[CANCER RESEARCH 61, 5979–5984, August 15, 2001]

Advances in Brief

Estrogen Receptor Status in Breast Cancer Is Associated with Remarkably Distinct Gene Expression Patterns[1]

Sofia Gruvberger, Markus Ringnér, Yidong Chen, Sujatha Panavally, Lao H. Saal, Åke Borg, Mårten Fernö, Carsten Peterson, and Paul S. Meltzer[2]

Department of Oncology [S. G., Å. B., M. F.] and Complex Systems Division, Department of Theoretical Physics [M. R., C. P.], Lund University, SE-221 00 Lund, Sweden, and Cancer Genetics Branch, National Human Genome Research Institute [S. G., M. R., Y. C., S. P., L. H. S., P. S. M.], NIH, Bethesda, Maryland 20892

Abstract

To investigate the phenotype associated with estrogen receptor α (ER) expression in breast carcinoma, gene expression profiles of 58 node-negative breast carcinomas discordant for ER status were determined using DNA microarray technology. Using artificial neural networks as well as standard hierarchical clustering techniques, the tumors could be classified according to ER status, and a list of genes which discriminate tumors according to ER status was generated. The artificial neural networks could accurately predict ER status even when excluding top discriminator genes, including ER itself. By reference to the serial analysis of gene expression database, we found that only a small proportion of the 100 most important ER discriminator genes were also regulated by estradiol in MCF-7 cells. The results provide evidence that ER+ and ER− tumors display remarkably different gene-expression phenotypes not solely explained by differences in estrogen responsiveness.

Introduction

Estrogens are important regulators of growth and differentiation in the normal mammary gland and are also important in the development and progression of breast carcinoma. Estrogens regulate gene expression via ER,[3] however the details of the estrogen effect on downstream gene targets, the role of cofactors, and cross-talk between other signaling pathways are far from fully understood. As approximately two-thirds of all breast cancers are ER+ at the time of diagnosis, the expression of the receptor has important implications for their biology and therapy (1). Opinions differ as to whether those breast cancers which lack ER expression at diagnosis arise from an ER− compartment within the mammary epithelium or represent evolution from an ER+ to an ER− state (2).

The cDNA microarray technology allows for parallel analysis of the expression of thousands of genes (3) to address complex questions in tumor biology. Statistical tools are required to analyze the large amount of expression data generated by this methodology. ANNs are computer-based algorithms for pattern recognition that are capable of learning from experience (4). The diagnosis of myocardial infarcts (5) and heart arrhythmias from electrocardiograms (6) are examples of applications of ANNs in medicine. We have recently demonstrated the utility of ANNs for the diagnostic classification of tumors using cDNA microarray data (7). In this study, we have applied ANNs as well as conventional methods to analyze cDNA microarray data from a selected group of node-negative breast cancers that differ with respect to their ER status. Here we report that ER+ and ER− tumors display remarkably different phenotypes, which may be attributable to their evolution from distinct cell lineages.

Materials and Methods

Tissues and Cells. Fifty-eight grossly dissected primary tumors from node-negative breast cancer patients, tumor size 20–50 mm, were collected at the University Hospital, Lund, Sweden. Microscopic examination of touch preparations verified the presence of cancer cells in all samples. To train the classifier described below, 47 tumors, all from two previous randomized studies (Ref. 8)[4] were selected so that roughly half, 23, were ER+ (range, 50–1900 fmol/mg protein; median, 160), whereas the remaining 24 were ER− (range, 0–9 fmol/mg protein, median 0.7). In addition, 14 of the patients were premenopausal (5 ER+ and 9 ER−) and 33 were postmenopausal (18 ER+ and 15 ER−). To obtain an independent test set, the remaining 11 of the 58 tumors were selected from an ongoing clinical trial and used here as a blinded test set. Of the 11 blinded samples, 5 were ER+ (range, 40–120 fmol/mg protein; median, 60), 6 were ER− (range, 0–3 fmol/mg protein; median, 1.5), and all were premenopausal. ER protein determinations were performed using standard methods in the routine clinical laboratory (9). BT-474 cells, obtained from American Type Culture Collection, were maintained in RPMI 1640 supplemented by 10% fetal bovine serum, penicillin, and streptomycin. Cells were harvested at 60–80% confluency and used as a reference in all hybridizations.

RNA Isolation and cDNA Microarrays. Total RNA was isolated from cell lines using the RNeasy kit (Qiagen, Valencia, CA) with subsequent Trizol (Life Technologies, Inc., Rockville, MD) purification. Total RNA from tumors was isolated using two successive rounds of Trizol. Microarrays were prepared and hybridized as described previously (3, 10, 11) and according to standard protocols.[5] Briefly, the arrays were spotted with 6,728 sequence-verified cDNA clones, of which ~4000 were named human genes and the remaining clones were expressed sequence tags. BT-474 RNA (200 μg) and 65–100 μg of tumor RNA were used to produce labeled cDNA by anchored oligo(dT)-primed reverse transcription using SuperScript II reverse transcriptase (Life Technologies, Inc.) in the presence of either Cy5-dUTP or Cy3-dUTP (Amersham Pharmacia, Piscataway, NJ), respectively. Fluorescence scanning and image analysis with DeArray software were performed as described previously (12, 13).

