ANTICANCER RESEARCH *27*: 1295-1300 (2007)

A Microarray Approach to Translational Medicine in Breast Cancer: How Representative are Cell Line Models of Clinical Conditions?

JAI PRAKASH MEHTA, LORRAINE O'DRISCOLL, NIALL BARRON, MARTIN CLYNES and PADRAIG DOOLAN

National Institute for Cellular Biotechnology, Dublin City University, Dublin, Ireland

Abstract. *The aim of this study was to examine whether the degree to which cell lines model corresponding cells in vivo is an important aspect of their value as models in studying disease processes. Materials and Methods: The work presented here utilizes gene expression data from two published microarray datasets to compare the differences and similarities among the two systems in order to identify major transcriptional changes in the adaptation process from a tissue to a cell line. Results: Gene ontology and pathway analyses of comparator gene lists showed that the cell cycle related genes were significantly up-regulated in cell lines and immune response related genes were significantly up-regulated in clinical specimens. Estrogen receptor analysis also indicated differences in the clustering patterns of cell lines relative to clinical specimens. Conclusion: These findings suggest that significant differences in gene expression exist between clinical conditions and their respective cell line models and that these differences should be taken into account when extrapolating cell line results to in vivo systems.*

Cell line models are routinely studied to understand particular biological phenomena, with the expectation that discoveries made in these models will provide insight into human biology. These models are widely used to explore potential causes and treatments for human disease, where experimentation on humans would be unfeasible or unethical. Breast cancer cell lines are generated from cells isolated from breast tumour specimens and have the capability to divide indefinitely when grown *in vitro* under stringent growth conditions. This potential makes these cell lines an excellent model of study for understanding the basic biology of breast cancer. Many studies which are not possible on animal models can be relatively easily done on these cell lines.

There is, however, a great difference in the growth environment of the cancer cells *in vivo* to that of *in vitro*. Despite the relatively large number of cancer cell lines currently under study in a variety of clinical settings worldwide, so far studies aiming at investigating the similarity of cell line models to their respective clinical conditions have been very limited. A previous study found that only a small subset of primary breast cancers that display certain features of advanced tumour and poor prognosis can be cultured for a lengthy time (1). This group also reported that there was an excellent correlation among the cell lines to their clinical specimens (2), in terms of morphological features, presence of aneuploidy, immunohistochemical expression of estrogen receptor, progesterone receptors, HER2/neu, p53 proteins, allelic loss at all of the chromosomal regions analysed and *TP53* gene mutations. A more recent study, concluded that most of the currently used cell lines are derived from metastatic sites rather than primary tumour and therefore may not be representative of the diverse nature of breast cancer (3).

The advent of large-scale expression profiling experiments heralded by developments in microarray technology have facilitated a whole-genome analysis approach to this question. Large-scale expression profiling has made it possible to quantify the gene expression profiles of thousands of genes in a single experiment, thereby allowing the comparison of different samples on the basis of their full genomic expression profile, rather than on a selected number of genes. A study by Chang *et al.* (4) reviewed the role of microarrays in management and treatment of breast cancer, and observed that a combined genomic approach should be taken to understand the heterogeneity of breast cancer.

Correspondence to: Jai Prakash Mehta, National Institute for Cellular Biotechnology, Dublin City University, Glasnevin, Dublin 9, Ireland. Tel: +35317005700, Fax: +35317005484, e-mail: mehtaj2@mail.dcu.ie

Key Words: Cell line, clinical specimens, cell cycle, immune response, representative model, estrogen receptor, hierarchical clustering, principal component analysis.

