A thirteen-gene set efficiently predicts the prognosis of glioblastoma

HUYIN YANG, LUHAO JIN and XIAOYANG SUN

Department of Neurosurgery, Affiliated Huaian No. 1 People's Hospital of Nanjing Medical University, Huai’an, Jiangsu 223300, P.R. China

Received January 21, 2018; Accepted September 6, 2018

DOI: 10.3892/mmr.2019.9801

Abstract. Glioblastoma multiforme (GBM) is the most common type of brain cancer; it usually recurs and patients have a short survival time. The present study aimed to construct a gene expression classifier and to screen key genes associated with GBM prognosis. GSE7696 microarray data set included samples from 10 recurrent GBM tissues, 70 primary GBM tissues and 4 normal brain tissues. Seed genes were identified by the ‘survival’ package in R and subjected to pathway enrichment analysis. Prognostic genes were selected from the seed genes using the ‘rb surv’ package in R, unsupervised hierarchical clustering, survival analysis and enrichment analysis. Multivariate survival analysis was performed for the prognostic genes, and the GBM data set from The Cancer Genome Atlas database was utilized to validate the prognostic genes. Of the 1,785 seed genes analyzed, 13 prognostic feature genes, including collagen type XXVIII α1 chain (COL28A1), PDS5 cohesin-associated factor A (PDS5A), zinc-finger DHHC-type containing 2 (ZDHHC2), zinc-finger protein 24 (ZNF24), myosin VA (MYO5A) and myeloid/lymphoid or mixed-lineage leukemia translocated to 4 (MLLT4), were identified. These genes performed well on sample classification and prognostic risk differentiation, and six pathways, including adherens junction, cyclic adenosine 3',5'-monophosphate signaling and Ras signaling pathways, were enriched for these feature genes. The high-risk group was slightly older compared with the low-risk group. The validation data set confirmed the prognostic value of the 13 feature genes for GBM; of these, COL28A1, PDS5A, ZDHHC2, ZNF24, MYO5A and MLLT4 may be crucial. These results may aid the understanding of the pathogenesis of GBM and provide important clues for the development of novel diagnostic markers or therapeutic targets.

Introduction

Glioblastoma multiforme (GBM) is the most common and the most invasive subtype of brain cancer; it is characterized by symptoms that include personality changes, headaches, nausea and unconsciousness (1,2). GBM originates from normal brain cells or low-grade astrocytoma, and may be induced by genetic disorders and radiation exposure (3,4). Clinical techniques for the treatment of GBM include surgery combined with radiation therapy or chemotherapy; however, the survival benefit is limited to ~12-15 months, or even shorter if the disease recurs (4).

Gene therapy is a novel strategy for treating cancers (5). Transient receptor potential genes are overexpressed in GBM, which promote the survival of patients (6,7). It has been previously reported that low expression of B cell-specific Moloney murine leukemia virus integration site 1 suppresses proliferation and promotes apoptosis of U251 GBM cells, and enhances the chemosensitivity of these cells to cisplatin (8,9). The expression level of epidermal growth factor-containing fibulin extracellular matrix protein 1 (EFEMP1) was associated with the survival of patients with GBM treated with temozolomide (TMZ) (10); thus, EFEMP1 is considered a target for overcoming TMZ-resistance in GBM. Enhancer of zeste homolog 2 (EZH2) overexpression was associated with tumor grade and predicts short overall survival in patients with GBM (11); thus, EZH2 may be a promising prognostic factor and therapeutic target for patients. Additionally, HOX transcript antisense RNA (HOTAIR) overexpression was associated with poor outcome in patients with GBM, and HOTAIR may be a therapeutic molecular target for this disease (12,13). O-6-Methylguanine-DNA methyltransferase (MGMT) methylation status and mutations in the isocitrate dehydrogenase 1 (IDH1) gene are two known clinicopathological factors linked to overall survival of patients with GBM (14). MGMT methylation is significantly associated with the clinical prognosis of GBM (15); IDH1 is a prognostic marker of GBM, and mutations in this gene diminish the malignant progression of glioma (16). In addition, high expression of interleukin-13R mRNA is strongly associated with poor prognosis of GBM (17). POZ/BTB and AT hook-containing zinc-finger 1 is another prognostic marker of GBM; it is overexpressed in GBM-derived glioma-initiating stem cells, and is associated with the characteristic stem cell capacity to grow as...
neurospheres in vitro (18). Despite this collective knowledge, the genes involved in the prognosis of GBM have not been comprehensively reported.

