GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

# Partial least squares-based gene expression analysis in preeclampsia

**F. Jiang[1]*, Y. Yang[1]*, J. Li[2], W. Li[3], Y. Luo[1], Y. Li[1], H. Zhao[1], X. Wang[1], G. Yin[1] and G. Wu[4]**

[1]Department of Gynecology and Obstetrics,
Tangdu Hospital, The Fourth Military Medical University, Xi'an, China
[2]Department of Gastrointestinal Surgery, Xijing Hospital,
The Fourth Military Medical University, Xi'an, China
[3]Department of thoracic surgery, Tangdu Hospital,
The Fourth Military Medical University, Xi'an, China
[4]Departmentof Urology, Xijing Hospital, The Fourth Military Medical University,
Xi'an, China

*These authors contributed equally to this study.
Corresponding authors: F. Jiang / G. Wu
E-mail: fjiang@fmmu.edu.cn / gunwu@sina.com

Genet. Mol. Res. 14 (2): 6598-6604 (2015)
Received January 27, 2015
Accepted May 16, 2015
Published June 18, 2015
DOI http://dx.doi.org/10.4238/2015.June.18.2

**ABSTRACT.** Preeclampsia is major cause of maternal and fetal morbidity and mortality. Currently, the etiology of preeclampsia is unclear. In this study, we investigated differences in gene expression between preeclampsia patients and controls using partial least squares-based analysis, which is more suitable than routine analysis. Expression profile data were downloaded from the Gene Expression Omnibus database. A total of 503 genes were found to be differentially expressed, including 248 downregulated genes and 255 overexpressed genes. Network analysis identified 5 hub genes, including *UBB*, *PIK3R1*, *MAPRE1*, *VEGFA*, and *ITGB1*. Three of these, *PIK3R1*, *VEGFA*, and *ITGB1,* are known to be associated with preeclampsia or preeclampsia-

 

related biological processes. Our findings shed light on expression signatures of preeclampsia patients that can be used as theoretical support in future therapeutic studies.

**Key words:** Gene expression; Partial least squares; Preeclampsia

## INTRODUCTION

Preeclampsia, which is a major cause of maternal and fetal morbidity and mortality, is a pregnancy-specific medical condition characterized by high blood pressure and proteinuria (Sibai et al., 2005). Currently, the etiology of preeclampsia is uncertain. Understanding the underlying molecular mechanisms of this disease would be helpful for further therapeutic studies.

Recently, the development of high-throughput gene expression profiles has greatly facilitated the identification of molecular signatures underlying the pathogenesis of complex diseases. Several gene expression studies (Lapaire et al., 2012; Louwen et al., 2012; Meng et al., 2012; Zhou et al., 2013) have been conducted to investigate expression differences between preeclampsia patients and healthy controls. The major challenge for microarray analysis is to develop a suitable statistical model to deal with the small sample number and large number of genes. Previous studies mainly implemented variance or regression analysis, omitting array-specific factors, including different biological and relevant environmental factors. Previous studies (Ji et al., 2011; Chakraborty et al., 2012) have proposed that partial least squares (PLS) analysis is effective for solving feature-selection problem with microarray data. Compared with routine analysis, the PLS method is more sensitive and shows high specificity and a low false discovery rate and false non-discovery rate. Characterizing gene expression signatures in preeclampsia patients with PLS analysis may increase the understandings of its pathogenesis.

In this study, we used PLS analysis to investigate gene expression differences between preeclampsia patients and healthy controls. Gene expression data were downloaded from the Gene Expression Omnibus (GEO) database. To identify the biologically relevant signatures, pathways and Gene Ontology (GO) items significantly overrepresented with dysregulated genes were determined. We also constructed an interaction network with proteins encoded by selected genes to identify key molecules related to gene expression differences. Our results will increase the understanding of the pathogenesis of preeclampsia.

