RESEARCH ARTICLE

Pathway and Network Analysis in Glioma with the Partial Least Squares Method

Wen-Tao Gu, Shi-Xin Gu, Jia-Jun Shou*

Abstract

Gene expression profiling facilitates the understanding of biological characteristics of gliomas. Previous studies mainly used regression/variance analysis without considering various background biological and environmental factors. The aim of this study was to investigate gene expression differences between grade III and IV gliomas through partial least squares (PLS) based analysis. The expression data set was from the Gene Expression Omnibus database. PLS based analysis was performed with the R statistical software. A total of 1,378 differentially expressed genes were identified. Survival analysis identified four pathways, including Prion diseases, colorectal cancer, CAMs, and PI3K-Akt signaling, which may be related with the prognosis of the patients. Network analysis identified two hub genes, ELAVL1 and FN1, which have been reported to be related with glioma previously. Our results provide new understanding of glioma pathogenesis and prognosis with the hope to offer theoretical support for future therapeutic studies.

Keywords: Glioma - partial least squares - gene expression - pathway - survival analysis - network

Asian Pac J Cancer Prev, **15 (7)**, 3145-3149

Introduction

Glioma accounts for about 30% of brain and central nervous system tumors and 80% of malignant brain tumors (Goodenberger and Jenkins, 2012). The prognosis of glioma, especially high-grade (III–IV) glioma, is usually poor. Despite the high incidence of glioma, the etiology of this disease remains largely unknown. Capture the molecular characteristics of glioma patients may help understanding the underlying mechanism.

Recently, the development of high throughput experimental strategies facilitates the exploration of characteristics that underlie the progression of cancers. Several studies have investigated the gene expression signature in glioma patients (Ljubimova et al., 2001; Kim et al., 2002; Freije et al., 2004; Kawaguchi et al., 2012). Previous studies mainly used regression or variance analysis to identify deregulated genes which may contribute to glioma pathomechanism. However, this procedure cannot handle unaccounted array specific factors, such as various background biological and environmental factors. Partial least squares (PLS) based analysis has been proposed to be an effective procedure in solving feature-selection problem on high-dimensional small sample (Ji et al., 2011; Chakraborty et al., 2012). Compared with regression or variance analysis, PLS analysis is more sensitive. Besides, its specificity is relatively high while the small false discovery rate and false non-discovery rate are reasonably small. Previous study using this method on other complex diseases has proved its feasibility (Gao et al., 2013; Wu et al., 2013). Therefore, capture the gene expression signature in glioma patients with PLS based analysis may conduce to new understanding of the pathogenesis.

In the current study, to investigate the gene expression signature in high-grade glioma patients, we performed PLS based analysis with microarray data downloaded from the gene expression omnibus (GEO) database. KEGG Pathways or Gene Ontology items with significantly over-representation of differentially expressed genes were acquired. In addition, pathway based survival analysis was carried out to identify key biological processes which may contribute to the prognosis of the patients. Network of the proteins encoded by differentially expressed genes were also constructed to identify key molecules among the dysregulated genes.

Materials and Methods

Microarray data

Gene expression array data set (GSE4412) was downloaded from the GEO (http://www.ncbi.nlm.nih.gov/geo/) database. This series includes transcription profile from 74 patients whose tumor was diagnosed as a grade III (n=24) or IV (n=50). All RNA samples were extracted from fresh frozen tumor tissues obtained from initial surgical treatment. The data set was based on two platforms: GPL96 [HG-U133A] Affymetrix Human Genome U133A Array and GPL97 [HG-U133B] Affymetrix Human Genome U133B Array.

*Department of Neurosurgery, Huashan Hospital, Fudan University, Shanghai, China *For correspondence: shoujiajundr@gmail.com*