Given the novelty of microarrays, the number of studies utilizing this technology to investigate similarity between the gene expression profiles of cell lines and clinical specimens are limited. It has been found that cell lines and tumour specimens have distinct gene expression patterns (5) which need to be considered for their appropriateness for each subtype of clinical condition. Another study compared gene expression profiles of early passage tumour cultures and immortal cell lines (6) and observed that epithelium cultures isolated from primary breast tumours retain the characteristics of the tumour, but these characteristics are eliminated following *in vitro* selection of the rapidly proliferating cell population. In a similar comparative study of gene expression profiles of lung cancer cell lines and their respective clinical specimens (7), it was observed that 51 of 59 cell lines represented their presumed tumours of origin.

This study aims to determine differences in gene expression profiles between human breast cancer specimens and cell lines to evaluate the clinical relevance of cell line models which will assist in our translation of results obtained from cell line studies to clinical conditions.

Materials and Methods

Gene expression profiles of 189 clinical breast specimens (GEO accession: GSE2990) (8) and 19 cell lines (GEO accession: GSE3156) (9) were obtained from Gene Expression Omnibus (10). These samples were pooled as a single experiment and normalized using the dCHIP (11) algorithm (www.dchip.org). Since the clinical specimens were analyzed using U133A and the cell lines were analyzed using U133_Plus 2.0 microarray chips, the genes not represented in U133A were removed from the cell line data, giving a total available probeset number of 22283. However, for the estrogen receptor analysis on cell lines, all the genes on the U133_Plus 2.0 chips were included (54675).

Data filtration. Gene level filtering was done to remove those genes which did not change across the experiment, prior to comparison of samples by hierarchical clustering. This was achieved by applying a standard deviation (SD) filter on individual genes across all samples to remove these non-changed genes. For hierarchical clustering and Principal Component Analysis (PCA) comparison of cell lines relative to clinical specimens, all samples were pooled and an SD of less than 0.5 across the samples was applied. For the hierarchical clustering by ER status comparison of cell lines and clinical specimens, an SD filter of less than 1.0 across each group of samples (cell lines vs. clinical specimens) was applied and the samples were clustered on each respective gene-list and coloured according to ER status.

Clustering. For hierarchical clustering, the individual genes were mean centered and divided by SD using dCHIP. The distance criteria were Euclidean distance and the type of clustering used was average linkage clustering. For PCA, the data was mean centered and scaled to average intensity. The first two components were used to plot the samples and the distribution of the samples was observed. PCA was carried out using Genespring software (http://www.silicongenetics.com.)

Significant genes. Average gene expression values were obtained for all 22283 probesets for both the cell line and clinical groups and these values were compared to identify genes which were significantly differentially-expressed (DE) between the two groups. A combination of filtration criteria was designed to identify genes which were significantly up- or down-regulated, as defined by the following criteria: $p<0.001$ (Student's *t*-test), fold change >2 and a difference of 100 units of expression on an Affymetrix scale. This was carried out using dCHIP software.

Gene ontology and pathway analysis. In order to identify ontology categories and canonical pathways affected by the changed genes, Gene Ontology analysis and Pathway analysis was performed on the DE lists using Genmapp (12). The GO categories and canonical pathways were ranked by Z-score significance. The top 10 ranked categories/pathways (where available) are listed in Tables I and II.

Estrogen receptor status. The estrogen receptor status of the cell lines was obtained from the Breast Cancer Cell Line Database (http://www.mdanderson.org) and the American Tissue Culture Collection (http://www.atcc.org), while the ER status of the clinical specimens was obtained from the GEO (http://www.ncbi.nlm.nih.gov/geo/).

Results

Data filtration. As described in Materials and Methods, two SD filters of 0.5 and 1.0 were applied to generate gene lists for hierarchical clustering. For the pooled comparison of cell lines and clinical specimens, the total number of DE genes identified following an SD filter of 0.5 was 8036. For the comparison of cell line and clinical clustering relative to ER status, the number of filtered genes following application of an SD filter of 1.0 was 7738 for the cell lines and 6643 for the clinical specimens.

