A comparison of gene expression signatures from breast tumors and breast tissue derived cell lines

Douglas T. Ross[a] and Charles M. Perou[b,*]
[a]*Applied Genomics, Inc., Sunnyvale, CA, USA*
[b]*Dept. of Genetics and Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, NC, USA*

Cell lines derived from human tumors have historically served as the primary experimental model system for exploration of tumor cell biology and pharmacology. Cell line studies, however, must be interpreted in the context of artifacts introduced by selection and establishment of cell lines *in vitro*. This complication has led to difficulty in the extrapolation of biology observed in cell lines to tumor biology *in vivo*. Modern genomic analysis tool like DNA microarrays and gene expression profiling now provide a platform for the systematic characterization and classification of both cell lines and tumor samples. Studies using clinical samples have begun to identify classes of tumors that appear both biologically and clinically unique as inferred from their distinctive patterns of expressed genes. In this review, we explore the relationships between patterns of gene expression in breast tumor derived cell lines to those from clinical tumor specimens. This analysis demonstrates that cell lines and tumor samples have distinctive gene expression patterns in common and underscores the need for careful assessment of the appropriateness of any given cell line as a model for a given tumor subtype.

1. Introduction

Oncologists rely upon clinical information, a morphologic assessment, and to a limited degree, immunohistochemical and molecular markers to classify malignancies into groups that have distinct clinical behavior. It is clear, however, that additional markers and/or technologies are needed for classifying tumors as current methods sometimes fail to accurately predict patient clinical course. In breast cancer for example, tens to hundreds of different genes/proteins have been shown to be of prognostic value, however, many of these markers co-vary, and hence, are not of independent prognostic value. In addition, progress in adopting these markers into clinical practice has been limited both by technical constraints in the number of markers that can be examined efficiently and by the difficulty in comparing and validating studies that use different reagents and clinical sample sets. In breast cancer, only three markers are typically scored for in the clinical setting which include the estrogen receptor (ER), the tyrosine kinase receptor ERBB2/HER2, and an assessment of tumor proliferation index (e.g. Ki-67 labeling fraction) [1].

The advent of modern genomic analysis tools, in particular DNA microarrays, has essentially created a new tool that is capable of collecting thousands of objective observations on clinical samples that can and are being used to characterize tumors and cell lines at a level of definition that was not possible even five years ago [2]. Many groups have begun to use microarrays to measure gene expression in hundreds of tumor samples with the expectation that the genomic scale measurement of gene expression will reveal a novel molecular based classification of malignant cells. In this review, we will first focus on the characterization of breast tissue and tumor derived cell lines using data obtained from cDNA microarrays. These data can be used to 1) identify which cell lines are the best models for different breast tumor subtypes, 2) define molecular signatures that distinguish the biology of different cell lines and tumor types, 3) identify new candidate markers for tumor diagnosis and classification, and 4) identify subtype specific targets for therapeutic intervention.

*Address for correspondence: Charles M. Perou, Lineberger Comprehensive Cancer Center, CB# 7295, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA. E-mail: cperou@med.unc.edu.

Disease Markers 17 (2001) 99–109
ISSN 0278-0240 / $8.00

Table 1

Cell line name	Array #	Previous description of cell line and (source)
184A1	svcc38	immortal derivative of 184Aa (M. Stampfer)
184Aa	svcc17	primary HMEC strain Aa (M. Stampfer)
184B5	svcc40	immortal derivative of 184Aa (M. Stampfer)
BT-474-ATCC	svcc128	ERBB2 and ER positive line (ATCC)
BT-474-Stanford	svj107	ERBB2 and ER positive line (Stanford)
BT-549	svcc69	papillary/ductal carcinoma derived (NCI)
Fibroblast-UTSW	shav146	hTERT immortalized stromal cell line (J. Shay/UTSW)
HB2	svcc37	SV40 immortalized breast epithelial line (H.S.Wiley)
HCC-1937	shaj046	BRCA1 mutant carcinoma derived line (J. Shay/UTSW)
HME31	shat023	primary HMEC strain 31 (J. Shay)
HMEC+IFNα	svcc500	HMEC-C strain plus IFNα (Clonetics)
HMEC-C	svcc94	primary HMEC strain C (Clonetics)
HMEC-C CON	svcc47	HMEC-C strain 2 days at 100% confluence (Clonetics)
HMS32	shaj058	primary breast stromal/fibroblast cell strain (J. Shay)
Hs578T-ATCC	shac095	breast carcinosarcoma derived line (ATCC)
Hs578T-NCI	svcc110	breast carcinosarcoma derived line (NCI)
MCF-10A	svn008	non-tumorigenic breast epithelial cell line (ATCC)
MCF-12A	sham103	non-tumorigenic breast epithelial cell line (ATCC)
MCF7-NCI	svcc1299	ER positive line isolated from a pleural effusion (NCI)
MCF7-UCLA	shat022	ER positive line from a pleural effusion (UCLA)
MDA-MB-231-NCI	svcc73	ER negative line from a pleural effusion (NCI)
MDA-MB-231-UTSW	shaj054	ER negative line from a pleural effusion (J. Shay)
SK-BR-3	svcc15	ERBB2 positive line from a pleural effusion (ATCC)
T47D	svcc71	ER positive line from a pleural effusion (ATCC)

2. Classification of breast cell lines

In general, cell lines established from breast tissues have been mainly characterized with respect to their expression of cytokeratins, the estrogen receptor (ER), and ERBB2/HER2 protein [3–5]. For some lines, the histology of xenografts has been compared to the pathology of their tumor of origin in order to confirm that the cell line has conserved features of its parental tumor. As part of our efforts to characterize the phenotypic diversity of human breast tumors, we have measured gene expression using cDNA microarrays in a number of breast tissue derived cell lines [6–8]. In this review, we have re-analyzed previously published data from thirteen cell lines and report new data from three additional independent lines. Furthermore, we explored cell line stability by measuring gene expression in the same cell line obtained from different sources and therefore propagated independently. Lastly, we included some instructive data derived from a normal mammary epithelial cell line treated with interferon, data derived from a confluent normal cell line culture, and a variant of a normal cell line immortalized *in vitro*. Gene expression in these cell lines was measured using spotted microarrays in comparison to a common reference sample in a manner that has been previously described and that allows all samples to be compared to one another [6,7].

