Genomics Data 5 (2015) 333–336

Contents lists available at ScienceDirect

Genomics Data

journal homepage: http://www.journals.elsevier.com/genomics-data/

Data in Brief

Gene expression profiling distinguishes proneural glioma stem cells from mesenchymal glioma stem cells

Uma R. Chandran [a], Soumya Luthra [a], Lucas Santana-Santos [a,b], Ping Mao [c,j], Sung-Hak Kim [c], Mutsuko Minata [c], Jianfeng Li [e,g,i], Panayiotis V. Benos [a,b], Mao DeWang [j], Bo Hu [f], Shi-Yuan Cheng [f], Ichiro Nakano [c,d], Robert W. Sobol [e,g,h,i,*]

[a] Department of Biomedical Informatics, University of Pittsburgh School of Medicine, Pittsburgh, PA 15213, USA
[b] Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, PA 15213, USA
[c] Department of Neurological Surgery, The Ohio State University, Columbus, OH 43210, USA
[d] James Comprehensive Cancer Center, The Ohio State University, Columbus, OH 43210, USA
[e] Department of Pharmacology & Chemical Biology, University of Pittsburgh School of Medicine, Pittsburgh, PA 15216, USA
[f] Department of Neurology & Northwestern Brain Tumor Institute, Center for Genetic Medicine, H. Robert Lurie Comprehensive Cancer Center, Northwestern University Feinberg School of Medicine, Chicago, IL 60611, USA
[g] University of Pittsburgh Cancer Institute, Hillman Cancer Center, Pittsburgh, PA 15213-1863, USA
[h] Department of Human Genetics, University of Pittsburgh Graduate School of Public Health, Pittsburgh, PA 15216, USA
[i] University of South Alabama Mitchell Cancer Institute, Mobile, AL 36604, USA
[j] Department of Neurosurgery, First Affiliated Hospital of Medical School, Xi'an Jiaotong University, Xi'an, Shaanxi 710061, China

ARTICLE INFO

Article history:
Received 9 July 2015
Accepted 12 July 2015
Available online 14 July 2015

Keywords:
Microarray
Normalization
The Cancer Genome Atlas Project
Glioblastoma

ABSTRACT

Tumor heterogeneity of high-grade glioma (HGG) is recognized by four clinically relevant subtypes based on core gene signatures. However, molecular signaling in glioma stem cells (GSCs) in individual HGG subtypes is poorly characterized. Previously we identified and characterized two mutually exclusive GSC subtypes with distinct activated signaling pathways and biological phenotypes. One GSC subtype presented with a gene signature resembling Proneural (PN) HGG, whereas the other was similar to mesenchymal (Mes) HGG. Classical HGG-derived GSCs were sub-classified as either one of these two subtypes. Differential mRNA expression analysis of PN and Mes GSCs identified 5796 differentially expressed genes, revealing a pronounced correlation with the corresponding PN or Mes HGGs. Mes GSCs displayed more aggressive phenotypes *in vitro* and as intracranial xenografts in mice. Further, Mes GSCs were markedly resistant to radiation compared with PN GSCs. Expression of ALDH1A3 — one of the most up-regulated Mes representative genes and a universal cancer stem cell marker in non-brain cancers — was associated with self-renewal and a multi-potent stem cell population in Mes but not PN samples. Moreover, inhibition of ALDH1A3 attenuated the growth of Mes but not PN GSCs *in vitro*. Lastly, radiation treatment of PN GSCs up-regulated Mes-associated markers and down-regulated PN-associated markers, whereas inhibition of ALDH1A3 attenuated an irradiation-induced gain of Mes identity in PN GSCs *in vitro*. Taken together, our data suggest that two subtypes of GSCs, harboring distinct metabolic signaling pathways, represent intertumoral glioma heterogeneity and highlight previously unidentified roles of ALDH1A3-associated signaling that promotes aberrant proliferation of Mes HGGs and GSCs. Inhibition of ALDH1A3-mediated pathways therefore might provide a promising therapeutic approach for a subset of HGGs with the Mes signature. Here, we describe the gene expression analysis, including pre-processing methods for the data published by Mao and colleagues in PNAS [1], integration of microarray data from this study with The Cancer Genome Atlas (TCGA) glioblastoma data and also with another published study.
The raw CEL files and processed data were submitted to Gene Expression Omnibus (GEO) under the accession GSE67089.

