Partial least squares based analysis of pathways in recurrent breast cancer

Q.-G. GAO, Z.-M. LI, K.-Q. WU

First People's Hospital, Wujiang, Jiangsu, China

Abstract. – PURPOSE: Breast cancer remains a major health problem even with all the recent technological advancements. Large-scale gene expression analysis has offered great ease for both biological characterization and therapeutic planning of breast cancer. Previous studies mostly used variance/regression analysis, which becomes fundamentally flawed when there are unaccounted array specific factors. Here we aim to investigate the underlying mechanism of breast cancer through partial least squares (PLS) based gene expression profile analysis.

MATERIALS AND METHODS: With a gene expression profile data set downloaded from the Gene Expression Omnibus database, we performed PLS based analysis.

RESULTS: We acquire 932 and 771 differentially expressed genes (DEGs) in breast cancer metastasis of estrogen-receptor (ER)-positive and ER-negative patients, respectively. For ER-positive patients, 32 pathways were found to be enriched with DEGs, including immune related pathways, cellular processes and environmental information processing pathways. Survival analysis demonstrated that 18 of them were closely related with non-recurrence rate along time after surgery. For ER negative patients, only three pathways including the folate biosynthesis pathway were enriched with DEGs and none of them overlapped with those of ER positive patients. Only the cholinergic synapse pathway was significantly associated with the non-recurrence rate according to the survival analysis.

CONCLUSIONS: Our findings shed light on pathways involved in breast cancer relapse with the hope to give some theoretical supports for further therapeutic study.

Key Words:
Breast cancer, Gene expression profile, Partial least squares, Pathway, Survival analysis.

Introduction

Breast cancer is a major health problem that causes greater than 1,300,000 cases and 450,000 deaths each year worldwide[1]. Breast cancer is categorized into two broad groups: estrogen receptor (ER) positive and ER-negative groups[2], which can be further subclassified into additional distinct biologically and clinically significant subgroups, indicating that breast cancer is highly heterogeneous at both the molecular and clinical level[3,4]. These subtypes which are evident at the ductal carcinoma in situ (DCIS) stage are conserved across different ethnic groups[5], suggesting that distinct tumor progression pathways may contribute to different tumor types. During the past decade, our understanding of the molecular mechanisms underlying breast cancer progression has greatly improved as the crucial roles of genetic alterations in the early detection, diagnosis, and treatment of breast cancer have been investigated widely. Therefore, more reliable and specific treatment options for breast cancer are now available to patients. It is now well known that 5% to 10% of breast cancer cases arise in individuals with variants in breast cancer candidate genes, such as *BRCA1* and *BRCA2*[6-8]. However, the heterogeneous nature of breast cancer makes it difficult not only to develop cure for this disease, but also to assess risk factors for relapse.

Recent advances in microarray technology and other strategies of large-scale gene expression analysis have offered great ease for both biological characterization and therapeutic planning of breast cancer[9]. Previous study[10] proposed that partial least squares (PLS) based analysis was robust in selecting disease related genes with expression profile data. Compared with variance/regression analysis, which becomes fundamentally flawed when there are unaccounted array specific factors that are not detectable and cannot be removed by any standard normalizing method, PLS based analysis is more sensitive while maintaining reasonable specificity, false discovery rate and false non-discovery rate. Understanding of the molecular biology and gene expression signatures of breast cancer using PLS based method may help the development of novel preventing and therapeutic approaches of the disease.

In this work, using a dataset downloaded from the gene expression omnibus (GEO) database, we

Corresponding Author: Zhimin Li, MD; e-mail: lizhimin3051@163.com

investigated the pathological mechanism in ER positive and ER-negative breast cancer and acquired corresponding signaling pathways. Moreover, biological pathways co-related with life span of breast cancer patients were also obtained through survival analysis.

Materials and Methods

Microarray Data

The transcription profile of GSE2034 was downloaded from GEO (http://www.ncbi.nlm.nih.gov/geo/) database. This series represents 208 lymph-node negative breast cancer patients, including 209 ER positive and 77 ER negative patients. For ER positive patients, 129 samples showed recurrence and 80 samples showed no recurrence within 4 years. For ER negative patients, 27 samples showed recurrence and 50 samples showed no recurrence. The data set was based on platform GPL96: [HG-U133A] Affymetrix Human Genome U133A Array.

