Partial least squares based gene expression analysis in estrogen receptor positive and negative breast tumors

W. MA[1], T.-F. ZHANG[1], P. LU[2], S.H. LU[1,3]

[1]Oncology Department, the First Affiliated Hospital of Zhengzhou University, China
[2]Gastrointestinal Surgery Department, People's Hospital of Zhengzhou, China
[3]Cancer Institute and Hospital, Chinese Academy of Medical Sciences, Beijing, China

Abstract. – BACKGROUND: Breast cancer is categorized into two broad groups: estrogen receptor positive (ER+) and ER negative (ER-) groups. Previous study proposed that under trastuzumab-based neoadjuvant chemotherapy, tumor initiating cell (TIC) featured ER- tumors response better than ER+ tumors. Exploration of the molecular difference of these two groups may help developing new therapeutic strategies, especially for ER- patients.

MATERIALS AND METHODS: With gene expression profile from the Gene Expression Omnibus (GEO) database, we performed partial least squares (PLS) based analysis, which is more sensitive than common variance/regression analysis.

RESULTS: We acquired 512 differentially expressed genes. Four pathways were found to be enriched with differentially expressed genes, involving immune system, metabolism and genetic information processing process. Network analysis identified five hub genes with degrees higher than 10, including APP, ESR1, SMAD3, HDAC2, and PRKAA1.

CONCLUSIONS: Our findings provide new understanding for the molecular difference between TIC featured ER- and ER+ breast tumors with the hope offer supports for therapeutic studies.

Key Words:
Partial least squares, Gene expression, Tumor initiating cell, Breast cancer.

Introduction

Breast cancer remains a major health problem worldwide[1]. It can be divided into two distinct biologically and clinically significant groups: estrogen receptor positive (ER+) and ER negative (ER-) groups[2]. Currently, except for surgical operation, ER+ group can be treated with endocrine therapy or chemotherapy while chemotherapy is the only option for ER- group[3-5]. Thus, development of new treatment strategy for the ER- group is urgent. Previous study[6] proposed that under trastuzumab-based neoadjuvant chemotherapy, tumor initiating cell (TIC) featured ER- tumors response better than ER+ tumors. Exploration of the molecular difference of these two groups may help developing new therapeutic strategies for ER- patients.

Large-scale gene expression analysis is powerful in biological characterization and therapeutic planning of complex diseases, including breast cancer[7]. Previous investigations of gene expression analysis mostly used common variance/regression analysis, slipping over unaccounted array specific factors. By contrast, partial least squares (PLS) based analysis has been proposed to be more sensitive and robust in gene expression analysis[8,9]. Previous report[10] on breast cancer using PLS analysis detected new pathways potentially contribute the recurrence rate of breast tumor patients, further suggesting the feasibility of this method on expression profile analysis. However, they investigated ER- and ER+ breast tumors separately and didn't compare the expression difference between these two groups. Understanding of the molecular difference between TIC featured ER- and ER+ breast tumors using PLS based method may develop novel preventing and therapeutic targets of the disease.

In this work, to investigate the gene expression difference between TIC featured ER- and ER+ breast tumors, we carried out PLS-based microarray data analysis using expression profile from the gene expression omnibus (GEO) database. Pathways or Gene Ontology items significantly enriched with differentially expressed genes were also acquired to capture the biologically relevant signature. To identify key molecules among the differentially expressed genes, a protein-protein interaction (PPI) network was constructed with proteins encoded by selected genes.

Corresponding Author: Shih Hsin Lu, MD; e-mail: shihsin_l@126.com;
Peng Lu, MD; e-mail: lupeng.surgeon@gmail.com.

Materials and Methods

Microarray Data

The microarray data set GSE37946 used in this study was downloaded from the GEO database. This series represents gene expression profile of 32 TIC featured HER2+: ER- and 18 HER2+: ER + breast tumor samples. The data set was based on platform GPL96: [HG-U133A] Affymetrix Human Genome U133A Array.

