

In-Silico Analysis and
Characterization of Differentially
Expressed Genes to Distinguish
Glioma Stem Cells from Normal
Neural Stem Cells

**¹Dr Sumanta Bhattacharya Research
Scholar at MAKAUT , Public–
Foreign-Defence Policy Analyst , C.E ,
CH.E , CCIO , M.Tech , MA in
Development Studies, LLB , DIA&D ,
DG&GS ,PGCPP&A , MPI(Oxford
University) ORCID ID : 0000-0003-
2563-2787**

**²Rituparna Mazumder, B.Pharm,
M.Pharma(pursuing), Bengal School of
Technology**

**³Dr Swarupananda Mukherjee,
Assistant professor, department of
pharmaceutics NSHM Knowledge
campus, Kolkata, West Bengal**

**⁴Dr Sparsha Moni Chatterjee Ex Vice
Chancellor of IIST Shibpur, Member
of Executive Council AICTE ,Former
Director of Technical Education,
Chairman of BOAA,IIHT , Ministry of
Textile (Govt. of India).Member of
Development commission (Govt of
India).**

Abstract:

To inspect and characterize Discriminatorily Explicit Genes in Cancer Stem Cells and Neural Stem Cells so as to apprehend Glioma Stem Cells from Normal Neural Stem Cells. In this research, researchers used RNA-Seq inquiry on NSC and GSC samples to grasp about the discrepancies in their gene interpretation profiles, learn about the origins of GBM, and find potential biomarkers that could be used to selectively target CSCs during thrifty normal NSCs via legibility medicine. To recognize the differences in gene expression RNA-Seq was used among glioma stem cells and normal neural stem cells. Our research results in the

revelation of novel genes and genes with unknown alliance with gliomagenesis. Further investigation of the outlier (SRR9200898 PE) would almost certainly affirm extensive inklings to the etiology of this catastrophic cancer.

Keywords: GBM, Glioma Stem Cells, Normal Neural Stem Cells, In Silico.

Introduction:

Glioblastomas (GBMs) are a type of primary brain tumor which develops from oligodendrocyte precursor cells, neural stem cells (NSCs), and astrocytes derived from NSCs. With an incidence rate of 3.21 cases per 100,000 people, a median survival rate of 12-18 months, and a

higher male predominance, they are the most lethal and common malignancy among all brain tumors. GBMs are most commonly diagnosed in elderly patients (median age 65), and their prevalence rises with age, peaking between 75 and 84 years old and then declining after 85 (1-3).

Both genetic and epigenetic mutations are acquired by GSCs. Xenogeneic changes, such as genetic changes, serve as catalysts for transformation or work in tandem with genetic events. In contrast to genetic changes, epigenetic changes are theoretically reversible, making them appealing therapeutic targets. DNA methylation (5mC, mediated by the DNA

methyltransferases DNMT1, 3A, and 3B) and DNA hydroxymethylation (5hmC, mediated by the ten-eleven translocation complex) are two different types of DNA methylation (4, 5). The TET1, 2, and 3 families) are heavily disrupted in GBM. The prognosis for GBM patients with G-CIMP (glioma-CpG island hypermethylator phenotype induced by IDH1/2 mutation) or overall increased 5mC and/or 5hmC levels is better. Although DNA methylation and transcriptional changes in GBM have been extensively studied, little is known about the epigenetic reprogramming processes that contribute to and/or characterize glioma stem cells,

particularly in terms of TET-regulated DNA marks and their relationship to enhancer activity (6-8).

In the present research, we looked at In Silico Analysis and Characterization of Differentially Expressed Genes to differentiate Glioma Stem Cells from Normal Neural Stem Cells. The researchers utilized RNA-Seq to investigate the differences in gene expression patterns between NSCs and GSCs.

Methods:

- **Acquisition of RNA-Seq Data**

To find the perfect dataset for the study, a thorough search of the Gene Expression

Omnibus (GEO) Database was conducted. The SRP200400 related Sequence Retrieval Archive (SRA) Study dataset GSE132172 was selected (9). Twenty samples of NSCs (SRR9200813 to SRR9200832) and 20 samples of GSCs (SRR9200895 to SRR9200914) were chosen from the 188 samples on the associated SRA Run Selector and downloaded as an SRA.

