RESEARCH ARTICLE

Partial Least Squares Based Gene Expression Analysis in EBV-Positive and EBV-Negative Posttransplant Lymphoproliferative Disorders

Sa Wu*, Xin Zhang*, Zhi-Ming Li, Yan-Xia Shi, Jia-Jia Huang, Yi Xia, Hang Yang, Wen-Qi Jiang*

Abstract

Post-transplant lymphoproliferative disorder (PTLD) is a common complication of therapeutic immunosuppression after organ transplantation. Gene expression profile facilitates the identification of biological difference between Epstein-Barr virus (EBV) positive and negative PTLDs. Previous studies mainly implemented variance/regression analysis without considering unaccounted array specific factors. The aim of this study is to investigate the gene expression difference between EBV positive and negative PTLDs through partial least squares (PLS) based analysis. With a microarray data set from the Gene Expression Omnibus database, we performed PLS based analysis. We acquired 1188 differentially expressed genes. Pathway and Gene Ontology enrichment analysis identified significantly over-representation of dysregulated genes in immune response and cancer related biological processes. Network analysis identified three hub genes with degrees higher than 15, including CREBBP, ATXN1, and PML. Proteins encoded by CREBBP and PML have been reported to be interact with EBV before. Our findings shed light on expression distinction of EBV positive and negative PTLDs with the hope to offer theoretical support for future therapeutic study.

Keywords: Post-transplant lymphoproliferative disorder - Epstein-Barr virus - partial least squares

Asian Pac J Cancer Prev, 14 (11), 6347-6350

Introduction

Post-transplant lymphoproliferative disorder (PTLD) is a common complication of therapeutic immunosuppression after organ transplantation. The majority of PTLD is of B-cell origin and over 90% of PTLDs are Epstein-Barr virus (EBV) positive (Paya et al., 1999; Gottschalk et al., 2005). Clinically, EBV negative patients tend to occur later and have an overall poorer prognosis(Nelson et al., 2000). Currently, the pathogenesis of EBV-negative PTLD is less defined. Capture the molecular characteristics of EBV negative patients may help understanding the underlying mechanism.

Recently, the innovation of high throughput experimental strategies facilitates the identification of signatures that underlie the pathogenesis of complex diseases. Several studies have investigated the gene expression difference between EBV-positive and EBV-negative patients using microarray analysis (Craig et al., 2007; Morscio et al., 2013; Shi et al., 2013). These studies implemented variance or regression analysis to identify differentially expressed genes. This procedure becomes fundamentally flawed when there are unaccounted array specific factors, such as different biological, environmental or other factors relevant in the context. Previous studies (Ji et al., 2011; Chakraborty et al., 2012) has proposed that partial least squares (PLS) based gene expression analysis is effective in solve feature-selection problem on high-dimensional small sample. Compared with variance/regression analysis, PLS analysis is more sensitive while maintaining high specificity, small false discovery rate and false non-discovery rate. Characterize the gene expression difference between EBV-positive and EBV-negative PTLD patients with PLS based analysis may conduce to new understanding of the pathogenesis and further facilitate the development of novel therapeutic strategies.

In the current study, to investigate the gene expression difference between EBV positive and EBV negative PTLD patients, we performed PLS-based analysis with a microarray data set downloaded from the gene expression omnibus (GEO) database. Pathways or Gene Ontology items with significantly over-represented differentially expressed genes were also acquired. In addition, a protein-protein interaction (PPI) network for the proteins encoded by differentially expressed genes was constructed to identify key molecules related with the gene expression difference.
Materials and Methods

Microarray data
The microarray data set GSE38885 from the GEO (http://www.ncbi.nlm.nih.gov/geo/) database was used in this study. This series represents transcription profile of 65 malignant post-transplant lymphomas, including 31 EBV positive and 34 EBV negative samples. All samples were taken from frozen tumor tissues. The data set was based on platform GPL570: [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array.

Identification of differentially expressed genes (DEGs)
Raw data for all samples were obtained from the GEO database. Normalization of raw intensity values was carried out using Robust Multi-array Analysis (RMA) (Irizarry et al., 2003). The generated log2-transformed expression value of each probe was used in subsequent PLS analysis to estimate the effect of each probe in EBV positive and negative samples. Briefly, PLS latent variables were firstly calculated by using the non-linear iterative partial least squares (NIPALS) algorithm (Martins et al., 2010). Then the importance of probe expression on the status of the subjects was evaluated according to the variable importance in the projection (VIP) (Gosselin et al., 2010). Finally, the empirical distribution of PLS-based VIP was got by using a permutation procedure (n=10000). False discovered rate (FDR) of each probe was then calculated according to the empirical distribution. Differentially expressed probes, which were subject to further analysis, were defined as those with FDR less than 0.01.

Enrichment analysis
All probes were annotated according to the downloaded simple omnibus format in text (SOFT) format files. To capture biologically relevant signature of the differentially expressed genes, we carried out enrichment analysis. All genes were further mapped to the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways database (http://www.genome.jp/kegg/) (Kanehisa and Goto, 2000) and Gene Ontology database (Ashburner et al., 2000). Hyper geometric distribution test was then carried out to identify biological processes significantly enriched with differentially expressed genes.

