Chipping away at breast cancer: insights from microarray studies of human and mouse mammary cancer

K V Desai¹, C J Kavanaugh¹,2, A Calvo¹ and J E Green¹

¹Laboratory of Cell Regulation and Carcinogenesis, National Cancer Institute, NIH, Building 41, Room C629, 41 Library Dr., Bethesda, Maryland 20892, USA
²Division of Cancer Prevention, National Cancer Institute, NIH, Bethesda, Maryland 20892, USA

(Requests for offprints should be addressed to J E Green; Email: JEGreen@nih.gov)

Abstract

Breast cancer is the most prevalent tumor in American woman. Multiple factors, including age, diet, genetics, environment, geographic location, parity, as well as race, influence the development of this heterogeneous disease. As the process of oncogenesis involves the disruption of diverse cellular pathways including cell cycle, growth, survival, and apoptosis, the high throughput technique of microarray analyses provides a powerful insight into multiple cellular processes. These techniques determine expression patterns of thousands of genes simultaneously. Recent studies have identified particular expression patterns that can classify tumors into new groups and aid in the prediction of the natural history of the disease and the therapeutic response. This wealth of information may also form the basis for the development of new types of targeted therapies. Studies to identify the earliest molecular events in oncogenesis and progressive changes in the human disease have been difficult to perform within the same patient. The use of transgenic mouse mammary cancer models provides an opportunity to decipher molecular changes that occur at progressive stages of tumor development. This paper reviews microarray technology, and the insights gained from published breast cancer microarray analyses, and considers the contribution of microarray studies in identifying mouse cancer models that may be appropriate for answering particular experimental questions.

Overview

Breast cancer is the most prevalent tumor in American women accounting for 30% of all cancers diagnosed in the female population of the USA (Greenlee et al. 2000). Despite an abundance of information about molecular alterations that occur in breast cancer, many fundamental questions regarding the etiology, epidemiology and prediction of the clinical course of the disease remain largely unanswered. Until recently, the clinical classification of breast cancer has relied primarily upon morphologic characteristics of the tumor, including tumor size, lymph node status, and histology, as well as the expression of particular markers associated with the clinical course. These include steroid hormone receptor levels for estrogen receptor (ER) and progesterone receptor (PR), expression of the her2/neu oncogene, biomarkers for cell proliferation (Ki-67 antigen, proliferating cell nuclear antigen (PCNA)) and angiogenesis (vascular endothelial growth factor (VEGF)), CD31 antigen, and factor VIII. While these parameters have proved useful in predicting the clinical course of the disease, they do not adequately stratify patients into groups that would further define the most appropriate course of therapy for individuals. Relatively few therapeutic options for treating breast cancer currently exist and are of limited success for treating advanced disease, especially in the case of ER-negative (ER−) tumors.

The recent development of gene expression profiling using microarray technology allows for the molecular dissection of gene expression in tumors on an enormous scale. The expression patterns of tens of thousands of genes in a tumor analyzed in a single experiment can be compared with microarray data from large numbers of histologically similar tumors. This approach promises to greatly improve tumor classification by using particular sets of molecular alterations to define tumor sub-types. By determining signature profiles for sub-types of breast tumors and correlating this information with the natural history of tumor progression and therapeutic response, treatment for breast cancer can become more appropriately individualized. Another expectation from this revolutionary approach is that new, fundamental insights
Desai et al.: Breast cancer microarrays

into breast cancer biology and molecular genetics may be gained, leading to the identification of additional disease markers and targets for therapeutic intervention. Since much basic research on breast cancer is conducted using various animal models of the disease, gene expression profiling of these models offers the ability to define pathway-specific oncogenic signatures and to identify models that may be best suited for particular types of preclinical studies.

This review will focus on the application of microarray technologies for classifying both human breast cancer as well as mouse models of human breast cancer. We will first provide an overview of the clinical variations known to occur in breast cancer and the critical problems that may be addressed by microarray studies.

Clinical diversity of breast cancer and important challenges

Heterogeneity in diet, age, race, environmental factors, geographic location, number of pregnancies, as well as genetic makeup, are some factors that determine the risk of breast cancer (Ketcham & Sindelar 1975, Russo & Russo 1999). The degree of complexity of the disease is further enhanced by chromosomal rearrangements frequently associated with the pre-malignant disease (Bieche & Lidereau 1995, Schwab & Amler 1990). The cellular pathways that are altered by these aberrations have been difficult to evaluate in patients, especially during early stages of the disease process.

A major defining characteristic of breast cancer is the status of ERα expression in the breast tumors and their sensitivity to anti-estrogen therapy. Estrogen plays a key role in the development, growth, and differentiation of the normal mammary gland, as well as in initiation and progression of breast cancer (Allred et al. 2001). The significance of estrogen in breast cancer has been underscored by the success of endocrine therapy using tamoxifen, an antagonist of estrogen action. Tamoxifen reduced the relapse of estrogen-responsive breast tumors and delayed or prevented cancer in 49% of the women considered to be at high risk for breast cancer (Peto & Mack 2000). However, tamoxifen is associated with several side-effects in cases of prolonged treatment (Fisher et al. 1998, Day et al. 1999). Raloxifene, another ER antagonist, has therapeutic effects similar to tamoxifen but has fewer side-effects (Cummins et al. 1999, Barrett-Connor 2001). Thus, additional selective ER modulators and down-regulators, as well as aromatase inhibitors, are being developed in order to treat ER− and/or estrogen non-responsive cancer (Wolf & Davidson 2001). While anti-estrogen therapy has proven effective in a significant portion of cases, only about 50% of breast tumors are ER positive (ER+) at presentation. A significant number of these, however, are not responsive to tamoxifen (Dorssers et al. 2001).