* Corresponding author at: University of South Alabama Mitchell Cancer Institute, 1660 Springhill Avenue, Mobile, AL 36604, USA.
E-mail address: rwsobol@health.southalabama.edu (R.W. Sobol).

http://dx.doi.org/10.1016/j.gdata.2015.07.007

Specifications	
Organism/cell line/tissue	Human glioma and normal human neurospheres were derived from 19 high-grade glioma (HGG) samples, 3 human fetal brain-derived astrocytes (such as 16wf) and human neural progenitors — see Table S1 in Mao et al., 2013 [1].
Sex	See Table S1 in Mao et al., 2013 [1]
Sequencer or array type	Affymetrix Human Genome U219 Array
Data format	Raw CEL files and RMA normalized data
Experimental factors	GSC (PN vs. Mes) and tumor (GSC) vs. normal
Experimental features	We performed transcriptome microarray analysis of 27 GSC samples (triplicate samples) from nine patient-derived GSC cultures, five glioma cell lines as well as normal human astrocytes and fetal neural progenitors (16wf) as the normal controls.
Consent	Level of consent allowed for reuse if applicable; approved by Ohio State IRB under NIH guidelines.
Sample source location	Nakano lab, Department of Neurological Surgery, The Ohio State University, Columbus, Ohio. Human fetal neural stem cell 16wf was established at the University of California, Los Angeles [2]. Microarrays experiments and analysis were performed in the Sobol lab at the University of Pittsburgh Cancer Institute, Pittsburgh, PA.

1. Direct link to deposited data

http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE67089

2. Experimental design, materials and methods

2.1. Glioma tumor-derived neurospheres

All the work related to human tissues was performed at The Ohio State University under an IRB-proved protocol according to NIH guidelines. Glioma and normal neurospheres were derived from 19 HGG samples, 3 fetal brain-derived astrocytes (such as 16wf) and neural progenitors (see Table S1 [1]) as described previously [3–6]. Briefly, freshly resected glioma tumor samples were dissociated into single cells using both mechanical (gently pipet neurospheres with P1000 pipet tips 4–5 times) and enzymatic methods (TrypLE™ Express; Invitrogen, San Diego, CA). The dissociated tumor cells were cultured in DMEM/F12 (Invitrogen) supplemented with B27 (1:50), heparin (5 mg/ml), bFGF (20 ng/ml) and EGF (20 ng/ml). Growth factors (bFGF and EGF) were added twice a week. To differentiate GSCs, neurospheres were cultured in DMEM/F12 supplemented with 10% FBS for 10 days. Phenotypic characterization of these primary cultures was performed as described previously [7,8]. The human fetal neural stem cell sample (16wf) was established at the University of California, Los Angeles as described previously [2]. All the neurospheres analyzed in this study were cultured less than 20 passages. Detailed characterization of the neurospheres was performed as previously described [3].

2.2. RNA Isolation

Cells were lysed with 1 ml Qiazol lysis reagent. Total RNA was then extracted and purified using the Qiagen RNeasy Mini kit (cat# 217004) according to the manufacturer's instructions. After a wash with buffer RWT followed by two washes with buffer RPE, RNA products were eluted from the column with 30 μl RNase-free water. For each cell culture, three independent RNA samples were prepared. RNA quality was determined using an Agilent 2100 Bioanalyzer at the Cancer Biomarkers Facility at the University of Pittsburgh Cancer Institute. In all sample preparations, the average RNA integrity number (RIN) was greater than 9.0. RNA concentration was determined using a Nanodrop 2000.