In 2008, Murat et al (19) established the GSE7696 gene expression profile, and demonstrated that high expression levels of epidermal growth factor receptor and stem cell–related ‘self-renewal’ signature are involved in the resistance to concomitant chemoradiotherapy of GBM. In 2011, Lambiv et al (20) used the GSE7696 data set to explore the action mechanism of tumor suppressor gene Wnt inhibitory factor 1 (WIF1) in GBM models, and concluded that WIF1 may have a tumor suppressive role in GBM through senescence.

However, this data set has not been fully explored. Using the GSE7696 data, additional key genes associated with the prognosis of GBM were investigated using comprehensive bioinformatics methods, such as survival analysis, enrichment analysis and hierarchical clustering. Results from the present study provided novel insights into the prognosis of GBM and may aid in the development of novel therapeutic approaches.

Materials and methods

Data source. The GSE7696 microarray data set, based on the GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array (Thermo Fisher Scientific, Inc.) platform, along with the corresponding clinical information was downloaded from the Gene Expression Omnibus database (www.ncbi.nlm.nih.gov/geo). The data set included 10 recurrent GBM tissue samples (2 females and 8 males; mean age, 51.31 years), 70 primary GBM tissue samples (19 females and 51 males; mean age, 48.07 years) and 4 normal brain tissue samples (sex and age information not available). All patients participated in a phase II or randomized phase III trial (21,22), with informed consent provided. The GSE7696 data set is available and the study was approved by the local ethics committee (19).

Selection of probes with expression changes in primary GBM tissue samples. The normalized data of GSE7696 were obtained, and the primary GBM tissue samples data were selected to use in the present study. Firstly, the unloaded probes were removed. Subsequently, the probes with varied expressions among 70 different patients with primary GBM were identified by the following steps. Firstly, the variance in probe expression level in each sample was calculated, and probes with variance <20% of the total probe variance were excluded. Secondly, the median probe expression in each sample was calculated and, as in the prior step, probes with a value <20% of the total median probe expression were removed. Finally, the probes with expression changes among these primary GBM tissue samples were selected.

Survival analysis and pathway enrichment analysis. Using the ‘survival’ package (23) in R (www.r-project.org), univariate survival analysis was conducted for the above probes with expression changes, and those with P<0.05 were considered as seed genes. The Kyoto Encyclopedia of Genes and Genomes (KEGG; www.genome.ad.jp/kegg) is a database used to search for gene functions, connecting genomic with functional information (24). Using the DAVID database (david.abcc.niccrf.gov) (25), KEGG pathway enrichment analysis was performed for the seed genes to explore their functions. The categories with Pc<0.05 were considered to indicate a statistically significant difference.

Screening of prognostic feature genes, unsupervised hierarchical clustering and analysis of prognostic characteristics. Using the ‘rbsurv’ package in R (bioconductor.org/packages/release/bioc/html/rbsurv.html) (26), robust likelihood-based survival modeling was constructed to identify prognostic feature genes (27). The samples were classified based on the expression profiles of the prognostic feature genes using unsupervised hierarchical clustering (28). Thereafter, the prognostic differences among the classified samples were analyzed by the Kaplan-Meier survival analysis (29). The expression differences of the feature genes between primary GBM and normal samples were analyzed; and the scatter plot of the gene expression levels were drawn using the corrplot package (https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html) in R (version 3.4.4).

Functional and pathway enrichment analysis for the feature genes. The Gene Ontology (GO) database (www.geneontology.org) is used to predict potential functions of genes and their products (30). The prognostic feature genes were examined by GO functional and KEGG pathway enrichment analysis by the ‘clusterProfiler’ package in R (bioconductor.org/packages/release/bioc/html/clusterProfiler.html), with the threshold q-value <0.05 (31).

Multivariate survival analysis. The prognostic feature genes underwent multivariate survival analysis to check their overall influences on prognosis. In addition, the ‘survivalROC’ package in R (cran.r-project.org/web/packages/survivalROC/index.html) was used to draw the receiver operating characteristic (ROC) curve and calculate the area under the ROC curve (AUC) (32). Specifically, the survivalROC package was used to calculate the ‘true positive rate’ and ‘false positive rate’ of each sample, and the differences between the ‘true positive rate’ and ‘false positive rate’ for each sample were subsequently calculated. The sample with the smallest difference value was set as the cut-off in the multifactorial cox regression analysis, and samples with a higher value than this cut-off were deemed as the high-risk group, while those with a lower value were the low-risk group.

