanks Aaron Diaz, Benjamin Deneen, Justin Lathia, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

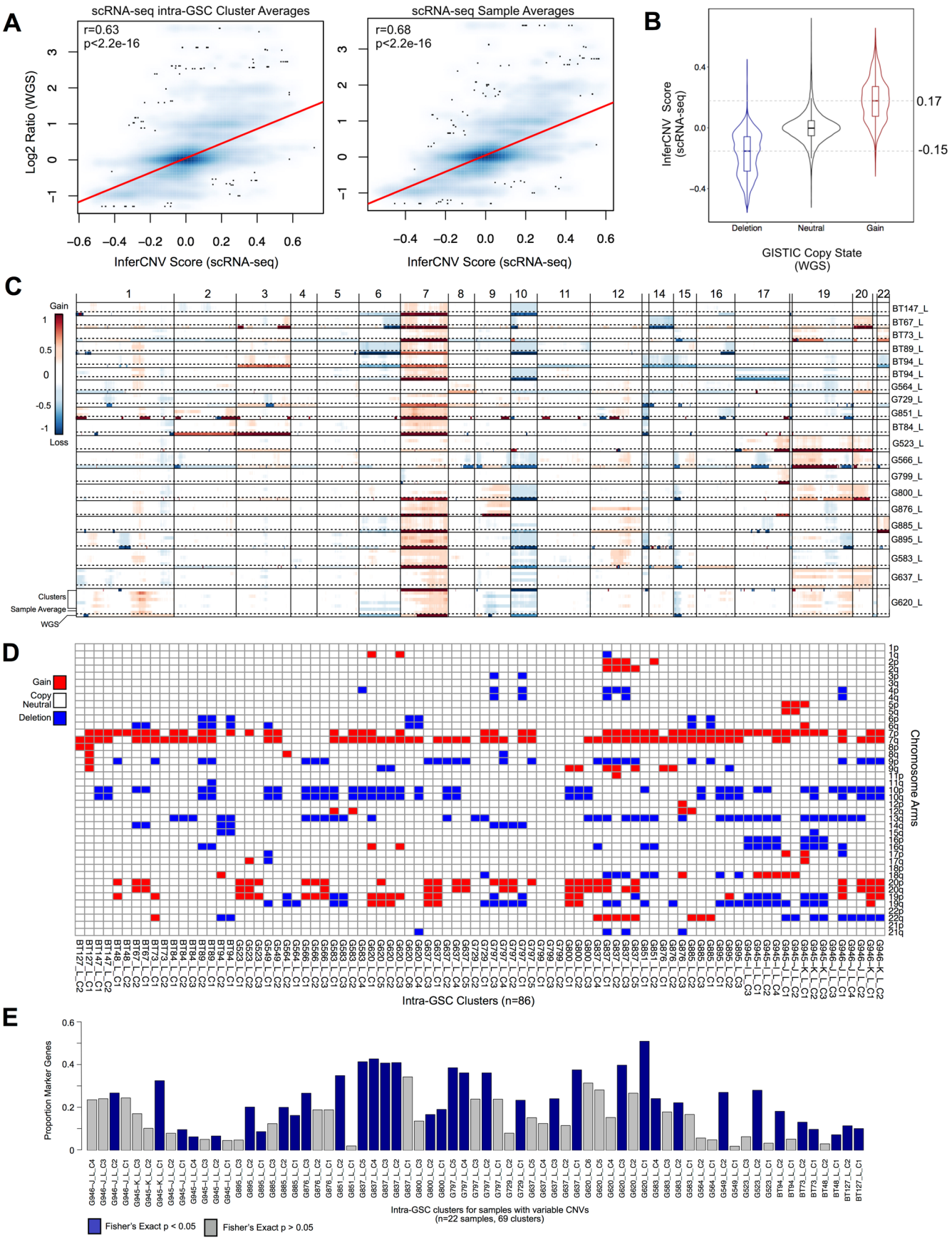


Extended Data Fig. 1 | Visualization and benchmarking of intra-GSC clustering. a, t-SNE representation of intra-GSC heterogeneity across 29 patient-derived GSCs. Cells are colored by transcriptional cluster. Samples ordered by number of clusters. **b**, Comparison of cluster number (top), marker genes per cluster (middle) and average silhouette width per cluster (bottom) between our original GSC smart local moving (SLM) clustering algorithm (blue), Louvain (yellow), Louvain with multilevel refinement (green), k-means (salmon) and spectral (pink). across 29 GSCs. The number of data points in the boxplots (middle, bottom) corresponds to the number of clusters in the matched histogram (top). Box plots represent the median, first and third quartiles of the distribution and whiskers represent either 1.5-times interquartile range or most extreme value.