Clustering. Hierarchical clustering and PCA were performed on these changed genes. Hierarchical clustering, using the filtered 8036 gene list, separated the sample set into two distinct clusters (Figure 1), one comprising the clinical specimens and the other comprising the cell line models. To examine whether the differences in hierarchical clustering between cell lines and tumour specimens were due to differences incorporated by sample processing at different sites, the clustering analysis was repeated substituting a separate 104-tumour dataset for the 189-tumour dataset detailed here. In this experiment, two separate clusters of cell lines and tumour specimens were again observed (data not shown).

PCA was also performed on the sample using the filtered 8036 gene list, which also separated the clinical specimens and cell lines into two distinct groups (Figure 2). As can be

Table I. *GO terms and pathways enrichment analysis for genes over-expressed in cell line models compared to clinical specimens. A higher Z-score represents a stronger association of that function to genes which have over-expressed in cell lines relative to clinical specimens.*

GO Name	Number of genes		Z-score
	Changed	Measured	
Mitotic cell cycle	61	281	15.236
Cell cycle	87	576	13.874
Mitosis	29	105	12.317
M phase of mitotic cell cycle	29	107	12.163
M phase	33	137	11.981
Nuclear division	31	132	11.404
Cell proliferation	94	877	10.501
DNA replication and chromosome cycle	30	154	9.791
Regulation of cell cycle	45	325	9.111
Mitotic anaphase	6	11	8.5
MAPP name			
Cell cycle (KEGG)	22	84	9.081
DNA replication reactome	11	42	6.347
G1 to S cell cycle reactome	9	65	3.311
Translation factors	7	48	3.069
Pentose phosphate pathway	2	7	2.856
mRNA processing reactome	12	115	2.739
Cholesterol biosynthesis	3	15	2.664

Table II. *GO terms and pathways enrichment analysis for genes over-expressed in clinical specimens compared to cell line models. A higher Z-score represents a stronger association of that function to genes which have over-expressed in clinical specimens compared to cell line models.*

GO Name	Number of genes		Z-score
	Changed	Measured	
Immune response	107	595	12.412
Defense response	110	650	11.847
Response to biotic stimulus	115	710	11.586
MHC class II receptor activity	9	11	10.458
Extracellular matrix	48	215	9.992
Antigen processing, exogenous antigen via MHC class II	8	10	9.731
Antigen presentation, exogenous antigen	8	10	9.731
Extracellular	105	742	9.451
Antigen presentation	12	23	9.204
Antigen processing	12	23	9.204
MAPP name			
Complement activation classical	7	16	5.405
Complement and coagulation cascades (KEGG)	11	49	3.897
Matrix metalloproteinases	7	30	3.216
Smooth muscle contraction	19	143	2.574
Inflammatory response pathway	6	31	2.435

seen on the axes, the total variance accounted for in the sample set was 27.95%. The clinical specimens also segregated into two further sub-groups, although not as distinct as those separating the cell lines from the clinical specimens.

Significant genes. The clinical specimens and the breast cell lines were compared for transcripts which were significantly up- or down-regulated in the two groups ($p<0.001$, fold change >2 and difference of 100 Affymetrix units). Of 2615 genes which passed the above filtration criteria, 1086 were up-regulated in cell lines relative to clinical specimens and 1529 genes were down-regulated in cell lines compared to clinical specimens.

Gene ontology and pathway analysis. Genmapp Gene Ontology and Pathway Analysis were performed on the up- and down-regulated gene lists and the over-represented GO categories/canonical pathways are outlined in Tables I and II. In cell lines relative to clinical specimens, many of the functions which were over-represented were related to cell cycle functions and nucleic acid processing (Table I). Where clinical specimens were compared to cell lines, the majority of categories and pathways affected were related to the immune response and related functions (Table II).