Validation of the prognostic feature genes using an independent data set. The GBM dataset in The Cancer Genome Atlas (TCGA; cancergenome.nih.gov) database (downloaded in January 27, 2015; based on the Illumina HiSeq platform, and included RNA-sequencing data in level 3 and clinical follow-up data) was obtained to validate the prognostic feature genes. The 172 samples in the data set comprised 13 recurrent GBM samples, 154 primary GBM samples and 5 adjacent normal tissue samples. Of the 154 primary GBM samples, 2 samples had no survival information. Finally, 152 primary GBM samples were selected for analysis. The data were transformed using log2(x+1), followed by Cox regression analysis using the ‘survival’ package in R to compare the differences of prognosis and recurrence among the samples in different
groups. Differences in the identified factors (age and sex distribution) were assessed using R software. The Mann-Whitney test analysis was performed using the respective internal function in R, ‘wilcox.test’. The $\chi^2$ value was calculated using the ‘chisq.test’ function in R.

Results

Survival analysis and pathway enrichment analysis for seed genes. The bioinformatics analysis process used in the present study is outlined in Fig. 1. Following calculation of the variance and median for the probes in the 70 primary GBM, a total of 38,370 probes with expression changes were identified. Based on the univariate survival analysis using the ‘survival’ package, a total of 1,785 gene probes significantly correlated with prognosis of GBM were selected as seed genes. KEGG pathway enrichment analysis identified six pathways that were significantly enriched for these seed genes, including ‘ribosome’, ‘regulation of actin cytoskeleton’, ‘endometrial cancer’, ‘pathways in cancer’, ‘amino sugar and nucleotide sugar metabolism’ and ‘non-small cell lung cancer’ (Table I).

Screening of prognostic feature genes, unsupervised hierarchical clustering and analysis of prognostic characteristics. Robust likelihood-based survival modeling identified 13 prognostic feature genes from the 1,785 seed genes, including collagen type XXVIII α1 chain (COL28A1), PDS5 cohesin-associated factor A (PDS5A; also known as sister chromatid cohesion protein 112), zinc-finger DHHC-type containing 2 (ZDHHC2), zinc-finger protein 24 (ZNF24), myosin VA (MYO5A) and myeloid/lymphoid or mixed-lineage leukemia translocated to 4 (MLLT4) (Table II). Unsupervised hierarchical clustering was conducted for these 13 prognostic feature genes. A heat map demonstrated that these genes may be used to classify the 70 primary GBM samples into two clusters: Cluster 1 contained 51 samples and Cluster 2 contained 19 samples (Fig. 2). Subsequently, Kaplan-Meier survival analysis was used to analyze the prognostic differences between the two clusters. Patients in the Cluster 1 and in Cluster 2 had significant differences in their prognosis, which indicated that the 13 prognostic feature genes may effectively differentiate between high-risk and low-risk patients in a clinical setting (Fig. 3A). In addition, the expression levels of the 13 prognostic feature genes were relatively low, which suggested that they had low redundancy (Fig. 3B).

Functional annotation of prognostic feature genes. Using the ‘clusterProfiler’ package, functional annotation of the prognostic feature genes were conducted. The results identified three genes, including, natriuretic peptide receptor...
Multivariate survival analysis. To check the overall influence of the prognostic feature genes on prognosis, multivariate survival analysis was carried out. These genes were classified as having a positive effect on prognosis (area under the ROC curve=0.826) for the 70 primary GBM samples (Fig. 4A). Furthermore, Kaplan-Meier survival analysis demonstrated that the samples in the high-risk and low-risk groups differed significantly in their prognosis (P=5.67x10^-02; Fig. 4B), which indicated that the 13 prognostic feature genes may effectively distinguish samples with a different prognostic risk.

Validation of the prognostic feature genes using another independent data set. The GBM data set was downloaded from the TCGA database and used to validate the prognostic feature genes. Survival analysis demonstrated that the 13 prognostic feature genes had good classification effects on samples in the validation data set (area under the ROC curve=0.886) (Fig. 5A). The high-risk and low-risk groups differed significantly in their prognosis (P=5.00x10^-02; Fig. 5B). Furthermore, age distribution analysis of the classified primary GBM samples demonstrated that samples in the high-risk and low-risk groups differed significantly in age, with the high-risk group being older than the low-risk group (61.54 vs. 56.96; P=0.04396; Table III). There was no significant difference concerning sex distribution of the classified samples.

In the present study, a total of 1,785 gene probes with significant prognostic differences were selected as seed genes. From these, 13 prognostic feature genes, including COL28A1, PD55A, ZDHHC2, MYOSA, MLLT4, NPR3, ANKH and OTUB1, were further screened. The prognostic feature genes performed well for the classification of samples into different prognostic risk categories. Additionally, the GBM data set downloaded from the TCGA database further confirmed the association of the 13 prognostic feature genes with the prognosis of GBM.