Data Analysis. For each gene, the fluorescent intensity of the most intense channel [red (Cy3) or green (Cy5)] for each sample, was averaged over all samples. All genes for which this average exceeded 2,000 fluorescence units (scale 0–65,535 units) were included in the analysis. In addition, we required, for all samples, that the red and green intensities both exceeded 20 fluorescence units and that the union (of the two channels) spot area exceeded 30 pixels. For the 58 (47 + 11) measured samples, these requirements left us with

Received 4/26/01; accepted 6/25/01.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

[1] Supported in part by the Swedish Research Council and the Knut and Alice Wallenberg Foundation through the SWEGENE consortium (to M. R.) and the Swedish Foundation for Strategic Research (to C. P.). This work was partly supported by grants from the Lund University Medical Faculty, the Swedish Cancer Society, Berta Kamprad's Foundation, the Gunnar Arvid and Elisabeth Nilsson Foundation, the Hospital of Lund Foundations, the E and F Bergqvist Foundation, and King Gustav V 's Jubilee Foundation.

[2] To whom requests for reprints should be addressed, at National Human Genome Research Institute, NIH, 49 Convent Drive, Bethesda, MD 20892-4470. Phone: (301) 594-5283; Fax: (301) 402-3281; E-mail: pmeltzer@nhgri.nih.gov.

[3] The abbreviations used are: ER, estrogen receptor α; ANN, artificial neural network; E2, estradiol; PCA, principal component analysis; ROC, receiver operating characteristic; MDS, multidimensional scaling; WGA, weighted gene analysis; SAGE, serial analysis of gene expression; GATA3, GATA-binding protein; 3 TFF3, trefoil factor 3.

[4] Å. Borg, M. Fernö, unpublished results.

[5] Internet address: http://www.nhgri.nih.gov/DIR/LCG/15K/HTML/protocol.html.

3,389 of the original 6,728 genes. We used multilayered perceptrons, a class of ANNs, which are powerful and versatile regression models (4) to predict the ER status of the tumors from their gene expression patterns and to determine the genes which were most important for this classification (Fig. 1*A*). To allow for a supervised regression model with no "over-training" (because we have a large number of genes as compared with the number of samples), the dimensionality (3,389) of the samples was reduced by the PCA (14). Thus, each sample was represented by 58 numbers, which resulted from a projection of the gene expressions using PCA eigenvectors. The samples were classified in two categories using a 3-fold cross-validation procedure, as follows. The 47 disclosed samples were randomly shuffled and split into three roughly equally sized groups. An ANN model was then calibrated with 8 or 10 PCA components as input variables using two of the groups (training), with the third group reserved for testing predictions (validation). This procedure was repeated three times, each time with a different group used for validation. The random shuffling was redone 200 times, and for each shuffling we analyzed three ANN models. Thus, in total, each disclosed sample belonged to a validation set 200 times, and 600 ANN models were calibrated. We selected the PCA components used as inputs based on the training set. For the ER− and ER+ classification, each ANN model gave an output between 0 (ER−) and 1 (ER+). For each validation sample, the 200 outputs were used as a committee: the average of all of the outputs (a committee vote) was calculated, and a validation sample was classified as ER− or ER+, depending on whether its committee vote was closer to 0 or 1 (the decision threshold was 0.5). All 600 models were used to classify the additional blinded samples. Different choices of the decision threshold correspond to different balances between the sensitivity and the specificity of the classification. All possible thresholds give rise to a so-called ROC curve in the (sensitivity, 1 − specificity)-plane. The area under this curve (ROC area) is a convenient measure of the classification performance. The sensitivity of the classification to individual genes was determined by the absolute value of the partial derivative of the output with respect to the gene expressions, averaged over samples and ANN models. A