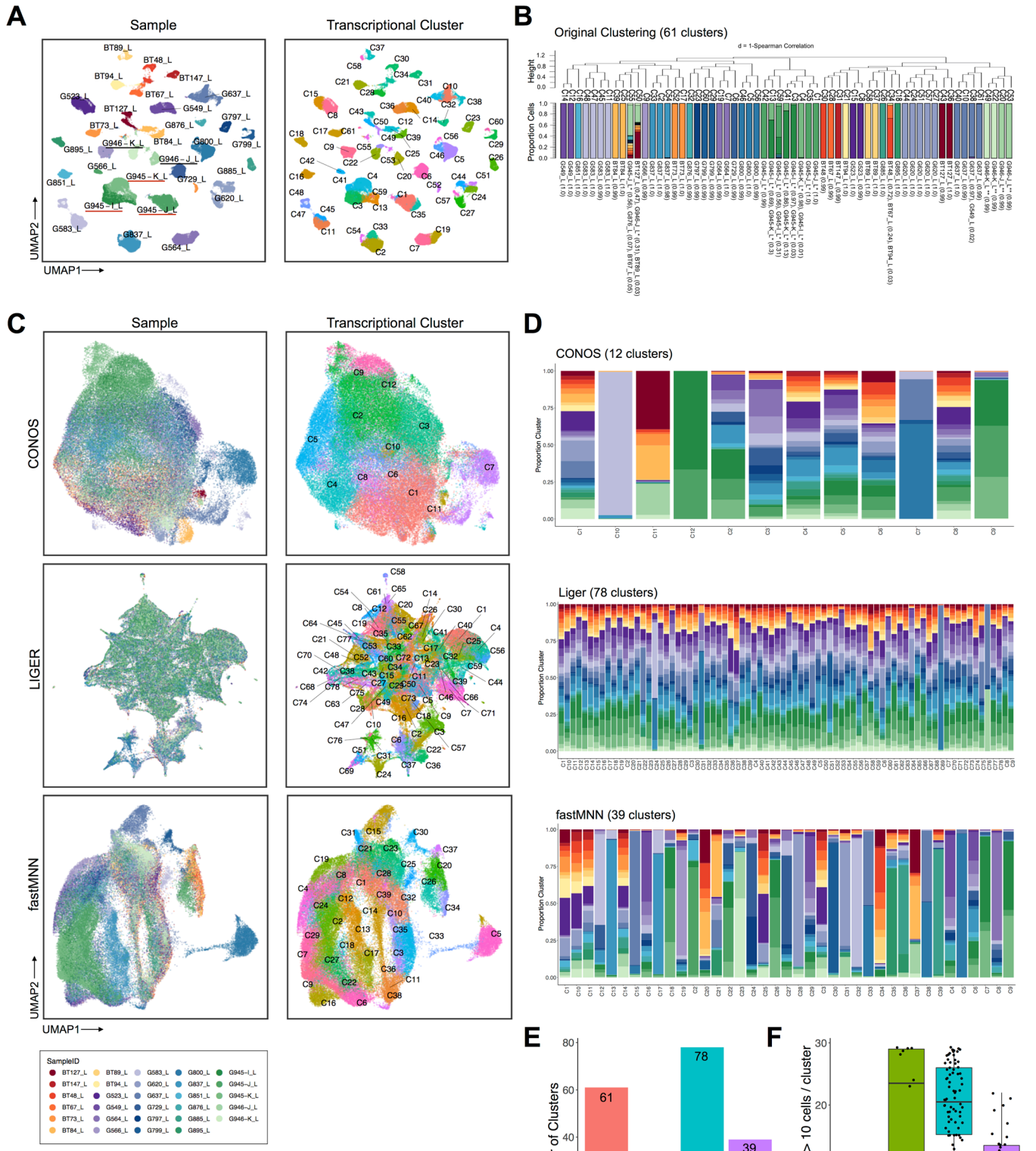
Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Defining intra-GSC transcriptional heterogeneity. a, Heat map of Jaccard Index (more similar = blue, less similar = white) between marker gene lists across 86 intra-GSC clusters. A subset of 14 clusters, from 13 samples, display increased similarity (labelled as Cluster 1). **b**, Enriched pathways from 358 genes common to all 14 clusters defined in Extended Data Fig. 2a. **c**, Spearman correlation between inferred proportion G2M cells from scRNA-seq data vs. survival in an orthotopic xenograft model (left; n = 18 independent GSC xenograft models) and doubling time *in vitro* (n = 15 GSC cultures) in adherent (green) or neurosphere (orange) GSCs. Red line represents a linear regression line. Shaded grey area represents 95% confidence interval. **d**, 14 intra-GSC clusters share increased marker gene overlap and define a core proliferation module shared across 13 patients. Expression of select marker genes common across all clusters. Columns separated by intra-GSC cluster, bolded labels represent clusters with upregulation of the proliferation module. **e**, Relative expression of top 5 significant marker genes (based on logFC, one-sided Wilcoxon rank-sum test, FDR < 0.05) for clusters C1 and C2 within G549_L (left). UMAP visualization of select marker genes of C2 (right). **f**, Relative expression of top 5 significant marker genes (based on logFC, one-sided Wilcoxon rank-sum test, FDR < 0.05) for clusters C1-C5 within G837_L (left). UMAP visualization of select marker gene of C5 (right).



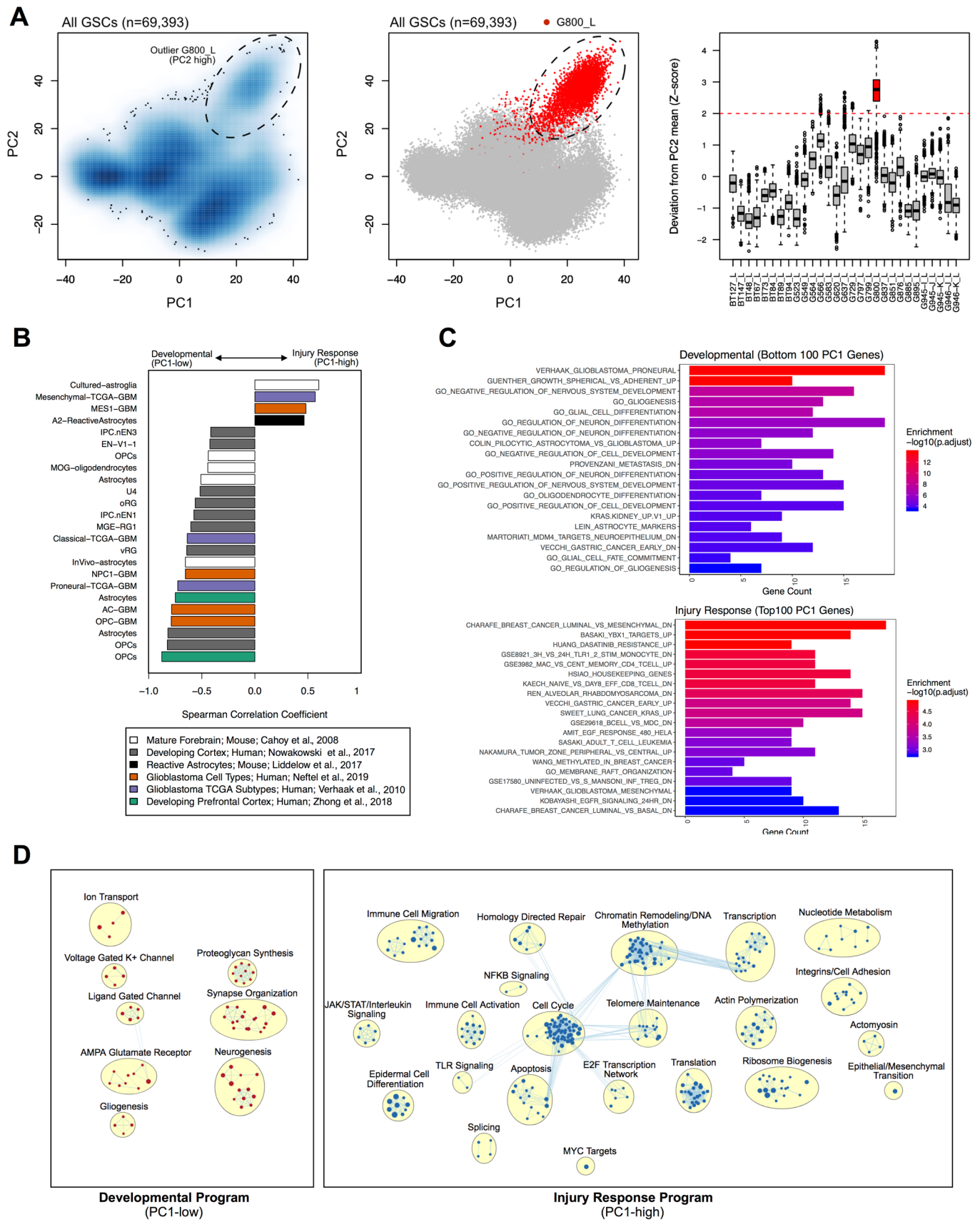
Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Validation of inferred single cell CNV profiles and impact on marker gene expression. a, Spearman correlation between inferred scRNA-seq CNV score from averaged intra-GSC clusters (left; n = 56 clusters from 20 GSC cultures) or averaged samples (right; n = 20 GSC cultures) and log2 ratios from matched genes from WGS of GSC samples (n = 20 GSC cultures). Each point represents a gene within a given sample. **b**, Distribution of InferCNV scores for genes labelled as deletion (<0; n = 11,617 genes), neutral (0; n = 100,426 genes) or amplified (>0; n = 12,777 genes across) by GISTIC from corresponding WGS data. Gene counts per GISTIC CNV state represent a cumulative number of genes across 20 GSCs. Median scores for deletions (-0.15) and gains (0.17) used as cut offs to classify InferCNV scores as at least single copy gains or losses. Box plots within the violin plot represent the median, upper and lower quartiles of the distribution and whiskers represent 1.5-times interquartile range. Tips of the violin plot extend to the minimum and maximum values of the distribution. **c**, Visualization of single cell CNV calls averaged by intra-GSC cluster (denoted "_C#"), averaged by sample ("SampleAverage") or results of matched WGS ("_WGS"). Samples (rows separated by solid lines) ordered by increasing cluster number. WGS CNV track below dashed line. Sample average above dashed line and cluster transcriptional profiles represent remaining rows. **d**, Binary heat map depicting chromosome arms (y-axis; sorted by genomic position) that are gained (red), deleted (blue) or copy-neutral (white) across intra-GSC clusters (x-axis; ordered alphabetically; n = 86 clusters from 29 GSC cultures). **e**, Proportion of cluster marker genes located within a variable CNV loci (y-axis) across intra-GSC clusters (x-axis; n = 69 clusters) from samples with variable cluster CNV profiles (n = 22 GSC cultures) as determined in Extended Data Fig. 3d. Clusters with significant (Fisher's Exact Test $p < 0.05$) enrichment of marker genes within variable CNV loci are colored dark blue.



