*Int. J. Cancer:* **123,** 1327–1338 (2008)


# Identification of new genes associated with breast cancer progression by gene expression analysis of predefined sets of neoplastic tissues

**Daniela Cimino[1]*, Luca Fuso[1,2], Christian Sfiligoi[1], Nicoletta Biglia[1,2], Riccardo Ponzone[1,2], Furio Maggiorotto[1,2], Giandomenico Russo[3], Luigi Cicatiello[4], Alessandro Weisz[4,5], Daniela Taverna[1,6,7], Piero Sismondi[1,2] and Michele De Bortoli[6,7]**

[1]*Institute for Cancer Research and Treatment, Unit of Gynecological Oncology, University of Turin Medical School, Candiolo, Turin, Italy*
[2]*Department of Gynecology and Obstetrics, University of Turin, Turin, Italy*
[3]*Istituto Dermopatico dell'Immacolata-Istituto di Ricovero e Cura a Carattere Scientifico, Rome, Italy*
[4]*Dipartimento di Patologia Generale, Seconda Università Degli Studi di Napoli, Naples, Italy*
[5]*AIRC Naples Oncogenomic Center, Naples, Italy*
[6]*Department of Oncological Sciences, University of Turin, Turin, Italy*
[7]*Center for Complex Systems in Molecular Biology and Medicine, University of Turin, Turin, Italy*

**Gene expression profiles were studied by microarray analysis in 2 sets of archival breast cancer tissues from patients with distinct clinical outcome. Seventy-seven differentially expressed genes were identified when comparing 30 cases with relapse and 30 cases without relapse within 72 months from surgery. These genes had a specific ontological distribution and some of them have been linked to breast cancer in previous studies: *AIB1*, the two keratin genes *KRT5* and *KRT15*, *RAF1*, *WIF1* and *MSH6*. Seven out of 77 differentially expressed genes were selected and analyzed by qRT-PCR in 127 cases of breast cancer. The expression levels of 6 upregulated genes (*CKMT1B*, *DDX21*, *PRKDC*, *PTPN1*, *SLPI*, *YWHAE*) showed a significant association to both disease-free and overall survival. Multivariate analysis using the significant factors (*i.e.*, estrogen receptor and lymph node status) as covariates confirmed the association with survival. There was no correlation between the expression level of these genes and other clinical parameters. In contrast, *SERPINA3*, the only downregulated gene examined, was not associated with survival, but correlated with steroid receptor status. An indirect validation of our genes was provided by calculating their association with survival in 3 publicly available microarray datasets. *CKMT1B* expression was an independent prognostic marker in all 3 datasets, whereas other genes confirmed their association with disease-free survival in at least 1 dataset. This work provides a novel set of genes that could be used as independent prognostic markers and potential drug targets for breast cancer.**
© *2008 Wiley-Liss, Inc.*

**Key words:** breast cancer; gene expression; microarray analysis; prognostic

The heterogeneous nature of breast cancer reflects the complexity of the molecular alterations that underlie the development and progression of this disease and poses serious problems to clinical management, also due to the lack of reliable pathological or molecular markers.

There are a number of major open questions, such as the evaluation of risk of distant metastasis in cases characterized by the presence of favorable indicators (negative axillary lymph nodes and/or positive estrogen receptors), the prediction of response to chemotherapy and/or antiestrogenic therapy and the prediction of metastases sites for high-risk cancers. Moreover, the principal molecular alterations leading to aggressive clinical behavior, representing potential therapeutic targets, still need to be identified in addition to the well-known factor *ERBB2* and the estrogen receptor pathways.

Molecular profiling using DNA microarrays have provided sound advancements in this field. The expression profile of gene clusters were useful in classifying breast tumors in biological subgroups with clinical relevance,[1] or in low- versus high-risk classes for relapse,[2–5] or to predict responsiveness to either hormonal- or chemo-therapy.[6,7] In a well-known study, van't Veer *et al.* addressed the risk of relapse in node-negative patients,[2] one of the most clinically relevant problems in breast cancer. One-third of these cases will progress and have poor outcome, but they cannot be distinguished on the basis of the most classical clinicopathological criteria. Microarray analysis of archival frozen tissues provided a 70-gene profile able to identify relapsing cases, outperforming traditional clinical prognostic factors. This signature maintained its prognostic power in larger cohorts of patients, including node-positive cases as well.[3] van de Vijver *et al.* dataset[3] is today largely used for data analysis and a published meta-analysis reported that other gene signatures, such as the wound-response model,[8] the intrinsic subtype model[1,9,10] and the rationale recurrence score model proposed by Paik *et al.*,[11] worked acceptably well in estimating the risk of relapse in this dataset.[12]

The first remarkable conclusion from these and other studies is that the primary tumor already possesses the hardwiring required to invade and metastasize.[13] The second, unexpected, conclusion is that the "signatures" obtained by other studies have impressive little overlap, still showing good power in other's datasets. Sets of genes identified in hierarchical cluster analysis showing the highest grade of internal coherence within the experiment are not necessary correlated in terms of biological significance. It has been recently demonstrated that the expression profile of any randomly selected set of genes with a sufficient width (around 100) can correctly identify the kind of human tissue or organ from which the RNA was extracted.[14]

Another unexpected conclusion is that these prognostic signatures very rarely include known progression-associated genes or clearly point to novel pathways or molecules associated with cancer progression. Therefore their exceptional prognostic and predictive value does not parallel an equal power to identify new potential targets for therapy and drug development.

Taking into account these results, we undertook a different approach, by directly comparing gene expression of high-risk *ver-*

---

This article contains supplementary material available via the Internet at http://www.interscience.wiley.com/jpages/0020-7136/suppmat

Christian Sfiligoi's current address is: Laboratorio Ematologia I, Az. Ospedaliera S. Giovanni Battista, Turin, Italy.

Daniela Cimino and Daniela Taverna's current address is: Molecular Biotechnology Center, Via Nizza 52 Turin, Italy.

Grant sponsor: Ministero della Salute; Grant number: RF.PE.2005. 147663; Grant sponsor: ABO Project; Grant number: TO47; Grant sponsor: MIUR; Grant number: PRIN 2006069030 003; Grant sponsor: UE (CRESCENDO IP); Grant number: LSHM-CT2005-018652; Grant sponsors: Regione Piemonte Ricerca Sanitaria; Ricerca Scientifica Applicata; AIRC (Italian Association for Cancer Research).

***Correspondence to:** Department of Oncological Sciences, University of Turin, Molecular Biotechnology Center, Via Nizza 52, Turin, Italy.
E-mail: daniela.cimino@unito.it

Received 27 November 2007; Accepted after revision 19 March 2008
DOI 10.1002/ijc.23660
Published online 17 June 2008 in Wiley InterScience (www.interscience.wiley.com).

*sus* low-risk patients from a cohort of archival breast cancer tissues, by microarray analysis of pooled samples, in order to minimize small and infrequent variations. These differentially expressed genes were found to associate individually with risk, when analyzed using quantitative fluorogenic RT-PCR (qRT-PCR) in a larger cohort, indicating new genes and pathways potentially associated with breast cancer progression.

## Material and methods

### *Patients and samples*

130 frozen tumor samples were selected from the Tumor Bank of the Department of Obstetrics and Gynecology, University of Turin. They were obtained from patients who underwent primary surgical treatment between 1988 and 2001 at a median age of 53 years (25–79). Eligibility criteria were the following: diagnosis of invasive breast cancer, all T and N stages, no distant metastasis at diagnosis (M0), complete clinical-pathological data and updated follow up. All patients were treated with radical modified mastectomy or quadrantectomy and axillary dissection plus breast irradiation. High-risk node-negative and node-positive patients received adjuvant treatments (generally 6 cycles of CMF, 600 mg/m$^2$ cyclophosphamide, 40 mg/m$^2$ Metotrexate, 600 mg/m$^2$ 5-Fluorouracil) and/or 20 mg tamoxifen daily for 5 years in ER+ cases. ER and PgR status were determined by immunohistochemical staining, patient stage distribution was assessed as prescribed by the UICC clinical staging guidelines and tumor grading was performed according to Elston and Ellis. Study design was approved by our medical ethical committee.

**TABLE I –** CLINICAL CHARACTERISTICS OF PATIENTS

| Characteristics | All patients | | Array set | |
|---|---|---|---|---|
| | No. | % | No. | % |
| No. of patients | 127 | | 60 | |
| Age (years) | | | | |
| Median | 54 | | 53 | |
| Menopausal status | | | | |
| Premenopausal | 53 | 42 | 30 | 50 |
| Postmenopausal | 74 | 58 | 30 | 50 |
| T stage | | | | |
| 1 | 35 | 28 | 13 | 22 |
| 2 | 80 | 63 | 42 | 70 |
| 3/4 | 12 | 9 | 5 | 8 |
| Grade | | | | |
| Poor | 58 | 45 | 32 | 53 |
| Good to moderate | 63 | 50 | 26 | 43 |
| Unknown | 6 | 5 | 2 | 3 |
| ER[1] | | | | |
| Positive | 78 | 61 | 40 | 67 |
| Negative | 49 | 39 | 20 | 33 |
| PgR[1] | | | | |
| Positive | 66 | 52 | 32 | 53 |
| Negative | 61 | 48 | 28 | 47 |
| LN | | | | |
| Positive | 86 | 68 | 41 | 68 |
| Negative | 41 | 32 | 19 | 32 |

ER, estrogen receptor; PgR, progesterone receptor; LN, lymph node status.
[1]ER and PgR are defined as positive when tumors contain more than 10 fmol/mg protein or less than 10% positive tumor cells.

