Distinguishing primary breast cancers and their lymph node metastases by knowledge-based multidimensional scaling analysis

Harri Lähdesmäki[1,2], Xishan Hao[3], Baocun Sun[3], Limei Hu[1], Olli Yli-Harja[2], Gregory N. Fuller[1], Ilya Shmulevich[1], Wei Zhang[1]

[1]The University of Texas M. D. Anderson Cancer Center, Houston, Texas [2]Tampere University of Technology, Tampere, Finland [3]Tianjin Medical University Cancer Hospital, Tianjin, China

Purpose of the study

- Metastasis is a major adverse factor in breast cancer prognosis and lymph node metastasis is often the first step in the metastasis process.
- The goal is to combine gene expression profiling with knowledge-based multidimensional scaling (MDS) analysis to identify key biological pathways or processes that distinguish between primary breast cancers and their lymph node metastases.

Data acquisition

- To capture the subtle differences and to avoid compounding factors associated with the heterogeneities of different patients, the most ideal materials for such a study are paired primary tumors and lymph node metastases from the same patient.
- The gene expression profiles of 9 matched primary and metastases tumors acquired using cDNA microarrays generated by the Cancer Genomics Core Laboratory at MDACC were used in our study.

Breast cancer tissues:

The primary and lymph node metastases tissues were surgically removed from the patients as part of the treatment and a fraction of the tissues were snap-frozen in liquid nitrogen immediately after surgical resection. The tissues were evaluated with H&E staining by pathologist (BS). Only tissues with more than 75% tumor cells were used for microarray studies.

Microarray assay:

RNA isolation, microarray production, hybridization, and image analysis were carried out as previously described (Hu et al., 2002; Shmulevich et al., 2002; Kobayashi et al., 2003). Two microarrays (1500 and 2300 genes, respectively) generated by the Cancer Genomics Core Laboratory, M. D. Anderson Cancer Center, were used in this study.

Tissue microarray and immunohistochemistry:

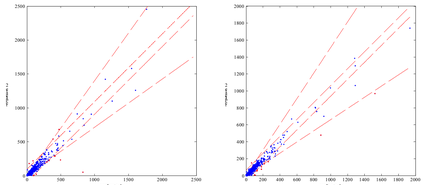
After screening H&E-stained slides for optimal tumor presence, a paired sample (primary/metastasis) tissue microarray consisting of 100 paired samples was constructed using 0.6 mm-diameter punch cores. Immunohistochemistry was performed as described previously (Wang et al., 2002) on tissue sections cut from the array block using polyclonal antibodies against IGFBP5 (Santa Cruz Biotechnology, Inc., Santa Cruz, CA), fibronectin (Zymed, Inc., S. San Francisco, CA), MMP2 (Zymed, Inc.), cyclin D1 (Zymed, Inc.) and mdm 2 (Zymed, Inc.). The staining index was evaluated and estimated by pathologists using established criteria (Wang et al., 2002).

Data Analysis

- The background-subtracted signal intensities were subjected to the following data analysis steps:
  - Preprocessing
  - Detecting differentially expressed genes
  - Multidimensional scaling
  - Multidimensional scaling together with *a priori* knowledge

Preprocessing

- Each gene expression value was replicated twice and “low quality” replicates were first detected by analyzing the differences between the replicates.
- For each set of primary and metastasis samples separately, histograms of differences and normalized differences were computed, which were further used to compute the standard deviation of the corresponding statistics.
- The standard deviations were used to define "quality checks" for the replicates: Each replicated measurement whose *(i)* absolute value of the difference was smaller than three times standard deviation of the differences, or *(ii)* absolute value of the normalized difference was smaller than two times standard deviation of the normalized differences passed the replicate quality control. Other replicates were considered to be of low quality, were flagged, and ignored in further analysis.

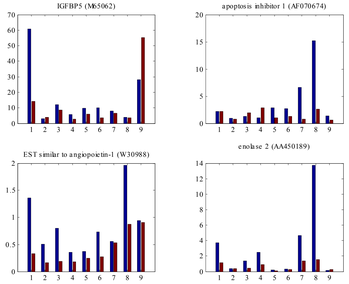


Example figures of replicated measurements. “Good spots” (blue) and “low quality spots” (red). Replicates of a primary tumor (left) and the corresponding metastases (right).

