ABSTRACT
A major objective of current cancer research is to develop a detailed molecular characterization of tumor cells and tissues that is linked to clinical information. Toward this end, we have identified approximately one-quarter of all genes that were aberrantly expressed in a breast cancer cell line using differential display. The cancer cells lost the expression of many genes involved in cell adhesion, communication, and maintenance of cell shape, while they gained the expression of many synthetic and metabolic enzymes important for cell proliferation. High-density, membrane-based hybridization arrays were used to study mRNA expression patterns of these genes in cultured cells and archived tumor tissue. Cluster analysis was then used to identify groups of genes, the expression patterns of which correlated with clinical information. Two clusters of genes, represented by p53 and maspin, had expression patterns that strongly associated with estrogen receptor status. A third cluster that included HSP-90 tended to be associated with clinical tumor stage, whereas a forth cluster that included keratin 14 tended to be associated with tumor size. Expression levels of these clinically relevant gene clusters allowed breast tumors to be grouped into distinct categories. Gene expression fingerprints that include these four gene clusters have the potential to improve prognostic accuracy and therapeutic outcomes for breast cancer patients.

INTRODUCTION
A detailed molecular characterization or fingerprint of cancer is an objective recently made possible by the development of several new high throughput analytical methods. These include techniques for the analysis of DNA, mRNA, and proteins within a cell (1–3). A goal that is now in sight is to build databases of detailed molecular information and to link them to clinical information (4). This approach has the potential to help patients by very accurately grouping tumor subtypes, which may enable clinicians to more accurately distinguish prognostic groups and predict the most effective therapies. In breast cancer, in particular, prognostic methods that are more informative than the current ones are needed to predict the benefit to patients of chemotherapy treatments.

In cancer prognosis and treatment, a shortcoming currently is the lack of methods that adequately address the complexity and diversity of the disease. Cancer is a highly heterogeneous disease, both morphologically and genetically (5). No simple relationship has been demonstrated between a mutation or the expression level of a given gene or protein and a certain etiology or extent of disease (6). Prognostic marker systems based on single parameters have generally proven inadequate. Thus, multiparametric methods, methods that rely on many pieces of information, are ideally suited to the grouping of tumor subtypes.

mRNA fingerprints that represent the expression patterns of large numbers of genes have the potential to allow precise and accurate grouping of tumor subtypes. cDNA microarrays recently were used to compare the changes in expression patterns of a set of 5000 randomly selected genes in breast tumor tissue and cultured breast epithelial cells (7). The results demonstrated two types of expression changes: (a) those relevant to the tumorigenic process of breast epithelial cells; and (b) those irrelevant to the disease and, in some cases, originating from nontumor cells that were present in varying amounts in the tumor biopsies. As a first step toward identifying physiologically relevant gene expression changes, clusters of genes that responded in a concerted manner to exogenously added growth factors were identified (7).

We report here a different approach to identify clinically relevant changes in gene expression. DD* reverse transcription-PCR (8, 9) was used as a primary screen to identify a set of 170 genes that were expressed by breast epithelial cells and were differentially expressed in a breast cancer cell line. DD comparisons were made between sorted normal breast epithelial cells (10) and a highly malignant breast tumor cell line, MDA-MB-435 (11). High-density, membrane-based hybridization arrays were then developed to assay expression patterns in cultured breast cells and biopsied tumor tissues. Using cluster analysis (12), a computer algorithm developed to analyze cDNA microarray data, groups of genes with expression patterns that correlated with clinical parameters were identified.

Four clusters of genes are reported here, the expression of which was associated with three major parameters used clinically to characterize breast tumors: ER status, tumor stage, and tumor size. Expression patterns of these clusters were then used to group breast cancer patients into distinct categories. The feasibility of building a large database of expression patterns linked to clinical information is demonstrated by these results. The four gene clusters described represent an initial step in the process of identifying useful sets of marker genes and will likely have their greatest utility when used in combination with additional sets of physiologically relevant genes.

