ORIGINAL ARTICLE

Whole-Genome mRNA Expression Profiling Identifies Functional and Prognostic Signatures in Patients with Mesenchymal Glioblastoma Multiforme

Zhao-Shi Bao,¹ Chuan-Bao Zhang,¹ Hong-Jun Wang,² Wei Yan,¹ Yan-Wei Liu,¹ Ming-Yang Li¹ & Wei Zhang¹

¹ Department of Neurosurgery, Beijing Tiantan Hospital, Capital Medical University, Beijing, China

² Department of Neurosurgery, The Second Affiliated Hospital of Harbin Medical University, Harbin, China

Keywords

Biomarker; Glioblastoma; Mesenchymal; Prognosis; Risk score.

Correspondence

W. Zhang, No.6, Tiantan Xili, Dongcheng District, Beijing 100050, China.

Tel./Fax: +86-10-6702-1832;

E-mail: zhangwei_vincent@yahoo.com.cn

Received 12 March 2013; revision 1 April 2013; accepted 1 April 2013

doi: 10.1111/cns.12118

The first three authors contributed equally to this work.

SUMMARY

Background: The Cancer Genome Atlas (TCGA) has divided patients with glioblastoma multiforme (GBM) into four subtypes based on mRNA expression microarray. The mesenchymal subtype, with a larger proportion, is considered a more lethal one. Clinical outcome prediction is required to better guide more personalized treatment for these patients. **Aims:** The objective of this study was to identify a mRNA expression signature to improve outcome prediction for patients with mesenchymal GBM. **Results:** For signature identification and validation, we downloaded mRNA expression microarray data from TCGA as training set and data from Rembrandt and GSE16011 as validation set. Cox regression and risk-score analysis were used to develop the 4 signatures, which were function and prognosis associated as revealed by Gene Ontology (GO) analysis and Gene Set Variation Analysis (GSVA). Patients who had high-risk scores according to the signatures had poor overall survival compared with patients who had low-risk scores. **Conclusions:** The signatures were identified as risk predictors that patients who had a high-risk score tended to have unfavorable outcome, demonstrating their potential for personalizing cancer management.

Introduction

Glioma is the most common and lethal intracranial tumor. Glioblastoma multiforme (GBM) is the most malignant type of glioma, the median overall survival (OS) of which, even after decades of efforts in developing new therapies, is still roughly a year [1]. That mainly due to the genetic heterogeneity of tumors. But with the current grading systems, problems such as the variability among observers, the ignorance to etiology, etc. will be obstacles. Thus, molecular classification-based personalized medicine represents a promising avenue for the treatment of gliomas in the future.

The Cancer Genome Atlas (TCGA) has divided GBM into four subtypes based on mRNA expression microarray [2]: proneural, neural, classical, and mesenchymal, which are widely accepted nowadays. The mesenchymal subtype was defined by an expression profile associated with mesenchyme and angiogenesis and overexpression of the *CHI3L1/YKL40* and *MET* genes, as well as astrocytic markers CD44 and MERTK and genes in the TNF superfamily and NF κ B pathways. Of the four subtypes, mesenchymal GBM had a poor survival tendency compared with other subtypes [3]. Although with poor prognosis, the mesenchymal GBM has not been stratified further. Therefore, in this study, we tried to

identify signatures based on mRNA expression microarray, which could divide patients in three datasets into two groups with different prognosis. Two of the signatures could assemble to two separate biological process, transcription and adhesion, which are vital for tumor development, invasion, and metastasis.

Methods**Datasets**

Whole-genome mRNA expression microarray data were downloaded from TCGA database (<http://cancergenome.nih.gov>) as training set. The repository for Molecular Brain Neoplasia Data (REMBRANDT, <http://caintegrator-info.nci.nih.gov/rembrandt>) and GSE16011 data (<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE16011>) were obtained as validation sets.

