

Distinguishing key biological pathways between primary breast cancers and their lymph node metastases by gene function-based clustering analysis

HARRI LÄHDESMÄKI^{1,2*}, XISHAN HAO^{3*}, BAOCUN SUN³, LIMEI HU¹,
OLLI YLI-HARJA², ILYA SHMULEVICH¹ and WEI ZHANG¹

¹Cancer Genomics Laboratory, Department of Pathology, The University of Texas M.D. Anderson Cancer Center, 1515 Holcombe Blvd., Unit 85, Houston, TX 77030, USA; ²Institute of Signal Processing, Tampere University of Technology, P.O. Box 553, 33101 Tampere, Finland; ³Tianjin Medical University Cancer Hospital, Tianjin 300060, P.R. China

Received September 1, 2003; Accepted October 1, 2003

Abstract. In order to identify key biological pathways that can distinguish between primary breast cancers and their lymph node metastases, we employed gene expression profiling together with gene function-based clustering analysis. We first acquired gene expression profiles of 9 matched primary tumors and the corresponding metastases that contained at least 75% of tumor cells. Then, we applied a clustering algorithm to the preprocessed data. In order to focus on the most informative genes, we ranked all the genes individually based on their abilities to separate the primary breast tumor and metastases samples. Further, we separated these genes into six functional groups according to the Stanford SOURCE database: 'cell cycle,' 'apoptosis,' 'metabolism,' 'cell adhesion and migration,' 'signal transduction,' and 'transcriptional factor and DNA binding molecules.' Unsupervised clustering analysis using all of the 2,303 genes on the microarrays was not able to separate the primary and metastases samples. Clustering analysis using the most informative genes revealed that primary tumors were more tightly clustered, whereas the metastases samples were relatively heterogeneous. The clustering analysis with the genes belonging to different functional groups showed that different functional gene sets varied in their abilities to separate

primary tumors and their metastases. Marked separations were found with genes involved in metabolism, signal transduction, cell cycle, and transcriptional factor and DNA binding molecules. In contrast, apoptosis and cell adhesion and migration genes did not provide a clear separation of the two groups of samples. These results suggest that metastatic cells have different metabolism and signal transduction activities, regulated by transcriptional events, from the primary tumor cells. The results also suggest that the altered cell adhesion and migration potentials that are required for tumors to metastasize already exist in the primary tumors as a whole.

Introduction

Breast cancer is a major health problem for women worldwide. Early detection of a local tumor before metastasis is a key to effective response to therapy by surgery and adjuvant chemotherapy and hormone treatment (1). The presence of lymph node metastases is often the first step of the disease progression from a more local and contained stage to a more aggressive stage, eventually leading to distant metastasis. Breast cancers with distant metastasis are poorly responsive to any therapy currently in use (2). Therefore, detection of sentinel lymph node metastases represents a major effort in breast cancer prognosis and disease management. Comparative molecular and genomic studies of primary breast cancers and their lymph node metastases, thus, may offer insight into the key events that underlie this important transition of the disease.

The development of genomic biology together with informatics has provided powerful tools for investigating the changes at gene levels in the disease progression. Because the disease is highly heterogeneous and multifaceted, involving many different cellular activities and genes, high-throughput gene expression profiling approaches have been used in recent years to identify the sets of genes that have diagnostic and/or prognostic value in breast cancers (2-4). Different computational methods, such as multidimensional scaling (MDS) (5), have been frequently used to illustrate the separation of different groups of disease based on gene

Correspondence to: Dr Wei Zhang, Cancer Genomics Core Laboratory, Department of Pathology, The University of Texas M.D. Anderson Cancer Center, 1515 Holcombe Blvd., Houston, TX 77030, USA

E-mail: wzhang@mdanderson.org

*Contributed equally

Key words: breast cancer, lymph node metastases, biological pathways, gene function-based clustering analysis

expression patterns. Using these techniques, normal breast epithelium and cancer or breast cancers that have or do not have mutations in BRCA gene have been successfully separated (6,7). Breast cancers with different subtypes such as luminal or basal types are also proposed to have different 'molecular portraits' (8). Further, using supervised statistical methods, 70 genes have recently been found to be informative for predicting breast cancer prognosis (2,3). In a similar fashion, Ramaswamy *et al* compared gene expression profiles of unmatched primary tumors and metastases from several cancer types and identified 17 signature genes for metastases containing eight upregulated and nine downregulated genes (4).

There are several motivations for comparing the gene expression profiles between paired primary breast cancer cells and their metastases. First, the results may reveal the key gene expression events that transform the primary local tumor into a systemic metastasis. Second, scrutiny of the differentially expressed genes may provide insight into the different cellular states between primary tumors and metastases. This is clinically pertinent because cellular status such as cell cycle and metabolism may affect how cells respond to certain chemotherapy drugs. Many drugs such as 5-fluorouracil and cisplatin preferentially kill proliferating cells (9). Finally, the use of matched primary and metastases samples from the same patients reduces the variability introduced by inter-patient heterogeneity and thus, improves the specificity and sensitivity of the analysis.