Estrogen receptor analysis. Hierarchical clustering was also performed separately on the two groups (*i.e.* cell lines and clinical specimens), to determine if either group clustered similarly when compared for ER status. This analysis segregated the cell lines into two distinct groups, which clustered largely according to their ER status (Figure 3). Exceptions to this rule included the ER-negative SK-BR-3 and MDA-MB-453 cell lines and the ER-positive HCC1428, which clustered with the opposite group. Hierarchical clustering performed on the 189 clinical sample dataset did not demonstrate any appreciable clustering according to ER status (data not shown), although there was a tendency for clinical specimens to cluster based on their grade.

Discussion

Cell lines are widely used as models of *in vivo* systems. However, limited studies have been carried out to establish if these models accurately reflect *in vivo* scenarios. Our study examined gene expression differences and similarities in a representative group of breast cancer cell lines and clinical specimens to estimate their approximate level of similarity.

Cell lines grow under very tight and well-optimized conditions, with enough space to grow and divide. In

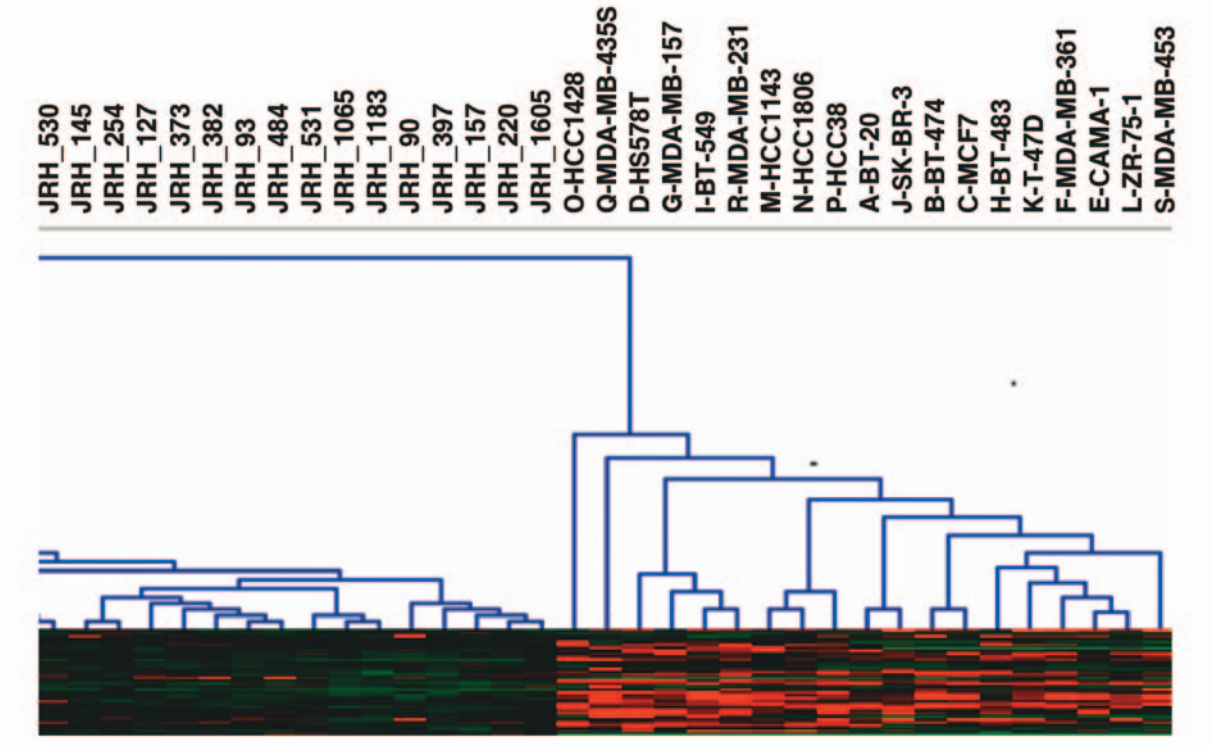

Figure 1. *Hierarchical clustering demonstrating that cell lines and clinical specimens form two discrete groups. The right cluster is of 19 cell lines included in the study. The left cluster (incompletely shown because of large number of tumour specimens) represents 189 breast tumours.*