Identification of differentially expressed genes

CEL and simple omnibus format in text (SOFT)-formatted files of all samples were obtained from the GEO database. After quality control, raw intensity values were normalized by using Robust Multi-array Analysis (RMA) (Irizarry et al., 2003) procedure. Briefly, background noise effects and processing artifacts were firstly neutralized by using model-based background correction; all expression values were then aligned to a common scale by using quantile normalization and expression value for each probe was generated by using an iterative median polishing procedure. The log2-transformed RMA values all probes were used in subsequent PLS analysis to estimate the effect of them in grade III and IV samples. Firstly, the non-linear iterative partial least squares (NIPALS) algorithm (Martins et al., 2010) was used to calculate PLS latent variables. Secondly, variable importance in the projection (VIP) (Gosselin. et al., 2010) was used to evaluate the importance of probe expression value on the disease status of the patients. Thirdly, a permutation procedure (n=10000) was used to generate the empirical distribution of PLS-based VIP. Finally, the empirical distribution was used to calculate the False discovered rate (FDR) of each probe. The threshold of significantly differentially expressed genes was set as 0.01. Fold change was used to define up or down-regulation. All above procedures were carried out by using the R software (version 3.0.0) in which BioConductor, limma packages (3.12.1) and libraries (Smyth et al., 2005) were included.

Enrichment analysis

To capture biologically relevant characteristics of the differentially expressed genes, the selected genes were then annotated according to the KEGG pathways database (http://www.genome.jp/kegg/) (Kanehisa and Goto, 2000) and Gene Ontology (GO) database (Ashburner et al., 2000). Hyper geometric distribution test was then performed to identify pathways and GO items enriched with dysregulated genes.

Survival analysis

To investigate the contribution of the pathways, which enriched with differentially expressed genes, to the survival time after surgery, we carried out survival analysis. For each pathway, the samples were separated into two classes (class 1 and class 2) with K-mean algorithm based on the expression values of all genes in the pathway. With the survival time or last follow-up time of all patients, survival rate of the two classes were drawn. Log-rank test were then used to investigate whether the two classes were significantly different from each other. Pathways with P values less than 0.05 were considered to be significantly related with the survival rate of the patients.

Network analysis

The interactions between proteins are important in all biological processes (Stelzl et al., 2005). Proteins encoded by differentially expressed genes which have more interactions may serve as more important molecules in the molecular difference between grade III and IV patients. To identify these key genes, we constructed a network with Cytoscape (V 2.8.3, http://www.cytoscape.org/) (Shannon et al., 2003) and the NCBI (http://ftp.ncbi.nlm.nih.gov/gene/GeneRIF/) database. To investigate the contribution of the genes included in the network, survival analysis was carried out for this gene set. In the network, proteins with degrees (links in the network) over 15 were considered as hub molecules.

Results

The results revealed that a total of 1378 genes were differentially expressed in the two stages of patients, including 637 down regulated and 741 overexpressed genes. For all well-characterized genes in the expression arrays, 5983 genes were mapped to various KEGG pathways, including 575 differentially expressed genes. Table 1 represents the top 15 pathways enriched with differentially expressed genes. These pathways include six nervous system pathways, four signaling pathways, three cellular processes pathways and two cancer related pathways. One of the nervous system pathways is Prion diseases, which is a Neurodegenerative disease. A total of 15854 genes in the arrays were annotated to various GO items, including 1265 differentially expressed genes. The top 15 GO items enriched with differentially expressed genes are listed in Table 2, including two nervous system

Table 1. The Top 15 Pathways Enriched with Differentially Expressed Genes

#KEGG_id	Pathway_description	Pathway_subclass	*p* value
hsa04145	Phagosome	Transport and catabolism	1.39E-13
hsa04514	Cell adhesion molecules (CAMs)	Signaling molecules and interaction	3.66E-08
hsa04512	ECM-receptor interaction	Signaling molecules and interaction	5.00E-08
hsa04510	Focal adhesion	Cell communication	1.19E-07
hsa04727	GABAergic synapse	Nervous system	2.34E-04
hsa05205	Proteoglycans in cancer	Cancers	2.39E-03
hsa04728	Dopaminergic synapse	Nervous system	4.02E-03
hsa04723	Retrograde endocannabinoid signaling	Nervous system	4.15E-03
hsa05020	Prion diseases	Neurodegenerative diseases	1.02E-02
hsa05210	Colorectal cancer	Cancers	1.38E-02
hsa04151	PI3K-Akt signaling pathway	Signal transduction	3.23E-02
hsa04915	Estrogen signaling pathway	Endocrine system	3.44E-02
hsa04724	Glutamatergic synapse	Nervous system	3.66E-02
hsa04726	Serotonergic synapse	Nervous system	3.77E-02
hsa04142	Lysosome	Transport and catabolism	4.58E-02

