A COMPUTATIONAL FRAMEWORK FOR THE COMPARATIVE ANALYSIS OF GLIOMA MODELS AND PATIENTS

DISSERTATION

zur Erlangung des akademischen Grades

Doctor of Philosophy (Ph.D.)

eingereicht an der Lebenswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin
von

Juan Carlos Company Nevado

Präsidentin der Humboldt-Universität zu Berlin

Prof. Dr. Julia von Blumenthal

Dekan der Lebenswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin
Prof. Dr. dr. Christian Ulrichs

Gutachter/innen

1. **Prof. Dr. Uwe Ohler**
2. **Prof. Dr. David Capper**
3. **Dr. Matthias König**

Tag der mündlichen Prüfung: **25 of March , 2023**

TABLE OF CONTENTS

TABLE OF CONTENTS 3

PROJECT……. 7

ABSTRACT 7

ZUSAMMENFASSUNG 9

AIMS OF THE PROJECT 11

BACKGROUND 13

1. EPIDEMIOLOGY, ORIGIN, AND CLASSIFICATION OF DIFFUSE GLIOMAS 13

1.1 Epidemiology of diffuse gliomas in adults 13

1.2 Origin and growth of adult-type diffuse gliomas 17

1.3 Molecular classification of adult-type diffuse gliomas 20

2. COMPUTATIONAL METHODS TO CHARACTERIZE TUMOR MODELS 23

2.1 Methods to integrate the expression profiles of tumor models and patients 23

2.2 Methods to integrate the single-cell profiles of tumor models and patients 27

RESULTS……. 31

1. COMPARISON OF THE MOLECULAR PROFILES OF PATIENTS AND ADULT-TYPE DIFFUSE GLIOMA MODELS 31

1.1 Genetic characterization of Glioma subtype avatar models (GSA) 31

1.1.1 Complete genetic profiling of GSA models -------- 31

1.1.2 Characterization of the Copy-number alterations in GSA models -------- 34

1.2 Computational framework to compare the expression profiles of tumor models and patients 37

1.2.1 Comparison of avatar models and patient expression profiles -------- 37

1.2.2 Comparison using CAPE of GSA and glioblastoma patients reveal a PN subtype identity in the models -------- 39

1.2.3 Benchmark CAPE results using different clinical cohorts and computational methods -------- 45

1.3 CAPE analysis reveals specific differences between GSA models and glioblastoma patients ... 50

1.3.1 Using CAPE to characterize gene module differences between GSA and glioblastoma patients -------- 50

1.3.2 Evaluation of cluster-specific transcriptional regulators to improve GSA models -------- 53

2. USING CAPE TO ASSESS THE GLIOBLASTOMA STATE ACQUISITION IN GSA MODELS 58

2.1 GSA models acquired new glioblastoma states after engraftment 58

2.1.1 CAPE analysis revealed differences in GSA model expression profiles in vitro and in vivo -------- 58

2.1.2 Using CAPE and genetic tracing reporter to evaluate the PN-to-ME transition in GSA models in vivo ------ 61

2.2 Evaluation of the pn-to-me transition upon external factor activation using GSA models 66

2.2.1 MGT#1 reporter activation showed in vitro the MES subtype acquisition in response to various stimuli ------ 66

2.2.2 Comparison of the effect of in vitro upstream regulators in the GSA models using CAPE ------ 70

3. CHARACTERIZATION OF THE SIMILARITY OF CELL POPULATIONS IN GSA MODELS AND PATIENTS 74

3.1 Identifying the cell populations in the GSA models under various growth conditions 74

3.1.1 Evaluation of the in vivo GSA cell populations using scRNA-seq revealed the presence of all glioblastoma states ------ 74

3.1.2 Analysis of the ex vivo GSA scRNA-seq profiles after treatment revealed the acquisition of specific glioblastoma states ------ 79

3.2 Comparison of the GSA populations identity to patients 84

3.2.1 Integration of ex vivo and in vivo scRNA-seq profiles revealed differences between cell population in GSA models ------ 84

3.2.2 Integration of scRNA-seq profiles of models and glioblastoma patients corroborated GSA cellular identity ------ 87

DISCUSSION ……… 93

1. COMPARISON OF THE MOLECULAR PROFILES OF PATIENTS AND ADULT-TYPE DIFFUSE GLIOMA MODELS 93

2. USING CAPE TO ASSESS THE GLIOBLASTOMA STATE ACQUISITION IN GSA MODELS 101

3. CHARACTERIZATION OF THE SIMILARITY OF CELL POPULATIONS IN GSA MODELS AND PATIENTS 106

AUTHOR CONTRIBUTIONS 111

PUBLICATIONS 113

METHODS…… 115

1. GENERATION OF THE MOLECULAR PROFILES 115

Generation of GSA models ------ 115

Generation and processing of WGS profiles for GSA models in vitro ------ 116

Generation of RNA-seq profiles of GSA models ------ 118

Generation GSA models scRNA-seq profiles ------ 119

2. ANALYSIS AND COMPUTATIONAL METHODS 120

Analysis the results of the CAPE integrations ---------- 120

Benchmark of the CAPE results ---------- 124

Analysis of RNA-seq profiles from genetic tracing reporter GSA models ---------- 126

Analysis of scRNA-seq profiles ---------- 129

SUPPLEMENTARY DATA 133

SAMPLES 133

GENE SET 135

REFERENCES……….153

PROJECT

ABSTRACT

Diffuse gliomas are the most aggressive and incurable type of adult brain cancer. Humanized mouse models are useful for understanding the molecular mechanisms of tumor types and finding new therapeutic targets. However, comparing tumor models and tumor samples from patients remains difficult. To overcome this challenge, I developed a novel computational framework called CAPE for comparing tumor models and patient expression profiles. This computational toolkit based on non-negative matrix factorization allowed the integration of samples and the equal evaluation of clusters and the associated gene modules. I used CAPE to compare the expression profiles of humanized mouse glioma subtype avatar models (GSA) generated in the laboratory and adult-type diffuse glioma patients. The analysis revealed a strong resemblance between the models and the proneural glioblastoma subtype. The integration of the expression profiles of in vitro and in vivo GSA using CAPE also revealed that transplantation improved the acquisition of new tumor states in the models. To further investigate the model changes, I combined novel genetic tracing reporter phenotypic selection with CAPE. The results showed that a subset of in vivo GSA populations selected using the reporters clustered with patients with astrocytic-like identities. Furthermore, CAPE showed that GSA models treated in vitro with human serum, TNFα, or ionizing radiation revealed changes in the cellular identity toward a mesenchymal state upon reporter selection. Ultimately, I annotated the GSA populations in different conditions using single-cell transcriptomics. The results showed the presence of all glioblastoma states in vivo and upon external factor activation. The comparison between the GSA single-cell populations and patients confirmed this identity. Overall, this outcome aligned with the CAPE results, with a strong acquisition of the astrocytic-like and

oligodendrocyte progenitor-like cells in vivo. In conclusion, this study established a comprehensive framework for testing and validating the improvement of tumor models to mimic patients, thereby opening up a new avenue for understanding tumor biology and treatment response.

ZUSAMMENFASSUNG

Diffuse Gliome sind die aggressivste und unheilbarste Form von Hirntumoren bei Erwachsenen. Humanisierte Mausmodelle sind nützlich, um die molekularen Mechanismen von Tumorarten zu verstehen und neue therapeutische Ziele zu finden. Der Vergleich von Tumormodellen und Patientenproben bleibt jedoch ein schwieriges Unterfangen. Um diese Herausforderung zu meistern, habe ich ein neuartiges computergestütztes System namens CAPE für den Vergleich von Tumormodellen und Expressionsprofilen von Patienten entwickelt. Dieses auf der nicht-negativen Matrixfaktorisierung basierende Toolkit ermöglicht die Integration von Proben und die entsprechende Einordnung von Clustern und den zugehörigen Genmodulen. Ich habe CAPE verwendet, um die Expressionsprofile von humanisierten Maus-Gliom-Subtyp-Avatarmodellen (GSA), die im Labor erzeugt wurden, und von Patienten mit diffusem Gliom vom Erwachsenentyp zu vergleichen. Die Analyse suggerierte eine starke Ähnlichkeit zwischen den Modellen und dem proneuralen Glioblastom-Subtyp. Die Integration der Expressionsprofile von in vitro und in vivo erzeugten Glioblastomen mit Hilfe von CAPE zeigte auch, dass die Transplantation die Entstehung neuer Tumorstadien in den Modellen verbesserte. Um die Dynamik der Modelle weiter zu untersuchen, kombinierte ich neuartige genetische Reporter für Zellzustandsänderungen und CAPE. Die Ergebnisse zeigten, dass eine Untergruppe von in vivo GSA-Populationen, die mit den Reportern selektiert wurden, mit Patienten mit astrozytären Identitäten geclustert wurden. Darüber hinaus zeigte CAPE, dass GSA-Modelle, die in vitro mit Humanserum, TNFα oder ionisierender Strahlung behandelt wurden, Veränderungen der zellulären Identität in Richtung eines mesenchymalen Zustands zeigten. Schließlich habe ich die GSA-Populationen unter verschiedenen Bedingungen mit Hilfe der Einzelzelltranskriptomik klassifiziert. Die Ergebnisse zeigten das Vorhandensein aller Glioblastom-Stadien in vivo und bei Aktivierung durch

externe Faktoren. Der Vergleich zwischen den GSA-Einzelzellpopulationen und Patienten bestätigte diese Identität. Insgesamt stimmte dieses Resultat mit den CAPE-Ergebnissen überein, die eine starke Anreicherung von Eigenschaften astrozytärer und oligodendrozytärer Vorläuferzellen in vivo zeigten. Zusammenfassend lässt sich sagen, dass mit dieser Studie ein umfassender Rahmen für die Erprobung und Validierung der Verbesserung von Tumormodellen zur Nachahmung von Patienten geschaffen wurde, wodurch sich ein neuer Weg zum Verständnis der Tumorbiologie und des Ansprechens auf die Behandlung eröffnet.

Aims of the project

- **Aim 1**: To compare the expression profiles of GSA and glioblastoma patients by defining the models changes under different conditions.

- **Aim 2:** To identify cell populations and asses tumor heterogeneity in the GSA models by contrasting single-cell expression profiles from glioblastoma patients and models.

- **Aim 3:** To determine and rank the potential transcriptional modules that could transform GSA models into a particular glioblastoma identity.

BACKGROUND

1. EPIDEMIOLOGY, ORIGIN, AND CLASSIFICATION OF DIFFUSE GLIOMAS

1.1 EPIDEMIOLOGY OF DIFFUSE GLIOMAS IN ADULTS

Cancer refers to a group of diseases defined by the uncontrolled growth of resident cells. This growth ultimately forms a mass known as a tumor. This abnormal cell development, or carcinogenesis, causes complications in the surrounding tissue and can lead to the death of the patient. Moreover, tumor cells can travel to foreign tissue, developing new masses, in a process known as metastasis. In some cases, tumors reappear after therapy, a process known as tumor recurrence. Thus, cancers require ongoing monitoring of patients throughout treatment and throughout their subsequent lives. Cancer is prevalent worldwide and is one of the leading causes of human death in developed countries [1]. However, despite its prevalence, a full understanding of tumor biology and an effective general therapeutic approach for cancers remain elusive.

In general, cancer types can be classified based on histology [2] and tissue of origin [3]. Hence, tumor types are usually named by their primary location (e.g., lung cancer, colorectal cancer, cancer of the central nervous system), even if they have been propagated to another region. The World Health Organization (WHO) keeps the classifications of cancer types up to date as new information emerges in the literature. Tumors of the CNS represent several diseases that affect the brain and neuronal tissue in the spinal cord. The most recent WHO classification of tumors of the CNS differentiates among eleven categories [3] spanning different ages, grades of severity (grades 1 to 4), and prognoses. In particular, diffuse gliomas are the most common brain cancer affecting adults. Diffuse gliomas are subdivided into six families, of which diffuse gliomas astrocytoma IDH^{mut} and glioblastoma IDH1^{wt} are the only grade 4 affecting adults (**Fig.B1**). These grade 4 adult-type diffuse gliomas are fatal, with a median overall survival rate with standard treatment of 14 to 17 months [4]. Tumors of the CNS are relatively infrequent, accounting for less than 1% of all tumors diagnosed in the USA [5], but are very aggressive. In fact, grade 4 adult-type diffuse gliomas alone account for 49.1% of the total malignant tumors detected [6].

In general, adult-type diffuse gliomas appear spontaneously, which makes their early detection and diagnosis complicated. Moreover, the growth location of the tumor in the brain imposes important difficulties in diagnosis and treatment. In fact, the most common primary detectable symptom is the observable neurological consequences affecting patients [7]. Although some studies suggest that certain germline variations are linked to tumorigenesis in gliomas, the significance of these variations is still being debated [8]. Furthermore, some neural pathologies, such as epilepsy and Li-Fraumeni syndrome, have been linked to tumorigenesis and can affect treatment [9] but not prevention. Therefore, diagnosis relies mostly on the observable symptoms in patients after the appearance of the tumor.

Following clinical examination and evaluation of the possibility of a brain tumor, preoperative diagnosis using MRI determines the size and location of the tumor. This step is followed by surgery, where tumor samples obtained through microsurgical resection or stereotactic biopsy help to complete the diagnosis. Afterward, the tumor histopathology and molecular assessment of the sample using the most recent classification [3, 10] will determine the type and grade of the tumor, as well as the best treatment to be provided to the patient [11].

The treatment of adult-type diffuse glioma patients is challenging due to the location in the brain and the heterogeneity of the tumor grade. The standard therapeutic strategy to treat these tumors involves a combination of radiotherapy and alkylating agent-based chemotherapy after tumor resection [7]. In this strategy, radiotherapy is administered at various doses. The radiation dose is defined by the type of tumor and the age of the patient (e.g., Grade 4, 50-60 Gy at 1.2-2 Gy/day). Radiotherapy is combined with chemotherapy using alkylating agents such as temozolomide or DNA alkylation compounds that can cross the blood–brain barrier (e.g., nitrosourea class agents such as lomustine). In some countries (e.g., USA), therapy also includes treatment with bevacizumab (an anti-VEGF agent). However, no studies support an increase in overall survival in patients. Currently, new therapies for patients are being investigated, such as tumor treating fields therapy [13] or cell-based immunotherapy and target therapy [12]. Although these techniques improve the prognosis of patients, adult-type diffuse gliomas remain incurable.

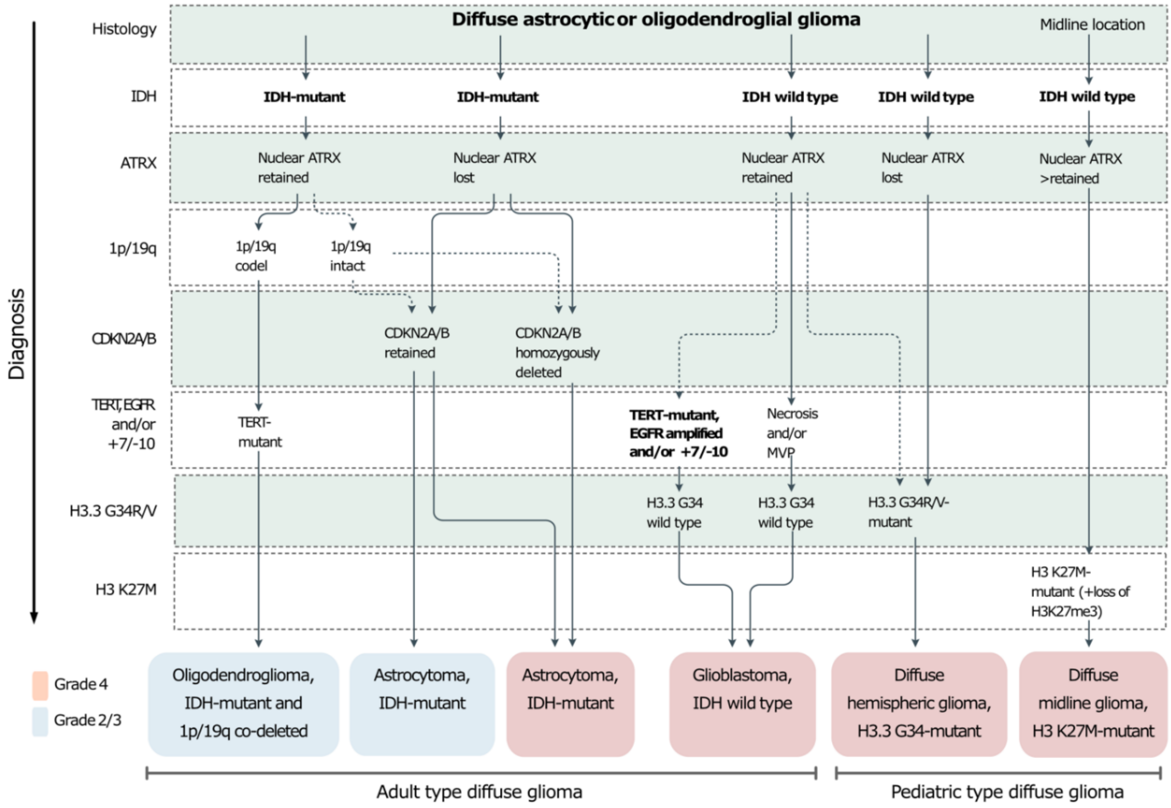

Figure B1. Characterization of diffuse gliomas

1.2 ORIGIN AND GROWTH OF ADULT-TYPE DIFFUSE GLIOMAS

A better understanding of glioma tumorigenesis is important to improve the prognoses of patients. Two types of processes contribute to tumor growth: genetic and nongenetic events [13]. Genetic changes influence the generation of tumor-initiating cells that sustain tumor growth. In particular, the accumulation of single-nucleotide mutations [14] and clusters of genomic aberration [15] transform normal cells, conferring beneficial traits for cell clonal selection. In addition, there are nongenetic determinants that play a paramount role in tumorigenesis, such as cell plasticity [16] and the tumor microenvironment (TME) [17]. In that sense, tumor growth depends on the genetic changes and the location of each cancer type [18]. A complete understanding of these processes in incurable malignancies, such as adult-type diffuse gliomas, will lead to therapies with better prognoses for patients [19].

Adult-type diffuse gliomas originate in the brain. The most accepted theory is that the tumor starts with genetic changes in neural stem cells (NSCs) or neural progenitor cells (NPCs) in the brain [20, 21]. In humans, only two adult brain regions sustain self-differentiating cells: the subventricular zone (SVZ) and the subgranular zone (SGZ) [22]. Currently, the SVZ is most accepted as the area where adult-type diffuse gliomas develop [23]. In 2018, Lee et al. showed the presence of shared glioma driver mutations (e.g., mutations in the TERT promoter) in both nonneoplastic and neoplastic tissues in the SVZ [23]. In contrast, it is uncertain whether the SGZ plays a role in the carcinogenesis of adult-type diffuse gliomas [24]. All these studies support the hypothesis that the SVZ location is the origin of this tumor, yet further work is required to fully understand all the steps that occur in early tumorigenesis.

Despite the available information, the exact cell-of-origin and the steps in early tumor growth in gliomas are still debated. Different studies using genetically engineered mouse models revealed the ability of different glial cells to form tumors, such as oligodendrocyte progenitor cells (OPCs) or astrocytes [25, 26]. In addition, different studies found similarities between the expression profiles of differentiated glioma cells to astrocytes, oligodendrocytes, or neural progenitor cells [27, 28]. Specifically, Couturier et al. showed that glioblastoma IDH1^{wt} cells recapitulate the differentiation process of glial cells in the fetal brain [29]. Relatedly, several studies adding GBM driver mutations such as TP53, NF1, or IDH to healthy human neural stem cells indicated tumor growth in animal models [30]. Another important point of study in adult-type diffuse gliomas is the sequence of the acquisition of genetic alterations. In 2019, Körber et al. showed that copy-number aberrations, such as chr7^{amp} or chr10^{del}, occur prior to the gain of point mutations, such as TERT [31]. These results were supported by the findings in Johnson et al. The authors used single-cell expression profiles and whole-genome sequencing to define the clonal evolution of the cells. The analysis showed similar results: copy-number alterations were acquired prior to the driver mutation [32]. Despite this, more additional studies are needed to understand the sequence of the early events that give rise to distinct adult-type diffuse gliomas.

The genetic changes in healthy cells are unable to explain all the tumor heterogeneity observed in adult-type diffuse gliomas [33]. Other nongenetic events, such as the TME and cell plasticity, play significant roles in the development and recurrence of GBM. Previous observations showed that even though early events are related to the mutation of the populations, later or recurrent events are more related to nongenetic determinants [34]. In gliomas, different cell types in the brain (e.g., microglia and neurons) interact with tumor cells, triggering new pathways and processes that contribute to tumorigenesis and increased tumor heterogeneity. The effect of the TME in gliomas is well known [35]. For example, several studies have shown how brain cells (e.g., astrocytes), resident microglia, or migrating immune cells (e.g., natural killer or T cells) interact and contribute to tumorigenesis in adult-type diffuse gliomas [36].

1.3 Molecular classification of adult-type diffuse gliomas

Microarray platforms opened a new era of high-throughput analysis in clinical and tumor research, helping to classify cancer types and subtypes [37–39]. Later, next-generation sequencing (or NGS) approaches permitted the examination of hundreds of samples at bulk [40] and single-cell levels [41], improving tumor classification. To better understand tumor heterogeneity, a large number of patients must be considered to account age groups, ethnicities, and tissue regions [42, 43]. This approach has been possible due to numerous large-scale consortiums, including TCGA [44], PCAWG [45], CPTAC [46], and TARGETx [47]. The use of different high-throughput methods helped provide new insights into tumorigenesis in adult-type diffuse glioma. This achievement led to the identification of different glioma types, subtypes, and states.

Molecular profiles are the best way to classify diffuse gliomas [48]. The most recent classification of tumors of the CNS [3] subdivides adult-type diffuse glioma into oligodendroma IDH^{mut}, astrocytoma IDH^{mut}, and glioblastoma IDH^{wt}. From those, only Astrocytoma IDH^{mut} and glioblastoma IDH^{wt} adult-type diffuse gliomas are categorized as grade 4 (i.e., the most severe type). Interestingly, previous classification based on histopathological traits [48] defined grade 4 adult-type glioma as glioblastoma multiforme or GBM. However, the addition of molecular profiles changed the CNS classification in 2016 [10]. In this classification, glioblastoma multiforme was divided as glioblastoma IDH^{wt} and glioblastoma IDH^{mut}.

Grade 4 adult-type diffuse glioma can be further classified into distinct subtypes using different molecular profiles [28, 44]. Specifically, gene expression levels are a good indicator of cellular activity and subtype-specific gene signatures. In 2006, Phillips et al. demonstrated that GBM expression profiles could be clustered into three subtypes: proneural (PN), mesenchymal (ME), and proliferative [27, 28]. The analysis based on the clustering of expression profiles (microarrays) defined the features of the GBM subtypes (e.g., survival and enriched pathways) and their similarities in specific brain cell types. In 2010, Verhaak et al. used expression profiles and consensus hierarchical clustering on microarray expression profiles to define four GBM subtypes: proneural (PN), classical (CL), mesenchymal (ME), and neural (NE) [28]. Additionally, the authors integrated genetic profiles into the analysis to associate mutations (e.g., IDH^{mut}) and copy-number profiles (e.g., PDGFRA) with the subtypes. This analysis showed that the PN subtype contained both IDH^{mut} and IDH^{wt} samples. A followed-up analysis in Brennan et al. defined the GBM genetic profiles [49]. Finally, in 2016, Ceccarelli et al. confirmed the difference between glioblastoma IDH^{mut} and IDH^{wt} using a multiomic evaluation [50]. This analysis confirmed that IDH^{mut} and IDH^{wt} constitute different types of adult-type diffuse glioma, and this division was included in the WHO 2016 classification [10].

The development of single-cell transcriptomics profiles allowed the comprehensive evaluation of GBM cell populations [51]. In 2017, Wang et al. [35] used nonnegative matrix factorization (NMF) to evaluate the expression profiles of glioblastoma IDH^{wt} patients to define three subtypes: PN, ME, and CL. In this case, the analysis included several microarrays, RNA-seq, and single-cell RNA-seq datasets. To define the genes used for the analysis, the authors identified the fraction of the genes that belong to neoplastic cells. Importantly, in 2019, Neftel et al. integrated several single-cell transcriptome profiles of glioblastoma IDH^{wt} patients and defined four distinct tumor states: OPC-like, NPC-like, MES-like, and astrocytic-like (AC) [52]. In this case, the state correlated with the previous subtypes: NPC/OPC corresponded to the PN subtype, AC to the CL subtype, and MES to the ME subtype. Furthermore, the authors revealed that these are transitory states that can interchange. The analysis also showed that all these states present cycling cells, indicating self-renewal capabilities for each tumor state. In addition, the analyses corroborated the copy-number amplifications associated with the subtypes. For example, the OPC state exhibited PDGFRA amplification, whereas the NPC exhibited CDK4 amplification.

Overall, the classification of grade 4 adult-type diffuse glioma into subtypes varies, whereas new studies reveal unknown subtype-related relationships and characteristics [53–55] .

2. COMPUTATIONAL METHODS TO CHARACTERIZE TUMOR MODELS

2.1 METHODS TO INTEGRATE THE EXPRESSION PROFILES OF TUMOR MODELS AND PATIENTS

Tumor models aim to mimic the molecular features of cancer. Therefore, tumor models must be compared to the type of tumor they recapitulate to understand its significance. In that sense, NGS molecular profiles have become essential in the characterization of existing tumor models. In particular, among the various profiles, gene expression has become critical in determining tumor identity and transcriptional signaling in various tumors, including grade 4 adult-type diffuse gliomas (see **Background 1.3**). It is essential to interpret these data to develop computational methods to assess the similarities between tumor models and patients. There are three general approaches to quantitatively evaluating tumor models using expression profiles: correlation, classification against a known set of genes and patients, and the integration of molecular profiles followed by clustering analysis.

The correlation between the gene expression values of patients and models aims to reflect the grade of similarity [56]. This method is based on quantifying the linear dependence in gene expression values between tumor models and patients (e.g., Pearson correlation). Therefore, the expression values should be corrected to avoid technical bias such as batch effect in order to compare the profiles. In the literature, there are several attempts to compare the expression profiles of models and patients based on correlation. For example, Chen et al. (2015) evaluated hepatocellular carcinoma cell lines and the expression profiles of patients and assessed the correlation of the most variable genes [57]. Similarly, Vincent et al. (2017) compared the expression profiles of several melanoma cell lines and patients [58]. The authors of this study established a ranking of cancer cell lines based on the average correlation coefficient of all genes in all cell line-malignant cell pairs. In another study, Cheng et al. (2018) compared the expression profiles of head and neck cancer cell lines and patients. For that purpose, they used the Spearman rank coefficient, a nonparametric measure of correlation [59]. Using this correlation, the authors focused on comparing copy number changes and expression profiles to assess models and patient similarities. In general, despite the simplicity of the approach, correlation only evaluates similarity but does not define individual differences.

Classification methods rely on the use of the expression profiles of patients as a reference to compare models. The classification can be performed either by the definition of subtype-specific gene signatures or by the implementation of classifiers with known tumor subtypes. In the first case, the evaluation using the gene signatures implies calculating a score that defines the subtype in the models. These signature genes determine the identity of the sample and help define the model. A simple method of applying this strategy is to evaluate the enrichment of upregulated subtype-specific genes in the model (e.g., hypergeometric test). Another approach is the individual evaluation of each sample subtype-specific score (e.g., ssGSEA). Overall, the major drawback of this approach is that they rely on previous knowledge regarding the tumor.

The evaluation using machine learning classifiers is the second type of classification method. In this strategy, the expression profiles from tumor patients associated with specific subtypes are applied as a reference to compare tumor models. Therefore, these expression profiles are used to train a model with annotated classes (i.e., tumor subtypes) that are used as a reference. This reference is then employed to define the identity of a new cohort of tumor patients or models. Usually, distinct correction or regularization methods are applied to avoid technical bias (e.g., using only the most-variable genes). There are multiple implementations of tumor classifiers [60–63]. For example, the web platform GlioVis incorporates the SubtypeME algorithm, which uses a combination of support vector machine (SVM), K-nearest neighbor, and ssGSEA to classify any input to the glioma subtypes [64]. Another example is MINT, which allows the identification of molecular signatures across experiments and platforms [65]. In general, classifiers are useful for differentiating between multiple tumor types. For example, Peng et al. (2021) developed a pancancer classifier using a top pair random forest approach that allows the classification of any tumor model within the cancers and subtypes defined in TCGA [66]. Despite the ability of the classifiers to easily integrate different expression profiles, it allows only the evaluation of similarities to a known phenotype.

The last strategy to compare tumor models and patients is the integration of the expression profiles. In this method, the values are first corrected to remove technical differences between patients and tumor molecular profiles, followed by unsupervised clustering of the samples. Specifically, the integration approach allows unbiased correlation of the sample identity and the expressed gene. For example, in 2021, Warren et al. developed Celligner [67]. The authors defined Celligner as a multistep procedure that removes variation and clusters integrated samples. Specifically, Celligner employs a contrastive PCA [68] and a modified version of mutual nearest neighbor [69] to correct technical differences. Once this step is finished, clusters are defined using algorithms that were originally built to cluster single-cell data (e.g., Louvain clustering from Seurat). In particular, the authors used this methodology to evaluate the similarities between the expression profiles of cancer cell lines [70] and TCGA samples. The major drawback of this approach is the requirement of the proper correction of the expression values to be integrated, which it might be challenging to define.

Although several methods have been proposed and used through years and technologies, the challenges in integrating the expression profiles from tumor models and patients still remain. For example, classifier methods or score systems, such as ssGSEA, can analyze only known subtypes. Therefore, the main objective is to develop methodologies that allow the simple evaluation of similarity between tumor models and patients as well as definition of the genes missing in the models. These methods will contribute to improving the evaluation of the specific differences between models and patients.

2.2 Methods to integrate the single-cell profiles of tumor models and patients

Evaluating expression profiles at the single-cell level has opened new avenues for assessing tumor heterogeneity. This technology can examine the transcriptomic profile of nonneoplastic and neoplastic populations within a tissue. Furthermore, the development and commercialization of droplet sequencing approaches enabled the generation of thousands of cellular profiles from a single sample. In parallel, the development of novel computational algorithms has been essential for analyzing the hundreds of thousands of transcriptomes produced from cells and yielding more accurate comparisons. The increased sample size and ability to focus on neoplastic cells alleviates some of the constraints imposed by bulk expression profiles. In recent years, several methods for integrating single-cell transcriptomic profiles across multiple conditions or sequencing runs have been developed. We can distinguish methods using cell population markers to differentiate populations, methods projecting a reference into the data to be explored, and methods integrating multiple datasets.

The first type of analysis relies on the definition of cell population markers. The expression profiles of cells that cluster together reveal specific marker genes that define their activity and differentiate them from other populations. Therefore, the simplest strategy for assessing the similarities between cell populations in tumor models and patients is to examine the enrichment of specific signatures upregulated in the tumor models. The enrichment of gene signatures can be evaluated at the population level (e.g., hypergeometric test) or in individual cells (e.g., ssGSEA). An example of this type of analysis would be the use of AUCell within SCENIC [71, 72] .AUCell estimates the enrichment of individual cells in the signature genes of tumor cell populations. Then, using the gene distribution, AUCell assigns each cell a defined identity. Interestingly, this method allows the definition of different cells with a shared identity, which is useful for observing transitions between tumor states. There are several examples of evaluating tumor models by defining cellular identity [73, 74]. However, this methodology relies on marker genes. Therefore, previous knowledge built on the cell population of patients is still needed.

Another approach to evaluating tumor models at the single-cell level is using single-cell profiles of tumor patients as a reference. In this case, single-cell expression profiles from patients are first analyzed to define the tumor-specific populations. Then, this dataset or "cell atlas" serves as a reference onto which new datasets are projected. In this case, there are two main approaches to projecting cells: similarity (e.g., scMAP, which is based on a scoring system that reflects correlation measures) and the generation of a classifier with the reference dataset (e.g., singleCellNet, which is based on a top pair random forest classifier). For example, using a projection approach, Couturier et al. (2020) integrated human tumor samples and fetal development [29]. For that purpose, they generated an atlas of fetal profiles using the PC space and then diffusion maps to integrate the tumor profile and evaluate various differentiation routes. In another study, Tan et al. (2019) developed a classifier by integrating several tumor data [75]. This method first builds a reference from different patients using a top pair random forest classifier, which allows the evaluation of the cell types in new profiles. The major drawback of these strategies is that they depend on the generation of a well-defined reference where to map tumor models.

A final approach to compare the single-cell profiles of models and patients is the integration of different datasets. This method allows the comparison without relying on reference datasets or defined populations. There are several approaches to integrating datasets, such as those based on decomposition methods (e.g., LIGER), neighborhood evaluation methods (e.g., KNN), correlation methods (e.g., CCA), and autoencoders (e.g., scVI). Despite the variability, the foundation of all these methods is the correction of the potential technical differences between datasets, followed by the analysis of the cell populations. Specifically, most of these methods rely on defined populations and the ability to differentiate between them. There are several examples in the literature of using this approach to compare single-cell profiles between tumor models and patients. Pine et al. integrated single-cell transcriptomics from GSC, patients, and GLICO models using batch-balanced k-nearest neighbors (BBKNN) to correct for technical differences and then evaluated the profiles [76]. Alternatively, Kiner et al. (2020) assessed the similarities between single-cell profiles of tumor cell lines and patients by integrating the corresponding single-cell samples [77]. In general, this method allows unbiased evaluation while preserving marker gene expression and cell populations. The main disadvantage of this methodology is that it is dependent on the assessment and correction of the technical differences. Therefore, as at the bulk level, a good control or a reference sample might be required for successful integration.

Overall, single-cell transcriptomics has improved the evaluation of tumor heterogeneity and the identification of specific events or tumor states. In addition, the evaluation of cell populations helped to understand the similarities and differences between models and patients. Despite this, there is still a need to develop new methods to compare datasets and populations from tumor models and patient samples. In particular, the generation of a large amount of single-cell profiles from many patients is still required to bring these tools to their full potential.

RESULTS

1. COMPARISON OF THE MOLECULAR PROFILES OF PATIENTS AND ADULT-TYPE DIFFUSE GLIOMA MODELS

1.1 GENETIC CHARACTERIZATION OF GLIOMA SUBTYPE AVATAR MODELS (GSA)

1.1.1 Complete genetic profiling of GSA models

To create an accurate representation of glioma patients, we generated subtype-specific adult-type diffuse glioma models, referred to as glioma subtype avatar models, or GSA models. To design the models, we modified human neural stem cells derived from human brain samples with genetic modifications specific to adult-type diffuse glioma patients [28, 49]. In total, we generated two GSA models, named after the mutational status of isocitrate dehydrogenase: IDH1^{wt} and IDH1^{mut}. However, a complete genetic characterization of these models was still lacking. Therefore, I evaluated the genetic profiles of the GSA models using whole genome sequencing and annotated the single nucleotide variations (SNVs).

The generation of GSA models started with the acquisition of neural stem cells from subventricular zone brain samples from human epilepsy patients. Then, we modified the neural stem cells by introducing a set of mutations and knockdowns (i.e., impaired gene expression) representative of two well-defined subsets of glioma patients (**Fig.R1**, see **Background 1.3**). The IDH1^{mut} GSA model contained the IDH1^{R132H} and TP53^{R273H} point mutations and PTEN knockdown. These mutations are frequently observed in patients with adult-type diffuse glioma IDH1^{mut} [49]. Comparison with the transcription-based classification (**Background 1.3**) showed that the IDH1 and TP53 genes in combination are frequently mutated in a subset of patients and are associated with the proneural subtype [28]. In contrast, the IDH1^{wt} GSA model contained knockdown of the PTEN, TP53, and NF1 genes. These mutations are frequently associated with the transcriptome-based ME subtype [49].

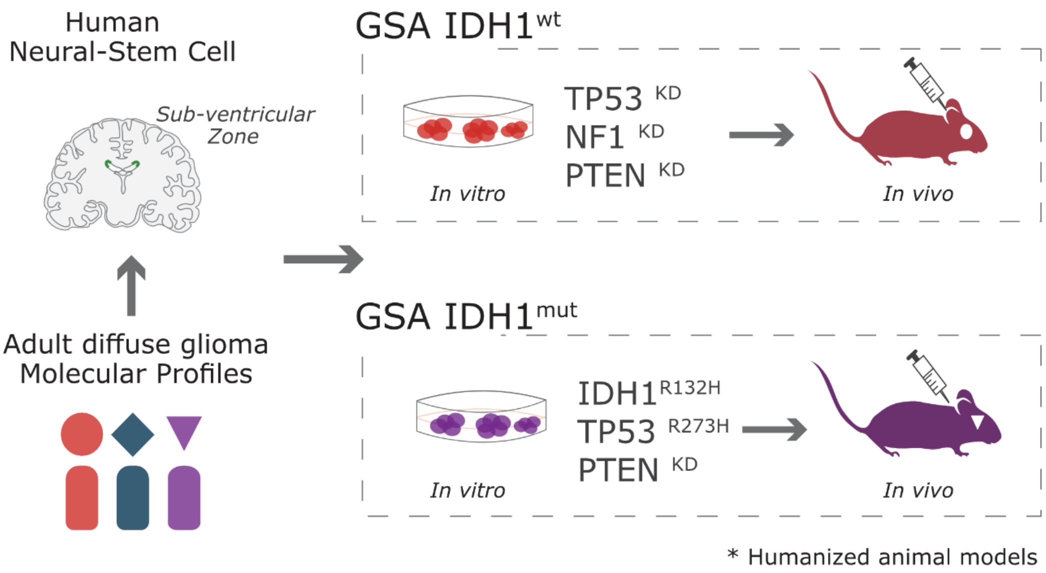

Figure R1. A diagrammatic representation of the Glioma subtype avatar models (GSA). GSA models of IDH1^{wt} (above) and IDH1^{mut} (below). The models are depicted in the two conditions in vitro and in vivo. The latter was obtained through orthotopic injection of in vitro samples into the brain of an NSD-Mouse.

To characterize the mutational status of the GSA models, we generated whole-genome sequencing profiles for each model (n=1). Following the sequencing of the samples, I used bwa [78] to map the reads to the GRCh38 assembly. I used the GATK v4.0 pipeline and Mutect2 variant calling to extract the SNVs associated with the GSA models [79]. Additionally, to retain only single nucleotide variations and not somatic variations, I filtered the called variants using the "1000G" dataset for positions annotated as single-nucleotide polymorphisms [80]. Then, using cancer-related SNV repositories such as COSMIC [81], I annotated the genetic profile of each model and evaluated the predicted effect on the protein using snpEff [82]. Overall, I discovered 2,950 SNVs in the IDH1^{mut} model and 2,623 SNVs in the IDH1^{wt} model. Only ~ 35% of the SNVs (918 SNVs) and ~50% of the genes (1050 genes) were common between the models.

Next, I examined the overlap between the SNV profiles of glioma patients and the GSA models. To achieve this, I correlated the SNVs in the GSA and TCGA-GBM samples (Mutect2 dataset, 392 samples, 2013), filtering by the exact genomic position and the nucleotides that changed in the sequence (**Table R1**). Then, I evaluated whether the results were consistent with the design of the GSA models and in comparison, to the genetic profiles of glioma patients. I noticed that the IDH1^{mut} model contained the expected point mutations IDH1 p.R132H and TP53 p.R273H, corroborating the experimental design of the model. Interestingly, both the IDH1^{wt} and IDH1^{mut} models shared the TP53 p.V216M missense mutation present also in patients.

Gene	**Coding**	**Model**	**Patient (%)**
TP53	p.R273H	IDH1^{mut}	35.7 %
TP53	p.V216M	IDH1$^{wt/mut}$	35.7 %
IDH1	p.R132H	IDH1^{mut}	6.1 %
SYT4	p.R288W	IDH1^{mut}	0.76 %

Table R1. The identified SNV from the genetic profiles of GSA models and TCGA-GBM (n= 392 samples). The correlation matched gene and protein modification.

Finally, to expand the evaluation of the status of glioma driver mutations, I examined the overlap between genes in the GSA models and the collection of pancancer drivers defined by TCGA [83]. In total, of the seventeen glioblastoma drivers identified in the pancancer study, ten (59%)

and six (35%) genes were mutated in the IDH1^{mut} and IDH1^{wt} GSA models, respectively (**Fig.R2**).

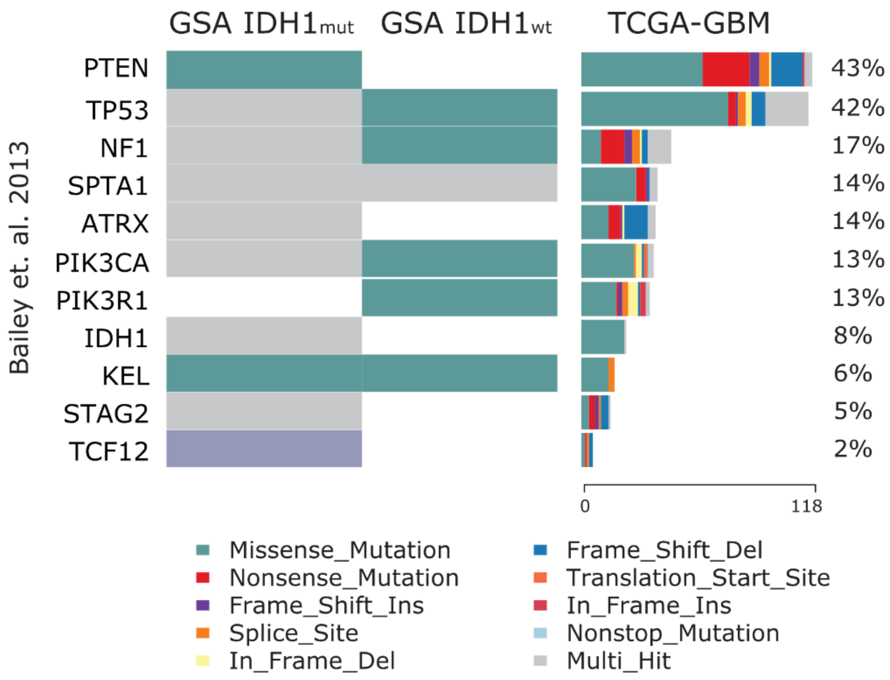

Figure R2. The heatmap displays the mutational status of GSA models and TCGA-GBM samples. The Y-axis label represents the correlated glioblastoma driver genes [83]. The color of the bar indicates the identified mutation type. The percentages in the TCGA-GBM column represent the relative number of patients for each gene.

1.1.2 Characterization of the Copy-number alterations in GSA models

Copy-number alterations (CNAs) are a hallmark of solid tumors and major contributors to adult-type diffuse glioma tumor heterogeneity [28, 49]. Therefore, I generated and analyzed the CNA profiles of the GSA models to complete their genetic evaluation. In addition, I assessed the overlap between the CNA profiles in patients and in the GSA models at the gene level.

I used the whole-genome sequencing profiles of the GSA models (see **Results 1.1.1**) to call the amplification and deletion regions present in each model. To call the CNA profiles, I used CNVKit [84]. Then, to improve the calling quality, I filtered out the black-listed human genome parts from the profiles [85] (i.e., regions in the genome defined by the high content of repetitive regions) and evaluated the CNAs that fell within gene regions. In total, I called nine CNAs that included twenty-two genes amplified (i.e., log2FC) in both the IDH1^{wt} and IDH1^{mut} GSA models. Surprisingly, I found an almost complete overlap between models copy number profiles. This finding suggests that the genomic aberrations occurred prior to the genetic modification of the neural stem cells.

The CNV profiles showed that both models carried focal CNV in the chr7 amplification and chr10 deletion hallmark also observed in adult-type diffuse glioma patients [28]. Interestingly, I also observed an amplification in the 4q12 chromosomal region (**Fig.R3A**) associated with the expression-based glioblastoma classification PN subtype [28]. Finally, I evaluated the overlap between the CNA profiles in the GSA models and patients using TCGA-GBM profiles at the gene level. To delimitate the number of genes, I only considered those genes listed as known glioblastoma amplifications/deletions [86]. In total, I found that our models shared three gene amplifications with patients: the CDK4, PDGFRA, and MYC genes (**Fig.R3B**). Notably, previous studies correlated CDK4 amplification with the NPC-like state and PDGFRA with the OPC-like state [52], indicating a potential similarity to these glioblastoma states of the GSA models.

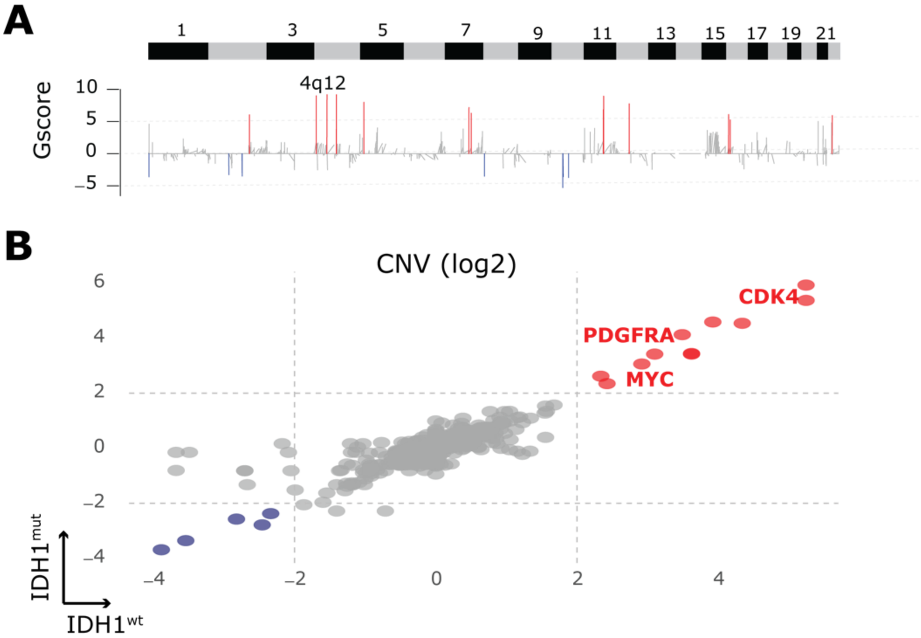

Figure R3. Copy number alterations in IDH1$^{wt\ /mut}$ GSA models. **(A)** The panel above displays the amplification and deletion identified in the GSA models (GScore, GISTIC2). The annotated chromosomal arm represents the focal 4q12 amplification [28] **(B)** The scatterplot depicts the comparison between log2FC of IDH1$^{wt\ /mut}$ GSA models CNA. The genes with amplification are annotated in red, while those with a deletion are annotated in blue. The graph labels correspond to genes associated with adult-type diffuse gliomas amplifications.

1.2 Computational framework to compare the expression profiles of tumor models and patients

1.2.1 Comparison of avatar models and patient expression profiles

Tumor models should represent their human analogous. Specifically, reliable models should be of use in drug testing or understanding complex tumor events, otherwise inaccessible from tumor samples only [78]. Publicly available databases, such as TCGA, allow the scientific community access to a wide set of molecular profiles of tumor patients [87]. This includes bulk expression profiles from thousands of donors, which can be used for tumor classification and subtype-specific pathway evaluation. However, the integration of the expression profiles from models and patients is still challenging. Currently, there are already proposed methods to integrate models and patients, including Celligner [67] (see **Background 2.1**). Although useful, these methods are still difficult when integrating multiple tumor cohorts and models or may give results that are difficult to interpret. To overcome these computational challenges, I developed an unbiased computational framework named CAPE (for Comparison of Avatar models and Patients expression profiles) that integrates current batch correction methods and NMF to compare the expression profiles of tumor models and patients.

In essence, CAPE integrates the bulk transcriptome profiles from tumor models and patients by defining the optimal sample aggregation of a corrected matrix, associating at the same time each group with a specific gene module (**Fig.R4**, see **Methods**). CAPE consists of three steps. First, CAPE takes the gene count matrix of models and patient samples and aggregates them into a SummarizedExperiment object [88]. In the process, it defines the metadata indicating the name, batch, and origin of the samples (i.e., model, or patient sample). After aggregation, a new function reduces the variability between conditions by withdrawing poorly represented genes and applying count-per-million normalization. Then, CAPE removes potential technical differences between datasets using RUV-Seq [89]. This method is based on a statistical framework using support vector decomposition and a set of control genes to correct the unwanted variation (i.e., technical bias). In CAPE, a function uses the indicated set of genes (e.g., housekeeping genes) or extracts the empirical control genes comparing dataset conditions as an input for the correction.

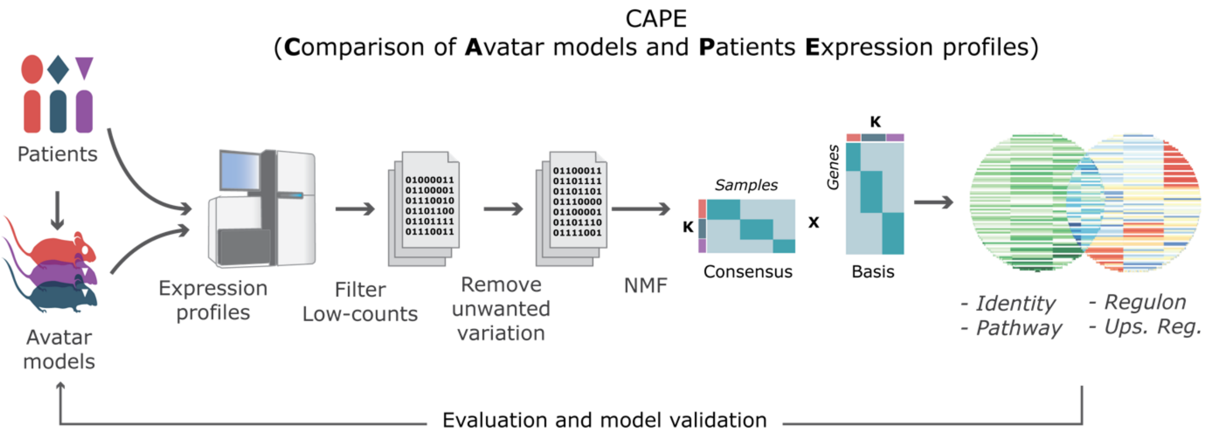

Figure R4. The scheme depicts the steps performed by CAPE in sequential order to integrate the expression profiles of tumor models and patients.

In the second part, CAPE evaluates the aggregation of the samples using NMF [90]. CAPE allows the user to assess the NMF decomposition of the matrices of combined patient-model samples and the samples for each individual dataset. NMF clustering is remarkably useful for dividing complex datasets and assigning clusters to their meta-modules. NMF works by decomposing the gene expression matrix n x m into the product of two matrices: n x k (basis matrix) and k x m (coefficient matrix), where n represents the genes, m the samples, and k the number of factors that define the decomposition. Hence, the selection of the factor k (or k-factor) is a critical part of the NMF algorithm and defines the data separation. CAPE selects this factor based on the selection of the best cophenetic value as defined in Brunet et al. [91]. This measure indicates the grade of correlation between integration after generating different runs of factorizations (nrun=10). By default, CAPE selects the highest cophenetic value across a range of distinct k-factors (e.g., k=2 to k=15) to set the final matrix decomposition. Despite this, an individual function has been set up to modify this selection manually.

In the last step, CAPE evaluates NMF decomposition by correlating each cluster with a specific gene module. In this step, the framework identifies the specific genes by comparing each group versus the rest using the basis matrix. The identification of each gene module is based on the selection of the genes above a log2FC threshold in comparison to the other clusters. In addition, I also included several extra individual functions to extract and evaluate the gene module information, such as annotating cluster tumor subtypes (see **Results 1.3**). Finally, to help the scalability of using this framework, CAPE was divided into different R functions and added to a public repository (https://gitlab.com/gargiulo_lab/cape) to be used by the scientific community.

1.2.2 Comparison using CAPE of GSA and glioblastoma patients reveal a PN subtype identity in the models

The genetic profiles of the GSA models recapitulated alterations observed in adult-type diffuse glioma patients (see **Results 1.1**). However, the grade of overlap between the transcriptome profiles of the tumor models and patients is still unknown. To evaluate whether our models recapitulated glioblastoma tumor formation and subtype specification, we generated several in vivo GSA model expression profiles. Then, I used the CAPE framework to integrate and compare the expression profiles of GSA models and glioblastoma patients.

We first generated several in vivo GSA expression profiles after orthotopic transplantation (i.e., introduced into the host the tumor initiating cells in the same location as the tumor growth) of in vitro IDH1$^{wt/mut}$ GSA model cells into the brains of immune-compromised mice. After three to four weeks, the mice started to show signals of tumor formation and thus were sacrificed to isolate the individual tumors. Consecutively, we generated RNA-seq profiles using individual tumors. In total, we generated several in vivo samples for each GSA model developed at two different times and sequencing locations (**Supplementary**). I integrated the expression profiles of the GSA models and TCGA-GBM [28, 35, 52] to assess the performance of CAPE (**Fig.R5**). To evaluate only the expression profiles of glioblastoma patients, I used only IDH1^{wt} samples [3].

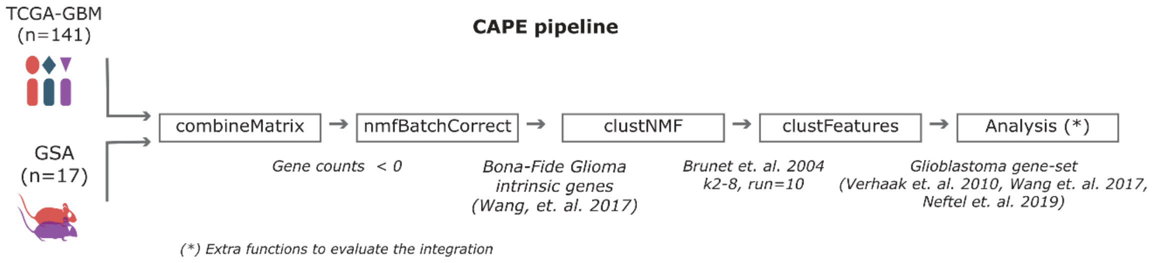
TCGA-GBM
(n=141)
GSA
(n=17)
CAPE pipeline
combineMatrix
nmfBatchCorrect
clustNMF
clustFeatures
Analysis (*)
Gene counts < 0
Bona-Fide Glioma intrinsic genes (Wang, et. al. 2017)
Brunet et. al. 2004 k2-8, run=10
Glioblastoma gene-set (Verhaak et. al. 2010, Wang et. al. 2017, Neftel et. al. 2019)
(*) Extra functions to evaluate the integration

Figure R5. Schematic depicting the integration between the expression profiles of in vivo IDH1$^{wt/mut}$ GSA models and IDH1^{wt} TCGA-GBM. The text within each box denotes the integration parameters.

CAPE contains several sequential steps (see **Results 1.2.1** and **Fig.R5**). First, I used the combined function (combineMatrix) to generate an integrated matrix of the transcriptome profiles of glioblastoma patients (n=141) and GSA models (n=18). This function also removed low-count genes from the analysis (i.e., genes with 0 counts in more than 95% of input samples). Second, I normalized the data and removed the unwanted variation between conditions using nfmBatchCorrect (see **Results 1.2.1**). To that end, I used a maximum of 9,334 genes empirically obtained by comparing the different datasets (**Fig.R6A**). Correction of the datasets can bias the results and affect NMF clustering. Then, in the third step, I assessed the correlation before and after the correction of the data to evaluate the coherence in the integration. The Spearman correlation for each sample between corrections scored a minimum sample correlation (or min. sample. cor) of 0.89 and an average sample correlation of 0.95, indicating a low impact of the correction of the data in the expression profiles (**Fig.R6B**). The comparison for each dataset was also significant (cor.test, adj.pvalue < 0.05). Finally, I assessed the sample distribution over a glioblastoma subtype gene set [35]. For that purpose, I computed the z-score for the CL, PN, and ME glioblastoma subtype gene sets for the corrected and noncorrected matrices [35]. Then, I evaluated the position of each sample in a ternary graph (which depicts the ratios of the three variables as positions inside an equilateral triangle), where each of the three axes corresponded to one of the glioblastoma subtypes. The analysis showed a clear distribution of both the IDH1^{wt} and IHD1^{mut} GSA models samples toward the PN axis before and after the correction (**Fig.R6C**). Intriguingly, the evaluation also shifted toward the CL subtype, suggesting that GSA models recapitulate certain grades of intratumor heterogeneity. Overall, these results indicated the ability of CAPE to generate a reliable correction of the data by maintaining the sample identity.

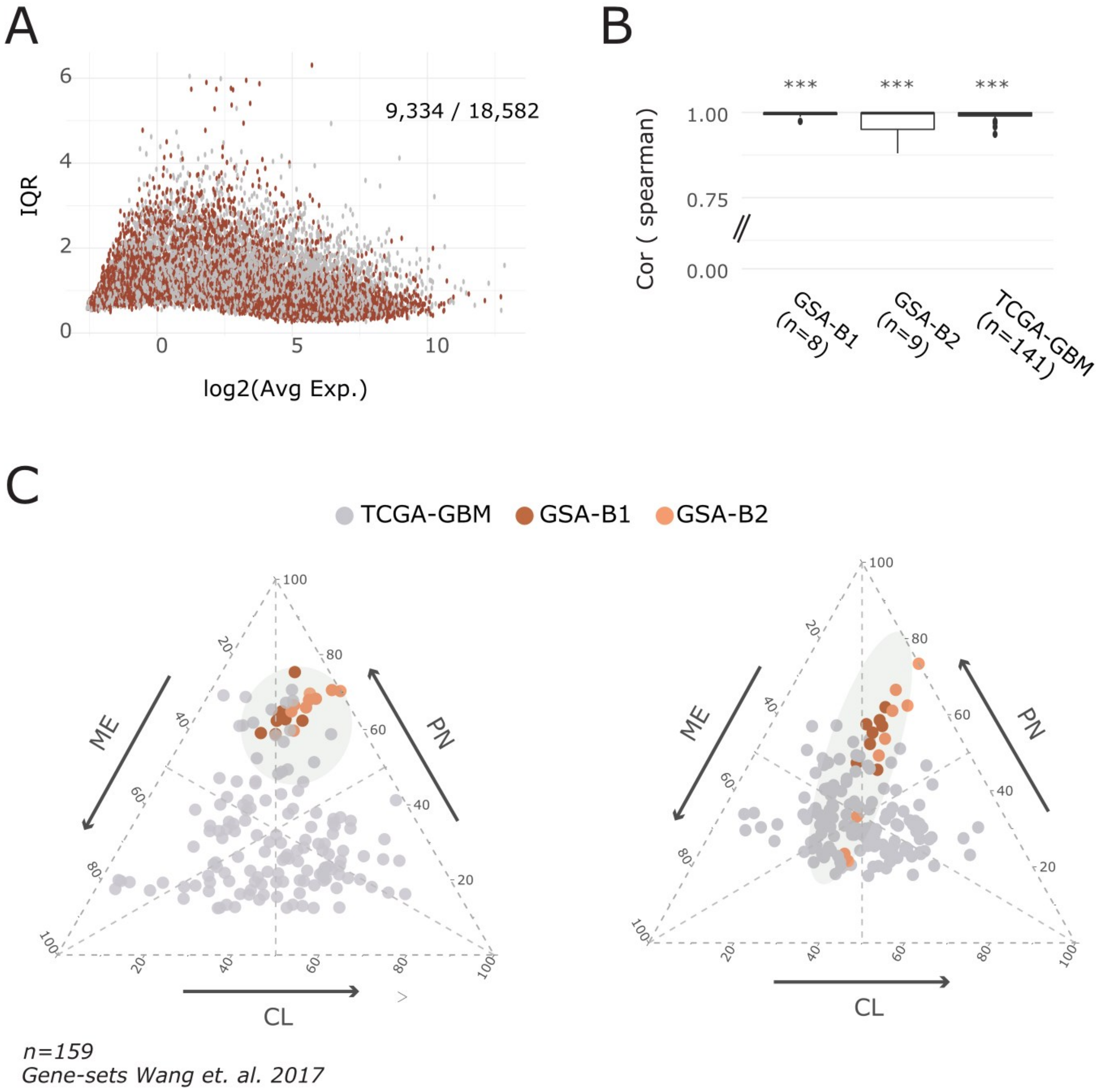

Figure R6. Evaluation of the correction step of the CAPE framework. **(A)** The scatter plot depicts the interquartile range (IQR) of the gene expression values distribution (avg. logCPM). The color represents the total number of empirical genes utilized by RUV-seq to generate the corrected matrix. The number displayed in the upper-right corner of the graph represents the ratio of empirical control versus total genes shared between datasets. **(B)** The boxplot describes the spearman correlation between the non-corrected matrix and the corrected for each sample grouped by dataset. **(C)** Ternary plot illustrating the non- (left) and corrected matrix (Right). Each axis represents a subtype of the glioblastoma gene set [35]. The color indicates the dataset of the sample.

After correction of the data, I evaluated the clustering using NMF. The CAPE framework integrates the NMF algorithm into the clustNMF function. Unlike tumor models, the expression profiles of tumor patients might contain contamination corresponding to the tumor microenvironment. Thus, to reduce the effect of the TME improving integration, I computed NMF clustering using only bona fide glioblastoma genes as defined in Wang et al. [35]. Then, I used k-factors from 2 to 8 as NMF parameters to determine the best decomposition of the samples (the brunet algorithm, which randomly initialized the decomposition, used Kullback–Leibler divergence as the loss function, and multiplicatively updated it to infer the decomposition). The evaluation using cophenetic values (**Fig.R7A**) showed that k=2 and k=4 represent the best k values (>0.99) to define the decomposition. Therefore, in line with the biological context, I selected k=4 to define the clusters. Interestingly, NMF decomposition revealed that GSA models clustered in the same group regardless of the sequencing run (**Fig.R7B**). Hence, the remaining clusters contained only profiles from glioblastoma patients. Finally, I evaluated the identity of each cluster using the given gene modules and glioblastoma subtype gene set [28, 35, 52] (**Fig.R7C-D**). The enrichment analysis (hypergeometric test) using several glioblastoma subtype gene sets indicated that each cluster was subtype specific. Particularly, the evaluation of the cluster that included the GSA model samples showed an enrichment of the PN subtype [28, 35].