### *RNA isolation*

After surgical removal, samples were macro-dissected by pathologists, quickly frozen and stored at −80°C. RNA was isolated with Concert Cytoplasmic RNA Reagent (Invitrogen, Carlsbad, CA) from 20 to 50 mg tumor tissues, according to the manufacturer's guidelines. Frozen tumors were placed in this reagent and homogenized using a ball mill (MM200, Retsch, Düsseldorf, Germany). The suspension was centrifuged at 14,000*g* for 5 min at 4°C, then lysed with 0.1 ml of 10% SDS followed by 0.3 ml of 5 M sodium chloride and 0.2 ml of chloroform per ml of reagent. The lysate was centrifuged at 14,000*g* for 15 min at 4°C and the upper aqueous phase was removed and combined with 0.8 vol of isopropyl alcohol for 10 min at room temperature. The RNA was recovered by centrifugation, washed with 75% ethanol and finally dissolved in RNase-free water. Ten-microgram aliquots of total RNA was treated with DNase I, using the "DNA free" kit (Ambion, Austin, TX) to eliminate genomic DNA contamination. The quantity and quality of the RNA samples were determined using the Agilent 2100 Bioanalyzer and the RNA 6000 Nano Assay kit (Agilent Technologies, Palo Alto, CA). Only high-quality RNA, having a 28S/18S rRNA band intensity ratio of 1.5–2 and an A260/280 absorbance ratio of 1.8–2, was used for subsequent analysis. RNA of optimal quality and quantity was recovered from 127 samples. The clinical characteristics of the respective patient set are summarized in Table I.

### *Microarray analysis of pooled samples*

Thirty cases of ductal invasive carcinoma with recurrence within 72 months from surgery (high-risk) and 30 without recurrence (low-risk) were selected for microarray analysis. From each group, 6 RNA pools of 5 samples each were prepared and analyzed on GeneChip Human Genome U133A oligonucleotide microarrays (HG-U133A, Affymetrix, Santa Clara, CA). Detailed information about the clinical characteristics of samples in the composed pools, including ER or lymph node status, are provided in the Supplementary Information, Table SI. Labeling was performed using 8 μg of total RNA with the One-Cycle Target Labeling Assay kit (Affymetrix), according to manufacturer's instructions. Double-stranded cDNA was purified and biotin-labeled by *in vitro* transcription. The biotinylated cRNA targets were then purified and fragmented to a length of 35–200 bases. The quality of *in vitro* transcription and fragmentation products was assessed using the Agilent 2100 Bioanalyzer. Hybridization to HG-U133A GeneChips (Affymetrix) and arrays scanning was carried out according to Affymetrix protocols. Hybridizations were performed in technical duplicates in 2 experimental sessions.

### *Microarray data analysis*

Data analysis was performed using the R statistical package (http://www.bioconductor.org). Array quality control was carried out using the affy library and the affyPLM package of R. Expression values were calculated from the raw.CEL files as GC Robust Multichip Analysis (gcRMA). Filtering procedure was done employing IQR filtering function using as cut an interquantile range within the various samples lower than 0.25. The first selection step of differentially expressed genes was done using a Bayesian *t*-test, implemented for DNA microarray data with a low replicate number.[15] The second step was made by the Significance Analysis of Microarrays program[16] and a two-dimensional unsupervised hierarchical clustering based on a centered Pearson correlation coefficient algorithm (TIGR MeV, www.tigr.org). The fold change threshold for SAM plot calculator was 2 and the median false discovery rate was lower than 5.0%. Statistical analyses of gene ontology (GO) terms was performed using the web-based tool DAVID Resource (http://david.abcc.ncifcrf.gov/); this tool provides GO terms and their significant probabilities of enrichment (*p*-values from Fisher Exact test) compared to the reference gene list (HG-U133A GeneChip Affymetrix list). Raw data and GCRMA intensity values are available in GEO (http://www.ncbi.nlm.nih.gov/geo/, accession number GSE9662).

Comparison among discrimination power of different gene lists[4,17–20] was performed using support vector machine (SVM) and partial decision trees, used to train and classify pooled samples (Weka software Version 3.4.8[21]).

### *Quantitative real time RT-PCR assays*

The RNA expression levels of individual genes (*CKMT1B*, *DDX21*, *PRKDC*, *PTPN1*, *SERPINA3*, *SLPI* and *YWHAE*) were

assayed using qRT-PCR with TaqMan® gene expression assays (Applied Biosystems, Foster City, CA) on the total set of 127 samples, including the 60 samples used for pooled microarray analysis. One microgram of total RNA was retrotranscribed in a 20 µl final reaction volume, using random decamer primers and the M-MLV Reverse Transcriptase (Ambion). Reaction conditions were recommended by the manufacturer. The amount of cDNA corresponding to 10 ng of RNA was used in 10 µl reactions with the TaqMan Universal PCR Master Mix (Applied Biosystems) and the corresponding sequence-specific primers/probes assay mix (Applied Biosystems). Six independent cDNA syntheses for each sample were made and each of them PCR amplified in single; analysis was performed on average values after filtering for outliers. Fluorescence detection was measured using an ABI Prism 7900 platform (Applied Biosystems) on 384-well plates. As reference sample we used human breast total RNA (Stratagene, La Jolla, CA) and, after the assessment of several constitutively expressed genes (Table SII), we chose the 18s ribosomal RNA as endogenous normalizer (Eukaryotic 18S rRNA Endogenous Control, VIC/TAMRA Probe, Primer Limited, Applied Biosystems). TaqMan gene expression assays are listed in the Supplementary Information, Table SII. Fold difference between samples was calculated for each gene by means of the Comparative CT Method ($\Delta\Delta$Ct), using the median normalized value as calibrator.

### *Statistics*

Statistical analysis was performed using the SPSS 13.0 statistical software (SPSS, Chicago, IL). The rank nonparametric statistical tests of Mann–Whitney and Kruskal–Wallis were used to examine associations between gene expression and clinicopathological data because no evidence of normal distribution was available (Kolmogorov–Smirnov $p < 0.0001$).

To compare mRNA expression level distribution among high-risk and low-risk samples in qRT-PCR, the Mann–Whitney nonparametric statistical test was used. Kaplan–Meier survival curves were used to estimate time-to-event models in the presence of censored cases. Risk differences between the 2 groups were assessed using the Mantel–Haenszel Log-rank test. Survival analysis was carried out in both univariate and multivariate setting using Cox's proportional hazard model. Variables that were significant at univariate level ($p < 0.05$) were considered to build multivariate model.

### *Analysis of published microarray datasets*

Genes analyzed by qRT-PCR were evaluated in 3 independent datasets available on line. Miller *et al.* dataset[18] and Sotiriou *et al.* dataset[20] were downloaded from Gene Expression Omnibus database (http://www.ncbi.nlm.nih.gov/geo/; accession numbers GSE3494 and GSE2990). For these datasets expression values were calculated from the raw.CEL files with gcRMA bioconductor function and intensities were scaled on median values for each gene; results were divided in 2 categories according to a cut-off value obtained from the Sensibility-Specificity ROC curve. van de Vijver *et al.* dataset[3] was downloaded from http://www.rii.com/publications/2002/nejm.html and, for each analyzed gene, the values of the column entitled "Log Ratio" in the array data file were divided in 2 categories according to a cut-off value obtained from the sensibility–specificity ROC curve.

## Results

### *Identification of genes potentially associated with breast cancer relapse*

Genes and pathways responsible for a more invasive and aggressive phenotype should be consistently activated in breast tumors which relapse within a few years from surgical treatment.

As a first step, we wanted to identify a group of genes showing consistently altered expression in tumors from patients relapsing within 72 months, as compared to tumors from relapse-free patients. With this aim we performed a gene expression analysis with a subpooling approach. We reasoned that genes showing a very wide range of expression or genes that are overexpressed only in rare cases should be "masked" by combining the samples in pools. The size of the pools was fixed in 5 because larger pool size will possibly completely smooth the differences among groups. This choice was also supported by literature evidence. By recalculating data from microarray analysis, it was shown that pooling samples by 5 gave similar sensitivity with the analysis of single samples.[22]

From a cohort of 127 cases of primary breast carcinomas with a median follow-up of 87 months, 2 groups of 30 samples each, showing or not disease recurrence within 72 months (high-risk and low-risk), with similar distribution of relevant prognostic factors, were selected (Supplementary Information, Table SI). As lobular carcinomas show a very distinctive expression profile[1] that could be prevalent and confounding, only carcinomas of the ductal histological type were included. Pools of 5 samples each were used for gene expression analysis by Affymetrix HG-U133A GeneChips. Detailed information about the clinical characteristics of samples in the composed pools, including ER or lymph node status, are provided in the Supplementary Information, Table SI. Differential expression was assessed by comparing pools from high-risk *versus* low-risk patients. Expression data from 2 analytical sessions were treated separately using a modified *t* test[15] and 2 gene sets were identified with a *p*-value cut-off of 0.05. A list of genes common to these 2 sets was then obtained and analyzed by applying a two-dimensional unsupervised hierarchical clustering and the Significance Analysis of Microarrays program,[16] 80 Affymetrix probe-sets representing 77 genes, whose expression patterns best distinguished high-risk from low-risk pools, with a false discovery rate <5% and a fold change threshold of 2.0, were identified. Unsupervised hierarchical clustering is shown in Figure 1 and the complete probe list is shown in Figure 2, together with fold changes and Affymetrix IDs.

As expected, in this set only 6 downregulated genes in high-risk pools were present. More advanced tumors showed ranges of gene expression wider than less advanced tumors or normal tissues. The comparison of arithmetical averages, intrinsic to the pooling approach, is expected to enhance detection of upregulated genes.

We looked for evidence of association with breast cancer in the literature. As shown in Figure 2 (column 5), 21 genes had some reported link to breast cancer, in agreement with their differential expression in high-risk *versus* low-risk tumors.[23–43] Some of them had also previously demonstrated to be independent prognostic factors, *i.e.*, *RAF1*,[24] *KRT15*,[25] *WIF1*,[26] *KRT5*,[40] *MSH6*[31] and *NCOA3/AIB-1*.[42] A GO analysis demonstrated significant probability of enrichment in biosynthetic and metabolic processes with an overrepresentation of amine metabolism classes, while pathway analysis showed significant enrichment in the urea cycle and in the amino group or proline metabolisms, in agreement with class representation (Table II). A borderline level of significance was also attained by cell cycle and insulin signaling pathways.

