Gene Expression Profiling of Breast Cancer

Maggie C.U. Cheang,[1] Matt van de Rijn,[2] and Torsten O. Nielsen[1]

[1]Genetic Pathology Evaluation Centre, Vancouver Coastal Health Research Institute, British Columbia Cancer Agency, Vancouver, British Columbia V6H 3Z6, Canada; email: chon@interchange.ubc.ca, torsten@interchange.ubc.ca

[2]Department of Pathology, Stanford University Medical Center, Stanford, California 94305; email: mrijn@stanford.edu

Annu. Rev. Pathol. Mech. Dis. 2008. 3:67–97

First published online as a Review in Advance on August 15, 2007

The *Annual Review of Pathology: Mechanisms of Disease* is online at pathmechdis.annualreviews.org

This article's doi:
10.1146/annurev.pathmechdis.3.121806.151505

1553-4006/08/0228-0067$20.00

Key Words

microarrays, biomarkers, molecular signatures, translational research

Abstract

DNA microarray platforms for gene expression profiling were invented relatively recently, and breast cancer has been among the earliest and most intensely studied diseases using this technology. The molecular signatures so identified help reveal the biologic spectrum of breast cancers, provide diagnostic tools as well as prognostic and predictive gene signatures, and may identify new therapeutic targets. Data are best presented in an open access format to facilitate external validation, the most crucial step in identifying robust, reproducible gene signatures suitable for clinical translation. Clinically practical applications derived from full expression profile studies already in use include reduced versions of microarrays representing key discriminatory genes and therapeutic targets, quantitative polymerase chain reaction assays, or immunohistochemical surrogate panels (suitable for application to standard pathology blocks). Prospective trials are now underway to determine the value of such tools for clinical decision making in breast cancer; these efforts may serve as a model for using such approaches in other tumor types.

INTRODUCTION

Since the introduction of DNA microarray (1, 2) technology in the mid-1990s, breast cancer has probably been the carcinoma most intensively studied by gene expression profiling. DNA microarrays allow researchers to measure the expression of tens of thousands of genes concurrently in one tissue sample. Enormous amounts of data are thereby generated, making microarrays an excellent tool for screening studies, for hypothesis generation, and for elucidating broad patterns of gene expression in health and disease. Over the past decade, there have been increasing numbers of reports describing gene signatures that help to explain the biology of breast cancer and/or have potential clinical value for prognosis, for predicting response to treatment, or for identifying therapeutic targets for drug development. Nevertheless, controversy remains about the validity and reproducibility of the findings, in large part because of the nature of these studies, in which thousands of genes are being assessed on a necessarily much smaller number of specimens (3). Owing to the risk, with this type of data set, of generating false inferences (4), an entire new field of bioinformatics has emerged, and the best data analysis tools to use are currently a matter of intense debate. Stringent study design and data interpretation and, most importantly, validation of results on independent cohorts of patient samples, are required for this technology to advance the clinical management of breast cancer (3). Such studies are now underway.

EXPRESSION PROFILING METHODOLOGY

Microarray Platforms Used in Breast Cancer Studies

In breast cancer research to date, the most commonly used platforms for gene expression profiling have been spotted cDNA (1) and oligonucleotide microarrays (5). Each platform yields measurements of mRNAs present in a tumor specimen, and requires high-quality total RNA that is best isolated from fresh-frozen tissue. The frozen tumors must be handled carefully and consistently during RNA extraction (6) to optimize the integrity and reliability of data for subsequent exploratory data analysis and to allow biologically relevant statistical inferences to be drawn.

Although several platforms for microarray measurements are in use, a mainstay of publicly available sequence information has been the spotted cDNA microarray. Probes are derived from established cDNA libraries, typically encoding 500 to 2000 nucleotides from the 3′ regions of transcripts. The sequences are amplified by polymerase chain reaction, then spotted on the surface of a chemically coated microscope glass slide in an array format. These immobilized cDNA sequence probes will bind to fluorescently labeled complementary sequences of corresponding gene targets present in a biological sample. Hybridization characteristics can be different between probes because of variations in spot size, probe length, and G-C content. As a result, the absolute measurements of staining intensities are not strictly comparable among probe spots and are therefore measured relative to a reference sample RNA pool. To label the genes expressed in tumor samples, the purified RNA sample is first reverse transcribed, incorporating a red-fluorescing nucleotide (Cy-5-dUTP), then cohybridized onto the microarray slide along with green fluorescent (Cy-3)–labeled reference RNA. The fluorescent signals are measured by emissions at the appropriate wavelengths, and the digitally scanned images are quantified. The two channel images are superimposed, to give a preliminary visual coloring readout representing the gene expression of each spot (**Figure 1*a***). In this manner, each sample analyzed can generate in the range of 40,000 data points.

The disadvantages of cDNA microarrays, which include repetitive DNA sequences, poorly characterized genes, and variable

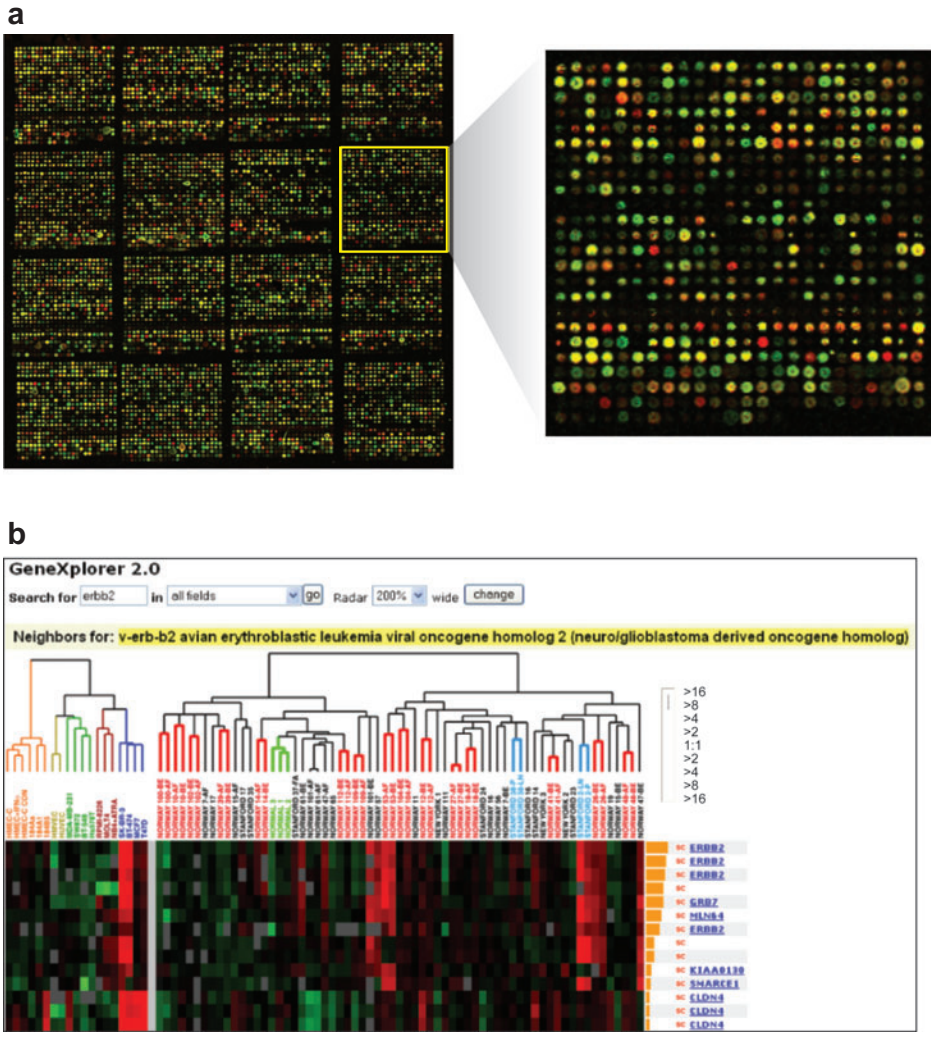


Figure 1

Publicly accessible primary microarray data. (*a*) Raw image data from Reference 31. The detail shows a magnified view of one sector of the array, with 576 hybridized cDNA spots. In total, the array consists of 9216 elements (more recent spotted arrays are higher density, with >42,000 elements). (*b*) GeneXPlorer allows the user to explore cluster diagrams interactively; in this view, genes whose expression is most closely correlated with ERBB2 are shown, including the neighboring GRB7, which is co-amplified with ERBB2 (*left*, cell lines; *right*, breast cancer specimens). Figure data from Reference 31 and available at **http://genome-www5.stanford.edu**.

hybridization kinetics, can be avoided by using fully defined oligonucleotides that represent unique sequences already mapped to the genome. Defined oligonucleotides, typically 30- to 70-mers, are spotted on glass slides using ink jet technology, and the readouts can be interpreted in a similar fashion as those from cDNA microarrays. Agilent-type arrays [e.g., MammaPrint (7)] and nonproprietary human exonic evidence-based oligonucleotide arrays being printed at several academic centers (**http://www.microarray.org/sfgf/heebo.do**) are examples currently in widespread use. High-density oligonucleotide microarrays, exemplified by Affymetrix GeneChip® (5) technology, use photolithography and combinatorial chemistry to enable synthesis of probes on a quartz wafer. Typical Affymetrix chips contain between 10^5 to 10^6 defined probe elements. In contrast to spotted oligonucleotide microarrays, the probes on these ultrahigh-density arrays are 25 nucleotides long, with genes represented by a set of 10–20 pairs of perfect match and mismatch probes. The perfect match probes provide quantitative measurements of fluorescent target sample binding, and the mismatch probes serve as internal

controls. These high-density microarrays are more expensive than spotted microarrays and are not readily customizable. Affymetrix GeneChip assays incorporate an amplification step and need less total RNA template from each tumor sample, 1–2 μg compared with 10–20 μg needed for spotted cDNA microarrays. Biostatisticians have also prepared open-source statistical procedures written especially for handling such data (e.g., BioConductor, **http://www.bioconductor.org**). Alternatively, all the above types of microarrays can be used to measure DNA copy number changes, a technique known as array-based comparative genomic hybridization (aCGH) (8, 9).

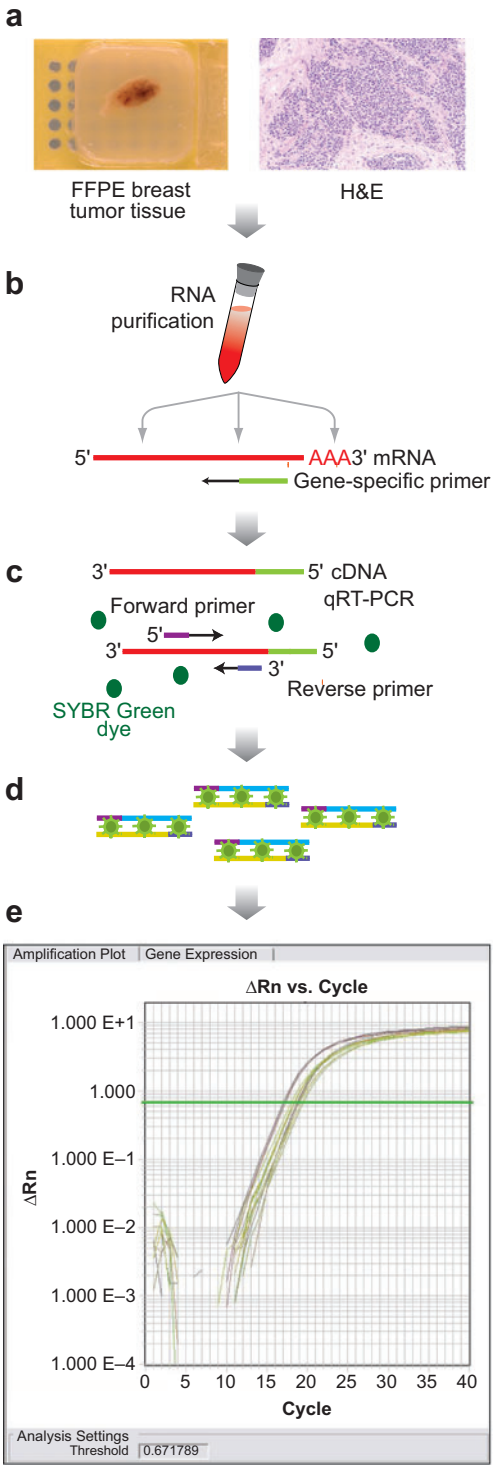
The identification of limited sets of key genes within expression signatures has provided a justification for collecting focused expression profiles (dozens to hundreds of genes) through quantitative real-time polymerase chain reaction (qRT-PCR) (10, 11). The most relevant genes of interest are deduced from microarray studies and other previous investigations. Using unique primer pairs specific for each gene, mRNA transcripts are amplified in a fashion that allows a very accurate quantization of expression (**Figure 2**). This methodology is applicable to very small specimens (e.g., fine needle aspirate or core biopsy material) and, with proper primer design and optimization, can be applied to formalin-fixed, paraffin-embedded tissues. The choice of genes to include in the assay [e.g., optimized for outcome prediction in a particular clinical setting (12) versus based primarily on biologic differences] can be difficult, but becomes critical for developing a clinically useful breast cancer assay by this approach.