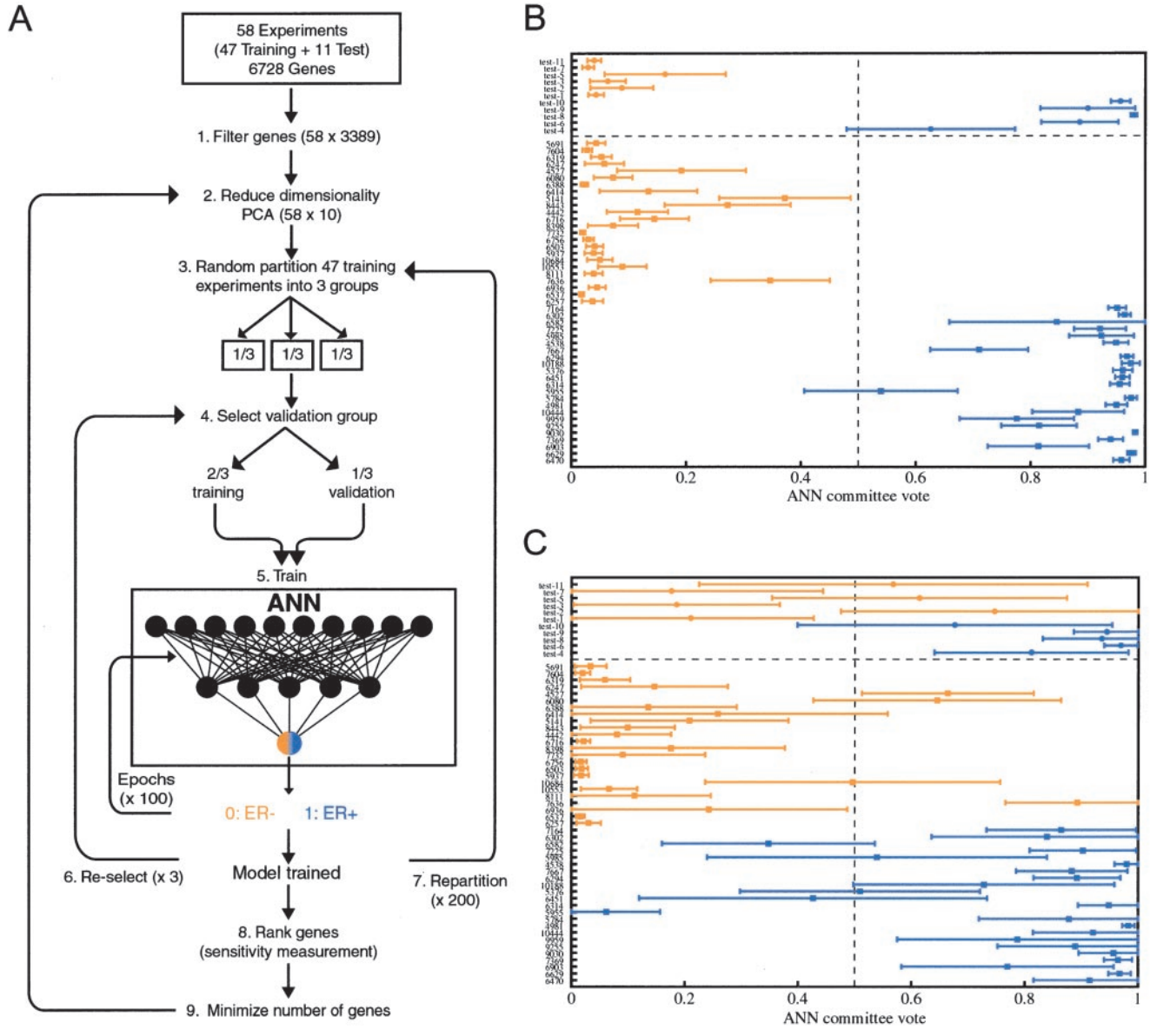

Fig. 1. Classification of ER+ and ER− tumors using ANNs and gene expression patterns. *A*, schematic illustration of the ANN models. Filtering of genes for a minimal level of expression and spot area reduced the number of genes to 3389 (1). PCA further reduced the dimensionality to 10 PCA components/sample (2). The samples were randomly partitioned into three groups (3). Two of these groups were used for training and one for validation (4) using an ANN model with five hidden nodes and the ER status as the output (5). The training process was repeated so that all three groups were used for validation (6). The random partition of samples was redone and the training procedure repeated 200 times so that a total of 600 models were trained (7). By measuring the sensitivity of the classification to a change in the expression level of each gene in the calibrated models, the genes could be ranked (8). The process (2–7) was repeated using fewer genes (9). The ANN output results from the committee of 600 models, including the SDs, using the top 100 (*B*) and the top 301–400 (*C*) discriminator genes. On the *X* axis, 0 and 1 represent ER− and ER+, respectively, with the decision threshold set at 0.5. The 47 training samples (■) and the 11 test samples (●) are plotted. *Yellow*, samples known to be ER− (23 + 6); *blue*, ER+ samples (24 + 5). Sample numbers (*labels*) are indicated.

large sensitivity for a gene implies that changing its expression influences the output significantly. In this way, the genes can be ranked. For comparison with the ANN method, we also analyzed the data and visualized the differences between tumors based on ER status using MDS (10), hierarchical clustering (15), and weighted gene list (16) techniques. To test whether genes which discriminate ER+ from ER− tumors demonstrated a response to E2 in MCF-7 cells (17) at either 3 h or 24 h after E2 treatment, we searched the SAGE database[6] using the xProfiler tool with default settings.

Results

Calibration and Validation of the ANN Models. To calibrate ANN models to classify the tumor samples, we used the gene expression data from cDNA microarrays containing 6728 genes (Fig. 1*A*). Filtering for a minimal level of expression and spot area reduced the number of genes to 3389. PCA further reduced the dimensionality, and we found that using 10 PCA components/sample as inputs, with one output and five hidden nodes, produced well-calibrated ANN models. The 3-fold cross-validation procedure (see "Materials and Methods") produced a total of 600 ANN models, and the training and validation was successful. In addition, inspection of the calibration curves showed that there was no sign of overtraining of the models (data not shown).

Optimization of Genes Used for Classification Using ANN Models. We next determined the contribution of each gene to the classification by the ANN models. This was done by measuring the sensitivity of the classification to a change in the expression level of each gene, using the 600 previously calibrated models (see "Materials and Methods"). In this way, we ranked the genes according to their significance for the classification. The 100 most important genes were then extracted and formed the input for another and final calibration. When using only 100 genes, we found that using eight PCA component inputs and four hidden nodes were sufficient. In this way, all 47 samples were correctly classified in the validation phase. The output of the models generated a number between 0 (ER−) and 1 (ER+), reflecting the crispness of the classification. A plot of the output values from the committee is shown for all 47 training samples (Fig. 1*B*). The majority of the samples, in both groups, obtain output values close to either 0 or 1, with small variations between the output results from the different models. Thus, the committee members agree in general on the classification of each sample, and the result is a clear separation between ER+ and ER− tumors. The top 50 genes extracted from the ANN models, which significantly contribute to the classification, represent a wide spectrum of cellular functions (Table 1). The *ER* gene, which, as expected, appears at the top of the ranked genes, is closely followed *by GATA-binding protein 3*, a transcription factor previously associated with ER+ tumors (18).

Prediction of ER Status of Blinded Breast Tumor Samples. To test whether the ANN classifications for ER status were generally applicable, the calibrated ANN models were also used to predict the outcome of 11 (5 ER+ and 6 ER−) blinded test samples from an independent data set of early stage primary breast cancers. These samples were also predicted with 100% accuracy. A plot of the committee output values, shown in Fig. 1*B*, displays a clear separation between the two categories.