- **RNA-Seq Data Analysis**

The PCR clean and Trimmomatic methods were used to remove adaptor sequences and low-quality sequencing data. The genome and junctions were mapped using TopHat2; To build

isoforms, Cufflinks was utilized; Cuffmerge was used to process GTFs; Bowtie2-t was used to map transcripts; and the final gene expression levels were produced using the RSEM method as RsemExpTable in Fragments Per Kilobase. For further analysis, the gene expression data was log transformed and quantile normalized.

- **Downstream Data Analysis**

The 40 samples were classified using Principal Component Analysis and Hierarchical Clustering based on their gene expression patterns (distance: Euclidean, linkage: ward.D2). Using differential gene expression, the DESeq2

pipeline was utilized to identify highly expressed genes in GSC and NSC samples. The differential gene expression data was filtered and retrieved if the threshold = TRUE, the p-adjusted value was 0.05, and the log2fold change value for GSCs was 3.0 (for GSCs) and -3.0 (for GSCs) (for NSCs). The top 25 most highly expressed genes in each kind of stem cell sample were further filtered, and a heat map and dendrogram were produced to analyze gene expression patterns across samples.

- **Gene Enrichment Analysis**

The gene lists were uploaded to the Database for Annotation, Visualization,

and Integrated Discovery (DAVID) v6.8 software to investigate the biological implications of the significant genes discovered via differential gene expression analysis. For pathway analysis, the Kyoto Encyclopedia of Genes and Genomes (KEGG) was used. Researchers searched Google Scholar, National Center for Biotechnology Information (NCBI) PubMed, and GeneCards® for the top 50 most significant differentially expressed genes in normal physiological circumstances and during glioma genesis to better understand their roles (which included the top 25 upregulated and top 25 downregulated genes in GSCs vs. NSCs).

Results:

After running the RNA-Seq Pipeline, the RSEMExp table included gene expression data in FPKM units for a total of 27,385 genes (Tuxedo Protocol). PCA plots showed different clustering of GSC and NSC data after quantile normalization and logscale transformation, with a principal component 1 (PC1) of 10.31 percent and a principal component 2 (PC2) of 8.9 percent. 1a (extra figure) A single outlier, a glioma stem cell sample (SRR9200898 PE), was discovered. As a result, a new PCA plot was created without the outlier. This plot yielded a PC1 of 88.03 percent and a PC2 of 2.13 percent (supplementary figure 1b). Hierarchical clustering was

used to confirm these findings, and it revealed that NSC samples (SRR9200813 to SRR9200832) and GSC samples (SRR9200895 to SRR9200897 and SRR9200899 to SRR9200914) clustered separately.

Between the two clusters, the outlier GSC sample (SRR9200898 PE) was sandwiched (supplementary figure 2).

Differential gene expression analysis found 12,437 differently expressed genes between the NSC and GSC samples (45.42 percent of the total genes). To illustrate the significant differences in gene expression between NSCs and GSCs, a volcano graphic was created

(figure 3). The genes that varied the most between NSCs and GSCs were filtered once more to find the genes that differed the most. For the log₂fold change value, a threshold of 3.0 (for GSCs) and -3.0 (for NSCs) was set. As a consequence, 348 genes had substantially varied levels of expression. We also chose the top 50 significantly different expressed genes in GSCs vs. NSCs, as well as the top 25 upregulated genes and downregulated genes (Table 1).

Table 1: Top 50 genes that vary in expression between GSCs and NSCs

Downregulated Genes			Upregulated	
ENSEMBL	Entrez	Log₂fold	ENSEMBL	I

Gene ID	Gene ID	change	Gene ID	C
ENSG00000163191.5	S100A11	-5.437646	ENSG00000160307.8	S
ENSG00000170315.12	UBB	-5.334275	ENSG00000197956.8	U
XLOC_029252	unknown	-5.0073837	ENSG00000189058.7	U
ENSG00000008394.11	MGST1	-4.9702519	ENSG00000135919.11	M
ENSG00000132386.9	SERPINF1	-4.9191335	ENSG00000261857.5	S
ENSG00000152583.11		-4.8928578	ENSG00000229344.1	

	SPARCL1			S
ENSG00000 1 97614.9	MFAP5	- 4.8696255	ENSG0000013 6 235.14	M
ENSG00000 2 13145.8	CRIP1	- 4.8111644	ENSG0000012 3 560.12	C
ENSG00000 1 39329.4	LUM	- 4.7388415	ENSG0000014 7 588.6	I
ENSG00000 1 01335.8	MYL9	- 4.6796897	ENSG0000007 8 596.9	M
ENSG00000 1 09113.16	RAB34	- 4.6787092	ENSG0000015 4 096.12	F
ENSG00000 1 98467.12		- 4.6243914	ENSG0000016 0 862.11	