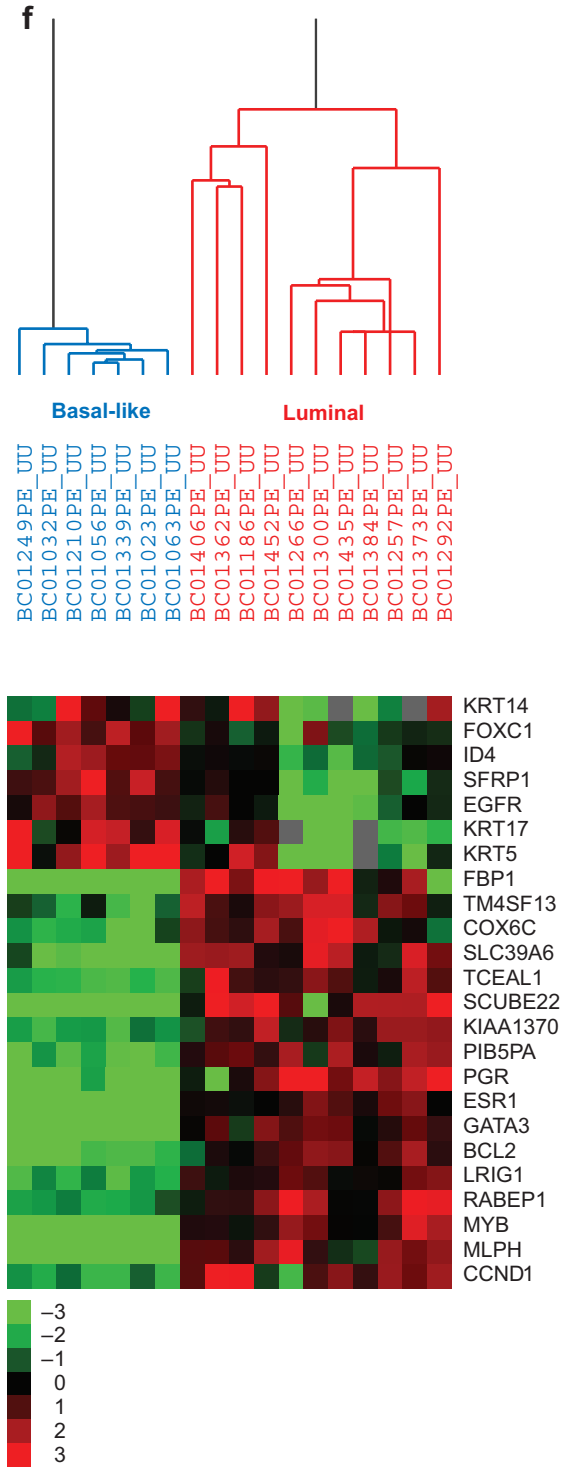
Data Processing

The massive amount of data created by microarray technologies fosters close collaboration between experimental biologists and statisticians. Commercial software (e.g., the GeneSpring Analysis Platform from Agilent Technologies) and open-source code (e.g., BioConductor, **http://www.bioconductor.org**) are available for analysis of genome-wide expression data. The open-source projects provide a wide range of powerful statistical and graphical methods for DNA microarray data analysis and promote high-quality documentation of the methods to facilitate reproducibility. Users thereby can apply the most up-to-date methods, without the worry of additional costs of add-on modules required with commercial software. However, scientists who want to engage in computational biology need to develop some programming skills. More user-friendly software freely available for academic users includes dChip (DNA-Chip Analyzer, **http://biosun1.harvard.edu/complab/dchip/**) and BRB ArrayTools (**http://linus.nci.nih.gov/BRB-ArrayTools.html**).

Figure 2

Expression profiling by quantitative real-time polymerase chain reaction (qRT-PCR). (*a*) Formalin-fixed, paraffin-embedded (FFPE) breast tumor tissue blocks are either sectioned (scroll method) or cored (guided by histology) to obtain a sample dominated by invasive breast cancer tissue. (*b*) Tissue is deparaffinized, followed by total RNA purification and genomic DNA removal. Approximately 0.5 μg from each sample is enough for 200 qRT-PCR runs. (*c*) The RNA template is split (e.g., on 96- or 384-well plates) for separate qRT-PCR reactions designed to measure individual genes of interest. In each well, mRNA transcripts are reverse transcribed using gene-specific primers into cDNA. (*d*) The purified cDNA template is mixed with optimized gene-specific forward and reverse primers for high-throughput SYBR Green-type qRT-PCR. (*e*) Expression of each gene is quantified from the incorporation curve. (*f*) Measurements are normalized (e.g., using housekeeper gene controls) and can then be plugged into any of the statistical analysis methods for microarray expression profiles. Unsupervised hierarchical clustering on 26 genes splits 18 breast tumor samples into basal-like and luminal breast cancer subtypes.





Both spotted cDNA and high-density oligonucleotide array data are subject to variations during array manufacturing, RNA isolation, reverse transcription and labeling, sample hybridization, and image analysis; each of these can impact the gene measurements and create artifactual expression signatures. Normalization is a vital first step to minimize such variation, but different normalization methods can lead to different interpretations of the data. Spotted cDNA and oligonucleotide microarrays use one probe per gene and generate two channel measurements that assume the intensity of each probe is proportional to the amount of target; it is therefore essential to correct measurements by subtracting the background intensity in the spot region. A normalization method is then done within each array to eliminate systemic variations in dye biases (13).

Because data analysis invariably compares results from multiple array experiments, a scaled normalization is carried out between arrays to adjust the distribution of logarithm intensity ratios to have a median of zero for each array. Most normalization methods are done by local regression (14, 15), although more sophisticated statistical methods (13, 16, 17) have also been proposed. The Affymetrix GeneChip employs multiple probes for each gene and a single-color detection method. Background intensity correction is also the first step (18, 19). Affymetrix GeneChip data can then be normalized across arrays, with various approaches available such as applying simple linear scaling in Affymetrix Microarray Suite software, nonlinear smooth curves using "rank invariant set" based on housekeeper control genes (20), and probe-level quantile normalization (19). As Affymetrix employs a set of several probes to represent one gene, summation of data is used to determine the corresponding mRNA quantities (19, 21). Compared with spotted cDNA and oligonucleotide microarrays, high-density oligonucleotide microarray data processing is less intuitive to many experimental biologists, and choosing the most appropriate normalization methods can be challenging.

Presentation of Primary Data

Clearly issues such as normalization and the correct interpretation of these very large data sets are a subject of ongoing discussion. This is one major reason that the consensus among the microarray research community is that raw data supporting published studies should be made publicly available, allowing researchers to use their own techniques to mine data sets and compare studies. In 2001, the Microarray Gene Expression Data Society proposed experimental annotation standards, known as minimum information about a microarray experiment, and these standards are now supported by leading journals. Studies undertaken to compare data from different gene expression microarray platforms have found poor correlation between mRNA measurements from matched genes (22), and differences in data analysis approaches also create apparent poor correlations across platforms (23). However, recent large-scale studies have supported technical consistency across platforms and laboratories (24).

In breast cancer, concerns have been raised about the disappointingly small number of overlapping genes in different published gene expression predictors, leading to questions about the applicability of using microarrays as a platform for developing consistent assays of cancer biology. The most powerful method to confirm the relevance of a particular gene set is to determine how expression levels for that gene set predict outcome in samples analyzed at a different institution using a different platform. In response, an important study tested five published gene expression predictors of breast cancer prognosis (25)—the 70-gene profile (26, 27), activated wound response (28, 29), recurrence score (12), intrinsic biological subtypes (30–33), and two-gene ratio (34)—against a common set of 295 primary breast cancers (25). Results suggest that four of the five tested models show

significant agreement in outcome predictions for individual patients.

Although some cancer microarray expression profile signatures fail to be confirmed, possibly because initial gene lists were developed using inadequate, purely internal training-validation designs (35), there are many that can be validated by comparisons with other gene array study sets or through complementary techniques such as tissue microarrays (TMAs). Below, these individual gene expression predictors are discussed in more detail. However, one very important conclusion that can be drawn is that public access to the primary data (e.g., **Figure 1*b***) not only makes that data set accessible for reanalysis with new bioinformatic tools, but perhaps even more importantly provides a data set useful for external validation studies of other signatures—the most critical step in discerning their potential clinical utility. Selected, large, publicly accessible repositories of breast cancer expression profiling data are shown in **Table 1**.

Statistical Methods

After proper data processing and normalization, microarray expression profiling data are explored to seek biologically and clinically meaningful patterns. Bioinformaticians have developed many statistical methods to deduce gene expression predictors, which must be carefully applied to avoid problems. In primary analyses without an external validation set, there is a risk that genes found to be of interest in the initial set may not be significant in other samples, a phenomenon referred to as overfitting (36).

Older papers in the field often merely extracted the genes with the most extreme differences in expression or associations with clinical parameters, but global analyses of patterns within the data take much better advantage of the genome-wide scale of microarray data sets. Cluster analysis (37–40) is one of the most widely used multivariate methods employed in breast cancer gene expression studies to organize genomic-scale data. This method groups genes on the basis of their overall similarity in expression patterns across specimens, and tumor samples on the basis of their similarities in gene expression. Two popular clustering algorithms are hierarchical clustering and k-means clustering.

Hierarchical clustering (41–43) starts by assigning each subject as a cluster, each containing one subject. Then the closest and most

Table 1 List of selected databases with publicly available breast cancer microarray data

Public Web-based database for breast cancer microarray data	URL	Organization	Description
Array Express	**http://www.ebi.ac.uk/arrayexpress/**	European Bioinformatics Institute (EBI)	Public data deposition and queries
GEO, Gene Expression Omnibus	**http://www.ncbi.nlm.nih.gov/geo/**	National Center for Biotechnology Information (NCBI)	Public data deposition and queries
ONCOMINE, Cancer Profiling Database	**http://www.oncomine.org/main/index.jsp**	University of Michigan	Public queries
PUMAdb, Princeton University MicroArray database	**http://puma.princeton.edu/**	Princeton University	Public queries
SMD, Stanford Microarray database	**http://genome-www5.stanford.edu/**	Stanford University	Public queries
UNC-Chapel Hill Microarray database	**https://genome.unc.edu/**	University of North Carolina at Chapel Hill	Public queries

similar pair of clusters are merged into one new cluster. The similarities between the new cluster and each of the old clusters are computed again, and this is repeated until all subjects are clustered into a single overarching cluster of all genes or all tumors. A dendrogram tree serves as a graphical summary of the data, with the lengths of branches representing the degree of relationship between genes or subjects (**Figure 2*f***). One weakness of hierarchical clustering is that small changes in the data can lead to quite different dendrogram relationships, and this kind of summary is valid only when the pair-wise dissimilarities between objects follow the hierarchical structure resulting from the algorithm (44, 45).

k-means clustering (46) attempts to segment *N* subjects into a predefined number of partitions. In brief, the operator first chooses the number of clusters, *K*. Then the algorithm randomly generates *K* clusters and computes the cluster centers in multidimensional space, known as centroids. Each subject is assigned to the nearest centroid, new centroids are computed, and these steps are iterated until convergence criteria are met (e.g., assignment of subjects does not change). The relative simplicity of this k-means algorithm is favored for its speed in handling large data sets. The disadvantage is that the initial selected number of clusters is somewhat arbitrary, and, if changed, cluster memberships change and may not be nested within the previously assigned clusters.

In 1998, Eisen et al. (47) implemented these procedures in a publicly available software known as Cluster, which allows experimental biologists with minimal knowledge of statistical programming to carry out these analyses. A Java-based software, Treeview (48), was similarly made available to display the clustered profiles in a visually appealing manner, often referred as a heatmap. This graphical display is a useful tool for a two-dimensional representation of the patterns of differential gene expression. The identification of breast cancer intrinsic biological subtypes (31, 33), discussed below, used these methods.