Prediction of ER Status when Excluding ER and Other Top Discriminators. When classifying the samples using the *ER* gene or the *GATA3* gene alone (without PCA), good classifications were obtained, indicating that, as expected, these genes carry sufficient information for successful classifications. An interesting issue is to what extent ER+ and ER− tumors can be separated when not explicitly including the expression values for ER itself. Repeating the ANN cross-validation and gene extraction procedure above, but excluding ER, only 1 ER+ sample, 6582, of the 47 samples was incorrectly classified. Interestingly, when using this calibration while masking the data for ER, again all 11 blinded test samples were correctly classified. Thus, a successful prediction does not occur for the trivial reason that ER mRNA expression is related to ER protein levels. These results led us to examine how far down on the discriminator list we could find genes carrying enough information for an accurate prediction. To test this we performed a series of classifications using different sets of 100 genes, starting from the top of the discriminator list by excluding the top 50 genes and following this by the stepwise exclusion of 50 additional genes for every classification (*i.e.*, excluding the top 50, 100, 150 . . . to 300 genes, respectively). The number of correctly classified samples and the ROC area for the predictions of both the 47 tumors in the validation set as well as for the 11 blinded test tumors were extracted (Table 2). Although the success of the predictions declined when using genes lower down on the discriminator list, the network performance was still fairly good. This was demonstrated as the 100 genes in positions 301 to 400 on the discriminator list achieved ROC areas of 93.7% and 96.7% for the validation set and the test set, respectively. However, the committee votes for these samples are now closer to the threshold value 0.5 and also display an increased variance (Fig. 1*C*), indicating that the classification is less stable and conclusive than when using the top 100 genes. Still, the results clearly demonstrate that the classification is not only controlled by a few very strong discriminator genes, but results from a far more complex expression pattern involving a substantial number of genes.

MDS and Hierarchical Clustering. We used two standard clustering techniques to illustrate further the differences, found by the ANN models, in gene expression profiles between the two tumor categories. An MDS plot was created displaying the position of each tumor sample in a three-dimensional Euclidean space, with the distance between the samples reflecting their approximate degree of correlation (Fig. 2*A*). This MDS clustering was based on a WGA (16) that generated a set of 113 genes that showed significantly differentiated expression levels between the two tumor categories. There was an ~50% overlap between the 100 most important discriminatory genes derived from WGA analysis and the ANN models, indicating that a substantial number of important discriminatory genes are revealed independent of the choice of analytical method. As can be seen from the MDS plot, the two categories (ER+ and ER−) are well separated, with the exception of one ER+ tumor, 6582, which clusters with the ER− tumors. This separation, consistent with the ANN analysis above, was confirmed additionally by hierarchical clustering of the 47 tumors and the 113 genes from the WGA (Fig. 2*B*), which organized the 47 tumor samples along the horizontal axis and created a dendrogram based on their similarities in gene expression profiles. The clustering organized all of the tumors into two separate dendrogram branches corresponding to ER+ and ER− tumors with the exception of two ER+ tumors, 6582 and 5955.

E2-Responsive Genes in the MCF-7 Cell Line. To examine the relationship between ER function and the genes discriminating ER+ and ER− tumors, we compared our results with SAGE gene expression data reported for E2 stimulated MCF-7 cells[7] (17). By reference to this data, 61 of the top 100 ER-discriminating genes uncovered by ANN analysis were represented by SAGE tags. Of these, only four genes (*CCND1*, *STC2*, *SLC7A5*, and *KRT18*) were regulated by E2 in MCF-7 cells, and one of these (*KRT18*, ranked 59 in the ANN sensitivity list) was regulated in a direction discordant with its relative

[6] Internet address: http://www.ncbi.nlm.nih.gov/SAGE/sagexpsetup.cgi.

[7] Internet address: http://sciencepark.mdanderson.org/ggeg.