## MATERIAL AND METHODS

### Microarray data

The expression profile GSE35574 used in this study was downloaded from the GEO (http://www.ncbi.nlm.nih.gov/geo/) database. This series represents the transcription profile of 19 preeclamptic and 40 control placentas. The data set was based on platform GPL6102: Illumina human-6 v2.0 expression beadchip.

### Detection of differentially expressed genes

Raw data for all 59 samples were downloaded from the GEO database and were normalized using robust multi-array analysis (Irizarry et al., 2003). PLS analysis was then carried

out to estimate the effect of each probe in preeclamptic and control samples as follows: first, a non-linear iterative PLS algorithm (Martins et al., 2010) was used to calculate PLS latent variables; second, the number of latent variables was determined using a 4-fold cross validation (30 times) to avoid model over-fitting; third, the variable importance in the projection (Gosselin et al., 2010) was used to assess the effect of probe expression on the disease status of the subjects. Finally, the false discovery rate of all probes was calculated according to the empirical distribution of PLS-based variable importance in the projection, which were obtained using a permutation procedure (N = 10000). Probes with a false discovery rate less than 0.01 were considered to be differentially expressed and were used for further analysis.

## Enrichment analysis

All probes were annotated based on the simple omnibus format in text format files. To capture biologically relevant signatures of the selected genes, we carried out Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway (Kanehisa and Goto, 2000) and GO (Ashburner et al., 2000) enrichment analysis. All genes were mapped to both databases, and the hypergeometric distribution test was used to identify biological processes that are overrepresented for the genes selected.

## Network analysis

Protein-protein interactions are important for all biological processes (Stelzl et al., 2005). Proteins encoded by genes selected with a large number of interactions may play more important roles in the biological differences between patients and healthy samples. To identify key molecules, network analysis was carried out using the Cytoscape software (V 2.8.3, http://www.cytoscape.org/) (Shannon et al., 2003) and the National Center for Biotechnology Information (NCBI) database (http://ftp.ncbi.nlm.nih.gov/gene/GeneRIF/). Proteins with a degree (number of interactions) of more than 10 were considered to be hub molecules.

# RESULTS

According to the cross-validation results, we selected 10 PLS latent variables that resulted in the highest classification accuracy (Figure 1). A total of 503 genes were identified to be differentially expressed, including 248 downregulated genes and 255 overexpressed genes. Pathway enrichment analysis revealed 8 pathways overrepresented with differentially expressed genes (Table 1), including 1 signal transduction pathway, the vascular endothelial growth factor (VEGF) signaling pathway. Other pathways mainly involved cellular processes and genetic information processing processes. GO analysis identified 10 items enriched with genes selected (Table 2). Consistent with pathway analysis, items involved in the cellular process were also identified. In addition to the pathway results, items related to the cell cycle were found to be overrepresented by differentially expressed genes.

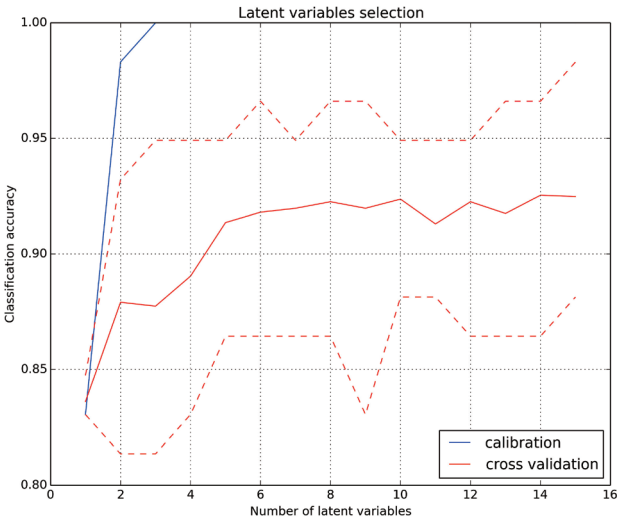


**Figure 1.** Latent variable selection. According to the cross-validation results, we selected 10 PLS latent variables that resulted in the highest classification accuracy.

