Original Article

Gene Expression Signatures in Breast Cancer Distinguish Phenotype Characteristics, Histologic Subtypes, and Tumor Invasiveness

Vincente Pedraza, PhD[1]; Jose A. Gomez-Capilla, PhD[2]; Georgia Escaramis, BsC[3]; Carolina Gomez, PhD[2]; Pablo Torné, MD[4]; Jose M. Rivera, MD[5]; Angel Gil, PhD[2]; Patricia Araque, PhD[1]; Nicolas Olea, PhD[1]; Xavier Estivill, PhD[3,6,7]; and M. Esther Fárez-Vidal, PhD[2]

BACKGROUND: The development of reliable gene expression profiling technology is having an increasing impact on the understanding of breast cancer biology. **METHODS:** In this study, microarray analysis was performed to establish gene signatures for different breast cancer phenotypes, to determine differentially expressed gene sequences at different stages of the disease, and to identify sequences with biologic significance for tumor progression. Samples were taken from patients before their treatment. After microarray analysis, the expression level of 153 selected genes was studied by real-time quantitative polymerase chain reaction analysis. **RESULTS:** Several gene sequences were expressed differentially in tumor samples versus control samples and also were associated with different breast cancer phenotypes, estrogen receptor status, tumor histology, and grade of tumor differentiation. In lymph node-negative tumors were identified a set of genes related to tumor differentiation grade. **CONCLUSIONS:** Several differentially expressed gene sequences were identified at different stages of breast cancer. ***Cancer*** **2010;116:486–96.**

KEYWORDS: breast cancer, gene expression signature, tumor invasiveness, microarrays, real-time quantitative polymerase chain reaction.

Breast cancer is a major cause of cancer-related morbidity and mortality among women worldwide. The disease is defined by a heterogeneous clinical course that encompasses a wide variety of pathologic entities. At the molecular level, a complex array of genetic alterations with influence on mammary cell functions characterizes the multistep nature of tumor progression.[1-3]

The recent introduction of DNA microarray technology has enabled the classification of malignant tumors on a genome-wide scale by simultaneously monitoring the expression of thousands of genes in study samples. Microarray gene expression profiling also is having a growing impact on other aspects of breast cancer, including prognosis, treatment, and prediction of response to therapy. The results of DNA microarray analyses performed to determine their prognostic value in breast cancer have revealed some discrepancies.[4-6] Differences in the microarray technology used, the selection of genes on each array, the data analysis methods, and the patient selection criteria most likely are responsible for the lack of an observed consensus.[7] In addition, gene expression signatures have been obtained from different sets of patients and for different objectives. For example, Sorlie et al[8] focused on subclassification, and van't Veer et al[9] focused on survival prediction. The clinical behavior and response to treatment of malignant breast tumors are equally diverse.[10] Consequently, as concluded by Jensen and Hovig,[11] there is a need for further research, clinical validation, and resolution of technologic issues before general agreement can be reached on the predictive value of gene profiles in breast cancer.

Corresponding author: M. Esther Fárez-Vidal, PhD, Biochemistry Department, University of Granada School of Medicine, Avda. de Madrid s/n, 18012 Granada, Spain; Fax: (011) 3458249015; efarez@ugr.es

[1]Department of Radiation Oncology, University of Granada School of Medicine, Granada, Spain; [2]Department of Biochemistry, University of Granada, Granada, Spain; [3]Center for Public Health Biomedical Research Network Research-Epidemiology and Public Health, Genes and Disease Program, Center for Genomic Regulation, Barcelona, Spain; [4]Department of Surgery, San Cecilio University Hospital, Granada, Spain; [5]Department of Pathology, Cruces University Hospital, Baracaldo (Vizcaya), Spain; [6]Experimental and Health Sciences Department, Pompeu Fabra University, Barcelona, Spain; [7]National Center of Genotyping, Barcelona Node, Barcelona, Spain

We thank Enrique Alava for help with the gathering of tumor samples, Professor Antonio Suarez for critical comments and contributions to the initial development of the study, and Richard Davies for assistance with the English version.

DOI: 10.1002/cncr.24805, **Received:** February 27, 2009; **Revised:** May 7, 2009; **Accepted:** June 10, 2009, **Published online** December 22, 2009 in Wiley InterScience (www.interscience.wiley.com)

In the current investigation, microarray analyses were performed in a controlled case-control study to establish gene signatures for 3 well-defined breast cancer phenotypes: 1) patients with tumors classified as T1, T2, or T3; lymph node-negative (N0); and estrogen receptor (ER)-positive; 2) patients with tumors classified as T1, T2, or T3; N0, and ER-negative; and 3) patients with tumors classified as T1, T2, or T3; lymph node-positive (N+) (N1, N2, or N3); and ER-positive or ER-negative. The general objectives were 1) to analyze the relation between these phenotypes and their respective gene expression profiles, 2) to determine the genetic sequences expressed differentially in different stages of the disease, 3) to investigate the metabolic gene pathways involved, and 4) to identify sequences with biologic significance for tumor progression among the group of genes with altered expression profiles.

MATERIALS AND METHODS

Tumor and Normal Tissue Samples

Tumor samples were taken from primary malignant breast tumors from nontreated patients by the excision of a fragment of tumor mass $\geq$100 mg during the initial surgical procedure. Samples were snap-frozen in liquid nitrogen within 10 minutes after excision and stored at -80°C until RNA extraction. Each tumor tissue sample had a paired normal tissue sample. Although preneoplastic molecular changes certainly could have occurred in the adjacent unaffected tissue, normal tissue samples (mass $\geq$500 mg) were taken from quadrants of the ipsilateral mammary gland that were clinically free of tumor. The rationale for not obtaining breast tissue from the unaffected contralateral breast was to reduce surgical handling to the diseased mammary gland. The "normal" samples were subjected to a meticulous histologic analysis to guarantee the total absence of epithelial tumor cells.

Before the study, all medical records and tumor sections were reviewed by an oncologist and a surgical pathologist. Informed consent was obtained from all patients for the study, which was approved by the ethics committee of our institution. Frozen tumor sections were stained with hematoxylin and eosin, and only those with $\geq$70% epithelial tumor cells and few infiltrating lymphocytes or necrotic tissue were selected for RNA extraction. For RNA isolation, 30-μm sections (RNA content $\geq$100 μg) of tumor and normal tissue were used.