Table 2. The top 15 GO Items Enriched with Differentially Expressed Genes

#GO_id	GO_description	GO_class	FDR
GO:0031012	extracellular matrix	Component	2.52E-23
GO:0005886	plasma membrane	Component	7.46E-15
GO:0030198	extracellular matrix organization	Process	2.37E-12
GO:0060333	interferon-gamma-mediated signaling pathway	Process	3.72E-11
GO:0007155	cell adhesion	Process	4.33E-10
GO:0071556	integral to lumenal side of endoplasmic reticulum membrane	Component	4.59E-08
GO:0005604	basement membrane	Component	7.39E-08
GO:0012507	ER to Golgi transport vesicle membrane	Component	2.11E-06
GO:0070062	extracellular vesicular exosome	Component	4.86E-06
GO:0030509	BMP signaling pathway	Process	4.86E-06
GO:0005515	protein binding	Function	4.86E-06
GO:0007179	transforming growth factor beta receptor signaling pathway	Process	6.18E-03
GO:0005509	calcium ion binding	Function	8.33E-03
GO:0007158	neuron cell-cell adhesion	Process	8.54E-03
GO:0043197	dendritic spine	Component	9.17E-03

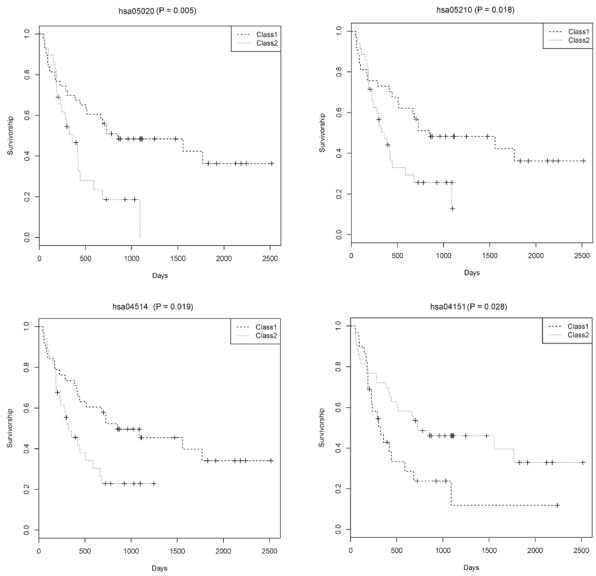

Figure 1. Survival Curve of the four Pathways which were Identified to be Related with the Survival Rate of the Patients

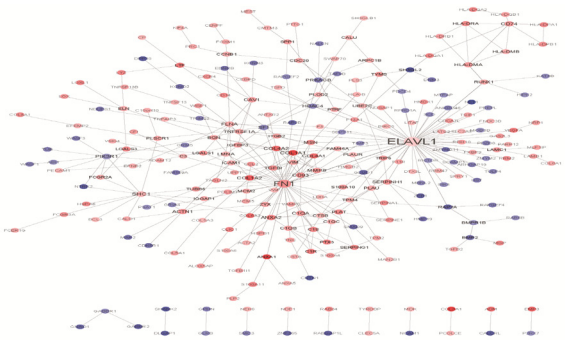

Figure 2. Interaction Network Constructed by Proteins Encoded by Dysregulated Genes. Proteins with more interactions are shown in bigger size. Proteins in red are encoded by overexpressed genes in grade IV patients while those in blue are encoded by down regulated genes in grade IV samples

related GO items: neuron cell-cell adhesion (GO:0007158) and dendritic spine (GO:0043197).

Survival analysis showed that four pathways were related with the survival rate of the patients, including the Prion diseases pathway (hsa05020, p=0.005), Colorectal cancer (hsa05210, p=0.018), Cell adhesion molecules (CAMs) (hsa04514, p=0.019), and PI3K-Akt signaling pathway (hsa04151, p=0.028). Figure 1 represents the survival curves of these pathways. Survival analysis of genes included in the network indicated their significant contributions to the survival rate of the patients (p=0.0014). Two proteins, ELAVL1 and FN1, were identified to be hub molecules with the degrees of 56 and 40, respectively (Figure 2).