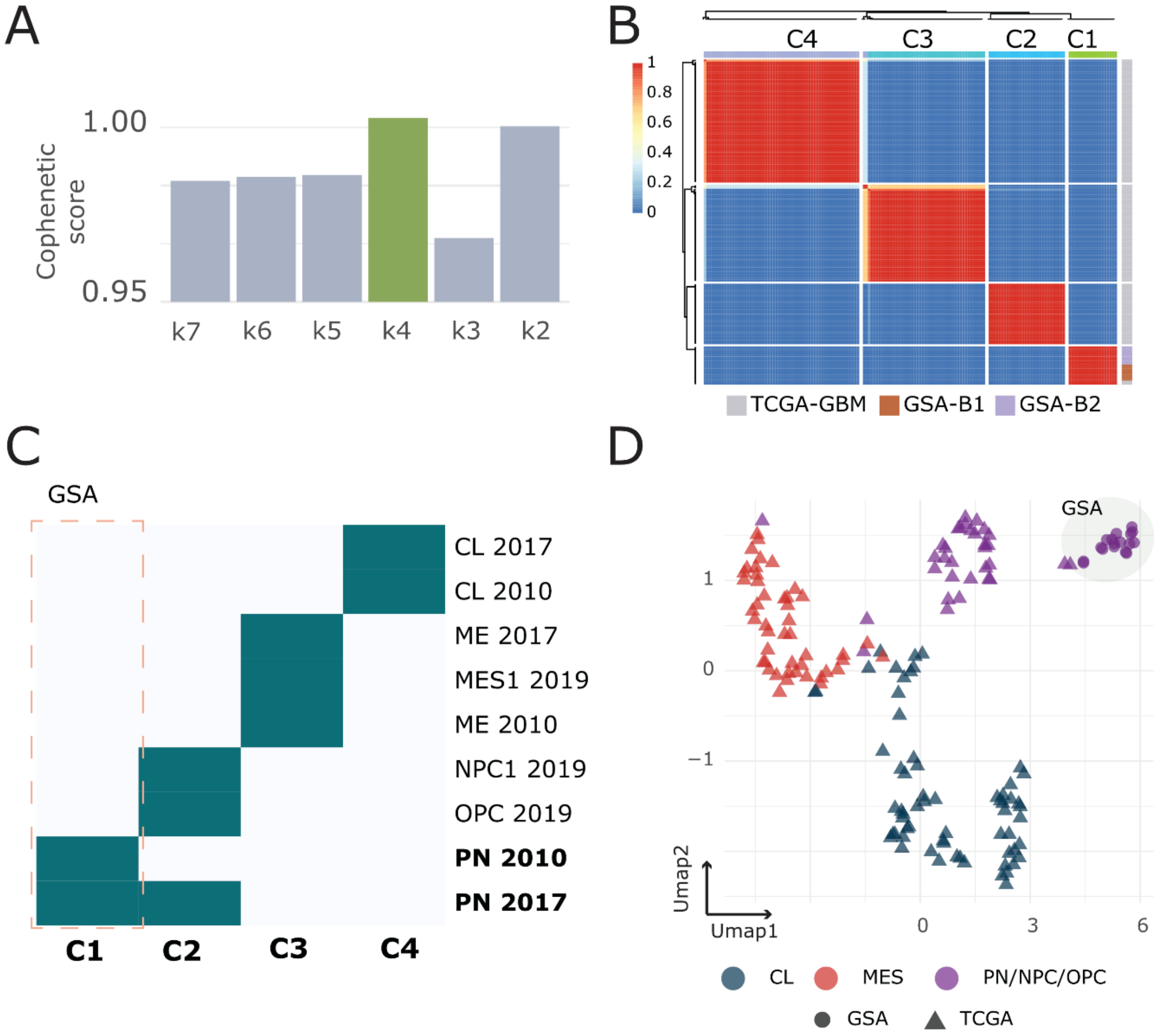

Figure R7. Evaluating the NMF decomposition step in the integration of expression profiles from in vivo GSA models and TCGA-GBM samples. **(A)** The barplot represents the cophenetic values for each k-factor in the NMF decomposition. The green bar represents the k selected for integration. **(B)** Consensus heatmap of the NMF decomposition. The values represent the connective score between n=10 runs for the selected NMF k-factor. The colored bar with annotations represents the samples. **(C)** The heatmap indicates the significant enrichment of glioblastoma subtype gene sets for each cluster-specific gene module [28, 35, 52]. The orange dashed box represents the model cluster. **(D)** UMAP for dimensional reduction using fitted values from the MMF decomposition. The shape of the dots indicates the origin of the sample. The color represents the enhanced glioblastoma gene set. The GSA samples were circled.

1.2.3 Benchmark CAPE results using different clinical cohorts and computational methods

In **Results 1.2.2**, I highlighted the ability of the CAPE framework to integrate the expression profiles of GSA models and glioblastoma patients. However, to establish the potential of CAPE to generate consistent integrations, I needed to demonstrate that it yielded similar results using different cohorts. For that purpose, I assessed the capabilities of CAPE by integrating the GSA models and different glioblastoma cohorts. In addition, I benchmarked the integration of the GSA models and TCGA-GBM samples using Celligner to evaluate the consistency of the results using different approaches [67]

First, I integrated the expression profiles of GSA models and three cohorts of glioblastoma patients as an alternative to TCGA [46, 92, 93]. To faithfully compare the previous GSA integration with the TCGA-GBM cohort, I used the same parameters as the previous analysis (see **Results 1.2.2**) and maintained glioma bona fide genes only [35]. First, I evaluated the consistency of the data before and after removing unwanted variations using Spearman correlation (min. sample cor. >0.95) and assessed the distribution of glioblastoma subtypes using a ternary representation (see **Results 1.2.2**). As previously observed, the analysis showed a distribution of GSA models toward the PN glioblastoma subtype and a shift toward the CL subtype (**Fig.R8A**). The comparison of the correlation at the dataset level also indicated significant values before and after the integration (**Fig.R8B**). Then, I computed the NMF decomposition using k-factors from 4 to 8. The cophenetic values of the integration showed the highest value at four well-defined clusters (cophenetic value >0.99; **Fig.R8C**). The annotation of the integration revealed the enrichment of PN markers in the GSA model clusters (C1 cluster) (**Fig.R8D**).

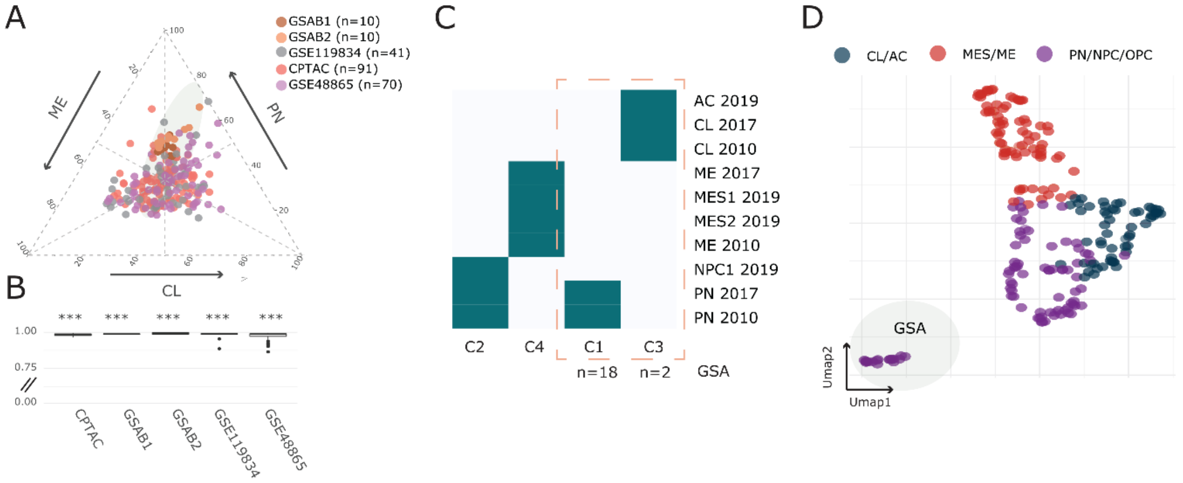

Figure R8. Evaluating the CAPE integration between the expression profiles of GSA models and GBM alternative cohorts. **(A)** Ternary plot illustrating the transform matrix. Each axis represents a subtype of glioblastoma [35]. The color indicates the origin of the sample. **(B)** The boxplot describes the correlation between the non-transform matrix and the transform matrix. **(C)** The heatmap shows the significant enrichment of glioblastoma subtype gene sets [28, 35, 52] for each cluster-specific gene module. The orange dashed box represents the model cluster. **(D)** UMAP dimensional reduction using fitted values from the MMF decomposition. The shape of the dots indicates the origin of the sample. The color represents the enhanced glioblastoma gene set. The GSA samples were circled.

Second, I compared the integration of GSA models with alternative selection of glioblastoma models and patient cohorts to assess the ability of CAPE to integrate different datasets. To this end, I used CAPE to evaluate the integration of the expression profiles of GSA, patient derived xenograft (PDX) [94, 95], and glioma stem cells (GSC) [92] with multiple glioblastoma cohorts [46, 92, 93]. I used similar settings to compare the results of the integration (see **Results 1.2.1**) and used bona fide glioma genes only [35]. As previously described in **Results 1.2.1**, the evaluation of the data transformation using a ternary plot showed that the GSA models were located toward the PN subtype axis (**Fig.R9A**). Furthermore, the correlation was also significant between the corrected and uncorrected matrix for each dataset (min. sample cor. > 0.67, cor.test < 0.05, **Fig.R9B**), with just one PDX sample having low correlation values. After the correction of the data, I evaluated the NMF decomposition. The clustering showed four well-defined groups (>0.99 cophenetic value). The analysis of GSA models still maintained the PN identity (**Fig.R9C**). Interestingly, in contrast to the GSA samples, the clustering showed that the GSC and PDX samples were completely integrated with glioblastoma samples (**Fig.R9D**). This indicates the ability of CAPE to define the clusters of different tumor models.

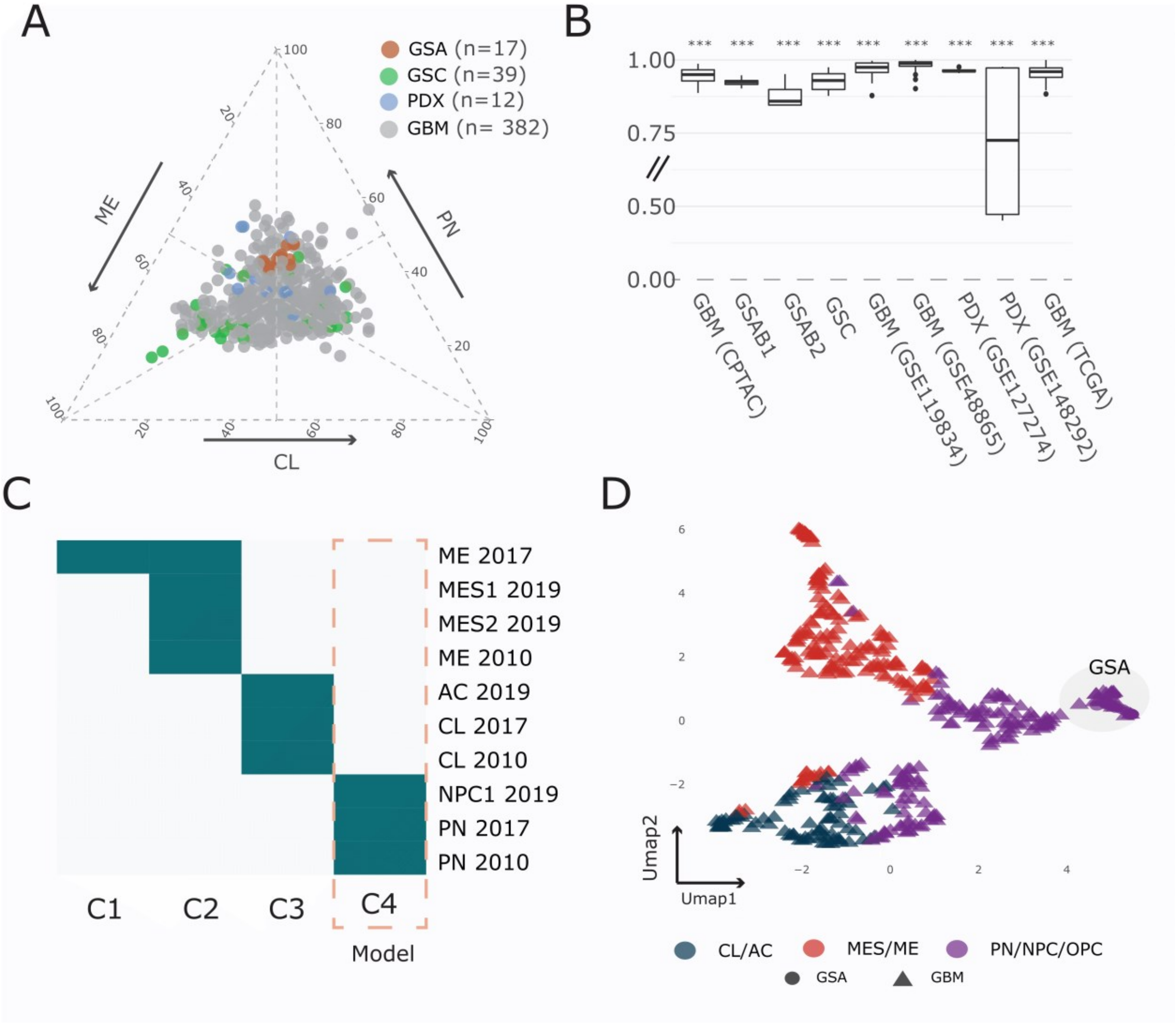

Figure R9. Evaluating the CAPE integration between the expression profiles of GSA models, PDX, GSC, and GBM cohorts. **(A)** Ternary plot illustrating the transform matrix. Each axis represents a subtype of glioblastoma [35]. The color indicates the origin of the sample. **(B)** The boxplot describes the correlation between the non-transform matrix and the transform matrix. **(C)** The heatmap shows the significant enrichment of glioblastoma subtype gene sets [28, 35, 52] for each cluster-specific gene module. The orange dashed box represents the model cluster. **(D)** UMAP dimensional reduction using fitted values from the MMF decomposition. The shape of the dots indicates the origin of the sample. The color represents the enhanced glioblastoma gene set. The GSA samples were circled.

Finally, I assessed the integration of the expression profiles of in vivo GSA models and glioblastoma patients (TCGA-GBM IHD1^{wt}) using Celligner [67] to benchmark the results from CAPE , I used only bona fide glioma genes only [35], after the integration. After following the Celligner pipeline [67], the integration yielded three defined clusters (**Fig.R10A**). Specifically, the GSA models were mostly included in the C0 cluster (n=14) and marginally included in the C2 cluster (n=4, **Fig.R10B**). Next, I extracted the cluster-specific markers using the FindAllmarkers function from Seurat as described in the methods. Then, I evaluated the enrichment of glioblastoma gene sets in each cluster-specific marker. The analysis showed that the C0 cluster was enriched in the CL, PN, and NPC1 glioblastoma states, while the C2 cluster was enriched in the ME subtype. In contrast, the C1 cluster did not present any enrichment (**Fig.R10C**). Finally, I used ssGSEA to define the individual score of the glioblastoma subtype gene sets [28, 35, 52]. The evaluation of the distribution of each score in UMAP dimensional reduction showed a high density of PN subtypes in the GSA models. In general, these results demonstrated that the GSA models correlated with the PN subtype. In addition, the comparison between the integration methods revealed that CAPE performed better than Celligner at defining discrete glioblastoma subtypes.

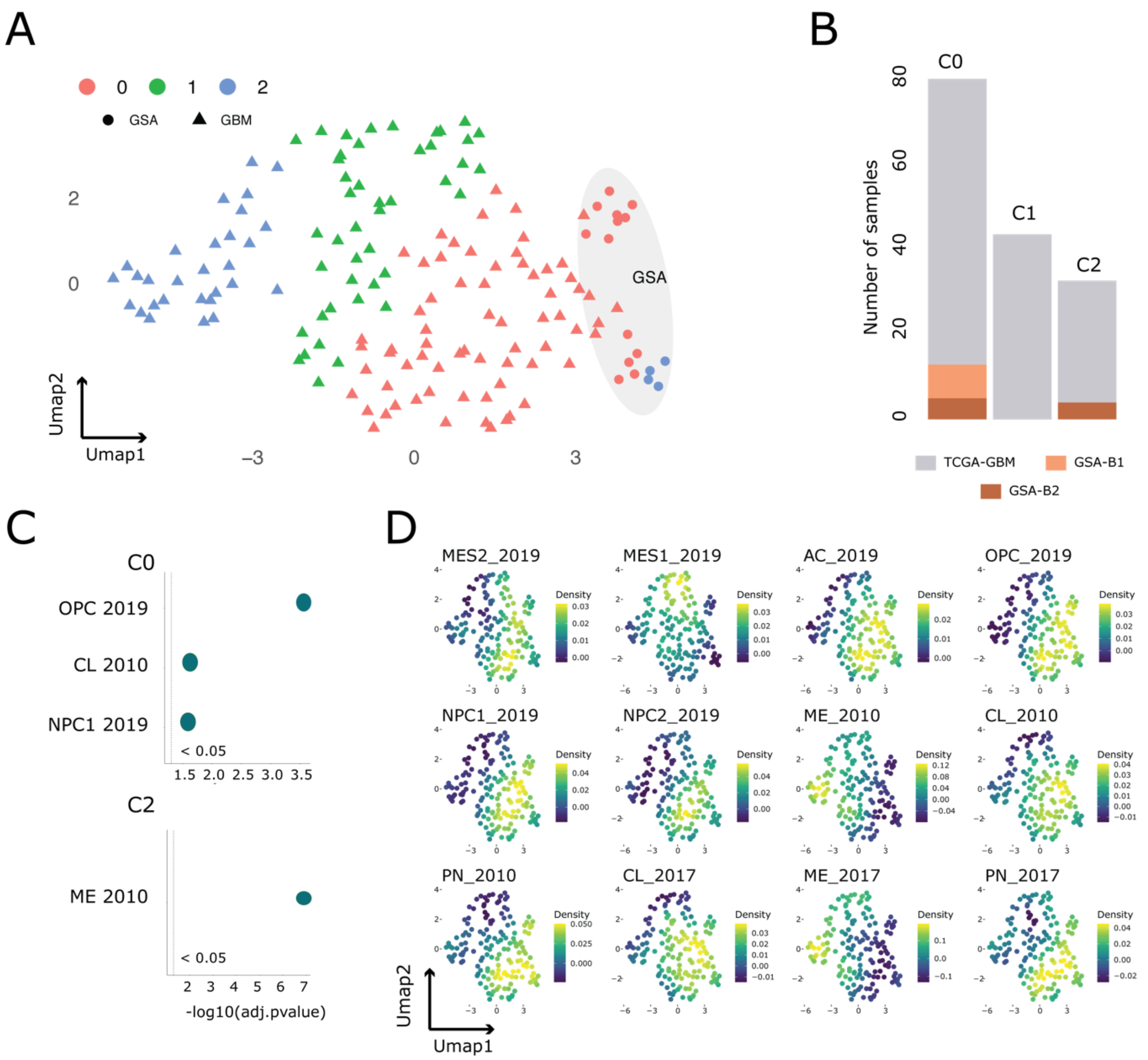

Figure R10. Integration using Celligner between expression profiles of in vivo IDH1$^{wt/mut}$ GSA and TCGA-GBM cohort. **(A)** UMAP dimensional reduction of the corrected matrix. The color of the dots corresponds to each cluster. The shape of the dots represents the origin of the samples. The circle represents the GSA model samples. **(B)** The bar plot indicates the total number of samples included in each cluster. The color represents the origin of the sample. **(C)** The dot-plot represents the glioblastoma subtype gene sets significantly enriched by each cluster. C1 is not displayed because it did not enrich for any glioblastoma gene set. **(D)** The UMAP dimensional reduction indicates the ssGSEA enrichment for each indicated glioblastoma gene set [28, 35, 52]. The color represents the score distribution across all samples.

1.3 CAPE ANALYSIS REVEALS SPECIFIC DIFFERENCES BETWEEN GSA MODELS AND GLIOBLASTOMA PATIENTS

1.3.1 Using CAPE to characterize gene module differences between GSA and glioblastoma patients

Matrix decomposition methods generate interpretable solutions after complex integrations (see **Results 1.2.1**). In particular, the CAPE framework generates two outputs: the cluster of samples and the cluster-specific gene modules. Both outputs are useful for interpreting the integration between the expression profiles of tumor models and patients. Specifically, the evaluation of the gene modules identifies the biological information that defines each cluster (e.g., activated pathways). Therefore, I used CAPE to identify the genes that differed between the GSA models and glioblastoma patients and evaluate the pathways contributing to the differences.

First, I evaluated the correlation between gene modules for each CAPE integration (see **Results 1.2.2 & 1.2.3**). After the clustering definition, CAPE obtains the cluster-specific gene modules (log2FC >1). For example, the integration of the expression profiles of GSA and TCGA-GBM showed four well-defined gene modules (**Fig.R11A**). To assess the grade of similarity between GSA models and glioblastoma cohorts at the gene level, I estimated the correlation of the gene modules obtained from different CAPE integrations using the Jaccard similarity coefficient. I applied the Jaccard index between pairs of gene modules from all integrations. Then, I evaluated the grouping using hierarchical clustering and heatmap representation (**Fig.R11B**). The analysis showed that gene modules were clustered by glioblastoma subtype [35].

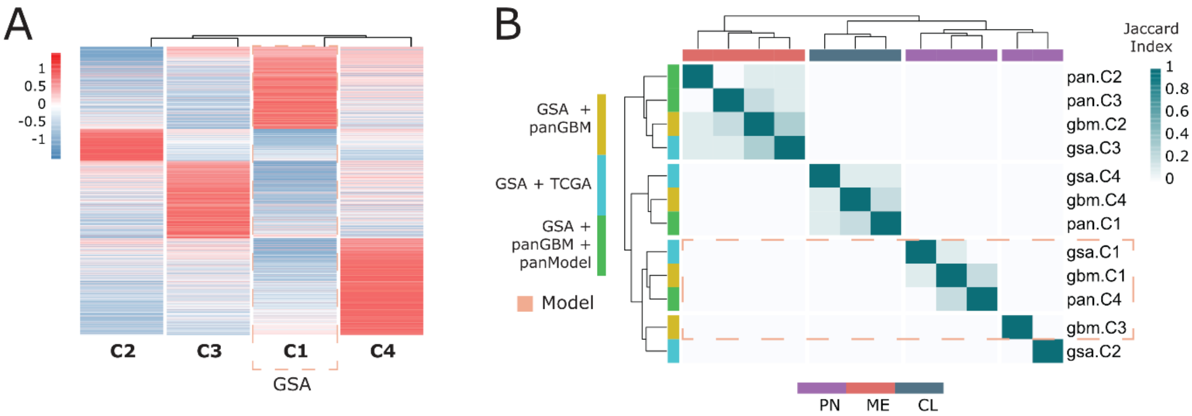

Figure R11. Evaluating gene module similarities across different CAPE integrations between expression profiles from GSA and glioblastoma patients. **(A)** Heatmap compares cluster-specific gene modules obtained from the NMF decomposition between the expression profiles from in vivo GSA and TCGA-GBM samples. The values represent the scaled values of the NMF basis matrix. The orange dashed box represents the GSA cluster. **(B)** The heatmap represents the correlation (Jaccard index values) between gene modules obtained from the indicated CAPE integrations (pan = GBM + panGBM + panModel, GSA = GSA + TCGA, gbm = GBM + panGBM). Color bars indicate the CAPE integration (left) and the enriched glioblastoma gene set (Right) [35]. The orange dashed box represents the model cluster.

Second, I evaluated the pathways represented in each cluster of the GSA and TCGA-GBM integration. For that purpose, I assessed the enrichment of hallmark gene sets using a hypergeometric test (MSigDB v7.2, adj.pvalue < 0.05, gene count). I observed that each cluster presented different associated pathways (**Fig.R12**). Interestingly, the C3 cluster (ME subtype) had the highest number of enriched pathways, indicating higher transcriptional activation. The C3 cluster (ME subtype) showed the enrichment of hallmarks such as those for the epithelial-to-mesenchymal transition (EMT), TNFα signaling via NF-kB, KRAS signaling, coagulation, and the inflammatory response. In contrast, the C4 cluster (CL subtype) was enriched in NOTCH signaling hallmarks and apical junctions. The C2 cluster (PN cluster associated with patients only) presented enrichment in hallmarks for MYC signaling and oxidative phosphorylation. The analysis of the hallmarks implicated in the cluster associated with the GSA models (C1 cluster) was related to proliferation.

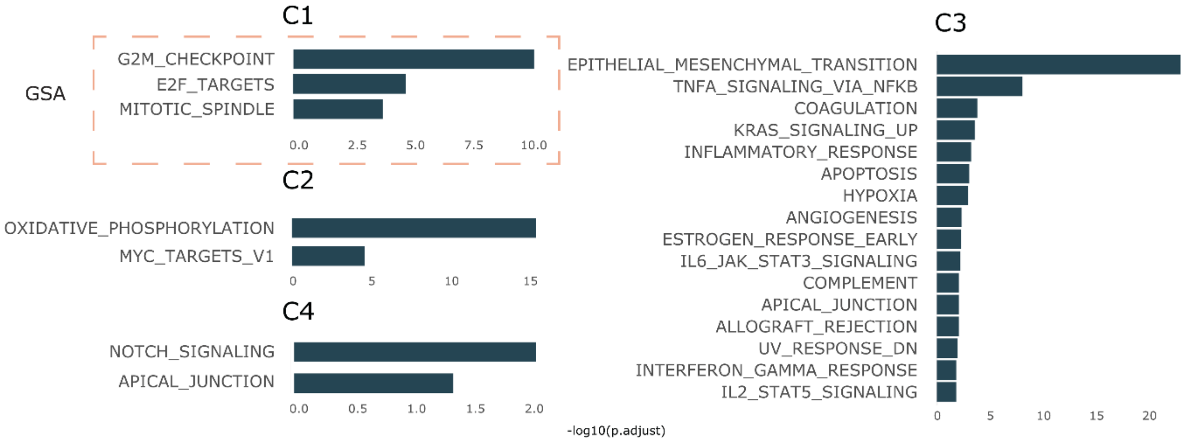

Figure R12. The bar plots represent the significantly enriched hallmarks (MSigDB v7.2, adj.pvalue <=0.05) in the cluster-specific gene modules for the CAPE integration between the expression values of GSA models and TCGA-GBM samples. The orange dashed box represents the model cluster.

Finally, I used PROGENy [96] to further define the cluster-specific pathways. Briefly, PROGENy fits a gene expression matrix to a precomputed reference model of cancer-associated pathways associating each sample and pathway to a value. First, to generate the PROGENy score in the CAPE integration, I filtered the corrected matrix using all the genes in the gene modules. Then, I applied PROGENy to the filtered matrix to obtain a matrix of cancer-associated pathways for each sample. To compare the pathways between clusters, I assessed the difference between cluster-specific samples for each pathway and a random selection of samples (adj.pvalue <0.01, **Fig.R13,** see **Methods**). As a result, the analysis showed different activations for each group in comparison to the other. C3 cluster (ME subtype) activity was associated with the TNFα, TP53, TGFß, and NF-kB pathways. On the other hand, the C1 cluster (GSA model-associated cluster) was enriched with the MAPK and PIK3 pathways.

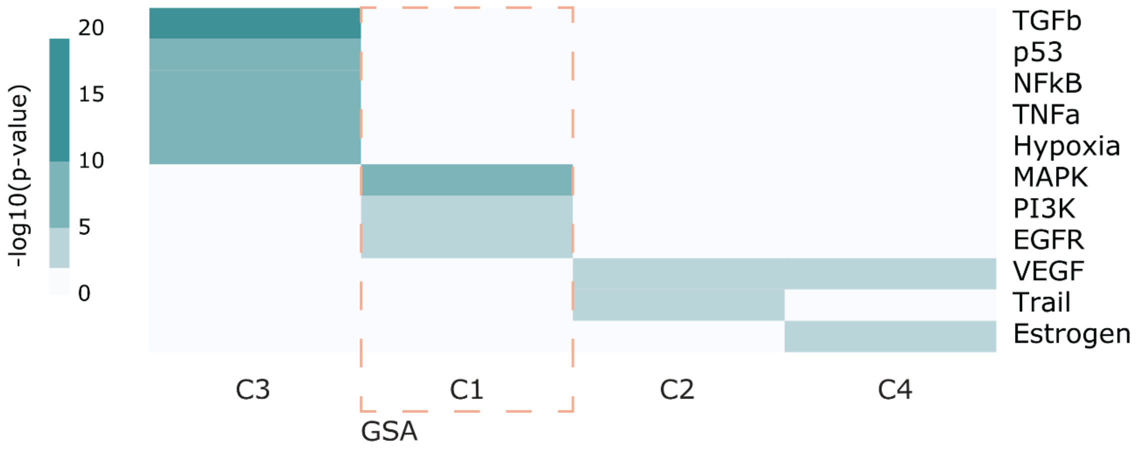

Figure R13. The heatmap illustrates the significant cancer-related pathways for each cluster in the CAPE integration between expression profiles of GSA models and TCGA-GBM. The value represents the t.test. Test difference between each cluster samples and the rest (adj.value < 0.05). The orange dashed box represents the model cluster.

1.3.2 Evaluation of cluster-specific transcriptional regulators to improve GSA models

The main objective for tumor models is to faithfully represent the molecular features observed in human samples. Modifying the growth conditions by adding external factors or changing transcriptional signaling through TF activation might be the simplest strategy to modulate the model identity. However, understanding the required conditions to promote these changes is still challenging. In that sense, the ability of CAPE to associate a set of genes with samples is useful to infer those factors. The comparison between GSA models and glioblastoma patients using CAPE showed the representation of each glioblastoma subtype. Therefore, I used the outcome of the CAPE integration between the expression profiles of GSA models and glioblastoma patients to define the elements activated in each subtype and not in the models.

First, I aimed to evaluate the transcription factors regulating glioblastoma subtypes. For that purpose, I generated the gene regulatory network (GRN) of glioblastoma IDH1^{wt} by integrating the expression profiles of different cohorts using CAPE (see **Results 1.2.3**). Additionally, to maintain the same conditions as before, I used bona fide glioma genes to reduce the effect of the microenvironment [35]. The integration showed a good correlation between samples (min. sample cor. >0.87, **Fig.R14A**), and NMF clustering divided the profiles into three clusters (cophenetic value 0.99; **Fig.R14B**). Then, the annotation of the cluster-specific gene modules indicated the enrichment of the glioblastoma subtypes for each cluster (**Fig.R14C**). To infer the GRN associated with glioblastoma patients, I used SCENIC [97]. This method uses the GRNBoost2 algorithm to infer the TF-gene interaction [97] and then defines the TF regulon (i.e., all the genes targeted by the TF).

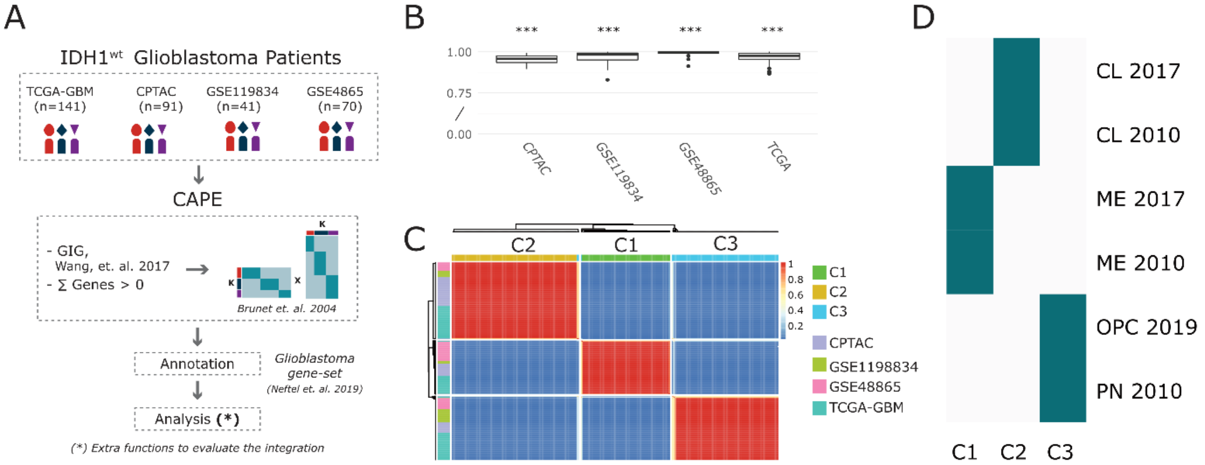

Figure R14. Evaluating the CAPE integration between the expression profiles of glioblastoma cohorts. **(A)** Schematic of the CAPE integration. **(B)** The boxplot describes the correlation between the non-transform matrix and the transform matrix. **(C)** Consensus heatmap of the NMF decomposition. The values represent the connective score between n=10 runs for the selected NMF k-factor. The colored bar with annotations represents the samples. **(D)** The heatmap indicates the significant enrichment of glioblastoma subtype gene sets for each cluster-specific gene module (adj.pvalue < 0.05) [28, 35, 52].

After the generation of GRN, I evaluated the TF enrichment for each gene module (**Fig.R11A,** adj.pvalue < 0.05, gene count >5). The evaluation of specific TFs showed a different enrichment for each subtype (**Fig.R15A**). The C3 cluster (ME subtype) was enriched in different TFs that formed a large TF network compared to other clusters. In particular, this cluster showed different regulon interconnections, such as FOSL1/2, RELB, NFKB2/1, CEBPB, ETS2, SP1, FLI1, IRF8/4, BCL3, and RUNX3, related to the ME subtype. On the other hand, the C1 cluster (PN cluster associated with GSA models only) revealed several enriched TFs related to the cell cycle, such as E2F2/8 and E2F2, and NPC-related TFs, such as SOX4/11 and MYC.

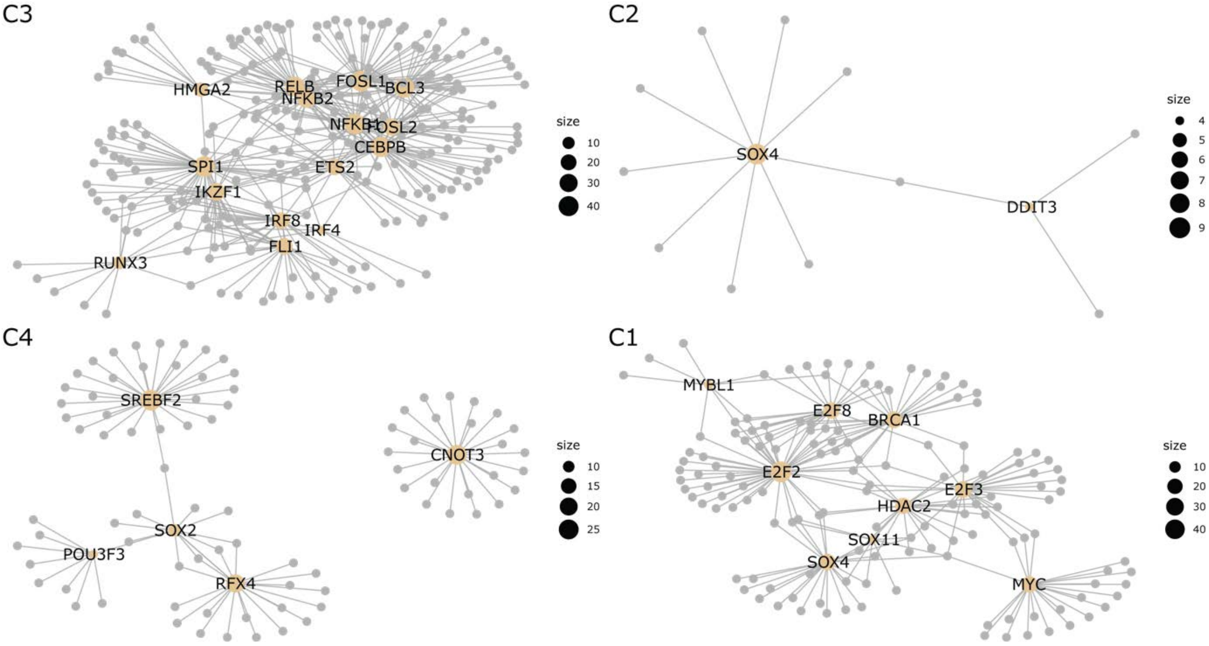

Figure R15. The connective networks represent the significantly enriched regulon for each cluster in the CAPE integration of expression profiles from GSA models and TCGA-GBM. The grey dots denote all genes associated with the orange-colored TF-regulon. The dot size indicates the number of genes associated with each regulon in the analysis. The orange dashed box represents the model cluster.

Finally, I evaluated which growth factors, cytokines, or hormones that appeared in each cluster gene module to define potential cluster-specific upstream regulators. For that purpose, I used the Omnipath database [98]. This package contains information about pathway-related databases focusing on ligand□receptor gene interactions. I extracted the values from databases with curated entries (CancerCellMap [99], CellPhoneDB [100], CellChatDB [101], and CellTalkDB [102]). Then, I ranked the upstream regulators present in each cluster using the NMF-based values (**Fig.R16A**). The analysis showed different upstream regulators linked to each cluster. For example, the C3 cluster (ME subtype) expressed the SPP1, CCL2, TGFB1, LIF, or IL1B gene. Next, I correlated all the upstream regulators in the clusters with the receptors in the GSA models to identify the best candidates to promote changes. To this end, I sorted the upstream regulators based on the expression of the receptor in the GSA models (avg. normalized gene expression, **Fig.R16B**). The analysis showed PDGFA (upregulated in the C4 cluster) and LIF (upregulated in the C3 cluster) as the top potential regulators. Finally, I evaluated the coexpression of different upstream regulators related to patients to simplify the selection. The analysis indicated that there were five clusters, including several upstream regulators and many clusters of just individual regulators. This result provides knowledge of how to improve the models toward specific subtypes by simply using already-published analytical packages and databases.

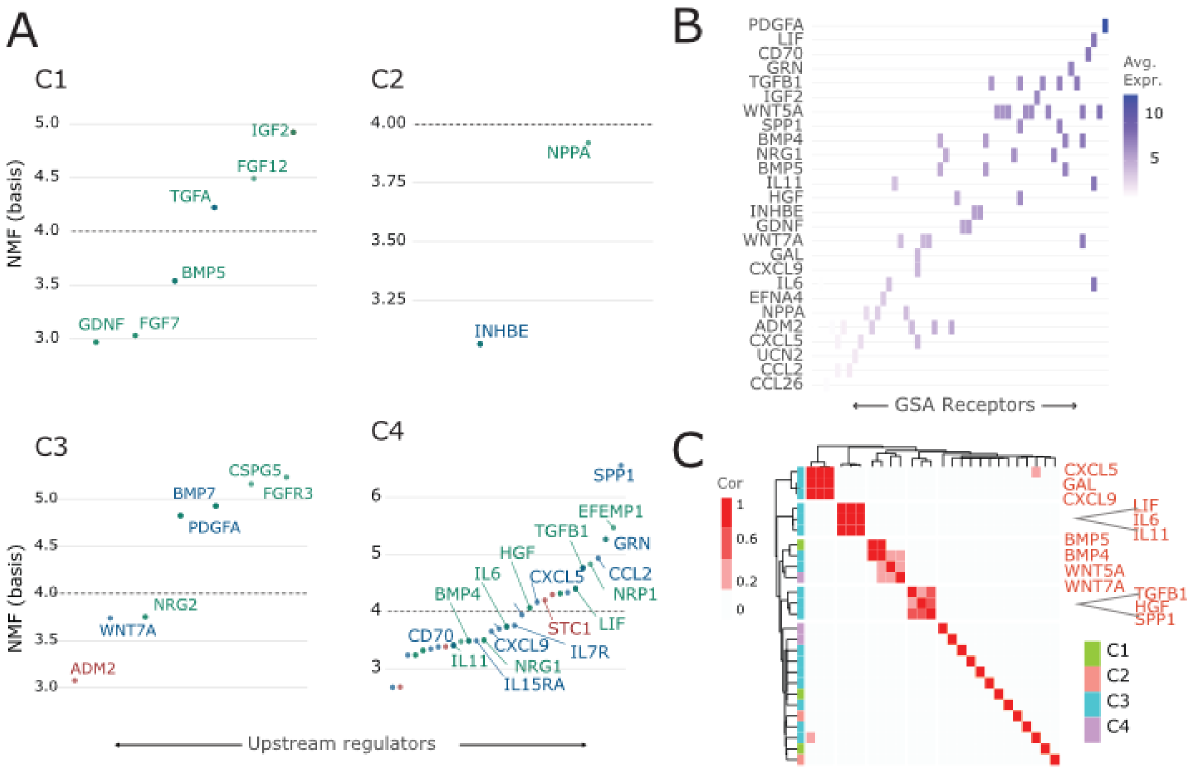

Figure R16. Evaluating the upstream regulators active in each cluster of the CAPE integration between expression profiles of in vivo GSA models and TCGA-GBM patients. **(A)** The scatter plots define the external factors detected in each cluster. The x-axis represents the growth factor, and the y-axis represents the NMF basis value in the indicated cluster. Color code indicates the type of upstream regulator according to OmniPath database . **(B)** The heatmap defines the relation between detected upstream regulators in the CAPE integration (y-axis) and their corresponding receptors expressed in the in vivo GSA models (x-axis). The scale indicates the average expression values in GSA model samples (quantile normalization). The average expression of the reporters determines the order. **(C)** The heatmap illustrates the correlation between growth factors based on the co-expression of their associated receptors in GSA samples. The annotation denotes the cluster in the CAPE integration.

2. Using CAPE to assess the glioblastoma state acquisition in GSA models

2.1 GSA models acquired new glioblastoma states after engraftment

2.1.1 CAPE analysis revealed differences in GSA model expression profiles in vitro and in vivo

Nongenetic features contribute to tumor heterogeneity (**Background 1.2**). In particular, in gliomas, the interaction between tumor cells and the brain microenvironment contributes to tumor differentiation [36]. GSA models can represent the growth conditions of adult-type diffuse gliomas in the brain. In that context, in vitro GSA models represent tumor-initiating cells with self-renewal capabilities with the potential to differentiate into tumors, and the in vivo GSA models recapitulate the tumor growth. However, the difference between in vitro and in vivo GSA models is still unknown. Therefore, I used CAPE to integrate and evaluate the main difference between the expression profiles of in vitro and in vivo GSA models.

We generated several transcriptome profiles of in vitro (n=8) and in vivo GSA (n=17) IDH1$^{wt/mut}$ models in three different sequencing runs. I used CAPE to integrate and assess the differences between GSA models under different conditions (**Fig.R17A**). First, I assessed the step of batch correction using the correlation between samples before and after correction. The minimum Spearman correlation for all the samples was 0.94, and the difference between datasets was statistically significant (adj.pvalue < 0.05, **Fig.R17B**). Then, I evaluated the clustering of the expression profiles using NMF decomposition. The analysis showed two well-defined clusters (**Fig.R17C**). Specifically, the integration separated between in vitro and in vivo samples in the C1 (8/4) and C2 (0/13) clusters, respectively (**Fig.R17D**). These results indicated that our models do not cluster based on the sequencing batch or the IDH1 mutational status. Finally, I used cluster-specific gene modules to evaluate the enrichment (adj.pvalue <0.05 and gene count >5) of glioblastoma subtype gene sets [28, 35, 52]. I observed that the C2 cluster (in vivo GSA) acquired AC, OPC, and NPC1 glioblastoma states [49], while the C1 cluster (in vitro GSA) was only enriched for the MES 2010 subtype genes (**Fig.R17E**). This result suggests increased tumor heterogeneity upon engraftment.

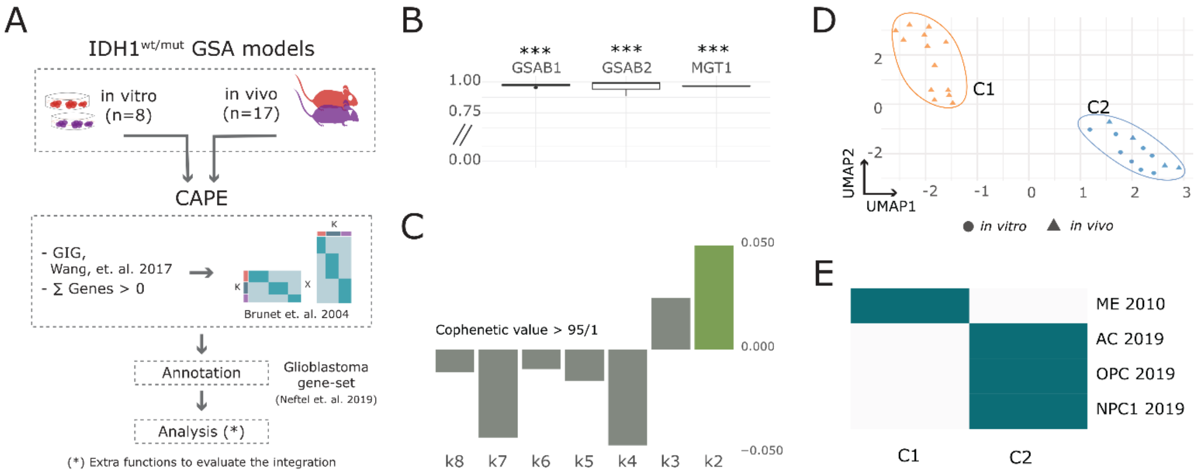

Figure R17. Evaluation of CAPE integration between expression profiles of in vivo and in vitro IDH1$^{wt/mut}$ GSA models. **(A)** Schematic of the CAPE integration. **(B)** The boxplot describes the correlation between the non-transform matrix and the transform matrix. **(C)** The barplot represents the cophenetic values for each k-factor in the NMF decomposition. The green bar represents the k selected for integration. **(D)** UMAP dimensional reduction using fitted values from the MMF decomposition. The annotation corresponds to the information in (E). The color indicates the NMF clusters. **(E)** The heatmap shows the significant enrichment of glioblastoma subtype gene sets for each cluster-specific gene module [28, 35, 52].

After integration, I assessed the functional differences between clusters. First, I evaluated the hallmark gene sets (MSigDB v7.2) enriched in each group using cluster-specific gene modules (**Fig.R18A**). The analysis showed that the C2 cluster (in vivo GSA) enriched hallmarks related to the IFN response, proliferation, and the EMT. Alternatively, the C1 cluster (in vitro GSA) was enriched in hallmarks of myogenesis, cholesterol biosynthesis, and hypoxia. Next, I used the Dorothea database [103] to evaluate the activated transcription factors in each cluster (**Fig.R18B**). I observed the enrichment of the TFs STAT1, SOX10, and NKFB1 in the C2 cluster (in vivo GSA). On the other hand, the regulon analysis in the in vitro cluster suggested changes related to adipogenesis, such as changes in SREBP1 [104] and KLF5 [105], and stemness, such as changes in PRDM14 [106]. These results indicate differences in the transcriptional regulation between tumor growth conditions recapitulated in the clusters.

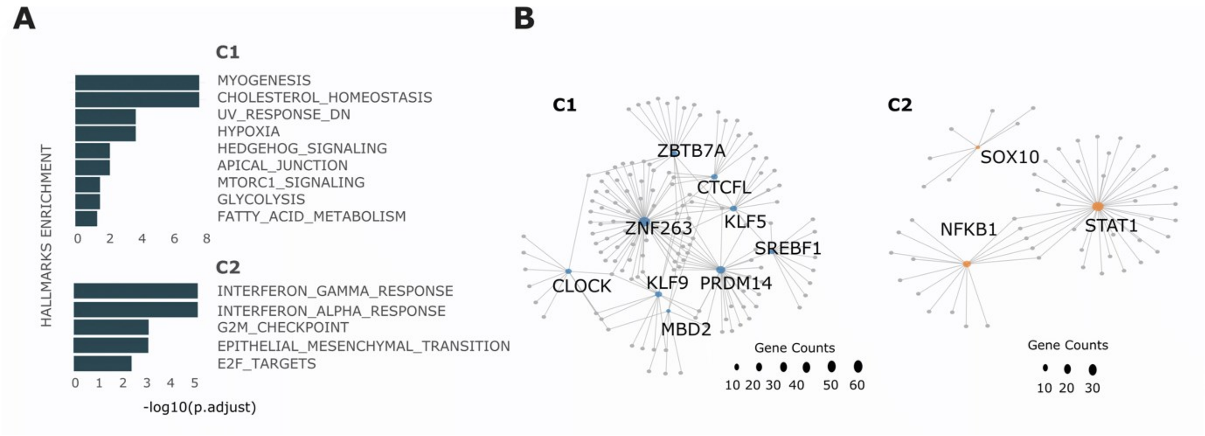

Figure R18. Evaluation of the enriched pathways and TF regulons in the CAPE integration between the expression profiles of in vivo and in vitro IDH1$^{wt/mut}$ GSA models. **(A)** The bar plots represent the significantly enriched hallmarks (MSigDB v7.2, adj.pvalue <=0.05) in the cluster-specific gene modules. **(B)** The connective networks represent the significantly enriched regulon for each cluster. The grey dots denote all genes associated with TF-regulon. The dot size indicates the number of genes associated with each regulon in the analysis.

Lastly, I used omnipath database [98] (see **Results 1.3.2**) to identify the upstream regulators expressed in each cluster that define cell homeostasis under different conditions. From the omnipath database, I selected only manually curated databases (CancerCellMap [99], CellPhoneDB [100], CellChatDB [101], and CellTalkDB [102]) to extract the cluster-specific upstream regulators. Then, I correlated the average expression of the samples in each cluster and the NMF-decomposition basis score to rank the upstream regulators (**Fig.R19).** The results showed that the C2 cluster (in vivo GSA) expressed PTN, IL6ST, JAG1, IL1RAP, and FGF12 as top upstream regulators (>q90), while the C1 cluster (in vitro GSA) expressed the molecules GPI, VEGFA, and VGF. Overall, the integration of the expression profiles of in vitro and in vivo GSA models using CAPE revealed the difference between differentiation and proliferation.

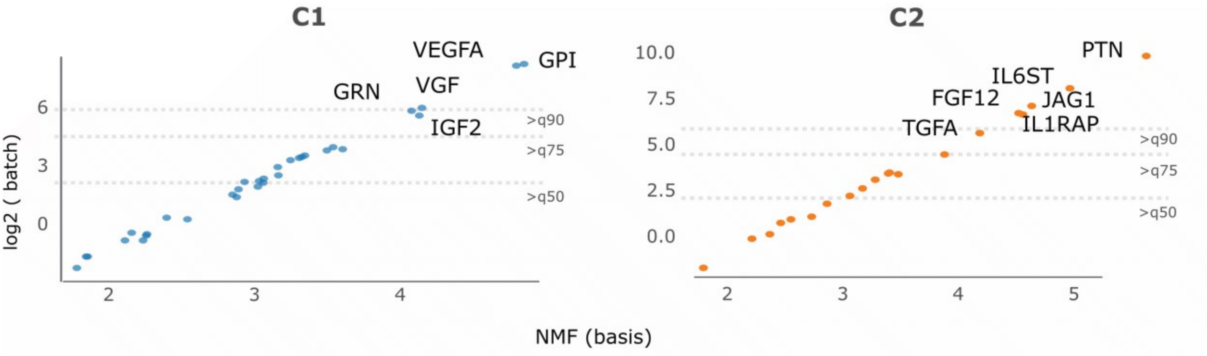

Figure R19. Evaluation of the upstream regulators active in each cluster of the CAPE integration between the expression profiles of in vivo and in vitro IDH1$^{wt/mut}$ GSA models. The scatter plots depict the external factors detected in each cluster gene module. The x-axis represents the external factor, while the y-axis represents the average log2 Expression. The dashed lines represent the expression values at the q50, q75, and q90. The labels indicate the genes with an expression above q90.

2.1.2 Using CAPE and genetic tracing reporter to evaluate the PN-to-ME transition in GSA models in vivo

The analysis of single-cell expression profiles of glioblastoma patients revealed multiple tumor states within a single patient [51, 52]. Similarly, the evaluation of the differences between in vitro and in vivo GSA models revealed an increase in tumor heterogeneity upon engraftment (see **Results 2.1.1**). However, it remains difficult to dissect and evaluate the individual glioblastoma subtypes in GSA models. To this end, we created a set of glioblastoma subtype-specific tracing reporters to assess the acquisition of different subtypes in the cells [107]. To validate these changes, we isolated cells expressing the genetic tracing reporter and generated expression profiles under different in vivo and in vitro conditions. Then, I evaluated these changes by comparing the expression profiles between conditions. To assess the model identity, I used CAPE to integrate the expression profiles of the new in vivo GSA models and glioblastoma patients.

We developed a genetic reporter that recapitulates the transcriptional signaling of glioblastoma subtypes [107]. Briefly, these reporters, named synthetic locus control regions (hereafter sLCRs), are formed by adding multiple sequences of cis-regulatory elements (or CREs) followed by a fluorescence reporter (**Fig.R20A**). CREs are derived by identifying DNA regions within the locus of a set of signature genes (e.g., MES signature genes) that are enriched in transcription factor binding sites (or TFBSs) associated with the target phenotype (e.g., MES TFs). Consequently, whenever the TF(s) bind in any of the CRE regions, the signal activates the fluorescence reporter, signifying the activation of the phenotype. To evaluate glioblastoma heterogeneity, we generated several IDH1^{wt} GSA cell lines integrating glioblastoma subtype sLCRs (see **Methods**). In particular, we used the MGT#1 reporter to assess the transition of the GSA models to the ME subtype. For that purpose, we generated several transcriptome profiles of in vitro and in vivo GSA models after selecting the MGT#1^{high} population using fluorescence-activated cell sorting (FACS) (**Fig.R20B**).

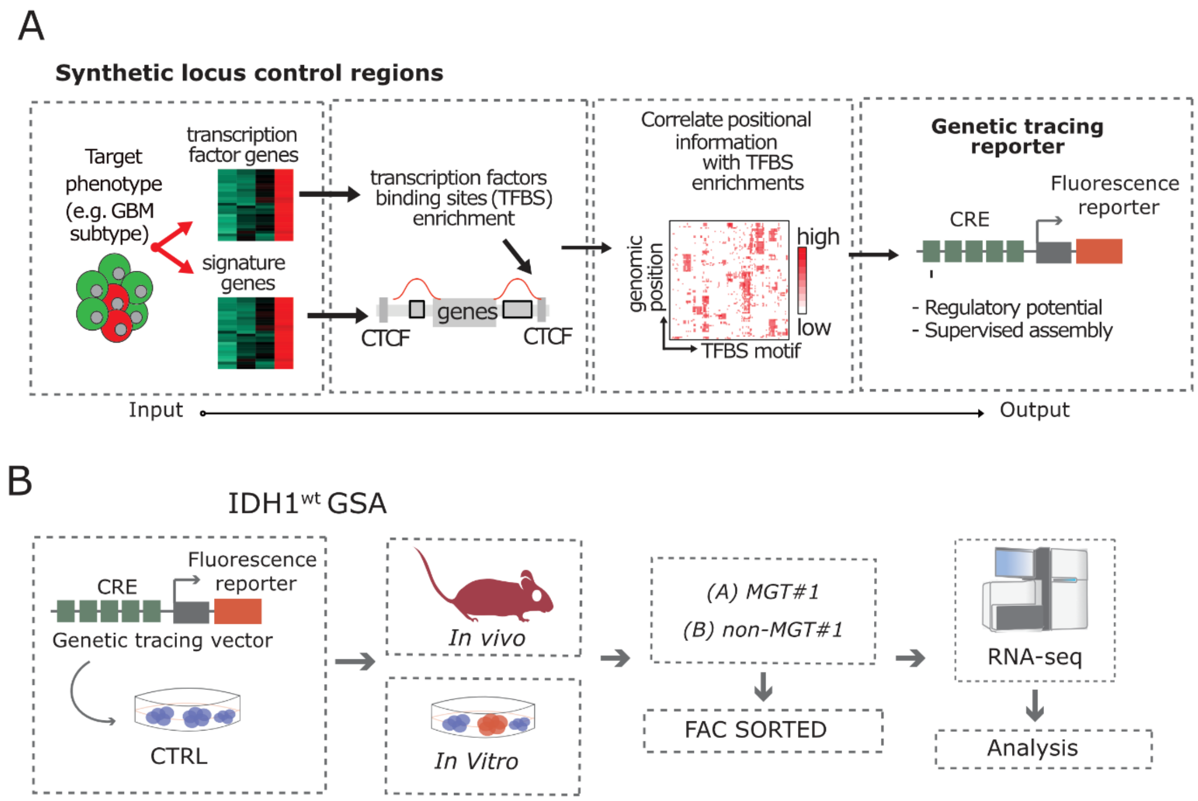

Figure R20. **(A)** Scheme of the sLCR conceptualization. (**B**) The scheme represents the generation and analysis of the expression profiles of in vivo/in vitro IDH1-$^{wt/mut}$ GSA non- and MGT#1 samples.

I compared the transcriptome profiles of MGT#1 and non-MGT#1 IDH1^{wt} GSA under in vitro and in vivo conditions to identify changes corresponding to ME subtype acquisition. First, I identified the upregulated genes of each condition using differential expression analysis (see **Methods**). Then, I performed gene sets enrichment analysis using glioblastoma gene sets [28, 35, 52] to evaluate changes in specific subtypes (**Fig.R21A**). The analysis showed that the activation of the MGT#1 reporter recapitulated the acquisition of the ME subtype under in vivo conditions. In contrast, a comparison of non-MGT#1 and MGT#1^{high} in vivo GSA revealed an enrichment of the PN subtype in the non-MGT#1 in vivo samples. Subsequently, we identified the upstream regulators activated in the in vivo MGT#1^{high} conditions in comparison to the in vitro non-MGT#1 conditions using ingenuity pathway analysis [108] (**Fig.R21B**).

The results showed the activation of the IFN□/□, NF-kB, TNF□, and IL6 upstream regulators under the in vivo MGT1^{high} conditions. In contrast, the analysis revealed the activation of the IKZF1, SREBF1/2, and SCAP upstream regulators under in vitro non-MGT1 conditions. Interestingly, these results correspond to the outcome of the evaluation of in vitro and in vivo GSA models using CAPE (see **Results 2.1.1).**

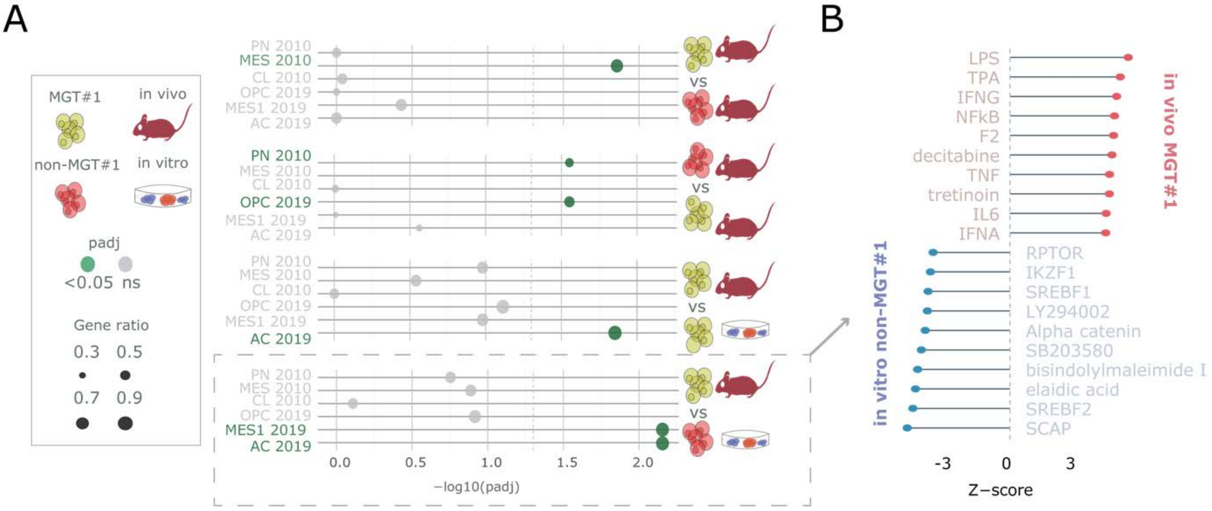

Figure R21. Comparison of MGT#1 in vivo and in vitro transcriptome profiles. **(A)** Bubble plot of GSEA adj.pvalue s for the specified glioblastoma subtypes/states and comparisons [28, 35, 52]. **(B)** Ingenuity pathway upstream regulator analysis of differential expression between in vitro non-MGT1 and in vivo MGT1 High.

After the comparison between conditions, I used CAPE to integrate the transcriptome profiles of the in vivo MGT#1^{high} GSA models and glioblastoma patients (**Fig.R22A**). Moreover, I included the in vivo GSA expression profiles (see **Results 1.2.2**) to compare the changes to previous results. I used CAPE with the same parameters as in previous integrations (see **Results 1.2.2**) and included bona fide genes only [35]. First, I evaluated the consistency of the correction before and after removing the batch effect using Spearman correlation (min. sample cor. >0.80, adj.pvalue < 0.05. Then, I used NMF decomposition to divide the dataset into four clusters based on the best cophenetic value (>0.99). The analysis showed that most of the GSA samples clustered together with glioblastoma patients in the C2 cluster. Interestingly, the analysis also revealed that some of the in vivo MGT#1^{high} samples clustered together with patients in the C1 (n=3) and C4 (n=3) clusters (**Fig.R22C**). Finally, I evaluated the enrichment of glioblastoma subtype gene sets [28, 35, 52] in each cluster-specific gene module (**Fig.R22B**). The analysis revealed that the C2 cluster (in vivo bulk GSA) was enriched with PN subtype markers, indicating similarities with previous CAPE integrations (see **Results 1.2.3**). In contrast, the C1 and C4 clusters (in vivo MGT#1^{high} GSA) were also enriched in the CL subtype and AC state, respectively. This result indicated that the FACS selection of GSA model cells using the MGT#1 reporters captured the change in glioblastoma subtype identity under in vivo conditions.