The function of some of these genes could be found to play a part in novel pathways involved in cancer. Several genes are in fact related to Acetyl-CoA and NAD+ metabolism (*ACLY*, *ACACA*, *KYNU*) and others to methyl group metabolism (*AHCY*), including the *AOF2* gene, also known as *LSD1*, the first discovered histone demethylase enzyme, whose function is strictly linked to chromatin remodeling as well as steroid receptor function.[44]

As recently reported by other authors, there is a very low concordance among gene-expression-based predictors for breast cancer.[12] Therefore we examined the presence of our genes in several published signatures, including the 231 prognostic reporter genes of van de Vijver *et al.*[2,3] the wound response signature of 677 genes,[45] the Oncotype DX assay list of 21 genes,[11] the intrinsic subtype gene group of 552 genes[1,9,10] and 2 signatures related to metastases sites, obtained from tumor cell lines with increased metastatic ability.[17,19] Twenty-six/77 genes were present in at

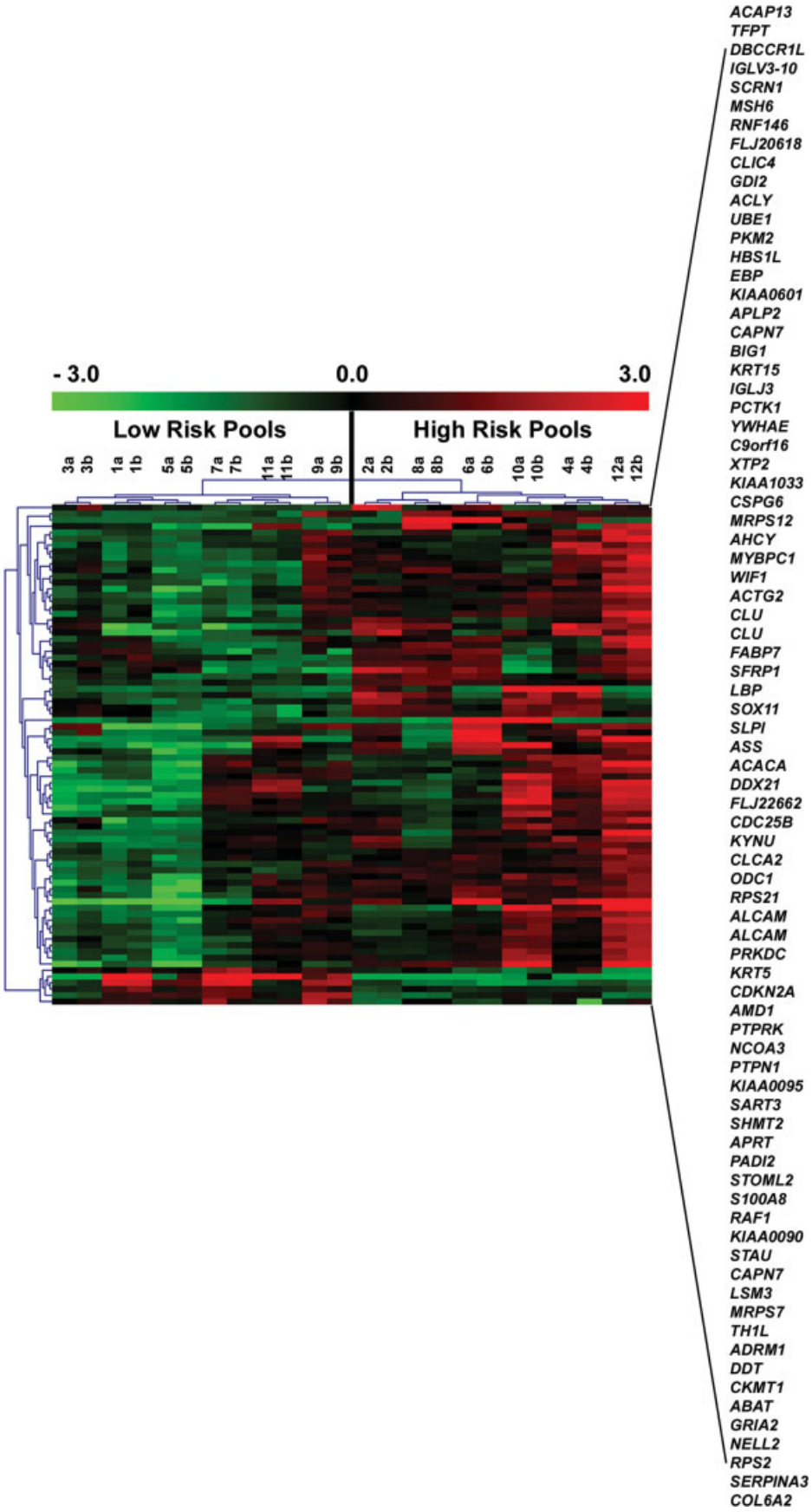


**FIGURE 1** – Expression pattern of the 80-gene signature obtained by unsupervised two-dimensional hierarchical clustering using TIGR MeV software and Pearson correlation. The data are shown in a table format in which each row represents a gene and each column corresponds to a pooled sample. The color in each cell reflects the expression level of the corresponding gene in the corresponding pooled sample; red color indicates a transcription level above the median expression of genes across all samples, green color indicates a transcription level below the median expression. Pool IDs are those provided in Table SI and each pool is represented by the 2 experimental replicates (a) and (b).

| | | | | | Breast Cancer related signature overlap | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| modulation | AFFYIDs | Gene Symbol | Fold Changes | Breast Cancer related | Sorlie T et al. (9) [552 genes] | van 't Veer LJ et al. (2) [231 genes] | Chang HY et al. (8) [677 genes] | Minn AJ et al. (19) [65 probes] | Kang Y et al. (17) [127 probes] | West M et al. (46) [100 genes] | Sotiriou C et al. (20) [97 genes] | Sotiriou C et al. (5) [485 probes] |
| up | 205030_at | FABP7 | 193.86 | | ■ | | | | | | | ■ |
| | 214087_s_at | MYBPC1 | 26.81 | | | | | | | | | |
| | 217562_at | FAM5C | 9.29 | | | | | | | | | |
| | 202917_s_at | S100A8 | 7.37 | 23 | | | | | | | | |
| | 202712_s_at | CKMT1B | 6.98 | | | | | | | | | |
| | 203021_at | SLPI | 4.13 | | ■ | | ■ | | | ■ | | |
| | 216853_x_at | IGLJ3 | 4.06 | | | | | | | | | |
| | 201244_s_at | RAF1 | 4.00 | 24 | | | | | | | | ■ |
| | 204734_at | KRT15 | 3.96 | 25 | | | | | | | | |
| | 209257_s_at | CSPG6 | 3.78 | | | | | | | | | |
| | 204712_at | WIF1 | 3.61 | 26 | | | | | | | | |
| | 214461_at | LBP | 3.44 | | | | | | | | | |
| | 207039_at | CDKN2A | 3.37 | 27 | | | | | | | | ■ |
| | 203357_s_at | CAPN7 | 3.25 | | | | | | | | | |
| | 209316_s_at | HBS1L | 3.22 | | | | | | | | | |
| | 204914_s_at | SOX11 | 3.22 | | | | | | | | | |
| | 214097_at | RPS21 | 3.20 | | | | | | | | | |
| | 202274_at | ACTG2 | 3.07 | | ■ | | | | | | | |
| | 216560_x_at | IGLV3-10 | 3.00 | | | | | | | | | |
| | 202929_s_at | DDT | 2.89 | | | | | | | | | |
| | 208152_s_at | DDX21 | 2.80 | | | | | | | | | |
| | 207076_s_at | ASS1 | 2.80 | | | | | | | ■ | | |
| | 208792_s_at | CLU | 2.69 | 28 | | | | | ■ | | | |
| | 208791_at | CLU | 2.69 | 28 | | | | | ■ | | | |
| | 200790_at | ODC1 | 2.58 | 29 | | | | | | | | ■ |
| | 201196_s_at | AMD1 | 2.52 | | | | | | | | | |
| | 210317_s_at | YWHAE | 2.50 | | | | | | | | | |
| | 204331_s_at | MRPS12 | 2.44 | | | | | | | | | ■ |
| | 201952_at | ALCAM | 2.44 | 30 | ■ | | | | | | | |
| | 221430_s_at | RNF146 | 2.43 | | | | ■ | | | | | |
| | 211450_s_at | MSH6 | 2.41 | 31 | | | | | | | | |
| | 208824_x_at | PCTK1 | 2.38 | | | ■ | | | | | | ■ |
| | 209791_at | PADI2 | 2.38 | | | | | | | | | |
| | 203219_s_at | APRT | 2.36 | | | | | | | | | |
| | 215416_s_at | STOML2 | 2.34 | | | | | | | | | |
| | 209535_s_at | AKAP13 | 2.33 | 32 | | | | | | | | |
| | 201281_at | ADRM1 | 2.32 | 33 | ■ | | | | | | | |
| | 200903_s_at | AHCY | 2.32 | | | | | | | | | |
| | 212337_at | TI-227H | 2.31 | | | | | | | | | |
| | 216266_s_at | ARFGEF1 | 2.27 | | | | | | | | | |
| | 209127_s_at | SART3 | 2.27 | 34 | | | | | | | | |
| | 202209_at | LSM3 | 2.24 | | | | ■ | | | | | |
| | 214437_s_at | SHMT2 | 2.24 | | | | | | | | ■ | ■ |
| | 203038_at | PTPRK | 2.24 | | ■ | | | | | | | |
| | 213787_s_at | EBP | 2.24 | | | | | | | | | |
| | 201853_s_at | CDC25B | 2.23 | 35 | | ■ | | | | | | ■ |
| | 200964_at | UBE1 | 2.22 | | ■ | | | | | | | |
| | 217388_s_at | KYNU | 2.21 | | | | | ■ | | | | |
| | 212396_s_at | KIAA0090 | 2.21 | | | | | | | | | |
| | 202188_at | NUP93 | 2.20 | | | | | | | | ■ | |
| | 201559_s_at | CLIC4 | 2.20 | 36 | | | | | | | | |
| | 201951_at | ALCAM | 2.19 | 30 | ■ | | | | | | | |
| | 200008_s_at | GDI2 | 2.18 | | | | | | | | | |
| | 201251_at | PKM2 | 2.17 | 37 | | | | | | | | |
| | 220607_x_at | TH1L | 2.17 | | | | | | | | | |
| | 217932_at | MRPS7 | 2.16 | | | | | | | | | |
| | 210337_s_at | ACLY | 2.14 | | | | | | | | | |
| | 202037_s_at | SFRP1 | 2.14 | 38 | | | | | | | | ■ |
| | 210543_s_at | PRKDC | 2.13 | | | | | | | | | |
| | 202716_at | PTPN1 | 2.12 | | | | | | | | | |
| | 217528_at | CLCA2 | 2.12 | 39 | | | | | | | | |
| | 201462_at | SCRN1 | 2.12 | | | | | | | | | |
| | 201820_at | KRT5 | 2.12 | 40 | ■ | | | | | | | ■ |
| | 212186_at | ACACA | 2.10 | 41 | | | | | | | | |
| | 218454_at | FLJ22662 | 2.09 | | | | | | | | | |
| | 208703_s_at | APLP2 | 2.07 | | | | | | | | | |
| | 204480_s_at | C9orf16 | 2.07 | | | | | | | | | |
| | 215936_s_at | KIAA1033 | 2.05 | | | | | | | | | |
| | 218996_at | TFPT | 2.04 | | | | | | | | | |
| | 211944_at | BAT2D1 | 2.03 | | | | | | | | | |
| | 211352_s_at | NCOA3 | 2.03 | 42 | | | | | | | | |
| | 211505_s_at | STAU1 | 2.02 | | | | | | | | | |
| | 203356_at | CAPN7 | 2.02 | | | | | | | | | |
| | 212348_s_at | AOF2 | 2.01 | | | | | | | | | |
| down | 205645_at | REPS2 | 0.47 | | | | | | | | | |
| | 209156_s_at | COL6A2 | 0.45 | | | | | | | | | |
| | 209459_s_at | ABAT | 0.43 | | | | | | | | | |
| | 205358_at | GRIA2 | 0.38 | | ■ | | | | | | | |
| | 202376_at | SERPINA3 | 0.38 | 43 | | | | | | ■ | | |
| | 203413_at | NELL2 | 0.14 | | | | | | | | | |