Discussion

The prognosis of high-grade glioma patients is poor. Understanding the key biological pathway which contributes to the progression of the disease may help developing new therapeutic strategies. Here with gene expression data for glioma patients, we carried out pathway based analysis to identify key biological processes which may contribute to the tumor progression and survival rate.

Enrichment analysis showed that 40% pathways with overrepresentation of differentially expressed genes were involved in the nervous system. In addition, one of the GO items enriched with differentially expressed genes is the neuron cell-cell adhesion (GO: 0007158). Dysregulation of the nervous system is expected since glioma is a neurologic tumor.

Survival analysis was further carried out for the pathways enriched with differentially expressed genes. The result revealed four pathways related with the survival rate of the patients. Among them, the Prion diseases pathway (hsa05020) showed the most significant correlation. Differentially expressed genes in this pathway are illustrated in Figure 1. The relationship between differential expression of the Prion protein and glioma was reported in previous studies (Shaochun, 1997). Several genes in this pathway may also be related with this disease. For example, LAMC1 was reported to be critical in the process of cell migration and tumor invasion (Nielsen et al., 1983). Fowler et al. has proposed the possible involvement of laminin in the migration and invasion of glioblastoma (Fowler et al., 2011). Consistent with our results, LAMC1 was up regulated in grade IV patients. One of the four pathways is a signaling pathway: PI3K-

Akt signaling pathway. This pathway is involved in many cellular functions, such as transcription, translation, cell proliferation, cell growth, and cell survival. Inhibition of this pathway was suggested to be a therapeutic strategy of glioma (Malla et al., 2010; Sami and Karsy, 2013). Genes in this pathway were also suggested to be related with glioma. Take CDK2 for example, this gene was reported to be suppressed during the inhibition of glioma cells (Harmalkar and Shirsat, 2006). This is consistent with our results, since CDK2 is overexpressed in grade IV patients. One of the pathways is a cancer related pathway: Colorectal cancer (hsa05210). Since all cancers share the common character of unregulated cell growth, it is not unexpected that other cancer related pathways were detected to be related with the survival rate of the patients. The last pathway is CAMs (hsa04514) and the important roles of CAMs in glioma cell adhesion and invasion have been reported in previous studies (Goldbrunner et al., 1998).

Network analysis revealed that ELAVL1was a hub gene with the highest degree (Figure 2). Protein encoded by this gene contains several RNA recognition motifs, which selectively bind AU-rich elements in the 3' untranslated regions of mRNAs. Dysregulation of this gene may interrupt the stabilization of ARE-containing mRNAs. Previous studies reported that this gene was unregulated in tumors compared with controls (Nabors et al., 2001). In addition, it has also reported that mRNA stability alterations mediated by this protein are necessary to sustain the rapid growth of glioma cells (Bolognani et al., 2012). Our results further revealed that this gene is unregulated in grade IV compared with grade III, suggesting its implication in the deterioration of this devastating type of cancer. Further therapeutic studies may consider it as a potential target. Another hub gene is FN1. Protein encoded by this gene is fibronectin, which is involved in cell adhesion and migration processes. Microarray expression analyses have showed the significant correlation between this gene and malignant glioma before (Wei et al., 2010). Our results further supported its contribution in the progression of the disease. Further therapeutic studies on this gene are warranted.

In summary, using gene expression profile data from the GEO database, we performed PLS based analysis to identify key biological processes contributing to tumor progression and patient's survival. Biological processes, including nervous system pathways, cancers, and signal transduction process, were identified to be enriched with differentially expressed genes. Further survival analysis identified four pathways which may be related with the prognosis of the patients. Network of differentially expressed genes identifies two hub genes, ELAVL1 and FN1. Our results here may offer potential targets for future molecular diagnostic studies.

References

Ashburner M, Ball CA, Blake JA, et al (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, **25**, 25-9.

Bolognani F, Gallani AI, Sokol L, et al (2012). mRNA stability alterations mediated by HuR are necessary to sustain the fast growth of glioma cells. *J Neurooncol*, **106**, 531-42.

Chakraborty S, Datta S, Datta S (2012). Surrogate variable analysis using partial least squares (SVA-PLS) in gene expression studies. *Bioinformatics*, **28**, 799-806.

Fowler A, Thomson D, Giles K, et al (2011). miR-124a is frequently down-regulated in glioblastoma and is involved in migration and invasion. *Eur J Cancer*, **47**, 953-63.