Identification of Differentially Expressed Genes (DEGs)

Robust Multiarray Analysis (RMA)[11] was used for raw intensity values normalization. The resulting RMA expression value (log2-transformed) of each probe was used for further analysis. PLS[12,13], which is a dimension reduction procedure for modeling without imposing strong assumptions were then used to estimate the effects for each gene in ER positive and ER negative patients. Briefly, NIPALS algorithm[14] was used to calculate PLS latent variables derived from expression profile on target trait and the statistics of variable importance on the projection (VIP)[15] was calculated to evaluate the importance of the expressed genes on the relapse status of the patients. Additionally, a permutation procedure (N=10,000 times) was implemented to obtain the empirical distribution of PLS-based VIP in each replicate. False discovered rate (FDR) of each gene was then evaluated according to the empirical distribution. The threshold of FDR was set as 0.05.

Pathway Enrichment Analysis

Identified differentially expressed probes were annotated according to the SOFT format files. All genes were mapped to the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database[16], which is a collection of wiring diagrams of molecular interactions, reactions, and relations. Hyper geometric distribution test was implemented to identify pathways significantly enriched with DEGs.

Survival Analysis

To investigate the contribution of these pathways to the relapse rate as time went after the surgery of the breast cancer patients, we carried out survival analysis. Take ER positive patients for example, for each pathway which was identified to be enriched with DEGs, the samples were divided into two categories (class 1 and class 2) using K-mean method according to the expressions of all genes in the pathway. With the relapse time or the last follow-up time of these patients, non-recurrence rate curve of the two classes were drawn. Log-rank test were used to compare the two curves and corresponding *p* values were obtained to investigate whether the two groups were significantly different from each other. The threshold was set as 0.05. That's, pathways with survival analysis *p* values less than 0.05 were considered to be significantly contribute to the non-recurrence rate along time went of the patients.

Results

According to the PLS based expression profile analysis, for ER positive patients, 1124 probes (932 genes) were identified to be differentially expressed in samples with or without relapse within 4 years. For ER negative patients, the number was 947 probes (771 genes). Only 96 DEGs were shared in the two groups. Pathway enrichment analysis showed that 32 pathways were enriched with DEGs in ER positive patients (Table I). Only three pathways showed enrichment in ER negative patients (Table II).

Further survival analysis showed that 18 pathways were related with the non-recurrence rate along time of ER positive patients, including the oocyte meiosis pathway (hsa04114), cell cycle pathway (hsa04110), p53 signaling pathway (hsa04115), pathways in cancer (hsa05200), small cell lung cancer (hsa05222) and other pathways which involve the immune system. Figure 1 illustrates the non-recurrence rate curve of the cell cycle pathway and the DEGs in this pathway. For ER negative patients, only the cholinergic synapse pathway (hsa04725) was identified to be related with the non-recurrence rate. The non-recurrence rate curve and the DEGs in this pathway are shown in Figure 2.

Table I. Pathways enriched with differentially expressed genes in samples with or without relapse within 4 years in estrogen receptor positive patients