Table 1 *Top 50 genes extracted from ANNs*

Rank[a]	Relative expression[b]	Gene symbol	Gene description	SAGE[c]	Clone ID no.
1^{+0}_{-0}	+	*ESR1*	Estrogen receptor 1	NS	725321
2^{+2}_{-0}	+	*TFF3*	Trefoil factor 3 (intestinal)	NT	298417
3^{+1}_{-1}	+	*GATA3*	GATA-binding protein 3	NS	214068
4^{+0}_{-2}	+		ESTs	NS	132140
5^{+4}_{-0}	−	*S100A8*	S100 calcium-binding protein A8 (calgranulin A)	NT	562729
6^{+4}_{-1}	−	*LCN2*	Lipocalin 2 (oncogene 24p3)	NS	741497
7^{+3}_{-2}	+		ESTs	NT	155072
8^{+2}_{-3}	−	*CDH3*	Cadherin 3, type 1, P-cadherin (placental)	NT	773301
9^{+2}_{-4}	+	*P28*	Dynein, axonemal, light intermediate polypeptide	NT	782688
10^{+3}_{-4}	−	*PFKP*	Phosphofructokinase, platelet	NS	26184
11^{+5}_{-2}	−	*LAD1*	Ladinin 1	NS	121551
12^{+6}_{-2}	−	*KCNN4*	Potassium intermediate/small conductance calcium-activated channel, subfamily N, member 4	NS	756708
13^{+8}_{-3}	+	*SFRS5*	Splicing factor, arginine/serine-rich 5	NT	143169
14^{+8}_{-3}	+	*IGFBP2*	Insulin-like growth factor binding protein 2 (36kD)	NS	233721
15^{+7}_{-4}	+	*HSPC195*	ESTs	NS	139354
16^{+6}_{-5}	+	*COX6C*	Cytochrome c oxidase subunit VIc	NS	838568
17^{+7}_{-5}	−	*PFKP*	Phosphofructokinase, platelet	NS	950682
18^{+7}_{-6}	+	*FBP1*	Fructose-1,6-bisphosphatase 1	NS	433253
19^{+7}_{-6}	−	*EGFR*	Epidermal growth factor receptor (avian erythroblastic leukemia viral (v-erb-b) oncogene homolog)	NS	324861
20^{+6}_{-7}	+	*CRIP1*	Cysteine-rich protein 1 (intestinal)	NS	1323448
21^{+6}_{-8}	+	*STC2*	Stanniocalcin 2	+	130057
22^{+5}_{-8}	−	*SCYD1*	Small inducible cytokine subfamily D (Cys-X3-Cys), member 1 (fractalkine, neurotactin)	NT	140574
23^{+6}_{-6}	−	*GALNT3*	UDP-N-acetyl-α-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 3 (GalNAc-T3)	NT	148225
24^{+5}_{-7}	−	*KRT7*	Keratin 7	NS	843321
25^{+5}_{-7}	−	*LMO4*	LIM domain only 4	NS	162533
26^{+7}_{-7}	+	*TIMP3*	Tissue inhibitor of metalloproteinase 3 (Sorsby fundus dystrophy, pseudoinflammatory)	NS	768370
27^{+6}_{-6}	+	*CCND1*	Cyclin D1 (PRAD1: parathyroid adenomatosis 1)	+	841641
28^{+13}_{-5}	−	*ATDC*	Ataxia-telangiectasia group D-associated protein	NT	377275
29^{+12}_{-6}	−	*ABP/ZF*	Alu-binding protein with zinc finger domain	NT	417424
30^{+14}_{-5}	+	*IGF2*	Insulin-like growth factor 2 (somatomedin A)	NT	207274
31^{+13}_{-5}	+	*IGHG3*	Immunoglobulin heavy constant γ 3 (Gm marker)	NT	855745
32^{+12}_{-6}	+	*PIB5PA*	Phosphatidylinositol (4,5) bisphosphate 5-phosphatase, A	NT	1359579
33^{+13}_{-7}	+	*ARHH*	*Ras* homolog gene family, member H	NS	302591
34^{+15}_{-6}	−	*MSE55*	Serum constituent protein	NS	214982
35^{+14}_{-7}	+	*SLC9A3R1*	Solute carrier family 9 (sodium/hydrogen exchanger), isoform 3 regulatory factor 1	NS	773286
36^{+13}_{-8}	+	*EIF3S4*	Eukaryotic translation initiation factor 3, subunit 4 (δ, 44kD)	NS	857319
37^{+12}_{-9}	−	*SOD3*	Superoxide dismutase 3, extracellular	NT	795309
38^{+11}_{-10}	−	*SLC7A5*	Solute carrier family 7 (cationic amino acid transporter, y+ system), member 5	−	755578
39^{+10}_{-11}	−	*NDRG1*	*N*-myc downstream regulated	NT	842863
40^{+13}_{-12}	−	*S100P*	S100 calcium-binding protein P	NS	135221
41^{+14}_{-13}	+	*THBS1*	Thrombospondin 1	NT	810512
42^{+14}_{-12}	−	*IMPA2*	Inositol(myo)-1(or 4)-monophosphatase 2	NS	32299
43^{+13}_{-13}	−	*CHI3L1*	Chitinase 3-like 1 (cartilage glycoprotein-39)	NT	770212
44^{+13}_{-14}	+	*SERPINI1*	Serine (or cysteine) proteinase inhibitor, clade I (neuroserpin), member 1	NT	564621
45^{+14}_{-12}	−	*HMGIY*	High-mobility group (nonhistone chromosomal) protein isoforms I and Y	NS	782811
46^{+14}_{-13}	+	*APOD*	Apolipoprotein D	NS	838611
47^{+16}_{-13}	+	*PVALB*	Parvalbumin	NT	430318
48^{+16}_{-14}	+	*A2M*	α-2-macroglobulin	NT	878182
49^{+16}_{-15}	+	*SELENBP1*	Selenium binding protein 1	NS	80338
50^{+27}_{-10}	−	*FABP7*	Fatty acid binding protein 7, brain	NS	345626

[a] The genes are ranked according to the sensitivity analysis (see "Materials and Methods"). The sensitivity is calculated averaged over all 600 ANN models. The errors for a rank are calculated from the SD of the sensitivities for the 600 ANN models.

[b] +, higher expression in ER+; −, higher expression in ER−. Defined as the sign of the ANN sensitivity.

[c] NT, no tag; NS, not significant; +, up-regulated by E2; −, down-regulated by E2.

expression in the tumor specimens (Table 1). This result suggests that the difference in gene expression profile between ER− and ER+ tumors can only in part be explained by the activity of a functional ER pathway in ER+ tumors.

Discussion

To characterize in more detail the phenotypic characteristics of ER+ and ER− breast cancers, the expression of 6728 genes was investigated in primary tumor tissues from 58 breast cancer patients. Gene expression profiles of breast tumors have been investigated previously (19–21) but not in a tumor set suited to specifically study the ER+ and ER− classification problem. Here we have identified a homogenous group of node-negative breast cancers, 20–50 mm in diameter, of which about one-half (28) were ER+, whereas the remaining 30 were ER−. We then classified the samples using ANN models. Compared with the majority of other methods used, our approach has the advantage that it takes nonlinear dependencies in the data into account. In addition, because we were interested in classifying two well-known cancer types, rather than discovering new classes, a supervised approach was optimal. The ANN models were successfully trained to recognize gene expression patterns generated by cDNA microarray analysis, inasmuch as they accurately classified both the 47 training samples and the 11 blinded test samples (Fig. 1*B*) using only 100 of the genes most important for the classification. Of note, the microarray analyses of the blinded test samples were performed separately using a different scanner and batch of microarray