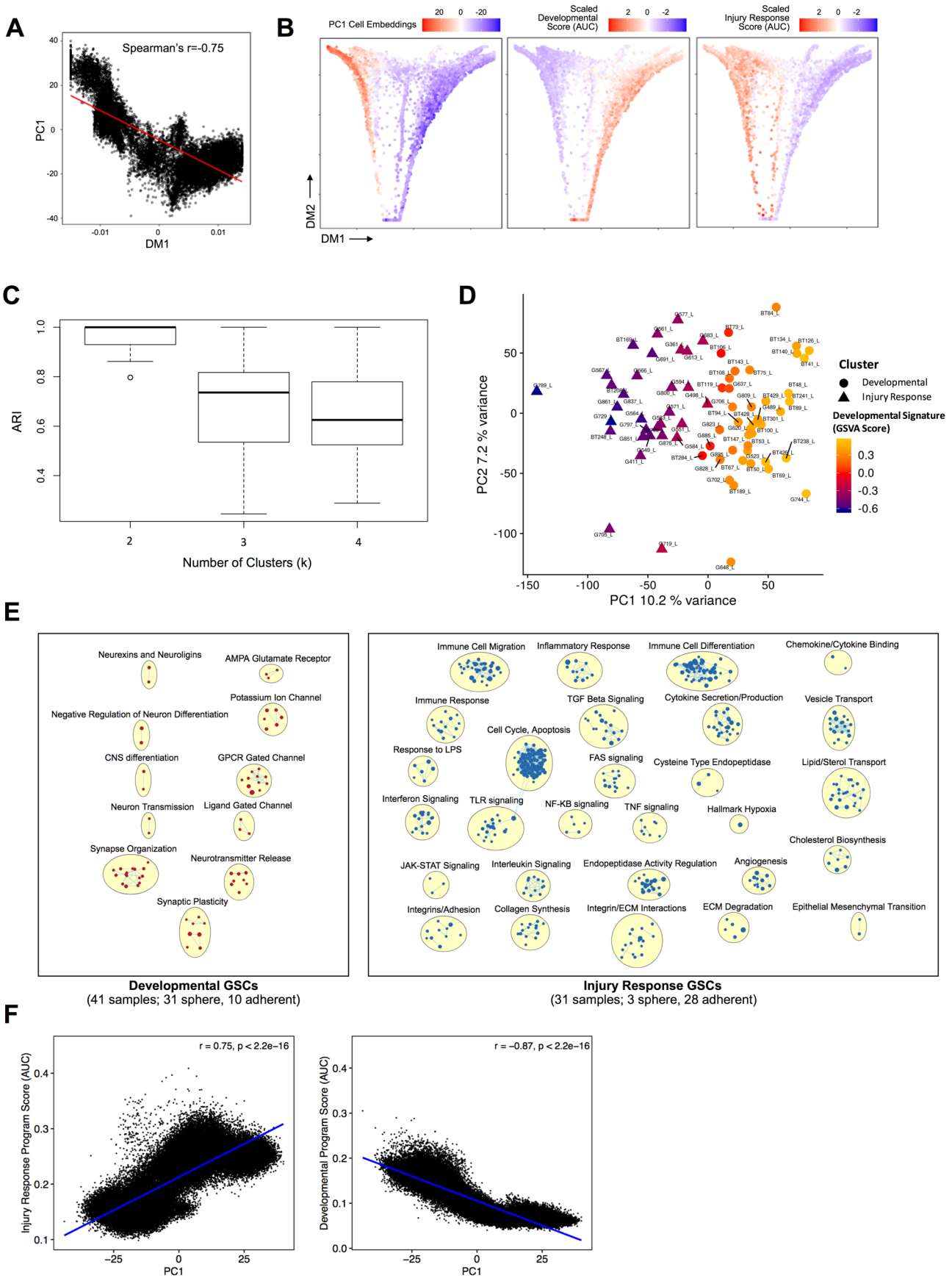
Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Defining global inter-GSC cluster relationships and evaluation of batch correction methods. a, UMAP projection of 69,393 GSC cells from 29 patients reveals patient-specific clustering patterns (left panel, cells colored by patient). Unbiased clustering reveals 61 transcriptional clusters (right panel, cells colored by transcriptional cluster). GSCs derived from different regions of the same tumor underlined with red (G945-I,J,K) and black (G946-J,K) bars. **b**, Transcriptional clusters from the same sample and patient are more similar to each other compared to cells from other samples. Dendrogram of average gene expression profiles of transcriptional clusters defined in Extended Data Fig. 4a based on distance (1-Spearman correlation) (top). Sample composition of transcriptional clusters (bottom). Vertical bars colored by sample. Labels at bottom depict sample identifier and proportion of sample for up to the top three samples/cluster. **c**, UMAP visualizations of global GSC clustering results with CONOS batch correction (top row), with Liger batch correction (middle row) and fastMNN batch correction (bottom row). Cells are colored by sample ID (left column) and transcriptional cluster (right column) (n = 69,393 cells from 29 GSC cultures). **d**, Proportion of cells (y-axis) corresponding to a given sample across transcriptional clusters (x-axis) across original and batch corrected datasets. **e**, Number of transcriptional clusters in original clustering pipeline vs. post-batch correction. **f**, Box plots representing the number of samples with >10 cells per transcriptional cluster across original and batch corrected clustering results (Original=61 clusters; Conos=12 clusters; Liger=78 clusters; fastMNN=39 clusters). Box plots represent the median, first and third quartiles of the distribution and whiskers represent either 1.5-times interquartile range or most extreme value. Outliers displayed as circles.