Statistical Analysis

For subtype annotation, we applied the previously published gene list and annotated the three datasets by prediction analysis of microarrays (PAM) in BRB Array Tools developed by Richard

Simon & BRB-ArrayTools Development Team. Of the 491 TCGA GBM samples, 159 samples were annotated as mesenchymal GBM. We then excluded samples without prognostic information or too short overall survival (OS) time (<30 days) which might due to other complications rather than GBM itself. The remaining 117 samples were used for signature identification. The other 91 and 67 mesenchymal samples in Rembrandt and GSE16011 were used as validation sets.

In next step, permutation tests were performed to identify genes that were associated significantly with OS. The permuted *P*-value for each gene was corrected by multiple comparison correction using the Benjamini–Hochberg false discovery rate (FDR). The genes with corrected permutation *P*-values < 0.01 were selected as the candidate genes. Here, 61 genes remained.

DAVID (<http://david.abcc.ncifcrf.gov/>) Gene Ontology (GO) analysis was performed for the 61 genes. Of the top four GO terms, 17 and eight genes were associated with transcription and adhesion, respectively. If we set the cutoff *P*-value as < 0.001, five genes remained. The four sets of signatures were used for further analysis.

To assess the genes that were identified for survival prediction, a risk-score formula for predicting survival was developed based on a linear combination of the gene expression level (*expr*) weighted by the regression coefficient derived from the univariate Cox regression analysis (β). The risk score for each patient was calculated as follows:

$$\text{Risk score} = \text{expr}_{\text{gene1}} \times \beta_{\text{gene1}} + \text{expr}_{\text{gene2}} \times \beta_{\text{gene2}} + \dots + \text{expr}_{\text{genen}} \times \beta_{\text{genen}}$$

Patients with high-risk scores were expected to have poor survival. According to the cutoff value (median risk score), patients in the training set were stratified into a high-risk group and a low-risk group. The same β was applied to the validation sets. For genes with multiple probes, we calculated the average expression value of them. The differences in OS between high-risk patients and low-risk patients were estimated using the Kaplan–Meier method and two-sided log-rank test.

For functional annotation, we also performed Gene Set Variation Analysis (GSVA) by GSVA package [4] of R. The gene lists of

Table 1 Sixty one genes associated significantly with OS

Symbol	Hazard ratio	Parametric <i>P</i> -value	Symbol	Hazard ratio	Parametric <i>P</i> -value
KCNF1	1.488	8.57E–05	KCNE4	1.264	0.006816
ALCAM ^b	0.647	0.000131	RORA ^a	1.296	0.007009
LHX6 ^a	0.467	0.000349	VRK3	1.832	0.00705
CST4	2.488	0.000483	LZTS1 ^a	2.891	0.007299
SPATA6	1.678	0.000582	PDE4B	1.25	0.007316
MEOX2 ^a	1.198	0.001403	C8orf44	0.381	0.007377
AMIGO2 ^b	0.816	0.00145	C1orf176	0.448	0.007418
TNFAIP6 ^b	1.243	0.001489	FXYS5	0.764	0.00753
PCDHGC3 ^b	3.32	0.001654	ZNF576 ^a	1.621	0.00754
BCAP29	2.432	0.002113	ZNF593 ^a	1.558	0.007646
C13orf18	1.287	0.002315	PDLIM1 ^a	0.797	0.007653
LAMB4 ^b	4.07	0.002333	C9orf53	4.454	0.00775
SIGLEC9 ^b	2.234	0.002648	RPS6KA5 ^a	0.652	0.007762
ITPKC	1.747	0.002694	STK17A	1.371	0.007899
EFNA5	0.148	0.003008	NPAS3 ^a	1.577	0.008039
ATP1B1	0.781	0.003912	MGST3	0.613	0.008215
AMOT	0.272	0.004074	FOXO4 ^a	0.333	0.008251
RG56	2.288	0.004491	GTF2A1 ^a	1.786	0.00879
MYO1D	0.707	0.004546	AP4M1	1.619	0.008795
BRF2 ^a	2.268	0.004803	NANS	0.616	0.008875
FLJ21963	1.303	0.0049	FKRP	2.769	0.008972
CHD3 ^a	0.471	0.005445	FZD7	1.236	0.009044
PLEKHA4	1.528	0.005584	ALKBH4	2.142	0.009071
HUS1	2.484	0.005645	TPBG ^b	0.846	0.009095
PTPRK ^{a,b}	0.705	0.005791	FBXO7	0.595	0.009225
AGTPBP1	0.636	0.005938	MNX1 ^a	0.629	0.00963
ARNTL ^a	1.396	0.006122	LFNG	2.177	0.009761
ZNF492 ^a	0.415	0.006151	SLC4A4	1.233	0.009805
SLC38A1	0.823	0.006728	SEMA6D	1.42	0.009949
CDC42EP2	0.37	0.006729	TMC6	0.593	0.009962
C7orf26	2.049	0.006778	–	–	–