MATERIALS AND METHODS

Cells and Tissues. Normal breast myoepithelial and luminal epithelial cells were sorted from primary cultures of mammaplasty tissue by immunomagnetic methods using epithelial membrane antigen and common acute lymphoblastic leukemia antigen antibodies (10). The 21T breast tumor progression series of cell lines originated from a primary tumor and metastatic pleural effusion from a single patient (13). 76N normal breast epithelial cells obtained from mammaplasty tissue express several myoepithelial cell markers and undergo senescence after 20 passages (14). Other cell lines were obtained from the American Type Culture Collection. All cells were grown in DFCI-1 medium (14) and harvested at 70% confluence. The PT series of breast tumor tissue samples was obtained within 30 min of surgery and frozen in liquid nitrogen, whereas the H series was OCT-cryofrozen archived tumor tissue. Tumor tissue was grossly dissected from normal adjacent tissue. All RNA was prepared by the CsCl-cushion method, as described (9, 15). ER status was determined clinically by the cytosolic ligand-based assay (H series) or immunoassay (PT series).

Received 10/15/99; accepted 2/16/00.

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 Funded by a grant from the Ludwig Institute for Cancer Research.
2 To whom requests for reprints should be addressed, at Dana-Farber Cancer Institute, Department of Cancer Biology, Cancer Institute, D610B. 44 Binney Street, Boston, MA 02115.
3 Deceased.

4 The abbreviations used are: DD, differential display; ER, estrogen receptor.
DD Reverse Transcription-PCR. DD was performed as described (9) to compare normal breast epithelial cells and a metastatic breast tumor cell line, MDA-MB-435 (11). The normal cells used were either 76N cultured breast epithelial cells (14) or sorted normal breast myoepithelial and luminal epithelial cells (10). Both up- and down-regulated cDNA bands were selected for analysis. Approximately 70 primer pairs were used, including: (a) LHA-1, -2, -3, -4, -5, -6, and -7 in combination with LHT1-G, -A, and -C (9); (b) E1-OPA-1, -2, -3, -7, and -8 and H3-OPA-5, -6, -9, and -10 in combination with LHT1-G, -A, -C (9); (c) ARP-1, -2, -3, and -4 in combination with AP-1, -2, -3, -4, -5, -6, -7, -8, -9, and -10 (Genomyx/Beckman Corp., Foster City, CA); and (d) AP-1, -2, -3, -4, and -5 in combination with T12 M-G, -A, -T, and -C (GenHunter Corp., Nashville, TN). PCR conditions were as recommended by the respective primer kit manufacturers or as described (9) for the LHA and OPAs series of primers. Gel electrophoresis was performed on the extended format of the programmable Genomyx LR apparatus (Genomyx/Beckman Corp.). DD bands were eluted and precipitated (16). Either of three different approaches was then used to identify the bands and obtain a cDNA clone: (a) cDNAs were TA-cloned into pCR2 or pCR2.1 (Invitrogen, Carlsbad, CA) directly and colonies were screened for differential expressors, which were then sequenced; (b) cDNAs were directly sequenced (16, 17), a gene-specific primer was made, and the PCR product was then TA-cloned and sequenced to confirm cloning of the correct cDNA. Genes that were measurable. Sets with SEs exceeding 150% of the mean were disregarded (not detected). Signal intensities for each membrane were normalized to the median signal of that membrane. For RNAs run multiple times geometric means of all non-BKG membrane-normalized values were calculated. A single median BKG value was determined from an entire set of membranes being compared, and this value was substituted for all BKG values. Signals for each individual gene were then normalized to the geometric mean of the expression level of that gene across the set of membranes being compared. Genes with consistently low signals across an entire set of comparison membranes were omitted from the analysis.

Cluster analysis was performed using publicly available software written by M. Eisen (Stanford University, Stanford, CA). Data sets were logarithmically transformed, and the similarity metric was an uncentered correlation. An image contrast of 3 or 4 was used. Raw data tables, full cluster diagrams with gene identities, and other materials are available.

RESULTS

DD Results. To begin to identify clinically relevant changes in gene expression in breast cancer, we have used DD as a primary screen. A DD method that produced fewer than 6% false positives (17) was used to compare normal breast myoepithelial, luminal epithelial (10), and 76N cells (14) with a highly malignant breast tumor cell line, MDA-MB-435 (11). Seventy primer combinations were used to screen ~7,000 genes, which, because cells express 15,000 mRNAs (19), represents over one-third of all expressed genes. Four and one-half percent of all mRNAs were differential; hence, a total of 700 genes were differentially expressed between the cancer and normal cells. This number is in agreement with other studies using different methods (20). A set of 170 genes that were differentially expressed was identified (a complete list of known genes with names and GenBank accession numbers is available).

