Transcriptional and Multi -Omic Heterogeneity in Glioblastoma Stem Cells

Owen Whitley

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy

Department of Molecular Genetics, University of Toronto

Transcriptional and Multi -Omic Heterogeneity in Glioblastoma Stem Cells

Owen Whitley

Doctor of Philosophy

Department of Molecular Genetics

University of Toronto

2022

Abstract

Glioblastoma Multiforme (GBM) is a disease with terrible prognosis, having a median survival time of ~12-15 months for patients undergoing current treatment regimens, and a 5-year survival rate under 10%. There is strong evidence to suggest the existence of stem like cells, which we term glioma stem cells (GSCs), that are capable of repopulating the tumor after surgery and chemotherapy. Thus, any effective treatment for GBM will likely have to target this population. Evidence of functional heterogeneity of GSCs at the level of the transcriptome and drug response motivates a full characterization of GSCs' biological variation. In this project, I use multiple -omic data types from both bulk and single cell resolution to explore how multiple biological processes such as transcription and epigenetic regulation play into GSC heterogeneity in therapeutic vulnerabilities. With single cell/nuclei RNA-sequencing, in collaboration with the Pugh and Dirks labs and others, I establish that transcriptional heterogeneity in GSCs and more generally GBM can be decomposed into two major axes of variation: a Developmental/Injury Response transcriptional axis present in both the stem fraction of GBM tumors as well as the tumors themselves, and a stem to astrocyte differentiation gradient present only in the tumor samples. Further functional characterization with CRISPR knockout data reveals that differential functional dependencies between Developmental and Injury Response GSCs largely match their differentially expressed genes, showing that this transcriptional axis of variation translates into functional consequences that could inform development of combination therapies. Following these results, I perform an integrated analysis of genomic, transcriptomic, epigenomic, and miRNA-seq data to obtain a more complete picture of both how this transcriptional axis is regulated as well as what other heterogeneity exists independent of transcription. Here, I find four major axes in multi -omics space: one corresponding to a hypermutation phenotype largely matching one previously characterized for GBM dependent on mismatch repair deficiency and temozolomide treatment, two others corresponding to apparent latent variation in regulation of inflammatory genes, and lastly a multi-omics axis corresponding to the coordinated regulation of the Developmental/Injury Response transcriptional axis by multiple biological layers. Collectively, the results presented in this thesis provide better mechanistic understanding of GSC heterogeneity and open the door to developing novel therapies.

Acknowledgements

This project would not have been possible without the support of a multitude of individuals that came in a variety of ways. First and foremost I would like to thank my supervisor, Dr. Gary Bader, for his guidance and support through these past four years. Not only was I given an interesting set of projects to work on, but the right combination of useful input from a seasoned computational biologist and cancer expert and the freedom to try new approaches to address the questions posed by this thesis. Considering the amount of other lab members' projects he was involved in, I am grateful for the amount of attention he was able to provide as well as the scientific training I received.

I would also like to acknowledge my committee members, Dr. Peter Dirks and Dr. Hannes Rost. Dr. Dirks provided a wealth of knowledge about the subject matter (cancer stem cells in glioblastoma) that is only acquired from decades of experience, and Dr. Rost provided additional expertise in signal processing and computational biology analyses. Neither of them held back honest feedback on ways to improve as a student, and were genuinely interested in seeing me grow as a scientist, and for that I have to thank them greatly.

Beyond those on my supervisory committee, I would like to thank all the members of the Stand Up to Cancer Canada Cancer Stem Cell Dream Team for their contributions to this project. I would especially like to thank Laura Richards, Fiona Coutinho, Michelle Kushida, Graham MacLeod, Florence Cavalli, and Paul Guilhamon for data acquisition, data processing, and analytical contributions. In all, this project was an enormous collaboration between many Canadian labs, and it would have been impossible for one person to do alone.

I can't go without mentioning the rest of the Bader Lab, which was a great environment to end up in. I was surrounded by smart people who were genuinely curious about the world around them, which perhaps comes naturally for an interdisciplinary field with biologists turning into data scientists and computer scientists becoming biologists. I'd especially like to thank Ruth Isserlin and Shraddha Pai for helping me out in my early years and providing the fun and obligatory lab banter.

Finally, I would like to thank my parents for supporting me with my decision to pursue a PhD, and being there throughout the whole process. Along with everyone else mentioned, you've added another scientist to the world, and I hope to take the training I have received and make an impact on people's lives.

Contents

List of Figures

List of Tables

List of Abbreviations

AC - astrocytic

ATAC-seq - assay for transposase-accessible chromatin using sequencing

$CaCl_2$ – Calcium Chloride

CGH – comparative genome hybridization

CNV – copy number variant

CO_2 – carbon dioxide

CONICS - Copy-number analysis in single-cell RNA-Sequencing

CONOS - Clustering On Network Of Samples

CrESCENT – Cancer Single Cell Expression Toolkit

CRISPR - clustered regularly interspaced short palindromic repeats

CSC – cancer stem cell

DAPI - 4′,6-diamidino-2-phenylindole

DE – differentially expressed

DM – diffusion map axis

DMSO - Dimethylsulfoxide

DNA – deoxyribonucleic acid

DNAm – DNA methylation

DNase – deoxyribonuclease

EDTA - ethylenediaminetetraacetic acid

EGA – European Genome/Phenome Archive

ESTIMATE - estimation of stromal and immune cells in malignant tumor tissues using expression data

FACS – flourescence activate cell sorting

FDR – false discovery rate

FPKM – fragments per kilobase million

GATK – Genome Analysis Toolkit

GBM - glioblastoma

GO – Gene Ontology

GRCh38 - Genome Reference Consortium human genome assembly 38

gRNA – guide RNA

GSC – glioma/glioblastoma stem sell

GSEA – Gene Set Enrichment Analysis

GSVA – Gene Set Variation Analysis

GTEx – Genome Tissue Expression Project

hg19 – human genome 19

hg38 – human genome 38

KO - knockout

KNN – K-nearest neighbors

MES - mesenchymal

miRNA – microRNA

$Mg(Ac)_2$ – Magnesium Acetate

MNN – mutual nearest neighbors

NOD/scid – non-obese diabetic/severe combined immunodeficiency

NPC – neural progenitor cell

OPC - oligodendrocyte progenitor cell

PBS – phosphate buffered saline

PC – principal component

PCA – principal components analysis

qBF – quantile-normalized Bayes Factor

qPCR – quantitative polymerase chain reaction

REB – research ethics board

RNA – Ribonucleic Acid

RNA-seq – RNA sequencing

SCID – severe combined immunodeficiency

scRNA-seq – single cell RNA sequencing

snRNA-seq – single nuclei RNA sequencing

SNN – shared nearest neighbor

SNP – single nucleotide polymorphism

SNV – single nucleotide variant

TAM – tumor associated macrophage/microglia

Tris-HCl – Tris Hydrochloride

t-SNE – t distributed stochastic neighbor embedding

TKOv3 – Toronto Knockout library version 3

UMAP – uniform manifold approximation

UMI – unique molecular identifiers

WGS – whole genome sequencing

1 Chapter 1: Introduction

1.1 Glioblastoma Disease and Treatment

Glioblastoma Multiforme (GBM) is a disease with a terrible prognosis, having a median survival time of ~12-15 months for patients undergoing current treatment regimens[1], with survival times of ~3 months for those undergoing surgery alone[2,3], and a 5 year survival rate under 10%[4]. Typically, GBM is diagnosed through MRI after patients present with symptoms ranging in severity from headaches to seizures[5]. The absence of specific symptoms early on makes the early detection of this disease difficult; for instance, only 2 in 1000 patients presenting with a headache will be diagnosed with brain tumors[5]. GBM thus represents a difficult disease to diagnose and treat.

A glioblastoma tumor forms a complex environment, with a hypoxic core[6] and infiltration by tumor associated macrophages/microglia (TAMs), which represent 30-50% of tumor mass[7]. GBM primarily manifests in the cerebrum[8] and there is evidence supporting astrocytes, oligodendrocytes, as well as neurons as potential cells of origin[9]. Histologically, GBM presents as a diffusely invasive tumor with irregular nuclei, astroglial appearance of cells, necrosis, visible mitoses, and/or angiogenesis[10]. Within this environment, TAMs promote glioma cell growth and invasiveness, at least in part due to interactions mediated by cytokines[7]. Conversely, glioblastoma cells promote the tumor beneficial phenotypes of TAMs through the secretion of cytokines, suppressing immune response and causing a feedback loop between TAMs and the tumor[7]. Hypoxia promotes angiogenesis and invasion, with tumors expressing high levels of VEGF[11]. The invasiveness of the tumor makes complete surgical resection impossible[12]. Given the complexity of glioblastoma tumors and their interactions with their environment, it is not surprising that they exhibit considerable phenotypic heterogeneity both within tumors and between patients[13–15] .

Despite the wealth of knowledge obtained over the past two decades, treatment of GBM is still largely unchanged from what it was in the early 2000s, with a combination of surgery, radiotherapy, and temozolomide being applied to tumors, only moderately extending survival time for patients[16]. Thus, solving the tumor recurrence problem at a mechanistic level is paramount to effectively treating glioblastoma.

Evidence has accumulated over the past two decades for the existence of glioblastoma cells that resemble neural stem cells (called glioma stem cells, or GSCs)[17–19] which are capable of repopulating patient tumors. Due to their role in tumor recurrence, they represent an enticing target, but development of targeted therapeutics is hampered by their heterogeneity both within and across patients[20]. Important questions remain as to the origin of GSCs, what sort of differentiation hierarchy exists among this population (if any), and how molecular variation translates into differential functional outcomes such as drug response.

In this chapter, I shall introduce genomic analyses and their utility for better understanding heterogeneity in human cancers, application of genomic analyses to glioblastoma, and finally, our current state of knowledge for GSC biology. With this information, it becomes clear that exploring GSC heterogeneity with a variety of -omics data types will further our understanding of this population as well as bring us closer to the goal of targeted therapeutics against it.

1.2 Cancer Genomics and Genomic Data Types

1.2.1 General Description

The problem of mechanistically understanding disease is general and must be addressed to develop better treatments. With the advent of gene expression microarrays, high throughput sequencing, as well as other genomic technologies, studying cancer no longer had to be done at the level of individual genes and researchers could be less constrained by prior knowledge in the questions they were able to ask. Over the past three decades, researchers have been able to perform tasks such as finding mutations likely to contribute to oncogenesis, stratifying patients by risk based on mutation status[21] or characterizing disease stages in prostate cancer based on RNA transcription[22]. Below, I will provide a brief overview of data types used in this thesis, as well as examples of how they may be useful for understanding cancer.

1.2.2 Whole Genome Sequencing (WGS)

Whole genome sequencing is typically done via the shotgun method, which involves sequencing short fragments of DNA and aligning these fragments to a reference genome[23]. This technique can be used to find mutations in cancer, such as single nucleotide variants (SNVs)[24], insertion/deletion mutations (indels)[24], copy number variation (CNVs)[25], and genomic rearrangements[26]. Since only SNVs and CNVs are presented in this thesis, techniques relevant to their detection are what I will discuss further.

In the context of cancer genome sequencing, SNVs and CNVs are typically detected through statistical techniques comparing tumor DNA sequence data to a normal genome reference[24]. For the SNV calling techniques used for this thesis, this is accomplished through modeling read counts as a function of tumor and normal allele fraction[27–30] or, in the case of VarScan2, through a combination of rules based calling of variants relative to reference and a fisher's exact test comparing tumor and normal allele frequencies[31]. For CNVs, segmentation regions can be performed based on probabilistic models of depth ratios[32], and cellularity, ploidy, and allele frequency can be estimated as well [33]. Beyond estimation of copy number profiles, copy number aberration events that produce them can be reconstructed as well[34].

SNVs may or may not affect the function of a gene or a noncoding region, depending on factors such as where the mutation occurs and, if in a gene, how a codon change affects protein structure and function. CNVs are typically associated with changes in gene expression due to increased or decreased availability of particular genes[35], and can be restricted to relatively small regions (down to 50bp)[36] or be present across large swaths of chromosomes (several Mb in size)[35,36].

It should be noted that SNVs and CNVs can also be detected through array based methods[37,38]. In the case of SNP arrays, the main signal measured is the intensity of hybridization signal for a particular allele[38]. With CNVs, a technique called comparative genome hybridization has been commonly used[39], which relies on the comparative hybridization of test (e.g. tumor) and reference (e.g. normal) DNA input to an array made to bind specific genomic regions. As the mutational calling data in this work primarily comes from WGS data (with the exception of estimated CNVs from scRNA-seq data, discussed in Chapter 2), I will not discuss the data processing relevant to these techniques further. However, many studies in the past have used these assays, so they are worth mentioning for the purposes of understanding the signal measured.

There are many examples of genome-wide mutational data providing either clinically valuable information or a better understanding of disease. For instance, Colli and colleagues were able to model response to checkpoint inhibitors (immunotherapy) as a function of the number of non-synonymous mutations detected in tumors with varying degrees of sensitivity and specificity depending on cancer type[40]. In another study of breast cancer tumors, the authors found that copy number events contributed to gene expression phenotype, additional driver mutations for breast cancer, and that a combination of mutational data and gene expression data could be used to cluster tumors into groups with differing survival profiles[41]. Overall, the ability to survey the mutational landscape of cancers has greatly advanced cancer research.

1.2.3 RNA-seq

RNA sequencing measures gene expression profiles from a biological sample using next generation sequencing to sequence cDNAs produced from a reverse transcriptase used on extracted RNA[42]. When the sequencing data is acquired, reads are mapped to the appropriate reference genome using alignment algorithms[42,43]. One can then obtain read counts per gene, which can be normalized in a variety of fashions, such as counts per million or reads per kilobase. Often, normalized count data is log transformed to produce relatively normally distributed data suitable for data exploration, but more sophisticated techniques have been developed to scale count data for sample depth and produce normally distributed, homoscedastic data[44]. In order to compare groups of samples (e.g. tumor and normal, or different apparent subtypes of tumor) for differentially expressed genes, sophisticated statistical models have been developed that can account for the negative binomial distribution of RNA-seq count data as well as other experimental factors in the data (e.g. batch) should these other factors not be totally confounded with the groupings of interest[44,45].

Examples of the utility of RNA-seq data for precision medicine in oncology are abundant, with applications such as stratifying tumors by clinical stage and response to chemotherapy[46] and defining transcriptional networks that distinguish different subsets of lung cancer[47].

1.2.4 ATAC-seq

ATAC-seq allows the profiling of chromatin accessibility, which is a continuous readout for how open chromatin is and likewise its availability for gene expression or binding of transcription factors or other DNA binding proteins[48,49]. This assay works by having a modified transposase with adapter sequences attempt to insert itself into a cell's genome, causing regions of open or accessible chromatin to have its DNA converted into fragments with adapters attached, thereby allowing sequencing of open regions[48]. After defining and mapping peaks of accessible chromatin, samples can be compared and explored through their ATAC-seq profiles in a similar manner as with other datatypes[49], with analyses such as dimensionality reduction, clustering and differential signal analysis. Given that this data often corresponds to sites of DNA binding proteins, analyses can also be performed to assess the overrepresentation of DNA binding motifs in regions of interest[49]. This allows for the profiling of open regions for the potential binding (or differential binding) of collections of DNA binding proteins rather than profiling for binding of one protein at a time as would be done for ChIP-seq.

An excellent example of the utility of ATAC-seq to learn about tumor biology was when The Cancer Genome Atlas project (TCGA) used ATAC-seq to profile 410 tumor samples from 23 cancer types. In their publication, the TCGA found that clusters of samples (based on chromatin accessibility) largely corresponded to tumor type and showed distinct patterns of transcription factor binding at distal regulatory elements, many near sites of disease associated SNPs[50]. They additionally predicted putative regulatory relationships between distal regulatory elements and genes, validating these relationships with a CRISPR based inhibition strategy in which a dead Cas9 protein linked with a KRAB domain caused heterochromatin to form at these elements and inhibited expression of a region's target genes if the region had regulatory function[50]. Overall, these results show the utility of ATAC-seq for understanding gene regulation at a mechanistic level beyond what can be seen through gene expression based analyses.

1.2.5 DNA Methylation

DNA methylation can serve a variety of functions, from promoting gene expression to suppressing it, under a variety of mechanisms. For example, promoter methylation is generally associated with the stabilization of heterochromatin and the prevention of transcription[51], while, in some cases it can lead to increased affinity of a DNA binding protein for its target[52]. The effects of DNA methylation on transcription can manifest through complex mechanisms, with CTCF binding to methylated 'insulator' regions to block the action of enhancer regions on their target genes[51]. In any case, DNA methylation is an epigenetic mechanism by which transcriptional states can be stabilized/potentiated.

DNA methylation can be measured using bisulfite sequencing, which relies on the conversion of cytosines to uracils for non-methylated cytosines in the 'bisulfite conversion' reaction, and the protection of methylated cytosines from this reaction[53]. While this technique allows for genome wide profiling of methylation status, it is cost prohibitive to perform this technique for many samples. A panel covering >850,000 methylation sites, the EPIC array, was developed by Illumina to balance the need for providing good coverage of the genome's potential methylation sites with that of keeping costs reasonably low[53,54]. It uses a similar bisulfite conversion step as with bisulfite sequencing, but with probes designed to capture methylated and unmethylated sequences, with the final readout represented as the ratio of methylated to unmethylated signal (beta-value). From our own analyses (see Chapter 3, *Methods*), we saw that approximately 20,000 gene promoters were covered by the EPIC array, demonstrating that most of the human genome's promoters can be profiled with this method. Additionally, enhancer regions, gene bodies, and intergenic regions are also profiled with these arrays.

DNA methylation data has been used on numerous occasions to link epigenetic regulation to gene expression, with examples in thyroid cancer[55], breast cancer[56], and glioblastoma[20]. In addition to studies linking DNA methylation to gene expression, others have used it to predict tumor types with great accuracy. For example DNA methylation profiles provided a feature set allowing accurate classification of 91 central nervous system tumor classes, with most instances of disagreement with histopathological label leading to reassignment upon a second histopathological assessment[57]. Separately, another group was able to develop a classifier that could distinguish head and neck cancer from lung squamous cell carcinomas[58]. Overall, DNA methylation has been shown on numerous occasions to have a regulatory role in affecting cancer phenotypes and to be reflective of tumor state.

1.2.6 miRNA-seq

miRNAs are small RNAs produced by cells to downregulate the translation of mRNAs into proteins, either through inhibition of translation or through degradation of the mRNA, with both actions mediated by the RISC complex[59]. miRNAs are capable of both promoting and suppressing tumorigenesis, depending on the genes targeted as well as cellular context[59]. Likewise, miRNAs have been proposed and researched as both targets and methods for anti-tumor therapies[60]. miRNA-seq experiments are typically similar to RNA-seq experiments, with the possibility for an enrichment step for short miRNA fragments[61]. Reads are processed and annotated in a similar manner to RNA-seq experiments, but with annotations for miRNA genomic features used[61]. An example where miRNA-seq provided the potential for novel therapies comes from a South Korean study in which certain miRNAs downregulated in lung cancer were found to decrease proliferation of lung cancer cell lines and be associated with longer survival in TCGA lung cancer data[62]. In general, miRNAs provide another lever by which cancer phenotypes can be modulated.

1.2.7 CRISPR Screens

CRISPR screens are a method that can be used to functionally profile the consequence of single gene knockouts for entire libraries of genes, allowing researchers to find dependencies within cell lines at scale[63]. In these screens, a library of guide RNAs are transduced to Cas9 expressing cells at a low multiplicity of infection via lentiviral vectors, leading to cells stably expressing guide RNAs used by Cas9 to cause cleavage (and likewise indel mutations from repair machinery) at targeted loci[63]. This technique was used by Hart and colleagues to show that different cancers have a convergent set of core fitness genes as well as cancer specific fitness genes, illustrating the utility of this method for characterizing biological dependencies in cancer[63,64].

1.2.8 scRNA-seq/snRNA-seq

A limitation of all of the aforementioned techniques is that they only capture population-wise averages of phenotypes in a sample and cannot capture variation within a population. The development of single cell RNA-sequencing addressed this issue, giving researchers the ability to profile gene expression of single cells and tease apart variation within individual cell populations. Briefly, single cells are lysed within droplets containing beads with barcoded adapters, which allow for sequencing and distinction of which RNA derived cDNA reads belong to which cells[65]. An alternate method, single nuclei RNA-seq, allows for the profiling of frozen tissues, though it misses some cytoplasmic RNA species and has a tendency to capture different proportions of cell types present in a population than scRNA-seq[66].

Examples of applications of scRNA-seq data include investigating the microenvironment of tumors, which frequently contain fibroblasts and immune cells[67], and in learning differentiation trajectories with tree/path based methods such as Monocle[68–70] and with ‘velocity’ methods which predict the direction of transcriptional change based on relative amounts of spliced and unspliced transcripts[71,72].

1.3 Genomic Landscape and Subtyping of GBM

The intractability of treating GBM with current methods has led to extensive efforts to molecularly characterize GBM tumors. Common mutations in GBM promoting tumorigenicity include chromosome 10 loss, PTEN mutation/deletion (on chromosome 10), chromosome 7 gain, EGFR mutation/amplification (on chromosome 7), P53 mutation (in secondary GBM, on chromosome 17), and CDKN2 loss/mutation (on chromosome 9)[4,73,74]. These mutations converge on the PI3K/MAPK, P53, and Rb pathways[4]. The vast majority of glioblastomas are primary tumors (>90%)[75], meaning that they are formed de-novo, while approximately 10% are secondary tumors, which tend to develop in younger patients. Almost all primary tumors are IDH1/2 wild type[74,76], and while most secondary glioblastomas, derived from lower grade tumors such as astrocytomas, are mutant for IDH1 or IDH2, with 88% of secondary glioblastomas having an IDH1 mutation[74]. IDH mutations have a tendency to lead to better prognosis, with longer survival[74,77]**.** IDH mutations are thought to exert their pro-survival effects through a variety of metabolic effects, including the depletion of alpha-ketoglutarate and the accumulation of reactive oxygen species[78]. Recently, due to the distinct phenotypes exhibited by IDH mutant tumors in comparison to IDH wild type tumors, the WHO has recommended restricting diagnosis of glioblastoma to IDH wild type tumors, with IDH wild type tumors lacking traditional histologic features of GBM but possessing TERT amplification, EGFR amplification, or the combination of chromosome 7 amplification and chromosome 10 deletion recommended to be diagnosed as GBM as well[79]. In addition to these mutations commonly associated with GBM, mutations inactivating mismatch repair proteins such as MSH6 have been associated with temozolomide resistance[80–83] and recently experimentally proven to be responsible in patient derived cell lines[81]. Additionally, defects in mismatch repair were associated with poorer response to checkpoint inhibitor immunotherapy in gliomas, and it is thought that the accumulation of subclonal mutations was responsible for this association[81].

Beyond genomic alterations, epigenetic state has bearing on survival as well. Methylation of MGMT is associated with longer patient survival[84–88]. The mechanism of this association is thought to be through methylation decreasing the expression of MGMT, a suicide enzyme that repairs alkylation damage caused by temozolomide[89]. MGMT gene expression is anticorrelated with and is thought to be suppressed by promoter methylation for that gene[89].

Due to the existence of biomarkers predictive of therapeutic resistance and survival, attempts have been made to develop molecular subtypes based off of a variety of -omics data types in GBM. The Cancer Genome Atlas (TCGA) project compiled a dataset of 200 GBM samples (subsequently built into a larger dataset) and described 4 transcription based subtypes: Proneural, Mesenchymal, Classical, and Neural[13]. The Proneural subgroup was defined by PDGFRA mutation or amplification and the expression of genes involved in oligodendrocyte specification such as OLIG2, as well as other genes involved in neural development such as TCF4, SOX genes, and ASCL1[13]. The Mesenchymal subgroup was defined by the expression of markers pertaining to epithelial-mesenchymal transition, such as NFKB[13], while the Classical subtype was defined by amplifications in EGFR/chromosome 7. The Neural subtype had neuron related biological pathways and neuron marker genes upregulated[13], but in later work it was found that this subtype was not able to be reproduced and was likely an artifact and was actually a signature of non-cancerous neural tissue[12]. Indeed, a further analysis of IDH wildtype GBMs from the TCGA using only genes that were enriched within tumors relative to tumor margin revealed only the Proneural, Mesenchymal, and Classical subtypes[90]. Despite the consistent reproducibility of at least the Proneural and Mesenchymal subtypes across bulk tumor[90,91] and scRNA-seq[14,15,92] based studies, the only major clinically actionable feature extracted from these transcriptional subtypes was the apparent survival advantage of Proneural GBMs relative to Mesenchymal GBMs[91], but this was confounded by the prevalence of IDH mutant GBMs among the Proneural subtype[13]. A later study of IDH wild type GBMs among the TCGA dataset showed no significant differences in survival between non-Mesenchymal and Mesenchymal GBMs unless analysis was restricted to patients with relatively pure populations of one transcriptional subtype[90]. Thus, there remains further work to be done in assessing the functional and clinical relevance of transcriptional subtypes and in extracting therapeutically valuable information from them.

In addition to transcriptional variation, characterization has been done for the epigenome of GBM. Methylation subtypes discovered by the TCGA[4] bore some resemblance to the transcriptional subtypes discovered earlier, but beyond genomic characterization and a survival advantage of Glioma CpG-Island Methylation Phenotype (G-CIMP) GBM samples over samples without that phenotype, little was done in the way of linking methylation to functional outcomes. More recently, Ma and colleagues re-analyzed methylation data from the TCGA and described 3 methylation based subtypes, again largely corresponding to the Proneural and Mesenchymal transcriptional subtypes[93]. In that study, it was found that the Mesenchymal cluster had shorter survival, but differences between Mesenchymal and Proneural GSCs were not investigated for IDH mutation or G-CIMP status[93]. A pathway analysis on promoter methylation associated with longer survival showed that genes involved in neural development tended to be differentially methylated among longer and shorter surviving patients' GBMs[93]. Additionally, a study on primary and recurrent GBM tumors revealed that promoters affected by neural developmental transcription factors were demethylated in Mesenchymal GBMs relative to Proneural GBMs, and that among recurrent GBMs, promoter methylation of genes involved in neural development and apoptosis increased while Wnt signaling and T-Cell activation genes saw promoter methylation decreased[94]. Overall, these results reveal that there is molecular variation beyond somatic mutations that impacts disease phenotype and progression.

Although studies of bulk tumors have been able to give us amazing insights as to heterogeneity between patients at the genetic, epigenetic, and transcriptional levels, in the past decade, single-cell RNA sequencing (scRNA-seq) studies and clonal genetic tracing work have presented a more complex picture of patient tumors by revealing heterogeneity within individual patients. In the case of scRNA-seq , they have shown patient tumors as admixtures of cells with varying transcriptional phenotypes ranging primarily from a neural/glial-progenitor-like state to a more mesenchymal-like state[14,15], with Neftel, Suva, and colleagues additionally identifying cells resembling astrocytes in the tumor[15]. Complementing this, it was shown that IDH wild type GBM tumors are mixtures of genetic subclones subject to selective pressures within the tumor microenvironment[73], demonstrating a lack of uniformity of GBM cells at the genetic level as well. In all, these results, in conjunction with prior results obtained on bulk -omics data, demonstrate substantial heterogeneity between and within individual patients, making the failure of monotherapies developed for GBM less surprising.

1.4 Cancer Stem Cells in Glioblastoma

A considerable amount of evidence has been accumulated for the existence of stem-like cells that are capable of initiating or re-initiating a tumor. The concept first gained traction with experiments performed by John Dick's lab showing evidence for stem-like cells being responsible for cancer initiation in leukemia[95,96] , showing similarity to normal hematopoietic stem cells in surface marker expression and differentiation potential and requiring only a small titre of cells for cancer initiation. Evidence for stem cells in brain tumors was first shown by Singh and colleagues for non-GBM diseases such as medulloblastoma and pediatric astrocytoma[97], with a CD133+, nestin positive population being able to be isolated in neural stem cell promoting culture conditions. Interestingly, this population could be differentiated to express markers such as GFAP (astrocytes) and Beta-Tubulin III (neurons), supporting its identity as a stem-like state[97]. Similar CD133+ populations in brain tumors were isolated by Singh and colleagues for glioblastoma and other brain tumors, and shown to have, in addition to differentiation potential, the ability to initiate tumors in mice resembling human tumors, and to be able to initiate tumors repeatedly through serial transplantation[18]. Further evidence for the existence of stem-like cells capable of causing tumor recurrence was shown by Bao and colleagues with tumor re-initiation in response to radiotherapy[19] and Chen and colleagues with a quiescent, nestin positive population repopulated tumors after chemotherapy, whose ablation resulted in extended survival of the mouse tumor models[17]. With the wealth of evidence accumulated for GSCs as the root of recurrence in GBM, there is a clear rationale for developing therapeutics against them.

1.5 Heterogeneity of Glioma Stem Cells

As enticing as GSCs are as a potential therapeutic target, the goal of targeting them is complicated by heterogeneity within this population. While CD133 was earlier thought to be a definitive marker for GSCs given the results of Singh and colleagues[18], it was later shown that CD133- cells that resembled a mesenchymal phenotype were capable of initiating tumors[98]. In establishing techniques for culturing GSCs, Pollard, Dirks, and others showed that GSCs showed patient level differences in markers expressed, such as markers for oligodendrocyte progenitor cells (OLIG2, PDGFRA) versus an astrocytic marker (GFAP-delta), as well as patient-specific drug sensitivity profiles[99]. Bhat and colleagues showed that cancer stem cells varied transcriptionally[100] with respect to the Proneural and Mesenchymal phenotypes previously described in bulk tumors described by the TCGA[13], and that a transition to a Mesenchymal phenotype promoted radioresistance[100]. Later Meyer, Dirks, and others showed heterogeneity with respect to sensitivity to a variety of drugs and transcription at the clonal level in patient derived GSC lines, with differential expression found between a TMZ sensitive and TMZ resistant clones[101]. Segerman and colleagues performed a much larger analysis on clonal patient derived GSC lines, and managed to show that, among GSCs, a Mesenchymal-like transcriptional phenotype was associated with multi-drug resistance and radioresistance, and that transcriptional heterogeneity was likely in part due to differential promoter methylation[20]. In addition to the results of Meyer and colleagues and Segerman and colleagues, while not restricted to GSCs, scRNA-seq analyses in GBM have revealed considerable transcriptional heterogeneity within tumors, with mixtures of the TCGA transcriptional subtypes[14,92] as well as putative neural progenitor-like, astrocyte-like, mesenchymal, and oligodendrocyte progenitor-like transcriptional states[15]. Those results for scRNA-seq analyses of GBM indicate the complexity of the tumor environment, and combined with Meyer and colleagues' results, raises the likelihood that similar scRNA-seq experiments in GSCs would reveal considerable transcriptional heterogeneity. Indeed, Bhat and colleagues showed that exposure to TNF-alpha, a cytokine secreted by tumor associated macrophages[7], could convert GSCs from a Proneural state to a Mesenchymal state *in-vitro*[100]. Thus, GSCs represent a highly heterogeneous population both within individual patients as well as across patients, necessitating a mechanistic understanding of GSC heterogeneity and how it impacts functional outcomes such as drug response.

More recent analyses have attempted to address this issue. For example, the HGCC consortium in Sweden applied a drug screen of 1544 compounds to 9 GSC lines, and further characterization of 92 compounds was done for 52 GSC lines with matched genomic, transcriptomic, and DNA methylation data[102]. In this work, the authors found among the most variably effective drugs selected for further characterization, sensitivity divided GSCs into proteasome inhibitor sensitive and resistant lines, and that this division primarily pertained to mutation status in P53 or

CDKN2[102]. Complementing this approach of linking expression, methylation, and mutation data to variably effective drugs was a series of CRISPR knockout screens conducted by MacLeod and colleagues, which revealed gene knockouts that could sensitize GSC lines to temozolomide or confer resistance[103]. In particular, they found that defects in double strand break repair, crosslink repair, and homologous recombination would sensitize GSCs to temozolomide, while defects in mismatch repair conferred resistance[103], the latter result consistent with previous results in GBM tumors and GSC lines[80,81]. Overall, these studies reveal that genomics data can be linked to functional outcomes in a high throughput manner, and that in addition to finding mechanistic insight for known effective compounds, GSC heterogeneity can be used to expand the search for targets for combination therapies.

1.6 Hypotheses on the Hierarchy, Differentiation Potential, and Growth Dynamics of the Glioma Stem Cell

The origin of glioblastoma tumors, as well as the growth and differentiation of GSCs as well as their relationship to the bulk tumor is still an area of active research. Pollard, Dirks, and others, in establishing techniques to isolate GSCs in adherent culture conditions, characterized the transcriptomes of patient derived GSC cell cultures and found that they were more similar to neural stem cells (NSCs) than to brain tissue but nonetheless distinct from NSCs[99]. There is also more direct experimental evidence for astrocytes, oligodendrocytes, NSCs, and neurons as the cell of origin for GBM[9,104]. Much of this evidence comes from mutating or knocking down genes *in vivo* (via Cre/Lox experiments in mice) or in cultured cells that could form GBM-like tumors in mice[105,106]. For instance, cultured astrocytes that initially expressed astrocyte markers, upon knockdown of P53 in combination with NF1 knockdown or an activated form of H-RAS, could form tumors in mice and gave an expression phenotype similar to neural stem cells[105]. Further supporting the notion of de-differentiation of committed CNS lineages, it has been shown that astrocytes can exhibit a dedifferentiated phenotype in response to stroke or stab wound injury *in vivo*[107,108]. In particular, modulation of signaling molecules such as Notch1 and TNF-alpha could potentiate this phenotype, allowing the production of new neurons *in vivo*[107] and differentiation into non-astrocytic lineages *in vitro*[108]. While these experimental approaches are valuable for elucidating how GBM tumors might arise, they do not necessarily prove a cell of origin in human GBMs.

In addition to attempting to assign a cell of origin to GBM, another area of exploration is the origins of cellular heterogeneity, such as clonal evolution within a patient and differentiation potential/cellular hierarchy (if such hierarchy exists). Recent work by Korber and colleagues showed clonal chromosome 7 amplifications, chromosome 9 or 10 deletions were very common among GBM tumors, and that among the most affected genes in these regions (EGFR, CDKN2A/B, PTEN), 81% of samples had a mutation in those genes[73]. Subclonality of TERT in some tumors provided evidence for selection of clones after tumor initiation[73]. Additional analyses performed by the authors led them to suggest that due to most tumor cells dying, only up to 31% of tumor cell divisions would contribute to tumor growth, and that given their estimated growth dynamics, the first GBM tumor cell could appear up to 7 years prior to diagnosis[73]. Additional work by Wang and colleagues showed that among non-hypermutated GBMs, about 45% of mutations were shared between the primary tumor and the recurrence, suggesting the selection or survival of a non-dominant clone upon temozolomide treatment[90]. They additionally found evidence for convergent evolution in several genes, such as P53 and PDGFRA, suggesting clonal divergence and convergent evolution of clones long before treatment is applied[90].

The concept of clonal evolution may also be extended to glioma stem cells. Lan and colleagues examined clone size distributions among tumor cells in serial xenograft transplantation experiments, and determined with a mathematical birth/death model that their negative binomial shape was consistent with a proliferative hierarchy in which slow cycling GSCs give rise to faster cycling progenitors which in turn output differentiated, post-mitotic cells[109]. They found that upon temozolomide treatment, most clones in tumors were eliminated, with surviving clones having size distributions that deviated from the negative binomial distribution but could be accounted for by giving additional resistance to apoptosis[109]. Given that only approximately 3% of barcoded cells were detected as clones within xenografts, with further reduction of surviving clones in subsequent passages[109], it is likely that only a small fraction of GBM cells possess tumorigenicity. Considering these results, as well as the prior evidence for a quiescent cancer stem cell population being required to repopulate a tumor after temozolomide treatment[17], it is likely that, whatever the cell of origin of GBM, and whatever driver mutations lead to its transformation to a malignant state, this cell or its descendants must adopt a GSC phenotype in order to sustain tumor growth.

Beyond clonal structure and evolution of GBM cells and GSCs, there is evidence for differential stemness and differentiation capacity based on factors such as transcriptional state and exposure to extracellular stimuli. Complementing the earlier discussed results involving astrocytes de-differentiating into stem-like cells upon stab-wound injury[107,108], it was recently shown that the number of tumor initiating cells increased in GBM cell cultures exposed to temozolomide relative to those that were not, and that the TMZ cells expressed higher levels of neural stem cell markers such as SOX2 and OCT4[110]. Interestingly, this effect was found to be dependent on the Damage Associated Molecular Pattern (DAMP) protein HMGB1and TLR signaling[110], suggesting that apoptosis caused by TMZ treatment might trigger repurposed inflammatory/damage related signaling pathways that in turn promote de-differentiation. Combined with earlier discussed results from Bhat and colleagues regarding a Proneural to Mesenchymal transition in transcriptional phenotype mediated by TNF-alpha[100], it is clear that GSC state is plastic and can be affected by exposure to external stresses and stimuli. In addition to plasticity in response to stimuli, differentiation capacity in response to stimuli can be affected by transcriptional state and the availability/absence of particular signaling pathways[111]. Early work by Pollard, Dirks, and others showed that there was patient specific variation in differentiation potential of GSCs in culture upon growth factor withdrawal[99] in cell lines showing inter-patient variation in expression of neural lineage markers. The Dirks lab, in collaboration with others, later showed that ASCL1 high GSCs (transcriptionally similar to the Proneural transcriptional program) could be induced to differentiate into neuron-like TUBB3 expressing cells via Notch inhibition, while ASCL1 low GSCs (transcriptionally similar to the Mesenchymal transcriptional program) could not, and this effect was thought to be mediated through modulation of chromatin state for promoters and enhancers of neuronal genes[111]. These results were followed up by a study showing that GSCs dependent on WNT for tumorigenesis and stem-like state were also transcriptionally similar to the Proneural phenotype, and that WNT inhibitor induced differentiation in this subset of GSCs was dependent on ASCL1, again suggesting a role for ASCL1 in potentiating differentiation into a neuron-like, less proliferative and tumorigenic state[112]. More recent work gives further examples illustrating the mechanisms of maintaining a GSC state and preventing differentiation. For example, it was recently shown that the absence of YAP and TAZ results in impaired initiation of GBM tumors, with the absence of these proteins resulting in impaired de-differentiation of patient derived cell lines and in increased differentiation as well as the inability for injected GBM cells to cause tumors in mouse xenograft models[113]. Additionally, SOX2 has recently been shown to be phosphorylated in order to protect from degradation, resulting in maintenance of GSC phenotype[114]. Overall, these results suggest that a balance exists between the GSC stem state and more differentiated states.

Overall, GSCs appear to represent a state analogous to a traditional stem cell hierarchy, capable of producing the rest of a GBM tumor, both initially and after treatment. Based on existing evidence, GSCs appear to be derived from a combination of a tumorigenic genetic background as well as stressors that can promote dedifferentiation of neural/glial cell types. Clonal heterogeneity at the genetic and transcriptomic levels are likely at least in part responsible for GBM tumors' resilience in response to treatment in cases beyond the obvious case of mismatch repair deficiency. Additionally, the activation of stemness programs in response to stress or (importantly) treatment, and the varied behavior and differentiation capacity of GSCs based on transcriptional state, reveal a complex system in which there is no statically defined population of GSCs, but rather a subset of phenotypic (i.e. transcriptomic, mutational, epigenetic, etc.) space that can be entered or left upon appropriate conditions and is primarily defined by the ability to initiate tumors.

1.7 Rationale for Project and Brief Results Summary

Given the wealth of evidence that GSCs are responsible for tumor recurrence, they present an enticing therapeutic target. However, their heterogeneity with respect to drug response presents a challenge to this goal. While there has been tremendous progress in the past two decades in characterizing GBM and GSC heterogeneity, the area still has open questions as to what sort of heterogeneity exists at the single cell level, and, mechanistically, how do different -omics layers function together and independently of one another to produce a phenotype in GSCs. The former question is worth addressing as it would inform us as to whether there are therapeutically relevant subpopulations within GSCs that might not be captured through bulk -omics assays (e.g. RNA-seq), or more broadly what the full transcriptional landscape of GSCs looks like. The latter question is important as, while there have been some multi -omics analyses performed in the past, a fuller picture of how different biological processes function together and independently in producing GSC heterogeneity would give a more holistic view of what integrated and independent biological circuits exist to produce GSC phenotypes. In the two data chapters of this thesis, I attempted to address both of these questions.

Work detailed in Chapter 2 of this thesis was aimed at characterizing GSC heterogeneity at the single cell level. Here, I worked with Laura M. Richards from Dr. Trevor Pugh’s lab in analyzing scRNA-seq data from 26 patient derived GSC cell lines from the Dirks and Weiss labs and found that GSC heterogeneity largely resided on one major transcriptional axis, defined by anticorrelated Developmental and Injury Response transcriptional programs. Further characterization was done using CRISPR-Cas9 screens from the Angers lab, which revealed differential perturbational sensitivity depending on location along this axis, and an integrated analysis with a combined dataset of GSC and patient tumor scRNA-seq data, which revealed the continued presence of the Developmental/Injury Response axis in patient tumors and a stem to astrocyte differentiation gradient progressing from GSCs to patient tumor cells orthogonal to that axis. We thus managed to show that GSC heterogeneity could be expressed as a continuum spanned by two major axes, the first carrying consequences for biological dependencies (e.g. metabolic, signaling) and the latter being analogous to a normal stem cell differentiation gradient.

