AD_____

AWARD NUMBER: W81XWH-04-1-0570

TITLE: Endocrine Therapy of Breast Cancer

PRINCIPAL INVESTIGATOR: Robert Clarke, Ph.D.

CONTRACTING ORGANIZATION: Georgetown University
Washington, DC 20057-1411

REPORT DATE: June 2006

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
01-06-2006	Annual	17 May 2005 – 16 May 2006

4. TITLE AND SUBTITLE

Endocrine Therapy of Breast Cancer

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-04-1-0570

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Robert Clarke, Ph.D.

E-Mail: clarker@georgetown.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Georgetown University
Washington, DC 20057-1411

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

A recent controversy in the treatment of estrogen receptor positive (ER+) breast cancers is whether an aromatase inhibitor, e.g., letrozole (LET) or TAM should be given as first line endocrine therapy. Unfortunately, response rates are lower, and response durations are shorter, on crossover than when these agents are given as first line therapies, e.g., ~40% of tumors show cross resistance to TAM or an aromatase inhibitor on crossover. Only 50% of ER+ tumors respond to endocrine therapy. Currently, we fail to predict endocrine responsiveness in about 66% of ER+/PgR- (progesterone receptor), 55% of ER-/PgR+, and 25% of ER+/PgR+ tumors. In this new Clinical Translational Research Award, we hypothesize that our analytical methods can extract expression profiles of breast tumors that define their responsiveness (sensitive vs. resistant) to endocrine therapy. These profiles, when combined with known predictive/prognostic factors, will support neural network and biostatistical classifiers or committee machines that predict each tumor's endocrine responsiveness. Our objectives are to array breast cancer cases, build classifiers of endocrine responsiveness (using microarray data), and validate these classifiers in independent data sets using mostly immunohistochemistry data (IHC). IHC will be done on cases with definitive outcomes data. In the long term, we will design custom arrays for use in clinical practice. Genes will be further studied using cellular and molecular methods, and their role as therapeutic targets explored.

15. SUBJECT TERMS
Antiestrogen, aromatase inhibitor, anastrazole, bioinformatics, biomarkers, biostatistics, breast cancer, class prediction, clinical trial, computer science, engineering, immunohistochemistry, letrozole, microarrays, molecular profiling, neural networks, recurrence, resistance, tamoxifen

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			USAMRMC
U	U	U	UU	36	**19b. TELEPHONE NUMBER** *(include area code)*

TABLE OF CONTENTS

1. Wang, Z., Wang, Y., Xuan, J., Dong, Y., Bakay, M., Khan, J., Clarke, R. & Hoffman, E.P. "Optimized multilayer perceptrons for molecular classification and diagnosis using genomic data." *Bioinformatics*, 22: 755-761, 2006.

2. Zhu, Y., Wang, A., Liu, M.C., Zwart, A., Lee, R.Y., Gallagher, A., Wang, Y., Miller, W.R., Dixon, J.M. & Clarke, R. "Estrogen receptor alpha (ER) positive breast tumors and breast cancer cell lines share similarities in their transcriptome data structures." *Int J Oncol*, 29: 15812-1589, 2006.

3. Zhu, Y., Singh, B., Hewitt, S., Liu, A., Gomez, B., Wang, A. & Clarke, R. "Expression patterns among proteins associated with endocrine responsiveness in breast cancer: interferon regulatory factor-1, human X-box binding protein-1, nuclear factor kappa B, nucleophosmin, estrogen receptor-alpha, and progesterone receptor." *Int J Oncol*, 28: 67-76, 2006.

INTRODUCTION

Endocrine therapy is often the least toxic and most effective treatment for hormone receptor positive invasive breast cancer. Such therapy includes antiestrogens (tamoxifen, fulvestrant) and aromatase inhibitors (anastrozole, letrozole, exemestane). Tamoxifen (TAM) increases disease free and overall survival in the adjuvant setting, reduces the incidence of estrogen receptor positive disease (ER+; unless otherwise noted ER=ERα) in high-risk women, and reduces the rate of bone loss secondary to osteoporosis in postmenopausal women [1,2]. Aromatase inhibitors are effective only in the absence of functioning ovaries - TAM can be used regardless of menopausal status. Recent studies suggest that anastrozole may be superior to TAM in the adjuvant treatment of postmenopausal women with ER+ breast cancer; other studies report higher overall response rates with letrozole (LET) *vs.* TAM as first line therapy in the metastatic setting. Thus, a recent controversy in the management of patients with ER+ disease is whether an aromatase inhibitor or TAM should be given as first line endocrine therapy [3-9].

In this Clinical Translational Research award, we will build classifiers that accurately separate antiestrogen sensitive from antiestrogen resistant breast tumors and begin to assist in the direction of specific endocrine treatments (antiestrogen *vs.* aromatase inhibitor) to *individual* patients. We hypothesize that endocrine responsiveness is affected by a gene network, rather than the activity of only one or two genes or signaling pathways [10-12]. Since the key components of such a network are unknown, we must study 10,000s of genes. We will use Affymetrix GeneChips. We will not identify mutational events, the presence of mRNA splice variants, or post-translational protein modifications. However, these factors have major effects on the transcriptome and their "footprints" should be identified by expression microarrays.

BODY

Overview: We will build classifiers that separate antiestrogen sensitive from antiestrogen resistant breast tumors and begin to assist in the direction of specific endocrine treatments (antiestrogen *vs.* aromatase inhibitor) to *individual* patients. To achieve this goal, and consistent with a CTR award, we will complete a 4-year, prospective, neoadjuvant study with Letrozole (LET) or TAM as the only systemic therapy. We will obtain molecular profiles from Affymetrix GeneChips and further develop and apply our innovative bioinformatic and biostatistic methods to explore these high dimensional data sets and build/validate new classifiers. A more accurate predictor of endocrine responsiveness would have widespread clinical use, allowing women and physicians to make more individualized and appropriate treatment decisions. For example, patients with tumors predicted to be resistant to antiestrogens and/or aromatase inhibitors would be strong candidates for an early intervention with cytotoxic chemotherapy.

In most predictive/prognostic marker studies investigators focus on a *single* factor and whether they obtain a p-value that reaches conventional statistical significance. Our approach is different because we will determine whether we can find joint gene subsets that can separate patients into sufficiently distinct groups that should differ in their treatment. We will **(1)** analyze >33,000 genes on retrospective and prospective material, **(2)** apply new biostatistical and bioinformatic methods to identify ~40 potentially informative "biomarkers," **(3)** build neural network and biostatistical model classifiers, **(4)** evaluate the joint discriminant power of selected genes concurrently rather than as single biomarkers, **(5)** focus on prediction for individual patients where the assessment of a p-value is less important than the classification rate of our predictors, **(6)** validate the classifiers in independent data sets, and **(7)** explore the ability of predictors to refine the targeting of *specific* endocrine therapies.

Evidence has begun to accumulate suggesting that an aromatase inhibitor might be a more effective first line endocrine therapy for some breast cancer patients than the current standard of care (Tamoxifen). These data have generated considerable interest and controversy, in part because unlike TAM, there are no long term studies with aromatase inhibitors where definitive survival data are available. Our study could provide new and

innovative insights into how to approach the more effective targeting of specific endocrine therapies to individual patients.

Specific Aims (from the original application)

We will complete two clinical studies and collect gene expression profiles from which to build predictors of endocrine responsiveness. Predictors will be built in Specific Aim 2 and validated in Specific Aim 3.

AIM 1: Clinical Studies - **Clinical Study-1** (retrospective) is of pretreatment, single, frozen samples where we will compare the molecular profiles of tumors that recurred on TAM with those of tumors that did not recur. Each resistant sample is matched with a TAM sensitive sample by age, stage, and duration of follow-up. We also have further, single (unmatched), frozen samples from patients already progressing on TAM. **Clinical Study-2** is a prospective study of breast tumor samples from patients treated with neoadjuvant TAM or LET.

AIM 2: We will apply novel bioinformatics and biostatistics to discover gene subsets that define the molecular differences between endocrine sensitive and resistant breast tumors. These genes will be used, in combination with established predictive/prognostic factors, *e.g.*, ER, PgR, stage, to build innovative classifiers that can better predict an individual tumor's endocrine responsiveness.

AIM 3: We will test, optimize, and validate the performance of the classifiers from Aim 2 in retrospective studies of human breast tumors. We must measure each gene individually by IHC, *in situ* RNA hybridization (ISH), or *real time* PCR (RT-PCR).

KEY RESEARCH ACCOMPLISHMENTS

Progress on the clinical goals for this award was greatly delayed because of the time taken to obtain DOD approval of our preexisting institutionally approved IRBs at Georgetown University and at the University of Edinburgh. All institutionally approved protocols and requested material were submitted to the DOD in July 2004; additional information was requested by the DOD several months later and submitted in November 2004. We did not receive final approval to proceed with the clinical studies until March 2005. Much of this delay seems to have been entirely unavoidable and due, in part, to major personnel changes at the DOD (within USAMRMC). Clearly, this has likely left us behind schedule in recruitment to the prospective studies. As noted in the previous report and as is again apparent in this report, this did not affect our ability to proceed with the informatics studies (algorithm development and optimization) and infrastructure development (database development and installation). We have now published the completed studies presented in our preliminary data and generated a novel optimization protocol for our multilayer perceptron-based classifiers (also now published). The *in silico* tissue heterogeneity correction method described in the application was developed sufficiently and submitted for publication – this is still in revision. We have been able to proceed with analysis of some of the retrospective studies and have obtained data and performed initial unsupervised analyses of the first 60 specimens. Publications supported since the commencement of this award are listed under "Reportable Outcomes"; these constitute some of our major accomplishments in the past year. These and other key research accomplishments are presented below.

Statement of Work (from the original application)

- **TASK 1.** Array breast tumor samples from Clinical Studies 1 (retrospective) and 2 (prospective)

To perform this task we will obtain breast tumor samples and clinical information from University of Edinburgh, collect and quality test RNA using validated tissue acquisition and processing protocol, and array RNA samples on oligonucleotide chips (*i.e.,* U133A Affymetrix GeneChips). ***Please note that we originally described analyses of approximately 12,000 genes in each sample and now indicate that we will measure***

almost 3-times as many genes. The increase is possible because Affymetrix improved their technology and now produce single chips with 40,000 probe sets representing 39,000 transcripts, of which 33,000 are well-substantiated genes. The cost of these chips, which essentially represent the probe sets previously included on two chips (U133A and U133B), is the same as the original U133A GeneChips described when the application was submitted. Since we were unable to start arraying in year 1 (see below), this proved very fortunate, since it greatly increases the power of our study to detect meaningful predictive patterns and genes or networks associated with the clinical outcomes.

We have received a total of 480 breast specimens (to date) from our collaborators at the University of Edinburgh; these have arrived at different times and been banked so that they could be processed in the most effective and logical manner. Of these 480 specimens, we have (to date) had 173 processed as frozen sections and analyzed by the study pathologist. A further 67 have been processed as frozen sections and will be analyzed by the study pathologist within days of the submission of this report. We have successfully extracted total RNA from 172 specimens, and labeled 160 for analysis. We have also completed the hybridization and assessment of microarray data quality control on 72 specimens; 60 were done on U133 plus 2 GeneChips (the other 12 were used on older chips to test and optimize our methods for these specimens), representing sufficient data for our first TAM study. We requested that the specimens be sent independent of the clinical information, so that we could adequately and appropriately randomize the RNA preparation, labeling and hybridization and minimize any operator-induced or technology-induced bias. All specimens were processed using our standard operating procedures; each manipulation being performed by the same individual to further reduced inter-operator variability.

Using industry-standard internal controls and spike-in controls as recommended by the manufacturer, the microarray data obtained appears to be of very high quality and reproducibility. In the absence of the clinical information required for supervised analysis these data, we have begun initial unsupervised analyses. These are only exploratory and the choice to publish the results of these ongoing (unsupervised) will largely depend on how well they capture treatment or recurrence status (when that information is available).

We used the current standard RMA algorithms to achieve a log2-based normalization of the data from the first 60 samples arrayed on the HU133 plus 2 Affymetrix GeneChip microarrays. The goal was to separate blindly (without supervision) the composite signatures. Samples were initially clustered (*Sample Clustering*) using two different methods – *K*-mean and Self Organizing Map (SOM) – and both methods produced consistent representations of the data as comprising three main clusters. Sample Clusters 1 and 3 each contained 14 specimens, the remaining 32 specimens comprising Sample Cluster 2. We then performed *Gene Clustering* by SOM and identified 65 potential gene clusters. From within these gene clusters, we identified those in which the genes are highly expressed in one sample cluster and exhibit consistently low expression in the remaining two sample clusters. Thus, we identified Gene Clusters that best define (by these criteria) Sample Cluster 1 (gene clusters 49, 50), Sample Cluster 2 (gene cluster 19), and Sample Cluster 3 (gene clusters 58 and 59). Currently, we are attempting to explore further these clusters and we will be in a better position to do so once we obtain the clinical information. We have now requested this information on these samples, since there is no longer a need to remain blinded to these clinical data.

We also attempted to apply the more common hierarchical clustering approach (also an unsupervised method) to this data set but this was entirely uninformative. We also tried clustering using only those genes previously reported to generate the groups commonly referred to as “luminal A”, “luminal B”, “basal-like”, “her2/neu2”, and “normal like”. None of these clusters were evident in our data set using this approach. To some degree, this may reflect the very high proportion of ER+ tumors, which would suggest that the “luminal A” and “luminal B” groups would be present (perhaps two major clusters), but these two clusters also were not immediately evident. However, this is a rather simplistic analysis method and is probably not capable of identifying what may be more subtle differences among the phenotypes represented in our data set.

- **TASK 2.** Store, process, and train/optimize classifiers from gene expression microarray data (modified to reflect our adoption of caArray)

To perform this task we will install and modify the MIAME Compliant caArray database. We will also collect and store de-identified clinical information and process gene expression data with "in house" state of art algorithms (we will also further develop and optimize these algorithms throughout this award period). For the initial studies, we will train/optimize initial neural network RNA classifier (MLP), the final classifier for the microarray data will be built when we have completed arraying all samples.

As noted in our previous report, we continue to make significant progress on addressing this task, largely as a consequence of our involvement in the National Cancer Institute Center for Bioinformatics (NCICB) led caBIG project. The PI (Dr. Clarke) leads the Lombardi Comprehensive Cancer Center's caBIG team and we have been actively involved in the development of caArray (NCICB's grid-enabled, MIAME compliant, microarray database). The caBIG program is open source-open access and is widely supported by NCICB and teams of collaborating scientists at other Cancer Centers across the country. We also have found the NCICB team highly responsive when we identify bugs or problems with the software. While NCICB has had some problems with the current version of caArray, we worked closely with their team and other Cancer Centers in caBIG to find and address some of these issues. A new version of caArray will likely be operational at our center before this report is fully reviewed, since we already are in the process of installation and testing. We anticipate that continued collaboration through the caBIG community will prove a more cost and time efficient approach to developing some components of the research infrastructure described in the original application. It is out intent to build any additional components in a manner consistent with the guidelines established by the caBIG community, since this will likely ensure long-term viability and the compatibility of our infrastructure.

With respect to the further development and optimization of data analysis algorithms, we have recently completed and published a new method for optimizing the use of multilayer perceptron (MLP) classifiers. MLPs are one of the most widely used and effective machine learning methods currently applied to diagnostic classification using high-dimensional genomic data. Based on Fisher linear discriminant analysis, we designed and implemented an MLP optimization scheme for a two-layer MLP that effectively optimizes the initialization of MLP parameters and the MLP architecture. In comparison with a conventional MLP using random initialization, we obtained significant improvements in major performance measures including Bayes classification accuracy, convergence properties, and area under the receiver operating characteristic curve (Az). This work is now published in the journal *Bioinformatics.*

We also continue to improve our existing algorithms. Our most recent studies in this regard have been to improve the VISDA algorithms described in the initial application and to begin developing novel approaches that will allow us to extract gene signaling networks from the microarray data. This will potentially allow us to obtain mechanistic insights from the data we are generating from the clinical specimens. While we had not included this possibility in the original application, which focused on classification, we see the potential to obtain novel mechanistic insights as a significant advantage to our ongoing studies. We will provide additional information in this regard in subsequent reports; relevant publications in this area are included below.

- **TASK 3.** Retrain/reoptimize classifiers using IHC data from Series 1 (Archival Tissues) and Series 2 (Scottish Adjuvant TAM Trial) for Validation

To perform this task we will obtain clinical information and breast tumor samples from University of Edinburgh (formalin fixed/paraffin embedded). We will rank and prioritize selected joint genes from RNA classifier built and optimized in TASK 2 (above) and retrain/reoptimize the initial neural network IHC classifier (MLP).

Finally, we will validate IHC classifier on independent data sets (data sets not used to build and train the MLP classifiers).

We will not be able to start this task on the timeframe as initially proposed here because of the delays in getting approval to work with the clinical specimens. However, we expect to receive clinical information on the samples already arrayed within the next 4-6 weeks. This will allow us to perform (and hopefully submit for publication) an initial analysis of the first retrospective TAM study.

REPORTABLE OUTCOMES

Papers and Meeting Reports*

Updates (cited as "in press" in the last report and now in print)

- Zhu, Y., Singh, B., Hewitt, S., Liu, A., Gomez, B., Wang, A. & Clarke, R. "Expression patterns among proteins associated with endocrine responsiveness in breast cancer: interferon regulatory factor-1, human X-box binding protein-1, nuclear factor kappa B, nucleophosmin, estrogen receptor-alpha, and progesterone receptor." *Int J Oncol*, 28: 67-76, 2006.

- Xuan, J., Dong, Y., Khan, J., Hoffman, E., Clarke, R. & Wang, Y. "Robust feature selection by weighted Fisher criterion for multiclass prediction in gene expression profiling." *Proc 17th Intl Conf Pattern Recon*, 2: 291-294, 2004.

- Riggins, R.B., Bouton, A.H., Liu, M.C. & Clarke, R. "Antiestrogens, aromatase inhibitors, and apoptosis in breast cancer." *Vit Horm*, 71: 202-237, 2005.

- Bouker, K.B., Skaar, T.C., Hamburger, D.S., Riggins, R.B., Fernandez, D.R., Zwart, A., Wang, A. & Clarke, R. "Tumor suppressor activities of interferon regulatory factor-1 in human breast cancer associated with caspase activation and induction of apoptosis." *Carcinogenesis*, 26:1527-1535, 2005.

New Publications (published and "in press" for the present reporting period)

- Wang, Z., Wang, Y., Xuan, J., Dong, Y., Bakay, M., Khan, J., Clarke, R. & Hoffman, E.P. "Optimized multilayer perceptrons for molecular classification and diagnosis using genomic data." *Bioinformatics*, 22: 755-761, 2006.

- Zhu, Y., Wang, A., Liu, M.C., Zwart, A., Lee, R.Y., Gallagher, A., Wang, Y., Miller, W.R., Dixon, J.M. & Clarke, R. "Estrogen receptor alpha (ER) positive breast tumors and breast cancer cell lines share similarities in their transcriptome data structures." *Int J Oncol*, 29: 15812-1589, 2006.

- Ressom, H.W., Zhang, Y., Xuan, J., Wang, Y. & Clarke, R. "Inference of gene regulatory networks from time course gene expression data using neural networks and swarm intelligence." *Proc IEEE Symp Compl Intel Bioinformatics Comput Biol*, in press.

- Ressom, H., Xuan, J., Wang, Y. & Clarke, R. "Classification of microarray data using machine learning methods." *TIBETS*, in press.

- Xuan, J., Wang, Y., Clarke, R. & Hoffman, E.P. "Normalization of microarray data by iterative nonlinear regression." *5th IEEE Symposium on Bioinformatics and Bioengineering*, Minneapolis, Minnesota, pp. 267-270, 2005

*We include in the appendix reprints of those papers that are already published. Manuscripts cited as "in press" will be included in the next annual report, once reprints are available. We do not list here or include in the appendices any published abstracts, but can do so if requested. Several other manuscripts related to our bioinformatic methods also are submitted and in preparation – these will be cited reported in the next report. Please note that the papers published in the engineering literature are different from most conference proceedings in the biomedical literature. These are not abstracts but fully peer-reviewed publications comparable to short communications in biomedical journals.

Comment on Subcontracts: Please also note that the majority of our publications include coauthors from one or both of our subcontracts. Thus, our program is working very effectively and collaboratively, this also should be apparent in the development of new informatics methods (Catholic University of America – now Virginia Polytechnic and State University subcontract – Dr. Xuan recently moved to Virginia Tech) and the large number of high quality breast tumor specimens we have obtained from the University of Edinburgh.

CONCLUSIONS

We have made good progress on the research infrastructure goals and in the development or optimization of the methods needed for data analysis. We also have completed and published most of the data presented as preliminary data in the initial application. The clinical studies were held up by an unexpectedly long delay in obtaining final approval for our existing protocols but this is now taken care of and we are poised to begin analysis of our first series of breast cancer specimens.

REFERENCES

1. Early Breast Cancer Trialists' Collaborative Group. Tamoxifen for early breast cancer: an overview of the randomized trials. Lancet, *351:* 1451-1467, 1998.

2. Early Breast Cancer Trialists Collaborative Group: Systemic treatment of early breast cancer by hormonal, cytotoxic, or immune therapy. Lancet, *399:* 1-15, 1992.

3. Winer, E. P., Hudis, C., Burstein, H. J., Chlebowski, R. T., Ingle, J. N., Edge, S. B., Mamounas, E. P., Gralow, J., Goldstein, L. J., Pritchard, K. I., Braun, S., Cobleigh, M. A., Langer, A. S., Perotti, J., Powles, T. J., Whelan, T. J., and Browman, G. P. American Society of Clinical Oncology technology assessment on the use of aromatase inhibitors as adjuvant therapy for women with hormone receptor-positive breast cancer: status report 2002. J Clin Oncol, *20:* 3317-3327, 2002.