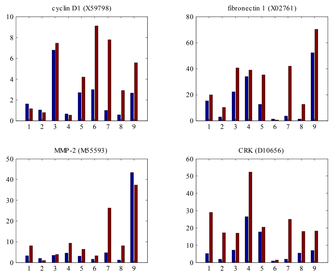
- Replicates for the measurements that passed the above test were averaged.
- The averaged measurements were further normalized by median normalization.

Differentially expressed genes

- A gene is defined to be significantly over-expressed (respectively, under-expressed) between the metastasis and primary sample if the ratio of expression level in metastasis sample to expression level in primary sample is at least two-fold (respectively, less than half), see e.g. (DeRisi et al. or Iyer et al.).
- In order to prevent the ratio from getting too "unstable", the minimum of non-normalized expression values in metastasis or primary samples is restricted to be at least five.
- The same test is applied to all genes and patient samples individually and the genes are then ranked based on the number of times they are either significantly over- or under-expressed.



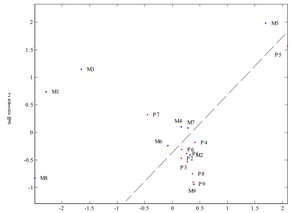
Examples of over-expressed genes in metastases (blue)



Examples of under-expressed genes in metastases (blue).

Multidimensional scaling

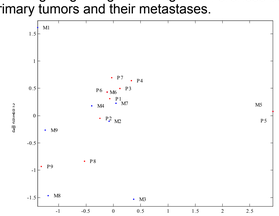
- The primary and metastases samples were clustered using the non-metric multidimensional scaling (Borg and Groenen, 1993).
- Unsupervised MDS clustering methods using all the 2,300 genes did not separate primary and metastases samples.
- Non-informative variables (genes) can have diminishing effects on clustering results. In order to reduce those effects, we reduced the number of genes before applying the MDS algorithm by considering only the genes that have significant variation in at least three patients out of nine. After that, we were left with 280 genes.
- MDS analysis using the 280 informative genes revealed that primary tumors were quite tightly clustered whereas the metastases samples were relatively heterogeneous.
- Dissimilarities for the MDS were computed as 1 - *r*, where *r* is the Spearman's correlation measure between gene expression profiles containing 280 genes.



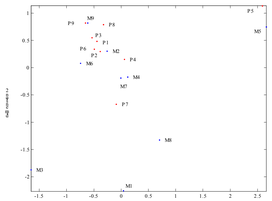
Result of applying multidimensional scaling to 9 primary and metastases samples with 280 genes.

Knowledge-based MDS

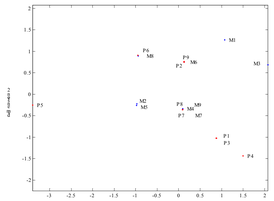
- The set of 280 informative genes was used in the second MDS analysis.
- That is, we separated the 280 genes into six functional groups based on prior knowledge of those genes: cell cycle, apoptosis, metabolism, cell adhesion and migration, signal transduction, and transcriptional factor, and DNA binding molecules.
- Each of the six functional gene sets was then used to perform knowledge-based MDS analyses and it showed that different functional gene sets vary in their ability to separate primary tumors and their metastases.
- The best separations were found with gene sets of "cell adhesion and migration," "metabolism," "signal transduction," and "transcription factor and DNA binding molecule."
- In contrast, cell cycle separated the two groups less well and apoptosis related genes did not separate the two groups at all.
- Results from the MDS analysis suggest that alteration in apoptosis regulation is not the key event leading to lymph node metastasis, whereas the alteration in cell migration and metabolism may be crucial. This is consistent with the prior knowledge regarding the biological differences between primary tumors and their metastases.



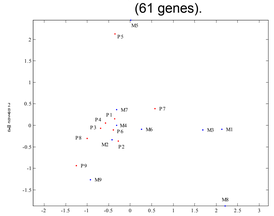
MDS result with genes related to apoptosis (8 genes).



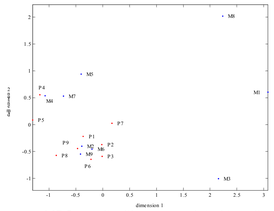
MDS result with genes related to metabolism (61 genes).



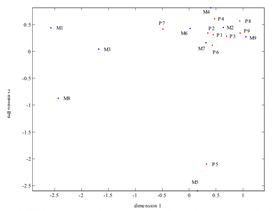
MDS result with genes related to cell cycle (6 genes).