Table 2 *Prediction of ER status*

	Validation		Test	
Genes	Correct[a]	ROC area	Correct[b]	ROC area
Top–100	47	100.00%	11	100.00%
51–150	43	97.80%	9	100.00%
101–200	45	99.30%	11	100.00%
151–250	44	97.50%	9	100.00%
201–300	41	93.70%	11	100.00%
251–350	39	95.30%	9	93.30%
301–400	41	93.10%	8	96.70%
Random	38.8 ± 0.2	91.8 ± 0.2%	5.5 ± 0.2	53.0 ± 2.6%

[a] Number of correct classifications of 47 samples.
[b] Number of correct classifications of 11 samples.

slides from those used for training, indicating the robustness of the methods involved, with regard to both measurements and analysis. Standard clustering algorithms such as MDS and hierarchical clustering showed similar results (Fig. 2), strengthening the conclusion based on ANN models that ER+ and ER− tumors exhibit distinct patterns of gene expression. We achieved our best classification results using the ANN models, demonstrating the potential to obtain improved results with nonlinear classification methods. This is also important for the extraction of relevant genes, because a nonlinear method may extract important genes that cannot be found by linear methods. However, this relatively small data set does not allow for a rigorous comparison of methods.

In addition to the accurate classification of disclosed as well as blinded samples based on the top 100 discriminatory genes, we found that a fairly good classification could be accomplished using lower-ranked genes. Although the reliability of the classification declined when using genes farther down on the list, the results indicate that information that contributes to the classification is carried by genes deep on this list. This is consistent with the gene expression profiles of ER+ and ER− tumors differing in a complex way, indicating the existence of two phenotypically very distinct groups of tumors. The ER status of breast tumors has been suggested to either reflect tumor progression with ER− tumors evolving from ER+ precursors, or to indicate a distinct origin from different types of epithelial cells in the mammary gland. Metastases from ER+ tumors may be ER− (22) supporting the former view. On the other hand, ER+ tumors have been suggested to exhibit the phenotype of luminal epithelial cells, whereas ER− tumors resemble myoepithelial (basal) cells (19). Recently, it has been proposed that myoepithelial cells derive from self-renewing luminal cell precursors, an observation which might explain the predominant luminal phenotype of breast cancers (23). Several of the ER status-discriminator genes are relevant to mammary gland histology. For example, we found that P-cadherin, characteristic of myoepithelial cells (24), was more highly expressed in ER− tumors. The correlation between the expression of P-cadherin and ER-negativity in tumors has been observed previously (25). The transcription factor C/EBP β, which has been suggested to control the cell-fate decision in the mammary gland (26), is more highly expressed in ER− tumors. Of interest, C/EPB β-null mice have a defect in lobuloalveolar development and an abnormally high proportion of cells expressing the progesterone receptor (27). We also identified lipocalin 2 as a gene associated with ER− tumors, consistent with a previous report (28). Another gene expressed more highly in ER− tumors, *ladinin*, though not previously studied in breast cancer, is a basement-membrane protein that may well be associated with the basal/myoepithelial compartment (29). Perou *et al.* (19) emphasized varying patterns of cytokeratin expression in breast cancer, and our results are consistent with that report, to the extent that the arrays used in these studies overlapped.

