Molecular Profiling of Breast Cancer

Vincent A. Funari, Ph.D. ¹ and Stefanie S. Jeffrey, M.D. ²

¹Department of Surgery, Stanford University School of Medicine, Medical School Lab-Surge Bldg. Room P229, 1201 Welch Road, Stanford, California 94305-5494, USA
phone (650) 724-3519, fax (650) 724-3229

²Department of Surgery, Stanford University School of Medicine, Medical School Lab-Surge Bldg. Room P214, 1201 Welch Road, Stanford, California 94305-5494, USA
phone (650) 723-0799, fax (650) 724-3229

Email addresses for authors:
   Vincent A. Funari - vfunari@stanford.edu
   Stefanie S. Jeffrey - ssj@stanford.edu

Corresponding author: Stefanie S. Jeffrey (email: ssj@stanford.edu)
The human genome project and the development of high throughput technology over the last 5-10 years have thrust biology and medicine into a new era. Characterizing, diagnosing, and treating breast cancer using new molecular profiling techniques is a powerful patient-specific approach to treating and even preventing breast cancer. The technology is advancing rapidly and changes in the field occur often. This chapter will focus on the promises, progress and problems of molecular profiling in breast cancer.

**PROBLEMS WITH CURRENT METHODS**

At the present time, only a limited set of tumor parameters are used to estimate prognosis for a patient with breast cancer. In general, these include tumor type (ductal, lobular, medullary, mucinous, etc), size of invasive component, grade of the invasive component, the expression of hormone receptors including Estrogen Receptor (ER) and Progesterone Receptor (PR), the expression of the growth factor receptor HER2/neu (ERBB2), the presence and number of lymph node metastases, and any evidence of distant disease. In many areas of the U.S., measures of tumor proliferation, such as S-phase analysis or Ki67 expression are also determined. From these data, risk of distant relapse is assessed and recommendations for systemic therapy are given. Generally, almost all patients with lymph node metastases and the great majority of patients with lymph node negative invasive tumors greater than one centimeter will be candidates for systemic therapy. As a result, many women with stage I and II breast cancer, who may be cured by surgery and/or radiation alone, are over treated by systemic therapy. Other women are treated with systemic therapies that are ineffective against their specific tumor type. Further, many chemotherapeutic agents are non-specific, killing rapidly dividing cells in other organs (eg bone marrow or the GI tract). In general, currently used tumor parameters do not provide sufficient tumor-specific predictions for survival, need for systemic therapy, and drug response.
Although individual gene measurements (such as ER, PR, HER2/neu) have provided insightful information, it is now possible to measure global genetic changes using new technologies that provide a unique molecular profile or a fingerprint of the tumor. These multiple gene measurements represent a more comprehensive tumor signature that should provide more precise insights into a tumor’s clinical behavior, response to systemic therapy, or offer possible targets for the development of novel tumor-specific therapeutics.

**CONSTRUCTING A MOLECULAR PROFILE**

For precise characterization, breast tumors must be analyzed at all molecular levels: DNA, RNA, and protein. The goal is to identify tumor-specific features that molecularly subtype a tumor and then to correlate clinical outcome with molecular features. This would enable a patient and her physician to make specific decisions as to whether systemic therapy is indicated, and if it is, to use targeted therapies for treatment specifically aimed to kill or immobilize the molecular type-specific tumor cells.

There are four major steps in achieving accurate molecular profiling data. The first step is to obtain samples (from cell cultures, human tissues, blood, or body fluids) and purify the molecules of interest (DNA, RNA, protein). In the second step the DNA, RNA, or protein from the sample is measured. This usually involves constructing or purchasing a high throughput assay device, such as a microarray or protein chip, that can measure the presence or absence of hundreds to tens of thousands of genes, expressed genes, or proteins in a single sample. The third step involves data analysis using bioinformatics tools. This entails information storage and application of data processing algorithms to analyze and visualize the complex data. Finally, conclusions must be reached, validated, and translated into clinical applications.
DNA MOLECULAR PROFILES

Changes in chromosomal DNA occur in breast cancer. Identifying specific sites of DNA copy number change may identify candidate oncogenes or tumor suppressor genes. In contrast to methods such as loss of heterozygosity (LOH) and sequencing that traditionally have measured genetic modifications in specific genes or loci, comparative genomic hybridization (CGH)\(^2\) and array based CGH\(^3\)–\(^6\) map chromosomal or gene copy number changes on a global genomic scale. In CGH, tumor DNA and control DNA (isolated from peripheral blood lymphocytes from a healthy donor) are differentially labeled with fluorescent dyes and cohybridized onto normal metaphase chromosomes, also obtained from peripheral blood lymphocytes stimulated in vitro. The image is digitized and bioinformatic tools calculate the fluorescence ratio of tumor to normal genomic DNA. The ratio of fluorescence along the chromosome identifies regions of amplifications (gains) and deletions (losses) in the tumor DNA.