In Chapter 3, I attempted to expand on our findings in Chapter 2 and address the question of how biological processes function in tandem or independently to produce GSC heterogeneity. Here, I analyzed six data types (RNA-seq, DNA methylation, ATAC-seq, miRNA-seq, Single Nucleotide Variants, Copy Number Variation) from 54 patient derived GSC lines from the Dirks and Weiss labs. I found four major multi -omics axes, one corresponding to temozolomide induced hypermutation in mismatch repair, two corresponding to transcription orthogonal variation in chromatin accessibility and promoter methylation, and a multi -omics axis revealing coordinated activity among transcription, miRNA based gene suppression, promoter methylation of Injury Response genes, and copy number variation associated with gene expression. Overall, these results support a model in which mismatch repair deficiency (which causes a hypermutation phenotype) and a multi -omics Developmental/Injury response axis represent two separate dimensions in which GSCs exhibit heterogeneous functional outcomes such as perturbational sensitivity and tumor aggression.

2 Chapter 2: Gradient of Developmental and Injury Response transcriptional states defines functional vulnerabilities underpinning glioblastoma heterogeneity

This work was originally published in Nature Cancer: Richards, L.M.*, Whitley, O.K.N.*, MacLeod, G., Cavalli, F.M.G., Coutinho, F.J., Jaramillo, J.E., Svergun, N., Riverin, M., Croucher, D.C., Kushida, M., Yu, K., Guilhamon, P., Rastegar, N., Ahmadi, M., Bhatti, J.K., Bozek, D.A., Li, N., Lee, L., Che, C., Luis, E., Park N.I., Xu, Z., Ketela, T., Moore, R.A., Marra, M.A., Spears, J., Cusimano, M.D., Das, S., Bernstein, M., Haibe-Kains, B., Lupien, M., Luchman, H.A., Weiss, S., Angers, S., Dirks, P.B.#, Bader, G.D.#, Pugh, T.J.#. Gradient of Developmental and Injury Response transcriptional states defines functional vulnerabilities underpinning glioblastoma heterogeneity. Nature Cancer 2, 157–173 (2021). https://doi.org/10.1038/s43018-020-00154-9

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. O.K.N.W. performed pathway, bulk RNA-seq, PCA analysis, signature analysis and outlier detection, and defined the injury response and stem cell-mature gene expression signatures. O.K.N.W. developed the cluster stability assessment procedure for bulk RNA-seq clustering. O.K.N.W. and L.M.R. performed scRNA-seq analyses on both internally generated and public datasets. O.K.N.W. performed malignant cell identification in publicly available datasets as well as signature scoring. O.K.N.W. developed a logistic regression classifier to identify stem-like tumour cells, and performed RNA-velocity to characterize a stem-to astrocyte trajectory. O.K.N.W. performed pathway analysis on genome wide CRISPR-Cas9 screen data, with L.M.R. and G.M. performing additional analyses. 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. 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. 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., O.K.N.W., G.D.B., P.B.D. and T.J.P. wrote the manuscript with feedback from all authors.

2.1 Abstract

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.

2.2 Introduction

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[4,14,15] . Molecular stratification of GBMs into transcriptional subgroups[13,90] (proneural, mesenchymal and classical) has not led to the development of successful targeted therapies [115], 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[14,15,92,116,117], 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[18,109,118,119]. GSCs hijack developmental stem cell programs to drive and maintain tumor growth, as well as acquire resistance mechanisms to evade chemotherapy and radiotherapy[17,19,120]. However, it is still unclear how diversity within the GSC pool may affect the cellular composition and growth of GBMs.

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.

2.3 Results

2.3.1 Transcriptional heterogeneity within GSCs

To enrich for rare stem-like cells within primary tumors, we used established serum-free culturing methods[99,121] to generate a collection of patient-derived GSCs capable of sustaining growth in vitro and initiating tumors in mice (*Methods*). This method supports the growth of a diversity of clones that closely matches human GBM xenografts[109] 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.

To explore GSC heterogeneity within individual patients, we clustered GSCs from each sample independently using extensive hyperparameter optimization and validation with multiple algorithms (*Methods*, Fig 2.1). We discovered substantial intra-GSC heterogeneity, uncovering two to six transcriptional subpopulations per GSC, totaling 86 clusters across 29 samples (Fig. 2.2A,B and Fig 2.1), 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. 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 (Fig. 2.3A-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[122] (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[123] and self-renewal in brain tumor stem cells[124]. GSCs with a larger fraction of actively cycling cells displayed increased aggressiveness and reduced survival upon implantation in an orthotopic xenograft model (Fig 2.3C). 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 (Fig 2.3A). 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[125], while the second cluster (C2) overexpressed ASCL1, a transcription factor critical for neuronal differentiation that suppresses tumorigenicity in GSCs[111] (Fig 2.3E,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 GBM.

Figure 2.1 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.

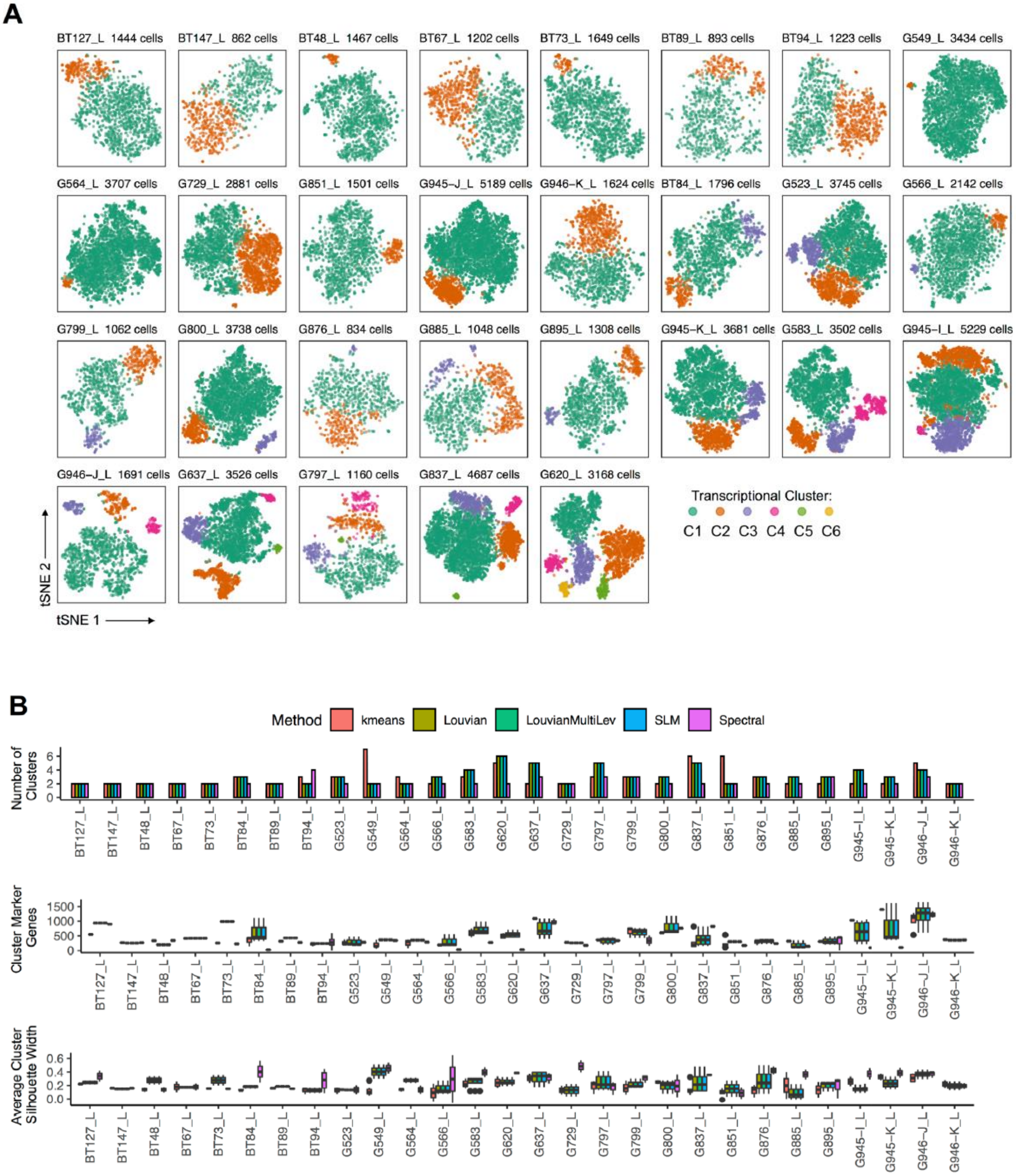

A
BT127_L 1444 cells
BT147_L 862 cells
BT48_L 1467 cells
BT67_L 1202 cells
BT73_L 1649 cells
BT89_L 893 cells
BT94_L 1223 cells
G549_L 3434 cells
G564_L 3707 cells
G729_L 2881 cells
G851_L 1501 cells
G945-J_L 5189 cells
G946-K_L 1624 cells
BT84_L 1796 cells
G523_L 3745 cells
G566_L 2142 cells
G799_L 1062 cells
G800_L 3738 cells
G876_L 834 cells
G885_L 1048 cells
G895_L 1308 cells
G945-K_L 3681 cells
G583_L 3502 cells
G945-I_L 5229 cells
G946-J_L 1691 cells
G637_L 3526 cells
G797_L 1160 cells
G837_L 4687 cells
G620_L 3168 cells
Transcriptional Cluster:
C1 C2 C3 C4 C5 C6
tSNE 2
tSNE 1
B
Method
kmeans
Louvian
LouvianMultiLev
SLM
Spectral
Number of Clusters
0 2 4 6
Cluster Marker Genes
0 500 1000 1500
Average Cluster Silhouette Width
0.0 0.2 0.4 0.6
BT127_L BT147_L BT48_L BT67_L BT73_L BT84_L BT89_L BT94_L G523_L G549_L G564_L G566_L G583_L G620_L G637_L G729_L G797_L G799_L G800_L G837_L G851_L G876_L G885_L G895_L G945-I_L G945-K_L G946-J_L G946-K_L

Figure 2.2 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 (**Fig 2.1A**). 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.

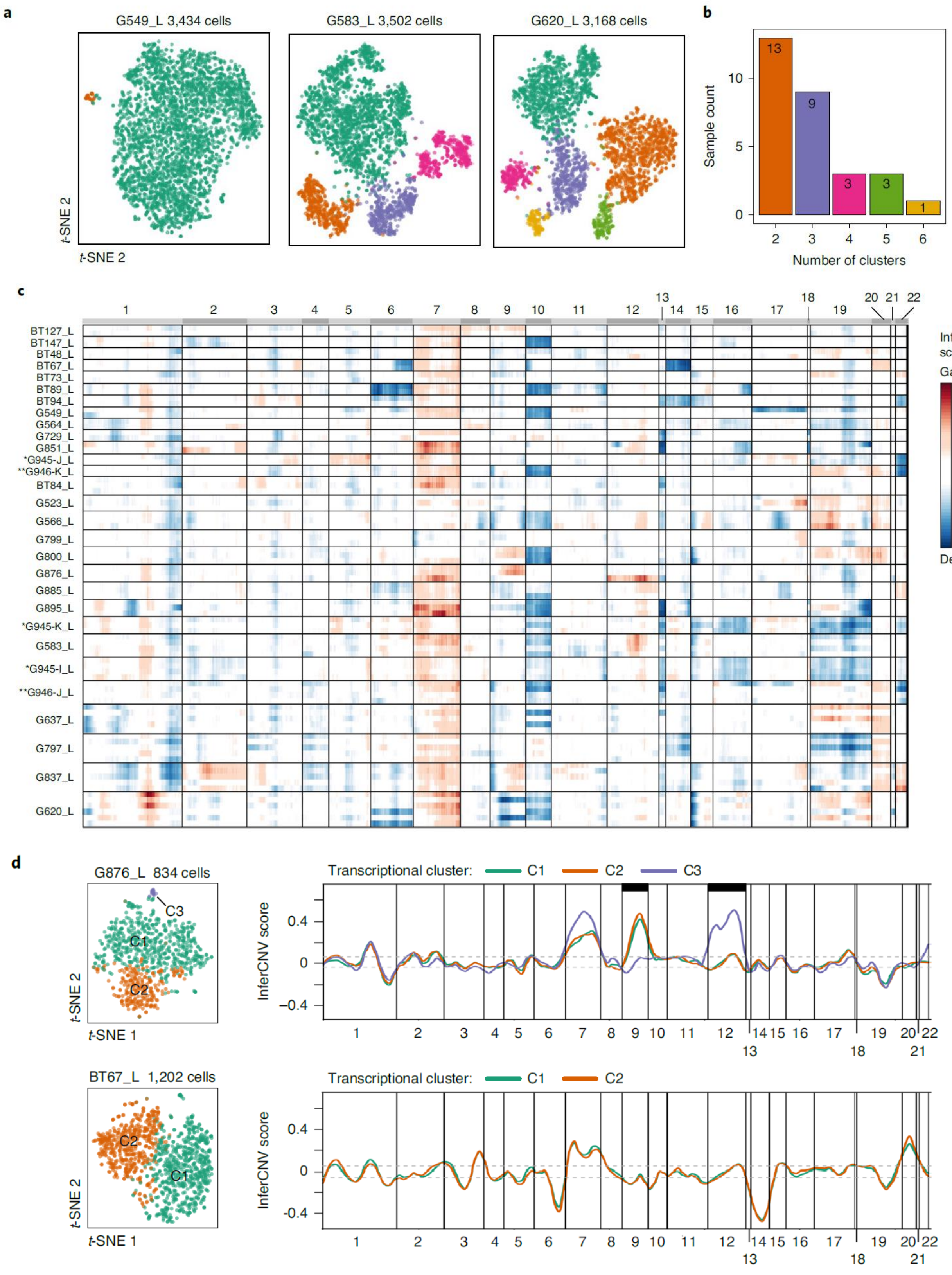
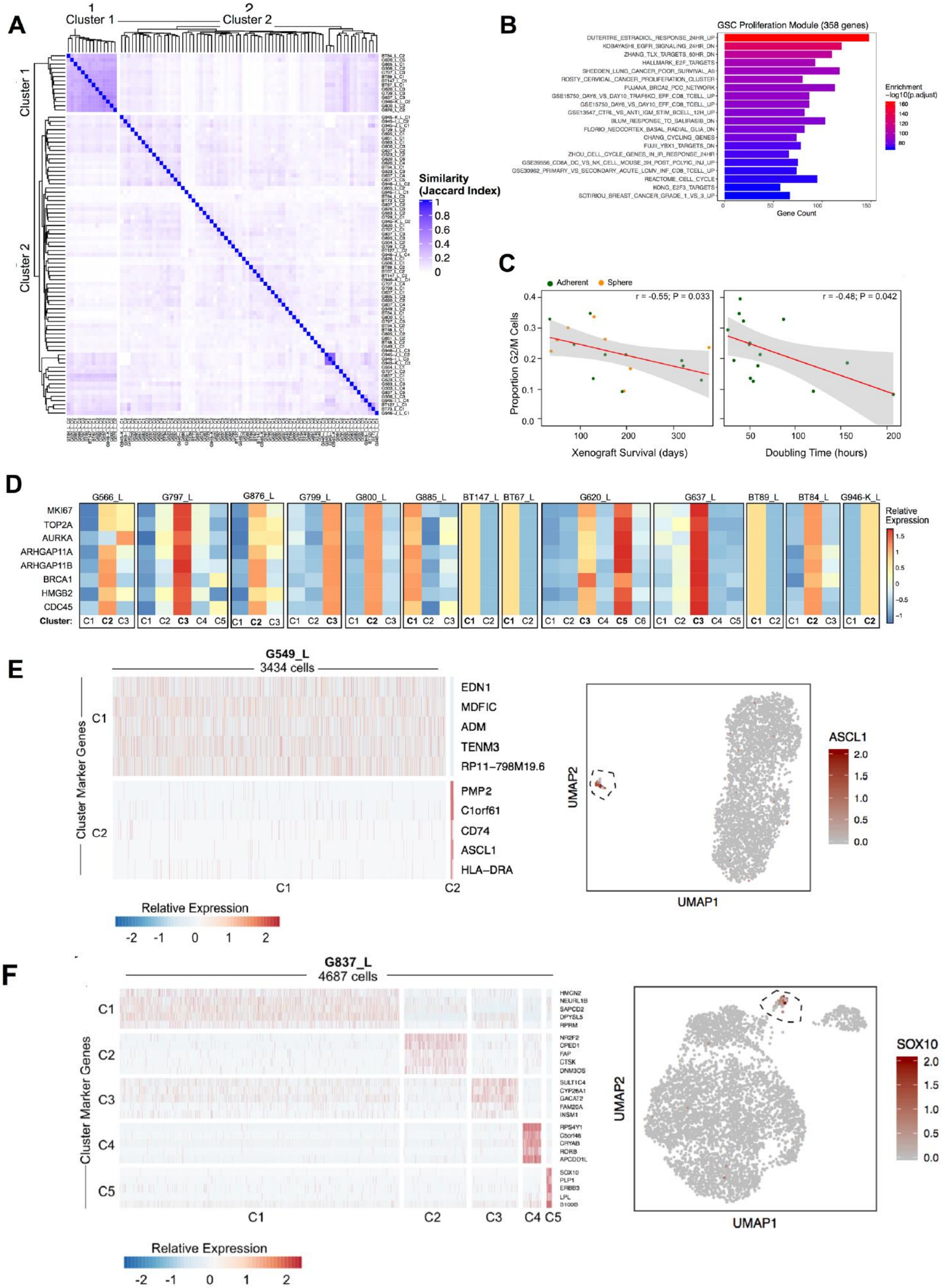

a
G549_L 3,434 cells
G583_L 3,502 cells
G620_L 3,168 cells
t-SNE 2
t-SNE 2
b
13
9
3
3
1
Sample count
0
5
10
2
3
4
5
6
Number of clusters
c
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
BT127_L
BT147_L
BT48_L
BT67_L
BT73_L
BT89_L
BT94_L
G549_L
G564_L
G729_L
G851_L
*G945-J_L
**G946-K_L
BT84_L
G523_L
G566_L
G799_L
G800_L
G876_L
G885_L
G895_L
*G945-K_L
G583_L
*G945-I_L
**G946-J_L
G637_L
G797_L
G837_L
G620_L
InferCNV score
Gain
0.4
0
−0.4
Deletion
d
G876_L 834 cells
C3
C1
C2
t-SNE 2
t-SNE 1
Transcriptional cluster: C1 C2 C3
InferCNV score
0.4
0
−0.4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
BT67_L 1,202 cells
C2
C1
t-SNE 2
t-SNE 1
Transcriptional cluster: C1 C2
InferCNV score
0.4
0
-0.4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22

Figure 2.3 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 **Fig. 2.3A**. **(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).



A
1
2
Cluster 1
Cluster 2
Similarity
(Jaccard Index)
1
0.8
0.6
0.4
0.2
0
B
GSC Proliferation Module (358 genes)
DUTERTRE_ESTRADIOL_RESPONSE_24HR_UP
KOBAYASHI_EGFR_SIGNALING_24HR_DN
ZHANG_TLX_TARGETS_60HR_DN
HALLMARK_E2F_TARGETS
SHEDDEN_LUNG_CANCER_POOR_SURVIVAL_A6
ROSTY_CERVICAL_CANCER_PROLIFERATION_CLUSTER
PUJANA_BRCA2_PCC_NETWORK
GSE15750_DAY6_VS_DAY10_TRAF6KO_EFF_CD8_TCELL_UP
GSE15750_DAY6_VS_DAY10_EFF_CD8_TCELL_UP
GSE13547_CTRL_VS_ANTI_IGM_STIM_BCELL_12H_UP
BLUM_RESPONSE_TO_SALIRASIB_DN
FLORIO_NEOCORTEX_BASAL_RADIAL_GLIA_DN
CHANG_CYCLING_GENES
FUJII_YBX1_TARGETS_DN
ZHOU_CELL_CYCLE_GENES_IN_IR_RESPONSE_24HR
GSE39556_CD8A_DC_VS_NK_CELL_MOUSE_3H_POST_POLYIC_INJ_UP
GSE30962_PRIMARY_VS_SECONDARY_ACUTE_LCMV_INF_CD8_TCELL_UP
REACTOME_CELL_CYCLE
KONG_E2F3_TARGETS
SOTIRIOU_BREAST_CANCER_GRADE_1_VS_3_UP
0
50
100
150
Gene Count
Enrichment
−log10(p.adjust)
160
140
120
100
80
C
● Adherent
● Sphere
r = -0.55; P = 0.033
r = -0.48; P = 0.042
Proportion G2/M Cells
0.4
0.3
0.2
0.1
0.0
100
200
300
Xenograft Survival (days)
50
100
150
200
Doubling Time (hours)
D
G566_L
G797_L
G876_L
G799_L
G800_L
G885_L
BT147_L
BT67_L
G620_L
G637_L
BT89_L
BT84_L
G946-K_L
MKI67
TOP2A
AURKA
ARHGAP11A
ARHGAP11B
BRCA1
HMGB2
CDC45
Cluster:
C1 C2 C3
C1 C2 C3 C4 C5
C1 C2 C3
C1 C2 C3
C1 C2 C3
C1 C2 C3
C1 C2
C1 C2
C1 C2 C3 C4 C5 C6
C1 C2 C3 C4 C5
C1 C2
C1 C2 C3
C1 C2
Relative
Expression
1.5
1
0.5
0
−0.5
−1
E
G549_L
3434 cells
Cluster Marker Genes
C1
C2
EDN1
MDFIC
ADM
TENM3
RP11−798M19.6
PMP2
C1orf61
CD74
ASCL1
HLA−DRA
C1
C2
Relative Expression
-2
-1
0
1
2
UMAP2
UMAP1
ASCL1
2.0
1.5
1.0
0.5
0.0
F
G837_L
4687 cells
Cluster Marker Genes
C1
C2
C3
C4
C5
HMGN2
NEURL1B
SAPCD2
DPYSL5
RPRM
NR2F2
CPED1
FAP
CTSK
DNM3OS
SULT1C4
CYP26A1
GACAT2
FAM20A
INSM1
RPS4Y1
C5orf46
CRYAB
RORB
APCDD1L
SOX10
PLP1
ERBB3
LPL
S100B
C1
C2
C3
C4 C5
Relative Expression
-2
-1
0
1
2
UMAP2
UMAP1
SOX10
2.0
1.5
1.0
0.5
0.0

2.3.2 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. 2.2C,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 (Fig 2.4).

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. 2.2C), consistent with reported frequency and evolutionary timing in GBMs[4,13,73,126]. 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 (Fig 2.4D). 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. 2.2D). 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 (Fig 2.4E). 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. 2.2D). 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.

2.3.3 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 (Fig 2.5A,B). Unsupervised clustering identified 61 transcriptional clusters, revealing striking patient-specific

Figure 2.4 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 **Fig 2.4D**. Clusters with significant (Fisher's Exact Test $p < 0.05$) enrichment of marker genes within variable CNV loci are colored dark blue.

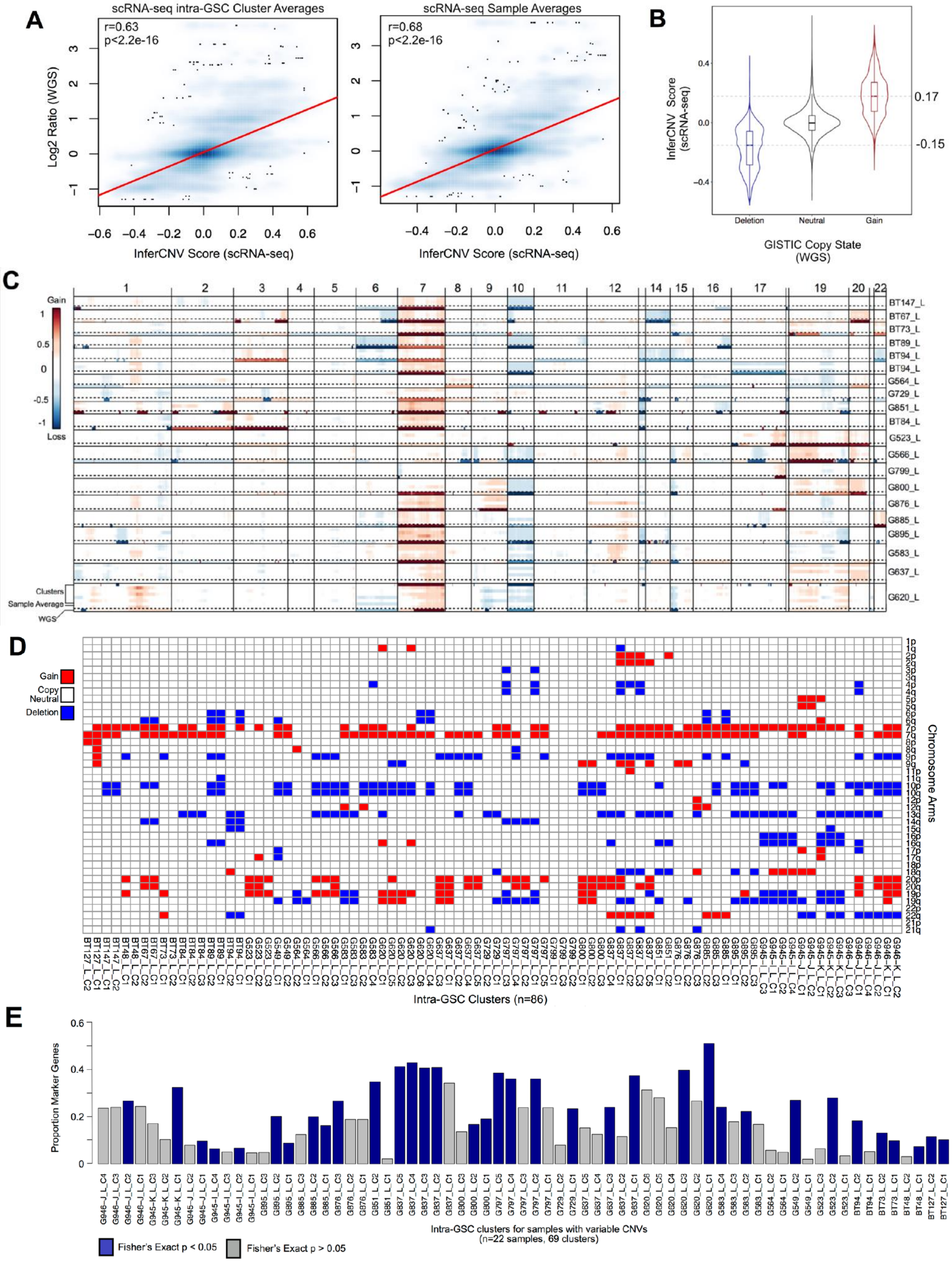
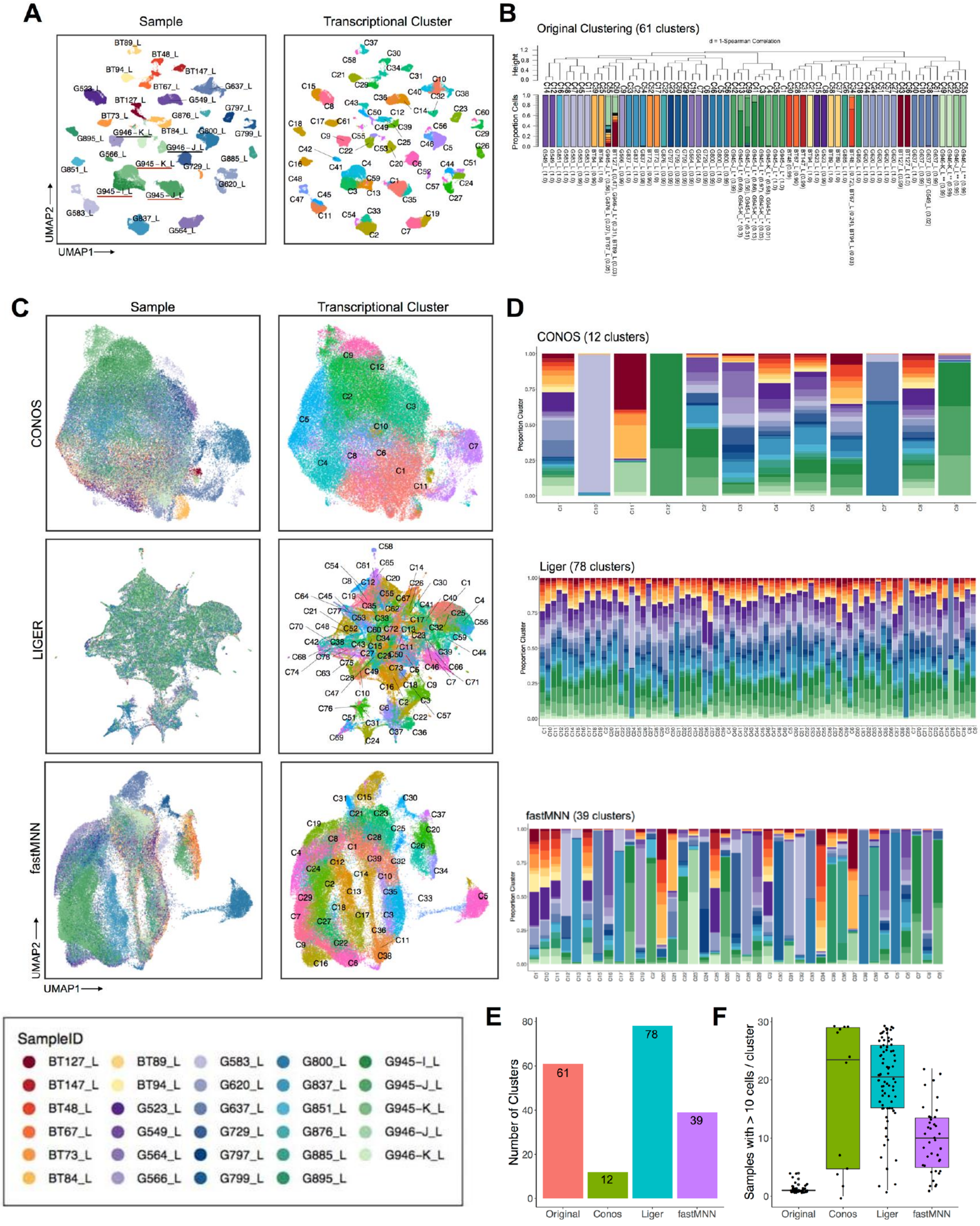

A
scRNA-seq intra-GSC Cluster Averages
r=0.63
p<2.2e-16
Log2 Ratio (WGS)
3
2
1
0
−1
−0.6 −0.4 −0.2 0.0 0.2 0.4 0.6
InferCNV Score (scRNA-seq)
scRNA-seq Sample Averages
r=0.68
p<2.2e-16
3
2
1
0
−1
−0.4 −0.2 0.0 0.2 0.4 0.6
InferCNV Score (scRNA-seq)
B
InferCNV Score
(scRNA-seq)
0.4
0.0
-0.4
0.17
-0.15
Deletion
Neutral
Gain
GISTIC Copy State
(WGS)
C
1
2
3
4
5
6
7
8
9
10
11
12
14
15
16
17
19
20
22
Gain
1
0.5
0
-0.5
-1
Loss
BT147_L
BT67_L
BT73_L
BT89_L
BT94_L
BT94_L
G564_L
G729_L
G851_L
BT84_L
G523_L
G566_L
G799_L
G800_L
G876_L
G885_L
G895_L
G583_L
G637_L
G620_L
Clusters
Sample Average
WGS
D
Gain
Copy
Neutral
Deletion
1p
1q
2p
2q
3p
3q
4p
4q
5p
5q
6p
6q
7p
7q
8p
8q
9p
9q
11p
11q
10p
10q
12p
12q
13q
14q
15q
16p
16q
17p
17q
18p
18q
20p
20q
19p
19q
22p
22q
21p
21q
Chromosome Arms
BT127_L_C2
BT127_L_C1
BT147_L_C1
BT147_L_C2
BT48_L_C1
BT48_L_C2
BT67_L_C2
BT67_L_C1
BT73_L_C1
BT73_L_C2
BT84_L_C1
BT84_L_C2
BT84_L_C3
BT89_L_C2
BT89_L_C1
BT94_L_C2
BT94_L_C1
G523_L_C1
G523_L_C2
G523_L_C3
G549_L_C1
G549_L_C2
G564_L_C2
G564_L_C1
G566_L_C1
G566_L_C2
G566_L_C3
G583_L_C1
G583_L_C3
G583_L_C2
G583_L_C4
G620_L_C1
G620_L_C5
G620_L_C2
G620_L_C3
G620_L_C6
G620_L_C4
G637_L_C3
G637_L_C1
G637_L_C2
G637_L_C4
G637_L_C5
G729_L_C2
G729_L_C1
G797_L_C3
G797_L_C4
G797_L_C2
G797_L_C1
G797_L_C5
G799_L_C1
G799_L_C3
G799_L_C2
G800_L_C1
G800_L_C2
G800_L_C3
G837_L_C4
G837_L_C1
G837_L_C2
G837_L_C3
G837_L_C5
G851_L_C1
G851_L_C2
G876_L_C1
G876_L_C2
G876_L_C3
G885_L_C2
G885_L_C3
G885_L_C1
G895_L_C1
G895_L_C2
G895_L_C3
G945-I_L_C3
G945-I_L_C1
G945-I_L_C2
G945-I_L_C4
G945-J_L_C1
G945-J_L_C2
G945-K_L_C1
G945-K_L_C2
G945-K_L_C3
G946-J_L_C3
G946-J_L_C1
G946-J_L_C4
G946-J_L_C2
G946-K_L_C1
G946-K_L_C2
Intra-GSC Clusters (n=86)
E
0.6
0.4
0.2
0
Proportion Marker Genes
G946-J_L_C4
G946-J_L_C3
G946-J_L_C2
G946-J_L_C1
G945-K_L_C3
G945-K_L_C2
G945-K_L_C1
G945-J_L_C2
G945-J_L_C1
G945-I_L_C4
G945-I_L_C3
G945-I_L_C2
G945-I_L_C1
G895_L_C3
G895_L_C2
G895_L_C1
G885_L_C3
G885_L_C2
G885_L_C1
G876_L_C3
G876_L_C2
G876_L_C1
G851_L_C2
G851_L_C1
G837_L_C5
G837_L_C4
G837_L_C3
G837_L_C2
G837_L_C1
G800_L_C3
G800_L_C2
G800_L_C1
G797_L_C5
G797_L_C4
G797_L_C3
G797_L_C2
G797_L_C1
G729_L_C2
G729_L_C1
G637_L_C5
G637_L_C4
G637_L_C3
G637_L_C2
G637_L_C1
G620_L_C6
G620_L_C5
G620_L_C4
G620_L_C3
G620_L_C2
G620_L_C1
G583_L_C4
G583_L_C3
G583_L_C2
G583_L_C1
G564_L_C2
G564_L_C1
G549_L_C2
G549_L_C1
G523_L_C3
G523_L_C2
G523_L_C1
BT94_L_C2
BT94_L_C1
BT73_L_C2
BT73_L_C1
BT48_L_C2
BT48_L_C1
BT127_L_C2
BT127_L_C1
Intra-GSC clusters for samples with variable CNVs
(n=22 samples, 69 clusters)
Fisher's Exact p < 0.05
Fisher's Exact p > 0.05

Figure 2.5 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 **Fig 2.5A** 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.


A
Sample
Transcriptional Cluster
BT89_L
BT48_L
BT94_L
BT147_L
G637_L
G523_L
BT67_L
BT127_L
G549_L
G797_L
BT73_L
G876_L
BT84_L
G800_L
G895_L
G946 - K_L
G946 - J_L
G799_L
G566_L
G945 - K_L
G729_L
G885_L
G851_L
G620_L
G945 - I_L
G945 - J_L
G583_L
G837_L
G564_L
UMAP2
UMAP1
B
Original Clustering (61 clusters)
d = 1-Spearman Correlation
Height
Proportion Cells
C
Sample
Transcriptional Cluster
CONOS
LIGER
fastMNN
UMAP2
UMAP1
D
CONOS (12 clusters)
Liger (78 clusters)
fastMNN (39 clusters)
Proportion Cluster
0.00
0.25
0.50
0.75
1.00
SampleID
BT127_L
BT147_L
BT48_L
BT67_L
BT73_L
BT84_L
BT89_L
BT94_L
G523_L
G549_L
G564_L
G566_L
G583_L
G620_L
G637_L
G729_L
G797_L
G799_L
G800_L
G837_L
G851_L
G876_L
G885_L
G895_L
G945-I_L
G945-J_L
G945-K_L
G946-J_L
G946-K_L
E
Number of Clusters
0
20
40
60
80
61
12
78
39
Original
Conos
Liger
fastMNN
F
Samples with > 10 cells / cluster
0
10
20
30
Original
Conos
Liger
fastMNN

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 (Fig 2.5C–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[14,15,127–134] and malignant cell lines[135–138] 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 (Fig 2.5B).

2.3.4 GSCs organize along a transcriptional gradient.

To identify core transcriptional programs underpinning inter-GSC heterogeneity, 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 (Fig. 2.6A). Re-running PCA without G800_L revealed a single axis of variation along PC1, separating cells into two prominent groups. (Fig 2.7A).

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 2.7B and Fig 2.6B,C). When compared to cell types found in developing fetal brain[139,140], mature adult brain[141–145] and malignant cell states in GBMs[13,15], these inflamed GSCs best resembled both the Cancer Genome Atlas (TCGA) Mesenchymal subtype and the mesenchymal-like cell state[15] in GBMs, as well as neuroprotective A2 reactive astrocytes (Fig. 2.7C, Fig 2.6B). Interestingly, A2 reactive astrocytes promote neuronal survival and tissue repair in response to ischemic injury[142,146], 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.

Figure 2.6 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. 2.7A**). 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. 2.7A**). **(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. 2.7A**). 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.