^a17 genes associated with transcription. ^beight genes associated with adhesion. Gene symbol in bold, top five genes associated with overall survival (OS).

adhesion and transcription were from the GO terms GO:0016337, GO:0051091, and GO:0043433.

Results

Detection of Signature and its Association with Survival in the Training Set

In the 117 TCGA mesenchymal GBM samples, we used Cox regression to analyze each of mRNA expression microarray data in the training set and identified 61 genes that were associated significantly with OS ($P < 0.01$, Table 1). By applying GO analysis of the 61 genes, eight genes and 17 genes were associated with adhesion and transcription, which were the top four GO terms (Table 2). Meanwhile, five genes were associated with OS with

even more strict P -value ($P < 0.001$, Table 1, gene symbol in bold).

We then applied the four gene groups—61, 17, eight, and five genes, respectively, to develop signatures using the risk-score method. The signature risk score was calculated for each of the 117 patients in the training set and then was used to divide them into a high-risk group and a low-risk group based on the cutoff value (median risk score of each signature). We observed that patients in high-risk group had shorter median OS than patients in low-risk group. The survival differences risk score and patients' survival distribution were also shown in Figure 1 and Figure 2.

Validation of the Prognostic Value of Gene Signatures in Validation Sets

For validation, we also downloaded whole-genome mRNA expression profiling of glioma patients from Rembrandt and GSE16011 and did subtype annotation as described in Statistical Analysis. For the remaining 91 and 67 mesenchymal GBM patients in the two datasets, we then used the same risk-score formulas obtained from the training set and got risk score for each individual in each respective situation. In each dataset, patients were divided into high-risk group and low-risk group according to the risk score, which was higher or lower than the cutoff. As the similar results were shown in Figure 3, the prognostic value of the signatures was validated in both of the datasets.

Table 2 Top four GO terms of the 61 genes

GO term	Biological process	Gene count	P -value
GO:0007155	Cell adhesion	8	0.009962
GO:0022610	Biological adhesion	8	0.010036
GO:0045449	Regulation of transcription	17	0.011749
GO:0006355	Regulation of transcription, DNA-dependent	13	0.015799

GO, gene ontology.

Figure 1 These Kaplan–Meier estimates of overall survival in patients with glioblastoma multiforme constructed by the gene signatures. (A) 61 genes signature; (B) 17 genes signature; (C) 8 genes signature; (D) 5 genes signature. P -values were indicated for the high-risk and low-risk groups stratified according to the median risk score in the Cancer Genome Atlas (TCGA) data. H, high-risk group; L, low-risk group.

Figure 2 Analysis of the four sets of signature risk score was illustrated for patients in the training set, including (Top) gene signature risk-score distribution and (Bottom) patient survival duration. (A) 61-gene signature; (B) 17-gene signature; (C) eight-gene signature; (D) five-gene signature. The blue vertical lines in the middle of each graph represented the gene signature cutoff (median risk score), each dot represented a single patient.

Functional Annotation of the Two Groups with Different Risk Scores

As we found that the functions of the 61 genes, which were able to divide patients into two subgroups with different prognosis, were mainly associated with adhesion and transcription, we further performed GSEA with the risk score going from low to high. The gene lists of adhesion and transcription were from the GO terms GO:0016337, GO:0051091, and GO:0043433. Patients with higher risk score tended to have a lower expression of

adhesion-associated genes and a higher expression of transcription-associated genes (Figure 4). That might be an explanation of the different prognosis of the two groups divided by risk score.