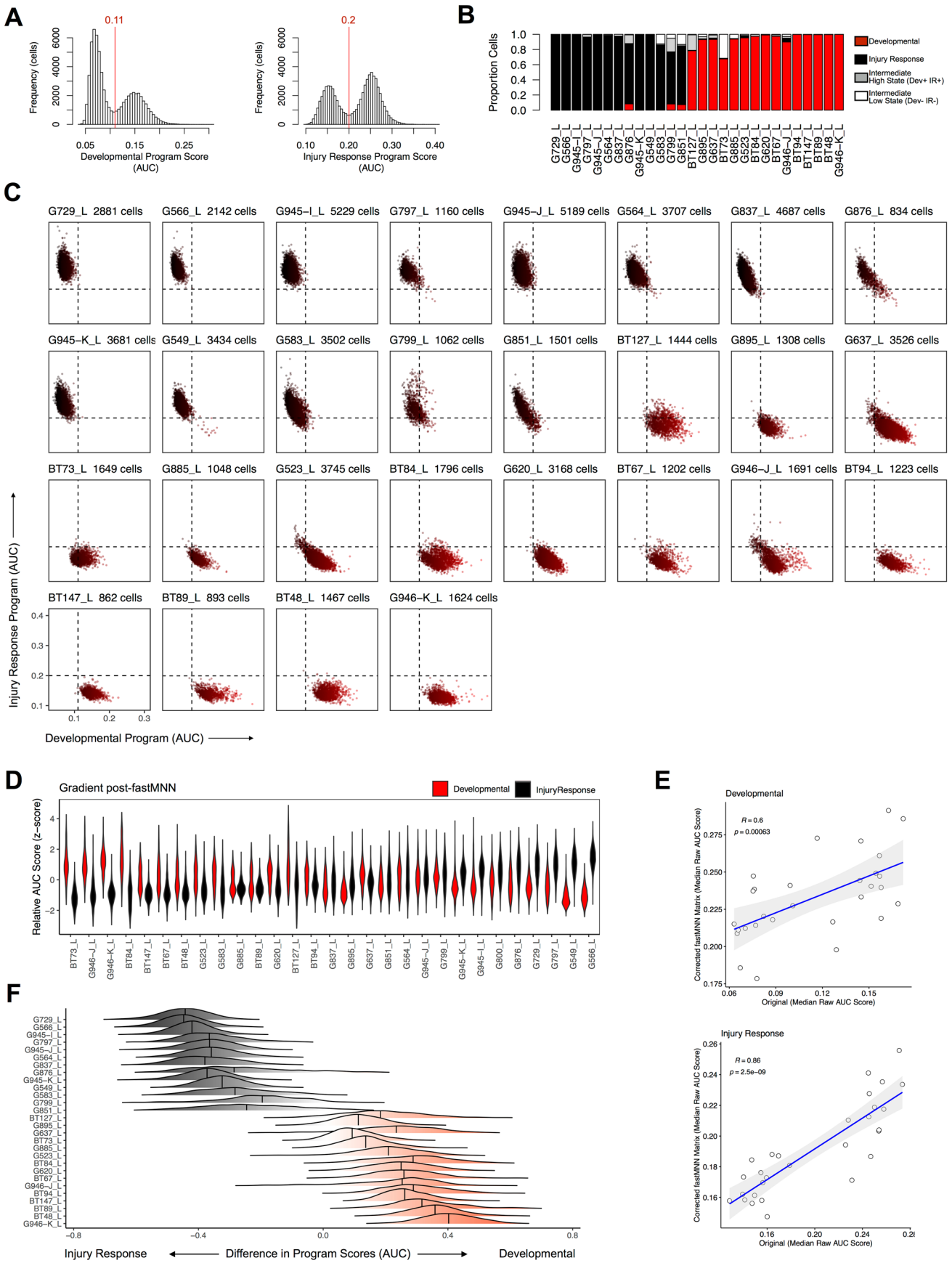
Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Characterization and interpretation of GSC transcriptional gradient. a, PCA plot of 69,393 cells from 29 GSC cultures. Plot colored by cell density (left). PCA plot with cells belonging to outlier sample G800_L, colored red. Remainder of cells colored grey (middle). Quantification of deviation from the mean of PC2 (y-axis) across samples. G800_L (red) represents an outlier with >95% of cells within the sample greater than two standard deviations from the mean. Horizontal dashed red line represents threshold of two standard deviations to determine outliers (right). Box plots represent the median, upper and lower quartiles of the distribution and whiskers represent 1.5-times interquartile range or the most extreme value. Outliers represented as circles. **b**, Correlation of cell type gene signature scores from PC1 cell embeddings (n = 65,655 cells from 28 GSC cultures; outlier G800_L removed as in Fig. 2a). Only correlations with Spearman correlation coefficient greater than |0.5| shown. Bars colored by gene signature source. **c**, Enriched MSigDB gene sets (FDR < 0.01) for top 100 and bottom 100 genes for PC1. (n = 65,655 cells from 28 GSC cultures; outlier G800_L removed as in Fig. 2a). **d**, Gene Set Enrichment Analysis (GSEA) on PC1 loadings (gene associations with PC1) visualized using EnrichmentMap (n = 65,655 cells from 28 GSC cultures; outlier G800_L removed as in Fig. 2a). Similar pathways (circles) are grouped into labeled clusters (larger bubbles). Blue circles denote positively associated pathways (Injury Response associated) and red circles denote negatively associated pathways (Developmental associated). Edges (lines) denote overlap between pathways.