	TPM2			T
ENSG000000 1 28610.10	FEZF1	- 4.5842385	ENSG00000012 9 824.14	F
ENSG000000 1 38829.9	FBN2	- 4.5675219	ENSG00000011 0 693.14	F
ENSG000000 2 25383.5	SFTA1P	- 4.5507508	ENSG00000016 4 434.10	S
ENSG000000 1 31435.11	PDLIM4	- 4.5340762	ENSG00000015 4 553.12	F
ENSG000000 1 04723.19	TUSC3	- 4.4994617	ENSG00000016 4 106.6	T
ENSG000000 1 29038.14		- 4.4942203	ENSG00000013 2 561.12	

	LOXL1			
ENSG00000067715.12	SYT1	-4.4821866	ENSG000000123610.4	S
ENSG00000105971.13	CAV2	-4.4721796	ENSG000000007237.17	C
ENSG00000137962.11	ARHGAP29	-4.4720021	ENSG0000000255737.2	A2
ENSG00000127083.7	OMD	-4.4647095	ENSG0000000240747.6	C
XLOC_012003	Proneural	-4.4462352	ENSG0000000184221.11	F
ENSG00000146411.5	SLC2A12	-4.4290885	ENSG0000000241990.4	S

XLOC_04736 7	p53- regulated lncRNAs	- 4.4287433	ENSG0000013 6 999.4	p r l
-----------------	------------------------------	----------------	------------------------	-------------

Next, we created PCA plots with only these significant genes with the outlier (Figure 1a) and without the outlier (Figure 1b) to determine the distinguishable potential of these top 50 significantly differentially expressed genes (Figure 1b). The differential expression pattern of these genes was then visualized in the heatmap (Figure 2) between two types of samples, i.e. GSCs vs NSCs.

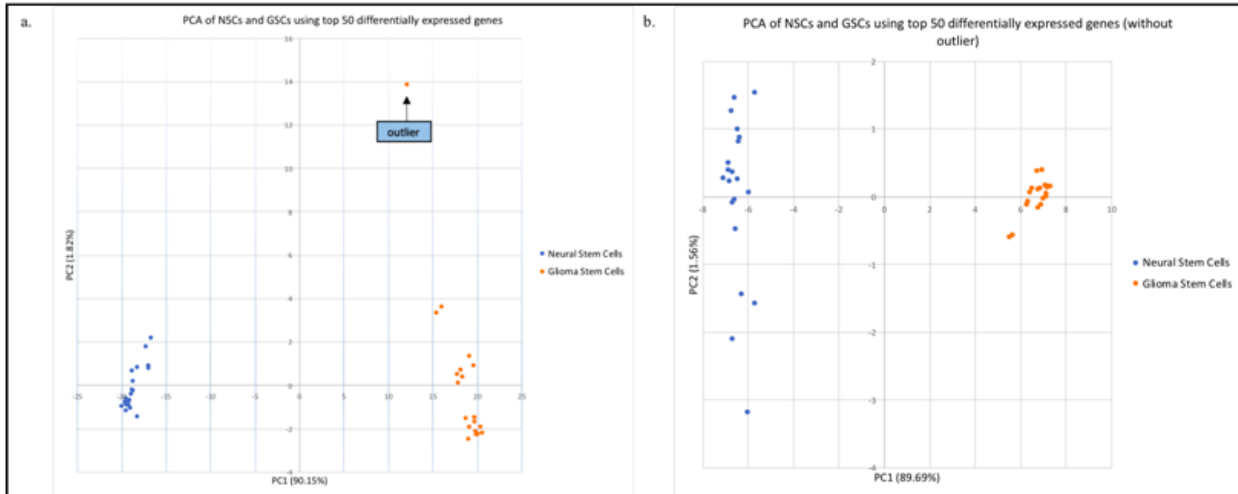


Figure 1: Exploratory data analysis using PCA based on the top 50 significantly differentially expressed genes between NSCs and GSCs. (a). PCA plot for all GSC and NSC samples, (b) PCA plot for GSC and NSC samples after removal of outlier