**FIGURE 2** – List of genes differentially expressed between high-risk and low-risk samples. For each gene, Affymetrix IDs and fold changes are shown. Gene symbols shown are those provided by Affymetrix. Breast cancer related column represents results of literature searching for correlation with breast cancer; each finding is indicated by a blue cell with the respective reference number. Breast cancer related signature overlap columns represent intersection with the published breast cancer signatures indicated by the column headers; shared genes are labeled as orange cells. Gene names are available in the Supplementary Information, Table SIII.

least 1 expression profile. Only 1 gene (*SLPI*) is shared by 3 published signatures[9,45,46] (Fig. 2). The general overlap was very low and proportional to the number of genes of each signature. Despite this little overlap, the combination of functional annotation and available literature is congruent with the involvement of the genes in our selection in breast cancer progression.

**TABLE II –** GO ENRICHMENT ANALYSIS OF DIFFERENTIALLY EXPRESSED GENES

| | Count[1] | %[2] | *p*-value[3] |
|---|---|---|---|
| Biological process | | | |
| Biosynthesis | 13 | 18.06 | 0.007 |
| Amino acid and derivative metabolism | 6 | 8.33 | 0.010 |
| Carboxylic acid metabolism | 7 | 9.72 | 0.018 |
| Organic acid metabolism | 7 | 9.72 | 0.019 |
| Amine metabolism | 6 | 8.33 | 0.022 |
| Biogenic amine metabolism | 3 | 4.17 | 0.023 |
| Cellular biosynthesis | 11 | 15.28 | 0.024 |
| Nitrogen compound metabolism | 6 | 8.33 | 0.029 |
| Amino acid derivative metabolism | 3 | 4.17 | 0.031 |
| Polyamine biosynthesis | 2 | 2.78 | 0.037 |
| Amine biosynthesis | 3 | 4.17 | 0.043 |
| Nitrogen compound biosynthesis | 3 | 4.17 | 0.043 |
| Polyamine metabolism | 2 | 2.78 | 0.052 |
| Primary metabolism | 36 | 50.00 | 0.070 |
| Aromatic compound metabolism | 3 | 4.17 | 0.092 |
| Amino acid metabolism | 4 | 5.56 | 0.098 |
| Molecular function | | | |
| Catalytic activity | 32 | 44.44 | 0.015 |
| Prenylated protein tyrosine phosphatase activity | 3 | 4.17 | 0.039 |
| Protein tyrosine phosphatase activity | 3 | 4.17 | 0.068 |
| Structural constituent of cytoskeleton | 3 | 4.17 | 0.091 |
| Pathways | | | |
| Arginine and proline metabolism | 4 | 5.56 | 0.005 |
| Urea cycle and metabolism of amino groups | 3 | 4.17 | 0.011 |
| Cell cycle | 4 | 5.56 | 0.043 |
| Insulin signaling pathway | 4 | 5.56 | 0.078 |

[1]Observed number of gene with a given GO annotation.–[2]Percentage of gene with a given GO annotation.–[3]*p* value for significance of GO term enrichment.

To investigate whether the gene signatures identified by other authors discriminate between high-risk and low-risk pools as does our gene list, we performed a classification analysis using the SVMs, a set of related supervised learning methods used for classification and regression. We analyzed only signatures obtained using Affymetrix microarrays in order not to alter their composition, due to an imperfect platform overlap. Sotiriou grading signature,[20] Miller gene list,[18] Minn genes associated to lung metastasis[19] and Kann bone signature[17] correctly classified each pool of ours, while Wang gene list[4] misclassified 1 pool out of 12. Concluding 4 out of the 5 analyzed signatures classified our samples correctly.

*Individual prognostic value of selected genes*

Next, we set out to ascertain if some of the selected genes could have the characteristics of individual prognostic markers. After exclusion of genes that were previously studied for association with survival in breast cancer, as discussed earlier, 6 upregulated genes were randomly selected: *CKMT1B* (*CKMT1B creatine kinase, mitochondrial 1B*), *SLPI* (*secretory leukocyte peptidase inhibitor*), *DDX21* (*DEAD box polypeptide 21*), *YWHAE* (*tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, epsilon polypeptide*), *PTPN1* (*protein tyrosine phosphatase, non-receptor type 1*) and *PRKDC* (*protein kinase, DNA-activated, catalytic polypeptide*). Among the small group of downregulated genes in high-risk samples, the *SERPINA3* (*serpin peptidase inhibitor, clade A, alpha-1 antiproteinase, antitrypsin, member 3*) gene was chosen because of conflicting literature data regarding breast cancer association.[43,47] mRNA expression level of these genes was analyzed by qRT-PCR in all 127 individual tumor samples (Table I). Six independent cDNA preparations were used for each sample and the values averaged.

qRT-PCR results related to the 60 samples employed for microarray analysis gave an indirect validation of microarray data. Fold changes among high- and low-risk samples was calculated and compared to those obtained from microarray data. In all case, comparable values were found, although qRT-PCR derived fold changes were generally wider (data not shown).

Next, we investigated whether these genes were also differentially expressed in the whole set of 127 samples, of which 55 were classified as low-risk (disease-free at 72 months) and 72 as high-risk (relapsing within 72 months). Distribution analysis of mRNA expression levels was performed with the Mann–Whitney nonparametric statistical test, and gave significant results for the 6 upregulated genes, as shown in Figure 3. Comparable results were obtained limiting the analysis to the 60 samples used for microarrays. Contrarily, the downregulated gene *SERPINA3* was not significant in both groups.

Next, the possible association of gene expression with survival was studied by the Kaplan–Meier estimate and Log-rank test. For each gene, "low" and "high" expression was defined by using the median as cut-off value.

For all 6 upregulated genes, high gene expression was significantly associated with early death (Fig. 4*a*) and a shorter disease-free survival (Fig. 4*b*), in keeping with microarray results. On the other hand, in both analyses, *SERPINA3* did not show any significant association with survival, in contrast with its lower expression found in poor prognosis pools by microarray analysis.

The 7 gene expression values were then subjected to statistical analysis to reveal significant associations with other clinicopathological data, using the Mann-Whitney and Kruskal–Wallis nonparametric statistics. No significant correlation with any parameter was observed, with the exception of *SLPI* and *SERPINA3* expression that correlated with steroid receptor status (*SLPI*: negative correlation, $p = 0.02$; *SERPINA3*: positive correlation, $p = 0.02$).

The association of gene expression with survival was further studied using Cox's proportional hazard model. In univariate analysis, expression values for each gene were significantly associated with survival (Table III), with the exception of *SERPINA3*. In a multivariate model that included as covariates the variables that were significant at univariate level (ER and lymph node status), all previously significant genes kept their statistical significance (Table III). This result confirmed that the analyzed genes are novel independent prognostic markers.