Tumor Phenotypes Analyzed

Primary breast tumor samples that we analyzed by microarray corresponded to the following phenotypes: Phenotype 1 (Ph1), tumors classified as N0; ER-positive; and T1, T2, or T3 (8 patients); Phenotype 2 (Ph2), tumors classified as N0; ER-negative; and T1, T2, or T3 (10 patients); and Phenotype 3 (Ph3), tumors classified as lymph node-positive; ER-positive; and T1, T2, or T3 (13 patients).

The TNM classification of tumors strictly followed the 2002 American Joint Committee on Cancer criteria. The categories N0 and N+ were established pathologically after surgical axillary lymph node dissection (N0, histologically tumor-free axillary lymph nodes; N+, axillary lymph nodes with tumor). Axillary invasion in 1 to 3 lymph nodes, in 4 to 9 lymph nodes, and in $\geq$10 lymph nodes was recorded as N1, N2, and N3, respectively. ERs were detected by immunohistochemistry. ER expression was detected by antibody clone 1D5 (M7047; DakoCytomation, Glostrup, Denmark). Tumors were regarded as ER positive when $\geq$10% of tumor cells had nuclear protein expression. Staining and scoring were performed on surgical specimens that were used for diagnosis in the Pathology Department of the San Cecilio University Hospital (Granada, Spain). Pathologic tumor classification was established according to the greatest dimension of the primary tumor in the surgical specimen: T1 (greatest primary tumor dimension, $\leq$19 mm), T2 (greatest primary tumor dimension, 20-49 mm), and T3 (greatest primary tumor dimension, $\geq$50 mm).

Analysis of Variables

In tumor samples, gene expression was analyzed according to the following criteria: 1) tumor size (T1, T2, or T3); 2) axillary lymph node involvement (N0, N1, N2, or N3); 3) ER status (ER-positive or ER-negative); 4) tumor phenotype (Ph1, N0; ER-positive; and T1, T2, or T3; Ph2, N0; ER-negative; and T1, T2, or T3; or Ph3, N+; ER-positive or ER-negative; and T1, T2, or T3); 5) grade of tumor differentiation (well differentiated [grade 1], moderately differentiated [grade 2], or poorly differentiated [grade 3]); and 6) tumor histology (ductal carcinoma [DC] or lobular carcinoma [LC])).

RNA Isolation and Microarray Procedures

Target preparation, microarray hybridization, and gene expression analysis were performed with the Affymetrix Genechip System (Affymetrix, Santa Clara, Calif) at Progenika Biopharma (Derio, Spain) and at San Cecilio University Hospital. Total RNA was extracted sequentially for microarray analysis from each frozen sample using Trizol (Invitrogen, Carlsbad, Calif) according to the

manufacturer's instructions and was purified using the RNeasy Mini Kit (Qiagen Inc., Valencia, Calif). RNA integrity was assessed by agarose gel electrophoresis. Gene expression profiles were determined by using the Affymetrix Human Genome U 133 Plus 2.0 Genechip according to the manufacturer's recommendations. In brief, 5 μg of total RNA were used in a reverse transcription reaction to synthesize complementary DNA (cDNA) with a primer containing poly-dT and T7 RNA polymerase promoter sequences. Double-stranded cDNA was purified with GeneChip Cleanup Module (Affymetrix) and used as a template for in vitro transcription.

The in vitro transcription reaction was performed using the IVT labeling kit (Affymetrix). Labeled RNA was purified with the GeneChip Cleanup Module (Affymetrix) and was quantified by spectrophotometry. Biotinylated complementary RNA (cRNA) was fragmented at 94°C for 35 minutes in fragmentation buffer (5 × buffer containing 200 mM Tris-acetate [pH 8.1]/500 mM KOAc, and 150 mM MgOAc) and hybridized to the microarrays in 200 μL of hybridization solution containing 15 μg labeled target in 1 × buffer containing 0.1 M Mes, 1.0 M NaCl, 20 mM ethylenediamine tetraacetic acid (EDTA), and 0.01%/ Tween-20 (Mes buffer); 0.1 mg/mL herring sperm DNA; 10% dimethyl sulfoxide; 0.5 mg/mL bovine serum albumin; 50 pM control oligonucleotide B2; and 1 × eukaryotic hybridization controls (bioB, bioC, bioD, cre). Both control oligonucleotide B2 and eukaryotic hybridization controls were purchased from Affymetrix.

Target cRNA was hybridized overnight to each oligonucleotide microarray containing 54,613 human gene probes and expressed sequence tags. Arrays were placed on a rotisserie and rotated at 60 revolutions per minute for 16 hours at 45°C. After hybridization, arrays were washed with 6 × 0.9 M NaCl, 60 mM NaH_2PO_4, 6 mM EDTA, and 0.01% Tween-20 (SSPE-T) at 30°C on a fluidics station (FS450; Affymetrix) for 10 cycles of 2 mixes per cycle and subsequently with 0.1 M Mes, 0.1 M NaCl, and 0.01% Tween-20 at 50°C for 6 cycles of 15 mixes per cycle. Then, the arrays were stained for 5 minutes at 35°C with a streptavidin-phycoerythrin conjugate (Molecular Probes, Eugene, Ore), followed by 10 washing cycles of 4 mixes per cycle with 6 × SSPE-T.

To enhance signals, arrays were stained further with antistreptavidin antibody solution for 5 minutes followed by a 5-minute staining with a streptavidin-phycoerythrin conjugate. After 15 washing cycles of 4 mixes per cycle, the hybridized arrays were scanned using the GeneChip Scanner 3000 (Affymetrix). The hybridization intensity (signal) of each transcript was determined using the GeneChip Operating Software (GCOS 1.4; Affymetrix). Intensity values were scaled such that the overall fluorescence intensity of each array was equivalent.

Microarray Data Analysis

Probe set measurements were generated from quantified Affymetrix image (.CEL) files using the Robust Multichip Average[12] from the Affy package (Bioconductor: available at: http://www.bioconductor.org accessed on December 20, 2008), following the 3 steps: background correction, quantile normalization, and log2-transformation.