4. Ravdin, P. Aromatase inhibitors for the endocrine adjuvant treatment of breast cancer. Lancet, *359:* 2126-2127, 2002.

5. Bonneterre, J., Buzdar, A., Nabholtz, J. M., Robertson, J. F., Thurlimann, B., von Euler, M., Sahmoud, T., Webster, A., and Steinberg, M. Anastrozole is superior to tamoxifen as first-line therapy in hormone receptor positive advanced breast carcinoma. Cancer, *92:* 2247-2258, 2001.

6. Baum, M., Buzdar, A. U., Cuzick, J., Forbes, J., Houghton, J. H., Klijn, J. G., Sahmoud, T., and ATAC Trialists Group Anastrozole alone or in combination with tamoxifen versus tamoxifen alone for adjuvant treatment of postmenopausal women with early breast cancer: first results of the ATAC randomised trial. Lancet, *359:* 2131-2139, 2002.

7. Ellis, M. J., Coop, A., Singh, B., Mauriac, L., Llombert-Cussac, A., Janicke, F., Miller, W. R., Evans, D. B., Dugan, M., Brady, C., Quebe-Fehling, E., and Borgs, M. Letrozole is more effective neoadjuvant

endocrine therapy than tamoxifen for ErbB-1- and/or ErbB-2-positive, estrogen receptor- positive primary breast cancer: evidence from a phase III randomized trial. J Clin Oncol, *19:* 3808-3816, 2001.

8. Miller, W. R., Anderson, T. J., and Dixon, J. M. Anti-tumor effects of letrozole. Cancer Invest, *20 Suppl 2:* 15-21, 2002.

9. Smith, I. E. and Dowsett, M. Aromatase inhibitors in breast cancer. N Engl J Med, *348:* 2431-2442, 2003.

10. Clarke, R., Leonessa, F., Welch, J. N., and Skaar, T. C. Cellular and molecular pharmacology of antiestrogen action and resistance. Pharmacol Rev, *53:* 25-71, 2001.

11. Clarke, R. and Brünner, N. Acquired estrogen independence and antiestrogen resistance in breast cancer: estrogen receptor-driven phenotypes? Trends Endocrinol Metab, *7:* 25-35, 1996.

12. Clarke, R., Skaar, T. C., Bouker, K. B., Davis, N., Lee, Y. R., Welch, J. N., and Leonessa, F. Molecular and pharmacological aspects of antiestrogen resistance. J Steroid Biochem Mol Biol, *76:* 71-84, 2001.

Bioinformatics Advance Access published January 10, 2006

BIOINFORMATICS

Optimized multilayer perceptrons for molecular classification and diagnosis using genomic data

Zuyi Wang[1,2], Yue Wang[3*], Jianhua Xuan[2], Yibin Dong[3], Marina Bakay[1], Yuanjian Feng[3], Robert Clarke[4] and Eric P. Hoffman[1]

[1]Center for Genetic Medicine, Children's National Medical Center, Washington, DC 20010, USA, [2]Department of Electrical Engineering and Computer Science, The Catholic University of America, Washington, DC 20064, USA, [3]The Bradley Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Arlington, VA 22203, USA, [4]Departments of Oncology, Physiology & Biophysics, Lombardi Comprehensive Cancer Center, Georgetown University, Washington, DC 20007, USA

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Multilayer Perceptrons (MLP) represent one of the widely used and effective machine learning methods currently applied to diagnostic classification based on high dimensional genomic data. Since the dimensionalities of the existing genomic data often exceed the available sample sizes by orders of magnitude, the MLP performance may degrade due to the curse of dimensionality and over-fitting, and may not provide acceptable prediction accuracy.

Results: Based on Fisher linear discriminant analysis, we designed and implemented an MLP optimization scheme for a two-layer MLP that effectively optimizes the initialization of MLP parameters and MLP architecture. The optimized MLP consistently demonstrated its ability in easing the cure of dimensionality in large microarray data sets. In comparison with a conventional MLP using random initialization, we obtained significant improvements in major performance measures including Bayes classification accuracy, convergence properties, and area under the receiver operating characteristic curve (Az).

Contact: yuewang@vt.edu

1 INTRODUCTION

Diagnostic classification with genomic data refers to the assignment of a particular unknown tissue sample to a known disease class based on its quantitative mRNA expression pattern from microarrays. This classification can be performed by a trained predictive classifier, such as a neural network classifier. This approach is particularly helpful for diagnosing complex genetic disease subtypes or stages whose subtle differences may be difficult to recognize by traditional clinical and pathological approaches (Bittner *et al.*, 2000; Brown *et al.*, 2000; Khan *et al.*, 2001; Mjolsness *et al.*, 2001; Ramaswamy *et al.*, 2001; Shipp *et al.*, 2002; West *et al.*, 2002; Linder, *et al.*, 2004; O'Neill, *et al.*, 2003; Wei, *et al.*, 2005). A common type of neural network classifier applied to diagnostic classification is feed-forward back-propagation Multilayer Perceptrons (MLP) (Figure 1). Input vectors and the corresponding target vectors are used to train an MLP, a process that updates the weights and biases until the MLP can approximate a mapping function that associates input vectors with specific output vectors. The generalization property makes it possible to train an MLP with a representative set of input/target pairs and get good results for predicting unseen input samples. The ability of an MLP to learn complex (nonlinear) and multidimensional mapping from a collection of examples makes it an ideal classifier for diagnostic classification (Haykin, 1999; Khan *et al.*, 2001; O'Neill, *et al.*, 2003; Wei, *et al.*, 2005).

Despite reported successful studies on applying MLPs to diagnoses with genomic data, such as gene expression microarray data (Khan *et al.*, 2001; Linder, *et al.*, 2004; O'Neill, *et al.*, 2003; Wei, *et al.*, 2005), the most critical problem, that of the curse of dimensionality, has not been effectively addressed. The curse of dimensionality is caused by the finite amount of training data available relative to the large input feature space. Accordingly, when the dimensionality increases considerably and the available information remains inadequate, the large number of model parameters in the classifier cannot be well-trained (Haykin, 1999, Jain *et al.* 2000). Consequently, the classifier performance may degrade beyond a certain point with the increasing inclusion of features or dimensions. In mRNA microarray experiments, there is typically an extremely ill-conditioned ratio of sample number (10's to 100's) to dimension number (probe or probe sets typically >10,000), which greatly augments the impact of the curse of dimensionality (Fukunaga, 1990; Haykin, 1999). In current studies, the approaches to avoiding the curse of dimensionality are generally limited to directly reducing the number of inputs. The commonly applied methods include conventional dimensionality reduction methods, such as principal component analysis (Khan, *et al.* 2001, Wei, *et al.*, 2005), t-statistics (Golub, *et al.*, 1999), correlation measure (van't Veer, *et al.*, 2002), and an MLP training-based gene selection procedure that selects genes with greater influence on the changes of outputs in an MLP (O'Neil *et al.*, 2003).

The design parameters in training an MLP include initial values of the model parameters (synaptic weights and biases), stopping rules, and MLP architecture, etc. Since no effective algorithms are available to search for a global optimum and traditional MLP initialization is done randomly, classification performance depends largely on the initial values of weights and biases. Furthermore, the higher complexity of the classifiers often results in more local minima in the error surface, and the classifier trainings can easily be trapped into such local minima (Raudys and Skurikhina, 1992; Raudys, 1994).

We hypothesized that developing an optimization of MLP initialization will allow the reduction of the curse of dimensionality,

*To whom correspondence should be addressed.

© The Author (2006). Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

and therefore improve performance of the MLP. Our goal was to find an effective nonrandom initialization scheme that places the initial state of an MLP closer to the optimal solution that is later sought by training (Wang, *et al.*, 2004). This approach is bolstered by previous studies in statistical pattern recognition field, where it has been shown that nonrandom initializations of MLP weights and biases resulted in the MLP with small generalization error even when the number of samples is smaller than the number of features or dimensions (Raudys, 1994; Raudys, 1997; Raudys and Skurikhina, 1992).

2 THEORY AND METHOD

2.1 wFC-Based MLP Initializations

2.1.1 Linear Dimension Reduction and MLP Feature Extraction – Hidden Layer Initialization

The MLP offers an integrated procedure for feature extraction and Bayes classification by learning the decision boundary (Haykin, 1999). Its feed-forward auto-associative architecture can also be used to construct nonlinear subspaces in a supervised or unsupervised mode (Haykin, 1999; Jain *et al.*, 2000). The output of the hidden layer may be interpreted as a set of new features presented to the output layer for classification (Haykin, 1999). On the other hand, multi-class linear discriminant analysis provides a multivariate prediction by estimating the density function. Its subspaces that are extracted based on the weighted Fisher Criterion (wFC), retain most closely the intrinsic Bayes separability (Loog, *et al.*, 2001). It can be shown that the determination of the linear dimension reduction (LDR) transformation is equivalent to finding the maximum-likelihood parameter estimates of a Standard Finite Normal Mixture (SFNM) model (Loog, *et al.*, 2001). This motivates an exploration of the connections between MLP and LDR. A natural hypothesis is that the class labels used as targets during supervised training force the outputs of the hidden layer to capture the most discriminatory components or subspaces for distinguishing the classes. Based on these theoretical observations, we suggest a wFC-based initialization mechanism for the MLP hidden layer (Wang, *et al.*, 2004). To limit the complexity of the MLP, we assume that the number of neurons in the hidden layer is smaller than the number of inputs.

Given an m_0-dimensional input **t**-space with K_0 classes, the multi-class LDR searches for a linear transformation **W** that transforms the original input space to a lower m_1-dimensional feature **x**-space ($m_1 < m_0$); the extracted **x**-space should preserve the maximum amount of class discriminatory information. Since it is too complex to directly use the Bayes error as a criterion, the most common technique for finding this transformation is LDR that is based on Fisher criterion (Jain *et al.,* 2000; Haykin, 1999). This method maximizes the ratio of the between-class scatter matrix to the within-class scatter matrix, thereby guaranteeing maximal separability. In this paper, we apply the wFC to the multi-class classification problem (Loog *et al.*, 2001), and the wFC is defined as,

$$J_{\mathrm{wFC}}(\mathbf{W}) = \sum_{k=1}^{K_0-1}\sum_{l=k+1}^{K_0} \pi_k \pi_l \omega(\Delta_{kl})\, trace(\mathbf{W}^{\mathrm{T}} \mathbf{S}_{\mathbf{t}w}^{-1} \mathbf{S}_{\mathbf{t}kl} \mathbf{W}), \qquad (1)$$

where **W** is the linear transformation matrix, π_k and π_l are the prior probabilities of classes k and l respectively, $\mathbf{S}_{\mathbf{t}w} = \sum_1^{K_0} \pi_k \mathbf{C}_{\mathbf{t}k}$ is the total within-class scatter matrix, and $\mathbf{S}_{\mathbf{t}kl} = (\boldsymbol{\mu}_{\mathbf{t}k} - \boldsymbol{\mu}_{\mathbf{t}l})(\boldsymbol{\mu}_{\mathbf{t}k} - \boldsymbol{\mu}_{\mathbf{t}l})^{\mathrm{T}}$ is the between-class scatter matrix for classes k and l. $\omega(\Delta_{kl})$ is the weighting function defined as,

$$\omega(\Delta_{kl}) = \frac{1}{2\Delta_{kl}^2} erf\left(\frac{\Delta_{kl}}{2\sqrt{2}}\right) \qquad (2)$$

where $\Delta_{kl} = [(\boldsymbol{\mu}_{\mathbf{t}k} - \boldsymbol{\mu}_{\mathbf{t}l})^{\mathrm{T}} \mathbf{S}_{\mathbf{t}w}^{-1} (\boldsymbol{\mu}_{\mathbf{t}k} - \boldsymbol{\mu}_{\mathbf{t}l})]^{1/2}$ is the Mahanalobis distance between classes k and l with class mean vector $\boldsymbol{\mu}_{\mathbf{t}}$ and covariance matrix $\mathbf{C}_{\mathbf{t}}$.

It has been shown that when there are more than two classes to be classified, the conventional multi-class Fisher criterion (cFC) for deriving dimension-reduced subspace is suboptimal with respect to classification (Loog *et al.*, 2001). The reason is that the cFC treats class pairs with various between-class distances equally. In contrast, the wFC incorporates a weight function that approximates the Bayes error rate between classes, and assigns larger weights to the closer class pairs and smaller weights to the distant pairs. Thus, in the extracted subspace found by wFC, the classes with heavy overlap gain adequate emphases, and the distant pairs remain well separated.

Finding a solution **W** that maximizes the wFC is essentially a problem of eigenvalue decomposition of the total Fisher scatter matrix,

$$\mathbf{S}_{\mathbf{t}w}^{-1} \sum_{k=1}^{K_0-1}\sum_{l=k+1}^{K_0} \pi_k \pi_l \omega(\Delta_{kl}) \mathbf{S}_{\mathbf{t}kl} \qquad (3)$$

By taking only the m_1 eigenvectors corresponding to the m_1 largest eigenvalues ($m_1 < m_0$), we can form a transformation that not only reduces the dimensionality of the original input space, but also retains maximal class separability information. We call this procedure wFC-Discriminatory Component Analysis (wFC-DCA).

With the transformation **W** ($m_0 \times m_1$) derived from LDR, the dimension-reduced feature subspace (**x**-space) with m_1 dimensions becomes $\mathbf{x}_i = \mathbf{W}^{\mathrm{T}}(\mathbf{t}_i - \mathbf{b}_{\mathbf{t}0})$, for $i = 1,\ldots,N$, where N is the number of samples, $\mathbf{x}_i$ is the representation of the sample vector $\mathbf{t}_i$ in the **x**-space with $x_{r,i} = \mathbf{w}_r^{\mathrm{T}}(\mathbf{t}_i - \mathbf{b}_{\mathbf{t}0})$ for $r = 1,\ldots,m_1$, and $\mathbf{b}_{\mathbf{t}0}$ is the global center of the data set. On the other hand, the outputs of the hidden layer in the MLP (Figure 1) can be acquired as, $a_n = \varphi(\mathbf{w}_n^{1\mathrm{T}}\mathbf{p} - b_{1,n})$,

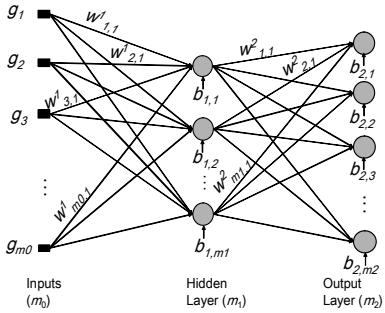


Figure 1. The general architecture of a two-layer MLP. The inputs and the layers of neurons are connected through sets of synaptic weights, e.g., $w_{1,1}^1$, and each neuron has an individual bias, e.g., $b_{1,1}$.

where $\mathbf{w}_n^1$ is the set of synaptic weights connecting m_0 inputs to neuron n at the hidden layer, a_n is the output of neuron n, $\mathbf{p}$ is the MLP input vector, $b_{1,n}$ is the bias of hidden neuron n, and $\varphi(\cdot)$ is an activation function (Haykin 1999). The connection between the LDR and the MLP feature extraction mechanism now becomes clearer, suggesting that the column vectors of the LDR matrix **W** can be used to initialize the weights between the input and hidden layer of an MLP, $\mathbf{w}_n^1 = \mathbf{w}_n$, and their biases can be initialized as,

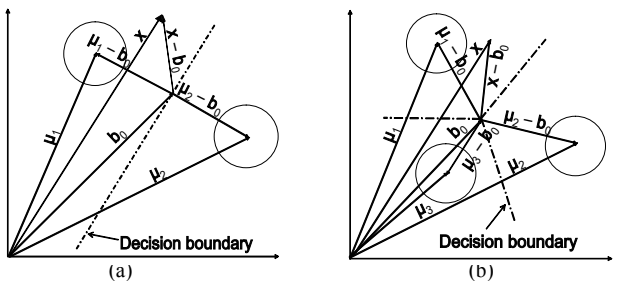


Figure 2. The illustrations of the MLP output layer initialization approach.

$\mathbf{b}_1 = \mathbf{W}^{\mathrm{T}}\mathbf{b}_{\mathbf{t}0}$. The new features are further scaled by the activation function $\varphi(\cdot)$ that could be linear or nonlinear. It has been theoretically shown that the minimization of the Bayes error with respect to the synaptic weights and biases of the MLP is equivalent to maximizing the wFC (Eq. (1)), and it can be entirely determined by the hidden neurons (Haykin 1999).

2.1.2 Linear Discriminant Analysis and Multi-class Perceptrons – Output Layer Initialization

Since the outputs of the hidden layer serve as new features, Fisher Linear Discriminant Analysis (LDA) determines a linear transformation for converting an m_1-dimensional problem to a one-dimensional problem (Haykin 1999). Consider the variable $y = \mathbf{w}^{\mathrm{T}}\mathbf{x} - b$ that is transformed from **x**-space to a one-dimensional space via LDA, and the LDA is defined by

$$J(\mathbf{w}) = \frac{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathbf{x}kl}\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathbf{x}w}\mathbf{w}} \qquad (4)$$

that is known as the generalized Rayleigh quotient. The solution that maximizes $J(\mathbf{w})$ is simply $\mathbf{w} = \mathbf{S}_{\mathbf{x}w}^{-1}(\boldsymbol{\mu}_{\mathbf{t}k} - \boldsymbol{\mu}_{\mathbf{t}l})$, which is also a generalized eigenvalue problem.

The neuron in the output layer behaves similarly as a perceptron that can be considered as a decision making element that bears a close resemblance to the Bayes classifier, and has been generalized to multiple classes (Haykin 1999). Specifically, the outputs of the neurons in the output layer are computed as, $y_i = \varphi(\mathbf{w}_i^{2\mathrm{T}}\mathbf{a} - b_{2,i})$ for $i = 1,\ldots,m_2$, where **a** is the output vector of the hidden layer, $\mathbf{w}_i$ is the set of weights connecting the hidden layer and the output neuron i, $b_{2,i}$ is the bias of output neuron i, and m_2 is the number of the output neurons, *i.e.*, the number of classes (Figure (1)). Consider a two-class case with a linear activation function $\varphi(\cdot)$, we have $y = \mathbf{w}^{\mathrm{T}}\mathbf{x} - b$ with $\mathbf{w} = \mathbf{S}_{\mathbf{x}w}^{-1}(\boldsymbol{\mu}_{\mathbf{t}1} - \boldsymbol{\mu}_{\mathbf{t}2})$ and $b = \mathbf{w}^{\mathrm{T}}\mathbf{b}_{\mathbf{x}0}$, where $\mathbf{b}_{\mathbf{x}0} = (\boldsymbol{\mu}_{\mathbf{t}1} + \boldsymbol{\mu}_{\mathbf{t}2})/2$. We can use two output neurons to derive a class-dependent representation by rearranging the output as, $y = \mathbf{w}^{\mathrm{T}}\mathbf{x} - b = (\mathbf{w}_1^{\mathrm{T}}\mathbf{x} - b_1) - (\mathbf{w}_2^{\mathrm{T}}\mathbf{x} - b_2) = y_1 - y_2$, where $\mathbf{w}_1 = \mathbf{S}_{\mathbf{x}w}^{-1}\boldsymbol{\mu}_{\mathbf{x}1}$, $\mathbf{w}_2 = \mathbf{S}_{\mathbf{x}w}^{-1}\boldsymbol{\mu}_{\mathbf{x}2}$, $b_1 = \mathbf{w}_1^{\mathrm{T}}\mathbf{b}_{\mathbf{x}0}$, and $b_2 = \mathbf{w}_2^{\mathrm{T}}\mathbf{b}_{\mathbf{x}0}$, so we have $y_1 = \mathbf{w}_1^{\mathrm{T}}\mathbf{x} - b_1$ and $y_2 = \mathbf{w}_2^{\mathrm{T}}\mathbf{x} - b_2$. Figure 2a illustrates such an interpretation. Based on the above derivation, the class-dependent Fisher linear discriminant transformation $\mathbf{w}_i$ can be again used to initialize the weights between the hidden and output neurons as, $\mathbf{w}_i^2 = \mathbf{S}_{\mathbf{x}w}^{-1}\boldsymbol{\mu}_{\mathbf{x}i}$, and the biases of the output neurons can be, $b_{2,i} = \mathbf{w}_i^{2\mathrm{T}}\mathbf{b}_{\mathbf{x}0}$ for $i = 1,2$. Accordingly, for a three-class case, it is straightforward to have $\mathbf{w}_1^2 = \mathbf{S}_{\mathbf{x}w}^{-1}\boldsymbol{\mu}_{\mathbf{x}1}$, $b_{2,1} = \mathbf{w}_1^{2\mathrm{T}}\mathbf{b}_{\mathbf{x}0}$, $\mathbf{w}_2^2 = \mathbf{S}_{\mathbf{x}w}^{-1}\boldsymbol{\mu}_{\mathbf{x}2}$, $b_{2,2} = \mathbf{w}_2^{2\mathrm{T}}\mathbf{b}_{\mathbf{x}0}$, and $\mathbf{w}_3^2 = \mathbf{S}_{\mathbf{x}w}^{-1}\boldsymbol{\mu}_{\mathbf{x}3}$, $b_{2,3} = \mathbf{w}_3^{2\mathrm{T}}\mathbf{b}_{\mathbf{x}0}$, where $\mathbf{b}_{\mathbf{x}0} = (\boldsymbol{\mu}_{\mathbf{x}1} + \boldsymbol{\mu}_{\mathbf{x}2} + \boldsymbol{\mu}_{\mathbf{x}3})/3$. Figure 2b depicts this case. Notice that such an initialization is readily applicable to single-layer perceptrons.