Chromosomal imbalances in breast cancer

CGH has identified multiple regions of chromosomal gains and losses in breast cancer. In primary breast cancers, chromosomal gains have been most frequently identified as whole arm gains in 1q and 8q and regional copy number increases at 17q and 20q.\(^7\) These data are, for example, consistent with known breast cancer oncogenes on chromosomes 8q (MYC) and 17q (HER2/neu [ERBB2]). In DCIS, chromosomal gains are observed in 1q, 8q, and 17q, whereas losses are most common in 8p, 11q, 13q, 14q, and 16q.\(^8\) In invasive breast cancer, gains of 1q, 6p, 8q, 11q, 16p, 17q, and 20q are most common. Chromosomal losses have been identified in 1p, 8p, 11q, 16q, 18q, and 22.\(^9\) Using CGH, Forozan and colleagues\(^10\) compared 38 established tumor cell lines to a meta analysis of CGH results from 698 primary tumors. In addition to the
chromosomal gains and losses mentioned for invasive tumors above, gains at 3q, 5p, 7p, 7q, 20p and losses at 4p, 18p, Xp, Xq were also found.

CGH may also be used to study tumor biology. Jain and colleagues\(^1\) studied the statistical relationship between CGH loci ratios and survival. Alterations in two loci, a gain at 8q24 and loss at 9q13, were associated with poor survival and were also associated with mutations in TP53, the tumor suppressor gene that codes for p53 protein. To study tamoxifen resistance, CGH has been used to compare a tamoxifen sensitive breast cancer cell line (MCF-7) and a tamoxifen resistant clone (CL-9).\(^1\) CGH findings revealed differential gains on chromosomes 2p, 2q, 3p, 12q, 13q, 17q, 20q, 21q and differential losses on chromosomes 6p, 7q, 11p, 13q, 17p, 18q, 19p, 22q. Neither ER-alpha on 6q25.1 nor ER-beta on 14q were involved in the differences. The authors suggest that this technique may be useful for identifying candidate genes involved in tamoxifen resistance.

*Characterizing cancer cell progression*

Beginning with usual ductal hyperplasia, there is evidence of accumulation of chromosomal aberrations that lead to invasive breast cancer.\(^13\)-\(^15\) Progression from hyperplasia to atypical hyperplasia to DCIS and finally to invasive breast cancer is thought to occur in a multi-step fashion.\(^16\) Consistent with this linear progression theory is that higher grade DCIS lesions demonstrate increased chromosomal aberrations with loss of differentiation.\(^17\)

However, others have argued against this linear continuum, and instead suggest alternative differentiation pathways of progenitor cells in the glandular tissue.\(^18\),\(^19\) In support of non-linear, independent pathways of genetic evolution in breast cancer, Buerger\(^8\) used CGH to study DCIS samples including all differentiation grades and some with associated invasive breast cancer. All cases showed chromosomal imbalances, identifying DCIS as a genetically advanced
lesion, with identical genetic lesions between the DCIS and invasive components in 83% of the cases. The most frequent chromosomal changes in well-differentiated DCIS were losses at 16q and gains at 1q. In contrast, high grade DCIS demonstrated losses at 8p, 11q, 13q, 14q and gains at 1q, 8q, 17q. Moreover, in 30% of DCIS cases with an invasive component, a gain of 11q13 was identified which was not present in pure DCIS. CGH was then performed on a larger population of intermediate and high grade invasive cancer. Chromosomal gains of 1q and 8q were seen in all invasive tumor grades. The loss of 16q, seen in well-differentiated DCIS, was not observed in the majority of poorly differentiated invasive cancers whereas more than half of intermediate grade DCIS showed this loss, suggesting that a subset evolved from well-differentiated DCIS and another subset evolved from poorly differentiated DCIS. Other chromosomal alterations, including gains at 8q, 17q and 20q and losses of 13q were found to be associated with poorly differentiated invasive carcinoma. Overall, this data suggests that invasive carcinoma recapitulates the genetic differentiation pattern of its precursor DCIS (low grade DCIS progresses to low grade invasive cancer and high grade DCIS progresses to high grade invasive cancer). Intermediate grade carcinoma may represent a mixture of DCIS subtypes evolving along different genetic pathways.

Array-based CGH analysis

While CGH provides a genome-wide view of chromosomal changes, its resolution is limited to measuring chromosomal imbalances of 10-20 megabases or more. Assuming about 10 genes per megabase, the resolution of conventional CGH spans about a 100-200 gene range. Array-based CGH is a high resolution alternative that can measure DNA copy number changes at the kilobase or gene level. For array CGH, tumor and normal genomic DNA are labeled with two different fluorescent dyes. The differentially labeled DNA is cohybridized to a microarray which
is a glass slide containing thousands of DNA elements. These elements can include either cDNAs (individual genes) or larger chromosomal segments that contain one or more genes with known chromosomal location, such as bacterial artificial chromosomes. The fluorescence ratio of tumor to normal DNA at each gene represents the copy number ratio between the two samples. Since gene expression studies may also be performed on similarly configured microarrays (see below), it is possible to directly correlate DNA copy number change and gene expression. 