The mechanism by which tumors change from a hormone-responsive to a hormone-independent state (generally associated with becoming ER−) is not clearly understood, although down-regulation or loss of ER expression may occur via methylation of the ER gene promoter (Ottaviano et al. 1994). It is also possible that ER− tumors arise from a population of ER− precursor cells and are, thus, resistant to anti-estrogen therapy (Ferguson et al. 1997, Russo et al. 1999).

In addition to surgical excision, traditional therapeutic strategies for breast cancer have been employed, including radiation therapy and chemotherapy using anti-neoplastic compounds (Table 1) (Dorr 1990, Clavel & Catimel 1993, Miller & Sledge 1999, Levi et al. 2001, Tan & Swain 2001). Although combinatorial therapies have decreased the overall risk of breast cancer relapse by 20–40%, significant side-effects as well as the development of chemoresistance may occur.

The first gene-based therapeutic approach for breast cancer was developed for her2/neu/erbB2, a receptor tyrosine kinase of the epidermal growth factor receptor (EGFR) family, which is amplified and/or overexpressed in 30% of early stage breast cancers. The expression of erbB2 is inversely correlated with patient survival (Slamon et al. 1987). A humanized monoclonal antibody to her2/erbB2 (Herceptin) has been developed which blocks the activity of this receptor tyrosine kinase. In combination with chemotherapy, Herceptin has been shown to increase the overall response rate, duration, and mean survival time by 25% (Slamon et al. 2001). However, Herceptin may lead to congestive heart failure when administered in combination with anthracyclines (Cook-Bruns 2001, Sparano 2001).

The results are encouraging and constitute a ‘proof-of-principle’ for the development of additional gene-based therapies. Therapies based upon other gene targets are currently

Table 1 Available therapeutic targets and those identified as novel candidates

<table>
<thead>
<tr>
<th>Endocrine therapy for ER+ tumours</th>
<th>SERMS</th>
<th>Tamoxifen</th>
<th>Raloxifene</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Aromatase inhibitor, GnRH agonists, isoflavones</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chemotherapy for ER+ or ER− tumors</td>
<td>Adjutant therapy with anthracyclines, anti-metabolites, alkylating agents, plant taxanes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-steroidal anti-inflammatory drugs, miscellaneous cytotoxic agents like sunlindac sulfone and monoterpenes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene-based therapy</td>
<td>Herceptin-anti-erbB2 antibody</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- anti-EGFR antibodies, cyclin D1 inhibitors (flavopiridol, staurosporine), HDAC inhibitors, anti-angiogenesis (anti-VEGF)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GnRH, gonadotropin releasing hormone; VEGF, vascular endothelial growth factor; HDAC, histone deacetylase.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
being evaluated, including those interfering with EGFR signaling, cell-cycle regulation (cyclin-dependent kinase inhibitors and serine-threonine kinase inhibitors such as staurosporine), histone acetylation (trichostatin A), and angiogenesis inhibitors (endostatin) (Bange et al. 2001). Since breast cancer is a clinically heterogeneous disease with a limited number of therapeutic choices available, a major challenge is to identify new potential therapeutic targets to treat the varied forms of the disease.

The most integrated approach towards understanding multiple molecular events and mechanisms by which cancer may develop is the application of gene expression profiling using microarray technologies. This review focuses on the recent advances in breast cancer research using microarray technology to (1) establish molecular profiles for tumor classification with correlation to clinical outcome, (2) identify critical genes and pathways of oncogenesis in order to identify novel biomarkers and candidate targets for intervention, and (3) predict drug response of human breast tumors. In addition, we discuss the application of gene expression profiling to mouse models of human breast cancer as a means of validating their use in studying human breast cancer, and their potential use in preclinical testing of novel drugs.

Gene expression profiles of breast cancer

It has become clear that the molecular alterations in breast cancer are complex and involve cross-talk between multiple cellular signaling pathways. Advances in genomic technologies and basic sequence of the human genome have paved the way for the identification of tens of thousands of genes. The advent of microarray technology provides a means of capturing and comparing the expression patterns of the entire genome of multiple samples in a high throughput manner. Entire networks of regulatory pathways can be studied using a systems biology approach and new genes associated with the oncogenic process may be identified. Several platforms for gene expression studies are currently available and are briefly discussed below.

**cDNA microarrays**

The first arrays with 31 104 cDNA clones were spotted onto Nylon membranes and radioactively labeled cDNAs from two sources were used for comparative hybridization (Drmanac & Drmanac 1994). Schena et al. (1995) first adapted printing cDNA clones onto glass slides and studied the expression of 45 Arabidopsis genes in two tissues, the root and the stem. Since then hundreds of thousands of cDNA clones representing named and unnamed expressed sequence tags (EST) have been identified, collected, and made publicly available, primarily through the Cancer Genome Anatomy Project of the National Cancer Institute. cDNA extracted from selected clones are robotically spotted onto a chemically treated glass slide to produce a cDNA array which may contain more than 25 000 different cDNAs (Eisen & Brown 1999). The arrays are used in a competitive hybridization reaction where a sample of interest (e.g. tumor) is directly compared with a reference sample (generally normal tissue or cell lines). In order to accomplish this, RNA isolated from the samples of interest and the reference sample is differentially labeled using two distinguishable fluorescent dye-conjugated deoxyuridine triphosphate (d-UTPs) by reverse transcription. The resultant labeled cDNAs are mixed in equal quantities and competitively hybridized to the cDNA microarray (Fig. 1).