One of the most striking findings from genomic studies of gene expression to date has been that tumors and/or cell lines have common characteristics of biological or clinical importance that can be identified in their patterns of expressed genes. Hierarchical clustering analysis has been used to identify systematic features in the patterns of variation of expression of genes across sample sets [9–12]. The functions of the known genes that are either relatively over-expressed or under-expressed in comparison between samples can give clues as to the differences in biology that are reflected in gene expression patterns [6–8,13–16]. In this breast cell line data set, we selected for analysis, the subset of genes that showed 1) a signal intensity of $>$70 arbitrary units in both the Cy3 and Cy5 channels, and 2) expression variation of at least 3-fold or more from average for that gene across the sample set in two or more of the 24 total experiments. This criterion selected 1287 genes out of the 8102 total original genes that were both well measured and changed in expression significantly between cell lines. All primary microarray data for the experiments presented here can be obtained from the Stanford Microarray Database at http://genome-www4.stanford.edu/MicroArray/SMD/, and all figures can be seen at our Supplementary Information website at http://genome-www.stanford.edu/breast_cancer/cell_line_review2001/.

As can be seen in Fig. 1A, hierarchical clustering analysis divided the cell lines into three main dendrogram branches. The first branch on the far left (Red) contained all three of the primary Human

Mammary Epithelial Cell (HMEC) lines, all HMEC immortal derivatives and the non-tumorigenic MCF-12A [17] cell line ostensibly derived from breast epithelium. The center dendrogram branch (Orange) contained both normal fibroblast derived cell lines, a carcinosarcoma derived line (Hs578T [18]), a line ostensibly derived from breast epithelium (MCF-10A [19]), and two lines derived from breast carcinoma specimens (BT-549/Coutinho and Lasfargues 1978, and MDA-MB-231 [20]). The far right dendrogram branch (Blue) contained all cell lines that were thought to be derived from luminal epithelial cells including two estrogen receptor expressing cell lines (MCF-7 [21] and T47D [22]), two ERBB2 over-expressing lines (SK-BR-3/Trempe and Old 1970, and BT-474 [23]), the SV-40 transformed epithelial derived cell line HB2 [5], and the BRCA1 mutant cell line HCC-1937 that was originally isolated by A. Gazdar and colleagues [24]. The biological functions of the sets of genes that were differentially expressed across different "branches" of the cell line dendrogram (Fig. 1B–E) suggested that the gene expression patterns identified cell lines with features that could be related to different types of normal breast cells. The cell lines sorted into those that either expressed HMEC/basal-cell characteristics, those that expressed stromal/mesenchymal-cell-like characteristics or those that expressed luminal-cell characteristics. It should be emphasized that this is an interpretation of the gene expression patterns and that alternative interpretations of these gene expression patterns are possible.

3. Breast basal epithelial cell signature

The group containing all HMEC lines (Red dendrogram branch) was distinguished by the very high expression of a set of genes that contained many markers of normal breast basal-epithelial cells including keratins 5 and 17 (Fig. 1C) [3,4,25,26]. This set also included many genes whose roles in cell physiology distinguish basal from luminal epithelial cells including the production of basal lamina components and interactions with the extracellular matrix (e.g. gamma and alpha-laminin, collagen type-XVII, integrins alpha-3, alpha-6 and beta-4). The cultured basal-like cell lines expressed variable amounts of smooth-muscle-actin but much less relative to the other lines that expressed the "stromal cell" gene expression signature (Fig. 1D). Therefore, these cultured cells appeared to express some, but not all, of the features of so-called "myo-epithelial cells" which are mature smooth-muscle-actin expressing cells that have a functional contractile apparatus [3,5,27]. This "basal" pattern of gene expression was not restricted to HMEC in that this set of genes were moderately expressed in three other lines (MCF-10A, BT-549 and HB2) that also expressed strong stromal-like gene expression signatures and therefore, did not fall into this class by clustering analysis (see below). It should also be noted that even immortal HMEC derivatives, like 184B5 and 184A1, showed the dominant "basal" cell gene expression pattern and not other signature patterns, and hence, this pattern was not dramatically influenced by immortalization or transformation (see Supplementary Information Fig. 3 – http://genome-www.stanford.edu/breast_cell_line_review2001/).

4. Luminal epithelial cell signature

Approximately 60–70% of sporadic breast tumors are estrogen receptor positive and are believed to be derived from breast luminal epithelial cells [1], which can be distinguished by their expression of cytokeratins 8 and 18 and by their location and function in lining breast secretory-ducts. The *in vitro* study of this cell type has been complicated by the difficulty in maintaining primary cultures of normal estrogen-receptor-positive luminal cells for longer than a few population doublings [28]. Therefore, most *in vitro* studies on breast luminal epithelial cells have been performed on cell lines derived from primary breast tumors or pleural effusions from breast cancer patients.

The pattern of gene expression that distinguished the luminal-like signature was comprised of a set of genes that were nearly exclusively expressed in all of the luminal like lines while very low to absent levels were seen in all of the other tested lines (Fig. 1E). Contained within this set of genes were genes/proteins that have been previously used to distinguish luminal breast epithelial cells including the estrogen receptor and keratins 8 and 18 [3]. However, two distinct subtypes of cell lines that expressed luminal characteristics could be distinguished, 1) those that expressed high levels of the estrogen receptor and essentially lacked either stromal or basal gene expression signatures (e.g. MCF7, BT-474 and T47D), and 2) those that expressed little or no estrogen receptor but also expressed genes characteristic of the basal signature (HCC-1937 and HB2). SK-BR-3 was unique in this set of cell lines in that it expressed low levels of the estrogen receptor but a strong luminal gene expression signature without the expression of basal cell characteristics.