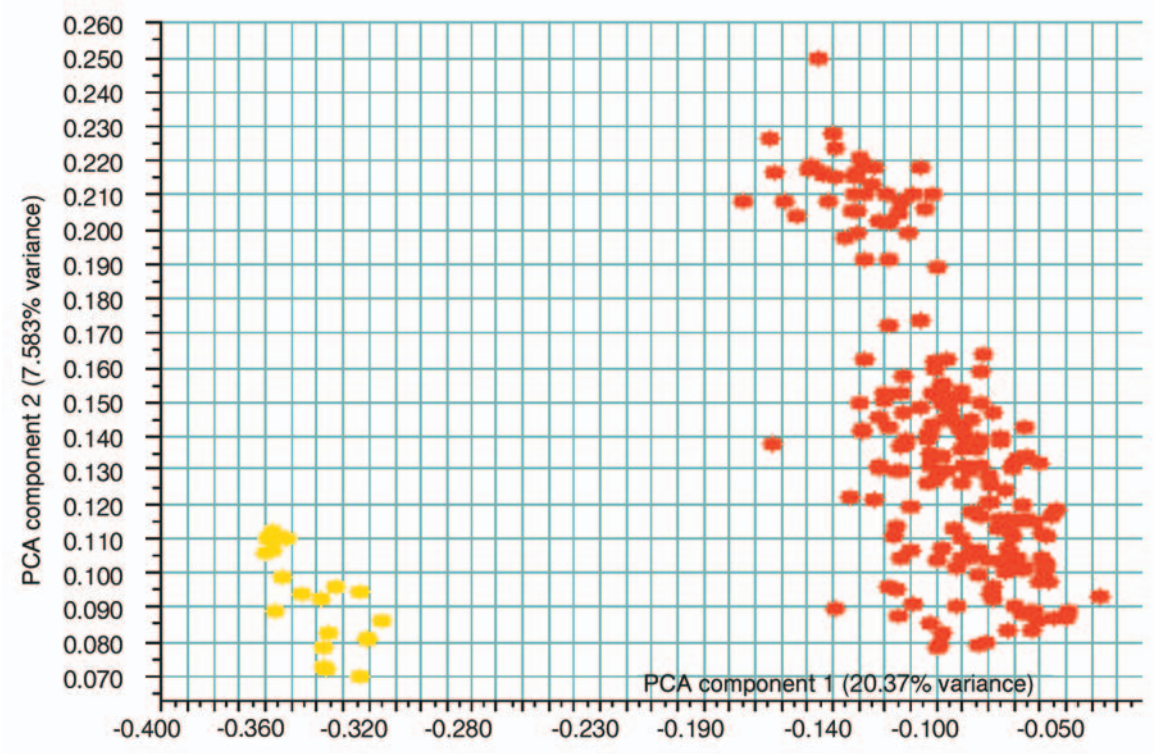

Figure 2. *Principal component analysis was performed on the samples and the two components were plotted. Clinical specimens are highlighted in red, cell line samples are highlighted in yellow.*