2.3. Quantitative real time polymerase chain reaction (qRT-PCR)

ImProm-II™ Reverse Transcription System (Promega, Madison, WI) was used to synthesize cDNA from the resulting RNAs according to the manufacturer's protocol. The reverse transcribed cDNA was analyzed by qRT-PCR and GAPDH was used as an internal control. Each qRT-PCR reaction included 25 μl reaction mixture per well that includes 2 μl cDNA, 1 μl forward primer (10 μM), 1 μl reverse primer (10 μM), 8.5 μl of DNase/RNase-free distilled water and 12.5 μl SYBR green reagent (QIAGEN, Valencia, CA). The following cycles were performed during DNA amplification: program started from heating to 94 °C for 2 min, then followed by 45 cycles of 94 °C (30 s), 60 °C (30 s) and 72 °C (40 s), ending with the addition of melt curves as an evaluation of quality. The primer sequences for various human genes used in this study include the following: CD133 forward: ACTCCCATAAAGCTGGAC CC; CD133 reverse: TCAATTTTGGATTCATATGCCTT; Olig2 forward: CTCCTCAAATCGCATCCAGA; Olig2 reverse: AGAAAAAGGTCATCGGGC TC; Sox2 forward: ACCGGCGGCAACCAGAAGAACAG; Sox2 reverse: GCGCCGCGGCCGGTATTTAT; Sox11 forward: GGCGTTAACCAGGTTCTC AA; Sox11 reverse: TACCACCAATGGCTGCATTA; Notch1 forward: AGTGTGAAGCGGCCAATG; Notch1 reverse: ATAGTCTGCCAC GCCTCTG; CD44 forward: CC CAGATGGAGAAAGCTCTG; CD44 reverse: ACTTGGCT TTCTGTCCTCCA; LYN forward: CTGAACTCAAGTCACCGTGG; LYN reverse: TCCATCGTCACTCAAGCTGT; WT1 forward: TTAAAGGGAGTTGC TGCTGG; WT1 reverse: GACACCGTGCGTGTGTATTC; BCL2A1forward: ATGGATAAGGCAAAACGGAG; BCL2A1 reverse: TGGAGTGTCCTTTCTGGT CA; Chek1 forward: TTGGGCTATCAATGGAAGAAA; Chek1 reverse: CCCTTAGAAAGCCGGAAGTC; Chek2 forward: CCTGAGGACCAAGAACCT GA; Chek2 reverse: TGTCCCTCCCAAACCAGTAG; Rad17 forward: TGCC TACCAGCTTTATGCCT; Rad17 reverse: AAAGTGTCGCTTCAGAGGGA; Rad51 forward: CTGAGGGTACCTTTAGGCCA; Rad51 reverse: CTGGTG GTCTGTGTTGAACG; GAPDH forward: GAAGGTGAAGGTCGGAGTCA; GAPDH reverse: TTGAG GTCAATGAAGGGGTC; Vimentin forward: GGAGGACATCTTCGAGCTTC; Vimentin reverse: ATGCCTGAGATGTAGA TGCG; CDH1 forward: GGAGGAGAGCGGTGGTCAAA; CDH1 reverse: TGTGCAGCTGGCTCAAGTCAA.

For the qRT-PCR analysis of the DNA damage-repair genes, TaqMan Gene Expression Assay probes from Life technologies were used and β-actin (cat# 4352935E) was used as an internal control. Each qRT-PCR assay was performed in a 20 μl volume with 4 μl cDNA, 1 μl TaqMan probe, 10 μl TaqMan® Fast Universal Master Mix (2×) (cat# 4367846) and 5 μl of DNase/RNase-free distilled water. The reactions were performed in an ABI StepOnePlus RT-PCR system according to the manufacturer's protocol. The probe IDs for this study are ATM: Hs01112307_m1; BRCA1: Hs01556193_m1; BRCA2: Hs00609073_m1; RAD50: Hs00990023_m1; RAD51: Hs00153418_m1; and CDC25C: Hs00156411_m1.

2.4. DNA microarray analysis

Comparative analysis of mRNA expression was performed using the Human U219 Array Strip and the Affymetrix GeneAtlas system, as per the manufacturer's instructions. Microarray analysis for each of the cell cultures (in triplicate) was accomplished with 100 ng purified total RNA (described above) as the initial material and the corresponding amplified and labeled antisense RNA (aRNA) using a GeneCHip 3'IVT Express kit (Affymetrix), as described by the manufacturer. The resulting aRNA was fragmented as described by the manufacturer. The labeled aRNAs were then mixed with hybridization master mix and the hybridization cocktails were then denatured at 95 °C for 5 min, followed by 45 °C for 5 min then kept at 45 °C until applied to the hybridization tray (GeneAtlas System; 120 μl hybridization cocktail of a cell culture was transferred into a well of a 4 well hybridization tray). The array strip was immersed into hybridization cocktail and incubated in the hybridization station at 45 °C for 16 h. After hybridization, the strip was washed and stained in the GeneAtlas Fluidics Station