**Table 1.** Pathways enriched with differentially expressed genes.

| KEGG_id | Pathway description | Pathway subclass | P value |
|---|---|---|---|
| hsa04142 | Lysosome | Transport and catabolism | 6.70E-04 |
| hsa04510 | Focal adhesion | Cell communication | 7.71E-03 |
| hsa04145 | Phagosome | Transport and catabolism | 9.86E-03 |
| hsa04520 | Adherens junction | Cell communication | 1.27E-02 |
| hsa03420 | Nucleotide excision repair | Replication and repair | 1.72E-02 |
| hsa04141 | Protein processing in endoplasmic reticulum | Folding, sorting and degradation | 1.73E-02 |
| hsa04810 | Regulation of actin cytoskeleton | Cell motility | 2.08E-02 |
| hsa04370 | VEGF signaling pathway | Signal transduction | 4.24E-02 |

**Table 2.** Gene Ontology (GO) items enriched with differentially expressed genes.

| GO_id | GO description | GO class | P value |
|---|---|---|---|
| GO:0044267 | Cellular protein metabolic process | Process | 1.22E-06 |
| GO:0005788 | Endoplasmic reticulum lumen | Component | 1.40E-04 |
| GO:0000278 | Mitotic cell cycle | Process | 1.42E-04 |
| GO:0005856 | Cytoskeleton | Component | 3.55E-04 |
| GO:0000086 | G2/M transition of mitotic cell cycle | Process | 9.00E-04 |
| GO:0051301 | Cell division | Process | 4.38E-03 |
| GO:0008083 | Growth factor activity | Function | 4.99E-03 |
| GO:0030036 | Actin cytoskeleton organization | Process | 5.76E-03 |
| GO:0030198 | Extracellular matrix organization | Process | 1.99E-02 |
| GO:0043065 | Positive regulation of apoptotic process | Process | 2.12E-02 |

The interaction network of proteins encoded by differentially expressed genes is illustrated in Figure 2. Five proteins were identified to be hub molecules, including UBB, PIK3R1, MAPRE1, VEGFA, and ITGB1. Three of these, PIK3R1, VEGFA, and ITGB1 are known to be associated with preeclampsia or preeclampsia-related biological processes (Table 3).

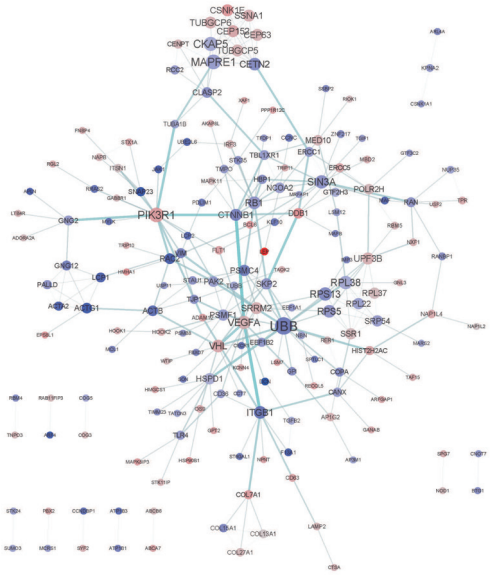


**Figure 2.** Interaction network constructed by proteins encoded by differentially expressed genes. Proteins with more links are shown in larger size. Proteins shown in red are encoded by overexpressed genes inpatients, while those in blue are encoded by downregulated genes.