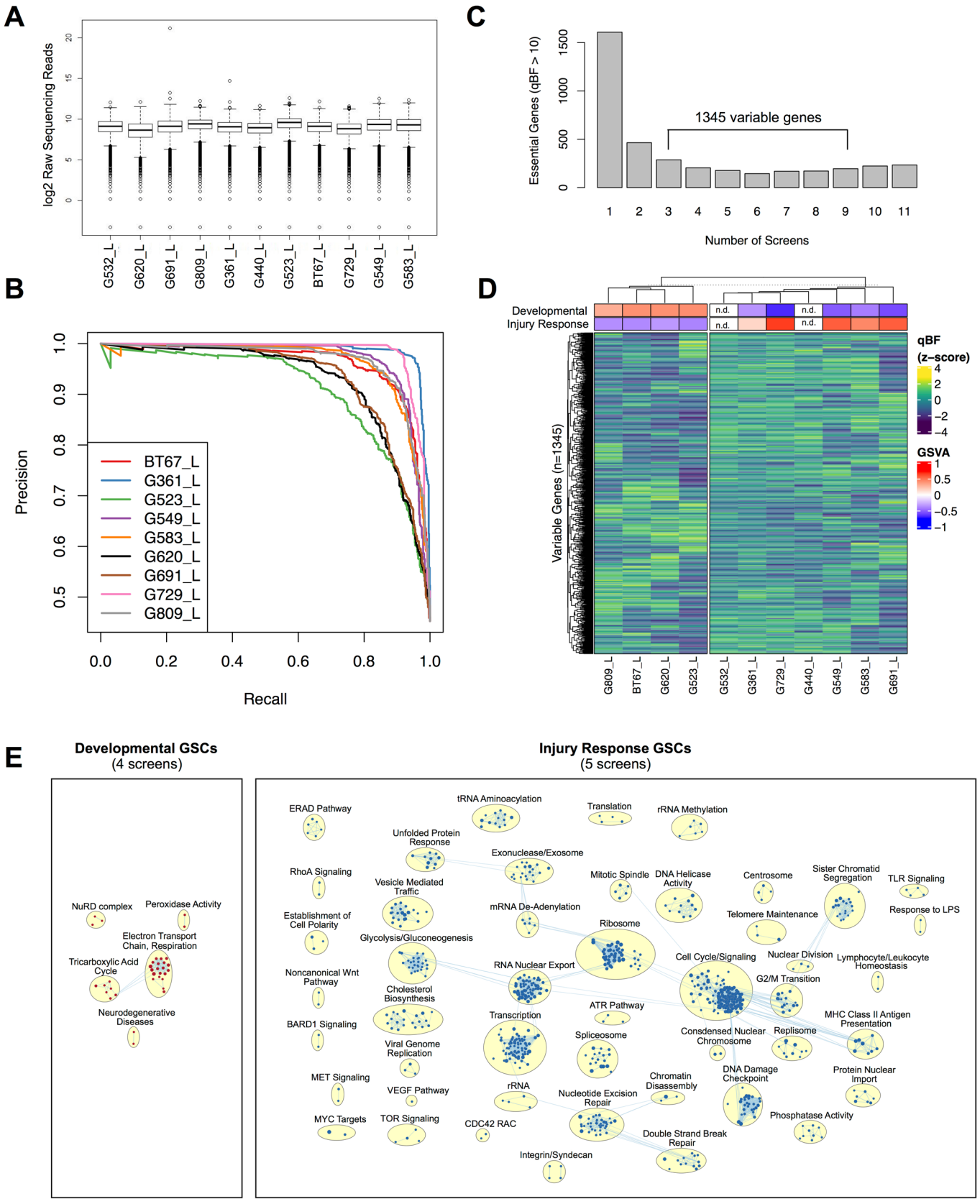
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Diffusion Map and bulk RNA-sequencing of 72 GSCs confirms Developmental and Injury Response transcriptional states. **a**, Spearman correlation between diffusion component 1 (DM1; x-axis) and principal component 1 (PC1; y-axis) cell embeddings for a subset of 14,000 GSCs (500 cells/sample). **b**, Diffusion Map of 14,000 GSCs. Cells coloured by PC1 cell embeddings (left; Related to Fig. 2a), scaled Developmental transcriptional program score (middle) and scaled Injury Response transcriptional program score (right). **c**, Spectral clustering determined GSCs (n = 72 GSC cultures) profiled with bulk RNA-sequencing separated into two stable clusters. For each cluster number (x-axis), boxplots depict 200 pairwise similarities (y-axis) (adjusted Rand index, ARI) between the solution obtained for the full dataset and random subsets of data containing 80% of samples. Box plots represent the median, first and third quartiles of the distribution and whiskers represent either 1.5-times interquartile range or most extreme value. Outliers displayed as circles. **d**, PCA plot of GSCs profiled with bulk RNA-sequencing colored by GSVA score for Developmental signature (n = 72 GSC cultures). Circles denote GSCs from the Developmental cluster, while triangles denote GSCs from the Injury Response Cluster. **e**, GSEA on differentially expressed genes between Developmental and Injury Response clusters as determined by bulk RNA-sequencing, visualized with EnrichmentMap. Similar pathways (circles) are grouped into labeled clusters (larger bubbles). Blue circles denote Injury Response associated pathways and red circles denote Developmental associated pathways. Edges (lines) denote overlap between pathways. **f**, Spearman correlation at the individual cell (n = 65,655) level between PC1 cell embeddings from scRNA-seq and Developmental and Injury Response gene signature scores derived from bulk RNA-sequencing.