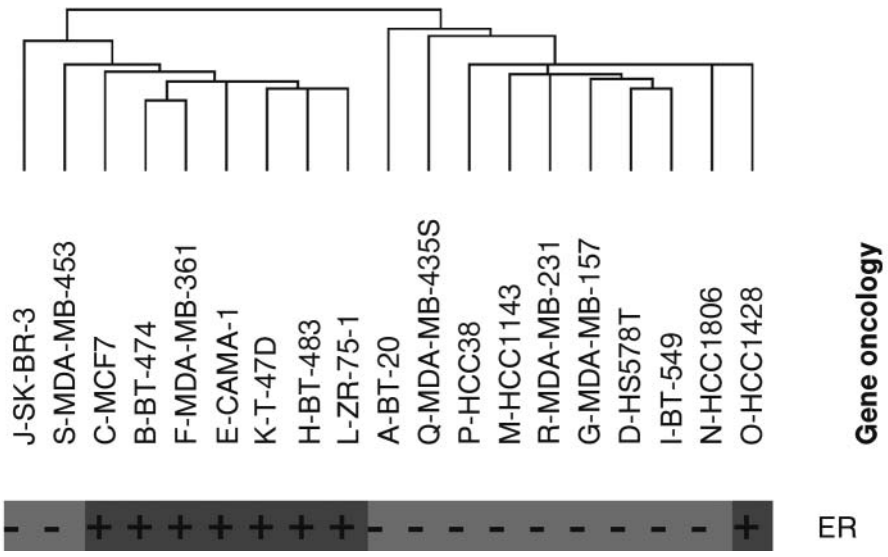

Figure 3. *Hierarchical clustering of cell lines. The + indicates ER-positive cell lines and the – represents ER-negative cell lines. The left cluster in enriched with ER-positive cell lines and the right cluster is enriched with ER-negative cell lines.*

comparison, tumours grow in a completely different environment and are influenced by a varied range of conditions. In this study, we show a clear segregation of the cell lines and clinical specimens by hierarchical clustering. This is in agreement with other similar studies where cell lines and clinical specimens tend to cluster separately from each other (5, 6). PCA also demonstrated a clear separation of the two groups, *i.e.* the cell lines from the clinical specimens. A segregation of the clinical specimens into two smaller sub-groups was also observed, although the clinical/biological basis for this has not been determined here. An earlier experiment also reported considerable data scatter among primary tumour cultures and cell lines compared to normal breast specimens using PCA as a comparison tool (6).

From the Genmapp analysis, cell cycle, mitosis, nuclear division, cell proliferation and other related functions are over-represented in cell line models in comparison to the clinical specimens, while functions related to immune response and defense response are over-represented in clinical specimens relative to cell lines. A recent study (13), also reported that genes related to proliferation and the cell cycle are over-represented in cell lines relative to clinical specimens, while cell communication, cell adhesion molecules and ECM-receptor interaction are down-regulated in cell lines compared to clinical specimens. Our study also indicated a decrease in expression of genes involved in cell adhesion in the cell lines compared to clinical specimens, although this data (not shown) did not make it into the top ten ontologies outlined in Table II.

While the analysis outlined above identified the macroscopic broad-based differences between breast cancer cell lines and clinical specimens, it was considered useful to assess the similarity relationships of the cell lines and clinical specimens with regard to their ER status. It was hoped that while differences had been observed when comparing cell lines and clinical specimens directly, both cell lines and clinical specimens would cluster similarly when ER status was used as the criteria. Previous studies had demonstrated that both cell lines (14) and clinical specimens (15) cluster largely on their ER status. To this end, unsupervised clustering of the cell lines and clinical specimens separately was carried out to determine if either group clustered according to ER status. However, while the cell lines largely clustered according to ER status, the clinical samples did not. This result indicated that, even on a single parameter basis, the differences between clinical specimens and their respective cell line models may remain considerable.

Conclusion

The findings reported here indicate that significant differences in gene expression between clinical specimens and their respective cell line models exist at both the large- and small-scale levels. The study of Dairkee *et al.* concluded that the results obtained from cell lines may act as good models for high-grade cancer, but may fail as useful models for most of the low- and medium-grade breast cancers (6). While our study does not indicate a specific clinical classification for which such cell line data may prove relevant, the data presented here demonstrate that these differences should be taken into account when extrapolating *in vitro* cell line results to clinically-relevant *in vivo* systems.

Acknowledgements

This work was supported by funding from Ireland's Higher Educational Authority Program for Research in Third Level Institutes (PRTLI) Cycle 3.