In conclusion, the analysis of association with survival showed that all upregulated genes examined correlated with progression,

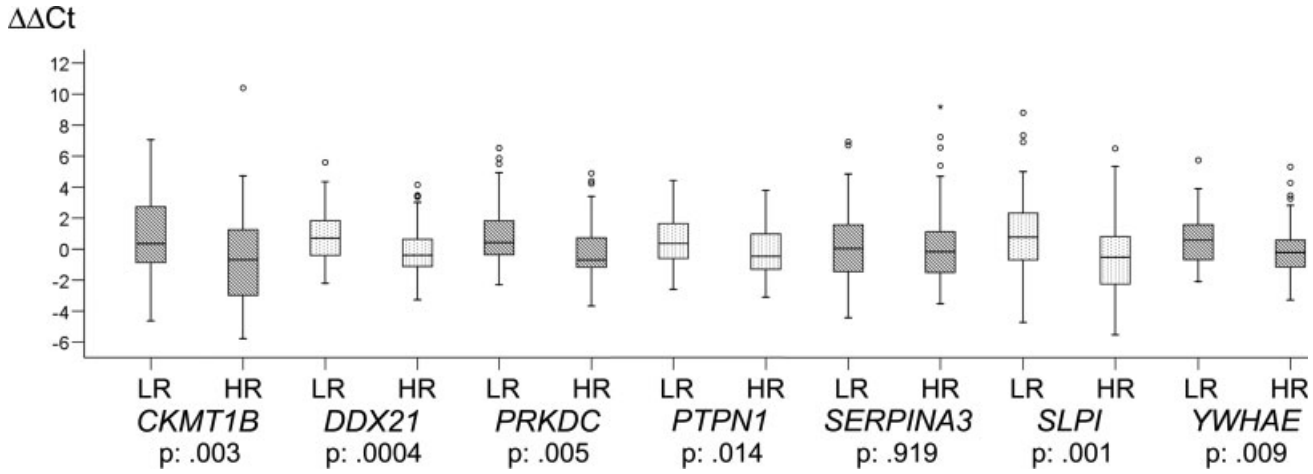


**FIGURE 3** – Distribution analysis of mRNA expression levels analyzed by box-and-whiskers plot. ΔΔCts obtained from the 2 risk categories of samples (LR, low risk; HR, high risk) are compared for each gene analyzed by qRT-PCR (lower ΔΔCts value means higher expression). Circles label outliers (cases with values between 1.5 and 3 interquartile range) and asterisks mark extreme values (cases with values more than 3 interquartile range). Shown *p*-values were obtained applying the Mann–Whitney nonparametric statistical test.

and in particular *DDX21* and *PRKDC* which displayed a very significant value also in Cox's multivariate analysis.

*Validation of selected genes on independent data sets*

One drawback of our study is that the 60 samples used for differential expression analysis were included in the 127-sample cohort used for validation. To reduce this problem, the prognostic significance of the genes studied by qRT-PCR was addressed on different patient cohorts. Three publicly available datasets were considered.[3,18,20] Expression data of each gene were obtained from these datasets and subjected to Kaplan–Meier and Cox analysis.

As shown in Figure 5*a*, in van de Vijver dataset results,[3] an association between poor prognosis and high expression of *CKMT1B* and *PRKDC* was revealed. Notably, only *SERPINA3* was found to be associated with long-term survival, as expected from our microarray results. By Cox univariate analysis, these 3 genes were significantly associated with survival (*CKMT1B*, $p = 0.004$; *PRKDC*, $p = 0.000$; *SERPINA3*, $p = 0.000$) and maintained the association in multivariate analysis adding as covariate the ER status (*CKMT1B*, $p = 0.037$; *PRKDC*, $p = 0.000$; *SERPINA3*, $p = 0.001$). We did not use lymph node status as covariate as it was not significant in univariate analysis ($p = 0.561$). A Kaplan Meier analysis was carried out sorting samples for estrogen receptor expression or lymph node invasion. *PTPN1* exhibited association with poor prognosis in node-positive samples ($p = 0.010$), while *SLPI* associated with shorter disease-free survival in node-negative samples ($p = 0.041$).

A Kaplan–Meier analysis of dataset from Sotirou *et al*.[20] (Fig. 5*b*) revealed significant correlation with a short disease-free survival for *CKMT1B* expression and *PRKDC* expression and this was confirmed by univariate Cox analysis (*CKMT1B*, $p = 0.002$; *PRKDC*, $p = 0.033$). Noteworthy, a high expression of *SLPI* was associated to long-term survival by both Kaplan–Meier and univariate Cox's models, contrary to our findings. Furthermore node-positive samples showed a correlation with short disease-free survival for *YWHAE* expression ($p = 0.005$).

Finally, the entire dataset of Miller[18] showed an association between an increased risk of relapse and *CKMT1B* expression (Fig. 5*c*), also confirmed by Cox's univariate model ($p = 0.022$) and multivariate analysis, with nodal status and p53 mutation as covariates ($p = 0.059$). According to Kaplan–Meier, the risk of recurrence was associated with *PTPN1* expression in node-positive cases ($p = 0.016$).

In conclusion, *CKMT1B* was identified as a prognostic factor in 3 independent cohorts of breast cancer patients for the first time in our study.

## Discussion

A direct comparison of gene expression in 2 balanced groups of breast tumors with different risk of relapse identified a novel group of genes that showed significant association with both disease-free and overall survival, when tested individually by a quantitative method. Some of these genes are novel independent prognostic markers and they represent novel genes linked to breast cancer as well, indicating pathways and targets that can be further studied and exploited for therapy.

To identify a set of genes differentially expressed in low- *versus* high-risk tumors by microarray analysis, the pooling strategy was preferred, for several reasons. First, since our goal was to identify novel genes rather than a prognostic signature, more resources were spent to guarantee high accuracy for the quantitative evaluation than for microarray detection step. Second, pooling RNAs will shield and reduce detection of genes showing small variations between groups but high variation between individual samples, as in the case of genes that are activated or repressed only in a small percentage of tumors (*e.g.* the *ERBB2* gene). Conversely, this will enhance detection of genes that are consistently upregulated or downregulated in 1 group. Thirdly, by measuring gene expression in pools, the average of individual values is measured. Since distribution of gene expression in tumor samples is often not Gaussian but follows a Poisson' distribution, this will possibly favor detection of upregulated genes in high-risk tumors.

It has been predicted that this approach would produce quite different results from those obtained by profiling a series of consecutive, unselected breast cancer cases individually.[22] Indeed, in our study, most of the genes are associated with the risk of relapse *per se*, rather than being part of a signature, as is often found in other studies.

This conclusion was justified not only by the fact that 6 out of the 7 genes individually tested by qRT-PCR displayed significant association with survival in our cohort of patients, but also because additional genes in this group had already been associated to prognosis. By inference, it is most likely that, within the 77-gene set, a number of other genes with similar characteristics will be found.

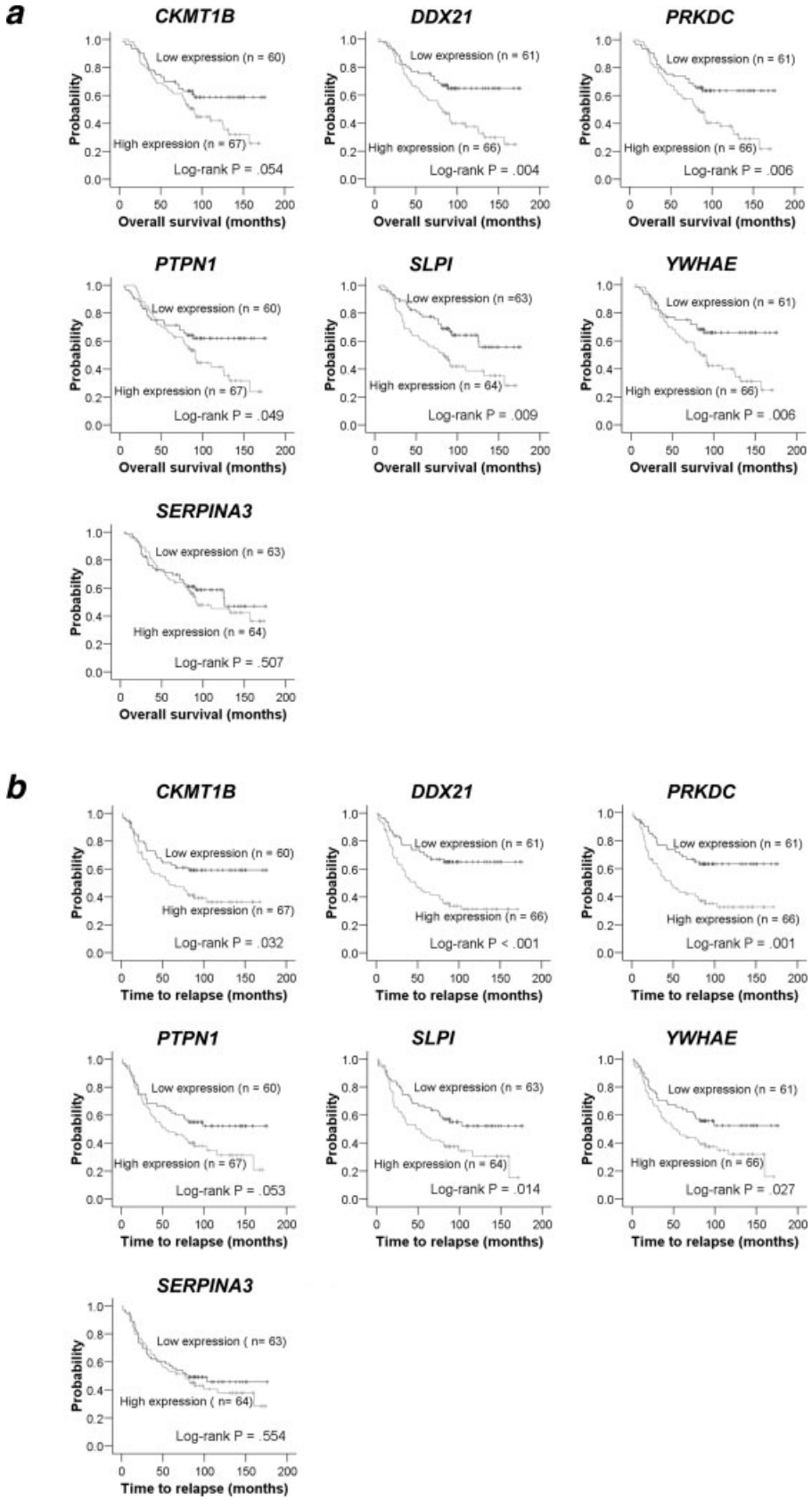


**FIGURE 4** – Kaplan–Meier analysis of the probability of overall survival (*a*) and of the probability that patients would remain free of disease (*b*) among all patients, as calculated on the expression of genes analyzed by qRT-PCR. Low expression and high expression are defined using the median value as cut-off.