Pair-wise fluorescence measurements are carried out for each set of samples, recorded as intensity units in two channels representing the individual dyes, and the ratios of the relative intensities are calculated for each feature of the microarray. This process is repeated for multiple samples. Since the reference RNA is constant throughout the set of experiments, it serves as an internal control and allows one to compare the relative ratios of gene expression between different array experiments. Most cDNA clones spotted onto cDNA microarrays represent the unique 3′ untranslated region of the genes, helping to minimize the chance of cross-hybridization, especially within large families of homologous genes. The use of such clones has several drawbacks, including clone mis-identification, the laborious effort required to determine the full-length cDNA sequences of the genes represented by unknown ESTs, and functional annotation of the genes. As additional sequence information is obtained and functional annotation of genes is improved, some of these problems can be better addressed.


**Oligonucleotide arrays**

Oligonucleotides designed to be gene specific have been used to produce arrays by several means. In one method, 25-mers are chemically synthesized onto a glass slide using photolithographic technology (Affymetrix Inc., Santa Clara, CA, USA) (Fodor et al. 1993). Another approach builds oligonucleotides on slides using inkjet printing technologies (Rosetta Inpharmatics, licensed to Agilent Technologies, Palo Alto, CA, USA) (Hughes et al. 2001). The direct spotting of oligonucleotides up to 70 bp in length is also being performed. Another technology allows for the deposition of oligos into a gel matrix applied to glass slides (Amersham Life Sciences) (Ramakrishnan et al. 2002).

Total RNA samples are reverse-transcribed into cDNA using an oligo-dT primer having the T7 polymerase promoter sequence at the 5′ end. Unlike the cDNA microarray protocols, uniform labeling of the cDNA is not performed in these
Platforms for microarray analysis: a schematic representation of sample preparation protocols for cDNA and oligonucleotide microarray are shown. (A) cDNA arrays: sample preparation involves direct incorporation of cyanine-3 (Cy3) and cyanine-5 (Cy5) dyes during the reverse transcription reaction using oligo-dT primer. The differentially labeled reference and tumor cDNAs are pooled and hybridized to a glass slide spotted with cDNA clones. A dual channel pair-wise fluorescence measurement is carried out for each spot and the relative abundance of the mRNA in the tumor and reference mRNA sample is obtained as a ratio of Cy5/Cy3 intensities. (B) Oligonucleotide arrays: sample preparation for this platform involves multiple steps: (1) reverse transcription using an oligo-dT-T7 RNA polymerase primer, (2) second strand synthesis, (3) in vitro transcription using T7 RNA polymerase in the presence of biotinylated rNTPs, (4) hybridization of cRNA to oligo-chip, and (5) detection of hybridized cRNA using a fluor- and streptavidin-conjugated antibody. Each reference and tumor sample is labeled similarly but hybridized to independent oligo-chip arrays. Thus, laser scanning and intensities of signal on each oligo-chip represent absolute data from a single sample. Data from two different oligo-chips are used to quantify the relative abundance of mRNA for each gene between the two samples. rNTPs, ribonucleoside triphosphates.
protocols. The cDNA is then converted to double-stranded DNA (dsDNA). This dsDNA is then used for in vitro transcription using T7 RNA polymerase to generate cRNA in the presence of biotinylated rNTPs. Each sample of biotinylated cRNA is individually hybridized to the oligo-chip. The signal is detected by the biotin–streptavidin detection system wherein streptavidin is conjugated to a suitable fluorophore, the abundance of which can be determined using a laser scanner (Fig. 1). Thus, oligonucleotide arrays have an additional step of target RNA amplification using in vitro transcription methods, leading to loss of the most linear relationship between the samples studied. In addition, a second loss of linearity occurs during the detection of the hybridized cDNA. Whether such processing leads to a bias in the overall data or a sub-set of genes is debatable. Amplification, however, allows one to perform array analyses on extremely small samples such as early human tumor lesions and biopsies and use of proper controls can reduce such bias in the data (Lockhart & Winzeler 2000).

Both cDNA and oligonucleotide-based microarrays have some technical difficulties and are limited by their range of sensitivity. In addition, saturation of the signal may occasionally cause a problem. Since cDNA microarray data are presented as a comparative analysis of two samples hybridized to the same array, the output is dependent on the ‘reference’ or the ‘control’ sample used. Comparisons of data between research groups becomes difficult and statistically challenging when different reference RNAs are used by different groups. Early generations of cDNA clone sets had high rates of clone failures due to incorrect clone identification and clone contamination. Since cDNA clones are amplified by PCR, the relative amounts of spotted DNA available for probe hybridization may vary from clone to clone, thus incorporating spots with a ‘weak’ signal that does not correspond to the actual amounts of the gene-specific transcript in the samples. A more uniform amount of DNA can be immobilized to the array for each oligo thus reducing ‘weak’ signals. The use of oligonucleotides for arrays promises to overcome some of these potential problems since they are chemically synthesized and there is less chance of cross-contamination. Some oligonucleotides offer the advantage of performing single sample hybridization. Each spot on the array corresponds to the abundance of that RNA species in an individual sample, unlike the relative measurement obtained in cDNA arrays. In theory, these single channel arrays can be comparatively analyzed with data from different research groups, although additional technical issues must be rigorously addressed.

With the wealth of sequence information available for the human and mouse genome, oligo-based arrays can be designed with greater precision. However, this information must be coupled with information from transcribed sequences to identify appropriate target sequences. An additional advantage of oligonucleotide arrays is that the strand specificity of the hybridizing probe is known, which may not be the case for double-stranded cDNA arrays. In certain cases, particular oligonucleotides have not performed well probably due to additional factors such as secondary structure. Despite the potential technical problems, oligonucleotide-based arrays may become the platform of choice in the near future.