Unsupervised forms of data analysis such as hierarchical clustering permit discovery of inherent data patterns and relationships among genes and tumors, for example, highlighting the previously unrecognized basal-like type of breast cancer (31). By comparison, supervised data analysis techniques identify genes whose expression most closely tracks with a known variable of interest such as tumor grade or patient outcome. An example of a result from supervised analysis is the 70-gene breast cancer prognostic signature (26), which linked tumor gene expression (among node-negative patients under age 55 with tumors less than 5 cm in size) to the presence or absence of distant metastases at 5 years.

The advantage of supervised approaches is that the extracted gene list will almost certainly link to the clinical question being addressed. The biggest disadvantage is that because gene selection is typically optimized against outcome from a set of patients with certain tumor characteristics and treatment regimens, the ability to generalize outside the patient set from which it is derived may be limited.

False discovery rates need to be quantified and limited to optimize gene lists (49), for example, through the sophisticated semisupervised clustering and supervised principal component methods, proposed by Bair & Tibshirani (50), using the nearest shrunken centroids. Such methods can serve, for example, as a method for developing a clinically practical customized mini-array or qRT-PCR assay sufficient to classify breast tumors into biological subtypes and to predict outcome. Other supervised microarray data analysis options include self-organizing maps (51), support vector machines (52), artificial neural networks (53), Significance Analysis of Microarrays (54), and Gene Set Enrichment Analysis (55). Regardless of the method chosen, for both unsupervised and, particularly, supervised approaches, results from genomic-scale experiments still need to be validated on independent series of tumors.

BREAST CANCER EXPRESSION PROFILES

Studies to Understand Breast Tumor Biology

The ability to measure concurrently the expression levels of tens of thousands of genes in cell lines and tumor specimens has, not surprisingly, expanded our understanding of the biology underlying breast cancer. The first published studies, from the Brown and Botstein labs at Stanford, showed that cDNA microarrays could identify expression signatures specific to breast cancer cells (56, 57). A similar study showed that gene expression patterns can predict the invasive capacity of breast cancer cell lines (58).

The first major published study on gene expression profiling of large numbers of clinically annotated breast tumor tissue specimens came out in 2000, wherein Perou et al. (31) demonstrated a biological classification of breast cancers on the basis of distinctive patterns of gene expression. This study applied unsupervised hierarchical clustering to group 65 breast surgical excision specimens from 42 unique patients, using a condensed list of intrinsic genes. The intrinsic genes were defined as those showing significant variation in expression across different tumors but not between paired samples from the same tumor; this is one method to limit analysis to the subset of genes whose variation is more likely a product of biological differences between tumor specimens from different patients than of changes induced by individual patient or technical variability. These molecular interpretations, subsequently validated by studies from Sorlie et al. (32, 33), correlated with patient outcome and, in addition to recognizing known clinically important subgroups, also indicated the existence of novel subtypes of breast carcinoma.

The five major biological breast cancer subtypes so derived are termed Luminal A, Luminal B, HER2 (human epidermal growth factor receptor 2 oncoprotein, product of the *ERBB2* gene) overexpressing, basal-like, and normal-like; their variations in growth rate, specific signaling pathways, and cellular composition of tumors can be explained, at least in part, by their corresponding gene expression patterns. In particular, the proliferation genes are elevated in the three clinically aggressive subtypes: Luminal B, basal-like, and HER2 overexpressing breast cancers. In subsequent studies, these signatures have shown significant reproducibility in predicting patient survival in different cohorts of patients, by different laboratories using different gene array platforms and by novel unsupervised statistical methods (59–62).

One particularly important finding of this research was the highlighting and characterization of the basal-like subgroup of breast cancers, a biologically distinctive group for which targeted therapies are currently unavailable (63). The existence of the novel subtype of basal breast carcinoma was confirmed not only by independent gene array studies but also by a series of immunohistochemical studies on TMAs containing breast carcinoma samples with known clinical features (64–66). The intrinsic subtypes identified by gene expression profiling also correlate with genomic copy number aberrations identified by aCGH (67).

Another interesting breast cancer gene expression pattern reflecting the underlying biology of cancer is the wound-response signature identified by Chang et al. (28, 29), which supports the concept [proposed by H. F. Dvorak in 1986 (68)] that a molecular program activated in normal wound healing is co-opted during cancer invasion and progression. This transcriptional signature is derived from cultured, serum-stimulated fibroblasts (29), and expression of this signature predicted poor clinical outcome among 295 early breast cancer tumors, improving the prediction of clinical outcome over that reflecting the risk stratification provided by clinical risk factors (28). The same signature also predicted poor outcome in gastric and lung carcinoma, indicating that the wound response is shared by a variety of tumor types and

highlighting how investigations of breast cancer expression profiles have implications for other kinds of cancer. A similar approach was used to identify a hypoxia response signature from cultured mammary and renal tubular epithelial cells, and when present in breast cancer specimens, this gene signature is significantly associated with poor clinical outcome, independent of the wound healing signature and of standard clinical parameters (69).

A recent study has proposed an "invasiveness" gene signature based on 186 genes, which includes apoptosis genes *BCL2* and *CASP8*, chemotaxis genes *PLP2* and *CXCL2*, and proliferation genes *SSR1*, *EMP1*, and *ERBB4*, selected because they are differentially expressed between tumor-derived $CD44^+/CD24^-$, highly tumorigenic breast cancer cells in comparison with reduction mammoplasty-derived normal breast epithelium cells (70). This signature is designed to track a minority population of breast cancer cells that behave like stem cells in mouse xenograft assays (71–73). The invasiveness gene signature shows strong association with clinical outcomes in breast cancers, consistent with the underlying idea that stem cells are present and important in breast cancer development (74, 75).

Intriguingly, similar to the wound-response signature mentioned above, this signature also has prognostic value when applied to other malignancies, including lung cancer, prostate cancer, and medulloblastoma data sets, suggesting it identifies a feature of general importance in cancer (70). The invasiveness gene signature probably tracks a different process (more likely to reflect cancer stem cells) than the wound healing response (a nonoverlapping signature more directly linkable to local invasion). Combining these two results in even more impressive cancer risk stratification (70), which can be interpreted as supporting the "seed and soil" hypothesis of cancer (76).

Hereditary cases account for 5%–10% of breast cancers. *BRCA1* and *BRCA2* are the two susceptibility genes whose germ-line mutations are involved in most familial breast cancers, and their mutation status is reflected in somatic tumor gene expression profiles (32, 77). A related gene array study compared *BRCA1*-mutated breast cancers with matched sporadic tumor specimens (78), and reported that a subset of genes involving cell adhesion and migration such as laminins, different collagens, and fibronectins characterizes the *BRCA1*-associated tumors. Familial non-*BRCA1/BRCA2*-mutated breast cancers segregate into two distinct subclasses on the basis of gene expression profiling, and these two subclasses do not cluster together with *BRCA1*- or *BRCA2*-mutated familial cases (79). These findings encourage the search for novel breast cancer predisposition genes among non-*BRCA1/BRCA2*-mutated familial breast cancer and should facilitate such studies, by allowing identification of similar cases within a heterogeneous background.

Studies to Improve the Diagnosis of Breast Cancer

Whereas gene expression profiling undoubtedly provides valuable biologic information to advance our understanding of breast cancer development and progression, its value as a clinical tool is less well proven. The diagnostic setting provides the most direct opportunity to translate gene expression information into clinical use. One relatively straightforward approach is to inspect the expression profile for biomarkers recognized by existing diagnostic immunohistochemistry antibodies. In this manner, for example, a simple immunohistochemistry panel was developed to identify the basal-like subtype of breast cancer (65). Breast cancer gene expression profiles are maintained at sites of distant metastasis (80), implying that such tests may even have some value in the setting of metastatic carcinoma of unknown primary sites.

Disrupted p53 function plays an important role in tumor progression and resistance to

therapy (81–83). However, conventional detection methods such as immunohistochemistry cannot reliably differentiate wild-type versus mutant versus deleted p53, leading to inconsistent and sometimes misleading research results. A 32-gene p53 signature, built from Affymetrix U133 expression profiles of 251 *TP53*-sequenced breast tumors, distinguishes p53-mutant from p53-wild-type tumors in multiple independent data sets, and does a better job predicting clinical outcome than mutation status (84). The relevant p53 signature genes are not canonical targets of p53, but rather are associated with cell proliferation and growth, transcription, ion transport, and breast cancer biology. The signature thus reflects net changes in several cancer biology pathways induced by p53 dysregulation, a type of analysis for which expression profiling is better suited than other methods.

Gene expression profiling studies have also been linked to breast cancer histology. Invasive lobular carcinomas generally proliferate more slowly than ductal tumor, lose E-cadherin expression (85), and are more often estrogen receptor (ER) and progesterone receptor (PR) positive (86). They can show histological overlap with ductal carcinomas and are not always reliably distinguished in practice; perhaps because of this, invasive lobular and ductal carcinomas are treated similarly on the basis of stage, ER, and HER2 status.

Using spotted cDNA microarrays, a supervised analysis of 17 lobular versus 106 ductal breast cancers reported a signature of 11 genes (*E-cadherin*, *survivin*, *cathespin B*, *TP11*, *SPRY1*, *SCYA14*, *TFAP2B*, *thrombospondin 4*, *osteopontin*, *HLA-G*, and *CHC1*) to distinguish lobular and ductal subtypes (87). This signature awaits external validation and clinical translation; a similar study found less clear-cut differences between lobular and ductal carcinomas (88).

Inflammatory breast cancer, in contrast, is easy to identify histologically and has underlying expression signatures that reflect the general biological subtypes seen in non-inflammatory cases (89, 90). Using an 18,000 feature cDNA array from the Netherlands Cancer Institute, Hannemann et al. (91) recently presented publicly available data from a comparison of forty cases of ductal carcinoma in situ with an equal number of invasive breast cancers. The authors present a classifier of 35 genes that could stably distinguish between groups with 91% accuracy, among which *transforming growth factor-β2* and *matrix metalloproteinase 11* were prominent members. The authors further described a classifier that distinguishes well-differentiated from poorly differentiated ductal carcinoma in situ. These findings were validated by internal statistical methods, but again await external validation and clinical translation.

Returning to the topic of hereditary breast cancer, the current diagnostic test for *BRCA1/BRCA2* mutation is time and labor intensive. Given that subtle changes at the DNA level may induce more profound and widespread changes at the RNA level, gene expression profiling offers a potential alternative to identifying mutation carriers. Indeed, not only do the hereditary breast tumors themselves bear characteristic expression signatures (77–79), non-neoplastic fibroblasts from *BRCA1* and *BRCA2* mutation carriers also present distinct gene expression signatures after radiation-induced damage, consistent with an altered DNA repair response (92). If validated, this information could be used to create a screening test based on the net functional defect in these patients, which cannot always be inferred even from sequencing.

Studies to Identify Prognostic Signatures

Histological grade, based on mitotic index, nuclear pleomorphism, and architectural differentiation, is one of the most important prognostic factors for breast cancers, yet has only a moderate level of interobserver agreement among pathologists (93). Thirty to sixty percent of breast cancers are graded as moderately differentiated (Grade 2), and this

group of tumors shows intermediate outcome, providing relatively little guidance for adjuvant treatment decisions.

Using Affymetrix U133A microarrays and a rigorous training/test/external validation strategy, Sotiriou et al. (94) identified distinct gene expression patterns from 97 unique genes differentiating Grade 1 from Grade 3 tumors. These genes, including *UBE2C*, *KPNA2*, *TPX2*, *FOXM1*, *STK6*, *CCNA2*, *BIRC5* (survivin), and *MYBL2* function in cell proliferation and progression, and are in general more highly expressed in Grade 3 tumors. A scoring system, termed gene expression grade index, was developed from this data and serves to classify Grade 2 breast carcinomas into two risk groups, which are significantly associated with distinct relapse-free survival, independent of standard clinical parameters and ER status. In multivariate analysis, gene expression grade index had a higher hazard ratio than other prognostic factors (including nodal status) and displaced histologic grade from the Cox model. Ivshina et al. (95) reported similar findings shortly thereafter from a concurrent, independent study.

Breast tumors identified as Luminal A in other microarray studies have a low gene expression grade index compared with other luminal tumors (96), supporting the idea that expression levels of proliferation genes are useful in defining the poor-outcome Luminal B subtype. This finding is consistent with Dai et al.'s (97) results, which demonstrated that proliferation gene expression signatures are most powerfully prognostic among older, ER-positive patients. Although pathologists can assess proliferation activity by mitotic counts and/or Ki67 immunohistochemistry, fixation issues and variability in visual counting methods probably explain why a gene expression panel seems to do a significantly better job at this task. For the moment, histological grading is so cheap and convenient that it will be hard to replace even by better measurements. However, if gene expression profiling in some form becomes more clinically accessible, proliferation signatures appear poised to be an important application in breast cancer analysis.