2.1.3 Determining the size of the hidden layer

The wFC-based MLP initialization method may also suggest a suitable number of hidden neurons, a key component of MLP architecture. Neural networks, like other flexible nonlinear estimation methods, are vulnerable to problems of under-fitting and over-fitting (Haykin, 1999; Ripley, 1996). The over-fitting problem occurs more easily when the number of samples in the training set is small and the network is relatively large, which is the case for most genomic data. Therefore, it is important to use a network that is just large enough to provide an adequate fit. The resulting subspace represented by the outputs of the hidden layer should maintain as much class separability as possible (Haykin 1999): the retained partial separability is given by $J_{\mathrm{wFC}} = (\mathbf{W})$ (Eq. (1)). Hence, it is appropriate to let the number of pseudo genes (*i.e.*, m_1, the number of hidden neurons) be the number of significant eigenvalues derived from wFC-DCA because the eigenvalues represent class separability in feature space. In this study, we select the dominant eigenvalue subset that contains 99% of the total separability, and let the number of hidden neurons be equal to the number of selected eigenvalues.

2.2 Selection of MLP Inputs

Input selection is a prerequisite for diagnostic classification using genomic data; we apply our newly developed two-step wFC-based input selection method (Xuan *et al.*, 2004) that shares the same theoretical basis (wFC) with the proposed MLP initialization ap-

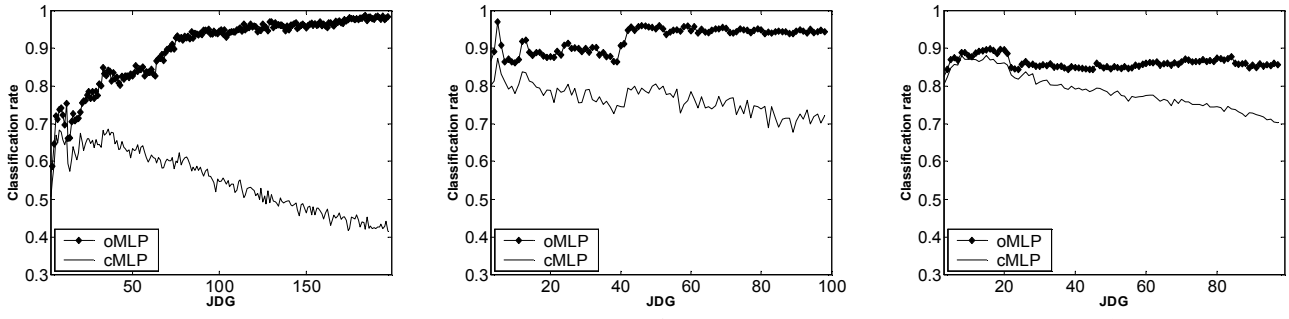


Figure 3. The classification rate curves of the oMLP and cMLP with various JDG sets as inputs, (a) LGMD, (b) leukemia, (c) CNS cancer. For all JDG sets, oMLP consistently outperformed cMLP. The classification rate for each JDG set is the average of the 100 iterations of 3-fold cross validations. The JDG set corresponding to the maximal classification rate of the oMLP is considered as the optimal JDG set.

proach. Firstly, we rank all genes based on their individual discriminatory power measured by the one-dimensional wFC (Xuan *et al.*, 2004); a gene will be selected as an Individually Discriminatory Gene (IDG) if its discriminatory power is above an empirical threshold. Secondly, from the IDG pool, we select Jointly Discriminatory Gene (JDG) subsets (with various sizes) whose joint discriminatory power is the maximum among all sets of the same size. The joint discriminatory power is also determined by the multi-dimensional version of wFC (Eq. (1)). Furthermore, the JDG sets are refined by testing on a trained MLP, which ultimately determines the "optimal" diagnostic gene subset that minimizes the generalization error. From the curve of classification rate vs. JDG subsets, we pick the optimal JDG subset that corresponds to the maximal classification rate as the final inputs for the MLP. This step boosts the MLP performance, and also determines its number of inputs (m_0, Figure 1).

3 EXPERIMENTAL VERIFICATION

3.1 Data

To highlight the biological and clinical relevance, we chose diagnostic tasks that are difficult for standard clinical and pathological methods alone. The following list summarizes the microarray data sets tested in this study.

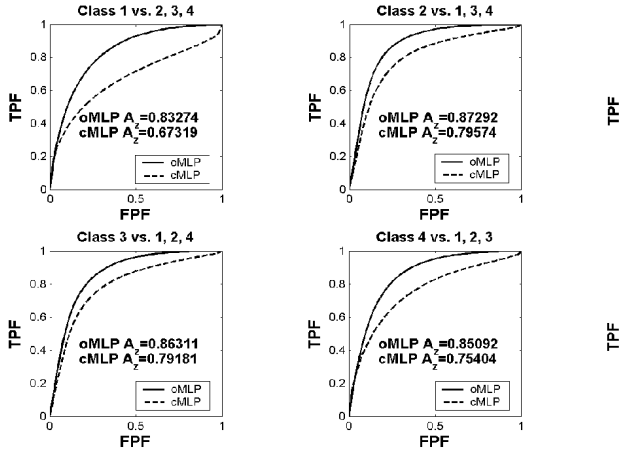
(1) Limb-girdle muscular dystrophy (LGMD, provided by Children National Medical Center, Center for Genetic Medicine): 4 diagnostic groups, Fukutin related protein deficiency (FKRP) (homozygous missense for glycosylation enzyme, limb-girdle muscular dystrophy sub-type, n = 7), Becker muscular dystrophy (BMD, hypomorphic for dystrophin, n = 5), Dysferlin deficiency (putative vesicle traffic defect, n = 9), and Calpain III deficiency (n = 11), total 32 samples, 22,283 genes.

(2) Leukemia (Kohlmann *et al.* 2004): 3 diagnostic groups, T-ALL (n = 9), MLL (n = 10), and BCR-ABL (n = 15), total 34 samples, 312 genes.

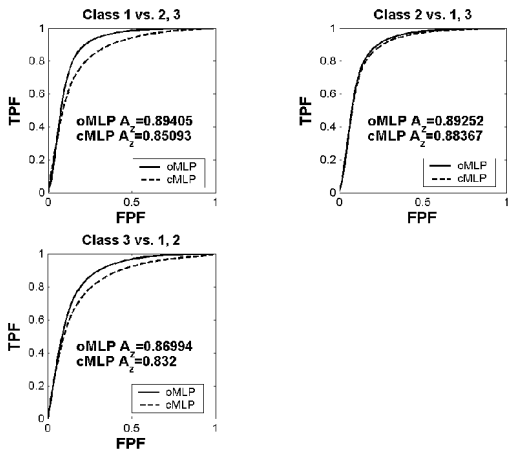
(3) Central nervous system (CNS) cancer (Pomeroy *et al.* 2002): 5 diagnostic groups, Medulloblastomas (n = 60), Malignant glioma (n = 10), Rhabdoid tumours (n = 10), Normal cerebella (n = 4), Supratentorial PNET (n = 6), total 90 samples, 7129 genes.

3.2 Results

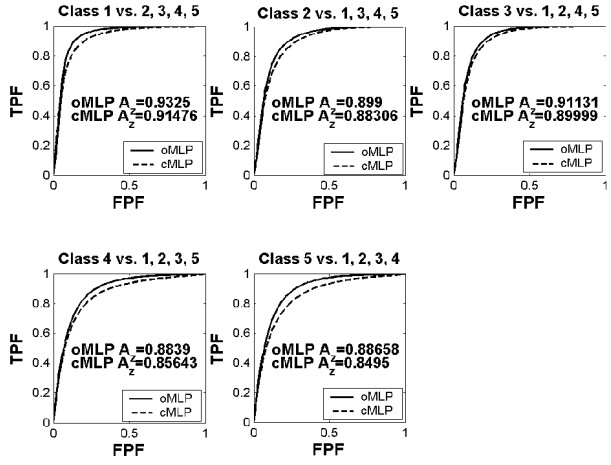
The experiments were designed to show the impact of the proposed MLP optimization method on two major aspects of MLP performance: prediction accuracy and training efficiency. For the prediction accuracy, we examined classification rate and A_z from Receiver Operating Characteristic (ROC) analysis; to probe the training property we recorded initial error (mean squared error, MSE)



(a)



(b)



(c)

Figure 4. The one-against-rest ROC curves of the oMLP and cMLP with the optimal JDG set as inputs, (a) LGMD, (b) leukemia, (c) CNS cancer. Each data set has several sets of ROC curves for the oMLP and cMLP, and the number of sets is equal to the number of classes in each data set. The corresponding A_z values for the oMLP and cMLP are displayed with the curves. The oMLP consistently showed superior performance over the cMLP for all data sets with larger A_z values.

between target and output before training, final error (MSE) after training, total number of epochs needed for convergence, and percentage of converged training.

In all experiments with MLP training and testing, we applied 100 iterations of stratified 3-fold cross validations in order to ensure reliability, and all performance measures were calculated based on the results from the cross validations. In the stratified 3-fold cross validation, the data set is randomly divided into three subsets of equal size, and the proportion of each class in each subset remains the same as that in the entire set. In each fold, one of the subsets is used for testing and the rest are combined for training; in each iteration, the training is repeated until all subsets have been used for testing.

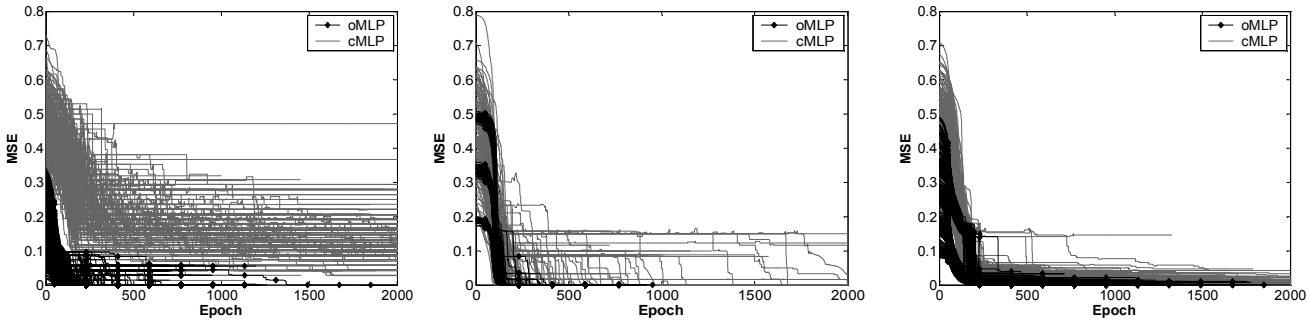
The optimized MLPs (oMLP, wFC-based initialization) consistently outperformed conventional MLPs (cMLP, conventional random initialization) for all different tested JDG subsets (Figure 3). We selected 200 JDG subsets consisting of 1 to 200 genes as inputs of the MLPs. Figure 3 plotted the curves of the classification rate from the test set (those samples not used for MLP training) vs. JDG subsets, which is part of the step 2 in the two-step input selection procedure. To determine the optimal JDG subset among the 200 candidate subsets, the oMLP and cMLP were trained with the same training set and tested with the same test set in each fold for fairness and reliability. The search of the optimal JDG subset was considered sufficient when the classification rate of the oMLP did not increase substantially and the classification rate of the cMLP decreased consistently over 20 JDG sets. The oMLP was able to maintain high classification rate as the size of the JDG increased, whereas the cMLP performance degraded. Moreover, the smaller standard deviation (STD) of the oMLP classification rate across all cross validations indicated that the oMLP provided more stable performance (Table 1).

Additionally, as ROC analysis (Metz, 1986) has been widely recognized as the most meaningful assessment of medical diagnostic performance (Metz, 1986), we also evaluated relative prediction performance of the oMLP and cMLP using a one-against-rest ROC analysis (Hand and Till, 2001) that was specifically designed for the multi-class classification. The ROC analysis offers a description of the tradeoffs between true positive fraction (TPF) and false positive fraction (FPF) of a detection test as the decision threshold varies. In the one-against-rest ROC analysis, the approximated posterior probabilities (the outputs of an MLP) of test samples were recorded, and a two-class ROC analysis was applied to all combinations of one class against the rest classes. For example, there will be n ROC curves for an n-class classification task. A ROC curve plots TPF vs. FPF; generally the larger the index, A_z (area under the curve), the better the prediction performance of the classifier. With the optimal JDG subset as inputs, the oMLPs had greater A_z values for all one-against-rest combinations than the cMLPs, therefore showed better overall performance (Figure 4). Within each individual case, the larger difference between the prediction accuracies of the oMLP and cMLP corresponds to the larger differences in A_z values (Figure 4 and Table 1).

The evaluations of training properties on the oMLP and cMLP with the optimal JDG subset as inputs clearly demonstrated the effectiveness of the proposed initialization approach (Table 1 and Figure 5). The smaller averages of initial and final MSE and the smaller STD of the final MSE in the oMLP trainings, also shown by the training curves (Figure 5), provided clear evidence that the proper initialization offered a better starting training point so that the trainings were led to a better and less diverse convergence point. In addition, we monitored whether each training process converged by recording the percentage of converged trainings. Note that a training process is considered as converged only if it meets the error goal or is stopped by a standard early stop procedure we applied in all MLP trainings to prevent over-training. The result showed that 100% of the oMLP trainings converged, but a number of cMLP trainings were eventually terminated by a preset maximal number of epochs (Table 1). Moreover, the smaller average and STD of the number of total epochs needed by the oMLP to achieve convergence further confirmed that the oMLP needed less computational resources to reach higher classification rate (Table 1).

The two-step input selection procedure is effective and computationally feasible in handling a large number of genes so that the curse of dimensionality problem is significantly reduced to a more manageable scale. The considerable change of the classification rate over the entire curve (Figure 3) confirmed that the content and size of the inputs strongly influenced MLP performance. Particularly, since it shares the same theoretical criterion with the proposed MLP initialization method (wFC), their joint influence is augmented.

We further compared the oMLP to two of the most commonly applied classifiers, K-Nearest Neighbor (KNN) and One-vs-Rest Support Vector Machine (OVR-SVM) that is a typical type of multi-class SVMs (Ramaswamy *et al.*, 2001; Statnikov *et al.*,



(a) (b) (c)

Figure 5. The training curves of the oMLP and cMLP with the optimal JDG set as inputs, (a) LGMD, (b) leukemia, (c) CNS cancer. The training properties, *e.g.*, the convergence speed, initial and the final errors, can be clearly observed from the figures. The trainings of the oMLP usually started from smaller initial errors and converged to smaller final errors, whereas the cMLP training started from and converged to larger and more diverse errors. All these improved properties supported the advantage of the wFC-based initialization over the conventional random initialization.

Table 1. The performance evaluation of the oMLP and cMLP with the optimal JDG set as inputs. MLP structure indicates the numbers of inputs (optimal JDG set), hidden neurons, and output neurons. The transfer functions in hidden neurons and output neurons are linear and log-sigmoid respectively.

Data	MLP structure (Input-hidden-output)	Classifier	Prediction accuracy		Initial error, MSE		Final error, MSE		Total epochs		Converged training, %
			Ave., %	STD, %	Ave.	STD	Ave.	STD	Ave.	STD	
LGMD	186-3-4	oMLP	98.69	4.39	0.1726	0.0796	0.0062	0.0146	761.0	131.4	100
		cMLP	42.05	18.96	0.4803	0.0718	0.1521	0.0747	1633.2	514.6	50
Leukemia	7-2-3	oMLP	96.96	5.27	0.3313	0.1086	$2.3\text{x}10^{-16}$	$2.1\text{x}10^{-17}$	726.9	41.5	100
		cMLP	87.37	15.77	0.4416	0.1085	0.0279	0.0465	981.0	368.9	93.3
CNS cancer	19-4-5	oMLP	89.82	4.46	0.2658	0.1076	0.0044	0.0057	873.4	227.3	100
		cMLP	86.86	7.19	0.4527	0.0896	0.0097	0.0125	1368.3	418.8	84.7

2005). Both KNN and OVR-SVM undertook rigorous optimizations for seeking optimal performance. We determined the parameter K in the KNN model based on 100 iterations of 3-fold cross validations. Each SVM unit in the OVR-SVM was tested for seven different kernel functions (linear, second and third order polymonials, and Gaussians with scale factors, 0.01, 0.1, 0.5, and 1.0), and five penalty values C = 0.001, 0.01, 0.1, 1.0, and 10.0. The KNN took the optimal JDG set as inputs; the OVR-SVM took two types of inputs, optimal JDG set and all genes. In summary, the OVR-SVM with optimal JDG set as inputs and the oMLP provided excellent and comparable accuracies and the OVR-SVM demonstrated small improvements, whereas the KNN and cMLP generally showed inferior performances. Detailed results of these comparative experiments can be found in Appendix A (all appendices are under http://www.cbil.ece.vt.edu/software).

Table 2. The classification rate of the model with the best performance for the KNNs and OVR-SVMs. The results are listed as average (STD).

Data	KNN	OVR-SVM	
		Optimal JDG set	All genes
LGMD	41.33 (12.66) K = 15	100.00 (0.00) Linear, Gaussian 10.0	50.94 (12.58) Gaussian 10.0
Leukemia	88.39 (8.73) K = 6	98.37 (3.76) Linear	95.34 (5.97) Linear
CNS cancer	86.59 (4.65) K = 4	95.59 (3.25) Gaussian 10.0	89.13 (3.49) Linear

4 CONCLUSIONS AND DISCUSSIONS

By suggesting an initialization technique based on the wFC and the link between the MLP mechanism and Fisher LDA, together with the input selection procedure, we offer an efficient and practical MLP prototype that can ease the curse of dimensionality in multi-class high dimensional genomic data classification and provide excellent generalization performance. The wFC-based initialization procedure initiates the MLP close to the optimal condition for decision-making, which increases the likelihood that the MLP may converge to a better local or global optimum. The curse of dimensionality is a significant problem because it can easily lead to poor predictions to test samples; classification using genomic data is more prone to this problem due to the small ratio of sample size to dimensionality. The reduction of curse of dimensionality in the oMLP is clearly shown by our experimental results: the oMLP was able to retain its classification rate to a very high level even when the number of the inputs significantly increased, while the cMLP performance degraded drastically. Besides, in the design of the wFC-based initialization, we discussed the close connection between the classification by MLP and by LDA, and made contributions in the theoretical insight and experimental validation on how the MLP actually works.

The improved performance of our optimized MLP approach does not imply that this method will be effective for any multi-class nonlinearly separable problem. Such a classification problem could be an intrinsically nonlinear problem, or may become a nonlinear problem after dimensionality reduction according to Cover's theorem on the separability of patterns. Therefore, the hidden layer of the MLP needs to perform the additional function of transforming a nonlinearly separable problem into a linear classification. This may be achieved by the existing hidden layer through dual-purpose training, or one additional hidden layer may be required. An elegant yet simple method is to apply divide-and-conquer principle to the data set and accordingly introduce some pseudo-classes to the output layer, such that all class-pairs become linearly separable. Notice that the discrete decision fusion can be readily and effortlessly done without using any combiner, since the pseudo-classes belong to some of the known classes a priori. It is important to note that a net reduction in MLP complexity can still be achieved when m_0 is large, since the total number of weights in a two-layer MLP is $m_1(m_0 + m_2)$ such that the reduction due to m_1 surpasses the generally limited increase due to m_2. Refinements, allowing a co-determination of m_1 and m_2, may further reduce the curse of dimensionality and improve the generalization performance.

A complex multi-class classification task is beyond the capability of a single classifier. It is remarkable that the single classifier, oMLP, can compete with the OVR-SVM built with a collection of single binary SVMs and show comparable outstanding performance when the number of classes is relatively small (<= 5, more experimental results in Appendix A). However, the OVR-SVM is generally expected to outperform most existing classifiers as the number of classes increases (Statnikov *et al.*, 2005).

As another verification of the effectiveness of the MLP initialization, we tested and compared the *untrained* oMLP and cMLP, and the results showed that the *untrained* oMLPs considerably outperformed the *untrained* cMLPs (Appendix B). Even without training, the hidden layer of an *untrained* oMLP is able to extract discriminant features derived from the wFC; then the neurons in the output layer can perform linear one-vs-rest classifications

based on these extracted features. We used linear transfer function in the hidden neurons and log-sigmoid transfer function in the output neurons. Hence, an *untrained* oMLP closely resembles LDA, and the initial condition of the oMLP (*i.e.*, performance of the *untrained* oMLP) reflects the performance of LDA.

Using simulated data, we demonstrated that the proposed MLP optimization method is not sensitive to the deviation of the distribution of a diagnostic group from a standard single multivariate Gaussian to a mixture of Gaussian (Appendix C). Although the wFC may only find less precise discriminant components when the distribution of each class cannot be closely modeled by a single Gaussian distribution, such loss of information is expected to be small and can be well compensated by further training of weights and biases that offers extensive degrees of freedom in modeling decision boundary.

ACKNOWLEDGEMENTS

This study was supported in part by the National Institutes of Health grants under CA109872, CA096483 and EB000830, and DOD/CDMRP grant under BC030280. Zuyi Wang was also supported by the Crystal Ball of Virginia Beach VA, and the Muscular Dystrophy Association.

REFERENCES

Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhinl, Z., Ben-Dor, A., Sampas, N., Dougherty, E., Wang, E., Marincola, F., Gooden, C., Lueders, J., Glatfelter, A., Pollock, P., Carpten, J., Gillanders, E., Leja, D., Dietrich, K., Beaudry, C., Berens, M., Alberts, D., Sondak, V., Hayward, N., and Trent, J. (2000) Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, **406**(3), 536-540.

Brown, M.P., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares, T. Jr., and Haussler, D. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci.*, **97**, 262-267.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., and Lander, E.S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531-537.

Hand, D.J. and Till, R.J. (2001) A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Machine Learning*, **45**, 171-186.

Haykin, S. (1999) Neural networks: a comprehensive foundation. 2nd Edition, Prentice-Hall, Inc.

Khan, J., Wei, J.S., Ringner, M., Saal, L.H., Lananyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C.R., Peterson, C., and Meltzer, P.S. (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, **7**(6), 673-679.