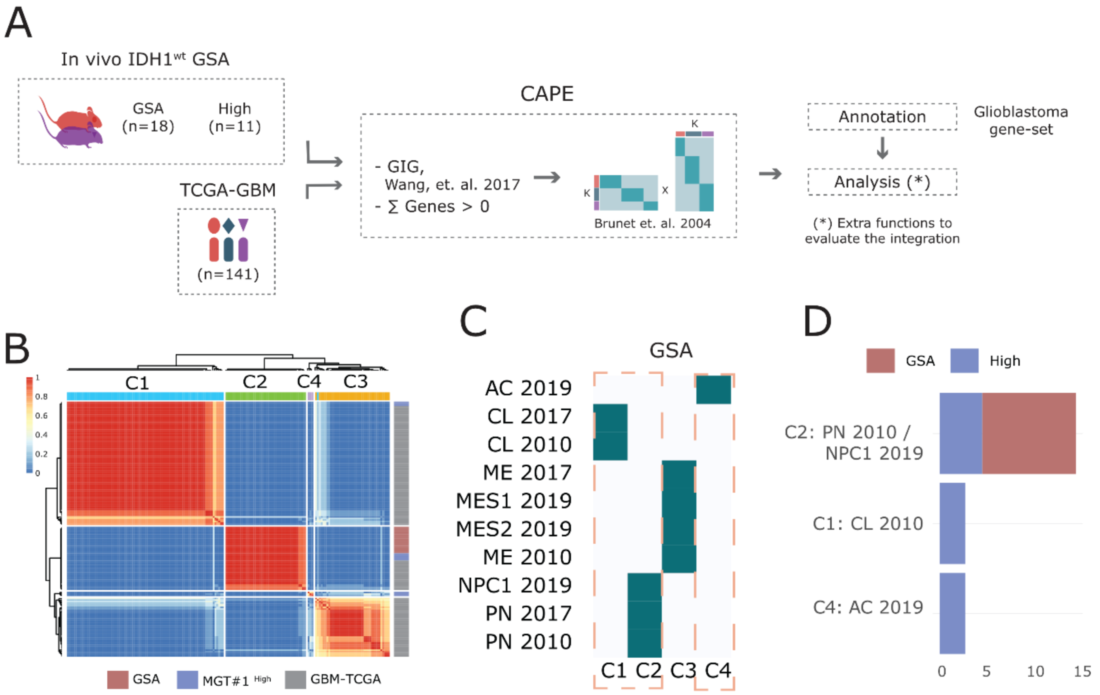

Figure R22. Evaluation of CAPE integration between expression profiles of in vivo GSA MGT1High models and TCGA-GBM patients. **(A)** Scheme of the CAPE integration. **(B)** Consensus heatmap of the NMF decomposition. The values represent the connective score between n=10 runs for the selected NMF k-factor. The colored bar with annotations represents the samples. **(C)** The heatmap indicates the significant enrichment of glioblastoma subtype gene sets for each cluster-specific gene module. The orange dashed box represents the model cluster [28, 35, 52]. **(D)** The barplot represents the total number of samples included in each group. The color indicates the dataset

2.2 Evaluation of the PN-to-ME transition upon external factor activation using GSA models

2.2.1 MGT#1 reporter activation showed in vitro the MES subtype acquisition in response to various stimuli

The glioblastoma cellular identities showed plasticity to transdifferentiation between states [52]. However, the complete set of transcriptional and regulatory elements leading to the emergence of each glioblastoma identity is still incomplete. In that sense, the GSA models displayed a solid PN identity (see **Results 1.2 & 1.3**). Thus, our models represent an excellent system to investigate changes between glioblastoma states from the PN subtype, particularly to investigate the PN-to-ME transition. To this end, we tested the effect of several external activators of the ME subtype using in vitro IDH1^{wt} GSA models and an MGT#1 genetic tracing reporter (**Fig.R23**). To evaluate the changes, we generated several transcriptome profiles of nontreated (control) and FACS-sorted MGT#1^{high} populations upon external factor activation.

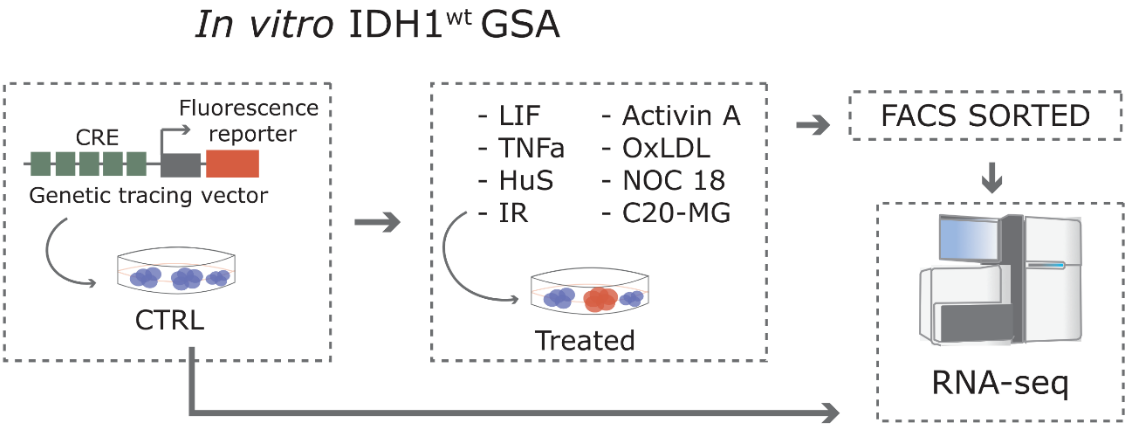

Figure R23. The scheme represents the generation and analysis of the expression profiles of in vitro IDH1^{wt} GSA naive and MGT#1 samples.

We evaluated the activation of the ME subtype in response to different external factors under in vitro conditions. These factors correspond to upstream regulators observed to be related to the ME subtype, such as LIF (see **Results 1.3.2**), TNFα (see **Results 1.3.2**), human serum (HuS) [109], oxidized low-density lipoprotein (OxLDL) [110], and activin A [111]. To include the effect of treatment and the microenvironment, we also evaluated the activation of the MGT#1 reporter under ionizing radiation (IR) [112], microglia (C20-MG [35, 113]), and nitric oxide (NOC_18) [114]. In total, the datasets of different conditions contained six control samples and twenty-five treated in vitro GSA IDH1^{wt} samples. I estimated the differentially regulated genes between the control and treated samples to assess the changes acquired upon external factor addition. The analysis showed common differentially expressed genes between several activation cues, indicating shared pathway activation. Then, I estimated sample aggregation using uniform manifold approximation and projection (UMAP) for dimensional reduction, integrating only the combination of all upregulated genes for each comparison (**Fig.R24B**).

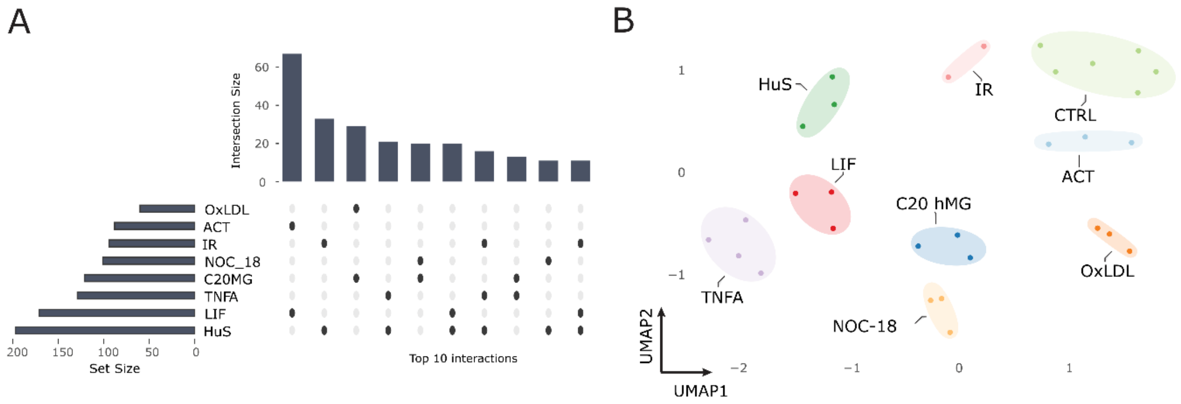

Figure R24. Analysis of upregulated genes between in vitro IDH1^{wt} GSA naive and MGT#1 samples upon activation by an external factor. **(A)** The upset graphic represents the common upregulated genes for each external factor activation. **(B)** UMAP of expression values. The dimension reduction includes all upregulated genes for each external factor.

Next, I evaluated the identity acquisition using gene sets enrichment analysis of glioblastoma gene sets [28, 35, 52] for each comparison. The results showed that, except for Activin-A, external activation of in vitro GSA models and FACS selection of MGT#1^{high} cells activated the ME subtype (**Fig.R25A**, adj.pvalue <0.05). Despite this, the analysis revealed that activation of some external factors also recapitulated markers of the CL subtype and AC state. This result is in line with previous observations using the MGT#1 reporter (see **Results 2.1.2**). Then, I evaluated the ability of the upregulated genes in each comparison to define the subtypes in glioblastoma patients. To achieve this, I computed the ssGSEA in the expression profiles of glioblastoma patients using the upregulated genes and the glioblastoma subtype gene sets as input [28, 35, 52]. The analysis indicated that OxLDL, NOC18, C20MG, IR, HuS, and TNFα upregulated genes similar to the current ME signature gene sets (**Fig.R25B**). In addition, the evaluation of the patients also revealed clustering based on glioblastoma subtypes.

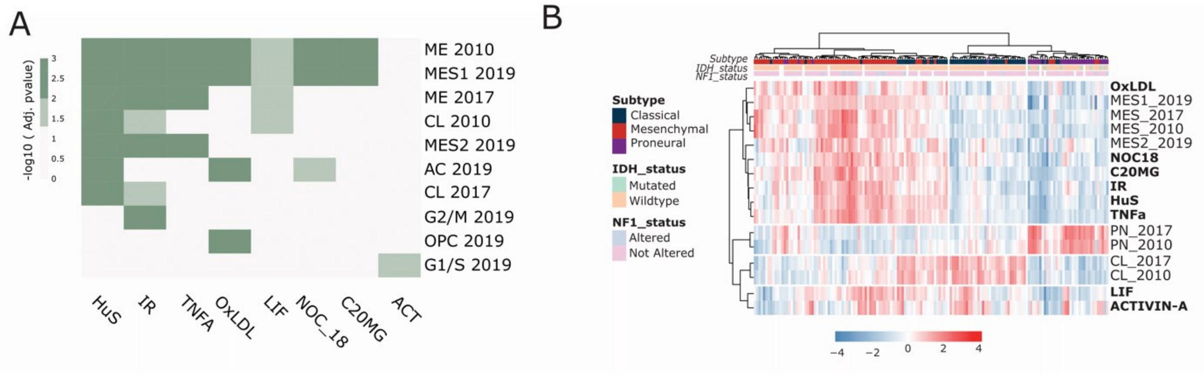

Figure R25. Analysis of upregulated glioblastoma genes in the comparison between in vitro IDH1^{wt} GSA naive and MGT#1^{High} samples upon activation by an external factor. **(A)** The heatmap represents the -log10 adj.value of GSEA results for the indicated glioblastoma subtypes/states and comparisons between the identified MES-inducing stimuli [28, 35, 52]. **(B)** The heatmap represents the relative ssGSEA normalized score for the indicated gene sets [28, 35, 52] in glioblastoma patients (TCGA-GBM). It includes the glioblastoma subtype gene sets and the upregulated genes from the comparison. The annotate bar indicates the status of the IDH1 mutation.

Next, I identified the pathways that were activated upon the addition of each external factor. First, I used PROGENy [96] to assess the expression of cancer-related pathways in the samples (**Fig.R26A**, see **Methods**). The analysis showed differences in the activated pathways upon the addition of distinct external factors. Specifically, I found that the addition of HuS distinctively activated TGFβ and MAPK signaling, and TNFα activated the TNFα and NF-kB pathways, consistent with previous results (see **Results 1.2.2**).

Finally, I identified the transcription factors controlling the activation of the MGT#1 reporter upon exposure to different stimuli using master regulator analysis [115]. This method uses an expression matrix (e.g., complete matrix) to infer the gene regulatory network and a list of genes to evaluate the MRA (e.g., upregulated genes for each external factor comparison). The assessment of MRA showed a difference in the number of TF-activated cells for each comparison (**Fig.R26B**). Then, I calculated the correlation between master regulators controlling each external factor activation using the Jaccard similarity coefficient and hierarchical clustering (**Fig.R26C**). The analysis revealed two major clusters, with one cluster containing control samples and the other not containing them. Specifically, the cluster without the control contains the HuS, IR, LIF, and TNFα TFs, indicating a similar process in response to the activation of these various external factors.

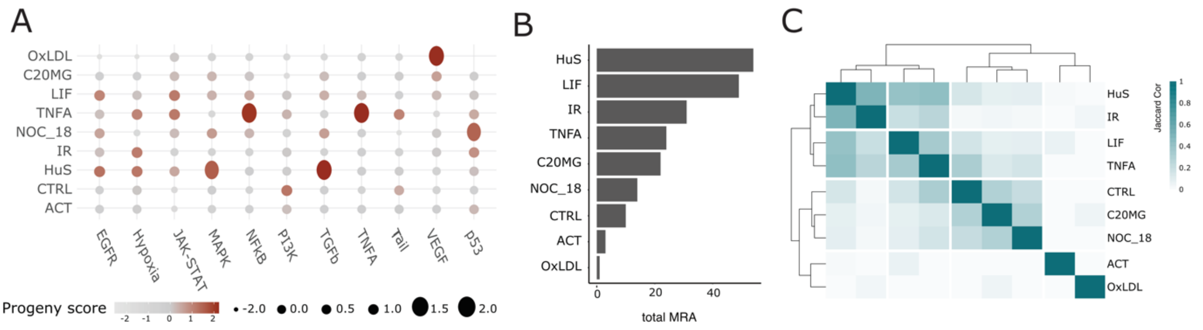

Figure R26. Analysis of the activated pathways and TF in the comparison between in vitro IDH1^{wt} GSA naive and MGT#1 samples upon activation by an external factor. **(A)** The dot plot displays the PROGENy score for each indicated pathway for each category of samples in the analysis. **(B)** Barplot with the total MRA in each comparison. **(C)** The heatmap

represents the correlation (Jaccard index values) between the master regulators identified in each category of samples in the analysis. Hierarchical clustering utilizing the Manhattan distance and complete clustering method.

2.2.2 Comparison of the effect of in vitro upstream regulators in the GSA models using CAPE

The analysis of the expression profiles of the in vitro MGT#1^{high} showed the activation of the ME subtype upon the addition of external factors (see **Results 2.2.1**). However, the pairwise comparison between control and MGT#1^{high} samples focused on individual changes in each factor and not on the comparison to patients. Therefore, I used the CAPE framework to integrate and evaluate the expression profiles of in vitro GSA MGT#1^{high} and glioblastoma patients.

The enrichment analysis of the pathways and the MRA showed differences in the effect of each external factor on the GSA models (see **Results 2.2.1**). I first used the CAPE framework to select the best external factors promoting the acquisition of the ME subtype in the models (**Fig.R27**). To this end, I used CAPE with default parameters (see **Results 1.2.2**) to integrate the expression profiles of the control and the in vitro IDH1^{wt} MGT#1^{high} GSA models. First, I evaluated the correlation for each sample before and after the transformation to assess the effect of the batch correction (min. sample cor. >.98). Then, I used NMF decomposition to evaluate the clustering of the samples. The analysis divided the expression profiles into two clusters (cophenetic value > 0.99). Specifically, the analysis revealed that the C1 cluster contained the expression profiles of the in vitro IDH1^{wt} GSA model treated with TNFα, HuS, and IR, while the C2 cluster contained the control and the remaining treated samples. I extracted the specific gene modules to evaluate the enrichment of glioblastoma subtype gene sets in each cluster. The enrichment analysis showed the acquisition of OPC/PN/CL markers in the C2 cluster (including the control samples) and the MES1/2 markers in the C1 cluster (not including the control samples) (**Fig.R28A**). This result indicated that the external factors that activated the MGT#1 reporter in the C1 cluster samples most effectively promoted the ME subtype.

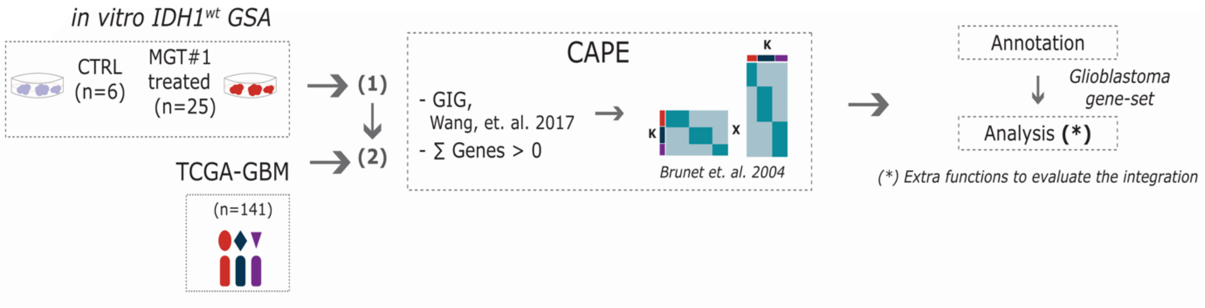

Figure R27. The scheme represents the CAPE generation between the expression profiles of in vitro IDH1-^{wt} GSA naive and MGT#1 High upon external factor activation samples.

Next, I evaluated the pathways and TF(s) correlated with each cluster. For that purpose, I analyzed the hallmark gene sets enriched in each cluster-specific gene module. The results showed that the C1 cluster (not including the control samples) was enriched in hallmarks of hypoxia, TNFα signaling, and EMT, while the C2 cluster (including the control samples) presented enrichment of hallmarks of cholesterol biosynthesis, the unfolded protein response, and oxidative phosphorylation. This result indicated distinct pathway activation for each cluster. Then, I evaluated the enrichment of TF regulon activation using the Dorothea database [103]. The results showed that the C1 cluster activated inflammatory TFs, such as RELA or NFKB1, and hypoxia TFs, such as HIF1A, while the C2 cluster was only enriched in the SREBP1 TF.

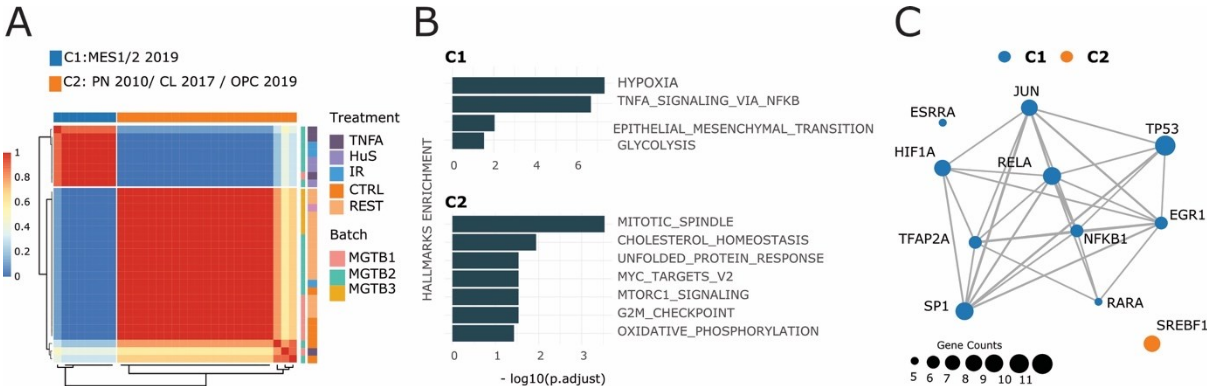

Figure R28 Evaluation of the CAPE integration between expression profiles of in vitro GSA MGT1^{High} and naïve samples. **(A)** Consensus heatmap of the NMF decomposition. The values represent the connective score between n=10 runs for the selected NMF k-factor. The colored bar with annotations represents the samples and conditions. **(B)** The bar plots represent the significantly enriched hallmarks (MSigDB v7.2, adj.pvalue <=0.05) in the cluster-specific gene modules. **(C)** The connective networks represent the significantly enriched regulon for each cluster. The grey dots denote all genes associated with TF-regulon. The dot size indicates the number of genes associated with each regulon in the analysis.

The CAPE integration showed that the HuS, IR, and TNFα samples were the best candidates to recapitulate the ME subtype identity. Therefore, I used CAPE to integrate the expression profiles of this selection of in vitro MGT#1^{high} samples and glioblastoma patients (TCGA-GBM, IDH1^{wt} only). To this end, I used CAPE with default parameters (see **Results 1.2.2**) and bona fide glioma genes only [28, 35, 52]. First, I calculated the correlation for each sample before and after transformation to assess the effect of the batch correction (min. sample cor. >0.88). Then, I used NMF decomposition to evaluate the clustering of the samples. The analysis indicated the division of the samples into four clusters (cophenetic value 0.99, **R29A**). The results showed that in vitro MGT#1^{high} samples clustered together with patients in the C3 cluster. Then, I estimated the enrichment of glioblastoma subtype gene sets in each gene module. The analysis indicated that the C3 cluster was enriched in ME subtype and MES2 state markers (**Fig.R29B**). Interestingly, the C1 cluster was also enriched in ME subtype markers. Finally, I calculated the enrichment of hallmark gene sets (MSigDB v7.2, adj.pvalue <=0.05) for each cluster-specific gene module. In particular, I evaluated the difference between the two clusters enriched in ME subtype markers (C1 and C3 clusters). The assessment of the enrichment showed that the C3 cluster (including GSA samples) is mainly driven by EMT and hypoxia, while the C1 cluster (only patients) is related to the IFNγ/α response and the inflammatory response. These analyses indicated the ability of these factors to promote the PN-to-ME transition in the GSA models.

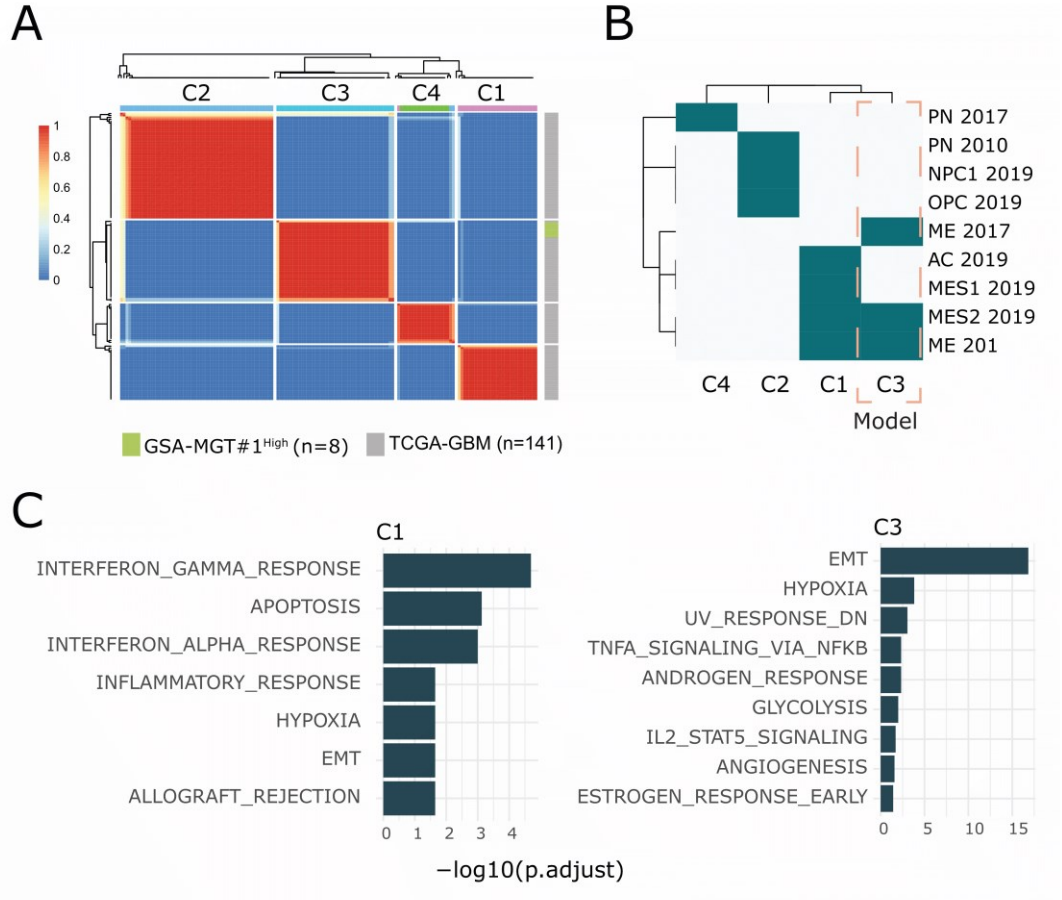

Figure R29. Evaluation of the CAPE integration between expression profiles of treated in vitro GSA MGT1High models and TCGA-GBM patients. **(A)** Consensus heatmap of the NMF decomposition. The values represent the connective score between n=10 runs for the selected NMF k-factor. The colored bar with annotations represents the samples. **(B)** The heatmap indicates the significant enrichment of glioblastoma subtype gene sets [28, 35, 52] for each cluster-specific gene module. The orange dashed box represents the model cluster. **(C)** The bar plots represent the significantly enriched hallmarks (MSigDB v7.2, adj.pvalue <=0.05) in the cluster-specific gene modules.

3. Characterization of the similarity of cell populations in GSA models and patients

3.1 Identifying the cell populations in the GSA models under various growth conditions

3.1.1 Evaluation of the in vivo GSA cell populations using scRNA-seq revealed the presence of all glioblastoma states

The comparison between the expression profiles of the in vitro and in vivo IDH1$^{wt/mut}$ GSA models showed the enrichment of several glioblastoma states upon engraftment (see **Results 2.1.1**). However, the identity of the cell populations within each model is still unknown. We generated several in vivo GSA IDH1$^{wt/mut}$ single-cell RNA sequencing (scRNA-seq) profiles to identify the cell populations that composed the models.

We obtained scRNA-seq profiles from three biological samples for each in vivo IDH1^{wt} and IHD1^{mut} GSA model (i.e., six different mouse hosts, **Fig**.**R30**). We generated the profiles using a barcoding strategy to identify biological replicates and GSA model cells within the samples (see **Methods**). scRNA-seq profiles of xenograft models (e.g., GSA models) might contain contaminating host cells (e.g., mouse) within the model cells that affect the analysis. Therefore, I mapped the scRNA-seq datasets to the human (CRCh38) and mouse (mm10) assemblies to evaluate the expression of mouse genes in the cells. To remove the source of contamination, I kept only those cells that did not express any mouse gene. In total, the scRNA-seq dataset retained 6,787 cells, from which 2,963 and 3,824 cells corresponded to in vivo GSA IDH1^{mut} and IDH1^{wt}, respectively.

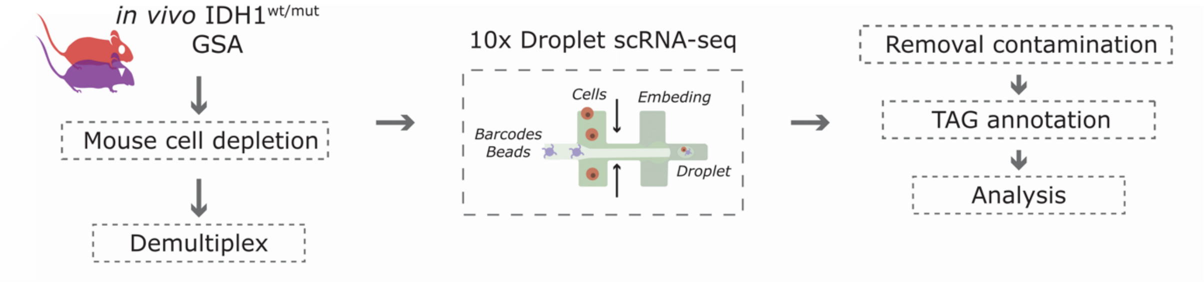

Figure R30. The scheme represents the analysis of the single-cell expression profiles of in vivo IDH1$^{wt/mut}$ GSA models (n=3 each)

I processed the scRNA-seq profiles using the Seurat v4 pipeline [116]. First, I removed low-quality cells (**Methods**) and doublets (i.e., two different cells encapsulated in the same droplet) from the scRNA-seq data using DoubletFinder [117]. In total, the dataset retained 3,578/3824 IDH1^{wt} and 2,700/2,963 IDH1^{mut} GSA cells. Then, I generated a shared nearest neighbor (SNN) graph and Louvain clustering to define the cell populations in the scRNA-seq profiles (**Fig.R31A**). The analysis showed six clusters (C0-5). To assign each cell to the IDH1$^{wt/mut}$ GSA model, I annotated the cells using the information defined by the barcoded metadata (**Fig.R31B**). The evaluation of the cells represented in each cluster indicated the presence of IDH1^{wt} and IDH1^{mut} GSA cells in all cell populations (**Fig.R31D**). To statistically evaluate this clustering, I computed the enrichment (hypergeometric test, adj.pvalue < 0.05) of IDH1$^{wt/mut}$ GSA cells in each cluster (**Fig.R31E**). The results indicated that the C1 and C2 clusters were enriched in IDH1^{mut} cells, while the C0, 3, 4, and 5 clusters were enriched in IDH1^{wt} cells.

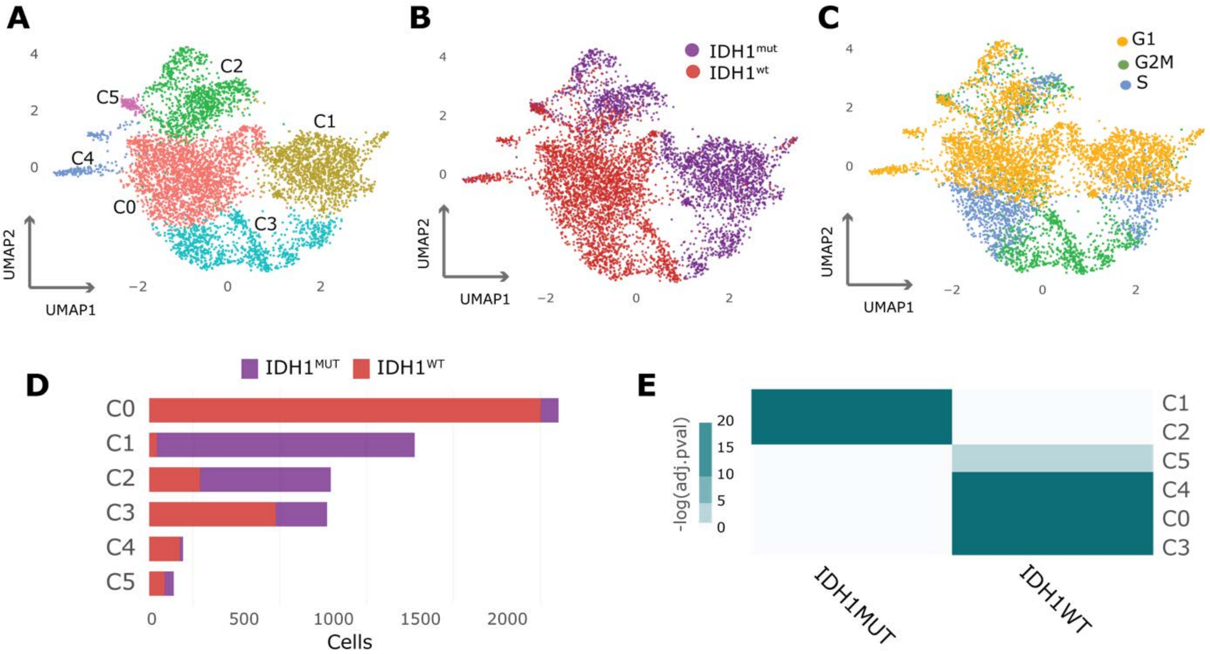

Figure R31. Evaluation of the analysis of single-cell transcriptome profiles of in vivo IDH1$^{wt/mut}$ GSA models. **(A)** UMAP representation of cell populations defined in single-cell expression profiles. **(B)** UMAP representation annotated by GSA model. **(C)** UMAP representation of individual cell-defined cell-cycle stage in the single-cell expression profiles. **(D)** The bar plot represents the number of cells for each GSA model included in the defined cell population in the single-cell profile. **(E)** The heatmap represents the enrichment (fisher.test, adj.pvalue < 0.05) for the GSA cell population in each cluster

Next, I evaluated the pathways and transcription factors activated in each cluster. To this end, I first computed the markers that defined each population using findAllMarkers [116]. Then, I estimated the enrichment of hallmark gene sets for each cluster-specific marker gene (MSigDB v7.2, adj.pvalue < 0.05, gene count >5, **Fig.R32A**). The analysis revealed distinct activation of pathways in each cluster. In particular, the C0 and C4 clusters (enriched in IDH1^{wt} cells) activated the EMT (C0 cluster), IFN□T□ response (C4 cluster), TNF□ signaling (C4 cluster) and IL6-JAK-STAT3 signaling (C4 cluster). In contrast, the C1 and C2 clusters (enriched in IDH1^{mut} cells) showed enrichment of the MYC target hallmark. Finally, I evaluated the enrichment of TF in each specific cluster marker gene using the Dorothea database [103] (**Fig.R32B**). The analysis revealed several TFs regulating each cluster, indicating specific activation.

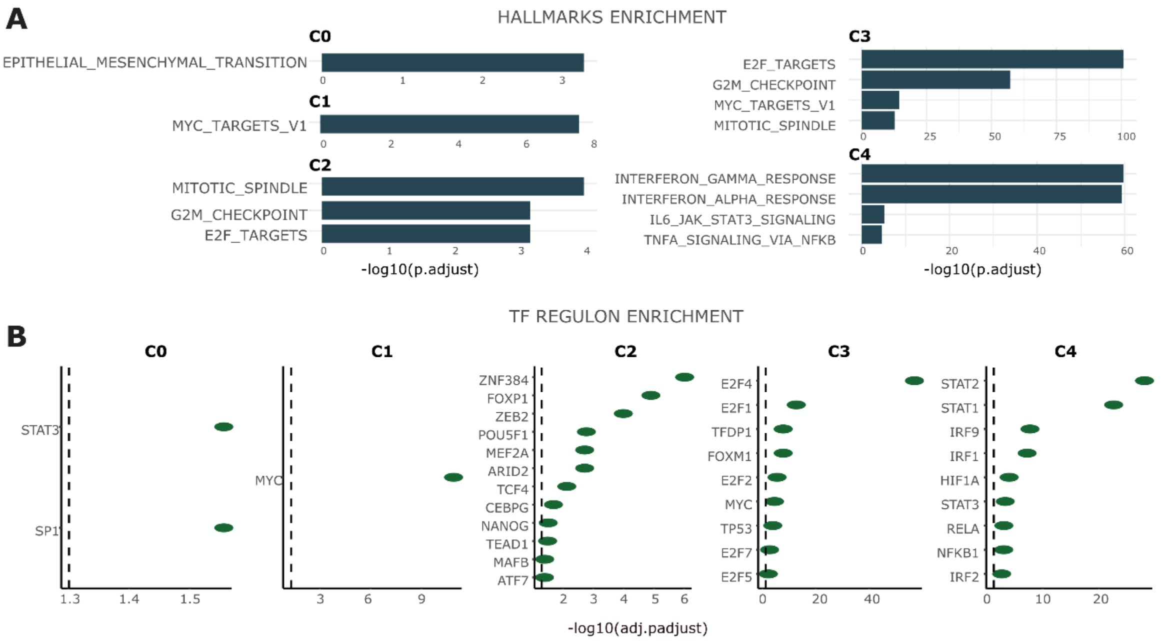

Figure R32. Evaluation of pathways and TF of the analysis of single-cell transcriptome profiles of in vivo IDH1$^{wt/mut}$ GSA models. **(A)** The bar plots represent the significantly enriched hallmarks (MSigDB v7.2, adj.pvalue <=0.05) in the cluster-specific marker genes. **(B)** The dot plots represent the significantly enriched TF-regulon (DorotheaDB [103], adj.pvalue <=0.05) in the cluster-specific gene modules.

After the evaluation of the cell populations in the in vivo IDH1$^{wt/mut}$ GSA models, I assessed the enrichment (hypergeometric test, adj.pvalue < 0.05) of each cell population in the glioblastoma gene sets. The analysis revealed that the C0 and C4 clusters (enriched in IDH1^{wt}) were enriched in AC and MES1/2 markers, while the C2 cluster (enriched in IDH1^{mut}) showed enrichment of PN subtype and NPC1-2 state markers (**Fig.R33A-B**). Interestingly, in comparison to the C4 cluster, C0 was also enriched in OPC markers. In contrast, C3 was enriched in cell cycle markers.

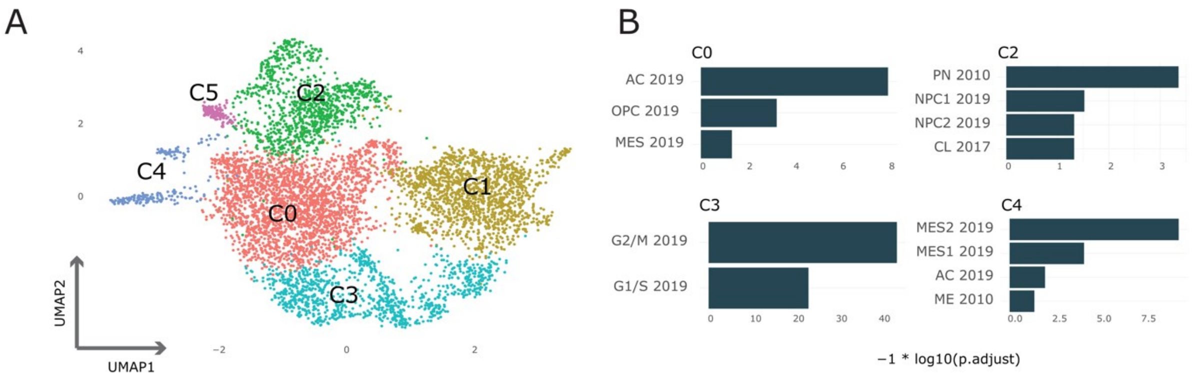

Figure R33. Evaluation of the enrichment of glioblastoma markers in the single-cell transcriptome profiles of in vivo$^{wt/mut}$ GSA models. **(A)** UMAP representation of cell populations defined in single-cell expression profiles. **(B)** The bar plots represent the significantly enriched glioblastoma subtype gene sets [28, 35, 52] (adj.pvalue <=0.05) in the cluster-specific marker genes

Finally, I used AUCell to individually evaluate the glioblastoma identity of the cells in the scRNA-seq profiles without considering clusters [72]. AUCell computes a rank-based score to assign each cell an identity defined by a specific set of marker genes. I ran AUCell to assign the glioblastoma state identity to the GSA model cells [52]. The analysis of the distribution of the AUCell score in the UMAP revealed different regions enriched (Fisher’s exact test, adj.pvalue < 0.05) for different glioblastoma states (**Fig.R34A-B**). The analysis showed that the C0 cluster (enriched in IDH1^{wt}) was enriched in the OPC, AC, and MES1 states, while the C1 (enriched in IDH1^{mut}) was enriched in the NPC1/2 states. Finally, I evaluated the enrichment of the assigned glioblastoma state by the GSA model (**Fig.R34C**). The results indicated an enrichment of OPC and NPC1-2 in IDH1^{mut}, while IDH1^{wt} was more enriched in the MES1-2 glioblastoma states.

In conclusion, for all the analyses, the dynamics of the populations of the in vivo IDH1$^{wt/mut}$ GSA models reflected the glioblastoma state distribution observed in patients.

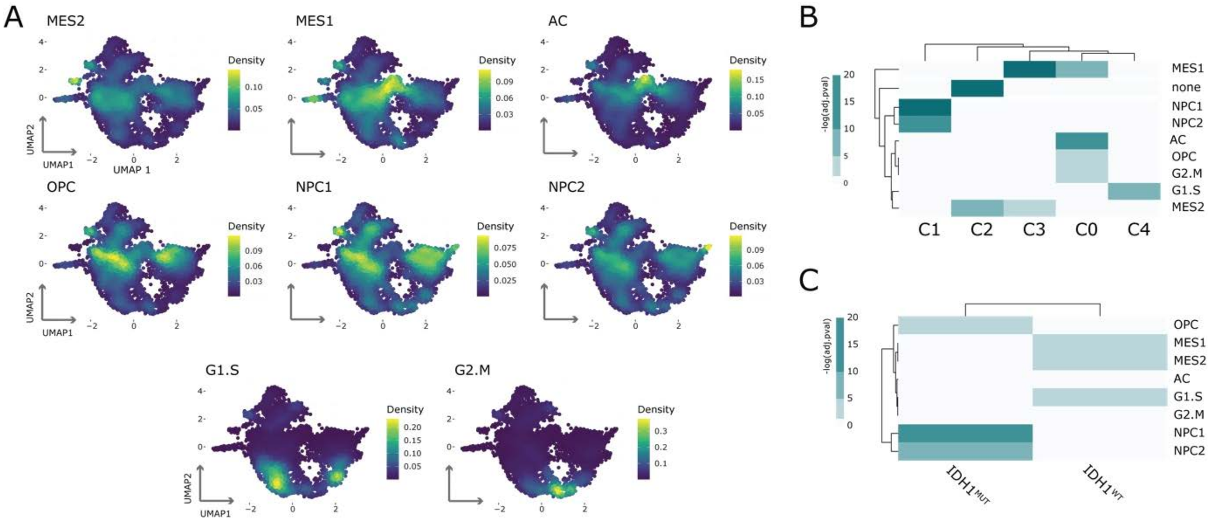

Figure R34. Evaluation of cell-specific glioblastoma gene sets in the single-cell transcriptome profiles of in vivo IDH1$^{wt/mut}$ GSA models. **(A)** UMAPs indicate the AUCell score for each indicated glioblastoma gene set. The color represents the score distribution across all cells. **(B)** The heatmap represents the enrichment (fisher.test, adj.pvalue < 0.05) for the glioblastoma gene sets [52] in each cluster identified in the single-cell expression profile using AUCell. **(C)** The heatmap represents the enrichment (fisher.test, adj.pvalue < 0.05) for the glioblastoma gene sets [52] in each GSA model identified in the single-cell expression profile using AUCell. Heatmaps hierarchical clustering was generated using Manhattan distance and complete clustering method.

3.1.2 Analysis of the ex vivo GSA scRNA-seq profiles after treatment revealed the acquisition of specific glioblastoma states

The analysis of the in vivo single-cell expression profiles of the IDH1$^{wt/mut}$ GSA models showed the predominance of the OPC and AC states (see **Results 3.1.1**). At the same time, we demonstrated the ability of the GSA models to acquire different glioblastoma states upon external factor activation (see **Results 2.2.1**). However, the changes in the GSA models at the single-cell level using external factor activation are still unknown. Therefore, we generated scRNA-seq profiles of GSA models treated with different external factors to evaluate the effect of the treatment on the heterogeneity of the models.

We generated two scRNA-seq profiles of ex vivo IDH1$^{wt/mut}$ GSA models upon external factor activation (**Fig.R35,** see **Methods**). Specifically, we treated ex vivo IDH1^{wt} GSA with external factors activating the ME subtype (TGFβ and TNFα) and ex vivo IDH1^{mut} treated with external factors activating the PN subtype (IGF2, BDNF, NRG1) (see **Methods**). In addition, we included in vitro nontreated cells as a control in the profiles.

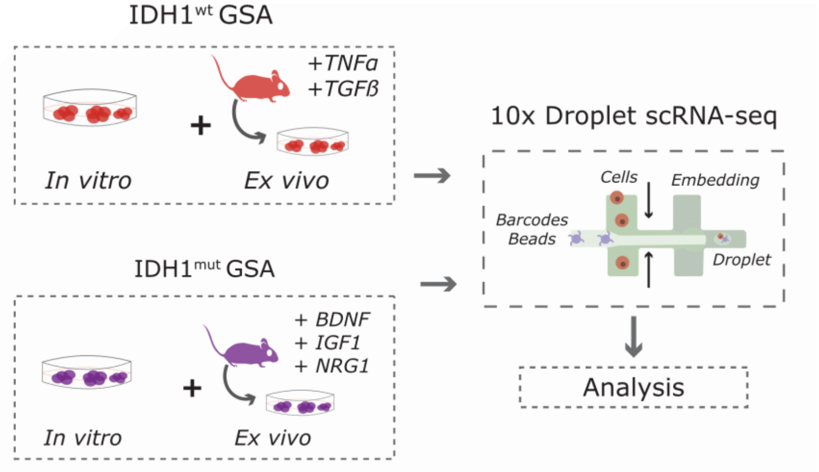

Figure R35. The scheme represents the analysis of the single-cell expression profiles of naïve and in vitro IDH1^{wt} and IDH1^{mut} GSA models upon external factor activation.

I individually analyzed each scRNA-seq dataset using the Seurat v4 pipeline [116]. First, I removed low-quality cells (see **Methods**) and doublets from the scRNA-seq profiles using DoubletFinder [117]. As a result, 3,683 ex vivo IDH1^{wt} cells and 2,885 ex vivo IDH1^{mut} GSA cells were retained. Then, I performed SNN graph and Louvain clustering to define the cell populations in each scRNA-seq profile. The analysis indicated eight clusters in the ex vivo IDH1^{wt} (**Fig.R36A**) and six clusters in the ex vivo IDH1^{mut} profiles (**Fig.R36C**).

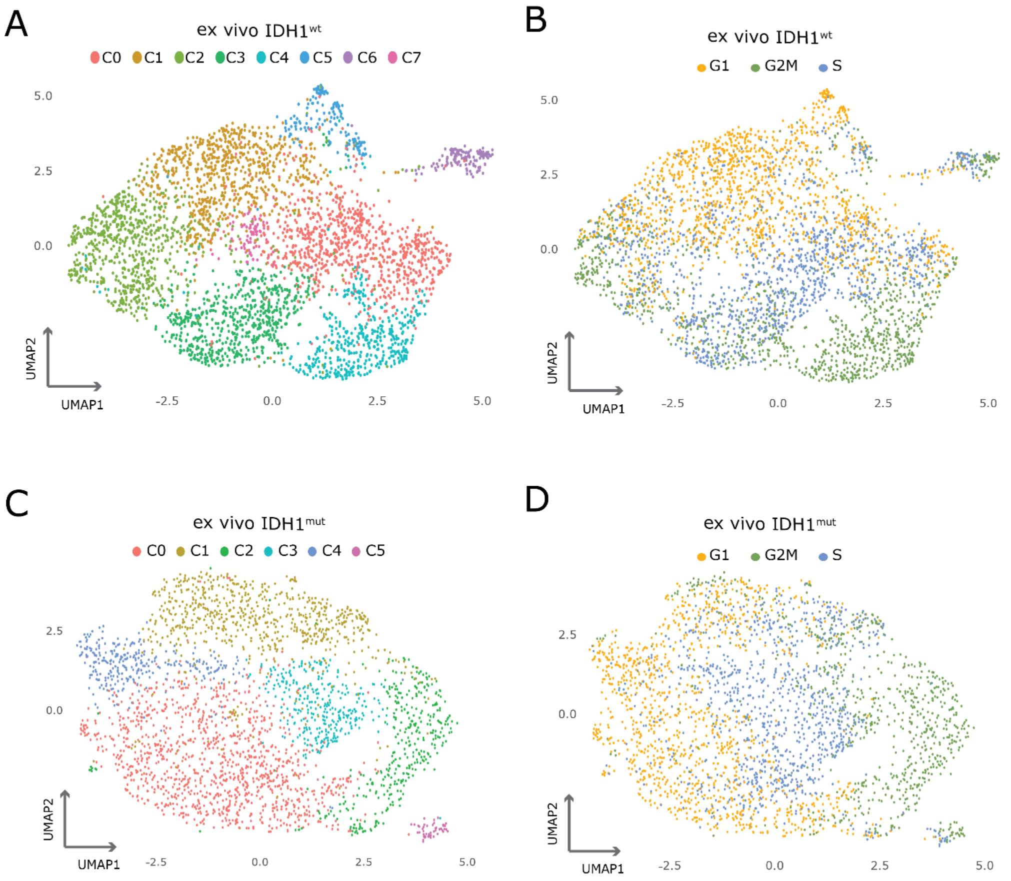

Figure R36. Evaluation of the analysis of single-cell transcriptome profiles of naïve and ex vivo IDH1$^{wt/mut}$ upon external factor activation GSA models. **(A)** UMAP representation of cell populations defined in single-cell expression profiles of IDH1^{wt} sample. **(B)** UMAP representation of individual cell-defined cell-cycle stage in the single-cell expression profiles of the IDH1^{wt} sample. **(C)** UMAP representation of cell populations defined in single-cell expression profiles of the IDH1^{mut} sample. **(D)** UMAP representation of individual cell-defined cell-cycle stage in the single-cell expression profiles of the IDH1^{mut} sample.

Next, I used AUCell to individually evaluate the glioblastoma identity of the cells in the scRNA-seq profiles without considering clusters [72]. I ran AUCell to assign the glioblastoma state [52] identity to each cell of the scRNA-seq profile of ex vivo GSA models. The evaluation of the AUCell score distribution in the UMAP of each profile revealed the activation of different glioblastoma states in the models (**Fig.R37A-D**). The analysis of the enrichment of glioblastoma states (/ ▯▯▯▯▯▯▯▯▯exact test, adj.pvalue < 0.05) showed the activation of MES1 (C7 cluster) and NPC1 (C5 cluster) states in the ex vivo IDH1^{wt} profiles and OPC/AC (C0 cluster) in the ex vivo IDH1^{mut}

profile. This result indicated the ability of the selected external factors to promote the MES (IDH1^{wt} treated) and PN identity (IDH1^{mut} treated) in the GSA models.

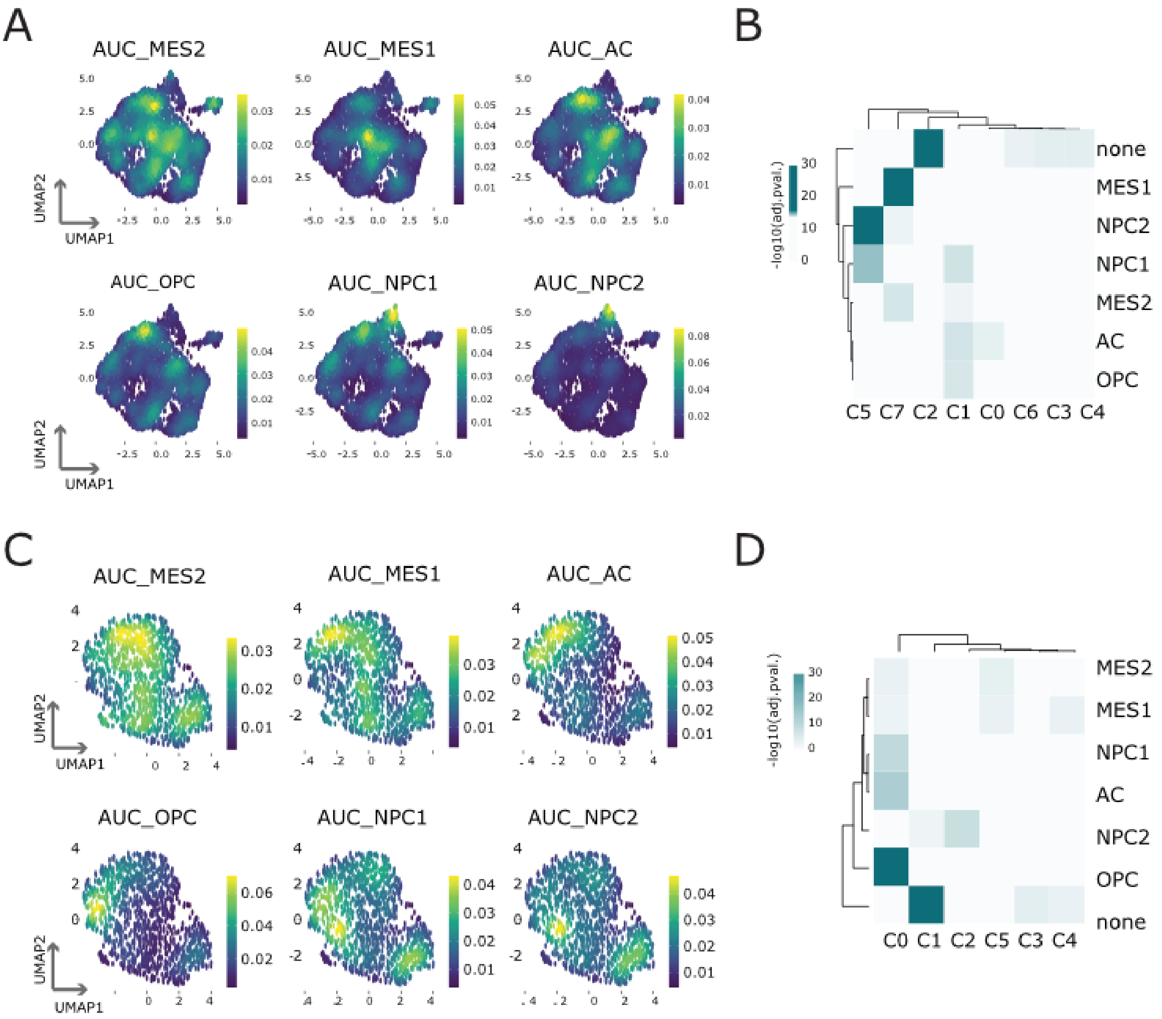

Figure R37. Evaluation of cell-specific glioblastoma gene sets in the single-cell transcriptome profiles of naïve and ex vivo IDH1$^{wt/mut}$ upon external factor activation GSA models. **(A)** UMAPs indicate the AUCell score for each indicated glioblastoma gene set of the IDH1^{wt} sample. The color represents the score distribution across all cells. **(B)** The heatmap represents the enrichment (fisher.test, adj.pvalue < 0.05) for the glioblastoma gene sets [52] in each cluster identified in the single-cell expression profile of the IDH1^{wt} sample using AUCell. **(C)** UMAPs indicate the AUCell score for each indicated glioblastoma gene set of the IDH1^{mut} sample. The color represents the score distribution across all cells. **(D)** The heatmap represents the enrichment (fisher. Test, adj.pvalue < 0.05) for the glioblastoma gene sets [52] in each cluster identified in the single-cell expression profile of the IDH1mut sample using AUCell.

Next, I assessed the activated pathways in each scRNA-seq profile to evaluate the effect of external factors. To this end, I first computed the markers that defined each population using findAllMarkers (Wilcox. Text, adj.pvalue < 0.05, log2FC > 0.5) [116]. Then, I estimated the enrichment of hallmark gene sets for each cluster-specific marker gene (MSigDB v7.2, adj.pvalue < 0.05, gene count >5, **Fig.R38A-B**). The scRNA-seq profiles of the ex vivo IDH1^{wt} GSA model showed enrichment of hallmarks for EMT, hypoxia, cholesterol homeostasis, and apoptosis in the C7 cluster (enriched in the MES1 state). On the other hand, the same analysis indicated that in the scRNA-seq profiles of the ex vivo IDH1^{mut} GSA model, MTORC1 signaling was enriched in C0 (enriched in OPCs). These results indicate the ability of the cell populations to promote the activation of a different signal for each model.

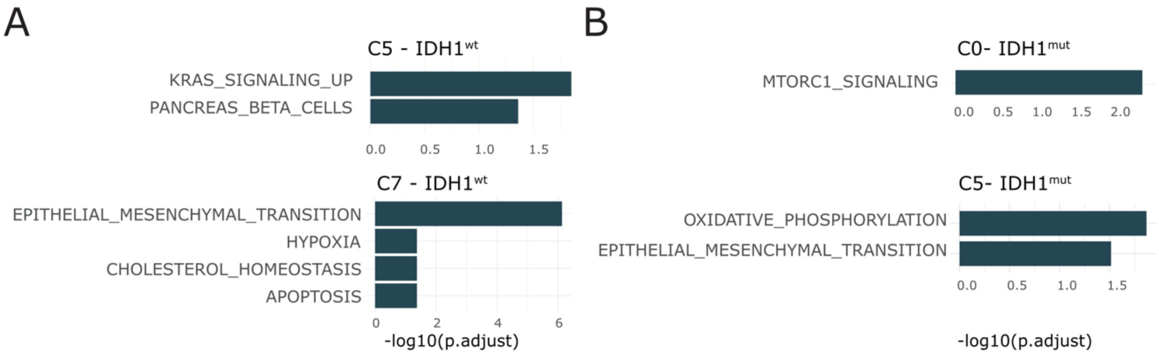

Figure R38. Evaluation of pathways and TF of the analysis of single-cell transcriptome profiles of naïve and ex vivo IDH1$^{wt/mut}$ upon external factor activation GSA models. **(A)** The bar plots represent the significantly enriched hallmarks (MSigDB v7.2, adj.pvalue <=0.05) in the cluster-specific marker genes of the IDH1^{wt} sample. **(B)** The bar plots represent the significantly enriched hallmarks (MSigDB v7.2, adj.pvalue <=0.05) in the cluster-specific marker genes of the IDH1^{mut} sample

3.2 Comparison of the GSA populations identity to patients

3.2.1 Integration of ex vivo and in vivo scRNA-seq profiles revealed differences between cell population in GSA models

The analysis of the single-cell expression profiles of in vivo and ex vivo GSA models showed the enrichment of specific glioblastoma states for each GSA model and treatment (see Results **3.1**). Despite this, a comparison of the ability to recapitulate glioblastoma states in different conditions is still lacking. To this end, I integrated the scRNA-seq profiles of the ex vivo and in vivo IDH1$^{wt/mut}$ GSA models to compare the identity of the cell populations in each condition.

I integrated and evaluated the scRNA-seq profiles of the in vivo and ex vivo IDH1$^{wt/mut}$ GSA models using Seurat v4 [108]. The high expression of cell cycling genes in the scRNA-seq profiles can mask the effects of marker genes, hindering the integration of the samples [118]. Therefore, I used only the noncycling cells to reduce the potential bias in the integration. I integrated the models using the CCA algorithm in Seurat v4 [116]. Then, I performed SNN graph and Louvain clustering to define the cell populations in the integration. The analysis revealed eight different clusters **(Fig.R39A)**. The evaluation of the dataset identity per cluster indicated a similar distribution for all the scRNA-seq profiles **(Fig.R39B)**. I calculated the enrichment (Fisher Test, adj.pvalue < 0.05) of each GSA profile within the clusters of the integration to evaluate the distribution of samples (**Fig.R39C)**. The analysis showed the enrichment of in vivo IDH1^{wt} cells in the C3-6 clusters, while the C1-2 clusters were enriched in the in vivo IDH1^{mut} and ex vivo IDH1$^{wt/mut}$ cells. Interestingly, the C0 cluster was enriched in ex vivo IDH1$^{wt/mut}$ cells, and the C7 cluster was enriched in ex vivo IDH1^{wt} cells only.

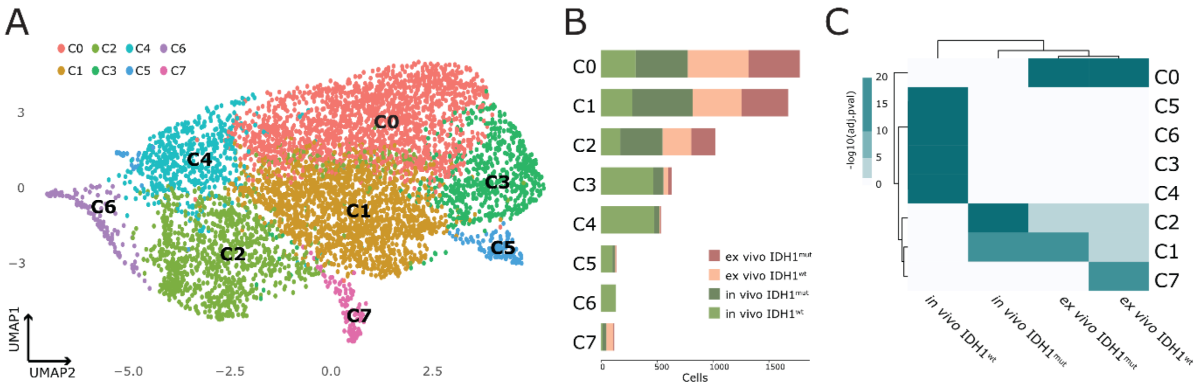

Figure R39 Evaluation of the integration between single-cell transcriptome profiles of in vivo and ex vivo upon external factor activation of IDH1$^{wt/mut}$ GSA models. **(A)** UMAP representation of cell populations defined in single-cell expression profiles**.** **(B)** The bar plot represents the number of cells included in the indicated cluster for each dataset. **(C)** The heatmap represents the enrichment (fisher.test, adj.pvalue < 0.05) for the dataset cells in each cluster

After integration, I evaluated the glioblastoma state for each cluster. To this end, I first computed the markers that defined each population using findAllMarkers (Wilcox. Test, adj.pvalue < 0.05, log2FC 0.5). Then, I assessed the enrichment of glioblastoma subtype gene sets for each cluster (**Fig.R40A**). The results indicated that the C5 cluster (enriched with in vivo IDH1^{wt} cells) is enriched in the MES1-2 state, the C7 cluster (enriched with ex vivo IDH1^{wt} cells) in the NPC1-2 state, the C4 cluster (enriched with in vivo IDH1^{wt} cells) in the AC/OPC state, and the C3 cluster (enriched with in vivo IDH1^{wt} cells) in the MES2 state (**Fig.R40B**). Then, I evaluated the enrichment by an integrated dataset. The analysis showed that NPC1-2 was enriched in ex vivo IDH1^{wt} and MES1-2/AC/OPC in in vivo IDH1^{wt} (**Fig.R40C**).

Finally, I used AUCell to individually estimate the glioblastoma identity of the cells in the scRNA-seq profiles without considering clusters [72]. I ran AUCell to assign the glioblastoma state identity to each cell in the integration of the GSA models [52]. The analysis of the distribution of the AUCell score in the UMAP showed that different glioblastoma states were enriched (/ □□□□□□exact test, adj.pvalue < 0.05) for each GSA model (**Fig.R40E-F**). The analysis showed that in vivo IDH1^{wt} was enriched in the AC/MES2 states, in vivo IDH1^{mut} in AC/OPC, ex vivo IDH1^{wt} in MES1/NPC1-2, and ex vivo IDH1^{mut} in the NPC1-2/OPC state. These results indicate that engraftment conditions promote the acquisition of AC, while the ex vivo condition recapitulates features of the NPC state.