Detection of Differentially Expressed Genes (DEGs)

Raw data of all samples were downloaded. Raw intensity values were normalized by using Robust Multi-array Analysis (RMA)[11]. Log2-transformed expression values generated from RMA were used in PLS analysis to evaluate their effect in the TIC featured HER2+: ER- and HER2+: ER+ samples. In brief, Firstly, non-linear iterative partial least squares (NIPALS) algorithm[12] was used to calculate PLS latent variables; Secondly, the effect of probe expression value on the samples was evaluated based on variable importance in the projection (VIP)[13]. Thirdly, a permutation procedure (n=10,000) was implemented to obtain the empirical distribution of PLS-based VIP and false discovered rate (FDR) of each probe was calculated. In this study, the threshold of differentially expression was defined as 0.05.

Enrichment Analysis

Annotation of the probes was carried out by using the simple omnibus format in text (SOFT) format files. The biological processes which involve the genes were obtained based on the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways database (http://www.genome.jp/kegg/)[14] and Gene Ontology (GO) database[15]. Pathway and GO enrichment analysis were performed to capture biologically relevant signature of the differentially expressed genes with hyper geometric distribution test.

Network Analysis

Most proteins function through its interactions with other proteins thus protein-protein interaction is essential for all biological processes[16]. Proteins encoded by differentially expressed genes with more interactions with other proteins are supposed to play more important roles in the biological difference of TIC featured HER2+: ER- and HER2+: ER + samples. To identify key genes among the differentially expressed genes, a network was constructed by using Cytoscape (V 2.8.3, http://www.cytoscape.org/)[17] and the National Center for Biotechnology Information (NCBI) database (http://ftp.ncbi.nlm.nih.gov/gene/GeneRIF/). The degree of each protein was defined as its number of interactions. Proteins with degrees over 10 were considered as hub molecules in this study.

Results

A total of 512 genes were differentially expressed between TIC featured HER2+: ER- and HER2+: ER + samples. For all genes in the microarray, 5339 genes can be mapped to the KEGG pathway database, including 223 DEGs. Pathways enriched with deregulated genes are listed in Table I. These four pathways involve the immune system, metabolism and genetic information processing. The complement and coagulation cascades pathway (hsa04610) was the most significant pathway with over represented selected genes. Of all genes in the array, 12221 genes were annotated according to the GO database, including 474 DEGs. Table II represents the top 10 GO items enriched with selected genes. Consistent with the pathway analysis, process items related with immune response such as chemokine production (GO:0032602) and chronic inflammatory response (GO:0002544) were also identified to be enriched with dysregulated genes. Two GO items, response to estrogen stimulus (GO:0043627) and estrogen-activated sequence-specific DNA binding RNA polymerase II transcription factor activity

Table I. Pathways enriched with differentially expressed genes.

#KEGG	Pathway description	Pathway class	*p* value
hsa04610	Complement and coagulation cascades	Immune systems	4.56E-04
hsa00760	Nicotinate and nicotinamide metabolism	Metabolism	7.00E-03
hsa03018	RNA degradation	Genetic Information Processing	2.40E-02
hsa00970	Aminoacyl-tRNA biosynthesis	Genetic Information Processing	3.85E-02

(GO:0038052), were associated with the activity of estrogen. Other items included processes related with transcription and apoptotic signaling.

Interaction network constructed by proteins encoded by DEGs is illustrated in Figure 1. A total of five hub molecules were identified, including APP, ESR1, SMAD3, HDAC2 and PRKAA1, with the degrees of 73, 34, 15, 13, and 12 respectively.

Discussion

For gene expression profile analysis, a major challenge is to create an effective mathematical model to handle the small sample and the relative large number of genes. Previous mostly used variance or regression analysis, which does not take unaccounted array specific factors into consider. Here, we used a PLS based model to identify differentially expressed genes between TIC featured ER- and ER+ breast tumors.

Pathway analysis revealed that the complement and coagulation cascades pathway (hsa04610) was the most significant pathway with over represented selected genes. Involvement of this pathway in breast cancer has been proposed in studies using both gene expression and proteomics data[18,19]. Our results further confirmed the relationship between this pathway and breast cancer, suggesting that this pathway may contribute to the clinical difference between the two groups. Consistent with pathway analysis, GO analysis also revealed the involvement of immune response in the biological difference between the two groups. GO items associated with the activity of estrogen were also identified to be with over represented selected genes. This is expected since a major clinical difference of the two groups is the presence status of the estrogen receptor.