Significance analysis of microarrays[13] (SAM) was used to identify differences in gene sequences among conditions. Two different strategies were used, depending on whether paired or unpaired samples were used. For comparisons of tumor phenotypes against normal samples, only paired tissue samples were used. The change in expression of each gene was calculated by determining the fold-change (FC) ratio as 2 to the power of the mean of the paired intensity differences. The moderated, Student *t* statistic for paired data was used to identify significant differences. For comparisons of different conditions in tumor samples, the FC was determined by the ratio of the mean intensity of each group, and the moderated Student *t* test for unpaired data was applied for significant purposes. To correct for multiple testing, the false discovery rate (FDR) was estimated for each comparison. When comparing tumor phenotypes against control samples, an estimated FDR of <0.01 was considered significant. In comparisons among conditions in tumor samples, an FDR of <0.05 was considered significant. A less stringent threshold for the FDR test was selected in the latter case, because some of the genes were validated further by quantitative reverse transcriptase polymerase chain reaction (RT-qPCR) analysis. All analyses were performed using the R software package Siggens (Bioconductor; available at: http://www.bioconductor.org).

Unsupervised hierarchical clustering analysis was performed on the most variable genes between tumor and control samples deduced from SAM. Pearson correlation was used as similarity measure, and complete linkage for the agglomerative method was applied. The microarray data have been deposited in the public repository (Gene Expression Omnibus,[14] accession number GSE10810).

Real-Time RT-PCR Analysis

On the basis of the array results, expression levels of 153 selected genes were evaluated by quantitative RT-PCR

(qPCR) analysis. cDNAs were reverse-transcribed from total RNA samples (10 ng per sample) using the High-Capacity cDNA Archive Kit (Applied Biosystems, Foster City, Calif) according to the manufacturer's instructions. TaqMan PCR reactions were performed on cDNA samples using the TaqMan Universal PCR Master Mix (Applied Biosystems) according to the manufacturer's instructions in conjunction with custom 7900 microfluidic cards (Applied Biosystems) and ABI PRISM 7900 HT Sequence Detection Systems. TaqMan strategies for each gene were developed as Assay-on-Demand by Applied Biosystems.

The set of genes (see Additional Data File 1 [available at: http://www.ugr.es/~efarez/]) contained 2 housekeeping genes: glyceraldehyde-3-phosphate dehydrogenase (*gapdh*) and ribosomal protein 18S (mandatory controls designed into each experiment by the manufacturer). In this study, 4 more housekeeping genes also were selected (ribosomal protein 6 [*rpl6*], *rpl10*, actin β [*actb*], and β-2 microglobulin [*b2m*]) based on no-change criteria among conditions in microarray experiments. Absolute threshold cycle values (Ct) were determined with SDS software (Applied Biosystems). Control genes were analyzed with GeNorm software, which was developed to determine optimal control genes from a longer list of putative control genes that had the lowest variation among conditions.

GeNorm[15] was used to obtain a normalization factor, and values were log2-transformed for the statistical analysis. Real-time quantitative PCR (qPCR) data were evaluated according to the recommendations of Yuan et al, [16] and ΔCt values were used as dependent variables in the statistical analysis.

Variables were examined for normality using the Shapiro-Wilk test (number of patients <30) and the Kolmogorov-Smirnof test (number of patients ≥ 30). When normality was plausible, the Student *t* test for unpaired data was used to compare between different conditions; otherwise, the Mann-Whitney *U* test was used. The FDR[17] was estimated to correct for multiple tests, and an FDR <0.05 was considered significant.

Concordance Between Microarray and Real-Time PCR Results

An agreement analysis between microarray and RT-PCR FC values was performed for each condition using the concordance correlation coefficient[18] (CCC), which yields values between 0 (independence) and 1 (perfect agreement).

Online Supplemental Material

Additional Data File 1, S1 (available at: http://www.ugr.es/~efarez/) lists sets of genes analyzed by real-time RT-PCR. Additional Data File 2, S2 (available at: http://www.ugr.es/~efarez/) provides the FC values associated with each sequence obtained in the microarray analysis of paired samples. Additional Data File 3, S3 (available at: http://www.ugr.es/~efarez/) provides the FC values associated with each sequence obtained in the microarray analysis of nonpaired samples. Additional Data File 4, S4 (available at: http://www.ugr.es/~efarez/) lists validated sequences that were obtained by qPCR in the different comparisons and indicates the *P* value and FC.

RESULTS

Tumor and Nontumor (Control) Samples

Forty-nine tumor samples and 43 samples of normal breast tissue were available for study. Of the 49 tumor samples, 18 were excluded from the analysis for insufficient tumor cell density (11 samples), or for an inadequate amount (5 samples), or for poor quality of extracted RNA (2 samples). Of the 43 normal tissue samples, 16 were excluded after histologic demonstration of neoplastic cells in the study specimen. Therefore, the final study sample of valid specimens comprised 31 tumor samples and 27 normal breast tissue samples, including 26 tumor samples that were paired with their corresponding 26 normal samples.

Microarray Results of Malignant Breast Tumors

Fifty-eight samples of human breast tissue (tumor and nontumor) were analyzed using the Human Genome U133 Plus 2.0 chip from Affymetrix. The unsupervised hierarchical clustering classified breast tissue samples into 2 main branches (Fig. 1), with the right branch containing all tumor breast samples (blue) and the left branch containing all nontumor (control) samples (red). It can be observed in Figure 1 that there was good homogeneity among normal control samples but wide variability among tumor samples.

Gene sequences that were expressed differentially in the analysis of paired samples are shown in Table 1 (Additional Data File 2 S2 [available at: http://www.ugr.es/~efarez/] includes the *P* value and FC value associated with each sequence). Numerous gene sequences were expressed differentially between nontumor control samples and Ph1, Ph2, or Ph3 tumors (FDR, <0.01).

Table 2 shows the gene sequences that were expressed differentially in the analysis of nonpaired samples

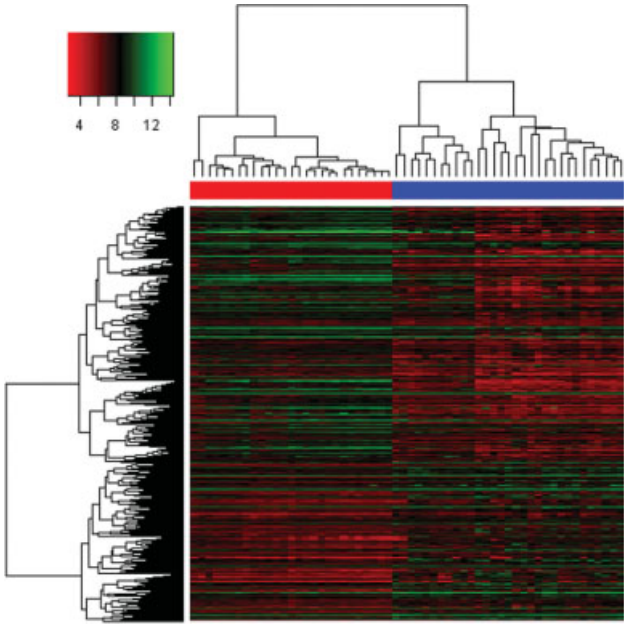

Figure 1. This is an unsupervised hierarchical cluster of 58 samples that was constructed using 8088 differentially expressed sequences. Each sample is color coded for its experimental condition. The red bar represents nontumor tissue samples, and the blue bar represents tumor tissue samples.