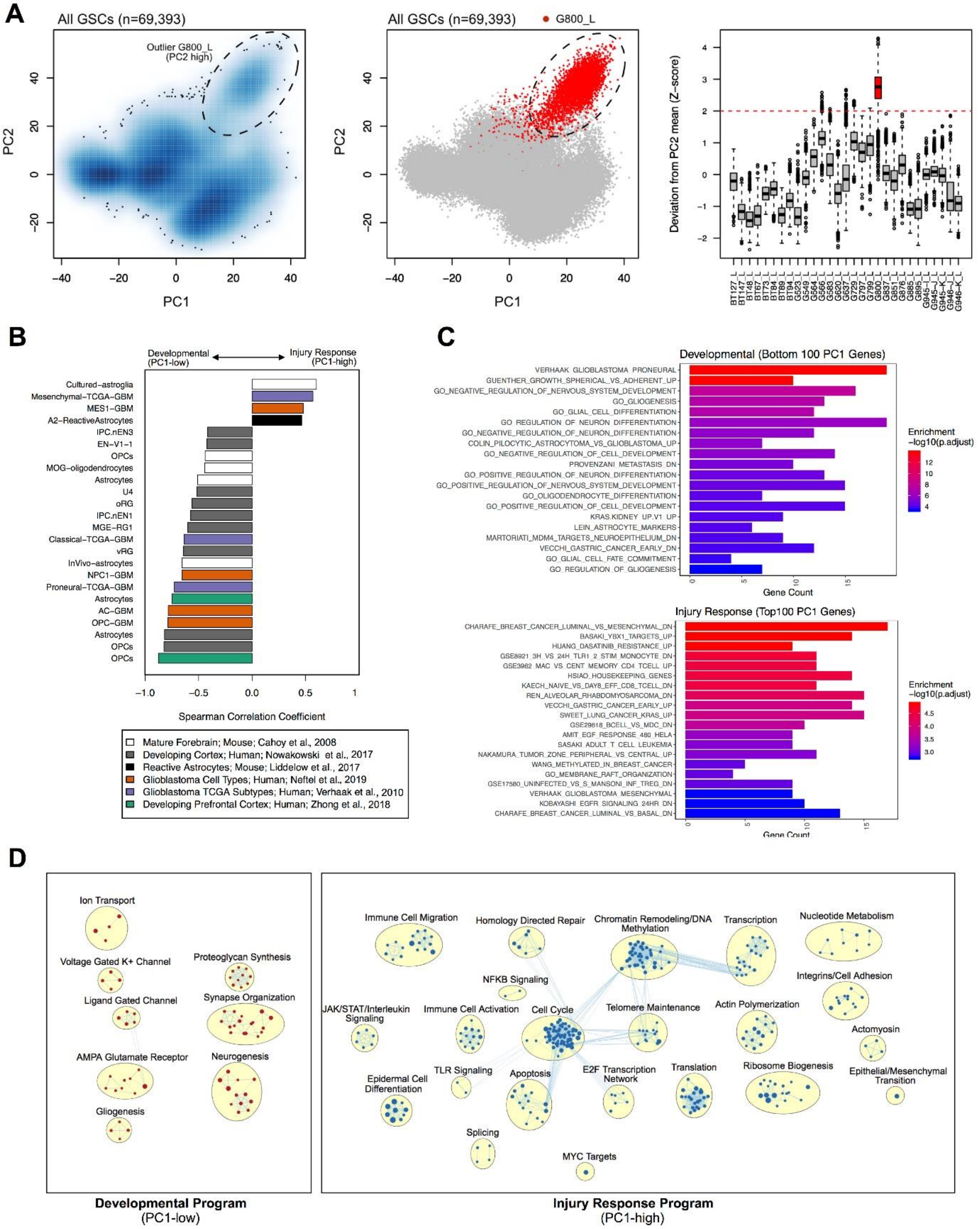
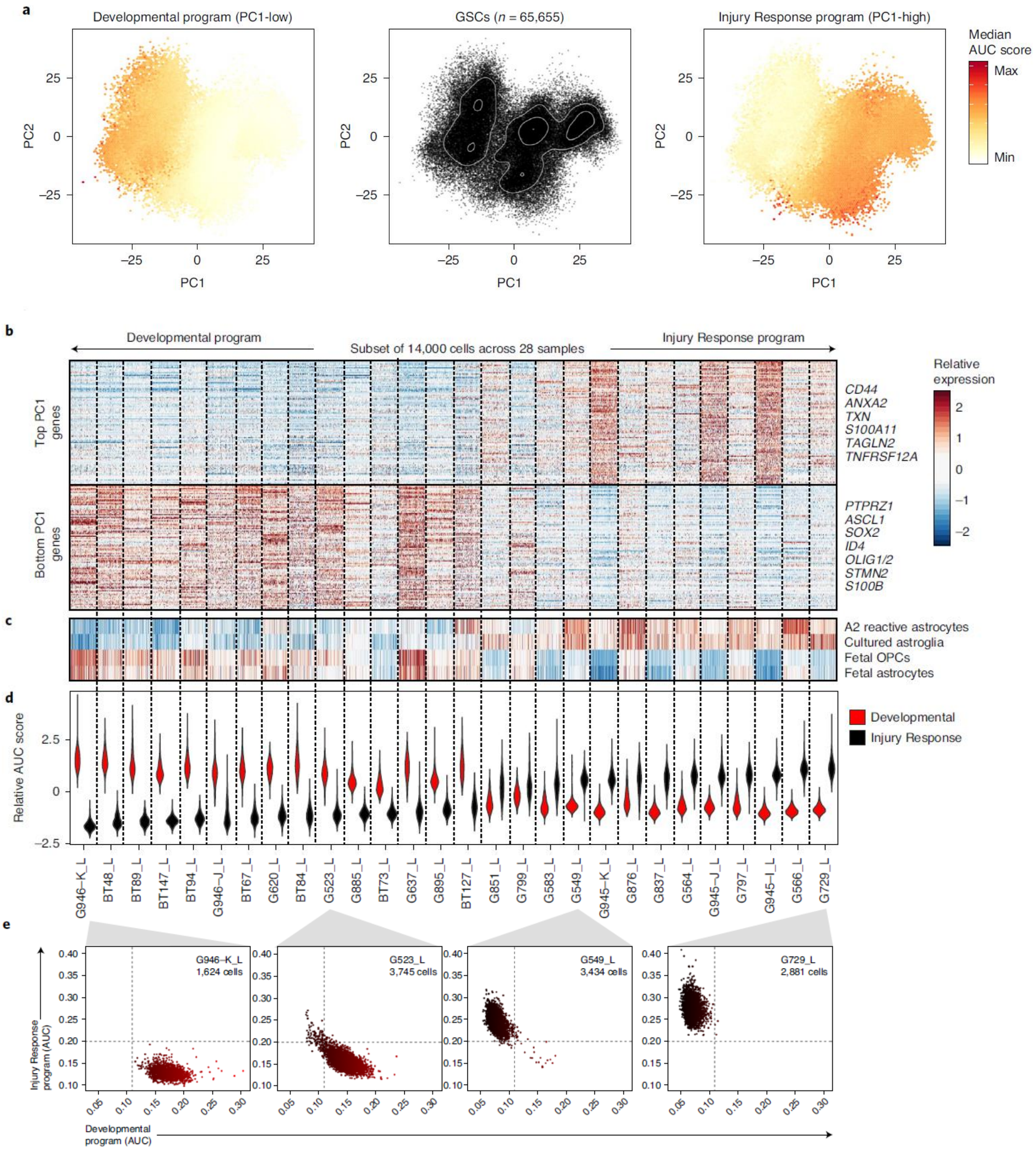

A
All GSCs (n=69,393)
Outlier G800_L
(PC2 high)
PC2
PC1
−40
−20
0
20
40
All GSCs (n=69,393)
G800_L
Deviation from PC2 mean (Z-score)
4
3
2
1
0
−1
−2
BT127_L
BT147_L
BT48_L
BT67_L
BT73_L
BT84_L
BT89_L
BT94_L
G523_L
G549_L
G564_L
G566_L
G583_L
G620_L
G637_L
G729_L
G797_L
G799_L
G800_L
G837_L
G851_L
G876_L
G885_L
G895_L
G945–I_L
G945–J_L
G945–K_L
G946–J_L
G946–K_L
B
Developmental
(PC1-low)
Injury Response
(PC1-high)
Cultured–astroglia
Mesenchymal–TCGA–GBM
MES1–GBM
A2–ReactiveAstrocytes
IPC.nEN3
EN–V1–1
OPCs
MOG–oligodendrocytes
Astrocytes
U4
oRG
IPC.nEN1
MGE–RG1
Classical–TCGA–GBM
vRG
InVivo–astrocytes
NPC1–GBM
Proneural–TCGA–GBM
Astrocytes
AC–GBM
OPC–GBM
Astrocytes
OPCs
OPCs
−1.0
−0.5
0.0
0.5
1
Spearman Correlation Coefficient
□ Mature Forebrain; Mouse; Cahoy et al., 2008
■ Developing Cortex; Human; Nowakowski et al., 2017
■ Reactive Astrocytes; Mouse; Liddelow et al., 2017
■ Glioblastoma Cell Types; Human; Neftel et al., 2019
■ Glioblastoma TCGA Subtypes; Human; Verhaak et al., 2010
■ Developing Prefrontal Cortex; Human; Zhong et al., 2018
C
Developmental (Bottom 100 PC1 Genes)
VERHAAK_GLIOBLASTOMA_PRONEURAL
GUENTHER_GROWTH_SPHERICAL_VS_ADHERENT_UP
GO_NEGATIVE_REGULATION_OF_NERVOUS_SYSTEM_DEVELOPMENT
GO_GLIOGENESIS
GO_GLIAL_CELL_DIFFERENTIATION
GO_REGULATION_OF_NEURON_DIFFERENTIATION
GO_NEGATIVE_REGULATION_OF_NEURON_DIFFERENTIATION
COLIN_PILOCYTIC_ASTROCYTOMA_VS_GLIOBLASTOMA_UP
GO_NEGATIVE_REGULATION_OF_CELL_DEVELOPMENT
PROVENZANI_METASTASIS_DN
GO_POSITIVE_REGULATION_OF_NEURON_DIFFERENTIATION
GO_POSITIVE_REGULATION_OF_NERVOUS_SYSTEM_DEVELOPMENT
GO_OLIGODENDROCYTE_DIFFERENTIATION
GO_POSITIVE_REGULATION_OF_CELL_DEVELOPMENT
KRAS.KIDNEY_UP.V1_UP
LEIN_ASTROCYTE_MARKERS
MARTORIATI_MDM4_TARGETS_NEUROEPITHELIUM_DN
VECCHI_GASTRIC_CANCER_EARLY_DN
GO_GLIAL_CELL_FATE_COMMITMENT
GO_REGULATION_OF_GLIOGENESIS
0
5
10
15
Gene Count
Enrichment
−log10(p.adjust)
12
10
8
6
4
Injury Response (Top100 PC1 Genes)
CHARAFE_BREAST_CANCER_LUMINAL_VS_MESENCHYMAL_DN
BASAKI_YBX1_TARGETS_UP
HUANG_DASATINIB_RESISTANCE_UP
GSE8921_3H_VS_24H_TLR1_2_STIM_MONOCYTE_DN
GSE3982_MAC_VS_CENT_MEMORY_CD4_TCELL_UP
HSIAO_HOUSEKEEPING_GENES
KAECH_NAIVE_VS_DAY8_EFF_CD8_TCELL_DN
REN_ALVEOLAR_RHABDOMYOSARCOMA_DN
VECCHI_GASTRIC_CANCER_EARLY_UP
SWEET_LUNG_CANCER_KRAS_UP
GSE29618_BCELL_VS_MDC_DN
AMIT_EGF_RESPONSE_480_HELA
SASAKI_ADULT_T_CELL_LEUKEMIA
NAKAMURA_TUMOR_ZONE_PERIPHERAL_VS_CENTRAL_UP
WANG_METHYLATED_IN_BREAST_CANCER
GO_MEMBRANE_RAFT_ORGANIZATION
GSE17580_UNINFECTED_VS_S_MANSONI_INF_TREG_DN
VERHAAK_GLIOBLASTOMA_MESENCHYMAL
KOBAYASHI_EGFR_SIGNALING_24HR_DN
CHARAFE_BREAST_CANCER_LUMINAL_VS_BASAL_DN
0
5
10
15
Gene Count
Enrichment
−log10(p.adjust)
4.5
4.0
3.5
3.0
D
Ion Transport
Voltage Gated K+ Channel
Proteoglycan Synthesis
Ligand Gated Channel
Synapse Organization
AMPA Glutamate Receptor
Neurogenesis
Gliogenesis
Immune Cell Migration
Homology Directed Repair
Chromatin Remodeling/DNA Methylation
Transcription
Nucleotide Metabolism
NFKB Signaling
Integrins/Cell Adhesion
JAK/STAT/Interleukin Signaling
Immune Cell Activation
Cell Cycle
Telomere Maintenance
Actin Polymerization
Actomyosin
Epidermal Cell Differentiation
TLR Signaling
Apoptosis
E2F Transcription Network
Translation
Ribosome Biogenesis
Epithelial/Mesenchymal Transition
Splicing
MYC Targets
Developmental Program
(PC1-low)
Injury Response Program
(PC1-high)

Figure 2.7 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. 2.7D**. **(C)** Relative program score for individual cells (500 cells per sample; same cells as in **Fig 2.7B**) for top-correlated cell-type signatures. GSC cultures ordered as in **Fig 2.7D**. **(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 **Fig 2.8 D,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.


a
Developmental program (PC1-low)
GSCs (n = 65,655)
Injury Response program (PC1-high)
Median AUC score
Max
Min
PC1
PC2
25
0
−25
b
Developmental program
Subset of 14,000 cells across 28 samples
Injury Response program
Top PC1 genes
Bottom PC1 genes
CD44
ANXA2
TXN
S100A11
TAGLN2
TNFRSF12A
PTPRZ1
ASCL1
SOX2
ID4
OLIG1/2
STMN2
S100B
Relative expression
2
1
0
−1
−2
c
A2 reactive astrocytes
Cultured astroglia
Fetal OPCs
Fetal astrocytes
d
Relative AUC score
2.5
0
−2.5
Developmental
Injury Response
G946−K_L
BT48_L
BT89_L
BT147_L
BT94_L
G946−J_L
BT67_L
G620_L
BT84_L
G523_L
G885_L
BT73_L
G637_L
G895_L
BT127_L
G851_L
G799_L
G583_L
G549_L
G945−K_L
G876_L
G837_L
G564_L
G945−J_L
G797_L
G945−I_L
G566_L
G729_L
e
G946−K_L
1,624 cells
G523_L
3,745 cells
G549_L
3,434 cells
G729_L
2,881 cells
Injury Response program (AUC)
Developmental program (AUC)
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05

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. 2.7B and Fig 2.6B–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[13] and recently reported neural precursor (NPC), astrocyte (AC) and OPC-like cell states[15] (Fig. 2.7C and Fig 2.6B). 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[147] (Fig 2.8A,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. 2.7A).

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 (Fig 2.8C,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) (Fig 2.8E,F)

At the population level using bulk RNA-seq profiling, GSCs were categorized discretely as Developmental or Injury Response (Fig 2.8d). However, at the single-cell level, we observed a transcriptional gradient between the two states (Fig. 2.9). For each patient, GSCs occupied a

Figure 2.8 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. 2.7A**), 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.

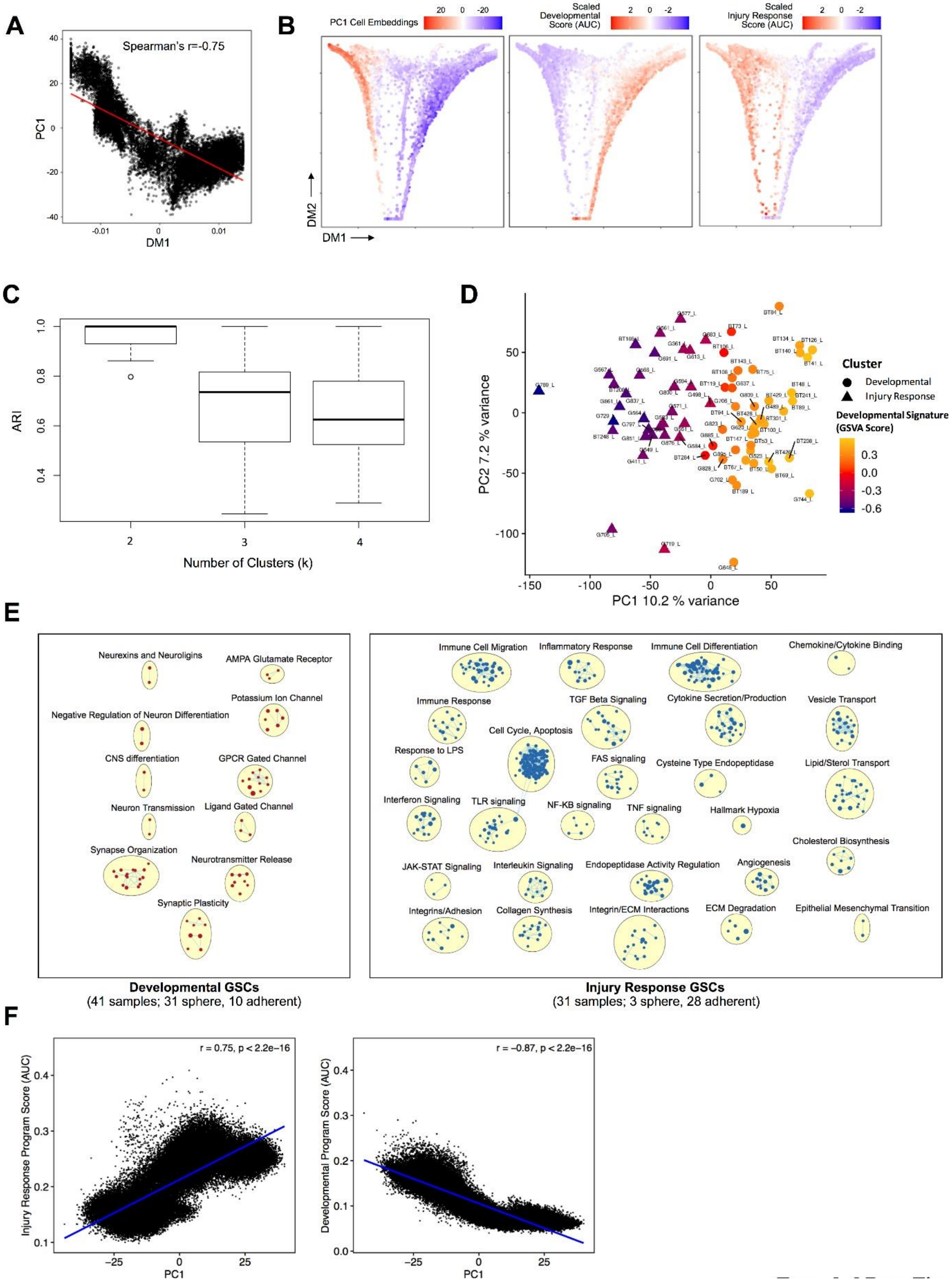
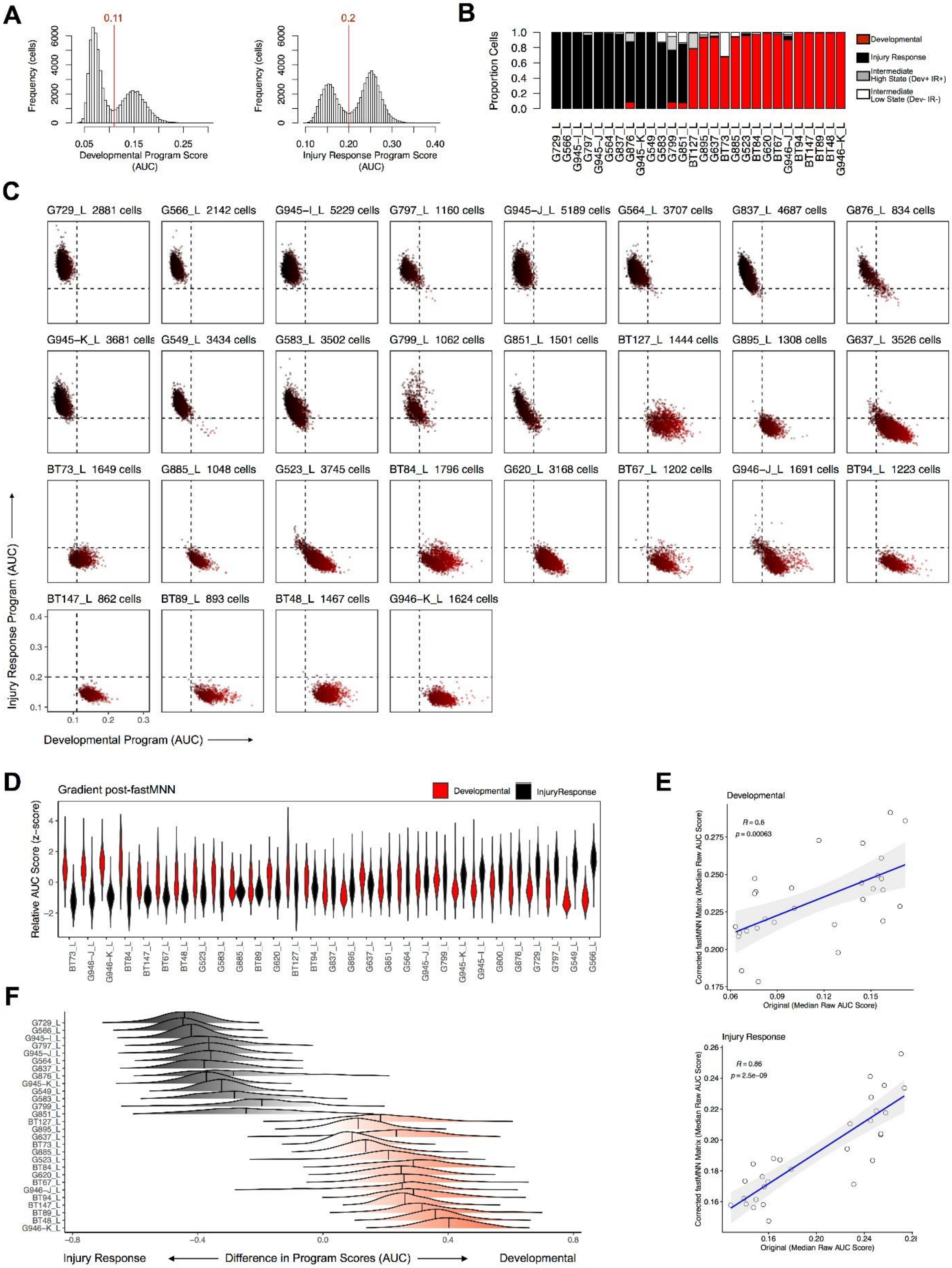

A
Spearman's r=-0.75
PC1
DM1
40
20
0
-20
-40
-0.01
0
0.01
B
PC1 Cell Embeddings
20 0 -20
Scaled Developmental Score (AUC)
2 0 -2
Scaled Injury Response Score (AUC)
2 0 -2
DM2
DM1
C
ARI
1.0
0.8
0.6
0.4
2
3
4
Number of Clusters (k)
D
PC2 7.2 % variance
PC1 10.2 % variance
50
0
-50
-100
-150
-100
-50
0
50
Cluster
● Developmental
▲ Injury Response
Developmental Signature (GSVA Score)
0.3
0.0
-0.3
-0.6
E
Neurexins and Neuroligins
AMPA Glutamate Receptor
Potassium Ion Channel
Negative Regulation of Neuron Differentiation
CNS differentiation
GPCR Gated Channel
Neuron Transmission
Ligand Gated Channel
Synapse Organization
Neurotransmitter Release
Synaptic Plasticity
Developmental GSCs
(41 samples; 31 sphere, 10 adherent)
Immune Cell Migration
Inflammatory Response
Immune Cell Differentiation
Chemokine/Cytokine Binding
Immune Response
TGF Beta Signaling
Cytokine Secretion/Production
Vesicle Transport
Response to LPS
Cell Cycle, Apoptosis
FAS signaling
Cysteine Type Endopeptidase
Lipid/Sterol Transport
Interferon Signaling
TLR signaling
NF-KB signaling
TNF signaling
Hallmark Hypoxia
Cholesterol Biosynthesis
JAK-STAT Signaling
Interleukin Signaling
Endopeptidase Activity Regulation
Angiogenesis
Integrins/Adhesion
Collagen Synthesis
Integrin/ECM Interactions
ECM Degradation
Epithelial Mesenchymal Transition
Injury Response GSCs
(31 samples; 3 sphere, 28 adherent)
F
r = 0.75, p < 2.2e−16
Injury Response Program Score (AUC)
0.4
0.3
0.2
0.1
−25
0
25
PC1
r = −0.87, p < 2.2e−16
Developmental Program Score (AUC)
0.4
0.3
0.2
0.1
0.0
−25
0
25
PC1

Figure 2.9 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 2.7D**. **(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.


A
0.11
0.2
Frequency (cells)
Developmental Program Score (AUC)
Injury Response Program Score (AUC)
B
Proportion Cells
Developmental
Injury Response
Intermediate High State (Dev+ IR+)
Intermediate Low State (Dev- IR-)
C
G729_L 2881 cells
G566_L 2142 cells
G945–I_L 5229 cells
G797_L 1160 cells
G945–J_L 5189 cells
G564_L 3707 cells
G837_L 4687 cells
G876_L 834 cells
G945–K_L 3681 cells
G549_L 3434 cells
G583_L 3502 cells
G799_L 1062 cells
G851_L 1501 cells
BT127_L 1444 cells
G895_L 1308 cells
G637_L 3526 cells
BT73_L 1649 cells
G885_L 1048 cells
G523_L 3745 cells
BT84_L 1796 cells
G620_L 3168 cells
BT67_L 1202 cells
G946–J_L 1691 cells
BT94_L 1223 cells
BT147_L 862 cells
BT89_L 893 cells
BT48_L 1467 cells
G946–K_L 1624 cells
Injury Response Program (AUC)
Developmental Program (AUC)
D
Gradient post-fastMNN
Developmental
InjuryResponse
Relative AUC Score (z-score)
E
Developmental
R = 0.6
p = 0.00063
Corrected fastMNN Matrix (Median Raw AUC Score)
Original (Median Raw AUC Score)
Injury Response
R = 0.86
p = 2.5e-09
F
Injury Response
Difference in Program Scores (AUC)
Developmental

discrete range within the Developmental and Injury Response spectrum. (Fig. 2.7D,E and Fig. 2.9C). 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[148] across samples (Fig. 2.9D,E). Furthermore, cells from multiple patients mapped to overlapping regions of the Injury Response–Developmental gradient, supporting common cellular phenotypes across patients (*Methods*; Fig. 2.9F). Thus, profiling GSCs from many samples is necessary to characterize the full spectrum of possible transcriptional states giving rise to bulk GBM.

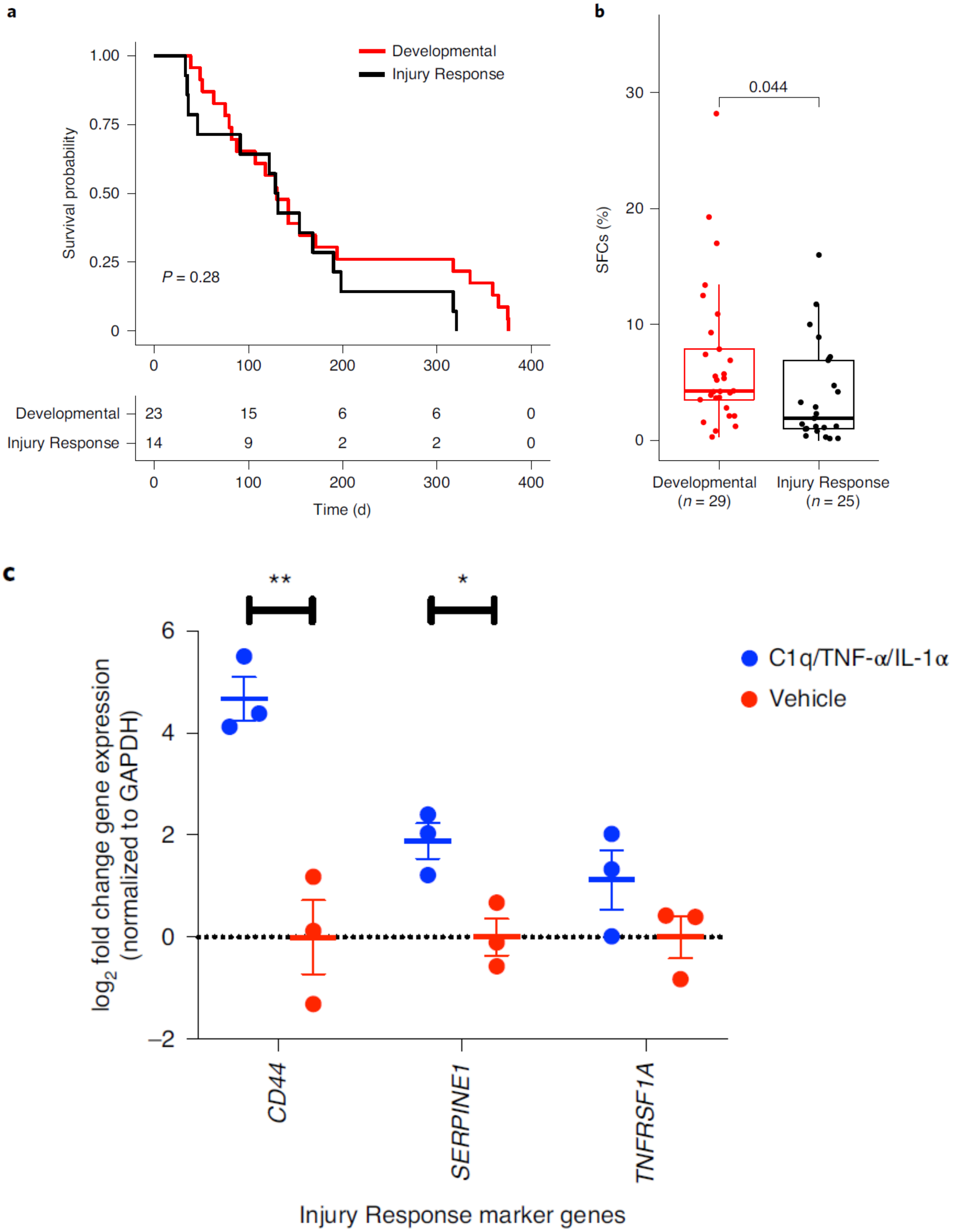
2.3.5 Developmental and Injury Response GSC states have functional differences and exhibit plasticity

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. 2.10B). 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[13] , 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. 2.10A). 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.

Figure 2.10 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.


a
1.00
0.75
0.50
0.25
0
Survival probability
Developmental
Injury Response
P = 0.28
0
100
200
300
400
Developmental 23 15 6 6 0
Injury Response 14 9 2 2 0
Time (d)
b
SFCs (%)
30
20
10
0
0.044
Developmental (n = 29)
Injury Response (n = 25)
c
**
*
6
4
2
0
−2
log2 fold change gene expression (normalized to GAPDH)
C1q/TNF-α/IL-1α
Vehicle
CD44
SERPINE1
TNFRSF1A
Injury Response marker genes

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. 2.10C). The cytokine cocktail induced expression of Injury Response genes after 48h, 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[142], may also induce the expression of Injury Response genes in Developmental subgroup GSCs.

2.3.6 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[149] (70,948 guides targeting 18,053 protein-coding genes) in 11 GSCs, a subset of which overlapped those profiled by bulk (n=9 of 11) and scRNA-seq (n=6 of 11). We used the BAGEL algorithm[103,150] 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. 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. 2.11A and Fig. 2.12). 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 essential genes. 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. 2.11B).

Figure 2.11 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. 2.7D**). 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 2.11B**) that are essential in each GSC. Samples are ordered by position on the transcriptional gradient (related to **Fig 2.7D**) and annotated with GSVA gene signature scores from matched bulk RNA-seq.

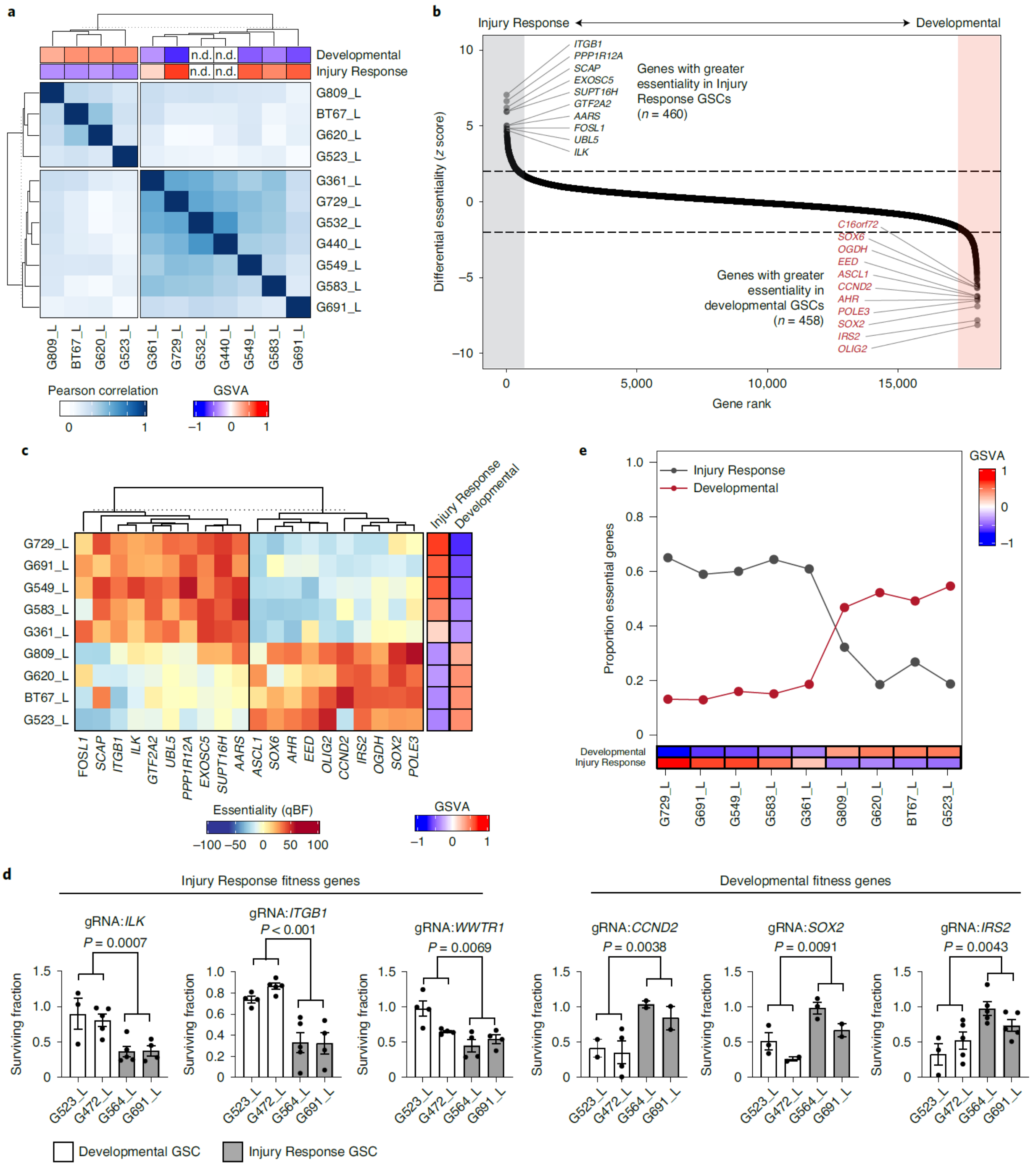
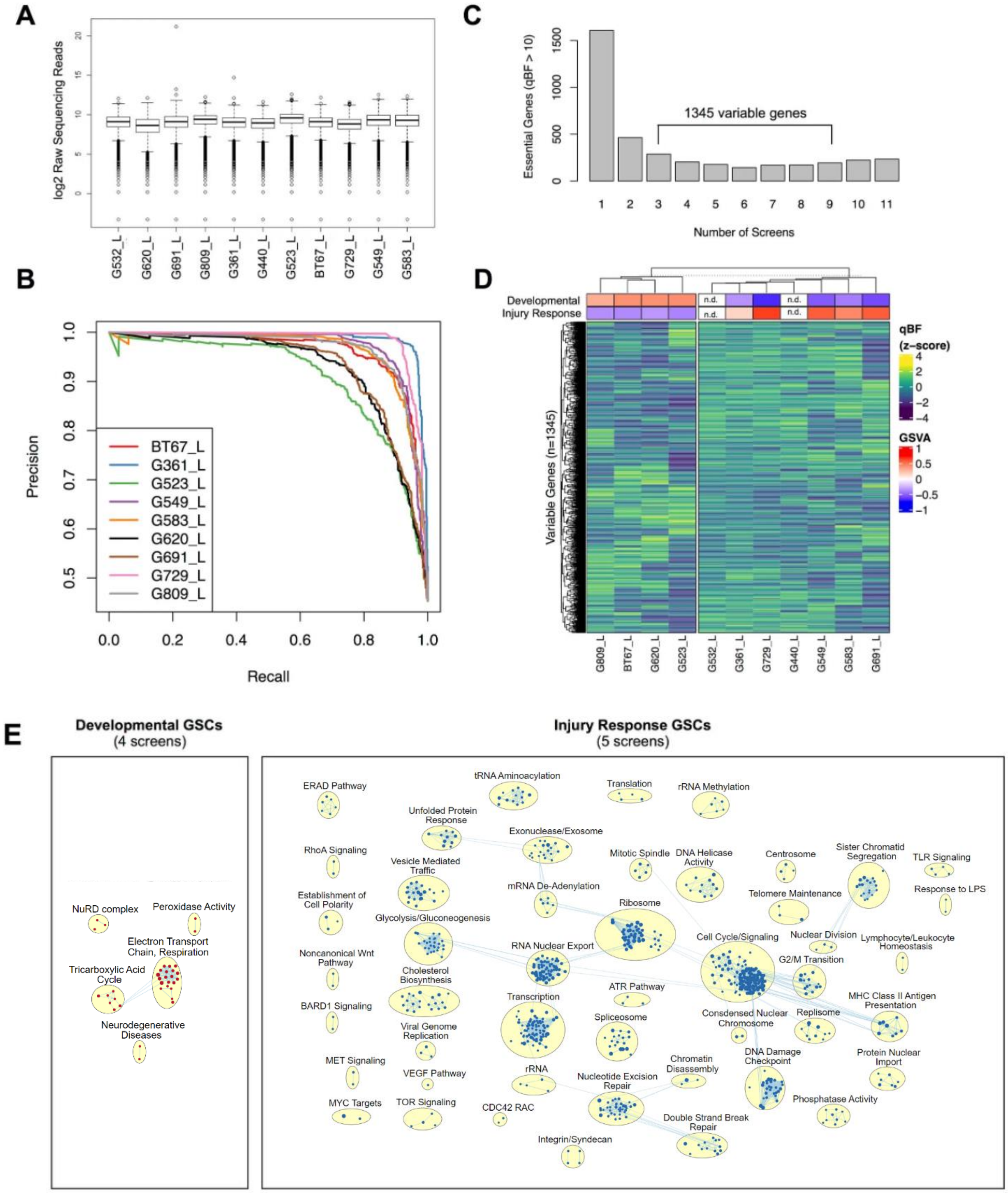

a
Developmental
Injury Response
n.d.
n.d.
n.d.
n.d.
G809_L
BT67_L
G620_L
G523_L
G361_L
G729_L
G532_L
G440_L
G549_L
G583_L
G691_L
Pearson correlation
0
1
GSVA
−1
0
1
b
Injury Response
Developmental
ITGB1
PPP1R12A
SCAP
EXOSC5
SUPT16H
GTF2A2
AARS
FOSL1
UBL5
ILK
Genes with greater essentiality in Injury Response GSCs (n = 460)
C16orf72
SOX6
OGDH
EED
ASCL1
CCND2
AHR
POLE3
SOX2
IRS2
OLIG2
Genes with greater essentiality in developmental GSCs (n = 458)
Differential essentiality (z score)
10
5
0
−5
−10
0
5,000
10,000
15,000
Gene rank
c
Injury Response
Developmental
G729_L
G691_L
G549_L
G583_L
G361_L
G809_L
G620_L
BT67_L
G523_L
FOSL1
SCAP
ITGB1
ILK
GTF2A2
UBL5
PPP1R12A
EXOSC5
SUPT16H
AARS
ASCL1
SOX6
AHR
EED
OLIG2
CCND2
IRS2
OGDH
SOX2
POLE3
Essentiality (qBF)
−100
−50
0
50
100
GSVA
−1
0
1
e
Injury Response
Developmental
GSVA
1
0
−1
Proportion essential genes
1.0
0.8
0.6
0.4
0.2
0
Developmental
Injury Response
G729_L
G691_L
G549_L
G583_L
G361_L
G809_L
G620_L
BT67_L
G523_L
d
Injury Response fitness genes
Developmental fitness genes
gRNA:ILK
P = 0.0007
gRNA:ITGB1
P < 0.001
gRNA:WWTR1
P = 0.0069
gRNA:CCND2
P = 0.0038
gRNA:SOX2
P = 0.0091
gRNA:IRS2
P = 0.0043
Surviving fraction
1.5
1.0
0.8
0.6
0.4
0.2
0.5
0
G523_L
G472_L
G564_L
G691_L
Developmental GSC
Injury Response GSC

Figure 2.12 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.


A
log2 Raw Sequencing Reads
0
5
10
15
20
G532_L
G620_L
G691_L
G809_L
G361_L
G440_L
G523_L
BT67_L
G729_L
G549_L
G583_L
B
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0
0.5
0.6
0.7
0.8
0.9
1.0
BT67_L
G361_L
G523_L
G549_L
G583_L
G620_L
G691_L
G729_L
G809_L
C
Essential Genes (qBF > 10)
0
500
1000
1500
1345 variable genes
1
2
3
4
5
6
7
8
9
10
11
Number of Screens
D
Developmental
Injury Response
n.d.
n.d.
n.d.
n.d.
qBF
(z-score)
4
2
0
−2
−4
GSVA
1
0.5
0
−0.5
−1
Variable Genes (n=1345)
G809_L
BT67_L
G620_L
G523_L
G532_L
G361_L
G729_L
G440_L
G549_L
G583_L
G691_L
E
Developmental GSCs
(4 screens)
Injury Response GSCs
(5 screens)
NuRD complex
Peroxidase Activity
Electron Transport Chain, Respiration
Tricarboxylic Acid Cycle
Neurodegenerative Diseases
ERAD Pathway
tRNA Aminoacylation
Translation
rRNA Methylation
Unfolded Protein Response
Exonuclease/Exosome
RhoA Signaling
Vesicle Mediated Traffic
Mitotic Spindle
DNA Helicase Activity
Centrosome
Sister Chromatid Segregation
TLR Signaling
Establishment of Cell Polarity
mRNA De-Adenylation
Telomere Maintenance
Response to LPS
Ribosome
Glycolysis/Gluconeogenesis
Nuclear Division
Lymphocyte/Leukocyte Homeostasis
Cell Cycle/Signaling
Noncanonical Wnt Pathway
RNA Nuclear Export
G2/M Transition
Cholesterol Biosynthesis
ATR Pathway
MHC Class II Antigen Presentation
BARD1 Signaling
Transcription
Consdensed Nuclear Chromosome
Replisome
Spliceosome
Viral Genome Replication
DNA Damage Checkpoint
Protein Nuclear Import
Chromatin Disassembly
MET Signaling
rRNA
VEGF Pathway
Nucleotide Excision Repair
Phosphatase Activity
MYC Targets
TOR Signaling
CDC42 RAC
Double Strand Break Repair
Integrin/Syndecan

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. 2.11c).

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. 2.11D). 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 (Fig. 2.12E). 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[151], 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 (Figs. 2.6D and 2.8E) 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. 2.11E). 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.

2.3.7 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. 2.13). 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. 2.13A). 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[73].

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. 2.13B). 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[73]. 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.

2.3.8 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. 2.14). We performed

Figure 2.13 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.