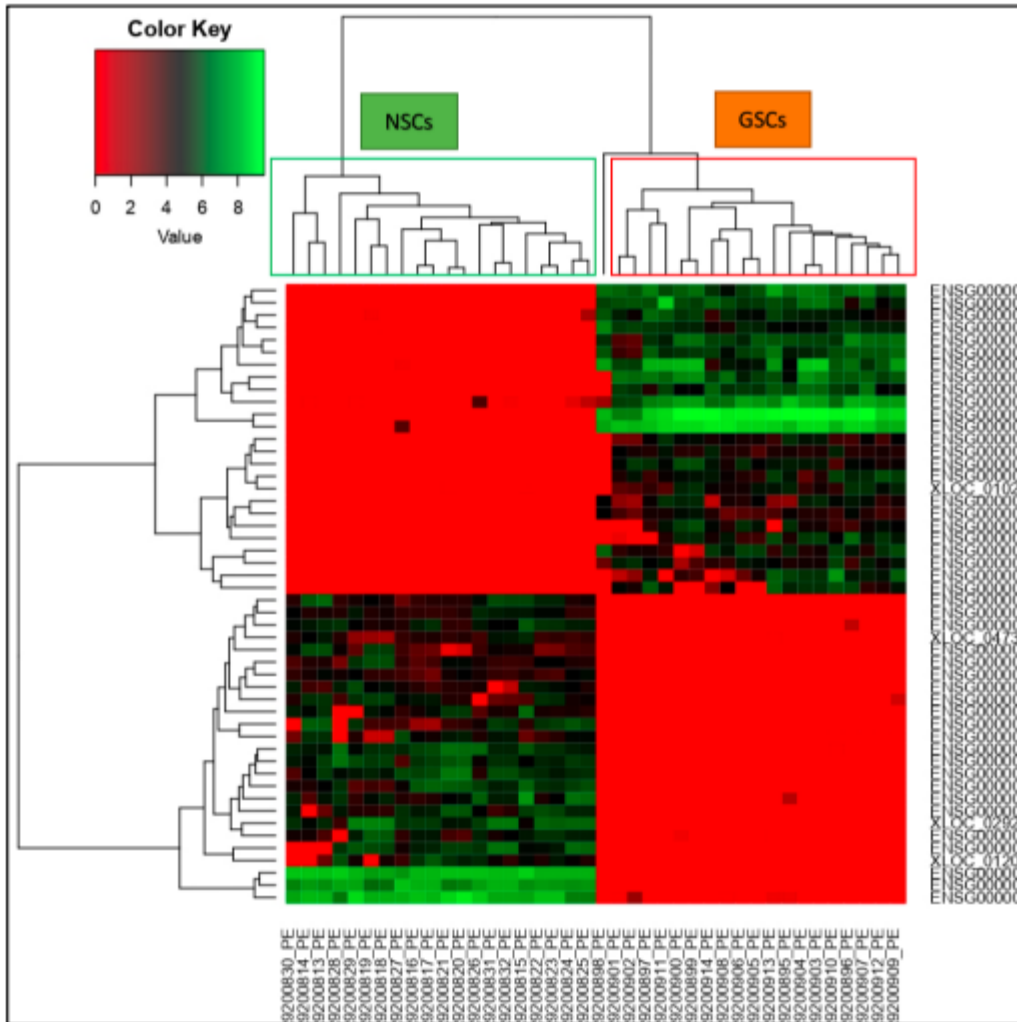


Figure 2: Heat map representing the expression pattern of the top 50 differentially expressed genes expressed genes between NSCs and GSCs.

Finally, we ran H-Clustering again using the top 50 genes (figure 3). The GSC and

NSC samples clustered independently, with the outlier sample branching in between.