**TABLE III –** UNIVARIATE AND MULTIVARIATE ANALYSIS OF ANALYZED GENE EXPRESSION, LYMPH NODE STATUS AND ESTROGEN RECEPTOR STATUS ($N$ = 127)

| Characteristic | HR[1] | CI[2] | $p$-value[3] |
|---|---|---|---|
| Univariate analysis | | | |
| *DDX21* | 1.8 | 1.4–4.2 | 0.019 |
| *CKMT1* | 2.4 | 1.1–3.1 | 0.001 |
| *SLPI* | 1.9 | 1.1–3.2 | 0.010 |
| *YWHAE* | 2.0 | 1.2–3.5 | 0.007 |
| *PTPN1* | 1.7 | 1.0–2.8 | 0.050 |
| *PRKDC* | 2.3 | 1.3–4.0 | 0.002 |
| *SERPINA3* | 1.2 | 0.7–1.9 | 0.517 |
| LN | 2.0 | 1.1–3.7 | 0.019 |
| ER | 1.7 | 1.0–2.8 | 0.045 |
| Multivariate analysis | | | |
| *DDX21* | 2.2 | 1.2–3.8 | 0.004 |
| LN | 2.0 | 1.1–3.7 | 0.018 |
| ER | 1.7 | 1.0–2.8 | 0.036 |
| *CKMT1* | 1.8 | 1.0–3.1 | 0.022 |
| LN | 2.1 | 1.2–4.0 | 0.010 |
| ER | 1.8 | 1.1–3.0 | 0.017 |
| *SLPI* | 1.6 | 0.9–2.7 | 0.072 |
| LN | 2.0 | 1.1–3.7 | 0.022 |
| ER | 1.6 | 0.9–2.7 | 0.056 |
| *YWHAE* | 1.9 | 1.1–3.3 | 0.012 |
| LN | 2.1 | 1.1–3.9 | 0.011 |
| ER | 1.8 | 1.1–3.0 | 0.021 |
| *PTPN1* | 1.6 | 0.9–3.8 | 0.055 |
| LN | 2.1 | 1.1–3.9 | 0.011 |
| ER | 1.8 | 1.1–3.1 | 0.014 |
| *PRKDC* | 2.0 | 1.2–3.6 | 0.007 |
| LN | 2.1 | 1.1–3.8 | 0.015 |
| ER | 1.6 | 1.0–2.7 | 0.049 |
| *SERPINA3* | 1.0 | 0.6–1.8 | 0.755 |
| LN | 2.1 | 1.1–4.0 | 0.012 |
| ER | 1.8 | 1.1–3.0 | 0.017 |

LN = positive lymph node risk; ER = negative receptor status risk. [1]Variable hazard ratio in the model.–[2]95% confidence interval.–[3]Based on Cox regression.

A critical point in our study regards the 60 samples used for the detection step by microarray analysis, also part of the overall cohort where individual associations with survival were calculated. The association with survival, however, was clear when the expression of these genes was extrapolated from 3 very popular publicly available datasets, *i.e.,* the Amsterdam study,[3] Miller *et al*. study[18] and the Sotiriou *et al*. study.[20] This was true in spite of these facts, *e.g*., (*i*) microarray expression data of single genes are definitely less quantitative than real-time RT-PCR data, (*ii*) regions targeted by microrray probes in different platforms and in TaqMan assays do not always correspond, (*iii*) patient cohorts differ in many regards, including N−/N+ ratio, age, treatment and other. Indeed it is noteworthy that at least 2 genes discovered by our study still show a significant association with both relapse-free and overall survival in at least 2 out of 3 studies, even in multivariate analysis, definitely confirming that these genes are novel independent prognostic markers for breast cancer.

The function of the 7 individually studied genes also suggests that they deserve further studies. The *CKMT1B* gene, which gave the most consistent results also in metanalysis, encodes a protein that transfers high energy phosphate from mitochondria to cytosolic creatine. It was overexpressed in cancers with poor prognosis and correlated to the high energy turnover that characterizes growing tumor tissue.[48] Moreover, in the octameric state, CKMT1B interacts with porin of mitochondrial membrane pore, reducing the probability of pore opening, thus interfering with the induction of apoptosis dependent on cytochrome *c* release.[49]

DDX21 is a putative nucleolar ATP-dependent RNA helicase which plays an important role in ribosomal RNA biogenesis, RNA editing and RNA transport.[50] When it moves to the nuclear compartment, it interacts with the *c-jun* oncogene and acts as a transcriptional coactivator.[51]

The 14-3-3ε protein encoded by the *YWHAE* gene binds to phosphoserine-containing proteins and mediates signal transduction. For example, 14-3-3 binding is required for the stabilization of active RAF-1[52] and CDC25-mediated cell cycle control,[53] whereas its interaction with BAD and BAX prevents their proapoptotic release to mitochondrial membrane.[54,55]

The PRKDC protein belongs to the PI3-K related kinase family and represents a key complex for DNA repair. It is involved in the nonhomologous end-joining process corresponding to the major activity responsible for cell survival when double strand breaks in the DNA are produced, after ionizing radiation or chemotherapeutic treatments.[56]

PTPN1 is a nonreceptor protein-tyrosine phosphatase that modulates protein phosphorylation in cell signaling networks. It is involved in leptin and insulin signaling and in several other signaling pathways such as growth factor and integrin mediated processes.[57] Several studies demonstrated that changes in abundance and distribution of PTPases could impair insulin signal transduction, causing insulin resistance.[58] Aberrant insulin signaling, which leads to insulin resistance, hyperinsulinaemia and increased concentrations of endogenous estrogen and androgen, was linked to high breast cancer risk by clinical and experimental evidence.[59]

The above mentioned functions confirmed the expression behavior of selected genes, in tumor samples of our study and in metanalysis results. A quite different situation was observed for the remaining upregulated gene, *SLPI*. SLPI is a secreted serine proteinase inhibitor that protects epithelial tissues from inflammation-induced damage caused by endogenous proteolytic enzymes and it exerts its activity against neutrophil elastase, cathepsin G, trypsin and chymotrypsin.[60] SLPI also exhibits proliferative effects, although its mechanism remains unknown.[61] Several studies demonstrated that SLPI is altered in cancer; it was found upregulated in ovarian and in lung carcinomas, and its serum level correlated with tumor stage and response to therapy.[62,63] The exact mechanism by which SLPI promotes malignancy is not yet known: in addition to tumor growth support, the enhancement of malignancy could be due to its effects on angiogenesis. In fact, SLPI prevents the formation of the antiangiogenic factor endostatin, by inhibiting elastase, its activator.[64]

In our cohort, qRT-PCR clearly indicated that high *SLPI* expression was correlated with high-risk of relapse and death. However, in the Sotiriou dataset, we observed a positive correlation of *SLPI* expression with low-risk patients and, in another study, a protective effect against liver metastasis was reported, explained through a reduced inflammatory response.[65] The SLPI effect in reducing the inflammatory response could also explain the results obtained in Sotiriou dataset. In fact, in this study the only adjuvant treatment used was tamoxifen and it is known that inflammation is linked to resistance to endocrine treatments.[66]

The only downregulated gene found by microarrays in high-risk patients studied by qRT-PCR was *SERPINA3*, because of uncertain results in the literature.[43,47] SERPINA3 or Alpha-1-antichymotrypsin (ACT) is a well-known serine protease inhibitor that regulates the activity of cathepsin G in neutrophils[67] and is an estrogen-induced gene.[43] In breast cancer, *SERPINA3* mRNA expression was reported as an indicator of good prognosis, but only in ER-positive tumors,[43] while SERPINA3 protein expression was reported as unfavorable factor.[47] Our qRT-PCR results did not confirm a prognostic role for *SERPINA3*, contrary to pooled microarray results, whereas metanalysis of the van de Vijver dataset confirmed a positive correlation with favorable prognosis. Nevertheless, our qRT-PCR data confirmed a positive correlation with steroid receptor status. Certainly, differences in the cohorts studied, as well as in the adjuvant treatments used in different studies could account for the observed discrepancies.