Several early studies have used all the genes included in the microarray for clustering analysis. Although this method can be successful for separating very distinct groups, such as normal and tumor, high-grade and low-grade tumors, or different sub-types of tumors, clustering analysis based on all genes may fail to separate groups with more subtle differences. An alternative approach is to first select informative genes (or the intrinsic gene set) based on differences in gene expression and then use this set of genes for clustering analysis (10). In this study, we moved one step further. We first used all genes followed by the informative genes for clustering. Then, we classified the informative genes into six functional groups based on known gene ontology information and used the functionally informative genes in the clustering analysis. Discrimination approaches utilizing similar knowledge-based methods have also been used by others (11). Using this knowledge-based approach, we identified the functional groups of genes that best separate primary breast cancers and their lymph node metastases.

Materials and methods

Breast cancer tissues. The primary and lymph node metastases tissues were surgically removed from the patients as part of the treatment. A fraction of the tissues were snap-frozen in liquid nitrogen immediately after surgical resection. The procedure of tissue collection and use of the material for research were approved by institutional reviewing committee of Tianjin Cancer Hospital. The tissues were evaluated with H&E staining by a pathologist (B. Sun). Only tissues with >75% tumor cells by pathological evaluation were used for microarray studies. Because of this stringent criterion, 9

paired samples obtained initially for this study were excluded for this analysis and 9 paired primary breast cancer (P) and metastases (M) were used in this study.

Microarray assay. RNA isolation, microarray production, hybridization, and image analysis were carried out as previously described (12-15). Each matched primary breast tumor and the corresponding metastases sample were hybridized on the same array, the samples being labeled with Cy3 and Cy5, respectively. A cDNA microarray generated by the Cancer Genomics Core Laboratory, M.D. Anderson Cancer Center, was used in this study. The array contains 2,303 known genes and ESTs printed in duplicate. The data set consisted of 9 matched primary breast tumor (P) and metastases (M) samples, resulting in 18 gene expression profiles in total. Each gene activity profile, a measurement from either P or M, contains the gene expression value for 2303 genes. All the measurements are replicated twice.

Preprocessing. The data set was preprocessed as follows: First, the replicated background-subtracted signal intensities z_{ij}^1 and z_{ij}^2 , for each sample $1 \leq j \leq 18$ and each gene $1 \leq i \leq 2303$, were combined by averaging. The averaged expression values z_{ij} were log-transformed (base 2): $y_{ij} = \log_2 z_{ij}$, and the log-arithmic-domain values were further processed. Since the data is from a two-color microarray system, the dye bias effect was corrected using the standard 'lowess' smoothing-based normalization with smoothing parameter $f=20\%$ (16). All genes were used in the local robust fits. Shift and scale normalization between different arrays was done for each array separately by subtracting the median and dividing by the mad (median absolute deviation from the median), i.e., $x_{ij} = (y_{ij} - \bar{y}_j) / \sigma_j$, where $\bar{y}_j = \text{median}(y_{1j}, \dots, y_{2303j})$ and $\sigma_j = \text{median}(|y_{1j} - \bar{y}_j|, \dots, |y_{2303j} - \bar{y}_j|)$.

Informative gene selection. Since we aim to distinguish between primary breast cancers and their corresponding metastases, based on the gene expression profiles only, we focused on the most informative genes for class prediction. For that purpose, we ranked the genes based on the standard paired two-sample t-test (or t-statistic). (Note that the standard paired two-sample t-test reduces to the one sample t-test of paired differences, with the null hypothesis being that the mean of the differences is zero.) The 25 most informative genes are shown in Fig. 1.

Before performing the knowledge-based clustering analysis, we should decide which genes are selected to be informative. In order to have a large enough number of genes in further clustering analysis, which would provide us the general picture of the selected biological processes, we define the informative genes to be the ones whose significance value does not exceed 0.1. We found in total 446 genes in our data set satisfying the above criterion.

Gene function-based analysis. Our exploratory data analysis showed that the expression values of the 5th sample, both from P and M, were significantly different from the expression values of the other samples. Since results of clustering algorithms may be significantly distorted by outliers, we excluded the 5th sample from further analysis.

