Genomics & Informatics Vol. 6(4) 210-222, December 2008

Microarray Data Analysis of Perturbed Pathways in Breast Cancer Tissues

Changsik Kim, Jiwon Choi and Sukjoon Yoon*

Department of Biological Sciences, Research Center for Women's Diseases (RCWD), Sookmyung Women's University, Seoul 140-742, Korea

Abstract

Due to the polygenic nature of cancer, it is believed that breast cancer is caused by the perturbation of multiple genes and their complex interactions, which contribute to the wide aspects of disease phenotypes. A systems biology approach for the identification of subnetworks of interconnected genes as functional modules is required to understand the complex nature of diseases such as breast cancer. In this study, we apply a 3-step strategy for the interpretation of microarray data, focusing on identifying significantly perturbed metabolic pathways rather than analyzing a large amount of overexpressed and underexpressed individual genes. The selected pathways are considered to be dysregulated functional modules that putatively contribute to the progression of disease. The subnetwork of protein-protein interactions for these dysregulated pathways are constructed for further detailed analysis. We evaluated the method by analyzing microarray datasets of breast cancer tissues; i.e., normal and invasive breast cancer tissues. Using the strategy of microarray analysis, we selected several significantly perturbed pathways that are implicated in the regulation of progression of breast cancers, including the extracellular matrix-receptor interaction pathway and the focal adhesion pathway. Moreover, these selected pathways include several known breast cancer-related genes. It is concluded from this study that the present strategy is capable of selecting interesting perturbed pathways that putatively play a role in the progression of breast cancer and provides an improved interpretability of networks of protein-protein interactions.

Keywords: breast cancer, microarray, metabolic pathways, protein-protein interactions, systems biology

*Corresponding author: E-mail yoonsj@sookmyung.ac.kr
Tel +82-2-710-9415, Fax +82-2-2077-7322
Accepted 21 November 2008

Introduction

Microarray experiments have been a popular approach for identifying marker genes that are related to the progression of disease by providing insights into genome-wide gene expression data. Conventional analysis of microarray data has focused on finding significantly overexpressed and underexpressed genes as putative markers of disease. This has been useful in discriminating the roles of various individual genes in the progression of disease and in correlating dissected expression signatures with clinical outcomes (Dhanasekaran *et al*., 2001; Beer *et al*., 2002; van't Veer *et al*., 2002; Glinsky *et al*., 2004). However, comparing expression data between normal and diseased conditions can typically yield thousands of genes that are differentially expressed between the conditions with a statistical confidence ($p < 0.05$) (Dhanasekaran *et al*., 2001). That is, the conventional method may not be sufficient to narrow down the target pathways and genes for discriminating disease states, because only a few significantly dysregulated candidates genes can be studied in detail at any given moment. Moreover, most proteins are known to mediate their functions within regulated complex networks or pathways of interconnected macromolecules by forming dynamic topological interactomes. Additionally, genes that are not significantly altered may play a critical role with other significantly dysregulated components in their biological pathways. Therefore, a systems biology approach that can identify pathways with these proteins would significantly improve the ability to find disease-associated genes from micorarray datasets. This also would be useful in understanding the relationship between pathways and various phenotypes.

There has been a tremendous increase in information for constructing large-scale protein-protein interaction networks from public interactome databases, such as HPRD (Peri *et al*., 2004). A number of approaches have been demonstrated for identifying subnetworks of protein-protein interactions, based on coherent expression patterns of their genes (Chen and Yuan, 2006; Chuang *et al*., 2007). There also is a study that has identified candidate genes that are related to certain diseases based only on the topological features of the network of disease-related protein-protein interactions (Hwang *et al*., 2008). Recently, several methods for integrating microarray data with metabolic pathways have been pre-

sented (Setlur *et al*., 2007; Grosu *et al*., 2008). None of these approaches has mapped transcriptional changes in both metabolic pathways and protein-protein interactions. Moreover, protein-protein interaction networks have very complex topological characteristics that sometimes impose difficulties in interpretation. Therefore, it will be convenient for the interpretation if the functional modules that have significantly perturbed genes are first identified to construct a subnetwork of protein-protein interactions in each functional module. It is believed that these subnetworks of protein-protein interactions in each functional module will provide greater interpretability than the genome-wide network of protein-protein interactions. Based on these points of view, we applied 3 steps of microarray data analysis. First, differentially expressed genes were selected using the standard t test. Second, significantly perturbed metabolic pathways were selected based on those differentially expressed genes. A test for the statistical significance of the selected pathways also is presented in this study. Third, subnetworks of protein-protein interactions in those perturbed metabolic pathways were constructed for further interpretation of pathways in detail.

Breast cancer is one of many complex progressive diseases. Due to its polygenic nature, it is believed that breast cancer is caused not by single genes but rather by perturbations of multiple genes and their complex interactions, which contribute to the wide aspects of disease phenotypes. Therefore, we apply the strategy of microarray analysis using the "score of perturbation" to identify significantly perturbed pathways. To this end, we identified significantly perturbed pathways in breast cancer tissues, thereby providing interesting pathways that putatively play roles in the progression of breast cancer. Furthermore, we constructed a subnetwork of protein-protein interactions in these significantly perturbed pathways for further interpretation of pathways in detail.

Methods

We used the dataset from Turashvili *et al*. (2007), which consists of 2 types of breast cancer tissues; i.e, invasive lobular and ductal carcinomas. This dataset includes a total of 30 samples that consist of normal ductal cells from 10 patients, normal lobular cells from 10 patients, invasive ductal carcinoma cells from 5 patients, and invasive lobular carcinoma cells from 5 patients, which were microdissected from cryosections of 10 mastectomy specimens from postmenopausal patients. In this dataset, 50 nanograms of total RNA was amplified and labeled by PCR and in vitro transcription, and samples were analyzed using Affymetrix U133 Plus 2.0 Arrays. Pathways from KEGG (http://www.genome.jp/kegg/pathway.html) databases were used as pathway references for analysis.

The basic idea of our approach is to identify perturbed pathways that have relatively large amounts of overexpressed or underexpressed genes. To begin, a p value that is calculated from the standard t test is assigned to every gene in each pathway, and the number of significantly perturbed genes ($p < 0.01$) is counted in each pathway. Note that the t test for each gene is conducted by comparing 2 mean values of gene expression between 20 samples of normal breast cancer cells and 10 samples of invasive cancer cells. The score of perturbation for each pathway is assigned with the probability that we can, by chance, expect at least the same number of significantly perturbed genes in each pathway, given the number of significantly perturbed genes in the background set of genes. This probability is calculated using the cumulative hypergeometric distribution as follows:

$$P(x,r,n,N) = \sum_{i=x}^{\min(r,n)} \frac{\binom{r}{i}\binom{N-r}{n-i}}{\binom{N}{n}}$$

where n is the number of genes in each pathway, N the number of genes in whole pathways, x the number of significantly perturbed genes in each pathway, and r is the number of significantly perturbed genes in whole pathways. The pathways that have $p < 0.01$ are selected as significantly perturbed genes in breast cancer tissues.