(Additional Data File 3, S3 [available at: http://www.ugr.es/~efarez/] includes the *P* value and FC value associated with each sequence). Significant differences in gene sequences were observed for the following comparisons: tumor phenotype, ER status (ER-positive vs ER-negative), histologic subtype (LC vs DC), and tumor differentiation (grade 1/2 vs grade 3). When samples without axillary lymph node involvement (N0) were studied, we observed 1311 differentially expressed gene sequences for tumor differentiation grade (N0/grade 1/2 tumors vs N0/grade 3 tumors; FDR, <0.05).

Database for Annotation, Visualization, and Integrated Discovery/Expression Analysis Systematic Explorer Analysis

The Database for the Annotation, Visualization, and Integrated Discovery (DAVID) classification system is a powerful bioinformatic tool for classifying genes according to their function. DAVID analysis identifies families of genes that may play significant roles in specific pathways, biologic processes, and molecular functions. In the current study, it was used to classify the 8088 differentially expressed sequences between tumor samples and nontumor samples (Table 2). The DAVID database annotated 72.05%, 76.75%, and 76.95% of these array sequences in relation to biologic processes, cellular

Table 1. The Number of Probe Sets That Showed Differentially Expressed Genes With a 0.01 False Discovery Rate According to Affymetrix Arrays for Paired Data

Comparison[a]	No. of Samples	No. of Significant Probes With FDR <0.01	
		UpR	DownR
Ph1 vs control	8 vs 8	2121	2049
Ph2 vs control	6 vs 6	7	261
Ph3 vs control	12 vs 12	2928	3757

FDR indicate false discovery rate; UpR, up-regulated; DownR, down-regulated; Ph, phenotype.
[a]Ph1: tumors classified as lymph node negative (N0), estrogen receptor positive (ER+), and T1, T2, or T3; Ph2: tumors classified as N0, ER negative (ER−), and T1, T2, or T3; Ph3: tumors classified as lymph node positive, ER+ or ER−, and T1, T2, or T3.

Table 2. The Number of Probe Sets That Had Differentially Expressed Genes With a 0.05 False Discovery Rate According to Affymetrix Arrays When Working With Unpaired Data

Samples Implicated	No. of Samples	No. of Significant Probes	
		UpR	DownR
All samples			
Tumor vs control[a]	31 vs 27	3752	4336
Tumor phenotypes[b]			
Ph1 vs Ph2	8 vs 10	2783	NS
Ph1 vs Ph3	8 vs 13	9	NS
Ph2 vs Ph3	10 vs 13	NS	NS
All tumors			
T1 vs T2	10 vs 15	NS	NS
T1/T2 vs T3	25 vs 6	NS	NS
N0 vs N+	18 vs 13	NS	NS
ER+ vs ER−	19 vs 12	1818	4
LC vs DC	9 vs 22	966	NS
Grade 1/2 vs grade 3	12 vs 10	584	NS
N0 tumors			
T1 vs T2	7 vs 7	NS	NS
T1/T2 vs T3	14 vs 4	NS	NS
LC vs DC	4 vs 14	NS	NS
Grade 1/2 vs grade 3	7 vs 7	1311	NS

UpR indicates up-regulated; DownR, down-regulated; Ph, phenotype; NS, none significantly differentially expressed; N0, negative lymph node status; ER, estrogen receptor; +, positive; −, negative; LC, lobular carcinoma; DC, ductal carcinoma.
[a]The false discovery rate used was <0.001.
[b]Ph1: tumors classified as N0, ER+, and T1, T2, or T3; Ph2: tumors classified as N0, ER−, and T1, T2, or T3; Ph3: tumors classified as N+, ER+ or ER−, and T1, T2, or T3.

components, and molecular functions, respectively, in Gene Ontology and *Kyoto Encyclopedia of Genes and Genomes* terms.

Figure 2 depicts results using an Expression Analysis Systematic Explorer Analysis (EASE) cutoff *P* value of .05 for the analysis. According to biologic processes, these sequences were involved in macromolecule metabolic processes (50.07%), primary metabolic processes (44.16%), cellular processes (28.81%), and cellular component organization and biogenesis (14.86%), among others. In the analysis of cellular components, differentially expressed sequences were mainly intracellular (59.89%), cell parts (43.72%), membrane-bound organelle (42.93%), and nuclei (26.59%), among others. In the analysis of molecular functions, sequences were involved in binding (66.61%), protein binding (35.46%), and transferase activity (8.69%).

Validation of Tumor-Associated Genes by qPCR Analysis

Findings of the expression array study were validated by using qPCR analysis to test 159 genes (including the 6 control housekeeping genes). These genes were selected according to the *P* values and the FC values for the differentially expressed sequences obtained in the comparisons shown in Table 2. The control housekeeping genes selected were *gapdh, rna 18s, rpl6, rpl10, actb*, and *b2m*. Validation was performed for 54 samples, including 40 tumor samples (11 samples with the Ph1 phenotype, 12 samples with the Ph2 phenotype, and 17 samples with the Ph3 phenotype) and 14 nontumor control samples. Twenty-nine of 40 tumor samples had been analyzed previously by microarray analysis, and 11 samples (4 samples with the Ph1 phenotype, 3 samples with the Ph2 phenotype, and 4 samples with the Ph3 phenotype) had not (new tumor samples).

Table 3 and Additional Data File 4, S4 (available at: http://www.ugr.es/~efarez/) show the validated sequences for each comparison, indicating the *P* value, the FC value, and number of samples that were included in each comparison. Table 3 shows up-regulated and down-regulated genes for each comparison and the percentage of differentially expressed genes (FDR, <0.05).

The FC values of tumor samples compared with normal tissue samples in microarray analysis were in excellent agreement with the FC values from qPCR results (CCC = 0.859; 95% confidence interval, 0.840-0.878). Acceptable agreement (CCC values of $\approx$0.6) was obtained between microarray FC values and qPCR FC values for each condition (Table 4).