These 170 genes represent approximately one-quarter of all of the genes that were differentially expressed in the normal and cancer cells compared, whereas the 107 genes included on the hybridization arrays represent one-seventh.

The great majority of the differential genes we observed were similarly expressed in all of the normal breast cell types, but were either down- or up-regulated in the tumor cells. A marked contrast was seen in the types of genes that comprised the down- and up-regulated categories. Nearly 70% of the genes that were down-regulated in the tumor cells were categorized as filamentous, cell surface, and secreted genes that play roles in adhesion, communication, and the maintenance of cell shape (Table 1). In contrast, 75% of the known genes, the expression of which was up-regulated in tumor cells, were enzymes involved in metabolism, macromolecular synthesis, and disruption of the extracellular matrix.

Hybridization Array Method. To further study mRNA expression patterns, we have used a membrane-based hybridization array spotted with tags representing 124 different genes. These gene tags included 89 DD-identified normal cell-specific genes, 18 DD-identified tumor cell-specific genes, as well as literature reported cancer genes and housekeeping genes (Fig. 1A). A full listing of the genes included on the arrays is available.

To assess the reproducibility of this set of hybridization array assays, experiments using the same RNA preparation were repeated on different days, and measurable, median-normalized expression values for each gene were compared. In two experiments using


MCF-10A RNA, expression levels of 95.5% of genes had repeated values that were within 4-fold of each other (Fig. 1B). In five experiments using MCF-10A RNA, 95% of values deviated by <5-fold (data not shown). Based on the level of reproducibility of this set of array experiments, individual gene expression changes of <5-fold were not considered significant. In addition, no conclusions based on a single data point have been made. The array assay, as performed, was highly sensitive for a non-PCR-based assay, as indicated by the mean expression level that was within 4-fold of each other (Fig. 1B). Furthermore, expression signals from >90% of the DD genes could be measured, which was considerably better than our detection rate of ~60% (14 of 22) with conventional Northern assays for the same set of DD genes (Ref. 17 and data not shown).

**Identification of Gene Clusters with Clinically Relevant Expression Patterns.** mRNA expression levels of 124 genes were assayed using hybridization arrays in breast tumor tissue samples obtained from 18 patients and seven breast cell lines. Color-coded relative expression levels are displayed in Fig. 2. Data from individual genes are shown in rows and tissues in columns. Cluster analysis was used to organize the genes and the tissues so that those with the most closely related expression patterns are positioned adjacent to each other (12). Cluster analysis is a computer method that calculates correlation coefficients between all gene and tissue data sets. It then organizes the positions of the rows and columns of the display and generates hierarchical trees that indicate the degree of relatedness by the height of nodes.

When cluster analysis was performed using expression patterns from all of the DD genes included on the arrays, the cells and tumor tissues fell into two major groups shown by the two branches of the dendrogram at the top of Fig. 2. The group at the right included predominantly ER-positive samples, whereas the group at the left included predominantly ER-negative samples. This division indicated that the arrays included many genes with expression patterns that reflected ER status.

Seven well characterized breast cell lines are included with the tumor tissues on Fig. 2. The manner in which cluster analysis organized these cell lines demonstrated its ability to accurately recognize and make groupings according to physiologically relevant characteristics. The 21T series of cell lines (21 MT-1, 21 MT-2, 21NT, and 21PT), which were all derived from a single breast cancer patient (13), were grouped together in a closely related cluster at the left. 21PT cells, which are unable to form tumors in nude mice (13), were positioned on a deeper node reflecting a more distant relationship to the three other 21T cell lines, which are all tumorigenic in mice (13). The highly metastatic breast tumor cell line MDA-MB-435 (MDA-435) (11) was found to be more similar to the 21T series of tumor cell lines than was the immortal but nonmalignant breast cell line MCF-10A (21). The single ER-positive cell line, MCF7 (22), was widely separated from the six ER-negative cell lines.