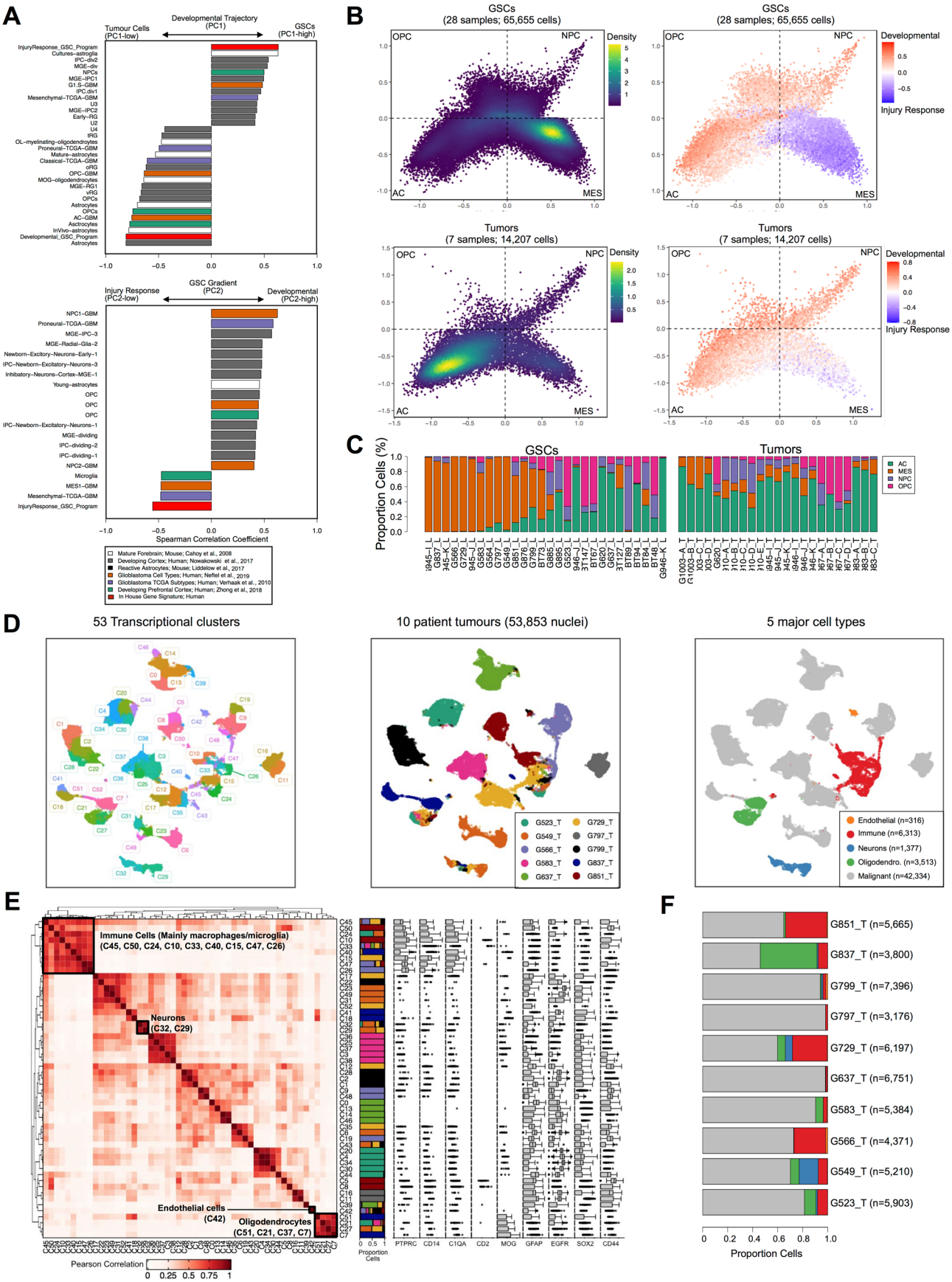
Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Continuous transcriptional gradient of Developmental and Injury Response cell states across patients. a, Distribution of AUC gene signature scores for Developmental (left) and Injury Response (right) programs across all GSC cells (n = 65,655 cells from 28 GSC cultures). Red line marks classification threshold to determine if a given program is active or not. **b**, Proportion of cells across samples categorized as being resembling Developmental or Injury Response states, as well as intermediate hybrid states. **c**, Position of cells on the Developmental (x-axis) and Injury Response (y-axis) gradient across all samples (n = 65,655 cells from 28 GSC cultures). Cells are colored by relative expression of the Developmental (red) and Injury Response (black) expression programs. GSC cultures with intermediate scores either contain subpopulations of both subtypes or middling scores for both states. Samples ordered as presented in Fig. 2d**. d**, Violin plots depicting the distribution of Developmental (red) and Injury Response (black) programs post-fastMNN correction for cells within samples. Samples sorted by increasing median Injury Response program score. **e**, Pearson correlation of median Developmental (top panel) and Injury Response (bottom panel) between transcriptional program scores derived from the original expression matrix (x-axis) and expression matrix post-fastMNN batch correction (y-axis). Blue line represents linear regression line, shaded grey area represents 95% confidence interval and each dot represents the median raw AUC score per GSC. **f**, Ridge plots depicting distribution of the difference in Developmental (red) and Injury Response (black) scores (x-axis) across cells within samples (y-axis) (n = 65,655 cells from 28 GSC cultures). Samples ordered as presented in Fig. 2d. Vertical black line represents the median.