Finally, an analysis of biologic processes using the DAVID/EASE system revealed that sequences were involved mainly in cellular processes (77.67%), primary metabolic processes (55.34%), and the regulation of biologic processes (36.89%). The cellular components that were identified were intracellular elements (70.87%), organelles (54.37%), and membrane-bound organelles (48.54%). Finally, differentially expressed genes related to molecular function were involved mainly in binding (81.55%), protein and nucleic binding (61.17% and 33.01%, respectively), and RNA binding (11.65%). The results are depicted in Figure 3.

DISCUSSION

Expression profiling of tumor tissue banks combined with long-term clinical follow-up is a current research line of considerable interest in oncology.[19] For the current study, we used fresh tissues obtained from breast cancer patients whose postoperative treatment planning was based on surgical results. The breast samples were subjected to rigorous clinical and histologic classification. Because the current study population comprised patients who had sporadic breast cancer and no direct family history of the disease, possible mutations of the breast cancer (BRCA) genes BRCA1 and BRCA2 were not analyzed. The hierarchical clustering that was used to identify significantly expressed gene sequences had good homogeneity among normal control samples that contrasted with wide variability among tumor samples. This variability most likely is because of the combination of tumor characteristics (tumor size, tumor grade, axillary lymph node involvement, ER status) that defines breast cancers.

Some studies[4,20] have demonstrated that malignant breast tumors can be classified and tumor subtypes can be defined by the analysis of gene expression patterns. The results obtained in the current study demonstrate that gene expression profiles differ significantly among the tumor phenotypes studied (Ph1, Ph2, and Ph3). The statistical analysis applied to select differentially expressed gene sequences also allowed us to identify expression profiles for distinct pathologic stages of breast cancer. Thus, significant differences in the gene expression profile were observed between ER-positive status versus ER-negative status, LC versus DC histologic subtype, and grades of tumor differentiation (Table 2).

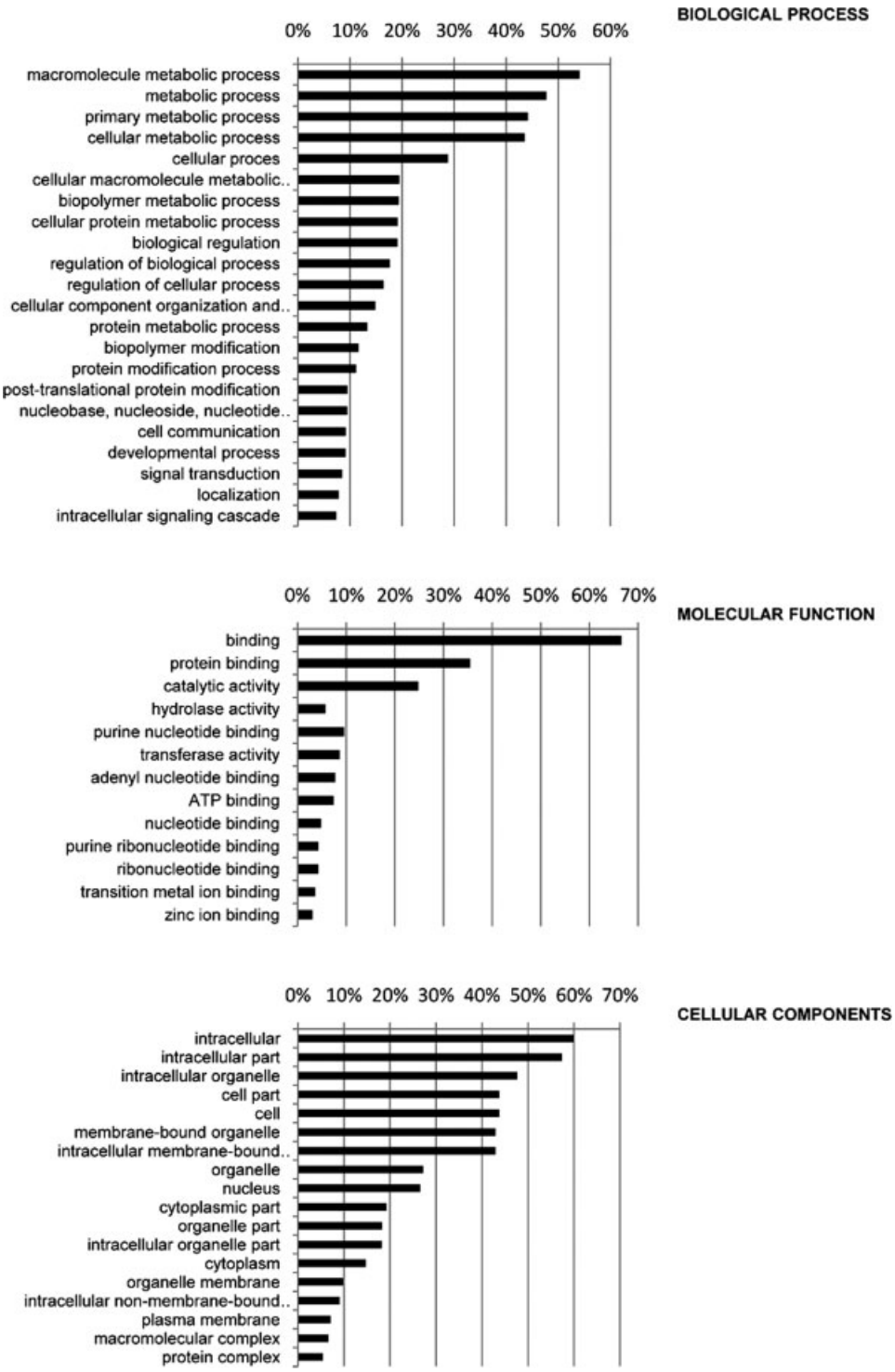

Figure 2. These charts illustrate Gene Ontology (GO) classification for the 8088 differentially expressed sequences from the tumor versus nontumor comparison. The percentage of coverage represents the percentage of genes annotated in GO and *Kyoto Encyclopedia of Genes and Genomes* terms. ATP indicates adenosine triphosphate.