Results and Discussion

To explore perturbed pathways in breast cancer tissues, we analyzed the microarray dataset from Turashvilli *et al*. (2007), which was downloaded from the NCBI GEO (http://www.ncbi.nlm.nih.gov/geo/) database. The dataset was standardized such that each sample array has a mean of 0 and a standard deviation of 1. The dataset contains samples of 2 breast cancer tissues and their corresponding normal cells. The standard t test was used to score genes for overexpression or underexpression in breast cancer tissues in comparison with their normal tissues. The list of significantly perturbed genes ($p < 0.01$) was classified into known biological pathways to select target pathways that are perturbed in breast cancer tissues, as described in Methods. As a result, it was found that 36 pathways were significantly perturbed, based on the score of perturbation ($p < 0.01$) in breast cancer tissues (see Supplementary Table S1).

Table 1 shows 36 significantly perturbed pathways,

Table 1. Top 36 perturbed pathways in which component genes are significantly perturbed ($p < 0.01$) in breast cancer tissues

Pathway	N1	N2	N3	N4	N5
Focal adhesion	199	19	16	35	5.08E-17
Cell junctions	134	12	17	29	1.09E-16
ECM-receptor interaction	87	15	7	22	2.44E-14
Systemic lupus erythematosus	125	23	1	24	1.05E-12
Regulation of actin cytoskeleton	217	11	14	25	2.91E-08
Axon guidance	129	8	7	15	1.94E-05
Prostate cancer	91	6	6	12	3.91E-05
Drug metabolism - cytochrome P450	67	2	8	10	6.05E-05
Colorectal cancer	85	3	8	11	9.90E-05
Cytokine-cytokine receptor interaction	273	7	14	21	0.00023
p53 signaling pathway	68	4	5	9	0.000365
Cell cycle	115	9	3	12	0.000384
Cell adhesion molecules (CAMs)	132	7	6	13	0.000388
Renal cell carcinoma	69	5	4	9	0.000408
Melanoma	71	5	4	9	0.000506
Glutathione metabolism	47	1	6	7	0.000812
Metabolism of xenobiotics by cytochrome P450	65	2	6	8	0.001268
Leukocyte transendothelial migration	116	7	4	11	0.001501
alpha-Linolenic acid metabolism	17	1	3	4	0.002006
Toll-like receptor signaling pathway	104	7	3	10	0.002214
Small cell lung cancer	87	5	4	9	0.002215
Bladder cancer	42	4	2	6	0.002405
Pancreatic cancer	73	5	3	8	0.002697
MAPK signaling pathway	269	5	13	18	0.003258
Vibrio cholerae infection	60	4	3	7	0.003474
Adherens junction	76	3	5	8	0.003477
Tight junction	135	3	8	11	0.004968
Glioma	65	4	3	7	0.005453
GnRH signaling pathway	100	5	4	9	0.005691
Biosynthesis of unsaturated fatty acids	23	1	3	4	0.00639
PPAR signaling pathway	68	2	5	7	0.006989
Nitrogen metabolism	24	1	3	4	0.007472
Complement and coagulation cascades	69	2	5	7	0.007566
TGF-beta signaling pathway	87	4	4	8	0.007923
Non-small cell lung cancer	54	4	2	6	0.008535
Neurodegenerative diseases	39	2	3	5	0.008856

Note that *N1* represents the total number of genes in each pathway, *N2* is the number of overexpressed genes ($p < 0.01$) in each pathway, *N3* is the number of underexpressed genes ($p < 0.01$) in each pathway, *N4* is the total number of significantly perturbed genes in each pathway (i.e., *N4=N2+N3*), and *N5* is the score of perturbation; i.e., p-values by the cumulative hypergeometric distribution.

including Cell Junctions (Fig. 1), the ECM-receptor interaction pathway (Fig. 2), the Focal Adhesion pathway (Fig. 3), and the p53 signaling pathway, which have been implicated to play a role in the progression of breast cancers (Behmoaram *et al.*, 2008; Fata *et al.*, 2004; Lin *et al.*, 2000; Ryan *et al.*, 2000). It is well known that most cancers lack active tumor suppressor p53, which inhibits cell growth through activation of cell cycle arrest and apoptosis and that the activation of NFKB1 is induced by p53 (Ryan *et al.*, 2000).

There are significant amounts of evidence that the ECM-receptor pathway is related to the progression of breast cancer. For instance, Fata *et al.* (2004) reviewed considerable research that indicated that mammary gland branching morphogenesis is dependent, in part, on the ECM; ECM-receptors, such as integrins and other ECM receptors; and ECM-degrading enzymes, including matrix metalloproteinases (MMPs) and their inhibitors, tissue inhibitors of metalloproteinases (TIMPs). They also provided some evidence that these ECM processes affect 1 or more of the following processes: cell survival, polarity, proliferation, differentiation, adhesion, and migration.

It is well known that breast carcinoma most often is

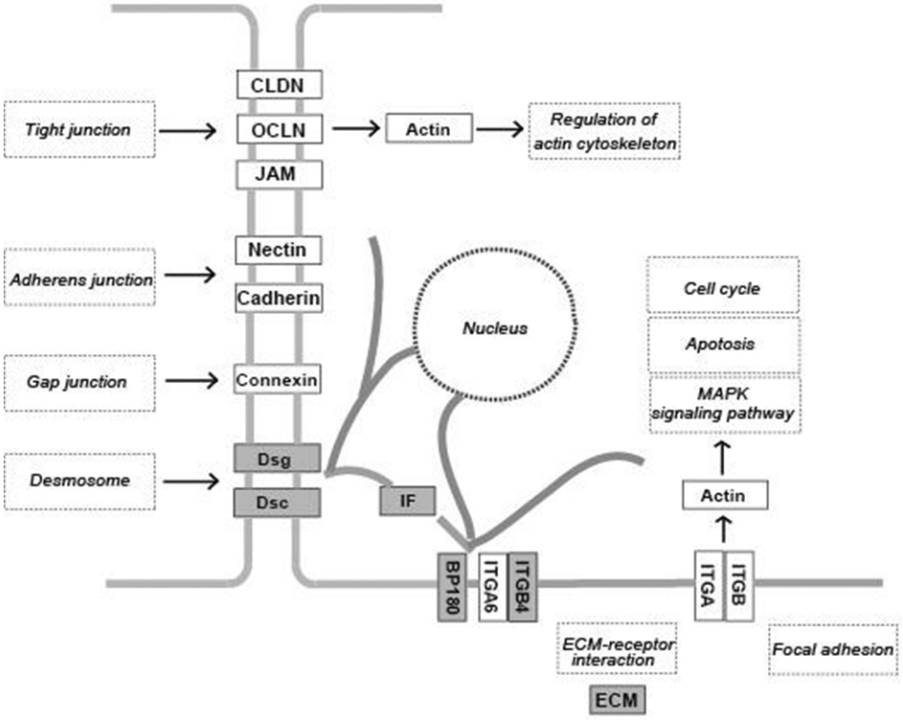

Fig. 1. Summary of the Cell *Junctions* pathway from the KEGG (http://www.genome.jp/kegg) database. Grey-colored boxes represent protein complexes with at least 1 significantly perturbed protein ($p < 0.01$) in breast cancer tissues. Note that the list of perturbed genes is tabulated in Table 2.