The 70-gene prognostic signature, identified from the Netherlands Cancer Institute primary breast cancer cases (26, 27), is an excellent example of applying supervised learning on gene expression profiles. This signature, showing prognostic value for distant metastasis within five years, was first identified using a cohort of 78 node-negative breast cancers occurring in women below age 55 who had not received systemic adjuvant therapy, using oligonucleotide microarrays (26). Many genes involving the hallmarks of cancer are represented: cell cycle, metastasis, angiogenesis, and invasion. This prognostic profile was then validated on a cohort of 295 young patients from the same institution, including node-positive and node-negative tumors with and without systemic treatment (27).

The prognostic value of the 70-gene signature was significantly better at predicting distant relapse-free survival than standard St. Gallen or National Institutes of Health clinical criteria. As this study included the original data set used to derive the prognostic value in its initial validation, a concern of possible data overfitting has generated debate in the microarray community (36), and it remains unclear if this signature really does perform better than the Nottingham prognostic index (98). Nevertheless, the authors of this study made their primary data publicly accessible, making this data set a particularly valuable resource for validation and comparative studies by others in the field (25, 28, 35, 69, 99, 100). A Danish group confirmed the prognostic value of the 70-gene signature in low-risk primary breast cancers, but presented a 32-gene classifier that appeared to be better among the very lowest-risk subset (101), a result that requires external validation.

The 76-gene Rotterdam signature, identified by Wang et al. (102) using Affymetrix U133A arrays on a set of 115 breast cancer patients, was developed to predict distant relapse

rate in untreated node-negative patients. This classification algorithm was trained and optimized in ER-negative and ER-positive patients separately, on the basis of the assumption that the mechanisms driving these two types of cancers are distinct. The genes forming the signature are involved mainly in DNA replication and repair, cell cycle and apoptosis, and immune response. The prognostic value of this 76-gene signature has been validated on an independent set of 180 node-negative untreated patients from different institutions (103) and is also available for public access.

Analysis of the particular genes expressed in a primary breast tumor may also be able to classify the risks for site of distant metastasis. The 70-gene and intrinsic subtype signatures have been shown to be preserved between primary and metastatic sites, supporting the idea that metastatic capacity is inherent in the primary tumor and that metastases closely reflect the biology of the primary tumor (80). Gene signatures that predict breast cancer metastasis to bone (104) and lung (105) have been published. Predicting risk of bone metastasis, in particular, could guide imaging follow-up (e.g., need for bone scans) and/or use of adjuvant bisphosphonates (104).

Prognostic models suffer from inherent limitations. In general, different patient populations and statistical methods will yield different optimal prognostic signatures (106, 107). Supervised analyses optimized against outcome will necessarily be overfitted to the data set from which they are derived and will be expected to lose some prognostic power in validation studies on other data sets. Gene lists need not be expected to overlap, as distinct genes may track similar biological processes (25). Finally, no matter how sophisticated and thorough a microarray analysis may be, there is a stochastic component to patient outcome that will prevent any prognostic model from being perfect—there is a certain unavoidable randomness to fate.

Studies to Predict Tumor Response to Adjuvant Systemic Therapy

The major clinical goal in applying gene expression profiling to breast cancers is to guide more individualized neoadjuvant or adjuvant systemic therapies; conventional biomarkers other than ER and HER2 have not been able to help greatly in this regard. Thus, many studies have sought to use microarray expression profiling to go beyond prognostic signatures to develop predictors of drug response.

Expression profiling of cell cultures before and after drug treatment can be used to develop drug response predictors that may be applied to patient tumor specimens. Doxorubicin- and 5-fluorouracil-induced changes in gene expression seen in cell lines are also seen in tumor specimens from treated patients (108). Cells with luminal-type expression profiles responded by inducing cell cycle checkpoint genes not induced by cells with a basal-like expression profile, indicating that the intrinsic biologic subtype could influence tumor response to conventional chemotherapy. Using various chemotherapy-sensitive carcinoma cell lines as models, a 79-gene doxorubicin resistance signature (including the P-glycoprotein efflux pump) was identified using 43,000 spot cDNA microarrays (109). This signature could be identified in publicly available primary breast cancer specimens and correlated with shorter patient survival, although it has yet to be validated in clinical material randomized to anthracycline versus nonanthracycline therapy.

Treatment of cell cultures with estrogens has also been used to identify an estrogen response signature, including *PR*, *RERG*, *CTSD*, and *PDZ1* and proliferation-associated genes such as *CCNB2*, *CCND1*, *MKI67*, *MYBL2*, *BIRC5* (survivin), and *STK6* (110). When applied retrospectively to multiple independent breast cancer expression profile data sets, this 822-gene classifier predicts outcome among ER-positive and tamoxifen-treated patients.

In the absence of long-term follow-up data, complete pathological response is used as a surrogate end point for patient benefit in a neoadjuvant setting, and exploratory microarray analyses have been incorporated into some clinical trial designs. A study using Affymetrix HgU95-Av2 arrays to compare expression profiles before and after neoadjuvant docetaxel reported that sensitive tumors have higher expression of genes involved in cell cycle, cytoskeleton, adhesion and protein transport, and modification (111), whereas resistant tumors express the mammalian target of rapamycin (mTOR) survival pathway. This small study ($N = 13$) requires confirmation, and the implication that mTOR inhibitors may make an effective combination therapy needs experimental validation.

In contrast, Hanneman et al. (112) were unable to generate expression predictors of pathologic response in 24 doxorubicin-cyclophosphamide and 24 doxorubicin-docetaxel neoadjuvant breast cancer patients ($T > 3$ cm and/or node positive) using 18,000 spot cDNA microarrays. A 31,000 cDNA microarray study of 42 patients (24 for classifier discovery and 18 for validation) yielded a 74-gene classifier to predict complete pathologic response to paclitaxel-fluorouracil-doxorubicin-cyclophosphamide (T/FAC) neoadjuvant chemotherapy (113). In this case, the training-validation split was a very appropriate design method, but left small numbers of events in each group, yielding validation results with borderline significance and low sensitivity (i.e., the study correctly predicted three of seven patients who did respond and eleven of eleven who did not).

Expression profiles from 37 locally advanced patients treated with neoadjuvant liposomal doxorubicin-paclitaxel chemotherapy were evaluated using U133 Plus 2.0 microarrays (114) and linked to outcome. No strong link was possible to clinical or primary site pathologic response in the primary tumor, although a gene list predicting nodal involvement at subsequent surgery was generated that had general prognostic value in univariate analyses in two other series.

For future studies, valuable, prospectively collected before-and-after neoadjuvant samples may best be used to validate preexisting hypotheses about gene predictors, rather than be sacrificed to deduce lists of novel predictor genes (an analysis that can and should still be performed secondarily from the data generated on such specimens). For example, in a study of 82 patients treated with neoadjuvant T/FAC profiled using U133A microarrays, pathologic complete response (pCR) was much higher in cases expressing the previously established basal-like and HER2-positive intrinsic subtype signatures (both with 45% pCR) than in tumors with luminal (2 of 30 pCR) or normal-like (0 of 10 pCR) expression profiles (115). Exploratory gene lists correlating with pCR were then generated by supervised analyses and, interestingly, are completely different between the basal-like and HER2 types, suggesting that tumors with different intrinsic biology employ different mechanisms of drug resistance.

One unanswered question is whether molecular profiles outperform available clinical and pathological parameters. A 28K cDNA microarray study showed gene expression profiles perform as well as (but no better than) classifiers based on clinical parameters and the Nottingham prognostic index to predict odds of recurrences in a cohort of 85 premenopausal, lymph-node-positive breast cancer patients treated with adjuvant CMF (cyclophosphamide, methotrexate, and 5-fluorouracil) (116). This group of patients generally receives aggressive treatment, and the potential to identify any who could be spared chemotherapy would be clinically relevant.

ER-positive tumors are typically associated with better clinical outcomes and good response to hormonal therapies such as tamoxifen (117). However, a subset of patients recurs, and up to 40% develop resistance to tamoxifen (118). Results from randomized trials suggest that aromatase

inhibitors can reduce the recurrence risk, especially at distant sites, for postmenopausal endocrine-responsive breast cancers (119–122). Response to such endocrine therapies can be correlated with global gene expression patterns among ER-positive breast cancers.

Using 18,000 cDNA microarrays, Jansen et al. (123) profiled 10- to 20-year-old frozen primary tumor specimens from patients who received adjuvant tamoxifen. They identified an 81-gene signature (26% involving estrogen action, 14% apoptosis, 9% extracellular matrix formation, and 6% immune response), discriminating 21 patients with objective response from 25 with progressive disease. This list was further reduced to a 44-gene optimal predictive signature and validated on a separate set of 66 breast tumors, confirming significant predictive value for longer time to progression. The overall accuracy to classify tamoxifen resistance was 80%.

A similar study by Ma et al. (34) used 20,000 oligonucleotide microarrays to profile 60 frozen primary tumor samples from women treated with adjuvant tamoxifen, from which they distilled a two-gene ratio predictor for tamoxifen resistance. The genes, *HOXB13* and *IL17BR*, did not have any previously characterized functional relevance to tamoxifen resistance. Initial validation was performed on 20 independent cases by qRT-PCR, which can be readily applied to formalin-fixed tissue. Subsequent validation studies found this ratio to be predictive only among node-negative breast tumors (124), and independent groups have not been able to verify strong concordance with other expression-profile-derived predictors (intrinsic subtype, 70-gene Amsterdam, wound healing, and recurrence score signatures) (25).

Standard pathology blocks of formalin-fixed, paraffin-embedded tissue are more amenable to qRT-PCR (**Figure 2**) assay than full microarray profiling. Use of such material facilitates clinical translation and allows retrospective analysis of previously collected large cohorts with available follow-up data. Building upon published microarray data sets including the intrinsic subtypes and 70-gene Amsterdam signatures, qRT-PCR tests for 250 genes (together effectively constituting a partial expression profile) were developed, and, following testing on three independent clinical series, an optimized 21-gene assay (16 discriminators and 5 internal controls) was developed to predict recurrence in ER-positive, node-negative patients treated with adjuvant tamoxifen. The resulting recurrence score gives differential weightings to the contributing genes (proliferation: *MKI67*, *STK15*, *BIRC5*/survivin, *CCNB1*, *MYBL2*; estrogen response: *ER*, *PGR*, *SCUBE2*; HER2 amplicon: *ERBB2*,*GRB7*; local invasion: *MMP11*, *CTSL2*; antiapoptotic: *BCL2*, *BAG1*; drug metabolism/antioxidant: *GSTM1*; macrophage response: *CD68*) on the basis of model building in the training sets, and was shown to give a quantitative assessment of the likelihood of distant recurrence in 668 cases of ER-positive and node-negative breast cancers treated by adjuvant tamoxifen in the National Surgical Adjuvant Breast and Bowel Project (NSABP) B14 protocol (12).

A later report evaluated the performance of this assay in 651 ER-positive and node-negative breast tumors, which as part of NSABP B20 had been randomized to tamoxifen or to tamoxifen plus chemotherapy (125). Patients with high recurrence scores had a large benefit from chemotherapy (0.26 relative risk and 27.6% mean absolute decrease in a 10-year distant recurrence rate), whereas the low recurrence score group (54% of patients) derived essentially no benefit. On the basis of these exciting results, this assay was promptly offered as a commercial test (Onco*type* Dx™). Controversy remains, however, as the 651 patient set included 227 cases from the tamoxifen arm of NSABP B20, data which were used during the model development phase and then reapplied to the larger combined data set, making the prediction of benefit from chemotherapy appear more impressive (126). As with other commercially available tests (such as MammaPrint), it also remains unclear whether it provides a cost-effective advance

over standard clinicopathologic factors routinely being obtained on all patients.

Predictive assays are in high demand to drive clinical care decisions. Ultimately, proof of the clinical value of expression signatures will require prospective clinical trials, ideally with decision making randomized as to whether decisions are based on standard clinicopathological factors or on the information derived from expression profile signatures. Prospective trials assessing the 70-gene Amsterdam signature and 21-gene recurrence score assays are now underway to help address these issues (described further below).

Studies to Identify Therapeutic Targets

Previous technologies used to analyze broad aspects of tumor biology, such as histomorphology and chromosomal karyotyping, do not directly provide information about drug sensitivity. Microarrays not only identify broad gene expression patterns important to tumor biology, but also link these patterns to specific genes that may be targeted by established or experimental drugs. The re-identification of *ESR1* (encoding ER) and *ERBB2* (encoding HER2) as central genes within identifiable subtypes of breast cancer, as defined by microarray expression profiling, serves as a validation of this technology as a means to identifying relevant drug targets (31). The basal-like molecular phenotype expresses neither ER nor HER2, but does characteristically overexpress epidermal growth factor receptor (32, 33, 56, 65), which is targeted by several new drugs used in colorectal, head and neck, and lung carcinomas, one of which (erlotinib) showed activity in a breast cancer pilot trial (127).