Table 3. The Number of Sequences Validated by Quantitative Polymerase Chain Reaction Analysis for Each Comparison, Including the Number of Genes Confirmed as Regulated in the Same Direction (Up-Regulated or Down-Regulated) With a *P* Value <.05 and the Number of These Genes Above the False Discovery Rate Threshold of 0.05[a]

		No. of Genes Tested by qPCR		No. of Genes With Confirmed $P < .05$ (%)		No. of Genes Significant at $P < .05$ FDR (%)	
Sample	No. of Samples	UpR	DownR	UpR	DownR	UpR	DownR
All samples							
Tumor vs control	40 vs 14	39	64	24 (61.54)	51 (79.69)	24 (61.54)	49 (76.56)
Tumor phenotypes[b]							
Ph1 vs Ph2	10 vs 12	85	—	70 (82.35)	—	70 (82.35)	—
Ph1 vs Ph3	10 vs 18	2	—	2 (100)	—	—	—
All tumors							
ER+ vs ER−	24 vs 16	75	1	62 (82.67)	1 (100)	60 (80.00)	1 (100)
LC vs DC	12 vs 28	46	—	22 (47.83)	—	8 (17.39)	—
Grade 1/2 vs grade 3	16 vs 12	25	—	16 (64.00)	—	13 (52.00)	—
N0 tumors							
Grade 1/2 vs grade 3	9 vs 7	52	—	18 (33.96)	—	0 (0)	—

qPCR indicates real-time quantitative polymerase chain reaction; FDR, false discovery rate; UpR, up-regulated; DownR, down-regulated; Ph, phenotype; ER, estrogen receptor; +, positive; −, negative; LC, lobular carcinoma; DC, ductal carcinoma; N0, negative lymph node status.
[a]Additional Data File 4, S4 contains gene names and the fold change for each sequence shown in this table (available at: http://www.ugr.es/~efarez/).
[b]Ph1: tumors classified as N0, ER+, and T1, T2, or T3; Ph2: tumors classified as N0, ER−, and T1, T2, or T3; Ph3: tumors classified as N+, ER+ or ER−, and T1, T2, or T3.

Table 4. Concordance Correlation Coefficients Between Microarray Fold Changes and Real-Time Polymerase Chain Reaction Fold Changes for Each Compared Condition

Samples	No. of Genes Implicated	CCC	CCC 95% CI
All			
Tumor vs control	103	0.859	0.840-0.878
Tumor phenotypes[a]			
Ph1 vs Ph2	85	0.682	0.640-0.725
Ph1 vs Ph3	2	—	—
All tumors			
ER+ vs ER−	75	0.643	0.595-0.692
LC vs DC	46	0.506	0.431-0.582
Grade 1/2 vs grade 3	25	0.610	0.512-0.708
N0 tumors			
Grade 1/2 vs grade 3	52	0.607	0.539-0.675

CCC indicates concordance correlation coefficients; 95% CI, 95% confidence interval; Ph, phenotype; ER, estrogen receptor; +, positive; −, negative; LC, lobular carcinoma; DC, ductal carcinoma; N0, negative lymph node status.
[a]Ph1: tumors classified as N0, ER+, and T1, T2, or T3; Ph2: tumors classified as N0, ER−, and T1, T2, or T3; Ph3: tumors classified as N+, ER+ or ER−, and T1, T2, or T3.

The presence of metastatic tumor foci in axillary lymph nodes is most likely the best available marker for the individual risk of distant metastasis in breast cancer.[19,21] Nevertheless, approximately 25% of patients who have N0 status harbor micrometastases and are destined to develop recurrence, and up to 50% of patients with N+ disease do not develop disease recurrence after many years of follow-up, even without adjuvant chemotherapy.[22] Furthermore, some reports[23,6] have indicated that patients with negative lymph nodes but metastasizable primary tumors may be at greater risk than patients with positive lymph nodes. The possibility that a given gene expression profile might predict the metastatic potential of malignant tumors, even in the absence of positive axillary lymph nodes, is of considerable interest in breast cancer.

The genetic basis of the transition from N0 to N+ in our patients was analyzed by relating gene expression profiles to axillary lymph node status. Among N0 tumors, significant differences in gene expression profile were observed between grade 1/2 tumors and grade 3 tumors (Table 2).

Table 3 and Additional Data File 4 (available at: http://www.ugr.es/~efarez/) show the validated qPCR gene sequences from the microarray analysis comparisons. According to these data, differentially expressed gene sequences distinguished Ph1 from Ph2 and from Ph3, distinguished well or moderately differentiated tumors from poorly differentiated tumors, distinguished ER-positive tumors from ER-negative tumors, and distinguished between LC from DC histologic subtypes.

The large number of validated gene sequences that distinguished N0/ER-positive tumor phenotypes from