Pathway ID	Pathway name	*p*-value	FDR	Survival *p*
hsa04145	Phagosome	3.45E-07	6.87E-06	2.14E-01
hsa04510	Focal adhesion	2.43E-06	3.94E-05	3.17E-01
hsa05332	Graft-versus-host disease	6.42E-12	8.31E-10	2.50E-02
hsa05330	Allograft rejection	5.86E-12	1.52E-09	2.50E-02
hsa04940	Type I diabetes mellitus	2.00E-10	1.73E-08	2.50E-02
hsa05416	Viral myocarditis	3.04E-09	1.97E-07	3.60E-02
hsa04514	Cell adhesion molecules (CAMs)	6.64E-09	3.44E-07	5.03E-02
hsa05320	Autoimmune thyroid disease	1.01E-08	4.36E-07	2.50E-02
hsa04512	ECM-receptor interaction	1.84E-08	6.81E-07	2.51E-01
hsa04672	Intestinal immune network for IgA production	2.31E-08	7.49E-07	1.50E-01
hsa05140	Leishmaniasis	3.07E-08	8.84E-07	2.62E-01
hsa04612	Antigen processing and presentation	9.12E-08	2.36E-06	3.26E-02
hsa05310	Asthma	2.20E-07	5.17E-06	1.60E-02
hsa05169	Epstein-Barr virus infection	2.63E-07	5.68E-06	3.17E-02
hsa05166	HTLV-I infection	3.90E-07	7.22E-06	5.93E-02
hsa05145	Toxoplasmosis	4.51E-07	7.79E-06	1.10E-02
hsa05168	Herpes simplex infection	4.58E-06	6.98E-05	4.65E-02
hsa05152	Tuberculosis	5.36E-05	7.71E-04	1.77E-01
hsa04114	Oocyte meiosis	6.79E-05	9.26E-04	1.57E-06
hsa05150	Staphylococcus aureus infection	8.07E-05	1.05E-03	1.77E-01
hsa05323	Rheumatoid arthritis	1.69E-04	2.09E-03	3.13E-02
hsa04974	Protein digestion and absorption	1.79E-04	2.11E-03	5.39E-01
hsa05146	Amoebiasis	3.29E-04	3.70E-03	8.36E-02
hsa04360	Axon guidance	7.79E-04	8.41E-03	4.97E-01
hsa04110	Cell cycle	9.80E-04	9.76E-03	5.06E-04
hsa04151	PI3K-Akt signaling pathway	9.48E-04	9.76E-03	2.27E-01
hsa05222	Small cell lung cancer	1.15E-03	1.10E-02	4.54E-02
hsa04115	p53 signaling pathway	1.32E-03	1.22E-02	7.02E-03
hsa05200	Pathways in cancer	2.79E-03	2.50E-02	3.63E-02
hsa00562	Inositol phosphate metabolism	3.87E-03	3.34E-02	2.27E-01
hsa05144	Malaria	4.18E-03	3.49E-02	9.35E-03
hsa05164	Influenza A	5.51E-03	4.46E-02	2.72E-02

Note: The third column indicated the *p* value of the enrichment analysis. The last column indicated the p value of the survival analysis.

Table II. Pathways enriched with differentially expressed genes in samples with or without relapse within 4 years in estrogen receptor negative patients.

Pathway ID	Pathway name	*p*-value	FDR	Survival *p*
hsa04725	Cholinergic synapse	1.39E-04	1.80E-02	6.35E-03
hsa00790	Folate biosynthesis	4.85E-04	4.19E-02	2.16E-01
hsa05145	Toxoplasmosis	7.02E-04	4.54E-02	6.62E-03

Note: The third column indicated the p value of the enrichment analysis. The last column indicated the p value of the survival analysis.

Discussion

Progressive pathophysiology of breast cancer is highly complex. Genome-wide expression profiling is a powerful procedure for investigating the underlying pathophysiological cascades in breast cancer patients. However, it is a challenge to create a suitable mathematical model to deal with relatively small sample size and large number of genes[17]. Previous gene expression studies for breast cancer mostly used variance/regression to identify the DEGs. With this procedure, the true picture may be blurred by hidden biological effects which cannot be removed through routine