Array-based CGH has been used to investigate previously recognized areas of amplification, such as chromosome 20q13, which had been characterized extensively by other techniques. The increased resolution of array CGH was able to identify two potential oncogenes, CYP24 and ZNF217, the former not previously associated with breast cancer. Pollack and colleagues used array CGH to study gene copy number changes and their correlation to gene expression. Interrogating 6,691 mapped human genes in locally advanced primary breast tumors and ten breast cancer cell lines, DNA chromosomal alterations were found in all samples with aberrations found in every chromosome. Gains were identified within 1q, 8q, 17q and 20q in a large proportion of tumors and cell lines; losses were observed within 1p, 3p, 8p, and 13q. A strong relationship between DNA copy number and gene expression was found, well exemplified by chromosome 17. Although gene amplification does not always yield an increase in gene expression, for highly amplified DNA regions, 42% were associated with high gene expression and 62% were associated with moderately high gene expression. This suggests that a tumor’s molecular phenotype is in large part impacted by underlying variation in DNA copy number. The authors estimate that overall 7-12% of variation in gene expression in breast tumors is due to variation in gene copy number. A study by Kallioneimi and colleagues had similar findings. Comparing DNA copy number and mRNA expression levels of 13,824 genes in 14 breast cancer
cell lines, they showed that 44% of highly amplified genes were overexpressed and 10.5% of the genes with high-level expression were amplified.

**RNA MOLECULAR PROFILES**

Since only a fraction of genes in a cell are expressed at any given time, the set of expressed genes (the gene expression profile) provides a snapshot reflecting that cell’s physiology and response to environmental influences. Differences in gene expression profiles can be used to define different molecular phenotypes of breast cancer, to predict the need for and responsiveness to systemic therapies, and to identify novel targets for tumor-specific therapies.

There are a number of reasons why RNA expression profiling has dominated the molecular profiling arena: (1) RNA is the product of an expressed gene and usually contains more functional significance than DNA, (2) protein assays are still in their infancy and sensitivity and precision require further optimization and validation, (3) classical RNA technologies were easily adapted to high throughput systems, and (4) conserved RNA properties facilitate amplification and measurement of minute amounts.

Before the genome project began, scientific methodology was candidate gene dependent, discovering and identifying one gene at a time was knowledge driven. High throughput technologies developed as part of the Human Genome Project changed this systematic methodology. Using these technologies, global gene expression profiles for thousands of known and unknown genes were determined in tissues before the genome was even sequenced.24, 25 The initial gene discovery methods, included Expressed Sequence Tags (ESTs, explained below), subtractive hybridization,26, 27 serial analysis of gene expression (SAGE),28 and differential display (DD),29 were developed based on universal RNA properties and available laboratory techniques without needing prior knowledge of an expressed gene’s function, sequence, or
chromosomal location. Using this technology, novel genes were identified at a more rapid pace than functions could be assigned. Today, approximately half of the expressed sequences (ie, genes) still have no assigned function, yet the abundance of gene sequence knowledge available from these techniques has enabled scientific focus to change from gene discovery to gene function. While these methods are powerful, they are technically difficult, require large-scale robotic sequencing instruments, and only allow study of a few different biological samples at one time. In contrast, DNA microarrays were developed in the mid-1990’s and have been used to measure RNA expression of thousands of genes from multiple samples at one time. They represent the quickest, easiest, and least expensive method to relate expressed genes to clinical data.

*ESTs*

An EST is a sequence of nucleotides that represents a portion of an expressed gene. It is obtained from automated sequencing of a cDNA library. A cDNA library is constructed by first isolating mRNA from a tissue sample of interest. The mRNA is reverse transcribed into complementary DNA (cDNA), which is then inserted into plasmids that are replicated in *E. coli* colonies on a nutrient-enriched plate. The colonies are randomly picked and the amplified cDNA is isolated and sequenced using an automated sequencer. A set of sequences from the same tissue sample is called an EST library.

If every cDNA clone is picked and sequenced, the entire transcript population of the cell (called a transcriptome) will be represented quantitatively and qualitatively in the EST library. The ESTs are matched by sequence identity to a database of known genes to determine if the expressed sequences have been previously identified. Thousands of unidentified genes have been discovered using EST technology. ESTs were the first successful functional molecular profiling
Corresponding Author: Stefanie S. Jeffrey

DNA Microarrays

Microarrays produce a gene expression profile by simultaneously measuring gene expression of hundreds to thousands of genes from a single sample. Known gene sequences are attached to membrane-based or glass arrays. Although more expensive than membrane-based arrays, glass arrays are smaller, easier to use, and allow a higher density of gene spots.

There are two types of glass arrays. One type is constructed using short 20-80 nucleotide fragments (oligonucleotides) to represent each gene. The oligonucleotides are synthesized in situ on the glass slide, using special lithographic\textsuperscript{32} or ink-jet printing\textsuperscript{33} technologies that were developed by Affymetrix Corporation or Agilent Technologies. Oligonucleotides can also be synthesized in batches prior to immobilization onto an array, which can reduce the cost. The second type of array, cDNA microarrays,\textsuperscript{30} contains partial to full length cDNAs, 500-5,000 nucleotides in length that are “spotted” on histological slides using robotics and fine print tips and then immobilized. The cDNAs consist of known and unknown genes, identified using EST technology. Oligonucleotide array technology is more expensive, but in general, demonstrated to be more precise and sensitive.