An early study using spotted membrane-based arrays to screen 124 genes on 18 breast tumors highlighted *ESR1* and *HSP90* as top targets (128). Studies now use microarrays measuring in excess of 40,000 transcripts, in over 100 specimens at a time; one more recent study that combined expression profiles with genomic aCGH expression profile data highlighted *ERBB2*, *FGFR1*, *IKBKB*, *PROCC*, *ADAM9*, *FNTA*, *ACACA*, *PNMT*, and *NR1D1* as top targets that are potentially druggable with available agents (67). Other studies have proposed matrix metalloproteinases (129) or NFκB (130) as key targets, in particular breast cancer histologic subtypes. One potentially exciting finding is that apparent estrogen independence and tamoxifen resistance may be conferred, in some cases, by upregulation of androgen receptor action (131, 132), suggesting there would be resistance to aromatase inhibitors but sensitivity to anti-androgens in these patients. Such findings highlight the ability of gene expression profiling to identify pathways that, although only relevant in a minority of breast cancer patients, can be targeted by existing drugs, facilitating the development of personalized medicine.

However, it is important to realize that gene expression, particularly in a microarray study, does not necessarily correlate with the identification of a useful therapeutic target (**Table 2**). Experimental validation, including typical in vitro drug sensitivity assays, is a necessary step even when a target of a well-established drug appears to be expressed in breast cancer specimens. Expression profiling is only a discovery-based screen, albeit a marvelously comprehensive one, for drug targets.

CHALLENGES AND FUTURE DIRECTIONS

Potential Limitations of Breast-Cancer-Profiling Studies

The emergence of prognostic (associated with clinical outcome) and predictive (associated with response to therapy) gene expression signatures holds promise for attempts to individualize breast cancer treatment; however, some studies have raised concerns about the clinical applicability of these published signatures. One concern was that the 76-gene signature, identified by Wang et al. (102) to

Table 2 Upregulated genes in a cancer expression profile

Reason upregulated	Example	Implication for targeted therapy
Fundamental to tumor oncogenesis	*ERBB2*, *ESR1*	Excellent, druggable targets
Secondarily-activated gene encoding, e.g., structural protein	*CK5*, *CK17*	Good diagnostic marker but may not be a useful target
Tertiary changes well downstream from primary event	Probably most genes	Less specific to tumor, more side effects
Compensatory change for oncogenesis	*CDKN*	Targeting this may worsen disease
Expressed by infiltrating normal cells	*CD4*	Not specific to tumor, would have side effects
Untranslated transcript	miRNAs, some expressed sequence tags	Would require new nucleic-acid-based drugs
False discovery	Could apply to any gene	Targets need validation (e.g., by analyzing TMAs)

predict distinct metastasis in untreated node-negative patients of all age groups shared only three genes with the 70-gene signature developed by the Amsterdam group (van't Veer, van de Vijver and colleagues) to predict a five-year distant metastasis risk for node-negative patients younger than 55 years of age (26, 27). This cannot be explained simply by the use of different microarray platforms (the Affymetrix and Agilent platforms share large numbers of overlapping features, particularly for characterized genes important in cancer) nor by the partial differences in patient selection criteria.

Rather, it appears that in breast cancer, multiple distinct gene sets, derived through a variety of approaches and perhaps representing distinct biological processes, can result in prognostic models with a high degree of significance. Thus, although there may seem to be little consistency between studies in the actual gene lists, a recent report comparing five independent published breast cancer gene signatures (25) validated four of them as correlating equally well with survival when applied to one common primary microarray data set (26–28). The four gene signatures showing significant agreement in outcome predictions are the 70-gene Amsterdam signature, wound-healing response, 21-gene qRT-PCR recurrence score, and intrinsic molecular subtype signatures. This important finding implies that, despite differences in gene lists, these separate signatures may contribute equally to the prognostic space in breast cancer (133). Only the two-gene ratio (*HOXB13:IL17BR*) failed to show prognostic value in this study, a result also found by others (134).

In a separate approach to devise more robust gene signatures, one study suggested that combining cross-platform microarray data sets (135, 136) into one single data set may yield gene signatures with a higher predictive power that may be more broadly generalized (137). This finding emphasizes, again, how valuable it is to make primary microarray raw data publicly accessible, to allow external validation, optimization, and meta-analysis studies. There can be a selection bias in extracting significant genes from among thousands of measurements (138), suggesting a need to estimate an error rate by cross-validation or bootstrapping during the gene-selection process. Using the same data set from which the 70-gene signature was originally derived (26), an independent group demonstrated that many predictive gene lists can be selected that correlate equally well with survival, but the correlation fluctuates when measured over different subsets of patients (100). This observation is shared by another study reporting that, using the same data set, the performance of molecular gene signatures can be unstable and depends upon the selection of patient subsets in creating the training sets (35).

After an initial burst of descriptive publications on breast cancer profiles, studies such as these have already become increasingly important, and elements of external validation and data accessibility are now obligatory for publication of gene expression profiling studies in major journals. A recent study, designed to assess the value of a mathematical method named Probably Approximately Correct to quantify the stability of gene signatures, reported that gene expression profiles from thousands of breast cancers may be needed during the discovery phase to deduce a truly definitive predictive list of genes that can have an overlap with more than 50% in a second list of genes derived from similar specimens (139). Discovery-based expression profiling studies on this scale will be almost impossible to realize. Nevertheless, it should be kept in mind that the original 70-gene signature, like several other signatures mentioned above, has now been shown to be a significant prognosticator independent of traditional prognostic markers in a variety of retrospective validation sets.

Concerning the issue of technical consistency across platforms, the intrinsic subtypes model, for example, has been validated to show consistent class assignments across three microarray platforms—Applied Biosystems 60-mer oligonucleotide microarrays, Stanford cDNA microarrays, and Agilent 60-mer oligonucleotide microarrays (140)—and further validated against qRT-PCR for a minimal subset of discriminator genes. Furthermore, the basal subtype of carcinoma recognized using these expression profiling platforms has been validated using immunohistochemistry on TMAs (as further discussed below).

Contribution of Noncancerous Elements to Breast Cancer Expression Profiles

Tumor specimens consist of a complex mixture of invading malignant epithelial cells and their surrounding stroma containing fibroblasts, myofibroblasts, endothelial cells, fat, mast cells, and inflammatory cells. In addition, the samples used for gene array studies may contain normal breast tissue or in situ carcinoma. A fundamental limitation of expression profiling is the correct attribution of the gene expression measurements to the cancer cells themselves. Although laborious, and therefore necessarily reducing sample throughput, this issue can be addressed by laser microdissection (60, 141, 142) or similar techniques. Allinen et al. (143) used a magnetic bead-based cell sorting coupled with serial analysis of gene expression to determine the transcriptional profile of stromal and epithelial cells in normal breast, in situ, and invasive breast carcinoma. At a minimum, frozen sections of the tumor tissue used for RNA isolation should be cut and examined by a pathologist to confirm that the sample is correctly diagnosed and representative of the tumor, as well as to determine the percentage of cancer cells.

To a certain extent, the signature contributions of stroma, normal epithelium, and inflammatory cells can be picked out by bioinformatics, but many genes may function differently in different cells and thus wholesale subtraction of such genes is probably inappropriate. In addition, the non-neoplastic elements adjacent to cancer have distinct, abnormal signatures (60) of their own. Stromal signatures in breast cancer can be matched to soft tissue tumor signatures (99): In a proof-of-principle study using expression profiles of two soft tissue tumors, it was shown that stroma in aggressive breast cancers resembles the solitary fibrous tumor signature (similar to epithelial support fibroblasts), whereas stroma in less-aggressive tumors express a fibromatosis-like signature [similar to scar and seen in normal breast stroma as well (60)].

Note that contributions by non-neoplastic stromal elements are still very important to overall tumor biology and clearly can relate to prognosis (144). However, if the expression data are being mined for diagnostic markers or therapeutic targets, genes expressed in stromal cells will not be as disease specific. Although drugs targeting a stromal cell gene

might be more likely to have side effects on normal tissues, such agents may also prove particularly effective inasmuch as the target cell lacks a cancer cell population's capacity to rapidly acquire resistant mutations. Moreover, the stromal reaction signatures seen in breast carcinoma may be shared at least in part with other carcinoma types, and thus could lead to the development of drug targets that could be effective in a variety of epithelial malignancies.

Given the issues with ascribing gene expression profiles derived from a complex mixture of cells to individual cell types within the specimen, TMA technology (145) has become a valuable tool for high-throughput validation of expression profiles with morphologic correlation (146).

TMAs are constructed by transferring small cores of paraffin-embedded tissue samples to a recipient block, which allows the assessment of one biomarker's expression across hundreds of patients in one experiment. This has the advantage of allowing access to large banks of existing tumor samples independent of those used for expression profiling, which may already be linkable to mature patient outcome data. The main limitation is the capacity to validate only one gene at a time, and TMA-based analyses are really limited to techniques applicable on formalin-fixed, paraffin-embedded tissue (primarily protein-level analysis by immunohistochemistry, and mRNA and DNA analysis by in situ hybridization).

Using TMAs linked to clinical outcome, surrogate immunohistochemical panels can be developed in an attempt to recapitulate the biological subgroupings of breast cancer derived from full gene expression profiles. The basal-like intrinsic subtype can be identified using a panel of established immunostains recognizing ER, PR, HER2, epidermal growth factor receptor, and cytokeratin 5/6 (**Figure 3**) (65). This panel remained prognostic on an independent cohort of patients, and by allowing assessment of standard pathology blocks, the basal-like breast cancer subtype was especially prevalent among premenopausal African American patients (147).

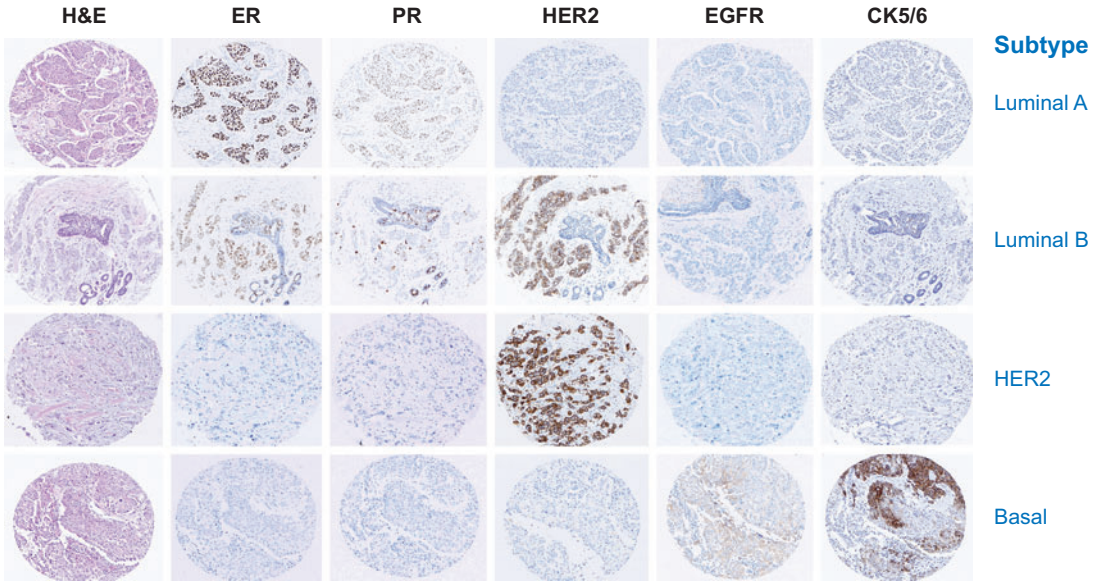


Figure 3

Immunohistochemical surrogate panel derived from gene expression profile data allows molecular subtyping of breast cancers on tissue microarrays.

Novel prognostic immunostaining panels can also be identified by using TMAs (148).

Gene Predictors as Diagnostic Tests for Clinical Use

The 21-gene assay is commercially available under the name Onco*type* Dx. This test requires sending paraffin-embedded tissue to the U.S.-based testing laboratory for qRT-PCR analysis, and is being evaluated prospectively in a large clinical trial named the Trial Assigning IndividuaLized Options for treatment (TAILORx). Sponsored by the National Cancer Institute and in collaboration with the Eastern Cooperative Oncology Group, TAILORx opened in May 2006 and is recruiting more than 10,000 breast cancer patients diagnosed with ER- or PR-positive, HER2-negative breast cancers in 900 sites across the United States. Each tumor will have a recurrence score determined by the multigene qRT-PCR assay, and those with moderate recurrence scores will be randomized to receive adjuvant hormonal therapy with or without chemotherapy. TAILORx is designed to evaluate if the patients with moderate recurrence score benefit from adjuvant chemotherapy and whether the 21-gene recurrence score test can help in treatment decision plans for these patients.