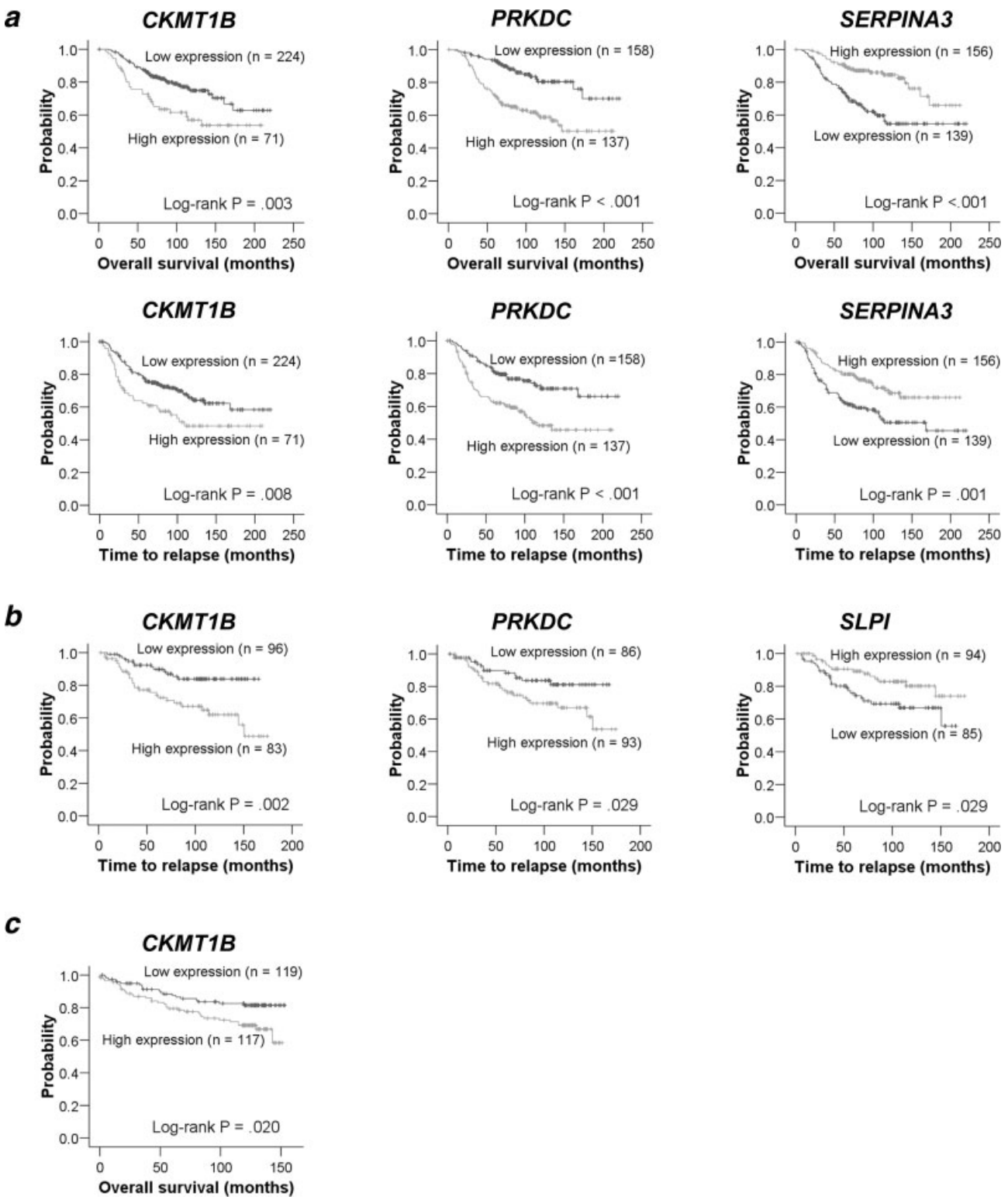


**FIGURE 5** – The association with survival of the 7 risk-associated genes was evaluated by Kaplan–Meier method in 3 datasets available online. Only genes showing significant or borderline association are shown. Panel (*a*) shows results from van de Vijver *et al*. dataset.[3] Panel (*b*) illustrates results of Sotiriou *et al*. dataset.[20] Panel (*c*) shows results obtained from Miller *et al*. dataset.[18]

In conclusion, we have performed a profiling experiment with some unusual characteristics, as compared to published breast cancer profiling studies. As expected, our approach allowed for the discovery of a number of verified or potential new prognostic markers and biological targets, rather than predictive signatures. Many of the genes found are novel, and as inferred from the results obtained on some of them, deserve further studies.

## Acknowledgements

The authors thank Mr. GianMario Milano, Dr. Barbara Martinoglio, Dr. Mauro Helmer Citterich and Dr. Olivier Friard for technical assistance, the surgeons, pathologists, and internists of the Saint Anna Hospital (Turin) and of the IRCC Hospital in Candiolo (Turin), for the supply of tumor tissues, for their assistance in the collection of the clinical follow-up data, or both, and Prof. Raffaele Calogero for microarray data analysis suggestions.

## References

1. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, et al. Molecular portraits of human breast tumours. Nature 2000;406:747–52.
2. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415:530–6.
3. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrish M, Atsma D, et al. A gene-expression signature as a predictor of survival in breast cancer. New Engl J Med 2002;347:1999–2009.
4. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, Talantov D, Timmermans M, Meijer-van Gelder ME, Yu J, Jatkoe T, Berns EM, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 2005;365:671–9.
5. Sotiriou C, Neo SY, McShane LM, Korn EL, Long PM, Jazaeri A, Martiat P, Fox SB, Harris AL, Liu ET. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. Proc Natl Acad Sci USA 2003;100:10393–8.
6. Sotiriou C, Powles TJ, Dowsett M, Jazaeri AA, Feldman AL, Assersohn L, Gadisetti C, Libutti SK, Liu ET. Gene expression profiles derived from fine needle aspiration correlate with response to systemic chemotherapy in breast cancer. Breast Cancer Res 2002;4:R3.
7. Jansen MP, Foekens JA, van Staveren IL, Dirkzwager-Kiel MM, Ritstier K, Look MP, Meijer-van Gelder ME, Sieuwerts AM, Portengen H, Dorssers LC, Klijn JG, Berns EM. Molecular classification of tamoxifen-resistant breast carcinomas by gene expression profiling. J Clin Oncol 2005;23:732–40.
8. Chang HY, Nuyten DS, Sneddon JB, Hastie T, Tibshirani R, Sorlie T, Dai H, He YD, van't Veer LJ, Bartelink H, van de Rijn M, Brown PO, et al. Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. Proc Natl Acad Sci USA 2005;102:3738–43.
9. Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S, Demeter J, Perou CM, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci USA 2003;100:8418–23.
10. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci USA 2001;98:10869–74.
11. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, Baehner FL, Walker MG, Watson D, Park T, Hiller W, Fisher ER, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. New Engl J Med 2004;351:2817–26.
12. Fan C, Oh DS, Wessels L, Weigelt B, Nuyten DS, Nobel AB, van't Veer LJ, Perou CM. Concordance among gene-expression-based predictors for breast cancer. New Engl J Med 2006;355:560–9.
13. Miller LD, Liu ET. Expression genomics in breast cancer research: microarrays at the crossroads of biology and medicine. Breast Cancer Res 2007;9:206.
14. Son CG, Bilke S, Davis S, Greer BT, Wei JS, Whiteford CC, Chen QR, Cenacchi N, Khan J. Database of mRNA gene expression profiles of multiple human organs. Genome Res 2005;15:443–50.
15. Baldi P, Long AD. A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. Bioinformatics 2001;17:509–19.
16. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA 2001;98:5116–21.
17. Kang Y, Siegel PM, Shu W, Drobnjak M, Kakonen SM, Cordon-Cardo C, Guise TA, Massague J. A multigenic program mediating breast cancer metastasis to bone. Cancer Cell 2003;3:537–49.
18. Miller LD, Smeds J, George J, Vega VB, Vergara L, Ploner A, Pawitan Y, Hall P, Klaar S, Liu ET, Bergh J. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. Proc Natl Acad Sci USA 2005;102:13550–5.
19. Minn AJ, Gupta GP, Siegel PM, Bos PD, Shu W, Giri DD, Viale A, Olshen AB, Gerald WL, Massague J. Genes that mediate breast cancer metastasis to lung. Nature 2005;436:518–24.
20. Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, Smeds J, Nordgren H, Farmer P, Praz V, Haibe-Kains B, Desmedt C, Larsimont D, et al. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. J Natl Cancer Inst 2006;98:262–72.
21. Frank E, Hall M, Trigg L, Holmes G, Witten IH. Data mining in bioinformatics using Weka. Bioinformatics 2004;20:2479–81.
22. Peng X, Wood CL, Blalock EM, Chen KC, Landfield PW, Stromberg AJ. Statistical implications of pooling RNA samples for microarray experiments. BMC Bioinformatics 2003;4:26.
23. Cross SS, Hamdy FC, Deloulme JC, Rehman I. Expression of S100 proteins in normal human tissues and common cancers using tissue microarrays: S100A6. S100A8, S100A9 and. S100A11 are all overexpressed in common cancers. Histopathology 2005;46:256–69.
24. Koster A, Landgraf S, Leipold A, Sachse R, Gebhart E, Tulusan AH, Ronay G, Schmidt C, Dingermann T. Expression of oncogenes in human breast cancer specimens. Anticancer Res 1991;11:193–201.
25. Folgueira MA, Brentani H, Katayama ML, Patrao DF, Carraro DM, Mourao Netto M, Barbosa EM, Caldeira JR, Abreu AP, Lyra EC, Kaiano JH, Mota LD, et al. Gene expression profiling of clinical stages II and III breast cancer. Braz J Med Biol Res 2006;39:1101–13.
26. Ai L, Tao Q, Zhong S, Fields CR, Kim WJ, Lee MW, Cui Y, Brown KD, Robertson KD. Inactivation of Wnt inhibitory factor-1 (WIF1) expression by epigenetic silencing is a common event in breast cancer. Carcinogenesis 2006;27:1341–8.
27. Debniak T, Scott R, Masojc B, Serrano-Fernandez P, Huzarski T, Byrski T, Debniak B, Gorski B, Cybulski C, Medrek K, Kurzawski G, van de Wetering T, et al. MC1R common variants. CDKN2A and their association with melanoma and breast cancer risk. Int J Cancer 2006;119:2597–602.
28. Redondo M, Villar E, Torres-Munoz J, Tellez T, Morell M, Petito CK. Overexpression of clusterin in human breast carcinoma. Am J Pathol 2000;157:393–9.
29. Manni A, Astrow SH, Gammon S, Thompson J, Mauger D, Washington S. Immunohistochemical detection of ornithine-decarboxylase in primary and metastatic human breast cancer specimens. Breast Cancer Res Treat 2001;67:147–56.
30. King JA, Ofori-Acquah SF, Stevens T, Al-Mehdi AB, Fodstad O, Jiang WG. Activated leukocyte cell adhesion molecule in breast cancer: prognostic indicator. Breast Cancer Res 2004;6:R478–87.
31. Chintamani Jha BP, Bhandari V, Bansal A, Saxena S, Bhatnagar D. The expression of mismatched repair genes and their correlation with clinicopathological parameters and response to neo-adjuvant chemotherapy in breast cancer. Int Semin Surg Oncol 2007;4:5.
32. Wirtenberger M, Tchatchou S, Hemminki K, Klaes R, Schmutzler RK, Bermejo JL, Chen B, Wappenschmidt B, Meindl A, Bartram CR, Burwinkel B. Association of genetic variants in the Rho guanine nucleotide exchange factor AKAP13 with familial breast cancer. Carcinogenesis 2006;27:593–8.
33. Simins AB, Weighardt H, Weidner KM, Weidle UH, Holzmann B. Functional cloning of ARM-1, an adhesion-regulating molecule upregulated in metastatic tumor cells. Clin Exp Metastasis 1999;17:641–8.
34. Suefuji Y, Sasatomi T, Shichijo S, Nakagawa S, Deguchi H, Koga T, Kameyama T, Itoh K. Expression of SART3 antigen and induction of CTLs by SART3-derived peptides in breast cancer patients. Br J Cancer 2001;84:915–9.
35. Ito Y, Yoshida H, Uruno T, Takamura Y, Miya A, Kuma K, Miyauchi A. Expression of cdc25A and cdc25B phosphatase in breast carcinoma. Breast Cancer 2004;11:295–300.
36. Suh KS, Crutchley JM, Koochek A, Ryscavage A, Bhat K, Tanaka T, Oshima A, Fitzgerald P, Yuspa SH. Reciprocal modifications of CLIC4 in tumor epithelium and stroma mark malignant progression of multiple human cancers. Clin Cancer Res 2007;13:121–31.
37. Luftner D, Mesterharm J, Akrivakis C, Geppert R, Petrides PE, Wernecke KD, Possinger K. Tumor type M2 pyruvate kinase expression in advanced breast cancer. Anticancer Res 2000;20:5077–82.
38. Ugolini F, Charafe-Jauffret E, Bardou VJ, Geneix J, Adelaide J, Labat-Moleur F, Penault-Llorca F, Longy M, Jacquemier J, Birnbaum D, Pebusque MJ. WNT pathway and mammary carcinogenesis: loss of expression of candidate tumor suppressor gene SFRP1 in most invasive carcinomas except of the medullary type. Oncogene 2001;20:5810–7.
39. Li X, Cowell JK, Sossey-Alaoui K. CLCA2 tumour suppressor gene in 1p31 is epigenetically regulated in breast cancer. Oncogene 2004;23:1474–80.