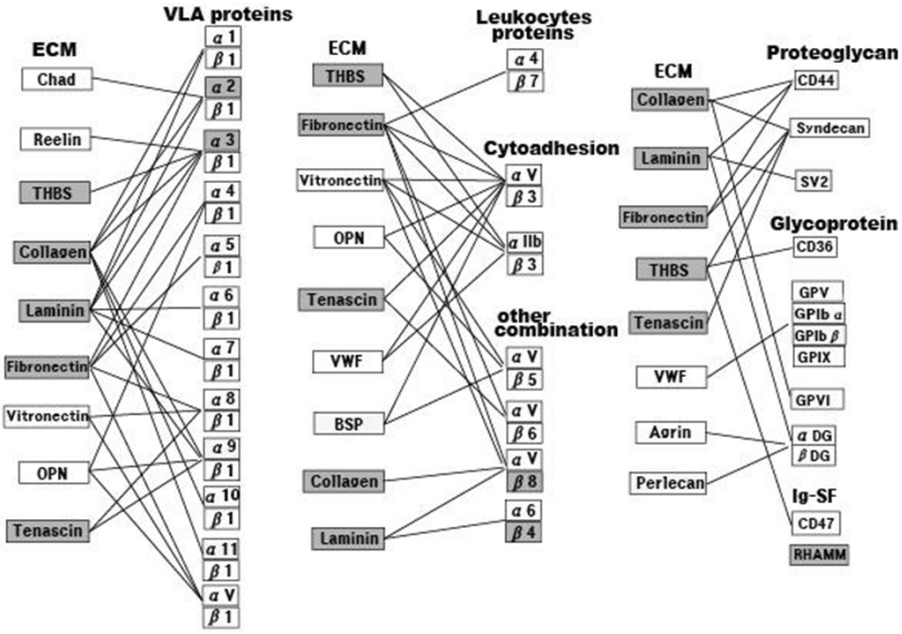

Fig. 2. Summary of the *ECM-receptor Interaction* pathway from the KEGG (http://www.genome.jp/kegg) database. Grey-colored boxes represent protein complexes with at least 1 significantly perturbed protein ($p < 0.01$) in breast cancer tissues. Note that the list of perturbed genes is tabulated in Table 2.

associated with an extensive 'stromal reaction', termed desmoplasia, in which excess collagen is deposited (Fata *et al*., 2004). It also has been shown that aberrations in the integrity, deposition, and composition of the ECM often are associated with breast cancer (Lochter and Bissell, 1995; Petersen *et al*., 2001).

In addition, upregulation of expression of the fibrillar collagen gene is an indicator of the metastatic phenotype (van't Veer *et al*., 2002; van de Vijver *et al*., 2002; Wang *et al*., 2002). The ECM and its receptors that attenuate or augment signaling regulate branching morphogenesis in a process that may be considered as controlled invasion. For instance, it has been shown that increased collagen type I upregulates activated MMP 2

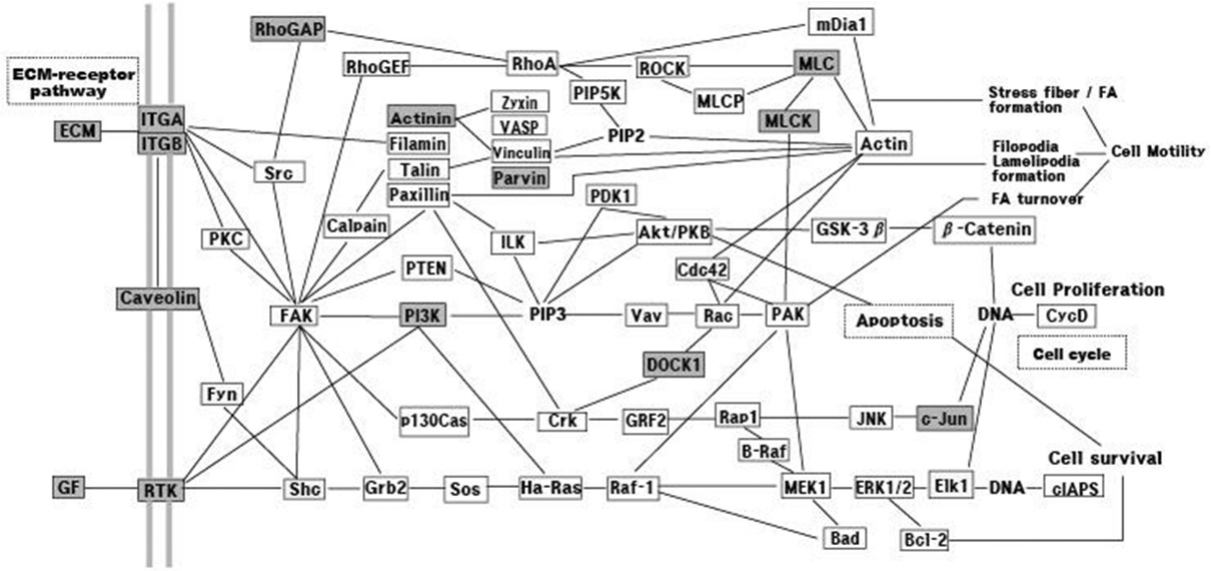

Fig. 3. Summary of the *Focal Adhesion* pathway from the KEGG (http://www.genome.jp/kegg) database. Grey-colored boxes represent protein complexes with at least 1 significantly perturbed protein ($p < 0.01$) in breast cancer tissues. Note that the list of perturbed genes is tabulated in Table 2.

Table 2. List of significantly perturbed genes ($p < 0.01$) in the *Cell Junctions, ECM-receptor interactions,* and *Focal Adhesion* pathways

Genes	Path1	Path2	Path3	Genes	Path1	Path2	Path3
COL11A1	△	△	△	KRT16	▽		
COL1A1	△	△	△	KRT17	▽		
COL1A2	△	△	△	KRT23	▽		
COL3A1	△	△	△	KRT5	▽		
COL4A1	△	△	△	KRT6B	▽		
COL5A1	△	△	△	KRT7	▽		
COL5A2	△	△	△	KRT81	▽		
COL6A1	△	△	△	FNDC1		△	
COL6A3	△	△	△	FNDC3A		△	
COMP	△	△	△	HMMR		△	
THBS2	△	△	△	CAV1			▽
FN1	△	△	△	ACTN1			▽
LAMB3	▽	▽	▽	ARHGAP5			▽
LAMC2	▽	▽	▽	DOCK1			△
TNR	▽	▽	▽	EGFR			▽
ITGB4	▽	▽	▽	JUN			▽
ITGA2		▽	▽	MET			▽
ITGA3		▽	▽	MYL9			▽
ITGB8		▽	▽	MYLK			▽
COL17A1	▽			PARVB			△
DSC3	▽			PDGFB			△
DSG3	▽			PDGFRA			▽
DSG4	▽			PIK3CB			△
KRT14	▽			VEGFA			△
KRT15	▽						

△ corresponds to significantly overexpressed genes and ▽ to significantly underexpressed genes. Note that *Path1* represents the *Cell Junctions* pathway, *Path2* is the *ECM-receptor interactions* pathway, and *Path3* is the *Focal Adhesion* pathway. It is noteworthy that there are several genes that are involved in more than 1 pathway, allowing crosstalk between pathways.

in human metastatic breast cancer cells (Thompson *et al*., 1994). Other collagens, such as types III, V, and VII, also are altered with regard to expression and deposition in breast cancer (Barsky *et al*., 1982; Fukuda *et al*., 2000; Lagace *et al*., 1985; Wetzels *et al*., 1991), triggering signals that lead to the loss of structure and function in breast.