The 70-gene signature is now available as a custom-designed, condensed microarray chip known as MammaPrint (7), and has recently obtained Food and Drug Administration approval. Testing requires a sample of fresh tumor tissue to be sent to the company's laboratory in The Netherlands. The 70-gene signature so obtained has been validated, in historical cohort analyses, to predict early distant relapse (149); to provide additional prognostic information beyond what can be determined from patient age and tumor grade, size, and ER status (149); and to perform at least equally well as outcome probabilities derived from Adjuvant! (**http://www.adjuvantonline.com**) (150, 151).

A large collaborative prospective trial named Microarray In Node-negative Disease may Avoid Chemotherapy Trial (MINDACT) opened in July 2006 and is currently recruiting 6000 node-negative early-stage breast cancer patients to examine the benefit and risk of chemotherapy among patients having discordant risk assessments from the 70-gene signature versus the clinicopathological risk factors based on Adjuvant! This trial is conducted by the Breast International Group and coordinated by the European Organisation for Research and Treatment of Cancer. The objective of MINDACT is to assess prospectively whether 10%–15% of low-risk breast cancer patients can be spared from adjuvant chemotherapy, on the basis of results obtained with the MammaPrint microarray (152). As this expression signature, which requires fresh tissue, could not first be applied to retrospective studies with treatment and outcome information, this is a necessary, bold step to carry such a gene expression signature forward into a phase III trial. Unlike in the TAILORx trial, wherein all patients are to receive the molecular test, patients in MINDACT will be randomized as to whether they receive the test or not, providing a particularly rigorous assessment as to whether the expression profile improves treatment decisions over current gold standard methods.

CONCLUSIONS

Gene expression profiling analysis is a fast-moving field, with new and improved platforms and data analysis methods coming out on almost a monthly basis. General principles that have emerged include the value of public access to primary data and the need for external validation studies, lessons that should also be applied to other high-throughput techniques such as aCGH and proteomics studies. Multiple gene expression signatures relating to breast cancer biology, diagnosis, prognosis, and prediction have been described, and the technique has excellent potential as a discovery tool for new therapies. With some

modifications, this technology is being applied to clinical specimens, and commercially available tests have been released. Such tests are being assessed in prospective trials, whose findings will not be available for several years. Nevertheless, gene expression profiling studies of breast cancer are to a large extent more advanced than those in other solid tumor types and are providing important information about how expression profiling technology can be used to help patients with cancer.

SUMMARY POINTS

1. Several technical platforms and statistical methods for generating and analyzing gene expression profiles exist, and the field continues to develop rapidly.
2. Public availability of primary microarray data in an accessible format is vital not only to allow alternative data analysis tools to be applied, but also to provide data sets for external validation studies.
3. Gene expression signatures relating to breast cancer biology, diagnosis, prognosis, and response to treatment have been published, several of which have been validated in independent series.
4. Large prospective clinical trials are now underway to assess the value of gene expression profiles for clinical decision making in breast cancer.
5. Lessons learned about study design and insights gained about the biology of breast cancer are relevant to other tumor types.

FUTURE ISSUES

1. External validation should be rigorously pursued to assess the clinical importance of particular signatures, and neoadjuvant studies in particular should include planned attempts to validate published predictive signatures, in addition to exploratory research to discover new signatures.
2. Proof that expression profiling technologies can contribute to the cost-effective improvement in patient care decision making over standard clinicopathologic methods will be required for such expression profiling technologies to achieve widespread clinical use.

DISCLOSURE STATEMENT

The Genetic Pathology Evaluation Centre is supported by an unrestricted educational grant from sanofi-aventis. M. van de Rijn is a member of the scientific advisory board of Agendia BV.

ACKNOWLEDGMENTS

T.O.N. is a Scholar of the Michael Smith Foundation for Health Research and supported by the NCI Strategic Partnering to Evaluate Cancer Signatures program (U01-CA114722). We thank Neal M. Poulin for assistance in figure preparation.

LITERATURE CITED

1. Schena M, Shalon D, Davis RW, Brown PO. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270:467–70
2. Shalon D, Smith SJ, Brown PO. 1996. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res.* 6:639–45
3. **Dupuy A, Simon RM. 2007. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *J. Natl. Cancer Inst.* 99:147–57**
4. Quackenbush J. 2006. Microarray analysis and tumor classification. *N. Engl. J. Med.* 354:2463–72
5. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, et al. 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* 14:1675–80
6. Naderi A, Ahmed AA, Barbosa-Morais NL, Aparicio S, Brenton JD, et al. 2004. Expression microarray reproducibility is improved by optimising purification steps in RNA amplification and labelling. *BMC Genomics* 5:9
7. Glas AM, Floore A, Delahaye LJ, Witteveen AT, Pover RC, et al. 2006. Converting a breast cancer microarray signature into a high-throughput diagnostic test. *BMC Genomics* 7:278
8. Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, et al. 1999. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.* 23:41–46
9. Coe BP, Ylstra B, Carvalho B, Meijer GA, Macaulay C, et al. 2007. Resolving the resolution of array CGH. *Genomics* 89:647–53
10. Perreard L, Fan C, Quackenbush JF, Mullins M, Gauthier NP, et al. 2006. Classification and risk stratification of invasive breast carcinomas using a real-time quantitative RT-PCR assay. *Breast Cancer Res.* 8:R23
11. Szabo A, Perou CM, Karaca M, Perreard L, Quackenbush JF, et al. 2004. Statistical modeling for selecting housekeeper genes. *Genome Biol.* 5:R59
12. **Paik S, Shak S, Tang G, Kim C, Baker J, et al. 2004. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N. Engl. J. Med.* 351:2817–26**
13. Dobbin KK, Kawasaki ES, Petersen DW, Simon RM. 2005. Characterizing dye bias in microarray experiments. *Bioinformatics* 21:2430–37
14. Holloway AJ, van Laar RK, Tothill RW, Bowtell DD. 2002. Options available—from start to finish—for obtaining data from DNA microarrays II. *Nat. Genet.* 32(Suppl.):481–89
15. Quackenbush J. 2002. Microarray data normalization and transformation. *Nat. Genet.* 32(Suppl.):496–501
16. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, et al. 2002. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* 30:e15
17. Goryachev AB, Macgregor PF, Edwards AM. 2001. Unfolding of microarray data. *J. Comput. Biol.* 8:443–61
18. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, et al. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4:249–64
19. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, et al. 2003. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* 31:e15

3. Provides a list of dos and don'ts for cancer microarray study design, identifying the most common errors in published studies.

12. Initial description of the 21-gene qRT-PCR recurrence score assay that forms the basis for the Onco*type* Dx assay and TAILORx trial.

20. Schadt EE, Li C, Su C, Wong WH. 2000. Analyzing high-density oligonucleotide gene expression array data. *J. Cell Biochem.* 80:192–202
21. Li C, Wong WH. 2001. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl. Acad. Sci. USA* 98:31–36
22. Kuo WP, Jenssen TK, Butte AJ, Ohno-Machado L, Kohane IS. 2002. Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics* 18:405–12
23. Jarvinen AK, Hautaniemi S, Edgren H, Auvinen P, Saarela J, et al. 2004. Are data from different gene expression microarray platforms comparable? *Genomics* 83:1164–68
24. Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, et al. 2006. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.* 24:1151–61
25. Fan C, Oh DS, Wessels L, Weigelt B, Nuyten DS, et al. 2006. Concordance among gene-expression-based predictors for breast cancer. *N. Engl. J. Med.* 355:560–69
26. van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, et al. 2002. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415:530–36
27. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, et al. 2002. A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* 347:1999–2009
28. Chang HY, Nuyten DS, Sneddon JB, Hastie T, Tibshirani R, et al. 2005. Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *Proc. Natl. Acad. Sci. USA* 102:3738–43
29. Chang HY, Sneddon JB, Alizadeh AA, Sood R, West RB, et al. 2004. Gene expression signature of fibroblast serum response predicts human cancer progression: similarities between tumors and wounds. *PLoS Biol.* 2:E7
30. Hu Z, Fan C, Oh DS, Marron JS, He X, et al. 2006. The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics* 7:96
31. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, et al. 2000. Molecular portraits of human breast tumours. *Nature* 406:747–52
32. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, et al. 2001. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. USA* 98:10869–74
33. Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, et al. 2003. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. Sci. USA* 100:8418–23
34. Ma XJ, Wang Z, Ryan PD, Isakoff SJ, Barmettler A, et al. 2004. A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. *Cancer Cell* 5:607–16
35. Michiels S, Koscielny S, Hill C. 2005. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet* 365:488–92
36. Ransohoff DF. 2004. Rules of evidence for cancer molecular-marker discovery and validation. *Nat. Rev. Cancer* 4:309–14
37. Datta S, Datta S. 2006. Evaluation of clustering algorithms for gene expression data. *BMC Bioinformatics* 7(Suppl. 4):S17
38. Datta S, Datta S. 2006. Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes. *BMC Bioinformatics* 7:397
39. Kaufman L, Rousseeuw P. 1990. *Finding Groups in Data. An Introduction to Cluster Analysis*. New York: Wiley
40. Dudoit S, Fridlyand J. 2002. A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biol.* 3:RESEARCH0036

25. Compares several published breast cancer prognostic signatures against a common data set and concludes that most make similar predictions about patient outcome.

26. Initial description of the 70-gene Amsterdam signature that forms the basis for the MammaPrint assay and MINDACT trial.

31. Original description of the intrinsic subtype signatures and one of the first large-scale microarray studies of human solid tumor specimens to be published.

41. Jobson J. 1992. *Applied Multivariate Data Analysis: Categorical and Multivariate Methods*. New York: Springer-Verlag
42. Hartigan J. 1975. *Clustering Algorithms*. New York: Wiley
43. Gordon AE. 1981. *Classification: Methods for the Exploratory Analysis of Multivariate Data*. New York: Chapman & Hall
44. Hastie T, Tibshirani R, Friedman J. 2001. *The Elements of Statistical Learning*. New York: Springer-Verlag
45. Datta S, Datta S. 2003. Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics* 19:459–66
46. Hartigan JA, Wong MA. 1979. A k-means clustering algorithm. *Appl. Stat.* 28:100–8
47. Eisen MB, Spellman PT, Brown PO, Botstein D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* 95:14863–68
48. Saldanha AJ. 2004. Java Treeview–extensible visualization of microarray data. *Bioinformatics* 20:3246–48
49. Storey JD. 2002. A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B* 64:479–98
50. Bair E, Tibshirani R. 2004. Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol.* 2:E108
51. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, et al. 1999. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA* 96:2907–12
52. Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, et al. 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. USA* 97:262–67
53. Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, et al. 2001. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.* 7:673–79
54. Tusher VG, Tibshirani R, Chu G. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* 98:5116–21
55. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* 102:15545–50
56. Perou CM, Jeffrey SS, van de Rijn M, Rees CA, Eisen MB, et al. 1999. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl. Acad. Sci. USA* 96:9212–17
57. Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, et al. 2000. Systematic variation in gene expression patterns in human cancer cell lines. *Nat. Genet.* 24:227–35
58. Zajchowski DA, Bartholdi MF, Gong Y, Webster L, Liu HL, et al. 2001. Identification of gene expression profiles that predict the aggressive behavior of breast cancer cells. *Cancer Res.* 61:5168–78
59. Calza S, Hall P, Auer G, Bjohle J, Klaar S, et al. 2006. Intrinsic molecular signature of breast cancer in a population-based cohort of 412 patients. *Breast Cancer Res.* 8:R34
60. Finak G, Sadekova S, Pepin F, Hallett M, Meterissian S, et al. 2006. Gene expression signatures of morphologically normal breast tissue identify basal-like tumors. *Breast Cancer Res.* 8:R58
61. Kapp AV, Jeffrey SS, Langerod A, Borresen-Dale AL, Han W, et al. 2006. Discovery and validation of breast cancer subtypes. *BMC Genomics* 7:231
62. Sotiriou C, Neo SY, McShane LM, Korn EL, Long PM, et al. 2003. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc. Natl. Acad. Sci. USA* 100:10393–98