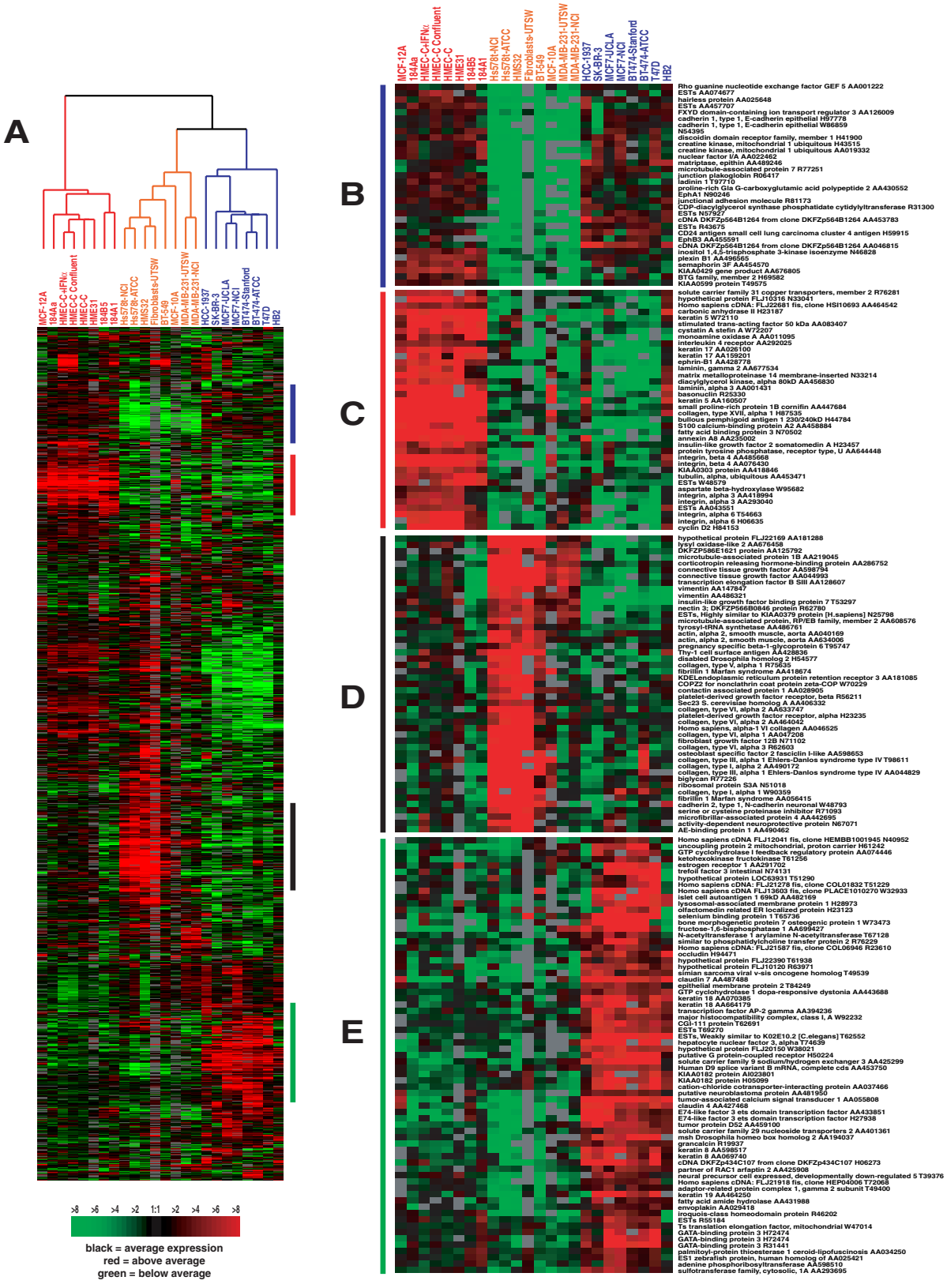
A
B
C
D
E
MCF-12A
184Aa
HMEC-C-nIFNα
HMEC-C Confluent
HMEC-C
HME31
184B5
184A1
Hs578t-NCI
Hs578t-ATCC
HMS32
Fibroblasts-UTSW
BT-549
MCF-10A
MDA-MB-231-UTSW
MDA-MB-231-NCI
HCC-1937
SK-BR-3
MCF7-UCLA
MCF7-NCI
BT474-Stanford
BT474-ATCC
T47D
HB2
>8 >6 >4 >2 1:1 >2 >4 >6 >8
black = average expression
red = above average
green = below average

Fig. 1. Cluster diagram depicting relative gene expression differences between cell lines. The red-green pseudocolor chart depicts gene expression data in comparison between different cell lines. Red blocks depict genes relatively over-expressed in comparison between the measured samples whereas green blocks depict genes relatively under-expressed. The data table has been organized by hierarchical clustering which groups the genes on the basis of their similarity in expression patterns across a set of experimental samples (e.g. cell lines), and groups the experimental samples together based upon their similarity in gene expression patterns across the set of chosen genes. The result of the analysis is a re-ordering of the data table such that genes with relatively similar patterns of expression across the sample set are adjacent to one another in the rows, and samples with similar patterns of expression in the set of chosen genes are adjacent to one another in the columns. The dendrogram above the color chart depicts the relative similarities of the cell lines to one another; terminal branches contain cell lines that express relatively similar patterns of gene expression across whereas those separated by longer branches express relatively less similar gene expression patterns [9]. A) Complete cluster diagram that depicts all 1287 transcripts across 18 independent cell lines including 24 different hybridizations. B) "Common epithelial" cell gene set that was expressed in both basal and luminal cells but was not expressed in the cells that have strong fibroblast-like characteristics. C) Breast basal epithelial cell gene set that was strongly expressed in all HMEC derived cell lines. D) Stromal-like/fibroblast gene set that was expressed in some fibroblasts as well as some breast cancer derived cell lines that were ostensibly mis-classified as carcinoma derived. E) Luminal epithelial gene set that was expressed in estrogen-receptor-positive cell lines as well as a few other lines. The color scale at the bottom left depicts the gene expression measured in each cell line relative to the average expression for each gene as determined in the 24 different cell line samples.

5. Mesenchymal/stromal cell signature

We have previously shown that a small, but significant, number of cell lines ostensibly of epithelial origins showed patterns of gene expression that were more consistent with characteristics expected of stromal cells (see [6] and http://genome-www.stanford.edu/nci60/). In order to further investigate the gene expression properties of these cell lines we compared them to two cell lines explicitly derived from breast stroma (HMS32 and Fibroblast-UTSW, both obtained from Jerry Shay/UTSW). The distinguishing gene expression signature for these strains/lines was comprised of the high expression of a number of genes with roles in remodeling of extracellular matrix including high expression of the genes encoding smooth muscle actin, vimentin, fibrillin, byglycan and collagen types I, III, V and VI (Fig. 1D), combined with the low expression of genes characteristic of epithelial cells (Fig. 1B). The cell lines contained within this branch of the dendrogram (Orange) were further subdivided into a branch that contained three similar lines that expressed the highest levels of this "stromal" gene expression signature and others that showed incrementally less expression of this set of genes. Consistent with the interpretation that this signature reflected expression of stromal cell physiology, this signature was the most strongly expressed in the carcinosarcoma derived line Hs578T and the two cell lines/strains established from fibroblasts, HMS32 (primary fibroblast strain with a finite lifespan) and Fibroblast-UTSW (telomerase immortalized breast derived fibroblast line). The remaining lines that clustered with these fibroblast-like lines, showed decreased overall expression of this set of genes with incrementally less expression in BT-549, less in MCF-10A, and finally, only a few stromal-signature genes expressed in the two independently propagated lines of MDA-MB-231 (Fig. 1D).