Jain, A.K., Duin, R.P.W., and Mao, J. (2000) Statistical pattern recognition: a review. *IEEE Trans. Pattern Anal. Machine Intell.*, **22**, 4-37.

Kohlmann, K., Schoch, C., Schnittger, S., Dugas, M., Hiddemann, W., Kern, W. and Haferlach, T. (2004) Pediatric acute lymphoblastic leukemia (ALL) gene expression signature classify an independent cohort of adult ALL patients. *Leukemia*, **18**, 63-71.

Linder, R., Dew, D., Sudhoff, H., Theegarten, D., Remberger, K., Pöppl, S.J, and Wagner, M. (2004) The 'subsequent artificial neural network' (SANN) approach might bring more classificatory power to ANN-based DNA microarray analyses. *Bioinformatics*, **20**(18), 3544-3552.

Loog, M., Duin, R.P.W., and. Haeb-Umbach, R (2001) Multiclass linear dimension reduction by weighted pairwise Fisher criteria. *IEEE Trans. Pattern Anal. Machine Intell.*, **23**(7), 762-766.

Metz, C. (1986) Statistical analysis of ROC data in evaluating diagnostic performance. *Multiple Regression Analysis*, 365-384.

Mjolsness, E. and DeCoste, D. (2001) Machine learning for science: state of the art and future prospects. *Science*, **293**, 2051-2055.

O'Neill, M.C., and Song, L. (2003) Neural network analysis of lymphoma microarray data: prognosis and diagnosis near-perfect. *BMC Bioinformatics*, 4:13.

Pomeroy, S., Tamayo, P., Gaasenbeek, M., Sturla, L.M., Angelo, M., McLaughlinM.E., Kim, J.Y., Goumnerovak, L.C., Blackk, P.M., Lau, C, Allen, J.C., Zagzagl, D., Olson, J.M., Curran, T., Wetmore, C., Biegel, J.A., Poggio, T., Mukherjee, S., Rifkin, R., Califanokk, A., Stolovitzkykk, G., Louis, D.N., Mesirov, J.P., Lander, E.S., and Golub, T.R. (2002) Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, **415**(24), 436-442.

Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J.P., Poggio, T., Gerald, W., Loda, M., Lander, E.S., and Golub, T.R. (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl. Acad. Sci.*, **98**(26), 15149-15154.

Raudys, S. (1992) Accuracy of feature selection and extraction in statistical and neural net pattern classification. *Proc. of 11th IAPR International Conf. on Pattern Recognition*, **2**, 62 – 70.

Raudys, S. (1994) Why do multilayer perceptrons have favorable small sample properties? *Pattern Recognition in Practice IV, Elsevier Science B. V.*, 287-298.

Raudys, S. (1997) On dimensionality, sample size, and classification error of nonparametric linear classification algorithms. *IEEE Trans on Pattern Analysis and Machine Intelligence*, **19**(6), 667 – 671.

Raudys, S., Jain, A.K. (1991) Small sample size effects in statistical pattern recognition: recommendations for practitioners. *IEEE Trans on Pattern Analysis and Machine Intelligence*, **13**(3), 252 – 264.

Raudys, S., Skurikhina, M. (1992) The role of the number of training samples on weight initialisation of artificial neural net classifier. *RNNS/IEEE Symposium on Neuroinformatics and Neurocomputers*, **1**, 343 – 353.

Ripley, B. (1996) Pattern Recognition and Neural Networks. Cambridge University Press.

Shipp, M.A., Ross, K.N., Tamayo, P., Weng, A.P., Kutok, J.L., Aguiar, R.C., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G.S., Ray, T.S., Koval, M.A., Last, K.W., Norton, A., Lister, T.A., Mesirov, J., Neuberg, D.S., Lander, E.S., Aster, J.C., and Golub, T.R. (2002) Diffuse large B-cell lymphoma outcome prediction by gene expression profiling and supervised machine learning. *Nature Medicine*, **8**(1), 68-74.

Statnikov, A, Aliferis, C.F., Tsamardinos, I., Hardin, D., and Levy, S. (2005) A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, **21**(5), 631-643.

van't Veer, L.J., Dai, H., vande Vijver, M.J., He, Y.D., Hart, A.M., Mao, M., Peterse, H.L., van derKooy, K., Marton, M.J., and Witteveen, A.T. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530-536.

Wang, Y., Wang, Z., Xuan, J., Zhang, J., Hoffman, E., Clarke, R., and Khan, J. (2004) Optimizing multilayer perceptrons by discriminatory component analysis. *Proc. IEEE Workshop on Machine Learning for Signal Processing*, 273-282, San Luis, Brazil.

Wei, J.S., Greer, B.T., Westermann, F., Steinberg, S.M., Son, C.G., Chen, Q.R., Whiteford, C.C., Bilke, S., Krasnoselsky, A.L., Cenacchi, N., Catchpoole, D., Berthold, F., Schwab, M., and Khan, J. (2005) Prediction of clinical outcome using gene expression profiling and artificial neural networks for patients with neuroblastoma. *Cancer Res.*, **65**(1), 6883-6891.

West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson, J.A., Marks, J.R., and Nevins, J.R. (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl. Acad. Sci.*, **98**, 11462-11467.

Xuan, J., Dong, Y., Khan, J., Hoffman, E.P., Clarke, R., and Wang, Y. (2004) Robust feature selection by weighted fisher criterion for multiclass prediction in gene expression profiling. *Proceedings of the 17th International Conference on Pattern Recognition*, **2**, 291-94.

Estrogen receptor alpha positive breast tumors and breast cancer cell lines share similarities in their transcriptome data structures

YUELIN ZHU[1], ANTAI WANG[1,2], MINETTA C. LIU[1], ALAN ZWART[1], RICHARD Y. LEE[1], ANN GALLAGHER[1], YUE WANG[4], WILLIAM R. MILLER[5], J. MICHAEL DIXON[5] and ROBERT CLARKE[1,3]

[1]Lombardi Comprehensive Cancer Center and Departments of Oncology, [2]Biostatistics, Bioinformatics and Biomathematics, and [3]Physiology and Biophysics, Georgetown University School of Medicine, Washington, DC; [4]Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Arlington, VA 22203, USA; [5]Department of Oncology, University of Edinburgh, Western General Hospital, Crewe Rd South, Edinburgh, Scotland, UK

Abstract. Established human breast cancer cell lines are widely used as experimental models in breast cancer research. While these cell lines and their variants share many phenotypic characteristics with human breast tumors, the extent to which they reflect the underlying molecular biology of breast cancer remains controversial. We explored this issue using a probabilistic rather than heuristic approach. Data from gene expression microarrays were used to compare the global structures of the transcriptomes of three estrogen receptor alpha positive (ER^+) human breast cancer cell lines (MCF-7, T47D, ZR-75-1) and 13 human breast tumors (11 ER^+; 2 ER^-). Linear representations of the respective data structures were obtained by deriving those top principal components (PCs) required to capture ≥80% of the cumulative variance for each data set (M PCs). We then identified those genes most highly correlated with the M PCs (Pearson's correlation coefficient $r \geq 0.800$) and identified a group of 36 genes commonly correlated with both the cell line (M = 5 PCs) and tumor (M = 6 PCs) data structures. All 36 common genes were correlated with PC1 from the breast tumor data: 21/36 genes were correlated with PC1, 14/36 genes correlated with PC2, and 1/36 genes correlated with PC3 from the cell line data. Genes important in defining the data structures include NFκB p65, IGFBP-6, ornithine decarboxylase-1, and paxillin. When data from MDA-MB-435 xenografts (ER^-) were included in the analysis, we were unable to find any common genes between these xenografts and the breast tumors. These data clearly imply that MCF-7, T47D, and ZR-75-1 cells and ER^+ breast tumors share substantial global similarities in the structures of their respective transcriptomes, and that these cell lines are good models in which to identify molecular events that are likely to be important in some ER^+ human breast cancers.

Correspondence to: Dr Robert Clarke, Department of Oncology, Room W405A Research Building, Georgetown University School of Medicine, 3970 Reservoir Rd, NW, Washington, DC 20057, USA
E-mail: clarker@georgetown.edu

Abbreviations: M, essential dimensionality; ER, estrogen receptor alpha; PC, principal component; PCA, principal component analysis; tgDNA, total genomic DNA

Key words: breast cancer, bioinformatics, cell lines, data structure, genomics, microarray analysis, principal component analysis

Introduction

Human breast cancer cell lines, whether growing *in vitro* or *in vivo* as xenografts in immunedeficient rodents, are among the most widely used experimental models in breast cancer research (1-4). These cell lines and their variants have been particularly useful as experimental models and enable investigators to address hypotheses in ways that would be technically difficult or ethically inappropriate in humans. We and others have extensively reviewed the characteristics of selected human breast cancer cell lines, their phenotypes, and the extent to which these phenotypes reflect key components of the human disease (2-4).

Almost 100 breast cancer cell lines have been described but Lacroix and Leclercq estimated that over two-thirds of studies involved work with only one or more of three models (4): the estrogen receptor alpha positive (ER^+), estrogen-dependent and antiestrogen sensitive MCF-7 and T47D cell lines, and the ER^-, estrogen-independent and antiestrogen resistant MDA-MB-231 cell line (3). All three cell lines are tumorigenic and locally invasive in immunedeficient rodents (2) but only orthotopic xenografts of MDA-MB-231 cells produce spontaneous metastases (5). ZR-75-1 is another commonly used ER^+ breast cancer cell line and is phenotypically similar to MCF-7 and T47D cells (3). MDA-MB-435 cells are widely used as a

metastatic ER- model (5); while these cells are phenotypically similar to MDA-MB-231 cells, the breast origin of the MDA-MB-435 cell line has been questioned (6,7).

It is evident that human breast cancer cell lines reflect important phenotypic characteristics present in the human disease, and they have been central to discovering and extending new knowledge in many areas of breast cancer research (4). However, the extent to which biological insights can be extrapolated from preclinical models to the human disease remains somewhat controversial. Established breast cancer cell lines exhibit substantial aneuploidy and genetic instability, and variants can arise spontaneously over time (8). While this is probably a reflection of the inherent molecular/genetic instability of breast tumors, it is unclear how well human breast cancer cell lines growing *in vitro* reflect the underlying molecular biology of breast tumors. For example, a study comparing the molecular profiles of 60 human cell lines showed that, by unsupervised hierarchical clustering, breast cancer cell lines do not cluster together but are scattered across the entire dendrogram (6). These investigators also reported a hierarchical clustering analysis of data restricted to five breast cancer cell lines, four leukemia cell lines, two breast tumors, one breast tumor metastasis, and one specimen of normal breast tissue. The breast cancer cell lines clustered together but this cluster was more similar to the leukemia cell lines than to the breast tumors. Moreover, the normal breast specimen and the breast tumors formed a single cluster distinct from all the cell lines (6). A subsequent review of these and other molecular profiles concluded that breast cancer cell lines and tumors shared some gene expression patterns in common. However, the authors took a largely intuitive rather than probabilistic approach, looking for commonalities in gene expression patterns in cell lines with predetermined cellular phenotypes/functions. The authors acknowledged that alternative interpretations of the data were possible (9).

In this study, our primary goal was to obtain a relatively unbiased probabilistic assessment of the global similarities in the transcriptomes of human breast cancer cell lines and breast tumors. Rather than compare broadly defined phenotypic or genetic characteristics, we asked directly whether similarities exist within the structures of their respective high dimensional gene expression microarray data. To address this goal, we first developed an application of principal components analysis (PCA) (10) based on the general approach described by Jolliffe (11). PCA is a technique that finds linear transformations of data such that the first principal component (PC) is that linear projection that best captures the greatest variance in the data. The second PC is orthogonal to the first and captures the second greatest variance, and so on. In this manner, PCA can be used to find those projections that best capture the overall structure of the data. We show that three of the most widely used ER+ human breast cancer cells lines (MCF-7, T47D, ZR-75-1) exhibit substantial similarities in their transcriptome data structures to a panel of mostly ER+ breast cancer specimens from patients.

Materials and methods

Human breast cancer cell lines. MCF-7 cells were originally obtained from the Barbara A. Karmanos Cancer Institute (Detroit, MI), T47D and ZR-75-1 cells were obtained from the American Type Culture Collection (Manassas, VA), and MB-MDA-435 cells were from Dr Janet Price (M.D. Anderson Cancer Center, Houston, TX). All cell lines were maintained at 37˚C in cell culture medium (improved minimal essential medium with phenol red and supplemented with 5% (v/v) heat-inactivated fetal bovine serum; Biofluids, Rockville, MD) in a 95% air/5% CO_2 atmosphere. All cell lines were shown to be free of contamination with *Mycoplasma* spp.

Human breast tumor specimens. The 13 breast tumor specimens and the associated microarray data used in this study have been previously reported (12). Five of the 13 specimens were obtained from patients undergoing a diagnostic core needle or excisional biopsy at Georgetown University Hospital. All patients signed a written consent approved by the Georgetown University Medical Center Institutional Review Board. Core needle biopsies were either obtained under mammographic or ultrasound guidance during a routine diagnostic procedure, or obtained intraoperatively after surgical exposure of the tumor. The study pathologist performed a routine histopathologic analysis of frozen sections from all biopsies as previously described (12); biopsies were released for microarray analysis only if they did not contain any new clinical information important for patient care. The other eight breast tumor specimens were obtained at the Department of Oncology, University of Edinburgh (Scotland, UK); samples were collected with appropriate patient consent, and all procedures were performed using guidelines consistent with the relevant UK legislation. Once released for study, all patient identifiers were removed from each sample. Information not already published on these samples is included in Table I. The clinical material, mostly frozen in OCT, was directly provided to the research laboratory for storage and/or processing, whereupon tissue was either stored at -80°C or processed immediately for RNA extraction.

MDA-MB-435 human breast cancer xenografts. Cells from subconfluent monolayers were removed by trypsinization. To establish xenografts, $1x10^6$ viable cells, as estimated by trypan-blue dye exclusion, were subcutaneously inoculated into the region of the mammary fat pad as previously described (2,13). Mice were 4-6 week old female, NCr *nu/nu* athymic mice (~20 g body weight) and were housed 4 or 5 per cage and fed sterilized, pelleted food and sterilized water *ad libitum.* Nude mice (38) were used and tumors were observed at each of the inoculation sites. Tumors were measured twice weekly for 4 weeks *post inoculum*; consistently proliferating tumors were identified and removed immediately *post mortem* using sterile scissors and forceps. Studies were performed by the Lombardi Comprehensive Cancer Center Animal Research Shared Resource in a pathogen-free environment within a central facility approved by the American Association for Accreditation of Laboratory Animal Care. All work that required the use of vertebrate animals was performed in accordance with the current regulations and standards described by the United States Department of Agriculture and the United States Department of Health and Human Services, and with the approval of the Georgetown University Animal Care and Use Committee.

Table I. Characteristics of the breast tumor specimens.

Tumor	ER	Lymph nodes	% Cancer	Source
1	+	+	90	GU
2	+	-	80	GU
3	+	+	90	GU
4	+	+	90	GU
5	-	ND	80	GU
6	+	+	90	EU
7	+	+	90	EU
8	+	ND	99	EU
9	+	-	90	EU
10	+	+	90	EU
11	+	-	70	EU
12	+	+	90	EU
13	-	+	90	EU

ER, estrogen receptor alpha (positive, +; or negative, -); lymph nodes, presence (+), absence (-) of involved lymph nodes, or no data (ND); % cancer, proportion of each specimen that contains neoplastic breast epithelial cells; Source, center at which cases were accrued; GU, Georgetown University; EU, Edinburgh University. Additional information on selected cases has been previously published (12).

RNA preparation and gene expression microarray studies. Study materials were collected over a prolonged period and were processed slightly differently. These differences replicate some of the methodologic variability anticipated across laboratories but might be expected to introduce some noise into the data. For cell lines growing *in vitro*, each cell line sample represents data from an independent cell culture grown on a different day; no cultures were pooled, nor were RNAs extracted from cultures grown at the same time. Sub-confluent monolayers were rapidly trypsinized, cells were centrifuged at 1,000 x g for 5 min in cell culture medium and total RNA was extracted from the cell pellets using the TRIzol reagent as described by the manufacturer (Invitrogen, Carlsbad, CA). For MDA-MB-435 xenografts in athymic nude mice, tumors were removed at necropsy, immediately placed in RNALater™ (Ambion, Austin, TX) and stored at -80°C as previously described (12). Frozen xenografts from mice were placed in '1x1' plastic bags, pulverized on dry ice, transferred to 35 ml conical Oakridge tubes (Nalgene, Rochester, NY), and weighed. Frozen tissues were homogenized in TRIzol using a polytron homogenizer (Brinkmann Instruments, Inc. Westbury, NY) and total RNA isolated using the TRIzol reagent. For the human tumors, frozen tissue was placed in a '1x1' plastic bag on dry ice, pulverized, and lysis buffer added (Qiagen RNeasy kit; Qiagen Inc., Valencia, CA). Each sample was then homogenized with a 1 ml syringe and 18 gauge needle, added to the Qiagen spin column, processed as described by the manufacturer, and RNA eluted with dH_2O. None of the RNAs was amplified or pooled.

RNA concentrations were determined by comparing the optical density ratios (OD_{260}/OD_{280}) obtained spectrophotometrically using a Beckman DU640 Spectrophotometer (Beckman, Fullerton, CA). RNA quality was assessed using an Agilent 2100 Bioanalyzer and RNA 6000 LabChip kits (Agilent Technologies, New Castle, DE), which allows for visual examination of both the 18S and 28S rRNA bands as a measure of RNA integrity. We used high quality RNA as assessed by standard measures (12).

NamedGenes GeneFilters (ResGen, part of the Invitrogen Corporation, Inc., Huntsville, AL) that contain 4,132 known cDNAs and 192 controls including total genomic DNAs (tgDNA) on each filter were used. Probes were generated as previously described (14). Briefly, total RNA (500 ng) from experimental samples was reverse transcribed and simultaneously radioactively labeled by incorporation of [α-^{33}P]ATP and [α-^{33}P]CTP. This method radiolabels both the sense and antisense probe strands. Probes were purified and hybridized to a GeneFilter, and incubated for 12-18 h at 42°C in a roller oven (Robbin Scientific, Sunnyvale, CA). Each hybridized GeneFilter was washed twice in 2X SSC, 1% SDS at 50°C for 20 min and once at 55°C in 0.5X SSC, 1% SDS for 15 min. Hybridization signals were detected by phosphorimage analysis using a Molecular Dynamics Storm PhosphorImager (Molecular Dynamics, Sunnyvale, CA).

Microarray data preprocessing. Pathways™ 4.0 software algorithms (Research Genetics, Inc.) were used to acquire data from microarray images. Briefly, this software geometrically quantifies the intensities of both the spot and local background for each gene. Local background correction is estimated by subtracting local signals from areas devoid of target from the raw intensity value of each target cDNA, and a value of one is added to all non-negative values to conserve the relative intensities with low expression values. Negative values resulting from background subtraction were adjusted to one. Background-corrected data were then normalized to account for differences in probe specific activity, hybridization, and other variables among replicates. The global mean method was used to normalize the data from each array.

A signal bleeding effect from neighboring cDNA spots, where signals from adjacent spots bleed into each other, is a major confounding factor with this microarray technology. To determine if a spot on the filter was affected by signal bleeding, we used an in-house algorithm (programmed in MatLab version R13SP1; Mathworks, Natick, MA; unpublished data). This algorithm calculates the difference between the respective local background for a gene and global background from the filter, expressed as a percentage of the raw intensity value for that gene. Values above a predetermined threshold indicate that the signal from neighboring spots bled into the spot of interest. The digitized images for all spots flagged by the algorithm were subjected to visual inspection to confirm any signal bleeding. Genes with signals determined visually and/or mathematically to be confounded by a bleed effect were excluded from further analysis.

We used several criteria to identify and exclude likely non-informative genes and construct a reduced dimensional data set for analysis. The goal of these preprocessing steps was to obtain a series of robust expression values for genes determined

to be present in all three groups to be compared in the study (3 ER+ human breast cancer cell lines; 13 breast tumors; 38 MDA-MB-435 xenografts). First, we excluded genes that have expression values consistently in the undetectable range in all microarrays or that have signals compromised by signal bleed as defined above. If a gene was found to be free of bleeding effects in at least 70% of arrays, data for this gene were retained for further study. Genes in the undetectable range were eliminated if their normalized expression levels were <0.1 in all experimental groups. We did not attempt to estimate and replace missing values. Application of these criteria across all microarrays from the cell lines, breast tumors, and MDA-MB-435 xenografts resulted in a list of 428 robust gene signals for further analysis.

Data analysis: comparison of high dimensional data structures. To estimate independently the data structures, we conducted separate PCA on the robust gene expression data set (n=428 genes) for each of the three groups and determined the essential dimensionality (M) for samples within the same group. PCA was performed using the covariance matrices for standardized gene expression levels. M is defined as the number of principal components (PCs) needed to account for the variation in the original data. Jolliffe proposed several strategies to determine M (11); we applied the most commonly used rule and selected those PCs that represent the smallest value of M that captures a high cumulative percentage of the total variance (≥80%).

Once the M PCs were identified for a group, we calculated the Pearson's correlation coefficient for each gene with each PC and selected those genes with an absolute correlation coefficient r≥0.800 with at least one of the top M PCs (top genes). While this approach is broadly comparable to the method proposed by Jolliffe (11), we ranked the PCs such that PC1 captured the highest proportion of data variance, PC1 + PC2 captured the next highest proportion and we continued until PC1 + PC2...PC_i captured ≥80% of the total variance. Thus, we placed more weight on the top PCs, whereas Jolliffe's method attributed equal importance to each of the M PCs. Our approach appears reasonable, since genes tend to have larger correlation coefficients with higher ranked compared with lower ranked PCs.