**Table 3.** Hub molecules in the network constructed by differentially expressed genes.

| Gene | Implications in preeclampsia or related biological processes (PMID) | Degree |
|---|---|---|
| *UBB* | Not known association with preeclampsia or related biological processes | 16 |
| *PIK3R1* | Angiogenesis (18449193) | 12 |
| *MAPRE1* | Not known association with preeclampsia or related biological processes | 12 |
| *VEGFA* | Preeclampsia (23522390) | 10 |
| *ITGB1* | Preeclampsia (22716646) | 10 |

## DISCUSSION

Gene expression profiling is a powerful method for investigating the pathogenesis of preeclampsia. It is important to create an effective statistical model to handle the small sample and large number of genes. In this study, we used a PLS-based model, which has been reported to be sensitive and reliable, to identify dysregulated genes in preeclampsia.

Pathway and GO item enrichment analysis revealed that dysregulated biological processes were mainly related to cellular processes. This is consistent with the findings of previous studies. For example, in the lysosome pathway, excretion of lysosomal enzymes is considered to be a potential diagnostic or predictive marker in preeclamptic women (Jackson et al., 1996) as proximal tubule epithelial injury in preeclampsia patients leads to the release of lysosomal enzymes.

Network analysis, which aims to identify key genes among the differentially expressed genes, identified that *UBB* was the hub gene with the highest degree (Figure 2, Table 3). This gene encodes the ubiquitin B protein, and this is the first report of the involvement of this gene in the pathogenesis of preeclampsia. Although no previous studies have reported

a relationship between this gene and preeclampsia, dysregulation of the ubiquitin-mediated degradation process has long been considered to be associated with the pathophysiology of preeclampsia (van Dijk et al., 2010; Cayli et al., 2012). Therefore, the potential association between this gene and preeclampsia should be further investigated. *PIK3R1* is another hub gene and showed the second highest degree (Figure2, Table 3). This gene encodes the regulatory subunit of the phosphoinositide-3-kinase (PI3K). Because of the importance of PI3K signaling in endothelial cell migration (Graupera et al., 2008) and endothelial dysfunction in preeclampsia patients, dysregulation of this gene may contribute to the pathogenesis. Two other genes, *VEGFA* and *ITGB1*, have been reported to be related to preeclampsia in previous studies (Zhao et al., 2013; Li et al., 2013). Identification of *VEGFA* agrees with our pathway analysis, as the VEGF signaling pathway showed significant overrepresentation of dysregulated genes. VEGF is a positive regulator of angiogenesis and plays important roles in the development of vascular endothelial cells, the growth of blood vessels, and the progression of vessel permeability (Ferrara, 2004). A previous study showed that hypoxia-driven disruption of the angiogenic balance involving VEGF may contribute to the symptoms of preeclampsia (Levine et al., 2004). Our results confirmed the involvement of this gene in preeclampsia, suggesting that as a hub gene, it may serve as a potential therapeutic target. *ITGB1* encodes the beta subunit of integrin. Depression of the mRNA of this gene in preeclampsia has been confirmed by reverse transcription-polymerase chain reaction in a previous study (Li et al., 2013), which is consistent with our results. In humans, during trophobalst invasion, trophoblast cells acquire variations in the integrin phenotype, acquiring integrins $\alpha5\beta1$ and $\alpha1\beta1$ (Jovanović et al., 2010). Regulation of normal human trophoblast cell invasion may be important in the mechanisms underlying preeclampsia (Li et al., 2013). Therefore, further therapeutic studies of *ITGB1* and preeclampsia are necessary.

In summary, we carried out PLS analysis to identify differentially expressed genes in preeclampsia. Pathway and GO enrichment analyses were also used to identify biologically relevant signatures of dysregulated genes. We also constructed an interaction network to identify key hub genes. Our results increase the understanding of the pathogenesis of preeclampsia.

## Conflicts of interest

The authors declare no conflict of interests.

## ACKNOWLEDGMENTS

Research supported by the National Natural Science Foundation of China (#81070496).

## REFERENCES

Ashburner M, Ball CA, Blake JA, Botstein D, et al. (2000). Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet*. 25: 25-29.