To identify genes with clinically relevant expression patterns, charts with tissues ordered by various clinical parameters were prepared. These were then screened for individual genes and gene clusters with expression patterns that increased or decreased across the chart. The significance of resulting Ps (Fisher’s exact test) was considered in the context of the multigene and multitest analysis. Because we assayed 124 genes, a P for an individual gene must be <1–0.95\(^{124}/124\) or <0.0004 to be considered statistically significant, if we had performed a single test. We tested for the association of gene expression with six parameters: ER status, tumor stage, grade, size, the percentage of S phase cells, and patient age. Hence, statistically meaningful Ps for individual genes must be <1–0.95\(^{6/124}\) or <0.00007. Failure to meet this level does not mean that the gene is not a good marker, but rather that the current experimental design did not prove that it is.

Seven individual genes were found to have P < 0.0004, which suggested a possible association with a clinical parameter. All seven were associated with ER status. They included keratin 19 (P < 0.0001 and 0.0002), PAG (P = 0.0002), unknown 94 (P = 0.0001), PDAC-2 (P = 0.0002), unknown 102f (P = 0.0002), and lysosomal sialyl-transferase (P = 0.0001). No other individual genes were found to be significantly associated with any other clinical parameter.

In a multigene and multitest study it may be more appropriate to address overall patterns of results (e.g., clustering). Gene clusters with mean expression levels that showed important associations with clinical parameters are shown (Fig. 2). Two clusters were associated with ER status, one with clinical stage, and one with tumor size. No

<table>
<thead>
<tr>
<th>Gene Category</th>
<th>Reduced in tumors cells (%)</th>
<th>Increased in tumor cells (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Secreted</td>
<td>27</td>
<td>5</td>
</tr>
<tr>
<td>Membrane associated</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>Filamentous</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>Metabolic enzymes</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>Regulatory</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>Protein synth and processing</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>Transcription</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Replication</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>70</td>
<td>35</td>
</tr>
</tbody>
</table>

**Table 1** Comparison of functional classes of genes reduced or increased in expression in tumor cells relative to normal cells.
clusters were associated with tumor grade, percentage of S phase cells, or patient age. $P$s (Fisher’s exact test) calculated using the average gene expression data shown in the top row of each cluster are shown in an attempt to summarize and compare the significance levels of the four clusters. Because of the complexity of the analysis, these $P$s are not statistically interpretable.

ER status is a major clinical grouping in breast cancer that is routinely measured to predict responsiveness to antihormone therapy.

Fig. 2. Expression profiles for breast tumor tissue and cultured tumor cells. Top, mRNA expression levels for 124 genes assayed in 18 biopsied breast tumor samples and seven cultured breast cell types. Relative expression levels are indicated in the top color key. Genes are shown in rows, and identities are indicated. Tissues/cells are shown in columns, and identities are indicated. The dendrogram at the top indicates the degree relatedness between tumor samples by the height of nodes. Similarly, the dendrogram at left indicates the relatedness between gene expression patterns. Gene clusters I, II, III, and IV, which were associated with clinical parameters, are indicated by magenta boxes on the left dendrogram. ER status of the cell lines and patient tissues is indicated following the sample name. Bottom, genes within clinically relevant clusters I, II, III, and IV. Each of the four panels shows color-coded gene expression levels for the indicated genes. The top row of each panel represents the mean gene expression level. $P$s (Fisher’s exact test) are indicated. Expression of the genes in clusters I and II associated with ER status. All 25 breast tumor cell lines and tumor biopsies were included in ER analysis. Expression of cluster III genes was associated with tumor stage. Eighteen tumor tissues were included in the analysis. Tumor stage is indicated after the tissue identification code. Expression of cluster IV genes was associated with tumor size. Eighteen tumor tissues were included in the analysis. Tumor size is indicated after the tissue identification.
Clusters I and II were strongly ER-associated and expressed inversely to each other \((r = -0.50, P = 0.012)\). Identities of genes in these clusters are shown (Fig. 2). Expression of the p53 cluster (cluster I) was higher in ER-positive tissues than ER-negative tissues. Expression of the maspin cluster (cluster II) was the inverse.