40. Turashvili G, Bouchal J, Baumforth K, Wei W, Dziechciarkova M, Ehrmann J, Klein J, Fridman E, Skarda J, Srovnal J, Hajduch M, Murray P, et al. Novel markers for differentiation of lobular and ductal invasive breast carcinomas by laser microdissection and microarray analysis. BMC Cancer 2007;7:55.
41. Chin K, DeVries S, Fridlyand J, Spellman PT, Roydasgupta R, Kuo WL, Lapuk A, Neve RM, Qian Z, Ryder T, Chen F, Feiler H, et al. Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. Cancer Cell 2006;10:529–41.
42. Zhao C, Yasui K, Lee CJ, Kurioka H, Hosokawa Y, Oka T, Inazawa J. Elevated expression levels of NCOA3. TOP1, and TFAP2C in breast tumors as predictors of poor prognosis. Cancer 2003;98:18–23.
43. Yamamura J, Miyoshi Y, Tamaki Y, Taguchi T, Iwao K, Monden M, Kato K, Noguchi S. mRNA expression level of estrogen-inducible gene, α1-antichymotrypsin, is a predictor of early tumor recurrence in patients with invasive breast cancers. Cancer sci 2004;95:887–92.
44. Garcia-Bassets I, Kwon YS, Telese F, Prefontaine GG, Hutt KR, Cheng CS, Ju BG, Ohgi KA, Wang J, Escoubet-Lozach L, Rose DW, Glass CK, et al. Histone methylation-dependent mechanisms impose ligand dependency for gene activation by nuclear receptors. Cell 2007;128:505–18.
45. Chang HY, Sneddon JB, Alizadeh AA, Sood R, West RB, Montgomery K, Chi JT, van de Rijn M, Botstein D, Brown PO. Gene expression signature of fibroblast serum response predicts human cancer progression: similarities between tumors and wounds. PLoS Biol 2004;2:E7.
46. West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, Zuzan H, Olson JA, Jr, Marks JR, Nevins JR. Predicting the clinical status of human breast cancer by using gene expression profiles. Proc Natl Acad Sci USA 2001;98:11462–7.
47. Hurlimann J, van Melle G. Prognostic value of serum proteins synthesized by breast carcinoma cells. Am J Clin Pathol 1991;95:835–43.
48. Joseph J, Cardesa A, Carreras J. Creatine kinase activity and isoenzymes in lung, colon and liver carcinomas. Br J Cancer 1997;76:600–5.
49. Vyssokikh MY, Brdiczka D. The function of complexes between the outer mitochondrial membrane pore (VDAC) and the adenine nucleotide translocase in regulation of energy metabolism and apoptosis. Acta Biochim Pol 2003;50:389–404.
50. Valdez BC, Wang W. Mouse RNA helicase II/Gu: cDNA and genomic sequences, chromosomal localization, and regulation of expression. Genomics 2000;66:184–94.
51. Westermarck J, Weiss C, Saffrich R, Kast J, Musti AM, Wessely M, Ansorge W, Seraphin B, Wilm M, Valdez BC, Bohmann D. The DEXD/H-box RNA helicase RHII/Gu is a co-factor for c-Jun-activated transcription. EMBO J 2002;21:451–60.
52. Roy S, McPherson RA, Apolloni A, Yan J, Lane A, Clyde-Smith J, Hancock JF. 14–3–3 facilitates Ras-dependent Raf-1 activation in vitro and in vivo. Mol Cell Biol 1998;18:3947–55.
53. Thorson JA, Yu LW, Hsu AL, Shih NY, Graves PR, Tanner JW, Allen PM, Piwnica-Worms H, Shaw AS. 14–3–3 proteins are required for maintenance of Raf-1 phosphorylation and kinase activity. Mol Cell Biol 1998;18:5229–38.
54. Won J, Kim DY, La M, Kim D, Meadows GG, Joe CO. Cleavage of 14–3–3 protein by caspase-3 facilitates bad interaction with Bcl-x(L) during apoptosis. J Biol Chem 2003;278:19347–51.
55. Nomura M, Shimizu S, Sugiyama T, Narita M, Ito T, Matsuda H, Tsujimoto Y. 14–3–3 Interacts directly with and negatively regulates pro-apoptotic Bax. J Biol Chem 2003;278:2058–65.
56. Salles B, Calsou P, Frit P, Muller C. The DNA repair complex DNA-PK, a pharmacological target in cancer chemotherapy and radiotherapy. Pathol Biol 2006;54:185–93.
57. Liang F, Lee SY, Liang J, Lawrence DS, Zhang ZY. The role of protein-tyrosine phosphatase 1B in integrin signaling. J Biol Chem 2005;280:24857–63.
58. Byon JC, Kusari AB, Kusari J. Protein-tyrosine phosphatase-1B acts as a negative regulator of insulin signal transduction. Mol Cell Biochem 1998;182:101–8.
59. Stoll BA. Upper abdominal obesity, insulin resistance and breast cancer risk. Int J Obes Relat Metab Disord 2002;26:747–53.
60. Boudier C, Cadene M, Bieth JG. Inhibition of neutrophil cathepsin G by oxidized mucus proteinase inhibitor. Effect of heparin. Biochemistry 1999;38:8451–7.
61. Zhang D, Simmen RC, Michel FJ, Zhao G, Vale-Cruz D, Simmen FA. Secretory leukocyte protease inhibitor mediates proliferation of human endometrial epithelial cells by positive and negative regulation of growth-associated genes. J Biol Chem 2002;277:29999–30009.
62. Devoogdt N, Hassanzadeh Ghassabeh G, Zhang J, Brys L, De Baetselier P, Revets H. Secretory leukocyte protease inhibitor promotes the tumorigenic and metastatic potential of cancer cells. Proc Natl Acad Sci USA 2003;100:5778–82.
63. Tsukishiro S, Suzumori N, Nishikawa H, Arakawa A, Suzumori K. Use of serum secretory leukocyte protease inhibitor levels in patients to improve specificity of ovarian cancer diagnosis. Gynecol Oncol 2005;96:516–9.
64. O'Reilly MS, Boehm T, Shing Y, Fukai N, Vasios G, Lane WS, Flynn E, Birkhead JR, Olsen BR, Folkman J. Endostatin: an endogenous inhibitor of angiogenesis and tumor growth. Cell 1997;88:277–85.
65. Wang N, Thuraisingam T, Fallavollita L, Ding A, Radzioch D, Brodt P. The secretory leukocyte protease inhibitor is a type 1 insulin-like growth factor receptor-regulated protein that protects against liver metastasis by attenuating the host proinflammatory response. Cancer Res 2006;66:3062–70.
66. Riggins RB, Schrecengost RS, Guerrero MS, Bouton AH. Pathways to tamoxifen resistance. Cancer lett 2007;256:1–24.
67. Laine A, Davril M, Hayem A. Interaction between human serum alpha 1-antichymotrypsin and human leukocyte cathepsin G. complex formation and production of a modified inhibitor. Biochem Biophys Res Commun 1982;105:186–93.