Network analysis was constructed to identify key molecules among the DEGs. *APP* was a hub gene with the highest degree (Figure 1). Protein encoded by this gene is amyloid precursor protein, a transmembrane protein that has been implicated in some human malignancies. Previous study has reported it as a potent prognostic factor in ER+ breast cancer patients, but not in ER- patients[20]. In our study, this gene was identified to be significantly down regulated in ER+ patients. Thus, this gene may involve in the molecular mechanism of ER+ patients, leading to distinct clinical manifestations of ER+ and ER- patients. *ESR1* was identified as a hub gene with the second highest degree (Figure 1). The differential expression of this gene is expected since the presence status of the estrogen receptor is the major clinical difference between the two groups. Enrichment analysis also revealed the involvement of estrogen related process in the biological difference between the two groups. *SMAD3* is another hub gene with the degree of 15. Protein encoded by this gene a critical intracellular mediator of TGF signaling, and has been considered as a potential prognostic and therapeutic target in breast cancer[21,22]. Another two hub genes are *HDAC2* and *PRKAA1*. Dysregulation of *HDAC2* has been associated with clinicopathological indicators of breast cancer progression[23]. *PRKAA1* was also considered as a candidate risk gene of breast cancer due to its involvement in the AMPK/mTOR pathway[24,25], which has been considered as a effective target in anti-cancer therapy[26]. Our results suggested that these genes may also play important roles in the difference between ER+ and ER- patients.

Table II. The top ten GO items enriched with differentially expressed genes.

$^{\#}$GO id	GO description	GO class	*p* value
GO:2001244	Positive regulation of intrinsic apoptotic signaling pathway	Process	7.32E-04
GO:0045926	Negative regulation of growth	Process	8.63E-04
GO:0032602	Chemokine production	Process	1.50E-03
GO:0001071	Nucleic acid binding transcription factor activity	Function	2.81E-03
GO:0002544	Chronic inflammatory response	Process	4.09E-03
GO:0043433	Negative regulation of sequence-specific DNA binding transcription factor activity	Process	4.62E-03
GO:0045893	Positive regulation of transcription, DNA-dependent	Process	5.38E-03
GO:0005154	Epidermal growth factor receptor binding	Function	5.45E-03
GO:0043627	Response to estrogen stimulus	Process	3.46E-02
GO:0038052	Estrogen-activated sequence-specific DNA binding RNA polymerase II transcription factor activity	Function	3.88E-02