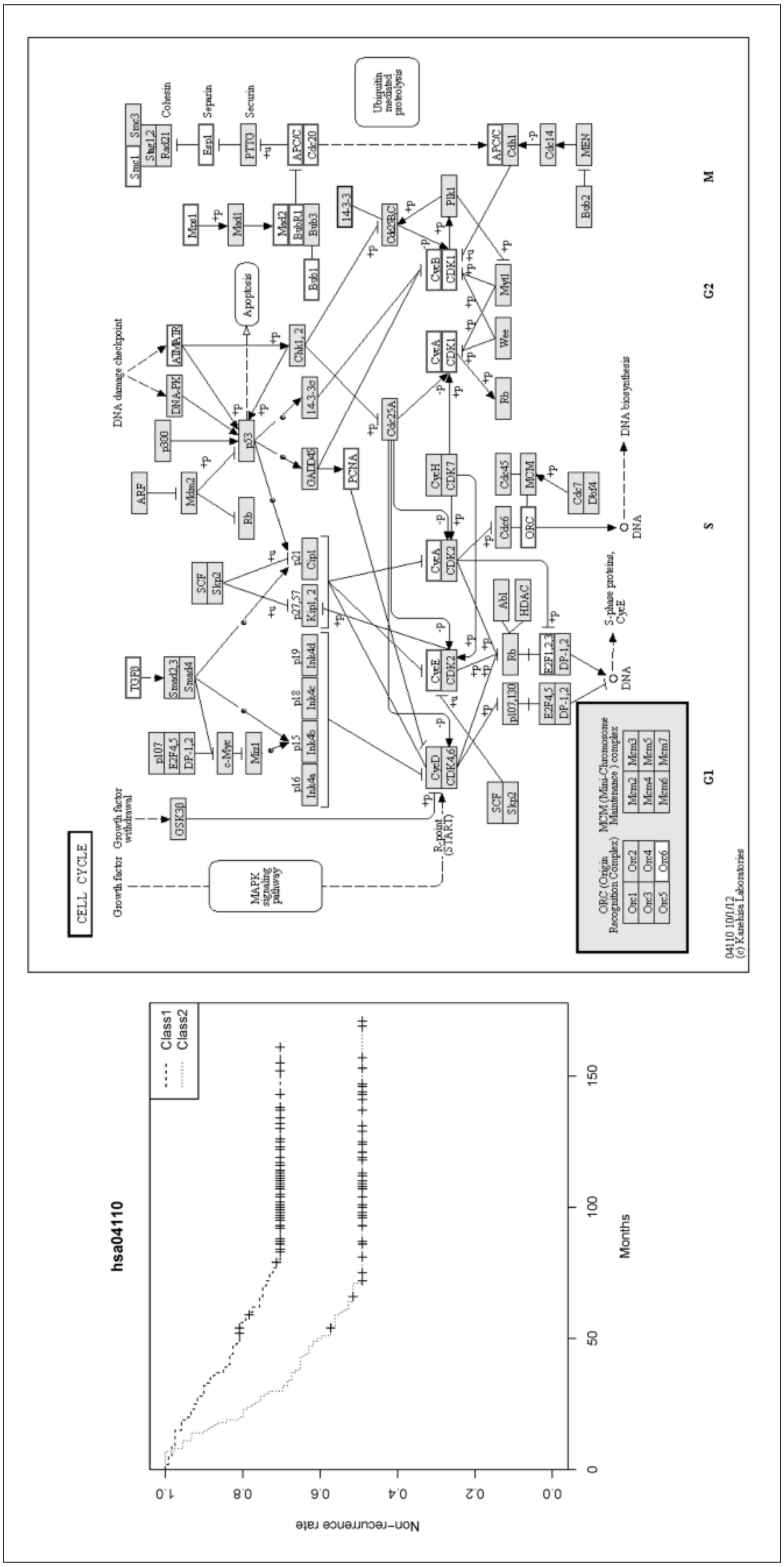

Figure 1. Survival analysis results of the cell cycle pathway in estrogen receptor positive patients. The top plot represents the survival curve. The below plot represents the deferentially expressed genes involved in this pathway. Proteins encoded by overexpressed genes are shown with white background. Proteins which are shown with grey background means no differential expression was detected.