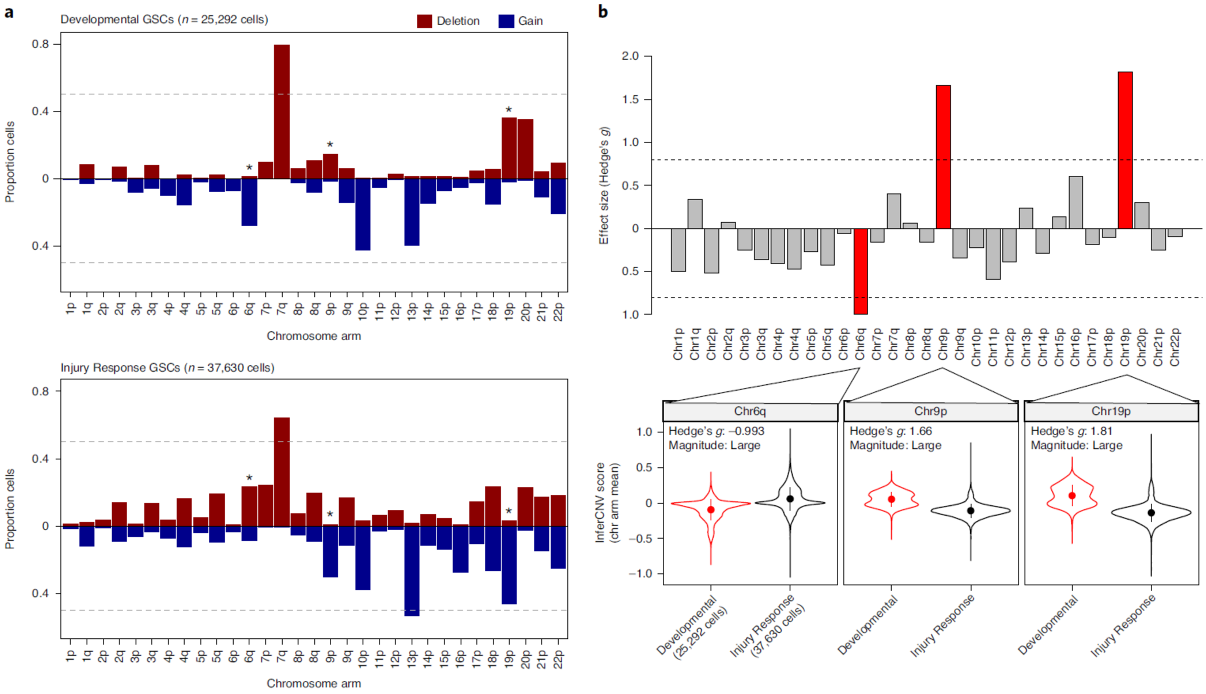
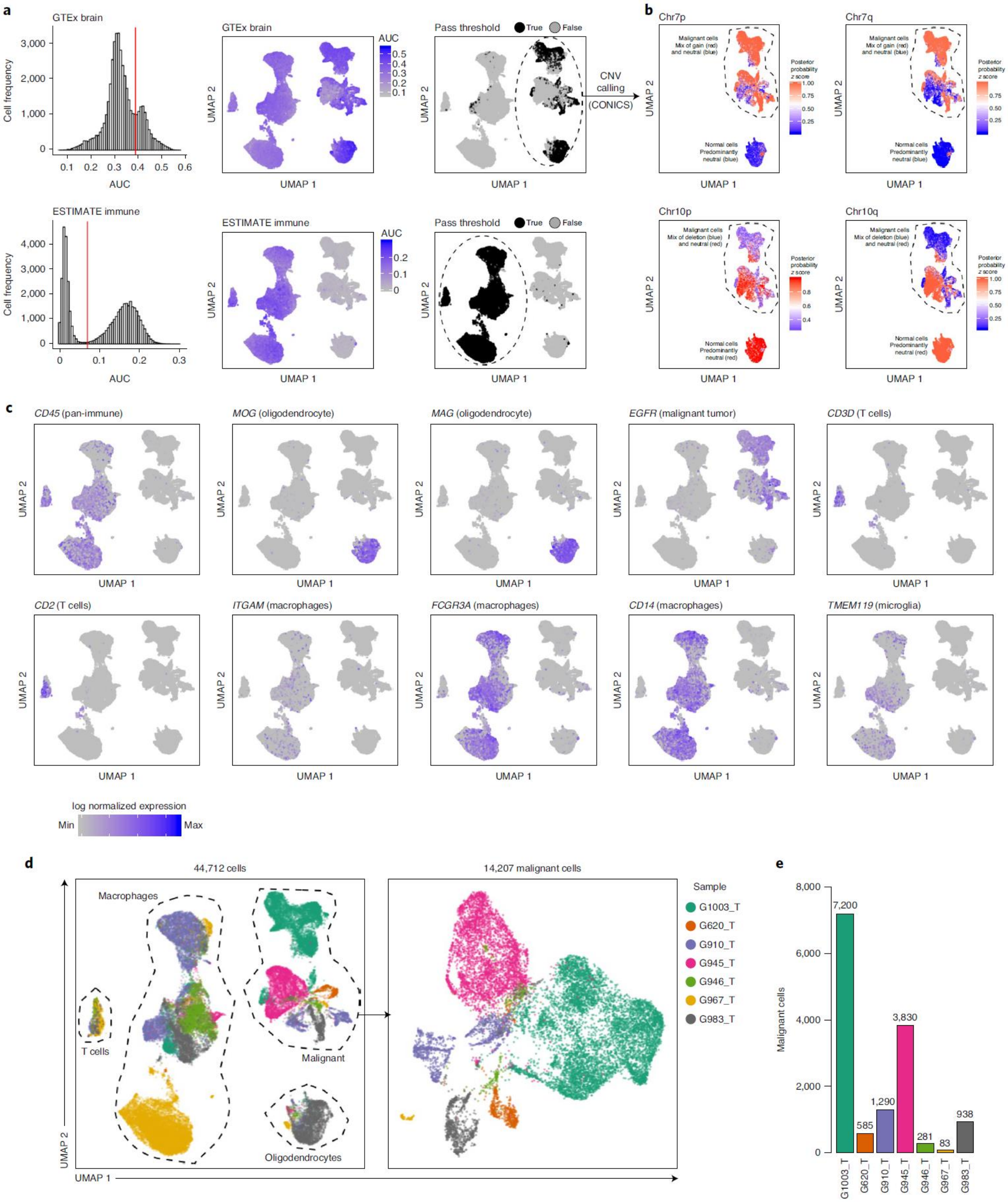

a
Developmental GSCs (n = 25,292 cells)
Deletion
Gain
Proportion cells
0.8
0.4
0
0.4
Chromosome arm
1p 1q 2p 2q 3p 3q 4p 4q 5p 5q 6p 6q 7p 7q 8p 8q 9p 9q 10p 11p 12p 13p 14p 15p 16p 17p 18p 19p 20p 21p 22p
*
Injury Response GSCs (n = 37,630 cells)
b
Effect size (Hedge's g)
2.0
1.5
1.0
0.5
0
0.5
1.0
Chr1p Chr1q Chr2p Chr2q Chr3p Chr3q Chr4p Chr4q Chr5p Chr5q Chr6p Chr6q Chr7p Chr7q Chr8p Chr8q Chr9p Chr9q Chr10p Chr11p Chr12p Chr13p Chr14p Chr15p Chr16p Chr17p Chr18p Chr19p Chr20p Chr21p Chr22p
Chr6q
Hedge's g: −0.993
Magnitude: Large
Chr9p
Hedge's g: 1.66
Magnitude: Large
Chr19p
Hedge's g: 1.81
Magnitude: Large
InferCNV score
(chr arm mean)
1.0
0.5
0
−0.5
−1.0
Developmental
(25,292 cells)
Injury Response
(37,630 cells)
Developmental
Injury Response
Developmental
Injury Response

Figure 2.14 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 **Fig 2.14A**). **(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 **Fig 2.14A**). **(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.


a
GTEx brain
Cell frequency
3,000
2,000
1,000
0
0.1 0.2 0.3 0.4 0.5 0.6
AUC
GTEx brain
AUC
0.5
0.4
0.3
0.2
0.1
UMAP 2
UMAP 1
Pass threshold
● True
● False
UMAP 2
UMAP 1
CNV calling (CONICS)
ESTIMATE immune
Cell frequency
4,000
3,000
2,000
1,000
0
0 0.1 0.2 0.3
AUC
ESTIMATE immune
AUC
0.2
0.1
0
UMAP 2
UMAP 1
Pass threshold
● True
● False
UMAP 2
UMAP 1
b
Chr7p
Malignant cells
Mix of gain (red)
and neutral (blue)
Normal cells
Predominantly
neutral (blue)
Posterior probability z score
1.00
0.75
0.50
0.25
UMAP 2
UMAP 1
Chr7q
Malignant cells
Mix of gain (red)
and neutral (blue)
Normal cells
Predominantly
neutral (blue)
Posterior probability z score
0.75
0.50
0.25
UMAP 2
UMAP 1
Chr10p
Malignant cells
Mix of deletion (blue)
and neutral (red)
Normal cells
Predominantly
neutral (red)
Posterior probability z score
0.8
0.6
0.4
UMAP 2
UMAP 1
Chr10q
Malignant cells
Mix of deletion (blue)
and neutral (red)
Normal cells
Predominantly
neutral (red)
Posterior probability z score
1.00
0.75
0.50
0.25
UMAP 2
UMAP 1
c
CD45 (pan-immune)
MOG (oligodendrocyte)
MAG (oligodendrocyte)
EGFR (malignant tumor)
CD3D (T cells)
CD2 (T cells)
ITGAM (macrophages)
FCGR3A (macrophages)
CD14 (macrophages)
TMEM119 (microglia)
UMAP 2
UMAP 1
log normalized expression
Min
Max
d
44,712 cells
Macrophages
T cells
Malignant
Oligodendrocytes
14,207 malignant cells
UMAP 2
UMAP 1
Sample
● G1003_T
● G620_T
● G910_T
● G945_T
● G946_T
● G967_T
● G983_T
e
Malignant cells
8,000
6,000
4,000
2,000
0
7,200
585
1,290
3,830
281
83
938
G1003_T
G620_T
G910_T
G945_T
G946_T
G967_T
G983_T

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. 2.15A). 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. 2.15A). 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. 2.15B). 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. 2.15A). Tumor cells most distant from GSCs, at the end of the differentiation trajectory, resemble mature nonproliferative astrocytes[141,152], expressing canonical markers such as GFAP, AQP4 and APOE (Fig. 2.15C,D and Fig. 2.16A). 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[153]. 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. 2.15C,D).

We further interpreted our two gradients in the context of previously described cell types in adult and pediatric GBM[15]. We projected GSCs and tumor cells onto a cellular state map consisting of NPC, OPC, astrocyte-like and mesenchymal-like quadrants (Fig. 2.16B). 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 (Fig. 2.16C). Together these findings demonstrate that, despite culture conditions and lack of microenvironment, GSCs

Figure 2.15 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 2.15A** (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.

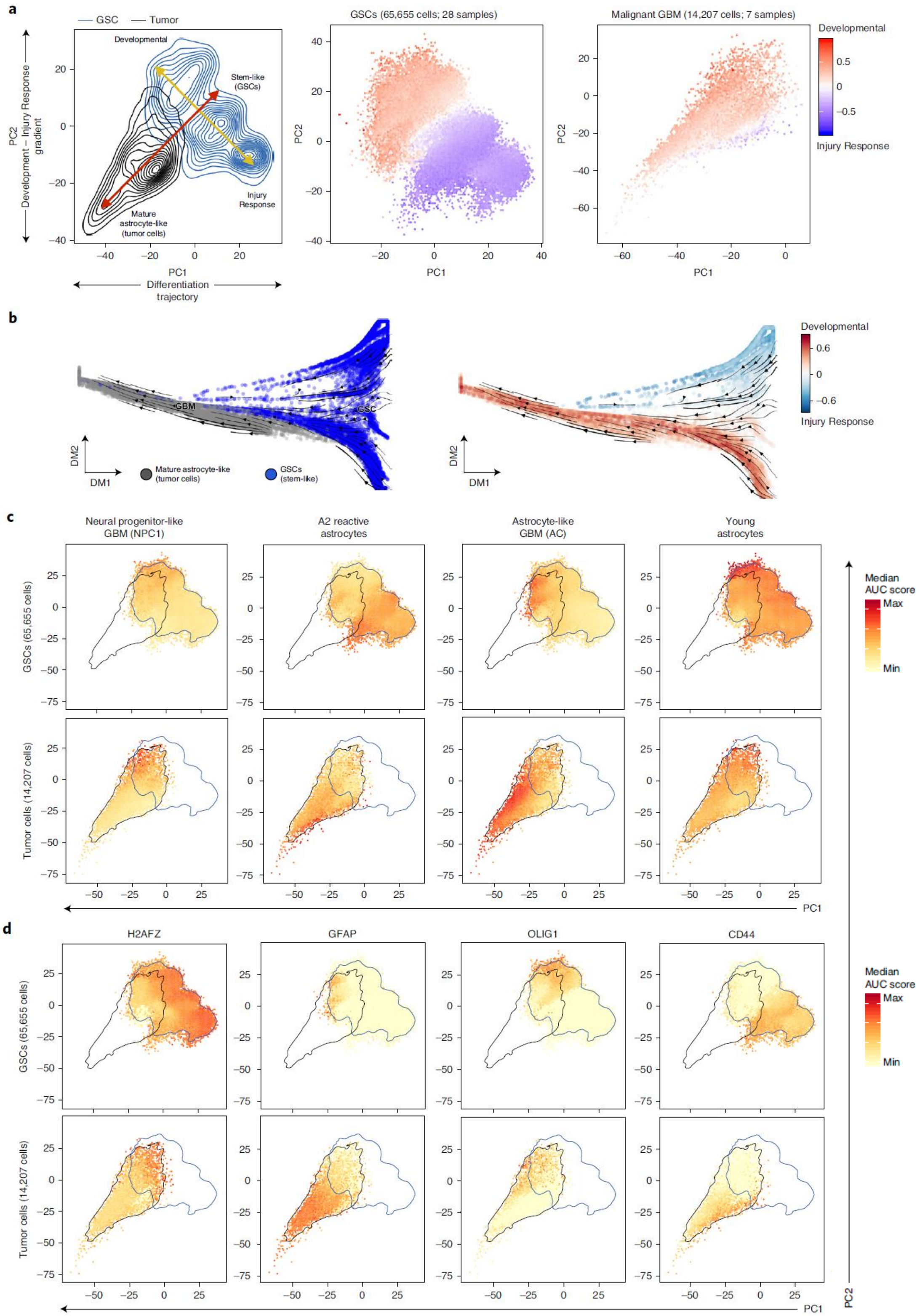
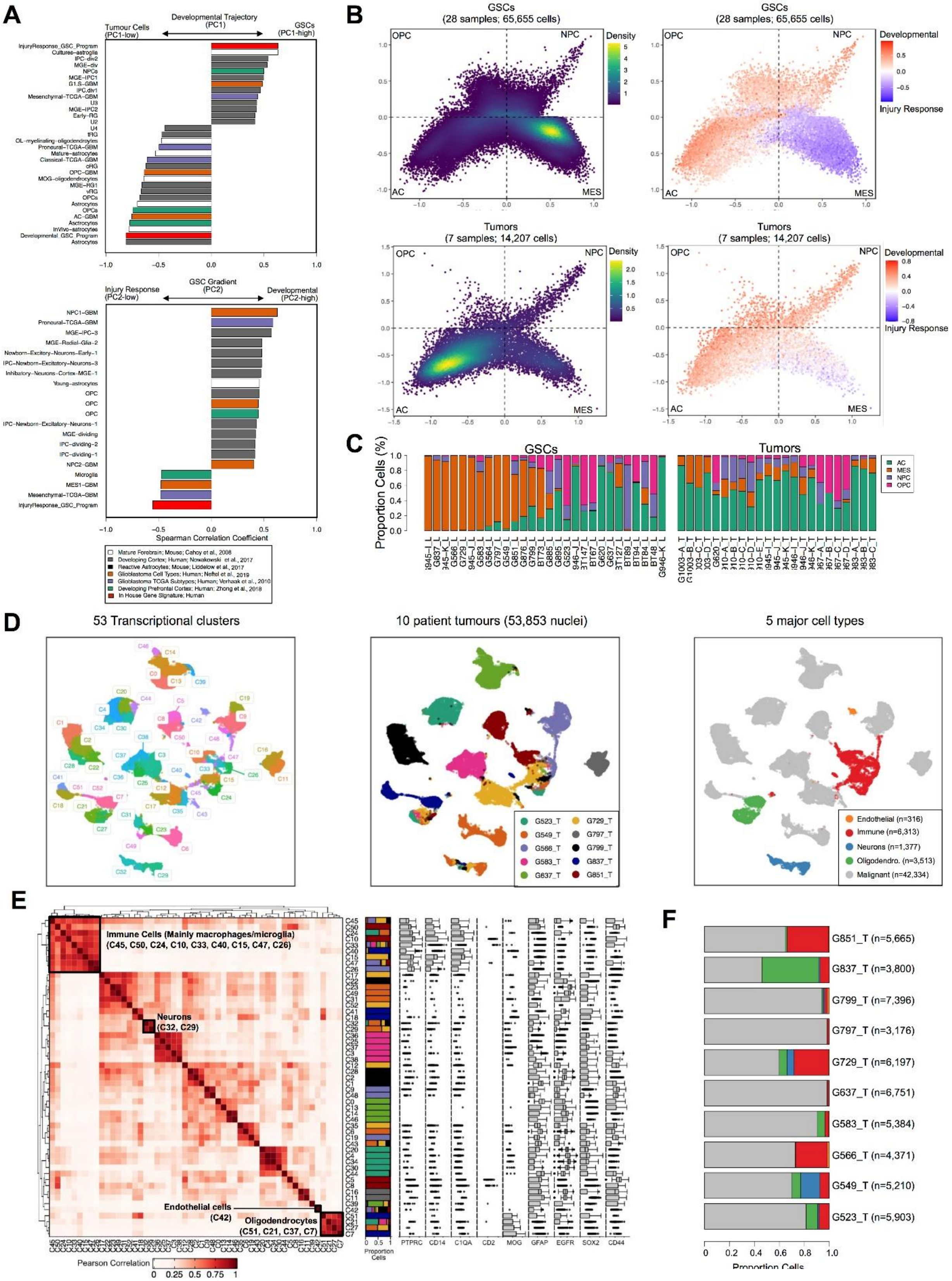

a
GSC
Tumor
Developmental
Stem-like (GSCs)
Injury Response
Mature astrocyte-like (tumor cells)
PC2
Development – Injury Response gradient
PC1
Differentiation trajectory
GSCs (65,655 cells; 28 samples)
Malignant GBM (14,207 cells; 7 samples)
Developmental
0.5
0
−0.5
Injury Response
b
GBM
GSC
DM2
DM1
Mature astrocyte-like (tumor cells)
GSCs (stem-like)
Developmental
0.6
0
−0.6
Injury Response
c
Neural progenitor-like GBM (NPC1)
A2 reactive astrocytes
Astrocyte-like GBM (AC)
Young astrocytes
GSCs (65,655 cells)
Tumor cells (14,207 cells)
Median AUC score
Max
Min
PC1
d
H2AFZ
GFAP
OLIG1
CD44
GSCs (65,655 cells)
Tumor cells (14,207 cells)
Median AUC score
Max
Min
PC1
PC2

Figure 2.16 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 **Fig 2.16D**). Numbers in brackets represent the total number of nuclei per tumor.


A
Developmental Trajectory (PC1)
Tumour Cells (PC1-low)
GSCs (PC1-high)
InjuryResponse_GSC_Program
Cultures-astroglia
IPC-div2
MGE-div
NPCs
MGE-IPC1
G1.S-GBM
IPC.div1
Mesenchymal-TCGA-GBM
U3
MGE-IPC2
Early-RG
U2
U4
tRG
OL-myelinating-oligodendroytes
Proneural-TCGA-GBM
Mature-astrocytes
Classical-TCGA-GBM
oRG
OPC-GBM
MOG-oligodendrocytes
MGE-RG1
vRG
OPCs
Astrocytes
OPCs
AC-GBM
Asctrocytes
InVivo-astrocytes
Developmental_GSC_Program
Astrocytes
−1.0
−0.5
0.0
0.5
1.0
GSC Gradient (PC2)
Injury Response (PC2-low)
Developmental (PC2-high)
NPC1-GBM
Proneural-TCGA-GBM
MGE-IPC-3
MGE-Radial-Glia-2
Newborn-Excitory-Neurons-Early-1
IPC-Newborn-Excitatory-Neurons-3
Inhibatory-Neurons-Cortex-MGE-1
Young-astrocytes
OPC
OPC
OPC
IPC-Newborn-Excitatory-Neurons-1
MGE-dividing
IPC-dividing-2
IPC-dividing-1
NPC2-GBM
Microglia
MES1-GBM
Mesenchymal-TCGA-GBM
InjuryResponse_GSC_Program
Spearman Correlation Coefficient
□ Mature Forebrain; Mouse; Cahoy et al., 2008
■ Developing Cortex; Human; Nowakowski et al., 2017
■ Reactive Astrocytes; Mouse; Liddelow et al., 2017
■ Glioblastoma Cell Types; Human; Neftel et al., 2019
■ Glioblastoma TCGA Subtypes; Human; Verhaak et al., 2010
■ Developing Prefrontal Cortex; Human; Zhong et al., 2018
■ In House Gene Signature; Human
B
GSCs (28 samples; 65,655 cells)
OPC
NPC
AC
MES
Density
Developmental
Injury Response
Tumors (7 samples; 14,207 cells)
C
Proportion Cells (%)
GSCs
Tumors
AC
MES
NPC
OPC
D
53 Transcriptional clusters
10 patient tumours (53,853 nuclei)
G523_T
G549_T
G566_T
G583_T
G637_T
G729_T
G797_T
G799_T
G837_T
G851_T
5 major cell types
Endothelial (n=316)
Immune (n=6,313)
Neurons (n=1,377)
Oligodendro. (n=3,513)
Malignant (n=42,334)
E
Immune Cells (Mainly macrophages/microglia)
(C45, C50, C24, C10, C33, C40, C15, C47, C26)
Neurons
(C32, C29)
Endothelial cells
(C42)
Oligodendrocytes
(C51, C21, C37, C7)
Pearson Correlation
0
0.25
0.5
0.75
1
Proportion Cells
PTPRC
CD14
C1QA
CD2
MOG
GFAP
EGFR
SOX2
CD44
F
G851_T (n=5,665)
G837_T (n=3,800)
G799_T (n=7,396)
G797_T (n=3,176)
G729_T (n=6,197)
G637_T (n=6,751)
G583_T (n=5,384)
G566_T (n=4,371)
G549_T (n=5,210)
G523_T (n=5,903)
0
0.2
0.4
0.6
0.8
1.0
Proportion Cells

mirror cell types found in primary tumors and represent a major transcriptional axis underpinning GBMs.

2.3.9 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. 2.17A). Tumor cells resembled the Developmental state more often, however Injury Response-like tumor cells were visible in every tumor (Fig. 2.17B–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. 2.17A and Fig. 2.16D–F) and analyzed four public GBM sc/snRNA-seq datasets[15,92,116,117] (52 tumors; 49,018 malignant cells or nuclei) (*Methods*). Across all datasets, Developmental and Injury Response programs were anti-correlated (mean Pearson's r=−0.70; Fig. 2.17E), 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. 2.17A,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[111]. 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 (Fig. 2.18A). 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 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 (Fig. 2.18B). These results further support the potential for CNVs to influence GSC and subsequent

Figure 2.17 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. 2.15A**). 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.

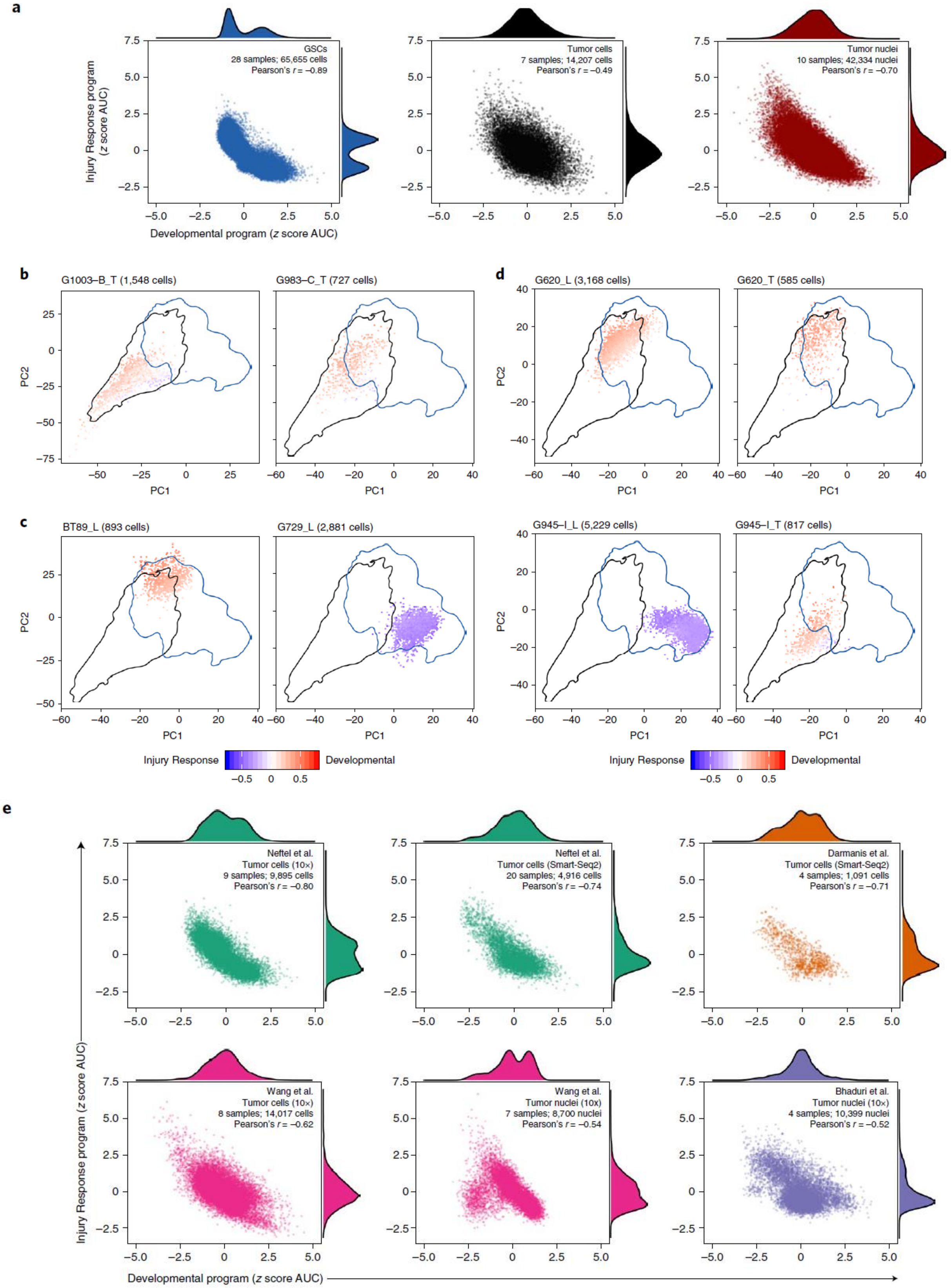
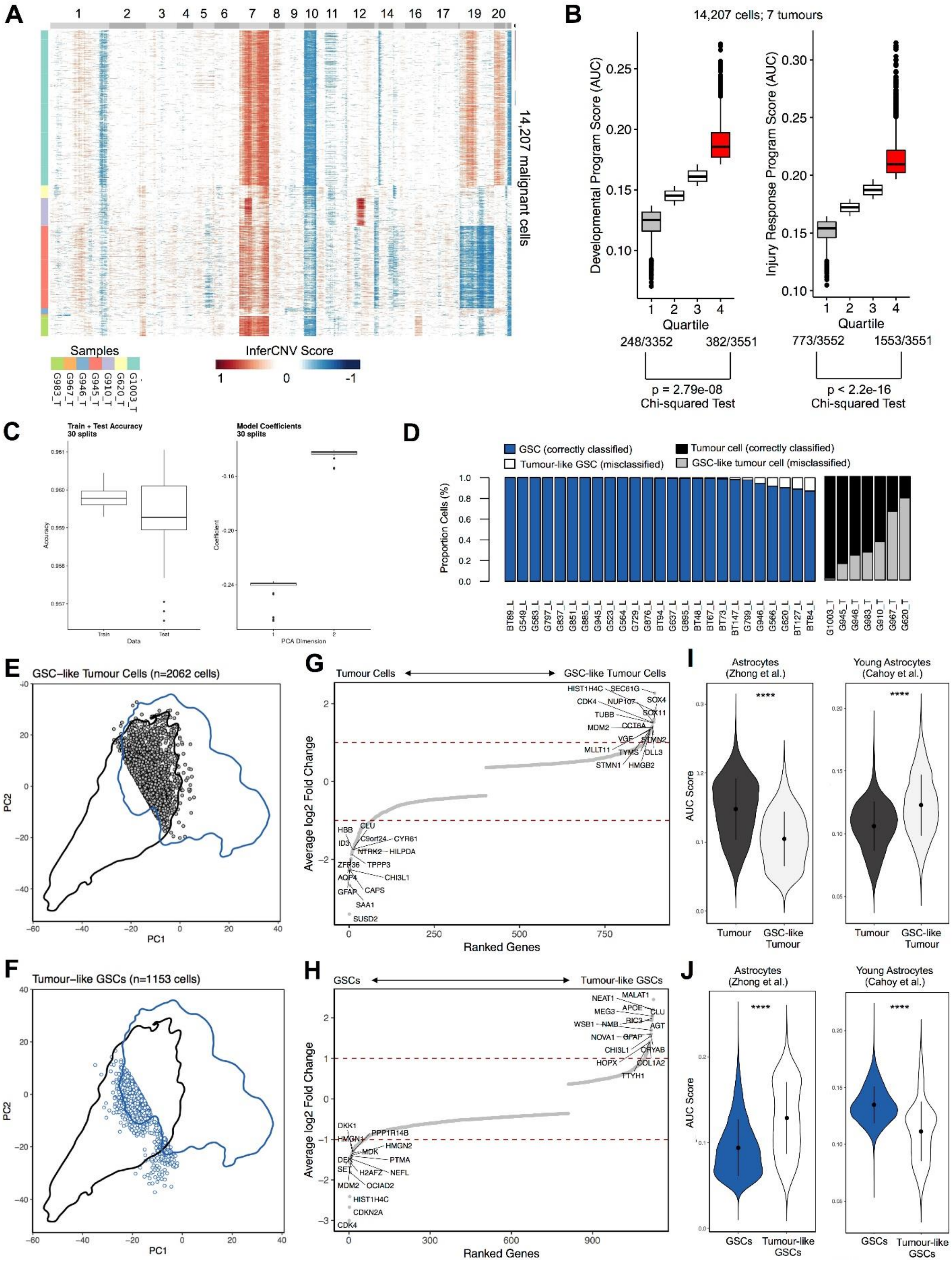

a
GSCs
28 samples; 65,655 cells
Pearson's r = −0.89
Tumor cells
7 samples; 14,207 cells
Pearson's r = −0.49
Tumor nuclei
10 samples; 42,334 nuclei
Pearson's r = −0.70
Injury Response program (z score AUC)
Developmental program (z score AUC)
7.5
5.0
2.5
0
−2.5
−5.0
−2.5
0
2.5
5.0
b
G1003–B_T (1,548 cells)
G983–C_T (727 cells)
c
BT89_L (893 cells)
G729_L (2,881 cells)
d
G620_L (3,168 cells)
G620_T (585 cells)
G945–I_L (5,229 cells)
G945–I_T (817 cells)
PC1
PC2
Injury Response
Developmental
−0.5
0
0.5
e
Neftel et al.
Tumor cells (10×)
9 samples; 9,895 cells
Pearson's r = −0.80
Neftel et al.
Tumor cells (Smart-Seq2)
20 samples; 4,916 cells
Pearson's r = −0.74
Darmanis et al.
Tumor cells (Smart-Seq2)
4 samples; 1,091 cells
Pearson's r = −0.71
Wang et al.
Tumor cells (10×)
8 samples; 14,017 cells
Pearson's r = −0.62
Wang et al.
Tumor nuclei (10x)
7 samples; 8,700 nuclei
Pearson's r = −0.54
Bhaduri et al.
Tumor nuclei (10×)
4 samples; 10,399 nuclei
Pearson's r = −0.52
Injury Response program (z score AUC)
Developmental program (z score AUC)

Figure 2.18 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. 2.15A**. 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).


A
1 2 3 4 5 6 7 8 9 10 11 12 14 16 17 19 20
14,207 malignant cells
Samples
G983_T G967_T G946_T G945_T G910_T G620_T G1003_T
InferCNV Score
1 0 -1
B
14,207 cells; 7 tumours
Developmental Program Score (AUC)
0.10 0.15 0.20 0.25
Quartile
1 2 3 4
248/3352 382/3551
p = 2.79e-08
Chi-squared Test
Injury Response Program Score (AUC)
0.10 0.15 0.20 0.25 0.30
Quartile
1 2 3 4
773/3552 1553/3551
p < 2.2e-16
Chi-squared Test
C
Train + Test Accuracy
30 splits
Accuracy
0.957 0.958 0.959 0.960 0.961
Train Test
Data
Model Coefficients
30 splits
Coefficient
-0.24 -0.20 -0.16
1 2
PCA Dimension
D
GSC (correctly classified)
Tumour-like GSC (misclassified)
Tumour cell (correctly classified)
GSC-like tumour cell (misclassified)
Proportion Cells (%)
0.0 0.2 0.4 0.6 0.8 1.0
BT89_L G549_L G583_L G797_L G837_L G851_L G885_L G945_L G523_L G564_L G729_L G876_L BT94_L G637_L G895_L BT48_L BT67_L BT73_L BT147_L G799_L G946_L G566_L G620_L BT127_L BT84_L
G1003_T G945_T G946_T G983_T G910_T G967_T G620_T
E
GSC–like Tumour Cells (n=2062 cells)
PC2
-40 -20 0 20 40
PC1
-60 -40 -20 0 20 40
F
Tumour–like GSCs (n=1153 cells)
PC2
-40 -20 0 20 40
PC1
-60 -40 -20 0 20 40
G
Tumour Cells
GSC-like Tumour Cells
Average log2 Fold Change
-2 0 2
Ranked Genes
0 250 500 750
HIST1H4C SEC61G CDK4 NUP107 SOX4 TUBB SOX11 MDM2 CCT6A VGF STMN2 MLLT11 TYMS DLL3 STMN1 HMGB2
HBB CLU ID3 C9orf24 CYR61 NTRK2 HILPDA ZFP36 TPPP3 AQP4 CHI3L1 CAPS GFAP SAA1 SUSD2
H
GSCs
Tumour-like GSCs
Average log2 Fold Change
-3 -2 -1 0 1 2
Ranked Genes
0 300 600 900
NEAT1 MALAT1 MEG3 APOE CLU WSB1 NMB RIC3 AGT NOVA1 GFAP CHI3L1 CRYAB HOPX COL1A2 TTYH1
DKK1 HMGN1 PPP1R14B HMGN2 MDK DEK PTMA SET H2AFZ NEFL MDM2 OCIAD2 HIST1H4C CDKN2A CDK4
I
Astrocytes
(Zhong et al.)
****
AUC Score
0.0 0.1 0.2 0.3
Tumour
GSC-like Tumour
Young Astrocytes
(Cahoy et al.)
****
0.05 0.10 0.15 0.20
Tumour
GSC-like Tumour
J
Astrocytes
(Zhong et al.)
****
AUC Score
0.0 0.1 0.2
GSCs
Tumour-like GSCs
Young Astrocytes
(Cahoy et al.)
****
0.05 0.10 0.15 0.20
GSCs
Tumour-like GSCs

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 (Fig. 2.18C; *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%) (Fig. 2.18D). Notably, the tumor with the highest proportion of GSC-like cells was the only IDH1 mutant (p.R100Q, G620_T) in the cohort (Fig. 2.17D). IDH1 mutations promote convergence toward a proneural phenotype[154], 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 (Fig. 2.18E–J). These data demonstrate that substantial overlap exists between GSCs cultured from patient tumors and GSCs found directly within surgical GBM samples.

2.4 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[14,15,92,116,117]. 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[13,90,100,155] or based on similarity to neural subtypes found in normal or fetal brain development[117]. 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[156,157]. 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[142] 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[73,126] 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[15,117] 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.

2.5 Methods

2.5.1 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).

Adherent GSC cultures (denoted “G###_L”) were cultured as previously described[85] . In brief, cells were grown adherently on culture plates coated with poly-L-ornithine and laminin. Serumfree NS cell self-renewal media (NS media) consisted of Neurocult NS-A Basal media, supplemented with 2 mmol/L L-glutamine, N2 and B27 supplements, 75 μg/mL bovine serum albumin, 10 ng/mL recombinant human EGF (rhEGF), 10 ng/mL basic fibroblast growth factor (bFGF), and 2 μg/mL heparin. All assays were completed with cultures between passages P8-12.

Non-adherent GSC lines were cultured as free-floating spheres (denoted “BT###_L”) , in serum-free media as previously described[121] . Briefly, GSC lines were maintained in Serum-Free Media (SFM) supplemented with EGF and bFGF (20ng/ml each, Peprotech) and heparan sulfate (2µg/ml, Sigma) until non-adherent spheres formed, typically after 7-21 days. Upon reaching 150-200 μm, spheres were dissociated via mechanical trituration or with AccumaxTM (Innovative Cell Technologies, Inc.) and re-plated as single cell suspensions in T25 flasks for routine maintenance.

2.5.2 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 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×(log n2−log n1), where t=time and n=number of cells.

2.5.3 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: 1mm anterior of bregma, 2mm to the right of the midline and 3mm deep. Mice were housed in groups of three to five and maintained on a 12h 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.

2.5.4 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 100µl 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% CO2 for 14 or 21d when all wells were scored for the presence or absence of spheres. The SFC was calculated using Extreme Limiting Dilution Analysis software[158]. Three biological replicates from each GSC culture were plated.

2.5.5 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 24h in NS medium, fresh medium containing vehicle or cytokines was added, with final concentrations as follows: TNF-α (30ngµl −1), C1q (400ngµl −1) and IL-1α (3ngµl −1). Cell pellets were collected after 48h 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).

2.5.6 Single-cell and single-nuclei RNA-seq.

2.5.6.1 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 snapfrozen tumors. Tissues were minced on dry ice and dissolved in lysis buffer (0.32M sucrose, 5mM $CaCl_2$, 3mM $Mg(Ac)_2$, 20mM Tris-HCl (pH 7.5), 0.1% Triton-X-100, 0.1mM 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%.

2.5.6.2 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).

2.5.6.3 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[159]. 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[160] (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[161,162] (v.2.3.4) as described below and defined detected genes as those with >0 UMIs.

2.5.6.4 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.

2.5.6.5 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.

2.5.6.6 Adjusting for cell-cycle signal.

Heterogeneity from cell cycle stage, particularly among cells grown in vitro, can contribute substantial transcriptomic variation and mask biological signal. However, removing all signal associated with cell cycle can blur the distinction between cell types where proliferation is a biological trait (e.g. mitotic and post-mitotic neural progenitors[163]).

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). First, we assigned cell-cycle scores to individual cells on the basis of expression of previously published G2/M and S-phase gene signatures[133], 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. Cell cycle phases for each cell are stored, so we can recall the mitotic phase for each cell post-regression.

2.5.6.7 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[164] 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.

2.5.6.8 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[165] (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. Clusters were visualized using t-SNE and UMAP.

We validated the ability of our clustering pipeline to accurately detect intra-GSC subpopulations by benchmarking four clustering algorithms: original Louvain algorithm[166] (implemented in Seurat), Louvain algorithm with multilevel refinement (implemented in Seurat), k-means and spectral clustering[167] (Fig 2.1B). For spectral clustering, we clustered cells across a range of possible centres (k=2-8) and picked the optimal number of clusters based on the highest average intra-cluster silhouette width. For k-means clustering (k=2-8), we selected the optimal number of clusters using majority rule across 30 comparative indices calculated in the NbClust R package[168] (v3.0). Both Louvain algorithms agreed completely with our original clustering across all samples. Spectral and k-means clustering agreed with our original clustering results in 67% and

78% of samples, respectively, when the top 2 solutions (k) were considered. Together, this demonstrates our ability to identify reproducible clusters, with limited technical bias.

2.5.6.9 RNA velocity

CellRanger BAM files were sorted with samtools[169] (v1.10). Per sample loom files were generated from sorted BAM files using Velocyto[71] (v0.17.13) and the GRCh38 annotation.gtf (10x Genomics). Loom files were merged with loompy (v3.0.6; https://github.com/linnarssonlab/loompy) and subset to contain cells used in Diffusion Map. For scvelo[72] (v0.2.2) analysis, genes in the merged loom file were filtered for a minimum count of 20 in both spliced and unspliced count matrices and with the filter_genes_dispersion function (n_top_genes=2000). Count matrices were normalized by counts per cell, and log transformed log(x + 1). The functions scvelo.tl.velocity and scvelo.tl.velocity_graph were run using default parameters. RNA velocity plots on Diffusion Map coordinates were generated with the scvelo.pl.velocity_embedding_stream function.