Total RNA (approximately 50 µg) or mRNA (3 µg) is used to measure expressed transcripts on cDNA microarrays. In general, total RNA or mRNA is isolated and reverse transcribed with fluorescently-tagged nucleotides to label the cDNA. For samples that do not
contain sufficient amounts of RNA for microarray hybridization, RNA amplification techniques can be employed.\(^\text{34}\)

Each spot (or feature) on a microarray corresponds to a specific gene or EST. Labeled cDNA from an experimental sample (eg, cDNA prepared from breast cancers and containing unknown quantities of specific genes, such as HER2/neu) is hybridized to the microarray. Excess or non-hybridized cDNA is washed off. Because of the specificity of base pairing at each feature, the abundance of a gene in the sample is measured. It is difficult to measure an absolute gene expression value on cDNA microarrays due to systematic differences in gene printing and hybridization kinetics. Therefore, reference RNA is used to generate a relative abundance ratio between the sample and a reference that allows gene-to-gene comparisons between different samples. Sample and reference RNA are labeled with different fluorophores (usually Cy5, which fluoresces red at 635 nm, and Cy3, which fluoresces green at 525 nm) and cohybridized to the microarray (Figure 86-1). The hybridized fluorescence signals can be read with an optical scanner. Using bioinformatics software, a fluorescence signal intensity ratio between the sample and the reference is computed. Signal intensity ratios provide a relative measure of gene abundance. Correlations can be made based on the gene expression similarity between independent samples.\(^\text{35}\) Genes or samples that demonstrate similar expression patterns are called clusters (Figure 86-2). Statistical analyses\(^\text{36-38}\) can be performed and related to pathological and clinical data to define samples or reactions to treatments.

**Characterizing breast cancer subtypes**

In 1999, human breast cancers were the first solid tumor to undergo global transcription analysis using microarrays.\(^\text{39, 40}\) Before these studies, it was not known whether the genetic and cellular diversity of solid tumors would preclude identifying gene expression patterns in breast
cancer. Despite the limited number of tumor samples consisting of different breast cancer types and grades, the small number of genes assayed, and lack of usual breast cancer-associated genes on the array (eg, HER2/neu and ER), Perou and colleagues\textsuperscript{39} identified multiple genes that were similarly expressed and implicated in the molecular phenotype of solid tumors. In a follow-up study, cDNA microarrays were used to molecularly subtype normal, benign and malignant breast tumors.\textsuperscript{41} Variations in growth rate, activity of specific signaling pathways, and cellular composition of the tumors were all reflected in gene expression profiles. This and follow-up studies\textsuperscript{42, 43} identified genes that divided the tumors into distinct molecular subtypes: two ER-overexpressing subtypes (denoted “Luminal A and B” due to presence of luminal epithelial cytokeratin markers) and three ER-negative subtypes: “basal-like” tumors that expressed cytokeratin markers characteristic of basal epithelial cells, “ERBB2 (HER2/neu)-overexpressing” tumors, and “normal-like” tumors that showed relatively high expression of genes characteristic of basal epithelial cells and adipocytes which clustered with normal breast tissue samples. The expression of known luminal and basal cytokeratin epithelial cell markers suggests that breast cancers may arise from at least two progenitor cell types through different mechanisms. Other studies\textsuperscript{44-48} have since demonstrated that ER and ER co-regulated gene expression (or lack thereof) provides a pervasive molecular signature marked by an abundant and robust gene expression. \textit{c-myc} is amplified in 15\% of breast cancers and is highly expressed in “basal-like” tumors, possibly regulating the expression of genes that play a role in the behavior of these tumors.\textsuperscript{48} Overall, these data suggest that groups of genes better characterize and refine tumor subtypes than single gene markers, like ER or HER2/neu.

Using 43,000 feature cDNA microarrays to profile histologically varied tumors from more racially diverse patient populations, our lab has identified additional molecular subtypes of
breast cancer. We have also shown that invasive lobular carcinomas may be classified into “typical” and “ductal-like” lobular tumors by their expression profiles.

**Subtype profiling of hereditary breast cancers**

Mutations in the breast cancer susceptibility genes, *BRCA1* and *BRCA2*, influence DNA repair and transcriptional regulation differently. Using microarrays, multiple genes were identified that distinguished *BRCA1* from *BRCA2* subtypes. Interestingly, a patient without a *BRCA1* mutation whose tumor expressed a *BRCA1* molecular phenotype, had DNA hypermethylation of the *BRCA1* promoter, silencing its expression. In another expression profiling study, 16 of 18 *BRCA1* tumors from lymph node negative patients under age 55 were characterized by downregulation of ER co-regulated genes and upregulation of lymphocytic genes, including those primarily expressed by B and T cells. All the *BRCA1* tumors from this study were also demonstrated in a different study to have a “basal-like” gene-expression phenotype, consistent with classical studies characterizing *BRCA1* tumors as mostly high grade ER, PR, and HER2/neu negative tumors that stain positive for basal cytokeratins and are often associated with a lymphocytic infiltrate. Tumors from patients with *BRCA2* mutations, however, appeared to have a luminal estrogen receptor positive expression profile, consistent with ER positive status and luminal keratin overexpression also found in another study. Prophylactic tamoxifen therapy significantly reduces the incidence of breast cancer in patients with *BRCA2* mutations and only modestly, if at all, in patients with *BRCA1* mutations, further supporting the hypothesis that these tumors arise from different epithelial origins, luminal ER-expressing and basal ER-negative cell types. Global profiling studies are also being used to evaluate familial non-*BRCA1/2* breast cancers, with preliminary studies suggesting a partition
into at least two subtypes that do not share gene expression profiles with \textit{BRCA1} or \textit{BRCA2} tumors.\textsuperscript{55} In summary, molecular profiling data suggest that \textit{BRCA1} and \textit{BRCA2} hereditary breast cancers originate from different progenitor cell populations, with independent malignant mechanisms, different prognosis, and different response to prophylactic tamoxifen treatment.