63. Yehiely F, Moyano JV, Evans JR, Nielsen TO, Cryns VL. 2006. Deconstructing the molecular portrait of basal-like breast cancer. *Trends Mol. Med.* 12:537–44
64. van de Rijn M, Perou CM, Tibshirani R, Haas P, Kallioniemi O, et al. 2002. Expression of cytokeratins 17 and 5 identifies a group of breast carcinomas with poor clinical outcome. *Am. J. Pathol.* 161:1991–96
65. Nielsen TO, Hsu FD, Jensen K, Cheang M, Karaca G, et al. 2004. Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma. *Clin. Cancer Res.* 10:5367–74
66. Rakha EA, Putti TC, Abd El-Rehim DM, Paish C, Green AR, et al. 2006. Morphological and immunophenotypic analysis of breast carcinomas with basal and myoepithelial differentiation. *J. Pathol.* 208:495–506
67. Chin K, DeVries S, Fridlyand J, Spellman PT, Roydasgupta R, et al. 2006. Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell* 10:529–41
68. Dvorak HF. 1986. Tumors: wounds that do not heal. Similarities between tumor stroma generation and wound healing. *N. Engl. J. Med.* 315:1650–59
69. Chi JT, Wang Z, Nuyten DS, Rodriguez EH, Schaner ME, et al. 2006. Gene expression programs in response to hypoxia: cell type specificity and prognostic significance in human cancers. *PLoS Med.* 3:e47
70. Liu R, Wang X, Chen GY, Dalerba P, Gurney A, et al. 2007. The prognostic role of a gene signature from tumorigenic breast-cancer cells. *N. Engl. J. Med.* 356:217–26
71. Al-Hajj M, Wicha MS, Benito-Hernandez A, Morrison SJ, Clarke MF. 2003. Prospective identification of tumorigenic breast cancer cells. *Proc. Natl. Acad. Sci. USA* 100:3983–88
72. Reya T, Morrison SJ, Clarke MF, Weissman IL. 2001. Stem cells, cancer, and cancer stem cells. *Nature* 414:105–11
73. Singh SK, Hawkins C, Clarke ID, Squire JA, Bayani J, et al. 2004. Identification of human brain tumour initiating cells. *Nature* 432:396–401
74. Shipitsin M, Campbell LL, Argani P, Weremowicz S, Bloushtain-Qimron N, et al. 2007. Molecular definition of breast tumor heterogeneity. *Cancer Cell* 11:259–73
75. Song LL, Meile L. 2007. Cancer stem cells—an old idea that's new again: implications for the diagnosis and treatment of breast cancer. *Expert Opin. Biol. Ther.* 4:431–38
76. Fidler IJ. 2003. The pathogenesis of cancer metastasis: the 'seed and soil' hypothesis revisited. *Nat. Rev. Cancer* 3:453–58
77. Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, et al. 2001. Gene-expression profiles in hereditary breast cancer. *N. Engl. J. Med.* 344:539–48
78. Berns EM, van Staveren IL, Verhoog L, van de Ouweland AM, Meijer-van Gelder M, et al. 2001. Molecular profiles of BRCA1-mutated and matched sporadic breast tumours: relation with clinico-pathological features. *Br. J. Cancer* 85:538–45
79. Hedenfalk I, Ringner M, Ben-Dor A, Yakhini Z, Chen Y, et al. 2003. Molecular classification of familial non-BRCA1/BRCA2 breast cancer. *Proc. Natl. Acad. Sci. USA* 100:2532–37
80. Weigelt B, Glas AM, Wessels LF, Witteveen AT, Peterse JL, et al. 2003. Gene expression profiles of primary breast tumors maintained in distant metastases. *Proc. Natl. Acad. Sci. USA* 100:15901–5
81. Berns EM, Foekens JA, Vossen R, Look MP, Devilee P, et al. 2000. Complete sequencing of TP53 predicts poor response to systemic therapy of advanced breast cancer. *Cancer Res.* 60:2155–62
82. Borresen-Dale AL. 2003. TP53 and breast cancer. *Hum. Mutat.* 21:292–300

67. Describes large-scale analysis of breast cancer cell lines and tumors by both expression profiling and aCGH.

83. Geisler S, Lonning PE, Aas T, Johnsen H, Fluge O, et al. 2001. Influence of *TP53* gene alterations and c-erbB-2 expression on the response to treatment with doxorubicin in locally advanced breast cancer. *Cancer Res.* 61:2505–12
84. Miller LD, Smeds J, George J, Vega VB, Vergara L, et al. 2005. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc. Natl. Acad. Sci. USA* 102:13550–55
85. Cleton-Jansen AM. 2002. E-cadherin and loss of heterozygosity at chromosome 16 in breast carcinogenesis: different genetic pathways in ductal and lobular breast cancer? *Breast Cancer Res.* 4:5–8
86. Coradini D, Pellizzaro C, Veneroni S, Ventura L, Daidone MG. 2002. Infiltrating ductal and lobular breast carcinomas are characterised by different interrelationships among markers related to angiogenesis and hormone dependence. *Br. J. Cancer* 87:1105–11
87. Korkola JE, DeVries S, Fridlyand J, Hwang ES, Estep AL, et al. 2003. Differentiation of lobular versus ductal breast carcinomas by expression microarray analysis. *Cancer Res.* 63:7167–75
88. Zhao H, Langerod A, Ji Y, Nowels KW, Nesland JM, et al. 2004. Different gene expression patterns in invasive lobular and ductal carcinomas of the breast. *Mol. Biol. Cell* 15:2523–36
89. Bertucci F, Finetti P, Rougemont J, Charafe-Jauffret E, Cervera N, et al. 2005. Gene expression profiling identifies molecular subtypes of inflammatory breast cancer. *Cancer Res.* 65:2170–78
90. Bertucci F, Finetti P, Rougemont J, Charafe-Jauffret E, Nasser V, et al. 2004. Gene expression profiling for molecular characterization of inflammatory breast cancer and prediction of response to chemotherapy. *Cancer Res.* 64:8558–65
91. Hannemann J, Velds A, Halfwerk JB, Kreike B, Peterse JL, van de Vijver MJ. 2006. Classification of ductal carcinoma in situ by gene expression profiling. *Breast Cancer Res.* 8:R61
92. Kote-Jarai Z, Matthews L, Osorio A, Shanley S, Giddings I, et al. 2006. Accurate prediction of *BRCA1* and *BRCA2* heterozygous genotype using expression profiling after induced DNA damage. *Clin. Cancer Res.* 12:3896–901
93. Robbins P, Pinder S, de Klerk N, Dawkins H, Harvey J, et al. 1995. Histological grading of breast carcinomas: a study of interobserver agreement. *Hum. Pathol.* 26:873–79
94. Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, et al. 2006. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J. Natl. Cancer Inst.* 98:262–72
95. Ivshina AV, George J, Senko O, Mow B, Putti TC, et al. 2006. Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer Res.* 66:10292–301
96. Sotiriou C, Wirapati P, Loi S, Desmedt C, Durbecq V, et al. 2005. Better characterization of estrogen receptor (ER) positive luminal subtypes using genomic grade. *Breast Cancer Res. Treat.* 94(Suppl. 1):S19
97. Dai H, van't Veer L, Lamb J, He YD, Mao M, et al. 2005. A cell proliferation signature is a marker of extremely poor outcome in a subpopulation of breast cancer patients. *Cancer Res.* 65:4059–66
98. Eden P, Ritz C, Rose C, Ferno M, Peterson C. 2004. "Good Old" clinical markers have similar power in breast cancer prognosis as microarray gene expression profilers. *Eur. J. Cancer* 40:1837–41
99. West RB, Nuyten DS, Subramanian S, Nielsen TO, Corless CL, et al. 2005. Determination of stromal signatures in breast carcinoma. *PLoS Biol.* 3:e187

100. Ein-Dor L, Kela I, Getz G, Givol D, Domany E. 2005. Outcome signature genes in breast cancer: Is there a unique set? *Bioinformatics* 21:171–78
101. Thomassen M, Tan Q, Eiriksdottir F, Bak M, Cold S, et al. 2007. Prediction of metastasis from low-malignant breast cancer by gene expression profiling. *Int. J. Cancer* 120:1070–75
102. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, et al. 2005. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* 365:671–79
103. Foekens JA, Atkins D, Zhang Y, Sweep FC, Harbeck N, et al. 2006. Multicenter validation of a gene expression-based prognostic signature in lymph node-negative primary breast cancer. *J. Clin. Oncol.* 24:1665–71
104. Smid M, Wang Y, Klijn JG, Sieuwerts AM, Zhang Y, et al. 2006. Genes associated with breast cancer metastatic to bone. *J. Clin. Oncol.* 24:2261–67
105. Minn AJ, Gupta GP, Siegel PM, Bos PD, Shu W, et al. 2005. Genes that mediate breast cancer metastasis to lung. *Nature* 436:518–24
106. Naderi A, Teschendorff AE, Barbosa-Morais NL, Pinder SE, Green AR, et al. 2006. A gene-expression signature to predict survival in breast cancer across independent data sets. *Oncogene* 26:1507–16
107. Huang E, Cheng SH, Dressman H, Pittman J, Tsou MH, et al. 2003. Gene expression predictors of breast cancer outcomes. *Lancet* 361:1590–96
108. Troester MA, Hoadley KA, Sorlie T, Herbert BS, Borresen-Dale AL, et al. 2004. Cell-type-specific responses to chemotherapeutics in breast cancer. *Cancer Res.* 64:4218–26
109. Gyorffy B, Serra V, Jurchott K, Abdul-Ghani R, Garber M, et al. 2005. Prediction of doxorubicin sensitivity in breast tumors based on gene expression profiles of drug-resistant cell lines correlates with patient survival. *Oncogene* 24:7542–51
110. Oh DS, Troester MA, Usary J, Hu Z, He X, et al. 2006. Estrogen-regulated genes predict survival in hormone receptor-positive breast cancers. *J. Clin. Oncol.* 24:1656–64
111. Chang JC, Wooten EC, Tsimelzon A, Hilsenbeck SG, Gutierrez MC, et al. 2005. Patterns of resistance and incomplete response to docetaxel by gene expression profiling in breast cancer patients. *J. Clin. Oncol.* 23:1169–77
112. Hannemann J, Oosterkamp HM, Bosch CA, Velds A, Wessels LF, et al. 2005. Changes in gene expression associated with response to neoadjuvant chemotherapy in breast cancer. *J. Clin. Oncol.* 23:3331–42
113. Ayers M, Symmans WF, Stec J, Damokosh AI, Clark E, et al. 2004. Gene expression profiles predict complete pathologic response to neoadjuvant paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide chemotherapy in breast cancer. *J. Clin. Oncol.* 22:2284–93
114. Dressman HK, Hans C, Bild A, Olson JA, Rosen E, et al. 2006. Gene expression profiles of multiple breast cancer phenotypes and response to neoadjuvant chemotherapy. *Clin. Cancer Res.* 12:819–26
115. Rouzier R, Perou CM, Symmans WF, Ibrahim N, Cristofanilli M, et al. 2005. Breast cancer molecular subtypes respond differently to preoperative chemotherapy. *Clin. Cancer Res.* 11:5678–85
116. Nimeus-Malmstrom E, Ritz C, Eden P, Johnsson A, Ohlsson M, et al. 2006. Gene expression profilers and conventional clinical markers to predict distant recurrences for premenopausal breast cancer patients after adjuvant chemotherapy. *Eur. J. Cancer* 42:2729–37
117. Osborne CK. 1998. Tamoxifen in the treatment of breast cancer. *N. Engl. J. Med.* 339:1609–18
118. Clarke R, Liu MC, Bouker KB, Gu Z, Lee RY, et al. 2003. Antiestrogen resistance in breast cancer and the role of estrogen receptor signaling. *Oncogene* 22:7316–39