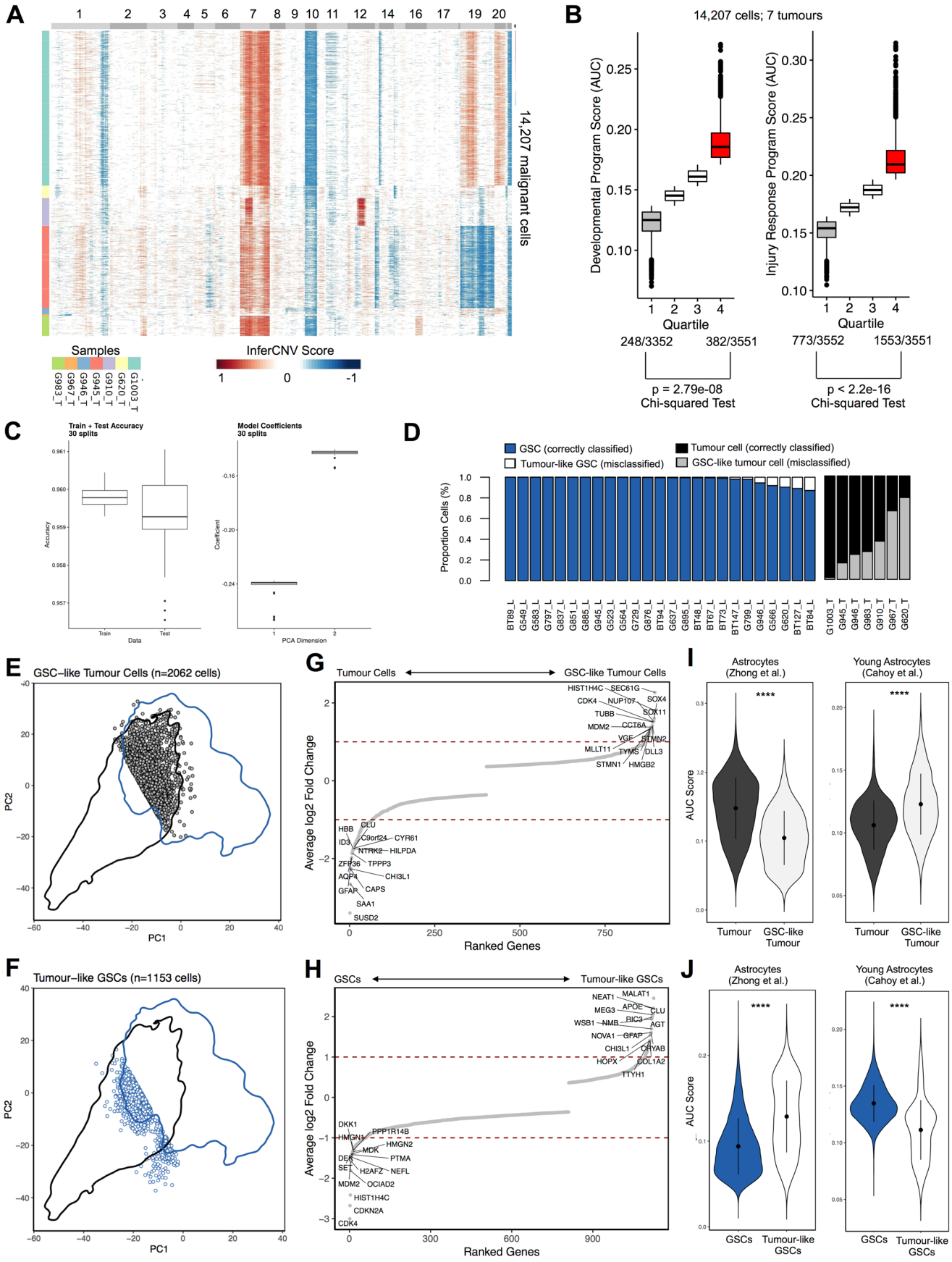
Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Genome-Wide CRISPR-Cas9 screens in GSCs. a, Box and whisker plots of TKOv3 gRNA library complexity in T0 populations for 70,948 individual gRNAs from a single independent screen per GSC (n = 11 screens in 11 GSC cultures). Box plots represent the median, first and third quartiles of the distribution and whiskers represent 1.5-times the interquartile range. Outliers displayed as circles. **b**, Precision-recall curves for 11 GSC CRISPR-Cas9 screen produced with BAGEL pipeline and v2 reference for essential/non-essential genes. **c**, Barplot depicting the number of shared fitness genes across GSC screens. **d**, Heatmap of quantile normalized gene fitness Bayes factor (qBF) scores for the 1,484 most variable genes across 11 GSC screens. Samples (columns) annotated with GSVA score for Developmental and Injury Response gene signature scores from bulk RNA-sequencing. **e**, GSEA on differentially essential genes between Developmental and Injury Response GSCs, visualized with EnrichmentMap. Similar pathways (circles) are grouped into labeled clusters (larger bubbles). Blue circles denote pathways more essential in Injury Response GSCs and red circles denote pathways more essential in Developmental GSCs. Edges (lines) denote overlap between pathways.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Characterization of axes of variation in glioblastoma and single nuclei RNA-sequencing of 53,853 nuclei from 10 patient tumors. **a**, Spearman correlation of cell type gene signature scores to PC1 and PC2 cell embeddings for combined PCA of GSC and tumor cells (n = 65,655 cells from 28 GSC cultures and 14,207 malignant cells from 7 tumors). Only correlations with Spearman correlation coefficient greater than |0.4| shown. Bars colored by gene signature source. **b**, Projection of GSCs (top row; n = 65,655 cells) and =patient tumor cells (bottom row; n = 14,207 cells) onto GBM cell state map: astrocyte-like (AC; bottom left quadrant), oligodendrocyte precursor cell-like (OPC; upper left quadrant), neural progenitor cell-like (NPC, upper right quadrant) and mesenchymal-like (MES; bottom right quadrant). Cells are colored by density (left panels) and Developmental - Injury Response gradient program scores (right panels). **c**, Proportion of cells across samples that map to each of the 4 GBM cell states. **d**, UMAP visualization of 53,853 nuclei from 10 patient tumors colored by transcriptional cluster (left), patient (middle) and cell type (right). **e**, Pearson correlation between average transcriptional cluster expression (left). Proportion patient cells per transcriptional cluster (middle), as colored in panel B. Box plots detailing expression of cell type marker genes per cluster (right). Box plots represent the median, first and third quartiles of the distribution and whiskers represent either 1.5-times interquartile range or most extreme value. Outliers are removed. **f**, Proportion of cell types across tumors (as colored in the right panel of Extended Data Fig. 9d). Numbers in brackets represent the total number of nuclei per tumor.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Validation of GSC-state CNVs in patient tumors and identification of GSC-like tumor cells. a, Genome-wide inferred CNV profiles for 14,207 malignant cells from 7 patient tumors. Columns represent genomic regions, ordered by genome position across all chromosomes. Rows represent CNVs for individual cells, annotated by sample. **b**, Developmental (left) and Injury Response (right) program scores across quartiles. Numbers underneath quartile labels depict the number of cells harbouring respective Developmental or Injury Response CNVs. Enrichment of CNVs between upper and lower quartiles was determined using a Chi-squared test. Box plots represent the median, first and third quartiles of the distribution and whiskers represent either 1.5-times interquartile range or most extreme value. Outliers are displayed as circles. **c**, Train and test accuracy for logistic regression model, 30 random 80:20 train test splits (left). Distributions of model coefficients corresponding to the 30 trained models (right). Model coefficients are weights by which the logistic regression model describes class likelihood as a function of PC1 and PC2. Box plots represent the median, first and third quartiles of the distribution and whiskers represent either 1.5-times interquartile range or most extreme value. Outliers displayed as circles. **d**, Proportion of cells in GSCs correctly classified as being GSCs (blue) or misclassified representing tumor-like GSCs (white). Proportion of tumor cells correctly classified as being tumor (black) or misclassified as being GSC-like (grey). **e**,**f**, PCA plot of all GSCs and tumor cells as in Fig. 5a. Black line represents contour encompassing 99% of tumor cells. Blue line represents contour encompassing 99% of GSCs. Grey dots represent tumor cells classified as being GSC-like. White dots with blue outline represent GSC cells classified as being tumor-like. **g**, Differential gene expression analysis between tumor cells and GSC-like tumor cells. Each dot represents a gene (x-axis) ordered by average log2 fold change (y-axis). Red dashed line represents a log2 fold change of double between groups. **h**, Differential gene expression analysis between GSCs and tumor-like tumor cells. Each dot represents a gene (x-axis) ordered by average log2 fold change (y-axis). Red dashed line represents a log2 fold change of double between groups. **i**, Expression of mature and young astrocyte gene signatures between tumor cells (black; n = 12,145 cells) and GSC-like tumor cells (grey; n = 2,062 cells). **j**, Expression of mature and young astrocyte gene signatures between GSCs (blue; n = 64,502 cells) and tumor-like GSCs (white; n = 1,153 cells).