Statistical analysis and validation of microarray data

The most critical aspect of microarray analysis is the retrieval of biologically relevant data and gene expression patterns that assist in harnessing the vast amount of information collected from a myriad of spots (Simon et al. 2002). Analysis of the data can be broadly divided into two steps. The first concerns the selection of ‘good spots’ that provide reproducible data and is influenced by slide quality, signal-to-noise ratio, sample preparation, and labeling. The criteria for selection are as simple as spot size and intensity, or as complex as issues of normalization of intensities in two channels. The latter aspect is of particular importance in cDNA arrays, as the two fluorophores have differential labeling and detection efficiencies, leading to dye bias and/or misleading data.

Normalization of data in both channels is therefore carried out either by several methods using global normalization (for example as in GenePix software, Axon Instruments, Foster City, CA, USA) (Alter et al. 2000, Wang et al. 2002) or by using several housekeeping genes and adjusting their ratio to 1 (Ermolaeva et al. 1998). Certain problems with microarray data may be reduced by generating data from duplicate samples, which may not always be possible because of prohibitive costs or lack of sufficient sample. In addition, the dye bias in experiments can be reduced by dye-swapping (reverse-labeling) experiments and incorporating this information into the analyses. Thus, good experimental design, number of samples tested with replication, relevance of the question asked, and stringent selection of spot data will considerably improve the quality and meaning of the data (Draghici et al. 2001, Kerr & Churchill 2001).

Following the selection of good spots, the second biggest challenge lies in discerning the biological meaning of the array data in order to derive patterns for classification and gene discovery and to formulate new hypotheses. No exact rules exist regarding the statistical approaches required for such analysis, although various methods and packages dedicated to microarray technology are continually being developed (Burke 2000, Bier & Kleinjung 2001, Li & Hong 2001, Miles 2001, Raychaudhuri et al. 2001). Software for microarray analysis is available at http://linus.nci.nih.gov/BBR-ArrayTools.html, http://genome-www4.stanford.edu/microarray/SMDDrestech.html and http://www.tigr.org/dbi/microarray. Commonly used statistical techniques have been extensively reviewed (Sherlock 2000, Quackenbush 2001).
and critical issues of microarray data analysis are more thoroughly considered elsewhere (Simon et al. 2002 and http://bioinformatics.duke.edu/camda). The extraction of relevant information from array experiments generally requires the rigorous use of an appropriate combination of statistical tests with due consideration to their biological relevance.

Validation of microarray data

Although microarrays are extremely powerful in documenting whole pathway changes and alterations in multiple genes simultaneously, technical difficulties do compromise the quality of data to a certain extent. In order to improve the certainty of the observation, validation of the gene expression pattern either by Northern blot analysis, real-time PCR or immunohistochemistry is advisable (Bustin 2000, Torhorst et al. 2001). Microarray data should be reproducible, although some variation is observed in the exact value of change between Northern blot analysis and fluorescence measurements.

Tumors from germ-line breast cancer related antigen (BRCA) 1 and BRCA2 mutation carriers

Less than 5% of all breast cancer cases are due to BRCA1 and BRCA2 genes which predispose to the disease, although it has been estimated that genetic factors contribute to about 35% of breast cancer cases (Nathanson et al. 2001). Common mutations associated with a high penetrance of breast cancer have been found in the BRCA1 and BRCA2 genes.

The BRCA1 gene encodes a nuclear protein with a zinc-finger RING motif at the N-terminal region and contains an acid-rich C-terminal region (Miki et al. 1994). BRCA1 interacts with several important regulatory proteins including p21WAF1/CIP1, murine double mutant 2 (MDM2), BclX, BRCA2, p53, and RAD51 (Scully & Livingston 2000). No recognizable structural protein motifs have been found in the BRCA2 protein, however, making it difficult to predict its cellular function. BRCA2 has been shown to associate with BRCA1 (Wooster et al. 1995). BRCA2 knockout mice have no thymocytes, are devoid of germ cells and display chromosomal segregation abnormalities. Although, mutations in either BRCA1 and BRCA2 genes confer a lifetime risk of 50–85% for breast cancer, BRCA1 mutation tumors differ histologically from BRCA2 mutant tumors (Ford et al. 1994, 1998, Lakhani et al. 1998, Thorlacius et al. 1998). The BRCA mutation-positive tumors are associated with ER- and PR-negative status, have lymphocyte infiltration, and a more aggressive clinical course (Loman et al. 1998).

Hedenfalk et al. (2001) compared gene expression profiles of seven tumors from BRCA1 mutation carriers, seven tumors from BRCA2-positive patients and seven sporadic tumors. The authors selected a sub-set of 51 genes by a modified F test, whose variance among samples best discriminated between the three groups. Secondly, the authors used a class-prediction method to determine if gene expression profiles can correctly classify BRCA1 and BRCA2 mutation-positive tumors. This test correctly classified all BRCA1 mutation-positive tumors, but misclassified one BRCA2 mutation-positive tumor. Nine genes were differentially expressed between BRCA1 mutation-positive and BRCA1 mutation-negative tumors, whereas 11 genes were differentially expressed in BRCA2-positive and BRCA2-negative tumors. As the authors conclude, analysis of a larger number of samples and larger gene sets may be required to improve the gene number in the class predictor and the robustness of class prediction tests. The authors then used three statistical tests (modified t-test, weighted gene analysis, and mutual-information scoring) to select 176 genes that distinguished between the BRCA1 mutation- and BRCA2 mutation-positive tumors. Of the genes selected, BRCA1 mutations appear to affect genes in the DNA repair and apoptotic pathway displaying a ‘stress’-type state. An important observation of these experiments is that BRCA1 and BRCA2 mutations dictate cellular pathways distinct from those observed in sporadic tumors.