6. Common epithelial cell signature

In addition to the gene expression signatures that distinguished the three major branches of the cell line dendrogram, both the basal-epithelial-like cell lines and the luminal-epithelial-like cell lines expressed a set of genes that were absent in those lines expressing the stromal gene expression signature (Fig. 1B). This set was comprised of many genes that play roles in cell-to-cell contacts that seal the lumen or extracellular space in epithelial tissues (e.g. E-cadherin, plakoglobin and junctional-adhesion protein). This cluster likely distinguished genes involved in functions common between subtypes of epithelial cells, and therefore, comprised a molecular signature of epithelial cells (Fig. 1B).

7. Cell lines of ambiguous origins

Of particular interest were the breast derived cell lines that lacked expression of the common epithelial genes and expressed some characteristics of the "stromal" expression signature including BT-549, MCF-10A and MDA-MB-231. BT-549 showed strong expression of most of the genes in the stromal cluster (Fig. 1D) and lacked expression of the other signature expression patterns including the common epithelial cell pattern. This suggests that BT-549 may represents a myofibroblast-like line transformed *in vitro* or a line like Hs578T that was originally derived from a stromal-like tumor *in vivo* [18]. The MCF-10A [19] line is a well-studied breast model system that expressed characteristics of both the basal epithelial signature including keratin expression, as well as genes that comprised the stromal signature, but significantly, lacked expression of the common epithelial cell pattern. Perhaps the most enigmatic cell line was MDA-MB-231 that did not

show strong characteristics of any of the three signature patterns of expression (Fig. 1C–E) except for a fraction of genes that comprised the stromal cluster. This cell line has been previously shown to be similar to renal carcinoma derived cell lines in a separate study that compared cell lines derived from a diverse set of tumor types, and therefore, may represent a de-differentiated cell type that has lost expression of the signature of its tissue of origin [6]. Further gene expression studies on these cell lines, including studies of their responses to extracellular matrix stimuli, might distinguish their potential for differentiation into cells with a more clear relationship to their *in vivo* counterparts.

In a previous study, we shown that another cell line that has been used as a model of aggressive-metastatic-breast tumors, MDA-MB-435, showed a pattern of gene expression that was very similar to the pattern seen in seven independent melanoma derived cell lines (see http://genome-www.stanford.edu/nci60/images/figure-2c.html and [6]). This distinctive pattern included strong expression of many genes characteristic of melanocytes including tyrosinase, dopachrome tautomerase and S100-β and therefore suggested that the tested cell line was derived from a Melanoma and not from a breast carcinoma. A number of different samples of MDA-MB-435 derived from different sources showed a similar pattern suggesting that most, if not all, examples of this cell line are similarly misclassified (data not shown). These finding suggest that this cell line is not an appropriate model system for the study of breast carcinoma.

8. Breast tumor gene expression patterns

One of the most useful aspects of microarray technology is its utility in the study of gene expression patterns in clinical tumor specimens [7,13,15,29–32]. We have previously published a study of gene expression profiles of forty breast cancer patients that included twenty samples from patient's tumors before and after a sixteen week course of doxorubicin chemotherapy [7]. In order to identify the best set of genes to use for tumor classification, we utilized a statistical approach to identify the subset of genes that showed significant variation in expression across different patients/tumors, but which varied little in expression within paired samples from the same tumor [7]; this set of genes, termed the "intrinsic" gene set, was enriched for those genes whose expression patterns were characteristic of each tumor as opposed to those that varied as a function of sampling error. An example of a gene that showed this "intrinsic" property was ERBB2/HER2, which was expressed at high levels in some tumors and not others (forty-fold difference within this sample set), but which was consistent in expression in comparison between multiple samples taken from the same tumor.

Hierarchical clustering analysis using this "intrinsic" gene set of 476 cDNA clones (including 426 different genes) resulted in a novel molecular classification of the tumor samples on the basis of gene expression patterns (see Fig. 2 of [7] at http://genome-www.stanford.edu/molecularportraits/images/figure2.html). Importantly, the tumor samples cluster dendrogram from this analysis showed that the repeat biopsies taken from the same patient (i.e. the 20 "before" and "after" doxorubicin sample pairs) were found to almost always be more similar to each other than either was to any of the other tumors tested (17/20 "before" and "after" pairs were paired and 2/2 tumor/lymph node metastasis pairs were paired). This implied that every tumor is unique and has a distinctive gene expression "signature" or "portrait". These gene expression patterns distinguished four discrete tumor subtypes that, similar to the cell line studies, could be related to features of normal breast cell type *in vivo*. The classes of tumor subtypes identified were 1) a "luminal epithelial/ER+" subtype that was distinguished by high expression of a set of approximately twenty genes that included the ER gene and other genes known to be regulated by estrogen, 2) a normal breast-like group of samples that contained the three normal breast samples, a single fibroadenoma and 5 tumor samples, 3) a group of tumors most of which expressed high levels of the ERBB2/HER2 gene, and 4) a group of tumors that had gene expression patterns reminiscent of breast basal epithelial cells.

Building upon these studies, we recently reported gene expression patterns of an expanded set of 78 different breast tumors [33]. Cluster analysis of data derived from this larger tumor set re-identified the same four tumor subtypes in addition to one additional subclass of ER+ tumors, such that five subtypes were now distinguished. To explore whether this classification of breast tumors was clinically significant, we assigned each of the 51 tumors that comprised the cohort of "before" and "after" doxorubicin patients [34], to a class based upon its location within the tumor sample associated dendrogram and did Kaplan-Meier survival analysis (see http://genome-www.stanford.edu/mopo/clinical/ and [33]). We found that these subtypes, as defined by gene expression pat-

terns, had statistically significant different overall patient survival and relapse free survival characteristics. The luminal/ER+ positive patients were sub-divided into two classes of which a novel subtype was identified (Luminal B/C) that had a significantly worse outcome when compared to the rest of the ER+ tumors (Luminal A), which showed the most favorable outcomes. Furthermore, the set of patients classified as "basal-like" had outcomes as poor as those that over-expressed ERBB2/HER2. The patterns of gene expression that were distinguished in these studies represent novel molecular signatures of breast tumors that can be used to 1) develop new clinical tests based upon gene expression patterns to score for these subtypes, 2) identify candidate markers for diagnosis, 3) identify genes important for understanding the biology that distinguishes basal and luminal epithelial cells, and 4) identify subtype specific targets for developing therapeutic interventions.