There also are several indications that the focal adhesion pathway is related to the progression of breast cancer. For instance, Lin *et al*. (2000) reported a direct effect of progesterone in inducing the spread and adhesion of breast cancer cells, with the conclusion that progesterone-induced cell spreading and focal adhesion may have significant implications in breast tumor metastasis. In addition, there is crosstalk between the ECM-receptor pathway and the focal adhesion pathway, in which several proteins bind to form ECMs that bind to their receptors, triggering signaling cascades within the focal adhesion pathway and leading to cell motility, cell proliferation, and cell survival (Fig. 2, 3).

Table 2 shows overexpressed and underexpressed genes in Cell Junctions, the ECM-receptor interaction pathway, and the Focal Adhesion pathway, including THBS2, PDGF, COL1A1, COLA2, COL3A1, COL5A1, and COL5A2. There are several indications that these genes are associated with cancer. For instance, THBS2 has been shown to function as a potent inhibitor of tumor growth and angiogenesis (Potikyan *et al*., 2007; Hawighorst *et al*., 2001). PDGF is known to activate the RAS/PIK3/AKT1/IKK/NFKB1 pathway, in which NFKB1 induces putative antiapoptotic genes (Romashkova *et al*., 1999). Collagen type I (COL1A1, COL1A2), type III (COL3A1), and type V (COL5A1, COL5A2) are implicated in playing roles in the progression of metastatic breast cancer (Barsky *et al*., 1982; Fukuda *et al*., 2000; Lagace *et al*., 1985; Thompson *et al*., 1994; Wetzels *et al*., 1991).

Based on the selected perturbed pathways, we combined selected metabolic pathways with protein-protein interaction information by constructing a subnetwork of protein-protein interactions (e.g., Fig. 4). To contruct a subnetwork of protein-protein interactions for each pathway, information on protein-protein interactions was extracted from the Human Protein Reference Database (HPRD) (Peri *et al*., 2004). Fig. 4 shows that protein complexes can be identified based on the definitions in individual metabolic pathways, in which the protein-protein interactions can be categorized into intra- or inter-pathway interactions. It also is possible to identify significantly perturbed proteins within the protein complexes for a more detailed analysis of the pathways.

Based on perturbation score, we present 36 significantly perturbed pathways, instead of collecting a large amount of significantly dysregulated individual genes. The selected pathways are then considered to be dysregulated functional modules that putatively con-

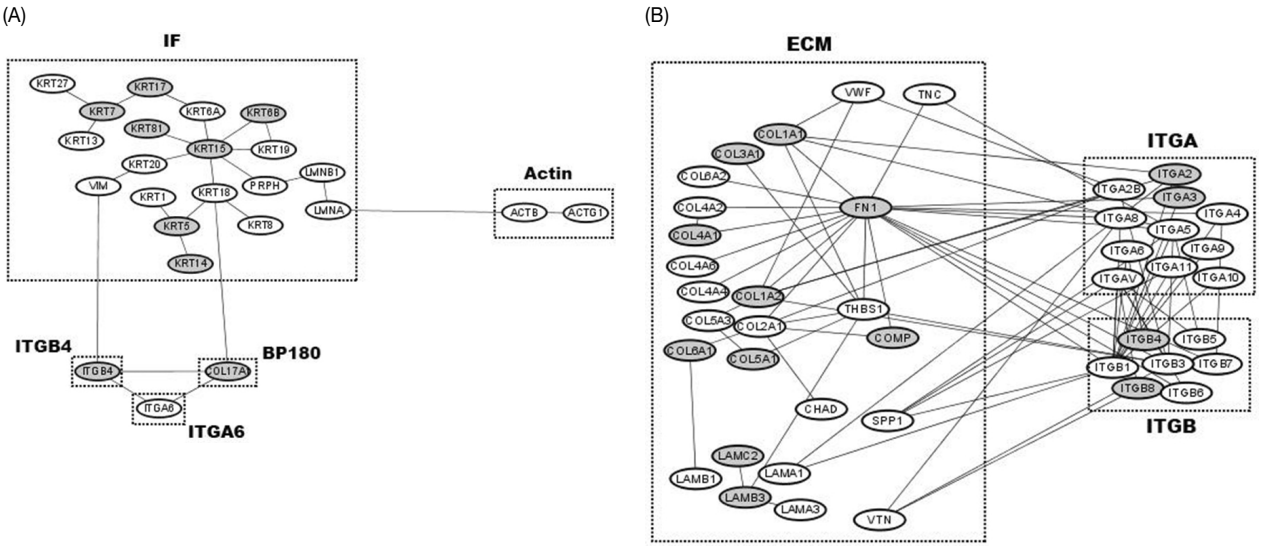

Fig. 4. (A) Subnetwork of protein-protein interactions in the *Cell Junctions* pathway. The rectangular boxes with dashed lines represent the protein complexes of *IF, Actin, ITGB4, ITGA6,* and *BP180,* respectively (see Fig. 1). (B) Subnetwork of protein-protein interactions in both the *ECM-receptor Interaction* and *Focal Adhesion* pathways. The rectangular boxes with dashed lines represent the protein complexes of *ECM, ITGA,* and *ITGB,* respectively (see Fig. 3). Note that 3 protein complexes in (B) can be subdivided into smaller protein complexes based on the definition of the *ECM-receptor* pathway (Fig. 2). Note also that grey-colored nodes represent significantly perturbed genes ($p < 0.01$) in breast cancer tissues.

tribute to the progression of disease. The result of this study suggests that the strategy of microarray analysis, using the score of perturbation, selects several interesting perturbed pathways that are implicated in the progression of breast cancer. It also was found that these selected pathways include several known breast cancer-related genes. Therefore, based on the selected pathways, this study sets the stage for further investigation of the basic mechanisms that serve as a basis for discriminating different breast cancer types to find new therapeutic drug targets.

Acknowledgments

This research was supported by Sookmyung Women's University Research Grants 1-0703-0148.

References

Barsky, S.H., Rao, C.N., Grotendorst, G.R., and Liotta, L.A. (1982). Increased content of Type V Collagen in desmoplasia of human breast carcinoma. *Am. J. Pathol.* 108, 276-283.

Beer, D.G., Kardia, S.L., Huang, C.C., *et al.* (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med.* 8, 816-824.

Behmoaram, E., Bijian, K., Jie, S., Xu, Y., Darnel, A., Bismar, T.A., and Alaoui-Jamali, M.A. (2008). Focal adhesion kinase-related proline-rich tyrosine kinase 2 and focal adhesion kinase are co-overexpressed in early-stage and invasive ErbB-2-positive breast cancer and cooperate for breast cancer cell tumorigenesis and invasiveness. *American Journal of Pathology* 173, 1540-1550.