Collagen XVI may promote the invasion of glioma cells by damaging cell-cell interactions or mediating the β1-integrin activation pattern, which provides new approaches for treating cancer in neuro-oncology (33). Fibrillar collagens and the collagen internalization receptor, endocytic receptor 180 (Endo180), are overexpressed in GBM (34). Additionally, Endo180 affects the invasion and progression of tumors (34). Collagen type I (COL1A1), collagen type III (COL3A1), and collagen type IV (COL4A1) were reported to serve a suppressive biological role in the progression of glioma, suggesting that it may be applied to treating the disease (35). The COL28A1 protein is a filament-forming collagen and is detected in the adult sciatic nerve (36). In our study, COL28A1 was enriched in ‘collagen biosynthesis and modifying enzymes’ pathway. Therefore, COL28A1 may serve a role in the progression of GBM.

*PD55A* is tissue-specific and has two-fold effects in tumorigenesis, acting as a tumor suppressor or an oncogenic activator.
Figure 2. Clustering heatmap for the 13 prognostic feature genes. The horizontal axis represents the samples, and the vertical axis represents the genes.

Figure 3. Kaplan-Meier survival curve and expression correlations analysis for the 13 prognostic feature genes. (A) The survival curve of the two clusters. (B) The lower-left part is the scatter plot of the gene expression levels. The red to blue color in the upper-right part represent correlation coefficients ranging from -1 to +1, respectively. The diagonal line represents the expression distribution histogram of each gene. Asterisks indicate a correlation coefficient ≥0.45.

Figure 4. Multivariate survival analysis for the 13 prognostic feature genes in the 70 primary GBM samples in GSE7696. (A) The AUC of multivariate survival analysis. (B) The Kaplan-Meier survival curve. AUC, area under the Receiver Operating Characteristic curve.
factor to promote tumor proliferation (37). *PDS5A* was previously reported to be overexpressed in high-grade gliomas and is positively associated with the World Health Organization grade of gliomas (37). Based on our enrichment result, *PDS5A* was involved in ‘separation of sister chromatids’ pathway, indicating it might influence the tumor’s DNA synthesis phase in GBM development.

*ZDHHC2* on chromosome 8p21.3-22 belongs to the DHHC-domain protein family of protein acyltransferases (38). As the largest transcription factor family in humans, zinc-finger proteins serve roles in multiple biological processes, including autophagy, metabolism, development and differentiation, and in cancer progression (39). In addition, *ZDHHC2* is associated with prognosis in a number of types of cancer. In gastric cancer, decreased *ZDHHC2* expression is associated with lymph node metastasis and indicates a poor prognosis (40). In the present study, *ZDHHC2* was highly associated with the ‘surfactant metabolism’ pathway and identified as a prognostic gene in GBM, suggesting that it might have an important role in GBM prognosis.

*ZNF24* plays a critical role in brain development and various cancer types. Furthermore, *ZNF24* may contribute to cell cycle promotion and maintenance of the progenitor stage of neural cells (41). Platelet-derived growth factor receptor (PDGFR) signaling is a crucial mechanism for the initiation and development of GBM (42). In GBM cell lines, *ZNF24* negatively regulates two transcription factors, vascular endothelial growth factor (VEGF) and PDGFR-β (43).

In the present study, the above three genes were among the 13 crucial prognostic genes of GBM. The collective findings implicate *PDS5A*, *ZDHHC2* and *ZNF24* are involved in the pathogenesis of GBM. Although no studies have reported the potential roles of *ZDHHC2* in GBM, this gene may also be associated with the prognosis of GBM, based on our results.

Myosin II is required for the invasion of glioma cells and is a promising target for the anti-invasive treatment of malignant brain tumors (44,45). In 1321N1 GBM cells, *MYO1C* is crucial for lamellipodia formation to produce a protein complex promoting cell migration (46). *MYO6* was implicated in human glioma and its inhibition may be a promising therapeutic method for the disease (47). The activation of myosin-associated contractility sensitizes GBM tumor-initiating cells, which subsequently weakens the invasive ability of the cells (48). In the present study, *MYO5A* was enriched in ‘regulation of actin dynamics for phagocytic cup formation’. Thus, *MYO5A* may serve a role in the progression of GBM.