9. An integrated cell line and breast tumor analysis

The set of genes that defined epithelial cell characteristics within the cell line panel was remarkably similar to the set of genes that distinguished tumor subtypes amongst the breast carcinomas. This suggested that certain cell lines may be very good models for specific subtypes of tumors. In order to more directly compare and contrast primary breast tumors and cell lines, we created a single hierarchical clustering diagram using the aforementioned "intrinsic" gene list and data from 16 cell lines discussed above and our previously published study on 40 breast tumors and three normal breast samples (Fig. 2, and see Supplementary Information Fig. 4 for the complete cluster diagram – http://genome-www.stanford.edu/breast_cancer/cell_line_review2001/). As expected, the results showed a similar grouping of the tumors samples into at least four subtypes including luminal/ER+ (dark blue), ERBB2/HER2+ (pink), normal breast-like (green), and a basal-like classes (dark red). The cell lines, regardless of their presumed cell-type of origin, clustered together on a single large dendrogram branch separate from all of the tumors, however, they were also similarly subdivided into the basal, luminal, and stromal-like groups described above (Fig. 2).

The luminal/ER+ signature was most strongly expressed in a large group of ER expressing tumors and the subset of cell lines that expressed the luminal gene expression signature (blue dendrogram branch and Fig. 2B). The pattern of expression of this set of genes showed both quantitative and qualitative differences between the luminal/ER+ tumors in comparison to the luminal cell lines. These tumor samples were comprised of at most 60–70% tumor cells, but expressed stronger and more consistent levels of this gene set when compared to the "pure" population of cell derived from a single cell line. Given the difficulty in establishing cultures from cells expressing luminal characteristics, the loss of expression of these genes may be related to the process of establishment or maintenance of cell lines *in vitro*. It was interesting that the ERBB2+ tumor subtype expressed fewer of these genes than the luminal-like/ER+ tumors, and in most cases, expression levels lower than the luminal cell lines. These patterns were consistent with the notion that the relative level of expression of this set of genes may reflected the degree of luminal differentiation of the tumor samples. Taken together, these results suggested that the best models for ER+ luminal epithelial cell derived tumors, choosing from the cell lines tested here, are MCF7, T47D and BT-474, with SK-BR-3 serving as a model of luminal-cell derived ER-negative/ERBB2 positive tumors.

We have previously shown that most of the genes highly expressed in a pattern similar to ERBB2 across large sets of breast tumors are all part of the co-amplified chromosomal region that contains ERBB2 [7, 35–38]. In this analysis, most of the tumors that expressed this gene expression signature as their primary distinguishing characteristic were clustered together and formed a discreet subtype (pink dendrogram branch, and Fig. 2D). A few other tumors that expressed the ERBB2 signature were present, however, they also expressed either the luminal or normal breast signatures and were clustered based upon those distinguishing characteristics. Both BT-474 and SK-BR-3 showed high level expression of the ERBB2 signature and therefore are likely appropriate cell line models for ER-positive and ER-negative, ERBB2 over-expressing tumors, respectively.

The basal-like tumors were distinguished by strong expression of keratin 5 and 17 relative to the luminal cell-like tumors (Fig. 2 and [7]). The basal epithelial gene expression signature expressed by HMEC lines was in part expressed by both the basal-like tumors and normal breast tissue, which appeared to be comprised predominantly of basal epithelial cells (dark red dendrogram branch and Fig. 2E). These results suggested that the basal like cell lines/HMEC cultures and their derivatives, are in general, appropriate model systems