Sporadic breast cancer

The high variability in breast tumor incidence, multiplicity, age of onset, and altered molecular pathways is attributed to differences in environmental, dietary, and genetic factors. Sporadic breast cancer is the most prevalent form of breast cancer and is thought to arise from the accumulation of multiple mutations or alterations in the expression of important cell regulatory genes, including oncogenes like erbB2/her2, c-myc, EGFR, and p53. Microarray analysis is a powerful method to assess the effects of these multiple molecular alterations on the expression of thousands of genes leading to more precise tumor classifications and predictions of clinical outcome.

Perou et al. (1999, 2000) performed gene expression profile studies on 20 matched pairs of breast tumors before and after doxorubicin treatment along with 44 other breast tumor samples. A set of 1753 genes was selected on the basis of a fourfold change in gene expression from the median abundance of a transcript in the sample set. In addition, another 476 clones (referred to as an ‘intrinsic’ gene set) were selected to show differences between tumors from different patients rather than those between paired tumor samples before and after doxorubicin chemotherapy. The tumors were segregated into four major groups based upon their expression patterns: (1) basal cell-like: expressing keratin 5/6 and 17, B4 and laminin, low ER, (2) erbB2-positive cluster: expressing high levels of erbB2 but low to absent ER, (3) normal breast-like: expressing genes characteristic of normal breast tissue, and (4) luminal cell-like: high ER levels, LIV-1
protein, GATA-binding protein 3, prolactin receptor and carnitine palmitoyl-transferase II. Interestingly, the gene expression profiles before and after chemotherapy were remarkably similar to each other in 15 of the 20 pairs studied. Of the remaining five pairs, three pairs of tumor samples that displayed ‘normal-like’ gene expression profiles had clinically ‘responded’ to the doxorubicin therapy.

In a subsequent study, Sorlie et al. (2001) studied 78 breast tumors, three fibroadenomas and four normal breast samples to determine if a correlation between microarray-based tumor classification and clinical outcome could be established. The intrinsic subset of 476 genes resulted in a similar pattern of separation of tumors as described by Perou et al. (1999, 2000) but the luminal subtype displayed two major sub-groups, the luminal type A, and another group, the luminal type B and type C. The luminal type A had high levels of ERα expression along with GATA-binding factor whereas the type C tumors had genes with no known coordinated function. The authors evaluated the association of these tumor subtypes with disease-free and overall survival. The status of p53 mutations and erbB2 over-expression in the tumors was also considered since these are associated with poor clinical outcome and resistance to chemotherapy. Of the luminal A sub-type, only 13% of the patients (4/30) showed mutations in p53, whereas 71% of the basal epithelial-like (5/7) and 82% erbB2-positive tumors (9/11) harbored mutations for p53. The latter two groups relapsed more quickly and had reduced overall survival compared with the group with fewer p53 mutations. Interestingly, the luminal sub-types differed significantly in their outcome of disease-free and overall survival, thereby suggesting that this classification had novel clinical implications.

Several other microarray studies have found that breast cancers tend to segregate into clusters depending upon whether tumors are ER+ or ER− (Gruvberger et al. 2001, Van ’t Veer et al. 2002), demonstrating that ER status is associated with distinctive gene expression patterns. To determine the minimum set of genes whose expression patterns could stratify tumors into ER+ and ER− groups, Gruvberger et al. (2001) analyzed the gene expression profiles of breast cancers from 58 node-negative patients, of which 23 were ER+ and 24 were ER−. The expression patterns of 3389 genes studied were altered in a statistically significant manner. This set of genes was used to train artificial neural networks (Khan et al. 2001) to define a sub-set of signature genes whose expression patterns could correctly predict whether a tumor is ER+ or ER−. After establishing a ranked order of these signature genes, the authors used five ER+ and 6 ER− tumors to test the robustness of their classification. This analysis resulted in the correct classification of the tumors as being ER+ or ER−. Using serial analysis of gene expression (SAGE) analysis data from MCF-7 cells, treated or untreated with estrogen, the authors demonstrated that several genes that belonged to the ER+ set of signature genes were not regulated directly by estrogen signaling, suggesting that other regulatory pathways may be involved in the expression of those genes (Khan et al. 2001). Alternatively, estrogen may affect gene expression of stromal cells or other cell types in the vicinity of the tumor that could not be demonstrated in the MCF-7 assay.

Although breast tumors may be classified similarly based upon traditional pathology criteria, response to treatment and clinical outcome can vary widely. The prediction of response to alternative therapies based upon gene expression profiling would be of immense clinical value and help determine the best care for individual patients. Another study has used microarray analyses to associate gene expression signatures with the natural history and clinical outcome of breast cancer. Van ’t Veer et al. (2002) studied 98 tumors from node-negative patients under the age of 55 years. Of the 98 tumors, 34 displayed distant metastasis in 5 years, 44 remained disease free, 18 had germ-line mutation in BRCA1 and two were BRCA2 carriers. Out of the 25 000 genes analyzed, about 5000 genes were regulated in a manner that led to the classification of the tumors into two distinct groups containing 62 and 36 tumors respectively. Most of the tumors in the first group retained ER expression and only 34% of the tumors displayed distant metastasis. The second group contained 34 ER-negative sporadic tumors, most with lymphocyte infiltration, and 16 of the 18 BRCA1 tumors. Seventy percent of the patients with sporadic tumors in the second group had progressive disease.