Cayli S, Demirturk F, Ocakli S, Aytan H, et al. (2012). Altered expression of COP9 signalosome proteins in preeclampsia. *Gynecol. Endocrinol*. 28: 488-491.

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

Ferrara N (2004). Vascular endothelial growth factor as a target for anticancer therapy. *Oncologist* 9 Suppl 1: 2-10.

Gosselin R, Rodrigue D and Duchesne C (2010). A bootstrap-VIP approach for selecting wavelength intervals in spectral

imaging applications. *Chemometr. Intell. Lab*. 100: 12-21.
Graupera M, Guillermet-Guibert J, Foukas LC, Phng LK, et al. (2008). Angiogenesis selectively requires the p110alpha isoform of PI3K to control endothelial cell migration. *Nature* 453: 662-666.
Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, et al. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4: 249-264.
Jackson DW, Sciscione A, Hartley TL, Haynes AL, et al. (1996). Lysosomal enzymuria in preeclampsia. *Am. J. Kidney Dis*. 27: 826-833.
Ji G, Yang Z and You W (2011). PLS-based gene selection and identification of tumor-specific genes. *IEEE Trans. Syst. Man. Cybern. C* 41: 830-841.
Jovanović M, Stefanoska I, Radojcić L and Vićovac L (2010). Interleukin-8 (CXCL8) stimulates trophoblast cell migration and invasion by increasing levels of matrix metalloproteinase (MMP)2 and MMP9 and integrins alpha5 and beta1. *Reproduction* 139: 789-798.
Kanehisa M and Goto S (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 28: 27-30.
Lapaire O, Grill S, Lalevee S, Kolla V, et al. (2012). Microarray screening for novel preeclampsia biomarker candidates. *Fetal Diagn. Ther*. 31: 147-153.
Levine RJ, Maynard SE, Qian C, Lim KH, et al. (2004). Circulating angiogenic factors and the risk of preeclampsia. *N. Engl. J. Med*. 350: 672-683.
Li P, Guo W, Du L, Zhao J, et al. (2013). microRNA-29b contributes to pre-eclampsia through its effects on apoptosis, invasion and angiogenesis of trophoblast cells. *Clin. Sci.* 124: 27-40.
Louwen F, Muschol-Steinmetz C, Reinhard J, Reitter A, et al. (2012). A lesson for cancer research: placental microarray gene analysis in preeclampsia. *Oncotarget* 3: 759-773.
Martins JPA, Teófilo RF and Ferreira MMC (2010). Computational performance and cross-validation error precision of five PLS algorithms using designed and real data sets. *J. Chemometr*. 24: 320-332.
Meng T, Chen H, Sun M, Wang H, et al. (2012). Identification of differential gene expression profiles in placentas from preeclamptic pregnancies versus normal pregnancies by DNA microarrays. *OMICS*16: 301-311.
Shannon P, Markiel A, Ozier O, Baliga NS, et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 13: 2498-2504.
Sibai B, Dekker G and Kupferminc M (2005). Pre-eclampsia. 365: 785-799.
Stelzl U, Worm U, Lalowski M, Haenig C, et al. (2005). A human protein-protein interaction network: a resource for annotating the proteome. *Cell*122: 957-968.
van Dijk M, van Bezu J, van Abel D, Dunk C, et al. (2010). The STOX1 genotype associated with pre-eclampsia leads to a reduction of trophoblast invasion by alpha-T-catenin upregulation. *Hum. Mol. Genet*. 19: 2658-2667.
Zhao M, Yin Y, Guo F, Wang J, et al. (2013). Placental expression of VEGF is increased in pregnancies with hydatidiform mole: possible association with developing very early onset preeclampsia. *Early Hum. Dev*. 89: 583-588.
Zhou Y, Gormley MJ, Hunkapiller NM, Kapidzic M, et al. (2013). Reversal of gene dysregulation in cultured cytotrophoblasts reveals possible causes of preeclampsia. *J. Clin. Invest*. 123: 2862-2872.