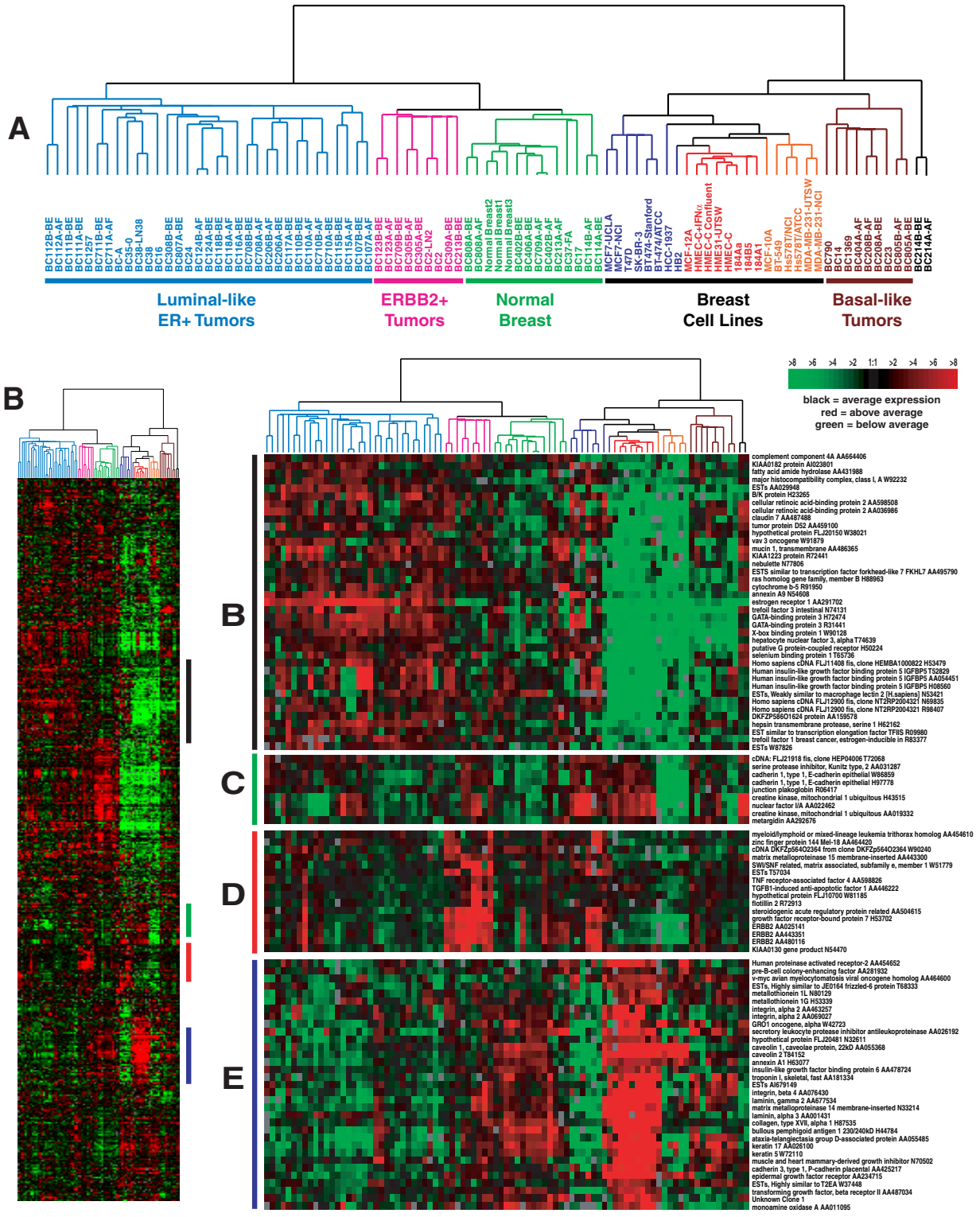
A
BC112B-BE
BC112A-AF
BC111B-BE
BC111A-BE
BC1257
BC711B-BE
BC711A-AF
BC-A
BC35-0
BC38-LN38
BC38
BC16
BC809B-BE
BC809A-BE
BC24
BC124B-AF
BC124A-BE
BC118B-BE
BC118A-AF
BC116A-BE
BC708B-BE
BC708A-AF
BC206A-BE
BC117A-BE
BC110B-BE
BC110A-AF
BC710B-AF
BC710A-BE
BC115B-BE
BC115A-AF
BC107B-AF
BC107A-AF
BC123B-BE
BC123A-AF
BC709B-BE
BC305B-AF
BC305A-BE
BC2-LN2
BC2
BC309A-BE
BC309B-BE
BC808A-AF
Normal Breast2
Normal Breast1
Normal Breast3
BC402B-BE
BC406A-BE
BC709A-AF
BC402B-AF
BC21-FA
BC17
BC114B-AF
BC114A-BE
MCF7-UCLA
MCF7-NCI
T47D
SK-BR-3
BT474-Stanford
BT474-ATCC
HCC-1937
HB2
MCF-12A
HMEC-C+IFNα
HMEC-C Confluent
HME31-UTSW
HMEC-C
184Aa
184B5
184A1
MCF-10A
BT-549
Hs578T/NCI
Hs578T/ATCC
MDA-MB-231-UTSW
MDA-MB-231-NCI
BC790
BC14
BC4369
BC404A-AF
BC208B-AF
BC208A-BE
BC23
BC805B-AF
BC805A-BE
BC214B-BE
BC214A-AF
Luminal-like ER+ Tumors
ERBB2+ Tumors
Normal Breast
Breast Cell Lines
Basal-like Tumors
B
>8 >6 >4 >2 1:1 >2 >4 >6 >8
black = average expression
red = above average
green = below average
B
C
D
E
complement component 4A AA664406
KIAA0182 protein AI023801
fatty acid amide hydrolase AA431988
major histocompatibility complex, class I, A W92232
ESTs AA029948
B/K protein H23265
cellular retinoic acid-binding protein 2 AA598508
cellular retinoic acid-binding protein 2 AA036986
claudin 7 AA487488
tumor protein DS2 AA459100
hypothetical protein FLJ20150 W38021
vav 3 oncogene W91879
mucin 1, transmembrane AA486365
KIAA1223 protein R72441
nebulette N77806
ESTS similar to transcription factor forkhead-like 7 FKHL7 AA495790
ras homolog gene family, member B H88963
cytochrome b-5 R91950
annexin A9 N54608
estrogen receptor 1 AA291702
trefoil factor 3 intestinal N74131
GATA-binding protein 3 H72474
GATA-binding protein 3 R31441
X-box binding protein 1 W90128
hepatocyte nuclear factor 3, alpha T74639
putative G protein-coupled receptor H50224
selenium binding protein 1 T65736
Homo sapiens cDNA FLJ11408 fis, clone HEMBA1000822 H53479
Human insulin-like growth factor binding protein 5 IGFBP5 T52829
Human insulin-like growth factor binding protein 5 IGFBP5 AA054451
Human insulin-like growth factor binding protein 5 IGFBP5 H08560
ESTs, Weakly similar to macrophage lectin 2 [H.sapiens] N53421
Homo sapiens cDNA FLJ12900 fis, clone NT2RP2004321 N69835
Homo sapiens cDNA FLJ12900 fis, clone NT2RP2004321 R98407
DKFZP586O1624 protein AA159578
hepsin transmembrane protease, serine 1 H62162
EST similar to transcription elongation factor TFIIS R09980
trefoil factor 1 breast cancer, estrogen-inducible in R83377
ESTs W87926
cDNA: FLJ21918 fis, clone HEP04006 T72068
serine protease inhibitor, Kunitz type, 2 AA031287
cadherin 1, type 1, E-cadherin epithelial W86859
cadherin 1, type 1, E-cadherin epithelial H97778
junction plakoglobin R06417
creatine kinase, mitochondrial 1 ubiquitous H43515
nuclear factor I/A AA022462
creatine kinase, mitochondrial 1 ubiquitous AA019332
metargidin AA292676
myeloid/lymphoid or mixed-lineage leukemia trithorax homolog AA454610
zinc finger protein 144 Mel-18 AA464420
cDNA DKFZp564O2364 from clone DKFZp564O2364 W90240
matrix metalloproteinase 15 membrane-inserted AA443300
SWI/SNF related, matrix associated, subfamily e, member 1 W51779
ESTs T57034
TNF receptor-associated factor 4 AA598826
TGFB1-induced anti-apoptotic factor 1 AA446222
hypothetical protein FLJ10700 W81185
flotillin 2 R72913
steroidogenic acute regulatory protein related AA504615
growth factor receptor-bound protein 7 H53702
ERBB2 AA025141
ERBB2 AA443351
ERBB2 AA480116
KIAA0130 gene product N54470
Human proteinase activated receptor-2 AA454652
pre-B-cell colony-enhancing factor AA281932
v-myc avian myelocytomatosis viral oncogene homolog AA464600
ESTs, Highly similar to JE0164 frizzled-6 protein T68333
metallothionein 1L N80129
metallothionein 1G H53339
integrin, alpha 2 AA463257
integrin, alpha 2 AA069027
GRO1 oncogene, alpha W42723
secretory leukocyte protease inhibitor antileukoproteinase AA026192
hypothetical protein FLJ20481 N32611
caveolin 1, caveolae protein, 22kD AA055368
caveolin 2 T84152
annexin A1 H63077
insulin-like growth factor binding protein 6 AA478724
troponin I, skeletal, fast AA181334
ESTs AI679149
integrin, beta 4 AA076430
laminin, gamma 2 AA677534
matrix metalloproteinase 14 membrane-inserted N33214
laminin, alpha 3 AA001431
collagen, type XVII, alpha 1 H87535
bullous pemphigoid antigen 1 230/240kD H44784
ataxia-telangiectasia group D-associated protein AA055485
keratin 17 AA026100
keratin 5 W72110
muscle and heart mammary-derived growth inhibitor N70502
cadherin 3, type 1, P-cadherin placental AA425217
epidermal growth factor receptor AA234715
ESTs, Highly similar to T2EA W37448
transforming growth factor, beta receptor II AA487034
Unknown Clone 1
monoamine oxidase A AA011095