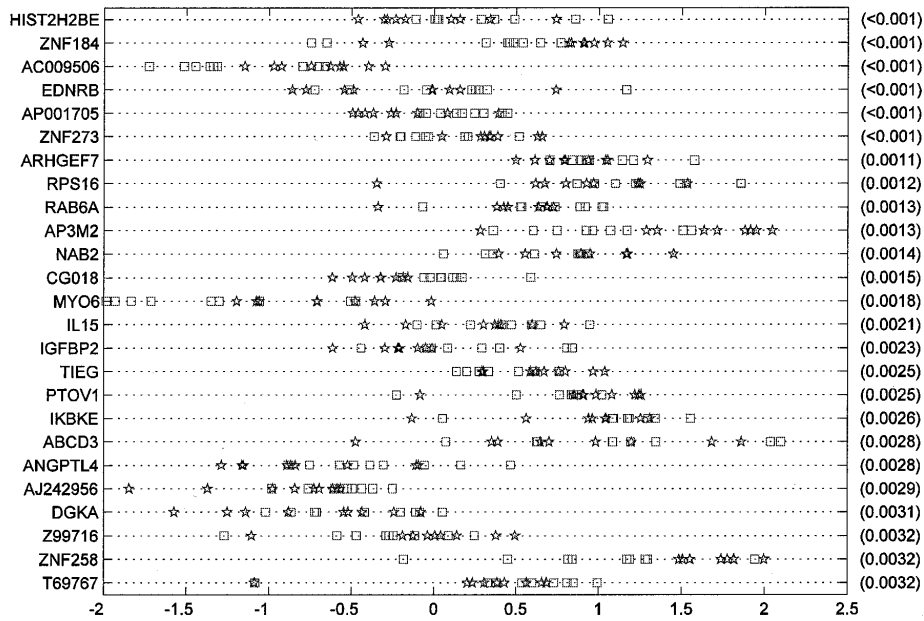


Figure 1. The 25 most informative genes as obtained by the paired t-statistic-based gene ranking. The normalized expression values of each candidate gene are plotted on the horizontal axis. Symbols: stars, primary breast tumor; and squares, corresponding metastases. Gene names (or accession no. if no standard gene name was found) are shown on the left. The corresponding t-test-based significance values are shown on the right.

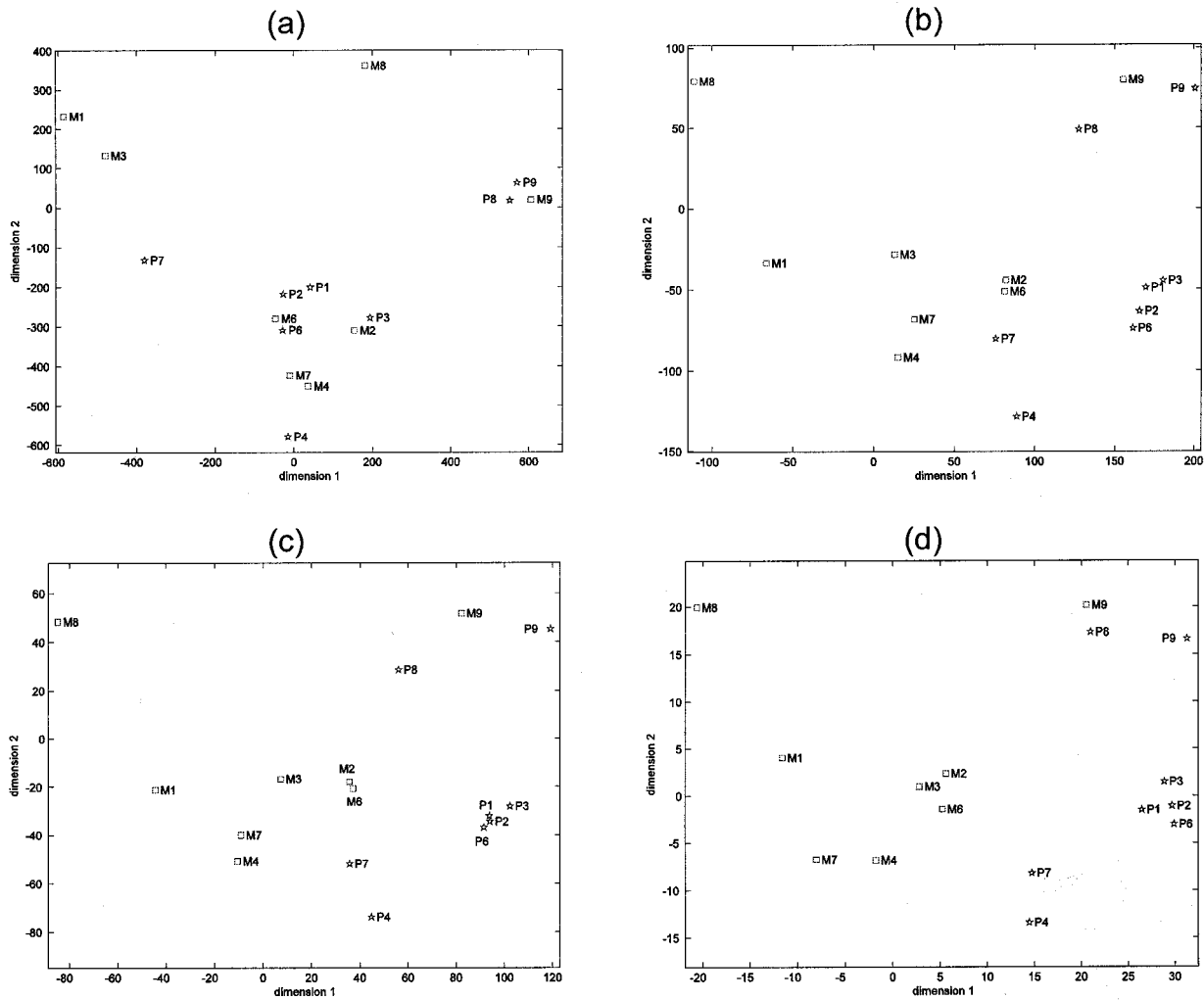


Figure 2. Clustering results for the primary and metastasis samples using principal component analysis (PCA). The number of genes used: (a), all the 2,303 genes, (b), the 446 informative genes (see text for details), (c), only the 255 most informative genes, and (d), the 72 most significant genes based on the paired two-sample t-statistic. Symbols: stars, primary breast tumor; and squares, corresponding metastases.