\textit{Characterizing cancer cell progression}

Specific changes in DCIS, atypical hyperplasias, usual hyperplasias, normal lobules or ducts, can be measured by isolating these cell populations from neighboring cells by microdissection. This can be done manually with a dissecting microscope\textsuperscript{13} or with newer techniques that, under microscopic guidance, apply laser energy to excise the cells of interest (laser microdissection, LMD) or melt a polymer onto the cells to be captured and extract only the targeted cells from the surrounding tissue (laser capture microdissection, LCM).\textsuperscript{56}

LCM has been used to extract pure populations of epithelial cells from normal lobules from reduction mammoplasties or breasts with associated cancer, atypical ductal hyperplasia (ADH), DCIS, and invasive ductal carcinoma (IDC)\textsuperscript{57} for microarray analysis. Expression profiling demonstrated that normal epithelial cells distant from cancers had similar transcriptional signatures to normal epithelial cells from reduction mammoplasties. Significant expression changes were observed in ADH and persisted in DCIS and IDC dissected from the same patient, showing patient-specific phenotypes and suggesting that ADH and DCIS are precursors to IDC. The authors found that Grade I expression signatures generally differed from Grade III signatures, but intermediate grade lesions shared either a hybrid signature or a distinct low grade or high grade signature.
In sum, global RNA profiling studies at the invasive\textsuperscript{41,43} and preinvasive\textsuperscript{57} stages suggest that breast cancers originate from progenitor cells with specific molecular subtypes. This corroborates earlier studies by Warnberg\textsuperscript{58} with traditional immunohistochemical (IHC) techniques and by Buerger\textsuperscript{8,59} who used CGH, fluorescence in situ hybridization (FISH), and IHC analyses.

**Molecular profiling in clinical use**

**Tailoring patient treatments using microarrays**

Several groups have shown that molecular profiling can be performed on minimally invasive breast biopsies taken prior to primary chemotherapy or from non-palpable lesions identified by breast imaging. Fine needle aspiration (FNA) biopsies and core needle biopsies have been used\textsuperscript{47,60-63} to successfully isolate RNA for microarray studies. In a small pilot study using core needle biopsies taken before and within the first 48 hours of different regimens of neoadjuvant chemotherapy, Buchholz and colleagues\textsuperscript{60} showed that expression profiles of tumors with and without a good pathological response clustered distinctly. Sotiriou and colleagues\textsuperscript{61} used FNA biopsies performed on ten patients before and during neoadjuvant chemotherapy to monitor patient response to doxorubicin and cyclophosphamide. Candidate gene expression profiles were identified that distinguished responders from nonresponders. Interestingly, the responders also showed expression changes in ten times the number of genes than the nonresponders after the first cycle of chemotherapy. Chang and colleagues\textsuperscript{63} performed core needle biopsies on 24 patients with locally advanced breast cancer. Using microarray analysis, they were able to define a set of 92 differentially expressed genes that characterized docetaxel sensitive tumors, defined as those that had 25\% or less residual disease following treatment. This
gene set showed a positive predictive value of 92% and negative predictive value of 83% and is currently being applied in a larger clinical trial.

**Prognosis Profiling**

Currently, most lymph node negative breast cancer patients with tumors over 1 cm and all lymph node positive patients are candidates for adjuvant systemic treatment, yet only 2-15% will benefit. Better diagnostic methods are necessary to successfully identify the patients that require treatment, predict who will benefit from specific therapies, and discover targets to serve as the basis for new therapies. Patients whose breast cancers are stratified by expression profiling into five molecular subtypes (ER-positive “luminal A and B”; ER-negative “basal”, “ERBB2 over-expressing” and “normal” breast subtypes) or simply into luminal and basal phenotypes demonstrate independent relapse-free survival curves. Of the five major subtypes, the basal-like and ERBB2 subtypes reveal the poorest prognosis. Although luminal A and B subtypes share gene expression similarities and both overexpress ER co-regulated genes, luminal A tumors show the best prognosis of all the subtypes, even in patients with locally advanced breast cancer, while luminal B tumors demonstrate poorer survival. Luminal B tumors express groups of known and unknown genes that are also expressed in ERBB2+ and basal-like tumors, and like these other subtypes, also exhibit TP53 mutations, possibly influencing the poorer prognosis of these three subtypes in initial studies. Since long-term survival in locally advanced breast cancer patients treated with 16 weeks of doxorubicin and tamoxifen was better for the luminal A phenotype, these results suggest that either the tumors possessed a favorable biology or it reflected their responsiveness to doxorubicin and/or tamoxifen treatment.

Breast cancer staging criteria is based on tumor size and the presence of lymph node metastases. Recent data, however, suggests that current staging criteria may need reevaluation. In
expression profiling studies, nodal status and tumor size appear to have less impact on gene expression and survival than tumor biology. Hormone receptor status and grade, however, appear to strongly impact gene expression.\textsuperscript{46, 48, 66} Metastatic potential may be pre-programmed in the biology of the tumor.\textsuperscript{46, 67, 68}

Using a 70 gene expression profile, van’t Veer and colleagues\textsuperscript{46} were able to successfully predict outcome in 81\% of women aged less than 55 years with lymph node negative Stage I and II breast cancers, most of whom did not receive systemic therapy: 91\% of the good prognosis group and only 27\% of the poor prognosis group were disease-free at five years. In a follow-up study by van de Vijver and colleagues,\textsuperscript{66} the 70 gene profile was retrospectively tested on tumors from patients less than 53 years of age, but this time with lymph node negative and positive Stage I and II disease, many of whom received treatment. Lymph node positive patients were evenly divided between good- and poor-prognosis signatures, suggesting that lymph node metastasis may be an independent event distinct from systemic metastasis. After ten years, 85\% in the good prognosis set remained distant metastases free compared to 51\% in the poor prognosis group, offering improvement over St. Gallen\textsuperscript{69} and NIH\textsuperscript{70} criteria. A clinical trial is now underway in Europe to prospectively compare this 70 gene profile to standard classification criteria as the basis for treatment decisions.