Chen, J., and Yuan, B. (2006). Detecting functional modules in the yeast protein-protein interaction network. *Bioinformatics* 22, 2283-2290.

Chuang, H.Y., Lee, E., Liu, Y.T., Lee, D., and Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Molecular Systems Biology* 3, 140.

Dhanasekaran, S.M., Barrette, T.R., Ghosh, D., *et al.* (2001). Delineation of prognostic biomarkers in prostate cancer. *Nature* 412, 822-826.

Fata, J.E., Werb, Z., and Bissell, M.J. (2004). Regulation of mammary gland branching morphogenesis by the extracellular matrix and its remodeling enzymes. *Breast Cancer Res.* 6, 1-11.

Fukuda, Y., Ishizaki, M., Okada, Y., Seiki, M., and Yamanaka, N. (2000). Matrix metalloproteinases and tissue inhibitor of metalloproteinase-2 in fetal rabbit lung. *Am. J. Physiol. Lung Cell Mol. Physiol.* 279, 555-561.

Glinsky, G.V., Glinskii, A.B., Stephenson, A.J., Hoffman, R.M., and Gerald, W.L. (2004). Gene expression profiling predicts clinical outcome of prostate cancer. *J. Clin. Invest.* 113, 913-923.

Grosu, P., Townsend, J.P., Hartl, D.L., and Cavalieri, D. (2008). Pathway processor: a tool for integrating whole-genome expression results into metabolic networks. *Genome Research* 12, 1121-1126.

Hawighorst, T., Velasco, P., Streit, M., Hong, Y.K., Kyriakides, T.R., Brown, L.R., Bornstein, P., and Detmar, M. (2001). Thrombospondin-2 plays a protective role in multistep carcinogenesis: a novel host anti-tumor defense mechanism. *The EMBO Journal* 20, 2631-2640.

Hwang, S., Son, S.W., Kim, S.C., Kim, Y.J., Jeong, H., and Lee, D. (2008). A protein interaction network associated with asthma. *Journal of Theoretical Biology* 252, 722-731.

Lagace, R., Grimaud, J.A., Schurch, W., and Seemayer, T.A. (1985). Myofibroblastic stromal reaction in carcinoma of the breast: variations of collagenous matrix and structural glycoproteins. *Virchows Arch. A. Pathol. Anat. Histopathol.* 408, 49-59.

Lin, V.C., Ng, E.H., Aw, S.E., Tan, M.G., Ng, E.H., and Bay, B.H. (2000). Progesterone Induces Focal Adhesion in Breast Cancer Cells MDA-MB-231 Transfected with progesterone receptor complementary DNA. *Molecular Endocrinology* 14, 348-358.

Lochter, A., and Bissell, M.J. (1995). Involvement of extracellular matrix constituents in breast cancer. *Semin. Cancer Biol.* 6, 165-173.

Peri, S., Navarro, J.D., Kristiansen, T.Z., Amanchy, R., Surendranath, V., Muthusamy, B., Gandhi, T.K., Chandrika, K.N., Deshpande, N., Suresh, S., Rashmi, B.P., Shanker, K., Padma, N., Niranjan, V., Harsha, H.C., Talreja, N., Vrushabendra, B.M., Ramya, M.A., Yatish, A.J., Joy, M., Shivashankar, H.N., Kavitha, M.P., Menezes, M., Choudhury, D.R., Ghosh, N., Saravana, R., Chandran, S., Mohan, S., Jonnalagadda, C.K., Prasad, C.K., Kumar-Sinha, C., Deshpande, K.S., and Pandey, A. (2004). Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res.* 32, D497-D501.

Petersen, O.W., Lind, N.H., Gudjonsson, T., Villadsen, R., Ronnov-Jessen, L., and Bissell, M.J. (2001). The plasticity of human breast carcinoma cells is more than epithelial to mesenchymal conversion. *Breast Cancer Res.* 3, 213-217.

Potikyan, G., Savene, O.V., Gaulden, J.M., France, K.A., Zhou, Z., Kleinerman, E.S., Lessnick, S.L., and Denny, C.T. (2007). EWS/FLI1 regulates tumor angiogenesis in Ewing's sarcoma via suppression of thrombospondins. *Cancer Research* 67, 6675-6684.

Romashkova, J.A., and Makarov, S.S. (1999). NF-kappa-B is a target of AKT in anti-apoptotic PDGF signalling. *Nature* 401, 86-90.

Ryan, K.M., Ernst, M.K., Rice, N.R., and Vousden, K.H. (2000). Role of NF-kappa-B in p53-mediated programmed cell death. *Nature* 404, 892-897.

Setlur, S.R., Royce, T.E., Sboner, A., Mosquera, J.M., Demichelis, F., Hofer, M.D., Mertz, K.D., Gerstein, M., and Rubin, M.A. (2007). Integrative microarray analysis of pathways dysregulated in metastatic prostate cancer. *Cancer Res.* 67, 10296-10303.

Thompson, E.W., Yu, M., Bueno, J., Jin, L., Maiti, S.N., Palao-Marco, F.L., Pulyaeva, H., Tamborlane, J.W., Tirgari, R., Wapnir, I., *et al.* (1994). Collagen induced MMP-2 activation in human breast cancer. *Breast Cancer*

Res. Treat. 31, 357-370.

Turashvili, G., Bouchal, J., Baumforth, K., Wei, W., *et al.* (2007). Novel markers for differentiation of lobular and ductal invasive breast carcinomas by laser microdissection and microarray analysis. *BMC Cancer* 7, 55.

van de Vijver, M.J., He, Y.D., van't Veer, L.J., Dai, H., Hart, A.A., Voskuil, D.W., Schreiber, G.J., Peterse, J.L., Roberts, C., Marton, M.J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E.T., Friend, S.H., and Bernards, R. (2002). A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* 347, 1999-2009.

van't Veer, L.J., Dai, H., van de Vijver, M.J., *et al.* (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415, 530-536.

van't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A., Bernards, R., and Friend, S.H. (2002). Expression profiling predicts outcome in breast cancer. *Breast Cancer Res.* 5, 57-58.

Wang, W., Wyckoff, J.B., Frohlich, V.C., Oleynikov, Y., Huttelmaier, S., Zavadil, J., Cermak, L., Bottinger, E.P., Singer, R.H., White, J.G., Segall, J.E., and Condeelis, J.S. (2002). Single cell behavior in metastatic primary mammary tumors correlated with gene expression patterns revealed by molecular profiling. *Cancer Res.* 62, 6278-6288.

Wetzels, R.H., Robben, H.C., Leigh, I.M., Schaafsma, H.E., Vooijs, G.P., and Ramaekers, F.C. (1991). Distribution patterns of type VII collagen in normal and malignant human tissues. *Am. J. Pathol.* 139, 451-459.