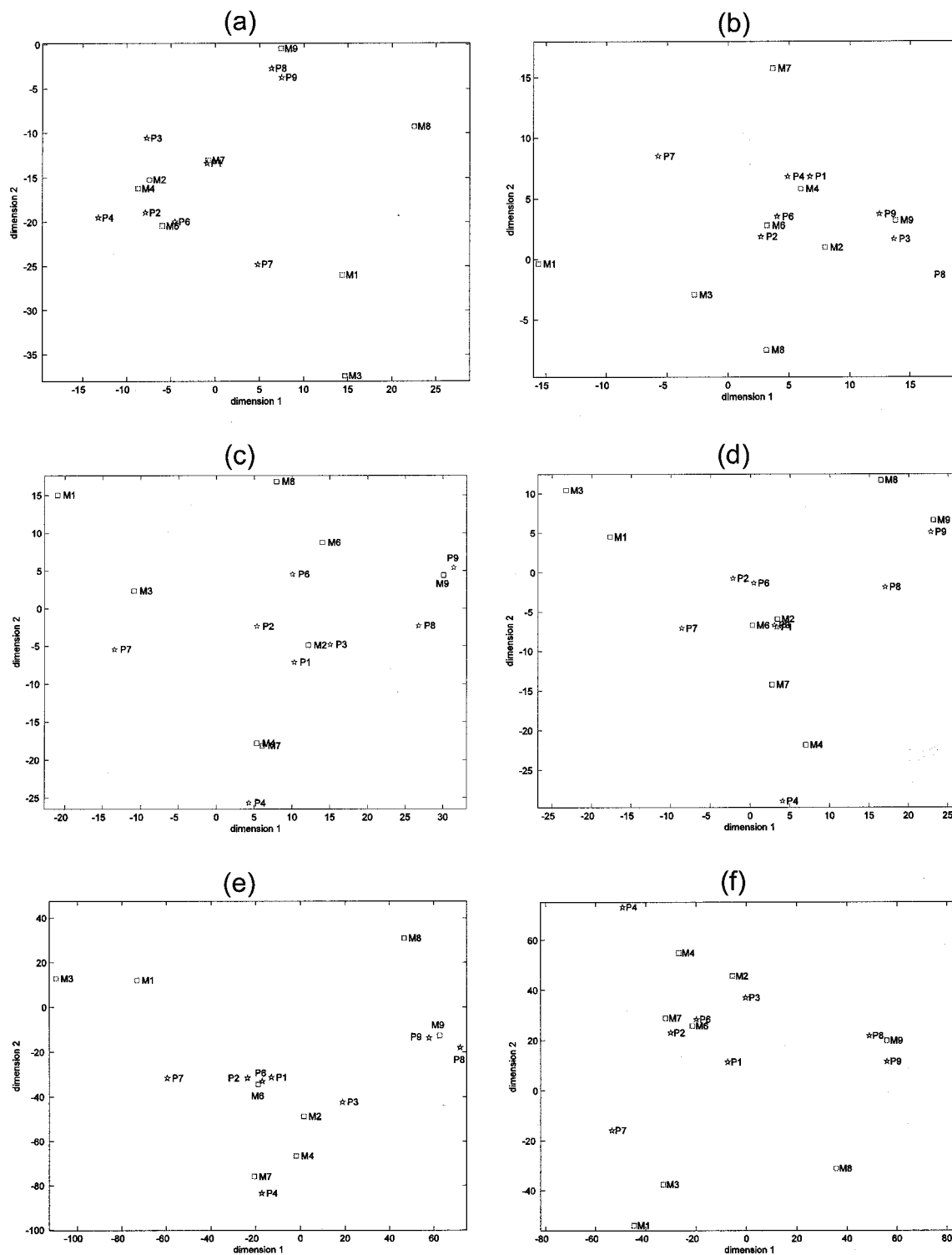


Figure 3. Knowledge-based clustering results for the primary and metastasis samples using all genes from one of the six functional categories at a time. (a), ‘cell cycle’ (65); (b), ‘apoptosis’ (44); (c), ‘metabolism’ (106); (d), ‘cell adhesion and migration’ (86); (e), ‘signal transduction’ (289); and (f), ‘transcriptional factor and DNA binding molecules’ (253). The number of genes in each category is shown in parentheses. Symbols: stars, primary breast tumor; and squares, corresponding metastases.

First, the primary and metastases samples were subjected to principal component analysis (PCA) (17). PCA-based clustering results using all the genes and using the set of informative genes for which the significance value is ≤ 0.1 , ≤ 0.05 , and ≤ 0.01 (see above) are shown in Fig. 3a, b, c, and d, respectively. Since the number of measurements is far smaller than the number of genes, we use a standard approach

when solving the eigenvalue-eigenvector problem for the sample covariance matrix of the measurements $x_i, i=1, \dots, N$. Let $\mathbf{A}^T = [\mathbf{x}_1 - \bar{\mathbf{x}}, \dots, \mathbf{x}_N - \bar{\mathbf{x}}]$ and $\bar{\mathbf{x}} = 1/N \sum_{i=1}^N \mathbf{x}_i$. Instead of finding the eigenvalues of the original sample covariance matrix $\Sigma = \mathbf{A}^T \mathbf{A}$, we compute them for the matrix $\hat{\Sigma} = \mathbf{A} \mathbf{A}^T$. The eigenvalues of Σ and $\hat{\Sigma}$ are the same and the eigenvectors of Σ can be obtained from the eigenvectors of $\hat{\Sigma}$ by multiplying them by \mathbf{A}^T .