References

1 Gazdar AF, Kurvari V, Virmani A, Gollahon L, Sakaguchi M, Westerfield M, Kodagoda D, Stasny V, Cunningham HT, Wistuba II, Tomlinson G, Tonk V, Ashfaq R, Leitch AM, Minna JD and Shay JW: Characterization of paired tumour and non-tumour cell lines established from patients with breast cancer. Int J Cancer *78(6)*: 766-774, 1998.

2 Wistuba II, Behrens C, Milchgrub S, Syed S, Ahmadian M, Virmani AK, Kurvari V, Cunningham TH, Ashfaq R, Minna JD and Gazdar AF: Comparison of features of human breast cancer cell lines and their corresponding tumours. Clin Cancer Res *4(12)*: 2931-2938, 1998.

3 Burdall SE, Hanby AM, Lansdown MRJ and Speirs V: Breast cancer cell lines: friend or foe? Breast Cancer Res *5*: 89-95, 2003.

4 Chang JC, Hilsenbeck SG and Fuqua SA: The promise of microarrays in the management and treatment of breast cancer. Breast Cancer Res *7(3)*: 100-104, 2005.

5 Ross DT and Perou CM: A comparison of gene expression signatures from breast tumours and breast tissue derived cell lines. Dis Markers *17(2)*: 99-109, 2001.

6 Dairkee SH, Ji Y, Ben Y, Moore DH, Meng Z and Jeffrey SS: A molecular signature of primary breast cancer cultures; patterns resembling tumour tissue. BMC Genomics doi:10.1186/1471-2164-5-47, 2004.

7 Wang H, Huang S, Shou J, Su EW, Onyia JE, Liao B and Li S: Comparative analysis and integrative classification of NCI60 cell lines and primary tumours using gene expression profiling data. BMC Genomics doi: 10.1186/1471-2164-7-166, 2006.

8 Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, Smeds J, Nordgren H, Farmer P, Praz V, Haibe-Kains B, Desmedt C, Larsimont D, Cardoso F, Peterse H, Nuyten D, Buyse M, Vijver MJV, Bergh J, Piccart M and Delorenzi M: Gene expression profiling in breast cancer: Understanding the molecular basis of histological grade to improve prognosis. Nat Can Inst *98(4)*: 262-272, 2006.

9 Bild AH, Yao G, Chang JT, Wang Q, Potti A, Chasse D, Joshi MB, Harpole D, Lancaster JM, Berchuck A, Jr JAO, Marks JR, Dressman HK, West M and Nevins JR: Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature *439*: 353-357, 2005.

10 Edgar R, Domrachev M and Lash AE: Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nuc Acids Res *30(1)*: 207-210, 2002.

11 Li C and Wong WH: Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. Prot Natl Acad Sci USA *98*: 31-36, 2001.

12 Dahlquist KD, Salomonis N, Vranizan K, Lawlor SC and Conklin BR: GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. Nat Gen *31(1)*: 19-20, 2002.

13 Ertel A, Verghese A, Byers SW, Ochs M and Tozeren A: Pathways-specific differences between tumour cell lines and normal and tumour tissue cells. Mol Cancer doi:10.1186/1476-4598-5-55, 2006.

14 Charafe-Jauffret E, Ginestier C, Monville F, Finetti P, Adelaide J, Cervera N, Fekairi S, Xerri L, Jacquemier J, Birnbaum D and Bertucci F: Gene expression profiling of breast cell lines identifies potential new basal markers. Oncogene *25(15)*: 2273-2284, 2006.

15 Sotiriou C, Neo SY, McShane LM, Korn EL, Long PM, Jazaeri A, Martiat P, Fox SB, Harris AL and Liu ET: Breast cancer classification and prognosis based on gene expression profiles from a population-based study. PNAS *100(18)*: 10393-10398, 2003.

Received December 19, 2006
Revised February 15, 2007
Accepted February 16, 2007