Fig. 2. Integrated analysis of breast tumors and cell lines using the "intrinsic" gene set. Cluster diagram (see legend to figure 1) depicting the relationships between gene expression patterns in breast cancer derived cell lines and tumor specimens. A) Experimental sample associated dendrogram showing distinct tumor and cell line subtypes. B) Luminal/ER+ gene expression cluster. C) Common epithelial cell cluster containing E-cadherin. D) ERBB2+ amplicon cluster. E) Basal epithelial cell cluster. The color scale at the top right depicts the gene expression measured in each sample relative to the average expression for each gene as determined in the 87 different samples. The full microarray data files for all 87 experiments can be obtained from the Stanford Microarray Database at http://genome-www4.stanford.edu/MicroArray/SMD/, and the full cluster diagrams for Figs 1 and 2 can be seen at http://genome-www.stanford.edu/breast_cancer/cell_line_review2001/.

for the breast basal-like tumors. The absence of expression of any genes characteristic of the luminal signature suggests that HMEC cultures and their derivatives are NOT appropriate models of hormone responsive breast tumors, which comprise the majority of sporadic breast tumors [1]. HB2 and HCC-1937, as noted above, expressed genes characteristic of both the basal and luminal signatures, and therefore may be similar to basal-like tumors that often co-express cytokeratin 17 and 18 *in vivo* (M. van de Rijn, personal communication).

The cell lines that expressed the stromal cell signature (orange dendrogram branch), with the exception of MCF-10A, showed very few gene expression characteristics in common with any of the breast tumors, even those that were highly metastatic when sampled for microarray analysis. The cellular origins of these cell lines are still enigmatic and it is not clear which, if any, of these cell lines are appropriate models for breast carcinoma. It is interesting to note, however, that some of the cell lines contained within this cluster, in particular MDA-MB-231, are some of the most tumorigenic and aggressive in nude mouse xenograft models [39]. Most interesting amongst these cell lines was MCF-10A, which expressed some genes characteristic of basal like tumors including keratin expression, but lacked expression of the common epithelial signature. Interestingly, this cell line is not tumorigenic in xenograft models.

10. Summary

The advent of the DNA microarray technology has enabled researchers to measure genomic scale gene expression in human cancers and cell lines. The exploration of these gene expression patterns is challenging oncologists and pathologists to re-assess traditional classifications of cancer and incorporate molecular features into treatment regimens and drug development strategies. The great strength of cDNA microarray studies coupled to hierarchical clustering analysis is the ability to objectively identify sets of coordinately expressed genes and display the data in a format that a biologist can utilize to form hypotheses.

The data presented here, and in our previous studies, have shown that many different gene selection criteria, across different sets of breast tumors and cell lines, consistently identified similar sets of genes that can serve as markers for probing the biology of breast cancer [6, 7,33]. The comparison of gene expression patterns between cell lines and tumors is dominated by differences related mostly to the proliferative index of the samples, with most cell lines growing at a much faster rate than *in vivo* tumor cells [6–8,13,15]). However, in the case of breast tumors, cell lines and tumors share many aspects of their gene expression patterns that can be related to the normal and pathological physiology that distinguishes breast cell types *in vivo*. These gene sets include 1) the basal epithelial cluster, 2) the luminal epithelial/ER+ cluster, 3) the ERBB2+ amplicon cluster, 4) the proliferation cluster, and 5) the interferon cluster [7,8]. Remarkably, the classes of tumors as defined by gene expression, in part, are consistent with current markers that are used for breast cancer stratification and prognostication (e.g. ER status, ERBB2 status, proliferative index) [33].

In addition to re-identifying and elucidating traditional classes of tumors, gene expression patterns are revealing novel subtypes of tumors that appear both biologically and clinically distinct. In the study by Sørlie et al., the class of tumors distinguished by expression of the basal epithelial signature, including expression of cytokeratins 5 and 17, showed an outcome as poor as ERBB2 over-expressing tumors and were as numerous. Similarly, this study also identified a sub-group of patients with ER-positive tumors, traditionally classified as having a good prognosis, that had very poor outcomes [33]. The similarities and differences between cell lines and tumors should allow a much more informed choice to be made about the appropriateness of any given cell line model for a particular aspect of tumor biology to be studied *in vitro*.

In addition to the dominant patterns of gene expression described in this review, there is tremendous additional variation in gene expression patterns in comparison between tumor subtypes (see Supplementary Information Figures at http://genome-www.stanford.edu/breast_cancer/cell_line_review2001/).

Gene expression studies of fifty to one hundred tumor specimens likely does not have the power to identify all of the inherent biologic diversity of tumors. It remains to be determined whether one, a few, or tens to hundreds of markers will be necessary to identify distinguishing characteristics of tumors in the clinical setting. It is likely that larger gene expression profiling studies, and/or large *in situ* or immunohistochemistry studies using candidate markers identified in microarray studies, will be necessary to distinguish all of the clinically relevant variation in biology that can be exploited to develop better patient management algorithms and targeted drug strategies.

Acknowledgements

We are grateful to David Botstein and Patrick O. Brown for guidance and for providing the resources that were used in this study. We also thank John C. Matese for his efforts in creating and maintaining the website that supports this paper, and the following individuals for their cell lines or mRNA samples that were used in this study (Jerry Shay, H.S. Wiley, Fuyuhiko Tamanoi, Martha Stampfer and Paul Yaswen) and Robert Strausberg for critical reading of this manuscript.