using the GeneAtlas Hybridization, Wash, and Stain Kit (Affymetrix #900720) and the intensity of each hybridized probe was generated using the GeneAtlas™ Imaging Station. Raw .CEL files from the Human U219 Array Strip were analyzed using the 'affy' package in R Bioconductor. The raw data were normalized and summarized using Robust Multichip Average method (RMA). At this point, each gene is represented by one or more probe sets. Several filtering steps were performed to remove uninformative probesets. Probesets expressed at less than 75 units across all samples are considered as non-expressors and were removed, but only if a gene had other probe sets that were expressed at greater than 75 units in at least one sample. If all probesets for a gene are expressed at less than 75 units across all samples, the probesets were not removed to avoid removing the gene altogether. For genes represented by multiple probe sets after filtering, the probe set with the highest inter-quartile range (IQR) was selected to represent the gene. IQR, calculated as the difference between the third and first quartiles, is a descriptive statistic used to summarize the extent of the spread of the data. This is a robust and widely recommended method to select the probeset that is most likely to detect differential expression of a gene [9]. It is important to note that although this probeset filtering method eliminates the complexity of interpreting results from multiple probe sets per gene, it does not address the issue of whether a probeset is annotated or mapped correctly to a gene. The IQR statistic is not directly correlated with probe quality or annotation. Affymetrix probesets may be remapped and re-annotated using a number of published methods whose results may disagree with the Affymetrix annotations. These alternate methods were not examined.

2.5. Differential expression and pathway analysis

Differentially expressed genes were detected between mesenchymal and proneural cells using a t-test. Genes with an FDR value <0.05 were considered to be differentially expressed. Hierarchical bi-clustering was performed on all 5,796 differentially expressed genes and 27 samples by independently clustering samples and genes. Euclidean distance and average linkage were used as similarity metric and clustering method, respectively. Clustering was done using the R statistical package (hclust function). The purpose of hierarchical bi-clustering was to identify similar groups and trends between samples and genes in the dataset. Differentially expressed genes were compared to all pathways listed in Kyoto Encyclopedia of Genes and Genomes (KEGG) and enrichment p-value was calculated using the Fisher's exact test. This analysis identifies those pathways, which have a statistically large number of genes in the differentially expressed set. Pathways that had a p-value less than 0.05 were considered significantly enriched. KEGG enrichment analysis was