Since we explored independently each group, the PCs and the genes that best define these PCs reflect only the structure of the data for that group. In this manner, we can compare the relative importance of each gene expression value across data structures. Thus, having selected the top genes from each of the three groups, we compared the respective M PC-derived gene lists among groups and created a 'common genes' list. For example, if *gene-1* was one of the top genes for both breast tumor and cell line samples, we considered *gene-1* as a common gene between these two groups.

Results

Cell line and tumor data structures share similar essential dimensionality. For this study, an unsupervised probabilistic approach applied to each experimental group should have the greatest potential to generate relatively unbiased, independent representations of data structures. Since we do not predetermine the number of PCs, only the percentage of cumulative variation, and the M PCs are independently identified within each group, the M PCs obtained should provide reasonable representations by which to compare data structures. While we might expect similar data structures to be defined by approximately similar numbers of M PCs, this is an inadequate single measure because the genes most closely correlated with each PC may be different. Conversely, it is possible that a different number of PCs may be required to satisfy M in each experimental group but the genes correlated with the respective M PCs may be very similar.

To address these issues, we compared the number of M PCs, ranked these by their relative ability to capture data variation, and then assessed the correlation of each gene with each ranked PC. Data sets that exhibit similarities may be defined by similar numbers of M PCs. More importantly, data structures with substantial similarities will have the same genes highly correlated with similarly ranked PCs; for example, *gene-1* is highly correlated with PC1 in one group and also is highly correlated with PC1, PC2, or PC3 in another experimental group. The higher proportions of genes

Table II. Principal component analysis and essential dimensionality.

PCA	Cell lines (%)	Tumors (%)	Cell lines/tumors (combined) (%)
M PCs	n=5	n=6	n=10
PC1	31.8	35.8	28.8
PC1 + PC2	51.0	48.8	42.9
PC1...PC3	65.7	58.2	50.5
PC1...PC4	74.6	60.7	57.0
PC1...PC5	**80.9**	74.5	62.6
PC1...PC6	85.7	**81.4**	67.5
PC1...PC7	89.9	87.1	71.8
PC1...PC8	93.0	91.3	75.9
PC1...PC9	95.8	94.2	79.1
PC1...PC10	98.1	96.5	**81.8**
r≥0.800 (M PC)	n=103 genes	n=106 genes	n=65 genes
Common genes	36 genes are common to the 103-genes (cell lines) and 106-genes (tumors)		31 of the 36 common genes correlate with PC1 of the combined group[a]

M PCs is the number of PCs required to capture ≥80% of the cumulative variances in the data set (essential dimensionality). Percentages are the cumulative variances captured by the sum of the M PCs as indicated. The final row shows the number of genes in each group that have a correlation coefficient r≥0.800 with at least one of the M PCs in that group. For example, there are 103 genes correlated either with PC1, PC2, PC3, PC4 or PC5 in the breast cancer cell line data set. [a]For 22 genes r≥0.800; for a further 9 genes r≥0.750.

that are highly correlated with top ranked PCs in both groups, the more similar are the data structures being compared.

In this data set there are over 420 possible orthogonal PCs that can be explored as projections of the high dimensional data. However, we would expect most of the data variation to be captured by a much smaller number of M PCs. Using our approach, we found that only six PCs are required to define the breast tumor data structure by our criterion of ≥80% cumulative variance (cumulative variance = 80.9%; Table II). Similarly, only five PCs are required to describe the breast cancer cell line data structure (cumulative variance = 81.4%; Table II).

The top principal components of cell line and tumor data share notable similarities. To compare the PCs, we then calculated the correlation coefficient of each gene with each of the M PCs and obtained two gene lists, one for each experimental group. Thirty-six genes are important in describing independently the data structures for both the tumor and cell line groups (Table III). Surprisingly, all 36 common genes were correlated with the top ranked PC (PC1) from the tumor data set. Of the genes from the ER^+ cell lines data set, 21 genes also correlated with its PC1. Thus, there are striking similarities between the top PC in both data sets; each of which capture almost one-third of the variation in their respective data sets (Table II). Of the remaining 15 genes, 14 genes are correlated with PC2; only one gene is correlated with PC3 in the ER^+ cell lines data. The sign of the correlation is less informative than the absolute value of the coefficient; since we would not expect the PCs to be identical, the direction of each gene's correlation with a PC may vary in each data set and its absolute value reflects the true significance. Nonetheless, 61% of the genes (22/36) show the same directional correlation. Twenty-one of these genes correlated with PC1, strongly suggesting substantial similarities in the top PC. Taken together, these data provide evidence of notable similarities between the human breast cancer cell line and breast tumor transcriptome data structures.

To further support these observations, we combined the cell line and tumor data sets and performed PCA on the combined group. Since the tumors are more heterogeneous than the cell lines, we would expect the combined data set to require a higher number of M PCs and that fewer of the previously identified common genes will be highly correlated with these M PCs. Consistent with the general similarities, Table II shows that only 10 PCs are required (cumulative variance = 81.8%) to define the structure of the combined data set. We then calculated the correlation coefficients for the previously identified 36 common genes with the top PCs derived from the combined group. Thirty-five genes could be evaluated since one gene was not correlated with the top PCs. Twenty-two of the common genes met the initial criterion of $r \geq 0.800$ and a further 9 genes had correlations of $r \geq 0.750$ (Table III). All 31 of these genes were correlated with PC1. The remaining 4 genes were correlated with PC2 (n=1) or PC3 (n=3) but their coefficients were much lower. Thus, most of the common genes important in separate group analysis also are important in combined group analysis.

Since PCA can be used to perform multidimensional scaling for visualization (12,15), we used the top two PCs to visualize the combined data group. Fig. 1 shows that the cell line and breast tumor samples do not form distinct separable clusters in 2-dimensional PC space. These projections are visually consistent with the PCA analysis described above.

We also performed similar independent M PC analyses using data from 38 MDA-MB-435 xenografts growing in the mammary fat pad regions of athymic nude mice (data not shown). Capturing the essential dimensionality of the data structure required 16 PCs and no genes met the criteria for commonality between these xenografts and the breast tumors. Only four genes were found to be in common with the three other breast cancer cell lines: S100A11 (S100 calcium binding protein A11), PTPN7 (protein tyrosine phosphatase, non-receptor type 7), MR1 (major histocompatibility complex, class I-related), and DCI (dodecenoyl-Coenzyme A delta isomerase; 3,2 trans-enoyl-Coenzyme A isomerase). The notable lack of similarity with the breast cancer cell lines and tumors is consistent with the putative non-breast cancer origin (16), although the ER^- status of this cell line and the predominantly ER^+ status of the breast tumors and breast cancer cell lines also may contribute to the lack of similarity in MDA-MB-435 xenograft data structure with the breast tumors and data sets of other cell lines.

While our approach was not designed to select genes for their functional relevance or differential association among specific breast cancer outcomes/phenotypes, we might expect some of these genes to represent functions implicated in other breast cancer studies. We used the six main gene ontology functional categories as defined in the GO database (http://www.geneontology.org) and applied by Pawitan *et al* (17), who compared two gene lists implicated in predicting breast cancer prognosis. This appears to be a reasonable comparison as our data set included both lymph node positive and negative cases (Table I); lymph node involvement is one of the strongest independent predictors of a poor prognosis (18,19). Since there are only three common genes between the 64-gene (Pawitan) and 70-gene (van't Veer) gene lists, despite the similarities between these two studies, it was not surprising that we did not find any of those genes in common with our 36 genes. However, we found 11/36 genes in 5 of the 6 functional categories (Table IV). Thus, 31% of the genes are represented in the 6 functional categories, compared with 37% of the van't Veer *et al* genes (20) and 45% of the Pawitan *et al* genes (21).

Discussion

Limitations in the ability of individual experimental models to reflect fully the complexity of their corresponding human cancer are widely acknowledged. For example, cellular signaling in rodent cells may not be similar to that in human cells. Human cells require notably more changes in genetic, epigenetic, or gene expression events for malignant transformation (22-24); the same may be true for post-transformation events that drive malignant progression. While established human breast cancer cell lines exhibit many phenotypic characteristics of the human disease, the ability to use these models to discover meaningful molecular insights into breast cancer biology also remains controversial (8). Thus, the primary goal of this study was to compare the transcriptome structures, as derived from gene expression microarray data,

Table III. Common genes correlated with the top M PCs in the breast tumors and cell lines.

Gene	Gene name	Cells	r	Tumors	r	Comb	R
METAP2	Methionyl aminopeptidase 2	PC2	0.964	PC1	-0.857	-0.394	PC3
A2M	Alpha-2-macroglobulin	PC1	-0.885	PC1	-0.835	-0.804	PC1
IGFBP6	Insulin-like growth factor binding protein 6	PC2	0.956	PC1	-0.900	0.844	PC1
KRT13	Keratin 13	PC2	0.869	PC1	-0.844	-0.822	PC1
DRAP1	DR1-associated protein 1 (negative cofactor 2 alpha)	PC2	0.979	PC1	-0.908	0.754	PC1
GPC1	Glypican 1	PC1	-0.805	PC1	-0.957	-0.901	PC1
PCOLN3	Procollagen (type III) N-endopeptidase	PC1	-0.825	PC1	-0.805	-0.786	PC1
ATP5J	ATP synthase, H^+ transporting, mitochondrial F0 complex, subunit F6	PC2	0.960	PC1	-0.874	0.778	PC1
MRPL49	Mitochondrial ribosomal protein L49	PC1	-0.875	PC1	-0.927	-0.903	PC1
RELA	NFκB (p65)	PC2	0.969	PC1	-0.861	-0.758	PC1
PTHR1	Parathyroid hormone receptor 1	PC1	-0.886	PC1	-0.857	-0.793	PC1
FST	Follistatin	PC2	0.930	PC1	-0.881	-0.349	PC3
POLA	Polymerase (DNA directed), alpha	PC1	-0.854	PC1	-0.858	-0.826	PC1
CREBL1	cAMP responsive element binding protein-like 1	PC1	-0.937	PC1	-0.862	-0.873	PC1
GOLGA2	Golgi autoantigen, golgin subfamily a, 2	PC2	0.809	PC1	-0.816	0.562	PC3
SF3A1	Splicing factor 3a, subunit 1, 120 kDa	PC2	0.962	PC1	-0.841	0.494	PC2
USP4	Ubiquitin specific protease 4 (proto-oncogene)	PC1	-0.889	PC1	-0.912	-0.898	PC1
CR2	Complement component (3d/Epstein-Barr virus) receptor 2	PC1	-0.835	PC1	-0.818	-	-
NR1D1	Nuclear receptor subfamily 1, group D, member 1	PC1	-0.885	PC1	-0.874	-0.884	PC1
ODC1	Ornithine decarboxylase 1	PC1	-0.870	PC1	-0.945	-0.901	PC1
ORM2	Orosomucoid 2	PC2	0.962	PC1	-0.972	0.876	PC1
AMFR	Autocrine motility factor receptor	PC1	-0.824	PC1	-0.883	-0.887	PC1
RYR1	Ryanodine receptor 1	PC2	0.981	PC1	-0.927	-0.772	PC1
PPM1F	Protein phosphatase 1F	PC1	-0.820	PC1	-0.845	-0.843	PC1
KCNN4	Potassium intermediate/small conductance calcium-activated channel, subfamily N, member 4	PC1	-0.802	PC1	-0.929	-0.886	PC1
NT5E	5' nucleotidase (CD73)	PC2	0.912	PC1	-0.819	-0.786	PC1
ITGB2	Integrin beta 2	PC1	-0.908	PC1	-0.885	-0.876	PC1
ABCC1	ATP-binding cassette, subfamily C (CFTR/MRP), member 1	PC2	0.953	PC1	-0.923	-0.827	PC1
PXN	Paxillin	PC1	-0.889	PC1	-0.951	-0.929	PC1
STAM	Signal transducing adaptor molecule (SH3 domain and ITAM motif) 1	PC2	0.899	PC1	-0.882	-0.817	PC1
COX6B1	Cytochrome c oxidase subunit VIb	PC3	-0.813	PC1	-0.980	-0.895	PC1
ACTR1A	Actin-related protein 1 homolog A	PC1	-0.837	PC1	-0.906	-0.882	PC1
LOC56311	Ankyrin repeat domain 7	PC1	-0.911	PC1	-0.854	-0.796	PC1
KIAA1641	Chronic lymphocytic leukemia-associated antigen KW-1	PC1	-0.845	PC1	-0.907	-0.797	PC1
CSTA	Cystatin A (stefin A)	PC1	-0.892	PC1	-0.928	-0.910	PC1
B7	B7 protein	PC1	-0.845	PC1	-0.850	-0.804	PC1

For comparison of the cell lines and tumors, each gene selected must exhibit a correlation coefficient of $r \geq 0.800$ with one of the top M PCs. For example, a gene in common between breast tumors and breast cancer cell lines must be correlated ($r \geq 0.800$) with PC1, PC2, PC3, PC4 or PC5; there are only 5 M PCs in the breast cancer cell line group (see Table II); there are 36 genes in common by these criteria. The gene CR2 was not associated with the top M PCs in the combined group. Gene, gene symbol as designated by the human gene ontology (HUGO) gene nomenclature committee. Comb, data from the combined cell line (MCF-7, T47D, ZR-75-1) and tumor data set. The four genes in the combined data set where $r < 0.75$ are indicated.

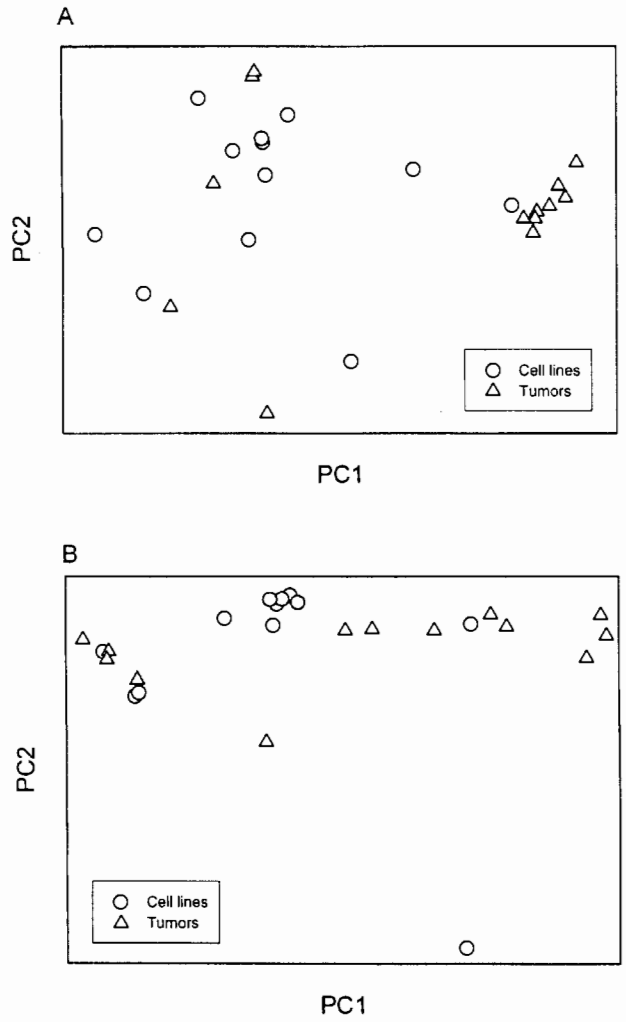


Figure 1. Multidimensional scaling of cell line and tumor data. (A), 428 dimensional data set; (B), 36 dimensional data set. Δ, breast tumors; O, cell lines.

Table IV. Gene functions among the 36 common genes.

Biological function	Genes
DNA replication/transcription	CREBL1, NR1D1, POLA, RELA, RYR1
Apoptosis	ITGB2, PPM1F, RELA, RYR1
Cell cycle/proliferation/growth	IGFBP6, RYR1
Cell adhesion/motility	AMFR, ITGB2, PXN, RYR1
Signal transduction	AMFR, CREBL1, IGFBP6, ITGB2, PTHR1, PXN, RELA, RYR1, STAM

The GO database was used to annotate the gene functions (http://www.geneontology.org). The GO categories are based on the six used by Pawitan *et al* (17) to compare their breast cancer predictive gene list with that of van't Veer *et al* (20). We found no genes in the 'angiogenesis' category; Pawitan *et al* reported only one gene from their 64-gene data set and a different gene from the 70-gene van't Veer *et al* data set in this category (17).

of predominately ER+ breast tumor specimens from patients and the three most widely used ER+ human breast cancer cell lines (MCF-7, T47D, ZR-75-1).

Breast tumor specimens can include multiple different cell types such as epithelial, myoepithelial, fibroblastic, myofibroblastic, and reticuloendothelial (25), whereas cell lines are, in comparison, biologically more homogeneous. Thus, the goal of comparing cell lines and tumor specimens, using direct comparisons of gene expression levels, is potentially confounded by tissue heterogeneity. In breast tumors, a gene's signal will reflect the sum of values from all cell types included in the specimen. Earlier microarray studies did not account for this heterogeneity and this may partly explain the greater similarity reported between normal breast and breast cancer specimens than between the breast cancer specimens and human breast cancer cell lines (6). Furthermore, earlier studies used unsupervised hierarchical clustering methods to solve the high dimensional data structures and identify putative relationships among samples. Since these hierarchies can be built using different distance measures and the data points linked by different measures (26-28), different clustering methods may provide different solutions to the same data sets (29,30). With no goodness-of-fit for the data solutions (29) or comparisons with other methods that may provide more accurate or more complete solutions, the inability of breast cancer cell lines to cluster together or to cluster with breast cancers may reflect the limitations inherent in the analytical approaches applied. The lack of consideration of specimen heterogeneity also may have confounded the analysis.

Rather than apply heuristic rules to deduce similarities or differences based on broad phenotypic characteristics or other observations, we applied a relatively unbiased probabilistic approach to compare transcriptomes. Unlike most prior microarray studies that focus upon finding differential gene expression patterns among groups, we were most interested in those genes that are commonly important in defining data structure. While we would expect differences in the absolute levels or patterns of expression of some genes, our main goal was to explore the similarities in overall data structures. Differences in absolute gene expression values could lead to the appearance of differential gene expression values that may more closely reflect the cellular rather than molecular differences between relatively homogeneous cell lines and heterogeneous tumors.

The probabilistic approach we used compares the M PC projections in the data sets and those genes that best define these respective PCs. Thus, the method should capture, in a largely unbiased manner, those PCs and genes that best define the structure of each high dimensional data set - at least as defined by its total variation. Our data show that the three most widely used ER+ human breast cancer cell lines, even when growing *in vitro*, exhibit marked similarities to a panel of ER+ breast tumor specimens. These molecular observations on the primary structure of the breast tumor and cell line transcriptomes appear consistent with the widely reported biological similarities between these cell lines, their variants, and the human disease (2-4).

The genes identified in these tumors and models reflect the specimens and microarray technology used; similar data collected from other breast tumors, cell lines, or microarray

platforms may or may not find the same genes to define the M PCs of those data sets. However, we would anticipate that such studies may find genes that exhibit similar statistical properties, or perhaps broadly similar molecular functions, to be associated with the top PCs. Since we identified genes that best define the PCs from a small but robust subset of expression measures, their selection reflects a probabilistic assessment only of their contribution to global data structure. Thus, there is no compelling biological rationale why these specific genes must reflect key biological processes in breast tumors. The use of PCA for gene selection in mechanistic studies is potentially flawed for several reasons, some of which are discussed elsewhere (31). Nonetheless, it is intuitively reasonable to expect some genes closely associated with data structure to broadly reflect key molecular processes and/or include genes already implicated in breast cancer.

Several of the genes or gene functions represented in the 36 common genes identified herein have been directly or indirectly implicated in affecting key breast cancer phenotypes. For example, we found 11 genes in 5 of the 6 gene function categories implicated in separating good prognosis from poor prognosis breast cancers (17). While our study would not be expected to find the same genes as these two previous studies - we did not look for such discriminant genes nor did we use similar microarray platforms - the data in Table IV suggest that the 36 common genes and/or the functional categories they represent are important in both human breast cancer and human breast cancer cell lines. Examples of specific genes from the 36 common gene list include RELA (NFκB p65), ornithine decarboxylase-1 (ODC1), paxillin (PXN), and insulin-like growth factor (IGF) binding protein-6 (IGFBP-6). RELA is implicated in estrogen independence (32,33) and acquired antiestrogen resistance in cell culture models (15,34,35), and is readily detected by immunohistochemistry in breast tumors (36). The polyamine ODC1 is estrogen regulated (37,38), is a target for drug development (38,39), and is a potential breast cancer biomarker (40). The focal adhesion protein PXN is regulated by heregulin, a key effector of breast cancer cell growth (16). PXN expression also is regulated by activation of the IGF-type 1 receptor (41). This receptor is activated by IGF-II, a major mitogen for breast cancer cells (42); IGFBP6 has a notably high affinity for binding IGF-II, inhibits its activity (43), and also is a candidate breast cancer biomarker (21).

The data we present here suggest that well-established ER+ human breast cancer cell lines and breast tumors share global similarities in the structures of their respective transcriptomes. The strong correlations of similar genes with the top PC projections in each data set clearly imply that MCF-7, T47D, and ZR-75-1 cells are good models in which to identify molecular events that also are important in some ER+ human breast cancers.

Acknowledgments

This work was supported in part by Public Health Service award R01-CA096483 (R. Clarke), and a Clinical Translational Award (W81XWH-04-1-0570) from the United States Army Medical Research and Materiel Command Breast Cancer Research Program (R.C.), BC030280 (R.C.). Technical services also were provided by the Histopathology Shared Resource funded through Public Health Service award P30-CA51008-14 (Lombardi Comprehensive Cancer Center Support Grant). The authors wish to thanks Dr Rebecca Riggins for critical reading of the manuscript.