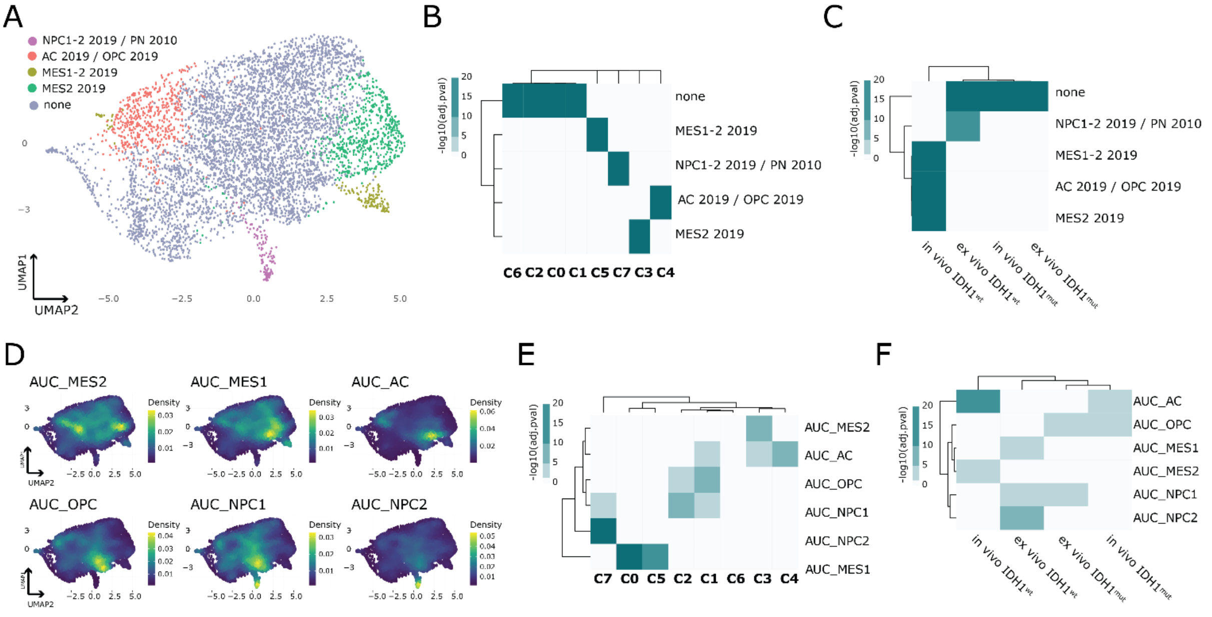

Figure R40. Evaluation of the analysis of the glioblastoma state representation of the integration between single-cell transcriptome profiles of in vivo and ex vivo upon external factor activation of IDH1$^{wt/mut}$ GSA models. **(A)** UMAP representation of glioblastoma states [28, 35, 52] enriched in the integration. **(B)** The heatmap represents the enrichment (fisher.test, adj.pvalue < 0.05) for the glioblastoma gene sets [28, 35, 52] in each cluster identified in the integration. **(C)** The heatmap represents the enrichment (fisher.test, adj.pvalue < 0.05) for the glioblastoma subtype gene sets [28, 35, 52] in each dataset included in the integration. **(D)** UMAPs indicate the AUCell score for each indicated glioblastoma gene set [52]. The color represents the score distribution across all cells. **(E)** The heatmap represents the enrichment (fisher.test, adj.pvalue < 0.05) for the glioblastoma subtype gene sets [52] in each dataset included in the integration using AUCell. **(F)** The heatmap represents the enrichment (fisher.test, adj.pvalue < 0.05) for the glioblastoma gene sets [52] in each dataset included in the integration using AUCell.

3.2.2 Integration of scRNA-seq profiles of models and glioblastoma patients corroborated GSA cellular identity

The integration of the expression profiles of in vivo GSA models and glioblastoma patients at the bulk level showed a predominant PN subtype identity in the models (see **Results 1.2**). At the same time, the modification of the growth conditions of the GSA models indicated their ability to modulate their identity upon external factor addition (see **Results 2.1.2 & 2.2.2**). In addition, the analysis of the single-cell expression profiles confirmed different glioblastoma states in vivo and upon external factor activation (see **Results 3.2.1**). However, a comparison of scRNA-seq profiles between GSA models and glioblastoma patients was still lacking. To that end, I integrated the scRNA-seq profiles of glioblastoma patients and GSA models to evaluate the correlation at the single-cell level.

First, I integrated several publicly available scRNA-seq datasets to generate a combined cohort of glioblastoma patients [29, 119, 120](see **Methods**). For that purpose, I processed each scRNA-seq profile using the Seurat v4 pipeline [108]. I removed low-quality cells and doublets from each scRNA-seq profile using DoubletFinder [117]. Then, I generated a shared nearest neighbor (SNN) graph and performed Louvain clustering to define the cell populations in each scRNA-seq profile. To integrate only tumor populations and not populations from the microenvironment, I evaluated the copy-number aberrations (i.e., hallmarks of tumor cells) using copyKat [121] and removed these cells from the profiles. Finally, to improve the integration (see **Results 3.2.1**), I retained only the noncycling cells from each scRNA-seq profile. In total, I obtained 14,745 cells from 21 scRNA-seq profiles of glioblastoma patients (8,919 cells from ten profiles [119], 4,085 from six profiles [29] and 1,741 from five profiles [120]).

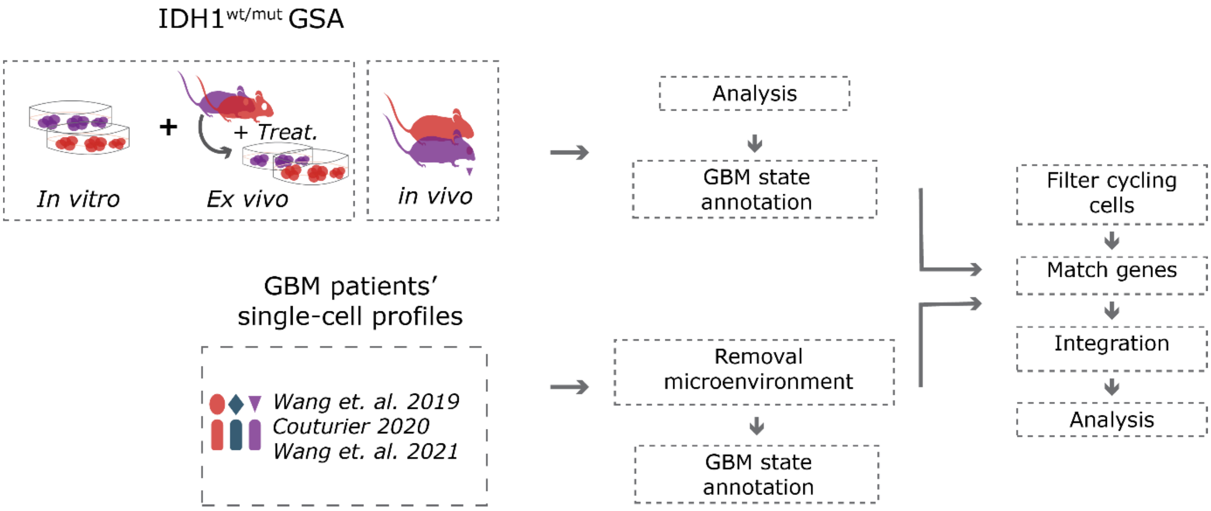

Figure R41. The scheme represents the integration between single-cell expression profiles from ex vivo/in vivo IDH1$^{wt/mut}$ GSA models and glioblastoma patients.

Next, I combined the scRNA-seq profiles of the GSA models and glioblastoma patients (**Fig.R41**). First, I combined the scRNA-seq matrix using the function createLiger from the LIGER package [122]. Then, I integrated the samples using the rPCA algorithm from the Seurat v4 pipeline [116]. After integration, I generated a SNN graph and performed Louvain clustering to define the cell populations in the integration. The analysis revealed twelve clusters (**Fig.R42A**). The evaluation of the datasets represented in each cluster revealed eight clusters with the GSA models and glioblastoma patient cells (**Fig.R42B-C**). The evaluation of the distribution of GSA model cells (**Fig.R42D**) showed the enrichment of in vivo IDH1$^{wt/mut}$ in the C5 clusters and ex vivo IDH1$^{wt/mut}$ in the C1-3 clusters. This analysis recapitulates previous differences between the models.

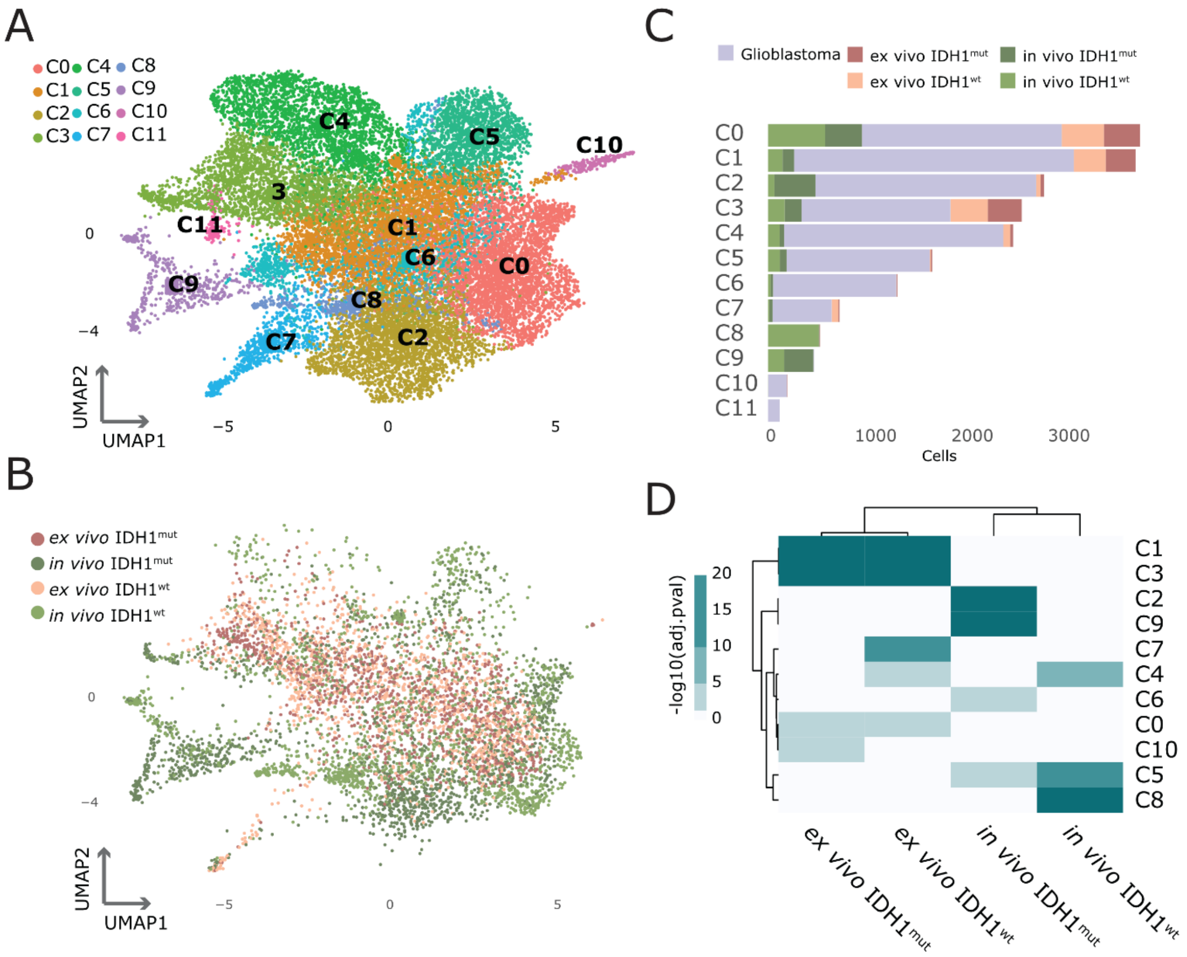

Figure R42. Evaluation of the integration between single-cell transcriptome profiles of glioblastoma patients, in vivo and ex vivo, upon external factor activation of IDH1$^{wt/mut}$ GSA models**. (A)** UMAP representation of the cell populations defined in the integration. **(B)** UMAP representation indicates the cell origin. **(C)** The bar plot represents the number of cells included in the indicated cluster for each dataset. **(D)** The heatmap represents the enrichment (fisher.test, adj.pvalue < 0.05) for the dataset cells in each cluster.

Next, I assessed the glioblastoma state for each cluster in the integration. To this end, I first computed the markers that define each population using findAllMarkers (Wilcox. Text, adj.pvalue < 0.05, log2FC >0.5) [116]. Then, I evaluated the enrichment of markers in various glioblastoma subtype gene sets (**Fig.R43A**). The analysis revealed a similar clustering of the enriched cell populations (see **Results 3.1**). In particular, the results showed that the in vivo IDH1^{wt} is associated with AC/NPC/OPC/MES1 states, in vivo IDH1^{mut} to OPC/NPC1 state, ex vivo

IDH1^{wt} to AC/NPC1-2/MES1, and ex vivo IDH1^{wt} to AC/MES1-2/NPC1-2 states.

Finally, I used AUCell [72] to assign the glioblastoma state identity to each cell in the integration [52]. The analysis of the distribution of the AUCell identity in the UMAP showed that different glioblastoma states were enriched (Fisher's exact test, adj.pvalue < 0.05) for each GSA model (**Fig.R43C-D**). The analysis revealed that the ex vivo IDH1^{wt} dataset was enriched in MES1-2/NPC1-2 cell populations, and ex vivo IDH1^{mut} was enriched in OPC/NPC, in vivo IDH1^{mut} in the AC/NPC1 cell populations, and in vivo IDH1^{wt} in AC/MES1-2. These results are similar to the previous selection validating the analysis (**Fig.R43D**).

Overall, the integration of scRNA-seq profiles of GSA models and glioblastoma patients showed similarities in the cell identity in the models and the patients. Furthermore, the analysis recapitulates previous findings from individual observations.

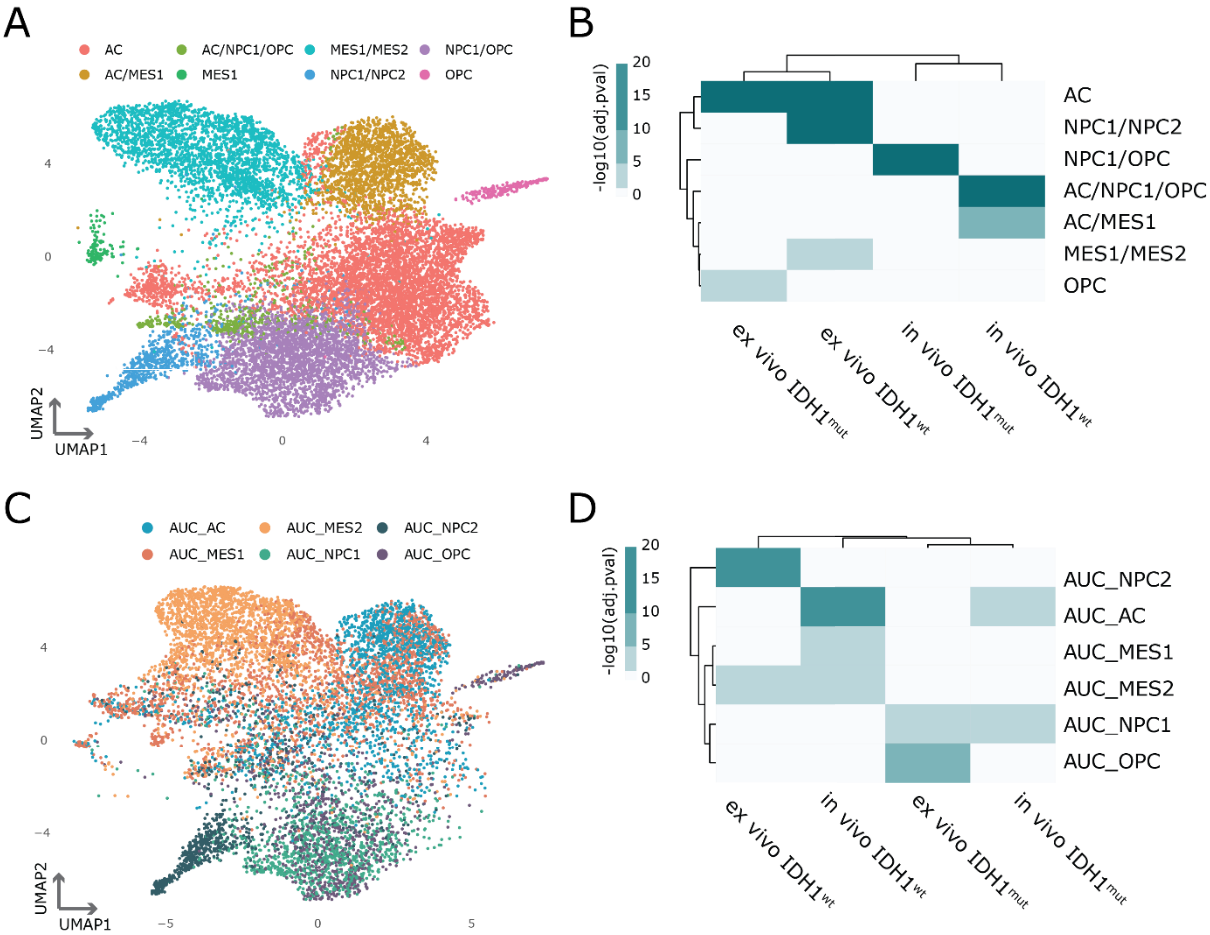

Figure R43. Evaluation of the glioblastoma state enrichment in the integration between single-cell transcriptome profiles of glioblastoma patients, in vivo and ex vivo upon external factor activation of IDH1$^{wt/mut}$ GSA models**. (A)** UMAP representation of the enrichment of glioblastoma states [52] in the integration. **(B)** The heatmap represents the enrichment (fisher.test, adj.pvalue < 0.05) for the glioblastoma state [52] for each dataset. **(C)** UMAP representation of the AUCell enrichment of glioblastoma states in [52] the integration. **(D)** The heatmap represents the enrichment (fisher.test, adj.pvalue < 0.05) for the glioblastoma state [52] for each dataset using AUCell.

DISCUSSION

1. COMPARISON OF THE MOLECULAR PROFILES OF PATIENTS AND ADULT-TYPE DIFFUSE GLIOMA MODELS

Humanized mouse tumor models exhibit tight control of the genetic profiles compared to patient-derived tumor models, thus aiding in controlled drug screenings and improving the understanding of the biology of the tumor. To answer those questions in adult-type diffuse glioma, we generated two models that represent the patient profiles by the genetic modification of nontumor-derived neural stem cells. First, I validated the presence of the genetic alterations introduced to generate the IDH1^{wt} and IDH1^{mut} models. Whole-genome sequencing analysis revealed a distinct mutational profile between the models corresponding to the expected profiles, validating the experimental design. In addition, the analysis revealed that both GSA models carried almost identical copy-number alterations (CNAs). Since the initial cells for the study originated from nontumor human cells, these findings implied that CNAs occurred during the early stages of cell immortalization. This observation might reveal processes related to the spontaneous acquisition of large genomic alterations such as Chromothripsis [15], the presence of external chromosomes [123] or altered clones in nontumor neural stem cells. Remarkably, the acquisition of CNAs parallels the early events governing the formation and evolution of adult-type diffuse gliomas [31]. Despite this, limited access to the cell of origin in these models prevents a more in-depth examination of the cause of these aberrations. Interestingly, the analysis also revealed that the 4q12 chromosomal region, which contains the PDGFRA gene, is amplified in both models. This gene is strongly associated with the glioblastoma PN subtype [28] and the OPC state [52]. In addition, GSA models also amplified the MYC and CDK4 genes associated with the NPC glioblastoma state [52]. Overall, the evaluation of the copy-number profiles suggested that GSA models were similar to the PN/NPC/OPC glioblastoma states.

Following the genetic characterization of the GSA models, the expression profiles of the models were compared to those of glioblastoma patients to determine their similarity level and identify the different active pathways. The difference between the expression profiles of tumor models and patient genomes presents a challenge that hinders the evaluation of the identity of the models. In particular, in avatar models, the similarities and differences between tumor models and patients are initially unknown. In that sense, the definition of the grade of overlap, as well as the exact tumor features of the models, remains difficult to assess at the bulk and single-cell levels. Another issue associated with this comparison is that the assessment of the expression profiles in different sequencing runs leads to technical differences that must be corrected prior to evaluating the models (e.g., batch correction). These issues imply that complex pipelines and orthogonal analysis should be used, but these methods are not always easy to implement, and the biological interpretation of the results can be difficult. To overcome these challenges, I developed a new computational framework: CAPE. This computational approach uses batch correction and NMF to integrate bulk RNA-seq datasets of models and patient genomes. Specifically, NMF deconvolution generates interpretable solutions that enable easy grouping of genes and samples into different clusters without establishing previous assumptions. To the best of my knowledge, there are only a few algorithms available to integrate expression profiles from models and patient at the bulk level, such as Celligner [67] or CancerCellNet [66]. In particular, Celligner searches for the most variable genes and applies a modified version of the mutual nearest neighbor to generate the correction. In that regard, Celligner will require the presence of similar sample types to integrate the sample [69], which may imply poor performance with only a few tumor model samples. If the correction step fails, Celligner will have difficulty defining the cluster markers and may miss some results. In contrast, CAPE relies on the identification of empirical control genes or the use of housekeeping genes to correct the batch effect in the dataset using RUV-seq [89] prior to computing NMF deconvolution. As a result, CAPE allows for the integration of smaller datasets (i.e., few model samples) but may perform worse when integrating multiple tumor types and models. In that sense, CAPE still has room for improvement by better defining the control genes used for batch correction. For example, an ideal experimental design to support the full potential in CAPE might include control genes whose expression is known, such as an RNA spike-in, to allow for correction. This approach will translate to more accurate NMF deconvolution. Despite the clear shortcomings, the CAPE framework still provides a novel, simple and adaptable method for evaluating tumor model bulk expression profiles and comparing clusters and gene modules to patients. This advantage might help us to understand complex interactions in specific tumor types or to facilitate studies such as the evaluation of PDX biobanks to improve personalized medicine-based cancer treatment.

To test the newly developed method, I used CAPE to integrate the expression profiles of in vivo GSA IDH1$^{wt/mut}$ isogenic models and TCGA-GBM IDH1^{wt} patients. The complete experimental evaluation of GSA models and the control over the cell-of-origin and genomic profiles makes these models suitable to assess the framework. First, I evaluated the batch correction using Spearman correlation and single-sample enrichment. The analysis revealed a significant correlation between the samples before and after data correction, even to potential outliers (min correlation value), confirming that, with the correct setup, the batch correction step avoided substantially altering the expression profiles. As an additional control, the glioblastoma subtype gene set score was independently estimated for each sample using a ternary graph and revealed that PN-type features were maintained after the correction step. Interestingly, the analysis showed a shift toward the

CL subtype and not to the MES subtype, indicating that the in vivo GSA models presented a certain grade of tumor heterogeneity toward a PN/CL subtype. This result might indicate that under in vivo conditions, oligodendrocyte (associated to PN) and astrocyte (associated to CL) fates are favored more than mesenchymal fate in GSA models. The clustering of the corrected matrix using NMF deconvolution revealed four clusters. One cluster grouped the IDH1$^{wt/mut}$ GSA samples, and the remaining clusters included samples from glioblastoma patients only. This result revealed that NMF finds the best deconvolution approach by including the models and the patients in separate clusters. Finally, NMF generates sample clustering and gene modules. These outputs can be used to evaluate the enrichment of the glioblastoma gene sets for each cluster. The analysis showed that PN, CL, and MES subtype markers were enriched in the clusters with glioblastoma patients, while the cluster containing the GSA models was enriched in the PN subtype. Together, the enrichment of PN markers and the division of patients according to glioblastoma subtype corroborated the predominant PN identity of GSA models.

Since CAPE relies on the identification of negative empirical genes, it might generate different outcomes depending on the input samples. To evaluate the extent of this bias over different conditions, I compared the outcome of integrating GSA models with different GBM cohorts and tumor models. Specifically, I performed two integrations, one with three patient cohort alternatives to the TCGA-GBM cohort and another integrating all patient cohorts and several tumor models. In both cases, the integration showed IDH1$^{wt/mut}$ GSA models clustered as the glioblastoma PN subtype. In particular, the integration of the expression profiles of the GSA models with other GBM models (PDX [94, 95] and GSC [92] models) and multiple cohorts of glioblastoma patients preserved the PN identity of the GSA models despite the number of different datasets. In this case, NMF deconvolution included PDX and GSC samples in the cluster with patients.

Despite this, the analysis suggested a different integration in which GSA models are integrated with patients. These findings could imply that when multiple models and cohorts are combined, the differences are established at the level of the glioblastoma subtype rather than with the models. This indicated that our models have enough similarities to PN patients but can still be improved. Overall, the outcome of these additional integrations indicated the ability of CAPE to differentiate between samples even when multiple cohorts are integrated and GSA models are clustered under the same phenotype.

In addition, to compare the results to other methods of integration, I benchmarked CAPE using Celligner [67]. In this case, the evaluation of the markers of each cluster revealed two predominant groups enriched in OPC/CL/NPC and MES glioblastoma states. Notably, most GSA model samples clustered with the OPC/CL/NPC cluster. These results showed that GSA samples still clustered together as the PN subtype using an alternative method. Despite this, Celligner had more difficulty distinguishing between the three glioblastoma subtypes. This might be explained by the method used to generate the integration, which benefits from the presence of more samples to generate the integration. In that sense, this analysis demonstrated that the cluster definition was more precise with CAPE in a smaller group of samples. This issue impacts the definition of cluster-specific gene modules, highlighting the value of using CAPE to investigate the differences between tumor models and patients.

As mentioned, CAPE NFM deconvolution also generates cluster-specific gene modules, and it presents a great advantage in comparison to other methods. In particular, the integration of models and patients helps to differentiate those pathways expressed at lower levels in the models. To confirm whether CAPE generated comparable results even with distinct cohorts, I evaluated the correlation between gene modules generated in different integrations of the GSA model and those found in glioblastoma patients using the Jaccard index and hierarchical clustering. Notably, the analysis corroborated the observation that different CAPE integrations generated similar results even with a variety of cohorts of patients. In particular, the analysis showed the clustering of the samples based on glioblastoma subtype identity. This result demonstrated the capability of CAPE to produce consistent gene modules in different integrations. However, despite this, further analysis of the correlation between gene sets might be necessary. For example, by analyzing the gene regulatory networks using different integration and conditions, CAPE might differentiate between processes and generate better results that can help improve the models.

After determining the identity of each cluster, I compared the differences between gene modules in models and patients. Since CAPE performance improved in direct comparisons, I used the in vivo GSA and GBM-TCGA integration to better evaluate the differences between the model and patient. Specifically, the evaluation of the active hallmarks in the cluster with only patients displayed an enrichment on oxidative phosphorylation hallmark not enriched in the cluster with GSA models. Remarkably, this characteristic defined a glioblastoma subtype based on metabolic classification [53]; thus, it may distinguish between PN identity in GSA models and a specific subset of glioblastoma patients. In addition to the enrichment, I estimated the tumor-related pathways in each cluster using Progeny and evaluated the enrichment of particular pathways in the samples that compose a cluster. Interestingly, this analysis showed that the GSA models activated the PIK3 and MAPK pathways, which, together with the hallmark enrichment, may indicate proliferative activity [124]. This might explain the lack of integration of our models in the patient-specific clusters. In comparison to the other cluster, the upregulation of TGFβ, TNFα, NF-kB, and hypoxia pathways in the MES subtype cluster, characteristic of the subtype, supported the validity of the hallmarks observed and indicated a more differentiated state.

The ability of CAPE to generate cluster-specific gene modules allowed us to define the difference between GSA and glioblastoma patients to improve the models. In particular, this feature enables us to evaluate specific differences in factors, such as TFs, druggable targets, and upstream regulators required for cluster activation. In comparison to other methods, such as Celligner or CancerCellNet, NMF deconvolution facilitates the generation of more interpretable solutions. Therefore, I used CAPE to integrate a cohort of glioblastoma patients using several datasets and generated a gene-regulatory network tailored to the patients [46, 92, 93]. Then, I analyzed the TF regulon enriched in the gene modules resulting from the CAPE integration of GSA models and glioblastoma patient. The analysis revealed that the model-only PN cluster was enriched in transcription factors associated with neural stem cells (i.e., SOX4 or SOX11), proliferation (i.e., E2F2-3-8), or CNV-amplified TF genes (i.e., MYC). Notably, the elements enriched in the MES cluster showed interconnected TFs, such as FOSL1/2, RELB, NFKB2/1, CEBPB, ETS2, and BCL3, related to the activation of TNFα signaling via NF-kB [125]. These TFs were previously identified in the MES subtype, validating the results. Finally, to assess the potential effect of upstream factors upregulated in the patients on modifying GSA models, I correlated the expression of specific receptors in the models with the list of growth factors from the integration. PDGFA emerged as the most important upstream regulator in the analysis. This factor has been associated with astrocytic secretion [126] activating the OPC differentiation process. In our analysis, its receptor, PDGFRA, was amplified in the GSA models. In addition, the ranked list also showed the upstream factors LIF (i.e., EMT [127]), CD70 (i.e., TNF ligand family [128]), (i.e., tumor growth), and TGFB1 (i.e., EMT). These factors are associated with the MES subtype cluster and are candidates for promoting the transition from the PN subtype to the MES subtype in the GSA models. Finally, I evaluated the coexpression of receptors and revealed that the selected factors could be grouped into five clusters. For example, the analysis revealed that LIF, IL6, or IL11 [129] could promote the transition from the PN subtype to the MES subtype in our models focused on EMT. These findings demonstrated the potential of CAPE to generate results that are conveniently accessible for improving and defining models compared to patient profiles.

Overall, in this section, I demonstrate the correlation between the expression profiles of IDH1$^{wt/mut}$ GSA models, which each have their own unique genetic profile, and the expression profiles of glioblastoma patients. To achieve this, I designed the CAPE framework. This framework evaluates the similarities between datasets more straightforwardly than currently available methods by simultaneously defining the cluster-specific identity and gene modules to improve the models. CAPE is generally useful for evaluating tumor models when the experimental design changes, as it identifies the cellular identity and the information required to improve them. This framework identified the adult diffuse glioma GSA models as PN and distinguished them from other glioblastoma subtypes in multiple comparisons. Simultaneously, it identified elements that could potentially improve GSA models. In future analysis, CAPE can benefit from a better experimental design in which external RNA is included to improve the correction. At the same time, the ability to generate gene modules can be used to build complex regulatory networks in which several factors are combined to uncover previously unknown connections between gene modules and sets of patients.

2. USING CAPE TO ASSESS THE GLIOBLASTOMA STATE ACQUISITION IN GSA MODELS

GSA models can replicate various conditions of glioblastoma tumor growth. In that sense, in vitro conditions represent a glioma stem cell state, whereas the in vivo conditions indicate a tumor differentiation state upon interaction with the environment. I integrated in vitro and in vivo IDH1$^{wt/mut}$ samples using CAPE to evaluate the transcriptional changes between growth conditions. Notably, the integration revealed a nearly perfect separation between in vitro and in vivo samples, regardless of IDH1 status or sequencing batch. This indicates that the main difference between the models is due to the environment and not genetic profiles. In addition, the evaluation also uncovered disparities in glioblastoma state acquisition. Specifically, only the in vivo cluster presented enriched markers for OPC/AC/NPC glioblastoma states. This result indicated an increase in the tumor heterogeneity of GSA models upon engraftment in comparison to the glioma initiating state. To evaluate the main differences between these changes, I focused on the evaluation of each cluster of activated pathways, TFs, and upstream regulators. The analysis showed that the in vivo cluster presented enrichment in hallmarks such as IFN response, EMT related and proliferation markers, whereas the in vitro cluster presented enrichment in hallmarks related to metabolic processes such as glucose metabolism. Similarly, the evaluation of TF enrichment linked the in vivo condition to glioblastoma states TF such as STAT1 [130], NFKB1 [28], and SOX10 [131]. In contrast, the analysis of in vitro TFs revealed regulons related to glioma stem cell maintenance, such as CLOCK [132], PRDM14 [133], and CTCFL [134]. Finally, the analysis of growth factors expressed in each cluster supported these differences between conditions. Analysis of the in vitro upstream regulators revealed that the most significant growth factor in the population is GPI, an autocrine motility factor involved in glucose metabolism [135]. Specifically, this factor has previously been identified as a factor that maintains stem cell homeostasis [135]. The most significate growth factor in the in vivo cluster was pleiotrophin (PTN), a chemotactic factor shown to promote angiogenesis and the migration of glioblastoma cells from the subventricular zone [136]. Interestingly, EGF, FGF, and PDFGA are active in vitro due to the composition of the cell culture media, which can explain the differences in proliferation and metabolic activation in vivo and in vitro, respectively. Despite this, in vitro GSA models still exhibited upregulation of another set of regulators such as IGF2, which was validated by mass spectrometry in the laboratory (data not shown), demonstrating the ability of CAPE to identify important activators. Overall, all these analyses demonstrated that in vitro populations maintain the stem cell population, whereas in vivo samples are susceptible to proliferation, EMT, and IFN modulation establishing more differentiated tumor states.

To study glioblastoma heterogeneity, the laboratory developed and tested a new approach to generate genetic tracing reporters known as synthetic locus control regions or sLCR [107]. We used this method to investigate the changes in the GSA models as they transitioned from the PN to the MES glioblastoma subtype. In particular, I assessed the differences between the expression profiles of in vitro and in vivo GSA after FACS selected samples carrying the sLCR reporter. The analysis of differentially expressed genes between conditions revealed the enriched genes for the predicted glioblastoma subtype validating the reporter selection. Notably, in the comparison between MGT#1^{high} in vitro and in vivo, the evaluation also showed an enrichment upon in vivo conditions of the AC glioblastoma state [137]. This result was in part expected due to the MES subtype including some features shared with the AC glioblastoma state [107]. Despite this, this outcome corroborated the acquisition in the GSA models of the AC glioblastoma state under in vivo conditions observed in the ternary graphics and CAPE integrations. Overall, these results indicated that GSA cells undergo astrocytic differentiation in the host brain.

Remarkably, the analysis of the upstream regulators activated by the in vivo MGT#1^{high} showed the upregulation of IL6, NF-kB and IFNγ/α signaling, similar to the results of CAPE integration of the in vitro and in vivo GSA samples. To test these changes compared to glioblastoma patients, I used CAPE to integrate in vivo MGT#1^{high} and bulk GSA samples together with TCGA-GBM IDH1^{wt} glioblastoma patient samples. Strikingly, the analysis indicated that most in vivo MGT#1^{high} GSA populations resembled the PN subtype and not the MES subtype. These results suggested a strong influence of the cell of origin or the CNA on the phenotype of the tumor models that might be difficult to overwrite without the addition of external factors. Despite this, some MGT#1^{high} samples also clustered within the CL subtype or AC state cluster. Interestingly, this outcome differs from in vivo GSA bulk integration and indicated that selecting GSA populations using sLCR helped to enrich the GSA models in some samples toward a CL/AC glioblastoma state.

The comparison between growth conditions demonstrated that in vitro GSA populations acted as glioma stem cells. In that sense, it is possible to use in vitro GSA to assess which changes contribute to the differentiation of glioblastoma subtypes. In particular, we evaluated the effect of external factors regulating the changes in in vitro GSA MGT#1^{high} cells to identify the factors that promote the PN-to-MES transition [107]. The differential expression analysis between treated and control cells revealed that, except for Activin A, all stimuli resulted in an increase in the expression of MES markers. Surprisingly, some external factors, such as IR, also activated several other glioblastoma states. For example, human serum (HuS) and ionizing radiation (IR) activated ME/MES1-2 and the CL/AC state. This finding is in line with previous analysis where in vivo GSA MGT#1^{high} also enriched CL features, indicating some granularity in the selection by the reporters. In addition, the examination of cancer-related pathways revealed disparities in the activation of various upstream regulators. Specifically, the analysis showed that HuS activated TGFβ, whereas TNFα primarily activated inflammatory pathways. In both cases, those pathways are related to the glioblastoma MES subtype. To define the TF controlling each upstream regulator activation, I analyzed the master regulator (i.e., TF) for each condition. Remarkably, the master regulator analysis (MRA) revealed a correlation between the TFs activated by each different upstream regulator. Specifically, the clustering of the correlation between upstream regulators revealed a distinct cluster of TNFα, HuS, and IR opposite to the other cluster containing the control samples. These results suggested that certain triggers promote the PN-to-MES transition more effectively than others.

I used CAPE to compare the expression profiles of in vitro IDH1^{wt} GSA control and MGT#1^{high} samples and assess the distinctions between external factor activation. Surprisingly, the integration displayed two well-defined clusters, where HuS, TNFα, and IR samples clustered together, while the remaining samples clustered with the control. Consequently, this analysis validated the previous findings made using master regulator analysis. In particular, the evaluation of the gene module indicated that the two clusters were enriched in MES1/2 and PN/OPC/CL markers. Interestingly, the MES1/2 cluster included samples treated with HuS, TNFα, and IR. Therefore, the CAPE integration showed that among all the external factors in the analysis these are the best strategies to promote the PN-to-MES transition in the GSA models. To fully evaluate this hypothesis, I used CAPE to integrate the models with more ME-like MGT#1^{high} expression profiles with glioblastoma patients. Remarkably, the integration showed that MGT#1^{high} GSA models clustered with patients in the MES subtype cluster, indicating that a change in identity in vitro in the GSA models is possible under certain conditions. This supported our previous hypothesis that MES is an activated state highly dependent on the activation of specific signaling pathways.

Overall, this analysis demonstrated the ability of CAPE to differentiate between conditions as well as to define the best strategy to promote our GSA models toward a more MES subtype identity. This section also highlights the practical usage of CAPE and the sLCRs to evaluate, improve and understand tumor models. Specifically, the analysis demonstrated the potential of combining sLCR genetic tracing reporters to select specific states in vitro upon external factor activation and CAPE to estimate their differences. In that sense, I also implemented an automatic pipeline for the generation and selection of genetic tracing reporters (see **Publications**). The integration of both newly developed frameworks can be coupled utilizing CAPE as a validation tool or to construct the input data required to design the reporters (e.g., integration of multiple cohorts of cancer patients). In general, this strategy will help to determine the model identity and allow for the evaluation of how different external factors influence the models.

3. CHARACTERIZATION OF THE SIMILARITY OF CELL POPULATIONS IN GSA MODELS AND PATIENTS

The integration of in vitro and in vivo GSA models using CAPE revealed the presence of several glioblastoma states upon engraftment. The integration at the bulk level has the advantage of the analysis of several cohorts of patients, as well as the ability to integrate new conditions of the same model in a relatively inexpensive manner. Despite this, it lacks a proper view of cell population heterogeneity and the ability to differentiate between intermediate states. To directly assess these changes, I examined the single-cell expression profiles of IDH1$^{wt/mut}$ in vivo GSA models. Interestingly, the analysis revealed that all defined clusters incorporated cellular profiles from both models. This observation indicated similarities between GSA models and explained the lack of separation in the NMF integrations at the bulk level. However, the evaluation of the enrichment of IDH1^{wt} and IDH1^{mut} GSA model cells in each cluster indicated a distinction between models. In combination, these results concluded that the GSA models showed specific differences but not enough to create a strong separation between models highly influenced by the cell of origin or the copy-number profiles. Remarkably, the enrichment of the glioblastoma gene set revealed these differences between clusters. In particular, IDH1^{wt}-enriched clusters expressed AC/MES markers, whereas IDH1^{mut}-enriched clusters expressed OPC/AC markers. Surprisingly, both clusters upregulated markers of the AC state, corroborating the previous observations at the bulk level to detect the AC state in vivo. Finally, to evaluate the cell identity without defining the clusters, I used AUCell software from SCENIC. Notably, the assessment of the AUCell score distribution in the UMAP indicated high OPC and AC scores in both models. Explicitly, the evaluation of the enrichment of cell identity within each group revealed that IDH1^{wt} enriched MES cells more than IDH1^{mut}, whereas IDH1^{mut} enriched OPC cells. Interestingly, the same evaluation by cluster showed enrichment of OPCs in IDH1^{wt}-enriched clusters, which explains the observations at the bulk level of the in vivo GSA PN/CL identity. Overall, the evaluation of single-cell profiles of in vivo GSA models corroborated the observations made at the bulk level using CAPE and genetic tracing reporters and highlighted some of the differences between models.

To assess whether the addition of external factors to the in vivo population could also generate changes at the single-cell level, we generated two ex vivo single-cell profiles by adding external factors activating the ME (TGFβ [28] and TNFα [28]) and PN (BDNF [138], IGF1 [139], NRG1 [94]) subtypes in IDH1$^{wt/mut}$ GSA models. The individual evaluation of the single-cell profiles showed different clusters for each integration, indicating the presence of different cell populations. Notably, the definition of cluster-specific markers for each individual dataset and the evaluation of glioblastoma state enrichment also indicated differences between ex vivo GSA profiles. In particular, the analysis revealed the enrichment of the OPC state markers in the IDH1^{mut} profile and in the MES1 state markers in the IDH1^{wt} profile-specific clusters. Furthermore, the evaluation of the hallmark enrichment analysis revealed different pathway activation. In the case of the MES-enriched clusters in ex vivo GSA models, the IDH1^{wt} profile activated EMT, hypoxia, and apoptosis. Specifically, EMT is enriched in the MES state. In contrast, the OPC cluster in the ex vivo IDH1^{mut} profile showed enrichment of MTORC1 signaling, which is part of the IGF1 signaling pathway [140]. Remarkably, mTOR signaling was also associated with OPC activation [141], indicating a specific activation pathway toward a more PN/OPC subtype identity. Therefore, the treatment of ex vivo GSA cells with external factors promoted the acquisition of the ME and PN subtype cell populations in the models.

To contextualize the changes in the ex vivo populations, I also evaluated the integration of scRNA-seq profiles of in vivo and ex vivo GSA models. Interestingly, the integration analysis showed cells from each single-cell profile in all the defined clusters and not individual clusters, indicating similarity between profiles. Despite this, a detailed evaluation of the enrichment of dataset-specific cells in each cluster revealed differences between GSA models. Specifically, the results indicated that in vivo GSA IDH1^{wt} cells presented more exclusive cell populations than the other profiles. In the same sense, the analysis of the enrichment of glioblastoma subtype gene sets in each cluster revealed that in vivo IDH1^{wt} was enriched in clusters upregulating MES1/2 and AC/OPC state markers, while the other profiles were enriched in individual glioblastoma states. In addition, the evaluation of the cell-specific glioblastoma state score using AUCell revealed the enrichment of MES cells in in vivo/ex vivo IDH1^{wt} profiles, whereas in vivo/ex vivo IDH1^{mut} profiles were enriched in OPC cells. Interestingly, the ex vivo IDH1^{wt} profile showed more enrichment of MES1 cells than the in vivo IDH1^{wt} profile, which can be explained by the treatment of the GSA models with TGFβ and TNFα, as was previously reported. Notably, the in vivo profiles presented enrichment for AC state cells, while ex vivo NPC. This observation corroborated the ability of the in vivo conditions to promote the AC state in the GSA models. Overall, the integration validated the individual analysis and showed the differences between the treatment, profiles and growth conditions.

Finally, the integration of the GSA model and glioblastoma cell populations was required to evaluate the identity of the single-cell profiles of GSA models. First, to generate a reference dataset of patients, I compiled and analyzed multiple publicly available glioblastoma single-cell expression profiles [29, 119, 120]. I combined only noncycling cells from each dataset to limit the effect of the cell cycle on the integration of scRNA-seq profiles of glioblastoma patients and GSA models. The integration revealed different clusters in which GSA models were integrated into nearly all clusters. Notably, the enrichment analysis indicated the presence of all glioblastoma states in the integration. In particular, the evaluation of the GSA cells enriched for each cluster indicated similarities with individual analysis. For example, the analysis revealed the enrichment of in vivo IDH1^{wt} GSA cells in AC/NPC1/OPC/MES1-enriched clusters. The enrichment analysis using AUCell showed similar results, corroborating the observation. In that sense, the evaluation of the enrichment demonstrated that in vivo/ex vivo IDH1^{wt} was enriched in MES state cells more than the in vivo/ex vivo IDH1^{mut} GSA models. Interestingly, I also observed enrichment of in vivo cell profiles in AC populations, which supported the hypothesis that AC is acquired during engraftment. This analysis corroborated the similarities between the single-cell profiles of GSA models and glioblastoma patient, indicating similar cell populations.

Overall, in this part, I demonstrated how single-cell profiles replicated previous observations at the bulk level using CAPE and sLCR selection. Interestingly, the analysis at the single-cell level reveals the predominance of some profiles for specific glioblastoma states not captured by the bulk profiles. In particular, the evaluation of the in vivo GSA model profiles revealed a predominant AC/OPC population in both models, similar to what was observed using CAPE but with a preference for the MES subtype in the in vivo profiles. This finding indicates that bulk RNA-seq helps identify the predominant population but not small changes. Nevertheless, CAPE was able to differentiate the PN-to-MES transition under the correct treatment, which still underscores the ability of the framework to generate results at a lower cost. Similarly, the ex vivo GSA-generated profiles showed enriched cell populations in the expected MES and PN subtypes, similar to the results obtained using genetic tracing reporters and external factor activation. Overall, the integration of in vivo and ex vivo GSA model and glioblastoma patient genome profiles confirmed the cellular identity observed in the individual analyses.

AUTHOR CONTRIBUTIONS

I declared that I wrote the dissertation, designed the figures and included and research all the information presented in the document. Despite this, the work presented in this dissertation are the result of a collaborative effort by several colleagues, and I would like to emphasise this fact and declare my own contribution to properly recognise the work of others:

- I developed the computational framework CAPE framework. Also, I created, implemented, and completed all computational analyses showed in this dissertation, as well as to their interpretation. I contributed to the experiment design of the different experiments, including whole-genome sequencing, single-cell transcriptomics and bulk transcriptome sequencing.

- Yuliia Dramaretska (MDC Berlin, AG Gargiulo) contributed experimentally to grow the cell, extraction the RNA/DNA material, and library preparation for all the sequencing experiments showed in the dissertation (see **Results**). Andreas Göhrig, Pilar Sanchez, Melanie Großmann, Michela Serresi, Gaetano Gargiulo, Matthias Schmitt (MDC Berlin, Gargiulo Lab) and Danielle Hulsman (NKI Amsterdam, Netherlands) contributed to the generation of the in vitro and in vivo experiments (see **Results 2**)

- Massimo Squatrito (CNIO Madrid, Spain) contributed to develop the ssGSEA analysis (see **Results 2**, **Fig.R25B**) and Iros Barozzi (Imperial College, London) originally developed the sLCR algorithm (see **Results 2**, **Fig.R20**) which was then adapted by me.

- Gaetano Gargiulo (MDC Berlin, Gargiulo Lab) developed the GSA models, GSA B1 RNA-seq , developed the sLCR concept, designed and supervised the experiments, and interpreted the data (see **Results**)

PUBLICATIONS

[1] **Company, C**.*, Schmitt, M. J.*, Dramaretska, Y.*, Kertalli, S. , Ben J., Barozzi, I., Gargiulo, G. (2022). Logical design of synthetic cis-regulatory DNA for genetic tracing of cell identities and state changes. In preparation

[2] Schmitt, M. J.*, **Company, C**.*, Dramaretska, Y.*, Barozzi, I., Göhrig, A., Kertalli, S., Großmann, M., Naumann, H., Sanchez-Bailon, M. P., Hulsman, D., Glass, R., Squatrito, M., Serresi, M., & Gargiulo, G. (2021). Phenotypic Mapping of Pathologic Cross-Talk between glioblastoma and Innate Immune Cells by Synthetic Genetic Tracing. Cancer Discovery, 11(3), 754–777.

[3] van den Berk, P., Lancini, C.*, **Company, C*.**, Serresi, M., Sanchez-Bailon, M. P., Hulsman, D., Pritchard, C., Song, J. Y., Schmitt, M. J., Tanger, E., Popp, O., Mertins, P., Huijbers, I. J., Jacobs, H., van Lohuizen, M., Gargiulo, G., & Citterio, E. (2020). USP15 Deubiquitinase Safeguards Hematopoiesis and Genome Integrity in Hematopoietic Stem Cells and Leukemia Cells. Cell reports, 33(13), 108533.

[4] Serresi, M., Siteur, B., Hulsman, D., **Company, C**., Schmitt, M. J., Lieftink, C., Morris, B., Cesaroni, M., Proost, N., Beijersbergen, R. L., van Lohuizen, M., & Gargiulo, G. (2018). Ezh2 inhibition in Kras-driven lung cancer amplifies inflammation and associated vulnerabilities. The Journal of experimental medicine, 215(12), 3115–3135.

(*) Authors contributed equally to the work

METHODS

1. GENERATION OF THE MOLECULAR PROFILES

Generation of GSA models

Generation of GSA models

Briefly, the laboratory-generated two adult-type diffuse glioma models: GSA IDH1mut and IDH1wt, by transforming human neural stem cells derived from human sub-ventricular zone samples of non-glioma patients (NPC-hSVZ, provided by R. Glass, LMU, Munich, Germany). The models contain the knockouts and knock-downs outlined in Results 1.1.1. In particular, the IDH1mut was modified with pLenti6.2/ V5-IDH1-R132H (supplied by Hai Yan, Duke University, Durham, North Carolina), p53R173H, and p53R273H (provided by D. Peeper). The IDH1wt GSA model was generated by transforming the same NPC-hSVZ with Prospero-sh-PTEN, pLKO.1-sh-TP53 (TRCN0000003754), and IRS-shNF1 constructs.

NOD mice (Jackson Laboratory, NOD.Cg-Prkdcscid Il2rgtm1Wjl/SzJ mice) were used for research involving orthotopic glioma xenografts. The study included both male and female mice. In general, the laboratory utilized mice aged 7 to 12 weeks. Xenograft studies of adulty-type diffuse glioma were conducted by transplanting in vitro GSA model cells orthotopically into the brains of mice. The tumor was removed if there was no neurological signal 5 to 8 weeks after the injection. Brains were collected immediately after euthanasia and examined with a fluorescence microscope to determine the tumor size and presence (data not shown). Xenografted tumors were processed the same day for FAC analysis and sorting, or cells were frozen in a medium containing 10% DMSO until needed. All in vivo experiments are conducted in accordance with a protocol approved by the

Institutional Animal Care and Use Committee and European Union regulations.

Generation and processing of WGS profiles for GSA models in vitro

Generation of GSA whole-genome sequencings

One In vitro sample of IDH1wt (100 ng/ul) and IDH1mut (120 ng/ul) GSA models were used to generate the whole-genome sequencing profiles. Briefly, the laboratory extracted gDNA from each GSA model and segmented the genome using Tn5 tagmentation. Then, the library was prepared using the fragmented DNA and sequenced using NovaSeq. The IDH1^{wt} sample contained a total of 459,515,606 reads, while IDH1^{mut} contained 356,367,591 reads.

Generation of the GSA models SNV profiles

After sequencing, I evaluated the raw data quality using FastQC v0.11.8. Then, I applied to trim galore (parameters: --paired --nextera) to remove the sequencing adapters. Next, I mapped the raw sequence to the human genome (GRch38) using the bwa mem algorithm [78]. After the alignment, I used the GATK v4 pipeline [79] to call the SNV. First, the mapping files were sorted using the SortSam function from Picard (parameters: VALIDATION_STRINGENCY = STRICT). Then, the duplicated reads were eliminated using the MarkDuplicates function. To identify known variants and re-align the sequence, I used the BaseRecalibrator and ApplyBQSR functions of the GATK v4 pipeline [79]. I excluded known variants (parameter: --known-sites) in both functions. Then, I used Mutect2 [79] to call the specific SNV (parameters: true -L -L somatic-hg38-only-gnomad.vcf.gz --panel-of-normals 1000g pon.hg38.vcf.gz). Specifically, I used the Funcotator repository (ftp://gsapubftpanonymous@ftp.broadinstitute.org/bundle/funcotator/) to define SNV variants and removed the known variants. After this step, I followed the pipeline to collect information to help the final variant calling. Then, I applied the GetPileupSummaries (parameters: --variant somatic-hg38-af-only-gnomad.vcf.gz) and LearnReadOrientationModel function to generate the files to clean the variants. Finally, I filter those low-quality variants using the output of the previous functions. As a result of running the GATK v4 pipeline, I generated two files, one in the VCF format and the other in MAF format, for each model.

Generation of the GSA models CNA profiles

Evaluation of the copy-number profiles used as input the deduplicated alignment file (above). To call for Copy-number alterations I used CNVKit [84]. Each final call was annotated with the UCSC GRch38 2013 value and To generate the CNV cnvkit.py batch was used (parameters: --normal control.bam --specifically targets exon.baits.annotate.bed --mark refFlat GRCh38 2013 UCSC.txt as annotated --fasta Human Assembly 38 Filter.fasta diagram --disperse -m wgs). I used as a control file the whole-genome sequencing iPS-derived NSC. To compare the calls with the TCGA whole-exome sequencing profiles, generate baits.py was used to generate only exome regions. Finally , to compare with TCGA-GBM the segmented files were re-process using GISTIC2 [142]. CNV for each model IDH1^{wt} 31 amplifications and 30 deletions and IDH1^{mut} 26 amplifications and 22 deletions.

I used the R v4.0 environment to evaluate and represent the data. In particular, the SNV profiles were processed using maftools v2.6.05 [143] using the maf files as input. The evaluation of glioblastoma patients was generated after downloading profiles using TCGAbiolink [144] (Mutect2 Samples=392).

Generation of RNA-seq profiles of GSA models

Generation of the RNA-seq profiles

The Bulk in vivo GSA samples (**Results 1.2, 1.3, and 2.1**) were processed in two separate sequencing runs. At NKI Amsterdam and at the MDC Berlin. Briefly, following orthotopic transplantation, six IDH1^{wt} and IDH1^{mut} samples were sequenced. RNA was extracted in this instance using the Trizol protocol. Then, NextSeq 500 was used to generate the profiles. Schmitt et al. [107] describe how RNA-seq samples for **Results 2.1.1** and **2.2** were produced.

After sequencing, fastq files were evaluated with FastQC v0.11.8. The adapters were trimmed using skewer v0.2.2 (default parameters). Then, I mapped the reads to the GRch38 human assembly (TCGA assembly G200 GRCh38.d1.fa) using STAR v2.6.0c (parameters: --out FilterMultimapscores within range 1 FilterMultimapNmax maximum of 20 --out --align FilterMismatchNmax to 10 --align IntronMax 5000000 1000000 --sjdbScore 2 --align MatesGapMax SJDBoverhang --limit Min 1 --genomeLoad NoSharedMemory BAMsortRAM --readFilesCommand output --out FilterMatchNminOverLread 0 --sjdbOverhang 200 --out FilterScoreMinOverLread 0 --sjdbOverhang 0 --out SAMstrandField intronMotif SAM attributes NH Within HI NM MD AS XS –out SAMunmapped --out SAMtype BAM SortedByCoordinate --twopass1reads --twopass1reads N -1 Basic --twopassMode). Finally, I used HTSeq package [145](parameters: -s reverse -i gene name) to extract the gene counts per samples (gencode.v22.annotation.gtf)

Generation of the glioblastoma patient's expression profiles

To generate a multi-dataset cohort of adult-type diffuse gliomas, I downloaded and processed the data associated with each indicated dataset [46, 92, 93]. First, the data was downloaded from SRA. For the TCGA data, HTSeq count profiles were downloaded from https://portal.gdc.cancer.gov/ and integrated as a count matrix. Then, I processed each patient cohort using the same pipeline to process the GSA models' RNA-seq profiles (above). Only IDH1^{wt} samples were considered in the integration.

Generation GSA models scRNA-seq profiles

Generation of the in vivo scRNA-seq

After orthotopic transplantation, In vivo GSA cells from three IDH1^{WT} and three IDH1^{mut} tumors were isolated, processed, and then, sequenced using 10x Genomics Chromium technology. The GEMs (Gel Bead-In EMulsions) protocol was used to maximize the number of cells. In total, six barcodes were applied to differentiate between the six samples. The same library was generated twice to increase the number of cells in the analysis. The raw sequences were mapped to the genome and counted using 10x Genomics Cell Ranger v6.1.2 multi-function. The GRCH38 and mm10 assemblies were used in this analysis to identify the host (mouse) and donor (human) cells.

Generation of the ex vivo scRNA-seq

Ex vivo GSA IDH1^{wt} cells were treated with TNFα 100 ng/ml and TGFβ 5 ng/ml over two days and then used for sequencing. In the same way, ex vivo GSA IDH1^{mut} were treated with NRG1 90 ng/ml, IGF1 10 ng/ml, and BDNF 100 ng/ml over two days and then used for sequencing. The library was created using 10x Genomics' Chromium technology v2. The raw sequences were mapped to the human genome (GRCH38) and counted using 10x Genomics Cell Ranger v.3.0.0.

2. ANALYSIS AND COMPUTATIONAL METHODS

Analysis the results of the CAPE integrations

The CAPE framework is detailed in the results section (see **Results 1.2**). The parameters for each CAPE integration are described in the results (see **Results 1.2 & 2.1.2 & 2.2.2**). The results were generated in the R v4.0 environment. The graphical representations were generated ggplot2 v3.35.

Description of the CAPE framework

Briefly, the framework starts by combining the input expression profiles and generating the metadata saving the information in a summarizedExperiment object using the function combineMatrix. The same function removes low-count genes defined as n counts in at least 90% of all samples (parameter: n = 0 counts). Then, it uses the cpm function from edgeR v3.26 R package [146] to correct the sequence depth and normalizes the output matrix using quantile normalization (normalize.quantiles.robust function from preprocessCore R package [147], parameters: remove.extreme=both).

As previously stated, the RUV-seq [89] algorithm within the nmfBatchCorrection function is used for batch correction of the datasets (see **Results 1.2**). The framework corrects the data after identifying the empirical genes using the RUVg function as described in the RUV-seq pipeline [89]. In brief, the CAPE generates comparisons between datasets and selects those genes with low changes in expression to define the list of genes for correction. The total number of genes can be defined and used for the correction. If the list of control genes (e.g., housekeeping genes) is known, it can be passed as input to correct the data-set instead of using the approach described in the RUV-seq documentation. After batch correction, if a list of genes is provided, only those genes are kept for further analysis. Otherwise it will use all the genes in the integration. For example, CAPE decompositions used only bona fide glioma genes from Wang et al. [35] (see **Results 1.2 & 2.1.2 & 2.2.2**). All the different corrections matrices are stored in the SummarizedExperiment and are accessible through the metadata.

CAPE clustering using the NMF v0.23.0 [90] R package. The clustNMF function uses as input the summarizedExperiment, and it generates the decomposition after defining the number of runs, the times the NMF approaches a solution, and the number of k to evaluate. In the dissertation results, the algorithm used as default Brunet et al. 2004 NMF algorithm [91], k2 to 8, and nrun=10 (see **Results 1.2 & 2.1.2 & 2.2.2**). However, these parameters are adjustable in the clustNMF function. More information about each parameter can be found in the NMF v0.23.0 [90] documentation. To evaluate the best decomposition, CAPE selects the best k by finding the highest cophenetic value (> 0.99) as defined in Brunet et. al. [91]. However, the k factor parameter can be changed by using the includeNMF function. As output, the NMF decomposition generates a consensus matrix defining the clusters of the samples and the matrix corresponding to the gene grouped by the cluster. The algorithm defines each cluster-specific gene module in CAPE as the log2FC > 1 between clusters. The includeNMF function also updates the metadata and the gene modules if a different k is indicated. All the information about the decomposition, as well as the results of the clustering, are stored within the SummarizedExperiment. The framework can be accessed through (https://gitlab.com/gargiulo_lab/cape)

Evaluation of the CAPE framework batch correction

The evaluation of the integration is determined in several steps. First, I used the normalized quantile matrix to compute the interquartile range

(IQR) by row (RowIQR) to evaluate the data distribution (see **Fig.R6**). Second, I calculated the spearman correlation using the cor R function between the quantile normalized matrix and the batch corrected matrix for all samples. These values defined the minimum and mean correlation between matrices of each dataset. Then, the significance of the correlation for each sample is calculated using the cor.test function (method=spearman). To compare the significance of each dataset, I used Fisher's combined probability test (poolr R package [148]). Finally, I compared the distribution of the samples using single-sample gene set enrichment (GSVA v1.38, method="zscore"[149]) of glioblastoma gene sets from Wang et. al. 2017 [35]. The data is scaled to the min value to remove negative values. To evaluate the results, I used ternary graphics (ggtern R package [150]).

Evaluation of the CAPE framework clustering

The NMF decomposition is evaluated in two steps. First, I extracted the NMF decomposition metrics from the summary function of the NMF v0.23.0 package [90]. The selection of the best cophenetic is obtained from this table in the summarizedExperiment metadata. Then, to graphically evaluate the cophenetic values I used ggplot2 v3.35. In the second step, the NMF decompositions are evaluated using the consensusmap function (parameters: method=complete, distance=Euclidean) of the NMF v0.23.0 R package [90].

The generation of the CAPE gene modules is described above. To evaluate the basis heatmap , I extracted the basis values using the basis function of the NMF v0.23.0 package [90], and filtered the matrix to represent only the genes defined in the gene modules. The heatmap is generated using the pheatmap function of the pheatmap v1.0.1 (clustering_method = "ward.D2", clustering_distance_cols = "correlation")

Evaluation of the CAPE framework gene modules

To evaluate the enrichment of the gene modules for different gene sets, I used the enricher function of the ClusterProfiler v.2.1.2 R package [151]. Then, the results were filtered by adj.pvalue < 0.05 and gene count >5. The gene sets used in the analysis were the glioblastoma subtype/state markers [28, 35, 52], molecular hallmarks from MSigDB v7.2 (https://www.gsea-msigdb.org/gsea/msigdb), and the TF regulons from Dorothea v1.3.3 R package [103] (I only retained used the A, B, C confidence levels from the dorothea_hs object). The representation of the enrichment results barplot used ggplot2 v3.35. The values represented the -log10(adj.pvalue) of each enriched gene set (above). To represent the TF networks, I used cnetplot function of the enrichplot v1.10.2, which uses the output of the enricher function as input.