Supplementary Table S1. The list of all pathways from KEGG database, which are sorted according to the scores of perturbation. Note that *N1* represents the total number of genes in each pathway, *N2* is the number of overexpressed genes ($p < 0.01$) in each pathway, *N3* is the number of underexpressed genes ($p < 0.01$) in each pathway, *N4* is the total number of significantly perturbed genes in each pathway (i.e., *N4=N2+N3*), and *N5* is the score of perturbation i.e., p-values by the cumulative hypergeometric distribution

KEGG Pathway Description	# of genes in pathway	# of genes linked to GPL570 Probes	$p < 0.05$ # of Over Expressed Genes		$p < 0.01$ # of Over Expressed Genes		P1	P2	P3
			Lobular Carcinoma	Ductal Carcinoma	Lobular Carcinoma	Ductal Carcinoma			
Systemic lupus erythematosus	134	122	6	55	1	34	0.8	27.9	14.3
Keratan sulfate biosynthesis	16	16	1	4	1	3	6.3	18.8	12.5
ECM-receptor interaction	88	87	18	25	8	11	9.2	12.6	10.9
Chondroitin sulfate biosynthesis	22	22	1	5	1	2	4.5	9.1	6.8
Cell junctions	138	134	15	26	8	10	6.0	7.5	6.7
Maturity onset diabetes of the young	25	23	2	7	1	2	4.3	8.7	6.5
alpha-Linolenic acid metabolism	17	17	2	3	1	1	5.9	5.9	5.9
Linoleic acid metabolism	29	29	1	4	1	2	3.4	6.9	5.2
Focal adhesion	200	198	28	29	10	10	5.1	5.1	5.1
Glycosphingolipid biosynthesis - lactoseries	10	10	0	3	0	1	0.0	10.0	5.0
Glycosphingolipid biosynthesis - neo-lactoseries	21	21	2	4	0	2	0.0	9.5	4.8
Reductive carboxylate cycle (CO2 fixation)	11	11	1	1	0	1	0.0	9.1	4.5
Ether lipid metabolism	33	33	4	4	2	1	6.1	3.0	4.5
Nitrogen metabolism	24	24	2	1	1	1	4.2	4.2	4.2
Valine, leucine and isoleucine biosynthesis	12	12	3	2	0	1	0.0	8.3	4.2
Glycan structures - biosynthesis 2	63	63	4	11	1	4	1.6	6.3	4.0
Prion disease	14	13	1	2	0	1	0.0	7.7	3.8
Protein export	15	14	1	2	1	0	7.1	0.0	3.6
Graft-versus-host disease	42	42	3	8	0	3	0.0	7.1	3.6
Glycosphingolipid biosynthesis - globoseries	14	14	0	2	0	1	0.0	7.1	3.6
Sulfur metabolism	14	14	4	2	0	1	0.0	7.1	3.6
Bladder cancer	42	42	8	9	2	1	4.8	2.4	3.6
GnRH signaling pathway	100	100	10	12	5	2	5.0	2.0	3.5
ABC transporters - General	44	43	3	4	2	1	4.7	2.3	3.5
Glycerophospholipid metabolism	72	72	8	5	2	3	2.8	4.2	3.5
VEGF signaling pathway	73	73	6	9	3	2	4.1	2.7	3.4
Cell adhesion molecules (CAMs)	133	132	7	24	1	8	0.8	6.1	3.4
Type I diabetes mellitus	44	44	4	8	0	3	0.0	6.8	3.4
Toll-like receptor signaling pathway	107	104	11	17	3	4	2.9	3.8	3.4
Pyrimidine metabolism	91	90	6	18	1	5	1.1	5.6	3.3
Riboflavin metabolism	16	16	1	1	1	0	6.3	0.0	3.1
One carbon pool by folate	16	16	2	4	0	1	0.0	6.3	3.1
Cell cycle	119	115	14	22	0	7	0.0	6.1	3.0
Propanoate metabolism	34	33	2	3	1	1	3.0	3.0	3.0
Renin-angiotensin system	17	17	2	2	1	0	5.9	0.0	2.9
Cysteine metabolism	17	17	2	1	0	1	0.0	5.9	2.9
Glycan structures - biosynthesis 1	123	121	8	21	2	5	1.7	4.1	2.9
Basal transcription factors	37	35	1	5	0	2	0.0	5.7	2.9
Autoimmune thyroid disease	53	53	5	10	0	3	0.0	5.7	2.8
Arachidonic acid metabolism	56	55	1	4	1	2	1.8	3.6	2.7
Nicotinate and nicotinamide metabolism	37	37	1	4	0	2	0.0	5.4	2.7
T cell receptor signaling pathway	93	93	6	17	1	4	1.1	4.3	2.7
Ribosome	91	75	0	15	0	4	0.0	5.3	2.7
Fructose and mannose metabolism	38	38	2	3	1	1	2.6	2.6	2.6
Allograft rejection	38	38	4	6	0	2	0.0	5.3	2.6
Fc epsilon RI signaling pathway	77	77	7	7	3	1	3.9	1.3	2.6