References

1. Ceriani RL, Peterson JA, Blank EW, Chan CM and Cailleau R: Development and characterization of breast carcinoma cell lines as *in vivo* models for breast cancer diagnosis and therapy. In Vitro Cell Dev Biol 28A: 397-402, 1992.
2. Clarke R: Human breast cancer cell line xenografts as models of breast cancer. The immunobiologies of recipient mice and the characteristics of several tumorigenic cell lines. Breast Cancer Res Treat 39: 69-86, 1996.
3. Clarke R, Leonessa F, Welch JN and Skaar TC: Cellular and molecular pharmacology of antiestrogen action and resistance. Pharmacol Rev 53: 25-71, 2001.
4. Lacroix M and Leclercq G: Relevance of breast cancer cell lines as models for breast tumours: an update. Breast Cancer Res Treat 83: 249-289, 2004.
5. Price JE, Polyzos A, Zhang RD and Daniels LM: Tumorigenicity and metastasis of human breast carcinoma cell lines in nude mice. Cancer Res 50: 717-721, 1990.
6. Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van de RM, Waltham M, Pergamenschikov A, Lee JC, Lashkari D, Shalon D, Myers TG, Weinstein JN, Botstein D and Brown PO: Systematic variation in gene expression patterns in human cancer cell lines. Nat Genet 24: 227-235, 2000.
7. Van't Veer LJ and Weigelt B: Road map to metastasis. Nat Med 9: 999-1000, 2003.
8. Burdall SE, Hanby AM, Lansdown MR and Speirs V: Breast cancer cell lines: friend or foe? Breast Cancer Res 5: 89-95, 2003.
9. Ross DT and Perou CM: A comparison of gene expression signatures from breast tumors and breast tissue derived cell lines. Dis Markers 17: 99-109, 2001.
10. Hotelling H: Analysis of a complex of statistical variables into principal components. J Educ Psychol 24: 417-441 and 498-520, 1933.
11. Jolliffe IT: Principal Component Analysis. 2nd edition. Springer-Verlag, New York, NY, 2002.
12. Ellis M, Davis N, Coop A, Liu M, Schumaker L, Lee RY, Srikanchana R, Russell CG, Singh B, Miller WR, Stearns V, Pennanen M, Tsangaris T, Gallagher A, Liu A, Zwart A, Hayes DF, Lippman ME, Wang Y and Clarke R: Development and validation of a method for using breast core needle biopsies for gene expression microarray analyses. Clin Cancer Res 8: 1155-1166, 2002.
13. James MR, Skaar TC, Lee RY, MacPherson A, Zwiebel JA, Ahluwalia BS, Ampy F and Clarke R: Constitutive expression of the steroid sulfatase gene supports the growth of MCF-7 human breast cancer cells *in vitro* and *in vivo*. Endocrinology 142: 1497-1505, 2001.
14. Sgroi DC, Teng S, Robinson G, Le Vangie R, Hudson JR and Elkahloun AG: *In vivo* gene expression profile analysis of human breast cancer progression. Cancer Res 59: 5656-5661, 1999.
15. Gu Z, Lee RY, Skaar TC, Bouker KB, Welch JN, Lu J, Liu A, Zhu Y, Davis N, Leonessa F, Brunner N, Wang Y and Clarke R: Association of interferon regulatory factor-1, nucleophosmin, nuclear factor-kappaB and cyclic AMP response element binding with acquired resistance to faslodex (ICI 182,780). Cancer Res 62: 3428-3437, 2002.
16. Rae JM, Ramus SJ, Waltham M, Armes JE, Campbell IG, Clarke R, Barndt RJ, Johnson MD and Thompson EW: Common origins of MDA-MB-435 cells from various sources with those shown to have melonoma properties. Clin Exp Metastasis 21: 543-552, 2004.
17. Pawitan Y, Bjohle J, Amler L, Borg AL, Egyhazi S, Hall P, Han X, Holmberg L, Huang F, Klaar S, Liu ET, Miller L, Nordgren H, Ploner A, Sandelin K, Shaw PM, Smeds J, Skoog L, Wedren S and Bergh J: Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. Breast Cancer Res 7: R953-R964, 2005.

18. Carter CL, Allen C and Henson DE: Relation of tumor size, lymph node status, and survival in 24,740 breast cancer cases. Cancer 63: 181-187, 1989.
19. Fisher B, Bauer M, Wickerham DL, Redmond CK, Fisher ER, Cruz AB, Foster R, Gardner B, Lerner H and Margolese R: Relation of number of positive axillary nodes to the prognosis of patients with primary breast cancer. An NSABP update. Cancer 52: 1551-1557, 1983.
20. Van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der KK, 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.
21. Kaulsay KK, Ng EH, Ji CY, Ho GH, Aw TC and Lee KO: Serum IGF-binding protein-6 and prostate specific antigen in breast cancer. Eur J Endocrinol 140: 164-168, 1999.
22. Rangarajan A, Hong SJ, Gifford A and Weinberg RA: Species- and cell type-specific requirements for cellular transformation. Cancer Cell 6: 171-183, 2004.
23. Rangarajan A and Weinberg RA: Opinion: comparative biology of mouse versus human cells: modelling human cancer in mice. Nat Rev Cancer 3: 952-959, 2003.
24. Hahn WC and Weinberg RA: Modelling the molecular circuitry of cancer. Nat Rev Cancer 2: 331-341, 2002.
25. Clarke R, Dickson RB and Lippman ME: Hormonal aspects of breast cancer: growth factors, drugs and stromal interactions. Crit Rev Oncol Hematol 12: 1-23, 1992.
26. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R and Staudt LM: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403: 503-511, 2000.
27. Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L, Kohn KW, Reinhold WC, Myers TG Andrews DT, Scudiero DA, Eisen MB, Sausville EA, Pommier Y, Botstein D, Brown PO and Weinstein JN: A gene expression database for the molecular pharmacology of cancer. Nat Genet 24: 236-244, 2000.
28. Thykjaer T, Workman C, Kruhoffer M, Demtroder K, Wolf H andersen LD, Frederiksen CM, Knudsen S and Orntoft TF: Identification of gene expression patterns in superficial and invasive human bladder cancer. Cancer Res 61: 2492-2499, 2001.
29. Satagopan JM and Panageas KS: A statistical perspective on gene expression data analysis. Stat Med 22: 481-499, 2003.
30. Hinneburg A and Keim DA: Optimal grid-clustering: towards breaking the curse of dimensionality in high-dimensional clustering. In: Proceedings of the 25th Conference on Very Large Databases. Atkinson MP, Orlowska ME, Valduriez P, Zdonik SB and Brodie ML (eds.) Morgan Kaufman, San Francisco, pp506-517, 1999.
31. Wittes J and Friedman HP: Searching for evidence of altered gene expression: a comment on statistical analysis of microarray data. J Natl Cancer Inst 91: 400-401, 1999.
32. Nakshatri H, Bhat-Nakshatri P, Martin DA, Goulet RJ and Sledge GW: Constitutive activation of NF-kappaB during progression of breast cancer to hormone-independent growth. Mol Cell Biol 17: 3629-3639, 1997.
33. Pratt MAC, Bishop TE, White D, Yasvinski G, Menard M, Niu MY and Clarke R: Estrogen withdrawal-induced NF-kappaB activity and bcl-3 expression in breast cancer cells: roles in growth and hormone independence. Mol Cell Biol 23: 6887-6900, 2003.
34. Riggins R, Zwart A, Nehra N, Agarwal P and Clarke R: The NFκB inhibitor parthenolide restores ICI 182,780 (Faslodex; Fulvestrant)-induced apoptosis in antiestrogen resistant breast cancer cells. Mol Cancer Ther 4: 33-41, 2005.
35. Osipo C, Gajdos C, Liu H, Chen B and Jordan VC: Paradoxical action of fulvestrant in estradiol-induced regression of tamoxifen-stimulated breast cancer. J Natl Cancer Inst 95: 1597-1608, 2003.
36. Zhu Y, Singh B, Hewitt S, Liu A, Gomez B, Wang A and Clarke R: Expression patterns among interferon regulatory factor-1, human X-box binding protein-1, nuclear factor kappa B, nucleophosmin, estrogen receptor alpha and progesterone receptor proteins in breast cancer tissue microarrays. Int J Oncol 28: 67-76, 2006.
37. Qin C, Samudio I, Ngwenya S and Safe S: Estrogen-dependent regulation of ornithine decarboxylase in breast cancer cells through activation of nongenomic cAMP-dependent pathways. Mol Carcinog 40: 160-170, 2004.
38. Manni A: Polyamine involvement in breast cancer phenotype. In Vivo 16: 493-500, 2002.
39. Manni A, Washington S, Mauger D, Hackett DA and Verderame MF: Cellular mechanisms mediating the anti-invasive properties of the ornithine decarboxylase inhibitor alpha-difluoromethylornithine (DFMO) in human breast cancer cells. Clin Exp Metastasis 21: 461-467, 2004.
40. Mimori K, Mori M, Shiraishi T, Tanaka S, Haraguchi M, Ueo H, Shirasaka C and Akiyoshi T: Expression of ornithine decarboxylase mRNA and c-myc mRNA in breast tumours. Int J Oncol 12: 597-601, 1998.
41. Guvakova MA and Surmacz E: The activated insulin-like growth factor I receptor induces depolarization in breast epithelial cells characterized by actin filament disassembly and tyrosine dephosphorylation of FAK, Cas, and paxillin. Exp Cell Res 251: 244-255, 1999.
42. Sachdev D and Yee D: The IGF system and breast cancer. Endocr Relat Cancer 8: 197-209, 2001.
43. Bach LA: IGFBP-6 five years on; not so 'forgotten'? Growth Horm IGF Res 15: 185-192, 2005.

Expression patterns among interferon regulatory factor-1, human X-box binding protein-1, nuclear factor kappa B, nucleophosmin, estrogen receptor-alpha and progesterone receptor proteins in breast cancer tissue microarrays

YUELING ZHU[1], BALJIT SINGH[3], STEPHEN HEWITT[4], AIYI LIU[5], BIANCA GOMEZ[1], ANTAI WANG[2] and ROBERT CLARKE[1]

Departments of [1]Oncology, and [2]Biostatistics and Biomathematics, Lombardi Comprehensive Cancer Center, Georgetown University Medical Center, Washington, DC; [3]Department of Pathology, New York University School of Medicine, New York, NY; [4]Tissue Array Research Program, Laboratory of Pathology, Center for Cancer Research, National Cancer Institute; [5]Biometry and Mathematical Statistics Branch, National Institute of Child Health and Development, National Institutes of Health, Bethesda, MD; USA

Abstract. Interferon regulatory factor-1 (IRF-1), human X-box binding protein-1 (hXBP-1), nuclear factor kappa B p65 (NFκB p65) and nucleophosmin (NPM) have been implicated in a signaling network of endocrine responsiveness. Expression of these proteins was measured by immunohistochemistry in tissue microarrays of 54 breast tumors. Correlations between each protein and established prognostic markers were assessed by Spearman's rank order correlation coefficient and partial correlation coefficient analyses. Moderate/strong staining is seen for hXBP-1 (79% of tumors) and NFκB p65 (57%). NPM exhibits nuclear staining (95%); IRF-1 exhibits both cytosolic (IRF-1c; 90%) and nuclear staining (IRF-1n; 51%). IRF-1c is associated with age (p=0.034); IRF-1n and PgR expression are correlated (p=0.014). NFκB p65 shows a borderline association with S phase (p=0.062). Coexpression of IRF-1c and hXBP1 (p=0.001), IRF-1c and NFκB (p=0.002), and hXBP-1 and NFκB (p=0.018) is observed. An inverse correlation exists between IRF-1n and NFκB (p=0.034). All four proteins are detected in breast tumors and their expression patterns support their role(s) in a key signaling network.

Introduction

Endocrine therapy, usually either the antiestrogen, Tamoxifen (TAM), ovariectomy or more recently an aromatase inhibitor or one of the newer selective estrogen receptor modulators (SERM) or 'pure' antiestrogens, is an effective means to manage hormone-dependent breast cancer (1-3). An understanding of the mechanisms of resistance to endocrine therapies could identify better ways to predict responsiveness. We have previously hypothesized that endocrine responsiveness is affected by a complex gene network, rather than the activity of only one or two genes or signaling pathways (4-6). To identify the key components of such a network, we first derived variants of the MCF-7 human breast cancer cell line with different estrogen (7,8) and antiestrogen response profiles (9,10). Initial transcriptome and proteome analyses of these variants implicate several genes in endocrine resistance, including interferon regulatory factor-1 (IRF-1) (11,12), nuclear factor kappa B p65 (NFκB) (11,13), human X-Box binding protein-1 (hXBP-1) (11) and nucleophosmin (NPM) (14,15), which appear to function as part of a broader gene expression network (Fig. 1).

In the proposed network, NPM is predicted to inhibit IRF-1 activity, which reduces the ability of IRF-1 to activate apoptosis, most likely through inducing a caspase cascade. Inhibition of IRF-1 activity also may eventually contribute to increased activity of the survival factor NFκB (Fig 1). Increased NFκB may, in turn, induce a second survival factor, hXBP-1 (16). Evidence from experimental models has begun to show the likely functional relevance of the altered IRF-1 (12), NFκB (11,13,17), and hXBP-1 activities (Gomez BP, *et al*, Proc Am Assoc Cancer Res, abs. 3498, 2004) in affecting endocrine responsiveness. Studies to identify other members of this network and their interrelationships are currently in progress. The known functions of the key components of the network are described below.

IRF-1 is a transcription factor that exhibits tumor suppressor activities in several cancers (18,19). In breast cancer cells, IRF-1 signaling can reduce both the rate of cell proliferation and the incidence of human breast cancer xenografts in athymic nude mice (manuscript submitted). We have shown that a

Correspondence to: Dr Robert Clarke, Department of Oncology, Georgetown University School of Medicine, Room W405A Research Building, 3970 Reservoir Rd, NW, Washington, DC 20057, USA
E-mail: clarker@georgetown.edu

Key words: immunohistochemistry, antiestrogen, Faslodex, tamoxifen, gene network

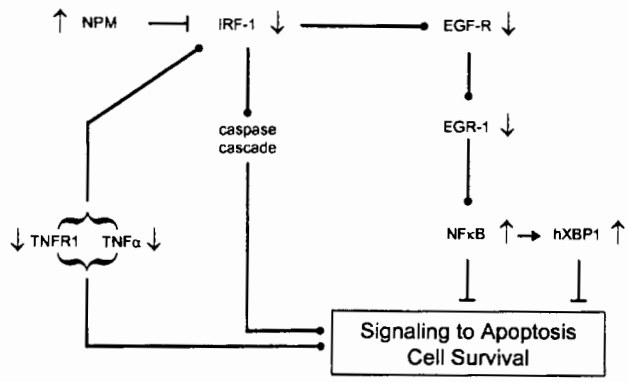


Figure 1. Components of a putative signaling network associated with endocrine responsiveness in breast cancer cells (adapted from ref. 6). ↑, increased; ↓, decreased; ⊥, blocks; —•, reduced ability to affect target. The network component incorporates the known protein/protein interactions between NPM and IRF-1 (NPM binds to IRF-1 and inhibits IRF-1 activity) and the predicted regulation of hXBP-1 by NFκB. Down-regulation of IRF-1 activity would reduce the activity of IRF-1/NFκB heterodimers that are known to regulate the transcription of several genes implicated in breast cancer, such as RANTES, VCAM-1, and IL-6. Down-regulation of TNFR1 (tumor necrosis factor receptor 1) and its ligand, TNFα (tumor necrosis factor alpha), were previously described (11), signaling from this complex is a major inducer of IRF-1 transcription. A full description of this network component and its anticipated function can be found in ref. 6.

dominant negative IRF-1 blocks antiestrogen-induced apoptosis in sensitive breast cancer cells and reduces their antiestrogen sensitivity (12); a similar role for IRF-1 has been recently reported in normal mammary cells (20). These activities of IRF-1 are probably mediated through its proapoptotic effects, which can occur in a p53-dependent or -independent manner (21,22) and involve its ability to induce a caspase cascade that includes caspase-1 (20,22), caspases-3/7 (20,23), caspase-8 (24) and/or Fas ligand (25). Caspases are known to affect antiestrogen responsiveness (26). Lower levels of IRF-1 protein have been reported in high-grade ductal carcinoma *in situ* or invasive ductal carcinoma of the breast when compared with adjacent normal breast epithelium (27).

The NFκB p50/p65 heterodimer complex comprises two homologous proteins encoded by different genes; the p105 precursor of p50 (NFκB1) is on chromosome 4, while the p65 (RelA) gene is on chromosome 11. The predominant form in human breast cancer cell lines is NFκB (p50/p65); another member of the family (p52) is also expressed in some breast cancers (28). NFκB (p50/p65) is implicated in several critical cellular functions including cell survival (29); these functions are often cell context specific (30). We have shown differential expression and activation of NFκB expression with both acquired antiestrogen resistance (11) and estrogen independence in breast cancer cells (13). Other studies also show increased expression of NFκB in endocrine-resistant breast cancer cell lines (31,32). We have begun to establish the functional relevance of these observations. For example, estrogen-independent cells significantly up-regulate NFκB; when its inhibitor IκBα is overexpressed in these cells their xenografts regress upon estrogen withdrawal (13). Anti-estrogen-resistant cells are more sensitive to growth inhibition by parthenolide, a small molecule inhibitor of NFκB, than their antiestrogen sensitive parental cells (11). Furthermore, parthenolide can reverse the antiestrogen resistance phenotype and synergistically interact with antiestrogens *in vitro* (17).

As a member of the ATF/CREB transcription factor family that activate promoters containing specific cyclic AMP responsive elements (CRE) (33), hXBP-1 regulates the expression of several tissue-specific genes, including tissue inhibitor of metalloproteinases, osteopontin and osteocalcin (34). Potentially downstream of NFκB activation in some cells (35), hXBP-1 is associated with increased proliferation and reduced apoptosis (16), which implies a survival function. Changes in cAMP concentrations and CRE activation have been widely implicated in carcinogenesis and endocrine signaling, including affecting signaling from ERα and PgR (36). While the role of hXBP-1 in the normal/neoplastic breast has not been studied in detail, hXBP-1 is part of a cluster of genes associated with some ERα-positive breast tumors (37,38) and a recent study suggests it may be expressed in breast cancer cells (39). We have previously implicated increased expression of hXBP-1 in acquired antiestrogen resistance (11). More recently, we have shown the ability of hXBP-1 to induce an estrogen-independent phenotype and to confer antiestrogen resistance (unpublished data).

The oncogenic nucleolar phosphoprotein, NPM, is a DNA-binding protein (40) that inhibits the ability of the YY1 (41) and IRF-1 transcription factors to regulate gene expression (42). NPM also serves as a substrate for several important serine-threonine kinases, including protein kinase C (43), $p34^{cdc2}$ kinase (44,45) and casein kinase II (46). Insulin, which is a major mitogen for breast cancer cells, also increases NPM phosphorylation (45). Overexpression of NPM is sufficient to transform NIH/3T3 fibroblasts (47), and chromosomal translocations fusing NPM to either an anaplastic lymphoma kinase (48) or the retinoic acid receptor-α (49) have been reported in some cancers. We have shown that NPM is induced by estradiol (14) and expressed at higher levels in estrogen-independent breast cancer cells (11); a putative estrogen responsive element in the NPM promoter has now been recently described (50). In breast cancer patients, auto-antibodies to NPM are lower in patients treated with the anti-estrogen Tamoxifen and increase six months prior to recurrence (15). Of particular relevance is the reduced expression of IRF-1 and concurrent increased expression of NPM, an endogenous IRF-1 inhibitor, in antiestrogen-resistant breast cancer cells (11).

We have now measured the expression of IRF-1, NFκB (p65), hXBP-1 and NPM in breast cancer specimens from women diagnosed at our institution. Using tissue microarrays and immunohistochemistry, we asked if these proteins could be detected in breast cancer, whether their expression might be correlated with other known prognostic markers, and whether the four proteins are expressed in patterns consistent with their known functions and/or our putative gene expression network. We find several proteins to be either coexpressed or inversely expressed in patterns consistent with our network hypothesis.

Materials and methods

Tissue specimens. Tissue microarrays were constructed using fifty-four, untreated, primary breast cancer cases diagnosed

between 1998 and 1999 from the breast cancer tumor bank at the Lombardi Comprehensive Cancer Center Histopathology Shared Resource at Georgetown University Medical Center. The cases were initially selected to determine the number of cores from a tumor needed to give the same estimation of ERα and PgR positivity as the entire section (51). Hence, the proportion of steroid hormone receptor-positive specimens (81% ERα; 44% PgR) are higher than might be expected from a random sampling of breast cancer cases. While we cannot exclude the possibility of some selection bias in these cases, this selection should have identified cases most relevant to our initial hypothesis, implicating the proteins of interest in endocrine responsiveness (52). Differentiation/nuclear grade (53), DNA index (54) and S Phase (55) were determined as previously described. The categories for each end-point in Table I were selected prior to data analysis and are consistent with other studies (56,57). All material and information was collected and used in accordance with approved Institutional Review Board protocols. Clinical-outcome data and additional prognostic marker data are not available for these cases; available data are shown in Table I.

Table I. Patient/tumor characteristics.

	n	Range
Age	54	36-85 (56)[a]
<50 years	25	
≥50 years	29	
Tumor grade	49	0-2
Grade 1	13	
Grade 2	26	
Grade 3	10	
Tumor size	54	0.2-6.8 (1.35)
<2 cm	39	
≥2 cm	15	
Lymph nodes	40	0-4
Negative (0)	28	
Positive (≥1)	12	
DNA index	47	1-2.89
<1.5%	27	
≥1.5%	20	
S-phase	28	1.64-27.00 (5.09)
<5% (low)	14	
≥5% (high)	14	

[a]Values in parentheses are median values.