References

[1] F.A. Tavassoli and S.J. Schnitt, *Pathology of the breast*, New York: Elsevier. xiii, 1992, pp. 669.
[2] P.O. Brown and D. Botstein, Exploring the new world of the genome with DNA microarrays, *Nat Genet* **21**(1 Suppl) (1999), 33–37.
[3] L. Ronnov-Jessen, O.W. Petersen and M.J. Bissell, Cellular changes involved in conversion of normal to malignant breast: importance of the stromal reaction, *Physiol Rev* **76**(1) (1996), 69–125.
[4] M.R. Stampfer and P. Yaswen, Culture systems for study of human mammary epithelial cell proliferation, differentiation and transformation [see comments], *Cancer Surv* **18** (1993), 7–34.
[5] J. Taylor-Papadimitriou et al., Keratin expression in human mammary epithelial cells cultured from normal and malignant tissue: relation to *in vivo* phenotypes and influence of medium, *J Cell Sci* **94**(Pt 3) (1989), 403–413.
[6] D.T. Ross et al., Systematic variation in gene expression patterns in human cancer cell lines [see comments], *Nat Genet* **24**(3) (2000), 227–235.
[7] C.M. Perou et al., Molecular Portraits of Human Breast Tumors, *Nature* **406** (2000), 747–752.
[8] C.M. Perou et al., Distinctive gene expression patterns in human mammary epithelial cells and breast cancers, *Proc Natl Acad Sci USA* **96**(16) (1999), 9212–9217.
[9] M.B. Eisen et al., Cluster analysis and display of genome-wide expression patterns, *Proc Natl Acad Sci USA* **95**(25) (1998), 14863–14868.
[10] V.R. Iyer et al., The transcriptional program in the response of human fibroblasts to serum [see comments], *Science* **283**(5398) (1999), 83–87.
[11] P.T. Spellman et al., Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization, *Mol Biol Cell* **9**(12) (1998), 3273–3297.
[12] J.N. Weinstein et al., An information-intensive approach to the molecular pharmacology of cancer, *Science* **275**(5298) (1999), 343–349.
[13] A.A. Alizadeh et al., Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling [see comments], *Nature* **403**(6769) (2000), 503–511.
[14] A.P. Gasch et al., Genomic expression programs in the response of yeast cells to environmental changes [In Process Citation], *Mol Biol Cell* **11**(12) (2000), 4241–4257.
[15] C.M. Perou, P.O. Brown and D. Botstein, *Tumor classification using gene expression patterns from DNA microarrays*, New Technologies for life sciences: A Trends Guide, 2000, pp. 67–76.
[16] T.R. Hughes et al., Functional discovery via a compendium of expression profiles, *Cell* **102**(1) (2000), 109–126.
[17] T.M. Paine et al., Characterization of epithelial phenotypes in mortal and immortal human breast cells, *Int J Cancer* **50**(3) (1992), 463–473.
[18] A.J. Hackett et al., Two syngeneic cell lines from human breast tissue: the aneuploid mammary epithelial (Hs578T) and the diploid myoepithelial (Hs578Bst) cell lines, *J Natl Cancer Inst* **58**(6) (1977), 1795–1806.
[19] H.D. Soule et al., Isolation and characterization of a spontaneously immortalized human breast epithelial cell line, MCF-10, *Cancer Res* **50**(18) (1990), 6075–6086.
[20] R. Cailleau, M. Olive and Q.V. Cruciger, Long-term human breast carcinoma cell lines of metastatic origin: preliminary characterization, *In Vitro* **14**(11) (1978), 911–915.
[21] H.D. Soule et al., A human cell line from a pleural effusion derived from a breast carcinoma, *J Natl Cancer Inst* **51**(5) (1973), 1409–1416.
[22] I. Keydar et al., Establishment and characterization of a cell line of human breast carcinoma origin, *Eur J Cancer* **15**(5) (1979), 659–670.
[23] E.Y. Lasfargues, W.G. Coutinho and E.S. Redfield, Isolation of two human tumor epithelial cell lines from solid breast carcinomas, *J Natl Cancer Inst* **61**(4) (1978), 967–978.
[24] G.E. Tomlinson et al., Characterization of a breast cancer cell line derived from a germ-line BRCA1 mutation carrier, *Cancer Res* **58**(15) (1998), 3237–3242.
[25] M.R. Stampfer et al., Gradual Phenotypic Conversion Associated with Immortalization of Cultured Human Mammary Epithelial Cells, *Mol Biol Cell* **8**(12) (1997), 2391–2405.
[26] L. Ronnov-Jessen et al., The origin of the myofibroblasts in breast cancer. Recapitulation of tumor environment in culture unravels diversity and implicates converted fibroblasts and recruited smooth muscle cells, *J Clin Invest* **95**(2) (1995), 859–873.
[27] M. Stampfer, The HMEC Homepage.
[28] C. Pechoux et al., Human mammary luminal epithelial cells contain progenitors to myoepithelial cells, *Dev Biol* **206**(1) (1999), 88–99.
[29] U. Alon et al., Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed

by oligonucleotide arrays, *Proc Natl Acad Sci USA* **96**(12) (1999), 6745–6750.

[30] T.R. Golub et al., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science* **286**(5439) (1999), 531–537.

[31] H. Okabe et al., Genome-wide analysis of gene expression in human hepatocellular carcinomas using cDNA microarray: identification of genes involved in viral carcinogenesis and tumor progression, *Cancer Res* **61**(5) (2001), 2129–2137.

[32] J.B. Welsh et al., Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer, *Proc Natl Acad Sci USA* **98**(3) (2001), 1176–1181.

[33] T. Sørlie et al., Gene expression patterns of breast carcinomas distinguish tumor subclasses with potential clinical implications, Submitted, 2001.

[34] S. Geisler et al., 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**(6) (2001), 2505–2512.

[35] C. Moog-Lutz et al., MLN64 exhibits homology with the steroidogenic acute regulatory protein (STAR) and is overexpressed in human breast carcinomas, *Int J Cancer* **71**(2) (1997), 183–191.

[36] J.R. Pollack et al., Genome-wide analysis of DNA copy-number changes using cDNA microarrays, *Nat Genet* **23**(1) (1999), 41–46.

[37] J.S. Ross and J.A. Fletcher, The HER-2/neu oncogene in breast cancer: prognostic factor, predictive factor, and target for therapy, *Stem Cells* **16**(6) (1998), 413–428.

[38] D. Stein et al., The SH2 domain protein GRB-7 is co-amplified, overexpressed and in a tight complex with HER2 in breast cancer, *Embo J* **13**(6) (1994), 1331–1340.

[39] H. Pulyaeva et al., MT1-MMP correlates with MMP-2 activation potential seen after epithelial to mesenchymal transition in human breast carcinoma cells, *Clin Exp Metastasis* **15**(2) (1997), 111–120.