Supplementary Table S1. Continued

KEGG Pathway Description	# of genes in pathway	# of genes linked to GPL570 Probes	p<0.05 # of Over Expressed Genes		p<0.01 # of Over Expressed Genes		P1	P2	P3
			Lobular Carcinoma	Ductal Carcinoma	Lobular Carcinoma	Ductal Carcinoma			
Long-term depression	78	78	4	10	2	2	2.6	2.6	2.6
Parkinson's disease	20	20	0	5	0	1	0.0	5.0	2.5
Glycosphingolipid biosynthesis - ganglioseries	21	20	1	4	0	1	0.0	5.0	2.5
Glycolysis / Gluconeogenesis	62	61	3	9	0	3	0.0	4.9	2.5
Methionine metabolism	21	21	1	3	0	1	0.0	4.8	2.4
Amyotrophic lateral sclerosis (ALS)	21	21	0	2	0	1	0.0	4.8	2.4
Glioma	65	65	8	9	2	1	3.1	1.5	2.3
Metabolism of xenobiotics by cytochrome P450	70	65	1	5	0	3	0.0	4.6	2.3
TGF-beta signaling pathway	90	87	9	10	1	3	1.1	3.4	2.3
Valine, leucine and isoleucine degradation	44	44	1	3	1	1	2.3	2.3	2.3
N-Glycan biosynthesis	45	44	2	2	1	1	2.3	2.3	2.3
Prostate cancer	91	91	13	13	2	2	2.2	2.2	2.2
Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	23	23	1	1	1	0	4.3	0.0	2.2
Epithelial cell signaling in Helicobacter pylori infection	69	69	7	9	2	1	2.9	1.4	2.2
Mismatch repair	23	23	1	2	0	1	0.0	4.3	2.2
Melanoma	71	71	8	13	2	1	2.8	1.4	2.1
Pantothenate and CoA biosynthesis	24	24	0	3	0	1	0.0	4.2	2.1
B cell receptor signaling pathway	73	73	4	16	1	2	1.4	2.7	2.1
mTOR signaling pathway	51	50	6	6	1	1	2.0	2.0	2.0
Inositol phosphate metabolism	51	51	7	5	0	2	0.0	3.9	2.0
Galactose metabolism	26	26	2	3	0	1	0.0	3.8	1.9
Phenylalanine metabolism	26	26	0	2	0	1	0.0	3.8	1.9
Non-small cell lung cancer	54	54	7	8	1	1	1.9	1.9	1.9
Hedgehog signaling pathway	57	55	2	9	0	2	0.0	3.6	1.8
Urea cycle and metabolism of amino groups	28	28	1	1	1	0	3.6	0.0	1.8
Hematopoietic cell lineage	87	86	6	16	0	3	0.0	3.5	1.7
Leukocyte transendothelial migration	116	115	9	13	2	2	1.7	1.7	1.7
Thyroid cancer	29	29	3	6	1	0	3.4	0.0	1.7
Small cell lung cancer	87	87	10	16	0	3	0.0	3.4	1.7
Aminosugars metabolism	29	29	2	2	0	1	0.0	3.4	1.7
Purine metabolism	147	146	10	20	1	4	0.7	2.7	1.7
Wnt signaling pathway	149	146	7	14	1	4	0.7	2.7	1.7
Antigen processing and presentation	88	88	7	11	0	3	0.0	3.4	1.7
Vibrio cholerae infection	59	59	4	10	1	1	1.7	1.7	1.7
Olfactory transduction	383	118	3	13	0	4	0.0	3.4	1.7
MAPK signaling pathway	269	268	15	37	4	5	1.5	1.9	1.7
Retinol metabolism	65	60	4	4	1	1	1.7	1.7	1.7
Glutamate metabolism	31	31	4	2	1	0	3.2	0.0	1.6
Gap junction	96	94	5	13	2	1	2.1	1.1	1.6
Huntington's disease	32	32	1	1	1	0	3.1	0.0	1.6
Axon guidance	128	128	10	18	1	3	0.8	2.3	1.6
Bile acid biosynthesis	33	33	1	2	0	1	0.0	3.0	1.5
Drug metabolism - cytochrome P450	72	67	1	6	0	2	0.0	3.0	1.5
p53 signaling pathway	69	68	9	16	0	2	0.0	2.9	1.5
Renal cell carcinoma	69	69	6	8	1	1	1.4	1.4	1.4
Natural killer cell mediated cytotoxicity	141	138	10	18	1	3	0.7	2.2	1.4
Complement and coagulation cascades	69	69	3	6	0	2	0.0	2.9	1.4
Long-term potentiation	70	70	1	10	1	1	1.4	1.4	1.4

Supplementary Table S1. Continued

KEGG Pathway Description	# of genes in pathway	# of genes linked to GPL570 Probes	p<0.05 # of Over Expressed Genes		p<0.01 # of Over Expressed Genes		P1	P2	P3
			Lobular Carcinoma	Ductal Carcinoma	Lobular Carcinoma	Ductal Carcinoma			
Base excision repair	35	35	1	7	0	1	0.0	2.9	1.4
Primary immunodeficiency	35	35	2	6	0	1	0.0	2.9	1.4
Arginine and proline metabolism	35	35	3	2	0	1	0.0	2.9	1.4
Regulation of actin cytoskeleton	219	216	19	28	3	3	1.4	1.4	1.4
DNA replication	36	36	2	5	0	1	0.0	2.8	1.4
Pancreatic cancer	73	73	8	12	2	0	2.7	0.0	1.4
Folate biosynthesis	39	38	3	4	1	0	2.6	0.0	1.3
Neurodegenerative Diseases	39	39	0	3	0	1	0.0	2.6	1.3
Phosphatidylinositol signaling system	80	80	7	7	0	2	0.0	2.5	1.3
Pyruvate metabolism	42	42	1	4	0	1	0.0	2.4	1.2
Colorectal cancer	85	85	9	7	2	0	2.4	0.0	1.2
Type II diabetes mellitus	44	43	2	8	0	1	0.0	2.3	1.2
Nucleotide excision repair	43	43	1	5	0	1	0.0	2.3	1.2
ErbB signaling pathway	87	87	5	7	2	0	2.3	0.0	1.1
Apoptosis	89	88	6	12	0	2	0.0	2.3	1.1
Notch signaling pathway	46	46	3	3	1	0	2.2	0.0	1.1
Taste transduction	53	46	2	8	0	1	0.0	2.2	1.1
Histidine metabolism	50	50	0	4	0	1	0.0	2.0	1.0
Endometrial cancer	52	52	6	5	1	0	1.9	0.0	1.0
Basal cell carcinoma	55	54	1	5	0	1	0.0	1.9	0.9
Acute myeloid leukemia	58	58	6	8	1	0	1.7	0.0	0.9
Tyrosine metabolism	58	58	3	3	0	1	0.0	1.7	0.9
Neuroactive ligand-receptor interaction	303	302	9	39	1	4	0.3	1.3	0.8
Oxidative phosphorylation	129	125	8	23	1	1	0.8	0.8	0.8
Tight junction	135	134	5	14	1	1	0.7	0.7	0.7
PPAR signaling pathway	69	68	7	4	0	1	0.0	1.5	0.7
Cytokine-cytokine receptor interaction	279	273	17	35	1	3	0.4	1.1	0.7
Insulin signaling pathway	139	138	6	12	2	0	1.4	0.0	0.7
Adherens junction	75	75	2	6	0	1	0.0	1.3	0.7
Chronic myeloid leukemia	76	76	7	12	1	0	1.3	0.0	0.7
Jak-STAT signaling pathway	155	155	15	22	1	1	0.6	0.6	0.6
Calcium signaling pathway	176	176	9	23	1	1	0.6	0.6	0.6
Melanogenesis	102	101	5	9	1	0	1.0	0.0	0.5
Ubiquitin mediated proteolysis	136	133	8	24	0	1	0.0	0.8	0.4
Adipocytokine signaling pathway	72	72	5	10	0	0	0.0	0.0	0.0
Starch and sucrose metabolism	79	75	3	7	0	0	0.0	0.0	0.0
Pathogenic Escherichia coli infection - EHEC	51	49	0	6	0	0	0.0	0.0	0.0
Pathogenic Escherichia coli infection - EPEC	51	49	0	6	0	0	0.0	0.0	0.0
Biosynthesis of steroids	24	24	0	5	0	0	0.0	0.0	0.0
O-Glycan biosynthesis	31	31	3	5	0	0	0.0	0.0	0.0
Heparan sulfate biosynthesis	20	19	2	5	0	0	0.0	0.0	0.0
Sphingolipid metabolism	39	38	3	5	0	0	0.0	0.0	0.0
Terpenoid biosynthesis	6	6	0	5	0	0	0.0	0.0	0.0
Glycine, serine and threonine metabolism	42	42	1	4	0	0	0.0	0.0	0.0
Tryptophan metabolism	58	58	0	4	0	0	0.0	0.0	0.0
Fatty acid metabolism	46	46	3	3	0	0	0.0	0.0	0.0
Alanine and aspartate metabolism	33	33	1	3	0	0	0.0	0.0	0.0
Lysine degradation	52	52	1	3	0	0	0.0	0.0	0.0
Glutathione metabolism	50	47	1	3	0	0	0.0	0.0	0.0
Glycosaminoglycan degradation	17	17	0	3	0	0	0.0	0.0	0.0