Although ER status of the tumors was not an independent prognostic factor in the van de Vijver study, it has been shown to be the most important clinico-pathological discriminator of expression subtype by Sotiriou and colleagues,\textsuperscript{48} who also showed that lymph node status has a minimal influence on expression profiling. Using overlapping gene expression data from the van’t Veer study, Sorlie\textsuperscript{43} showed that basal-like tumors were a prominent subtype with rapid development of metastases within five years. It is possible that the relatively homogeneous
expression pattern shared by basal-like tumors strongly influenced the 70 gene poor-prognosis signature.

**Validation**

The advantage of high throughput global gene expression is the precision afforded in measuring thousands of genes simultaneously; precision at the individual gene level, however, can sometimes be sacrificed to perform global assessments. Therefore, validation must be performed to confirm gene expression and, if desired, to identify the cell type expressing the gene.

A high throughput validation technology is the tissue microarray (TMA). This is a paraffin block made up of hundreds of cores from paraffin-embedded tissues from different patients. When the TMA is sectioned, placed on a slide and combined with traditional validation technologies like IHC and RNA in situ hybridization, gene expression can be validated over hundreds of samples at one time. Another validation tool is real-time quantitative polymerase chain reaction (qPCR) (also called TaqMan PCR). In this technique, RNA from a tissue sample is purified and amplified under optimized gene-specific conditions. Fluorescence molecules are discharged with each amplification cycle and the amount of fluorescence released is dependent on the abundance of RNA in the sample. Using this technology with plates containing multiple sample wells, hundreds of genes can be rapidly measured with high precision and sensitivity.

**Proteomics**
Proteomics is the study of expressed proteins from a genome. The proteome is potentially the most important molecular profile because proteins are the actuators of the genome and a cell’s proteins should determine its phenotype at a given moment. Like DNA and RNA, comprehensive proteomics can be studied in a quantitative (abundance) and qualitative (presence or absence) manner. Unlike DNA and RNA, protein function is also influenced by other factors that shape protein activity, such as protein-protein interactions, subcellular location, conformational changes, half-life changes, and post-translational modifications. To resolve these changes, proteomic techniques include separation and identification techniques.\textsuperscript{33} Due to the increased biochemical and structural diversity of proteins relative to DNA and RNA, these two tasks are difficult. Current techniques are still in development and have not been able to construct a genome-wide proteome to describe breast cancer phenotypes. However, proteomic patterns in the breast cancer serum and ductal fluid already show promise for clinical use in early diagnosis of breast cancer.

\textit{Proteomic Techniques—Protein separation and identification}

Two-dimensional gel electrophoresis (2-DE) sequentially separates proteins by their charge and mass. The separation on a single gel can show thousands of proteins, including proteins that may undergo post-translational modification (such as by phosphorylation, glycosylation, lipid attachment, or peptide cleavage) and be represented by multiple spots on a gel. 2-DE can be utilized to identify protein patterns or to separate proteins prior to identification by mass spectrometry.

Mass spectrometry (MS) is a sensitive and precise approach to identify proteins that are first separated, digested into peptides, and then ionized. Protein separation can be accomplished with 2-DE or other methods such as high performance liquid chromatography (HPLC), 2-D liquid chromatography (2D-LC or LC/LC), capillary electrophoresis, or by biochip chromatography. Proteins are then individually ionized into a protonated gas phase using multiple techniques. Electrospray ionization (ESI)
creates a fine spray of charged droplets from a liquid sample that evaporates, producing gaseous ionized molecules. For samples in a solid state, matrix-assisted laser desorption/ionization (MALDI) is a technique that mixes proteins digested by sequence-specific proteases with a light-absorbing organic acid matrix that catapults the peptides into an ionized form when irradiated by an ultraviolet laser. Surface-enhanced laser desorption/ionization (SELDI) uses resin biochips with different chromatographic properties on their surface to fractionate and isolate proteins through affinity capture. After washing, retained proteins are mixed with energy absorbing molecules and ionized by laser pulsation. A newer modification places the energy absorbing molecules directly on the chip. After ionization by any of these methods, protein fragments are propelled and accelerated by magnetic or electrostatic forces through a time of flight (TOF) mass spectrometer, which separates them by their specific mass to charge (m/z) ratio, forming a peptide mass fingerprint. For MALDI, protein identification is typically accomplished by searching large protein databases and comparing the masses of collections of peptides (peptide mass fingerprint) to those predicted from digestion of protein sequences. For LC-ESI analysis, tandem mass spectrometry (MS/MS), in which individual peptides are fragmented in the mass spectrometer, is utilized to determine the identity of proteins by their amino acid sequences. For SELDI, in which proteins are analyzed in intact form, there is as of yet no straightforward method to identify proteins from mass spectra.