2.5.6.10 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[170]. To determine if a given gene signature is on or off in a cell, AUCell examines the distribution of AUC scores across cells and nominates a score threshold based on the best fit of multiple distributions to the data (i.e. bimodal, normal). When directly comparing the difference of AUCell scores between two gene signatures, such as in **Fig. 2.15A**, AUCell scores were normalized between [0,1] by subtracting the minimum and dividing by the range.

We curated a collection of gene signatures from cluster and cell type markers in published single-cell and bulk RNA-sequencing of developing human cortex[139,140], glioblastoma[13,15], mature forebrain[141] and developmental astrocyte states[141,142]. We additionally used ‘Injury Response’ (4399 genes) and ‘Developmental’ (3968 genes) signatures derived from differential expression between the Injury Response and Developmental clusters identified with bulk RNA-seq of GSC samples (for further information see *Bulk RNA-sequencing* section of this chapter).

Marker genes and principal component top/bottom-loading gene lists were annotated using over-representation analysis in clusterProfiler[171] (v.3.10.1). For this analysis, all expressed genes in the dataset defined the “universe” of background genes. Over-representation analysis was performed with the ‘enricher’ function from the clusterProfiler R package using a hypergeometric test, with a q-value cutoff of 0.01 after multiple comparison adjustment with the Benjamini and Hochberg (BH) procedure. Hallmark (H), curated (C2), gene ontology (GO, C5), oncogenic (C6) and immunologic (C7) gene sets from MSigDB18 (v7.0) were used for annotation. The enrichment map was generated using genes ranked by the negative of PC1 loadings as described for *Bulk RNA Sequencing* (**Fig. 2.8D**).

2.5.6.11 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. 2.7B**. To validate the accuracy of our single-cell CNV calls, we compared inferCNV scores and WGS CNV log2 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[34] (v.2.0.23) on matched WGS data (**Fig. 2.4A-C**).

2.5.6.12 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 **Fig. 2.4D**. 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 (**Fig. 2.4E**).

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≥0.8; **Fig. 2.13B**).

2.5.6.13 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. The "brain origin" gene signature was derived from the union of the top 50 expressed genes for each brain subregion in GTEx[172], resulting in 129 genes specific to normal brain tissue but agnostic of anatomical region. We used the ESTIMATE[173] Immune score gene list as the "immune origin" signature.

We then identified malignant tumor cells within the brain fraction using single-cell CNV inference. We used groups of co-clustered cells that appeared brain-like but not immune-like as input for CONICS[174] (v0.0.0.1), a tool to identify CNVs in single cell RNA-sequencing data without the use of an a priori reference cell dataset by fitting a two component Gaussian Mixture Model to the average gene expression across all genes on each chromosome arm. The posterior probabilities for each cell belonging to the component with the higher mean can be used to visualize copy number states across cells and discern malignant brain tumor cells from normal brain cells. Using this information and prior knowledge of GBM CNVs[4,13,73], we conclude that chromosome 7p/q has two copy number states, cells colored red have high probability of harbouring amplifications (higher mean gene expression), compared to low probability cells (blue) that are likely diploid at that locus. Opposite patterns are observed with chromosome 10. Combining CNV state prediction from chromosome 7 and 10, we were able to define a subset of cells that co-localize on UMAP and harbour CNVs at canonical GBM loci, as well as a copy-quiet subset of normal brain cells (**Fig. 2.14B**).

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. 2.14**).

2.5.6.14 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.

2.5.6.15 Projection onto GBM cell-state map.

The two-dimensional cell-state representation map was created as described by Neftel et al.[15] (**Fig. 2.16B**). 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 $\log2(SC_{OPC}-SC_{NPC}+1)$ and for astrocyte/mesenchymal cells ($D<0$), the x axis was computed as $\log2(SC_{AC}-SC_{MES}+1)$.

2.5.6.16 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. 2.15A**) were used as inputs to be mapped to labels.

On the training set, a logistic regression classifier was trained using the best hyperparameter determined from 5 fold cross validation. The following functions/classes from sklearn (v0.21.2) were called in R using the reticulate package (v1.13): “linear_model.LogisticRegressionCV”, “metrics.accuracy_score” and “model_selection.train_test_split”. The range of hyperparameters (argument Cs for LogisticRegressionCV) considered for model training was from 1e-4 to 1e4. L2 regularization was used. To assess model stability, we repeated the 80/20 train test split 30 times with different random (stratified) splits and examined the distributions of test accuracy and model coefficients (**Fig. 2.18C**). We found that the model was robust to sampling effects. We picked the best performing model (highest test accuracy) and used it to predict class labels for the entire dataset.

2.5.7 Bulk RNA-seq

2.5.7.1 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 500ng 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.

2.5.7.2 Data pre-processing and clustering

Strand-specific 75-bp paired-end reads were aligned to hg38 reference using STAR[43] (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. [45]) (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. Following VST transformation of counts, batch correction was performed using the ComBat[175] function in the sva package[176] (v3.30.1) to incorporate, sequencing cohort, culture condition (adherent vs. neurosphere), sex, age, and primary/recurrent status into the model to avoid removing these factors' contributions to variance in the data

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 cluster numbers (k=2-4), we used kernlab[177] (v0.9-27) to perform 200 runs of spectral clustering on the full dataset and do the same on randomly subsampled datasets containing 80% of samples. We then calculated the Adjusted Rand Index (ARI) between clustering solutions computed on the full and subsampled datasets. For each value k (number of clusters) we determined which solution obtained on the full dataset had the highest median ARI with respect to the solutions from subsampled data, and called it the optimal solution for the given k. The optimal solution for k = 2 had the highest median ARI with respect to subsampled solutions out of all values k, so we used this solution for further analyses of bulk RNA-seq data (**Fig. 2.8C**).

2.5.7.3 Differential gene expression and pathway analysis

Differential gene expression analysis was carried out on count data using DESeq2 (ref. [45]) 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 (**Fig. 2.6 C-D**). GSVA[178] (v.1.30.0) was used to assess the activity of gene signatures across samples. Gene Set Enrichment Analysis (GSEA)[179] (v3.0) was performed on differentially expressed genes ranked by sign(LFC)*(-log10(FDR)) (LFC = log fold change, FDR = Benjamini Hochberg False Discovery Rate). Pathways used for GSEA were obtained from Human_GO_AllPathways_no_GO_iea_April_01_2018_symbol.gmt (http://download.baderlab.org/EM_Genesets/) and filtered with a minimum size of 15 and a maximum size of 200 genes. Enriched pathways were filtered at FDR<0.10. Enriched pathways were grouped into functional themes by AutoAnnotate[180] (v1.3.2) and visualized using Enrichment Map[181] (v3.1) in Cytoscape[182] (v3.7.2) (**Fig. 2.6D**)

2.5.8 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 500ng 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[183] (v.0.7.15).

We used Genome Analysis Toolkit (GATK)[184,185] (v3.5) best practices to pre-process BAM files using the markduplicate command, IndelRealignment and BQSR functions. Cellularity, ploidy and allele-specific copy number profiles were estimated from GSC-blood pairs using VarScan2[31] (v2.3.8) calls as input for Sequenza[33] (v2.1.2). Log2 copy number ratios between -0.35 and 0.3 were set to assign genome losses and gains, respectively, using GISTIC[34] (v2.0.23) with the maximum number of segments (-maxseg) set to 24000.

2.5.9 Genome-wide CRISPR–Cas9 screens

We performed CRISPR–Cas9 screens using the 70-k TKOv3 library[149] (Addgene, 90294) using previously established protocols[103] with cells cultured as described above. A minimum of 8×10^7 cells were transduced with gRNA library-expressing lentivirus in the presence of 0.8μg polybrene at a multiplicity of infection of 0.3. At 24h after transduction, lentiviral medium was removed and cells were cultured with 2 μgml^{-1} puromycin for 48–72h 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 the subsection *gDNA extraction, library preparation and sequencing.*

2.5.9.1 gDNA extraction, library preparation and sequencing

Genomic DNA was extracted from T0 and screen cell pellets using the QiaAMP DNA Blood Maxi Kit (QIAGEN) according to the manufacturer’s protocol. Following ethanol precipitation to concentrate gDNA samples, we amplified 50 μg of gDNA from each sample in 20 individual PCR reactions using Kapa HiFi Master Mix (Kapa Biosystems) and 1 μM of TKOv3 library primers (forward: 5’-GAGGGCCTATTTCCCATGATTC-3’, reverse: 5’-GTTGCGAAAAAGAACGTTCACGG-3’) for 19 amplification cycles. Individual sets of PCR reactions were pooled and 5 μL was used as a template for a second PCR reaction to attach TrueSeq adaptor sequences and unique i5/i7 barcode combinations for each sample in Kapa HiFi Master Mix with 1μM primer concentration for 17 amplification cycles. Barcoded PCR product was gel purified and sequenced on an Illumina NextSeq500 instrument with 4 x 10^7 reads for T0 samples and 1.5 x 10^7 reads for screen samples.

2.5.9.2 Analysis of genome-wide screen data

DNA sequencing reads for each CRISPR screen were mapped to TKOv3 library gRNAs using MAGECK (v0.5.8)[186]. The BAGEL[150] (v0.91) pipeline was used to normalize gRNA reads for sample sequencing depth, calculate fold-change for each gRNA from the T0 baseline and compute a Bayes Factor (BF) for each gene representing a confidence score that gene knockout produces a fitness defect[103,149]. As a quality control step, we assessed gRNA complexity in the T0 population. For all 11 CRISPR screens, greater than 98% of the 71,090 gRNAs in the library were detected at least 30 sequencing reads, with a similar distribution across all samples (**Fig. 2.12A**). Quantile normalization was performed for comparison of BF (qBF) scores across screens. Variable essential genes were defined as those with a qBF > 10 in 3-9 of the 11 screens (**Fig. 2.12B**)

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. 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 for bulk RNA-seq.

2.5.9.3 Competitive proliferation assays

For validation of gene knockouts producing fitness defects, Cas9-expressing GSCs were first engineered via lentiviral transduction as previously described[103]. 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 24h after transduction, cells were selected with 2 $\mu g ml^{-1}$ puromycin for 48–72h. 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 14d 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.

2.5.10 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 (**Fig. 2.6A**). 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).

2.5.11 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[187] (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. Previously published scRNA-seq data that were re-analyzed in this study are available from the following sources: Wang et al.[92] (GSE138794), Bhaduri et al.[117] (http://cells.ucsc.edu/?ds=gbm), Neftel et al.[15] (https://singlecell.broadinstitute.org/single_cell/study/SCP393/) and Darmanis et al.[116] (http://gbmseq.org/). Source data are provided with this paper.

3 Chapter 3: Glioma Stem Cell Heterogeneity Explained by Transcription-Coupled and Transcription Orthogonal Axes

Owen K.N. Whitley, Florence M.G. Cavalli, Paul Guilhamon, Laura M. Richards, Graham MacLeod, Fiona J. Coutinho, Michelle Kushida, Julia E. Jaramillo, Claire Che, Naghmeh Rastegar, Lilian Lee, Moloud Ahmadi, Jasmine K. Bhatti, Danielle A. Bozek, Christopher Arlidge, Yussanne Ma, Richard A. Moore, Marco A. Marra, Julian Spears, Michael D. Cusimano, Sunit Das, Mark Bernstein, Stephane Angers, Trevor J. Pugh, Mathieu Lupien, H. Artee Luchman, Gary D. Bader, Samuel Weiss, Peter B. Dirks

Authors' contributions

O.K.N.W., F.M.G.C., P.B.D., M.L., S.W., G.B., H.A.L., T.J.P., S.A., S.D., M.C., Y.M., R.M., M.B. conceived of and designed this study. O.K.N.W., performed data analyses, supervised by G.D.B and both conceived of the analysis methods. F.J.C., F.M.G.C., L.M.R. and P.G. generated and pre-processed bulk RNA-seq, miRNA-seq, or WGS data. P.G. and C.A. performed DNA Methylation experiments. P.G. performed ATAC-seq experiments for this project and generated pre-processed DNA methylation and ATAC-seq data for the study. 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. 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. O.K.N.W., G.D.B., M.L., P.B.D., and T.J.P. wrote the manuscript with feedback from all authors.

3.1 Abstract:

Glioma stem cells (GSCs), thought to be the root of glioblastoma and responsible for relapse after the failure of standard therapies, exhibit considerable heterogeneity at the molecular level. This molecular variation is associated with GSC functional properties and variability in response to therapeutics. Here, we explore six different genomic data measurements (RNA, miRNA, Single Nucleotide Variation, Copy Number Variation, ATAC-seq, DNA methylation) of a panel of patient-derived GSCs to gain a holistic and mechanistic model of GSC biology. We identify four major axes of variation in GSCs at the multi-omic level. The major biological signal is an axis of variation among genetic and non-genetic features, including RNA, miRNA, chromatin accessibility, DNA methylation, and DNA copy number corresponding to a developmental to injury response transcriptional gradient correlated with CRISPR-based gene essentiality measurements. The SNV data captures an additional axis of variation corresponding to a hypermutation phenotype associated with shorter tumor model survival times. Our results provide new insight into GSC biology that will be useful to identify novel therapeutic targets for glioblastoma.

3.2 Introduction:

Glioblastoma has a terrible prognosis, with a median survival time of 12-15 months[1] and a survival rate of under 10%[4]. Strong evidence, accumulated over the past two decades, points to the existence of stem-like glioma stem cells (GSCs), capable of repopulating a tumor after surgery and radio/chemotherapy[17–19]. Thus, any effective treatment of this lethal disease likely requires targeting the GSC population. Efforts to characterize GSCs have revealed variability in drug response[20,101] and functional outputs[188,189] correlated with genetic and non-genetic phenotypes, complicating this potential therapeutic strategy. Recently, we showed that transcriptional heterogeneity in GSCs can largely be decomposed into a single axis of variation, namely the Developmental/Injury Response axis, and that this axis delineates differential functional dependencies in GSCs[188]. Here, we further investigate this functional axis using a multi -omics approach.

To expand our picture of molecular heterogeneity in GSCs, we collected two genetic (copy number variation (CNV) and single nucleotide variants (SNVs) from whole genome sequencing (WGS)) and four functional (RNA-seq, DNA methylation, ATAC-seq, and miRNA-seq) -omics data types for 54 patient-derived GSC lines (Figure 3.1). Performing an integrated analysis of variation common among data types and specific to individual data types, we discovered four major axes in the multi-omics data which explained at least 10% of variation in at least one datatype. In addition to SNV, DNA methylation, and chromatin accessibility-specific axes, we report a multi-omic axis found in all data layers except SNVs. This multi -omic axis represents the recently discovered developmental and injury response GSC phenotype[188] that correlates with GSC sensitivity to gene deletions, suggesting it is a useful mechanism to target therapeutically. Our work highlights the concerted effects of epigenetic suppression and mutually opposed miRNA regulation programs on transcription within this axis, identifying new GSC mechanisms for targeting.

3.3 Results

3.3.1 Integrated principal component analysis of 6 -omics data types reveals heterogeneity associated with and orthogonal to the Developmental to Injury Response transcriptional gradient

To identify major sources of orthogonal signal across our six -omics data types from 54 GSC lines (Figure 3.1, Figure 3.2), we applied principal component analysis (PCA) on the most variable features for each data type. Immediately, we noticed that principal component 1 (PC1) was accounted for virtually entirely by SNV signal, while PCs 2+3 explained a large portion of variance in all other -omics data types (Figure 3.3A). PC1 is associated with a known hypermutation phenotype[80,81] (Figure 3.3B), while PCs 2 + 3 are capable of separating GSCs into Developmental and Injury Response clusters previously identified by Richards and colleagues[188] (Figure 3.3C). PC4 and PC5 explain a substantial portion of variation in the ATAC-seq data (20%) and DNAm data (12%), respectively. PC4 is associated with average ATAC-seq signal per sample (Figure 3.3D), while PC5 is associated with average promoter DNA methylation (Figure 3.3E). Together, these results have identified four major axes of variation in our data: a multi-omics axis representing the Developmental and Injury Response, a known hypermutator

Figure 3.1 Assembly of multi -omics glioma stem cell dataset

Schematic of sample collection and data generation. Glioma stem cell (GSC) cultures were created as described in *Methods*, and were aliquoted for runs through five -omics experiments, producing six data types: DNA methylation, ATAC-seq, miRNA-seq, RNA-seq, whole genome sequence-derived copy number variation (CNVs) and single nucleotide variation (SNV). SNV signal was aggregated at the level of biological pathways to reduce noise before analysis.

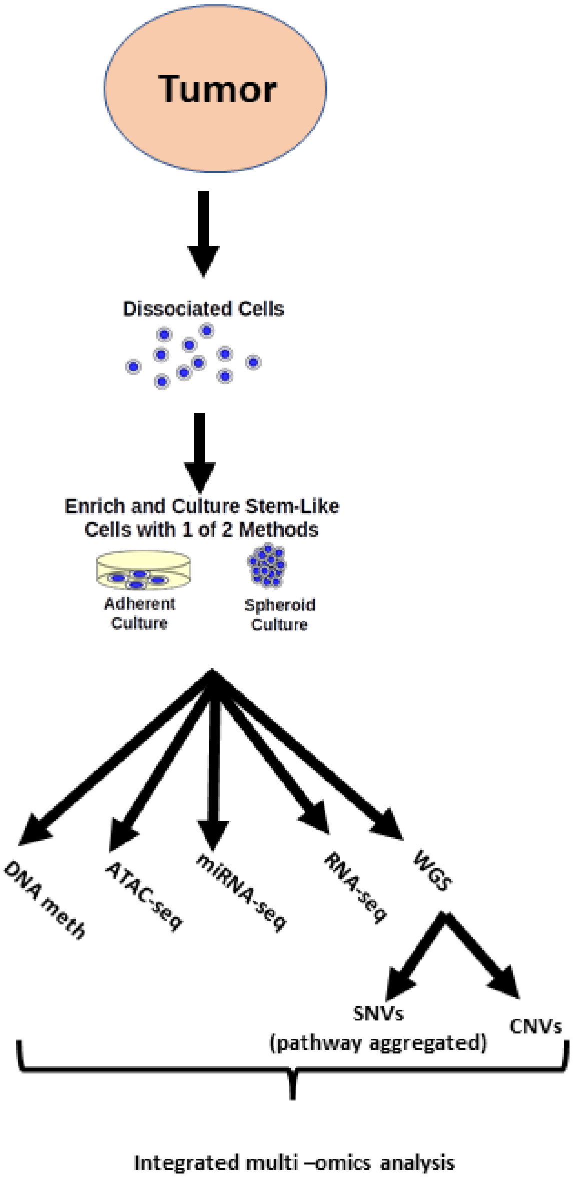

Tumor
Dissociated Cells
Enrich and Culture Stem-Like Cells with 1 of 2 Methods
Adherent Culture
Spheroid Culture
DNA meth
ATAC-seq
miRNA-seq
RNA-seq
WGS
SNVs (pathway aggregated)
CNVs
Integrated multi –omics analysis

Figure 3.2 Data availability and imputation strategy

Table of data availability for the 54 GSC samples analyzed in this study. Samples with missing data had data imputed via k-nearest neighbors (KNN) before PCA and further downstream analyses. For miRNA, differential miRNA expression and correlation analyses were performed using only samples for which miRNA data existed, without imputation. For ATAC-seq, differential chromatin accessibility was assessed using only samples for which ATAC-seq data existed, without imputation.

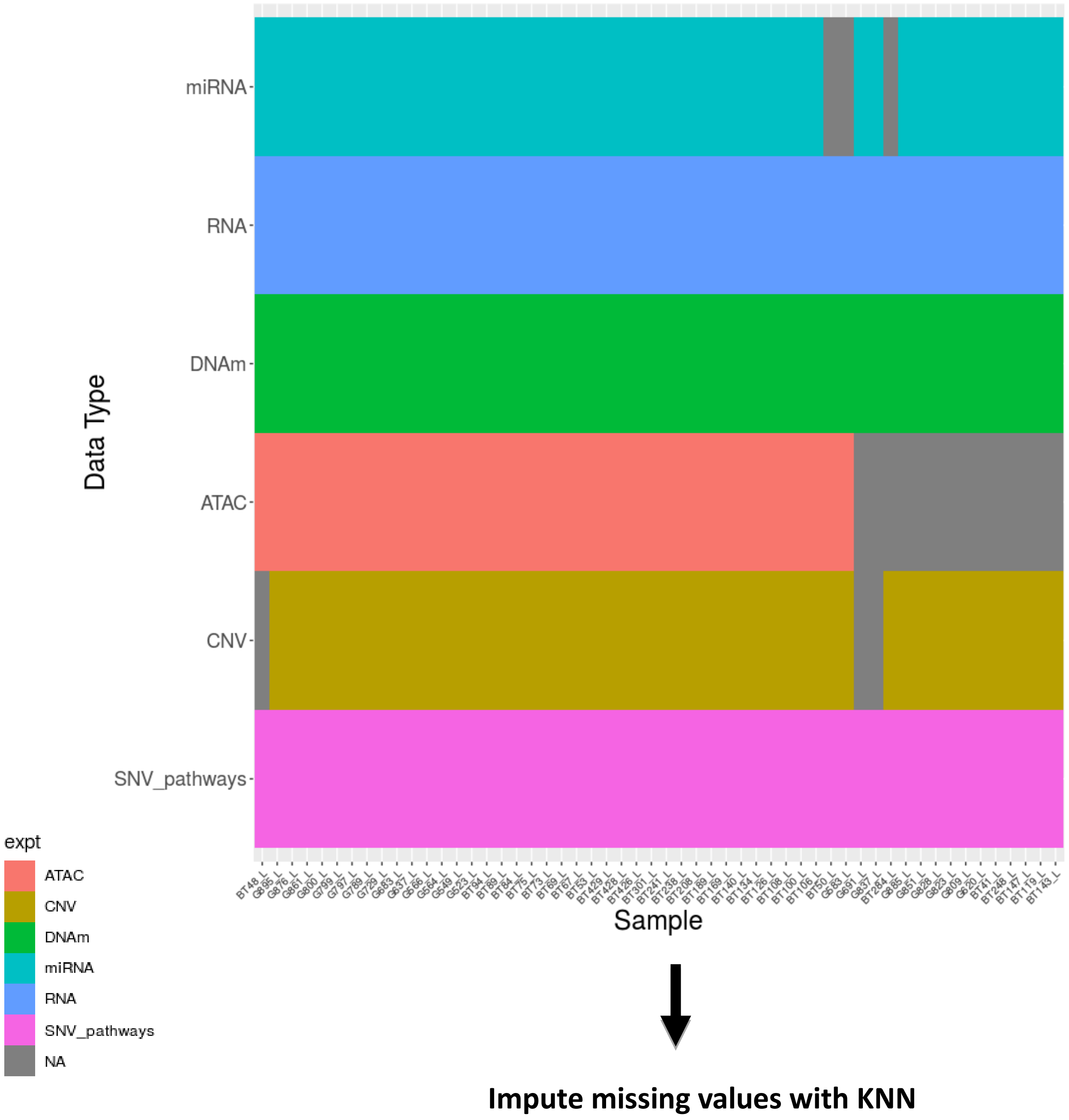
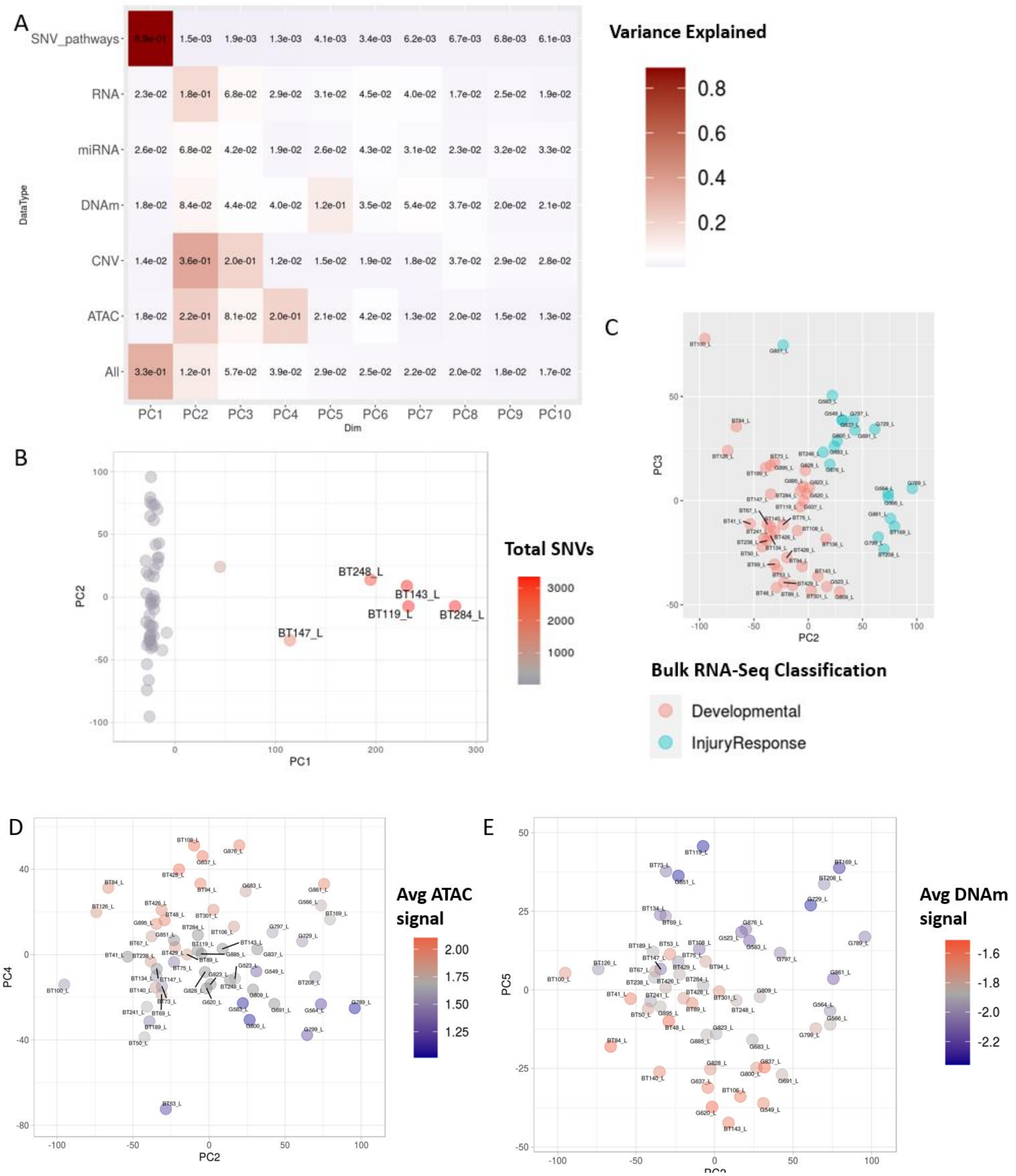


Figure 3.3 PCA on multi-omics data

(A) Variance explained (coefficient of determination/r-squared) per datatype (or for all datatypes) per principal component. **(B)** Total SNV signal (total number of called somatic mutations) per sample, overlaid on PCA plot. Hypermutator samples are labeled. **(C)** Developmental and Injury Response cluster labels overlaid on PCA plot for PCs 2 and 3. **(D)** Averaged ATAC-seq signal across all ATAC-seq features. ATAC-seq signal for individual features calculated as log(average region signal + 1). **(E)** Averaged DNAm signal across all DNAm gene features. Feature signal calculated as average m-value for a promoter at a given gene.


A
Variance Explained
0.8
0.6
0.4
0.2
DataType
SNV_pathways 8.9e-01 1.5e-03 1.9e-03 1.3e-03 4.1e-03 3.4e-03 6.2e-03 6.7e-03 6.8e-03 6.1e-03
RNA 2.3e-02 1.8e-01 6.8e-02 2.9e-02 3.1e-02 4.5e-02 4.0e-02 1.7e-02 2.5e-02 1.9e-02
miRNA 2.6e-02 6.8e-02 4.2e-02 1.9e-02 2.6e-02 4.3e-02 3.1e-02 2.3e-02 3.2e-02 3.3e-02
DNAm 1.8e-02 8.4e-02 4.4e-02 4.0e-02 1.2e-01 3.5e-02 5.4e-02 3.7e-02 2.0e-02 2.1e-02
CNV 1.4e-02 3.6e-01 2.0e-01 1.2e-02 1.5e-02 1.9e-02 1.8e-02 3.7e-02 2.9e-02 2.8e-02
ATAC 1.8e-02 2.2e-01 8.1e-02 2.0e-01 2.1e-02 4.2e-02 1.3e-02 2.0e-02 1.5e-02 1.3e-02
All 3.3e-01 1.2e-01 5.7e-02 3.9e-02 2.9e-02 2.5e-02 2.2e-02 2.0e-02 1.8e-02 1.7e-02
PC1 PC2 PC3 PC4 PC5 PC6 PC7 PC8 PC9 PC10
Dim
B
Total SNVs
3000
2000
1000
BT248_L
BT143_L
BT119_L
BT284_L
BT147_L
PC2
PC1
C
PC3
PC2
Bulk RNA-Seq Classification
● Developmental
● InjuryResponse
D
Avg ATAC signal
2.00
1.75
1.50
1.25
PC4
PC2
E
Avg DNAm signal
-1.6
-1.8
-2.0
-2.2
PC5
PC2

phenotype explaining SNV variation, and two transcription-orthogonal axes across promoter DNA methylation and chromatin accessibility.

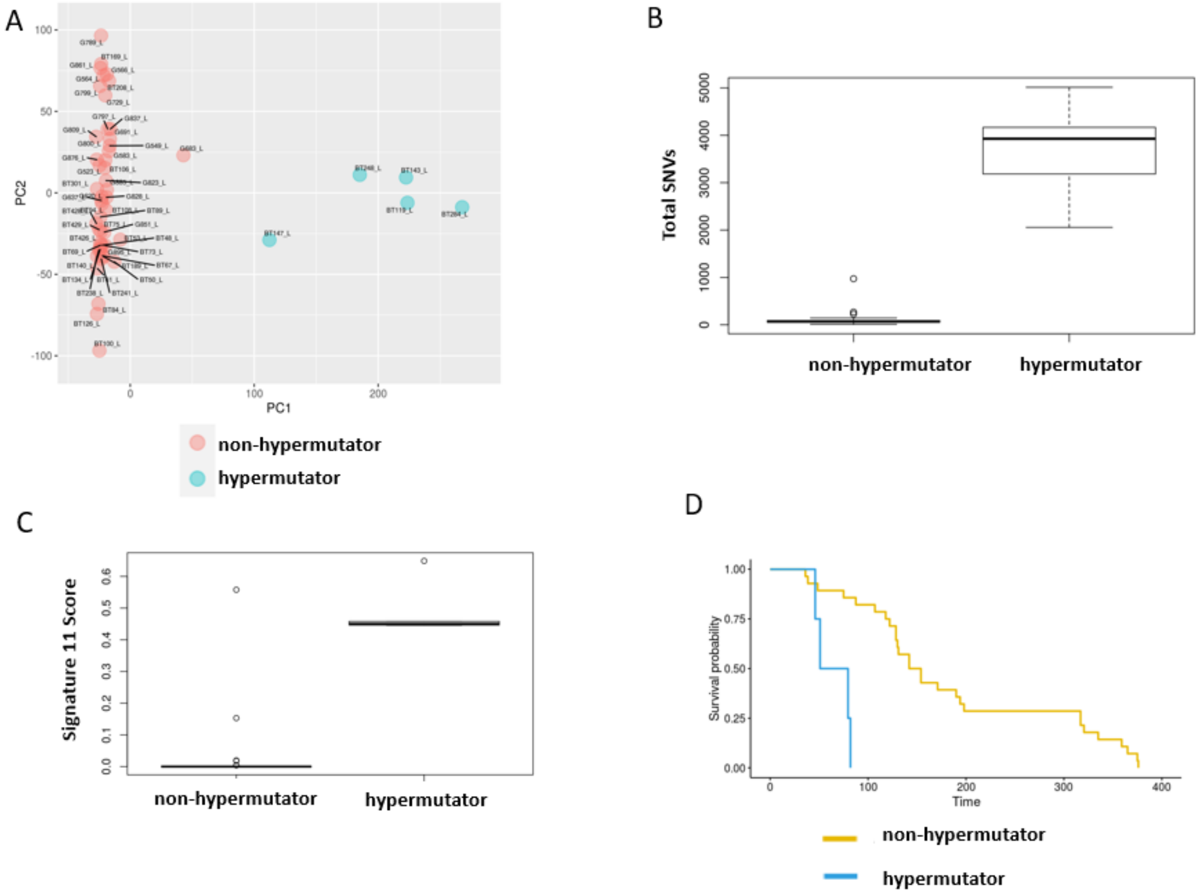
3.3.2 Hypermutated GSCs defined by signature 11 mutations and mismatch repair deficiency while non-hypermutators defined by common GBM mutations

From the multi -omics PCA analysis, we immediately saw that samples separated by the presence or absence of a hypermutator phenotype associated with mutation signature 11, indicative of temozolomide treatment[81,190] and captured by PC1. Indeed, we saw that MSH6 mutation, which causes a signature 11 hypermutation phenotype[81], was present in all hypermutated lines (Table 3.1), which we identified using k-means clustering (k=2) on SNV count (Figure 3.4A). Hypermutation was also associated with mutation signature 11 scoring of our SNV data (Figure 3.4C) using the mutSignatures package and COSMIC signatures (see *Methods* for details). GSCs derived from ten recurrent patient tumors significantly overlapped (though were not exclusive to) the hypermutated cluster (Table 3.2, $p < 8.0$ e-5, Fisher's exact test). Hypermutated GSCs were also associated with shorter xenograft survival time (Figure 3.4D). Collectively, these results are consistent with prior work showing temozolomide based hypermutation to result in more aggressive tumors[81,83].

Individual mutations enriched in hypermutators relative to non-hypermutators (Fisher’s exact test, two-sided, FDR < 0.05) associated with genes in PIP signaling, cell junctions/adhesion, neuron differentiation, and histone modification (Fisher’s exact test, one-sided, FDR < 0.10). These mutations corresponded to clustering along an axis orthogonal to transcription, indicating that there may be changes in cellular signaling that do not result in large transcriptional changes but do affect cellular aggression properties. TTN, P53, PTEN, and EGFR were among the most commonly mutated genes in non-hypermutator GSC samples, consistent with prior results[4,191] (Appendix Table 1). There were no differentially mutated genes between Developmental and Injury Response GSCs (Fisher’s exact test, FDR > 0.05), suggesting that while SNVs could help a cell towards becoming cancerous, they are not a determinant in transcriptional phenotype along the Developmental/Injury Response Axis.

Figure 3.4 Hypermutation axis associated with drop in xenograft survival

(A) GSCs clustered by total SNV signal with k-means clustering (k=2), overlaid on PCA coordinates from Figure 2. Blue dots represent hypermutated samples. **(B)** Total SNV signal in non-hypermutators (SNV_C1, n = 49), and non-hypermutators (SNV_C2, n = 5). Comparison: one-sided t-test, alternative hypothesis of SNV_C2 mean greater than SNV_C1 mean. **(C)** Signature 11 mutation score in hypermutator (n = 5) compared to non-hypermutator samples (n = 49) (p-value = 6.4e-5, by one-sided t-test). **(D)** Xenograft survival (in days) among mice injected with hypermutator (n = 4) and non-hypermutator (n = 28) GSCs. Hypermutator GSCs result in shorter survival (p = 0.00013, log-rank test)



3.3.3 Copy Number Variation may modulate Developmental and Injury Response gene expression

While the expected chromosomal patterns of copy number changes in GSCs (i.e. chromosome 7 amplification, chromosome 10 deletion, CDKN2 deletion on chromosome 9p) were present (Figure 3.5, Appendix Table 2), consistent with existing knowledge of GBMs[4,73], the majority of variation in the set of variable CNV features was explained by PC2 and PC3, suggesting association with the Developmental/Injury Response axis. We thus investigated the correlation of protein-coding gene expression with copy number status, suggestive of copy number driven gene expression changes (Figure 3.6A). The vast majority of significant CNV-gene expression correlation (examining all genes common to the CNV and RNA-seq data, n = 21833) is positive (n = 6633, FDR = 0.05), with a few negative correlation relationships (n = 45). Thus, there is a large portion of the protein coding genome in GSCs with expression potentially affected by copy number. When examining how these CNV-correlated genes relate to the Developmental to Injury Response gradient, we found 1036 Developmental associated and 1020 Injury Response associated genes (compared to signature sizes of 3968 and 4399 genes for the Developmental and Injury Response signatures, respectively) whose expression is positively correlated with copy number (Figure 3.6A). Developmental copy number affected genes are enriched in chromatin modification related pathways (Appendix Table 3) and Injury Response copy number affected genes are enriched for inflammatory and hypoxia related pathways (Appendix Table 4). These results suggest that CNV status can affect specific pathways that are part of the Developmental to Injury Response gradient.

3.3.4 Injury Response genes and pathways are methylation-repressed in Developmental GSCs

We next investigated DNA methylation variation associated with the Developmental to Injury Response gradient, which was captured by PC2 and PC3 in the multi -omics PCA. DNA methylation at promoters often is associated with stabilization of heterochromatin and prevention of transcription, although at times transcription factors can specifically prefer a methylated promoter for gene expression activation[192]. We looked for Developmental to Injury Response gradient genes whose expression levels are associated with promoter methylation (Figure 3.7A).

Figure 3.5 Copy Number Profiles of GSCs

Gene-wise averages of estimated log2(Tumor/Blood) genomic copy number ratio, averages done per chromosome arm.

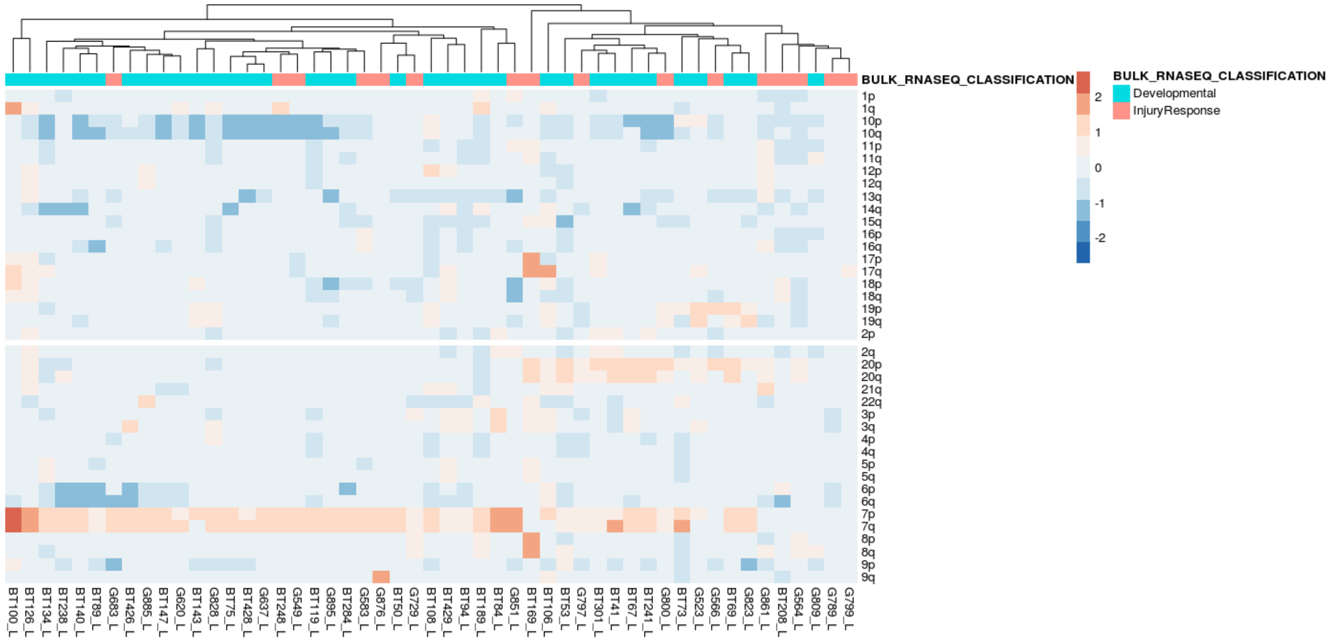


Figure 3.6 Association of copy number variation with Developmental/Injury Response axis

(A) Genes significantly correlated (FDR < 0.05) with CNV signal (log2 (tumor/normal), GISTIC 2.0) and with Developmental or Injury Response state. Pathways significantly enriched (FDR < 0.05) in genes that have positive association of gene expression with CNV signal and association with the Developmental or Injury Response transcriptional phenotypes are shown.