Several genes previously associated with ER positivity or a ductal/

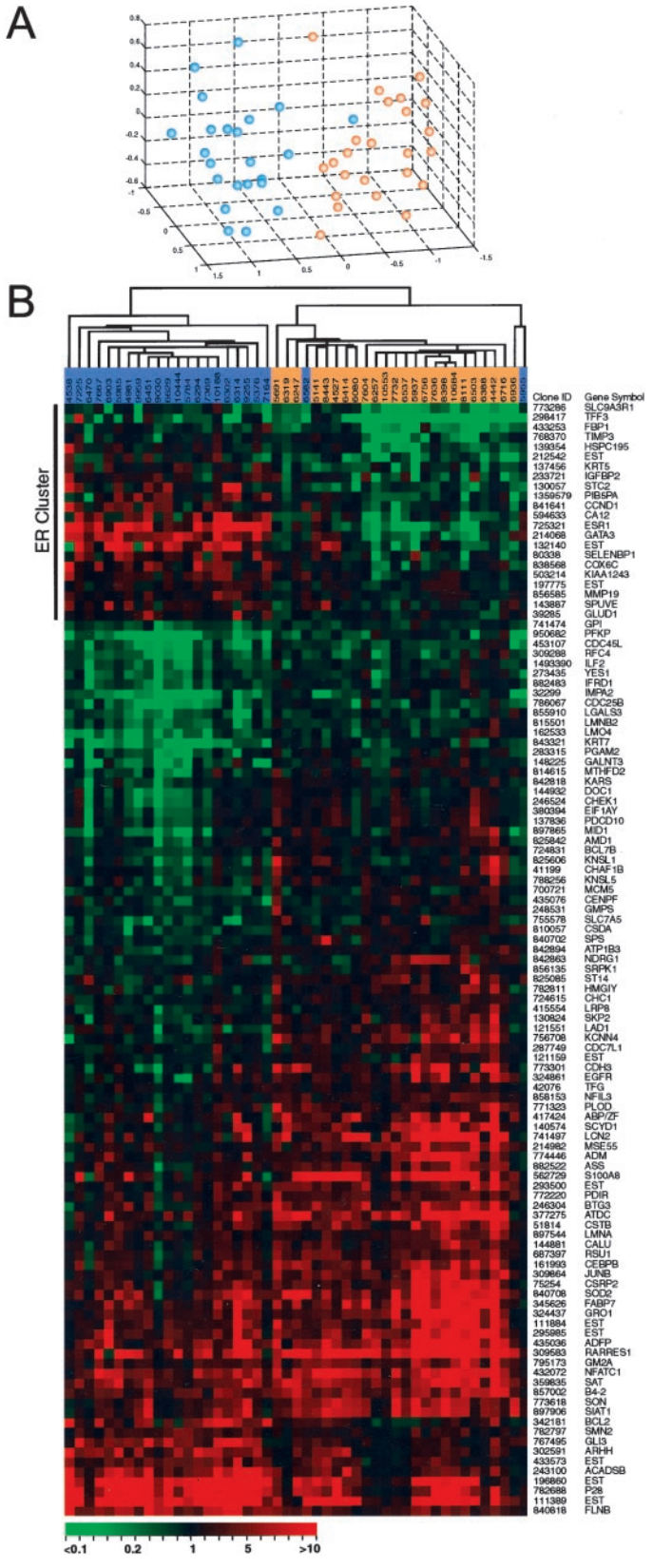

Fig. 2. Clustering of gene expression data from ER+ and ER− tumors. *Blue*, ER+ tumors; *yellow*, ER−. MDS (*A*) displays the position of each tumor sample in a three-dimensional Euclidean space, with the distance between the samples reflecting their approximate degree of correlation. Hierarchical clustering (*B*) presents the clustered samples in columns and the clustered genes in rows. A pseudocolored representation of gene expression ratios is shown, with the scale below. The genes in the cluster, which included the *ER* gene, are denoted. The 113 genes used for the two clustering methods were generated by WGA.

luminal localization were also identified as more highly expressed in this group of tumors. Among these were not only *GATA3*, but also *TFF3*, belonging to the same family of trefoil factors as *pS2*, a gene whose expression is regulated by *ER*. Although *TFF3* was not present in the SAGE data, its induction by estrogen has been reported previously in MCF-7 cells (30). *Cyclin D1*, a gene that is strongly associated with ER expression in breast cancer in this and other studies (31), is strongly induced by E2 in MCF7. Carbonic anhydrase XII has very recently been localized to the ductal epithelium where it may promote tumor invasion by modifying the extracellular pH (32). It is striking, though, that only a few genes on our discriminator list are E2-responsive in cell culture. This observation is consistent with the unique patterns of gene expression being largely explained on the basis of cell lineage, with a component of the ER+ pattern resulting from the function of an ER signaling pathway. In addition, the *in vitro* response of a single cell line to E2 may not faithfully reproduce the physiological effects of ER signaling *in vivo*, and the role of genes regulated by the progesterone receptor remains to be explored.

In conclusion, we have found that ER+ and ER− tumors display very different gene expression phenotypes. From examining expression patterns alone, we cannot establish whether the ER+ and ER− phenotypes reflect tumorigenesis from populations which diverged during normal differentiation or represent a phenotypic interconversion during tumorigenesis. Notably, only a small proportion of cells in the normal mammary epithelium express ER (33), in sharp contrast to the high proportion of ER+ tumors. The underlying biology of the mammary epithelium is complex and the distinct cellular compartments, which give rise to cancers, are not fully defined. The mechanisms, which regulate these distinct gene expression programs, remain to be investigated, and are of importance for future breast cancer research.

Acknowledgments

We are indebted to participating departments of the South Sweden Breast Cancer Group for providing us with breast cancer samples and to Mattias Ohlsson for valuable discussions regarding the properties of ANNs.