Freije WA, Castro-Vargas FE, Fang Z, et al (2004). Gene expression profiling of gliomas strongly predicts survival. *Cancer Res*, **64**, 6503-10.

Gao QG, Li ZM, Wu KQ (2013). Partial least squares based analysis of pathways in recurrent breast cancer. *Eur Rev Med Pharmacol Sci*, **17**, 2159-65.

Goldbrunner RH, Bernstein JJ, Tonn JC (1998). ECM-mediated glioma cell invasion. *Microsc Res Tech*, **43**, 250-7.

Goodenberger ML, Jenkins RB (2012). Genetics of adult glioma. *Cancer Genet*, **205**, 613-21.

Gosselin R, Rodrigue D, Duchesne C (2010). A Bootstrap-VIP approach for selecting wavelength intervals in spectral imaging applications. *Chem Intel Lab Sys*, **100**, 12-21.

Harmalkar MN, Shirsat NV (2006). Staurosporine-induced growth inhibition of glioma cells is accompanied by altered expression of cyclins, CDKs and CDK inhibitors. *Neurochem Res*, **31**, 685-92.

Irizarry RA, Hobbs B, Collin F, et al (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249-64.

Ji G, Yang Z, You W (2011). PLS-Based Gene Selection and Identification of Tumor-Specific Genes. Ieee Transactions On Systems, Man, And Cybernetics-Part C: *Appl Rev*, **41**, 830-41.

Kanehisa M, Goto S (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, **28**, 27-30.

Kawaguchi A, Yajima N, Komohara Y, et al (2012). Identification and validation of a gene expression signature that predicts outcome in malignant glioma patients. *Int J Oncol*, **40**, 721-30.

Kim S, Dougherty ER, Shmulevich I, et al (2002). Identification of combination gene sets for glioma classification. *Mol Cancer Ther*, **1**, 1229-36.

Ljubimova JY, Lakhter AJ, Loksh A, et al (2001). Overexpression of alpha4 chain-containing laminins in human glial tumors identified by gene microarray analysis. *Cancer Res*, **61**, 5601-10.

Malla R, Gopinath S, Alapati K, et al (2010). Downregulation of uPAR and cathepsin B induces apoptosis via regulation of Bcl-2 and Bax and inhibition of the PI3K/Akt pathway in gliomas. *PLoS One*, **5**, 13731.

Martins JPA, Teofilo RF, Ferreira MMC (2010). Computational performance and cross-validation error precision of five PLS algorithms using designed and real data sets. *J Chemomet*, **24**, 320-32.

Nabors LB, Gillespie GY, Harkins L, King PH (2001). HuR, a RNA stability factor, is expressed in malignant brain tumors and binds to adenine- and uridine-rich elements within the 3' untranslated regions of cytokine and angiogenic factor mRNAs. *Cancer Res*, **61**, 2154-61.

Nielsen M, Christensen L, Albrechtsen R (1983). The basement membrane component laminin in breast carcinomas and axillary lymph node metastases. *Acta Pathol Microbiol Immunol Scand A*, **91**, 257-64.

Sami A, Karsy M (2013). Targeting the PI3K/AKT/mTOR signaling pathway in glioblastoma: novel therapeutic agents and advances in understanding. *Tumour Biol*, **34**, 1991-2002.

Shannon P, Markiel A, Ozier O, et al (2003). Cytoscape: a software environment for integrated models of biomolecular

interaction networks. *Genome Res*, **13**, 2498-504.
Shaochun YXWKL (1997). A Study of Prion Protein Expression in Gliomas. Acta universitatis medictnae tangji, 3.
Smyth GK, Michaud J, Scott HS (2005). Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics*, **21**, 2067-75.
Stelzl U, Worm U, Lalowski M, et al (2005). A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, **122**, 957-68.
Wei KC, Huang CY, Chen PY, et al (2010). Evaluation of the prognostic value of CD44 in glioblastoma multiforme. *Anticancer Res*, **30**, 253-9.
Wu S, Zhang X, Li ZM, et al (2013). Partial least squares based gene expression analysis in EBV- positive and EBV-negative posttransplant lymphoproliferative disorders. *Asian Pac J Cancer Prev,* **14**, 6347-50.