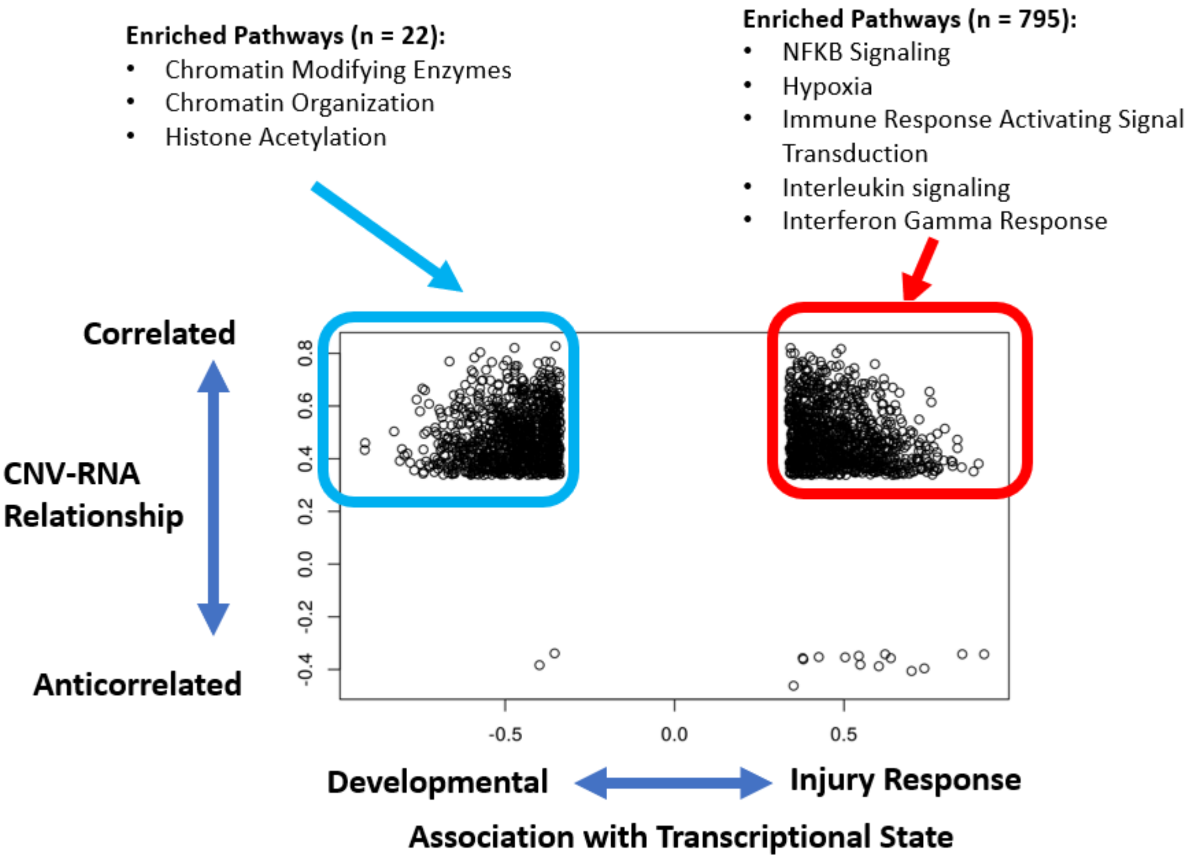
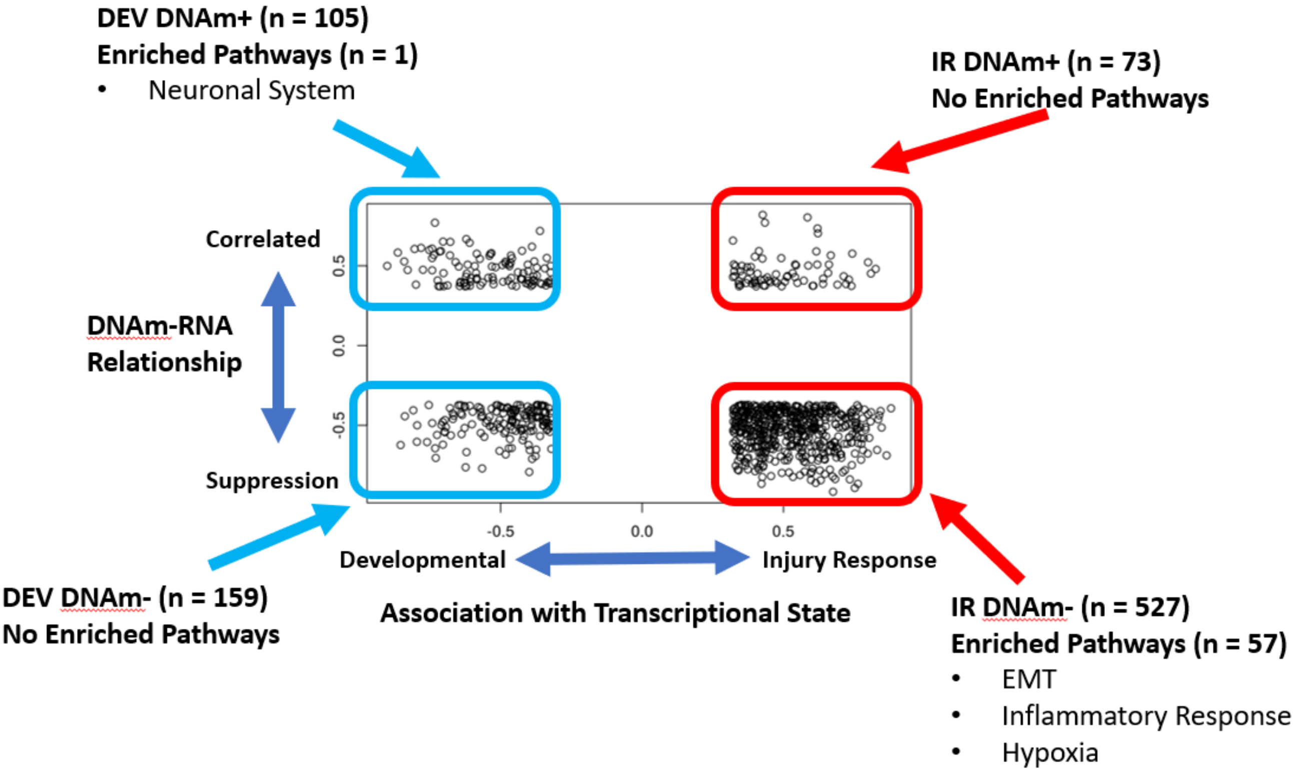


Figure 3.7 DNA methylation as a suppressor of Injury Response phenotype

(A) Genes significantly correlated (FDR < 0.05) with promoter methylation signal and with Developmental or Injury Response state. Pathways significantly enriched (FDR < 0.05) in DEV DNAm+ and IR DNAm- genes are shown.



DEV DNAm+ (n = 105)
Enriched Pathways (n = 1)

- Neuronal System

IR DNAm+ (n = 73)
No Enriched Pathways

DEV DNAm- (n = 159)
No Enriched Pathways

IR DNAm- (n = 527)
Enriched Pathways (n = 57)

- EMT
- Inflammatory Response
- Hypoxia

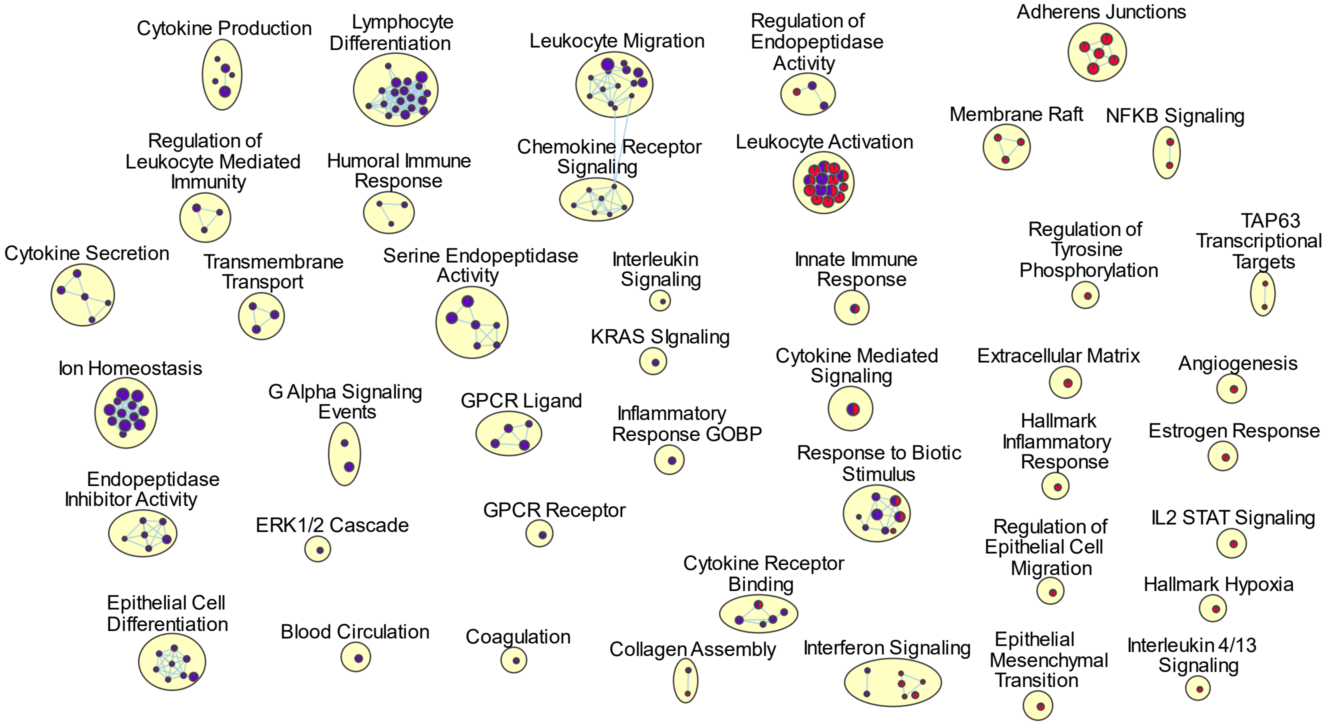
Anticorrelation of promoter methylation and gene expression is potentially indicative of methylation stabilizing a suppressive gene expression state, while positive association could represent gene expression activation. This created four categories of genes depending on association with Developmental (DEV) or Injury Response (IR) transcriptional gradient and association with promoter methylation (DNAm+ for promoter methylation positive association or alternatively, DNAm-). Each category contained genes: DEV DNAm+ (n = 105), DEV DNAm- (n = 159), IR DNAm+ (n = 73), IR DNAm- (n = 527); however only IR DNAm- genes are strongly enriched in a large number of biological pathways, grouped into epithelial mesenchymal transition, inflammatory response, and hypoxia themes (Figure 3.7A, Figure 3.8A). These results suggest that in Developmental GSCs, promoter methylation and suppression of Injury Response genes contributes to the maintenance of a Developmental transcriptional state.

We additionally found a principal component associated with aggregate promoter methylation signal (PC5) (Fig 3.3E) orthogonal to the PC2/PC3 Developmental/Injury Response axis. Pathway analysis of genes correlated with PC5 showed that genes whose promoter methylation was negatively associated with PC5 were mainly enriched for immune cell differentiation and inflammation related pathways (Figure 3.8). Enriched biological pathways were generally complementary to those enriched in IR DNAm- genes, with only 11 pathways of 57 PC5 associated pathways and 208 IR DNAm- pathways intersecting between the two sets. The gene sets themselves generally did not involve the same genes, with only 60 of 2966 PC5 negatively associated genes also in the IR DNAm- category, and none of these 60 genes enriched in pathways (FDR < 0.05). These results suggest that there may be a general DNA methylation-based regulation (potentially suppression, thought we cannot tell from the data presented) of a subset of inflammatory related signaling that does not manifest in changes in transcriptional state.

Overall, the results we have obtained for the DNA methylation data are suggestive of promoter methylation as an epigenetic method of suppressing Injury Response gene expression, contributing to Developmental/Injury Response transcriptional variation,, and any effects PC5's varied promoter methylation might have on transcription dominated by the Developmental/Injury Response axis.

Figure 3.8 Pathway enrichment comparison of PC5 DNAm associated genes and IR DNAm- genes

(A) Pathway enrichment analyses were performed using Fisher's exact test for pathway membership using (separately) PC5 correlated genes (w.r.t. promoter DNAm) and IR DNAm-genes as query sets (FDR < 0.05). Results are visualized as an enrichment map, where pathways are represented as nodes connected if they share genes, and colored by the query set(s) which they are enriched for (either PC2/PC3 – red – or PC5 related – purple).


Cytokine Production
Lymphocyte Differentiation
Leukocyte Migration
Regulation of Endopeptidase Activity
Adherens Junctions
Membrane Raft
NFKB Signaling
Regulation of Leukocyte Mediated Immunity
Humoral Immune Response
Chemokine Receptor Signaling
Leukocyte Activation
Cytokine Secretion
Transmembrane Transport
Serine Endopeptidase Activity
Interleukin Signaling
Innate Immune Response
Regulation of Tyrosine Phosphorylation
TAP63 Transcriptional Targets
KRAS SIgnaling
Ion Homeostasis
G Alpha Signaling Events
GPCR Ligand
Inflammatory Response GOBP
Cytokine Mediated Signaling
Extracellular Matrix
Angiogenesis
Hallmark Inflammatory Response
Estrogen Response
Response to Biotic Stimulus
Endopeptidase Inhibitor Activity
ERK1/2 Cascade
GPCR Receptor
Regulation of Epithelial Cell Migration
IL2 STAT Signaling
Cytokine Receptor Binding
Hallmark Hypoxia
Epithelial Cell Differentiation
Blood Circulation
Coagulation
Collagen Assembly
Interferon Signaling
Epithelial Mesenchymal Transition
Interleukin 4/13 Signaling

🔴 Injury Response Genes w/ negative association with promoter DNAm

🟣 Genes with promoter DNAm negatively associated with PC5

3.3.5 Chromatin Accessibility has transcription coupled and transcription independent components

We then explored the GSCs within the space of the ATAC-seq data. After merging peaks across samples, peak features covered regions ranging in size from 200-3257 bp, with 90% of regions under 888bp (Figure 3.9A). This distribution of peak lengths is overall consistent with typical ATAC-seq fragment lengths being between less than 1000 bp[48,49], though the peak detection method we used allowed for the detection of open chromatin regions larger than that size. Generally, variation in chromatin accessibility was greater between individual samples than across chromosomes within an individual sample (Figure 3.9B). Given the presence of a principal component (PC4) responsible for much variation along aggregate ATAC-seq signal (Figure 3.3D), and another set of principal components (PC2, PC3) responsible for a considerable portion of the rest of ATAC-seq related variance (Figure 3.3A), we asked if variation along the Developmental/Injury Response Axis could be the result of variation in specific sets of chromatin regions.

We identified differentially accessible chromatin regions between Developmental and Injury Response GSCs (using cluster labels from Richards and colleagues[188]) using ATAC-seq peak calls for 40 GSC lines (n = 14 for Developmental, n = 26 for Injury Response). We found that differentially accessible regions (n = 20527 for Developmental lines, n = 4336 for Injury Response lines, FDR = 0.20) were mutually opposing, generally open in one transcriptional state and closed in the other (Figure 3.10 A,B). We then investigated if these differentially accessible chromatin regions are enriched for transcription factor binding motifs, and found that regions accessible in Developmental GSCs were enriched for RFX, CTCF/BORIS, and OCT/POU motifs (Figure 3.10C), while regions open in Injury Response GSCs were enriched for NFKB, FOSL1 (FRA1)/FOSL2 (FRA2), IRF, NRF2, BACH1, BACH2, and NFE2 motifs (Figure 3.10D). Interestingly, RFX2 and POU3F3 were shown by Richards and colleagues[188] to be upregulated in Developmental GSCs, while IRF1, IRF7, IRF9, NFKB, NRF2 (also known as NFE2L2), NFE2, FOSL1, and FOSL2 are upregulated in Injury Response GSCs. POU3F3 and RFX2 are involved in neural development and stemness[193,194] , while NFKB, the FOS/JUN (AP1) transcription factors, and IRF transcription factors are involved in promoting epithelial-mesenchymal transition and inflammatory signaling[100,195–198]. CTCF is involved in neural development and is highly expressed during embryonic development[199], making its binding of

Figure 3.9 General Characteristics of ATAC-seq data for 40 GSC lines

(A) Distribution of peak widths for detected peaks; **(B)** Averaged log chromatin accessibility signal across chromosomes.

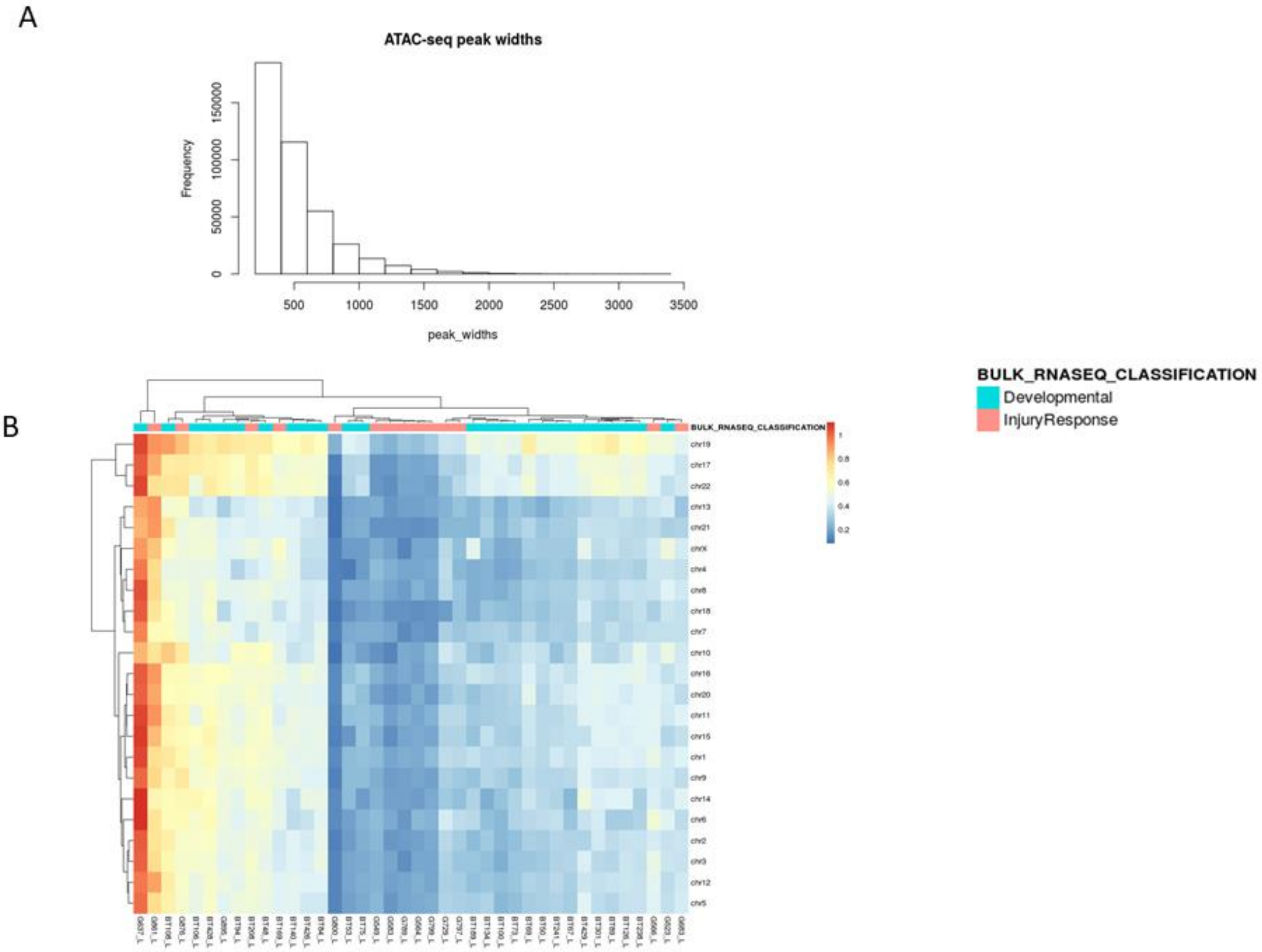
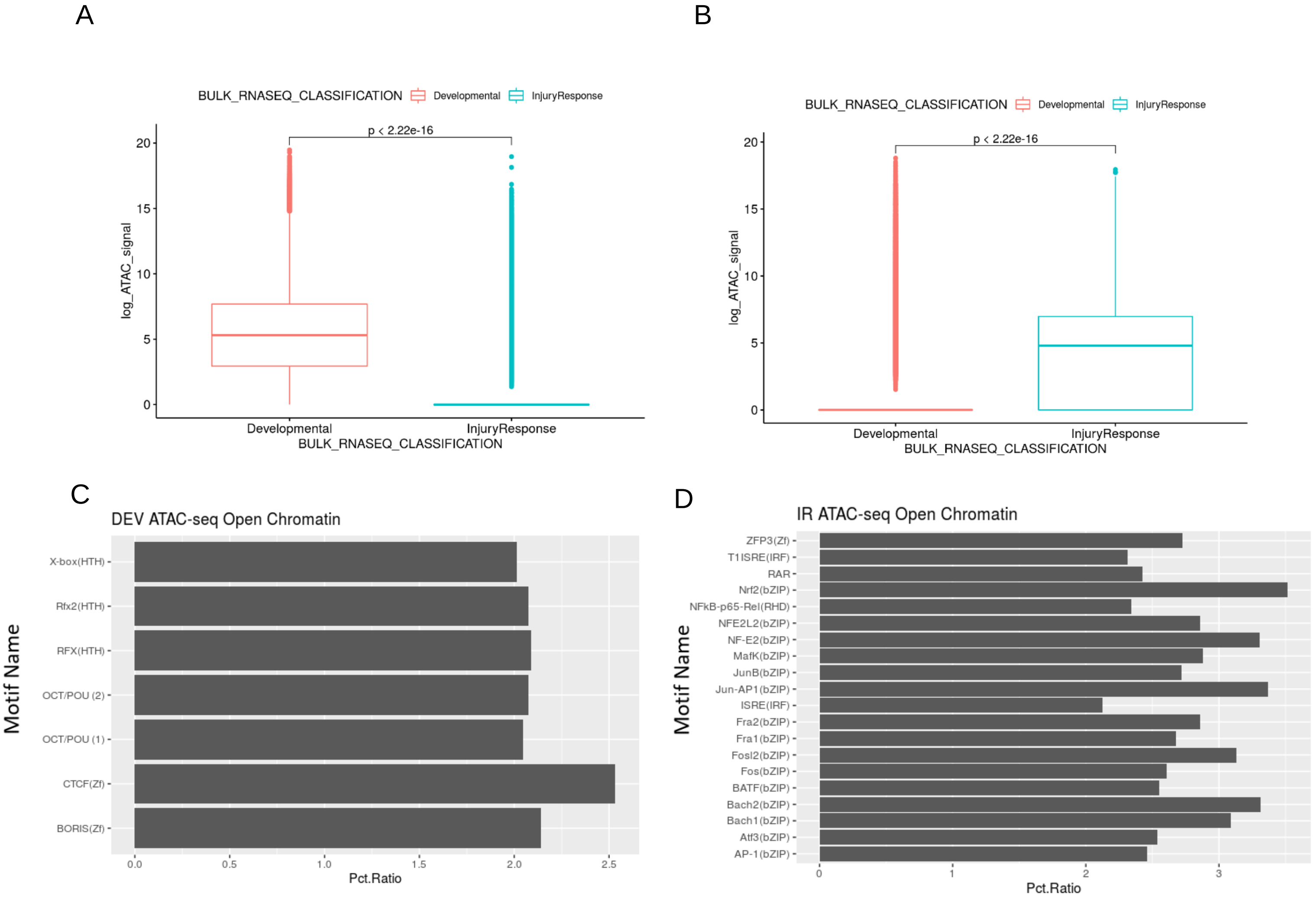


Figure 3.10 Differential chromatin accessibility corresponds to differentially expressed TFs

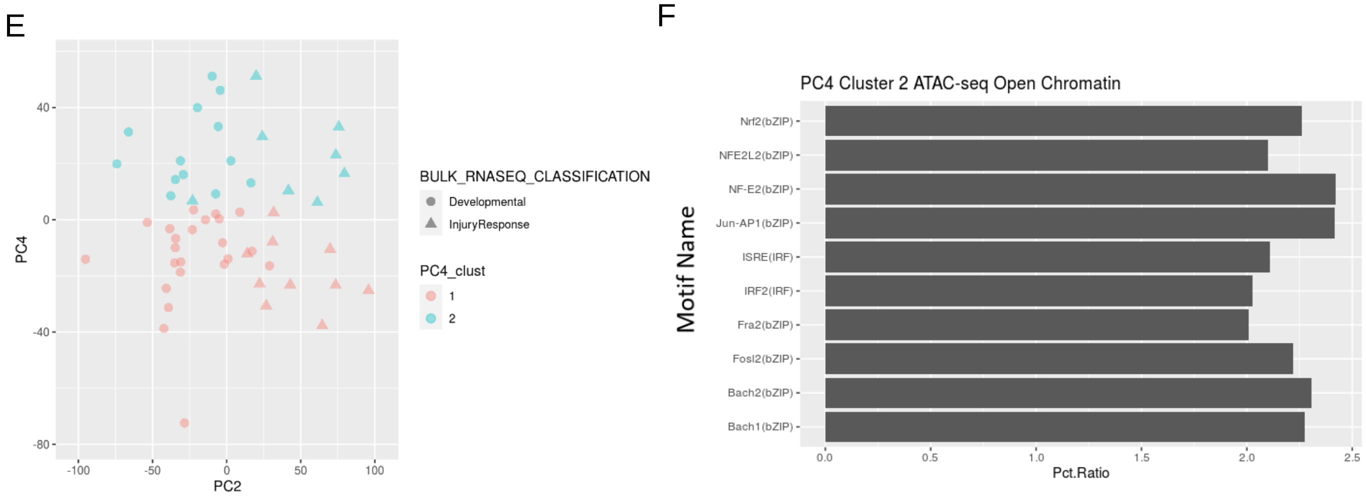
Motif enrichment for TF motifs found by HOMER given open regions found in Developmental and Injury Response GSCs or PC4 Cluster 2 by ATAC-seq. Bars represent the ratio of percentage query region hits to percentage background region hits. **(A-B)** Chromatin accessibility signal for regions found to be more open in Developmental **(A)** and Injury Response **(B)** GSCs. P-values are from two-sided Wilcoxon rank sum tests. **(C)** Enriched motifs found in Developmental enriched open chromatin regions by HOMER. HOMER motif names are mapped to the names shown in this figure via Table 3.4. **(D)** Enriched motifs found in Injury Response enriched open chromatin regions by HOMER. HOMER motif names are mapped to the names shown in this figure via Table 3.5. **(E)** Clustering of GSCs along PC4 using k-means clustering. **(F)** Enriched motifs found in PC4 Cluster 2 enriched open chromatin regions by HOMER. HOMER motif names are mapped to the names shown in this figure via Table 3.6.

DEV Open Regions

IR Open Regions



PC4 Associated Chromatin Regions



regions in the Developmental GSCs consistent with the notion that they represent a phenotype similar to fetal brain cell types[188]. Additionally, NRF2 regulates response to oxidative stress[200], and NFE2, while normally pertaining to hematopoietic cell differentiation into megakaryocytes[201], has been linked to metastasis in breast cancer and shown *in vitro* to give advantages in a hypoxic environment[202]. As for BACH1 and BACH2, given the similarity of NFE2 to BACH1 and BACH2 binding motifs as found by homer[203] (similarity scores of 0.98 and 0.95, respectively), it is unsurprising that these transcription factors appear as hits despite their lack of upregulation in Injury Response GSCs relative to Developmental GSCs[188]. We then ran a pathway analysis on genes with a promoter region or gene body in differentially accessible regions in Developmental and Injury Response GSCs and found that Developmental accessible chromatin regions pertained to neural developmental and ion channel related pathways, consistent with RNA-seq expression results presented by Richards and colleagues[188] (Appendix Table 5). The genes found in accessible chromatin regions enriched in Injury Response GSCs did not yield statistically significant pathway enrichment, likely due to a lack of statistical power given the small number of genes mapped (n = 126). Nonetheless, the presence of FOSL1/2 binding motifs within the Injury Response accessible regions suggests that epithelial mesenchymal transition and invasive behaviors may be promoted[195,196,204]. Thus, chromatin accessibility and transcription factor activity likely act to promote or stabilize the Developmental and Injury Response transcriptional programs.

To examine PC4, which also explained a substantial portion of variance in the ATAC-seq data, we asked if its broad association with average ATAC-seq signal would be enriched for particular sets of transcription factors. We clustered samples along PC4, splitting samples by that PC (Figure 3.10E). Using a similar approach to finding differentially accessible regions as was done for Developmental and Injury Response GSCs above, we found no regions enriched in PC4 Cluster 1 and 22640 regions enriched in PC4 Cluster 2. Interestingly, PC4 Cluster 2's open regions were enriched for some of the same transcription factor motifs found in Injury Response open regions (e.g. JUN/AP1, BACH1, BACH2, FOSL2 and NFE2, and IRF motif sequences, Figure 3.10F). These results suggest that there may be differential transcription factor activity related to inflammatory signaling that does not result in broad changes in transcription. Taken together with the predicted differential transcription factor occupancy associated with the Developmental and Injury Response axis, the transcription-orthogonal chromatin accessibility axis (PC5) suggests that, like promoter methylation, transcription factor occupancy is a contributor to, but not the only determinant of transcriptional output.

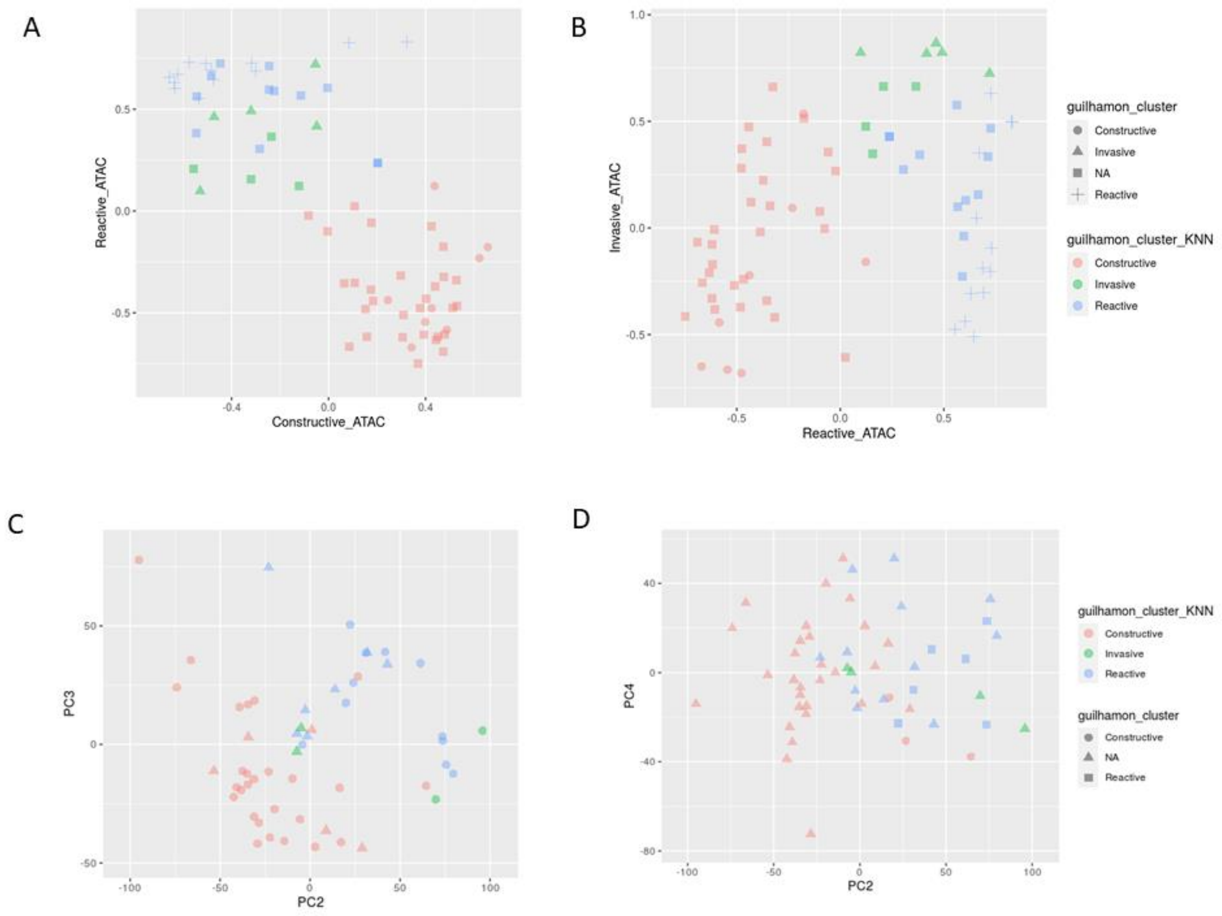
Previous work by Guilhamon and colleagues identified three GSC states in a dataset of 27 GSC lines[189]: Constructive, Reactive, and Invasive based on ATAC-seq, DNA methylation and RNA-seq data. Given their use of similar information to identify a state (Invasive) that displayed shorter survival than the two others identified, we wondered how these three states distributed in the space of our integrated PCA. Having data pertaining to 26 of the 27 GSC lines used in Guilhamon et al., 9 of which overlapped with the set of 54 GSC lines used for multi -omics PCA, we took ATAC-seq data from the union (n = 71) of the 26 GSCs available from the Guilhamon publication and 54 GSC lines used for this study (40 with measured data, 14 with imputed data), and performed KNN imputation of unknown Constructive, Reactive, and Invasive labels (45 unlabeled GSCs) using known labels from the Guilhamon et al. publication[189]. To obtain a low dimensional feature space appropriate for label imputation, we obtained three open chromatin signatures (one each for Constructive, Reactive, Invasive, see *Methods*) that we applied to labeled and unlabeled data, and fit a KNN classifier to the 26 samples with known labels, predicting the labels of the 45 unlabeled GSCs (Figure 3.11 A,B). The Reactive and Invasive states were for the most part separated from the Constructive state along PC2 and PC3 (Figure 3.11C). The Injury Response transcriptional phenotype broadly overlapped with the Reactive and Invasive states, whereas the Developmental transcriptional phenotype broadly corresponded to the Constructive state (Table 3.3). Additionally, the Invasive state was rare (n = 4) in our 54 sample dataset and did not appear to be well separated by the major components of variation for ATAC seq (Figure 3.11C, D). Additional study of Invasive lines will be needed to link them to the mechanisms in the multi-omics analysis presented here.

3.3.6 Mutually Opposing miRNA suppressive programs target Developmental and Injury Response genes

The strongest source of signal in the miRNA data correlated with the Developmental/Injury Response axis suggesting a regulatory role for miRNAs in the two transcriptional programs. To identify potential miRNA regulators of the Developmental/Injury Response axis, we performed differential miRNA expression analysis between Developmental and Injury Response GSCs

Figure 3.11 Signature based imputation of Guilhamon et al.[189]

(A-B) Distribution of samples with known Constructive/Reactive/Invasive labels and those imputed via KNN (n = 71 total), shown in space of determined ATAC-seq based signatures. **(C-D)**: Distribution of GSCs (n = 54) in PCA space from multi -omics PCA. Sample points are shaped by known GSC state (if available, NA otherwise) and colored by KNN imputed GSC state (GSCs with known state are colored by known state).

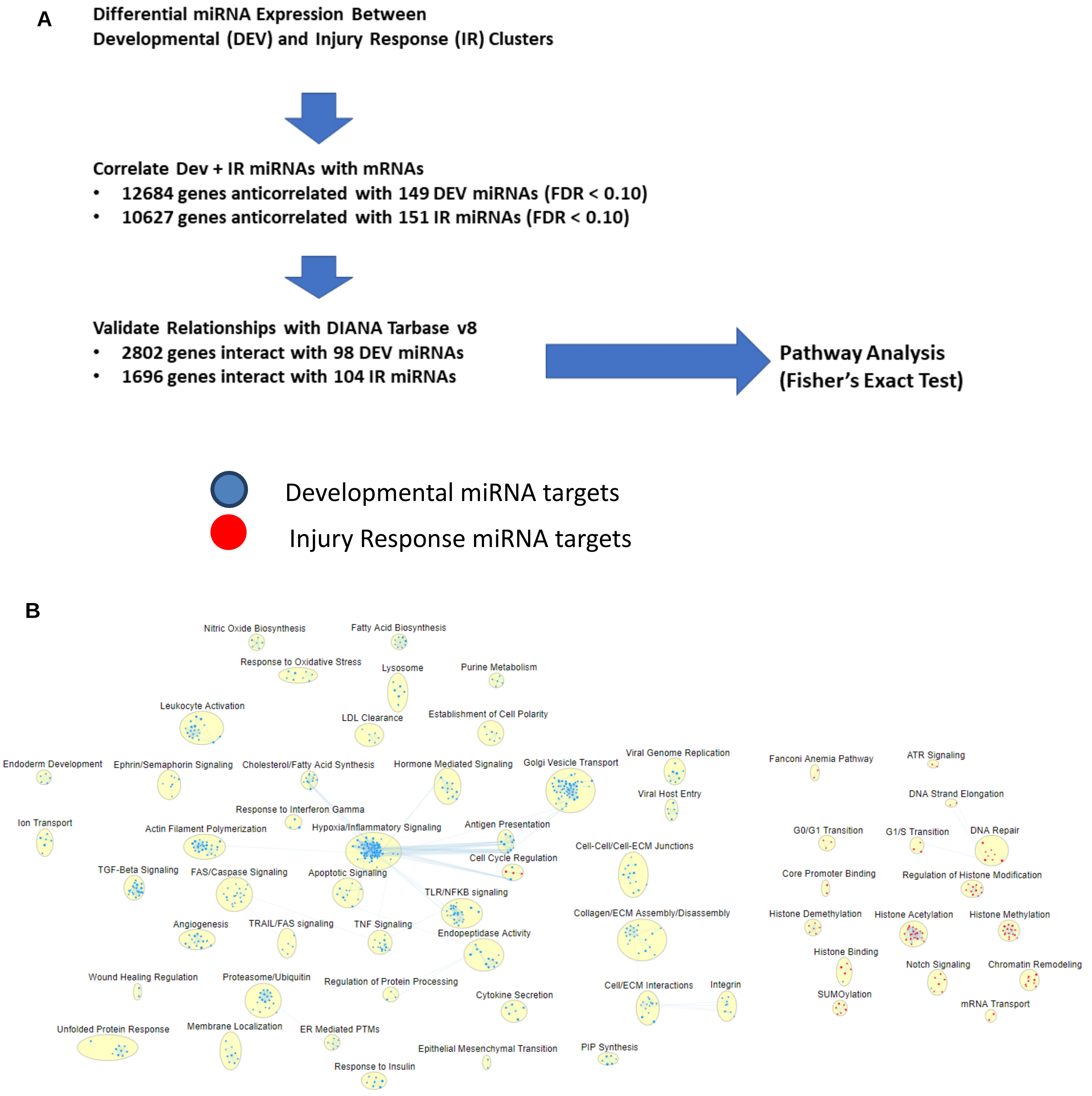


(FDR < 0.05) filtering these for anticorrelation between miRNAs and RNA expression (FDR < 0.10), and further filtering these for experimentally known miRNA-RNA relationships present in DIANA-TarBase v8[205] (Figure 3.12A). This resulted in interactions between 98 Developmental upregulated miRNAs and 2802 genes (Appendix Table 6) and between 104 Injury Response upregulated miRNAs and 1696 genes (Appendix Table 7) A pathway analysis on the targeted genes (Fisher's exact test for overrepresentation, one-sided, FDR < 0.05) revealed that Developmental upregulated miRNAs predominantly targeted genes in Injury Response associated pathways, while Injury Response miRNAs targeted genes in chromatin remodeling (shown by Richards and colleagues to be more expressed and essential in Developmental GSCs[188]), as well as DNA repair and cell cycle regulation (Figure 3.12B). Thus, we predict that the Developmental and Injury Response mRNA transcriptional programs are associated with mutually opposed miRNA based suppression programs, with Developmental miRNAs suppressing aspects of the Injury Response transcriptional program and Injury Response miRNAs conversely suppressing aspects of the Developmental transcriptional program. These results support a role for miRNAs in stabilizing and maintaining the transcriptional phenotype along the Developmental/Injury Response transcriptional axis.

Richards and colleagues showed that many of the biological pathways affected by differential expression along the Developmental/Injury Response transcriptional axis were also differentially essential between Developmental and Injury Response GSCs[188]. To identify potentially important miRNA regulators, we compared our miRNAs and their targets to sensitive genes from CRISPR screens. We re-analyzed our published CRISPR screening data[188], defining uniquely essential genes for Developmental and Injury Response GSC lines (Appendix Table 8, Appendix Table 9). We then intersected this list with Developmental and Injury Response differentially expressed genes[188], and further intersected these with the union of DEV and IR miRNA targets (Figure 3.13). Interestingly, the resulting predicted Developmental miRNAs only targeted Injury Response genes and Injury Response miRNAs only targeted Developmental genes. These results suggest that Developmental GSCs express miRNAs that would render Injury Response GSCs less fit and that the converse would be true for Injury Response GSCs' miRNAs, suggesting that miRNAs are a possible therapeutic target for GSCs. Some of the miRNAs found to be differentially expressed between Developmental and Injury Response GSCs have roles as tumor suppressors (e.g. miR-128)[206] or tumor enhancers (e.g. miR-10b, miR-21)[206], and in particular

Figure 3.12 Mutually opposed miRNA based suppression programs in Developmental and Injury Response GSCs

(A) Schematic showing workflow for discovering and filtering differentially expressed miRNAs and their targets in Developmental and Injury Response GSCs. Differentially expressed miRNAs were filtered at FDR 0.05. Developmental and Injury Response upregulated miRNAs (separately) underwent correlation analysis to find correlated genes (FDR 0.10), and negative correlation pairs were filtered against DIANA-TarBase and then fed to a pathway analysis (Fisher's exact test for overrepresentation, one-sided, FDR < 0.05) **(B)** Pathway analysis of filtered miRNA gene targets as an enrichment map. Blue nodes represent pathways enriched for targets of Developmental miRNAs, red nodes represent pathways enriched for targets of Injury Response miRNAs. Edges represent overlap relationships between pathways (average of Jaccard and overlap coefficients).