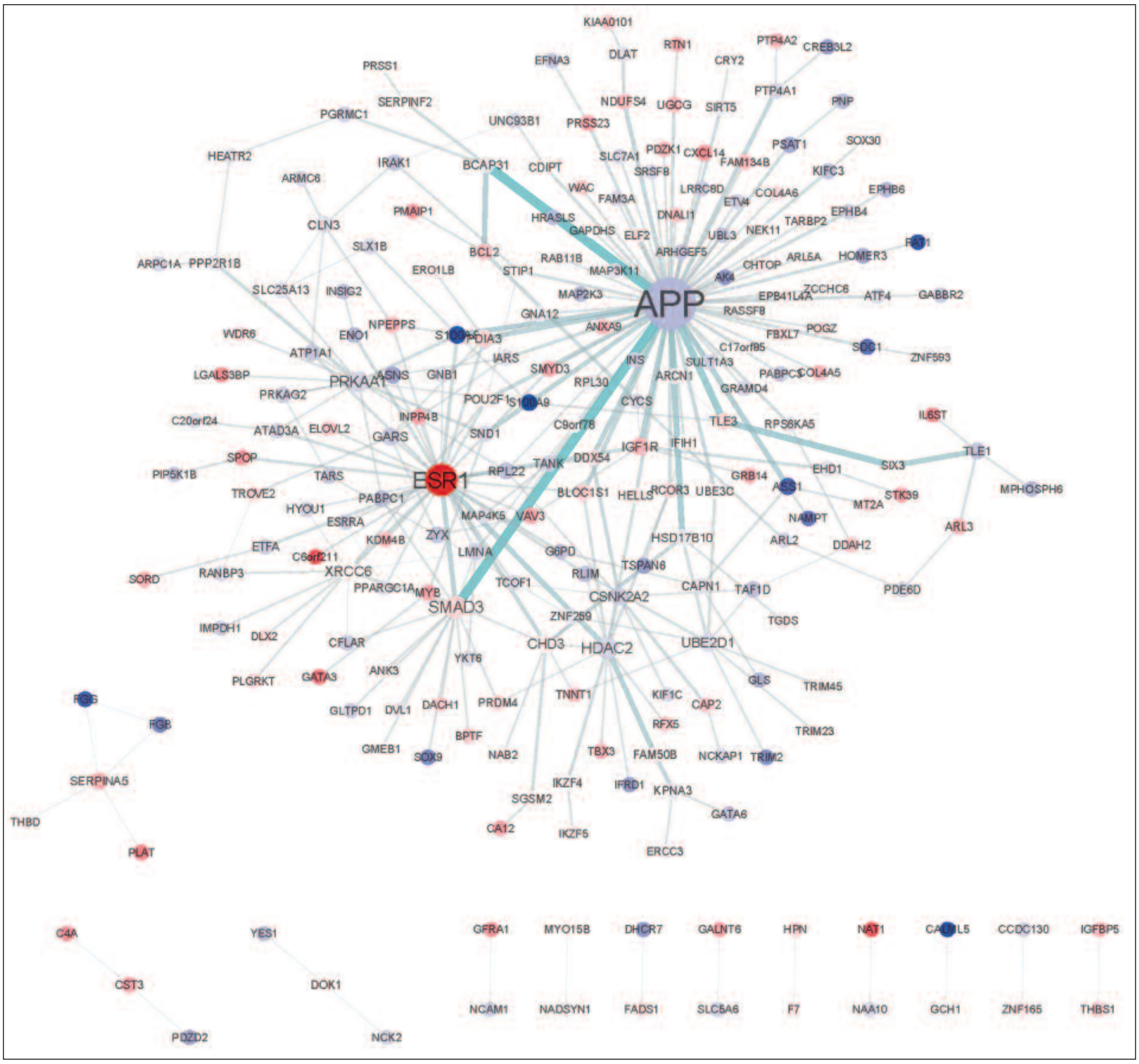

Figure 1. Interaction network constructed by proteins encoded by differentially expressed genes. Proteins with more interactions are shown in bigger size. Proteins in red are encoded by overexpressed genes in ER+ patients while those in blue are encoded by depressed genes in ER+ samples.

Conclusions

With gene expression profile from the GEO database, we implemented PLS based analysis to identify differentially expressed genes TIC featured ER- and ER+ breast cancer patients. Enrichment analysis identified the involvement of immune system, metabolism and genetic information processing process. Network analysis identified five hub genes. Our results provide new understanding of the molecular mechanism underlying the distinct clinical manifestations of ER- and ER+ patients.

Conflict of Interest

The Authors declare that there are no conflicts of interest.

References

1) NETWORK TCGA. Comprehensive molecular portraits of human breast tumours. Nature 2012; 490: 61-70.

2) RAKHA EA, REIS-FILHO JS, ELLIS IO. Basal-like breast cancer: a critical review. J Clin Oncol 2008; 26: 2568-2581.

3) PAIK S, SHAK S, TANG G, KIM C, BAKER J, CRONIN M, BAEHNER FL, WALKER MG, WATSON D, PARK T, HILLER

W, Fisher ER, Wickerham DL, Bryant J, Wolmark N. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 2004; 351: 2817-2826.

4) van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002; 415: 530-536.

5) Perou CM. Molecular stratification of triple-negative breast cancers. Oncologist 2011; 16(Suppl 1): 61-70.

6) Liu JC, Voisin V, Bader GD, Deng T, Pusztai L, Symmans WF, Esteva FJ, Egan SE, Zacksenhaus E. Seventeen-gene signature from enriched Her2/Neu mammary tumor-initiating cells predicts clinical outcome for human HER2+:ERalpha- breast cancer. Proc Natl Acad Sci U S A 2012; 109: 5832-5837.

7) Olopade OI, Grushko TA, Nanda R, Huo D. Advances in breast cancer: pathways to personalized medicine. Clin Cancer Res 2008; 14: 7988-7999.

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

9) Ji G, Yang Z, You W. PLS-based gene selection and identification of tumor-specific genes. ieee transactions on systems, man, and cybernetics-Part C. Appl Rev 2011; 41: 830-841.