The Gene Function Interpretation of the Five Genes

The five genes, as described above, were KCNF1, ALCAM, LHX6, CST4, and SPATA6. With combined risk-score analysis, they could

Figure 3 Kaplan-Meier estimates of overall survival in patients with mesenchymal glioblastoma multiforme (GBMs) illustrated the risk-score analysis using the four separate gene signatures. Except for the 17-gene signature in GSE16011, which got marginal P -value, patients could be divided into two groups with significantly different prognosis. H, high-risk group; L, low-risk group. Cutoff value, (A-C, G) median risk score; (D-F, H), patient proportion, 29:38, 59:32, 39:28, 42:25 (low-risk vs. high-risk).

Figure 4 Gene Set Variation Analysis of the Cancer Genome Atlas (TCGA) samples with risk score of 61-gene signature. The risk score (upper panel) was calculated with the formula described above and ranked from left to right. Gene set enrichment scores (lower panel) of adhesion and transcription were analyzed by Gene Set Variation Analysis (GSVA) package of R. Patients with higher risk score tended to have a lower expression of adhesion-associated genes and higher expression of transcription-associated ones.

divide patients into high- and low-risk groups with significant prognosis.

KCNF1, voltage-gated potassium channel, subfamily F, member 1, was first isolated, characterized, and mapped in the year 1997 [5]. Voltage-gated potassium channels represent the most complex class of voltage-gated ion channels from both functional and structural standpoints. Their diverse functions in the nervous system include regulating neurotransmitter release, neuronal excitability, cell volume, and so on. This gene, which is intronless and expressed in all tissues tested, encodes a member of the potassium channel, voltage-gated, and subfamily Lubec et al. [6] found that there was no statistically significant difference between the control and the Down Syndrome group by applying an array of 96 brain RNAs mainly consisting of channels and transporters to show their expression levels in the two groups. Somatic mutation has also been found in intraductal papillary mucinous neoplasm by applying whole-exome sequencing [7].

ALCAM, activated leukocyte cell adhesion molecule, also known as CD166, was first cloned, mapped, and characterized in the year 1995. It was found to implicate in cell adhesion and migration [8,9]. Carbotti et al. [10] also identified serum ALCAM as a potential biomarker of epithelial ovarian cancer in type II tumors. There are similar findings [11–14] in other tumor types.

LHX6 encodes a protein which may function as a transcriptional regulator and may be involved in the control of differentiation and development of neural cells [15,16]. Zhang et al. [17] found some of the mechanisms of LHX6 involved in normal craniofacial and tooth development. It was also been found a tumor suppressor in human cervical cancer [18] and a methylation marker in head and neck carcinomas [19].

CST4, cystatin S, is a secreted protein, which is commonly found in saliva, tears, and seminal plasma. Secreted proteins are known to influence pathological interactions between metastatic cancer cells and the original site. Blanco MA found that high expression of CST4 was associated with bone metastasis by applying nonquantitative mass spectrometry [20]. Wiech T also identified its association with esophageal adenocarcinoma by applying single nucleotide polymorphism (SNP) microarrays [21].

SPATA6, a spermatogenesis associated gene, has rarely been reported. It was first identified in rat by Yamano et al. [22] in 2001. So far, we have known little about its association with cancer.

Discussion and Conclusion

Glioma is the most common type of intracranial tumor, about half of which is GBM, the most malignant one. Based on the current histopathologic classification system, there is a high rate of divergent diagnoses, inexact prognostic capabilities, and poor therapeutic predictive properties, which all reveal the urgent need for an objective, molecular-based classification system.

Molecular classification of gliomas, especially GBM has developed rapidly these years. Several groups have reported their classification systems based on mRNA expression [2,3,23,24], microRNA expression [25], or methylation [26,27]. Among them, there are more reports focused on mRNA expression classification systems, and the mesenchymal and proneural subtypes were identified consistently through the various datasets. Because of the molecular features of mesenchymal GBM I mentioned above, it had a poor survival tendency compared with other subtypes [3], which revealed the urgency for further stratification. In this study, we performed Cox regression and risk-score analysis on three mRNA expression microarray data from patients in western countries and identified a series of gene signatures, which could divide patients into high- and low-risk groups with different prognosis. And the prognostic value of the signatures was validated in additional datasets. Because of the potential differences from platforms, population, and so on, the marginal *P*-value of 17-gene signature in GSE16011 is acceptable.