Table I. The number of genes in each functional group.^a

Functional category	No. of genes among all the 2,303 genes	No. of genes among all the 'informative' (446) genes
'Cell cycle'	65	11
'Apoptosis'	44	7
'Metabolism'	106	19
'Cell adhesion and migration'	86	12
'Signal transduction'	289	52
'Transcriptional factor and DNA binding molecules'	253	51

^aThe first column shows the number of genes in each functional group when all the 2,303 genes are considered. The second column shows the number of genes in each functional group when only the informative 446 genes are counted.

The clustering significance values were obtained by a simple permutation test (18). For the test statistic, we used the trace of the within-cluster scatter matrix. Let $c_1, \dots, c_{N/2}$ and $c_{N/2+1}, \dots, c_N$ be the configurations of the primary breast cancer and the corresponding metastases samples after the clustering (i.e. in the eigenvector space), and $\mathbf{C}_P^T = [c_1 - \bar{c}_P, \dots, c_{N/2} - \bar{c}_P]$, where $\bar{c}_P = 2 / N \sum_{i=1}^{N/2} c_i$ (similarly for the metastases samples). The scatter matrices for both classes are $\mathbf{S}_P = \mathbf{C}_P^T \mathbf{C}_P$ and $\mathbf{S}_M = \mathbf{C}_M^T \mathbf{C}_M$, the within-cluster scatter matrix is $\mathbf{S} = \mathbf{S}_P + \mathbf{S}_M$, and the actual scatter criterion itself is trace (S) (see ref. 17 for further details). Permutation distribution of the test statistic was computed based on 10,000 times repeated random permutation of the class labels and the significance value was taken to be the fraction of samples not exceeding the test statistic for the original sample.

The same discrimination as described above was then repeated, but now using only genes from a certain functional group at a time. Genes were categorized into functional groups using the SOURCE database (online-service of the Genetics Department in Stanford University) (19). Table I shows the number of genes in each functional category for all the 2,303 genes (the first column) and for the informative genes (the second column). Note that some genes may belong to more than one functional category. In further analysis, we allow genes to belong to several categories, partly because the decision of the main biological process is ambiguous for some genes. This results in a type of fuzzy memberships. The final knowledge-based discrimination results are shown in Figs. 3 and 4. In Fig. 3, all the 2,303 genes are categorized into functional groups and used in the analysis, whereas in Fig. 4 only the informative genes are categorized into functional groups and used in the PCA analysis.

Results

The informative genes were identified using the standard paired t-statistic. The results are shown in Fig. 1. The

normalized expression values of each candidate gene are plotted on the horizontal axis. Expression values of the primary breast tumor and the corresponding metastases are shown as stars and squares, respectively. The most promising genes for class separation are on the top of the graph.

In order to decide which genes to use in further analyses, we looked at the t-test-based significance values of the genes. One would like the number of identified 'informative' genes to be high enough to see the general picture of the differences in the expression profiles between the two classes (P and M), and especially to identify the biological processes that exhibit major divergence at the transcriptome level. The subsequent clustering methods based on only a few genes would not provide the general picture of the biological processes, but merely reflect the behavior of only a few differentially expressed genes. Thus, we define the 'informative' genes to be the ones whose significance values do not exceed 0.1. We found 446 genes in our data set satisfying this criterion.

Using all the genes, principal component analysis (PCA)-based clustering did not reveal a clear separation between the primary breast cancer and the corresponding metastases samples (Fig. 2a). However, when the same method was applied to the set of 446 informative genes selected as described above, a better separation between the two groups of samples was observed, as expected, due to the supervised nature of the gene selection. More distinct separation was achieved when using only the genes for which the significance value is ≤ 0.05 (Fig. 2c) and ≤ 0.01 (Fig. 2d). In the latter cases, there are 255 and 72 genes, respectively. Moreover, the primary samples form a fairly tight cluster whereas the metastases samples are relatively more spread out. In order to get a more quantitative picture, a significance value of the clustering result was computed by randomly permuting the class labels (primary breast tumor and the corresponding lymph node metastases) for each gene set (for details, see Materials and methods). The significance values for the gene sets are 0.5547, 0.0039, 0.0017 and 0.0004, respectively. Because the set of informative genes separate P and M better than the set including all genes, we expect that the same holds when the gene function-based clustering is applied.