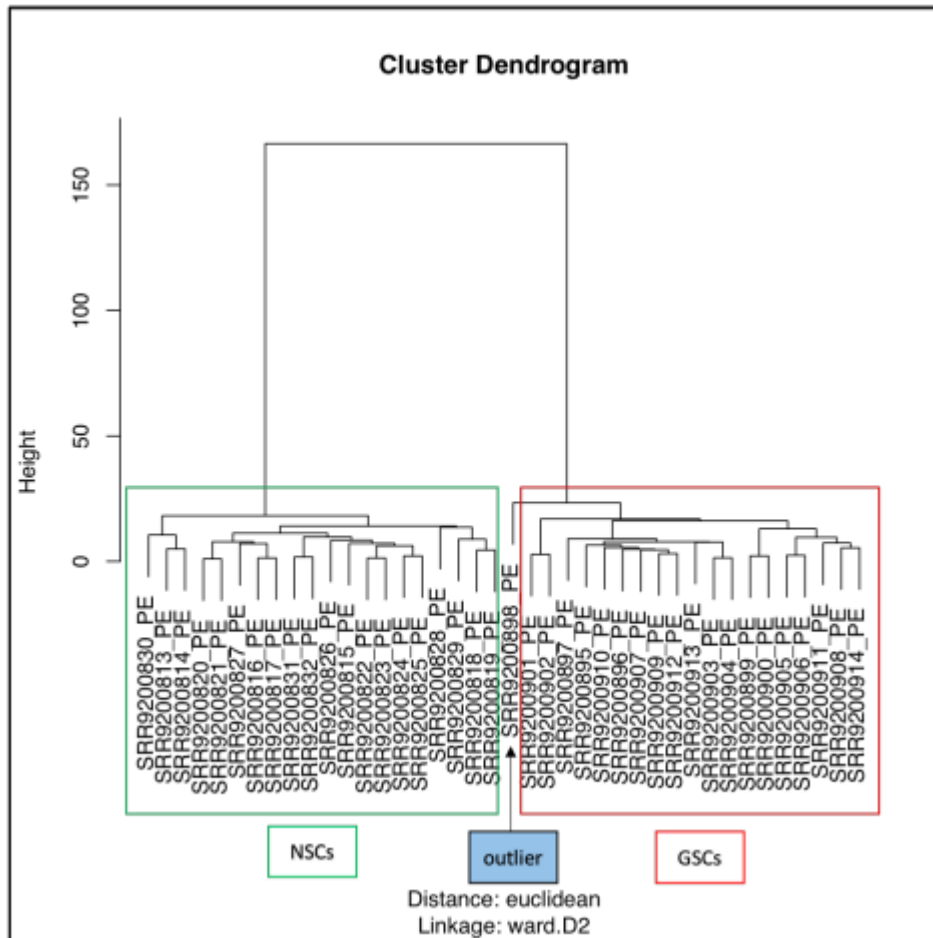


Figure 3: The unique groups of GSCs and NSCs samples are shown in a dendrogram produced by hierarchical clustering based on the top 50

substantially differentially expressed genes

Gene Enrichment analysis

On DAVID, the gene list for NSC (156 genes) and GSC (192 genes) was uploaded, and functional annotation and clustering were performed. Alzheimer's disease, respiratory function tests, Tobacco Use Disorder, body weight, bone mineral density, macular degeneration, alcoholism, and other diseases were among the top hits with the NSC gene list. Extracellular matrix, secreted, glycoprotein, signal, calcium, disulphide bond, disease mutation, membrane, and transmembrane were among the top

keyword annotations. Signal peptide, N-linked glycosylation, Leucine Rich Repeats (LRR 6), EGF-like 6, LRR 5, LRR 1, LRR 2, LRR 7, LRR 4, disulphide bond, and topological domain were among the top sequence features found (cytoplasmic). Functional clustering (medium stringency) revealed 27 clusters with 137 DAVID IDs, with the top three clusters involved in extracellular matrix, glycoprotein, and LRRs, respectively, and enrichment scores of 5.99, 4.95, and 2.05. Schizophrenia, Attention deficit hyperactivity disorder (ADHD), tobacco use disorder, depression, body height, body weight, autism, and myocardial infarction were among the top diseases

found in the GSC gene list, which contained 192 genes. Disulphide bond, glycoprotein, signal, MHC-II, cell adhesion, cell membrane, alternative splicing, polymorphism, and transmembrane were among the top keyword annotations. Glycosylation (N-linked), signal peptide, disulphide bond, topological domain (extracellular and cytoplasmic), transmembrane, splice variant, and sequence variant were among the top hits for sequence features. “The top three clusters were involved in glycoprotein disulphide bonds, melanocyte differentiation metallothionein domain, and had enrichment scores of 6.17, 2.35, and 2.15,

respectively”, according to functional clustering (medium stringency).

Ankyrin, 21, Lumican, Wnt, 51, SDC-4, and Caveolin were identified as being involved in various tumorigenesis pathways by KEGG Pathway analysis. PVRL3, PVRL2, NCAM, L1CAM, IGSF4, CDH3, and NEO1, which are enriched in the nervous system, were also shown. In the pathway analysis, ECM Receptor Interaction genes such as Fibronectin, Collagen, Laminin, Tenascin, V, 8, Thrombospondin (THBS), and Osteopontin (OPN) were found. Genes involved in axon guidance pathways, such as FAK, PAK, RasGAP, Ephrin A, Ephrin B, Robo1, Robo3, ERK, Cofilin, GSK3,