By applying GO analysis of the first one, 61-gene signature, obtained by setting the *P*-value as < 0.01 , we further selected another two sets of signatures based on their function. 17 and eight of the 61 genes are associated with transcription and adhesion, respectively. Those are vital roles for tumor development, invasion, and metastasis. We also performed GSVA with risk score going from low to high, where we found that patients with higher risk score tended to have a lower expression of adhesion-associated genes and a higher expression of transcription-associated genes. Therefore, we could infer that the reason for the divisibility of mesenchymal GBM patients is mainly the two biological processes.

It has been reported that E-cadherin, for example, a key cell-to-cell adhesion molecule, involved the loss in carcinoma cells. By forming adherens junctions with adjacent epithelial cells, E-cadherin helps to assemble epithelial cell sheets and maintains the quiescence of the cells within these sheets. Increased expression of E-cadherin was well established as an antagonist of invasion and metastasis, whereas reduction of its expression was known to potentiate these phenotypes [28,29]. Expression of genes encoding other cell-to-cell adhesion molecules is demonstrably altered in some highly aggressive carcinomas, with those favoring cytostasis typically being downregulated. Let alone transcription-associated genes, which could directly up- or low-regulate oncogenes or tumor suppressors. Thus, that might be a good explanation for why patients with high -risk

score (i.e., poor prognosis) had a low expression of adhesion-associated genes and high expression of transcription-associated ones.

In summary, by applying Cox regression and risk-score analysis on mesenchymal GBM both in the training and validation data, we identified four sets of signatures, which could divide patients into subgroups with significantly different prognosis. Meanwhile, we found that patients in the separate group had different gene set variation, which might partially explain the prognosis difference. To our knowledge, this is the first study focusing on substratifications of current TCGA mRNA classification. Although the results are all obtained by bioinformatic analysis without *in vivo* or *in vitro* assays, which are needed in the future, the high concordance among the three datasets still revealed the significance and

potential of the classification system for personalizing cancer management.

Acknowledgments

This work was supported by grants from National High Technology Research and Development Program (No. 2012AA02A508), International Science and Technology Cooperation Program (No. 2012DFA30470) and National Natural Science Foundation of China (No. 81201993).

Conflict of Interest

The authors declare no conflict of interest.