**Characterizing breast tissues proteomes and identifying biomarkers and targets**

2-DE has been used to differentiate protein patterns in normal breast tissue, benign breast tissue, and breast cancer.\(^\text{74}\) A 2-DE technique called difference gel electrophoresis (DIGE), which compares samples from multiple sources differentially labeled with fluorescent dyes by using post-run fluorescent imaging, has been used to differentiate lysates of breast cancer cell lines to identify proteins associated with ERBB2 overexpression.\(^\text{75}\) Bergman and colleagues\(^\text{76}\) used 2-DE combined with ESI-MS and
MALDI-MS to identify polypeptides differentially expressed in solid tumor cell extracts made from scrapings of benign and malignant breast tumors. Some of the overexpressed proteins in breast cancer included nuclear matrix proteins, cytoskeletal and redox proteins, while the known oncogene product DJ-1 was identified in a breast fibroadenomas, not malignant tissue. Truncated forms of overexpressed proteins were also identified, suggesting proteolytic processing in both benign and malignant tissue.

Cell type heterogeneity in breast tissue adds complexity to the characterization of protein populations. Page and colleagues\textsuperscript{77} grew primary epithelial cell cultures derived from reduction mammoplasties and used cell sorting techniques to separate luminal and myoepithelial cells. Protein differences were studied with 2-DE, MALDI and MS/MS technology; a fraction of the differentially expressed proteins were annotated. Many of these corresponded to known cytokeratin markers that distinguish the two cell types. Luminal and myoepithelial cell types also demonstrated significant global homology in their protein profiles, which the authors believed was consistent with derivation from a common stem cell. Several groups have purified epithelial cells from breast cancers and normal tissue using LCM and then performed comparative proteomic analyses.\textsuperscript{78, 79} Wulfkuhle and colleagues\textsuperscript{80} isolated DCIS and normal ductal epithelium by LCM and identified proteins in DCIS involved in intracellular trafficking of lipids, vesicles, and membranes. They also found changes in proteins involved in cell motility and genomic instability, suggesting that DCIS is an already advanced preinvasive lesion.

In the future, sets of cancer-associated biomarkers identified in nipple aspirate fluid and serum may prove useful as clinical diagnostic tools. Varnum and colleagues\textsuperscript{81} collected nipple aspirate fluid (NAF) in healthy women and identified 64 proteins, showing that NAF is a highly concentrated source of biomarkers. Paweletz and colleagues\textsuperscript{82} used SELDI-TOF to analyze NAF, and found protein profiles that appeared to distinguish women with breast cancer from healthy controls. Reasoning that the breast
is a paired organ, Kuerer and colleagues found much higher spot variation comparing protein profiles by 2-DE of paired NAF samples between matched malignant and normal breasts in women with unilateral breast cancer. Applying SELDI-TOF technology to NAF, Sauter and colleagues identified five proteins differentially expressed in women with and without breast cancer that are now being tested in a prospective clinical trial.

At present, investigators are searching for accurate blood tests to diagnose breast cancer. They are hoping that serum protein profiles may be eventually applied to clinical practice. Using SELDI technology, Li and colleagues identified three biomarkers in breast cancer serum. Together these markers can differentiate over 90% of serum samples obtained from women with and without breast cancer. This test did not, however, discriminate serum samples on the basis of tumor size or lymph node metastases. Following up on studies suggesting distinct serum markers in women with ovarian cancer, Petricoin, Liotta and colleagues are using serum protein profiles to develop a blood test to screen for early breast cancer.

**New techniques to more accurately characterize subpopulations of the proteome**

Since current technologies are not able to measure the entire proteome, scientists have also focused on developing proteomic technologies to analyze protein subpopulations. These techniques promise a more detailed and complete view of interesting proteins (membrane proteins or biomarkers) or protein characteristics (protein activity).

A protein microarray measuring comparative fluorescence can be constructed based on protein (eg, antibody) and ligand interactions, analogous to a high throughput enzyme-linked immunosorbant assay (ELISA). Used as an antibody array to detect antigens or an autoantigen array to detect antibodies, this high density array can separate and identify proteins related to breast cancer in complex solutions such as serum or NAF in a fast, efficient, and cost effective manner.
Adam and colleagues\textsuperscript{91} combined membrane isolation techniques, gel electrophoresis and mass spectrometry to gain insight into the enriched membrane protein fractions of breast cancer cell lines, which traditionally have been poorly defined by current global proteomic techniques because of their hydrophobic properties. In addition to many membrane proteins with known significance in breast cancer, such as MUC1 and the HER2/neu and EGF receptors, three novel genes were identified: BCMP11, BCMP84, BCMP101. Protein and mRNA expression of BCMP101 was low in normal tissues in contrast to high levels in many breast cancers confirming BCMP101 as a potential breast cancer marker.

Le Naour and colleagues\textsuperscript{92} used a novel proteomics approach to identify secreted breast cancer proteins in serum using antibodies from patients’ serum. Antibodies in breast cancer serum identified a reactive protein in lysates of human breast cancer tissues and cell lines spotted on a 2-D gel. MALDI-TOF was then used to identify the protein as RS/DJ-1, which was detected at high levels in the sera of 37\% of patients diagnosed with breast cancer. The combined use of autoantibodies and proteomics to discover and identify secreted proteins in cancer remains a promising methodology.