Van ’t Veer et al. subsequently used 78 tumors with known clinical outcome to determine gene expression signatures predictive of good and poor prognosis via artificial neural networks. A subset of 231 genes appeared to be associated with disease outcome and were used to further define an optimal ‘classifier’ gene set capable of correctly differentiating between ‘good’ prognosis and ‘poor’ prognosis patients. The selected 70 genes that defined this optimal set misclassified three tumors with poor prognosis and 12 tumors from disease-free patients into opposite groups. Thus, despite having a ‘good’ prognosis expression profile, three patients developed distant metastases within 5 years. This observation suggests that small primary tumors negative for node metastases can display a poor prognosis signature and may already be programmed for metastasis. However, individual genes previously used to predict breast disease progression like Erbb2 and c-myc were not included in this list. Thus, gene expression profiles may be utilized to determine prognostic classifiers and to predict candidacy for certain therapeutic treatments. The major features of human microarray studies are summarized in Table 2.

**Gene expression profiles of mouse models of human breast cancer**

The development of genetically altered mouse models for human cancer has provided key insights into genetic regu-


Desai et al.: Breast cancer microarrays

<table>
<thead>
<tr>
<th>Study</th>
<th>Patient cohort</th>
<th>Unique identifiers</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hedenfalk et al. (2001)</td>
<td>7 BRCA1 mutation-positive; 7 BRCA2 mutation-positive; 7 sporadic breast tumors (ER+ or ER−)</td>
<td>(1) 51 genes to distinguish BRCA1 mutation-positive, BRCA2 mutation-positive and sporadic tumors (2) 176 genes to distinguish BRCA1 mutation-positive from BRCA2 mutation-positive tumors</td>
<td>Mutation in BRCA1 or BRCA2 gene results in gene expression profiles distinct from sporadic tumors</td>
</tr>
<tr>
<td>cDNA microarray, 6512 features</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perou et al. (2000)</td>
<td>42 patients all with locally advanced disease (36 infiltrating ductal carcinomas, 2 lobular, 1 DCIS, 1 fibroadenoma, 3 normal breast), of these, 20 pairs of tumors were sampled before and after 16 weeks of doxorubicin chemotherapy</td>
<td>Tumors before and after therapy had similar expression patterns: intrinsic gene set of 476 cDNAs could classify the tumors into four major groups (see text for details)</td>
<td>Identification of a novel sub-group of basal-like breast tumors</td>
</tr>
<tr>
<td>cDNA microarray, 8102 features</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sorlie et al. (2001)</td>
<td>Total 78 breast carcinomas (71 ductal, 5 lobular, 2 DCIS); 3 fibroadenomas; 4 normal breast samples</td>
<td>Using the ‘intrinsic’ gene set, identification of novel luminal-type sub-classes luminal A, luminal B and luminal C</td>
<td>Basal-like, erbB2, luminal sub-type B-C had worse clinical outcome: luminal A sub-type had good clinical outcome</td>
</tr>
<tr>
<td>cDNA microarray, 8102 features</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gruvberger et al. (2001)</td>
<td>58 node-negative patients, 23 were ER+ and 24 were ER−</td>
<td>Developed a class predictor of 100 genes that best distinguish ER+ and ER− breast tumors</td>
<td>ER presence or absence determines distinct gene expression patterns</td>
</tr>
<tr>
<td>cDNA microarrays 6728 features</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Van ’t Veer et al. (2002)</td>
<td>98 tumors, 34 with metastasis in 5 years, 44 disease free, 18 with BRCA mutation, 2 with BRCA2 mutation: sporadic patients were lymph-node negative and under 55 years of age</td>
<td>(1) 5000 genes that classified the tumors into 2 dominant ER+ and ER− groups (2) Identified a class predictor with 231 genes that could predict clinical outcome</td>
<td>Predictor may be useful in defining therapy regimens</td>
</tr>
<tr>
<td>Oligonucleotide array with 25 000 genes</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

DCIS, Ductal carcinoma in situ.

About 70% of human breast cancers have deletions in the short arm of chromosome 17 in the vicinity of the p53 gene. An association between poor disease prognosis and p53 mutations, like those found in familial breast cancer (Li-Fraumeni syndrome), and retinoblastoma gene 1 (Rb-1) mutations in ER+ tumors has been documented earlier (Runnebaum et al. 1991, Berns et al. 1995, Lehman et al. 2000). The targeted overexpression of the SV40-T-antigen results in mammary cancer, at least in part, by the inactivation of both the p53 and Rb genes (Tzeng et al. 1993, Maroulakou et al. 1994). Overexpression of another viral oncoprotein, polyoma middle T-antigen, also results in mammary cancer in transgenic mice and requires the src signaling pathway (Guy et al. 1992, 1994). Important mammary cancer models have been developed by the functional deletion of p53 through homologous recombination. Similarly, conditional loss of BRCA1 with a concomitant loss of p53 leads to mammary tumor formation (Koh et al. 1995, Barrington et al. 1998, Brodie et al. 2001, Kavanaugh et al. 2002). The pathologic characteristics of these and several other transgenic models of mammary cancer have been classified and compared with those arising in human breast cancers (Cardiff et al. 2000b). It is clear, however, that no existing
model identically matches all of the histological and biological properties of human breast cancer, but that they can be validated for representing specific aspects of human breast cancer.

The use of these mouse models enables investigators to study the initiation and progression of spontaneously occurring tumors in animals with intact immune systems and uniform genetic backgrounds, thus avoiding some of the challenges in studying human cancer. Moreover, molecular events elaborated by a single oncogenic stimulus can be studied singly or in combination with other genetic alterations. Histopathologically, a number of similarities as well as differences have been observed between mouse mammary tumors and human breast cancer (Cardiff et al. 2000). As opposed to heterogeneous lesions that appear in human disease, the transgenic mouse tumors generally have characteristic histologies depending upon the expressed event. This character may be attributed to the presence of a uniform genetic lesion in all the tumor cells. Moreover, the mouse strain background contributes significantly to the biology, onset and growth of the tumor (Lifsted et al. 1998).