The same clustering analysis was carried out incorporating prior biological information. That is, the genes were first categorized into six functional groups, 'cell cycle,' 'apoptosis,' 'metabolism,' 'cell adhesion and migration,' 'signal transduction,' and 'transcriptional factor and DNA binding molecules,' and each of the six functional gene sets was then used to perform knowledge-based clustering analysis. The numbers of genes in each functional category are shown in Table I. Note that some genes may belong to more than one functional category. In Fig. 3, all of the 2,303 genes are categorized into functional groups and used for discrimination, whereas in Fig. 4, only the 446 informative genes categorized into functional groups are used for clustering. As expected, the gene function-based clustering results using only the informative genes reveal clearer separation between P and M than that obtained when all of the 2,303 genes were used. Also, it is worth noting that none of the functional categories provide a 'perfect' separation between the two classes; the results in Fig. 4 only give a general picture of the

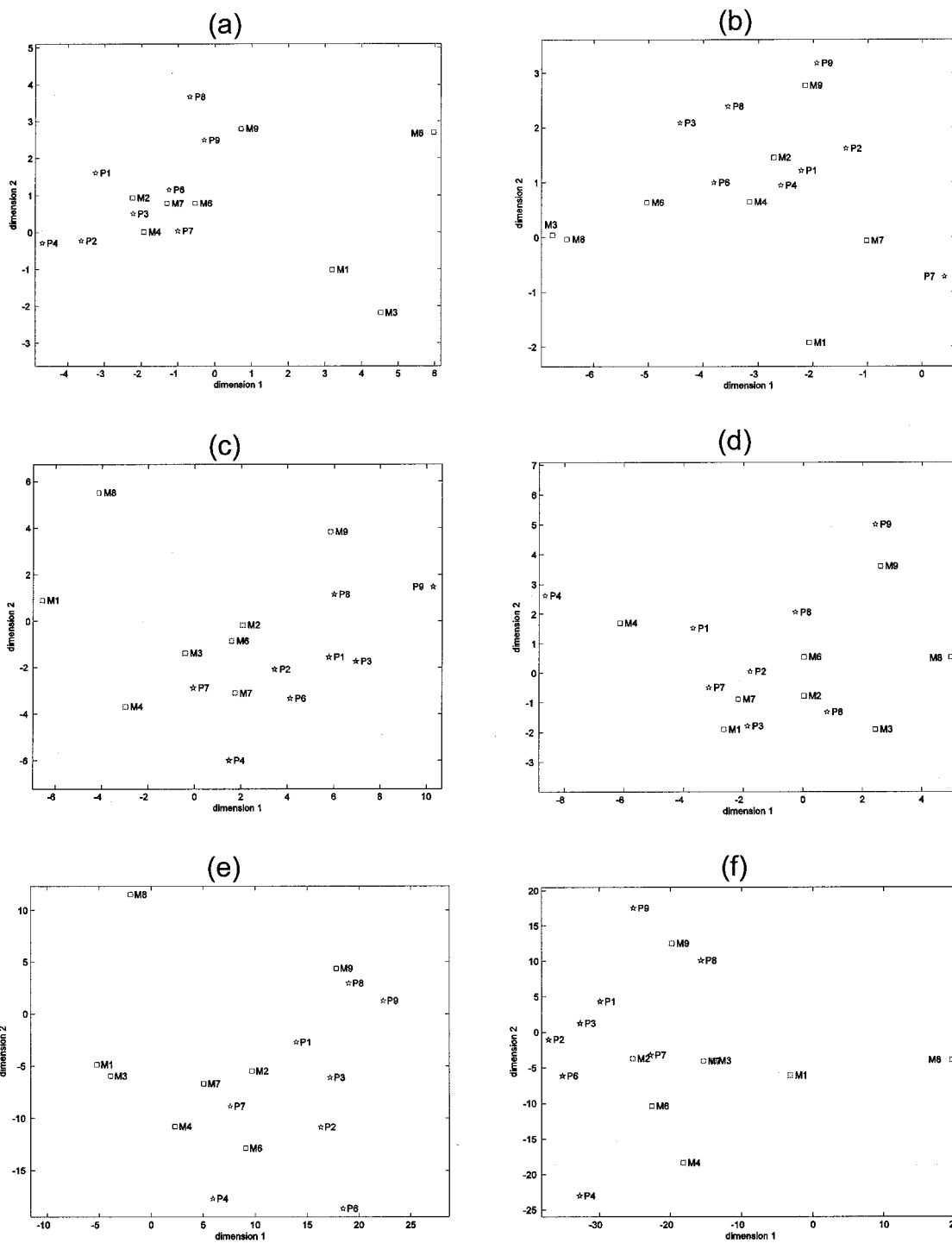


Figure 4. Knowledge-based clustering results for the primary and metastasis samples using the informative genes from one of the six functional categories at a time. (a), ‘cell cycle’ (11); (b), ‘apoptosis’ (7); (c), ‘metabolism’ (19); (d), ‘cell adhesion and migration’ (12); (e), ‘signal transduction’ (52); and (f), ‘transcriptional factor and DNA binding molecules’ (51). The number of genes in each category is shown in parentheses. Symbols: stars, primary breast tumor; and squares, corresponding metastases.

underlying differences between primary breast tumors and the corresponding lymph node metastases for each functional group of genes. The significance values for the knowledge-based discrimination results in Fig. 3 are: ‘cell cycle’ (0.3706), ‘apoptosis’ (0.3529), ‘metabolism’ (0.4991), ‘cell adhesion and migration’ (0.6981), ‘signal transduction’ (0.6208), and ‘transcriptional factor and DNA binding molecules’ (0.6450). The same significance values for the corresponding knowledge-based discrimination, when incorporating only

the informative genes, are (Fig. 4): ‘cell cycle’ (0.0227), ‘apoptosis’ (0.1571), ‘metabolism’ (0.0107), ‘cell adhesion and migration’ (0.3465), ‘signal transduction’ (0.0118), and ‘transcriptional factor and DNA binding molecules’ (0.0049).