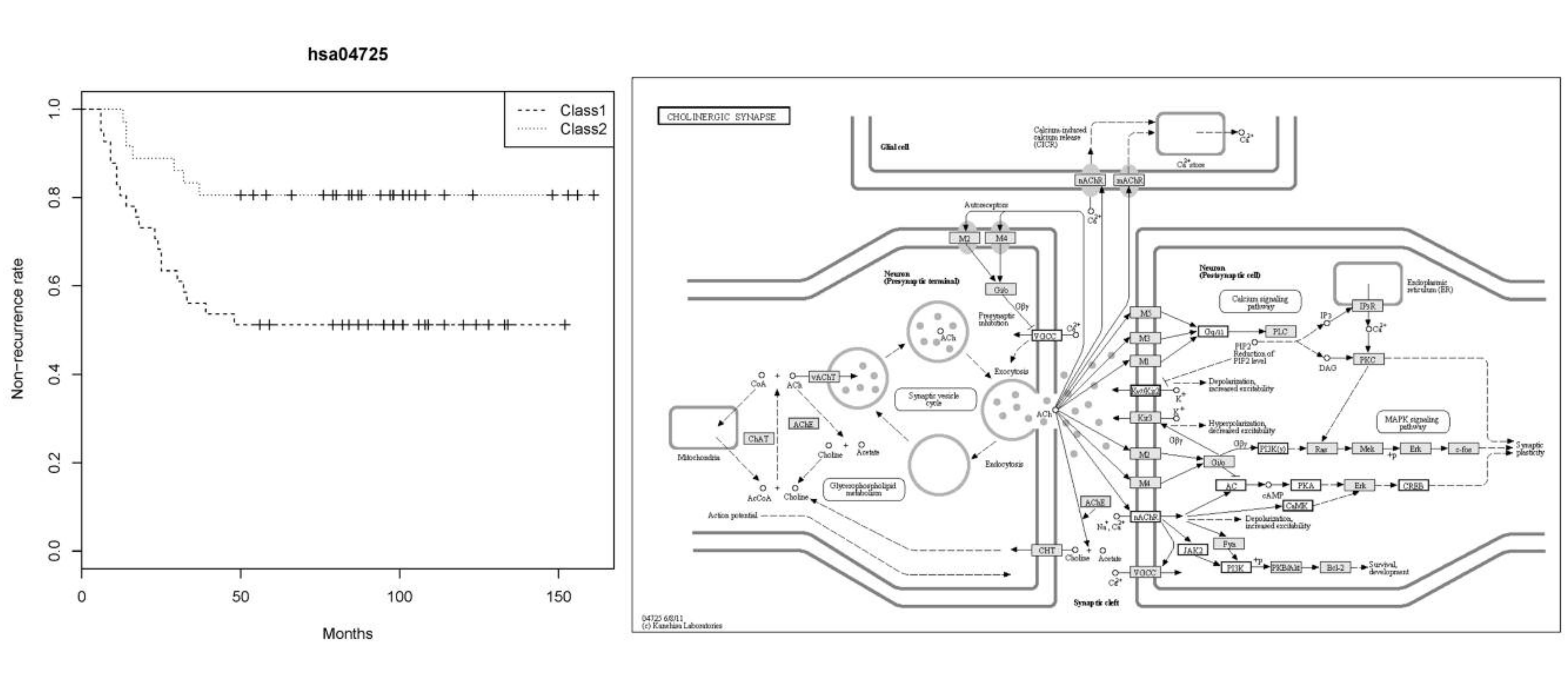

Figure 2. Survival analysis of the cholinergic synapse pathway in estrogen receptor negative patients. The top plot represents the survival curve. The below plot represents the differentially expressed genes involved in this pathway. Proteins encoded by down regulated genes are shown with white background. Proteins which are shown with grey background means no differential expression was detected.

normalizing method. Here, we used a PLS based multivariate linear model to identify DEGs in sample with or without relapse in both ER positive and negative patients. Subsequent pathway analysis was also used to explore underlying mechanism of recurrent breast cancer.

According to the results, 932 and 771 genes were identified to be differentially expressed in samples with or without relapse within 4 years in ER positive and ER negative patients, respectively. Only 96 DEGs, less than 13% of the total DEGs, were shared in the two groups and this is not unexpected since the two groups are biologically distinct from each other.

For ER positive patients, pathway enrichment analysis identified 32 pathways enriched with DEGs. Except for the pathways involve immune system, several pathways are related with the cellular processes and environmental information processing (Table I). For example, the extracellular matrix (ECM)-receptor interaction signaling pathways were reported to be closely related with breast cancer. Among the deregulated genes in this pathway, laminin was shown to play important roles in regulating cell migration, facilitating tumor invasion[18]. Type IV collagen is involved in regulation process of mammary cell proliferation, cell attachment and migration[19]. Besides, disruption of cell-ECM interactions is a consistent feature of mammary tumors and cells in vivo and *in vitro*[20,21]. Subsequent survival analysis identified 18 of these pathways which may contribute to the non-recurrence rate of patients as time goes after the surgery. Among them, several pathways are well known to be related with cancers, such as p53 signaling pathway (hsa04115) and pathways in cancer (hsa05200). Besides, identification of the cell cycle pathway is also reasonable since deregulation of this pathway may contribute to the malignant proliferation of the tumor cells. As shown in Figure 1, most DEGs in this pathway are up regulated. Take *CDK1* for example, previous studies have reported that it is related with tumor aggressiveness and poor prognosis[22-24]. This is consistent with our results, since *CDK1* is up-regulated in patients with relapse according to our analysis results.