References

1. Osborne, C. K. Steroid hormone receptors in breast cancer management. Breast Cancer Res. Treat., *51:* 227–238, 1998.
2. Parl, F. F. Estrogens, Estrogen Receptor, and Breast Cancer. Amsterdam: IOS Press, 2000.
3. Shalon, D., Smith, S. J., and Brown, P. O. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. Genome Res., *6:* 639–645, 1996.
4. Bishop, C. M. Neural Networks for Pattern Recognition. Oxford: Clarendon Press, 1995.
5. Heden, B., Ohlin, H., Rittner, R., and Edenbrandt, L. Acute myocardial infarction detected in the 12-lead ECG by artificial neural networks. Circulation, *96:* 1798–1802, 1997.
6. Silipo, R., Gori, M., Taddei, A., Varanini, M., and Marchesi, C. Classification of arrhythmic events in ambulatory electrocardiogram using artificial neural networks. Comput. Biomed. Res., *28:* 305–318, 1995.
7. Khan, J., Wei, J., Ringnér, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C., and Meltzer, P. S. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat. Med., *7:* 673–679, 2001.
8. Swedish Breast Cancer Cooperative Group. Randomized trial of two *versus* five years of adjuvant tamoxifen for postmenopausal early stage breast cancer. J. Natl. Cancer Inst., *88:* 1543–1549, 1996.
9. Fernö, M., Stål, O., Baldetorp, B., Hatschek, T., Källström, A. C., Malmström, P., Nordenskjöld, B., and Rydén, S. Results of two or five years of adjuvant tamoxifen correlated to steroid receptor and S-phase levels. Breast Cancer Res. Treat., *59:* 69–76, 2000.
10. Khan, J., Simon, R., Bittner, M., Chen, Y., Leighton, S. B., Pohida, T., Smith, P. D., Jiang, Y., Gooden, G. C., Trent, J. M., and Meltzer, P. S. Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. Cancer Res., *58:* 5009–5013, 1998.
11. DeRisi, J., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., Chen, Y., Su, Y. A., and Trent, J. M. Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat. Genet., *14:* 457–460, 1996.
12. Khan, J., Bittner, M. L., Chen, Y., Meltzer, P. S., and Trent, J. M. DNA microarray technology: the anticipated impact on the study of human disease. Biochim. Biophys. Acta, *1423:* M17–M28, 1999.
13. Chen, Y., Dougherty, E. R., and Bittner, M. L. Ratio-based decisions and the quantitative analysis of cDNA microarray images. Biomed. Optics, *2:* 364–374, 1997.
14. Jollife, I. T. Principal Component Analysis. New York: Springer-Verlag New York, Inc., 1986.
15. Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA, *95:* 14863–14868, 1998.
16. Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, 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., and Sondak, V. Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature (Lond.)., *406:* 536–540, 2000.
17. Charpentier, A. H., Bednarek, A. K., Daniel, R. L., Hawkins, K. A., Laflin, K. J., Gaddis, S., MacLeod, M. C., and Aldaz, C. M. Effects of estrogen on global gene expression: identification of novel targets of estrogen action. Cancer Res., *60:* 5977–5983, 2000.
18. Yang, G. P., Ross, D. T., Kuang, W. W., Brown, P. O., and Weigel, R. J. Combining SSH and cDNA microarrays for rapid identification of differentially expressed genes. Nucleic Acids Res., *27:* 1517–1523, 1999.
19. Perou, C. M., Sorlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S. X., Lonning, P. E., Borresen-Dale, A. L., Brown, P. O., and Botstein, D. Molecular portraits of human breast tumours. Nature (Lond.), *406:* 747–752, 2000.
20. Martin, K. J., Kritzman, B. M., Price, L. M., Koh, B., Kwan, C. P., Zhang, X., Mackay, A., O'Hare, M. J., Kaelin, C. M., Mutter, G. L., Pardee, A. B., and Sager, R. Linking gene expression patterns to therapeutic groups in breast cancer. Cancer Res., *60:* 2232–2238, 2000.
21. Bertucci, F., Houlgatte, R., Benziane, A., Granjeaud, S., Adelaide, J., Tagett, R., Loriod, B., Jacquemier, J., Viens, P., Jordan, B., Birnbaum, D., and Nguyen, C. Gene expression profiling of primary breast carcinomas using arrays of candidate genes. Hum. Mol. Genet., *9:* 2981–2991, 2000.
22. Kuukasjarvi, T., Kononen, J., Helin, H., Holli, K., and Isola, J. Loss of estrogen receptor in recurrent breast cancer is associated with poor response to endocrine therapy. J. Clin. Oncol., *14:* 2584–2589, 1996.
23. Pechoux, C., Gudjonsson, T., Ronnov-Jessen, L., Bissell, M. J., and Petersen, O. W. Human mammary luminal epithelial cells contain progenitors to myoepithelial cells. Dev. Biol., *206:* 88–99, 1999.
24. Palacios, J., Benito, N., Pizarro, A., Suarez, A., Espada, J., Cano, A., and Gamallo, C. Anomalous expression of P-cadherin in breast carcinoma. Correlation with E-cadherin expression and pathological features. Am. J. Pathol., *146:* 605–612, 1995.
25. Peralta Soler, A., Knudsen, K. A., Salazar, H., Han, A. C., and Keshgegian, A. A. P-cadherin expression in breast carcinoma indicates poor survival. Cancer (Phila.), *86:* 1263–1272, 1999.
26. Seagroves, T. N., Lydon, J. P., Hovey, R. C., Vonderhaar, B. K., and Rosen, J. M. C/EBPβ (CCAAT/enhancer binding protein) controls cell fate determination during mammary gland development. Mol. Endocrinol., *14:* 359–368, 2000.
27. Seagroves, T. N., Krnacik, S., Raught, B., Gay, J., Burgess-Beusse, B., Darlington, G. J., and Rosen, J. M. C/EBPβ, but not C/EBPα, is essential for ductal morphogenesis, lobuloalveolar proliferation, and functional differentiation in the mouse mammary gland. Genes Dev., *12:* 1917–1928, 1998.
28. Stoesz, S. P., Friedl, A., Haag, J. D., Lindstrom, M. J., Clark, G. M., and Gould, M. N. Heterogeneous expression of the lipocalin NGAL in primary breast cancers. Int. J. Cancer, *79:* 565–572, 1998.
29. Marinkovich, M. P., Taylor, T. B., Keene, D. R., Burgeson, R. E., and Zone, J. J. LAD-1, the linear IgA bullous dermatosis autoantigen, is a novel 120-kDa anchoring filament protein synthesized by epidermal cells. J. Investig. Dermatol., *106:* 734–738, 1996.
30. May, F. E., and Westley, B. R. Expression of human intestinal trefoil factor in malignant cells and its regulation by oestrogen in breast cancer cells. J. Pathol., *182:* 404–413, 1997.
31. Courjal, F., Louason, G., Speiser, P., Katsaros, D., Zeillinger, R., and Theillet, C. *Cyclin* gene amplification and overexpression in breast and ovarian cancers: evidence for the selection of *cyclin D1* in breast and *cyclin E* in ovarian tumors. Int. J. Cancer, *69:* 247–253, 1996.
32. Ivanov, S., Liao, S. Y., Ivanova, A., Danilkovitch-Miagkova, A., Tarasova, N., Weirich, G., Merrill, M. J., Proescholdt, M. A., Oldfield, E. H., Lee, J., Zavada, J., Waheed, A., Sly, W., Lerman, M. I., and Stanbridge, E. J. Expression of hypoxia-inducible cell-surface transmembrane carbonic anhydrases in human cancer. Am. J. Pathol., *158:* 905–919, 2001.
33. Petersen, O. W., Hoyer, P. E., and van Deurs, B. Frequency and distribution of estrogen receptor-positive cells in normal, nonlactating human breast tissue. Cancer Res., *47:* 5748–5751, 1987.