A second major clinical grouping applied to breast cancer is tumor stage, which takes into account information on tumor size, nodal status, and distant metastases (23). Gene cluster III, which included HSP-90, tended to be overexpressed in stage IV tumors relative to stages I, II, and III tumors. Stage IV breast tumors are distinguished from earlier stage tumors by the presence of distant metastases. Clinical stage is currently the best indicator of disease prognosis (23), and, hence, this cluster may represent a valuable set of markers that provide prognostic information.

Tumor size is also an important independent predictor of disease prognosis (23). Gene cluster IV, which included keratin 14, was reduced in expression in tumors larger than 1.5 cm relative to smaller tumors.

**Using the Clinically Relevant Gene Clusters to Categorize Breast Tumors.** We have used cluster analysis and the expression patterns of the four clinically relevant gene clusters to group breast tumor tissue into categories. Results are shown in Fig. 3. This analysis sorted the tumors into two major groups that differed in their ER status \((P = 0.0002)\). The ER-negative group is the top, gray group in Fig. 3. Grouping by other clinical parameters is also apparent. For example, two highly related groups of tumors (group 1: H16, H4, and H43, group 2: PT-10 and PT-6) included tumors with similar clinical data. The former group was advanced stage, ER-negative tumors, whereas the latter group was ER-positive, stage II invasive carcinomas from women of similar ages.

Although cluster analysis was able to group breast tumors by their ER status with high accuracy, those grouped apparently inappropriately may represent the most interesting cases. Three of the 18 tumors seemed to be grouped inappropriately; a single ER-positive tumor (H16) grouped in the ER-negative cluster and two ER-negative tumors (H10 and H33) grouped in the ER-positive cluster. The gene expression patterns may reflect clinically important characteristics of these tumors. For instance, tumor H16 may express a receptor capable of binding to estrogen but otherwise nonfunctional and, hence, unable to activate ER-responsive genes. Such a tumor would be predicted to be unresponsive to treatment with antiestrogens. We note that H16 was an unusual tumor specimen. It was derived from a hip metastasis that seemed to originate from a primary breast tumor surgically removed 18 yr prior. Following total hip replacement, the patient was treated with radiation and tamoxifen. After 2 yr, the patient is without evidence of recurrent cancer. It is not possible to make conclusions regarding the responsiveness of the tumor to tamoxifen, because tamoxifen was not the sole treatment. The other two inappropriately grouped tumors, H10 and H33, may harbor constitutive oncogenic mutations, either in the ER itself or in downstream components of the pathway, that result in ER pathway activation in the absence of ER ligand-binding activity. Tumors of this type have been reported previously (24).

**DISCUSSION**

We have used an approach involving DD and hybridization arrays to link detailed gene expression patterns to clinical information in breast cancer. Four clusters of genes were identified that had expression patterns associated with three major parameters used clinically to characterize breast tumors: ER status, tumor stage, and tumor size. The two ER status-associated clusters are especially significant because they are large clusters with a high degree of statistical significance that have the potential to assist in identifying ER-positive tumors that are unresponsive to tamoxifen. We have used the expression patterns of all four clusters together to group breast cancer patients into distinct categories. These results demonstrate the feasibility of building large databases of expression patterns that can be linked to clinical information to assist in determining prognosis and making therapeutic decisions.

We describe here an approach that used a DD prescreening step and allowed us to identify a number of genes with physiologically relevant expression patterns. The DD step increased the probability that the arrayed tags detected genes that were expressed by breast epithelial cells and were differential in breast cancer. Other methods of preselecting tissue-appropriate genes for inclusion on arrays have been reported (25). These have also reported markedly better rates of differential expression in cancer cells than array methods using random gene collections. Recent studies have described the results of cDNA microarray hybridizations and cluster analysis of breast and colon tumor tissues (7, 26). These studies did not report the clustering of tissues based on ER status or clinical stage. Both of these studies used large unselected collections of arrayed gene tags that represented genes expressed in a wide variety of cell types, including fibroblasts,stromal cells, adipocytes, and epithelial cells. Whereas we point out that our current study was not an exhaustive analysis and that many...
important genes were certainly missed by limiting the comparison to a single tumor cell line, our results nevertheless demonstrate the usefulness of the DD preselection approach and identify several potentially useful marker genes.