The evaluation of active pathways was generated using progeny (parameters: perm = 100, scale = FALSE, z-scores = FALSE) function from progeny v.1.12.0 R package [96]. I used as input the corrected matrix filtered by the gene module. To compare the results for each cluster, I compared the pathway score between a cluster (formed by all the samples in the cluster) and a control (formed by samples randomly selected from the other cluster) using the t.test function of the R stats package (alternative=”greater”). Then, I correct the p-value using the p.adjust function of the R stats package (method=‶bonferroni"). The graphical evaluation was generated using the pheatmap function of the pheatmap v1.0.1 R package (clustering_method = "complete", clustering_distance_cols = " row"). The cancer-related pathways are sorted by cluster and the graphic represents the -log10(adj.pvalue).

The evaluation of the upstream regulators was generated using the Omnipath v2 database [94]. In detail, I used the import_omnipath_intercell function to import the database (https://omnipathdb.org/intercell) and focused the comparison on manually curated repositories only. Therefore, I included only the entries from CancerCellMap [99], CellPhoneDB [100], CellChatDB [101], and CellTalkDB [102] databases. Then, I filtered the tables for those annotated as cytokine, growth factor, and hormone. Next, I selected from each gene module all the ligand genes in the Omnipath table (“from” is the ligand , while “to” is the receptor). The rank of each upstream regulator was generated based on the NMF basis value (**Fig.R16**). The representation in **Fig.R19** represents the correlation between the NMF basis values and average gene expression for each cluster. Finally, to compare the ligand-receptor between GSA models and patients (**Fig.R16**), I evaluated all the selected ligands (y-axis) and the expressed receptors in the models (x-axis). The heatmap was generated using ggplot2 v3.35 and it represented the mean of Log2 gene expression values of the receptors in the model. The co-expressed ligand was evaluated based on the correlation (parameters: method=pearson) of the receptor expressed in the models. The heatmap was generated using the pheatmap function of the pheatmap v1.0.1 (parameters: clustering_method = "average",clustering_distance_cols="manhattan", clustering_distance_rows= "manhattan").

Benchmark of the CAPE results

Comparison between the CAPE integration gene modules

To correlate the gene modules from different CAPE integrations (see **Results 1.3.1** & **Fig.R11**), I computed the Jaccard similarity coefficients (jaccard function of the Jaccard v0.1.0 R package). The correlation was calculated using only the genes of each gene module in the comparison. The representation of the correlation between modules was generated using the pheatmap function of the pheatmap v1.0.1 (parameters: clustering_method = "ward.D2", clustering_distance_cols="manhattan", clustering_distance_rows= "manhattan"). The input of the heatmap was the matrix of the Jaccard similarity coefficients for comparison. To annotate the subtype, I computed the enrichment of the glioblastoma subtypes gene set in wang et. al. [35] using the enricher function from the clusterProfiler v.2.1.2 R package [151].

Integration of models and patients using Celligner

To benchmark CAPE results, I evaluated the integration of the expression profiles of GSA models and TCGA-GBM patients using Celligner [67]. In order to compare the results I integrated the expression profiles using the cpm matrix of the CAPE integration in **Results 1.2.2**. I used Celligner following the steps described in the algorithm repository (https://github.com/broadinstitute/celligner). The Global variables modified in the analysis were n_PC_dims = 5, mod_clust_res = 0.6, fast_cPCA = 20. I used the default for the other parameters. Only (not reduced to bona fide genes) The output is a Seurat v4 [116] object with the integration. To generate the Uniform manifold approximation and projection (UMAP) dimensional reduction I used the corrected matrix (**Fig.R10A**). Then, I used the umap function (parameters: n_neighbors = 20, metric = 'manhattan', min_dist = 0.1) from the uwot v0.1.11 R package. The graphical representation was generated using ggplot2 v3.35.

After the integration, I used the FindAllMarkers (parameters: test.use = 'LR') Seurat v4 R package [116] to evaluate the markers for each cluster. I filtered the results using p_val_adj < 0.05 & avg_log2FC > 1 & pct.2 < 0.75 to define the gene markers for each cluster. Then, I used enricher function (p.adjust < 0.05 & Count >= 3) of clusterProfiler v.2.1.2 R package [151] to evaluate the glioblastoma gene sets [28, 35, 52] enriched in each cluster. I used ggplot2 v3.35 to generate the representation of the enriched gene sets for each cluster (**Fig.R10C**). Finally, to evaluate the glioblastoma gene sets enrichment of each sample (**Fig.R10C**), I used ssGSEA. I used the gsva function (parameters: method = 'ssgsea',

ssgsea.norm=TRUE) of the GSVA v1.38 R package [149] to generate the ssGSEA values. Then, to generate the graphical representation of the scores in the UMAP, I used the plot_density function of the Nebulosa v1.6 R package [152]

Analysis of RNA-seq profiles from genetic tracing reporter GSA models

Analysis of the sLCR high expression profiles of in vitro and in vitro GSA models

The RNA-seq analysis of in vivo and in vitro GSA sLCR $^{high/low}$ (non-MGT#1 and MGT#1) data set was conducted using R v3.6 (see **Results 2.1.2**). After the data processing step (above), the quality of each sample was individually assessed using dupRadar v1.18 R package [153] (default parameters) and subsequently evaluated by the correlation between the number of genes and average counts for each sample (data showed in [107]). Then, differential expression analyses between specific sLCR activation, high/low, and in vivo/in vitro were conducted using DESeq2 v1.24 [154] on raw prefiltered counts (>100 and >50). Of note, principal component analysis was used to identify potential outliers in the in vivo samples data set and only MGT#1^{high} homogeneous samples were used in different comparisons (data showed in [107]). Differential upregulated regulated genes were considered if log2FC >1, adj.pval < 0.05 and base mean > 5.

Then, to evaluate the gene set enrichment of glioblastoma gene sets [28, 35, 52],I used runGSA function of piano v2.0.2 R package [155] (parameters: geneSetStat="page," signifMethod="geneSampling" , nPerm=1000). The graphical representations of this analysis were generated using ggplot2 v3.3.2. The values represented the -log10(adj.pvalue) of the indicated comparisons (**Fig.R21A**). Then, we computed Ingenuity pathway analysis to generate the upstream regulators and selected the top10 for the representation (**Fig R21B**).

Analysis of the MGT#1 high expression profiles upon external factor activation

The analysis of RNA-seq samples upon external factor activation (data showed in [107]) of the MGT#1 sLCR was generated using R v3.6 (see **Results 2.2.1**). First, I filtered low-count genes in the matrix using the filterByExpr function from the edgeR v3.26 R package [146] . Next, I used DESeq2 v1.24 [154] to evaluate the differential expression analysis of each MGT#1^{High} condition versus the control. The external factor TNFα, Leukemia inhibitory factor (LIF), Human Serum (HuS), Ionized radiation (IR), Activin A (ACT), NOC-18, oxidized LDL (OxLDL), and C20-human microglia co-culture [113] were individually compared to control samples (CTRL). Of note, the analysis includes four different sequencing runs, so the sva v3.32 R package [156] was applied for batch correction if necessary. The upregulated genes were considered as log2FC > 1, adj.pvalue > 0.05, and base mean 5. The control upregulated genes were defined by comparing the control and all remaining samples. I used the upset function of the UpSetR v1.4 R package to identify the genes shared between all the comparisons (**Fig.R24A**).

Then, I evaluated the upregulated genes. First, I used GSEA to evaluate the enrichment of glioblastoma gene sets [28, 35, 52] for each comparison (**Fig.R25A**). The analysis was generated using the runGSA (parameters: geneSetStat="page," signifMethod="gene sampling," and nPerm=1000) of the piano v2.0.2 R package [155]. Then, I generated the graphical representation using pheatmap function (parameters: clustring_method="ward.D2", clustering_distance_row & clustering_distance_cols ="manhattan") of the pheatmap v1.0.12 R package. Second, I computed the UMAP dimensional reduction (**Fig R25B**) using the umap function (parameters: n_neighbors = 10, metric = "manhattan", search k = 100) from the uwot v0.1.11 R package. I used as input the matrix of all the filtered samples to retain only the upregulated genes. To remove the batch effect between samples, I used removeBatchEffect function of the limma v3.46 R package [157]. Finally, I used the gsva function (parameters: method=”ssgsea”, ssgsea.norm=TRUE) from GSVA v1.32.0 [149] R package to obtain the enrichment of glioblastoma gene sets and MGT#1^{High} upregulated genes. As input, I used the normalized TCGA-GBM matrix. The heatmap representation (**Fig.R25**) was generated using the pheatmap function (parameters: clustring_method=”complete”, clustering_distance_row & clustering_distance_cols =”euclidean”) of pheatmap v1.0.12 R package.

The master regulator analysis was generated using the RTN v2.2 R package [115] and following the default pipeline. Briefly, I used the rtni function to generate the gene regulatory network. As input, I used the batch-corrected matrix. The list of TF to evaluate was defined in the humantfs.ccbr database v1.01 [158]. Then, I evaluated the master regulator analysis using the mra function (parameters: permutations=1000). To individually assessed the master regulator for each upstream regulator, I used the upregulated genes from each comparison as “hits” as described RTN v2.2 R package [115]. Then, I evaluated the total MRA for each external factor activation using ggplot2 v3.35 (**Fig.R26B**). Finally, I assessed the correlation between MRA using the Jaccard similarity coefficients. The graphical representation was generated using the pheatmap function (parameters: clustring_method=”complete”,clustering_distance_row & clustering_distance_cols =”manhattan”) of pheatmap v1.0.12 R package (**Fig.R26C**).

Analysis of scRNA-seq profiles

Analysis of the scRNA-seq profiles of in vivo GSA models

To evaluate the IDH1$^{wt/mut}$ in vivo GSA scRNA-seq profiles, I used the Seurat v4.0.3 R package [116]. First, I evaluated the raw data and used mouse genes to account for the potential host contamination. I kept only those cells with 0 counts of mouse genes for further analysis. Then, I filtered low-quality cells from the non-contaminated samples using nFeature RNA < 100 & nFeature RNA > 7000 & percent.mt < 10 & nCount RNA > 6000 parameters. Next, I used DoubletFinder v2.0.3 [117] to remove the doublets and kept only singletons. To process the data, I followed the SCTtransform pipeline described in the Seurat v4.0.3 R package [116]. After processing, I computed the UMAP using the RunUMAP function with default parameters (**Fig.R31**). The evaluation of the cell cycle was generated by using the cellCyleScoring function. To define the cells of the IDH1^{wt} and IDH1^{mut}, I annotated the data based on the barcodes defined in the 10x protocol.

Analysis of the scRNA-seq profiles of ex vivo GSA models

To evaluate the ex vivo GSA profiles, I used the Seurat v4.0.3 R package [116]. The profiles were analyzed individually. First, I filtered the low-quality cells for the ex vivo IDH1^{wt} (parameters: nFeature_RNA 700 & nFeature_RNA < 4200 & percent.mt < 7 & nCount_RNA < 20000) and IDH1^{mut} (parameters: nFeature_RNA 1000 & nFeature_RNA < 3800 & percent.mt < 7 & nCount_RNA < 15,000) profiles. Next, I used DoubletFinder v2.0.3 [117] to remove the doublets and kept only the singletons. Finally, I generated the transformation using SCTransform. After processing, I computed the UMAP using the RunUMAP function with default parameters (**Fig.R36**). The evaluation of the cell cycle was generated by using the cellCyleScoring function.

Integration of the scRNA-seq profiles of in vivo and ex vivo GSA models

I used the Seurat v4.0.3 R pipeline [116] to integrate the scRNA-seq profiles of in vivo GSA and the ex vivo GSA models. To avoid potential unspecific cell integrations, I integrated only cells annotated in the cell cycle G1 phase. I used the Canonical Correlation Analysis (CCA) algorithm following the default pipeline described in the Seurat v4.0.3 R pipeline [116] to integrate the single-cell profiles. Specifically, I used the FindIntegrationAnchors (parameters: normalization.method = 'SCT') and IntegrateData (parameters: dims=1:30) function as described in the pipeline. The number of features used for the integration is 30,000 to include all the genes.

Integration of the scRNA-seq profiles of in vivo and ex vivo GSA models and patients

To process the integration of in vivo GSA, ex vivo GSA and glioblastoma patients scRNA-seq profiles, I used the Seurat v4.0.3 pipeline [116].

First, I downloaded several publicly available scRNA-seq of glioblastoma patients from [29, 119, 120]. Using the metadata from the publications, I kept only the samples identified as IDH1^{wt} . Next, I used the Seurat v4.0.3 pipeline [116] to process the profiles individually. For each comparison, I applied the filters min.cells=500, min.features=1000. Then, I filtered the low-quality cells using nFeature_RNA > 200 & nFeature_RNA < q75 and percent.mt < q75. The q75 represents the value of those parameters in the percentile 75% of their distribution. Next, I used the SCTransform pipeline [116] of the Seurat v4.0.3 R pipeline [116] to process the profiles. In addition, the nontumor cells of each sample were annotated and filtered using a combination of gene markers and copy-number alterations. The copy-number alterations were identified using the copyKat

R package [121] (default parameters). Only those profiles that contained more than 500 cells and more than 1000 genes/cell were kept for further analysis.

To integrate the GSA and patients datasets, only the G1 cells were kept for the integration. First, the datasets were combined using LIGER [159], then converted to a Seurat object using the ligerToSeurat function. Then, the datasets were integrated following the Seurat v4.0.3 R pipeline [116]. In this case, the integration used the rPCA algorithm. Finally, the processing was generated following the default (above).

Evaluation of the scRNA-seq profiles results and enrichment

The evaluation of the population markers in the different single-cell analyses and integrations using the FindAllMarkers function of the Seurat v4 R package [116]. The cluster-specific gene markers were considered after filtering the results using adj.pvalue < 0.05, avg.log2FC >.25 , and pct.2 < .5. I used the enricher function of the ClusterProfiler v.2.1.2 R package [151] to evaluate the enrichment of different genesets. The gene sets used in the analysis were the glioblastoma subtype/state markers [28, 35, 52], molecular hallmarks from MSigDB v7.2 (https://www.gsea-msigdb.org/gsea/msigdb), and the TF regulons from Dorothea v1.3.3 R package [103] (I only retained used the A, B, C confidence levels from the dorothea_hs object). The graphical representation was generated using the pheatmap function (parameters: clustring_method=”complete”,clustering_distance_row/_cols=”manhattan”) of pheatmap v1.0.12 R package.

Finally, to evaluate the enrichment of the glioblastoma gene sets [28, 35, 52], I used AUCCell [72] (default parameters). I generated a graphical representation of the AUC scores distribution within the UMAP dimensional reduction using the plot_density function of the Nebulosa v1.6

R package [152]. Finally, I used the fisher exact test (adj.pvalue < 0.01) to evaluate the enrichment of cells within each indicated condition (e.g., datasets). I used ggplot2 v.3.3.5 to generate the different representations. The heatmaps were generated using the pheatmap function (parameters: clustring_method=”complete”, clustering_distance_row/_cols=”manhattan”) of pheatmap v1.0.12 R package.

SUPPLEMENTARY DATA

SAMPLES

ID	Fig.	IDH1	Ori.	Rep.	NGS	Cond.	Treat.
-	R2,3	IDH1^{wt}	GSA	1	WGS	In vitro	-
-	R2,3	IDH1^{mut}	GSA	1	WGS	In vitro	-
TCGA	R6,7,9,10,11,12,13,14,15,16,22,29	IDH1^{wt}	GBM	141	RNA-seq	-	-
CPTAC	R8,9,11,14	IDH1^{wt}	GBM	91	RNA-seq	-	-
GSE119834	R8,9,11,14	IDH1^{wt}	GBM	41	RNA-seq	-	-
GSE48865	R8,9,11,14	IDH1^{wt}	GBM	70	RNA-seq	-	-
GSC	R9,11,14	IDH1^{wt}	GSC	39	RNA-seq	In vitro	-
GSE148292	R9,11,14	IDH1^{wt}	PDX	8	RNA-seq	In vivo	-
GSE127274	R9,11,14	IDH1^{wt}	PDX	4	RNA-seq	In vivo	-
GSA-B1	R6,7,10,11,12,13,15,16,17,18,19,22	IDH1^{wt}	GSA	6	RNA-seq	In vivo	-
GSA-B1	R6,7,10,11,12,13,15,16,17,18,19,22	IDH1^{mut}	GSA	2	RNA-seq	In vivo	-
GSA-B1	R17,18,19,22	IDH1^{wt}	GSA	1	RNA-seq	In vitro	-
GSA-B1	R17,18,19,22	IDH1^{mut}	GSA	1	RNA-seq	In vitro	-
GSA-B2	R6,7,10,11,12,13,15,16,17,18,19	IDH1^{wt}	GSA	4	RNA-seq	In vivo	-
GSA-B2	R6,7,10,11,12,13,15,16,17,18,19	IDH1^{mut}	GSA	6	RNA-seq	In vivo	-
MGT1-B1	R17,18,19,26,25,26,28,29	IDH1^{wt}	GSA	3	RNA-seq	In vitro	-
MGT1-B1	R26,25,26,28,29	IDH1^{wt}	GSA	3	RNA-seq	In vitro	C20MG
MGT1-B1	R26,25,26,28,29	IDH1^{wt}	GSA	2	RNA-seq	In vitro	TNFa

MGT1-B2	R17,18,19,26,25,26,28,29	IDH1^{wt}	GSA	3	RNA-seq	In vitro	-
MGT1-B2	R26,25,26,28,29	IDH1^{wt}	GSA	3	RNA-seq	In vitro	TNFa
MGT1-B2	R26,25,26,28,29	IDH1^{wt}	GSA	3	RNA-seq	In vitro	HuS
MGT1-B2	R26,25,26,28,29	IDH1^{wt}	GSA	3	RNA-seq	In vitro	NOC_18
MGT1-B2	R26,25,26,28,29	IDH1^{wt}	GSA	3	RNA-seq	In vitro	IR
MGT1-B2	R26,25,26,28,29	IDH1^{wt}	GSA	3	RNA-seq	In vitro	OxLDL
MGT1-B3	R26,25,26,28,29	IDH1^{wt}	GSA	3	RNA-seq	In vitro	LIF
MGT1-B3	R26,25,26,28,29	IDH1^{wt}	GSA	3	RNA-seq	In vitro	ACT
-	R21,22	IDH1^{wt}	GSA	34	RNA-seq	In vivo	MGT1^{High}
-	R21,22	IDH1^{wt}	GSA	36	RNA-seq	In vivo	Non-MGT1^{High}
-	R21,22	IDH1^{wt}	GSA	3	RNA-seq	In vitro	MGT1^{High}
-	R21,22	IDH1^{wt}	GSA	3	RNA-seq	In vitro	Non-MGT1^{High}
In vivo IDH1^{mut} GSA	R31,32,33,34,39,40,42,43	IDH1^{mut}	GSA	3	scRNA-seq	In vivo	-
In vivo IDH1^{wt} GSA	R36,37,38,39,40,42,43	IDH1^{wt}	GSA	3	scRNA-seq	In vivo	-
Ex vivo IDH1^{mut} GSA	R36,37,38,39,40,42,43	IDH1^{mut}	GSA	1	scRNA-seq	Ex vivo	IGF1, NRG1, BDNF
Ex vivo IDH1^{wt} GSA	R36,37,38,39,40,42,43	IDH1^{wt}	GSA	1	scRNA-seq	Ex vivo	TNFa,TGFB
-	R40,42,43	IDH1^{wt}	GBM		scRNA-seq	-	-

[Rep.=Replicate; Fig.=Figure; Cond.=Condition; Treat.=treatment; WGS = Whole-Genome-Sequencing]

Gene set

Dataset	Group	Gene set
Wang et al. 2017	ME	S100A11, ARPC1B, CTSC, NPC2, GLIPR1, VDR, BCL3, PLAUR, PRSS23, TGFBI, LY96, RAB27A, P4HA2, TNFAIP8, CLEC2B, IGFBP6, S100A4, BACE2, RUNX1, CAV1, TDO2, GCNT1, IL7R, ITGB1, FTL, DKK1, SLPI, SOCS3, ACPP, LOX, CDCP1, COL1A2, IKBKE, SLC16A3, SYNGR2, SDC1, CD72, CNN2, LUM, PTGS2, FHL2, BNC2, COL5A1, PDK3, ANPEP, COL15A1, LGALS8, SFT2D2, ECGF1, UAP1, TGM2, CXCL6, LOXL2, FAP, PTGES, FTH1, DSC2, BST1, FTHP1, CTSW, LOC57228, DYRK3, PAPPA, DCBLD2, IL1A, CMKLR1, NFKB2, AFP, ITGBL1, CPZ
Wang et al. 2017	CL	PTPRA, ELOVL2, SOX9, MLC1, CENTD1, PAX6, ARNTL, BBS1, DENND2A, SGEF, PLCG1, VAV3, ZHX3, RASGRP1, BBOX1, EYA2, ZC3H14, C14orf159, ACSL3, LHFP, MYO6, NCOA1, CDH4, PLCE1, USP8, METTL8, ACSBG1, TP53BP2, FGFR3, SLC20A2, CST3, ZFHX4, ZNF45, DTNA, SEPT11, TJP1, MEOX2, ZNF211, SALL1, UPF1, STXBP3, MYO5C, MOSC2, KIAA0329, KIAA0355, SUOX, EGFR, PPARGC1A, SLC4A4, POLRMT, SPRY2, GRIK1, RBCK1, LPIN2, C5orf4, PNPLA6, NPEPL1, ST5, BCKDHB, PHKB, CAMK2B, BAG5, SCAMP4, SLC3A2, MAP4, SSFA2, TMEM131, PTPN11, VAPB, SLTM
Wang et al. 2017	PN	TMSL8, MLLT11, HN1, RAB33A, MYT1, FAM77C, HOXD3, HDAC2, KLRC3, C1QL1, LOC81691, NPPA, MNX1, CA10, PTTG1, HRASLS, UGT8, PFN2, MTSS1, TBPL1, EPHB1, TCP1, DCTN3, PAK7, PTTG3, ERBB3, RASL11B, SOX10, H2AFZ, SMPD3, MYB, SLC1A1, CAMKV, NARF, C2orf27, CDKN1C, ZNF804A, PDGFRA, BCL11A, ANKS1B, NDUFB11, NMU, DYNC1I1, JPH3, GABRA3, FA2H, MAST1, IL1RAPL1, B4GALNT1, C20orf42, SIM2, GPR23, TNRC4, ACOT7, REC8, SLC17A6, MAGEL2, BRSK2, PKMYT1, KLRK1, DCT, SUSD5, GABRB3, GBX2, CENPJ, KLRC4, GRID2, CENTG1, DAZ4, DAZ1
Verhaak et. al. 2010	MESVSALL_UP	TCIRG1, ARPC1B, DOK3, VDR, DSE, S100A11, SHC1, CTSC, TGFBR2, MAN2A1, LY96, CAST, CSTA, ALDH3B1, AIM1, TNFAIP8, CD14, IL4R, SLC16A3, MVP, CEBPB, SQRDL, TRADD, FCGR2B, STAB1, FMNL1, TGFBI, TNFRSF1B, CTSZ, PLAUR, SLC10A3, LCP2, FES, ARHGAP4, ITGA5, RHBDF2, AMPD3, NPC2, SERPINB1, GCNT1, MYO1F, PLAU, CLCF1, RAB27A, ITGB2, IQGAP1, ELF4, LRRFIP1, GNA15, DPYD, MAN1A1, CD2AP, CTSB, TIMP1, MAFB, LHFPL2, CAPG, JAK3, RABGAP1L, FNDC3B, CYBB, RELB, LAPTM5, NRP1, P4HA2, DENND1C, TGOLN2, RBMS1, CLIC1, SLC11A1, C5AR1, MYH9, ZNF217, ANXA2, LCP1, FCGR2A, PLK3, STAT6, IL1R1, FURIN, LOX, NOD2, CTSA, TNFRSF14, CASP1, HEXA, FOSL2, MAN2B1, ADAM12, POLD4, SLA, HCLS1, NCKAP1L, GRN, CSF1, RUNX2, CFI, MGAT1, PIK3CD, ABCC3, RHOG, FCER1G, TPP1, LTBP2, TMBIM1, BNC2, STEAP3, CASP4, DEF6, WAS, SLC7A7, IL15RA, SERPINE1, GFPT2, SAT1, IL13RA1, S100A4, CTSS, TAPBP, C1S, LNPEP, APOBEC3C, PTPN6, SCPEP1, SP100, IL7R, SH2B3, FHOD1, BACE2, TNFAIP3, LAMB1, IRAK1, CLEC7A, NFKB2, LIMS1, PTPN18, ST3GAL1, NOD1, CD44, SLC15A3, UPP1, FXYD5, WIPF1, RRAS, BCL3, NAGA, MAPK13, CDCP1, ANXA1, CD4, SYNPO, RUNX1, NNMT, MAP2K3, CCR1, THBS1, ADAM8, ZC3H12A, PTRF, ICAM1, SH3TC1, MANBA, IRAK3, MAP3K6, STXBP2, NAGPA, TRPV2, TLR1, ITGA4, MR1, TNFRSF1A, PRDM1, PRKCD, PROCR, SPI1, COL1A2, ITGB1, FAM129A, APOBEC3F, PTPN7, C2
Verhaak et. al. 2010	CLVSALL_UP	MLC1, PIPOX, SOX9, DENND2A, PDGFA, NES, KCNF1, PTPRA, LAMB2, GLG1, SEMA6D, CDH4, SPRY2, CDH2, NPAS3, ITGB8, ACSBG1, SLC12A4, LMO2, TRIB2, KLHL4, SEMA6A, SIPA1L1, PTPN21, SLC4A4, MEGF8, ACSL3, NR2E1, CD97, IRS2, WSCD1, EYA2, ITGA7, JAG1, ADAM19, RBM42, GAS1, SALL1, TRIM9, FZD3, SCRN1, BLM, MEIS1, SMO, DAG1, LHFP, SPRY4, EGFR, FABP7, EPHB4, PLCG1, DLC1, MCC, B3GALT1, POMT2, DDR1, GRIK5, SPRY1, ATP1B2, FGFR3, TMEM161A, ZNF45, CDH6, ANXA5, ZNF227, PTN, LFNG, HS3ST3B1, ERBB2, SDC3, CITED1, ZNF211, MEX3C, FBXO17, ARSJ, PRKD2, RGS6, TLE1, POLRMT, CCND2, RFX2, MED16, UPF1, CC2D1A, NPEPL1, PEPD, ADAMTS9, CREB5, GLI2, POFUT1, THSD1, ZNF134, SOCS2, NOTCH3, ERCC2, FJX1, ARNTL, KLHDC8A, LRP1, TTYH1, TEAD3, AASS, CDK6, SLC1A3, CLIP2, VPS16, ABCD2, TTC23, SPRED2, ACTN4, UNC45A, EHD2, PLEKHA4, JUND, SAMD4A, ETV4, EMP3, ZNF264, FKBP10, YAP1, KIAA0355, SLC6A9, EXTL3, CD151, GNG7, SCAMP4, TMEM158, SLC4A3, PCDHGC3, ZYX, TNPO2, ARHGEF18, ZNF20, TMED1, CKB, ALDH7A1, MTMR3, BBOX1, TGIF2, SLC20A2, CHMP2A, MRC2, AKT2, QTRT1, TMEM147, HMG20B, EFEMP2, GPR56, CST3, TRIP6, SMAD1, LRFN3, RIN1, HES1, ITPR2, RASGRP1, TYK2, RFXANK, SLC1A2, FZR1, POU3F2, SSH3, DOCK6, TBX2, LAMA5, MYO5C, TRPM3, ZNF444, NCLN, PTPRZ1, ZNF419, KEAP1, GPR125, ZNF471, LRP5, BTBD2, STMN3, ARHGEF12, ALK, DCHS1, CCDC130, AGT, SEMA5A, RRBP1, SHOX2, BICD1, LAMA2, CADM4, BBS1, AP3D1, GNAS, RGS12, PRPF31, ADORA1, INHBB, ABI2, SSFA2, SIN3B, PTRF, WIZ, ZNF512B, CALM1, STK11, TLK1, CAMK2B, PVR, IRF3, ZFHX4, HRH1, SLC12A9, PTPN11, SGSM2, TLE2, ETV5, MARK1, PACSIN3, KCNIP1, CHERP, FOXG1, RHBDF1, F3
Verhaak et. al. 2010	PNVSALL_UP	DCX, EPHB1, SCN3A, DNM3, KLRC3, SATB1, MYT1, NOL4, ALCAM, C1QL1, CXXC4, FHOD3, CRMP1, GSTA4, RAB33A, WASF1, DUSP26, CLASP2, MLLT11, CHD7, CBX1, SOX4, UGT8, TOX3, DPYSL4, PPM1E, SLCO5A1, GPR17, TTC3, BCL7A, STMN4, GRIA2, KIF21B, MTSS1, IL1RAPL1, DGKI, CNTN1, RBPJ, GNG4, AMOTL2, ELAVL3, MARCKSL1, RALGPS1, SOX11, MAST1, TOP2B, PAK3, PODXL2, OLIG2, PAK7, ERBB3, C1orf106, PLCB1, TMEFF1, SORCS3, MAP2, CDK5R1, NKX2-2, NRXN2, KLHL23, ZNF711, RALGPS2, RUFY3, FLRT1, LPHN3, LRP6, SLC1A1, RAP2A, ICK, VAX2, HN1, TRO, PFN2, ATP1A3, TMCC1, HRASLS, DBN1, GSK3B, TAGLN3, MMP16, KHDRBS3, MAGEH1, BRD3, KIF5C, PELI1, ACACA, BCAN, BEX1, RNF144A, EFS, FBXO21, CSNK1E, MICAL3, BASP1, ZEB2, ADD2, ZBTB5, BRSK2, MATR3, PDE10A, STMN1, TCEAL2, NCAM1, CA10, DPF1, HOXD3, KCND2, ARHGEF9, OPCML, BAHCC1, DLL3, FSD1, UBE2O, KLF12, PCBP4, BEX4, NR0B1, DSCAM, ENAH, SPTBN2, NXN, FXYD6, EYA1, MAPT, MNX1, ZCCHC14, HNRNPR, ASRGL1, FGF12, TMEM57, KIF1A, NFIB, CAMTA1, CIT, REEP1, TDRKH, HDAC2, ANKRD28, SMC3, VEZF1, SCN1A, LSAMP, ZNF248, DNAJB5, SNAP91, MKRN3, FAM110B, YPEL1, TSPYL4, SOX10, DRP2, CELSR3, P2RX7, SEC61A2, ZNF510, ZNF10, BCOR, NRXN1, SLC38A1, REC8, TNK2,

		LBR, ZNF286A, GNAI1, GRID2, PTPRS, CACNG4, H3F3A, RBM12B, PLCB4, RAF1, CBX5, PSIP1, TMEM35, RPRM, ZNF804A, CDC7, LRRC20, RUNDC3A, TOPBP1, TUB, CASC3, GPSM2, MMP15, H1FX, ZBED4, ARVCF, CASK, MED24, ZNF184, BOP1, PHF2, CLGN, HLTF, PCDH11X, CRB1, ZNF292, ATRNL1, STXBP1, BPTF, RAPGEF2, FGF14, IKBKAP, SMPD3, GADD45G, SETD5, PCGF2, MED13L, LARGE, PATZ1, ACVR2B, AGPAT4, HNRNPA1, PCDH7, TCF4, DZIP3, MTMR9, PDGFRA
Neftel et. al. 2019	MES1	CHI3L1, ANXA2, ANXA1, CD44, VIM, MT2A, C1S, NAMPT, EFEMP1, C1R, SOD2, IFITM3, TIMP1, SPP1, A2M, S100A11, MT1X, S100A10, FN1, LGALS1, S100A16, CLIC1, MGST1, RCAN1, TAGLN2, NPC2, SERPING1, C8orf4, EMP1, APOE, CTSB, C3, LGALS3, MT1E, EMP3, SERPINA3, ACTN1, PRDX6, IGFBP7, SERPINE1, PLP2, MGP, CLIC4, GFPT2, GSN, NNMT, TUBA1C, GJA1, TNFRSF1A, WWTR1
Neftel et. al. 2019	MES2	HILPDA, ADM, DDIT3, NDRG1, HERPUD1, DNAJB9, TRIB3, ENO2, AKAP12, SQSTM1, MT1X, ATF3, NAMPT, NRN1, SLC2A1, BNIP3, LGALS3, INSIG2, IGFBP3, PPP1R15A, VIM, PLOD2, GBE1, SLC2A3, FTL, WARS, ERO1L, XPOT, HSPA5, GDF15, ANXA2, EPAS1, LDHA, P4HA1, SERTAD1, PFKP, PGK1, EGLN3, SLC6A6, CA9, BNIP3L, RPL21, TRAM1, UFM1, ASNS, GOLT1B, ANGPTL4, SLC39A14, CDKN1A, HSPA9
Neftel et. al. 2019	AC	CST3, S100B, SLC1A3, HEPN1, HOPX, MT3, SPARCL1, MLC1, GFAP, FABP7, BCAN, PON2, METTL7B, SPARC, GATM, RAMP1, PMP2, AQP4, DBI, EDNRB, PTPRZ1, CLU, PMP22, ATP1A2, S100A16, HEY1, PCDHGC3, TTYH1, NDRG2, PRCP, ATP1B2, AGT, PLTP, GPM6B, F3, RAB31, PPAP2B, ANXA5, TSPAN7
Neftel et. al. 2019	OPC	BCAN, PLP1, GPR17, FIBIN, LHFPL3, OLIG1, PSAT1, SCRG1, OMG, APOD, SIRT2, TNR, THY1, PHYHIPL, SOX2-OT, NKAIN4, LPPR1, PTPRZ1, VCAN, DBI, PMP2, CNP, TNS3, LIMA1, CA10, PCDHGC3, CNTN1, SCD5, P2RX7, CADM2, TTYH1, FGF12, TMEM206, NEU4, FXYD6, RNF13, RTKN, GPM6B, LMF1, ALCAM, PGRMC1, HRASLS, BCAS1, RAB31, PLLP, FABP5, NLGN3, SERINC5, EPB41L2, GPR37L1
Neftel et. al. 2019	NPC1	DLL3, DLL1, SOX4, TUBB3, HES6, TAGLN3, NEU4, MARCKSL1, CD24, STMN1, TCF12, BEX1, OLIG1, MAP2, FXYD6, PTPRS, MLLT11, NPPA, BCAN, MEST, ASCL1, BTG2, DCX, NXPH1, HN1, PFN2, SCG3, MYT1, CHD7, GPR56, TUBA1A, PCBP4, ETV1, SHD, TNR, AMOTL2, DBN1, HIP1, ABAT, ELAVL4, LMF1, GRIK2, SERINC5, TSPAN13, ELMO1, GLCCI1, SEZ6L, LRRN1, SEZ6, SOX11
Neftel et. al. 2019	NPC2	STMN2, CD24, RND3, HMP19, TUBB3, MIAT, DCX, NSG1, ELAVL4, MLLT11, DLX6-AS1, SOX11, NREP, FNBP1L, TAGLN3, STMN4, DLX5, SOX4, MAP1B, RBFOX2, IGFBPL1, STMN1, HN1, TMEM161B-AS1, DPYSL3, SEPT3, PKIA, ATP1B1, DYNC1I1, CD200, SNAP25, PAK3, NDRG4, KIF5A, UCHL1, ENO2, KIF5C, DDAH2, TUBB2A, LBH, LOC150568, TCF4, GNG3, NFIB, DPYSL5, CRABP1, DBN1, NFIX, CEP170, BLCAP
CAPE GSA-TCGA (*)	C1	OS9, PDGFRA, RPL4, DST, NPM1, VCAN, EIF4B, PHLDA1, MYC, PRKDC, CENPF, TTC3, COL11A1, CNTN1, PEG10, FAT1, SACS, TOP2A, CCT5, PRRC2C, CDK6, MATR3, MMP16, HAPLN1, HMGA1, TPR, REV3L, CCND1, CHIC2, TBL1XR1, ATRX, RSL1D1, SLC38A1, SMARCC1, MKI67, SOX11, NUDT3, MAZ, BIRC6, IGF2, EEF1D, OSBPL8, MAP3K1, SMG1, USP34, GNB4, RBPJ, FUBP1, SOX6, VPS13C, TTC37, CHD7, MDN1, ATP13A3, MBNL3, ASPM, SFT2D2, ERBB3, CHD1, PPFIBP1, GNL3, LRP6, PTPRJ, DDX21, MAN1A2, SMC4, DKC1, RNF144A, AKAP9, ZNF644, THOC2, HEATR1, MIB1, LBR, SH3KBP1, PHIP, SMCHD1, SLC26A2, NIPBL, FANCI, ATM, DCX, PSME4, RIF1, MEX3A, NUP155, SOX5, RC3H2, CEP350, IGF1R, FAM20B, NIN, BDP1, ASCC3, TOPBP1, KCND2, RRP1B, FAT3, PPIP5K2, AHCTF1
CAPE GSA-TCGA (*)	C2	TUBA1A, RPLP0, RPS18, RPS11, RPS6, RPSA, RPS3A, RPL8, RPS8, RPS3, RPS5, RPS19, RPL32, S100B, RPL9, RPS14, RPLP1, RPL27, RPS16, RPS9, RPL27A, RPL18, TSPAN31, RPL24, RPL35, RPL18A, MDM2, NDUFA4, RPL29, RPLP2, DCTN2, RPS15A, PFN2, RPS15, STMN1, RPS21, FAU, RPL36, SGCB, RPS13, RPS29, DTX3, HSBP1, RPS10, SIRT2, RPS26, ATP6V0B, SLC29A1, SNRPD2, RPS28, UBL5, SCRG1, PSMA7, COX6A1, COX7A2, COX6B1, UQCRQ, COX5B, COX7B, CYC1, EIF3K, NDUFB7, ZNHIT1, UQCRH, CPM, UQCR11, FUNDC2, NDUFB2, ANAPC11, PSMB3, C1QBP, PHPT1, PSMD13, DYNLT1, GTF3A, NME1, DCTN3, STARD3NL, SNRPE, C12orf57, NDUFB11, C19orf53, ROMO1, B4GALNT1, HAX1, SLC35E3, DDIT3, SUCLG1, STOML2, NDUFA2, RBX1, UQCR10, SF3B5, COX5A, PIP4K2C, BUD31, NUP107, COMMD7, NDUFA3, LYPLA1
CAPE GSA-TCGA (*)	C3	CLU, CHI3L1, SPP1, FN1, CTSB, SOD2, A2M, LTF, COL1A2, COL1A1, COL3A1, CD44, TIMP1, COL6A1, CTSD, PLTP, GSN, COL6A2, ANXA2, NDRG1, EFEMP1, SERPINE1, TGFBI, HSPB1, SPOCK2, GRN, S100A10, ITGB2, FLNC, SCG2, S100A11, GPNMB, TMBIM1, SYNPO, ITGB1, JUNB, CAPG, THBS1, SOCS3, FBLN1, COL5A1, GAA, COL18A1, COL5A2, CADM3, CAV1, ITGA3, CCL2, ARPC1B, TNS3, SPOCD1, SLC39A14, MAN1C1, TPPP3, BHLHE40, SHC1, PCOLCE, IL13RA1, SDC4, LGMN, EMILIN1, MVP, FOSL2, TPM2, GM2A, ITGA5, NNMT, ACTA2, RGS2, CA12, NRP1, SLC20A1, RNASET2, LOXL2, ST6GAL1, ANKH, ANGPTL4, SNX10, SLC2A3, PLP2, BHLHE41, RDH10, TGM2, CP, SPOCK1, TGFB1, LHFPL2, TGFBR2, OLFML3, CFI, GFPT2, CERCAM, MYLK, PRSS23, GBP2, MFSD1, DCBLD2, LOX, PROS1, HSPA1B,

CAPE GSA-TCGA (*)	C4	EGFR, PTPRZ1, BCAN, SLC1A3, SDC3, NES, DDR1, ATP1A2, CCT6A, NCAN, MLC1, FADS2, SPTAN1, AHCYL1, NFIX, TTYH3, CLSTN1, CLIP2, CCND2, PCDHGC3, RGMA, SCRN1, SOX2, NCAM1, ADCYAP1R1, SNRNP200, PRRC2A, CTNND2, VOPP1, CIRBP, AGRN, RASSF2, GPR37L1, SOX9, PREX1, TRIM9, ITGB8, DCLK2, DAG1, ITGA7, HNRNPUL1, ABAT, FYN, CSPG5, SLC44A2, APC2, CDC42EP4, OLFM2, FAM168A, EDNRB, ID4, KHSRP, CLIP3, PLCG1, FGFR3, WSCD1PTPRA, SNRNP70, NDRG4, PTPRS, SLC4A4, CDH2, DPP6, STAT1, SLC7A5, MAPT, MSI2, NOTCH1, LENG8, CCDC80, NLGN3, PTPN11, KLHDC8A, MEGF8, AP3D1, DENND2A, CELSR2, LPL, CENPB, TNPO2, ASTN1, WDR6, MIDN, LFNG, HIP1, PHGDH, PLXNB1, MYO9B, PNMA2, EXTL3, SUPT5H, ARHGEF6, JAM2, FLOT2, AKT1, NOVA1, AKT2, SEC14L1, REPIN1, KCNJ10
CAPE GSA-GBM (*)	C1	OS9, PDGFRA, RPL4, TUBB, HNRNPA1, RPS2, RPL3, CDK4, RPL19, RPL10, NCL, RPL5, RPL7A, RPS4X, TRIO, RPS11, YBX1, PTMA, NPM1, HNRNPU, HNRNPH1, RPSA, NONO, RPLP1, LDHB, RPL32, RPL37A, RPL11, RPS3, RPS27, RPL37, RPS24, ILF3, RPS23, RPL12, NAP1L1, RPS27A, SET, SERBP1, RPL23, NACA, RPS7, PEG10, RPS14, HSPD1, RPL9, CHD4, CENPF, RPS12, RPL14, PARP1, RPL10A, RBMX, RPS5, FUS, MARCKSL1, BTF3, TOP2A, HNRNPA3, ACLY, CCT3, ODC1, RPL30, RPL35, RPL27, FASN, RPS25, CBX3, RPL18A, TOMM20, SOX4, SRSF3, RBM3, HNRNPR, COL11A1, RPS16, RPL35A, XRCC6, MCM7, RPL23A, CCT5, SYNCRIP, PAICS, SRSF11, RPL38, MKI67, ILF2, RPS13, DHX15, SRSF1, EEF1B2, HDGF, SOX11, CCT7, RPL34, TRA2B, LAPTM4B, RCC2, RBBP7, MYC
CAPE GSA-GBM (*)	C2	CHI3L1, FN1, SOD2, FTL, TNC, COL3A1, A2M, CTSB, FLNA, COL1A2, LTF, COL1A1, VEGFA, CD44, TIMP1, COL6A1, MYH9, ANXA2, COL6A2, EFEMP1, TMSB10, SAT1, NDRG1, SERPINE1, TGFBI, POSTN, IGFBP3, VMP1, TLN1, COL5A2, MCL1, THBS1, AKAP12, MYL6, EMP1, CALD1, MMP14, IQGAP1, ITGB1, CTSD, RCAN1, HIF1A, PLTP, ACTN1, ADAM9, GPNMB, MRC2, LAMB1, LGALS1, NRP2, GNS, COL5A1, SOCS3, GRN, DNAJB1, SLC39A14, NRP1, FOSL2, SERPINH1, FAM20C, COL8A1, ITGA3, PLXND1, DCBLD2, SLC2A3, JUNB, CP, CA12, CLIC1, MTRNR2L1, GPX1, HSPG2, PLOD1, PLOD2, ITGA5, SPOCD1, ITGB2, CAV1, FBN1, BHLHE40, FNDC3B, COL18A1, DPYD, VCL, CCL2, PRRX1, SLC2A1, ARL4C, TNFAIP2, NNMT, CAST, OSMR, SHC1, IFI16, MYOF, S100A10, SULF1, NPC2, MVP, GBP2
CAPE GSA-GBM (*)	C3	DST, SPTBN1, QKI, NTRK2, DYNC1H1, SPTAN1, MAP2, MDM2, PCDH9, KIF1A, KIF1B, APC, SCD, CNP, TAOK1, CPSF6, ZEB2, TF, VPS13C, CCDC88A, BIRC6, WDFY3, DCLK1, UBB, USP34, SYNE2, SESN3, KIDINS220, MKLN1, NF1, ABCA2, PCM1, LRP1B, ASH1L, CHD9, FAT3, NUDT3, HECTD1, VPS13D, RUFY3, MED13L, ATM, CLDND1, SECISBP2L, DOCK4, ZBTB20, CEP350, TNRC6B, AKAP9, LYST, CPEB4, FRYL, APLP1, ARAP2, CDC42BPA, PPP3CA, AKT3, PTAR1, MGA, RIF1, DOCK10, ALCAM, DICER1, TPPP, JMJD1C, BAZ2B, NCDN, PKP4, PURA, ZNF91, ANK3, LNPEP, DMXL1, MYO9A, PHF3, DSEL, NTRK3, RB1CC1, VPS13B, TMEM132B, CPLX2, SLAIN1, TMEM106B, HOOK3, ANKRD12, ANLN, CADM3, GUCY1A2, RABEP1, ALMS1, N4BP2L2, PHC3, ADAM22, BSN, KLHL24, GRIN2B, PIKFYVE, NCAM2, CLSTN2, UNC80
CAPE GSA-GBM (*)	C4	EGFR, CST3, ATP1A2, DDR1, SDC3, PEA15, NCAN, MLC1, ATP1B2, LANCL2, SCRN1, AGT, PDPN, PCDHGC3, SLC4A4, CLIP2, CHCHD2, SEC61G, GATM, TSC22D4, APC2, RGMA, TRIM9, VOPP1, SOX9, EDNRB, TTYH1, FABP7, STAT1, WSCD1, PON2, GPR37L1, CDC42EP4, KLHDC8A, DCLK2, PBXIP1, WLS, MEGF8, SLC44A2, GNA12, CCDC80, FGFR3, LRP4, PTPRA, IFI6, DENND2A, SRI, OLFM2, NLRP1, ELOVL2, SUMF2, DPP6, HOPX, ALDH1L1, PRCP, JAM2, LFNG, OAS3, CSPG5, CORO2B, RHBDD2, WBP2, PSRC1, STMN3, DNAJB2, MGLL, ARHGEF26, ALDH7A1, RND2, FAM181B, CPNE2, CLIP3, NPAS3, PDGFA, MED29, ARHGEF10L, CRIP2, P2RY1, GNG7, SGSM2, LRRC4B, CD82, NRBP2, PHKG1, SMARCD3, MOXD1, C2orf72, ARC, RAMP1, FIBIN, RPH3A, ADORA1, RGS12, CDH4, REEP2, SALL2, FOXG1, CTSF, TOB2, SEZ6
CAPE panGBM (*)	C1	CLU, EGFR, PTPRZ1, SLC1A3, CST3, PMP2, CCND2, NES, MLC1, CHCHD2, AGT, ITGB8, SEC61G, GJA1, CCT6A, ADCYAP1R1, DTNA, RASSF2, PCDHGC3, GATM, FABP7, LANCL2, SCRN1, SLC4A4, PDPN, VOPP1, SOX9, PON2, EDNRB, TRIM9, HOPX, KLHDC8A, CCDC80, RGMA, WSCD1, LIFR, LPL, SUMF2, GPR37L1, ID4, CDK6, F3, PNMA2, PRCP, ARHGEF6, ELOVL2, HEPACAM, IFI6, FGF1, CPSF6, ID3, DENND2A, OLFM2, NLRP1, FGFR3, LFNG, LRP4, ALDH1L1, CNR1, ELN, FAM181B, RFX4, FREM2, FAT3, SPRY4, GRIA1, SEMA6D, METTL7B, FJX1, LDLRAD3, DSEL, OAS3, RNF180, PDGFA, SLC20A2, RAMP1, PSPH, NPAS3, MASP1, TIMP4, FIBIN, TNFRSF19, ARHGEF26, FRS2, ACSS3, CD82, PDZD2, ALDH6A1, KLHL4, CDH6, CDH4, PHKG1, SOCS2, ST8SIA5, ATP13A4, C21orf62, PIPOX, LRRN3, P2RY1, TRIM69
CAPE panGBM (*)	C2	FN1, COL3A1, COL1A2, COL1A1, AHNAK, COL6A2, THBS1, COL5A2, COL5A1, HSPG2, EFEMP1, POSTN, FSTL1, FBN1, LAMB1, IGF2, LUM, CP, CAV1, COL8A1, COL12A1, NID1, CDH11, COL18A1, MGP, LOXL2, PCOLCE, AXL, ADAMTS1, UACA, MYO1B, COLEC12, FBLN1, CFH, COL7A1, MFAP4, LOX, SNED1, IER3, UNC5B, THSD4, HMCN1, NID2, OLFML3, ANPEP, SERPINF1, OLFML2A, COL14A1, P4HA2, ACSS2, LRIG3, NEDD4, PCDH18, ZFHX3, CMKLR1, FAT4, FBN2, TTN, RCN3, FZD1, KDR, VLDLR, SEMA3C, CTSK, ITGA8, CDCP1, ADAMTS12, S100A4, MYO1D, ITGA4, KLF4, SFRP1, LOXL1, COL15A1, PLA2R1, LEPR, PHACTR2, GPRC5C, PHLDB2, TPBG, BICC1, TSHZ3, CRABP2, BNC2, HGF, ENO3, ECM1, RNF152, SNAI2, MMP11, ALDH1L2, CLMP, ATP8B1, FAP, LRRK1, PPIC, COPZ2, MYH11, SPTLC3, EBF1
CAPE panGBM (*)	C3	CHI3L1, SPP1, FTL, SOD2, FLNA, CTSB, A2M, TNC, CD44, TMSB10, TIMP1, LTF, IGFBP3, GPNMB, SAT1, SQSTM1, CALU, VMP1, IQGAP1, RCAN1, ACTN1, AKAP12, CPD, CLIC1, NRP2, DCBLD2, RNF213, DNAJB1, CA12, SLC39A14, WWTR1, FAM20C, SOCS3, CTSC, CAST, SHC1, ATP13A3, SLC2A3, MMP9, CCL2, SLC20A1, LMAN1, SLC5A3, TIPARP, NPC2, SPOCD1, LHFPL2, TUBA1C, DPYD, MYL12A, RDH10, OSMR, NNMT, PYGL, STEAP3, CSF1, IL13RA1, MAN2A1, SLC4A7, HSPA1B, ZNF436, ITGB2, GBP2, NRIP1, PTX3, PROS1, SDC4, RGS2, GFPT2, TNFAIP2, LIMS1, DNAJC3, MAN1C1, PARP4, ERRFI1, RASSF8, PPP1R15A, RND3, PLP2, TNFRSF12A, BACH1, ICAM1, TNFRSF10B, UGCG, GLIPR1, GPX3, SOAT1,

		CHST2, HSPA1A, UPP1, CXCR4, CFI, RUNX1, SBNO2, LIF, SLC25A37, NQO1, BACE2, SLC39A8, GEM
CAPE panGBM (*)	C4	OS9, HNRNPA2B1, HNRNPA1, PDGFRA, SRRM2, CDK4, RPL5, YBX1, NCL, GNB1, KIF1A, PABPC1, CNP, MARCKSL1, STMN1, HIPK2, FUS, ILF3, SFPQ, RPL28, HNRNPA3, MRFAP1, PFN2, RPS16, SLC44A1, CNTN1, EWSR1, TSPAN31, RAB7A, ABCA2, SNRNP70, ENO2, SOX4, SF1, ODC1, FASN, HNRNPM, LENG8, DYNLL2, HNRNPD, SIRT2, EPB41L2, APLP1, PCBP2, MCM7, TOP2A, PTPRF, DCTN1, LUC7L3, DBN1, OLIG1, SRSF11, WDR6, U2AF2, HNRNPA0, NASP, SLC7A5, PSIP1, SLC6A8, PHLDB1, CSNK1E, UBAP2L, ACAP3, KHDRBS1, MAPK8IP3, SBF1, CCNL2, PGRMC1, ZMIZ1, SRCAP, PHGDH, PUM2, NCDN, DDX42, TAF15, RAB14, CHD3, RERE, SRRT, TARDBP, CSNK1D, EEF1A2, SNX1, KDM1A, ANP32A, COL20A1, ATXN2L, PURB, ALCAM, CIC, PPP3CA, GNAO1, HMGB2, HNRNPAB, SOX11, PNISR, NLGN2, PKP4, RBPJ, SMARCC1
CAPE In vitro-in vivo GSA (*)	C1	ALDOA, MT-RNR1, FASN, SCD, KIF1A, HMGCS1, FADS2, MT-ND2, IGF1R, SREBF2, FDPS, SPTBN2, INSIG1, SV2A, ALDOC, FDFT1, LSS, EEF1A2, ENO2, COL6A1, PI4KA, ACAT2, ZMYM3, PLEC, TCF20, FABP7, DHCR7, PRUNE2, PPP1R9B, THRA, HECTD4, ABCA2, ATP1A3, BAZ2A, PCSK2, DNM1, GRINA, NDRG1, STXBP1, ADCY1, AIFM1, CAMKV, KIF5A, NCDN, ITPR1, SPRY4, HK1, NRN1, NLGN2, MSMO1, MINK1, WBP2, LPIN1, DDX19A, KIF3C, GABBR1, SLC25A23, TMEM132A, TUBA4A, FBXL16, SIPA1L1, CNTNAP1, SYNGAP1, ACSS2, MADD, NFASC, NDRG4, CLCN6, BRSK1, TECR, BHLHE40, MAPK3, ADCY9, EPB41L1, DLG4, L1CAM, MTATP6P1, SYNGR1, GAA, IGF2, RAPGEF5, TENM4, CLSTN3, SPOCK1, VAMP2, RNF44, NOTCH3, ATP2B2, DTNA, PDZD2, GGT7, SREBF1, NRP2, ACSL1, GNAL, PCYT2, CMIP, RTN1, PER1, PRKACA
CAPE In vitro-in vivo GSA (*)	C2	PDGFRA, DST, PTN, VCAN, MATR3, C1orf61, RPL9, BCAN, CNTN1, HAPLN1, ODC1, RHOBTB3, GPM6A, EDIL3, RBM3, PMP2, NCAN, APBB2, ETV1, MBNL3, SOX6, EID1, ASPM, LIMA1, GNB4, SOX2, LINC00461, CCDC50, HMGN2, GNG12, RSL24D1, SNORA73B, DCX, GNG2, SEMA5A, MRPL42, SPARCL1, SPCS3, NOVA1, FGD5-AS1, GUCY1A2, ILDR2, KRR1, ARRDC3, GAB1, SYNE2, FAM171B, CCDC14, USP1, TMEM106B, RAB31, CCSAP, SMC2, STAT1, MAD2L1, ANXA5, UST, CENPE, GPSM2, NRXN1, SCD5, ROBO2, IL1RAP, FGF12, CADM2, HDAC9, ETS1, FAM131B, LRRK2, SCRG1, GRIA4, BRCA2, NLGN1, SEMA6D, SNORD17, PSMA2, RCAN1, APOD, LRRN1, ELMOD2, TNR, MEGF10, SKIL, PIK3R2, SRSF8, DUT, NLGN4X, CMTM6, CDK1, ZNF300, ZNF562, BRINP3, RPS27L, ARL4A, KAT2B, RBBP8, SCARNA10, KIAA1143, FER, RPAP2
CAPE MGT1 in vivo - TCGA (*)	C1	BCAN, APP, NES, ITM2C, SDC3, FLNA, LRP1, ATP1B2, CKB, ATP1A2, NCAN, VEGFA, AEBP1, CCT6A, TIMP2, NFIX, FADS2, MLC1, GJA1, TTYH1, CLSTN1, GPC1, TTYH3, PRRC2A, PLEC, JUN, CLIP2, CTNNA1, AGRN, MRFAP1, ACTN4, ATN1, SOX2, SCRN1, PREX1, CIRBP, LTBP3, EDNRB, WLS, NCOR2, TLN1, PRKCSH, LRIG1, LAMB2, SOX9, RASSF2, DAG1, SLC3A2, APC2, SF1, TRIM9, ITGA7, CLIP3, DCLK2, ID4, KHSRP, SLC44A2, NFE2L1, NFIC, ATP6V0E2, SNRNP70, FAM168A, USP11, ITGB8, EZR, LENG8, PLCG1, ATP2B4, CDC42EP4, UBE2H, STAT1, PLXNB2, STAT3, PYGB, PBXIP1, MAPT, ITPKB, GNA12, AP3D1, PTPRF, EPAS1, GLG1, SCARA3, CTNND1, TMEM132A, MEGF8, CLPTM1, MYO9B, TGOLN2, SLC4A4, CDH2, MLEC, PTPRA, FGFR3, CSNK1D, AHNAK, NFASC, SUPT5H, WSCD1, RNF187
CAPE MGT1 in vivo - TCGA (*)	C2	OS9, CDK4, PDGFRA, MAP2, DST, TSPAN31, VCAN, STMN1, MDM2, ODC1, CNTN1, TTC3, HSPD1, SLC44A1, PRKDC, SOX4, DCTN2, TOP2B, RBMX, TOP2A, LAPTM4B, USP9X, CENPF, TNPO1, CCT5, TSFM, DTX3, SACS, MYC, OGT, PPP3CA, FAT1, SRSF1, SOX11, HMGA1, CCND1, COL11A1, BCLAF1, SLC29A1, MMP16, TBL1XR1, TMPO, MATR3, DHX15, SLC38A1, LRPPRC, ATRX, REV3L, PUM2, KDM1A, CASK, NCDN, METTL1, PURB, BIRC6, APBB2, KPNA2, DCX, GNG2, XPOT, USP34, SPTBN2, ALCAM, MAZ, RAP2A, RBPJ, OSBPL8, TP53, CPSF6, CSE1L, CHD7, ATCAY, SMG1, MED13, B4GALNT1, TARDBP, HAPLN1, HDAC2, GNB4, FUBP1, MKI67, SOX6, ZNF638, CHIC2, NUDT3, GSK3B, ERBB3, U2SURP, AKT3, CEP170, TTC37, ABCE1, UBE3A, OPA1, SLC35F1, VPS13C, STXBP1, ZMYM2, DSEL, MIB1
CAPE MGT1 in vivo - TCGA (*)	C3	FTL, TPT1, B2M, CTSB, RPLP0, FN1, RPS18, COL1A2, TMSB10, COL3A1, CD63, COL1A1, RPS8, RPS3A, MYL6, RPLP1, RPS14, RPS19, RPS5, LTF, SEC61G, TIMP1, PGK1, RPL27, RPL27A, RPS9, RPS16, COL6A2, S100A6, RPL24, ANXA2, SAT1, GPX1, SERF2, RPS15A, HOPX, IGFBP3, RPL36, RPS21, TGFBI, FAU, POSTN, SERPINE1, RPS29, CALU, LGALS3, MMP14, CLIC1, GUK1, RPS10, ARPC5, SERPINH1, CD164, ATP6V0B, SNRPD2, MGP, PSMA7, ARPC3, UBL5, TRAM1, S100A11, SH3BGRL3, SSR4, PCOLCE, LAMB1, CLDND1, PRDX4, COL5A2, SRPX, NNMT, MYL12A, THBS1, COL5A1, UQCRQ, CAV1, ARPC4, CCL2, ZNHIT1, MAGT1, ARPC1B, IFNGR1, EMILIN1, RGS2, SPOCD1, TMEM14C, DYNLT1, LUM, UQCR11, SDC2, CA12, LOXL2, AP2S1, POMP, ITGA5, TMEM219, SLC40A1, IL13RA1, SYNGR2, PLP2, OLFML3
CAPE MGT1 in vivo - TCGA (*)	C4	CLU, CHI3L1, AGT, EGFR, LPL, RGMA, CRB2, SRP9, APOBEC3C, DTNA, NLRP1, PMP22, ADCYAP1R1, PPT1, NPC2, SULF1, B3GNT9, PLCD3, PCSK1N, LEPROT, PNMA2, CTSF, TAF10, PARP9, ZNF395, RAMP1, ABLIM1, MT1X, PLCD1, ALDH1L1, RWDD1, EFEMP2, C2orf68, PRELP, PSMD6, TGM2, GDE1, GAS2L1, RRM2B, TXN2, TMEM170A, DPM1, RNF139, MPDU1, FAM131A, RRAGA, HSD17B12, ECI2, MYOF, BTN3A2, MRPL47, SLC43A2, ZNF529, DOLPP1, RND3, ST3GAL3, ARFIP2, SMG9, ELAVL3, PARP4, MRPS17, PHLDA3, IL17RC, SUMF1, KRBA1, BCAM, SIAE, SIPA1, NICN1, ARRDC1, YRDC, ZNF140, CEND1, BUB1, TMEM208, ALKBH7, C1GALT1, PAN2, IL33, TMEM134, SP100, LRRCC1, ALDH5A1, ST3GAL5, RENBP, PTPMT1, NAGLU, TRNAU1AP, DCK, FBXO4, FRG1, DRG2, GPD1, SCML1, PEX11A, SHMT1, FBXO10, PARP3, PPIL3, MIIP