Tissue microarrays. Tissue microarrays were constructed with a Beecher Instruments manual tissue arrayer (Beecher Instruments, Inc., Sun Prairie, WI) as previously described (58,59). The instrument punches holes in the recipient paraffin block and acquires tissue cores from the donor block. Briefly, a thin-walled needle with an inner diameter of 0.6 mm was held in an X-Y precision guide. The cylindrical sample was retrieved from the selected region in the donor block and extruded directly into the recipient block with defined array coordinates. A solid steel wire, which closely fits the tube, was then used to transfer the tissue cores into the recipient block. The transfer was made under direct visual control with a stereotactic microscope using an additional bright light source. This cycle was repeated to obtain the appropriate number of cores. An adhesive-coated tape system (Instrumedics, Inc., Hackensack, NJ) was then used to cut 5 μm sections of the tissue microarray block. The microtome knife cut underneath tape placed over the block surface. Thin tissue sections adhered to the tape, which was then rolled on an adhesive-coated microscope slide to transfer the section onto the slide. For this study, tissue microarrays were built with 480 cores from fifty-four breast carcinomas. Regions of invasive carcinoma were marked on each hematoxylin-and-eosin-stained slide. Ten cores were made from these areas of the paraffin block for 42 cases; for 12 additional cases, five cores were made. Thus, either 10 or 5 cores represented each tumor.

Antibodies and immunohistochemistry. The following commercial antibodies were used: ERα (ER1D5, Immunotech) (60), erbB2 (CB11; Zymed, San Francisco, CA) (61), hXBP-1 (sc-7160; Santa Cruz); IRF-1 (sc-497; Santa Cruz) (27), and NFαB p65 (sc-109; Santa Cruz) (28). The NPM monoclonal antibody was kindly donated by Dr P-K Chan (62). Tissue microarray sections were deparaffinized in two 5-min changes of xylene and rehydrated through graded alcohol to distilled water. Immunohistochemistry was performed by a standard biotin-streptavidin-horseradish peroxidase method (63,64). Briefly, microarrays were treated with 1% H_2O_2 in methanol for 30 min to block endogenous peroxidase activity. Before applying the primary antibody, microarrays were boiled for antigen retrieval in 10 mM citrate buffer (pH 6.0) for a total 10 min. Microarrays were washed in phosphate-buffered saline containing 3% biotinylated goat antiserum to the appropriate IgG and 0.3% Triton X-100 (pH 7.4) for 30 min. Subsequently, tissue microarrays were incubated with the primary antibody at a 1:500 dilution (or as appropriate for the antibody) in PBS for 48 h at 4°C. After several washes, microarrays were treated with the appropriate secondary antibody (1:800; Vector Laboratories, Burlingame, CA) for 2 h, followed by a 1 h incubation with streptavidin-peroxidase conjugate (Vector Laboratories). Antigen-antibody complex was visualized by incubation with the VIP Kit (DAB Kit; Vector Laboratories). Finally, microarrays were counter-stained with either methyl green or hematoxylin, mounted and examined. All immunostaining was first optimized in single tissue slides. Negative controls were obtained using a standard method where microarrays are processed as described above but without incubation with the appropriate primary antibody.

Data analysis. The level of specific immunostaining, as determined relative to negative controls, was measured as an ordinal variable according to the nominal scale 0, 1+, 2+, 3+; where 0 is undetectable, 1+ refers to weak (barely perceptible) staining, 2+ to moderate staining, and 3+ to strong staining.

For nuclear staining, the scale applied was 1 = 0-25% of nuclei with detectable staining, 2 = 26-50%, 3 = 51-75%, 4 = ≥76%. The average score for all cores representing a tumor was used for data analysis. The relationships among staining values for each protein were compared using Spearman's rank order correlation coefficient analysis. All statistical tests are two-sided. We considered comparisons where $p<0.05$ to be statistically significant; estimates of $p\geq0.05$ and $p\leq0.10$ were considered to indicate borderline statistical significance and potential biological relevance; comparisons where $p>0.10$ were considered to be insignificant.

Pairwise correlation analyses could not account for the possibility that the associations of IRF-1n or IRF-1c may confound each other, since the expression of these two IRF measures may be correlated. To address this issue, we applied a novel use of partial correlation coefficient analysis, the partial correlations being calculated as shown in Eq 1:

$$r_{xy.z} = \frac{r_{xy} - (r_{xz})(r_{yz})}{\sqrt{(1-r_{xz}^2)(1-r_{yz}^2)}} \qquad \textbf{Eq 1}$$

Where $r_{xy.z}$ = the correlation coefficient between x and y while controlling for the correlations between x and z and between y and z.

Partial correlations are most widely applied in the analysis of small signaling networks of 3-5 variables, and allow an estimate of the correlation between two variables while controlling for a third, fourth and/or fifth. Since we make functional assessments based upon cellular location, the use of partial correlations appears reasonable in the context of IRF-1n and IRF-1c. For the correlations between IRF-1c or IRF-1n and age, ER, PgR, NFκB and hXBP-1 the partial correlations were calculated with either IRF-1c or IRF-1n as the controlling variable.

Results

ERα and PgR expression. Measurements of ERα and PgR expression are the most widely used predictive factors in directing breast cancer therapy. The specimens in this study were originally selected to study ERα and PgR expression (51) and these two proteins are coexpressed in a substantial proportion of breast cancers. The expression of ERα (81%) and PgR (44%) using our criteria (Table II), and their significant coexpression in the tumors used in this study (Table V; $p<0.001$) implies that the samples are likely to be broadly representative of ERα-positive breast cancers and appropriate for exploring protein expression patterns in cases likely to be selected for endocrine therapy.

IRF-1 expression. As a putative tumor suppressor, we might expect activated IRF-1 protein to be in the nucleus (IRF-1n) and inactive protein to be in the cytosol (IRF-1c). Whether these relationships are true for the IRF-1 signals we have measured is not known but we might expect the inactive form to predominate. In this context, and consistent with its putative tumor suppressor activities, the primary form of IRF-1 in breast tumors in this study appears to be IRF-1c (Fig. 2A). Of the tumors, 90% express detectable (2+ or 3+) IRF-1c in their neoplastic cells, almost half of which have 3+ IRF-1 staining in the cytosol. In contrast, only 51% of the tumors in our study express detectable IRF-1n in >50% of the tumor cells and no tumors express IRF-1n in >75% of cells (Tables II and III). While 98% of the specimens express both detectable IRF-1c and IRF-1n, only 2% express IRF-1c alone and none express only IRF-1n. The inverse relationship between IRF-1n and IRF-1c ($p=0.088$), while of borderline statistical significance, suggests that some breast tumors may differentially regulate the activation state of IRF-1 (Table V). This potential correlation raises the possibility that some associations implicating IRF-1c or IRF-1n may be confounded by the effect of the other. Our observations also are broadly consistent with a study reporting higher levels of IRF-1 protein in adjacent normal breast epithelium when compared with high-grade ductal carcinoma *in situ* or lymph node-positive invasive ductal carcinoma of the breast (27).

Table II. Immunohistochemical staining scores of five proteins detected in the cytosol.

Score	ERα	PgR	NFκB	[b]IRF-1c	hXBP-1
0	[a]9	22	2	0	1
1+	1	7	18	4	9
2+	5	3	14	21	27
3+	37	20	13	17	10
Total	52	52	47	42	47
[c]Detected	81% (42/52)	44% (23/52)	57% (27/47)	90% (38/42)	79% (37/47)

[a]Values represent the number of cases in each category; scoring categories are described in Materials and methods. [b]IRF-1c, IRF-1 cytoplasmic staining. [c]Detected, proportion of cases with weak or stronger cytosolic staining.

NFκB expression. We measured NFκB p65 expression, which is the predominant form of NFκB in human breast cancer cells (28) and the form associated with both estrogen independence (13) and acquired antiestrogen resistance (11). While active in breast cancer cell lines, NFκB p65 has been reported as being cytosolic (potentially inactive) whereas NFκB p50 has been reported to be primarily nuclear (active) in a prior study of n=17 breast tumors (28). While the pattern of NFκB p65 staining is broadly consistent with this observation in many of our breast tumors (Fig. 2C), we found 57% of the tumors to express detectable (2+ or stronger) NFκB in their neoplastic cells (Table II).

hXBP-1 expression. Increased expression of hXBP-1, a nuclear transcription factor that activates cyclic AMP responsive elements (33), is associated with some forms of acquired anti-estrogen resistance (11). hXBP-1 expression is detected in 79% of the breast tumors in this study (Table II), with the strongest staining seen in the cytosol (Fig. 2E). This observation

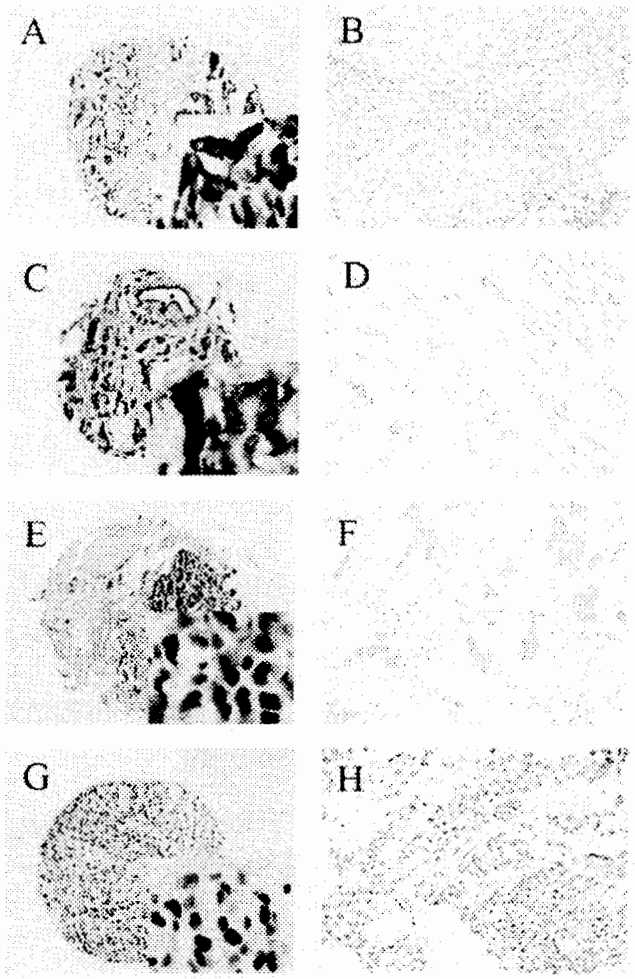


Figure 2. Representative immunostaining of IRF-1, NFκB p65, hXBP-1 and NPM in breast cancer. (A) IRF-1 staining where the inset shows typical patterns of cytosolic and nuclear staining; (B) Control for IRF-1 staining; (C) NFκB p65 staining where inset shows the primarily cytosolic staining pattern; (D) Control for NFκB p65 staining; (E) hXBP-1 staining where inset shows the primarily cytosolic staining pattern; (F) Control for hXBP-1 staining; (G) NPM staining where the inset shows the nuclear staining pattern for NPM; (H) Control for NPM staining. Figures of tissue microarray cores are at x10 magnification; inset at x100 magnification; control at x40 magnification.

is consistent with a small study of hXBP-1 expression in primary breast cancers (n=11) and breast cancer cell lines (n=5). In this recent study, expression was detected in all tumors and cell lines studied but hXBP-1 was almost undetectable in non-cancerous breast tissue (33).

NPM expression. NPM is a nucleolar phosphoprotein that is induced by estradiol (14) and expressed at higher levels in breast cancer cells with acquired antiestrogen resistance (11). In breast cancer patients, autoantibodies to NPM increase six months prior to recurrence and are lower in patients treated with TAM (15). Consistent with its nucleolar localization in cell culture, NPM staining is strongly nuclear in breast tumors (Fig. 2G). Of the breast cancers in this study, 95% express NPM in >50% of their neoplastic cell nuclei, the majority expressing NPM in >75% of their cells (Table II).

Table III. Immunohistochemical nuclear staining scores of IRF-1 and NPM.

Score	IRF-1n	NPM
1	2	1
2	22	1
3	25	3
4	0	39
Total	49	44
[c]>50%	51% (25/49)	95% (42/44)

[a]Values represent the number of cases in each category; scoring categories are described in Materials and methods. [b]IRF-1n, IRF-1 nuclear staining. [c]>50%, proportion of cases where data are available that exhibit >50% of cell nuclei staining positive relative to the negative controls (NPM and IRF-1n are data for nuclear staining).

Correlation among proteins and patient/tumor characteristics. Several correlations among existing prognostic markers are known and are apparent in our data set (Table IV). Both PgR-positive (p=0.03) and ERα-positive tumors (borderline) are associated with a greater degree of differentiation and better prognosis (65). Borderline relationships between DNA index and both PgR-positivity and ERα-positivity (inverse correlation), and between NFκB and S phase (direct correlation) are also evident. A higher incidence of ERα-positive tumors is seen in older women (66) but our study was probably underpowered to detect this relationship. Nonetheless, the significant association between IRF-1c and age (Table 1V; p=0.034) and the potential association between IRF-1c and ERα (p=0.079), may reflect the underlying relationship between ERα and age. We found no other associations among IRF-1, NFκB, hXBP-1 and NPM with either tumor grade, tumor size, DNA index, lymph node status, or S-phase fraction.

Correlation among protein expression patterns. Expression of several of the four proteins is correlated in breast tumors. Since our study is limited in size and power, we present those associations that reach conventional statistical significance and those where the association is of borderline statistical significance but of potential biological relevance. The data in Table V show coexpression of ER and IRF-1c (borderline), PgR and IRF-1n (p=0.014), IRF-1c and hXBP1 (p=0.001), IRF-1c and NFκB (p=0.002), and hXBP-1 and NFκB (0.018). Inverse correlations were seen between NPM and erbB2 (not shown; p=0.016), IRF-1n and NFκB (p=0.034), IRF-1n and IRF-1c (borderline), and IRF-1n and hXBP-1 (borderline). We estimated the partial correlations for each IRF-1n and IRF-1c correlation of interest; no effect is present when the sign and magnitude of the partial correlation coefficient is comparable to the original correlation coefficient. In each case, the partial coefficients were very similar to the original coefficients and shared the same sign. Hence, IRF-1c and IRF-1n are not antecedent, intervening, or suppressing variables

Table IV. Correlation among proteins and patient/tumor characteristics.

	ERα	PgR	IRF-1c	IRF-1n	NFκB	hXBP-1	NPM
Age	-	-	P=0.034 (r=0.28)	-	-	-	-
Tumor grade	P=0.067 (r=0.22)[a]	P=0.028 (r=0.28)	-	-	-	-	-
Tumor size	-	-	-	-	-		
Lymph nodes	-	-	-	-	-	-	-
DNA index	P=0.077 (r=0.22)	P=0.086 (r=-0.21)	-	-	-	-	-
S phase	-	P=0.016 (r=0.41)	-	-	P=0.062 (r=0.33)	-	-

[a]Values in parentheses are Spearman rank correlation coefficients. Comparisons where $p<0.05$ are statistically significant; estimates of $p\geq0.05$ and $p\leq0.10$ are considered of borderline statistical significance and of potential biological relevance; comparisons where $p>0.10$ were considered to be insignificant.

Table V. Correlation among protein expression patterns.

	ERα	PgR	[b]IRF-1c	IRF-1n	NFκB	hXBP-1
PgR	P<0.001 (r=0.46)[a]	1				
IRF-1c	P=0.079 (r=0.23)	-	1			
IRF-1n	-	P=0.014 (r=0.32)	P=0.088 (r=-0.21)	1		
NFκB	-	-	P=0.002 (r=0.44) (r=0.42)	P=0.034 (r=-0.27) (r=0.22)	1	
hXBP-1	-	-	P=0.001 (r=0.49) (r=0.40)	P=0.082 (r=-0.21) (r=-0.23)	P=0.018 (r=0.31)	1
NPM	-	-	-	-		-

[a]Values in parentheses are Spearman rank correlation coefficients; comparisons where $p<0.05$ are statistically significant; estimates of $p\geq0.05$ and $p\leq0.10$ are considered of borderline statistical significance and of potential biological relevance; comparisons where $p>0.10$ were considered to be insignificant. Values in parentheses are the estimated partial correlation coefficients. Use of partial correlation coefficients in networks can be found in De la Fuenta, *et al*: Bioinformatics 20: 3565-3575, 2004. [b]IRF-1c, cytoplasmic staining; IRF-1n, nuclear staining.

for the correlations indicated, they exhibit respectively with hXBP-1 or NFκB (Table V).

Discussion

One approach to exploring the potential relevance of observations from experimental models is to determine whether similar relationships may also arise in tumors from patients. While not directly informative in a mechanistic sense, identification of expression patterns in tumors that reflect patterns seen in xenografts and cell cultures can support mechanistic observations in these models. Furthermore, such studies may identify candidate biomarkers for further investigation. We have explored the expression levels and patterns of coexpression of a subset of four genes (IRF-1, NFκB p65, hXBP-1, NPM) implicated in endocrine resistance

from our prior studies in experimental models (11-14).

Of the four genes we have previously implicated, IRF-1, NFκB and hXBP-1 are transcription factors and NPM is a DNA-binding nucleolar phosphoprotein. Knowledge of a signal's cellular localization can provide mechanistic insight and all four proteins would be expected to exhibit some degree of nuclear staining that could reflect active protein. For example, NFκB is maintained in the cytosol in an inactive state, complexed with members of the IκB family (67). However, correctly identifying subcellular localization by immunohistochemistry can be confounded by fixation artifacts, leading to nuclear antigen redistribution during tissue processing. A fixation artifact is responsible for the apparent cytosolic localization of the NPM-anaplastic lymphoma kinase fusion antigen (68) but NPM staining is robust and primarily nuclear in our breast cancer specimens. In contrast, the activation state of hXBP-1 and NFκB (p65) is difficult to determine because the staining is primarily cytosolic and any weak nuclear staining was not sufficient for further analysis. In this study, we chose to focus on the localization of IRF-1, which exhibits readily detectable nuclear and cytosolic staining patterns that appear inversely correlated (borderline; p=0.088). Furthermore, these patterns of staining for IRF-1 also make biological sense when considered in the context of putative active (nuclear) and inactive (cytosolic) states (see below).

Since there are only very limited published data on the expression of IRF-1, NFκB, hXBP-1 and NPM in breast cancer, we first determined whether we could detect these proteins and estimate the extent to which they are expressed in this series of predominately ER+ breast tumors. All four proteins are detectable in the cases used in this study; NPM expression is detected in >25% of the neoplastic cells in almost all the breast cancers (95%). Consistent with recent reports, IRF-1 also is detected in breast tumors (90%) (27,69), as is hXBP-1 (79%) (39). NFκB (p65) expression is the least frequently detected among the four proteins yet is detectable in 57% of the tumors. Thus, these four proteins are present in high proportions of breast cancer and are candidate biomarkers that merit further evaluation as both independent biomarkers and as a possible panel to be concurrently measured.

We have previously hypothesized that the proteins of interest are associated with affecting endocrine responsiveness (6,11,12). Acquired antiestrogen resistance primarily occurs in tumors that continue to express sufficient levels of ERα to be considered ERα-positive. While the primary form of *de novo* endocrine resistance is the absence of both ERα and PgR, a significant proportion of *de novo*-resistant tumors also are ERα-positive (52). The functional importance of continued receptor expression in either acquired or *de novo* endocrine resistance is unclear, but we might expect to find some of our network members to be coexpressed in the breast tumors used in this study (6).

A significant positive association between IRF-1n and PgR (p=0.014) and a borderline positive association between ERα and IRF-1c (p=0.079) are evident. Those PgR-positive tumors that coexpress IRF-1n may have a better prognosis and/or a better response to antiestrogens. For example, we have recently shown that the ability of the steroidal antiestrogen ICI 182,780 (Faslodex; Fulvestrant) to signal apoptosis is mechanistically related to its ability to regulate IRF-1 expression and function in breast cancer cells (12).

We could not confirm coexpression of ERα and hXBP-1 (p=0.244), an association predicted from hierarchical cluster analysis of cDNA expression microarray data from human breast tumors (37,38). Several explanations for this outcome are possible. The nature of the signals from gene expression microarrays that measure mRNA and tissue microarrays that measure protein are very different. It also is not clear how closely the levels of mRNA and protein are related for hXBP-1. Furthermore, some of the associations/relationships identified in gene expression microarray studies were found by simple hierarchical clustering and these may not be correct or complete. The use of these clustering methods to identify gene expression patterns from within the very high dimensional data spaces generated by gene expression microarrays has been seriously questioned (70,71).

hXBP-1 expression is positively correlated with IRF-1c expression (p=0.001) but inversely associated with IRF-1n (borderline; p=0.082). These observations suggest a balance between IRF-1's inhibitory activity and hXBP-1's mitogenic activity. For example, tumors where hXBP-1 activity predominates may have a poor prognosis and/or poor response to antiestrogens. Some antiestrogen-resistant cells exhibit down-regulated IRF-1 activation and up-regulated hXBP-1 activity (western; promoter-reporter data) (11,12).

Expression of hXBP-1 and NFκB (p65) are positively correlated (p=0.018). If we assume that NFκB is inactive because of its cytosolic location, the coexpression of hXBP-1 might compensate for any lack of NFκB in affecting cell survival since both are antiapoptotic. However, hXBP-1 expression appears to be downstream of NFκB, at least in plasma cell differentiation (35), implying a potential induction of hXBP-1 by NFκB. If this occurs in breast cancer cells and NFκB is active, as suggested by the potential correlation between NFκB and S-phase (p=0.062), it may explain the coexpression of hXBP-1 and NFκB in Table V. We also cannot exclude the possibility that NFκB p50 and/or NFκB p52 expression are activated and may compensate for any loss of NFκB p65 activity (28).