The top 25 highly upregulated and 25 highly downregulated genes in GSCs and NSCs, as well as their roles and functions in normal pathophysiology and glioma genesis, are tabulated in Tables (see Supplementary tables S1 and S2). Upregulated genes like S100B, S100A6, GPNMB, CD90, and SOX10, to name a few, have been linked to glioma genesis, whereas genes like ITM2A, KRBOX1, PRR34-AS1, SLCO4A1, and the novel gene ENSG00000154553 have not been linked to the disease (10). S100A11, UBB, MGST1, SERPINF1, and SPARCL1 are similarly downregulated genes with established associations, whereas no previous association has been

established for XLOC 02 9252 (novel gene), FBN2, SFTA1P, OMD, XLOC 01 2003 (novel gene), XLOC 04 7367. (novel gene) (11).

Conclusion:

Epigenetic changes, like genetic changes, act as catalysts for transformation or work in tandem with genetic events. In contrast to genetic changes, epigenetic changes are theoretically reversible, making them appealing therapeutic targets. Glycosylation (N-linked), signal peptide, disulphide bond, topological domain (extracellular and cytoplasmic), transmembrane, splice variant, and sequence variant were among the top hits

for sequence features. The top three clusters were involved in glycoprotein disulphide bonds, melanocyte differentiation metallothionein domain, and had enrichment scores of 6.17, 2.35, and 2.15, respectively, according to functional clustering (medium stringency).

References:

1. 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**:1396–1401. doi: 10.1126/science.1254257

2. Towner, R. A. et al. Experimental validation of 5 in-silico predicted glioma biomarkers. *Neuro Oncol.* **15**, 1625–1634 (2013).
3. Louis DN, Perry A, Reifenberger G, von Deimling A, 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: 803–820, doi: 10.1007/s00401-016-1545-1.
4. Gamazon ER, Stranger BE. The impact of human copy number variation on gene expression. *Brief Funct Genomics* 2015; 14: 352–357, doi: 10.1093/bfgp/elv017.

5. Veeriah S, Brennan C, Meng S, Singh B, Fagin JA, Solit DB, et al. The tyrosine phosphatase PTPRD is a tumor suppressor that is frequently inactivated and mutated in glioblastoma and other human cancers. *Proc Natl Acad Sci USA* 2009; 106: 9435–9440, doi: 10.1073/pnas.0900571106.
6. Mukherjee, S. (2020). Quiescent stem cell marker genes in glioma gene networks are sufficient to distinguish between normal and glioblastoma (GBM) samples. *Scientific Reports*, 10(1), 1-18.
7. Gagliardi, F., Narayanan, A., Gallotti, A. L., Pieri, V., Mazzoleni, S.,

Cominelli, M., ... & Galli, R. (2020). Enhanced SPARCL1 expression in cancer stem cells improves preclinical modeling of glioblastoma by promoting both tumor infiltration and angiogenesis. *Neurobiology of disease*, 134, 104705.

8. Ramaswamy, V.; Remke, M.; Bouffet, E.; Bailey, S.; Clifford, S.C.; Doz, F.; Kool, M.; Dufour, C.; Vassal, G.; Milde, T.; et al. Risk stratification of childhood medulloblastoma in the molecular era: The current consensus. *Acta Neuropathol.* 2016, 131, 821–831.

9. Zhao, Y., Carter, R., Natarajan, S., Varn, F.S., Compton, D.A., Gawad,

C., Cheng, C., Godek, K.M., 2019. Single-cell RNA sequencing reveals the impact of chromosomal instability on glioblastoma cancer stem cells. BMC Med. Genomics 12, 79. <https://doi.org/10.1186/s12920-019-0532-5>.

10. Azimi, I.; Milevskiy, M.J.G.; Chalmers, S.B.; Yapa, K.T.D.S.; Robitaille, M.; Henry, C.; Baillie, G.J.; Thompson, E.W.; RobertsThomson, S.J.; Monteith, G.R. ORAI1 and ORAI3 in Breast Cancer Molecular Subtypes and the Identification of ORAI3 as a Hypoxia Sensitive Gene and a Regulator of

Hypoxia Responses. *Cancers* 2019, 11, 208.

11. Brandalise, F.; Ratto, D.; Leone, R.; Olivero, F.; Roda, E.; Locatelli, C.A.; Grazia Bottone, M.; Rossi, P. Deeper and Deeper on the Role of BK and Kir4.1 Channels in Glioblastoma Invasiveness: A Novel Summative Mechanism? *Front. Neurosci.* 2020, 14, 1237.