CAPE MGT1^{High} in vitro (*)	C1	PGK1, NDRG1, IGFBP3, IGFBP5, FLNB, GRB10, RNF19A, SLC2A3, PFKFB4, PHF19, SLC2A1, ICAM1, GBE1, CA9, L1CAM, GSN, ZNF395, IL1RAP, ADAM19, CAPN5, AHNAK2, LRP2, ADAMTS9, TMEM45A, FOSL2, NFKB2, EPAS1, LIF, CACNA1H, MATN2, UNC13A, LOXL2, DPYSL4, PCDH1, DNAH11, GAP43, TAGLN2, DUSP5, GABRB3, RELB, BEND5, SPAG4, ANXA1, TFPI, STC1, PSMD5, SYNPO, PIK3CD, NCAN, PAQR5, IER3, NOVA2, MICALL2, SLC4A4, EHD2, TRAF1, ESPN, SYT12, CDKN1A, LRRC4B, TMEM158, ZNF440, PRSS35, CRABP2, FZD8, CHI3L1, SERPINE1, HLA-B, ALPL, TTN, GPRC5A, TGFB1, SDC4, BCL3, MAPK8IP2, SEMA5B, ZNF256, TIFA, EGR2, SGCD, EMILIN1, EMP1, AMPD3, CHRNB2, PTPRU, CXCL16, VDR, EMILIN3, FANK1, ZIK1, COL5A2, INHBE, ZNF610, ZNF525, UNC5B, PTGS2, NOL4, TBX2, MATN1, EPHA10
CAPE MGT1^{High} in vitro (*)	C2	FASN, PTPRZ1, TRIO, REV3L, FTL, SRRM2, SLC26A2, MMP16, LSS, SQLE, FDFT1, DDX5, SAT1, LRP6, MAP2, NUFIP2, SLC44A1, ACACA, RNF213, SRSF11, SREK1, PCDH15, FABP7, TJP1, ELOVL6, CD164, TAF1D, FUBP1, LUC7L3, ELOVL5, TXNRD1, LRP8, SOX4, PRDX1, CHD2, SEL1L3, NKTR, DOCK10, KIF21A, ITGB8, PNISR, SERINC5, CAD, PDE4B, CYCS, APC, SOX6, PMP2, UTRN, TMEM167A, AIFM1, APBB2, BCAP29, ANKRD50, SUGP2, MVD, HSPA4, RNF157, CCNL2, STARD9, TBC1D14, ANKRD10, COL19A1, BCHE, FNBP4, ACOX1, KCND2, OLIG2, PTPRK, QPRT, TIA1, KLHL24, COG5, PIKFYVE, SALL3, TMEM33, RGMB, CCDC14, CREBZF, MAPK8IP3, VPS13B, SNX13, SYNE2, AGPAT4, CREB5, FGFBP3, MCF2L, DSEL, CELF2, GAB1, CCNL1, LUC7L, ATR, MBNL3, RDH11, CLK1, PDXDC1, CDHR1, TTL, ARHGEF7
CAPE MGT1^{High}-TCGA (*)	C1	CLU, CHI3L1, SPP1, SPARC, EGFR, CST3, SOD2, HLA-B, LTF, UBB, S100A6, CD44, EFEMP1, ANXA2, GATM, AEBP1, TAGLN2, ANXA1, HOPX, EMP1, S100A16, HSPB1, NLRP1, SPOCK2, PDPN, S100A10, LGALS3, TMBIM1, FGF1, SERPINB6, ITGB2, CAPG, S100A11, DPP7, TPPP3, CBR1, CDKN1A, MDK, CAV1, RAMP1, CSRP2, BST2, NNMT, GPX3, ABCA1, MT1X, MGP, IFI16, FABP5, SDC4, PLP2, CCL2, TRIP6, SYNM, DHRS3, S1PR3, MOXD1, PSMB8, PHKG1, VAMP5, CRB2, SLC40A1, VCAM1, CP, HSPA1A, CFLAR, HSPB6, GBP2, AXL, HNMT, ANGPTL4, CXCR4, MICALL2, HSPA1B, TGFB2, AVIL, BBOX1, PARP10, ADM, PHLDA3, PDLIM4, FBLN5, CYP27A1, SAMD9L, CAPS, RHOG, FAM111A, UBA7, SCPEP1, NME3, TCIRG1, SAA1, IAH1, UPP1, IL33, GSTM2, SP100, GNG11, IFI35, HAP1
CAPE MGT1^{High}-TCGA (*)	C2	BCAN, PTPRZ1, SDC3, CKB, NCAN, CCT6A, FADS2, MARCKSL1, TTYH1, TUBB2B, SPTAN1, NFIX, FXYD6, RTN3, NCAM1, CLSTN1, CNP, MAP2, MAGED1, CLIP2, CTDSP2, RASSF2, OLIG1, AP2B1, SOX2, SGCB, KIF1A, ABAT, FYN, CSPG5, DCLK2, SOX8, HNRNPUL1, TRIM9, CADM4, KIF1B, TCF12, DAG1, CRMP1, FAM168A, MAPT, WSCD1, USP11, PTPRS, ETV1, DPP6, NLGN3, MAPK8IP1, PTPRA, NOTCH1, STMN3, OLIG2, AIF1L, DYNLL2, MCM7, ZNF664, GRIK3, DBN1, SOX4, PLXNB1, CELSR2, DPYSL5, ABCA2, TNPO2, SCG3, ASTN1, DENND2A, PIK3R1, HIP1, SEZ6L, MIDN, DGCR2, PNMA2, KCNJ10, LSAMP, CACNG4, SEMA5A, NOVA1, REPIN1, ARC, SLC25A23, CPNE2, NLGN2, GNAO1, PSIP1, VGF, LRRN2, BTBD2, CORO2B, LRRC4B, NFIB, LRP4, ZEB1, SEZ6, ITPK1, ACAP3, DVL3, APBA2, SEMA6A, ZFAND3
CAPE MGT1^{High}-TCGA (*)	C3	VIM, A2M, FLNA, TNC, COL1A2, COL6A1, COL3A1, COL1A1, HSPA5, MAP1B, HSP90B1, NDRG1, VEGFA, COL6A2, IGFBP3, P4HB, CLIC4, MYH9, TPM4, CALU, PDIA3, PDIA4, PLOD1, CALD1, LAMC1, MRC2, AKAP12, FKBP9, LAMB1, PAM, IQGAP1, WWTR1, ADAM9, FKBP10, SLC2A1, BCAT1, ITGB1, GALNT2, PXDN, CA12, PLOD2, PRUNE2, HSPG2, SHC1, FGFR1, COL5A2, NID1, CTSC, FAT1, GRB10, BHLHE40, USP9X, SLC39A14, PLOD3, FBLN1, CPD, NRP2, PCOLCE, OSMR, SLC2A3, CAST, ITGB5, COL5A1, LMAN1, IGF2R, LAMA4, PYGL, SLC20A1, PABPC4, FOSL2, LOXL2, QSOX1, IL13RA1, ERRFI1, TPM2, HK2, ADAM10, AKAP13, FLNB, ATP13A3, SLC5A3, VCL, ANO6, TPM1, PROS1, ESYT2, FGFRL1, LOX, DCBLD2, PTK7, FNDC3B, DDR2, ANKH, FURIN, IGF1R, LHFPL2, LTBP1, SFT2D2, LUM, TUBB6
CAPE MGT1^{High}-TCGA (*)	C4	TUBA1A, TPT1, RPL13A, RPS18, CDK4, RPS11, RPS6, RPS3A, RPSA, RPS8, RPS3, RPL7A, RPL9, RPS5, RPS14, OS9, RPS19, RPL32, RPLP1, RPL37A, RPS7, PPIA, RPL27, RPL27A, RPS16, RPS9, RPL13, TSPAN31, RPL18, RPL35, RPL18A, RPL24, RPL10A, RPS15A, RPL29, RPLP2, RPL35A, RPS21, RPL36, NDUFA4, RPS13, RPS15, FAU, RPL34, RPL38, RPL41, RPS29, PRDX2, RPS10, SLC25A5, SUB1, C1orf43, SRP9, HSBP1, SNRPD2, COX7C, RPS26, RPL39, PHB2, PRMT1, COX7A2, TSFM, RPS4Y1, CYCS, PSMA7, SSR4, CAND1, COX7B, UQCRQ, UQCRH, RSL24D1, CYC1, PSMB1, C1QBP, SNRPB, EIF3K, METTL1, FUNDC2, GTF3A, NME1, ERH, PSMB3, PHB, TBCA, CUTA, JTB, SNRPE, C12orf57, TMEM106C, HSPE1, STOML2, ATP6V1G1, TOMM22, PDCD6, NDUFB11, HAX1, DCTN3, LYPLA1, PSMA5, RBX1
GSA-TCGA (Celligner)	C0	TUBB2B, OLIG1, IGF2, OLFM2, RNFT2, TTYH1, SOX2-OT, SERPINH1, DDIT3, LGALS3, MEST, DLL3, GAS5, ATP1B2, SGCB, SCG3, EDNRB, PCDHGC3, EPDR1, NDFIP1, LBH, AIF1L, DPP6, SCRG1, KLHL4, BEX1, GAP43, RPH3A, GNG7, TUBB4A, SEMA3F, FZD8, PLAT, CAPN5, NLGN3, ADAMTS9, WSCD1, IL1RAP, PLLP, KCNIP1, NGFR, STK17A, COL26A1, FAM168A, ACAN, CRB1, CITED1, UGT8, ADAMTSL2, SLC35A5, GDAP1L1, LRRC4B, TAGLN2, MDFI, SCRN1, MAPK8IP1, ITGA8, SEMA5B, MRAS, DPYSL5, PNMA2, NPPA, RET, CPXM1, CA10, MAPT, NEK6, ELAVL3, DRAXIN, RDX, EMC10, B3GAT1, SUSD1, PCDHGC4, STK32A, PODXL, TUBB2A, TNFRSF12A, NPAS3, SLC4A8, FOXF2, CSPG5, PIEZO2, NOVA2, NACAD, FGFBP3, ARC, SPRY4, FEZ1, CDH2, NXPH1, PCDH17, SYTL2, TMEM132B, PDGFA, CASP3, TRPM8, SPSB4, B3GNT7, SEMA5A, GSX1, PHLPP1, SCHIP1, RPE65, NMU, ECI2, LHFPL3, DOK5, AFAP1L1, GALR1, SOX9, C1orf226, TCEAL5, SHISA9, POU3F1, TMEM51, DPF3, C3orf70, CRYZ, CERS4, PNMA1, SYNDIG1, GLDC, SPNS2, EYS, FLT1, ELFN2, CACNG8, TNFRSF19, MMD2, GRIK4, FEM1C, NXPH3, FAM86DP, SLC39A3, FLRT1, SPON2, TMEM9B, MYO16, THSD1, MAP6D1, TMEM132C, C2orf80, PCDH1, ADAMTS18, SHROOM2, RHPN2, SERPIND1, PEPD, TMCC3, NTNG2, MIR4458HG, YBX1P1, F12, LINC00888, C11orf49, SMAD1, PRIMA1, TBCB, ZIM2-AS1, GREB1, LIPE, ADAMTS9-AS2, ZNF613, TUBB2BP1

GSA-TCGA (Celligner)	C1	CAPN6, PCDHGA12, BAALC-AS1, LINC01152, DLGAP1-AS5, MT1M, FAM184B, DMBX1, WNT4, MYOZ2
GSA-TCGA (Celligner)	C2	GFAP, MTRNR2L12, MTRNR2L2, RN7SL396P, MYBPH, PLXND1, SNORA53, RN7SL674P, RN7SL767P, MTRNR2L3, NLRP1, POM121L9P, SERPINA5, CCDC3, NWD2, PREX2, ADAMTSL4, RN7SL128P, LAMC2, MUC1, GALNT5, CD180, FGR, USP6, CXCL3, SFN, ANO4, NFAM1, MTATP8P1, LINC01579, POU2F2, APOBR, FER1L4, RPL18AP15, CCSER1, SOX21, CLDN23, KCNJ2, PRAM1, STXBP2, UNC13D, POLD4, P2RY6, CEACAM21, SH3TC1, SERPINB9, FERMT3, PTPN6, MLKL, LILRA6, MYO1G, PIK3CD, EEF1A1P33, ITGAM, KCNK6, ALDH3B1, SPN, ZNF826P, PIK3CG, CSF2RB, EEF1A1P27, TLR6, LDLRAD2, OSCAR, ADAMTS16, CACNA2D4, DLX4, DENND1C, RASAL3, PRICKLE3, AQP3, FAT2, EIF4A1P6, TNFRSF11A, AC132812.1, CFB, KLHL6, PPARG, IL6R, TMEM52B, SIGLEC11, NLRP3, IPO4, RPL7P10, NPEPL1, CYP2S1, NOD2, GPR132, ATP8B4, CD300E, WDFY4, TMEM150B, VSIG10L, GAS6-AS1, IL15RA, EPHA1, VENTX, SLC22A18, SLC9A7P1, EIF4A2P3, LILRB3, SLC46A3, TUBB1, HSPD1P6, GPR84, KYNU, CD300C, PTPN22, RN7SL138P, CD7, INTS6P1, RIPK3, PTPRH, IL2RB, DOK2, RASGRP4, GLIS1, STX11, CD28, PLEKHS1, PARP15, IL12RB1, BATF, MIR222HG, IFI30, TNFRSF9, CD68, CLDN7, MUC3A, NCF1C, HRH2, FCAR, SELE, SNORA60, SH3RF2, LACC1, FFAR4, NLRP4, NMD3P1, C17orf102, SUCNR1, ESR1, LINC01127, FRK, ACTA2-AS1, TNFSF15, PTCRA, RAC2, FCGR2C, CYP26C1, PKD2L1, AL021807.1, ADAM32, EPS8L1, PCDHB18P, TNFRSF10A, NCF1B, ICOSLG, ADTRP, TMEM26, S100P, RCVRN, TMEM236, GPBAR1, PIK3R5, LINC00968, TIMD4, MESTP1, EMB, PIK3R6, C11orf45, SRMS, CXorf21, MS4A4E, FAM3B, TRIM58, LINC00944, NCF2, TRIM63, ELANE, CD300LF, PCDHA8, DMBT1, CHST13, FLG, GUCY2D, PPATP1, SNORD89, LCNL1
In vivo GSA scRNA-seq	C0	APOE, APOC1, CLU, SERPINE2, BASP1, EGFR, LUZP2, cons3, EMP3, CST3, C1QL1, TPM2, MIA, SEMA5A, IGFBP5, TTYH1, PLAT, RPS5, LGALS1, RPL22L1, RPL28, TIMP1, HSPB1, SELM, CITED1, ITM2B, VGF, C1orf61, PCSK1N, CTHRC1, CYB5D2, SCRG1, APOD, PPAP2B, GRID2, ATP1B2, EMP2, TSC22D1, ZNF667-AS1, SEPW1, CHMP2A, RABAC1, TCEAL3, NKAIN4, YWHAE, NDUFA3, PHLDA2, PFN1, POLR2I, HEY1, RAB2A, AKAP12, C4orf48, NTRK2, BEX4, TCEAL4, TXNIP, CNRIP1, METRN, TXNDC17, GAP43, PDLIM2, RPL18, APLP2, PLP1, TMEM98, SSR4, KDELR1, ZFP36L2, SVIP, MDK, TSC22D4, BEX1, CSPG5, CLEC11A, TMEM147, MT-RNR1, DBI, CD81, TBCB, CD63, DHRS7, cons1, SPRY1, SCAND1, CETN2, ECH1, PCDH9, PEG10, CD59, RPS16, FTL, SARAF, PSMD8, SYT11, PTN, SPCS2, DSTN, C1QBP, NGFRAP1, RPS19, ARL4A, TROVE2, TMCO1, MT3, FCGRT, DPP6, FEZ1, SAT2, TMEM256, RBM3, PMP22, CTTNBP2, SCD5, NAA38, ZNF580, SEC61G, MT-ND1, GAL, SOCS2, PDCD5, CALM1, NUDT4, SFRP1, PLD3, PSENEN, C12orf57, CPNE3, EIF3K, MESP1, RHOB, OST4, RHOC, RPS9, TCEAL8, NENF, CLTA, TSEN34, UBE2M, MED15P9, HEY2, S100A13, TSPAN7, TCEA1, TMSB4X, MRPL36, SPP1, XYLT1, HSP90B1, MPZL1, MLF2, PGF, MT-ND4L, ASAH1, URI1, UQCRFS1, GCSH, COX6B1, RPL7, EIF4EBP1, ETFB, AC012146.7, ERLEC1, RAMP2, ATP2B1, MT-ND5, MT-ND2, LAMTOR1, FTH1, LINC00152, ATP6AP2, TM4SF1, PTMS, IGFBP2, RPS11, SUB1, STARD3NL, GSTP1, RP11-660L16.2, ZFP36, TAGLN2, IL6ST, LAMTOR2, SULF2, RAB31, SNRPD2, IFI27L2, CFL1, BARX1, BAALC, TRMT10C, SNHG6, DYRK4, TRAPPC1, SEPT7, mVenus, TNFRSF12A, CD9, SNX10, IFT57, PDCD6, CCSAP, ITM2C, RAC1, LDHB, SEPT9, DRAP1, CCDC85B, TMEM219, TIMM8B, PAFAH1B3, GADD45GIP1, CD99, NUCB2, BCAP31, RCN1, RGS10, SEC62, BCAP29, CHMP4B, TMBIM6, ARL6IP5, MAGED2, RAMP1, MTPN, LEF1, LAMTOR5, WBP5, NPW, LAPTM4A, BEX2, CCDC107, TMED2, TSPO, PRMT1, IGFBP3, GYPC, RPL13A, PSMC4, AP2S1, PPIB, ZNHIT1, PGRMC1, LRPAP1, ATP5E, EDNRB, TRIM24, PPAPDC1B, PLEKHB1, MLEC, CCDC47, RNF19A, STOML2, LPL, IFITM2, ISOC2
In vivo GSA scRNA-seq	C1	CDKN2A, CDK4, TSPAN31, METTL1, RPS4Y1, TSFM, MARCH9, OS9, GAS5, RPS8, EEF1A1, RPS12, SOX4, METTL21B, RPL12, RPS18, GNB2L1, AKAP7, RPL5, RPS28, NUPR1, RPS23, RPL11, RPS14, SLC25A6, RPL36, RP11-231C18.1, ZFAS1, UQCRH, RPS10, ARL4C, NDUFA11, RPL22, RPS17, EEF2, COX7A2, PRDX1, LAMA4, C6orf48, NUP93, RPLP1, HSP90AB1, RPL35A, EEF1B2, PSTPIP1, LRRC75A-AS1, NEUROD1, EIF3E, CCND1, RPS13, RPL10A, RPL4, CUTA, RPL39, RPL32, RPL18A, SEPP1, RPL10, RPS27L, RPL34, RPL30, S100B, RPL31, EIF3H, NACA, YBX1, GNG5, RPL23, SNHG5, EIF3G, TIMM13, WWTR1, HIST1H2AC, BTF3, HIGD2A, NDUFS5, MYL12A, RPL14, RPS2, MARCKSL1, RGS16, PVT1, RPL6, RHOBTB3, WDR83OS, RPS7, RPL26, EIF1AY, LYRM4, PABPC1, MTAP, UQCRB, SRM, RPL21, HES6, RPL37, PFDN6, ATP6V0B, LSM7, SF3B5, EPB41L4A-AS1, RPF1, RP3-428L16.2, RPL24, HINT1, HAPLN1, LPPR5, UQCR11.1, CTD-3014M21.1, RPS15A, GLIPR1, JUNB, C19orf70, CAP1, DANCR, RPS3A, NDUFAF4, RPL29, ZYX, RWDD1, RPS3, BAG1, RPS4X, TSPAN3, CSAG1, SNHG19, NSA2, NRXN1, PFDN5, TPT1, WASF1, MRPL54, COX7C, RPS24, RPL27A, SOD3, RP11-425L10.1, NPM1, GAD1, EDIL3, IGBP1, CAP2, SH3BGRL3, PDCD2, RPL3, SCP2, RPP40, HNRNPA1, DPH5, ALKBH7, AKIRIN1, OAZ1, LMO4, MIF, CRIPT, MRPS10, UBL5, C19orf43, RSL1D1, COX5A, ID2, CDKN2B, HIST3H2A, MAGEA1, FOS, UGT8, TXN, ARC, RPS25, THSD4, RPL41.1, IER2, ATP5G2, CTSC, RPL15, ETV1, C15orf61, HMGN3, RPL13, RSL24D1, RP11-698N11.2, RNF5, MAGEA6, TIAM2, MAP1LC3B, ATP6V1G1, RPS6, UBA52, C11orf96, RPS26, RPL19, ABRACL, INSM1, YTHDF2, NREP, NHP2, SFT2D1, RPF2, PTRHD1, RPL9, CKB, ADAMTS1, PRDX2, BTF3L4, CCND2, MRTO4, PDGFA, TCEB2, PBX3, TAF12, UFC1, PMP2, VTA1, OLIG1, EEF1D, FAHD1, RPLP2, FAM173A, EIF3I, SPATC1L, CHIC2, ATP5O, GOPC, YRDC, SYF2, KDM5B, PNRC1, RPL37A, ATP6V0E1, AVIL, PSMB1, RPS27A, SSU72, PHYHIPL, GNL2, LINC01158, RPS27, EBNA1BP2, MTIF3, HDAC2, IQGAP2, PPP1R11, SNHG8, MRPL14, MRPS18A, LSM10, TATDN1, ATP5L, HMOX2, PRAME, GADD45G, ICK, RPL8, RPL21P44, ACOT13, KIF21A, DHCR7, EPB41, MRPS15, SERBP1, RP11-231C18.2, RIOK1, MPC1, INPP4B, POLR3K, RPL36A, SYT1
In vivo GSA scRNA-seq	C2	NEAT1, MALAT1, GOLGA8A, GOLGA8B, KCNQ1OT1, DST, Rn45s, SREK1, SRRM2, OS9, TRIO, PAXBP1, NKTR, GABPB1-AS1, CCNL2, MACF1, RPL21P44, REV3L, COL20A1, FUS, PTPRZ1, SNRNP70, DDX17, MT-CO1, WSB1, GPR98, OGT, TXLNGY, MT-RNR2, CCDC144B, AKAP9, MDM4, TSPAN31, PNISR, LUC7L3, VPS13C, RP11-161M6.2, N4BP2L2, MT-CO3, POLR2J3, PVT1, SOX6, LUC7L, SACS, MT-ND3, ANKRD11, DDX5, FAM49B, TMEM259, CYP27B1, MAT2A, PRPF4B, ZCCHC11, BDP1, MT-CO2, MMP16, NOVA1, SON, chrHS-22-38-28785274-29006793.1, BOD1L1, ATM, ARGLU1, ZRANB2, ZNF292, COL9A3, HNRNPH1, RERE, MT-CYB, SALL3, ZFYVE16, COL11A1, ATRX, ANKRD10, SRSF11, MT-ATP6, PEAK1, NAIP, SMG1, FLNA, CNKSR3, CCNL1, CHD9, ZNF207, MAPK8IP3, CCDC14, SUGP2, CHTF18, ZNF638, LL22NC03-2H8.5, LRP6, SF1, PPP6R2, EWSR1, ILF3, MARCH9, SERINC5, LINC00461, MAN2A2, KMT2C, ASH1L, USP34, SPAG9, UBE2G2, LINC00969,

		CELF1, KIF21A, RP5-1039K5.19, GOLGB1, FAT1, FNBP4, HNRNPU-AS1, VCAN, THOC2, GLS, SPEN, MALAT1.1, DIP2A, SIPA1L2, AGO3, RBBP6, MT-ND4, RSRP1, NUFIP2, ANKRD36C, CASC21, EPB41L2, PLEKHH2, DYNC1LI2, RIF1, EIF3B, CHD7, SSFA2, ACAP3, BPTF, PDGFRA, PTPRS, VMP1, TNRC6A, SLC16A1-AS1, PNN, ZKSCAN1, SNHG14, AC159540.1, SPTBN1, HCG18, HERC2P2, HIPK2, CENPF, TTC3, AHI1, MT-ND5, LARS, MAP3K4, MT-ND4L, CTD-2537I9.12, PKN2, KMT2A, GPBP1, EIF4G3, GTF2H2, PRRC2B, CHD4, TRIM44, ANKRD12, SNRNP200, CSPG4, UTRN, ZC3H11A, PRKDC, TNRC6B, RBM5, CDC37, MDN1, LRPPRC, SFPQ, CAMSAP2, RNPC3, RHOT2, SPPL2B, SRGAP2, GPATCH2L, MAP3K1, BIRC6, KIF1B, METTL1, MLLT4, RCC2, TNRC6C, MT-RNR1, MCM3AP, ZMYM2, IGF2BP2, GABBR1, DDX3Y, TAOK3, PCSK7, SRSF1, HUWE1, MCL1, KANK1, EIF4G1, MARCH6, SKIV2L2, SYNE2, MAML2, CPNE7, PABPC1L, QKI, SF3B1, RPAP2, ATP1A1, FBXO22, ITPR2, PPIP5K2, WDR60, TMEM161B-AS1, RBM25, WDR90, SHPRH, MSI2, NCOR1, SOX11, LARP1, CLCN7, IGF1R, IQGAP1, TTC14, BRWD1, RP11-1023L17.1, MT-ND2, CREB5, CHD1, TPR, LPHN3, TCF4, ZEB2, PHLDB1, PLCG1, UBE3A, SPRED1, EZH2, KNOP1, MCAM, AKAP13, MYSM1, EGR1, PLXNB1, TAF1D, RP11-382A18.3, NPEPPS, RBM39, PRPF38B, CPSF6, SETX, PSMA3-AS1, SYNC, CTC-444N24.7, CC2D1A, TRIM9, LENG8, NCAN, HGS, GOLGA4, FAM133B, AGO2, SLC26A2, SMCHD1, MT-ND6, AGAP2, MAP2, DYNC1H1, HNRNPA2B1, FAM118A, XRCC2, DDX46, MAP1B, PAPD4, ZNF37BP, GGA1, NOM1, FASN, KIAA1109, USP9X, RANBP2, KANSL1, DDX39B, YEATS2, UPF3A, TTL, IKBKB, DLGAP1, RNMT, TRIP12, XPO1, NASP, PCDH7, DSEL, AHSA2, ABCF1, MT-ATP8, SMARCA4, EXOC7, KIAA0907, NUDT3, CTB-89H12.4, IGF2BP3, BRD7P4, SLTM, ZNF451, DDX3X, PKD1, TAOK1, ZMYND8, SCAF11, NCKAP1, WDR11, VPS13A, IL1RAPL1, KAT2A, TRPM7, SLC4A7, TBL1XR1, USP8, ANKRD10-IT1, RP11-571I18.5, TNPO1, HERC2, PHIP, GUSBP3, RP11-571M6.7, ABI2, MTATP6P1, RUFY3, RSBN1L, TUG1, RBM28, RP3-368A4.6, SRSF10, MT-ND1, ATP13A3, CEP152, ADAR, XXbac-BPG283O16.9, GRIK3, NSUN5P1, ADNP, MAGI2, CLTC, INPPL1, CHRM3, DHX9, HPS4, MBNL1, MUS81, BRD1, AFF4, TCF25, TIA1, MYH10, PIAS2, DKFZp434P228, TSFM, POU3F2, ZNF644, MEF2A, RP11-631M6.2, MAP4K4, PCDH9, NIPBL, LINC01578, TTC37, USP22, RP11-444D3.1, HERC2P9, MYO9A, CREBZF, MYO19, NFATC2IP, PUM1, FUBP1, AFG3L2, SORL1, CEP350, CDK12, MIA3, DNMT1, PCNXL4, RP11-366L20.2, PPFIBP1, USP7, PRKY, WDR73, EIF4A2, CCDC88A, ATP9B, ZC3H7A, HOMER1, FTX, WHSC1L1, PJA2, KRR1, ANKRD36, ICE1, MYO10, SMC1A, OCLN, PCNXL2, TRA2B, SRSF2, DDX42, INTS1, NES, OIP5-AS1, APC, EML4, EPN2, LRRC58, SENP6, RRBP1, SLC38A2, KIAA0020, MYEF2, DMXL2, IFT80, DNAJC10, RP11-513I15.6, RP11-315A16.1, AVIL, PWAR6, JMJD1C, DHX36, AFG3L1P, AC005154.6, HNRNPU, FRYL, ATG16L2, SPTAN1, MAP4, ANKS3, BAI3, MYO9B, TCERG1, TARBP1, PLXNA3, TJP1, NRBP2, PTCHD1, LRP8, PHF3, PAPD7, GATAD1, ANKDD1A, NPIPB5, PIEZO1, APBB2, TRA2A, RNF213, RP11-421E14.2, SLC44A1, FGFR1, ZBTB37, MTCH1, SOX4, MLLT6, BAZ1B, CBX5, SLC25A36, EIF2AK4, ARID4B, ONECUT2, PHF14, FUT9, PCMTD2, RP3-368A4.5, ANKRD17, CNTN1, TBC1D9B, NSRP1, PRRC2C, SGSM3, GTF2I, DCP2, SNHG9, RBMX, MEF2C, NRD1, ZNF891, LINC00657, SUPT16H, ZNF621, PTPN13, NR2F1, MMS22L, PRMT2, C5orf24, HERC4, CTD-2228K2.7, PPP1R12A, UNC80, RP3-394A18.1, KMT2E, DDX21, BTAF1, RP11-463O12.5, APP, GOLIM4, EIF3A, KDM5D, AC078842.4
In vivo GSA scRNA-seq	C3	HIST1H4C, TOP2A, NUSAP1, UBE2C, HMGB2, MKI67, CCNB1, CENPF, TPX2, ASPM, BIRC5, HIST1H1C, HIST1H1E, CENPE, PRC1, PTTG1, GTSE1, DLGAP5, KIF4A, UBE2S, TUBA1B, PBK, UBE2T, CCNB2, HMGN2, DHRS2, SGOL2, CDK1, SMC4, ARL6IP1, HIST2H2AC, CKS2, CASC5, TYMS, CKAP2, KPNA2, H2AFZ, SPC25, MAD2L1, CDKN3, AURKA, HMMR, ECT2, PLK1, KIAA0101, NDC80, CKS1B, NUF2, HIST1H3B, KIF14, CCNA2, PRR11, NCAPG, TACC3, HMGB1, KIF11, H2AFX, HIST1H2AH, TUBB, DEPDC1, ESCO2, KIF23, BUB1B, RRM2, SGOL1, STMN1, CDCA8, CENPK, DTYMK, CENPA, SMC2, CENPU, KIF2C, HMGB3, TMSB15A, MIS18BP1, AURKB, CENPW, DHFR, ARHGAP11A, CDC20, CENPH, CENPM, CEP55, DBF4, RAD21, KIF20B, HIST1H3D, TK1, ANP32E, LMNB1, DUT, CCDC34, TMPO, TUBB4B, CALM2, FAM64A, RNASEH2A, ATAD2, FANCI, NUCKS1, CDCA5, HJURP, RAD51AP1, CDCA2, BUB1, RPA3, DEK, HIST1H2AG, FOXM1, RTKN2, TTK, DIAPH3, DEPDC1B, RANBP1, NCAPD2, BRCA2, H2AFV, CKAP5, FBXO5, RRM1, KIF22, NEK2, NCAPH, ANLN, MND1, SKA3, KNSTRN, KIF15, CDCA3, BUB3, MXD3, SHCBP1, RACGAP1, GGH, SKA2, DDX39A, KIFC1, TMEM106C, GMNN, TUBA1C, HN1, KIF18A, CLSPN, TYMSOS, LSM4, MELK, NCAPG2, LBR, SPDL1, HIRIP3, ZWINT, BRCA1, MGME1, PCNA, CHEK1, SPAG5, MYBL2, SMC1A, ORC6, RFC3, MZT1, NUDT1, NCAPD3, PSIP1, CKAP2L, G2E3, USP1, LGALS1, TRIP13, H1FX, DCXR, ANP32B, PSRC1, RAD51C, C21orf58, CENPN, VRK1, EZH2, GAS2L3, BARD1, PHGDH, CDC25C, ATAD5, MIS18A, HP1BP3, KIAA1524, CEP135, SNRPG, POLD3, CEP152, TROAP, MNS1, NRGN, SAE1, PRAC2, ALYREF, RHNO1, ASF1B, MCM4, EMP2, LSM5, E2F7, CCDC18, RFC2, SYNE2, TMEM97, EXOSC8, COX8A, OIP5, SMC3, NEIL3, WHSC1, FEN1, MRPL51, KIF20A, MCM10, ITGB3BP, STRA13, HSPB11, CIT, PSMC3, DNMT1, RFC4, ANAPC11, DNAJC9, FBLN1, PRIM1, CENPJ, CDC25B, SNRPB, GNG4, PTMA, HNRNPD, RAN, SKA1, CSE1L, HINT2, RMI2, SETD8, TUBB6, DSN1, CDK5RAP2, CTCF, LSM3, COMMD4, CCNF, YWHAH, CDKN2C, PIF1, CTHRC1, FANCD2, UBALD2, LIG1, DDIAS, NMU, CARHSP1, ERH, CMC2, MCM8, SUMO2, MAGOHB, RDX, RBBP8, PGP, CDC45, ARHGAP11B, HMGN5, MCM7, PSMC3IP, TUBA1A, FANCB, SAP30, PPIA, ACYP1, BTG3, COX17, HIST1H2AM, CDC6, PARPBP, CCP110, GINS1, CKB, SCLT1, CBX1, NASP, WDR34, TOPBP1, CENPL, XRCC2, RHEB, RNASEH2B, TTF2, TUBG1, TFDP1, KIF5B, SLC25A5, TMX4, HIST2H2AA4.1, RBBP7, CNIH4, MYBL1, ILF2, HELLS, FAM133A, C12orf75, PKMYT1, SIVA1, SPC24, BANF1, LRR1, PDS5B, CNTRL, POLE3, HIST2H2BF, CKLF, ARHGEF39, ASRGL1, HMGN1, POLA1, DCTN3, BCL7C, SAC3D1, RANGAP1, ACAT2, CALM3, NDE1, ZWILCH, NSL1, KIF18B, HNRNPA2B1, HNRNPUL1, CDCA4, PARP1, EXO1, YEATS4, PPP2R3C, TP53I11, MRE11A, PPIH, HNRNPA3, H2AFY, HAUS1, CENPO, CEP57, PXMP2, CDC27, PRPSAP1, DTL, BLM, DCP2, SAPCD2, HAUS8, CHAF1A, SH3KBP1, CBX5, PTGES3, CLGN, TPRKB, HAT1, SUZ12, SMS, MTHFD2, IDH2, HADH, HIST1H4H, HIST1H2BF, CSRP2, SASS6, RPL22L1, DESI2, HNRNPH3, UPF3B, CENPC, GINS2, NFYB, CHRAC1, PHF19, ANP32A, CDKN1B, C12orf65, BRIP1, ITGB1BP1, SNRPD3, HIST1H4E, INCENP, CNRIP1, PPIF, SRSF3, SFPQ, CACYBP, GEN1, LDHA, RBL1, ACTL6A, CEP295, ATP2A1-AS1, STIL, PFN1, RAD23A, CDCA7L, RPA2, TIMM10, NUP50, HLTF, HDGF, TXNDC12, XRCC4, HAPLN1, CCNG2, C19orf48, BOLA3, TUBB2B, IQGAP3, DCAF7, HPRT1, WDR76, PLK4, PPDPF, CEP78, FOPNL, PA2G4, MT2A, PMF1, ALDH7A1, TPI1, NT5DC2, ANKRD32, TMEM237, AP2S1, NUP85, LMO7, STAG1, CDT1, PGRMC1, AKR1B1, SRP9, SEPT10, BRD8, KIAA0586, SUPT16H, NAA38, CENPQ, ZMYM1, UQCC3, POP7, SPA17, COQ2, GPSM2, ZGRF1, DCK, WBP11, KPNB1, HIST1H2BC, TIMELESS, MSH2, NRM, KHDRBS1, FADS1, GAL, CEP70, PITHD1, IKBIP, CHEK2, VPS29, NDUFA6, JADE1, MCM3, MASTL, FAM136A, ODC1, APOLD1, EMC9, SSRP1, C14orf80, CHCHD2, PARP2, PCNT, PNRC2, PPM1G, XPO1, GPN3, ADD3, MPHOSPH9, MAD2L2, ERI2, CRNDE, IFI27L1, EZR, C5orf34, POC1A, CEP57L1, HIST2H4B, POLQ, RAD1, PKM, EIF5A, UCHL5, RAD18, RSRC1, RPA1, RNF26, UACA, ASCL1, HIST1H2AE, BASP1, NUP37, PHIP, FAM83D, TPGS2, RNASEH2C, ARL6IP6, METTL4, PAICS, VDAC3, MZT2B, PHTF2, NUDT15, EXOSC9, YWHAQ, XRCC5, SHMT2, TFAM, PIN1, PSMG2, RAB8A, HIST1H3H, HMGXB4,EIF4EBP1, PAFAH1B3, PKIA, HNRNPR, SEPT7, ZNF714, SKP2, PCM1, ACAA2, CETN3, EME1, HES6, HNRNPAB, VBP1, TOMM5, NFATC2IP, FXYD6, VKORC1, FAM76B, GLRX5, NUDCD2, COPS3, ZNF724P, C4orf27, H3F3B, SET, POLR3K, FDPS, RALY, NUDT21, WDR54, PRDX3,

		BAZ1B, RBX1, RBMX, FAM122B, HN1L, MRPL22, SNRPD1, GGCT, SNRPE, E2F8, CCDC167, PCBP2, PGAM1, PRPS1, HRASLS, H3F3A, CNTLN, C9orf142, ICT1, TCTEX1D2, RMI1, CEP128, ENDOG, LIN9, PRKAR2B, CENPI, LKAAEAR1, NUDT5, WEE1, CISD2, CCDC150, SNX5, DERL3, CEBPG, NUP155, TAF15, POLA2, TBC1D1, HMGA1, KDELC2, MTFR2, MED21, CTSV, NCAPH2, GSTP1, MRPL17, RECQL, HSP90AA1, SOGA1, NPW, CBX3, WDHD1, CTDSPL2, RELL1, XRCC6, OXCT1, LIN54, CCT5, DNAJC21, SLC25A11, PSPH, TECR, RAD51, COMMD2, UBA2, C2orf69, TMEM60, MORC4, DYNLL1, ITGAE, SLBP, HIST1H2BO, PDIA4, CTD-2194L12.3, CHRNA5, RFC1, MYH10, UQCC2, HIST2H2AB, CPSF3, FGD5-AS1, SLC29A1, CBR3, CDKN2D, QSER1, CYB5A, SNRNP40, TMX1, CCDC77, PPIG, RFC5, PFN2, PPP2R5E, RUVBL2, EIF4EBP2, FAM111B, NSMCE4A, GMPS, FUZ, DPM2, FAM161A, CEP112, GPD2, MAGOH, KATNBL1, NEDD1, PPP1CA, TMEM160, LYAR, PTPLAD1, GAMT, TMEM107, RP11-620J15.3, ARMC1, MEA1, MZT2A, RP11-480I12.5, POC5, RPN2, NCAM1, SMCHD1, RCC1, GINS4, MORF4L2, SQLE, FN3KRP, IMMP1L, CFL1, GRK6, MYEF2, SEPT11, TNFAIP6, FAM96A, DLEU2, USP13, MRPL23, RFWD3, IER3IP1.1, PIGX, CEP192, HSPA14, LSM8, NDUFC2, PDCD5, CADM1, NAP1L4, PPP2R5C, UBL7-AS1, LRRK2, GLO1, NUP62, RHOA, PCBP1, ENC1, DONSON, MMD, IFT81, HAUS6, ARL6IP4, NIF3L1, ASNS, VEZF1, PRKDC, MIF4GD
In vivo GSA scRNA-seq	C4	CXCL10, ISG15, IFI6, IFITM3, BST2, HLA-B, IFI44L, IFI27, GBP1, IFIT3, IGFBP5, IFIT1, MT2A, PARP14, IFIT2, HLA-E, STAT1, B2M, WARS, HLA-C, VEGFA, PLSCR1, IFITM1, RSAD2, RNF213, HERC5, IFITM2, NRN1, APOL6, RDH10, MX2, SP100, LY6E, IGFBP2, DDX58, IFI35, PDGFRA, BNIP3, LGALS3BP, EIF2AK2, SAMD9, IFI44, IFIH1, LAP3, SPP1, PSMB9, XAF1, TAP1, HLA-A, GAPDH, IRF1, A2M, RARRES3, NDRG1, SOCS2, SAMHD1, DLK1, SOCS1, ISG20, OAS1, ARID5B, USP18, HILPDA, LYPD1, TRIM22, SERPING1, SLC2A3, TMEM158, PSME1, EGFR, PARP9, CCSAP, SLC5A3, SAMD9L, DTX3L, ATP1B2, UBE2L6, AKAP12, CD9, FOLH1, NUPR1, CXCL11, FTL, SELM, LGMN, TMSB10, OAS3, FN1, C5orf56, C19orf66, IFIT5, ERAP2, CD74, MT3, VGF, ADAR, ALDOA, PSME2, ERRFI1, STAT2, GLRX, NUB1, TNFSF10, MDK, HERC6, TAPBP, BANCR, ENPP2, IFI16, CHPF, CALD1, RNF19A, PTGDS, APOL2, PLEKHA4, BNIP3L, TRIM56, SOCS3, BRI3, ST13, NAMPT, NCOA7, OPTN, RBCK1, SCD, RABAC1, UGP2, GBP3, CMPK2, APOL1, PNPT1, PLOD2, HES4, TRIM25, CTSS, HELZ2, PPM1K, TNC, P4HA1, IL18BP, CD47, NT5C3A, OASL, APP, SQSTM1, AGT, IL13RA1, DRAP1, LRP2, ITM2B, TAP2, PML, MAF, RP9P, WSB1, CD59, ARRDC3, PDE4B, HIST1H2AC, RPL28, DDX60L, OAS2, GBP2, CFH, RAB4A, GSTK1, SPATS2L, EIF1, ZNFX1, CALCOCO2, PSENEN, HM13, SMIM14, DDIT4, OSMR, TNFSF13B, NANOS1, FTH1, BHLHE40, C4orf3, CHMP5, XRN1, GLTSCR2, CALR, LGALS1, CEBPB, DNAJB9, IFI27L2, HSPB1, BAALC, DNAJA1, CIR1, SP110, ART3, PCGF5, IDH1, SAT2, ERAP1, MED15P9, EPB41L4B, TIMP1, SEC61G, CLU, KDELR1, RTP4, CHIC2, PMP22, SCAMP1-AS1, ARMCX3, EGLN3, PTP4A3, ACSBG1, LAPTM4A, SCN9A, SSR4, OBSL1, ASPH, CEBPD, LNPEP, TRIM69, GAS5, ARHGAP42, RNF114, RPS5, ZFAS1, GOLGB1, NUCB1, UBALD2, DDX60, RPL36AL, FKBP9, WBP5, NFIL3, FBXL5, FNDC3B, BATF3, SOCS2-AS1, CAPNS1, YPEL5, PLD3, TMEM219, CSTB, LST1, C8orf59, ANKRD20A4, SARAF, CASP4, C4orf33, RPL29, PSAP, ZNF581, SH3GLB1, GSDMD, C12orf57, GUK1, OST4, RPS19, SH3BGRL3, APOL4, CSF1, TMEM45A, GBP1P1, TRIM5, RHOU, FAM45A, RPL18, PDLIM2, FCGRT, STAT3, SHISA5, IL6ST, PTPRA, TRANK1, CARD16, CLIC2, AC007246.3, KDM7A, CASP1, C1orf53, KIAA0040, ZNF608, DHRS2, RGS10, PCMTD1, RAB13, BATF2, FAM89A, ARMCX1, TMEM178B, NFKBIA, NPC2, AGTRAP, TNFRSF1A, EFNA1, LHX9, ARPC1B, RORA, COMMD6, MITD1, FAU, TNIP2, SRPX, PLTP, SERF2, CST3, TSHZ1, CA9, ATP6V1F, GLIS3, TLR3, RNF7, RPL39, AFF3, INSIG2, LINC00152, IRF2, FAM43A, OGFR, TPP1, MIR4435-1HG, TRAFD1, HAS2, HIST1H2BD, TFPT, DHX58, PTPN2, IGF2, PARP10, TRIM21, SKAP2, RBM43
In vivo GSA scRNA-seq	C5	SNHG9, RP11-329L6.2, RPS17, RPS29, RPL13A, RPL38, OS9, CDK4, RPL35, RPL36, RPL37A, RPL27A, RPL15, ZSCAN16-AS1, RPS11, MRPS27, NT5C3B, COL20A1, RABGGTB, AKR1B1, SLMO2, DAP3, BCYRN1, NELFE, NAA20, RPL32, AK2, THOC7, SRSF6, ACBD3, CYB5R3, RAB5C, RPL21, CCDC25, MFGE8, GNB2, MRPL1, BIRC2, EEF1A2, TMEM126A, CGGBP1, CNOT7, EEF2, RPL34, CIB1, ISOC2, ATP6AP1, C6orf48, RAB8B, YES1, RPS27, RPLP2, RPL3, TARDBP, RAP1B, POLE4, AHCY, USP3, TMEM208, HINT2, IRF2BP2, GTF2F2, RPL23, FSCN1, FAM192A, CHCHD5, KRAS, RPS28, DDX10, CLIC4, IPO5, NUDT5, COPZ1, MPST, TRIM28, MRPS18B, NOP14, ADI1, PPP2R1A, ELAVL1, TMEM259, SF3A3, ATP5E, CUEDC2, MAP2K2, SDCBP, PSME2, SDHD, RPL14, BAX, BNIP3, KPNA4, ETF1, TMEM161B-AS1, ZNF770, POP7, C7orf55, RPL22, MRPS15, MCM7, AP3B1, PPP1CA, PPP2CB, PGRMC2, MTHFD2L, NUDT4, MPC1, RAD51C, NOC3L, TMOD3, TXNDC12, UBE2E3, ITGAE, CLPTM1L, UTP11L, PBDC1, DNAJC21, CNIH1, ARPC1A, TIPRL, SDF4, KCTD20, RPL24, NDUFV1, FAM98A, FAM177A1, DPM3, MRPL27, RPL37, PCNP, TTC19, APMAP, GDI1, CEBPZOS, SEC61A1, PREPL, RNF181, REEP5, ERCC1, BCAS2, RPS18, ETV5, SELT, CCDC90B, RPL23A, NMT1, C1orf131, TFRC, KPNA2, RPS12, KANK1, SSU72
Ex vivo IDH1^{wt} GSA scRNA-seq	C0	MYC, ACAT2, LDHA, ENO1, TSFM, CCT5, SRSF7, PRDX1, HSP90AA1, NPM1, HNRNPA2B1, C1QBP, RAN, ATP1B3, PSMB1, PDIA3, CD9, PPIB, PRDX4, FDPS, PRMT1, HSP90AB1, SSB, PKM, PCNA, PDIA6, VDAC1, PSMA5, CD63, PSMC4, TUBA1A, PGK1, PSMA4, PSMB6, SRI, ATP5C1, HSPD1, PSMB3, ATP5B, OLIG2, MRPL13, NQO1, SLBP, FDFT1, PHGDH, CCT7, SPCS2, EIF4A3, LAPTM4A, LDHB, SKP1, EMC7, MDH1, SDHB, NDUFB6, HNRNPH3, CCT8, ILF2, XRCC6, HNRNPM, NASP, PSMA1.1, SNRPB2, UBC, RPN2, ID4, TMED9, MRPL3, DDX5, PSMA2.1, CTSC, PSMC5, SPCS1, CACYBP, PTTG1, ALDOA, DCAF13, HLA-A, RUVBL1, COPS6, SSBP1, EBNA1BP2, MCM7, TMBIM6, NME1, YWHAQ, FKBP4, PGRMC1, PGAM1, UQCRC2, WBSCR22, PSMD2, NDUFC2, HSPE1, PSMD8, LAPTM4B, PDHB, PFDN2, ADH5, EIF3I, HSPA9, UQCRFS1, MCM3, PSMD14, DKC1, PSMA7, CFL1, EIF5A, PRDX6, EID1, POMP, DHCR7, PSMA3, PRDX3, C14orf166, SNRPA1, B2M, PPT1, ETFA, UCHL1, CD81, NUDC, PPIA, PPA1, HNRNPF, CCT3, ATP6V0B, PCMT1, EMC4, TRAP1, MRPL47, PSMB2, PDHA1, MRPL40, TXNDC12, BCAP31, RTCB, RAB7A, MYL12A, TM4SF1, APLP2, SCRG1, MRPS18C, AIFM1, GLRX3, PARK7, OLIG1, DBI, HNRNPD, MAGOH, DPM1, IMP3, MDH2, EIF6, CNIH4
Ex vivo IDH1^{wt} GSA scRNA-seq	C1	GAS5, C1orf61, RPL13A, DDIT4, SLC25A6, OLIG1, EEF2, BTG1, MXD4, EEF1A1, LMO4, NEAT1, RP11-231C18.1, RPL28, ZFAS1, RPL3, EPB41L4A-AS1, VIM, PNRC1, RPL5, C6orf48, RPL10, RPS18, RPS11, LRRC75A-AS1, RPS27, HNRNPA1, RPS9, RPS28, RPS19, SAT1, MARCH9, RPL7A, RP11-329L6.2, RPS12, RPS5, RPL11, MT-CO2, GLTSCR2, IGBP1, EIF3H, RPL12, RPL10A, RPL18, MT-CYB, SNHG5, RPS15A, RPL13, SNHG8, RPL34, RPS16, RPL14, RPL15, RPL18A, RPS14, EIF4B, RPL22, RPL7, RPL4, TPT1, GNB2L1, RPL3P4,

		RPS2, MT-ND2, PFDN5, RPS25, RPS27A, IGFBP5, RPL35A, MT-ND4, HIST3H2A, RPS29, EEF1B2, RPL30, RP11-234A1.1, LINC01158, RPS4X, RPL39, RPL37, ATP5G2
Ex vivo IDH1^{wt} GSA scRNA-seq	C2	CTNNB1, HNRNPH1, SOX11, SET, PTPRS, CDC42, C1orf56, GOLIM4, CAPZA1, WTAP, PHKG1, RQCD1, CTTN, FTL, MDM4, YWHAE, MT-ND6, EIF2S3, TSPYL1, PPP1CB, ANKRD40, METRN, SRSF6, MT-ND1, MT-ND4, MT-ND5, RP11-538P18.2, PPP1R14B, LRRC75A, OS9, RPS29, MT-CO2, CDC42SE1, PPP3CA, WDR43, CDK6, HMGCS1, MT-CO1, RPL41.1, KCNQ1OT1, RP11-294J22.6, NUCKS1, LINC01578, RPS27, SOX4, MT-ATP6, KPNB1, MT-CO3, HMGN2, GTF3A, HMGA1, LETM1, ADNP, NUP155, LSM12, RPS28, TOR1AIP2, OSBPL8, NEAT1, PTP4A2, IAH1
Ex vivo IDH1^{wt} GSA scRNA-seq	C3	HIST1H4C, HIST1H1C, KIAA0101, HIST1H1E, TYMS, DUT, CDK4, UBE2T, TK1, PCNA, RRM2, DHFR, TSFM, CARHSP1, HMGB2, STRA13, TUBB, CKS1B, PSMC3IP, ZWINT, HIST2H2AC, SDF2L1, FEN1, MYBL2, RANBP1, CDCA5, UBE2C, IDH2, HSPB11, TOP2A, CENPK, METRN, ALYREF, CENPM, GINS2, TMEM106C, CLSPN, DCTPP1, TUBA1B, RFC2, AURKB, TNFRSF12A, METTL1, HSPA5, CENPU, RPA3, GMNN, DNMT1, CDK1, MCM7, RAD51AP1, RAD51C, ASF1B, WDR34, CSRP2, DNAJB11, SIVA1, MCM3, RFC4, USP1, CDC45, RNASEH2A, H2AFZ, POLR3K, DEK, PSMG1, HSP90B1, SMC2, MAD2L1, TUBB4B, ATAD2, CISD2, UQCRC1, CACYBP, COX5A, MRPS34, MRPL37, POP7, SMC4, EMP3, PPP1CA, ADRM1, BUB3, CD320, HAUS1, PPM1G, SMC1A, TUBA4A, CISD1, NTMT1, LSM6, ORC6, E2F1, PGP, FBXO5, IER2, TTYH1, PIN1, UBE2M, HAPLN1, DYNLL1, PHF5A, UBE2I, XRCC5, TUBG1, KIF22, TOMM40, CCND3, NETO2, PPIF, H2AFX, MAD2L2, VPS29, EXOSC9, CDCA4, ENO1, NDUFS3, TMPO, CALR, MELK, UQCC2, DTYMK, SLBP, PPIB, COMMD4, VKORC1, SAP30, DNAJC9, PA2G4, CHCHD3, GGCT, SNRNP70, PSMC3, RRM1, NCAPH2, MEA1, ARPC5L, NASP, CCNA2, AK2, IMPA2, ACAT2
Ex vivo IDH1^{wt} GSA scRNA-seq	C4	UBE2C, CENPF, CCNB1, TOP2A, CKS2, HMGB2, PLK1, NUSAP1, MKI67, ARL6IP1, AURKA, KPNA2, UBE2S, TPX2, CDC20, PTTG1, PRC1, CKS1B, ASPM, CENPA, CENPE, BIRC5, CCNB2, TUBB4B, CALM2, SGOL2, HMMR, PIF1, AURKB, CCNA2, GTSE1, KNSTRN, DLGAP5, HN1, KIF23, NUCKS1, DEPDC1, PSRC1, CDK1, CDKN3, SMC4, KIF14, HAPLN1, MAD2L1, SGOL1, KIF4A, CDCA3, SQLE, SCRG1, HNRNPDL, DEPDC1B, MSMO1, HMGB3, BUB3, SFPQ, NUF2, PDIA6, TUBB, DTYMK, KIF2C, UBE2T, CDCA8, TUBA1C, HSPA8, G2E3, CKAP2, TACC3, KIF5B, RCN2, DBF4, NDC80, LBR, DHX15, HSP90B1, PBK, MALAT1, SAPCD2, SAP30, ECT2, BUB1B, TMBIM6, RACGAP1, SYNCRIP, TUBB2B, UBALD2, KIF22, XRCC6, SPC25, CADM1, ARHGAP11A, PDIA4, RANBP1, HNRNPA2B1, OIP5, PHF19, XRCC5, ACSL3, BRIX1, APLP2, CD9, FAM64A, PRR11, ENC1, GPSM2, PSMA4, HNRNPR, SLTM, KIF20B, HSPD1, CASC5, NEK2, SMC2, UGP2, MORF4L1, PDIA3, HSP90AA1, SMC1A, SPCS2, IDI1, HMGB1, PCMT1, ESF1, INSIG1, TMEM97, APMAP, DHCR7, EXOSC8, DDX39A, ACTL6A, LDHA, NUDCD2, CENPW, HJURP, HMGN2, PSMC2, VRK1, DCAF13, CCDC47, TM4SF1, PRDX4, CEP55, CANX, ATP1B3, SNW1, METTL1, SLC3A2, SEPT7, MESDC2, MAPRE1, RNF26, SON, TROAP, SMC3, CDC27, CCT5, PTPLAD1, PHIP, CCND1, METTL5, MIS18BP1, DCAF7, ZMPSTE24, P4HB, NCAPH, HMGCR, UGDH, PSMD11, TNFAIP6, ANP32E, LARP7, CCT6A, SLC39A10, CCNF, RBM25, RAC1, GTPBP4, NUDC, RAD21, WDR43, RHNO1, PSMC3, HNRNPH3, USP14, DDX1, ALG8, BUB1, PUF60, EIF3A, HNRNPA3, CDCA2, RPL7L1, FBXO7, MRTO4, H2AFV, TTK, SMARCC1, KIF20A, U2SURP, ZC3H13, KIAA0020, RAN, NCL, HNRNPU, PTN, TMX1, DNAJB11, SERBP1, KDELR2, THOC6, OCIAD1, CKAP5, NIN, CNIH4, NCAPG, COL11A1, SAE1, CDCA5
Ex vivo IDH1^{wt} GSA scRNA-seq	C5	MLLT11, MAP1B, ID2, MALAT1, CPE, HES6, ID1, TUBA1A, SOX4, MFAP4, TUBB2B, CD24, S100A2, MGP, INSM1, NEUROD1, AKAP7, GAP43, TMSB4X, SYT1, STMN2, TCF4, STMN1, DCX, CXXC5, NHLH1, H3F3B, BEX1, ETV1, DDAH2, TMSB10, MARCKS, GPM6A, TTC3, SCG3, ID3, ACTG1, MARCKSL1, CD63, SH3BGRL, NREP, ROBO1, GKAP1, RGS16, TUBB, DLL3, CRMP1, FXYD6, TSPAN31, EIF4A2, TAGLN3, CDKAL1, LAPTM4A, DUSP6, SSBP2, C1orf61, STMN4, BLCAP, SNN, H3F3A, UBE2E3, PCBP4, ITM2B, PCSK1N, HLA-A, THSD7A, EDIL3, CCNI, CALM1, RTN4, CDKN2C, ATRX, SH3BGRL3, MYL6, CDKN1A, MAGED2, LIMD2, NUPR1, TEX9, TMEM59, CHD9, PNRC1, ATP6V1G1, PRDX1, KIF5C, MAP2, NKAIN4, FNBP1L, APOE, C12orf57, EIF1, CMTM7, UBC, ATP2B1, GABARAPL2, EBF1, TRDC, BTG1, YWHAZ, FAM127A, ANKRD65, SERINC2, DNER, TSPAN5, SLC25A6, SAMD5, CNR1, RGMB, PNISR, RAB26, TCEAL4, IGBP1, SDC2, WBP5, CHRNA3, GAS5, SHF, COMMD6, NRXN1, SVIL, DUSP5, CHRNA5, RALGAPA2, CHIC2, APLP1, EIF3H, GSTA4, GSN, MORF4L1, EIF4G2, IGFBP2, EEF2, FAM215B, PTP4A3, AUTS2, SESN3, PPP1R17, YPEL5, SBK1, MEIS3, IDS, SUMO2, AKIRIN2, FAM57B, RPRM, HIST1H2AC, DCC, SKIL, ROBO2, GAS2, SRP14, CNIH2, SNAP25, RGS2, ELAVL3, PAK3, DPYSL3, UQCRB, OST4, GSE1, ARMCX1, PERP, PHLDA3, DDR1, BTG2, MIAT, ARRDC3, MAGED1, D4S234E, AKT3, CTNND2, BTBD17
Ex vivo IDH1^{wt} GSA scRNA-seq	C6	CALM2, ANXA1, ACTB, PRDX1, PTGES3, UBC, HSP90AB1, KPNA2, CKS2, TXN, PSMA4, UBE2C, NCL, TUBB, SNRPD2, EIF4E, MAD2L1, SMS, SNRPB2, HNRNPK, VBP1, PA2G4, SOD1, UCHL1, PKM, SKP1, CYCS, EIF4A3, SEPT7, YWHAQ, GNL3, RPL35, MRPL42, RSL24D1, TUBB4B, HSP90AA1, SUB1, SNRPC, ARF4, SNRPD1, RPL31, MORF4L2, EIF3E, POLR2K, UBE2N, PTMA, ENO1, HMGB1, PGK1, H3F3A, CCNB1, CCT5, CMSS1, SNRPD3, EIF2S2, UQCRB, SSB, DSTN, CACYBP, PAICS, METAP2, MORF4L1, HSPD1, ANP32B, CCT3, DKC1, CCT4, COX7B, SSBP1, LDHA, PPIA, XRCC6, DDX5, CCT8, ATP5B, VDAC1, TCEB1, EIF1AY, SRSF3, PSMB3, NDUFA5, SNRPG, CKB, SEC61G, ZC3H15, MYL12B, DNAJC8, HNRNPA2B1, EIF5A, PFDN2, BRIX1, PFDN4, HNRNPC, NUDC, EDF1, ENY2, U2SURP, RSL1D1, UBE2L3, RPL36A, ATP5J2, SF3B6, SERBP1, LSM5, RPL21, PSMB7, NSA2, RANBP1, SUMO2, YWHAE, C1QBP, ARPC5, TPI1, LDHB, PARP1, NPM1, CBX3, MYC, NDUFB2, TBCA, HSPA8, H2AFZ, HNRNPM, SRP9, ALDOA, NME1, UQCR10, PSMB1, LSM3, RPS8, HNRNPDL, PRDX2, PFN1, GAPDH, HSPE1, NDUFS5, GSTP1, RAN, RPS4Y1, POLR2F, RPL38, NHP2L1, ERH, LAMTOR5, SNRPE, PPA1, NDUFA1, MINOS1, STMN1, RPL22, RPL26, SRP14, RPL27, RPLP0