IRF-1 and NFκB proteins form heterodimers that can regulate gene expression and we might expect to find these coexpressed in the same tumors. We found a significant coexpression of NFκB and IRF-1c (p=0.002) and an inverse association between IRF-1n and NFκB (p=0.034). Where both proteins are primarily sequestered in the cytosol, the ability of IRF-1:NFκB heterodimers to regulate gene transcription could be inhibited. Several genes regulated by these heterodimers are implicated in breast cancer, including RANTES (regulated upon activation, normally T-Expressed and presumably secreted) (72), VCAM-1 (vascular cell adhesion molecule-1) (73) and IL-6 (interleukin-6) (74). RANTES expression correlates with a poor prognosis in breast cancer (75). VCAM-1 is involved in angiogenesis and metastasis in breast tumors (76), and an autocrine production of IL-6 is associated with drug resistance in breast cancer cells (77). The inverse relationship between IRF-1n and NFκB suggests that some tumors may have activated IRF-1 in the absence of active NFκB; such tumors may have a good prognosis and/or be sensitive to antiestrogens.

We obtained limited expression data for erbB2 (not shown). We detected a significant inverse association between erbB2 and NPM (p=0.016), suggesting that the oncogenic properties of NPM may be important in erbB2 non-overexpressing breast tumors, which represent the majority of breast cancer. No association was seen between erbB2 and either IRF-1, hXBP-1 or NFκB.

The present study represents the first analysis of the coexpression patterns of a subset of genes associated with acquired endocrine resistance in breast cancer cells. We could not adequately assess the activation state of each of the proteins and clinical-outcome data are not available in this data set. Despite these limitations, the data clearly show that all four proteins are detectable in a high proportion of the breast tumors used in this study. The data are consistent with a role for IRF-1, NFκB, hXBP-1 and NPM and their interactions in breast cancer, and are broadly supportive of the proposed component of a larger signaling network as outlined in Fig. 1. Further analysis of the expression patterns of IRF-1, NFκB, hXBP-1 and NPM as potential biomarkers for further defining endocrine response profiles in some breast cancer patients is warranted.

Acknowledgements

This study was supported in part by Public Health Service award R01-CA/AG58022-10 (R. Clarke), R01-CA096483-01 (R. Clarke) and Department of Defense awards 17-00-1-0256 (R. Clarke), BC010619 (R. Clarke), BC030280 (R. Clarke) and BC010531 (B. Gomez) from the United States Army Medical Research and Material Command. Technical services also were provided by the Histopathology Shared Resource funded through Public Health Service award P30-CA51008-14 (Lombardi Comprehensive Cancer Center Support Grant; B. Singh).

References

1. Early Breast Cancer Trialists' Collaborative Group: Tamoxifen for early breast cancer: an overview of the randomized trials. Lancet 351: 1451-1467, 1998.
2. Early Breast Cancer Trialists' Collaborative Group: Systemic treatment of early breast cancer by hormonal, cytotoxic or immune therapy. Lancet 399: 1-15, 1992.
3. Miller WR: Biological rationale for endocrine therapy in breast cancer. Best Pract Res Clin Endocrinol Metab 18: 1-32, 2004.
4. Clarke R and Brünner N: Acquired estrogen independence and antiestrogen resistance in breast cancer: estrogen receptor-driven phenotypes? Trends Endocrinol Metab 7: 25-35, 1996.
5. Clarke R, Skaar TC, Bouker KB, Davis N, Lee YR, Welch JN and Leonessa F: Molecular and pharmacological aspects of antiestrogen resistance. J Steroid Biochem Mol Biol 76: 71-84, 2001.
6. Clarke R, Liu MC, Bouker KB, Gu Z, Lee RY, Zhu Y, Skaar TC, Gomez B, O'Brien K, Wang Y and Hilakivi-Clarke LA: Antiestrogen resistance in breast cancer and the role of estrogen receptor signaling. Oncogene 22: 7316-7339, 2003.
7. Clarke R, Brunner N, Katzenellenbogen BS, Thompson EW, Norman MJ, Koppi C, Paik S, Lippman ME and Dickson RB: Progression from hormone dependent to hormone independent growth in MCF-7 human breast cancer cells. Proc Natl Acad Sci USA 86: 3649-3653, 1989.
8. Brunner N, Boulay V, Fojo A, Freter CE, Lippman ME and Clarke R: Acquisition of hormone-independent growth in MCF-7 cells is accompanied by increased expression of estrogen-regulated genes but without detectable DNA amplifications. Cancer Res 53: 283-290, 1993.
9. Brunner N, Frandsen TL, Holst-Hansen C, Bei M, Thompson EW, Wakeling AE, Lippman ME and Clarke R: MCF7/LCC2: a 4-hydroxytamoxifen resistant human breast cancer variant which retains sensitivity to the steroidal antiestrogen ICI 182,780. Cancer Res 53: 3229-3232, 1993.
10. Brunner N, Boysen B, Jirus S, Skaar TC, Holst-Hansen C, Lippman J, Frandsen T, Spang-Thomsen M, Fuqua SA and Clarke R: MCF7/LCC9: an antiestrogen resistant MCF-7 variant in which acquired resistance to the steroidal antiestrogen ICI 182,780 confers an early crossresistance to the non-steroidal antiestrogen tamoxifen. Cancer Res 57: 3486-3493, 1997.
11. Gu Z, Lee RY, Skaar TC, Bouker KB, Welch JN, Lu J, Liu A, Zhu Y, Davis N, Leonessa F, Brunner N, Wang Y and Clarke R: Association of interferon regulatory factor-1, nucleophosmin, nuclear factor-kappaB and cyclic AMP response element binding with acquired resistance to faslodex (ICI 182,780). Cancer Res 62: 3428-3437, 2002.
12. Bouker KB, Skaar TC, Fernandez DR, O'Brien KA, Riggins RB, Cao D and Clarke R: Interferon regulatory factor-1 mediates the proapoptotic but not cell cycle arrest effects of the steroidal anti-estrogen ICI 182,780 (Faslodex, Fulvestrant). Cancer Res 264: 4030-4039, 2004.
13. Pratt MA, Bishop TE, White D, Yasvinski G, Menard M, Niu MY and Clarke R: Estrogen withdrawal-induced NF-kappaB activity and bcl-3 expression in breast cancer cells: roles in growth and hormone independence. Mol Cell Biol 23: 6887-6900, 2003.
14. Skaar TC, Prasad SC, Sharareh S, Lippman ME, Brunner N and Clarke R: Two-dimensional gel electrophoresis analyses identify nucleophosmin as an estrogen regulated protein associated with acquired estrogen-independence in human breast cancer cells. J Steroid Biochem Mol Biol 67: 391-402, 1998.
15. Brankin B, Skaar TC, Brotzman M, Trock B and Clarke R: Autoantibodies to numatrin: an early predictor for relapse in breast cancer. Cancer Epidemiol Biomarkers Prev 7: 1109-1115, 1998.
16. Reimold AM, Etkin A, Clauss I, Perkins A, Friend DS, Zhang J, Horton HF, Scott A, Orkin SH, Byrne MC, Grusby MJ and Glimcher LH: An essential role in liver development for transcription factor XBP-1. Genes Dev 14: 152-157, 2000.
17. Riggins RB, Zwart A, Nehra R and Clarke R: The NFκB inhibitor parthenolide restores ICI 182,780 (Faslodex; Fulvestrant)-induced apoptosis in antiestrogen resistant breast cancer cells. Mol Cancer Ther 4: 33-41, 2005.
18. Taniguchi T, Lamphier MS and Tanaka N: IRF-1: the transcription factor linking the interferon response and oncogenesis. Biochim Biophys Acta 1333: M9-M17, 1997.
19. Tanaka N, Ishihara M and Taniguchi T: Suppression of c-myc or fosB-induced cell transformation by the transcription factor IRF-1. Cancer Lett 83: 191-196, 1994.
20. Bowie ML, Dietze EC, Delrow J, Bean GR, Troch MM, Marjoram RJ and Seewaldt VL: Interferon-regulatory factor-1 is critical for tamoxifen-mediated apoptosis in human mammary epithelial cells. Oncogene 23: 8743-8755, 2004.
21. Tanaka N, Ishihara M, Lamphier MS, Nozawa H, Matsuyama T, Mak TW, Aizawa S, Tokino T, Oren M and Taniguchi T: Cooperation of the tumour suppressors IRF-1 and p53 in response to DNA damage. Nature 382: 816-818, 1996.
22. Tamura T, Ishihara M, Lamphier MS, Tanaka N, Oishi I, Aizawa S, Matsuyama T, Mak TW, Taki S and Taniguchi T: An IRF-1-dependent pathway of DNA damage-induced apoptosis in mitogen-activated T lymphocytes. Nature 376: 596-599, 1995.
23. Sanceau J, Hiscott J, Delattre O and Wietzerbin J: IFN-beta induces serine phosphorylation of Stat-1 in Ewing's sarcoma cells and mediates apoptosis via induction of IRF-1 and activation of caspase-7. Oncogene 19: 3372-3383, 2000.
24. Suk K, Chang I, Kim YH, Kim S, Kim JY, Kim H and Lee MS: Interferon gamma (IFNgamma) and tumor necrosis factor alpha synergism in ME-180 cervical cancer cell apoptosis and necrosis. IFNgamma inhibits cytoprotective NF-kappa B through STAT1/IRF-1 pathways. J Biol Chem 276: 13153-13159, 2001.
25. Chow WA, Fang JJ and Yee JK: The IFN regulatory factor family participates in regulation of Fas ligand gene expression in T cells. J Immunol 164: 3512-3518, 2000.
26. Mandlekar S, Hebbar V, Christov K and Kong AN: Pharmacodynamics of tamoxifen and its 4-hydroxy and N-desmethyl metabolites: activation of caspases and induction of apoptosis in rat mammary tumors and in human breast cancer cell lines. Cancer Res 60: 6601-6606, 2000.

27. Doherty GM, Boucher L, Sorenson K and Lowney J: Interferon regulatory factor expression in human breast cancer. Ann Surg 233: 623-629, 2001.
28. Cogswell PC, Guttridge DC, Funkhouser WK and Baldwin AS Jr: Selective activation of NF-kappa B subunits in human breast cancer: potential roles for NF-kappa B2/p52 and for Bcl-3. Oncogene 19: 1123-1131, 2000.
29. Bours V, Bentires-Alj M, Hellin AC, Viatour P, Robe P, Delhalle S, Benoit V and Merville MP: Nuclear factor-kappa B, cancer and apoptosis. Biochem Pharmacol 60: 1085-1089, 2000.
30. Voegel JJ, Heine MJ, Zechel C, Chambon P and Gronemeyer H: TIF2, a 160 kDa transcriptional mediator for the ligand-dependent activation function AF-2 of nuclear receptors. EMBO J 15: 3667-3675, 1996.
31. Nakshatri H, Bhat-Nakshatri P, Martin DA, Goulet RJ and Sledge GW: Constitutive activation of NF-kappaB during progression of breast cancer to hormone-independent growth. Mol Cell Biol 17: 3629-3639, 1997.
32. Osipo C, Gajdos C, Liu H, Chen B and Jordan VC: Paradoxical action of fulvestrant in estradiol-induced regression of tamoxifen-stimulated breast cancer. J Natl Cancer Inst 95: 1597-1608, 1997.
33. Clauss IM, Chu M, Zhao JL and Glimcher LH: The basic domain/leucine zipper protein hXBP-1 preferentially binds to and transactivates CRE-like sequences containing an ACGT core. Nucleic Acids Res 24: 1855-1864, 1996.
34. Clauss IM, Gravallese EM, Darling JM, Shapiro F, Glimcher MJ and Glimcher LH: *In situ* hybridization studies suggest a role for the basic region- leucine zipper protein hXBP-1 in exocrine gland and skeletal development during mouse embryogenesis. Dev Dyn 197: 146-156, 1993.
35. Reimold AM, Iwakoshi NN, Manis J, Vallabhajosyula P, Szomolanyi-Tsuda E, Gravallese EM, Friend D, Grusby MJ, Alt F and Glimcher LH: Plasma cell differentiation requires the transcription factor XBP-1. Nature 412: 300-307, 2001.
36. Cho H, Aronica SM and Katzenellenbogen BS: Regulation of progesterone receptor gene expression in MCF-7 breast cancer cells: a comparison of the effects of cyclic adenosine 3',5'-monophosphate, estradiol, insulin-like growth factor-I and serum factors. Endocrinology 134: 658-664, 1994.
37. West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, Zuzan H, Olson JA Jr, Marks JR and Nevins JR: Predicting the clinical status of human breast cancer by using gene expression profiles. Proc Natl Acad Sci USA 98: 11462-11467, 2001.
39. 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.
39. Fujimoto T, Onda M, Nagai H, Nagahata T, Ogawa K and Emi M: Up-regulation and overexpression of human X-box binding protein 1 (hXBP-1) gene in primary breast cancers. Breast Cancer 10: 301-306, 2003.
40. Feuerstein N, Mond JJ, Kinchington PR, Hickey R, Karjalainen Lindsberg ML, Hay I and Ruyechan WT: Evidence for DNA binding activity of numatrin (B23), a cell cycle-regulated nuclear matrix protein. Biochim Biophys Acta 1087: 127-136, 1990.
41. Inouye CJ and Seto E: Relief of YY1-induced transcriptional repression by protein-protein interaction with the nucleolar phosphoprotein B23. J Biol Chem 269: 6506-6510, 1994.
42. Kondo T: Identification and analysis of IRF-1 binding protein. Hokkaido Igaku Zasshi 71: 509-516, 1996.
43. Beckmann R, Buchner K, Jungblut PR, Eckerskorn C, Weise C, Hilbert R and Hucho F: Nuclear substrates of protein kinase C. Eur J Biochem 210: 45-51, 1992.
44. Peter M, Nakagawa J, Doree M, Labbe JC and Nigg EA: Identification of major nucleolar proteins as candidate mitotic substrates of cdc2 kinase. Cell 60: 791-801, 1990.
45. Feuerstein N and Randazzo PA: *In vivo* and *in vitro* phosphorylation studies of numatrin, a cell cycle regulated nuclear protein, in insulin-stimulated NIH 3T3 HIR cells. Exp Cell Res 194: 289-296, 1991.
46. Chan PK, Liu QR and Durban E. The major phosphorylation site of nucleophosmin (B23) is phosphorylated by a nuclear kinase II. Biochem J 270: 549-552, 1990.
47. Kondo T, Minamino N, Nagamura-Inoue T, Matsumoto M, Taniguchi T and Tanaka N: Identification and characterization of nucleophosmin/B23/numatrin which binds the anti-oncogenic transcription factor IRF-1 and manifests oncogenic activity. Oncogene 15: 1275-1281, 1997.
48. Morris SW, Kirstein MN, Valentine MB, Dittmer K, Shapiro DN, Look AT and Saltman DL: Fusion of a kinase gene, ALK, to a nucleolar protein gene, NPM, in non-Hodgkin's lymphoma. Science 263: 1281-1284, 1994.
49. Redner RL, Rush EA, Faas S, Rudert WA and Corey SJ: The t(5;17) variant of acute promyelocytic leukemia expresses a nucleophosmin-retinoic acid receptor fusion. Blood 87: 882-886, 1996.
50. Bourdeau V, Deschenes J, Metivier R, Nagai Y, Nguyen D, Bretschneider N, Gannon F, White JH and Mader S: Genome-wide identification of high-affinity estrogen response elements in human and mouse. Mol Endocrinol 18: 1411-1427, 2004.
51. Singh B, To L, Ossandon M, Tefft M and Liu A: Representation of the heterogeneity of breast cancer and hormone receptor expression in a tissue microarray. Mod Pathol 13: 225A, 2000.
52. Clarke R, Leonessa F, Welch JN and Skaar TC: Cellular and molecular pharmacology of antiestrogen action and resistance. Pharmacol Rev 53: 25-71, 2001.
53. Elston CW and Ellis IO: Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. Histopathology 19: 403-410, 1991.
54. Atkin NB: Modal deoxyribonucleic acid value and survival in carcinoma of the breast. Br Med J : 271-272, 1972.
55. Fossa SD, Thorud E, Vaage S and Shoaib MC: DNA cytometry of primary breast cancer. Comparison of microspectrophotometry and flow cytometry, and different preparation methods for flow cytometric measurements. Acta Pathol Microbiol Immunol Scand 91A: 235-243, 1983.
56. Schneeweiss A, Sinn HP, Ehemann V, Khbeis T, Neben K, Krause U, Ho AD, Bastert G and Kramer A: Centrosomal aberrations in primary invasive breast cancer are associated with nodal status and hormone receptor expression. Int J Cancer 107: 346-352, 2003.
57. Konecny G, Pauletti G, Pegram M, Untch M, Dandekar S, Aguilar Z, Wilson C, Rong HM, Bauerfeind I, Felber M, Wang HJ, Beryt M, Seshadri R, Hepp H and Slamon DJ: Quantitative association between HER-2/neu and steroid hormone receptors in hormone receptor-positive primary breast cancer. J Natl Cancer Inst 95: 142-153, 2003.
58. Kononen J, Bubendorf L, Kallioniemi A, Barlund M, Schraml P, Leighton S, Torhorst J, Mihatsch MJ, Sauter G and Kallioniemi OP: Tissue microarrays for high-throughput molecular profiling of tumor specimens. Nat Med 4: 844-847, 1998.
59. Moch H, Schraml P, Bubendorf L, Mirlacher M, Kononen J, Gasser T, Mihatsch MJ, Kallioniemi OP and Sauter G: High-throughput tissue microarray analysis to evaluate genes uncovered by cDNA microarray screening in renal cell carcinoma. Am J Pathol 154: 981-986, 1999.
60. Ellis MJ, Coop A, Singh B, Mauriac L, Llombert-Cussac A, Janicke F, Miller WR, Evans DB, Dugan M, Brady C, Quebe-Fehling E and Borgs M: Letrozole is more effective neoadjuvant endocrine therapy than tamoxifen for ErbB-1- and/or ErbB-2-positive, estrogen receptor- positive primary breast cancer: evidence from a phase III randomized trial. J Clin Oncol 19: 3808-3816, 2001.
61. Rhodes A, Jasani B, Anderson E, Dodson AR and Balaton AJ: Evaluation of HER-2/neu immunohistochemical assay sensitivity and scoring on formalin-fixed and paraffin-processed cell lines and breast tumors: a comparative study involving results from laboratories in 21 countries. Am J Clin Pathol 118: 408-417, 2002.
62. Nozawa Y, van Belzen N, van der Made ACJ, Dinjens WNM and Bosman FT: Expression of nucleophosmin/B23 in normal and neoplastic colorectal mucosa. J Pathol 178: 48-52, 1996.
63. Harlow E and Lane D: Antibodies. A Laboratory Manual. CSH, Cold Spring Harbor, 1988.
64. Mayer RJ and Walker JH: Immunohistochemical procedures. In: Immunochemical Methods in Cell and Molecular Biology. Academic Press, London, pp259-281, 1987.
65. Allred DC, Harvey JM, Berardo M and Clark GM: Prognostic and predictive factors in breast cancer by immunohistochemical analysis. Mod Pathol 11: 155-168, 1998.
66. Tarone RE and Chu KC: The greater impact of menopause on ER- than ER+ breast cancer incidence: a possible explanation (United States). Cancer Causes Control 13: 7-14, 2002.
67. Tam WF and Sen R: IkappaB family members function by different mechanisms. J Biol Chem 276: 7701-7704, 2001.

68. Mason DY, Pulford KA, Bischof D, Kuefer MU, Butler LH, Lamant L, Delsol G and Morris SW: Nucleolar localization of the nucleophosmin-anaplastic lymphoma kinase is not required for malignant transformation. Cancer Res 58: 1057-1062, 1998.
69. Hoshiya Y, Gupta V, Kawakubo H, Brachtel E, Carey JL, Sasur L, Scott A, Donahoe PK and Maheswaran S: Mullerian inhibiting substance promotes interferon gamma-induced gene expression and apoptosis in breast cancer cells. J Biol Chem 278: 51703-51712, 2003.
70. Hinneburg A and Keim DA: Optimal grid-clustering: towards breaking the curse of dimensionality in high-dimensional clustering. In: Proceedings of the 25th Conference on Very Large Databases. Atkinson MP, Orlowska ME, Valduriez P, Zdonik SB and Brodie ML (eds). Morgan Kaufman, San Francisco, pp506-517, 1999.
71. Satagopan JM and Panageas KS: A statistical perspective on gene expression data analysis. Stat Med 22: 481-499, 2003.
72. Lee AH, Hong JH and Seo YS: Tumour necrosis factor-alpha and interferon-gamma synergistically activate the RANTES promoter through nuclear factor kappaB and interferon regulatory factor 1 (IRF-1) transcription factors. Biochem J 350: 131-138, 2000.
73. Neish AS, Read MA, Thanos D, Pine R, Maniatis T and Collins T: Endothelial interferon regulatory factor 1 cooperates with NF-kappa B as a transcriptional activator of vascular cell adhesion molecule 1. Mol Cell Biol 15: 2558-2569, 1995.
74. Sanceau, J, Kaisho T, Hirano T and Wietzerbin J: Triggering of the human interleukin-6 gene by interferon-gamma and tumor necrosis factor-alpha in monocytic cells involves cooperation between interferon regulatory factor-1, NF kappa B and Sp1 transcription factors. J Biol Chem 270: 27920-27931, 1995.
75. Luboshits G, Shina S, Kaplan O, Engelberg S, Nass D, Lifshitz-Mercer B, Chaitchik S, Keydar I and Ben-Baruch A: Elevated expression of the CC chemokine regulated on activation, normal T cell expressed and secreted (RANTES) in advanced breast carcinoma. Cancer Res 59: 4681-4687, 1999.
76. Byrne GJ, Ghellal A, Iddon J, Blann AD, Venizelos V, Kumar S, Howell A and Bundred NJ: Serum soluble vascular cell adhesion molecule-1: role as a surrogate marker of angiogenesis. J Natl Cancer Inst 92: 1329-1336, 2000.
77. Conze D, Weiss L, Regen PS, Bhushan A, Weaver D, Johnson P and Rincon M: Autocrine production of interleukin 6 causes multidrug resistance in breast cancer cells. Cancer Res 61: 8851-8858, 2001.