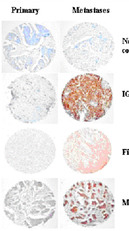
MDS result with genes related to signal transduction (48 genes).



MDS result with genes related to cell movement (28 genes).



MDS result with genes related to transcription (41 genes).



Analysis of the tissue microarray data

- One major obstacle is to obtain sufficient amount of paired surgical specimens from a large number of patients for genomics studies.
- Therefore, expression results of five genes were validated by a tissue microarray study using 100 paired samples measuring protein abundance levels between primary and metastases samples.
- Since the discrete-valued pair-wise tissue microarray measurements may be viewed as being subjective, different pairs may not necessarily be comparable. We circumvented that problem by comparing the differences between protein levels in primary and metastases only pair-wise. Moreover, in order to alleviate the non-quantitative nature of the data, we only used the information of the sign of the difference, i.e., the measured protein level is either higher or lower in metastases (or equal between primary and metastases).
- Since the measurements are not continuous-valued and, therefore, quite a large portion of the pair-wise differences are zero, the standard non-parametric tests, such as Wilcoxon's and Fisher's sign test, could not be reliably used. However, we found the bootstrap-based hypothesis testing method applicable and useful. As the null hypothesis, we selected the mean of pair-wise differences to be zero. The *p*-values were obtained using 100 000 bootstrap iterations. (For details of the bootstrap-based hypothesis testing, see Efron et al.)
- Summary of tissue array results for the five proteins tested

Protein name	Number of paired samples counted	Number of paired samples with staining levels M=P, M<P, M>P	*p*-value
IGFBP5	57	22, 9, 26	0.00332
fibronectin	84	17, 12, 55	0
MMP2	81	29, 15, 37	0.00197
cyclin D1	86	61, 16, 9	0.08052
mdm2	79	32, 20, 27	0.15011

M, metastases; P, primary. M=P, level of staining is equal between M and P; M<P, the level of staining in M is less than P; M>P, the level of staining in M is more than P.

Conclusions

- Incorporation of prior biological knowledge into MDS analysis may help reveal important biological mechanisms from gene expression profiles.
- Our MDS analysis showed that lymph node metastases represent a heterogeneous group of tumors with major alterations in cell adhesion and migration, metabolism, signal transduction, and transcription regulation.

References

- Hu, L., Wang, J., Baggerly, K., Wang, H., Fuller, G.N., Hamilton, S.R., Coombes, K.R., Zhang, W. Obtaining Reliable Information from Minute Amounts of RNA Using cDNA Microarrays. *BMC Genomics*, 3:16, 2002.
- Shmulevich, I., Hunt, K., El-Naggar, A., Taylor, E., Ramdas, L., Labordé, P., Hess, K.R., Pollock, R., Zhang, W. Tumor Specific Gene Expression Profiles in Human Leiomyosarcoma: an Evaluation of Intra-Tumor Heterogeneity. *Cancer*, 94:2069-75, 2002.
- 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 tumor and vascular specific gene expression in de novo CD5+ and CD5- diffuse large B-cell lymphomas. *Cancer Res.*, 63:60-66, 2003.
- Wang HM, Wang H, Zhang W, and Fuller GN. Tissue microarrays: applications in neuropathology research, diagnosis, and education. *Brain Pathology* 12:95-107, 2002.
- DeRisi, J., van den Hazel, B., Marc, P., Balzi, E., Brown, P., Jacq, C., and Goffeau, A. Genome microarray analysis of transcriptional activation in multidrug resistance yeast mutants. *FEBS Letters*, 470:156-160, 2000.
- Iyer, V. R., Eisen, M. B., Ross, D. T., Schuler, G., Moore, T., Lee, J. C. F., Trent, J. M., Staudt, L. M., Hudson Jr., J., Boguski, M. S., Lashkari, D., Shalon, D., Botstein, D., and Brown, P. O. The transcriptional program in the response of human fibroblasts to serum. *Science*, 283:83-87, 1999.
- Borg, I., and Groenen, P. *Modern Multidimensional Scaling: Theory and Application*, Springer, New York, 1997.
- Efron, B., and Tibshirani, R. J., *An Introduction to the Bootstrap*, Chapman & Hall, 1993.