Another potential drawback of using transgenic mouse models is the lack of tumor metastases to bone and other common sites of breast cancer dissemination (Pattengale et al. 1989). Transgenic mouse tumors metastasize primarily to the lung. Direct injection of cells into the left ventricle to mimic metastasis resulted in specific homing of the tumor cells to the pulmonary site and orthotopic models to mimic metastasis resulted in specific homing of the tumor cells. At least at the late stages of tumor lesions, the transgenic tumor metastases to bone and other common sites of breast cancer dissemination (Pattengale et al. 1989). However, though transgenic mice overexpressing activated neu oncogene develop ER− tumors may be better suited to study hormone-dependent cancer (Nandi et al. 1995). However, though transgenic mice overexpressing activated neu oncogene develop ER− tumor growth. On the other hand, a similar treatment of animals at 12 weeks of age resulted in reduction of tumor incidence by 50% (Menard et al. 2000). Thus, the status of ER and its effect on tumor development in transgenic mice may differ. In addition, hormone profiles and their differential influence on mammary gland development across species may contribute to fundamentally different tumor biology (Nandi et al. 1995).

In an effort to begin the classification of the mammary cancer models based upon gene expression profiling and validate them as systems to study human breast cancer, we performed cDNA microarray analysis on late-stage tumors excised from six mouse cancer models (Desai et al. 2002). Our study has revealed sets of genes whose expression was commonly altered among all of the tumor models, as well as gene expression signatures that were specific to the oncogenes used to initiate the cancer.

Cancer genes

We compared RNA isolated from multiple mouse tumors from each model with RNA from normal mouse mammary glands in order to determine mammary cancer-specific gene expression changes. A total of 8668 unique genes were analyzed by microarray using this approach. A twofold change in gene expression was used as the cut-off to identify genes differentially expressed between the normal mouse mammary glands and mammary tumors from different mouse models. A total of 930 genes were selected and hierarchical clustering analysis demonstrated that the mouse models had extremely similar gene expression profiles irrespective of the initiating oncogenic event. The ‘cancer-gene’ profile included many genes involved in cellular metabolism, cell-cycle regulation, adhesion, angiogenesis and apoptosis. In addition, numerous ESTs were uniquely regulated in the tumors, thereby identifying several novel potential targets for intervention. Induction of several zinc-finger transcription factors, protein tyrosine phosphatases and suppression of many protein tyrosine kinases appeared to be a common phenomenon in all of the mammary tumors studied.

Oncogenic signatures

Initial cluster analysis of the mouse tumor data sets using the subset of ‘cancer genes’ revealed that, in general, the gene expression profiles between the tumor models were quite similar. This indicated that, at least at the late stages of tumorigenesis, many molecular changes occurred in common, despite the fact that the oncogenic event that initiated the tumor process affected different pathways. In order to determine what features could be utilized to distinguish the tumor models in a more meaningful and insightful manner, a second statistical analysis was performed using the F-test. This approach identified a subset of 900 genes whose expression varied significantly between the different classes of tumor models, suggesting that oncogene-related signature genes could be identified.

Hierarchical clustering of this subset of genes resulted in the segregation of the mouse mammary tumors into three groups: (1) the SV40 T/t-antigen (T-ag) tumors, (2) myc tumors, and (3) Neu/ras/PyMT-derived tumors. The T-ag and myc tumors displayed some overlapping gene expression patterns limited to the cell-cycle pathway and were more closely related to each other than to the Neu/ras/PyMT group. The tight clustering of the latter group may be a result of the convergence of the neu and PyMT signaling pathways into the ras pathway (Dankort & Muller 2000).
Both T-ag and myc induce expression of cyclin B1, cdc25, and mcm6 family genes, a profile characteristic of loss of both G1/S and G2/M cell-cycle transition checkpoints. Moreover, higher S-phase fractions in the myc tumors have been documented and both T-ag and myc-derived tumors have a higher mitotic index as compared with other tumor models (Hundley et al. 1997, Maroulakou et al. 1999, Cardiff 2001). However, the neu-ras-PyMT group had elevated expression of cyclin D1-, cyclin E- and E2F-related genes, suggesting a passage through the G1/S check-point (Frame & Balmain 2000). Of the models studied, T-antigen appeared to have the largest and most highly diverse set of signature genes spanning several cellular pathways, including the unique regulation of calcium pathway genes. The most unique features of the oncogene signatures are summarized in Table 3.

The application of microarray technology now allows for the detailed molecular dissection of mouse models of mammary cancer, which will provide important information to help select appropriate models for particular experimental purposes. These analyses will provide important information to further validate these models of human breast cancer.