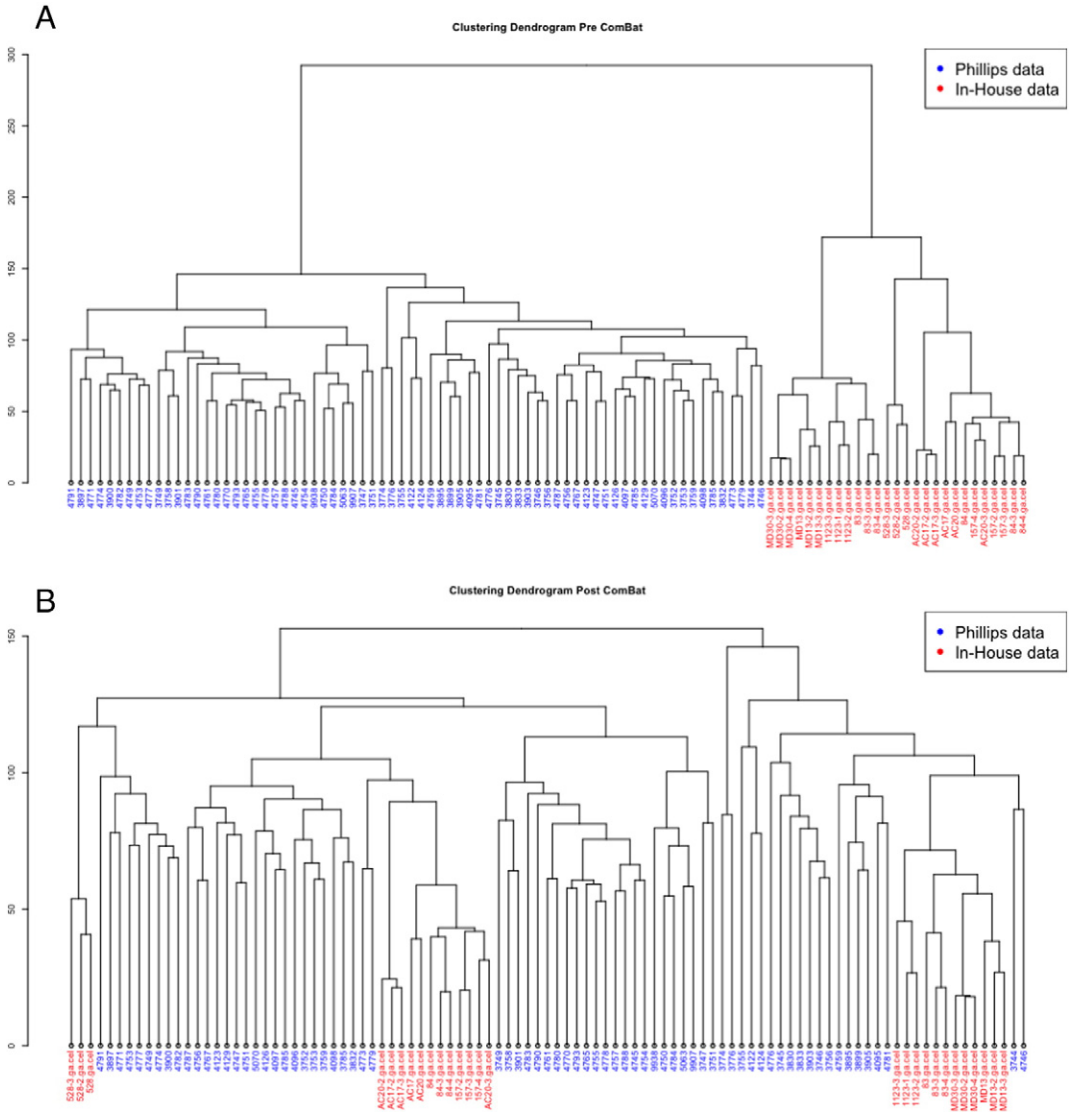

Fig. 1. A) Clustering dendrogram of the combined dataset Pre ComBat Normalization. B) Clustering dendrogram of the combined dataset Post ComBat Normalization.

Table 1
15 PN and 15 MES signature genes from Phillips Paper.

Probe	Gene symbol	Signature gene
209981_at	PIPPIN	Proneural
207723_s_at	KLRC3	Proneural
227984_at	SRRM2	Proneural
219537_x_at	DLL3	Proneural
218796_at	C20orf42	Proneural
243779_at	GALNT13	Proneural
214952_at	NCAM1	Proneural
206850_at	RRP22	Proneural
204953_at	SNAP91	Proneural
214279_s_at	NDRG2	Proneural
226913_s_at	SOX8	Proneural
232833_at	dA201G10.1	Proneural
214762_at	ATP6V1G2	Proneural
203146_s_at	GABBR1	Proneural
219196_at	SCG3	Proneural
205266_at	LIF	Mesenchymal
235417_at	FLJ25348	Mesenchymal
223333_s_at	ANGPTL4	Mesenchymal
205547_s_at	TAGLN	Mesenchymal
202628_s_at	SERPINE1	Mesenchymal
201058_s_at	MYL9	Mesenchymal
211966_at	COL4A2	Mesenchymal
226658_at	T1A-2	Mesenchymal
211981_at	COL4A1	Mesenchymal
229438_at	FAM20C	Mesenchymal
201666_at	TIMP1	Mesenchymal
209396_s_at	CHI3L1	Mesenchymal
215870_s_at	PLA2G5	Mesenchymal
211564_s_at	RIL	Mesenchymal
218880_at	FOSL2	Mesenchymal

done using custom scripts in R, pathway figures were created using the R package KEGG graph.

2.6. Comparison to TCGA GBM and Phillips HGG dataset

The Cancer Genome Atlas (TCGA) gene expression data (level 3) for 58 mesenchymal and 57 proneural tumors [10] was downloaded from the TCGA web site (https://tcga-data.nci.nih.gov/docs/publications/gbm_exp/) on July 10th 2012. TCGA data level 3 is post-normalized gene-level data, so no further normalization was performed. Since these are two independent datasets, TCGA data and in-house dataset were combined using Combat normalization [12]. The ComBat algorithm uses an empirical Bayes approach to adjust for potential batch effects that are introduced while combining data from different sources. Data was Z-scored after the removal of batch effects and hierarchical clustering was performed in R. Pearson correlation between TCGA and in-house datasets were performed in R ('cor' function) to verify that TCGA and in-house subtype expression profiles agree with each other.

As a part of the Phillips high-grade glioma (HGG) study [11], 77 primary HGGs and 23 matched recurrent HGGs were profiled on Affymetrix Human Genome U133A and U133B Arrays. The raw CEL files were downloaded from GEO (GSE4271) and RMA normalized using R. The RMA normalized data from the two chips is then put together and processed as described in the above section on DNA microarray analysis. As described above, the in-house data set and the Phillips dataset are combined using ComBat Normalization. Fig. 1 shows how the batch effect observed when the combined dataset is clustered Pre Normalization (Fig. 1A) and is adjusted for by ComBat normalization (Fig. 1B). Once the two datasets are combined, data for the 15 PN and 15 MES signature genes from the Phillips paper (Table 1) were extracted and hierarchical clustering was performed in R.