References

- Wang Y, Jiang T. Understanding high grade glioma: Molecular mechanism, therapy and comprehensive management. *Cancer Lett* 2013;**331**:139–146.
- Verhaak RG, Hoadley KA, Purdom E, 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.
- Yan W, Zhang W, You G, et al. Molecular classification of gliomas based on whole genome gene expression: A systematic report of 225 samples from the Chinese Glioma Cooperative Group. *Neuro Oncol* 2012;**14**:1432–1440.
- Hanzelmann S, Castelo R, Guinney J. GSEA: Gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics* 2013;**14**:7.
- Su K, Kyaw H, Fan P, et al. Isolation, characterization, and mapping of two human potassium channels. *Biochem Biophys Res Commun* 1997;**241**:675–681.
- Lubec G, Sohn SY. RNA microarray analysis of channels and transporters in normal and fetal Down syndrome (trisomy 21) brain. *J Neural Transm Suppl* 2003; 215–224.
- Furukawa T, Kuboki Y, Tanji E, et al. Whole-exome sequencing uncovers frequent GNAS mutations in intraductal papillary mucinous neoplasms of the pancreas. *Sci Rep* 2011;**1**:161.
- Thelen K, Jaehrling S, Spatz JP, Pollerberg GE. Depending on its nano-spacing, ALCAM promotes cell attachment and axon growth. *PLoS ONE* 2012;**7**: e40493.
- Iolyeva M, Karaman S, Willrodt AH, Weingartner S, Vigl B, Halin C. Novel role for ALCAM in lymphatic network formation and function. *FASEB J* 2013;**27**:978–990.
- Carbotti G, Orengo AM, Mezzanzanica D, et al. Activated leukocyte cell adhesion molecule soluble form: A potential biomarker of epithelial ovarian cancer is increased in type II tumors. *Int J Cancer* 2012;**132**: 2597–2605.
- Chaker S, Kak I, MacMillan C, Ralhan R, Walfish PG. Activated leukocyte cell adhesion molecule is a marker for thyroid carcinoma aggressiveness and disease-free survival. *Thyroid* 2013;**23**:201–208.
- Federman N, Chan J, Nagy JO, et al. Enhanced growth inhibition of osteosarcoma by cytotoxic polymerized liposomal nanoparticles targeting the alcam cell surface receptor. *Sarcoma* 2012;**2012**:126906.
- Ishiguro F, Murakami H, Mizuno T, et al. Membranous expression of activated leukocyte cell adhesion molecule contributes to poor prognosis and malignant phenotypes of non-small-cell lung cancer. *J Surg Res* 2013;**179**:24–32.
- Jiao J, Hindoyan A, Wang S, et al. Identification of CD166 as a surface marker for enriching prostate stem/progenitor and cancer initiating cells. *PLoS ONE* 2012;**7**: e42564.
- Volk DW, Matsubara T, Li S, et al. Deficits in transcriptional regulators of cortical parvalbumin neurons in schizophrenia. *Am J Psychiatry* 2012;**169**:1082–1091.
- Neves G, Shah MM, Liodis P, et al. The LIM Homeodomain Protein Lhx6 Regulates Maturation of Interneurons and Network Excitability in the Mammalian Cortex. *Cereb Cortex* 2012; doi: 10.1093/cercor/bhs159.
- Zhang Z, Gutierrez D, Li X, et al. The LIM Homeodomain Transcription Factor LHX6: A Transcriptional Repressor That Interacts With pituitary homeobox 2 (PITX2) to regulate odontogenesis. *J Biol Chem* 2013;**288**: 2485–2500.
- Jung S, Jeong D, Kim J, et al. Epigenetic regulation of the potential tumor suppressor gene, hLHX6.1, in human cervical cancer. *Int J Oncol* 2011;**38**:859–869.
- Estecio MR, Youssef EM, Rahal P, et al. LHX6 is a sensitive methylation marker in head and neck carcinomas. *Oncogene* 2006;**25**:5018–5026.
- Blanco MA, LeRoy G, Khan Z, et al. Global secretome analysis identifies novel mediators of bone metastasis. *Cell Res* 2012;**22**:1339–1355.
- Wiech T, Nikolopoulos E, Weis R, et al. Genome-wide analysis of genetic alterations in Barrett's adenocarcinoma using single nucleotide polymorphism arrays. *Lab Invest* 2009;**89**:385–397.
- Yamano Y, Ohyama K, Sano T, et al. A novel spermatogenesis-related factor-1 gene expressed in maturing rat testis. *Biochem Biophys Res Commun* 2001;**289**:888–893.
- Phillips HS, Kharbada S, Chen R, 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**:157–173.
- Li A, Walling J, Ahn S, et al. Unsupervised analysis of transcriptomic profiles reveals six glioma subtypes. *Cancer Res* 2009;**69**:2091–2099.
- Zhang W, Zhang J, Yan W, et al. Whole-genome microRNA expression profiling identifies a 5-microRNA signature as a prognostic biomarker in Chinese patients with primary glioblastoma multiforme. *Cancer* 2013;**119**:814–824.
- Noushmehr H, Weisenberger DJ, Diefes K, et al. Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell* 2010;**17**:510–522.
- Zhang W, Yan W, You G, et al. Genome-wide DNA methylation profiling identifies ALDH1A3 promoter methylation as a prognostic predictor in G-CIMP- primary glioblastoma. *Cancer Lett* 2013;**328**:120–125.
- Berx G, van Roy F. Involvement of members of the cadherin superfamily in cancer. *Cold Spring Harb Perspect Biol* 2009;**1**:a003129.
- Cavallaro U, Christofori G. Cell adhesion and signalling by cadherins and Ig-CAMs in cancer. *Nat Rev Cancer* 2004;**4**:118–132.