In order to test the discrimination between the two classes (P and M) at a finer level of abstraction, we also applied the same methods to the subcategories of the ‘cell cycle’ genes, namely ‘cell cycle regulation,’ ‘cell cycle arrest’ and ‘mitotic cell cycle.’ The obtained significance values for the

separateness were 0.0504, 0.0154 and 0.0165, respectively, indicating that all the subcategories could play an important role. (The significance value for the group 'mitotic cell cycle' is based on the paired t-test because there is only one gene in that subcategory). An issue to be taken into account is, however, that the number of genes in each subcategory gets fairly small (8, 3, 1) which may no longer provide a general picture of those finer functional groups.

Finally, genes for functional groups 'transcriptional factor and DNA binding molecules,' 'metabolism,' 'signal transduction' and 'cell cycle,' exhibit a perspicuous separation (Fig. 4a, c, e and f). On the other hand, 'apoptosis' and 'cell adhesion and migration' provides practically no separation between P and M (Fig. 4b and d).

Discussion

Understanding of metastasis is very important in cancer biology because metastasis represents a turning point in cancer progression. With metastasis, a local disease becomes a systemic disease and a relatively stationary disease becomes a moving target in terms of treatment (1). Thus, metastatic tumors invariably represent more aggressive tumors with much less response to treatment. A fundamental question is: what occurs at the molecular level, enabling cancer cells to leave their original environment and move to other sites in the body? Using transcriptome profiling technologies, we are beginning to acquire some valuable information. Ramaswamy *et al*, found 17 signature genes for the metastases in their recent study (4). For comparison purposes, we looked at those genes in our data set. Only four of those 17 genes were printed on our arrays, but none of them were found to be significantly differentially expressed. Van't Veer *et al* compared gene expression profiles of non-metastatic breast cancers and metastatic breast cancers and identified a set of 70 genes that show differential expression between the two groups (2). Only 6 of those 70 identified genes are found among our set of genes, and 5 of them are differentially expressed, although the 5 genes are not ranked at the top. Among the 70 genes is MMP9, which is known to be involved in metastasis. Although such an approach can identify some specific target genes, we are lacking methods for understanding this process through a more global and functional perspective. Thus, an important question is: which cellular pathways are most affected when cancer cells become metastatic?

In the present study, we attempted to address this issue. Our results shed some light on the cellular processes that may be crucial for breast cancer metastasis. There are several key features in our experimental design. First, because of the tremendous heterogeneity involved in different cancers of different patients, we believe that paired tissue samples of primary breast cancer and lymph node metastasis from the same patients are likely to provide more informative results. We also recognize the potential problem of different cell populations in different samples because surgically removed cancer tissues often contain normal surrounding tissues and some tissues are filled with infiltrating lymphocytes. All those factors will increase the 'noise' and potentially make the data less interpretable. Thus, among the initial 18 paired samples we accumulated, a pathologist first evaluated

the tissues and removed the tissues that have less than 75% of tumor cells. Although we only have 9 paired samples subsequent to this analysis, we believe the results that we obtained with this sample set are more reliable than if we were to use all 18 paired samples.

Second, it was not our intention to find the best classifier genes that separate the primary breast cancers and their metastases, although this is a valuable aim addressed in other studies. We focused on a systems perspective and explored the most affected molecular pathways before and after the cancer cells moved to the lymph node. To do so, we first identified a set of informative genes that exhibited a statistical difference between the two groups. Although the trend of change can be different in different samples, the fact that the expression levels were significantly changed suggests the molecular pathways involving these genes are perturbed and those genes can be informative as regards the alterations in the biological processes.

A third feature of our approach is that we incorporate prior biological knowledge of genes into the data analysis. Clustering analysis has been widely used to gain a global picture of the groupings of the biological specimens used in transcriptome studies. However, this analysis is typically affected by 'noise' or non-informative genes. Therefore, some pre-screening approach is usually needed, especially when the differences in the informative genes are subtle. Our studies clearly showed that the use of informative gene sets provide much better separation between primary breast cancers and their metastases. Looking at the 446 informative genes, some have very little information regarding their cellular functions whereas others have clear functional ontology information. To gain more insight, we separated the 446 informative genes into different functional groups and used each group of genes in the further analysis. This knowledge-based analysis provides us with additional insight. Among the six functional groups, 'apoptosis' and 'cell adhesion and migration' genes did not separate primary breast cancers and their lymph-node metastases. In contrast, 'transcriptional factor and DNA binding factor' 'metabolism,' 'signal transduction' and 'cell cycle,' groups all separated the primary breast cancer and metastases much better, although the separation was never complete. This heterogeneity may reflect the different stages of the metastasis of different patient cancers.

The fact that apoptosis genes did not separate the two groups is consistent with the fact that there is no evidence in the literature showing that apoptosis status of primary breast cancer and their lymph node metastases are different. However, it is intriguing that cell migration and invasion genes as a group did not separate two groups of tissues because one might envision that cells that migrate to the second site would have different gene activities in migration and invasion. This would particularly be true if only a small clone of cells in the primary tumors is metastatic. Therefore, this result would argue that the primary tumors at the metastasis stage have systematically gained the potential to invade and metastasize to other sites.