119. Boccardo F, Rubagotti A, Puntoni M, Guglielmini P, Amoroso D, et al. 2005. Switching to anastrozole versus continued tamoxifen treatment of early breast cancer: preliminary results of the Italian Tamoxifen Anastrozole Trial. *J. Clin. Oncol.* 23:5138–47
120. Jakesz R, Jonat W, Gnant M, Mittlboeck M, Greil R, et al. 2005. Switching of postmenopausal women with endocrine-responsive early breast cancer to anastrozole after 2 years' adjuvant tamoxifen: combined results of ABCSG trial 8 and ARNO 95 trial. *Lancet* 366:455–62
121. Thurlimann B, Keshaviah A, Coates AS, Mouridsen H, Mauriac L, et al. 2005. A comparison of letrozole and tamoxifen in postmenopausal women with early breast cancer. *N. Engl. J. Med.* 353:2747–57
122. Buzdar AU, Guastalla JP, Nabholtz JM, Cuzick J, Group AT. 2006. Impact of chemotherapy regimens prior to endocrine therapy: results from the ATAC (Anastrozole and Tamoxifen, Alone or in Combination) trial. *Cancer* 107:472–80
123. Jansen MP, Foekens JA, van Staveren IL, Dirkzwager-Kiel MM, Ritstier K, et al. 2005. Molecular classification of tamoxifen-resistant breast carcinomas by gene expression profiling. *J. Clin. Oncol.* 23:732–40
124. Goetz MP, Suman VJ, Ingle JN, Nibbe AM, Visscher DW, et al. 2006. A two-gene expression ratio of homeobox 13 and interleukin-17B receptor for prediction of recurrence and survival in women receiving adjuvant tamoxifen. *Clin. Cancer Res.* 12:2080–87
125. Paik S, Tang G, Shak S, Kim C, Baker J, et al. 2006. Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer. *J. Clin. Oncol.* 24:3726–34
126. Ioannidis JP. 2006. Gene expression profiling for individualized breast cancer chemotherapy: success or not? *Nat. Clin. Pract. Oncol.* 3:538–39
127. Yang SX, Simon RM, Tan AR, Nguyen D, Swain SM. 2005. Gene expression patterns and profile changes pre- and post-erlotinib treatment in patients with metastatic breast cancer. *Clin. Cancer Res.* 11:6226–32
128. Martin KJ, Kritzman BM, Price LM, Koh B, Kwan CP, et al. 2000. Linking gene expression patterns to therapeutic groups in breast cancer. *Cancer Res.* 60:2232–38
129. Pusztai L, Sotiriou C, Buchholz TA, Meric F, Symmans WF, et al. 2003. Molecular profiles of invasive mucinous and ductal carcinomas of the breast: a molecular case study. *Cancer Genet. Cytogenet.* 141:148–53
130. Van Laere S, Van der Auwera I, Van den Eynden GG, Fox SB, Bianchi F, et al. 2005. Distinct molecular signature of inflammatory breast cancer by cDNA microarray analysis. *Breast Cancer Res. Treat.* 93:237–46
131. Becker M, Sommer A, Kratzschmar JR, Seidel H, Pohlenz HD, et al. 2005. Distinct gene expression patterns in a tamoxifen-sensitive human mammary carcinoma xenograft and its tamoxifen-resistant subline MaCa 3366/TAM. *Mol. Cancer Ther.* 4:151–68
132. Doane AS, Danso M, Lal P, Donaton M, Zhang L, et al. 2006. An estrogen receptor-negative breast cancer subset characterized by a hormonally regulated transcriptional program and response to androgen. *Oncogene* 25:3994–4008
133. Massague J. 2007. Sorting out breast-cancer gene signatures. *N. Engl. J. Med.* 356:294–97
134. Reid JF, Lusa L, De Cecco L, Coradini D, Veneroni S, et al. 2005. Limits of predictive models using microarray data for breast cancer clinical treatment outcome. *J. Natl. Cancer Inst.* 97:927–30
135. Gruvberger S, Ringner M, Chen Y, Panavally S, Saal LH, et al. 2001. Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer Res.* 61:5979–84

136. West M, Blanchette C, Dressman H, Huang E, Ishida S, et al. 2001. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl. Acad. Sci. USA* 98:11462–67

137. Warnat P, Eils R, Brors B. 2005. Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. *BMC Bioinformatics* 6:265

138. Ambroise C, McLachlan GJ. 2002. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl. Acad. Sci. USA* 99:6562–66

139. Ein-Dor L, Zuk O, Domany E. 2006. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc. Natl. Acad. Sci. USA* 103:5923–28

140. Sorlie T, Wang Y, Xiao C, Johnsen H, Naume B, et al. 2006. Distinct molecular mechanisms underlying clinically relevant subtypes of breast cancer: gene expression analyses across three different platforms. *BMC Genomics* 7:127

141. Zhu G, Reynolds L, Crnogorac-Jurcevic T, Gillett CE, Dublin EA, et al. 2003. Combination of microdissection and microarray analysis to identify gene expression changes between differentially located tumour cells in breast cancer. *Oncogene* 22:3742–48

142. Fuller AP, Palmer-Toy D, Erlander MG, Sgroi DC. 2003. Laser capture microdissection and advanced molecular analysis of human breast cancer. *J. Mammary Gland Biol. Neoplasia* 8:335–45

143. Allinen M, Beroukhim R, Cai L, Brennan C, Lahti-Domenici J, et al. 2004. Molecular characterization of the tumor microenvironment in breast cancer. *Cancer Cell* 6:17–32

144. Nelson CM, Bissell MJ. 2006. Of extracellular matrix, scaffolds, and signaling: Tissue architecture regulates development, homeostasis, and cancer. *Annu. Rev. Cell Dev. Biol.* 22:287–309

145. Kononen J, Bubendorf L, Kallioniemi A, Barlund M, Schraml P, et al. 1998. Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nat. Med.* 4:844–47

146. van de Rijn M, Gilks CB. 2004. Applications of microarrays to histopathology. *Histopathology* 44:97–108

147. Carey LA, Perou CM, Livasy CA, Dressler LG, Cowan D, et al. 2006. Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. *JAMA* 295:2492–502

148. Ring BZ, Seitz RS, Beck R, Shasteen WJ, Tarr SM, et al. 2006. Novel prognostic immunohistochemical biomarker panel for estrogen receptor-positive breast cancer. *J. Clin. Oncol.* 24:3039–47

149. Buyse M, Loi S, van't Veer L, Viale G, Delorenzi M, et al. 2006. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *J. Natl. Cancer Inst.* 98:1183–92

150. Ravdin PM, Siminoff LA, Davis GJ, Mercer MB, Hewlett J, et al. 2001. Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer. *J. Clin. Oncol.* 19:980–91

151. Olivotto IA, Bajdik CD, Ravdin PM, Speers CH, Coldman AJ, et al. 2005. Population-based validation of the prognostic model ADJUVANT! for early breast cancer. *J. Clin. Oncol.* 23:2716–25

152. Bogaerts J, Cardoso F, Buyse M, Braga S, Loi S, et al. 2006. Gene signature evaluation as a prognostic tool: challenges in the design of the MINDACT trial. *Nat. Clin. Pract. Oncol.* 3:540–51

145. The original paper describing TMA technology.

150. Original description of the Adjuvant! online tool, which is the current gold standard for systemic treatment decisions based on standard clinical and pathological parameters.

TERMS AND DEFINITIONS

Adjuvant systemic therapy: drugs given following cancer surgery in an attempt to inhibit growth of residual or metastatic disease

Array-based comparative genomic hybridization (aCGH): a technique similar to microarray gene expression profiling of mRNA, but that measures DNA copy number at thousands of genomic loci

Bootstrapping: a tool to assess statistical accuracy and model overfitting by resampling the data with replacement

Cluster analysis: mathematical means of determining relationships among genes or among tumors based on a large matrix of gene expression measurements

Complete pathological response: tumor excised after neoadjuvant therapy that has no viable cancer cells visible when examined under the microscope

Dendrogram tree: a graphical display of the relationship of data elements as a series of binary tree branches

DNA microarray: two-dimensional matrix of distinct DNA probes (to measure thousands of genes) or thin tissue sections (to represent hundreds of tumors)

ER: estrogen receptor

External validation set: an independent data set used to validate hypotheses initially derived from another patient population

False discovery rate: a measure of the proportion of false positives among all apparently significant genes as a result of multiple hypothesis testing

Focused expression profiles: measurement of a reduced panel of genes, selected on the basis of biology and/or potential clinical value

Gene expression grade index: a score based on gene expression levels designed to segregate Grade 2 breast cancers into good and poor prognosis groups

Gene expression profiling: measurement of mRNA levels from thousands of genes in a biological sample

Heatmap: visual display of clustered gene expression data, where color represents degree of gene expression

HER2: human epidermal growth factor receptor 2 oncoprotein, product of the *ERBB2* gene

Hierarchical clustering: a method to group data objects into clusters based on their similarities (e.g., in gene expression)

High-density oligonucleotide microarrays: microchips representing millions of unique nucleotide sequences chemically synthesized at specific locations on the surface

Intrinsic genes: genes showing large variations between tumor samples but not paired samples from the same tumor (favors biologic over technical variability)

"Invasiveness" gene signature: 186 genes involved in invasion and metastasis, identified from breast cancer cells

k-means clustering: statistical method to fit genes into a predefined number of groups on the basis of similarities in their expression patterns

Minimum information about a microarray experiment: a data standard defining information that should be presented in published microarray studies to facilitate external analysis of data

Neoadjuvant: a cancer treatment given prior to surgical excision

Normalization: data preprocessing step employed to reduce technical variations in microarray data unrelated to the biology being measured

NSABP: National Surgical Adjuvant Breast and Bowel Project

Oligonucleotide microarrays: microarrays of any type comprising fully defined synthetic nucleotide sequences, typically 25–70 bases long

Overfitting: a statistical model too closely optimized to the data set from which it was derived that therefore extrapolates poorly

pCR: pathologic complete response

Quantitative real-time polymerase chain reaction (qRT-PCR): a particularly accurate method for measuring gene expression applicable on tiny samples, but typically measuring only one gene per reaction

Reference RNA: a high-quality internal control used in some microarray studies for standardization

Spotted cDNA microarray: probes selected from cDNA libraries spotted on chemically coated glass slides

Supervised data analysis: means to link a set of genes to a particular clinical feature or patient outcome

T/FAC: paclitaxel-fluorouracil-doxorubicin-cyclophosphamide

TMA: tissue microarray

Training sets: a data set used for hypothesis generation and statistical model building, often used for initial identification of gene signatures

Unsupervised forms of data analysis: means to discovering inherent patterns and relationships among genes and tumor specimens based on the expression data

Wound-response signature: genes involved cell migration, tissue remodeling, blood vessel growth, and wound-healing processes

Annual Review of Pathology: Mechanisms of Disease

Volume 3, 2008

Contents

The Relevance of Research on Red Cell Membranes to the Understanding of Complex Human Disease: A Personal Perspective
Vincent T. Marchesi 1

Molecular Mechanisms of Prion Pathogenesis
Adriano Aguzzi, Christina Sigurdson, and Mathias Heikenwalder 11

The Aging Brain
Bruce A. Yankner, Tao Lu, and Patrick Loerch 41

Gene Expression Profiling of Breast Cancer
Maggie C.U. Cheang, Matt van de Rijn, and Torsten O. Nielsen 67

The Inflammatory Response to Cell Death
Kenneth L. Rock and Hajime Kono 99

Molecular Biology and Pathogenesis of Viral Myocarditis
Mitra Esfandiarei and Bruce M. McManus 127

Pancreatic Cancer
Anirban Maitra and Ralph H. Hruban 157

Kidney Transplantation: Mechanisms of Rejection and Acceptance
Lynn D. Cornell, R. Neal Smith, and Robert B. Colvin 189

Metastatic Cancer Cell
Marina Bacac and Ivan Stamenkovic 221

Pathogenesis of Thrombotic Microangiopathies
X. Long Zheng and J. Evan Sadler 249

Anti-Inflammatory and Proresolving Lipid Mediators
Charles N. Serhan, Stephanie Yacoubian, and Rong Yang 279

Modeling Morphogenesis and Oncogenesis in Three-Dimensional Breast Epithelial Cultures
Christy Hebner, Valerie M. Weaver, and Jayanta Debnath 313

The Origins of Medulloblastoma Subtypes
Richard J. Gilbertson and David W. Ellison 341

Molecular Biology and Pathology of Lymphangiogenesis
Terhi Karpanen and Kari Alitalo 367

Endoplasmic Reticulum Stress in Disease Pathogenesis
Jonathan H. Lin, Peter Walter, and T.S. Benedict Yen 399

Autophagy: Basic Principles and Relevance to Disease
Mondira Kundu and Craig B. Thompson 427

The Osteoclast: Friend or Foe?
Deborah V. Novack and Steven L. Teitelbaum 457

Applications of Proteomics to Lab Diagnosis
Raghothama Chaerkady and Akhilesh Pandey 485

The Pathology of Influenza Virus Infections
Jeffrey K. Taubenberger and David M. Morens 499

Airway Smooth Muscle in Asthma
Marc B. Hershenson, Melanie Brown, Blanca Camoretti-Mercado, and Julian Solway 523

Molecular Pathobiology of Gastrointestinal Stromal Sarcomas
Christopher L. Corless and Michael C. Heinrich 557

Notch Signaling in Leukemia
Jon C. Aster, Warren S. Pear, and Stephen C. Blacklow 587

The Role of Hypoxia in Vascular Injury and Repair
Tony E. Walshe and Patricia A. D'Amore 615

Indexes

Cumulative Index of Contributing Authors, Volumes 1–3 645

Cumulative Index of Chapter Titles, Volumes 1–3 647

Errata

An online log of corrections to *Annual Review of Pathology: Mechanisms of Disease* articles may be found at http://pathol.annualreviews.org