The types of genes identified here by DD gives information on the process of tumorigenesis. The tumor cells generally lost the expression of genes involved in cell adhesion, communication, and the maintenance of cell shape, whereas they generally gained the expression of synthetic and metabolic enzymes important for cell proliferation. These general processes are well studied as important events in tumorigenesis.

A large number of the genes isolated by DD were associated with ER status. Of the 107 DD-identified genes tested here, 31 (or 29%) were included in clusters with expression levels that were significantly correlated with ER status. Twenty-two genes were expressed at high levels in ER-positive tumors, whereas nine genes were expressed at low levels in ER-negative tumors. The finding of many ER status-associated genes was surprising based on literature reports that generally describe remarkably few direct target genes of the ER. These include the ER gene itself, the progesterone receptor, p52, AR, LIV1, keratin 19 (27), α1-antichymotrypsin (28), complement component 3 (29), and HSP-27 (30). In explanation, the genes we found likely include directly and indirectly regulated genes, whereas genes previously identified include only directly regulated genes. The direct regulation by ER of only a few proliferative factors could result in many indirect changes in gene expression.

Several of the genes identified here as ER status-associated were previously reported to be ER-regulated or differentially expressed as a function of ER status. Keratin 19 gene expression and filament organization are regulated by estrogen (27). Keratin 19 mRNA was identified as one of 10 highly overexpressed mRNAs in a cDNA microarray comparison of two ER-positive breast tumor cell lines relative to two ER-negative lines (31). Studies of p53 have generally addressed protein expression levels, and very few studies have addressed p53 mRNA levels. A tendency toward the association of p53 mRNA expression levels with ER status was noted in one study (32), although it is well known that p53 protein levels are not associated with ER status (see, for example, Ref. 33). CD59, a cell surface component of the complement system, protects cells from complement attack. Protein levels of CD59 are high in the ER-positive breast tumor cell line MCF7 (34). Other complement component genes are known to be estrogen responsive (35). α1-Antichymotrypsin was the first estradiol-induced protein to be identified (28), although its levels of expression in ER-positive versus ER-negative cell lines or tissues have not been reported. Likewise, CC3 is also an estrogen-responsive gene (29), the mRNA expression levels of which have not been studied in tumor tissue. Histone H4 is a commonly used proliferation marker, and proliferation rates are generally reduced in ER-positive relative to ER-negative breast tumors (36). Maspin and elafin are both protease inhibitors that are down-regulated in breast cancer cells (37, 38). Neither has been previously studied in relation to ER status.

It is reasonable to think that ER-positive and ER-negative cells represent two very different tumor cell types in breast cancer. Not only do our results indicate that ER status is associated with widespread changes in gene expression patterns in tumor cells, but ER status is commonly used as a major clinical grouping in breast cancer. One can postulate that cancer cells with ER-positive gene expression patterns arise either from specialized cells that naturally require estrogen for their growth or by aberrantly using the hormone system to facilitate their growth. Cancer cells with ER-negative gene expression patterns, on the other hand, may include tumors that arise from nonhormone-responsive breast cells through mechanisms commonly used by tumor cells in nonhormonal tissues.

Useful clinical information may be provided by the expression patterns of ER status-associated genes in individual tumors. In particular, these patterns may help to identify tamoxifen nonresponsive patients. Patients with ER-positive tumors, as assayed clinically by standard ligand-binding or immunostaining assays, are typically treated with the antiestrogen tamoxifen. Response rates of up to 60% have been reported (39). Nonresponders may include patients with tumors that express an ER that can bind to ligands but does not function as a transcription activator. The expression patterns may show if a functional ER is present. In our study, one clinically ER-positive tumor (H16) was grouped with ER-tumors. This tumor may express a receptor capable of binding to estrogen but otherwise nonfunctional and, hence, unable to activate downstream genes. Such a tumor would be predicted to be unresponsive to treatment with antiestrogens.