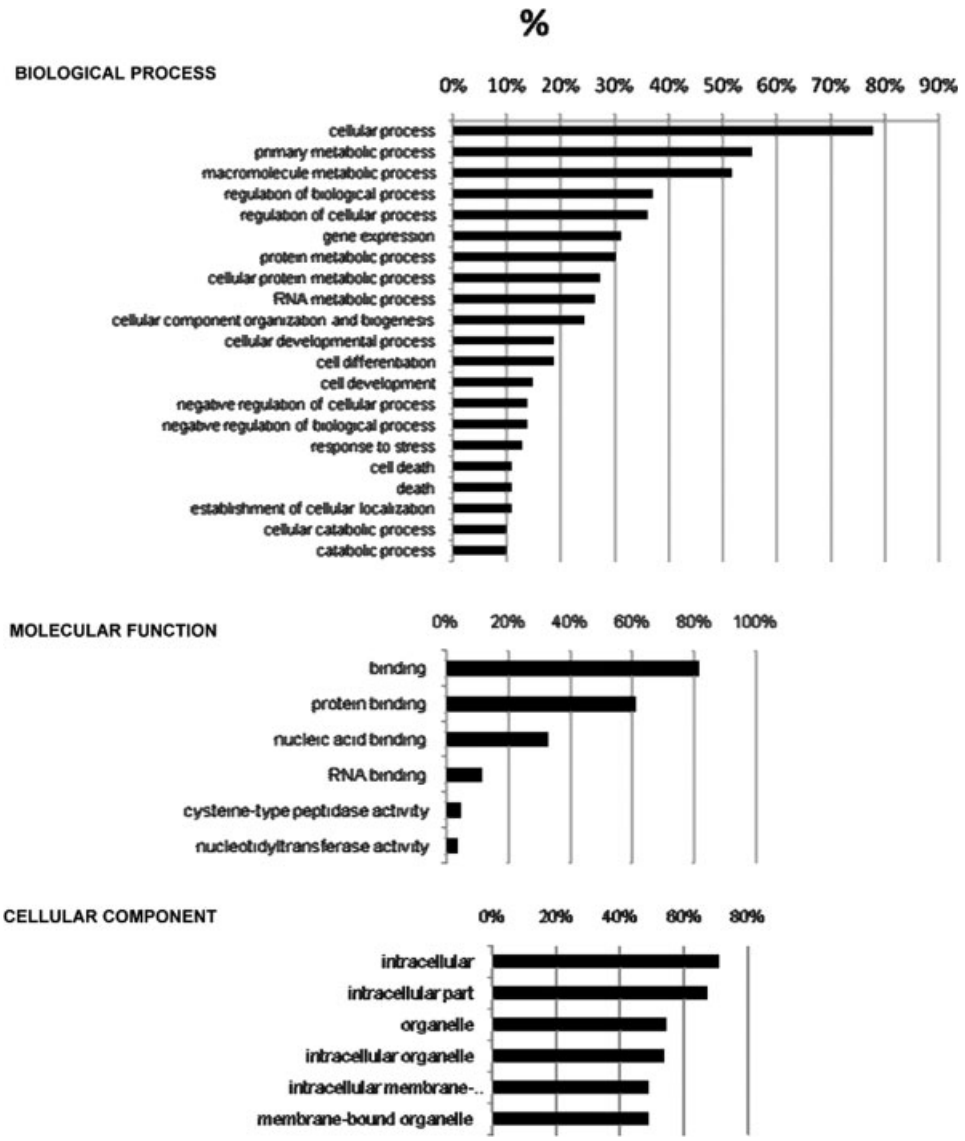

Figure 3. These charts illustrate the gene ontology (GO) classification of differentially expressed real-time quantitative polymerase chain reaction sequences. The percentage of coverage represents the percentage of genes annotated in GO and *Kyoto Encyclopedia of Genes and Genomes* terms.

N0/ER-negative tumor phenotypes suggest that ER status (a prognostic indicator and tumor response marker) is an important biologic parameter in breast cancer. Indeed, it is well documented that patients with ER-positive tumors have a longer disease-free interval and overall survival versus patients that lack ER expression.[24] Furthermore, Perou et al [4] reported that ER-positive tumors (defined as "luminal epithelial") and ER-negative tumors (defined as "basal epithelial") were associated with better and worse prognoses, respectively.

It has been established that invasive intraductal and lobular breast tumors have distinct histologies and clinical presentations. Lobular tumors often are ER-positive, tend to be more slowly proliferating,[25] and differ in DNA copy number changes from ductal cancers, suggesting a separate tumor progression pathway.[26,27] In the current study, cDNA microarrays were used to identify differential gene sequences in both tumor subtypes. Our findings indicate that 46 validated gene sequences were expressed differentially in lobular tumors versus ductal tumors in the global series (Table 3). These data are in agreement with those published by Korkola et al [28] and permit the genetic differentiation of these 2 histologic subtypes of breast cancer.

Differences in gene expression between high-grade tumors and low-grade tumors have been studied in relation to tumor progression,[29-31] and significant differences have been observed in gene expression profiles as a function of tumor differentiation grade. The N0 tumors in the current study revealed 52 up-regulated gene sequences in the comparison between grade 1 tumors and grade 2/3 tumors.

An acceptable agreement (CCC)[18] was observed between the microarray results and the qPCR results. There is a very small or no overlap between the list of validated genes identified in the current study and those reported previously.[4,5,9] Therefore, we believe that our analysis contributes new genetic information to the risk assessment of patients with breast cancer. Among the qPCR-validated genes, RAS-like estrogen-regulated growth inhibitor gene (*rerg*), solute carrier family 39 (zinc transporter) member 6 (*slc39a6*), and ENA/Vasp-like (*evl*) were the most highly overexpressed genes in N0, grade 1/2, Ph1 tumors; *rerg* and *slc39a6* were the most highly overexpressed genes in N0, grade 1/2, Ph1, ER-positive tumors; and signal transducer and activator of transcription 3 interacting protein 1 (*statip1*) was the most highly overexpressed gene (FC value, 5.28) in N0, grade 1/2 tumors. The genes par-6 partitioning defective 6 homolog β (*pard6b*), serum/glucocorticoid-regulated kinase family member 3 (*sgkl*), ER 1 α (*esr1*), and trefoil factor 1 (*tff1*) (FC values: 5.52, 6.63, 5.76, and 8.95, respectively) were the most highly overexpressed in Ph1 tumors.

slc39a6 belongs to a subfamily of proteins that have the structural characteristics of zinc transporters.[32] *pard6b*[33] most likely plays an important role in the cell polarization of mammalian cells by functioning as an adaptor protein that links the activated small GTPases *Rac* and *Cdc42* to atypical protein kinase C signaling. The signaling pathway that involves *sgkl*[34] plays an essential role in mammalian hair development. *tff1*, also known as *ps2*, is an estrogen (E2)-responsive gene that is methylated when actively transcribed and has been isolated from the human breast cancer cell line MCF-7.[35] Amplification of the *esr1* gene has been described in breast cancer.[36] *rerg*, also known as *xiap* (X-linked inhibitor of apoptosis), is a copper-binding protein.[37] *evl* is a cytoskeletal protein in the actin family that has functions in cell structure and motility. The *statip1* gene is involved in unclassified molecular functions and biologic processes.

Some years ago, investigators questioned[38] whether quantification of the level of dozens or hundreds of genes yielded more useful information regarding the metastatic potential of an individual cancer than an optimal analysis of standard histopathologic prognostic factors. If recommendations for potentially life-saving treatment are to be based on biologic parameters, then their assessment evidently must be reproducible and reliable.[39] This was 1 of the main objectives of the current study. To date, it has been demonstrated[5,40,41] that: 1) gene expression profiles are in good agreement with conventional predictive factors of a poor prognosis (ER-negative or ER-positive with high-grade tumors), 2) gene expression profiles in patients who have N0/ER-positive breast cancer can distinguish among levels of recurrence risk (high risk, intermediate risk, or low risk), and 3) gene expression profiles clearly distinguish between DC and LC and can be used to follow disease progression from benign to invasive malignant tumors. In our view, the differential gene expression profiles observed in our study offer new insights into the molecular basis of breast cancer, especially the data obtained in comparisons of ER-positive versus ER-negative, DC versus LC, and grade 1/2 versus grade 3 and in the genetic differences underlying the tumor phenotypes that we studied.

CONFLICT OF INTEREST DISCLOSURES

Supported by grant PI021806 from the Carlos III Institute, Madrid. Dr. Farez-Vidal received a Spanish Fondo de Investigación Sanitaria grant from the Carlos III Institute.