For ER negative patients, only three pathways were found to be enriched with DEGs and none of them were shared with the ER positive group (Table II). The folate biosynthesis pathway (hsa00790) is noteworthy since dietary folate intake has been reported to be protective for ER negative breast cancer[25]. Subsequent survival analysis showed that only the cholinergic synapse pathway was associated with the non-recurrence rate of ER negative patients. Detail investigation of the relationship between cholinergic synapse and breast cancer are still lacking. However, previous study showed that parathion, which can induce the hydrolysis acetylcholine at cholinergic synapses, is an initiator factor in the transformation process in breast cancer[26].

Conclusions

Using a gene expression profile data set from the GEO database, we carried out a PLS based analysis to identify DEGs contribute to the recurrence pathology of ER positive and negative patients. Further survival analysis were also implemented to identify pathways closely related with the non-recurrence rate along time went after surgery. Our results prove the feasibility of PLS-based method in detecting disease related genes with gene expression profile data and facilitate the disclosure of the molecular mechanism for breast cancer recurrence.

Conflict of Interest

The Authors have no financial conflicts of interest.

Reference

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) Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO, Botstein D. Molecular portraits of human breast tumours. Nature 2000; 406: 747-752.

4) Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Lonning PE, Borresen-Dale AL. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A 2001; 98: 10869-10874.

5) Yu K, Lee CH, Tan PH, Tan P. Conservation of breast cancer molecular subtypes and transcriptional patterns of tumor progression across distinct ethnic populations. Clin Cancer Res 2004; 10: 5508-5517.

6) Claus EB, Risch N, Thompson WD. Genetic analysis of breast cancer in the cancer and steroid hormone study. Am J Hum Genet 1991; 48: 232-242.

7) Easton DF, Ford D, Bishop DT. Breast and ovarian cancer incidence in BRCA1-mutation carriers. Breast Cancer Linkage Consortium. Am J Hum Genet 1995; 56: 265-271.

8) Schubert EL, Mefford HC, Dann JL, Argonza RH, Hull J, King MC. BRCA1 and BRCA2 mutations in Ashkenazi Jewish families with breast and ovarian cancer. Genet Test 1997; 1: 41-46.

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

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

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) Helland IS. On the structure of partial least squares regression. Commun Stat-Simul Comput 1988; 17: 581-607.

13) Helland IS. Partial least squares regression and statistical models. Scand J Stat 1990; 17: 97-114.

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

15) 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.

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

17) Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999; 286: 531-537.

18) Nielsen M, Christensen L, Albrechtsen R. The basement membrane component laminin in breast carcinomas and axillary lymph node metastases. Acta Pathol Microbiol Immunol Scand A 1983; 91: 257-264.

19) Kim JP, Chen JD, Wilke MS, Schall TJ, Woodley DT. Human keratinocyte migration on type IV collagen. Roles of heparin-binding site and alpha 2 beta 1 integrin. Lab Invest 1994; 71: 401-408.

20) Petersen OW, Ronnov-Jessen L, Howlett AR, Bissell MJ. Interaction with basement membrane serves to rapidly distinguish growth and differentiation pattern of normal and malignant human breast epithelial cells. Proc Natl Acad Sci U S A 1992; 89: 9064-9068.

21) Bernfield M, Hinkes MT, Gallo RL. Developmental expression of the syndecans: possible function and regulation. Dev Suppl 1993; 205-212.

22) Moroy T, Geisen C. Cyclin E. Int J Biochem Cell Biol 2004; 36: 1424-1439.

23) Sutherland RL, Musgrove EA. Cyclins and breast cancer. J Mammary Gland Biol Neoplasia 2004; 9: 95-104.

24) Lee MH, Yang HY. Regulators of G1 cyclin-dependent kinases and cancers. Cancer Metastasis Rev 2003; 22: 435-449.

25) Harris HR, Bergkvist L, Wolk A. Folate intake and breast cancer mortality in a cohort of Swedish women. Breast Cancer Res Treat 2012; 132: 243-250.

26) Calaf GM, Roy D. Gene expression signature of parathion-transformed human breast epithelial cells. Int J Mol Med 2007; 19: 741-750.