Supplementary Table S1. Continued

KEGG Pathway Description	# of genes in pathway	# of genes linked to GPL570 Probes	p < 0.05 # of Over Expressed Genes		p < 0.01 # of Over Expressed Genes		P1	P2	P3
			Lobular Carci-noma	Ductal Carci-noma	Lobular Carci-noma	Ductal Carci-noma			
Benzoate degradation via CoA ligation	23	23	0	3	0	0	0.0	0.0	0.0
Butanoate metabolism	36	36	0	3	0	0	0.0	0.0	0.0
Carbon fixation	24	24	0	3	0	0	0.0	0.0	0.0
Atrazine degradation	9	9	0	3	0	0	0.0	0.0	0.0
Porphyrin and chlorophyll metabolism	41	37	0	3	0	0	0.0	0.0	0.0
Aminoacyl-tRNA biosynthesis	39	39	4	3	0	0	0.0	0.0	0.0
Glycan structures - degradation	30	30	0	3	0	0	0.0	0.0	0.0
Biosynthesis of unsaturated fatty acids	23	23	2	3	0	0	0.0	0.0	0.0
Homologous recombination	28	28	0	3	0	0	0.0	0.0	0.0
Regulation of autophagy	34	33	2	3	0	0	0.0	0.0	0.0
Pentose phosphate pathway	26	26	0	2	0	0	0.0	0.0	0.0
Androgen and estrogen metabolism	55	52	1	2	0	0	0.0	0.0	0.0
gamma-Hexachlorocyclohexane degradation	18	18	0	2	0	0	0.0	0.0	0.0
Glycerolipid metabolism	51	50	4	2	0	0	0.0	0.0	0.0
Alkaloid biosynthesis II	20	20	0	2	0	0	0.0	0.0	0.0
Drug metabolism - other enzymes	52	49	0	2	0	0	0.0	0.0	0.0
RNA polymerase	25	25	2	2	0	0	0.0	0.0	0.0
Proteasome	35	35	1	2	0	0	0.0	0.0	0.0
Non-homologous end-joining	14	13	0	2	0	0	0.0	0.0	0.0
Alzheimer's disease	28	28	1	2	0	0	0.0	0.0	0.0
Dentatorubropallidoluysian atrophy (DRPLA)	15	15	1	2	0	0	0.0	0.0	0.0
Citrate cycle (TCA cycle)	28	27	1	1	0	0	0.0	0.0	0.0
Synthesis and degradation of ketone bodies	9	9	0	1	0	0	0.0	0.0	0.0
C21-Steroid hormone metabolism	11	11	0	1	0	0	0.0	0.0	0.0
Lysine biosynthesis	5	5	0	1	0	0	0.0	0.0	0.0
beta-Alanine metabolism	24	24	0	1	0	0	0.0	0.0	0.0
Selenoamino acid metabolism	32	32	1	1	0	0	0.0	0.0	0.0
1- and 2-Methylnaphthalene degradation	19	19	1	1	0	0	0.0	0.0	0.0
Glyoxylate and dicarboxylate metabolism	15	15	2	1	0	0	0.0	0.0	0.0
3-Chloroacrylic acid degradation	14	14	1	1	0	0	0.0	0.0	0.0
Limonene and pinene degradation	24	24	0	1	0	0	0.0	0.0	0.0
Caprolactam degradation	7	7	0	1	0	0	0.0	0.0	0.0
SNARE interactions in vesicular transport	38	38	1	1	0	0	0.0	0.0	0.0
Asthma	30	30	1	1	0	0	0.0	0.0	0.0
Inositol metabolism	2	2	0	0	0	0	0.0	0.0	0.0
Pentose and glucuronate interconversions	25	22	0	0	0	0	0.0	0.0	0.0
Ascorbate and aldarate metabolism	9	9	0	0	0	0	0.0	0.0	0.0
Fatty acid biosynthesis	6	6	0	0	0	0	0.0	0.0	0.0
Fatty acid elongation in mitochondria	10	10	0	0	0	0	0.0	0.0	0.0
Ubiquinone biosynthesis	15	13	1	0	0	0	0.0	0.0	0.0
Caffeine metabolism	7	7	0	0	0	0	0.0	0.0	0.0
Geraniol degradation	11	11	0	0	0	0	0.0	0.0	0.0
Bisphenol A degradation	5	5	0	0	0	0	0.0	0.0	0.0
Fluorobenzoate degradation	1	1	0	0	0	0	0.0	0.0	0.0
Phenylalanine, tyrosine and tryptophan biosyn-thesis	9	9	0	0	0	0	0.0	0.0	0.0
Novobiocin biosynthesis	3	3	0	0	0	0	0.0	0.0	0.0
Taurine and hypotaurine metabolism	10	10	0	0	0	0	0.0	0.0	0.0
Aminophosphonate metabolism	17	17	0	0	0	0	0.0	0.0	0.0
Cyanoamino acid metabolism	9	9	0	0	0	0	0.0	0.0	0.0

Supplementary Table S1. Continued

KEGG Pathway Description	# of genes in pathway	# of genes linked to GPL570 Probes	p<0.05 # of Over Expressed Genes		p<0.01 # of Over Expressed Genes		P1	P2	P3
			Lobular Carci-noma	Ductal Carci-noma	Lobular Carci-noma	Ductal Carci-noma			
D-Glutamine and D-glutamate metabolism	4	4	1	0	0	0	0.0	0.0	0.0
D-Arginine and D-ornithine metabolism	1	1	0	0	0	0	0.0	0.0	0.0
N-Glycan degradation	16	16	0	0	0	0	0.0	0.0	0.0
Nucleotide sugars metabolism	6	6	3	0	0	0	0.0	0.0	0.0
Streptomycin biosynthesis	10	10	1	0	0	0	0.0	0.0	0.0
Peptidoglycan biosynthesis	5	5	0	0	0	0	0.0	0.0	0.0
Tetrachloroethene degradation	3	3	0	0	0	0	0.0	0.0	0.0
1,4-Dichlorobenzene degradation	1	1	0	0	0	0	0.0	0.0	0.0
Styrene degradation	3	3	1	0	0	0	0.0	0.0	0.0
C5-Branched dibasic acid metabolism	2	2	0	0	0	0	0.0	0.0	0.0
Methane metabolism	7	7	0	0	0	0	0.0	0.0	0.0
Thiamine metabolism	8	8	0	0	0	0	0.0	0.0	0.0
Vitamin B6 metabolism	5	5	0	0	0	0	0.0	0.0	0.0
Biotin metabolism	4	4	0	0	0	0	0.0	0.0	0.0
Lipoic acid metabolism	2	2	0	0	0	0	0.0	0.0	0.0
Monoterpenoid biosynthesis	2	2	0	0	0	0	0.0	0.0	0.0
Phenylpropanoid biosynthesis	4	4	0	0	0	0	0.0	0.0	0.0
Alkaloid biosynthesis I	5	5	0	0	0	0	0.0	0.0	0.0
Circadian rhythm	13	13	0	0	0	0	0.0	0.0	0.0
						Mean:	0.9	2.1	1.5