The results show that metabolism and signal transduction genes separate primary tumors and their metastases. Because the tumors cells are located in two different cellular

environments, it is not surprising for them to have two different gene activities in signal transduction. The difference in metabolism is also meaningful. The cells that have to trek through circulation and grow in a foreign environment must have perhaps enhanced metabolism to achieve such a feat. Of note, metabolism has long been suspected to be a key cellular process involved in cancer. In the middle of the last century, Warburg proposed a hypothesis that the cause of cancer is primarily a defect in energy metabolism (20).

In summary, our gene function-based computational analysis has revealed some important cellular pathways that are important or affected when primary cells become lymph node metastases.

Acknowledgements

Authors wish to thank Sampsa Hautaniemi for stimulating discussions and useful suggestions. This study was partially supported by Tampere Graduate School in Information Science and Engineering (TISE) (H.L.), Academy of Finland (H.L., O.Y.-H.), the Tobacco Settlement Fund as appropriated to M.D. Anderson Cancer Center by the Texas Legislature, a generous donation from Koodorie Foundation, and the Tianjin Science and Technology Fund and a donation from TaiJi Co. Ltd.

References

1. Fisher ER, Palekar A, Rockette H, Redmond C and Fisher B: Pathologic findings from the National Surgical Adjuvant Breast Project (Protocol No. 4). V. Significance of axillary nodal micro- and macrometastases. *Cancer* 42: 2032-2038, 1978.
2. 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 and Friend SH: Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415: 530-536, 2002.
3. Van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrish M, Atsma D, Witteveen A, Glas A, Delahaye L, van der Velde T, Bartelink H, Rodenhuis S, Rutgers ET, Friend SH and Bernards R: A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* 347: 1999-2009, 2002.
4. Ramaswamy S, Ross KN, Lander ES and Golub TR: A molecular signature of metastasis in primary solid tumors. *Nat Genet* 33: 49-54, 2003.
5. Borg I and Groenen P: *Modern Multidimensional Scaling: Theory and Application*. Springer, New York, 1997.
6. Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, Meltzer P, Gusterson B, Esteller M, Kallioniemi OP, Wilfond B, Borg A and Trent J: Gene-expression profiles in hereditary breast cancer. *N Engl J Med* 344: 539-548, 2001.
7. Perou CM, Jeffrey SS, van de Rijn M, Rees CA, Eisen MB, Ross DT, Pergamenschikov A, Williams CF, Zhu SX, Lee JC, Lashkari D, Shalon D, Brown PO and Botstein D: Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc Natl Acad Sci USA* 96: 9212-9217, 1999.
8. 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 and Botstein D: Molecular portraits of human breast tumours. *Nature* 406: 747-752, 2000.
9. Valeriote F and van Putten L: Proliferation-dependent cytotoxicity of anticancer agents: a review. *Cancer Res* 35: 2619-2630, 1975.
10. 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, Eystein Lonning P and Borresen-Dale AL: Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci USA* 98: 10869-10874, 2001.
11. Badea L: Functional discrimination of gene expression patterns in terms of the gene ontology. In: *Pacific Symposium on Bio-computing* 8, pp565-576, 2003.
12. Hu L, Wang J, Baggerly K, Wang H, Fuller GN, Hamilton SR, Coombes KR and Zhang W: Obtaining reliable information from minute amounts of RNA using cDNA microarrays. *BMC Genomics* 3: 16, 2002.
13. Shmulevich I, Hunt K, El-Naggar A, Taylor E, Ramdas L, Laborde P, Hess KR, Pollock R and Zhang W: Tumor specific gene expression profiles in human leiomyosarcoma: an evaluation of intra-tumor heterogeneity. *Cancer* 94: 2069-2075, 2002.
14. Kobayashi T, Yamaguchi M, Kim S, Morikawa J, Ogawa S, Ueno S, Suh E, Dougherty E, Shmulevich I, Shiku H and Zhang W: Microarray reveals differences in both tumors and vascular specific gene expression in *de novo* CD5⁺ and CD5⁻ diffuse large B-cell lymphomas. *Cancer Res* 63: 60-66, 2003.
15. Wang H, Wang H, Shen W, Huang H, Hu L, Ramdas L, Zhou Y, Liao WSL, Fuller GN and Zhang W: Insulin-like growth factor binding protein 2 enhances glioblastoma invasion via activation of invasion enhancing genes. *Cancer Res* 63: 4315-4321, 2003.
16. Yang YH, Dudoit S, Luu P and Speed TP: Normalization for cDNA Microarray Data. In: *SPIE BiOS 2001*, San Jose, CA, 2001.
17. Duda RO, Hart PE and Stork DG: *Pattern Classification*. 2nd edition. Wiley, New York, NY, 2001.
18. Good P: *Permutation tests: a practical guide to resampling methods for testing hypotheses*. 2nd edition. Springer-Verlag, 2000.
19. Diehn M, Sherlock G, Binkley G, Jin H, Matese JC, Hernandez-Boussard T, Rees CA, Cherry JM, Botstein D, Brown PO and Alizadeh AA: Title. *Nucleic Acids Res* 31: 219-223, 2003.
20. Warburg O: On the origin of cancer cells. *Science* 123: 309-314, 1956.