REFERENCES

1. Hanahan D, Weinberg R. The hallmarks of cancer. *Cell*. 2000;100:57-70.
2. Farabegoli F, Champeme M, Bieche I, et al. Genetic pathways in the evolution of breast ductal carcinoma in situ. *J Pathol*. 2002;196:280-286.
3. Reis-Filho JS, Lakhani S. The diagnosis and management of preinvasive breast disease: genetic alterations in preinvasive lesions. *Breast Cancer Res*. 2003;5:313-319.
4. Perou CM, Sorlie T, Eisen MB, et al. Molecular portraits of human breast tumours. *Nature*. 2000;406:747-752.
5. van de Vijver MJ, He Y, van't Veer LJ, et al. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med*. 2002;25:1999-2009.
6. Wang Y, Klijn J, Zhang Y, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*. 2005;365:671-679.
7. Ein-Dor L, Kela I, Getz G, Givol D, Domany E. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*. 2005;21:271-278.
8. Sorlie T, Perou C, Tibshirani R, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A*. 2001;98:10869-10874.

9. van't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*. 2002;415:530-536.
10. Simpson PT, Reis-Filho J, Gale Y, Lakhani SR. Molecular evolution of breast cancer. *J Pathol*. 2005;205:248-254.
11. Jenssen TK, Hovig E. Gene-expression profiling in breast cancer. *Lancet*. 2005;365:634-635.
12. Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*. 2003;4:249-264.
13. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*. 2001;98:5116-5121.
14. Chang HY, Nuyten D, Sneddon JB, et al. Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *Proc Natl Acad Sci U S A*. 2005;102:3738-3743.
15. Vandesompele J, De Preter K, Pattyn F, et al. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol*. 2002;3:34.31-34.11.
16. Yuan JS, Reed A, Chen F, Stewart CN. Statistical analysis of real-time PCR data. *BMC Bioinformatics*. 2006;7:85-97.
17. Benjamini Y, Hochberg Y. Controlling the false discovery rate. A practical and powerful approach to multiple testing. *J R Stat Soc B*. 1995;57:284-300.
18. Carrasco JL, Jover L. Estimating the generalized concordance correlation coefficient through variance components. *Biometrics*. 2003;59:849-858.
19. Murphy N, Millar E, Lee CS. Gene expression profiling in breast cancer: toward individualising patient management. *Pathology*. 2005;37:271-277.
20. Hedenfalk I, Duggan D, Chen Y, et al. Gene-expression profiles in hereditary breast cancer. *N Engl J Med*. 2001;344:539-548.
21. [No authors listed] Primary Therapy of Early Breast Cancer, 9th International Conference. January 26-29, 2005. St. Gallen, Switzerland. Abstracts. *Breast*. 2005;14(suppl 1):S1-S56.
22. Early Breast Cancer Trialists Collaborative Group. Tamoxifen for early breast cancer: an overview of the randomized trials. *Lancet*. 1988;351:1451-1467.
23. West M, Blanchette C, Dressman H, et al. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc Natl Acad Sci U S A*. 2001;98:11462-11467.
24. Carlson RW, Brown E, Burstein HJ, et al. NCCN Task Force Report: adjuvant therapy for breast cancer. *J Natl Comp Canc Netw*. 2006;4(suppl 1):S1-S26.
25. Coradini D, Pellizaro C, Veneroni S, Ventura L, Daidone MG. Infiltrating ductal and lobular breast carcinomas are characterised by different interrelationships among markers related to angiogenesis and hormone dependence. *Br J Cancer*. 2002;87:1105-1111.
26. Nishizaki T, Chew K, Chu L, et al. Genetic alterations in lobular breast cancer by comparative genomic hybridization. *Int J Cancer*. 1997;74:513-517.
27. Gunther K, Merkelbach-Bruse S, Amo-Takyi BK, Handt S, Schroder W, Tietze L. Differences in genetic alterations between primary lobular and ductal breast cancers detected by comparative genomic hybridization. *J Pathol*. 2001;193:40-47.
28. Korkola JE, DeVries S, Fridlyand J, et al. Differentiation of lobular versus ductal breast carcinomas by expression microarray analysis. *Cancer Res*. 2003;63:7167-7175.
29. Porter DA, Krop I, Nasser S, et al. A SAGE (Serial Analysis of Gene Expression) view of breast tumor progression. *Cancer Res*. 2001;61:5697-5702.
30. Warnberg F, Nordgren H, Bergkvist L, Holmberg L. Tumour markers in breast carcinoma correlate with grade rather than with invasiveness. *Br J Cancer*. 2001;85:869-874.
31. Ma XJ, Salunga R, Tuggle JT, et al. Gene expression profiles of human breast cancer progression. *Proc Natl Acad Sci U S A*. 2003;100:5974-5979.
32. Taylor KM, Nicholson RI. The LZT proteins; the LIV-1 subfamily of zinc transporters. *Biochim Biophys Acta*. 2003;1611:16-30.
33. Noda Y, Takeya R, Ohno S, Naito S, Ito T, Sumimoto H. Human homologues of the *Caenorhabditis elegans* cell polarity protein PAR6 as an adaptor that links the small GTPases Rac and Cdc42 to atypical protein kinase C. *Genes Cells*. 2001;6:107-119.
34. Masujin K, Okada T, Tsuji T, et al. Mutation in the serum and glucocorticoid-inducible kinase-like kinase (Sgkl) gene is associated with defective hair growth in mice. *DNA Res*. 2004;11:371-379.
35. Kangaspeska S, Stride B, Metivier R, et al. Transient cyclical methylation of promoter DNA. *Nature*. 2008;452:112-115.
36. Brown LA, Hoog J, Chin SF, et al. ESR1 gene amplification in breast cancer: a common phenomenon [letter]?*Nat Genet*. 2008;40:806-807.
37. Mufti AR, Burstein E, Csomos RA, et al. XIAP is a copper binding protein deregulated in Wilson's disease and other copper toxicosis disorders. *Mol Cell*. 2006;21:775-785.
38. Heimann R, Hellman S. Clinical progression of breast cancer malignant behavior: what to expect and when to expect it. *J Clin Oncol*. 2000;18:591-599.
39. O'Shaughnessy JA. Molecular signatures predict outcomes of breast cancer. *N Engl J Med*. 2006;355:615-617.
40. Paik S, Shak S, Tang G, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med*. 2004;351:2817-2826.
41. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. 2002;30:207-210.