nature research

Corresponding author(s): Dr. Trevor Pugh
Dr. Gary Bader
Dr. Peter Dirks

Last updated by author(s): Nov 9, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used.
Data analysis	The following packages were used for processing and data analysis of single cell and single nuclei RNA-sequencing data: R v3.5.0 and v3.6.1, CellRanger v2, Dropbead v0.3.1, Seurat v2.3.4, clusterProfiler v3.10.1, scClustViz v1.2.1, AUCell v1.4.1, inferCNV v0.3, CONICS v0.0.0.1, NbClust v3.0, cluster v2.1.0, velocyto v0.17.13, scvelo v0.2.2, loompy v3.0.6, GSEA v3.0, AutoAnnotate v1.3.2, EnrichmentMap v3.1, Cytoscape v3.7.2. The logistic regression model was trained using sklearn v0.21.2, reticulate v1.13, R v3.5.2 and python v3.6.9 (Anaconda). The following packages were used for processing and data analysis of bulk RNA-sequencing data: R v3.5.2, STAR v2.4.2a, DEseq2 v1.22.2, sva v3.30.1, kernlab v0.9-29, GSVA v1.30.0, GSEA v3.0, AutoAnnotate v1.3.2, EnrichmentMap v3.1, Cytoscape v3.7.2. The following packages were used for processing and data analysis of whole genome sequencing data: FastQC v0.11.5, bwa v0.7.15, samtools v1.3.1, GATK v3.5, Picard Tools v2.6.0, VarScan v2.3.8 , Sequenza v2.1.2, GISTIC v2.0.23. The following packages were used for processing and data analysis of genome-wide CRISPR-Cas9 screening data: R v3.5.0, BAGEL v0.91, MAGECK v0.5.8, GSEA v3.0, AutoAnnotate v1.3.2, EnrichmentMap v3.1, Cytoscape v3.7.2. Extreme Limiting Dilution Analysis (ELDA) software was used to calculate sphere formation capacity (http://bioinf.wehi.edu.au/software/elda/). Plotting and statistical analysis was performed in the R statistical environment (v3.5.0, v3.6.1) and GraphPad Prism (v8).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Bulk RNA-sequencing (EGAS00001003070 & EGAS00001004395), whole genome sequencing (EGAS00001004395), single cell and single nuclei RNA-sequencing (EGAS00001004656) datasets generated and analysed in this study are available through the European Genome-Phenome Archive (https://www.ebi.ac.uk/ega/studies/) repository, in the form of FASTQ or BAM files. Processed single cell RNA-sequencing data are publically available through the Broad Institute Single-Cell Portal (https://singlecell.broadinstitute.org/single_cell/study/SCP503).

Previously published single-cell RNA-sequencing data that were re-analyzed in this study are available from the following sources: Wang et al. 2019 (GSE138794; doi:10.1158/2159-8290.CD-19-0329, Bhaduri et al. 2020 (doi:10.1016/j.stem.2019.11.015, URL: http://cells.ucsc.edu/?ds=gbm), Neftel et al. 2019 (doi: , url: https://singlecell.broadinstitute.org/single_cell/study/SCP393/), Darmanis et al. 2018 (doi: 10.1016/j.celrep.2017.10.030, URL: http://storage.googleapis.com/gbmseqrawdata/)

Source data for Figures 1-7 and Extended Data Figures 1-14 are available for this study. All other data supporting the findings of this study are available from the corresponding author on reasonable request.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size was determined by the availability of the human samples and GSC cultures.
Data exclusions	All of the data acquired was utilized for analysis, unless specified otherwise in the figure legends or text. We identified one GSC sample, G800_L, as an outlier on the basis of inflated PC2 signal and therefore removed it from downstream analysis (Extended Data Figure 6A). The exclusion criteria was not pre-established.
Replication	Technical replicates were not performed for sequencing. However, a subset of samples were profiled by multiple technologies (single cell RNA-sequencing, bulk RNA-sequencing, CRISPR and whole genome sequencing), as outlined in Supplementary Table 1.
Randomization	There is no randomization as part of this study.
Blinding	Tissue processing and sequencing (single cell, single nuclei, bulk RNA, CRISPR) were were performed without knowledge of patient outcomes or phenotypes.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	Patient-derived GSC cell lines were derived from surgical specimens from patients with Glioblastoma.
Authentication	Authentication was done using GenePrint10 STR profiling system (or comparable) polymorphic marker panel to confirm that cell lines matched the patient tumour from which they were derived.
Mycoplasma contamination	All cell lines were tested for mycoplasma contamination and they were confirmed to be mycoplasma-free.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used in this study.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Six- to 16-week-old female NOD/scid gamma or CB17/SCID mice were used in this study. Mice were housed in groups of three to five and maintained on a 12 hrs light/dark schedule with a temperature of 22°C ± 1°C and a relative humidity of 50% ± 5%. Food and water were available ad libitum.
Wild animals	No wild animals were used in this study.
Field-collected samples	No field collected samples were used in this study.
Ethics oversight	All attempts are made to minimize the handling time during surgery and treatment so as not to unduly stress the animals. Animals are observed daily after surgery to ensure there are no unexpected complications. All animal protocols described in this study were approved by the Animal Care Committee at The Hospital for Sick Children (Toronto, ON) and the University of Calgary operating under the Guidelines of the Canadian Council on Animal Care (Calgary, AB). All animal work procedures were in accordance with the Guide to the Care and Use of Experimental Animals published by the Canadian Council on Animal Care and the Guide for the Care and Use of Laboratory Animals issued by NIH (Bethesda, MD)

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics	Patients with newly diagnosed and recurrent glioblastoma were recruited at University Health Network (Toronto, Ontario, Canada), St.Michael's Hospital (Toronto, Ontario, Canada) and Foothills Hospital (Calgary, Alberta, Canada). Tumour specimens and control tissues were obtained from patients at the Tumour Tissue Bank within the Arnie Charbonneau Cancer Institute (Calgary, Alberta, Canada), University Health Network and St. Michael's Hospital (Toronto, Ontario, Canada), with the help of the Toronto Brain Tumour Biobank at The Hospital for Sick Children (Toronto, Ontario, Canada) and following informed consent. Population characteristics of patients are described in Supplementary Table 1.
Recruitment	Patients requiring tumour resection were recruited and consented by the surgeon or a tumour bank manager. There are no self-selection of other biases involved in recruitment that will impact results.
Ethics oversight	REB 1000025582 and REB 0020010404 approved by the Hospital for Sick Children Research Ethics Oversight committee (Toronto, Ontario, Canada). REB HREBA-CC-160762 approved by the Arnie Charbonneau Cancer Institute Research Ethics Board (Calgary, Alberta, Canada).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	We generated single-nuclei suspensions from snap-frozen tumors. Tissues were minced on dry ice and dissolved in lysis buffer (0.32 M sucrose, 5 mM CaCl2, 3 mM Mg(Ac)2, 20 mM Tris-HCl pH 7.5, 0.1% Triton-X-100, 0.1 mM EDTA pH 8.0), followed by homogenization with a pellet pestle. Nuclei integrity and quantity was assessed with SYBR Green II RNA Gel Stain (ThermoFisher Scientific). Nuclei were filtered through a 40 µm cell strainer and sorted for intact nuclei using DAPI (Sigma-Aldrich) on a BD Influx FACS sorter.
Instrument	BD Influx FACS sorter
Software	FlowJo (https://www.flowjo.com/) was used to analyze data.
Cell population abundance	DAPI was used to flow sort intact, high quality nuclei for single-nuclei RNA-sequencing. The proportion and quantity of high quality nuclei varied across tumours depending on quality and size of the biopsy.
Gating strategy	Nuclei were isolated from cell debris by gating for FSC and SSC. DAPI+ nuclei had higher fluorescence versus debris (DAPI-). DAPI + nuclei were sorted for and used as input for single nuclei RNA-sequencing.

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.