10) Gao QG, Li ZM, Wu KQ. Partial least squares based analysis of pathways in recurrent breast cancer. Eur Rev Med Pharmacol Sci 2013; 17: 2159-2165.

11) Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 2003; 4: 249-264.

12) Martins JPA, Teofilo RF, Ferreira MMC. Computational performance and cross-validation error precision of five PLS algorithms using designed and real data sets. J Chemometric 2010; 24: 320-332.

13) Gosselin R, Rodrigue D, Duchesne C. A Bootstrap-VIP approach for selecting wavelength intervals in spectral imaging applications. Chemometr Intell Lab Syst 2010; 100: 12-21.

14) Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000; 28: 27-30.

15) Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000; 25: 25-29.

16) Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, Timm J, Mintzlaff S, Abraham C, Bock N, Kietzmann S, Goedde A, Toksoz E, Droege A, Krobitsch S, Korn B, Birchmeier W, Lehrach, Wanker EE. A human protein-protein interaction network: a resource for annotating the proteome. Cell 2005; 122: 957-968.

17) Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003; 13: 2498-2504.

18) Zhang F, Chen JY. Discovery of pathway biomarkers from coupled proteomics and systems biology methods. BMC Genomics 2010; 11(Suppl 2): S12.

19) Maia AT, Spiteri I, Lee AJ, O'Reilly M, Jones L, Caldas C, Ponder BA. Extent of differential allelic expression of candidate breast cancer genes is similar in blood and breast. Breast Cancer Res 2009; 11: R88.

20) Takagi K, Ito S, Miyazaki T, Miki Y, Shibahara Y, Ishida T, Watanabe M, Inoue S, Sasano H, Suzuki T. Amyloid precursor protein in human breast cancer: An androgen-induced gene associated with cell proliferation. Cancer Sci 2013 Jul 24. [Epub ahead of print].

21) Tarasewicz E, Jeruss JS. Phospho-specific Smad3 signaling: impact on breast oncogenesis. Cell Cycle 2012; 11: 2443-2451.

22) Zhang X, Li Y, Zhang Y, Song J, Wang Q, Zheng L, Liu D. Beta-elemene blocks epithelial-mesenchymal transition in human breast cancer cell line MCF-7 through Smad3-mediated down-regulation of nuclear transcription factors. PLoS One 2013; 8: e58719.

23) Muller BM, Jana L, Kasajima A, Lehmann A, Prinzler J, Budczies J, Winzer KJ, Dietel M, Weichert W, Denkert C. Differential expression of histone deacetylases HDAC1, 2 and 3 in human breast cancer--overexpression of HDAC2 and HDAC3 is associated with clinicopathological indicators of disease progression. BMC Cancer 2013; 13: 215.

24) Campa D, Claus R, Dostal L, Stein A, Chang-Claude J, Meidtner K, Boeing H, Olsen A, Tjonneland A, Overvad K, Rodriguez L, Bonet C, Sanchez MJ, Amiano P, Huerta JM, Barricarte A, Khaw KT, Wareham N, Travis RC, Allen NE, Trichopoulou A, Bamia C, Benetou V, Palli D, Agnoli C, Panico S, Tumino R, Sacerdote C, van Kranen H, Bas Bueno-de-Mesquita H, Peeters PH, van Gils CH, Lenner P, Sund M, Lund E, Gram IT, Rinaldi S, Chajes V, Romieu I, Engel P, Boutron-Ruault MC, Clavel-Chapelon F, Siddiq A, Riboli E, Canzian F, Kaaks R. Variation in genes coding for AMP-activated protein kinase (AMPK) and breast cancer risk in the European Prospective Investigation on Cancer (EPIC). Breast Cancer Res Treat 2011; 127: 761-767.

25) Slattery ML, John EM, Torres-Mejia G, Lundgreen A, Herrick JS, Baumgartner KB, Hines LM, Stern MC, Wolff RK. Genetic variation in genes involved in hormones, inflammation and energetic factors and breast cancer risk in an admixed population. Carcinogenesis 2012; 33: 1512-1521.

26) Sun G, Shan MH, Ma BL, Geng ZL, Alibiyati A, Zhong H, Wang J, Ren GH, Li HT, Dong C. Identifying crosstalk of mTOR signaling pathway of lobular breast carcinomas. Eur Rev Med Pharmacol Sci 2012; 16: 1355-1361.