Network analysis
Protein-protein interaction (PPI) is crucial for all biological processes (Stelzl et al., 2005). Differentially expressed genes with more interactions with others may play more important roles in the biological difference of EBV positive and EBV negative samples. To visualize the interaction among these genes and identify key genes, a network was constructed with the software Cytoscape (V 2.8.3, http://www.cytoscape.org/) (Shannon et al., 2003) and the NCBI database (http://ftp.ncbi.nlm.nih.gov/gene/GeneRIF/) database. The degree of each protein is its number of links (interactions). Those with degrees more than 15 were considered as hub molecules in this study.

Results
PLS analysis revealed that 1188 genes were differentially expressed between EBV positive and EBV negative samples. For all well-characterized genes in the array, 6072 genes can be mapped to various pathways while 403 differentially expressed genes were mapped to KEGG pathways. The top ten pathways enriched with differentially expressed genes are listed in Table 1. These pathways mainly involve immune processes, such as immune system, immune diseases and infectious
diseases. In addition, two cancer pathways, transcriptional misregulation in cancers (hsa05202) and p53 signaling pathway (hsa04115) were also enriched with differentially expressed genes. Of all genes in the array, 16635 genes were annotated based on the GO database, including 1049 selected genes. Table 2 represents the top 10 GO items enriched with selected genes. Defense response to virus (GO: 0051607) was the most significant GO item with over represented selected genes. A cancer related item, DNA damage response, signal transduction by p53 class mediator resulting in cell cycle arrest (GO: 0006977) was also detected to be enriched with differentially expressed genes.

Interaction network of proteins encoded by differentially expressed genes is illustrated in Figure 1. Three proteins, CREBBP, ATXN1, and PML were identified to be hub molecules, with degrees of 23, 16, and 15 respectively.

**Discussion**

Microarray technology has offered great ease for investigating the gene expression difference between EBV-positive and EBV-negative PTLDs. However, creating an effective mathematical model to deal with the small sample and large number of genes is challenging. Previous gene expression studies mainly implemented variance or regression analysis, which cannot deal with unaccounted array specific factors. Here, we used a PLS based model to identify differentially expressed genes EBV-positive and EBV-negative PTLDs.

Pathway and GO item enrichment analysis revealed that biological processes related with the immune system, such as response to the virus and antigen processing, were over-represented with selected genes. This is generally consistent with previous studies, since both immune and EBV status have been described as main discriminating factors (Craig et al., 2007; Morscio et al., 2013). In addition, we also cancer related pathways, such as transcriptional misregulation in cancers (hsa05202) and p53 signaling pathway (hsa04115) to be enriched with differentially expressed genes. A cancer related GO item (GO:0006977) was also identified to be over-represented with selected genes. These identified biological signatures may contribute to the clinical differences between the two groups.

To identify key molecules among the differentially expressed genes, we carried out network analysis. The results revealed that CREBBP was a hub gene with the
highest degree (Figure 1). Protein encoded by this gene is involved in the transcriptional coactivation of many different transcription factors. Previous study (Adamson and Kenney, 1999) showed that CREBBP could bind to the EBV immediate early protein BZLF1-mediated and enhance viral early gene transcription. EBV nuclear protein 2 also interacts with CREBBP in activation of the virus oncogene LMP1 (Wang et al., 2000). Therefore, CREBBP may play important roles in the clinical difference between the EBV positive and EBV negative patients. ATXN1 was also identified as a hub gene with the second highest degree (Figure 1). No previous report of the relationship between this gene and EBV or lymphoma has been proposed up to date. However, polymorphisms with prognostic significance of ATXN1 have been identified in familial and sporadic chronic lymphocytic leukemia (Auer et al., 2007). Therefore, the potential roles of this gene warrant further investigation. PML was another hub gene with degree more than 15. Protein encoded by this gene is a member of the tripartite motif (TRIM) family. PML nuclear body protein was shown to physically and functionally interact with the EBV protein SM, increasing the stability of lytic EBV transcripts (Nicewonger et al., 2004). Thus, this gene may involve in the molecular mechanism of EBV positive lymphoma, leading to distinct clinical manifestations of EBV positive and negative patients.

In summary, with microarray data set downloaded from the GEO database, we carried out PLS based analysis to identify differentially expressed genes in EBV positive and negative PTLD patients. Enrichment analysis was also carried out the capture biological relevant signatures. A network of differentially expressed genes was constructed to identify key hub genes. Our results facilitate the disclosure of the molecular mechanism underlying the distinct clinical manifestations of EBV positive and negative patients.

Acknowledgements

This study was supported by the special funds for the development of strategic emerging industries in Shenzhen, China (JCYJ2012061311328732). The authors have no financial conflicts of interest.

References

Shi M, Gan YJ, Davis TO, Scott RS (2013). Downregulation of the polyamine regulator spermidine/spermine N-acetyltransferase by Epstein-Barr virus in a Burkitt’s lymphoma cell line. Virus Res.