### Table 3 Oncogenic signatures in mouse models of cancer.

<table>
<thead>
<tr>
<th>Mouse model</th>
<th>Gene%</th>
<th>Oncogenic signature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tag</td>
<td>19</td>
<td>Largest gene set spanning all cellular pathways. Unique regulation of calcium pathway genes.</td>
</tr>
<tr>
<td>Tag and Myc</td>
<td>9.9</td>
<td>Mostly cell cycle-related genes – cyclin B, PCNA, ribosomal RNA genes, transcription factors, homeobox genes</td>
</tr>
<tr>
<td>Myc</td>
<td>2.7</td>
<td>Ribosomal RNA genes, transcription factors, homeobox genes</td>
</tr>
<tr>
<td>Neu</td>
<td>2.4</td>
<td>Breast tumor antigens of tetraspanin family, proteases-matrix metalloproteinases, calpains, transcription factors</td>
</tr>
<tr>
<td>Neu, Ras and PyMT</td>
<td>6.6</td>
<td>Group specific signature – including cyclin D1, GTPases, GAP</td>
</tr>
<tr>
<td>Ras</td>
<td>0.5</td>
<td>No pathway-specific changes</td>
</tr>
<tr>
<td>Neu and PyMT</td>
<td>1.3</td>
<td>No pathway-specific changes</td>
</tr>
<tr>
<td></td>
<td>57.6</td>
<td>Genes that do not show specific patterns of gene expression but had variance between models</td>
</tr>
</tbody>
</table>

GAP, GTPase activating protein

Our mouse microarray studies have identified changes in gene expression that have been previously implicated in human breast cancer. Although the transgenic mouse tumors are largely ER−, the genes identified in the mouse models encompass gene expression changes observed in both ER+ as well as ER− human breast tumors. Despite these differences in the mouse models and human breast disease, it is important to note that genetic lesions identified in human breast disease often lead to mouse mammary cancer. Interestingly, the molecular events associated with the oncogenes/genetic event are often identical in both species (Webster & Muller 1994). Thus the comparison of gene expression profiles elicited in the transgenic mouse models will significantly improve our understanding of the breast disease at a molecular level.

### Comparative analysis of mouse mammary cancer and human breast tumors: challenges

Gene expression profiles of both human and mouse mammary tumors have now been established. However, the principle aims of the existing mouse and human gene profiling studies have not been directed towards inter-species comparisons of gene expression. For example, early versions of human and mouse microarrays have not been designed to contain orthologous genes or EST clones. Consequently, there are a relatively small number of genes common between the mouse and human microarray platforms. A second major technical obstacle in comparing mouse and human array data sets is related to the use of different reference RNAs in the mouse and human experiments. However, efforts are ongoing to meet these challenges. New generation mouse and human arrays will have significantly more gene overlap as arrays are designed with the latest genomic information.

Along with improvements in unigene clustering, genome sequences, gene annotations and chromosome mapping, direct comparisons between mouse and human expression arrays will be possible. These comparisons require additional bioinformatic approaches that we are currently developing.

### Future directions

The exciting advances made possible by microarray analysis in the genome era have resulted in an unprecedented amount of information, which can be overwhelming and difficult to summarize (Simon et al. 2002). Although microarray analysis of breast cancer has provided valuable information for classifying tumors on a molecular basis and in predicting the clinical outcome and response to therapies, the identification of novel targets for intervention using this data is not straightforward. Each experiment may lead to the selection of a large set of genes with unknown functional significance.
in oncogenesis. Knowing the functional relevance of these genes, however, is critical for choosing potential therapeutic targets. Other groups of genes may be useful ‘biomarkers’ for cancer or predictors of outcome. Moreover, lack of complete sequence information and gene function for thousands of ESTs further limit the interpretation of data.

At present, data from microarray analysis of grossly dissected human breast or mouse mammary tumors are available. Different stages of tumors often contain a mixture of varying proportions of epithelial and stromal cells, blood vessels, adipose and connective tissue. Isolation of cells obtained from fresh tissue in culture may define a purer population of cells, but pertinent information regarding intercellular cross-talk may be compromised. Thus, gene expression patterns derived from either of these samples may not accurately represent the molecular events associated with the disease. A more refined sampling of cells using laser capture microdissection (LCM) at particular stages may improve our basic knowledge of the initial processes involved in cancer ( Emmert-Buck et al. 1996, Bonner et al. 1997). Using this technique, highly homogenous populations of cells can be obtained from a complex milieu of ‘diseased’ and ‘normal’ disease-free tissue. In addition, sampling of adjacent cells may address questions pertaining to cell–cell interactions. Moreover, LCM offers the opportunity to isolate cells that may constitute only a microscopic fraction of the tissue. LCM samples have been used for further genetic analyses using RT-PCR, microarray loss of heterozygosity studies in LCM samples have been used for further genetic analyses of the same tumors by proteomics and genomics. These developments will further define molecular mechanisms of tumor development, which will lead to an increased understanding of breast cancer and the identification of novel targets for therapy.

References


Bouchard L, Lamarre L, Tremblay PJ & Jolicoeur P 1989 Molecular analysis of fine-needle aspiration and core biopsy specimens have been evaluated by investigators (Ellis et al. 2002, Sotiriou et al. 2002a,b). Reproducible RNA amplification protocols that utilize small amounts of RNA obtained by such methods are now available (Van Gelder et al. 1990, Wang et al. 2000).

Microarray technology, though powerful in determining changes at the transcription level, provides no information at the protein translation level. Thus, future studies need to be geared towards functional annotation of genes and parallel analyses of the same tumors by proteomics and genomics. These developments will further define molecular mechanisms of tumor development, which will lead to an increased understanding of breast cancer and the identification of novel targets for therapy.
Desai et al.: Breast cancer microarrays


Li H & Hong F 2001 Cluster-Rasch models for microarray gene expression data. *Genome Biology* 2 RESEARCH0031.


Desai et al.: Breast cancer microarrays


Sherlock G 2000 Analysis of large-scale gene expression data. *Current Opinion in Immunology* 12 201–205.


Sotiriou C, Khanna C, Jazaeri AA, Petersen D & Liu ET 2002a Core biopsies can be used to distinguish differences in expression profiling by cDNA microarrays. *Journal of Molecular Diagnostics* 4 30–36.