A

Differential miRNA Expression Between Developmental (DEV) and Injury Response (IR) Clusters

Correlate Dev + IR miRNAs with mRNAs

- 12684 genes anticorrelated with 149 DEV miRNAs (FDR < 0.10)
- 10627 genes anticorrelated with 151 IR miRNAs (FDR < 0.10)

Validate Relationships with DIANA Tarbase v8

- 2802 genes interact with 98 DEV miRNAs
- 1696 genes interact with 104 IR miRNAs

Pathway Analysis (Fisher's Exact Test)

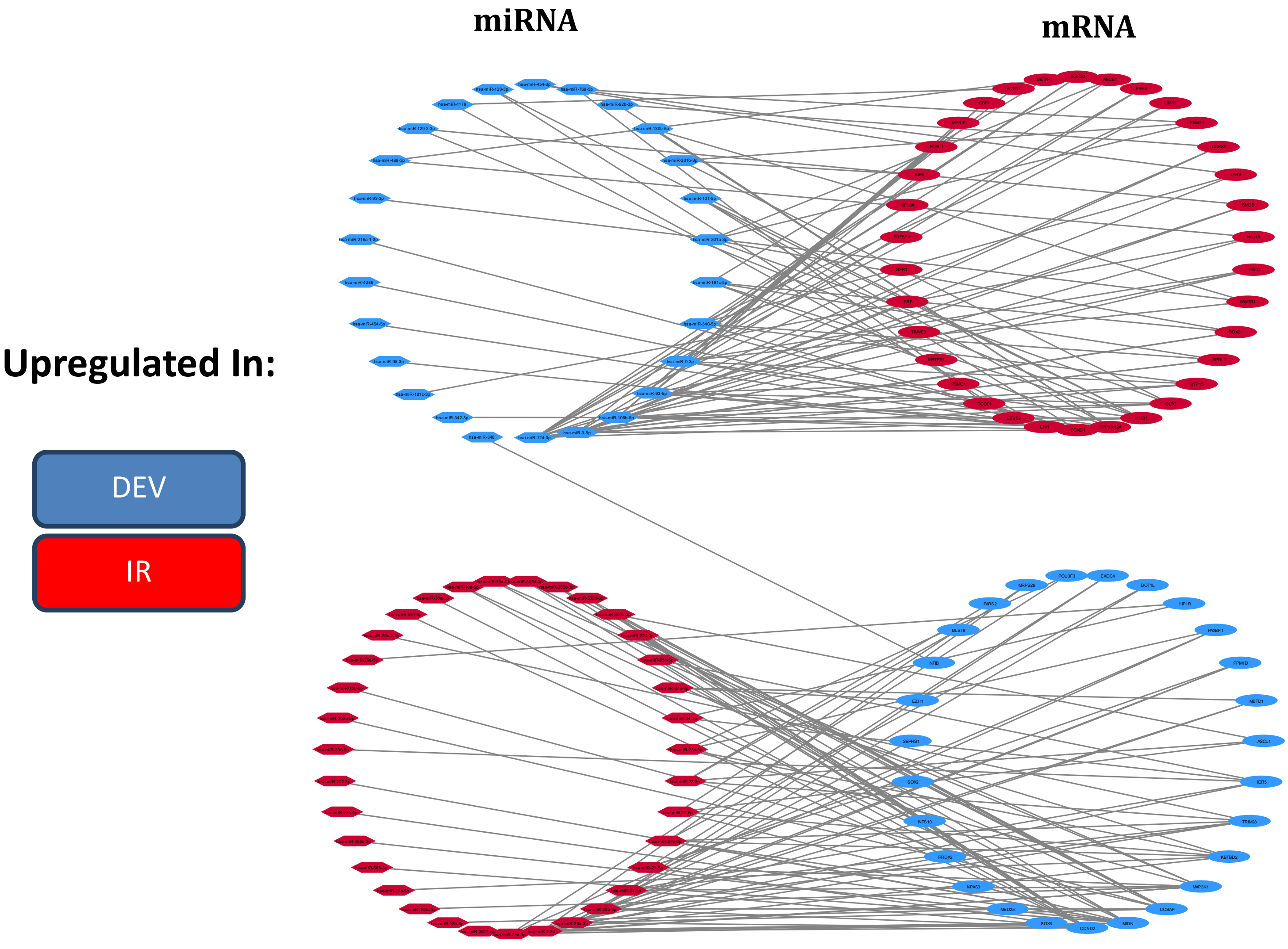
● Developmental miRNA targets

● Injury Response miRNA targets

B

Figure 3.13 miRNAs target differentially expressed, differentially essential genes for Developmental and Injury Response GSCs

Network of interactions between differentially expressed, differentially essential genes and differentially expressed miRNAs targeting them. Red nodes represent Injury Response genes and blue nodes represent Developmental genes. Circular nodes represent RNA and hexagonal nodes represent miRNAs. Edges represent predicted suppression relationships between miRNA and mRNA.



there are examples of miRNAs shown to interact with differentially essential genes having roles in regulating neural developmental (miR-21, upregulated in Developmental GSCs)[207] and invasive/metastatic phenotypes (miR-598, upregulated in Injury Response GSCs)[208]. While the one-to-many and many-to-one relationships between miRNAs and their targets complicate the prediction of phenotypes upon overexpression or knockdown (e.g. with the risk of transforming Developmental GSCs to Injury Response GSCs, or vice versa, without decreasing proliferation or survival), understanding the role of miRNAs in maintaining heterogeneous gene expression programs provides another handle by which we can manipulate GSC biology and attempt to downregulate those programs essential to GSCs' survival.

3.4 Discussion

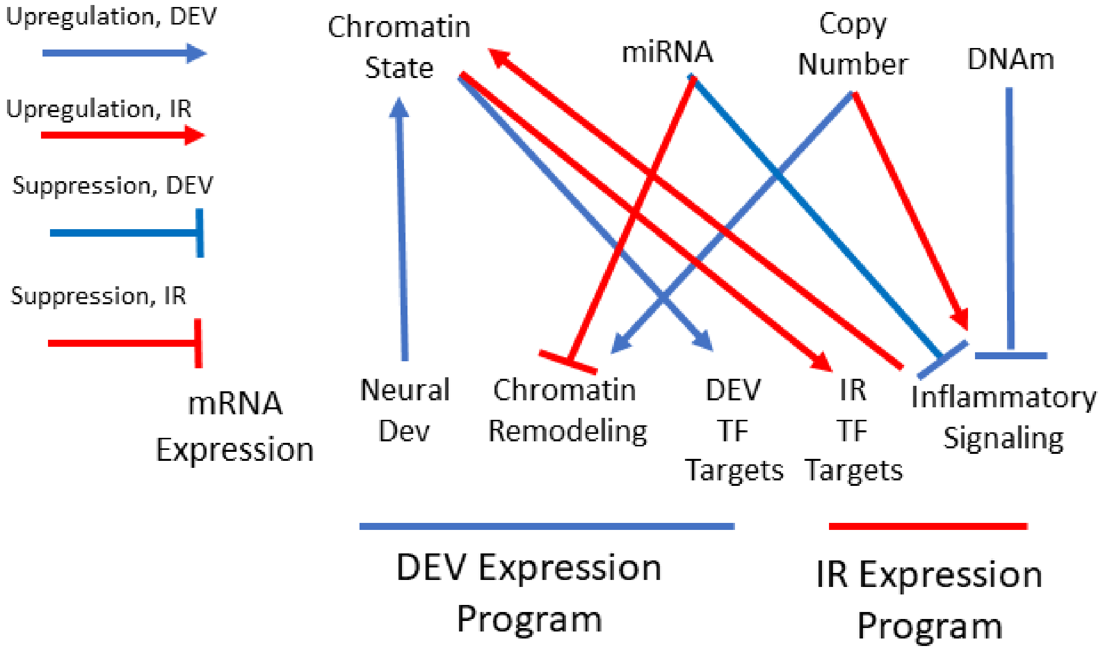
Here, leveraging the information in six -omics data types from five large scale -omics data collection efforts, we show the existence of four major components of variation in GSCs across multiple -omics layers, one defined by the Developmental/Injury Response transcriptional axis and the other three defined by hypermutation, promoter methylation, and chromatin state. With the results presented here, as well as the differential essentiality results among Developmental and Injury Response GSCs described by Richards and colleagues[188,189], we present evidence that the space represented by two of these axes translates into differences in functional behavior depending on a GSC's position within it. In particular, we have found that hypermutation phenotype and its associated reduction in xenograft survival occur largely independently of position on the Developmental/Injury Response transcriptional axis, whereas that axis, with its associated variation in gene essentiality, is affected through all other molecular processes we measured in this study, whether through miRNAs, promoter methylation, chromatin accessibility, or DNA copy number.

We hypothesize that separate biological processes are involved for hypermutation and the Developmental/Injury Response axis, while promoter methylation and transcription factor binding may make partial contributions to inflammatory transcriptional phenotype. Touat and colleagues[81], observing that recurrent GBMs were enriched for but not exclusively composed of hypermutated tumors, suggested that temozolomide treatment selected for mismatch repair mutated cells, which appears to be the case in our GSC lines given the enrichment of hypermutated GSCs among those derived from recurrent tumors. For the Developmental/Injury Response axis, from the data presented alone it cannot be determined which biological process, if any, precedes all others in promoting a Developmental or Injury Response transcriptional state. Copy number variation and non-genetic processes (i.e. chromatin state or promoter methylation) may precede transcriptional state, or may reinforce it and be selected due to known associations between biological pathways' differential expression and their differential essentiality among Developmental and Injury Response GSCs[188]. GSC state is plastic and Developmental GSCs can be induced to Injury Response phenotype through exposure to cytokines, activation of transcription factors, or miRNAs[100,188,207], and while the reverse transformation, to our knowledge, has yet to be demonstrated, there are at the very least multiple possible pathways to an Injury Response-like state. We also hypothesize that variation in inflammatory transcription factor occupancy and promoter methylation state for inflammatory genes can contribute to the expression of Injury Response transcriptional phenotype without being the sole determinants of final transcriptional output, explaining the apparent transcription-associated and transcription-orthogonal axes observed for promoter methylation and chromatin state. Thus, we hypothesize that a GSC's position on the Developmental and Injury Response axis are affected by a variety of mechanisms, and while coordinated effects of various biochemical processes may act to stabilize position in one state or another, intrinsic or environmental signals are capable of shifting transcriptional and epigenetic state. More broadly, our results are suggestive of a recurrence enriched hypermutation phenotype due to mismatch repair, and a bidirectional transcriptional axis with stable state preceded by a variety of non-genetic events that can be potentiated or stabilized by CNVs (Figure 3.14).

Given the complexity and heterogeneity demonstrated here at multiple levels within the cancer stem cell population in glioblastoma, it is unsurprising that previous efforts at targeted single-agent therapeutics have been largely unsuccessful. With the knowledge of what axes of variation exist and which molecular processes act in a coordinated fashion or independently of one another, future studies should be able to better develop targeted therapies for GSCs. In particular, miRNAs may represent a potential lever to modulate transcriptional phenotype and target differentially essential genes. Given the heterogeneity and plasticity of GSCs[100,188,189] and apparent redundancy in the molecular mechanisms observed, it is likely that future GBM therapeutics will need to target multiple aspects of these molecular mechanisms in order to prevent or delay recurrence.

Figure 3.14 Summary of proposed model of GSC Multi -Omic heterogeneity based on presented findings

Arrows represent activation/positive regulation relationships, while lines indicate suppression/negative regulation relationships. Blue relationships indicate events occurring in Developmental GSCs, while red relationships indicate events occurring in Injury Response GSCs.



3.5 Methods

3.5.1 RNA-sequencing

RNA-sequencing, count data generation, and VST transformation were performed as described by Richards and colleagues[188] for 54 GSC samples, minus the batch correction step as we excluded 18 samples from the Richards et al. study that displayed large batch effects. A log2 FPKM matrix was created as well for this 54 GSC dataset.

3.5.2 DNA Methylation

Bisulfite conversion and Illumina EPIC array experiments were performed as described by Guilhamon and colleagues[189] for 54 GSC samples. Data preprocessing was performed using the ChaMP package (v2.6.4)[209] as described by Guilhamon and colleagues[189]. Processed beta values were transformed to m-values, and average m-values were calculated for gene promoters using annotations from gencode-v12[210] to map probes to ENST identifiers where the probe location was within 200 bp of the transcription start site and using BioMart[211,212] to map ENST identifiers to HGNC symbols from Ensembl[213] for the GRCh38 human genome assembly (ENSEMBL 103). M-values (averaged for each probe for a promoter) were used for integrated PCA and correlation analyses.

3.5.3 Whole Genome Sequencing

Whole genome sequencing for matched GSC and patient blood samples was performed as described by Richards and colleagues[188]. CNV calls were performed as described by Richards and colleagues[188]. For SNV calls, SNVs were kept if 3 or more out of 5 callers (Strelka[27] v 1.0.14, VarScan2[31] v2.3.8, Mutect2[28] from GATK v3.8, MuSE[29] 1.0rc, and multiSNV[30] 2.3-12) reported the SNV. To generate a total count of SNVs for each GSC line, the total number of nonsynonymous, stopgain, or stoploss events detected for a GSC line (i.e. change of amino acid, truncation, or extension of polypeptide encoded). Due to the sparse nature of SNVs, and the desire to find pathway-level signal for mutations, for the SNV data input to KNN imputation and PCA, we performed pathway level aggregation of signal, using the same GMT file described in the *Pathway Analysis* section. To do this, we summed the number of times a gene in a pathway had one or more non-synonymous, stopgain, or stoploss mutations.

Noticing that there were only a handful of samples with high mutation frequency, we decided to divide samples by total SNV signal (defined as total number of nonsynonymous, stopgain, or stoploss mutations) using k-means clustering (k = 2).

For mutational signature scoring, we used the mutSignatures package (v 2.1.1) and COSMIC v3.2 SBS signatures (available at https://cancer.sanger.ac.uk/signatures/documents/453/COSMIC_v3.2_SBS_GRCh38.txt)

3.5.4 ATAC-seq

ATAC-seq experiments and peak calling were performed as described by Guilhamon et al.[189], with the exception that reads were aligned to hg38. In initial preprocessing steps, we used the union of 40 of 54 GSCs used for PCA that had non-imputed ATAC-seq data and the 26 samples available from Guilhamon and colleagues' work[189]. The bedtools software (v2.29.2) was used to create a merged set of peaks (bedtools merge) across all samples, and peak intensities for each merged peak for a given sample was the average of the sample's peaks occurring within the area of the merged peak. Features present in 50% or more samples (from PCA set/Guilhamon paper union) were kept for further analysis. Log-transformed ATAC-seq signal from this matrix was used for the integrated PCA (described in *Data Integration*).

To find differentially accessible regions, we used the 40 samples for which ATAC-seq peak calls were made and performed Fisher's Exact Test to assess overrepresentation of called peaks in Developmental (n = 26) and Injury Response (n = 14) GSCs. Differentially open regions were filtered at FDR 0.20, and fed through HOMER[203] (v 4.11.1) for motif enrichment analysis, with a background set defined as all called ATAC-seq peaks (n = 410316) in the set of the 40 samples (not just those present in 50% or more samples). Enriched motifs were filtered to have a ratio >= 2 of percent target region hits to percent background region hits.

For pathway analysis on enriched open regions, we obtained all genes with a transcription start site inside an enriched open region or within 2000 bp downstream of an enriched region. We repeated this process on the background set of open regions to obtain a background set of genes to use for pathway analysis.

To assess how our results related to those of Guilhamon et al.[189] we developed three signatures (one each for the Constructive, Reactive, and Invasive states) by using limma (v3.38.3) to find differentially accessible regions between the three clusters (FDR 0.10). 9 of 26 samples available from this study were present in the 54 samples used for multi-omics PCA. To classify the 45 unclassified GSC lines, we used the combination of imputed and non-imputed ATAC-seq data (see *Data Imputation and Integration* for more details) for these samples as well as the data for the 26 GSC labeled lines from the Guilhamon publication as input for GSVA (v 1.3.0) signature scoring. We then trained a KNN classifier (scikit-learn, v 0.22.2, ported through reticulate v 1.15) on the 26 labeled GSC lines and used that classifier to predict the label of all 45 unlabeled GSC lines.

3.5.5 miRNA

Samples were processed and count data generated as described by Chu and colleagues[61]. Once counts were obtained, data was VST transformed and normalized across cohorts using ComBAT[176] using the Surrogate Variable Analysis (sva) package (v 3.30.1). This data was used as input for the integrated PCA. Since the Developmental/Injury Response axis accounted for the strongest signal in miRNA data, we looked for miRNAs that varied along this axis as well as candidate interactors. To do this, we performed differential miRNA expression using DESeq2[45] (v 1.22.2) with miRNA count data to obtain differentially expressed miRNAs. We then correlated VST transformed miRNA data with log-transformed RNA FPKM data, and filtered anticorrelation relationships through DIANA-TarBase v8[205], using miRBase v22[214] accession/name mappings to use the most up to date names.

For pathway analysis, Fisher's exact test was used, with Developmental miRNA targets and Injury Response miRNA targets used as queries. The background set of genes used for this analysis was the set of genes present in the set of DIANA-TarBase interactions filtered for miRNAs present in our miRNA count matrix and genes present in the RNA log FPKM matrix.

To create the network shown in Figure 3.13, we filtered the set of putative miRNA/RNA interactions we found for the union of genes uniquely essential in Developmental or Injury Response GSCs.

For details on determining uniquely essential genes that putative miRNA targets were compared against, see *CRISPR Screens*.

3.5.6 Data Imputation and Integration

Data imputation was performed using K-nearest neighbors (KNN) based imputation (scikit-learn, v 0.22.2, ported through reticulate v 1.15) using all available features for the six data types present (Figure 3.2). To prevent scale of features and number of features from a given data type from skewing the imputation (i.e. to give each data type equal weighting in imputation), we zero centered all features and scaled each centered feature f for data type d by scale factor c_f = 1/(s_f * sqrt(p_d)), where s_f is the standard deviation of f and p_d is the number of features present in data type d. After this centering and scaling, KNN imputation was performed, with imputed data divided by the appropriate scale factor and added to the original feature mean to obtain our final imputed data.

For PCA, imputed data was filtered as follows: take top 2000 most variable features for each data type with the exception of pathway aggregated SNV signal and miRNA and then standard normalized prior to PCA. The subsetting of other genomic features was done to avoid a situation with data types which included an order of magnitude more features (ATAC-seq: 53267 features, CNVs: 25988 features, DNAm: 21482 features, RNA: 23005 features) drowning out signal from ones with fewer features – pathway aggregated SNV signal (5410 features) and miRNA signal (1729 features).

3.5.7 CRISPR Screens

The dataset of quantile-normalized Bayes-factor (qBF) scores for essentiality was the same used by Richards and colleagues[188]. Genes were determined to be essential for a group of GSC lines if the median qBF was greater than 10, and genes were determined to be uniquely essential for that group if they were essential for that group and no other group of cell lines. We identified uniquely essential genes for both Developmental (n = 4 lines) and Injury Response (n = 5 lines) GSCs in this manner.

3.5.8 Network Visualization

All biological networks were visualized using Cytoscape (v3.7.2)[182]. Enrichment Map (v3.1)[181] and AutoAnnotate (v1.3.2)[180] were used to visualize pathway enrichment analysis results.

3.5.9 Xenograft Experiments

Xenograft experiments were performed and survival times determined as described by Richards and colleagues[188], using a subset (n = 32) of data used for that publication.

3.5.10 Pathway Analysis

For all pathway analyses described in this paper, the following GMT file was used:

Human_GO_AllPathways_no_GO_iea_April_01_2018_symbol.gmt, available at the URL:

http://download.baderlab.org/EM_Genesets/April_01_2018/Human/symbol/Human_GO_AllPathways_no_GO_iea_April_01_2018_symbol.gmt

3.5.11 Other Software

All statistical analyses were performed using R version 3.5.2, and, where python dependencies were used, python 3.6.9 from Anaconda was used.

3.6 Tables

Table 3.1 MSH6 mutation status in hypermutator and non-hypermutator GSCs

Shown is MSH6 mutation status (WT = wild type, MUT = nonsynonymous, stoploss, or stopgain mutation) in SNV_C1 (non-hypermutator) and SNV_C2 (hypermutator) samples. Overrepresentation in SNV_C2 was tested with a one-sided fisher's exact test (p = 3e-7).

	SNV_C1	SNV_C2
MUT	0	5
WT	49	0

Table 3.2 Recurrence status of GSCs' source tumors vs SNV cluster

SNV Cluster	Primary	Recurrent
1	44	5
2 (hypermutated)	0	5

Table 3.3 Comparison of Guilhamon et al. 2021 chromatin based GSC states and Richards et al. 2020 Developmental and Injury Response transcriptional phenotypes

	Developmental	Injury Response
Constructive	30	2
Invasive	2	2
Reactive	4	14

Table 3.4 Name Mapping for Developmental Motifs

Homer Motif	Shortened Name
BORIS(Zf)/K562-CTCFL-ChIP-Seq(GSE32465)/Homer	BORIS(Zf)
CTCF(Zf)/CD4+-CTCF-ChIP-Seq(Barski_et_al.)/Homer	CTCF(Zf)
OCT:OCT(POU,Homeobox)/NPC-Brn1-ChIP-Seq(GSE35496)/Homer	OCT/POU (2)
OCT:OCT(POU,Homeobox)/NPC-OCT6-ChIP-Seq(GSE43916)/Homer	OCT/POU (1)
RFX(HTH)/K562-RFX3-ChIP-Seq(SRA012198)/Homer	RFX(HTH)
Rfx2(HTH)/LoVo-RFX2-ChIP-Seq(GSE49402)/Homer	Rfx2(HTH)
X-box(HTH)/NPC-H3K4me1-ChIP-Seq(GSE16256)/Homer	X-box(HTH)

Table 3.5 Name Mapping for Injury Response Motifs

Homer Motif	Shortened Name
AP-1(bZIP)/ThioMac-PU.1-ChIP-Seq(GSE21512)/Homer	AP-1(bZIP)
Atf3(bZIP)/GBM-ATF3-ChIP-Seq(GSE33912)/Homer	Atf3(bZIP)
Bach1(bZIP)/K562-Bach1-ChIP-Seq(GSE31477)/Homer	Bach1(bZIP)
Bach2(bZIP)/OCILy7-Bach2-ChIP-Seq(GSE44420)/Homer	Bach2(bZIP)
BATF(bZIP)/Th17-BATF-ChIP-Seq(GSE39756)/Homer	BATF(bZIP)
Fos(bZIP)/TSC-Fos-ChIP-Seq(GSE110950)/Homer	Fos(bZIP)
Fosl2(bZIP)/3T3L1-Fosl2-ChIP-Seq(GSE56872)/Homer	Fosl2(bZIP)
Fra1(bZIP)/BT549-Fra1-ChIP-Seq(GSE46166)/Homer	Fra1(bZIP)
Fra2(bZIP)/Striatum-Fra2-ChIP-Seq(GSE43429)/Homer	Fra2(bZIP)
ISRE(IRF)/ThioMac-LPS-Expression(GSE23622)/Homer	ISRE(IRF)
Jun-AP1(bZIP)/K562-cJun-ChIP-Seq(GSE31477)/Homer	Jun-AP1(bZIP)
JunB(bZIP)/DendriticCells-Junb-ChIP-Seq(GSE36099)/Homer	JunB(bZIP)
MafK(bZIP)/C2C12-MafK-ChIP-Seq(GSE36030)/Homer	MafK(bZIP)
NF-E2(bZIP)/K562-NFE2-ChIP-Seq(GSE31477)/Homer	NF-E2(bZIP)
NFE2L2(bZIP)/HepG2-NFE2L2-ChIP-Seq(Encode)/Homer	NFE2L2(bZIP)
NFkB-p65-Rel(RHD)/ThioMac-LPS-Expression(GSE23622)/Homer	NFkB-p65-Rel(RHD)
Nrf2(bZIP)/Lymphoblast-Nrf2-ChIP-Seq(GSE37589)/Homer	Nrf2(bZIP)
RAR:RXR(NR),DR5/ES-RAR-ChIP-Seq(GSE56893)/Homer	RAR
T1ISRE(IRF)/ThioMac-Ifnb-Expression/Homer	T1ISRE(IRF)
ZFP3(Zf)/HEK293-ZFP3.GFP-ChIP-Seq(GSE58341)/Homer	ZFP3(Zf)

Table 3.6 Name Mapping for PC4 Cluster 2 Motifs

Homer Motif	Shortened Name
Bach1(bZIP)/K562-Bach1-ChIP-Seq(GSE31477)/Homer	Bach1(bZIP)
Bach2(bZIP)/OCILy7-Bach2-ChIP-Seq(GSE44420)/Homer	Bach2(bZIP)
Fosl2(bZIP)/3T3L1-Fosl2-ChIP-Seq(GSE56872)/Homer	Fosl2(bZIP)
Fra2(bZIP)/Striatum-Fra2-ChIP-Seq(GSE43429)/Homer	Fra2(bZIP)
IRF2(IRF)/Erythroblas-IRF2-ChIP-Seq(GSE36985)/Homer	IRF2(IRF)
ISRE(IRF)/ThioMac-LPS-Expression(GSE23622)/Homer	ISRE(IRF)
Jun-AP1(bZIP)/K562-cJun-ChIP-Seq(GSE31477)/Homer	Jun-AP1(bZIP)
NF-E2(bZIP)/K562-NFE2-ChIP-Seq(GSE31477)/Homer	NF-E2(bZIP)
NFE2L2(bZIP)/HepG2-NFE2L2-ChIP-Seq(Encode)/Homer	NFE2L2(bZIP)
Nrf2(bZIP)/Lymphoblast-Nrf2-ChIP-Seq(GSE37589)/Homer	Nrf2(bZIP)

4 Chapter 4: Discussion

GBM is an aggressive and lethal disease, and previous attempts at linking genomic and non-genetic phenotypes to clinically relevant consequences have had limited success, with the exception of a handful of mutations and MGMT methylation status[81,84–88]. Given that glioma stem cells are responsible for the regeneration of tumors after therapy and presumably the formation of the primary tumor, they thus present an enticing treatment target to prevent recurrence in GBM. However, their heterogeneity leads to variable responses to individual therapies such as temozolomide and other small molecules[20]. It is thus imperative that we understand the underlying biology of this heterogeneity to understand why treatment so often fails, especially given that only 4 FDA approved drugs for treating GBM have been developed over the past 5 decades providing only marginal benefit to survival[215]. In this section, I discuss the implications of the work presented in this thesis towards addressing this area in the context of others' work, and propose future experiments that would logically follow to address new and remaining questions.

4.1 The Transcriptomic Landscape of GSCs and Their Relation to Bulk Tumors

The potential originating events of GBM and GSCs, as well as their mutational and clonal evolution have been known before our work[73,109]. Bulk RNA-sequencing data has also suggested plasticity between two transcriptional states, which we later interpreted in Chapter 2 as a Developmental and Injury Response[20,100,101] , and differential capacity for differentiation and radioresistance between these phenotypes. However, the distribution of GSC transcriptional phenotypes at a single cell level was previously unknown. Additionally, the relationships between different transcriptional phenotypes and differentiated tumors was still incomplete. Potential hypotheses for the landscape of transcriptional phenotype could include a singular apex cell with more differentiated progeny[216,217], or a variety of potential phenotypes that could enter a more differentiated state[109], with the potential for very distinct clusters of cells in this landscape or for a continuum of cell states. Beyond this, our understanding of the consequences of transcription on functional dependencies was still incomplete. Thus, there was a rationale for both performing scRNA-seq on patient derived GSC lines and patient tumors and performing functional characterization of GSCs based on transcriptomic state.

In Chapter 2, we performed experiments and analyses to address these open questions, and found that GSCs lie on a continuum between a Developmental and Injury Response axis and that GBM cells in general lie on a stem to astrocyte differentiation gradient orthogonal to this axis. We saw that Developmental and Injury Response GSCs could serve as the source of GBM tumors, and that position along the Developmental/Injury Response gradient affected functional dependencies in a manner consistent with the biological pathways associated with each transcriptional program. We additionally showed that the Developmental/Injury Response axis, while more pronounced in GSC cultures, was present in patient tumors as well, which harbored a stem-like fraction of cells that overlapped cultured GSC in the two-axis transcriptional space.

Interestingly, additional efforts to characterize the differentiation axis and hierarchy of GSCs have also shown that GSCs and the rest of the tumor exist along a continuum of differentiation, in which an Injury Response/mesenchymal phenotype resides within a subset of the GSC portion. In work by Castellan and colleagues, a bifurcation event between a more astrocytic phenotype and a more oligodendrocytic phenotype was found to be preceded by a stem-like state[113]. Interestingly, they found that in the stem-like state (which they termed G-STEM), cells exhibited a more mesenchymal phenotype, suggesting that the pool of GSCs in a patient's tumor exhibits phenotypes that are especially accentuated in Injury Response GSCs[113]. In all, our results, combined with recent results from others, support the presence of inflammatory phenotypes in addition to stem-like properties within the GSC compartment of GBM tumors.

4.2 Wound Healing as a Mechanism Behind GSC Formation

It will be very difficult to know exactly how a GBM forms due to its location in the brain and its early start of an estimated 1-7 years prior to diagnosis[73,218], though recurring founder mutations in genes such as EGFR, PTEN, and CDKN2A/B have been identified through lineage tracing work[73], and evidence exists for multiple neuronal and glial cell types as the cell of origin for GBM[9,104,105,126]. In particular, Lee and colleagues showed the existence of astrocyte-like stem cells in the sub-ventricular zone (SVZ) which in over half of patients studied shared driver mutations with GBM tumors (e.g. in EGFR, P53, PTEN) and which, if carrying those driver mutations, could initiated GBM tumors in mice[126]. Additionally, others have demonstrated the dedifferentiation of astrocytic cells to a stem like state upon injury[107,108]. Interestingly, there is evidence in other cancers (e.g. breast, lung, colon cancer) for inflammatory signaling and wound healing to be mechanisms by which cancer stem cells can be induced[219–223]. For instance, it was shown that in patient derived colon cancer cell lines, CD44 mediated STAT3 signaling upregulated the expression of stemness markers such as SOX2 and OCT4, and CD44 enhanced tumorigenicity of patient derived cell lines, in addition to promoting mesenchymal properties[222]. Additionally, in a lung cancer model of cancer stem cells, it was shown that expression of stem markers as well as tumorigenicity was considerably reduced upon inhibition of NFKB[221], while in cell line and xenograft models of breast cancer, breast cancer stem cell markers and mammosphere formation were upregulated upon chemotherapy in a TGF-beta dependent manner[223]. Thus, there is evidence both within and outside the context of the brain that injury related or inflammatory signaling could serve as a means by which cells with an appropriate mutational background could acquire a stem-like, tumorigenic, phenotype. In the context of GBM, this mutational background would likely involve the mutation of genes such as EGFR, PTEN, and CDKN2A/B, with chromosomal instability leading to common copy number aberrations such as chromosome 7 amplification and/or 10 deletion. This background, combined with injury or inflammation induced gains in self-renewal/pluripotency capacity, could lead to the formation of cancer stem cells in GBM.

From the results presented in Chapter 2, we see that a large portion of GSCs adopt an Injury Response transcriptional phenotype. While they adopt this phenotype to a more extreme extent than seen in bulk tumors, we saw that in tumors themselves there was still variation with regards to which cells adopted the Developmental and Injury Response transcriptional phenotypes in a manner similar to that seen in GSCs (i.e. anticorrelation). Importantly, we showed the similarity of the Injury Response expression program to that of reactive astrocytes[142]. There is, however, an apparent discrepancy between the increased tumorigenicity of Developmental GSCs relative to Injury Response GSCs in our work and the apparent requirement of wound response related signaling pathways for the formation of stem-like cells in other cancers or in astrocytes. It may well be that in the context of glioblastoma, inflammatory signaling could result in de-differentiation, but excessive activity of this signaling could impair differentiation capacity (and from there tumorigenesis) as evidenced by results from Park and colleagues[111]. Indeed, there is already considerable evidence that inflammatory signaling mediates induction or maintenance of GSCs. As early as 2012, it was shown that IRF7 could induce stem-like characteristics and tumorigenicity in glioblastoma cell lines, and its loss resulted in a decrease in the ratio of GSCs to differentiated cells in patient derived GSC lines[224]. More recently, Gao and colleagues showed that TMZ can induce GSCs in a DAMP/TLR dependent fashion[110]. Overall, the results presented in this thesis, combined with prior work by others, suggests Injury Response related signaling as a mechanism by which a de-differentiated, stem-like state in GBM cells can be obtained. Our work in describing a Developmental/Injury response axis as the major axis of variation in GSCs and the similarity of the Injury Response phenotype to an existing wound response program broadens the conceptualization of GSC origination from a traditional neural stem cell hierarchy to that of plasticity and de-differentiation within adult neural tissue.

4.3 Mechanisms Behind GSC Phenotype Adoption

Beyond the origin of GSCs, the question arises of how GSCs branch off along relatively discrete phenotypes such as mismatch repair associated hypermutation and more continuous spectra such as the Developmental/Injury Response axis. This question, with respect to mismatch repair deficiency and hypermutation, appears to have been largely answered for GBM tumors and with GSC cultures[80,81,103] , with mismatch repair deficiency conferring resistance to temozolomide treatment, and a hypermutation phenotype upon temozolomide treatment likely accounting for increased immune invasion. In Chapter 3, we were able to partially address the question of how the Developmental and Injury Response transcriptional phenotypes come to bifurcate by identifying the coordinated actions of DNA methylation, miRNA based suppression, and (likely) transcription factor activity as processes that can stabilize the Developmental and Injury Response transcriptional states, with copy number variation likely driving GSCs towards one transcriptional pole or the other. The coordination of regulatory activities across a variety of biological processes suggests that normal processes to ensure a neural developmental or inflammatory signaling state are tightly regulated and are present in GSCs as pre-existing circuitry. The capacity of differentiated astrocytes to de-differentiate upon stimulation with inflammatory signaling or in response to stresses such as stab wounds supports this notion[107,108]. Moreover, it has been shown that miR-21, upregulated in Injury Response GSCs, is upregulated upon spinal cord injury near the site of a lesion concomitant with the expression of stem markers, suggesting a combination of stem and inflammation/epithelial mesenchymal signaling may be involved in tissue repair[225]. Interestingly we found gene expression independent components of variation in chromatin accessibility as well as promoter methylation, specifically pertaining to variable accessibility in regions enriched for Injury Response related transcription factor motifs and variable promoter methylation for genes pertaining to certain biological pathways upregulated in Injury Response GSCs. This result suggests that there may be variable underlying influences from transcription factor activity and epigenetic state on the Developmental/Injury Response axis that do not manifest into the transcriptional state we observe and are instead overshadowed by other influences such as DNA copy number and miRNAs. While this result does argue for potential latent epigenetic regulation that can be regulated, future work will need to further investigate this, as detailed in the Future Directions section. Overall, we managed to identify the previously characterized mismatch repair deficiency dependent hypermutation phenotype as well as many novel aspects of the coordinated action of multiple biological processes to regulate the Developmental/Injury Response transcriptional axis identified in Chapter 2.

4.4 Future Directions

4.4.1 Using Spatial Transcriptomics Data to Assess Immune/GSC Interactions

There is considerable evidence from recent studies that Injury Response-like GBM cells (GSCs or otherwise) are products of interaction with immune cells, both through *in vitro* cytokine treatment experiments done by ourselves and others[100], as well as through spatial transcriptomics data[226] and direct *in vivo* experiments[227]. Additionally, cells with an Injury Response phenotype are thought to in turn contribute to immunosuppression through driving cytotoxic T-cells to an exhausted state[226]. This raises the question of where GSCs fit into the immune interaction and suppression paradigm. Future studies to address this question could include the application of GSC gene expression signatures to spatial transcriptomics data in GBM tumors, and assessing the localization of GSCs relative to immune infiltrates, as well as more generally the spatial distribution of the Developmental and Injury Response transcriptional programs within the GSC and differentiated compartments of the tumor. Ideally, we would use a technique with true single cell resolution such as Seq-FISH[228] or the more recently developed sci-Space method[229], though the lack of single cell resolution in spot based methods such as the 10X Visium assay[230,231] can be dealt with (albeit imperfectly) using deconvolution techniques[232]. Under the wound response hypothesis, we would expect at least a large subset of GSCs to be spatially close to infiltrating immune cells. Indeed, it has already been shown that GSCs can interact with immune cells as well as blood vessels[7]. Overall, further understanding of the place of GSCs within tumors will require understanding of their interactions with both the rest of the tumor and the tumor microenvironment.

4.4.2 Mechanistically Investigating the Wound Healing Hypothesis

The results presented in this thesis and in other work suggest that wound response is a natural mechanism by which dedifferentiation of cells into cancer stem cells can take place, particularly in the context of GBM. Future experiments should look to go beyond initial *in vitro* and xenograft based results for GBM and attempt to recapitulate the early events of tumorigenesis in this context, perhaps with co-culture (with 2D culture or 3D organoids) and experiments involving macrophages and differentiated neural/glial cell types with founder mutations such as chromosome 7 amplification and/or chromosome 10 deletion, with xenograft transplantation as well as scRNA-seq assays performed. Through these experiments, we can assess if wound response signaling in the context of oncogenic mutations is sufficient to induce de-differentiation and tumorigenicity. If the co-culture based experiments provide evidence for this hypothesis, scRNA-seq experiments could also be performed with genetically engineered mice harboring such mutations in the neural, oligodendrocytic, or astrocytic compartments of the brain, with stab wound injury, middle cerebral artery occlusion, or cytokine treatment used to induce inflammatory signaling based de-differentiation, with rates of tumor formation compared in between injured and non-injured mice. Overall, these proposed experiments would aim to mechanistically assess the hypothesis that wound response can be responsible for tumor origination and if this can be shown *in vivo*.

4.4.3 Finding the Origin of GSC Heterogeneity

What our results in Chapter 3 do not answer concerns earlier origination of the steady state data we see in our GSC lines. For example, does epigenetic and genetic state serve as an origin event driving GSCs to one transcriptional state or the other? Does epigenetic state act as a stabilizer of transcriptional state preceded by environmental influences on transcription? Is copy number variation an early contributor to the transcriptional phenotype adopted by GSCs, and is it a trait selected for under microenvironmental conditions pushing GSCs one way or the other, given the differential essentiality of biological pathways such as aerobic/anaerobic respiration and immune related signaling? Potential future experiments to address these questions could include performing the aforementioned co-culture experiments (assuming we are able to induce de-differentiation and tumorigenic activity this way) with and without treatment of cytokines known to provoke a Developmental to Injury Response transition (e.g. OMS[227], TNF-alpha[100]), or with transient overexpression of Developmental transcription factors such as SOX2 and NOTCH, and the profiling of treated and untreated cell lines over time for RNA-seq, ATAC-seq, DNA methylation, WGS, and miRNA. This would allow us to assess the degree to which pre-existing genetic and non-genetic potentiation (i.e. from the start of tumorigenesis) influence transcriptional phenotype, and assess if there is selective activity upon copy number variation depending on transcriptional state over time as well as microenvironmental cues.

4.4.4 Characterizing stability of Developmental and Injury Response Transcriptional States Given Genetic and Epigenetic Influences

It was shown in Chapter 2 that GSC transcriptional state could be plastic in response to cytokines, and in Chapter 3 that there was apparent latent variation in methylation for promoters of inflammatory signaling related genes, as well variable chromatin accessibility in regions enriched for Injury Response transcription factor binding motifs. This raises the question of how robust the Developmental and Injury Response transcriptional states are to perturbation given influences such as genetic background (i.e. copy number for a variety of Developmental and Injury Response genes) and epigenetic state. Experiments that could be done to address this could include overexpression of miRNAs or transcription factors that could potentially change transcriptional state from Developmental to Injury Response (or vice versa) and the examination of both the transcriptome and the epigenome over time. While some of the experiments proposed in this subsection might be similar to those proposed in the previous subsection (e.g. assessing GSCs' response to cytokines as a function of prior genetic/epigenetic state), I would be very interested in assessing the plasticity of transcriptional and epigenetic state along the Developmental and Injury response axis once established (i.e. well after any 'origination' event), meaning that these experiments could be performed with patient derived GSC lines derived as was done in this thesis (i.e. without attempting to de-differentiate neuronal/glial cell types). This work has to some extent been started, with Park and colleagues overexpressing ASCL1 (a Developmental transcription factor) in ASCL1 KO cells (which failed to differentiated upon Notch inhibition, similar to ASCL1 low cells that resembled Injury Response GSCs) and finding that genes involved in neural differentiation have their chromatin opened and an increase in cell differentiation occurs under these conditions[111]. Given our richer understanding since this work of both genetic and non-genetic influences on the Developmental and Injury Response state, it will be interesting to see to what extent influences from DNA methylation and chromatin accessibility can be overridden by other influences such as miRNA or transcription factor activity, how much DNA methylation and chromatin accessibility change in response to overexpression of those items, and how quickly.