Ex vivo IDH1^{wt} GSA scRNA-seq	C7	CYR61, ANXA1, IGFBP3, TNFRSF12A, CNN3, CRYAB, TMSB4X, ERRFI1, GADD45A, LGALS1, IGFBP5, PMP22, SERPINE2, TIMP1, RAB13, CITED2, ARID5B, MYL12B, TPM1, AKAP12, NES, CD63, CHI3L1, ANXA5, ZYX, SDC2, VIM, PHLDA2, CDKN2B, PDGFA, MYL12A, IER3, HBEGF, PLK2, RTN4, TAGLN2, RGS2, MCAM, ARF4, LRRN3, IL32, CTGF, CD59, ZFP36L2, LMNA, SH3BGRL3, TSPAN5, CLIC1, C9orf3, EIF1, EMP3, SNAPC1, RND3, HLA-A, SPARC, SNHG10, EIF4A2, ACTG1, VMP1, UGCG, ADAMTS1, PTRF, GLRX, WWTR1, MAP1LC3B, EMP1, SQSTM1, ITM2B, TMSB10, RGS3, B2M, ACTN1, LAPTM4A, AKIRIN1, NEXN, RASSF1, S100A10, IL6ST, GLIPR1, PPP1R15A, DNM3OS, PRSS23, ELL2, LINC00152, RCAN1, CD44, MIR4435-1HG, FHL3, NEDD9, TSC22D2, GADD45B, LIMCH1, SPP1, PCF11, YPEL5, LHFP, SAMD4A, TOB2, PDP1, GSN, VOPP1, FSTL1, ETS1, METTL7B, TUFT1, PRNP, TMEM43, EVA1A, HEXIM1, COL4A1, SEMA3A, H3F3B, RRAS2, KLF5, S100A16, PVRL2, RUNX1, FOSL2, RBPMS
Ex vivo IDH1^{mut} GSA scRNA-seq	C0	SCRG1, OLIG1, PRDX1, APOD, CD63, UBC, BCAN, ENO1, HAPLN1, S100B, HLA-A, LDHB, OLIG2, NQO1, TSFM, CD9, GAPDH, ACTG1, MYC, MAGED2, PHGDH, PSAT1, TMBIM6, TSPAN31, ATP5B, LAPTM4A, CCND1, NPM1, DDX5, SKP1, EID1, MDH1, MRPS18A, ID2, DDOST, LDHA, BSG, UCHL1, LAPTM4B, ACAT2, ITM2C, NEU4, MGST3, ALDOA, EIF3D, GNB2L1, EEF2, CCT7, PSMB6, TUBA1A, MRFAP1, HSP90AB1, MGLL, ITM2B, DBI, ARPC3, RTN3, RAMP1, HSPA5, COL9A3, FABP7
Ex vivo IDH1^{mut} GSA scRNA-seq	C1	CTNNB1, HNRNPH1, SET, SOX11, HIST1H4C, YWHAE, C1orf56, PHKG1, CDC42, CAPZA1, MT-ND6, PTPRS, MDM4, METRN, CBX6, GOLIM4, TSPYL1, MT-ND1, SRSF6, RPL41.1, TOR1AIP2, GTF3A, EIF2S3, CTTN, HMGA1, CDC42SE1, WTAP, PPP1R14B, MT-ATP6, ENAH, JTB, RPS29, SOX4, MT-ND4, TTC3, MT-CO3, TMSB4X, RPLP1, RPS27, RPS28, MT-CO2, MT-RNR1, MT-RNR2, CDK6, MT-ND2, NEAT1, FTL, PPP3CA, OS9, HMGN1, KCNQ1OT1, RQCD1
Ex vivo IDH1^{mut} GSA scRNA-seq	C2	CENPF, CCNB1, UBE2C, HMGB2, CKS2, TOP2A, PTTG1, MKI67, UBE2S, NUSAP1, ASPM, PLK1, ARL6IP1, TPX2, BIRC5, AURKA, CENPE, PRC1, CDC20, CCNB2, DLGAP5, CKS1B, MALAT1, NUCKS1, GTSE1, KPNA2, CENPA, PSRC1, TUBB4B, HMMR, CALM2, SGOL2, NUF2, HMGN2, TUBB, PRR11, KIF14, CDKN3, ANP32E, PIF1, HMGB3, KIF4A, TUBA1B, DEPDC1, SGOL1, AURKB, TUBB2B, HSP90AA1, KIF23, FAM64A, TUBA1A, KIF22, HSPA8, DEPDC1B, BUB1B, SFPQ, TACC3, LBR, CKAP2, KNSTRN, HN1, CASC5, MIS18BP1, SMC4, CDCA3, STMN1, G2E3, HNRNPR, HNRNPH1, DBF4, HMGB1, H2AFV, CCNA2, HP1BP3, TCF4, ARHGAP11A, HES6, KIF5B, PBK, SQLE, SLTM, SOX4, GPSM2, NDC80, ECT2, DTYMK, HNRNPA2B1, KIF2C, HMGCR, RACGAP1, OIP5, MARCKS, KIF20B, TOP1, CENPW, MAD2L1, CDK1, U2SURP, HNRNPA3, NFIA, SON, CDCA8, FANCI, HNRNPH3, NEK2, CDKN2C, SMC2, MSMO1, MAPRE1, H2AFZ, TMPO, CALM3, AVIL, DDX39A, MORF4L1, RCN2, TUBA1C, SAPCD2, PSMC2, HNRNPM, UBE2T, MZT1, SMC1A, BUB3, DNAJC8, SEPT7, PDIA4
Ex vivo IDH1^{mut} GSA scRNA-seq	C3	HIST1H4C, TYMS, HIST1H1C, TUBB, UBE2T, DUT, UBE2C, PCNA, HIST1H1E, KIAA0101, TK1, HMGB2, RRM2, MYBL2, CDCA5, TUBB4B, MCM7, TUBA1B, HIST2H2AC, CDK1, RANBP1, HSPB11, AURKB, PSMC3, STRA13, MKI67, FEN1, ALYREF, CKS1B, CDK4, SMC4, CENPM, DHFR, ZWINT, CDCA4, TUBB2B, H2AFZ, HSPA8, DTYMK, RFC2, CLSPN, TOP2A, CENPU, CDC45, PSMC3IP, SPC25, WDR34, IDH2, CARHSP1, CACYBP, ORC6, RAD51C, ATP6V0B, RFC4, FBXO5, NUSAP1, DNMT1, TUBG1, DEK, SMC2, SLBP, MAD2L1, NUDC, SAE1, HAUS1, CENPK, RAD51AP1, RNASEH2B, SIVA1, RNASEH2A, CSRP2, ASF1B, TOMM40, GGCT, SMC1A, METRN, COMMD4, DCTPP1, VPS29, PDIA6, ID1, GINS2, BIRC5, PTBP1, VRK1, PPM1G, CALM2, TNFRSF12A, USP1, SAC3D1, TSFM, TUBA4A, CD320, SRM, SGOL1, CCNA2, PSMG1, TRA2B, KIF22, GMNN, ADRM1, CHAF1A, TUBB2A, RRM1, SNRPB, DNAJB11, ST3GAL4, CRELD2, TRIP13, CENPN, TMEM106C, NCAPH2, PA2G4, BTG3, SDF2L1, MTCH1, MRPS34, C20orf24, PBK, H2AFX, MRPL37, CKB, PIN1, ENO1, YWHAH, ACTG2, HPRT1, IFRD1, ODC1, HADH, TUBB6, GTSE1, DNAJC9, MTHFD2, SNRPA1, TSEN34, SLC25A5, FH, SMC3, FAF1, COX5A, ABHD14A, HIRIP3, CHCHD3, MTHFD1, LRR1, THOP1, PPIF, LRPAP1, GNL1, CLN6, EZH2, CBR3, PRPS1, FAM50A, ENOPH1, EXOSC9, MEA1, POLR3K, C3orf14, RRP1, ARPC5L, UBE2M, RPA3, BUB3, C9orf142, KPNA2, MPDU1, OXCT1, CTNNAL1, MND1, MAD2L2, ACAT2, COPRS, SMCHD1, MCM3, PGP, IER2, SCCPDH, NDC80, DNAJA1, NTMT1, IMPA2, CALR, CKLF, UBA2, RPF2, FOXM1, CYB5A, NT5DC2, DHRS13, DSN1, ATAD2, EXO1, EIF5, TFDP1, MRPL39, RPN1, HNRNPD, TPX2, CCT2, GAPDH, KEAP1, TMPO, PHF19, PITHD1, PHGDH, HAT1, HNRNPDL, RAN, NDUFS3, TTYH1, CLIC1, SAP30, CCND3, NDUFA9, RFC3, UBE2A, FARSA, NKAIN4, SLC25A1, RRP7A, POP7, PPP1CA, AUP1, TRAP1, AP2S1, C14orf80, RHNO1, CDKN3, MRPL20, NCAPG, ATAD3A, LRRC59, MGME1, DDRGK1, DNAJC1, PDIA4, EXOSC8, GAMT, GLRX2, NIPA2, SNRPG, SVIP, COPS3, BAX, KIF23, UQCRC1, CDC6, GINS1, NSL1, SNRPD3, NUDT5, DPM2, FANCI, CKAP2, CMSS1, CCT4, CDT1, C19orf43, SLC29A1, SLC39A1, PPP5C, PTGES3, SGOL2, MIS18A, WBSCR22, PKMYT1, UQCC2, CHEK1, LSM4, RNPS1, EIF4EBP1, AK2, ADSL, PUSL1, POC1A, CMC2, EIF4A3, CCDC34, CD47, VKORC1, TOPBP1, UBE2I, ARHGDIA, CEP152, WBP11, SRRT, SEC11C, PAFAH1B3, YEATS4, ARHGAP11A, COPS8, CDKN2C, H2AFY, DUSP12, UFD1L, RER1, ABHD12, IER5, WDR61, LSM3, ARF6, AIP, PRMT1, NUF2, PDXK, HEXB, NUCB2, HMBS, PSMD11, AKR7A2, APMAP, SNRNP70, KIAA1524, KLHDC3, TRMT6, HARS, SSRP1, PIGX, CD63, XRCC5, FTSJ2, BLM
Ex vivo IDH1^{mut} GSA scRNA-seq	C4	S100A4, TMSB4X, RPS4X, SLC25A6, TIMP1, RPL10, BEX1, ACTG2, PGK1, VIM, RPL39, PPIB, HSPB1, IGFBP2, PRDX4, LGALS1, BEX4, SSR4, RPL5, HSD17B10, ATRX, NES, GAS5, MAGED2, SNHG5, NDUFA1, WBP5, RPS4Y1, IGFBP5, ETFA, DDIT4, LINC01420, TCEAL4, SERPINF1, MCTS1, LAMTOR5, COX7B, AIFM1, IFITM3, C1orf61, SSR2, BCAP31, SLC25A5, RPL4, HLA-C, PDHA1, LMNA, MORF4L2, TIMM17A, PKM, ATP6AP2, SMS, FLNA, NGFRAP1, TPM1, TRIP6, AP1S2, PSMA5, LDOC1, SNHG3, NPW, ID3, RBM8A, EEF1A1, RHOC, FHL1, SERF2, PNRC1, PARP1, CNN3, RPL36A, RPS17, DPM3, LINC01315, PCSK2, CRYAB, HTATSF1, TIMM17B, HPRT1, IGBP1, AKAP9, OAZ2, CTD-2192J16.15, MYL6, RABAC1, RP11-466H18.1, UTP14A, NKAP, SRP14, PSMD4, UBA1, MT-ATP6, DUSP9, RAB9A, OCIAD2, EEF1A2, RPL10A, CHIC2, CETN2, MAGEH1, TMEM9, EFNA4, DSTN, PDZD11, SLC38A5

Ex vivo $IDH1^{mut}$ GSA scRNA-seq	C5	CALM2, UBE2C, KPNA2, HMGB2, TUBB, CDK1, SNRPB2, NUSAP1, CCNB1, AURKA, ACTB, SEPT7, HSP90AB1, PSMA4, CKS2, HSPD1, BIRC5, PRDX1, HMGB1, TUBB4B, PTGES3, NCL, MORF4L1, HNRNPA2B1, PTMA, SF3B6, RANBP1, RSL24D1, SNRPD2, H3F3A, SPC25, CYCS, NME1, HNRNPK, UBC, C14orf166, HSP90AA1, VDAC1, SSB, HSPA9, ENO1, YWHAQ, PPIA, CCT5, SRP9, PKM, RPL31, TUBA1B, SLC25A5, HSPE1, TTK, RPL21, NPM1, TPI1, DSTN, C1QBP, GAPDH, RPL35
GSA scRNA-seq integration	C0	NME1, MYC, EIF5A, SNRPB, NPW, TUBA1B, IGFBP3, PFN1, C1QBP, H2AFZ, RPL22L1, PFDN2, EBNA1BP2, NPM1, PA2G4, PTMA, PSMA7, RAN, HSPE1, HSPD1, RANBP1, ODC1, NHP2, CCT5, NDUFAB1, DCTPP1, UBE2S, HN1, SNRPD1, VGF, TXNDC17, HMGB1, PRMT1, GTF3A, MRPL36, ATP5G3, EIF4EBP1, HMGN2, POLR2F, ZNF593, SLIRP, PPIA, LDHA, PRDX6, MDH2, DDX21, SNRPB2, ERH, NHP2L1, NDUFS6, CYCS, NOP58, CKS2, ALYREF, PHB, HSP90B1, SPP1, MRPL12, TXN, CD320, TNFRSF12A, RPA3, PNO1, DYNLL1, PPA1, HNRNPAB, ATP5G1, LSM7, DKC1, BRIX1, COTL1, SLC25A5, PPP1R14B, LDHB, GRPEL1, PRDX4, NCL, PDCD5, CHCHD2, GPATCH4, SNRPF, NDUFS5, ACTB, PRELID1, GSTP1, SNRPG, YBX3, NDUFB9, MRPS23, TXNL4A, LYAR, EXOSC4, YBX1, LSM4, SSBP1, MRPL20, TPI1, PAK1IP1, CYC1, MRPL3, PTTG1, SNRPD2, TOMM40, ANP32B, MIF, CCT2, BSG, SET, RAC3, NDUFC2, HMGB3, HSP90AB1, CFL1, STMN1, CCDC85B, DANCR, CMSS1, TRMT112, TMSB10, SDF2L1, UBE2M, MINOS1, GADD45GIP1, TUFM, SRM, EEF1B2, SNRPD3, HINT1, PSMD8, LSM5, NOP10, NUDT1, RUVBL1, PAICS, FAM173A, UQCRH, MYDGF, BOLA3, ATP5J2, NUDC, SNRPE, EIF2S2, FBL, METTL5, MRPS7, FAM60A
GSA scRNA-seq integration	C1	APOD, CST3, SCRG1, PNRC1, C1orf61, FABP7
GSA scRNA-seq integration	C2	MALAT1, NEAT1, KCNQ1OT1, DST, MT-CO3, TRIO, MT-RNR2, GOLGA8A, MT-ATP6, GOLGA8B, MT-CO1, DDX17, REV3L, PTPRZ1, CCDC144B, MT-CYB, NKTR, FUS, MMP16, MT-ND6, MDM4, SNRNP70, MT-ND3, COL20A1, VMP1, GABPB1-AS1, SREK1, N4BP2L2, SOX11, MT-CO2, MT-ND5, POLR2J3, MT-ND4, LUC7L3, MT-ND1, CHRM3, MACF1, CHD4, LRP6, SF1, VPS13C, MT-ND2, ARGLU1, MARCH6, AKAP9, HIPK2, ANKRD11, HNRNPH1, WSB1, PAXBP1, SOX6, DDX5, OS9, MTATP6P1, SACS, PRKDC, CELF1, ZC3H11A, BDP1, CAMSAP2, MT-RNR1, CHD7, ITPR2, KMT2A, SRRM2, PNISR, BPTF, LINC00461, SALL3, SON, TMEM259, CENPF, IGF1R, SLC26A2, MCAM, PCDH9, SPTBN1, ZNF37BP, OGT, KMT2C, COL11A1, VCAN, MLLT4, PHLDB1, MT-ND4L, XYLT1, ASH1L, LINC00969, ZNF638, ATM, NAIP, USP34
GSA scRNA-seq integration	C3	APOE, CHI3L1, APOC1, NEAT1, MALAT1, KCNQ1OT1, DDIT4, DDIT3, CEBPG, GADD45A, PLAT, SQSTM1, GOLGB1, STK17A, JUN, GARS, CAMK2D, BOD1L1, CREB5, AKAP12, EGR1, JAG1, MAP1B, GOLGA4, SLC3A2, FOS, MAFG, CITED1, ZNF704, IER2, LINC00657, HERPUD1, HOMER1, MAGI2, SIPA1L2, LPP, ARF4, SOX11, TSPYL2, CKAP4, CCNL1, RTN4, MAP1LC3B, RLIM, MCL1, ZFP36L1, TARS, TEAD1, REV3L, TPM4, HSPH1, DYNC1LI2, IDS, KPNA4, ANXA5, SARS, SECISBP2L, PSME4, GNG12, ACBD3, EPRS, ATP1A1, TAOK1, APC, N4BP2, CALU, XBP1, ETS1, HIPK2, RPL21P44, UFM1, DHRS2, CDR1-AS, ANKRD12, COPB1, F2R, EIF1, EBLN3, SLC38A1, TRIO, PIK3CA, KDM5B, VMP1, TOPORS, RAI14, HMOX1, AFF4, NIPBL, MEF2A, SETD2, ACTR2, CSNK1A1, COPB2, MSH6, CASK, CDK12, CRK, ANKRD11, AAK1, TRA2B, ZDBF2, NAMPT, NAV1, BTG1, CHD1, DSEL, SMG1, COPA, TNRC6C, TSR1, KRAS, RPS6KB1, DHX33, PNPLA8, AGO2, ZNF131, SERINC1, MAGEA10, BDP1, ANKRD17, DST, REST, MIA3, RUFY3, ZNF326, KMT2A, MTPN, PGRMC2, EIF5, RABEP1, PRKACB, SUV420H1, CEBPB, KMT2C, CTC-444N24.11, SEC31A, DNAJC3, RPAP2, ABHD2, ARHGAP5, TNIK, SETX, ITGB8, FAM199X, KANSL1, ZNF721, MSANTD4, XIAP, JMY, HMGCS1, DLGAP1, SERINC3, PUM1, ZCCHC7, CEP350, UBN2, SOS1, PTGDS, ZNF638, FAT1, SIKE1, KIDINS220, MED19, ARMCX3, WDR82, TMX3, BEX1, LPHN3, POGK, BCL10, ADAM17, YES1, HNRNPH2, ARIH1, SLAIN2, IL6ST, PURB, SEC63, PEAK1, ARHGAP21, ZNF652, TGOLN2, MED10, TMSB4X, MRFAP1L1, MIB1, SPECC1, MAP9, TUBB2A, MED1, PPFIA1, TAF15, LSG1, RB1CC1, NUFIP2, CEP170, KIF1B, SV2A, ARCN1, IARS, SMEK1, EIF4G1, TRIM23, ZNF711, ASH1L, PBRM1, SENP6, SETD5, NRIP1, SLC38A2, FN1, RIF1, BICD1, ZKSCAN1, SYT11, IQGAP1, TXNRD1, VEZF1, RNF115, SBDS, FAM63B, RAB1A, PTPN13, ZFAND5, SEMA6A, FKBP14, POLR3D, CHD2, APOOL, FBXO22, ZNF24, ZNF770
GSA scRNA-seq integration	C4	APOC1, NKAIN4, RPS26, GAPDH, METRN, MIA, RPS13, DBI, APOD, RPL18A, NACA, LRRC75A-AS1, RPL34, FAU, RPL30, S100B, RPL29, C1orf61, SCRG1, RPS24, RPL6, RPL10, RPL39, RPL32, COMMD6, RPL13, RPS4Y1, RPS8, FTH1, RPS12, RPL8, RPS15A, RPS3A, RPS27A, RPL14, RPL11, RABAC1, RPLP1, OLIG1, RPL12, RPS7, RPL28, LGALS1, ALDOA, RPL41.1, RPL24, MT3, RPS19, TCEAL3, RPS5, RPL18, SNHG6, RPS15, C12orf57, BCAN, PPDPF, RPS28, RPL10A, RPL19, RPS27, RPL26, UQCRB, PFDN5, BASP1, GAS5, EEF1D, SERF2, RPL35A, TCEAL4, TMEM258, ITM2B, BCHE, RPS23, TIMP1, ZFAS1, B2M, RPS9, SNHG8, RPS4X, RPL5, TPT1, VEGFA, RPS3, PTN
GSA scRNA-seq integration	C5	ISG15, IFI6, IFITM3, BST2, HLA-B, IFI44L, IFI27, IFIT1, IFIT3, IFIT2, PARP14, MT2A, B2M, STAT1, WARS, HLA-E, HERC5, RNF213, HLA-C, PLSCR1, LY6E, IFITM2, DDX58, RDH10, EIF2AK2, LAP3, SP100, LGALS3BP, IFI35, IFI44, IRF1, HLA-A, SAMD9, IFIH1, PDGFRA, PSMB9, SAMHD1, TAP1, RP11-231C18.1, SOCS1, SOCS2, USP18, PSME1, TMEM158, ARID5B, LGMN, PMP2, CCSAP, ISG20, PARP9, NUPR1, LYPD1, UBE2L6, PSME2, DTX3L, ERAP2, OAS3, ATP1B2, A2M, C19orf66, IFIT5, UGP2, C5orf56, PNPT1, ENPP2, ADAR, NCOA7, STAT2, NUB1, TAPBP, HERC6, VGF, TMSB10, PPM1K, HES4, SOD3, FOLH1, APOL2, PLEKHA4, DNAJA1, HELZ2, TRIM25, SELM, TNC, DRAP1, IL13RA1, APP, CD47, GSTK1, NAMPT, RBCK1, BAALC, OPTN, TRIM56, CD59, DDX60L, PML, MESDC2, TXNIP, TAP2, BRI3, IGFBP5, PPIB, CALR, ZNFX1, RHOBTB3, CHIC2, RAB4A, RNF19A, HSPA5, PCGF5, HM13, ERAP1, CST3, IGFBP2, RABAC1, PDE4B, HERPUD1, DTNBP1, CALD1, SMIM14, SPATS2L, IDH1, SP110, MT3, PSENEN, SCAMP1-AS1, PDIA6, NKAIN4, TOP1, UBALD2, IFNGR2, PTP4A3, CNP, SRGAP2C, RNF114, SPTBN1, MTSS1, C1orf53, IFI27L2, FAM45A, LAPTM4A, GSDMD, C4orf33, ARMCX3, SDF2L1, TRIM69, NUCB2, SCARB2,

		GPBP1, CALCOCO2, UBE2A, IL6ST, GOLGB1, NUCB1, TSPO, FBXL5, PLOD2, TMEM50A, CHMP5, PDIA4, ITM2B, ARMCX1, ARL6IP5, SERF2, SSR4, SMARCA1, PTPRA, TIMP1, CSTB, STAT3, FCGRT, CIR1, C4orf48, HSP90B1, PSAP, P4HB, GOLM1, TMEM219, RRBP1, COX14, GUK1, EIF3A, TMED10, TSC22D4, OST4, PHPT1, SH3BGRL3, SSR3, OGFR, AGTRAP, BTN3A2, RCN1, PDIA3, PGLS, GRN, GNB4, CNPY2, FKBP2, SH3GLB1, RAB3B, CD9, SEC61B, TXN, KDELR1, FTL, TMEM59, FMR1, RAB13, ERLEC1, C12orf57, EIF1, RPL36AL, RPL28, PSRC1, NRN1, ATP6V1F, STAU1, BRD7
GSA scRNA-seq integration	C6	NEAT1, OS9, COL20A1, CDK4, MALAT1, TSFM, GRID2, SNHG9, VPS13C, SNRNP70, METTL1, SREK1, MARCH9, PTPRZ1, KCNQ1OT1, RPS17, WSB1, LUC7L, CCDC144B, BCYRN1, VMP1, CCNL1, TOP2A, TSPAN31, SF1, SPRED1, MAZ, TRIM9, USP34, UBE3A, RPL36, DDX27, LRP6, MAPK8IP3, IGF2BP2, SMCHD1, KRR1, LRPPRC, CLTC, CDC5L, MACF1, COL9A3, ZFAND5, CNTN1, KMT2C, YTHDC1, MT-ATP8, RBBP6, XPO1, CAMSAP2, COPS2, TBL1XR1, ABI2, CHRM3, ST6GALNAC2, DSEL, NCKAP1, ADAR, USP11, TTC37, HUWE1, HDGFRP3, ZNF638, PIK3R1, TNRC6B, SMC1A, DYNLL2, ADNP, SEC31A
GSA scRNA-seq integration	C7	NEUROD1, SOX4, HES6, NHLH1, SSTR2, MAP1B, INSM1, NFIA, SOX11, BASP1, TMSB4X, SCG3, TUBA1A, NFIB, DLL3, DCX, THSD7A, SYT1, GAP43, MLLT11, ID2, CD24, GADD45G, GKAP1, MALAT1, TCF4, KCNQ1OT1, NREP, ZBTB18, LRRN3, MARCKS, CPE, MEX3A, EBF1, FNBP1L, BCL7A, CBFA2T2, H3F3B, BTG1, PTPRS, RND3, IER2, TCAF1, MAP2, PBX1, DDAH2, RTN4, TP53BP1, KLHL24, ARL4C, AFAP1, SPAG9, SHOX2, NRXN1, EYA2, KIF5C, MAP1A, CHGB, TUBB2B, TTC3, KLHL35, GPSM2, MAP4K4, TFDP2, STMN1, CEP170, BEX1, UBE2E3, PNRC1, NEDD4L, DYNC1I2, EZR, PKIA, SSBP2, GLCE, MARCH6, CSNK1E, TUBB, ELAVL2, ATRX, PROX1, MAP1LC3B, TNPO1, ATCAY, DUSP1, CDKN1C, BTG2, NFIX, GLCCI1, PCP4, IRF2BP2, BSDC1, CHMP1B, DYNLT1, FBXO11, TOX3, SEPT3, GRB2, SPIRE1, FAM107B, YTHDF2, GPBP1, GPM6A, ANKRD12, KIAA1598, COPA, HIST1H2AC, ARID4B, PNISR, HIST1H1C, GLRX, GABRB3, JUND, RBFOX2, YWHAZ, BICD1, MAPT, EIF4G2, PRKX, GABARAPL2, BCL11A, RUFY3, SUMO2, RHOBTB3, SESTD1, CSDE1, SOX9, ELAVL3, H3F3A, PLEKHO1, KIDINS220, PAIP2, CDH7, SMIM14, RAD21, TCF12, LBH, TMSB10, USP48, CALM1, GPR56, RB1CC1, INA, KMT2E, YPEL5, RSF1, MIB1, RTF1, NOVA1, SEMA6A, KIF1B, USP22, PHF6, N4BP2L2, ATP6V1G1, DLX6-AS1, LHX1, IDS, GDI1, FZD3, SESN3, CBX1, DHRS2, GSK3B, HN1, RAD23B, KLHDC10, AC004158.3, POLR2K, CCNI, SYP, CREB1, ARID1A, JAKMIP2, RND2, CSNK2A2, BEX2, KIAA0430, MARCKSL1, ARHGAP21, VPS72, MIR99AHG, ANK2, SRP14, MARCH1, ZBTB20, PHF14, SEZ6L2, RHOU, TMSB15A, PIK3R3, CGGBP1, CCNG2, UBE2H, KAT6B, RERE, PHF3, MIAT, RSBN1, CLK1, TSPAN5, PPP1R14C, ATF7IP, CIR1, JARID2, KLF12, GRIA4, LHFP, NDUFS1, DYNC1LI1, TAOK1, SPOCK1, SAT2, TRIM36, DPYSL4, RNF165, RASGEF1B, BAZ2B, SPATS2, STARD4-AS1, CRMP1, USP9X, FLRT3, NIPBL, SEMA6D, CPLX1, ZEB1, HEATR5B, TMOD2, ASNSD1, ZKSCAN1, DNAJB6, PTX3, NRCAM, KIF3C, KPNA6, CACNA2D1, CHN2, UBC, RHOB, CCDC136, RALGAPA2, F2R, SSBP3, HIST1H2BD, SCG2, EPC1, PSD3, GPC2, BNIP3L, KRAS, GNAI3, TGFB3, DCP2, POU2F1, KIAA1107, LRCH1, TTC28, SOGA1, ING4, ZNF766, ACVR2A, MXD4, RP3-449O17.1, SORBS2, SCAMP1, MTURN, BRD9, MLLT4, GPATCH2L, GALNT1, CLASP2, EVL, PCBP1-AS1, DOCK11, USP34, CRK, B3GALT2, PPP1R10, DAZAP2, CELF4, FRYL, ZEB2, TRIP12, CDK5R1, TRIM23, KLHDC2, GNG2, RP3-525N10.2, EYA1, YTHDC1, GPR161, TNRC6C, WDR47, EPHA4, N4BP2, CHD2, RALGDS, ZBTB10, GATS, TSC22D3, RBM41, AAK1, ZNF131, C8orf46, TUBGCP4, ZMYM2, HRK, TM9SF2, SECISBP2, LINC01158, XPR1, YPEL3, BCAS2, RBM15B, THOC1, RP3-368A4.6, RAPGEF5, MUM1, DNAJC12, SH3BP5, EXOC5, ILKAP, MAPKAPK5-AS1, TGOLN2, KIAA0907, PHF21A, ZNF528-AS1, DNAJC18, PLCB1, CCER2, C14orf37, PIAS1, CAPN10, PTPN12, NSRP1, ESCO1, PBRM1, LRRTM2, SLC22A17, CAB39, MECP2, NCOR2, NCAN, USP3, GAA, AFF3, CALCOCO1, TOPORS, FJX1, TDG, ZNF708, HIST3H2A, MGEA5, HPCAL1, SRRM3, EPB41, RBM12B, CLIP3, SHOC2, AC004540.4, MTF2, SORBS1
GSA & GBM scRNA-seq integration	C0	TUBA1B, HMGN2, KCNF1, FABP5, TNFRSF12A, CCND1, DBI, VGF, RPL22L1, COL9A3, TMEM158, S100A6, NME1, LGALS1, NES, CCT5, HMGB3, AGT, SNRPB, UBE2S, SRM, MDK, GPATCH4, ODC1, GAS1, METRN, PSMA7, HMGB1, PPP1R14B, TUBA1A, BIRC5, MINOS1, CD63, C1QBP, GSTP1, PPIA, TXNDC17, RAB13, KCNQ2, PRMT1, NUDT1, TOMM40, TIMM10, POLR2L, LMNA, PHB, CENPF, OCIAD2, PFN1, LSM7, DYNLL1, SNRPD1, RPL35, HNRNPAB, RAN, LSM4, ATP5MF, EIF5A, SRSF7, RANBP1, PTMA, HINT1, NIFK, POLR2F, PSMC5, LDHB, YBX1, S100A16, PSMD8, LDHA, PSMB3, NDUFS6, ANP32A, CCT2, JPT1, CYCS, NDUFC2, NDUFA6, C19orf48, NDUFB2, SLIRP, PRDX2, GAP43, METTL7B, H2AFV, TUBB, S100B, CKB, CCDC85B, PPDPF, GTF3A, RPS26, RUVBL1, F12, NCL, MDH2, NDUFB9, NDUFAF8, CD320, TUBB4B, UQCRFS1, EIF4A1, ADRM1, PSMC4, PLAT, STOML2, RPN2, HNRNPM, ATP5MC3, DCXR, SET, MZT2B, MRPL3, CITED1, MRTO4, PSMB1, PARK7, PSMB6, RPL8, TAF10, PSMA4, COX8A, EIF3G, UBE2M, ILF2, EBNA1BP2, PAK1IP1, EIF3K, HSD17B10, ROMO1, POLR3K, PA2G4, POLR2E, SNRPG, UFD1, MRPL12, PSMA2, HSP90AB1, TIMM13, NDUFA4, CCT3, MRPL4, SNF8, BUD23, RFXANK, ZYX, AHCY, ATP5F1B, PIN1, MRPL52, ATP5MG, RPL23A, PSMB2, HSPE1, LAPTM4B
GSA & GBM scRNA-seq integration	C1	ISG15, HLA-E, GBP1, PCDH9, IFI6
GSA & GBM scRNA-seq integration	C2	DLL3, AC009041.2, SOX8, OLIG1, NKAIN4, CDKN1C, OLIG2, ETV1, GRIA2, SOX4, CADM2, GADD45G, SHD, SOX6, C1QL1, CCND1, TCF12, VCAN, HES6, NNAT, ATCAY, MAP2, SMOC1, COL20A1, GLCCI1, FERMT1, CHD7, NEU4, KHDRBS3, HIP1, PLPPR1, DSEL, MARCKSL1, EPN2, LIMA1, GPM6A, ARL4A, ASCL1, ZNF462, NXPH1, FXYD6, SCD5, FIBIN, CCND2, CNTN1, TNK2, TMEFF2, RBPJ, KCNQ1OT1, PDE4B, POLR2F, PHLDA1, MARCKS, SIRT2, NRCAM, PLLP, ALCAM, BEX1, ASIC1, LSAMP, KCND2, SCRG1, MIDN, CD82, MAML2, CTTNBP2, ADGRL3, MTSS1, ACAP3, FAM110B, NCAM1, ZEB2, NRXN1, SNX22, BCAN, SOX11, GPR17, TPP2, STMN1, NUPL2, TCF4, PCDH17, C11orf96, TMEM121, KIZ, CENPV, ARPP21, PTPRZ1, UGT8, JPT1, ASIC4, MLLT11, PTMA, TNR, ABHD2, TRAF4, LHFPL3, TFDP2, ANGPTL2, LRRN1, RBP1, MYT1, THY1, NOVA1, MYO10, MDFI, H3F3A, SLC2A13, SRGAP1, LINGO1, ELMO1, MAP1A, P2RX7, BTG2, RAP2A, CSPG4, PDGFRA, SEMA5A, TIMM50, METRN, HES5, DNER, SNTG1, TNS3, LRRC4C, LAPTM4B, SH3D19, ODC1, PODXL2, S100B, KLHL7, PIK3R1, GTF2I, C3orf70, ATF7IP, DCX, APC2, FAM181B, ST3GAL5, EPHB1, DNM3, FOS, PPP1R14B, NKX2-2, KCNQ2, ARC, HIPK2, MMP16, RFTN2,

		PMP2, ZKSCAN1, KIF2A, PHYHIPL, SMARCC1, CHST11, KIF13A, RASSF2, FAM107B, LDHB, BCL7A, LDLRAD3, RAB33A, ZNF649, CADM4, TUBA1A, REPIN1, SCG3, NCALD, BCHE, ARHGEF7, EIF4G3, ZFYVE16, TUBB4A, CRB1, SOGA1, RTKN, SLC22A17, DUSP6, AMOTL2, ERF, MEX3A, ZEB1, WNK3, UHRF1, RCN2, PCDH15, SHC3, GSK3B, ZCCHC24, HNRNPA1, PGRMC1, NOTCH1, CRMP1, CXADR, REC8, ABAT, DST, OPCML, CCDC88A, ID2, CSPG5, SERPINE2, PRDX2, NME1, SFPQ, DPYSL3, NCAM2, CASK, BAALC, CLASP2, CELF2, PHF14, DSCAM, ZNF326, SMARCD1, ZNF708, RPAIN, MALAT1, KMT2E, ANKRD10, PID1, JMJD1C, ZNF322, RSF1, MXD4, KDM1A, RBMX, NIN, KIF3A, ZNF431, MKLN1, TCAF1, C2orf80, APEX1, NFIX, MAP4K4, DDX5, TTC3, ZBTB20, EFS, WSCD1, REV3L, NLGN1, TOP2B, MYCBP2, ANP32A, RGCC, CBX5, GNAI1, RCOR2, SAPCD2, RTN3, MEGF11, DOCK10, LRP3, VXN, MAGI1, ELAVL3, UCHL1, PFN2, DBN1, SHISA4, SET, PRKDC, GPC2, GNG4, MUM1, STMN4, RIC3, SGTA, SNRPE, KANK1, SATB1, GSTA4, IGF2BP3, HNRNPAB, ZNF428, MAP4K5, BZW2, ANTXR1, MAD2L2, SCN3A, SCG5, FAM222A, KLHL24, CAMSAP2, TCEAL2, PRDX1, SETD5, GDAP1, GNG2, SLC44A1, TXNIP, CHD3, KLF13, TSC22D1, RBM25, ATP1B3, HOXA7, MRPS7, SULF2, HNRNPA3, PTP4A3, RICTOR, SERINC5, DHX36, PPP1R12A, HNRNPM, MYO5A, CELF5, PRPF40A, TAOK3, NASP, BAG1, SKIL, MAPK10, ETS1, VEZF1, ATRX, FYN, ILF3, UBB, SRSF3, ARHGEF2, NAV1, MAGEH1, CSNK1E, ZMAT3, TMEM100, MT-CO2, H2AFY, TMEM206, HNRNPD, ZNF738, CKB, SMARCA4, RSPRY1, KIF21A, AKAP9, ARHGAP35, CMTM5, SIK3, BPTF, MFF, LRP6, KDM4B, HNRNPA0, MNAT1, TSHZ1, PBRM1, FGF14, ATXN7L3B, SIX1, TMCC1, BRINP3, GRIK3, ARID1A, PCMTD2, YBX1, BASP1, HRASLS, BRD8, CEP170, CCSER2, CHD4, BEST3, OLFM2, PLCB1, PIK3R3, AL391807.1, CACNG4, NCBP2, ARID4B, YTHDC1, BAX, NTM, SALL3, NSD3, KIAA0232, EPB41L2, MAPT, RERE, BTG1, USP24, PCBP4, TCF7L2, GDAP1L1, KLHL23, RNF130, CASTOR3, HNRNPDL, DLL1, POU3F3, GOLM1, NPPA, PPP2R2A, SYBU, UQCRB, MAP3K1, TSPO, ADGRG1, SOX9, KIAA1958, CHD6, BTF3L4
GSA & GBM scRNA-seq integration	C3	RPS18, ZFAS1, RPL41, RPL32, RPL34, RPL12, RPL11, RPS28, RPS19, RPLP1, RPL9, RPL3, TPT1, RPS14, RPL36, RPL30, NMB, RPS27, TOMM7, RPS4X, RPS12, RPL29, RPL37A, RPS27A, RPL39, RPL28, RPL26, RPS29, RPL27A, RPLP0, RPS8, RPS2, RPL18A, RPS15, RPL22, RPS15A, RPL13, RPL8, RPS25, FAU, RPS3, RPS16, RPS24, RPS21, RPL19, RPL10, RPS9, RPS6, RPS5, RPL35, RPS23, RPS3A, RPL21, RPL7A, LGALS1, RPL5, RPL35A, RPL13A, RACK1, COMMD6, RPL38, TMSB10, RPS10, RPL37, NACA, RPL6, PFDN5, RPL14, RPS13, RPL7, FTL, RPL18, RPL10A, UBA52, RPL36A, RPS11, RPL24, PHPT1, MZT2B, RPLP2, FTH1, COX7C, HIGD2A, GNG5, RPL23A, RPL31, EEF1A1, EEF1B2, PPDPF, NOP53, S100A11, UBL5, TPI1, VIM, SNRPD2, EPB41L4A-AS1, DDT, NENF, RPL15, COX7A2L, C19orf53, RPL17, RPSA, RPS7, MIF, C6orf48, GAPDH, TRMT112, TAGLN, EIF4B, PRDX6, RPL27
GSA & GBM scRNA-seq integration	C4	VEGFA, NDRG1, IGFBP5, IGFBP3, AKAP12, MT1X, HILPDA, LGALS3, IGFBP2, MT2A, NRN1, OLFM1, PGK1, VIM, SPP1, SCG2, YBX3, PLOD2, EIF1, NTRK2, CEBPD, GAPDH, HSPA5, TMSB10, CHI3L1, S100A11, SLC2A1, LMAN1, DDIT3, BNIP3L, BNIP3, ZFAS1, CEBPB, NAMPT, HSPA1A, BTG1, SERPINE1, HSP90B1, P4HA1, CAV1, ENO2, MCUB, IER5L, SSR3, ERO1A, P4HB, TPI1, CALR, EPAS1, TRIB3, CA12, XBP1, DNAJB9, SELENOS, SQSTM1, SEC61G, S100A10, SLC2A3, PDPN, CAMK2N1, LDHA, ASNS, EPB41L4A-AS1, SLC3A2, CAST, TPT1, CHPF, CRYAB, AK4, MAP1LC3B, GRB10, ENO1, RASSF8, ATF4, RACK1, NOP53, PFKP, FN1, DNAJB1, GARS, JAG1, ACTG1, MYDGF, GPI, MYO9B, EIF4EBP1, SARS, CANX, SCD, HSPB1, DDIT4, SLC16A1, NUPR1, SLC6A6, CXCL8, ZNF395, BHLHE40, RPL10, EMP1, CLIC1, HERPUD1, RPL9, UFM1, PYGL, FTL, EPRS, CYTOR, SERP1, FLNA, MEG3, WSB1, RPL34, PDIA6, CNN3, ABCA1, SELENOK, FAM162A, RDH10, ERRFI1, VOPP1, EIF2S2, SOD2, LIMS1, ANXA2, PGM2L1, TPM4, TIMP1, IQGAP1, HSPA9, HSPA1B, SPOCD1, SMIM3, CDK2AP2, MIR4435-2HG, SVIP, RPL26, SERTAD1, OAZ1, YWHAH, AL590617.2, ITGB1, VKORC1, ATP13A3, SUCO, TAGLN, DNAJC3, SLC38A1, MT3, TNFRSF12A, MAP1B, SYTL2, SLC7A5, BLVRB, CALU, CEBPG, TMED2, PTX3, TCIM, ADM, PGM3, TENT5A, KDSR, INSIG2, WWTR1, DOK5, SUGT1, SEC31A, PLP2, HMOX1, ARF4, ACTN1, EIF3E, VGF, TRAM1, BRI3, FAM20C, RPS25, PADI2, TOMM20, TCEAL9, ARRDC3, ARL4C, PDIA3, SH3GLB1, RPS27, EIF4B, TAGLN2, RPL3, ARID5B, NUCB2, FNBP1, CAVIN1, ZFAND2A, HSPA13, TCEA1, ANXA1, TMF1, GAP43, BET1, HDLBP, TLN1, TMEM45A, MANF, FAT1, KDELR2, PJA2, RSRP1, CALD1, PTPRF, RPS18, PPIB, MEF2A, OXR1, UAP1, ATF5, IDS, METTL26, RPS13, ARL1, CFAP36, PPP1R15A, NACA, CIB1, CD9, FTH1, RORA, WDR45B, RPL5, OBSL1, RPS8, ANKRD12, SHMT2, ALDOA, IFRD1, RBCK1, PAM, SERPINH1, SDC4, RNH1, PMEPA1, HSPH1, NUMA1, DDX3Y, LONP1, ACBD3, RPL14, SPCS2, EBLN3P, NFKBIA, NOL3, STK4, HSP90AA1, ATF3, UGCG, DAPK3, FOSL2, GOLT1B, RPS14, GBE1, PNRC1, CYR61, INAFM1, PDIA4, RRBP1, NARS, MORF4L2, TUBA1C, TGFB1I1, SLC25A37, ASPH, SLC25A36, SARAF, SIVA1, GADD45B, RPL11, RND3, CDV3, XPOT, UFL1, CCDC107, TAF1D, TMEM70, TAX1BP1, AL133453.1, EIF1B, RPL35A, PLOD1, GLUL, TNIP1, UPP1, RPL12, DHRS3, RSL1D1, RPL17, CCNI, CD44, BACE2, PHGDH, RPL7, FGFR1, MYADM, DSTN, H2AFZ, SNHG7, EIF3D, EEF2, ELOC, YIPF2, SRSF5, SNHG8, NFIL3, RPS28, WEE1, COPB1, ANKRD28, PCOLCE2, ARF1, RPL32, SEL1L, PKD1, SERINC1, SLC5A3, CARS, AC093673.1, FNDC3B, DHPS, FAM114A1, DYNLT3, RNMT, SBDS, RPS3, FXR1, COX7A2L, MTHFD2, TPM3, BACH1, MKNK2, SAT1, IL1RAP, TMED9, RRAS, MAP2K1, MYO1E, MXRA7, ARL6IP4, FAU, MIF, FERMT2, PRPF6, PRDX4, ELL2, OSTC, AC027031.2
GSA & GBM scRNA-seq integration	C5	CHI3L1, ID3, AQP4, HOPX, IGFBP7, CLU, GFAP, CST3, APOE, ID1, CRYAB, NMB, SLC1A3, FABP7, LPL, NCAN, ATP1B2, MT2A, CXCL14, MT3, LGALS3, TIMP1, RCAN1, SPP1, RAMP1, TRIM47, GATM, A2M, VIM, CNN3, PTN, SPARC, ANXA2, HTRA1, CITED1, NAMPT, CSRP2, SPARCL1, FJX1, PLTP, F3, ZFP36L2, SLC1A2, TTYH1, METTL7B, CD44, GLUL, C1orf61, TNC, PMP22, IFITM3, LMO2, MT1X, EMP3, CD99, AGT, TUBB2A, EMP1, CEBPD, LRIG1, C1R, TTYH3, ID4, UGP2, PLA2G16, CADM1, TSC22D4, GADD45B, ITM2C, FERMT2, PDLIM4, GPC1, RGMA, GAP43, S100A11, DCLK1, RHOB, HEY1, TAGLN, CAMK2N1, CLIC4, ITPR2, SERPINB6, LYPD1, B2M, ANXA1, S100A10, WLS, CD63, ARHGEF26, TMSB4X, IRS2, SLC4A4, ZFP36, EDNRB, PODXL, FLNA, SDC3, SFXN5, NEK6, TAGLN2, IFI27L2, TIMP3, SPON1, EFEMP2, NUDT4, JUNB, HSPB1, RHOC, CTSA, NACC2, DTNA, PROS1, GADD45A, MT1E, PLIN3, PLPP3, DCLK2, TMEM132A, TPM2, FAM107A, TIMP2, GPM6B, LHFPL6, MLC1, KCNF1, PPP1CB, RASD1, HLA-C, PON2, LIFR, SEMA6A, MYL6, DPY19L1, PDPN, PHLDA3, MAN1C1, LGALS3BP, STK17A, HEPN1, NME3, SEMA6D, APC, HES1, BCAN, TSTD1, TNFRSF12A, BST2, CD81, S100A16, ITGA6, S100A13, SOCS3, ECI2, COMT, CALM2, NLRP1, RAB31, MT-ND3, ETFB, GNG5, TNFRSF1A, PLEC, PTTG1IP, SRPX, HIF1A, RARRES3, NPDC1, MAP1B, PCDH9, CTSZ, CYR61, HLA-E, ATP1A1, DPP7, SCG2, FSTL1, ITM2B, PSRC1, TRIP6, OCIAD2, SEPT7, FAM181A, IFI16, SLC3A2, FABP5, CD59, NRP2, NEDD9, SPOCD1, SELENON, HEPACAM, ADAM9, RFX4, AP1S2, SYPL1, HS6ST1, SNX3, HRH1, CSRP1, CALM1, RDX, CHPT1, TIPARP, CBR1, ACSL3, GJA1, LGALS1, EVA1C, LAMP1, SSFA2, CDH4, GPR37L1, PRDX6, LRRN3, CHL1, SDCBP, CDC42EP4, TNIK, SORT1, EFHD2, SAT1, TMEM205, SYMPK, CAMK2D, ANXA5, BBOX1, APLP2, GALK1, ITGA7, NFIA, MOXD1, MAOB, CA2, SIRPA, MT1M, NAT8L, TSPAN3, SMOX, SCARA3, CTSB, DKK3,

		IFITM2, KAZN, SYNM, DPYSL2, ERBIN, DAG1, IQGAP1, S100A6, PEPD, PRR7, BAALC, ZNF436, CYSTM1, TUBB6, CDH2, CD151, GAS7, SEC14L2, SPTAN1, BLVRB, LAMB2, GLUD1, SIPA1L1, INAFM1, MRC2, NEAT1, GRN, DOK5, FEZ2, KDELR1, CERS1, CHCHD10, SNTA1, PRRX1, SEPT11, TMEM132B, ELOVL2, DNPH1, HLA-B, FAM69C, AIF1L, SMAD1, CFI, HSDL2, NPAS3, LMNA, NPC2, SPATS2L, METTL7A, EZR, PDLIM7, CLIC1, LTBP3, PLAU, DNASE2, IFT22, TSPAN5, ATP6AP2, TGFB2, DLC1, CPNE5, RYR3, ACTN1, PTPRA, TUBA1C, BCAP29, FAM20C, GSTK1, SCRN1, ADGRG1, VAMP5, PDLIM2, IFI6, ATP1A2, IL6ST, CARHSP1, TUBB2B, LRP10, TPST1, PLSCR1, PCSK1N, BICD1, ROM1, COL4A1, PKM, APC2, PLP2, LAPTM4A, RFTN1, DDAH1, FCGRT, ABCD3, CTSF, ACAA1, ALDH6A1, PTGR1, TMEM59L, DDRGK1, RAB34, MT-CYB, PRAF2, B4GAT1, C19orf70, TMEM147, CAMTA1, GLG1, PLA2G5, C1orf122, ENDOD1, OXTR, SHISA5, ENAH, LAMP2, FOXO1, SELENOW, DOCK7, FADS3, PRSS23, SERTAD1, C19orf53, TAPBP, SOD2, PSAP, BATF3, SBDS, SORBS1, PPIC, ERP29, DDR1, GCSH, ADD1, MAGED1, FKBP2, TMEM179B, CIB1, SRI, AEBP1, SPTBN1, PEA15, POU3F2, IQGAP2, GLIS3, SEC14L1, TMBIM6, RRBP1, SMIM29, SLC25A23, CRISPLD1, PSENEN, GALNT2, LAMA4, DPP6, RETREG1, ACTB, GSTT2B, NFKBIA, PPP2CB, CETN2, NAA38, CROT, CTNND2, MYL12A, FZD7, DBI, PARP9, PPT1, PGLS, POR, ACTG1, TXNL4B, DPF3, PAXX, GNAI2, NUCB1, ALKBH7, COL4A2, AHCYL1, CTSL, CSTB, PFN1, TMEM9B, GADD45GIP1, ANK2, NMT1, RHBDD2, MIR4435-2HG, GNG12, WWTR1, ASAH1, FAM3C, SPIRE1, PFKFB3, ZHX3, SSBP4, SLC25A18, UQCR11, ARL6IP5, ACTN4, LZTS1, ARL6IP1, PBXIP1, CCDC106, KTN1, SORL1, HES4, REX1BD, FGFR3, TGFB1, CPE, ERLEC1, CCL2, MT-CO3, KCTD12
GSA & GBM scRNA-seq integration	C6	AGT, SLC1A3, SOX9, SDC3, ITGB8, MALAT1, ATP1B2, GFAP, PTPRZ1, TNC, NEAT1, SLC1A2, LFNG, DCLK2, PCDH9, CLU, EDNRB, PIK3R1, BCAN, SEMA5A, AQP4, ACSL3, NPAS3, RGMA, LUZP2, LPL, CHL1, ATP1A2, CST3, GATM, EGFR, GPR37L1, MACF1, GAS1, DPP6, TRIB2, WLS, NTM, TRIM24, TANC2, CDK6, PCDHGC3, LMO2, NFIA, RRBP1, PLPP3, LRIG1, CSPG5, BICD1, ZNF708, FOSB, POLR2J3.1, TTYH3, ARHGEF26, PREX1, HOPX, KCNQ1OT1, SEMA6A, SRGAP2, EPN2, ARGLU1, ZBTB20, HIP1, ADGRL3, MSI2, MYO10, POU3F2, APC2, PMP2, DPY19L1, SPARC, LIFR, COL9A3, PTPRA, ILDR2, EGR1, LRP1, NCAN, DDR1, KMT2C, DTNA, TRIM9, FAM181B, QKI, IRS2, RASSF2, COL4A2, MRC2, B4GALT5, VCAN, MLC1, F3, IL6ST, SPON1, KCNF1, NORAD, NOTCH1, FJX1, KIF1B, ITGA7, TMEM132B, ROBO2, COL4A1, GLG1, SESN3, DAG1, SPATA6, CDH2, SPECC1, RAB31, CDH4, METTL7A, RCAN1, CRISPLD1, CAMK2D, GNA12, GTF2I, APOE, MAP3K1, ZFP36L2, ZNF254, UBL3, RFX4, PRRC2B, VMP1, POU3F3, SPTBN1, FOXG1, TRPS1, ENC1, GOLGB1, ZFP36L1, PTAR1, RAB3IP, DCLK1, MEST, ITPR2, GNB4, SRGAP2C, LYPD1, PLEKHA4, NLGN4X, SORT1, SMC5, ERBIN, SPRED1, MATN2, PODXL, ZHX3, CPEB4, IDH1, FZD3, ADGRG1, ADD1, ADAM9, ZNF91, ACTN4, IFI16, SCRN1, GRIK3, NUMBL, NFIB, RIC3, NAV1, SIPA1L1, NKTR, ATRX, SPRY4, BAZ2B, PAG1, TNIK, RNF180, EIF4G3, LRRC17, WDR60, BPTF, TUT4, NES, EXTL3, ACSS3, PAXBP1, SFPQ, CCDC88A, MT-ND4L, CCNL2, STIM2, SEPT7, RB1, BMP7, WSCD1, CUX1, C1R, NEDD9, MAGED1, PTN, ARHGAP21, ELOVL2, TNPO1, SUGP2, ANK2, CHD1, REST, GPATCH2L, FAM69C, TNRC6B, STAT3, PCM1, TMEM131, SORBS1, NSD1, TRIL, FAT1, PTGFRN, ARNT2, PRPF4B, KCND3, CADM4, PMP22, SETD2, GRIA2, PIK3C2A, PDE4B, CCND2, SSFA2, CADM1, CREB5, RHOJ, CASK, TSPAN3, HEPN1, MMP2, SPAG9, CHD4, RYR3, BCAP29, ZNF106, CHD9, APP, NTRK3, PDGFRA, CELSR2, ITGAV, TRIM47, ARAP2, SEMA6D, PLAT, ETV1, RBM26, METTL7B, ZFHX4, UHRF1, LDLRAD3, ADAR, ANKRD36C, APC, MAML2, ADAM23, ZNF43, CLIP2, NPTXR, CDCA7L, ANKRD17, PGAP1, KLF9, TJP1, C3orf70, IPO9, MCL1, VEZT, SACS, DDX17, LRRC58, TFRC, ARHGEF6, ABHD2, HTRA1, CEBPD, TCF12, PCDH10, SRGAP2B, DST, PRRX1, PON2, SPEN, USP8, FOS, FABP7, CARMIL1, PJA2, SSR1, GPRC5B, SULF2, ANKRD10, BMPR2, LRRTM3, MT-ND5, ADAMTS6, PPP1CB, ZEB1, NCOA1, ATM, ABCD3, CANX, PUM1, CREBZF, ACER3, SKI, COL6A1, UBE2G2, MYEF2, CORO2B, ERF, NCAM1, NIN, SLC6A9, APLP2, ZNF638, FMNL2, LINC00461, RSRP1, SERPINB6, ZNF493, ZNF827, CPXM1, CASC4, ARMCX3, KMT2B, HEPACAM, FAM20C, ANKIB1, NRDC, MKLN1, TAOK1, MUM1, MYO6, HEY1, ADGRB2, LRP1B, JUN, TAOK3, CYFIP1, LPP, MYH10, ABCA1, LHFPL3, ZKSCAN1, SCD5, KDM2A, SCARB2, RDX, KAT2B, PHF3, CLTC, SETD5, GRAMD2B, TCF4, PRRC2C, ARHGEF40, STK17A, FBXL5, AP3D1, QSER1, OGA, KIF21A, CAPN5, DOCK7, NOTCH3, TRIM73, C6orf62, ARID2, ZNF148, INTU, PHACTR4, FAM208A, ZNF260, REV3L, PLCE1, SPTAN1, TLK1, FTX, NCOR2, ATXN2, SOGA1, ASTN1, PURB, ACIN1, BCLAF1, LRRC4B, GLDC, CTNND2, DNAJC10, CALD1, USP34, MAP2, ETV5, PNN, ZNF431, MT-ATP8, RPGR, SREK1, ABI2, PARP14, ATP1A1, SWAP70, CLSTN1, TNKS, GNG7, NOVA1, PSRC1, PDGFA, PRKCA, SOX5, TMEM30A, ARHGAP35, PXDN, CHD7, SMARCD3, SPRY2, ARHGAP5, RFX3, FRYL, IGF2BP3, GOLGA4, LENG8, TP53
GSA & GBM scRNA-seq integration	C7	SOX11, NNAT, DCX, SOX4, CD24, STMN1, RND3, THSD7A, BASP1, INSM1, GADD45G, STMN4, NRXN1, NFIB, ELAVL3, FNBP1L, BTG1, AUTS2, TCF4, ARL4D, MAP1B, RBFOX2, HES6, CRMP1, RBP1, TAGLN3, NFIA, KIF5C, PAK3, C4orf48, CDK5R1, BEX1, JPT1, PKIA, CBX1, MLLT11, BCL7A, NREP, NFIX, PCSK1N, MARCKSL1, RPAIN, TTC3, ZC2HC1A, UCHL1, MAP2, SBK1, SCG3, GPM6A, MAPT, TSPAN13, KLHL7, AKAP9, ATP9A, PLK2, NEUROD1, CPE, BEX2, ZNF292, MEX3A, CEP170, TUBB, TMSB15A, CELF5, SRGAP3, KIF21A, EIF1B, MEG3, SLAIN1, PAFAH1B3, GPD1, KDM1A, PBX1, GSTA4, KIDINS220, APLP1, HIST1H4C, H3F3A, DBN1, GPC2, ARL4C, TMEM161B-AS1, ATCAY, THRA, CAMK2N1, MAP1LC3A, MEIS2, RERE, TUBA1A, TTC9B, ATRX, TOP2B, OLA1, TERF2IP, DLX6-AS1, DPF1, KHDRBS1, VAMP2, ZBTB18, UBE2E3, PROX1, GTF2I, KLC1, BLCAP, DLGAP4, MIDN, MAP4K4, NEDD4L, FOXG1, YWHAG, MYCBP2, ID2, TCEAL2, RUFY3, ENC1, SMARCD1, TCAF1, DDAH2, SOBP, PBRM1, TXNIP, DRAXIN, CBFA2T2, CHGB, APC, ZNF704, ATP6V1G1, DPYSL2, KDM6B, MARCKS, ZNF91, CCDC112, PGAP1, IGSF21, SSBP3, RPL7L1, GDAP1L1, SLC38A1, SNN, LINC00461, USP22, ARID4B, PTMS, LBH, C3orf14, PRKAR2B, SEPT3, TOX3, RSBN1L, MAPK10, GNG2, DDX24, EVL, KLHL23, DAAM1, SRPK2, CYTH2, CNIH2, ARID4A, CNOT2, PLCB1, EIF4G2, HMGB3, LANCL2, ZC3H13, REEP1, SRGAP1, MYT1, SEZ6L2, ZEB2, AKT3, TUBB2B, ZNF428, SPAST, SCN3A, TFDP2, CITED2, SMC3, MEX3B, SCX, RBX1, KLHDC8A, SEPT11, CCNG2, SNTG1, WDR82, NCAM1, BLOC1S6, CHD3, RB1CC1, DYNLL2, ASRGL1, DPYSL3, MALAT1, TSC22D1, PHF6, CENPV, LRP8, TXN, CACNA2D1, NUDT3, GNAI1, AES, DTD1, HDGFL3, BCAS2, STXBP1, RTF1, BAZ2B, STRBP, TIA1, RNF165, HNRNPH2, CLASP1, SPIN1, TAF11, AFDN, KIF3A, TCEAL5, ANKRD46, VPS28, MAP6, SUMO2, DHX29, CD200, SHOX2, RTN2, UBE2S, PALM, CELF1, GSE1, SRRM3, ASB8, SPOCK1, TCEAL7, MBIP, AHDC1, PHF14, RCN2, LINC00662, RTN3, SNRPN, YPEL5, TUBB4A, MORF4L1, MORF4L2, TMX4, FOXN3, HPCAL1, DDX5, C14orf132, DPYSL5, KIFAP3, BRD3, ARX, KMT2E, DDX6, TNRC6C, MRPL21, CERS6, FAM89B, PSMB7, BPTF, RSBN1, NSFL1C, ELMO1, TNRC6B, CAMK2N2, TRIM36, ATP1B3, PIK3R3, CXADR, MKRN1, NCAN, MARCH6, ATP1A3, NDUFA8, ENAH, CHD6, DYNC1I2, SESTD1, RNF24, GNAQ, CLASP2, TRAPPC4, COX7A2, ASIC4, SATB1, NUAK1, PNPLA8, CALM1, BEX4, CCDC167, NRP1, BZW2, IFRD1, SPTAN1, ANK3, RCOR2, GOLM1, ZNF793, ATF7IP, CDO1, ACTR10, LCORL, MRPS36, DHPS, FEM1A, PCDHB9, PAFAH1B1, KAT6B, PODXL2, HIPK2, EPB41, DHX36, HSP90AA1, MAGED2, TOMM20, MPHOSPH8, KRAS, SPOP, CCSAP, RRM1, SPATS2, RBBP4, MRPL44, SYF2, ARL6IP6, SCG5, ZMAT2, PFN2, PTMA, BEND5, PPP2R5E, HNRNPAB, FAM49B, CBX3, TXNRD1, AKAP6, CAMLG, CDKN2D, MTSS1, CHD7, ABAT,