In contrast to other proteomic techniques that measure protein abundance, Jessani and colleagues\textsuperscript{93} used a technique called activity-based protein profiling (ABPP) that detects enzymes only in their active states. Specific active site-directed probes that covalently labeled serine hydrolases, a large enzyme superfamily that comprises approximately 1\% of all proteins in the human proteome, allowed detection of activity in different subcellular locations and glycosylation states in various cancer cell lines. The authors identified proteases, lipases and esterases differentially regulated specific to tissue origin, including breast cancer. The most invasive cell lines, as demonstrated by matrigel assay, showed downregulation of these enzyme activities while a different set of secreted and membrane-associated serine hydrolases showed activation, possibly representing new markers of tumor aggressiveness.
Conclusions:

The progress achieved through molecular profiling tools has allowed us to reevaluate concepts involved in breast tumor evolution, diagnosis, and treatment. DNA molecular profiles have shown tumor progression is associated with accumulating genetic alterations and have exposed DCIS as an advanced lesion; one model suggests specific genetic lesions in DCIS can determine progression of invasive carcinomas; ie, that the differentiation status of the invasive cancer recapitulates that of the in situ lesion. In breast tumors, when RNA expression was compared to changes in the DNA, gene expression signatures were most often related to increases in DNA copy number. Furthermore, single mutations or events are probably not entirely culpable for carcinogenesis since global DNA profiling shows that among multiple breast cancers, a wide range of tumor genotypes (different chromosomal amplifications and deletions) exist. RNA molecular profiles are not quite as diverse, and at least five different expressed phenotypes exist, each with independent survival characteristics. This evidence suggests breast cancer treatments may need to be tailored to different tumor biologies.

RNA expression profiles indicate breast cancers may arise from progenitor cells that occur along basal or luminal differentiation pathways, with basal-like tumors associated with a worse prognosis. *BRCA1* breast cancers exclusively carry a basal-like expression signature that is easily identified using molecular profiling. The profiling also takes into account *BRCA1* methylation, which is not measured by mutation analysis. Importantly, expression profiling also shows that a tumor’s ability to metastasize may not be reliably measured by lymph node metastasis or size. This is in contradistinction to hormone receptor status and grade that play greater roles in distinguishing expression phenotypes.
Promising proteomic studies have utilized nipple aspirate fluid and serum to identify several breast cancer biomarkers. These non-invasive approaches are being tested in clinical trials. Functional proteomics, a new field that measures protein activity within tumor specimens, may identify biomarkers and therapeutic targets not discoverable by other techniques.

Despite the clear impact molecular profiling has made to improving our understanding of breast cancer, there is still a great deal of work ahead. It is important to note that nucleotide mutations in many key genes associated with breast cancer (e.g., TP53) are not distinguished using the global DNA, RNA, or protein molecular profiling methodologies discussed here, but are being studied using other techniques such as single nucleotide polymorphism (SNP) arrays. SNP arrays may also augment our understanding of the affects of chromosomal vs. nucleotide instability on tumor evolution and progression. Furthermore, other areas that may strongly impact breast cancer biology, such as racial/ethnic differences or stromal-epithelial cell interactions, are now being explored.

RNA expression profiling currently holds the most translational promise in breast cancer, but may ultimately be superceded by proteomic techniques. This technique appears to predict clinical outcome and response to systemic therapy better than classical staging criteria in initial studies. Recruitment for large prospective clinical trials to better assess molecular prognostic and predictive gene lists is now underway. It is anticipated that as new global profiling technologies are applied to clinical care, breast cancer diagnosis and care will be more precise and individualized than current methods and will lead to the development of novel tumor-specific therapeutics.
**Figure Legends**

**Figure 1. A general illustration of a cDNA microarray protocol.** A cDNA microarray can be used to determine either gene expression (RNA) or gene copy number (DNA) changes. After purification, the tumor and reference samples are labeled with Cy5 and Cy3, respectively. The mixture is hybridized to a microarray and scanned with two wavelengths to measure the relative intensities of red and green fluorescence at each feature. The relative intensities of features can be compared among tumors to identify changes in expression associated with a tumor subtype. Reproduced with permission from the American Society for Pharmacology and Experimental Therapeutics (ASPET) 94

**Figure 2. Gene expression patterns of 85 breast samples.** Seventy-eight carcinomas, three benign tumors, and four normal breast tissues cluster into 5 subtypes: Luminal A (ER positive, favorable survival); Luminal B (ER positive, poor survival); Normal breast-like; ERBB2 (HER2/Neu) amplicon; Basal epithelial-like cluster. (A) Tumors clusters are represented by branched dendrograms at the upper figure which indicate degree of similarity between samples. Genes are clustered by rows with genes that are expressed most similarly clustered together. Red indicates high relative gene expression compared to reference; green indicates more expression in reference RNA than in tumor sample (low relative expression). Representative gene clusters expressed by the five tumor subtypes above are shown: (B) the ERBB2 amplicon cluster; (C) genes coexpressed by the Luminal B tumors and the basal and ERBB2 tumors; (D) basal epithelial cluster containing keratins 5, 17; (E) normal breast-like cluster; and (F) Luminal A cluster containing ER-associated genes with lower relative expression of these genes by the Luminal B tumors. Permission requested by the Proceedings of the National Academy of Sciences 42