A recent study has described four subclasses of ER status based on the presence and functionality of the receptor (24). These include: (a) fully normal ER that binds to its ligands and exhibits ligand-dependent DNA-binding and transcription activation; (b) ER that bind to ligands but is not functional; (c) ER that does not bind to ligands, but constitutively binds to DNA and activates transcription; and (d) cells devoid of ER. The first three subtypes would be categorized as ER positive by clinical immunohistochemical assays, although the latter three subtypes may not respond to antiestrogen therapy. Assays that assess receptor function, such as the DNA-binding assay described previously (24) or the ER status-associated expression patterns described here, may help to identify the 40% of ER-positive breast cancer patients (39) who do not respond to tamoxifen.

The clinical stage-associated cluster identified here included HSP-90. Literature evidence supports a role for this gene in advanced breast tumors. HSP-90 was elevated in breast tumor tissue, and antibodies to HSP-90 were associated with poor survival in breast cancer (40, 41). Furthermore, HSP-90 plays a role in ER signal transduction, apparently by increasing ER transcriptional activity in conditions of low estrogen (42).

The tumor size-associated cluster included keratin 14, CD44, keratin 5, and glutathione S-transferase π. Keratin 14 is expressed specifically by normal breast myoepithelial cells and not by breast carcinoma cells, which typically express markers of luminal epithelial cells (43). Markers of myoepithelial cells are present in normal tissue, benign lesions, and ductal carcinoma in situ, but are absent from invasive tumors (44, 45). Keratin 14 loss in larger tumors likely reflects a reduced proportion of normal cells and ductal carcinoma in situ in biopsy specimens from larger tumors. The other genes in this cluster are also expressed preferentially by normal myoepithelial cells (46). And data not shown).

We have used undissected tissue biopsies because this material is readily available for clinical analysis. Tissue biopsies, however, are composed of a variety of different cell types. The presence of nontumor cell types may allow us to gain contextual information on the tumor. It is important to consider, however, that the expression patterns reported did not necessarily originate from the tumor cells themselves. For example, maspin and elafin were found to be highly expressed in a number of advanced stage, ER-negative tumors. These two genes are generally down-regulated in metastatic breast tumor cell lines relative to normal breast myoepithelial and luminal epithelial cells (37, 38). Furthermore, maspin acts as a tumor suppressor in breast cancer (37). These and other genes may be increased in expression in normal cells that are adjacent to tumor cells as a part of a natural tumor defense system (47).

Detailed molecular characterization or fingerprinting methods are versatile in their potential to improve therapeutic decisions in many diseases, not only cancer. In the cancer field, these methods are applicable to any malignancy for which malignant cells are available for analysis. Newer methods are being developed to use very small samples of tissue. It is also feasible to begin to look at disseminated
tumor cells circulating in the blood. A general approach for the development fingerprinting methods is to assemble a sufficient collection of marker genes relevant to a particular disease. Using these markers, a sizable database of expression patterns is then built from sources for which clinical histories are available. This database allows an assessment of the use of the method and also provides the backbone of information against which to compare each incoming sample for comparative prognostic and predictive information.

ACKNOWLEDGMENTS

We are grateful to Rebecca Gelman for biostatistics advice, Heide Ford for critical reading of the manuscript, Micaela Coady and Edith Kabingu for technical assistance, and Paul Morrison for website construction.

REFERENCES


16. Martin, K. J., Kwan, C. P., and Sager, R. A direct-sequencing-based strategy for the development of fingerprinting methods is to assemble a sufficient collection of marker genes relevant to a particular disease. Using these markers, a sizable database of expression patterns is then built from sources for which clinical histories are available. This database allows an assessment of the use of the method and also provides the backbone of information against which to compare each incoming sample for comparative prognostic and predictive information.


Linking Gene Expression Patterns to Therapeutic Groups in Breast Cancer

Katherine J. Martin, Brian M. Kritzman, Laura M. Price, et al.


**Updated version**
Access the most recent version of this article at:
http://cancerres.aacrjournals.org/content/60/8/2232

**Cited articles**
This article cites 41 articles, 21 of which you can access for free at:
http://cancerres.aacrjournals.org/content/60/8/2232.full.html#ref-list-1

**Citing articles**
This article has been cited by 31 HighWire-hosted articles. Access the articles at:
/content/60/8/2232.full.html#related-urls

**E-mail alerts**
Sign up to receive free email-alerts related to this article or journal.

**Reprints and Subscriptions**
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

**Permissions**
To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.