		GNB1, SHF, ARF4, FZD3, GABARAPL2, CPEB4, SLC25A29, WDR47, MIEN1, BPGM, SMARCA4, CCDC28B, SNRPB2, GNG4, CDC34, DHX9, GKAP1, ATXN7L3B, POGZ, ARHGAP33, NLRP1, RAD21, DYNC1LI1, ELAVL4, PRKX, NAGK, HOXA7, USP11, MT-ND1, CCNI, SRRM1, ATAT1, ZNF711, CEMIP2, CHD9, ATP6V0E2, SYT11, NXPH1, CXXC5, CSNK1E, NIPBL, CWC27, DDX17, CNR1, UBE2I, ROBO2, AMER2, BRSK2, FAXC, NSD3, GRIN2B, DCTN3, DNAJA4, PPHLN1, KLHDC2, GABPB1-AS1, NOL4, PTP4A1, UBQLN1, FSD1, HIST1H1E, BRWD1, OCIAD1, FBXO11, KIF2A, CHMP5, MZT1, KLHL42, RASA1, BTF3L4, H1FX, RALGDS, C12orf65, NIPSNAP1, HSDL1, CCDC88A, SRSF10, CNKSR2, OSBPL8, RAB11B, PHF20, ZBTB41, SETBP1, YWHAQ, DCTN2, CNOT7, ARL8A, ZNF3, C12orf73, MXD4, MAP4, TTC28, VEZF1, DNAJC7, PJA1, VAT1, CUL1, PPM1L, UBE2V2, CALM3, VDAC3, KMT5B, ATL1, EMSY, SHD, ACTR6, PPP2CA, XRCC5, RAB6A, ARHGEF9, RPRD1A, MAPRE3, TCP1, TNRC6A, ENO2, G3BP2, MTURN, RTN4, EBAG9, ZNF281, PTBP2, TOX, YTHDF2, PDS5B, RAB2A, IK, AFAP1, TMSB10, RAB14, DNAJB6, POU2F2, RSF1, DPYSL4, ARPP21, SH3BP5, NR2F1, IRF2BP2, PRRC2B, RUNDC3A, ANKMY2, GRIA4, TRIM13, SOGA1, ORC4, GON4L, ZNF536, PINK1, HDAC9, FOXN2, PSMD7, PAIP1, BAG6, HEXDC, GALNT1, OSGEP, TBC1D7, ZBTB20, TIPRL, USP16, RBM25, WIPI2, NPPA, SEZ6L, AK1, SUB1, CTNNBIP1, HTATSF1, SNX4, PRDX2, MOSPD3, CDKN1B, KHDRBS3, SEC11C, CBLB, ZNF667-AS1, AMN1, RAB5C, PBX3, EPHB2, ZBTB33, LRRC4, DEK, ROBO3, TERF1, RALA, RIOK3, REC8, ARID1A, NCBP2, LRRC47, DUSP12, MACO1, AKAP17A, ZSCAN18, ZC3H6, KIAA0232, POLR2B, SVBP, APBA2, PLPPR2, GATAD2B, AASDHPPT, STRAP, ATP6V1H, MARCH1, AP2M1, GLCE, KIAA1549, LSAMP, SYNE2, CLIP3, B4GALNT1, NAP1L4, KMT2A, CNTFR, APC2, UPF3A, POLB, PABPN1, TLE2, ANAPC15
GSA & GBM scRNA-seq integration	C8	PTPRZ1, EPN2, KCNQ1OT1, KCND2, C1QL1, SERPINE2, BCAN, PMP2, SEMA5A, ARL4A, S100B, SCRG1, NKAIN4, CYB5D2, OLIG1, LIMA1, CADM4, LUZP2, GRIA2, RAB31, LHFPL3, FAM107B, CADM2, DSEL, GRID2, VCAN, PDE4B, SOX6, PCDH9, KHDRBS3, NOVA1, SPATA6, MYO10, CNTN1, COL20A1, TSPAN7, ABAT, TRIM9, PDGFRA, MALAT1, METRN, BAALC, NLGN4X, PGRMC1, ABHD2, CSPG5, ATP1B2, TRIM24, MARCKS, PCDH10, SCD5, CTTNBP2, MAML2, LRP1, MT-ATP8, REV3L, SLC35F1, SHC3, PPAP2B, ZNF462, MAP3K1, KANK1, BCAP29, MTSS1, PCSK1N, TANC2, SESN3, TAOK3, ETV1, TCAF1, EIF4G3, CASK, RNF180, FXYD6, MMP16, LINC00511, SEPT7, APC, ATF7IP, PIK3R1, SCP2, C11orf96, APOD, NIPBL, MAP2, LSAMP, GALNT1, FAM181B, SULF2, ITM2C, FOS, PCMTD2, UQCR11.1, EDIL3, ARL2BP, ITGB8, LRRC4C, SERINC5, LINGO1, LYPD1, C1orf21, RAB8B, TCF12, PAK2, ZEB1, GLCCI1, CCND1, ATCAY, PHLDB1, LRP6, NCAM1, FAM110B, MT-ATP6, FIBIN, CBX5, TIMP2, DCX, PRKDC, GPSM2, OLIG2, LINC00461, GTF2I, ATP6V0E2, APBB2, NRCAM, HNRNPA3, TTYH1, KCNQ2, RHOQ, RP11-146D12.2, BRINP3, CELF2, GPM6A, RASSF2, PPFIBP1, NRXN1, GNAI2, DBI, SMARCD1, CHST11, PBX1, SCARB2, CREB5, NPAS3, MTATP6P1, TRIP12, B3GAT2, ADAM10, SCG3, EVI5, TRIB2, SEPT2, DSG2, CCDC88A, RNF157, LDLRAD3, GPM6B, DDAH1, GSK3B, PTPRA, CDK14, COL11A1, BRD7, PTGFRN, SDC3, FBXW11, ARL6IP5, SWAP70, AHCYL1, RAB3IP, SNX22, WNK3, AGO3, KIF1B, CTNND2, SLC44A1, CYTL1, UGP2, MT-ND3, ATRX, SON, SOX5, MAN1A2, CAMSAP2, RPN2, EDNRB, SMARCC1, SOX9, C3orf70, VMA21, TXNIP, ZNF480, SUPT16H, FEZ1, MAGED1, SEL1L3, BEX1, TTC37, FADS2, CEP170
GSA & GBM scRNA-seq integration	C9	GOLGA8A, GOLGA8B, NKTR, NEAT1, MALAT1, SNRNP70, GPR98, PAXBP1, DST, TRIO, WSB1, MYC, VPS13C, RP5-1039K5.19, ANKRD36C, TSPAN31, LL22NC03-2H8.5, MAN2A2, MARCH6, RBM6, MAPK8IP3, CD46, MCM3AP, TANC1, CAND1, CAPRIN2, APOE, SREK1, GABBR1, MACF1, THOC2, ARGLU1, TOP2A, SORL1, COL9A1, PCSK7, CYP27B1, LENG8, EZH2, KIAA2026, SUN1, TTC17, HPS4, CTB-89H12.4, PTK2, N4BP2L2, ZC3H7A, CHRM3, MMP16, DIP2A, RNPC3, IPO9, COL4A1, DPY19L3, KIAA1598, HNRNPU-AS1, LINC00969, TTL, PLCG1, CENPF, P2RX7, DDX17, ZC3H11A, KAT2A, NCOR2, SLC35E3, WDR90, MRC2, C5orf42, ZC3H14, HECTD4, PLXNA3, PPP6R2, UGGT2, PHTF2, DIP2C, SRF, BCYRN1, CNKSR3, IKBKB, ST6GALNAC2, KIAA0020, THBS2, CTTN, AHSA2, OS9, FUS, MYSM1, DGKI, MAP3K4, DDX3Y, SPECC1, SRRM2, FARP2, FMNL2, EML4, BIRC6, STAT5B, MTR, C7orf55-LUC7L2, ZBTB37, PGAP1, MALAT1.1, FUT9, PRRC2C, SLC35E2B, KRI1, USP36, TTLL3, MSL1, RP11-631M6.2, CPEB2, POLR3E, NOC2L, TULP4, SAMD4A, ADAMTS1, AKAP6, PPIP5K2, MTA1, CWF19L2, NOVA1, PGM2L1, TRPM7, AHI1, EP300, BRD1, FAT3, NRBP2, ASXL1, IGF2BP3, RP11-382A18.3, RNF213, UVSSA, HMGA2, SEL1L3, ZNF280D, WHSC1, SEMA3A, PCNT, ANKDD1A, TYRO3, QKI, SLC5A3, SNHG17, DDX39B, DDX55, SRGAP2, SETX, PRC1, PAPD4, TCERG1, CHKA, FUBP1, IFI44L, RP11-366L20.2, SACS, PARP14, ULK1, FOXK1, KIAA1109, SYNC, INPPL1, TEAD1, XXbac-BPG283O16.9, LUC7L3, FGFR1, HERC4, IGFBP5, TOR1AIP2, PABPC1L, PCDH15, XYLT1, MT-RNR2, SLC16A1-AS1, TARBP1, RHOT2, CAMSAP2, ANKRD36, NR2F2-AS1, PCBP1-AS1, AKAP13, CHERP, REXO4, NUPL1, AGAP2, DNAJC10, PCAT1, RP11-444D3.1, SUGP2, IFT80, LAS1L, SERINC5, MBNL3, AGRN, YTHDC2, CLCN7, YEATS2, POLK, CLDN12, FARP1, UNKL, GPR126, USP3, CCDC14, RP11-1023L17.1, COL6A1, KIAA0101, SOGA1, HMGXB3, ANKRD50, MFGE8, ARHGEF7, SON, ZNF334, RALGPS2, GPSM2, CRAMP1L, DVL1, BRD9, MGEA5, UBAP2L, HERC2P2, HARS, EIF3B, DCAF16, KIAA0430, PCNX, SHC2, NFATC2IP, GOLGA2P7, WHSC1L1, INTS1, PIEZO1, ST5, KIAA0907, STK36, AGAP3, MAGI2-AS3, GRIA3, LEF1, NPLOC4, CTC-444N24.7, DGKD, FAM118A, STAT1, AXIN1, ATRX, NSUN5P1, RP3-368A4.6, TMEM161B, IL1RAP, PPWD1, RNF111, UTRN, UPF2, TSPYL2, DUS1L, SAFB2, RNF144A, EMC10, GART, JMY, OPA1, PLEKHA4, SH3D19, LARP1, ACIN1, ATXN2L, LPHN1, CENPK, DNMT3A, GTF2H2, NRD1, DNASE1, CCAR1, PTPN13, PRPF3, TSHZ2, LINC01296, CPNE7, LIFR, PKD2, TNC, HNRNPA2B1, EIF2S3, TRIB1, CHTF18, ERBB3, PLXNB2, PVT1, KDM5D, SNHG14, GUSBP3, RIF1, CIC, PLXNB1, HERC2P9, SIPA1L2, GPR56, NSUN5P2, UTY, CDKN2A, FLVCR1, RABGGTB, GRID2, VIMP, FAM184A, BTAF1, USP7, NABP1, LMNB1, GPR125, PUM2, ZSCAN30, EPB41L2, REV3L, MPHOSPH9, PLEC, PGM5P2, ZNF562, NBEA, JAG1, APOC1, NCOA6, HCFC1, SEH1L, JADE1, NIPBL, FAM214A, EDRF1, TIA1, ATG16L2, DLGAP1, RC3H1, NXF1, CTA-29F11.1, AFG3L1P, BRWD3, GLCCI1, FAM92A1, DGKH, GIGYF1, PCMTD2, STAG3L3, TRMT11, COL16A1, PTCH1, LAMB1, DYNC2H1, ARFGAP1, DENND4B, ANKRD36BP1, TRIP12, TRO, RP11-315A16.1, SYNE2, ZNF37BP, ATP9B, LRIG2, PTPRZ1, ZNF529, PPP1R9A, PURB, PRKY, KANSL1, PCYT1A, BICD1, SIK3, FAM135A, RP3-394A18.1, ZNF133, BMPR1A, UHRF2, ERC1, ANKRD10-IT1, XPOT, PPFIA1, FCGRT, PROM1, LSG1, TP53BP1, NAIP, FANCI, NLGN1, ARHGEF40, NNT, TSPAN14, POM121, ZMYND8, ATP2B4, SMCHD1, SPEN, SCN1A, RBM12, WDR73, SLC38A1, PLAT, SLC6A6, CTD-3014M21.1, PLEKHH1, ZNF528-AS1, ZNF609, ERMARD, USP9X, HGS, KIAA0368, SOX10, STRA13, VGF, CEP85L, PHF3, NUP160, PCNXL4, DNAJC16, LINC00685, GAK, ASNS, DDX10, SH2B1, RP11-166D19.1, PCDH7, SNHG5, ZNF559, ARHGEF12, CSAD, NSD1, HELLS, ATP13A3, LL22NC03-N14H11.1, DMXL2, NUSAP1, IRX2, CTDSPL2, MYO5A, POLR3G, MED13L, MINA, SUV420H1, CLASP2, FAM212B, SLC7A1, FAM219B, NAA40, CTA-941F9.10, UBE4A, PREX1, APBB2, ADAT2, ATN1, NAT8L, RP11-317M11.1, PDE3B, ZNF451, ZNF711, RP11-463O12.5, BAI3, ISYNA1, PDXDC1, ARHGAP32, RIMKLB, IQGAP2, PEX1, KDM2A, MYT1, UPF3B, ZZZ3, HERC2, MYO9A, UBN1, YOD1, PPIL2, MYCBP2, SMARCAD1, TBL1XR1, GRIK2, WDR36, SPG11, TBCD, PCDH10, LMBRD2, CDK10, GGA3, TBC1D9B, KIF22, GLTSCR2, PXDN, C17orf85, AC005154.6, PBX3, NTRK3, DTWD1, XPO4, SS18L1, PHF12, TNRC6C, REST, AC159540.1, CYP20A1, LPAR4, APBA2, HOMER1, SGK3, CDK11B,

		CACNA2D1, WDR82, DHX57, ESRRA, C1orf27, NDE1, GLIPR1, D2HGDH, CCDC84, HECTD1, RP5-984P4.6, CSNK1G3, ANKLE2, CLK2, CCNB1, WDR75, RAP1GDS1, UBTF, ZNF276, PRRX1, TTF2, KIAA1244, CREB3L2, CTC-444N24.11, ARID1B, FOXK2, BAHCC1, RBM4, CTD-2017F17.2, PCNXL2, SCML1, AMDHD2, AFG3L2, DOT1L, SLC26A2, DFFA, PUS7L, RNF44, ISG20L2, MAN2B2, CEP290, RBM28, MGA, UBE3C, GALT, PTPRK, COL12A1, XPO6, ATP5A1, ICE1, SETD2, OTUD4, LRP8, LHX4-AS1, TCF3, CYP4V2, ZNF780B, FAN1, SPTAN1, SHPRH, SEC31B, PHF20L1, VPS13B, RNF157, TRIM33, RRN3, PYGB, SPRY4, CTSC, EPHA3, PCGF3, NBPF1, CREBBP, FAM13A, HDGFRP2, ABCC5, PHF10, MIA3, ANAPC7, PAPD7, COL11A2, NUFIP2, ZMYM4, BAZ1A, BAI1, RASSF8-AS1, AGAP2-AS1, DIS3, MAP4K4, ANK3, MBD6, HAPLN1, MTAP, SPTLC2, MEGF10, ATP11C, HDAC6, TM9SF3, ZNF131, UBA1, NLN, STRIP1, USP22, PYGO1, AKAP1, TSC1, HMGB2, SCAF4, CEP192, ZDBF2, ULK3, RECQL, RCC2, RBM15, GPATCH2L, NR1D2, ATR, CASP8AP2, CLK1, CCP110, DKC1, SUPT16H, FRYL, KCND2, GTF3C1, LDOC1L, AMOTL1, CBX8, SLC38A2, FBRSL1, KLHL18, SCAF11, KIAA0195, GPR82, USP21, RP11-571I18.4, SYT1, RBM12B, FBXO22, FAM73A, ENPP1, FRMD6, NCOA2, ITFG3, CABIN1, LOXL3, SH3YL1, LZTR1, SLC1A1, C21orf59, TNRC18, UBA6, ZBTB11, UNC80, CEP95, VPS45, CLSPN, ZNF124, PSMA2.1, TMEM192, CNTNAP3B, CTBP2, RP11-146D12.2, AMPD2, ATF6B, DDI2, WNK3, CEP295, SMEK2, ZNF37A, WASF3, RP5-1014D13.2, CROCCP2, CDC42BPB, PLCE1, OCLN, ITGA2, HCG18, FMNL3, SLC38A10, AGFG1, GPNMB, KLHL17, MORC2, OXCT1, MTOR, METTL21B, CTD-2228K2.7, ATP7A, MTHFD1L, RHPN1, RP4-717I23.3, EIF2AK2, TMEM260, PAWR, EPB41L3, ENTPD6, ENC1, ZNF493, FEM1B, STRN4, CASKIN1, CCDC136, ORC2, MAP3K7, SBF2, ZYG11B, SMN1, SOX5, CUL4A, SLC25A29, KIAA1328, MPDZ, DOCK7, RGL2, HIPK1, ELK4, SGSM2, SLITRK2, ZHX3, ARHGAP5, OTUD7B, SKIL, PFAS, PDK1, CPD, ICE2, NF1, MATR3.1, MCM7, MAGI2, N4BP2, SLC1A5, AK9, TMEM161B-AS1, SLC2A13, PRR14L, PRR14, C16orf13, TNKS2, CCDC57, TPT1-AS1, FAM208B, TP53BP2, CELF1, NIN, CLCN5, MAP3K5, TRIM11, EEF2K, SLC6A8, KIF13A, CXCL10, RANGAP1, ORC6, ST6GAL1, PAM, PLEKHG2, ATP5B, PHF1, AL589743.1, ZNF721, ZNHIT6, ST3GAL6, ZBED6, GAD1, MED15, HES4, RP11-1H8.5, NCKAP1, SLC25A46, NEDD1, NADK, GRIA4, NSUN6, CCT6P3, C2CD2, ETNK1, PMM2, AUTS2, GPR173, TLE4, LRRC58, ANKRD26, LCORL, CDS2, CTB-50L17.8, CHD8, POLA1, KDM6B, EIF4G3, URB1, KDM5B, PHF21B, TTC28, CRKL, SMC4, RNF217, C11orf30, DDX54, TMEM181, CAPN7, AC004951.6, GCN1L1, EXOC6, FNDC3A, LRRK2, RFFL, GPRIN3, CAD, SUPT5H, RP11-87H9.2, TDRD3, GALNT13, TEX10, ZZEF1, DNAJA3, RAD50, RMND5A, PIBF1, AQR, KIAA0930, PHLDB1, SCN3A, UBR3, NSMF, VPS37B, RP1-78B3.1, PWAR6, CLUHP3, SEC22C, B4GALNT4, MYO1C, ANKS6, PACSIN2, HEY2, HPS3, MYO19, ZBTB1, BAZ1B, CENPC, SCARF2, NUMA1, RASAL2, YES1, UBE2Q2P1, PRKRIR, EHD3, TAF1, MFSD10, RRAGB, PTPRJ, UTP20, STXBP4, NOP2, ZFY, EEA1, IGF2R, DOCK1, VPS13D, RP3-368A4.5, SUN2, SSH2, NOM1, RAD54L2, LRRC37A2, EGR2, ATP8B2, THADA, BTBD7, PDZD8, ZBTB25, ERN1, KDM4B, CSGALNACT1, UBE2Q1, XRCC2, THOC1, DDR2, ZNF84, SPATA6, MCM2, TTBK2, PRPF8, RP11-70L8.5, WDR19, CSNK1E, AC141586.5, CNOT1, RP11-513I15.6, KAT6B, KIAA0226, TRRAP, PTAR1, GFM1, FMR1, ABCD4, HIST1H1C, RBM26, DOCK10, PLXNA2, RN7SK, SMIM8, CNST, ARNT2, HERC5, INO80D, LRRFIP1P1, PABPN1, ING5, ILDR2, KIAA1033, EPM2AIP1, UBN2, NSUN2, U2AF1.1, ZSCAN29, SLC35A3, TRIM56, ABCA3, AEBP1, SNHG1, CMTR1, SFSWAP, TTTY15, CNKSR2, CEP104, SFXN4, OIP5-AS1, AP3B1, TTC3P1, PLCB4, TRAK1, SOX2-OT, SGSM3, MLXIP, CLOCK, SKIV2L2, USP8, LAMA5, COL4A5, PDCD11, ZSWIM8, TANC2, SNHG3, CCS, ZFHX4, IGDCC4, MIB1, TBX1, TRIM73, DAP3, MIR503HG, KIAA1549, PNKP, WDR26, NPHP3, RP11-258C19.7, VARS, AGPAT6, TCEB3, CEP131, TRIT1, GAL3ST4, TRABD, ADRBK1, ACBD3, KLHL42, ZNF692, NFKB1, COG8, WNK2, ASPSCR1, SETD8, VWA9, TP53, PHLPP1, SETDB1, MAD2L1, TSC2, TRAPPC13, DOCK4, DKFZp434P228, QTRTD1, SLC12A2, PPM1B, MCM3AP-AS1, ZNF236, SLC9B2, ZNF587, EZH1, UBXN2A, TNFRSF19, HAGHL, KLHL23, KANK1, GOLGA3, SLC7A11, E2F3, CSPG4, SMARCA4, RP11-797A18.5, PELI1, LINC00963, CKAP5, CLEC2D, RP11-159D12.2, ZC3H12C, SFI1, UAP1L1, FAM91A1, BACH1, TIAM2, AAK1, CSNK1G1, NLGN2, SH3KBP1, PTPLB, CCNH, C3orf17, NRP2, GCC2, CPSF1, CNTN1, RRP1B, UBXN7, C5orf63, RP11-104N10.2, NARF, TMUB2, AGK, ANKIB1, CTD-2270P14.1, SPCS3, NR2F1, PEX26, SOS1, NR0B1, MAP4K5, MAMLD1, ENTPD4, LMBR1, MLLT10, ADCY6, EPT1, AC009133.12, ARHGAP17, SUZ12P1, SCAMP4, ZC3H7B, GTPBP3, UPF3A, FAM21C, OPHN1, PCF11, ACSS1, BRD8, ZCCHC3, CYTH2, HIST2H2AC, RP11-421E14.2, FAT1, LPIN2, ERCC5, SPICE1, C7orf73, GDF11, SOCS7, SRCAP, PDS5B, SATB1, RRN3P1, ERBB4, TBC1D16, COPA, STX16, AFTPH, GATAD2A, MON2, DNM1, NUP98, PARP6, KLF7, ERO1LB, ZDHHC8, RABL2B, LINC01314, ZMIZ2, ZNF557, TBC1D10B, TIGD7, IRF2BP2, ZNF506, SCAND2P, CRYBG3, NUP155, CCDC137, C21orf58, COL19A1, ZNF320, PCID2, RAD17, PRDM2, ENGASE, CERK, GNPTAB, UBE3D, RP4-605O3.4, PYGL, ZNF26, AGPAT4, KDM7A, MNT, STK4, ERBB2IP, CLIP1, SNX5, COBL, RBBP7, PTPRM, PPP2R3B, RP3-525N10.2, PITRM1, FAM208A, KMT2B, MARK1, DISC1, CASKIN2, ZNF519, YIPF4, SPATA13, TARS2, WWP2, CCDC85C, RP11-571M6.7, WAPAL, RP11-138A9.1, ACTN2, DNAJC2, GGT7, TET3, U2AF2, TPM1, MTMR12, NUS1, TSEN54, RP11-268G12.1, LINC00470, FAM193B, RP11-544A12.8, RHBDD3, ANKRD13B, MED1, ZNF407, DUXAP8, ATG9A, KLHL36, ODF2L, PPP1R12C, FAM27E3, RBMXL1, CAPN15, BCL11A, RP13-638C3.4, RPS6KB2, CRTAP, KIAA1919, GARS, FAM160A2, TENM3, ZBED4, CEP170, QPRT, AKAP11, CWC27, XPC, CA12, AGAP1, FKBP10, NUP50, POLH, MDC1, CRHR1-IT1, AMMECR1, KIAA0586, TTC39C, FAM20B, ZNF770, KIAA0141, ARHGAP10, ERLIN1, ZNF33A, CAMSAP1, ATP1A1, ASAP2, ZMYM3, SLC1A3, MCM3, CBWD5, PTCD2, FAM195A, PDE12, ADARB1, UFD1L, TPCN2, ZNF121, RASA1, RQCD1, EMC1, ZNF17, STAG2, DENND5B, MIA, AKT2, CNTNAP1, RABGAP1, KCTD7, MKI67, TTLL12, TAF9.1, CEP164, C19orf26, RECQL5, DHFR, ADAM12, SCAPER, AP003900.6, MPPED2, PARP2, DONSON, LATS1, GAB1, EXOC1, FAM178A, ANKRD20A4, ADAMTS6, FAM179B, SEC24B, GTPBP2, TRIM27, ASPM, CDH11, ANKRD54, TTC37, pk, ZFYVE27, RPAP1, ZNF827, HBS1L, RRP7B, ZNF736, CDH10, SPATA5, STAT3, HSPG2, PLEKHG1, CHIC1, USP33, FOXO3, FAM13B, ZMYM1, SLC39A10, ENO2, ITGB5, CTD-2587M2.1, MCPH1, SNTB1, PDCD4, TRIP11, STXBP5, MREG, RP11-458F8.4, CSNK1D, GAB2, SURF6, RP11-545I5.3, SPG21, PSMA1.1, FNBP1, ZNF217, R3HDM1, TPX2, IKZF2, WBP5, PTPLAD1, CYLD, RP11-1114A5.4, TRIM25, RABL6, ATG2B, TMEM257, ZNF598, KIAA0232, ARFGAP3, ZNF316, ATXN7, NAF1, AC015813.1, IKBIP, USP24, IFRD2, FRMD4A, GTF3C4, RP11-226M10.3, EPHB3, NPTN, NPIPB4, TNPO1, SCAF8, TBC1D9, UBAP2, TRMU, AKT3, PM20D2, WDR27, PIK3C2A, CCDC150
GSA & GBM scRNA-seq integration	C10	PLP1, PTGDS, MBP, GSN, SELENOP, CLDN11, ANLN, LARP6, TF, PPP1R14A, IFIT3, NKX6-2, CRYAB, RNASE1, TMEM144, ABCA2, SGK1, QDPR, LGALS3BP, KCNMB4, CLDND1, MAG, RARRES3, GSTO1, HSPA2, HAPLN2, SHTN1, APLP1, AMD1, IFIT2, ENPP2, LHPP, SLC44A1, RAB40B, UGT8, PSAP, HLA-C, AMER2, HLA-B, NCOA7, ADIPOR2, OSBPL1A, CERCAM, AATK, SUN2, EDIL3, PLLP, CNP, PIP4K2A, B2M, ERMN, BCAS1, TUBB4A, TRIM2, DNER, S100A1, MOG, NDRG1, DOCK5, ALCAM, DPYSL2, GNAO1, CNTN2, SEMA3B, PDK4, PAQR8, EMC10, WNK1, RGCC, CDKN1C, ATP1B1, CBR1, VWA1, BEX1, GPRC5B, KIF1A, SPOCK3, RAPGEF5, SERINC1, HLA-E, CHADL, FAM102A, CFL2, ELOVL1, NACAD, B3GAT1, SIRT2, FAM107B, REEP3, MTUS1, AIF1L, SLC24A2, CTSD, PACS2, RNF13, ELOVL5, TPPP, NCAM2, AGFG1, NFASC, MAP4K4, FA2H, PXK, DNAJC6, DYNC1LI2, B3GAT3, CD9, PLA2G16, FAIM2, DAZAP2, SLC48A1, PRNP, MPC1, CD82, NDRG2, SOX10, SFT2D1, SPTLC2, SPOCK1, ARHGAP21, ARL8A, CDK18, TTLL7, SLAIN1, SEPT4, SCG5, DDR1, LAMP1, SCD5, GTPBP6, NFKBIA, MTURN, RNF130, MYRF, DICER1, FGFR2, DST, ARL6IP5,

		CAMK2N1, ANKH, GBP2, PLCL1, CTTNBP2, VAMP5, PRUNE2, CAPNS1, PLXDC2, MBNL2, CNDP1, EFHD1, APBB1, HLA-A, SCARB2, SLC22A17, SYT11, SHISA4, MAL, GLUL, PTCD3, MYLIP, ANKS1B, FRMD4B, TFEB, HIPK2, BIN1, BEX4, RETREG1, SREBF1, FBXO7, SMOC1, YWHAZ, MAP7, PCSK6, APOL2, MAP1LC3B, SGMS1, CARNS1, EPB41L2, SORT1, USP31, PPP2R2A, LAMP2, KLK6, SESTD1, CLASP2, MAP4K5, SVIP, HSD17B12, CD47, ARHGAP22, SCD, NF1, SASH1, FUT8, PADI2, KLF13, PPM1K, SERPINI1, PPA1, RFTN2, DNM3, MAPK8IP1, TSPAN15, PSMB9, OTUB1, NPC1, SNX30, PNKD, CPD, SPG7, SYNGR2, MAPK1, ARRDC2, SLC12A2, SNAPIN, PLPP2, NFE2L2, SAP30BP, CERS2, CERS4, PTP4A2, ERBB3, TMTC4, ATL1, SCN1B, MRPS18B, HSBP1, S100B, TMBIM4, PIM3, GRHPR, RNF216, RAB14, MVB12B, VRK2, STMN1, SEPT10, AIF1, ATP6V1A, APP, ZMAT3, EVI2A, CA2, MOBP, NISCH, FTH1, DCTN6, PCDH9, CD59, BEX2, OMG, ZNF652, RPS6KA5, PHLDA3, UBALD2, TMEM178A, PRKACB, DIP2C, C4orf48, ATP8A1, TGOLN2, VXN, PRDX1, ADAR, TRAPPC10, RAB30, LIMCH1, NXPE3, CARD19, SOX8, SLC25A13, PPP4R3B, RDX, MPG, RAPGEF2, WDFY3, PDZD8, TBC1D12, SH3GL3, SMPD1, KCTD3, SPP1, STRN, PLEKHB1, PSMD2, KANK1, COMMD4, HHIP, PKP4, CDH19, IRF1, NCBP3, NIPAL3, MEF2A, TNRC6C, PTPN13, SLC31A2, KIF13B, CYB5B, HBB, NINJ2, IGSF8, C11orf96, FEZ1, RNH1, LINC00844, TCEA2, TAOK1, PTMA, GPR37, TMEM208, ZDHHC20, HCN2, ZEB2, SQSTM1, TPRN, ATOX1, HERC3, HDAC5, EML2, KIF1BP, CTNNBIP1, NKAIN4, USP54, DAPK3, PTPRD, EEA1, ARHGDIA, SLC22A23, PDE1C, TTC37, ZBTB16, TLN1, COBL, SWI5, MIDN, LRCH3, ACACA, ITPKB, PARP14, MGRN1, LEPROT, HBA1, SH3GLB2, JOSD2, CCDC115, BTRC, DNM1L, TMED7, PLEKHH1, AK2, COX5B, LGMN, TTYH2, MFGE8, ARL2, ANKRD13A, NT5DC1, DAAM2, PEX16, RTN1, FMNL2, MAGT1, SEMA4D, PPP1R15B, PIGP, PRPF31, RBX1, NCOA1, KCNH8, RNF126, TAPBPL, OGFOD3, ERBIN, AC100810.1, PEF1, MAGOHB, NDUFS1, INPP5F, TTLL5, BACE1, ATCAY, ZDHHC2, ANAPC13, MFSD12, UBR3, NDFIP2, JAK1, GDAP1, GPR155, KLHL32, AKT3, IDH3G, EMILIN2, N4BP2L1, SLC23A2, AGPS, YBEY, ZDHHC14, GPM6B, DPYSL5, MLLT3, MFSD6, ITSN2, PPT1, SLCO3A1, PDXK, COQ4, GNAI1, GTDC1, AK6, TRIP11, MOB3B, ISG20, FAM120A, SNAPC5, ST18, RNF103, ZFP90, GPNMB, CAMKMT, NUDT16L1, RAB33A, TNR, PIK3IP1, ZDHHC17, PRPF8, FLNB, SLCO1A2, VPS26C, AC009041.2, NUP98, SPHK2, MIGA1, CTBP1, LPAR1, SYNJ2, MON2, HERC1, CLCN4, EPS8, FLYWCH2, POGK, PNPLA2, CLTA, TMEM125, PBDC1, PRKCE, RGL1, DCBLD2, OLIG2, LRRC1, WDFY2, CPSF2, GPS1, PTDSS2, NDST1, MPDU1, MPDZ, ARAP2, NLGN1, PLA2G4C, AGTPBP1, HBEGF, IRF9, SCAF8, RAD51B, MED27, RTN4, REPS2, ASIC1, IP6K1, ATP9A, MAFF, SURF2, FTL, DMAP1, TCF7L2, SEC61A1, NRBP1, NRDE2, NKX2-2, C11orf71, HIBCH, LUC7L, GPSM1, IFI16, FIS1, GAB1, NUMA1, MRPS28, ACTR1B, UCK1, HACL1, ARL2BP, SLC39A11, WDPCP, PHF19, MBTPS1, GK5, DSTYK, MRPS9, PEX19, ATMIN, DLL3, SECISBP2L, EXOC6B, SLC25A46, IPO13, CHST11
GSA & GBM scRNA-seq integration	C11	CD74, HLA-DRA, FTL, HLA-DRB1, APOC1, SPP1, CYBA, HLA-DPB1, C1QB, APOE, AIF1, HLA-DPA1, C1QA, TYROBP, MT1G, NPC2, SAT1, NUPR1, HLA-B, B2M, S100A11, CTSD, RNASET2, C1QC, HLA-C, LAPTM5, S100A4, CTSB, HLA-E, HSPA1A, DUSP1, SOD2, ALOX5AP, RGS1, NFKBIA, PSAP, XBP1, SRGN, TXNIP, CAPG, TSC22D3, FTH1, FCER1G, CTSS, ARHGDIB, HMOX1, CD68, FCGRT, VAMP8, MS4A7, SMAP2, GLRX, A2M, CXCR4, CD37, C3, TPT1, RPS6, KLF6, GRN, LY96, MAFB, HLA-DMA, S100A9, PLXDC2, TMSB4X, CXCL8, HERPUD1, GPNMB, BTG1, RPS12, SAMSN1, CTSC, RGS10, ITGB2, COTL1, PYCARD, MYL12A, CD53, SH3BGRL3, CCL3, TMEM176B, CORO1A, ST6GAL1, KLF2, MS4A6A, TYMP, RPS24, LIMS1, TMIGD3, PDK4, BIN1, LITAF, SUMO3, CD52, CSTB, SLCO2B1, REL, RPL22, LYZ, MS4A4A, FKBP11, GPR34, HCST, C4orf3, IFNGR1, CEBPB, CD14, HLA-A, SERPINB1, GYPC, TREM2, RPL12, LGALS1, RPS27, RPLP1, YPEL5, BST2, RPL13, DAB2, ARF6, SORL1, RPS3, CLIC1, GLIPR1, RPL26, CEBPD, RPL21, MEF2A, PABPC1, TNFRSF14, HLA-DQB1, HCLS1, LILRB4, IRF1, FCGR2A, RHOA, FUCA1, S100A8, FXYD5, SKAP2, ARRB2, IGSF6, LCP1, VIM, RPS29, OTUD1, LST1, ARPC1B, CFD, GPR183, CREG1, PSME2, GM2A, FKBP5, TMSB10, MGST2, MSR1, TNFAIP3, FAU, LGALS9, LTC4S, ANXA2, CHCHD10, RPL27A, TPM3, PSME1, FYB1, SLA, RPS15A, GIMAP4, RPL11, TMEM219, BRI3, CPVL, FCGR3A, GIMAP1, CTSH, RPL34, IER3, GIMAP7, ISG20, NCK2, ADAP2, PARP14, APBB1IP, RPL37, RPS23, EEF1B2, RCSD1, FCGR1A, HLA-DQA1, IL1B, YBX3, SERF2, IL18, RPL23A, RPS7, CD163, SERPINA1, WASF2, CSF1R, ACAA2, RPL39, HAVCR2, EIF1, CD83, HEXA, CD4, PDCD4, RPL10, LINC01736, RPL19, THEMIS2, CD164, CDKN1A, RPS25, RPS4X, YPEL3, TGFB1, MRPL18, MRPS6, RPL41, LAIR1, PIM3, FCGR1B, ARPC3, OLR1, CARD16, RPS18, IKZF1, CD84, RPL14, CYBB, NANS, EMB, CASP8, PARVB, RNASE6, NACA, RPS28, PPP1R10, GALM, GAS6, PELI1, SLC11A1, PLAUR, RPS27A, CASP1, MERTK, DIAPH2, VSIG4, FMNL1, RPS14, BHLHE41, GMFG, RPL3, RPL28, LGMN, NCF2, EVI2B, LCP2, USP53, RAB20, PER2, ARHGAP24, TPST2, GCHFR, CSTA, ACSL4, NAAA, ADAM28, RPS8

(*) Top 100 CAPE selected genes

REFERENCES

[1] CDC. https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm. 2022. https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm. Accessed 2 Jun 2022.

[2] NIH. https://training.seer.cancer.gov/disease/categories/classification.html. 2022. https://training.seer.cancer.gov/disease/categories/classification.html. Accessed 2 Jun 2022.

[3] Louis DN, Perry A, Wesseling P, Brat DJ, Cree IA, Figarella-Branger D, et al. The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. Neuro-oncology. 2021;23:1231–51.

[4] Molinaro AM, Taylor JW, Wiencke JK, Wrensch MR. Genetic and molecular epidemiology of adult diffuse glioma. Nat Rev Neurol. 2019;15:405–17.

[5] Miller KD, Ostrom QT, Kruchko C, Patil N, Tihan T, Cioffi G, et al. Brain and other central nervous system tumor statistics, 2021. Ca Cancer J Clin. 2021;71:381–406.

[6] Ostrom QT, Cioffi G, Waite K, Kruchko C, Barnholtz-Sloan JS. CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2014–2018. Neuro-oncology. 2021;23 Supplement_3:iii1–105.

[7] Weller M, Bent M van den, Preusser M, Rhun EL, Tonn JC, Minniti G, et al. EANO guidelines on the diagnosis and treatment of diffuse gliomas of adulthood. Nat Rev Clin Oncol. 2021;18:170–86.

[8] Eckel-Passow JE, Lachance DH, Decker PA, Kollmeyer TM, Kosel ML, Drucker KL, et al. Inherited genetics of adult diffuse glioma and polygenic risk scores—a review. Neuro-oncology Pract. 2022. https://doi.org/10.1093/nop/npac017.

[9] Ostrom QT, Bauchet L, Davis FG, Deltour I, Fisher JL, Langer CE, et al. The epidemiology of glioma in adults: a “state of the science” review. Neuro-oncology. 2014;16:896–913.

[10] Louis DN, Perry A, Reifenberger G, Deimling A von, Figarella-Branger D, Cavenee WK, et al. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. Acta Neuropathol. 2016;131:803–20.

[11] Miller AM, Shah RH, Pentsova EI, Pourmaleki M, Briggs S, Distefano N, et al. Tracking Tumor Evolution in Glioma through Liquid Biopsies of Cerebrospinal Fluid. Nature. 2019;565:654–8.

[12] Maggs L, Cattaneo G, Dal AE, Moghaddam AS, Ferrone S. CAR T Cell-Based Immunotherapy for the Treatment of Glioblastoma. Front Neurosci-switz. 2021;15:662064.

[13] Vendramin R, Litchfield K, Swanton C. Cancer evolution: Darwin and beyond. Embo J. 2021;40:e108389.

[14] Greaves M, Maley CC. Clonal evolution in cancer. Nature. 2012;481:306–13.

[15] Forment JV, Kaidi A, Jackson SP. Chromothripsis and cancer: causes and consequences of chromosome shattering. Nat Rev Cancer. 2012;12:663–70.

[16] Pogrebniak KL, Curtis C. Harnessing Tumor Evolution to Circumvent Resistance. Trends Genet. 2018;34:639–51.

[17] Parker TM, Gupta K, Palma AM, Yekelchyk M, Fisher PB, Grossman SR, et al. Cell competition in intratumoral and tumor microenvironment interactions. Embo J. 2021;40:e107271.

[18] Black JRM, McGranahan N. Genetic and non-genetic clonal diversity in cancer evolution. Nat Rev Cancer. 2021;21:379–92.

[19] Bozic I, Wu CJ. Delineating the evolutionary dynamics of cancer from theory to reality. Nat Cancer. 2020;1:580–8.

[20] Beiriger J, Habib A, Jovanovich N, Kodavali CV, Edwards L, Amankulor N, et al. The Subventricular Zone in Glioblastoma: Genesis, Maintenance, and Modeling. Frontiers Oncol. 2022;12:790976.

[21] Llaguno SA, Chen J, Kwon C-H, Jackson EL, Li Y, Burns DK, et al. Malignant Astrocytomas Originate from Neural Stem/Progenitor Cells in a Somatic Tumor Suppressor Mouse Model. Cancer Cell. 2009;15:45–56.

[22] Ma DK, Bonaguidi MA, Ming G, Song H. Adult neural stem cells in the mammalian central nervous system. Cell Res. 2009;19:672–82.

[23] Lee JH, Lee JE, Kahng JY, Kim SH, Park JS, Yoon SJ, et al. Human glioblastoma arises from subventricular zone cells with low-level driver mutations. Nature. 2018;560:243–7.

[24] Wang X, Zhou R, Xiong Y, Zhou L, Yan X, Wang M, et al. Sequential fate-switches in stem-like cells drive the tumorigenic trajectory from human neural stem cells to malignant glioma. Cell Res. 2021;31:684–702.

[25] Oldrini B, Curiel-García Á, Marques C, Matia V, Uluçkan Ö, Graña-Castro O, et al. Somatic genome editing with the RCAS-TVA-CRISPR-Cas9 system for precision tumor modeling. Nat Commun. 2018;9:1466.

[26] Alcantara Llaguno SR, Wang Z, Sun D, Chen J, Xu J, Kim E, et al. Adult Lineage-Restricted CNS Progenitors Specify Distinct Glioblastoma Subtypes. Cancer Cell. 2015;28:429–40.

[27] Phillips HS, Kharbanda S, Chen R, Forrest WF, Soriano RH, Wu TD, et al. Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. Cancer Cell. 2006;9:157–73.

[28] Verhaak RGW, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, et al. Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell. 2010;17:98–110.

[29] Couturier CP, Ayyadhury S, Le PU, Nadaf J, Monlong J, Riva G, et al. Single-cell RNA-seq reveals that glioblastoma recapitulates a normal neurodevelopmental hierarchy. Nat Commun. 2020;11:3406.

[30] Zuckermann M, Hovestadt V, Knobbe-Thomsen CB, Zapatka M, Northcott PA, Schramm K, et al. Somatic CRISPR/Cas9-mediated tumour suppressor disruption enables versatile brain tumour modelling. Nat Commun. 2015;6:7391.

[31] Körber V, Yang J, Barah P, Wu Y, Stichel D, Gu Z, et al. Evolutionary Trajectories of IDHWT Glioblastomas Reveal a Common Path of Early Tumorigenesis Instigated Years ahead of Initial Diagnosis. Cancer Cell. 2019;35:692-704.e12.

[32] Johnson KC, Anderson KJ, Courtois ET, Gujar AD, Barthel FP, Varn FS, et al. Single-cell multimodal glioma analyses identify epigenetic regulators of cellular plasticity and environmental stress response. Nat Genet. 2021;53:1456–68.

[33] Yabo YA, Niclou SP, Golebiewska A. Cancer cell heterogeneity and plasticity: A paradigm shift in glioblastoma. Neuro-oncology. 2021;24:669–82.

[34] Jackson M, Hassiotou F, Nowak A. Glioblastoma stem-like cells: at the root of tumor recurrence and a therapeutic target. Carcinogenesis. 2015;36:177–85.

[35] Wang Q, Hu B, Hu X, Kim H, Squatrito M, Scarpace L, et al. Tumor Evolution of Glioma-Intrinsic Gene Expression Subtypes Associates with Immunological Changes in the Microenvironment. Cancer Cell. 2017;32:42-56.e6.

[36] Andersen BM, Akl CF, Wheeler MA, Chiocca EA, Reardon DA, Quintana FJ. Glial and myeloid heterogeneity in the brain tumour microenvironment. Nat Rev Cancer. 2021;21:786–802.

[37] Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the Use of DNA Microarray Data for Diagnostic and Prognostic Classification. Jnci J National Cancer Inst. 2003;95:14–8.

[38] Lin DW, Nelson PS. Microarray Analysis and Tumor Classification. New Engl J Medicine. 2006;355:960; author reply 960.

[39] Ramaswamy S, Golub TR. DNA microarrays in clinical oncology. J Clin Oncol Official J Am Soc Clin Oncol. 2002;20:1932–41.

[40] Chang K, Creighton CJ, Davis C, Donehower L, Drummond J, Wheeler D, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45:1113–20.

[41] Rozenblatt-Rosen O, Regev A, Oberdoerffer P, Nawy T, Hupalowska A, Rood JE, et al. The Human Tumor Atlas Network: Charting Tumor Transitions across Space and Time at Single-Cell Resolution. Cell. 2020;181:236–49.

[42] LeBlanc VG, Trinh DL, Aslanpour S, Hughes M, Livingstone D, Jin D, et al. Single-cell landscapes of primary glioblastomas and matched explants and cell lines show variable retention of inter- and intratumor heterogeneity. Cancer Cell. 2022;40:379-392.e9.

[43] Marusyk A, Janiszewska M, Polyak K. Intratumor Heterogeneity: The Rosetta Stone of Therapy Resistance. Cancer Cell. 2020;37:471–84.

[44] Chang K, Creighton CJ, Davis C, Donehower L, Drummond J, Wheeler D, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45:1113–20.

[45] Campbell PJ, Getz G, Korbel JO, Stuart JM, Jennings JL, Stein LD, et al. Pan-cancer analysis of whole genomes. Nature. 2020;578:82–93.

[46] Wang L-B, Karpova A, Gritsenko MA, Kyle JE, Cao S, Li Y, et al. Proteogenomic and metabolomic characterization of human glioblastoma. Cancer Cell. 2021;39:509-528.e20.

[47] Ma X, Liu Y, Liu Y, Alexandrov LB, Edmonson MN, Gawad C, et al. Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. Nature. 2018;555:371–6.

[48] Louis DN, Ohgaki H, Wiestler OD, Cavenee WK, Burger PC, Jouvet A, et al. The 2007 WHO Classification of Tumours of the Central Nervous System. Acta Neuropathol. 2007;114:97–109.

[49] Brennan CW, Verhaak RGW, McKenna A, Campos B, Noushmehr H, Salama SR, et al. The Somatic Genomic Landscape of Glioblastoma. Cell. 2013;155:462–77.

[50] Ceccarelli M, Barthel FP, Malta TM, Sabedot TS, Salama SR, Murray BA, et al. Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. Cell. 2016;164:550–63.

[51] Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science. 2014;344:1396–401.

[52] Neftel C, Laffy J, Filbin MG, Hara T, Shore ME, Rahme GJ, et al. An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma. Cell. 2019;178:835-849.e21.

[53] Garofano L, Migliozzi S, Oh YT, D’Angelo F, Najac RD, Ko A, et al. Pathway-based classification of glioblastoma uncovers a mitochondrial subtype with therapeutic vulnerabilities. Nat Cancer. 2021;2:141–56.

[54] Venteicher AS, Tirosh I, Hebert C, Yizhak K, Neftel C, Filbin MG, et al. Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. Science. 2017;355.

[55] Varn FS, Johnson KC, Martinek J, Huse JT, Nasrallah MP, Wesseling P, et al. Glioma progression is shaped by genetic evolution and microenvironment interactions. Cell. 185:2184-2199.e16.

[56] Liesecke F, Daudu D, Bernonville RD de, Besseau S, Clastre M, Courdavault V, et al. Ranking genome-wide correlation measurements improves microarray and RNA-seq based global and targeted co-expression networks. Sci Rep-uk. 2018;8:10885.

[57] Chen B, Sirota M, Fan-Minogue H, Hadley D, Butte AJ. Relating hepatocellular carcinoma tumor samples and cell lines using gene expression data in translational research. Bmc Med Genomics. 2015;8 Suppl 2:S5.

[58] Vincent KM, Postovit L-M. Investigating the utility of human melanoma cell lines as tumour models. Oncotarget. 2017;8:10498–509.

[59] Cheng H, Yang X, Si H, Saleh AD, Xiao W, Coupar J, et al. Genomic and Transcriptomic Characterization Links Cell Lines with Aggressive Head and Neck Cancers. Cell Reports. 2018;25:1332-1345.e5.

[60] Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, et al. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. Cell. 2018;173:291-304.e6.

[61] Li Y, Kang K, Krahn JM, Croutwater N, Lee K, Umbach DM, et al. A comprehensive genomic pan-cancer classification using The Cancer Genome Atlas gene expression data. Bmc Genomics. 2017;18:508.

[62] Mostavi M, Chiu Y-C, Huang Y, Chen Y. Convolutional neural network models for cancer type prediction based on gene expression. Bmc Med Genomics. 2020;13 Suppl 5:44.

[63] Divate M, Tyagi A, Richard DJ, Prasad PA, Gowda H, Nagaraj SH. Deep Learning-Based Pan-Cancer Classification Model Reveals Tissue-of-Origin Specific Gene Expression Signatures. Cancers. 2022;14:1185.

[64] Bowman RL, Wang Q, Carro A, Verhaak RGW, Squatrito M. GlioVis data portal for visualization and analysis of brain tumor expression datasets. Neuro-oncology. 2016;19:139–41.

[65] Rohart F, Eslami A, Matigian N, Bougeard S, Cao K-AL. MINT: a multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms. Bmc Bioinformatics. 2017;18:128.

[66] Peng D, Gleyzer R, Tai W-H, Kumar P, Bian Q, Isaacs B, et al. Evaluating the transcriptional fidelity of cancer models. Genome Med. 2021;13:73.

[67] Warren A, Chen Y, Jones A, Shibue T, Hahn WC, Boehm JS, et al. Global computational alignment of tumor and cell line transcriptional profiles. Nat Commun. 2021;12:22.

[68] Abid A, Zhang MJ, Bagaria VK, Zou J. Exploring patterns enriched in a dataset with contrastive principal component analysis. Nat Commun. 2018;9:2134.

[69] Haghverdi L, Lun ATL, Morgan MD, Marioni JC. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol. 2018;36:421–7.

[70] Ghandi M, Huang FW, Jané-Valbuena J, Kryukov GV, Lo CC, McDonald ER, et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature. 2019;569:503–8.

[71] Sande BV de, Flerin C, Davie K, Waegeneer MD, Hulselmans G, Aibar S, et al. A scalable SCENIC workflow for single-cell gene regulatory network analysis. Nat Protoc. 2020;15:2247–76.

[72] Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, et al. SCENIC: single-cell regulatory network inference and clustering. Nat Methods. 2017;14:1083–6.

[73] Davie K, Janssens J, Koldere D, Waegeneer MD, Pech U, Kreft Ł, et al. A Single-Cell Transcriptome Atlas of the Aging Drosophila Brain. Cell. 2018;174:982-998.e20.

[74] Geron I, Savino AM, Fishman H, Tal N, Brown J, Turati VA, et al. An instructive role for Interleukin-7 receptor α in the development of human B-cell precursor leukemia. Nat Commun. 2022;13:659.

[75] Tan Y, Cahan P. SingleCellNet: A Computational Tool to Classify Single Cell RNA-Seq Data Across Platforms and Across Species. Cell Syst. 2019;9:207-213.e2.

[76] Pine AR, Cirigliano SM, Nicholson JG, Hu Y, Linkous A, Miyaguchi K, et al. Tumor Microenvironment Is Critical for the Maintenance of Cellular States Found in Primary Glioblastomas. Cancer Discov. 2020;10:964–79.

[77] Kinker GS, Greenwald AC, Tal R, Orlova Z, Cuoco MS, McFarland JM, et al. Pan-cancer single-cell RNA-seq identifies recurring programs of cellular heterogeneity. Nat Genet. 2020;52:1208–18.

[78] Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics. 2010;26:589–95.

[79] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.

[80] Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. Nature. 2015;526:68–74.

[81] Bindal N, Forbes SA, Beare D, Gunasekaran P, Leung K, Kok CY, et al. COSMIC: the catalogue of somatic mutations in cancer. Genome Biol. 2011;12 Suppl 1:P3–P3.

[82] Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly. 2012;6:80–92.

[83] Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell. 2018;173:371-385.e18.

[84] Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. Plos Comput Biol. 2016;12:e1004873.

[85] Amemiya HM, Kundaje A, Boyle AP. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Sci Rep-uk. 2019;9:9354.

[86] Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, et al. Pan-cancer patterns of somatic copy-number alteration. Nat Genet. 2013;45:1134–40.

[87] Hutter C, Zenklusen JC. The Cancer Genome Atlas: Creating Lasting Value beyond Its Data. Cell. 2018;173:283–5.

[88] Morgan M, Obenchain V, Hester J, Pagès H. SummarizedExperiment container. R package version 1.26.1. 2022. https://bioconductor.org/packages/SummarizedExperiment. Accessed 2 Jun 2022.

[89] Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. Nat Biotechnol. 2014;32:896–902.

[90] Gaujoux R, Seoighe C. A flexible R package for nonnegative matrix factorization. Bmc Bioinformatics. 2010;11:367.

[91] Brunet J-P, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. Proc National Acad Sci. 2004;101:4164–9.

[92] Mack SC, Singh I, Wang X, Hirsch R, Wu Q, Villagomez R, et al. Chromatin landscapes reveal developmentally encoded transcriptional states that define human glioblastoma. J Exp Medicine. 2019;216:1071–90.

[93] Bao Z-S, Chen H-M, Yang M-Y, Zhang C-B, Yu K, Ye W-L, et al. RNA-seq of 272 gliomas revealed a novel, recurrent PTPRZ1-MET fusion transcript in secondary glioblastomas. Genome Res. 2014;24:1765–73.

[94] Wang Z, Sun D, Chen Y-J, Xie X, Shi Y, Tabar V, et al. Cell Lineage-Based Stratification for Glioblastoma. Cancer Cell. 2020;38:366-379.e8.

[95] Xue W, Zhang J, Tong H, Xie T, Chen X, Zhou B, et al. Effects of BMPER, CXCL10, and HOXA9 on Neovascularization During Early-Growth Stage of Primary High-Grade Glioma and Their Corresponding MRI Biomarkers. Frontiers Oncol. 2020;10:711.

[96] Schubert M, Klinger B, Klünemann M, Sieber A, Uhlitz F, Sauer S, et al. Perturbation-response genes reveal signaling footprints in cancer gene expression. Nat Commun. 2018;9:20.

[97] Moerman T, Santos SA, González-Blas CB, Simm J, Moreau Y, Aerts J, et al. GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. Bioinformatics. 2018;35:2159–61.

[98] Türei D, Valdeolivas A, Gul L, Palacio-Escat N, Klein M, Ivanova O, et al. Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. Mol Syst Biol. 2021;17:e9923.

[99] Krogan NJ, Lippman S, Agard DA, Ashworth A, Ideker T. The Cancer Cell Map Initiative: Defining the Hallmark Networks of Cancer. Mol Cell. 2015;58:690–8.

[100] Efremova M, Vento-Tormo M, Teichmann SA, Vento-Tormo R. CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. Nat Protoc. 2020;15:1484–506.

[101] Jin S, Guerrero-Juarez CF, Zhang L, Chang I, Ramos R, Kuan C-H, et al. Inference and analysis of cell-cell communication using CellChat. Nat Commun. 2021;12:1088.

[102] Shao X, Liao J, Li C, Lu X, Cheng J, Fan X. CellTalkDB: a manually curated database of ligand–receptor interactions in humans and mice. Brief Bioinform. 2020;22.

[103] Garcia-Alonso L, Holland CH, Ibrahim MM, Turei D, Saez-Rodriguez J. Benchmark and integration of resources for the estimation of human transcription factor activities. Genome Res. 2019;29:1363–75.

[104] Kim JB, Spiegelman BM. ADD1/SREBP1 promotes adipocyte differentiation and gene expression linked to fatty acid metabolism. Gene Dev. 1996;10:1096–107.

[105] Oishi Y, Manabe I, Tobe K, Tsushima K, Shindo T, Fujiu K, et al. Krüppel-like transcription factor KLF5 is a key regulator of adipocyte differentiation. Cell Metab. 2005;1:27–39.

[106] Sybirna A, Tang WWC, Smela MP, Dietmann S, Gruhn WH, Brosh R, et al. A critical role of PRDM14 in human primordial germ cell fate revealed by inducible degrons. Nat Commun. 2020;11:1282.

[107] Schmitt MJ, Company C, Dramaretska Y, Barozzi I, Göhrig A, Kertalli S, et al. Phenotypic Mapping of Pathologic Cross-Talk between Glioblastoma and Innate Immune Cells by Synthetic Genetic Tracing. Cancer Discov. 2021;11:754–77.

[108] Krämer A, Green J, Pollard J, Tugendreich S. Causal analysis approaches in Ingenuity Pathway Analysis. Bioinformatics. 2014;30:523–30.

[109] Popov A, Scotchford C, Grant D, Sottile V. Impact of Serum Source on Human Mesenchymal Stem Cell Osteogenic Differentiation in Culture. Int J Mol Sci. 2019;20:5051.

[110] Bitorina AV, Oligschlaeger Y, Shiri-Sverdlov R, Theys J. Low profile high value target: The role of OxLDL in cancer. Biochimica Et Biophysica Acta Bba - Mol Cell Biology Lipids. 2019;1864:158518.

[111] Wamsley JJ, Kumar M, Allison DF, Clift SH, Holzknecht CM, Szymura SJ, et al. Activin Upregulation by NF-κB Is Required to Maintain Mesenchymal Features of Cancer Stem–like Cells in Non–Small Cell Lung Cancer. Cancer Res. 2015;75:426–35.

[112] Mao P, Joshi K, Li J, Kim S-H, Li P, Santana-Santos L, et al. Mesenchymal glioma stem cells are maintained by activated glycolytic metabolism involving aldehyde dehydrogenase 1A3. Proc National Acad Sci. 2013;110:8644–9.

[113] Garcia-Mesa Y, Jay TR, Checkley MA, Luttge B, Dobrowolski C, Valadkhan S, et al. Immortalization of primary microglia: a new platform to study HIV regulation in the central nervous system. J Neurovirol. 2017;23:47–66.

[114] Sato K, Ozaki K, Oh I, Meguro A, Hatanaka K, Nagai T, et al. Nitric oxide plays a critical role in suppression of T-cell proliferation by mesenchymal stem cells. Blood. 2006;109:228–34.

[115] Castro MA, Santiago I de, Campbell TM, Vaughn C, Hickey TE, Ross E, et al. Regulators of genetic risk of breast cancer identified by integrative network analysis. Nat Genet. 2016;48:12–21.

[116] Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021;184:3573-3587.e29.

[117] McGinnis CS, Murrow LM, Gartner ZJ. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. Cell Syst. 2019;8:329-337.e4.

[118] Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. Nat Rev Genet. 2015;16:133–45.

[119] Abdelfattah N, Kumar P, Wang C, Leu J-S, Flynn WF, Gao R, et al. Single-cell analysis of human glioma and immune cells identifies S100A4 as an immunotherapy target. Nat Commun. 2022;13:767.

[120] Wang L, Babikir H, Müller S, Yagnik G, Shamardani K, Catalan F, et al. The Phenotypes of Proliferating Glioblastoma Cells Reside on a Single Axis of Variation. Cancer Discov. 2019;9:1708–19.

[121] Gao R, Bai S, Henderson YC, Lin Y, Schalck A, Yan Y, et al. Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. Nat Biotechnol. 2021;39:599–608.

[122] Tran HTN, Ang KS, Chevrier M, Zhang X, Lee NYS, Goh M, et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. Genome Biol. 2020;21:12.

[123] deCarvalho AC, Kim H, Poisson LM, Winn ME, Mueller C, Cherba D, et al. Discordant inheritance of chromosomal and extrachromosomal DNA elements contributes to dynamic disease evolution in glioblastoma. Nat Genet. 2018;50:708–17.

[124] ZHANG W, LIU HT. MAPK signal pathways in the regulation of cell proliferation in mammalian cells. Cell Res. 2002;12:9–18.

[125] Bhat KPL, Balasubramaniyan V, Vaillant B, Ezhilarasan R, Hummelink K, Hollingsworth F, et al. Mesenchymal Differentiation Mediated by NF-κB Promotes Radiation Resistance in Glioblastoma. Cancer Cell. 2013;24:331–46.

[126] Fruttiger M, Calver AR, Richardson WD. Platelet-derived growth factor is constitutively secreted from neuronal cell bodies but not from axons. Curr Biol. 2000;10:1283–6.

[127] Peñuelas S, Anido J, Prieto-Sánchez RM, Folch G, Barba I, Cuartas I, et al. TGF-beta increases glioma-initiating cell self-renewal through the induction of LIF in human glioblastoma. Cancer Cell. 2008;15:315–27.

[128] Ortiz-Cuaran S, Swalduz A, Foy J-P, Marteau S, Morel A-P, Fauvet F, et al. Epithelial-to-mesenchymal transition promotes immune escape by inducing CD70 in non-small cell lung cancer. Eur J Cancer. 2022;169:106–22.

[129] Stahl N, Yancopoulos GD. IL6, IL11, LIF, OSM, cardiotrophin-1, and CNTF An example of a cytokine family sharing signal transducing receptor components. Growth Factors Cytokines Heal Dis. 1997;2:777–809.

[130] Turnquist C, Wang Y, Severson DT, Zhong S, Sun B, Ma J, et al. STAT1-induced ASPP2 transcription identifies a link between neuroinflammation, cell polarity, and tumor suppression. Proc National Acad Sci. 2014;111:9834–9.

[131] Wu Y, Fletcher M, Gu Z, Wang Q, Costa B, Bertoni A, et al. Glioblastoma epigenome profiling identifies SOX10 as a master regulator of molecular tumour subtype. Nat Commun. 2020;11:6434.

[132] Dong Z, Zhang G, Qu M, Gimple RC, Wu Q, Qiu Z, et al. Targeting Glioblastoma Stem Cells through Disruption of the Circadian Clock. Cancer Discov. 2019;9:1556–73.

[133] Taniguchi H, Hoshino D, Moriya C, Zembutsu H, Nishiyama N, Yamamoto H, et al. Silencing PRDM14 expression by an innovative RNAi therapy inhibits stemness, tumorigenicity, and metastasis of breast cancer. Oncotarget. 2017;8:46856–74.

[134] Soltanian S, Dehghani H. BORIS: a key regulator of cancer stemness. Cancer Cell Int. 2018;18:154.

[135] Kathagen-Buhmann A, Maire CL, Weller J, Schulte A, Matschke J, Holz M, et al. The secreted glycolytic enzyme GPI/AMF stimulates glioblastoma cell migration and invasion in an autocrine fashion but can have anti-proliferative effects. Neuro-oncology. 2018;20:1594–605.

[136] Qin EY, Cooper DD, Abbott KL, Lennon J, Nagaraja S, Mackay A, et al. Neural Precursor-Derived Pleiotrophin Mediates Subventricular Zone Invasion by Glioma. Cell. 2017;170:845-859.e19.

[137] Schmitt MJ, Company C, Dramaretska Y, Barozzi I, Göhrig A, Kertalli S, et al. Phenotypic Mapping of Pathologic Cross-Talk between Glioblastoma and Innate Immune Cells by Synthetic Genetic Tracing. Cancer Discov. 2021;11:754–77.

[138] Wang X, Prager BC, Wu Q, Kim LJY, Gimple RC, Shi Y, et al. Reciprocal Signaling between Glioblastoma Stem Cells and Differentiated Tumor Cells Promotes Malignant Progression. Cell Stem Cell. 2018;22:514-528.e5.

[139] Tian A, Kang B, Li B, Qiu B, Jiang W, Shao F, et al. Oncogenic State and Cell Identity Combinatorially Dictate the Susceptibility of Cells within Glioma Development Hierarchy to IGF1R Targeting. Adv Sci. 2020;7:2001724.

[140] Feng Z, Levine AJ. The regulation of energy metabolism and the IGF-1/mTOR pathways by the p53 protein. Trends Cell Biol. 2010;20:427–34.

[141] Galvao RP, Kasina A, McNeill RS, Harbin JE, Foreman O, Verhaak RGW, et al. Transformation of quiescent adult oligodendrocyte precursor cells into malignant glioma through a multistep reactivation process. Proc National Acad Sci. 2014;111:E4214–23.

[142] Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011;12:R41.

[143] Mayakonda A, Lin D-C, Assenov Y, Plass C, Koeffler HP. Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome Res. 2018;28:1747–56.

[144] Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res. 2016;44:e71–e71.

[145] Putri GH, Anders S, Pyl PT, Pimanda JE, Zanini F. Analysing high-throughput sequencing data in Python with HTSeq 2.0. Bioinformatics. 2022;38:2943–5.

[146] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40.

[147] B B. preprocessCore: A collection of pre-processing function. 2022.

[148] Cinar O, Viechtbauer W. The poolr Package for Combining Independent and Dependent p Values. J Stat Softw. 2022;101.

[149] Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-Seq data. Bmc Bioinformatics. 2013;14:7–7.

[150] Hamilton NE, Ferry M. ggtern : Ternary Diagrams Using ggplot2. J Stat Softw. 2018;87 Code Snippet 3.

[151] Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. Omics J Integr Biology. 2012;16:284–7.

[152] Alquicira-Hernandez J, Powell JE. Nebulosa recovers single-cell gene expression signals by kernel density estimation. Bioinformatics. 2021;37:2485–7.

[153] Sayols S, Scherzinger D, Klein H. dupRadar: a Bioconductor package for the assessment of PCR artifacts in RNA-Seq data. Bmc Bioinformatics. 2016;17:428.

[154] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.

[155] Väremo L, Nielsen J, Nookaew I. Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. Nucleic Acids Res. 2013;41:4378–91.

[156] Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012;28:882–3.

[157] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:e47–e47.

[158] Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, et al. The Human Transcription Factors. Cell. 2018;172:650–65.

[159] Welch JD, Kozareva V, Ferreira A, Vanderburg C, Martin C, Macosko EZ. Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. Cell. 2019;177:1873-1887.e17.