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

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

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 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...
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 ELAVL1 was 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 HuR are necessary to sustain the fast growth of glioma cells. J Neurooncol, 106, 531-42.