3. Discussion

We describe here the gene expression dataset used in the isolation and characterization of human glioma stem cells that exhibit characteristics of the different glioma subtypes from which they were isolated. The presence of the stem cells, which have the potential to drive glioma to different subtypes, is an important finding for understanding glioma tumor initiation and propagation. The publication from which this data set is derived has been cited in high impact journals; the microarray data are of high quality and methods we describe here will enable comparison of this data to other published studies including TCGA.

Acknowledgments

We thank Drs. H. Kornblum (University of California at Los Angeles) and K. Palanichamy (Ohio State University) for sharing their tumor samples for this study. This work was supported by the National Institutes of Health (NIH) Grants CA148629, GM087798, ES019498, and GM099213 (to R.W.S.), CA135013 (to I.N.), LM009657 (to P.V.B.), UL1RR024153 (Reis, Clinical and Translational Science Institute University of Pittsburgh), CA130966 and CA158911 (to S.-Y.C. and B.H.) and P30 CA047904 for the University of Pittsburgh Cancer Institute Core Facility (the Lentiviral Facility and the Cancer Biomarkers Facility, to R.W.S.); a Brain Cancer Research award from the James S. McDonnell Foundation (to B.H.); a Zell Scholar award from the Zell Family Foundation; funds from Northwestern Brain Tumor Institute and Department of Neurology Northwestern University (to S.-Y.C.); and the Basic Science Research Program through National Research Foundation of Korea (NRF) Grant 2011-0024089 (to S.-H.K.). P.M. was supported by the China Scholarship Council. This project also used the UPCI Cancer Bioinformatics Services which is supported in part by the National Cancer Institute award P30CA047904.

References

[1] P. Mao, K. Joshi, J. Li, et al., Mesenchymal glioma stem cells are maintained by activated glycolytic metabolism involving aldehyde dehydrogenase 1A3. Proc. Natl. Acad. Sci. U. S. A. 110 (21) (2013) 8644–8649, http://dx.doi.org/10.1073/pnas.1221478110.
[2] I. Nakano, K. Joshi, K. Visnyei, et al., Siomycin A targets brain tumor stem cells partially through a MELK-mediated pathway. Neuro-Oncology 13 (6) (2011) 622–634, http://dx.doi.org/10.1093/neuonc/nor023.
[3] Jijiwa M, Demir H, Gupta S, et al. CD44v6 regulates growth of brain tumor stem cells partially through the AKT-mediated pathway. Lesniak MS, PLoS One. 2011;6(9): e24217, http://dx.doi.org/10.1371/journal.pone.0024217
[4] G. Aad, et al., Search for production of resonant states in the photon-jet mass distribution using pp collisions at radicals = 7 TeV collected by the ATLAS detector. Phys. Rev. Lett. 108 (21) (2012) 211802.
[5] I. Nakano, et al., Maternal embryonic leucine zipper kinase is a key regulator of the proliferation of malignant brain tumors, including brain tumor stem cells. J. Neurosci. Res. 86 (1) (2008) 48–60, http://dx.doi.org/10.1002/jnr.21471.
[6] J.D. Dougherty, et al., PBK/TOPK, a proliferating neural progenitor-specific mitogen-activatedprotein kinase kinase. J. Neurosci. 25 (46) (2005) 10773–10785.
[7] T. Miyazaki, et al., Telomestatin impairs glioma stem cell survival and growth through the disruption of telomeric G-quadruplex and inhibition of the proto-oncogene, c-Myb. Clin. Cancer Res. 18 (5) (2012) 1268–1280.
[8] M. Jijiwa, et al., CD44v6 regulates growth of brain tumor stem cells partially through theAKT-mediated pathway. PLoS One 6 (9) (2011) e24217.
[9] X. Wang, D.D. Kang, K. Shen, et al., An R package suite for microarray meta-analysis in quality control, differentially expressed gene analysis and pathway enrichment detection. Bioinformatics 28 (19) (2012) 2534–2536, http://dx.doi.org/10.1093/bioinformatics/bts485.
[10] R.G. Verhaak, et al., Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell 17 (1) (2010) 98–110.
[11] H.S. Phillips, et al., Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. Cancer Cell 9 (3) (2006) 157–173.
[12] W.E. Johnson, C. Li, A. Rabinovic, Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8 (1) (2007) 118–127.