**MLLT4** is one of the aliases of the gene, afadin, adherens junction formation factor, which belongs to an adhesion system and encodes a protein participating in cell junctions during embryogenesis. Neurofibromin 1 (*NF1*) mutations have been identified in several cancer types, including GBM (49). Furthermore, in a mouse model of *NF1*-associated optic pathway gliomas (OPG), cyclic adenosine 3’,5’-monophosphate (cAMP) inhibits the growth of OPG (50). Notably, **MLLT4** is one of the *NF1*-regulated effectors downstream of RAS (51). Results from the present studies support the possible

---

**Table III. Age distribution of the samples in the high-risk and low-risk groups.**

<table>
<thead>
<tr>
<th>Sample group</th>
<th>Min age</th>
<th>Mean age</th>
<th>Max age</th>
<th>Mann-Whitney test</th>
<th>Female</th>
<th>Male</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>High risk (n=95)</td>
<td>21</td>
<td>61.54</td>
<td>89</td>
<td>P=0.04396</td>
<td>32</td>
<td>64</td>
<td>0.51</td>
</tr>
<tr>
<td>Low risk (n=57)</td>
<td>21</td>
<td>56.96</td>
<td>85</td>
<td></td>
<td>22</td>
<td>35</td>
<td></td>
</tr>
</tbody>
</table>

---

**Figure 5.** Multivariate survival analysis for the prognostic feature genes in the 152 primary GBM samples in TCGA (validation data set). (A) The AUC multivariate survival analysis. (B) The Kaplan-Meier survival curve. AUC, area under the Receiver Operating Characteristic curve.
importance of MLLT4 in GBM pathogenesis through the participation in the adherens junction, cAMP signaling and Ras signaling pathways. In addition, a number of pathways, such as leukocyte transendothelial migration, are enriched in GBM-derived extracellular vesicles (EVs), which suggested that GBM cells may exhibit mechanisms to selectively combine these proteins in EVs (52). In the present study, leukocyte transendothelial migration was one of the six enriched pathways identified for MLLT3; therefore, this pathway may also be an important regulation target in GBM development. With regard to the tight junction pathway, junctional adhesion molecule (JAM) is one family of the immunoglobulin-like superfamily expressed in tight junctions, and abnormal JAM-A expression was reported to contribute to the progression of GBM (53). This may be evidence of the gene expression alterations in the tight junction pathway in the development of GBM.

NPR3 encodes a natriuretic peptide receptor, which regulates metabolic processes. A long noncoding RNA (lncRNA), BCYRN1, has an oncogenic role in colorectal cancer cells by upregulating NPR3 expression (54). In clear cell renal cell carcinoma (ccRCC), the lncRNA MRCCAT1 has been observed to promote ccRCC metastasis by inhibiting NPR3 (55). ANKH is identified as a novel putative oncogene in small cell lung cancer cell lines (56). In cervical cancer, ANKH is also suggested as an oncogene and the upregulation is validated by reverse transcription-quantitative polymerase chain reaction (57). To the best of our knowledge, there have been no studies reporting an association between these two genes, NPR3 and ANKH, and glioma. However, in the present study, these were important prognostic genes of GBM and were enriched in crucial pathways, suggesting that they may be novel markers in GBM and correlate with prognosis.

The OTUB1 protein is a deubiquitinating enzyme. In glioma tissues, expression of OTUB1 is increased and the expression level is associated with the glioma grade; on the other hand, knockdown of OTUB1 suppresses the tumor cell migration (58). This finding suggests that OTUB1 serves an important role in the etiology of glioma (59).

However, all these predictive results need to be further validated by in vitro and in vivo experiments. Notably, patients in the high-risk group were slightly older compared with patients in the low-risk group; although the difference was minor, it was significant. A previous study identified age as a risk factor for the process of dexamethasone-induced leukocyte cytosis, which is associated with poor survival in the newly diagnosed GBM (59). Therefore, the present study suggested that age may be a risk factor of GBM, but this needs to be validated using larger data sets.

In conclusion, the 13-gene set was tested and verified, and very efficiently predicted the prognosis of GBM in independent data sets. In addition, COL28A1, PDS5A, ZDHHC2, ZNF24, MYOS5A and MLLT4 were implicated as key genes involved in the prognosis of GBM. The adherens junction, cAMP and Ras signaling pathways may be important in the progression of GBM, and age may be a risk factor for prognosis.

Acknowledgements

Not applicable.

Funding

No funding was received.

Availability of data and materials

The analyzed data sets generated during the present study are available from the corresponding author on reasonable request.

Authors' contributions

HY performed data analyses and wrote the manuscript. LJ contributed significantly to the data analysis and manuscript revision. XS conceived and designed the study. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Patient consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References


This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) License.