4.4.5 Novel Avenues for Potential Therapeutics

With a mechanistic characterization of how the Developmental and Injury Response programs are maintained as well as their functional importance, we have opened the door to the development of new therapeutic targets and methods, particularly with the potential for miRNAs to be used to target differentially essential genes (e.g. FOSL1, OLIG2). Future experiments should be conducted to follow up the results regarding differentially expressed miRNAs targeting differentially essential genes among Developmental/Injury Response GSCs. Particularly, an experiment that should be attempted is the overexpression in Injury Response GSCs of Developmental miRNAs targeting Injury Response essential genes, as well as the overexpression in Developmental GSCs of Injury Response miRNAs targeting Developmental essential genes. Additionally, inhibitors of differentially essential biological pathways such as inflammatory signaling and the NuRD complex could be used to assess differential efficacy among compounds in Developmental and Injury Response GSCs. Indeed, in the context of breast cancer, others have used antibodies to inhibit STAT3 signaling by targeting OSMR[233], recently shown to be responsible for mesenchymal conversion in GBM[227], and HDAC inhibitors have already been shown to reduce proliferation in stem-like glioblastoma cell lines[234]. In addition to single agent experiments, combination therapies could be assessed as well, due to both the functional heterogeneity we have described in this thesis and the likelihood that aberrantly used wound response and neural developmental programs have built-in redundancies to handle perturbations. Overall, the greater understanding of the Developmental/Injury Response axis gleaned from the work presented in this thesis has allowed for new hypotheses on novel therapeutics.

Taking a broader look at actual and potential therapeutics applied to GBM, in some cases applied specifically to GSCs, it is clear that single agents are insufficient to treat GBM. An obvious example concerns the case of mismatch repair deficiency and resulting resistance to temozolomide treatment, with resulting hypermutated tumors displaying worse response to immunotherapy than to systemic agents[81]. More specifically applied to GSCs, a multi -omics study that matched transcriptomic, methylomic, and mutation data in GSC lines similar to those used in this thesis found that out of the 94 compounds they identified as variably effective (out of an initial screen of 1544 compounds of varying mechanism of action), the major identifiable component of variation in drug response corresponded to variable response to proteasome inhibitors depending on activity in the P53 pathway[102], showing an entire class of compounds to be less effective in a large subset of GSCs compared to sensitive GSCs. Applying cocktails of therapeutics designed to be effective against large, partially overlapping, subsets of GSCs could thus represent a strategy to destroy or inhibit growth for as much of the GSC population as possible. For example, given that mismatch repair deficiency dependent temozolomide resistance and is largely orthogonal to the Developmental/Injury Response axis, one could imagine a cocktail of temozolomide paired with specific targeting of Developmental and Injury Response GSCs. Indeed, there have already been numerous studies involving various potential combination therapies showing increased efficacy relative to monotherapy in cell culture or xenograft models, such as combination treatments involving temozolomide, receptor tyrosine kinases, and radiotherapy[235]. Likewise, there is a clear rationale for future studies to incorporate results from mechanistic and functional studies of GSC heterogeneity into the search for combination treatments.

4.5 Concluding Remarks

Glioblastoma is a lethal disease, with GSCs responsible for recurrence. Our understanding of GSCs has shifted from that of a relatively homogenous population to that of a subset of transcriptional space that is associated with the ability to initiate or repopulate tumors, with heterogeneity with respect to drug response posing a considerable challenge towards developing new therapies. Prior to this project, there remained questions regarding the nature of transcriptional heterogeneity of GSCs at the single cell level as well as their relation to the bulk tumor. Additionally, there was a lack of understanding of how biological heterogeneity beyond transcription in GSCs manifested in terms of coordinated biological processes and independent sources of heterogeneity. In the work presented in this thesis, I addressed these questions by performing analyses on RNA-seq , CRISPR screening, and scRNA-seq data from GSC cultures as well as bulk patient tumors, and performing an integrated multi -omics analysis of transcriptional, genomic, and epigenetic data. With the work presented in Chapter 2, we discovered that at the single cell level, GSCs lie on a continuum between the Developmental and Injury Response transcriptional programs, an axis which was orthogonal to a stem-astrocyte differentiation axis. I further validated and characterized the Developmental/Injury Response Axis using an expanded bulk RNA-seq dataset of GSCs. We found that in addition to GSCs, bulk tumors also exhibited a continuum of expression between the Developmental and Injury Response axis. Importantly, this transcriptional axis could be linked to differential functional dependencies, which carries implications for the development of combination therapies against the GSC fraction of GBM tumors. With the multi -omics characterization of GSC heterogeneity presented in Chapter 3, I found that GSCs could largely be partitioned among a hypermutation axis pertaining to mismatch repair deficiency and temozolomide treatment as well as a coordinated multi -omics axis pertaining to the Developmental and Injury Response transcriptional programs and the genetic and non-genetic influences on the transcriptional state a cell adopts. The discovery of the multi -omics axis gives insight as to the fixed (genetic) and active (non-genetic) means by which a GSC might tend towards a transcriptional state and how that state is maintained. A particularly exciting part of these results was the discovery of differentially expressed miRNAs among Developmental and Injury Response GSC lines that target genes that are upregulated in and exclusively essential in cell lines expressing the opposite transcriptional program (Developmental is opposite to Injury Response, as these transcriptional programs are highly anticorrelated). Overall, these results give us a fuller picture of the heterogeneity of GSCs, the underlying mechanisms behind that heterogeneity, and a larger repertoire of handles by which we can target therapies at multiple subpopulations of GSCs.

5 Appendix

NOTE: All tables described in appendix are not displayed here, but included in a supplementary file appendix_files.zip

Table 5.1 Nonsynonymous Mutation Status in Genes, Ranked by Frequency in SNV C1 (Non-Hypermutator) GSCs

Listed are gene, frequency of having >=1 nonsynonymous mutation in an SNV C1 GSC, and, where available, frequency of mutation according to TCGA[4] (downloaded from https://www.cbioportal.org/study/summary?id=gbm_tcga_pub2013)

Table 5.2: Gene-Wise Copy Number Statistics

Columns: Gene.Symbol: Gene symbol; Cytoband: cytoband; chr.arm: chromosome arm; chr: chromosome; arm: arm; avg_call: mean of amplification and deletion calls across all GSC lines with CNA data present, for a matrix with 1 = amplification, 0 = no amplification or deletion, -1 = deletion. CNV calls were thresholded at 0.30 log2(tumor/blood) for amplification and -0.30 log2(tumor/blood) for deletion; amp_freq: amplification frequency, for call threshold as described frequency; del_freq: deletion frequency, for call threshold as described previously.

Table 5.3: Enriched Pathways for Developmental Genes with CNV Association

Developmental genes with CNV association (association with Injury Response GSVA score < 0, FDR < 0.05; association with CNV signal > 0, FDR < 0.05) underwent pathway analysis using a fisher's exact test for overrepresentation, at FDR 0.05.

Table 5.4: Enriched Pathways for Injury Response Genes with CNV Association

Developmental genes with CNV association (association with Injury Response GSVA score > 0, FDR < 0.05; association with CNV signal > 0, FDR < 0.05) underwent pathway analysis using a fisher's exact test for overrepresentation, at FDR 0.05.

Table 5.5: Pathways Enriched in Regions More Accessible in Developmental GSC lines

ATAC-seq open chromatin regions were mapped to HGNC symbols as described in *Methods*, and resulting HGNC symbols were used for a fisher's exact test for overrepresentation. Pathways were filtered at FDR 0.10.

Table 5.6: DIANA-TarBase miRNA/RNA interactions matched to anticorrelation relationships in GSC data, Developmental upregulated miRNAs

Table 5.7: DIANA-TarBase miRNA/RNA interactions matched to anticorrelation relationships in GSC data, Injury Response upregulated miRNAs

Table 5.8: Genes Uniquely Essential in Developmental GSCs

See methods in Chapter 3 for details.

Table 5.9: Genes Uniquely Essential in Injury Response GSCs

See methods in Chapter 3 for details.

6 References

1. Stupp, R. *et al.* Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *N. Engl. J. Med.* **352**, 987–996 (2005).

2. Scott, J., Tsai, Y.-Y., Chinnaiyan, P. & Yu, H.-H. M. Effectiveness of radiotherapy for elderly patients with glioblastoma. *Int. J. Radiat. Oncol. Biol. Phys.* **81**, 206–210 (2011).

3. Zinn, P. O., Colen, R. R., Kasper, E. M. & Burkhardt, J.-K. Extent of resection and radiotherapy in GBM: A 1973 to 2007 surveillance, epidemiology and end results analysis of 21,783 patients. *Int. J. Oncol.* **42**, 929–934 (2013).

4. Brennan, C. W. *et al.* The somatic genomic landscape of glioblastoma. *Cell* **155**, 462–477 (2013).

5. McKinnon, C., Nandhabalan, M., Murray, S. A. & Plaha, P. Glioblastoma: clinical presentation, diagnosis, and management. *BMJ* **374**, n1560 (2021).

6. Monteiro, A. R., Hill, R., Pilkington, G. J. & Madureira, P. A. The role of hypoxia in glioblastoma invasion. *Cells* **6**, (2017).

7. Hambardzumyan, D., Gutmann, D. H. & Kettenmann, H. The role of microglia and macrophages in glioma maintenance and progression. *Nat. Neurosci.* **19**, 20–27 (2016).

8. Oronsky, B., Reid, T. R., Oronsky, A., Sandhu, N. & Knox, S. J. A review of newly diagnosed glioblastoma. *Front. Oncol.* **10**, 574012 (2020).

9. Zong, H., Parada, L. F. & Baker, S. J. Cell of origin for malignant gliomas and its implication in therapeutic development. *Cold Spring Harb. Perspect. Biol.* **7**, 1–13 (2015).

10. Wen, P. Y. *et al.* Glioblastoma in adults: a Society for Neuro-Oncology (SNO) and European Society of Neuro-Oncology (EANO) consensus review on current management and future directions. *Neuro Oncol.* **22**, 1073–1113 (2020).

11. Nicolas, S., Abdellatef, S., Haddad, M. A., Fakhoury, I. & El-Sibai, M. Hypoxia and EGF Stimulation Regulate VEGF Expression in Human Glioblastoma Multiforme (GBM) Cells by Differential Regulation of the PI3K/Rho-GTPase and MAPK Pathways. *Cells* **8**, (2019).

12. Gill, B. J. *et al.* MRI-localized biopsies reveal subtype-specific differences in molecular and cellular composition at the margins of glioblastoma. *Proc Natl Acad Sci USA* **111**, 12550–12555 (2014).

13. Verhaak, R. G. W. *et al.* Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* **17**, 98–110 (2010).

14. Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396–1401 (2014).

15. Neftel, C. *et al.* An integrative model of cellular states, plasticity, and genetics for glioblastoma. *Cell* **178**, 835-849.e21 (2019).

16. Stupp, R. *et al.* Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase III study: 5-year analysis of the EORTC-NCIC trial. *Lancet Oncol.* **10**, 459–466 (2009).

17. Chen, J. *et al.* A restricted cell population propagates glioblastoma growth after chemotherapy. *Nature* **488**, 522–526 (2012).

18. Singh, S. K. *et al.* Identification of human brain tumour initiating cells. *Nature* **432**, 396–401 (2004).

19. Bao, S. *et al.* Glioma stem cells promote radioresistance by preferential activation of the DNA damage response. *Nature* **444**, 756–760 (2006).

20. Segerman, A. *et al.* Clonal Variation in Drug and Radiation Response among Glioma-Initiating Cells Is Linked to Proneural-Mesenchymal Transition. *Cell Rep.* **17**, 2994–3009 (2016).

21. Mavaddat, N. *et al.* Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *Am. J. Hum. Genet.* **104**, 21–34 (2019).

22. Marzec, J. *et al.* The transcriptomic landscape of prostate cancer development and progression: an integrative analysis. *Cancers (Basel)* **13**, (2021).

23. Ng, P. C. & Kirkness, E. F. Whole genome sequencing. *Methods Mol. Biol.* **628**, 215–226 (2010).

24. Xu, C. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Comput. Struct. Biotechnol. J.* **16**, 15–24 (2018).

25. Whitford, W., Lehnert, K., Snell, R. G. & Jacobsen, J. C. Evaluation of the performance of copy number variant prediction tools for the detection of deletions from whole genome sequencing data. *J. Biomed. Inform.* **94**, 103174 (2019).

26. Gu, W., Zhang, F. & Lupski, J. R. Mechanisms for human genomic rearrangements. *Pathogenetics* **1**, 4 (2008).

27. Saunders, C. T. *et al.* Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* **28**, 1811–1817 (2012).

28. Benjamin, D. I. *et al.* Calling Somatic SNVs and Indels with Mutect2. *BioRxiv* (2019) doi:10.1101/861054.

29. Fan, Y. *et al.* MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biol.* **17**, 178 (2016).

30. Josephidou, M., Lynch, A. G. & Tavaré, S. multiSNV: a probabilistic approach for improving detection of somatic point mutations from multiple related tumour samples. *Nucleic Acids Res.* **43**, e61 (2015).

31. Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).

32. Nilsen, G. *et al.* Copynumber: Efficient algorithms for single- and multi-track copy number segmentation. *BMC Genomics* **13**, 591 (2012).

33. Favero, F. *et al.* Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann. Oncol.* **26**, 64–70 (2015).

34. 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).

35. Shao, X. *et al.* Copy number variation is highly correlated with differential gene expression: a pan-cancer study. *BMC Med. Genet.* **20**, 175 (2019).

36. Pös, O. *et al.* DNA copy number variation: Main characteristics, evolutionary significance, and pathological aspects. *Biomed. J.* **44**, 548–559 (2021).

37. Fanciulli, M., Petretto, E. & Aitman, T. J. Gene copy number variation and common human disease. *Clin. Genet.* **77**, 201–213 (2010).

38. LaFramboise, T. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Res.* **37**, 4181–4193 (2009).

39. Shinawi, M. & Cheung, S. W. The array CGH and its clinical applications. *Drug Discov. Today* **13**, 760–770 (2008).

40. Colli, L. M. *et al.* Burden of Nonsynonymous Mutations among TCGA Cancers and Candidate Immune Checkpoint Inhibitor Responses. *Cancer Res.* **76**, 3767–3772 (2016).

41. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).

42. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).

43. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

44. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).

45. 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).

46. Chan-Seng-Yue, M. *et al.* Transcription phenotypes of pancreatic cancer are driven by genomic events during tumor evolution. *Nat. Genet.* **52**, 231–240 (2020).

47. Chang, H.-H., Dreyfuss, J. M. & Ramoni, M. F. A transcriptional network signature characterizes lung cancer subtypes. *Cancer* **117**, 353–360 (2011).

48. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).

49. Yan, F., Powell, D. R., Curtis, D. J. & Wong, N. C. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biol.* **21**, 22 (2020).

50. Corces, M. R. *et al.* The chromatin accessibility landscape of primary human cancers. *Science* **362**, (2018).

51. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **13**, 484–492 (2012).

52. Kribelbauer, J. F. *et al.* Quantitative analysis of the DNA methylation sensitivity of transcription factor complexes. *Cell Rep.* **19**, 2383–2395 (2017).

53. Shafi, A., Mitrea, C., Nguyen, T. & Draghici, S. A survey of the approaches for identifying differential methylation using bisulfite sequencing data. *Brief. Bioinformatics* **19**, 737–753 (2018).

54. Pidsley, R. *et al.* Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol.* **17**, 208 (2016).

55. Zhao, L. *et al.* Integrated bioinformatics analysis of DNA methylation biomarkers in thyroid cancer based on TCGA database. *Biochem. Genet.* **60**, 629–639 (2022).

56. de Almeida, B. P., Apolónio, J. D., Binnie, A. & Castelo-Branco, P. Roadmap of DNA methylation in breast cancer identifies novel prognostic biomarkers. *BMC Cancer* **19**, 219 (2019).

57. Capper, D. *et al.* DNA methylation-based classification of central nervous system tumours. *Nature* **555**, 469–474 (2018).

58. Jurmeister, P. *et al.* Machine learning analysis of DNA methylation profiles distinguishes primary lung squamous cell carcinomas from head and neck metastases. *Sci. Transl. Med.* **11**, (2019).

59. Inoue, J. & Inazawa, J. Cancer-associated miRNAs and their therapeutic potential. *J. Hum. Genet.* **66**, 937–945 (2021).

60. Rupaimoole, R. & Slack, F. J. MicroRNA therapeutics: towards a new era for the management of cancer and other diseases. *Nat. Rev. Drug Discov.* **16**, 203–222 (2017).

61. Chu, A. *et al.* Large-scale profiling of microRNAs for The Cancer Genome Atlas. *Nucleic Acids Res.* **44**, e3 (2016).

62. Yu, N. *et al.* Identification of tumor suppressor miRNAs by integrative miRNA and mRNA sequencing of matched tumor-normal samples in lung adenocarcinoma. *Mol. Oncol.* **13**, 1356–1368 (2019).

63. Wang, T., Wei, J. J., Sabatini, D. M. & Lander, E. S. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**, 80–84 (2014).

64. Hart, T. *et al.* High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. *Cell* **163**, 1515–1526 (2015).

65. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).

66. Slyper, M. *et al.* A single-cell and single-nucleus RNA-Seq toolbox for fresh and frozen human tumors. *Nat. Med.* **26**, 792–802 (2020).

67. González-Silva, L., Quevedo, L. & Varela, I. Tumor Functional Heterogeneity Unraveled by scRNA-seq Technologies. *Trends Cancer* **6**, 13–19 (2020).

68. Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).

69. Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979–982 (2017).

70. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).

71. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494–498 (2018).

72. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.* **38**, 1408–1414 (2020).

73. Körber, V. *et al.* Evolutionary Trajectories of IDHWT Glioblastomas Reveal a Common Path of Early Tumorigenesis Instigated Years ahead of Initial Diagnosis. *Cancer Cell* **35**, 692-704.e12 (2019).

74. Dang, L., Jin, S. & Su, S. M. IDH mutations in glioma and acute myeloid leukemia. *Trends Mol. Med.* **16**, 387–397 (2010).

75. Ohgaki, H. & Kleihues, P. The definition of primary and secondary glioblastoma. *Clin. Cancer Res.* **19**, 764–772 (2013).

76. Yan, H. *et al.* IDH1 and IDH2 mutations in gliomas. *N. Engl. J. Med.* **360**, 765–773 (2009).

77. SongTao, Q. *et al.* IDH mutations predict longer survival and response to temozolomide in secondary glioblastoma. *Cancer Sci.* **103**, 269–273 (2012).

78. Maus, A. & Peters, G. J. Glutamate and α-ketoglutarate: key players in glioma metabolism. *Amino Acids* **49**, 21–32 (2017).

79. Louis, D. N. *et al.* The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. *Neuro Oncol.* **23**, 1231–1251 (2021).

80. Choi, S. *et al.* Temozolomide-associated hypermutation in gliomas. *Neuro Oncol.* **20**, 1300–1309 (2018).

81. Touat, M. *et al.* Mechanisms and therapeutic implications of hypermutation in gliomas. *Nature* **580**, 517–523 (2020).

82. Hunter, C. *et al.* A hypermutation phenotype and somatic MSH6 mutations in recurrent human malignant gliomas after alkylator chemotherapy. *Cancer Res.* **66**, 3987–3991 (2006).

83. Cahill, D. P. *et al.* Loss of the mismatch repair protein MSH6 in human glioblastomas is associated with tumor progression during temozolomide treatment. *Clin. Cancer Res.* **13**, 2038–2045 (2007).

84. Molenaar, R. J. *et al.* The combination of IDH1 mutations and MGMT methylation status predicts survival in glioblastoma better than either IDH1 or MGMT alone. *Neuro Oncol.* **16**, 1263–1273 (2014).

85. Hegi, M. E. *et al.* MGMT gene silencing and benefit from temozolomide in glioblastoma. *N. Engl. J. Med.* **352**, 997–1003 (2005).

86. Yang, P. *et al.* IDH mutation and MGMT promoter methylation in glioblastoma: results of a prospective registry. *Oncotarget* **6**, 40896–40906 (2015).

87. Delgado-López, P. D. & Corrales-García, E. M. Survival in glioblastoma: a review on the impact of treatment modalities. *Clin. Transl. Oncol.* **18**, 1062–1071 (2016).

88. Mikkelsen, V. E. *et al.* MGMT Promoter Methylation Status Is Not Related to Histological or Radiological Features in IDH Wild-type Glioblastomas. *J. Neuropathol. Exp. Neurol.* **79**, 855–862 (2020).

89. Messaoudi, K., Clavreul, A. & Lagarce, F. Toward an effective strategy in glioblastoma treatment. Part I: resistance mechanisms and strategies to overcome resistance of glioblastoma to temozolomide. *Drug Discov. Today* **20**, 899–905 (2015).

90. Wang, Q. *et al.* Tumor Evolution of Glioma-Intrinsic Gene Expression Subtypes Associates with Immunological Changes in the Microenvironment. *Cancer Cell* **33**, 152 (2018).

91. Huse, J. T., Phillips, H. S. & Brennan, C. W. Molecular subclassification of diffuse gliomas: seeing order in the chaos. *Glia* **59**, 1190–1199 (2011).

92. Wang, L. *et al.* The phenotypes of proliferating glioblastoma cells reside on a single axis of variation. *Cancer Discov.* **9**, 1708–1719 (2019).

93. Ma, H. *et al.* Specific glioblastoma multiforme prognostic-subtype distinctions based on DNA methylation patterns. *Cancer Gene Ther.* **27**, 702–714 (2020).

94. Klughammer, J. *et al.* The DNA methylation landscape of glioblastoma disease progression shows extensive heterogeneity in time and space. *Nat. Med.* **24**, 1611–1624 (2018).

95. Bonnet, D. & Dick, J. E. Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. *Nat. Med.* **3**, 730–737 (1997).

96. Lapidot, T. *et al.* A cell initiating human acute myeloid leukaemia after transplantation into SCID mice. *Nature* **367**, 645–648 (1994).

97. Singh, S. K. *et al.* Identification of a cancer stem cell in human brain tumors. *Cancer Res.* **63**, 5821–5828 (2003).

98. Stopschinski, B. E., Beier, C. P. & Beier, D. Glioblastoma cancer stem cells--from concept to clinical application. *Cancer Lett.* **338**, 32–40 (2013).

99. 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).

100. Bhat, K. P. L. *et al.* Mesenchymal differentiation mediated by NF-κB promotes radiation resistance in glioblastoma. *Cancer Cell* **24**, 331–346 (2013).

101. 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).

102. Johansson, P. *et al.* A Patient-Derived Cell Atlas Informs Precision Targeting of Glioblastoma. *Cell Rep.* **32**, 107897 (2020).

103. 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.e9 (2019).

104. Habib, A. *et al.* Letter: glioblastoma cell of origin. *Stem Cell Rev and Rep* **18**, 691–693 (2022).

105. Friedmann-Morvinski, D. *et al.* Dedifferentiation of neurons and astrocytes by oncogenes can induce gliomas in mice. *Science* **338**, 1080–1084 (2012).

106. Jiang, Y. *et al.* Glioblastoma cell malignancy and drug sensitivity are affected by the cell of origin. *Cell Rep.* **19**, 1080–1081 (2017).

107. Magnusson, J. P. *et al.* A latent neurogenic program in astrocytes regulated by Notch signaling in the mouse. *Science* **346**, 237–241 (2014).

108. Gabel, S. *et al.* Inflammation Promotes a Conversion of Astrocytes into Neural Progenitor Cells via NF-κB Activation. *Mol. Neurobiol.* **53**, 5041–5055 (2016).

109. Lan, X. *et al.* Fate mapping of human glioblastoma reveals an invariant stem cell hierarchy. *Nature* **549**, 227–232 (2017).

110. Gao, X.-Y. *et al.* Temozolomide treatment induces HMGB1 to promote the formation of glioma stem cells via the tlr2/neat1/wnt pathway in glioblastoma. *Front. Cell Dev. Biol.* **9**, 620883 (2021).

111. Park, N. I. *et al.* ASCL1 reorganizes chromatin to direct neuronal fate and suppress tumorigenicity of glioblastoma stem cells. *Cell Stem Cell* **21**, 411 (2017).

112. Rajakulendran, N. *et al.* Wnt and Notch signaling govern self-renewal and differentiation in a subset of human glioblastoma stem cells. *Genes Dev.* **33**, 498–510 (2019).

113. Castellan, M. *et al.* Single-cell analyses reveal YAP/TAZ as regulators of stemness and cell plasticity in Glioblastoma. *Nat. Cancer* **2**, 174–188 (2021).

114. Fang, X. *et al.* Inhibiting DNA-PK induces glioma stem cell differentiation and sensitizes glioblastoma to radiation in mice. *Sci. Transl. Med.* **13**, (2021).

115. Carlsson, S. K., Brothers, S. P. & Wahlestedt, C. Emerging treatment strategies for glioblastoma multiforme. *EMBO Mol. Med.* **6**, 1359–1370 (2014).

116. 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).

117. Bhaduri, A. *et al.* Outer Radial Glia-like Cancer Stem Cells Contribute to Heterogeneity of Glioblastoma. *Cell Stem Cell* **26**, 48-63.e6 (2020).

118. Berezovsky, A. D. *et al.* Sox2 promotes malignancy in glioblastoma by regulating plasticity and astrocytic differentiation. *Neoplasia* **16**, 193–206, 206.e19 (2014).

119. Natsume, A. *et al.* Chromatin regulator PRC2 is a key regulator of epigenetic plasticity in glioblastoma. *Cancer Res.* **73**, 4559–4570 (2013).

120. Liu, G. *et al.* Analysis of gene expression and chemoresistance of CD133+ cancer stem cells in glioblastoma. *Mol. Cancer* **5**, 67 (2006).

121. Kelly, J. J. P. *et al.* Proliferation of human glioblastoma stem cells occurs independently of exogenous mitogens. *Stem Cells* **27**, 1722–1733 (2009).

122. Florio, M. *et al.* Human-specific gene ARHGAP11B promotes basal progenitor amplification and neocortex expansion. *Science* **347**, 1465–1470 (2015).

123. 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).

124. 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).

125. 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).

126. Lee, J. H. *et al.* Human glioblastoma arises from subventricular zone cells with low-level driver mutations. *Nature* **560**, 243–247 (2018).

127. Filbin, M. G. *et al.* Developmental and oncogenic programs in H3K27M gliomas dissected by single-cell RNA-seq. *Science* **360**, 331–335 (2018).

128. Gojo, J. *et al.* Single-Cell RNA-Seq Reveals Cellular Hierarchies and Impaired Developmental Trajectories in Pediatric Ependymoma. *Cancer Cell* **38**, 44-59.e9 (2020).

129. Hovestadt, V. *et al.* Resolving medulloblastoma cellular architecture by single-cell genomics. *Nature* **572**, 74–79 (2019).

130. Izar, B. *et al.* A single-cell landscape of high-grade serous ovarian cancer. *Nat. Med.* **26**, 1271–1279 (2020).

131. Ledergor, G. *et al.* Single cell dissection of plasma cell heterogeneity in symptomatic and asymptomatic myeloma. *Nat. Med.* **24**, 1867–1876 (2018).

132. Puram, S. V. *et al.* Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell* **171**, 1611-1624.e24 (2017).

133. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).

134. Tirosh, I. *et al.* Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. *Nature* **539**, 309–313 (2016).

135. Ben-David, U. *et al.* Genetic and transcriptional evolution alters cancer cell line drug response. *Nature* **560**, 325–330 (2018).

136. Kinker, G. S. *et al.* Pan-cancer single-cell RNA-seq identifies recurring programs of cellular heterogeneity. *Nat. Genet.* **52**, 1208–1218 (2020).

137. Krieger, T. G. *et al.* Single-cell analysis of patient-derived PDAC organoids reveals cell state heterogeneity and a conserved developmental hierarchy. *Nat. Commun.* **12**, 5826 (2021).

138. 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).

139. Nowakowski, T. J. *et al.* Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science* **358**, 1318–1323 (2017).

140. Zhong, S. *et al.* A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex. *Nature* **555**, 524–528 (2018).

141. 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).

142. Liddelow, S. A. *et al.* Neurotoxic reactive astrocytes are induced by activated microglia. *Nature* **541**, 481–487 (2017).

143. John Lin, C.-C. *et al.* Identification of diverse astrocyte populations and their malignant analogs. *Nat. Neurosci.* **20**, 396–405 (2017).

144. Chai, H. *et al.* Neural Circuit-Specialized Astrocytes: Transcriptomic, Proteomic, Morphological, and Functional Evidence. *Neuron* **95**, 531-549.e9 (2017).

145. Morel, L. *et al.* Molecular and functional properties of regional astrocytes in the adult brain. *J. Neurosci.* **37**, 8706–8717 (2017).

146. Miller, S. J. Astrocyte heterogeneity in the adult central nervous system. *Front. Cell. Neurosci.* **12**, 401 (2018).

147. Haghverdi, L., Buettner, F. & Theis, F. J. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* **31**, 2989–2998 (2015).

148. 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).

149. Hart, T. *et al.* Evaluation and Design of Genome-Wide CRISPR/SpCas9 Knockout Screens. *G3 (Bethesda)* **7**, 2719–2727 (2017).

150. Hart, T. & Moffat, J. BAGEL: a computational framework for identifying essential genes from pooled library screens. *BMC Bioinformatics* **17**, 164 (2016).

151. 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).

152. Lein, E. S. *et al.* Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445**, 168–176 (2007).

153. 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).

154. Philip, B. *et al.* Mutant IDH1 promotes glioma formation in vivo. *Cell Rep.* **23**, 1553–1564 (2018).

155. Xie, Y. *et al.* The human glioblastoma cell culture resource: validated cell models representing all molecular subtypes. *EBioMedicine* **2**, 1351–1363 (2015).

156. Sirko, S. *et al.* Reactive glia in the injured brain acquire stem cell properties in response to sonic hedgehog. [corrected]. *Cell Stem Cell* **12**, 426–439 (2013).

157. Robel, S., Berninger, B. & Götz, M. The stem cell potential of glia: lessons from reactive gliosis. *Nat. Rev. Neurosci.* **12**, 88–104 (2011).

158. 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).

159. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202–1214 (2015).

160. Alles, J. *et al.* Cell fixation and preservation for droplet-based single-cell transcriptomics. *BMC Biol.* **15**, 44 (2017).

161. 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).

162. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

163. Mayer, C. *et al.* Developmental diversification of cortical inhibitory interneurons. *Nature* **555**, 457–462 (2018).

164. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13**, 845–848 (2016).

165. Innes, B. T. & Bader, G. D. scClustViz - Single-cell RNAseq cluster assessment and visualization. [version 2; peer review: 2 approved]. *F1000Res.* **7**, (2018).

166. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.* **2008**, P10008 (2008).

167. Ng, A., Jordan, M. & Weiss, Y. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems* **14**, (2001).

168. Charrad, M., Ghazzali, N., Boiteau, V. & Niknafs, A. NbClust: an R package for determining the relevant number of clusters in a data set. *Journal of statistical software* **61**, 1–36 (2014).

169. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

170. Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).

171. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).

172. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).

173. Yoshihara, K. *et al.* Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4**, 2612 (2013).

174. Müller, S., Cho, A., Liu, S. J., Lim, D. A. & Diaz, A. CONICS integrates scRNA-seq with DNA sequencing to map gene expression to tumor sub-clones. *Bioinformatics* **34**, 3217–3219 (2018).

175. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).

176. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882–883 (2012).

177. Karatzoglou, A., Smola, A., Hornik, K. & Zeileis, A. kernlab-an S4 package for kernel methods in R. *Journal of statistical software* **11**, 1–20 (2004).

178. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14**, 7 (2013).

179. 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).

180. Kucera, M., Isserlin, R., Arkhangorodsky, A. & Bader, G. D. AutoAnnotate: A Cytoscape app for summarizing networks with semantic annotations. [version 1; peer review: 2 approved]. *F1000Res.* **5**, 1717 (2016).

181. Merico, D., Isserlin, R., Stueker, O., Emili, A. & Bader, G. D. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS ONE* **5**, e13984 (2010).

182. Otasek, D., Morris, J. H., Bouças, J., Pico, A. R. & Demchak, B. Cytoscape Automation: empowering workflow-based network analysis. *Genome Biol.* **20**, 185 (2019).

183. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

184. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).

185. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **11**, 11.10.1-11.10.33 (2013).

186. Li, W. *et al.* MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biol.* **15**, 554 (2014).

187. Mohanraj, S. *et al.* Crescent: cancer single cell expression toolkit. *Nucleic Acids Res.* **48**, W372–W379 (2020).

188. Richards, L. M. *et al.* Gradient of Developmental and Injury Response transcriptional states defines functional vulnerabilities underpinning glioblastoma heterogeneity. *Nat. Cancer* **2**, 157–173 (2021).

189. Guilhamon, P. *et al.* Single-cell chromatin accessibility profiling of glioblastoma identifies an invasive cancer stem cell population associated with lower survival. *eLife* **10**, (2021).

190. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).

191. Shpak, M., Goldberg, M. M. & Cowperthwaite, M. C. Rapid and convergent evolution in the glioblastoma multiforme genome. *Genomics* **105**, 159–167 (2015).

192. Greenberg, M. V. C. & Bourc'his, D. The diverse roles of DNA methylation in mammalian development and disease. *Nat. Rev. Mol. Cell Biol.* **20**, 590–607 (2019).

193. Dominguez, M. H., Ayoub, A. E. & Rakic, P. POU-III transcription factors (Brn1, Brn2, and Oct6) influence neurogenesis, molecular identity, and migratory destination of upper-layer cells of the cerebral cortex. *Cereb. Cortex* **23**, 2632–2643 (2013).

194. Hsu, Y.-C., Kao, C.-Y., Chung, Y.-F., Chen, M.-S. & Chiu, I.-M. Ciliogenic RFX transcription factors regulate FGF1 gene promoter. *J. Cell. Biochem.* **113**, 2511–2522 (2012).

195. Carro, M. S. *et al.* The transcriptional network for mesenchymal transformation of brain tumours. *Nature* **463**, 318–325 (2010).

196. Andreolas, C., Kalogeropoulou, M., Voulgari, A. & Pintzas, A. Fra-1 regulates vimentin during Ha-RAS-induced epithelial mesenchymal transition in human colon carcinoma cells. *Int. J. Cancer* **122**, 1745–1756 (2008).

197. Diesch, J. *et al.* Widespread FRA1-dependent control of mesenchymal transdifferentiation programs in colorectal cancer cells. *PLoS ONE* **9**, e88950 (2014).

198. Lee, C.-K. & Bluyssen, H. A. R. Editorial: stats and irfs in innate immunity: from transcriptional regulators to therapeutic targets. *Front. Immunol.* **10**, 1829 (2019).

199. Arzate-Mejía, R. G., Recillas-Targa, F. & Corces, V. G. Developing in 3D: the role of CTCF in cell differentiation. *Development* **145**, (2018).

200. Rojo de la Vega, M., Chapman, E. & Zhang, D. D. NRF2 and the hallmarks of cancer. *Cancer Cell* **34**, 21–43 (2018).

201. Gasiorek, J. J. & Blank, V. Regulation and function of the NFE2 transcription factor in hematopoietic and non-hematopoietic cells. *Cell. Mol. Life Sci.* **72**, 2323–2335 (2015).

202. Zhang, D. *et al.* Involvement of a Transcription factor, Nfe2, in Breast Cancer Metastasis to Bone. *Cancers (Basel)* **12**, (2020).

203. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).

204. Li, S., Fang, X.-D., Wang, X.-Y. & Fei, B.-Y. Fos-like antigen 2 (FOSL2) promotes metastasis in colon cancer. *Exp. Cell Res.* **373**, 57–61 (2018).

205. Karagkouni, D. *et al.* DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA-gene interactions. *Nucleic Acids Res.* **46**, D239–D245 (2018).

206. Chen, M., Medarova, Z. & Moore, A. Role of microRNAs in glioblastoma. *Oncotarget* **12**, 1707–1723 (2021).

207. Sathyan, P. *et al.* Mir-21-Sox2 Axis Delineates Glioblastoma Subtypes with Prognostic Impact. *J. Neurosci.* **35**, 15097–15112 (2015).

208. Wang, N., Zhang, Y. & Liang, H. MicroRNA-598 Inhibits Cell Proliferation and Invasion of Glioblastoma by Directly Targeting Metastasis Associated in Colon Cancer-1 (MACC1). *Oncol. Res.* **26**, 1275–1283 (2018).

209. Morris, T. J. *et al.* Champ: 450k chip analysis methylation pipeline. *Bioinformatics* **30**, 428–430 (2014).

210. Frankish, A. *et al.* GENCODE 2021. *Nucleic Acids Res.* **49**, D916–D923 (2021).

211. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **4**, 1184–1191 (2009).

212. Durinck, S. *et al.* BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**, 3439–3440 (2005).

213. Cunningham, F. *et al.* Ensembl 2022. *Nucleic Acids Res.* **50**, D988–D995 (2022).

214. Kozomara, A., Birgaoanu, M. & Griffiths-Jones, S. miRBase: from microRNA sequences to function. *Nucleic Acids Res.* **47**, D155–D162 (2019).

215. Fisher, J. P. & Adamson, D. C. Current FDA-Approved Therapies for High-Grade Malignant Gliomas. *Biomedicines* **9**, (2021).

216. Couturier, C. P. *et al.* Single-cell RNA-seq reveals that glioblastoma recapitulates a normal neurodevelopmental hierarchy. *Nat. Commun.* **11**, 3406 (2020).

217. Suvà, M. L. *et al.* Reconstructing and reprogramming the tumor-propagating potential of glioblastoma stem-like cells. *Cell* **157**, 580–594 (2014).

218. Stensjøen, A. L., Berntsen, E. M., Jakola, A. S. & Solheim, O. When did the glioblastoma start growing, and how much time can be gained from surgical resection? A model based on the pattern of glioblastoma growth in vivo. *Clin. Neurol. Neurosurg.* **170**, 38–42 (2018).

219. Arnold, K. M., Opdenaker, L. M., Flynn, D. & Sims-Mourtada, J. Wound healing and cancer stem cells: inflammation as a driver of treatment resistance in breast cancer. *Cancer Growth Metastasis* **8**, 1–13 (2015).

220. Jin, W. Role of JAK/STAT3 Signaling in the Regulation of Metastasis, the Transition of Cancer Stem Cells, and Chemoresistance of Cancer by Epithelial-Mesenchymal Transition. *Cells* **9**, (2020).

221. Zakaria, N., Mohd Yusoff, N., Zakaria, Z., Widera, D. & Yahaya, B. H. Inhibition of NF-κB Signaling Reduces the Stemness Characteristics of Lung Cancer Stem Cells. *Front. Oncol.* **8**, 166 (2018).

222. Su, Y.-J., Lai, H.-M., Chang, Y.-W., Chen, G.-Y. & Lee, J.-L. Direct reprogramming of stem cell properties in colon cancer cells by CD44. *EMBO J.* **30**, 3186–3199 (2011).

223. Bhola, N. E. *et al.* TGF-β inhibition enhances chemotherapy action against triple-negative breast cancer. *J. Clin. Invest.* **123**, 1348–1358 (2013).

224. Jin, X. *et al.* Interferon regulatory factor 7 regulates glioma stem cells via interleukin-6 and Notch signalling. *Brain* **135**, 1055–1069 (2012).

225. Strickland, E. R. *et al.* MicroRNA dysregulation following spinal cord contusion: implications for neural plasticity and repair. *Neuroscience* **186**, 146–160 (2011).

226. Ravi, V. M. *et al.* T-cell dysfunction in the glioblastoma microenvironment is mediated by myeloid cells releasing interleukin-10. *Nat. Commun.* **13**, 925 (2022).

227. Hara, T. *et al.* Interactions between cancer cells and immune cells drive transitions to mesenchymal-like states in glioblastoma. *Cancer Cell* **39**, 779-792.e11 (2021).

228. Eng, C.-H. L. *et al.* Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature* **568**, 235–239 (2019).

229. Srivatsan, S. R. *et al.* Embryo-scale, single-cell spatial transcriptomics. *Science* **373**, 111–117 (2021).

230. Moses, L. & Pachter, L. Museum of spatial transcriptomics. *Nat. Methods* (2022) doi:10.1038/s41592-022-01409-2.

231. Dries, R. *et al.* Advances in spatial transcriptomic data analysis. *Genome Res.* **31**, 1706–1718 (2021).

232. Elosua-Bayes, M., Nieto, P., Mereu, E., Gut, I. & Heyn, H. SPOTlight: seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic Acids Res.* **49**, e50 (2021).

233. Geethadevi, A. *et al.* Oncostatin M Receptor-Targeted Antibodies Suppress STAT3 Signaling and Inhibit Ovarian Cancer Growth. *Cancer Res.* **81**, 5336–5352 (2021).

234. Was, H. *et al.* Histone deacetylase inhibitors exert anti-tumor effects on human adherent and stem-like glioma cells. *Clin. Epigenetics* **11**, 11 (2019).

235. Ghosh, D., Nandi, S. & Bhattacharjee, S. Combination therapy to checkmate Glioblastoma: clinical challenges and advances. *Clin. Transl. Med.* **7**, 33 (2018).