An Innovative Microarray Strategy Identities Informative Molecular Markers for the Detection of Micrometastatic Breast Cancer

Kaidi Mikhitarian,[1] William E. Gillanders,[1] Jonas S. Almeida,[2] Renee Hebert Martin,[2] Juan C. Varela,[4] John S. Metcalf,[3] David J. Cole,[1] and Michael Mitas[1]

Abstract There is increasing evidence that molecular detection of micrometastatic breast cancer in the axillary lymph nodes (ALN) of breast cancer patients can improve staging. Molecular analyses of samples obtained from the Minimally Invasive Molecular Staging of Breast Cancer Trial (n = 489 patients) indicate that whereas the majority of molecular markers are informative for the detection of metastatic breast cancer (significant disease burden), only a few are sensitive for the detection of micrometastatic disease (limited disease burden). Frequency distribution and linear regression analyses reveal that relative levels of gene expression are highly correlated with apparent sensitivity for the detection of micrometastic breast cancer ($P < 0.05$). These data provides statistical validation of the concept that the most informative markers for detection of micrometastatic disease are those that are most highly expressed in metastatic disease. To test this hypothesis, we developed an innovative microarray strategy. RNA from a metastatic breast cancer ALN was diluted into RNA from a normal lymph node and analyzed using Affymetrix microarrays. Expression analysis indicated that only two genes [*mammaglobin* (*mam*) and *trefoil factor 1* (*TFF1*)] were significantly overexpressed at a dilution of 1:50. Real-time reverse transcription-PCR analysis of pathology-negative ALN (n = 72) confirm that of all the markers tested, *mam* and *TFF1* have the highest apparent sensitivity for detection of micrometastatic breast cancer. We conclude that a dilutional microarray approach is a simple and reliable method for the identification of informative molecular markers for the detection of micrometastatic cancer.

A major focus of current research is the application of recent advances in molecular genetics to the field of surgical pathology and the detection of metastatic and micrometastatic cancer. In this context, *metastatic* cancer is defined as metastatic cancer of significant disease burden that is readily detected by standard histopathologic techniques, whereas *micrometastatic* cancer is defined as metastatic cancer of limited disease burden that cannot be detected by standard histopathologic techniques. In the setting of breast cancer, the presence of metastatic breast cancer in axillary lymph nodes (ALN) remains one of the most important prognostic factors for predicting disease recurrence (1–3). Unfortunately, standard H&E histopathologic analysis of ALN has limitations, as it is subjective in nature and often lacks the sensitivity to detect small amounts of clinically relevant disease. A number of studies have shown that doing additional tissue sections and/or immunohistochemical staining (IHC) of ALN increases the ability to detect breast cancer metastases by up to 25% (4–6). Furthermore, these retrospective studies suggest that the prognosis for patients with micrometastatic is similar to patients with pathology-positive ALN (4, 5, 7). Taken together, these findings suggest that the development of more sensitive histopathologic or molecular techniques for the detection of micrometastatic breast cancer in ALN could significantly improve breast cancer staging.

We recently reported interim results of the Minimally Invasive Molecular Staging of Breast Cancer Trial (MIMS), a prospective cohort study designed to define the clinical significance of molecular detection of micrometastatic breast cancer in ALN (8). ALN from 489 patients with T1 to T3 primary breast cancers were analyzed by standard histopathology (H&E staining) and by multimarker, real-time reverse transcription-PCR (RT-PCR) for the following genes: *mam*, *mamB*, *muc1*, *CEA*, *PDEF*, *CK19*, and *PIP*. The interim results indicate that overexpression of breast cancer–associated genes in breast cancer subjects with pathology-negative ALN correlates with traditional predictors of disease progression, providing strong evidence that molecular markers serve as valid surrogates for the detection of micrometastatic breast cancer (8). Despite these positive results, one of the surprising findings from the interim analysis is that whereas the majority of

Authors' Affiliations: Departments of [1]Surgery, [2]Biostatistics Bioinformatics, and Epidemiology and [3]Pathology and Laboratory Medicine, [4]College of Medicine, Medical University of South Carolina, Charleston, South Carolina
Received 10/25/04; revised 2/4/05; accepted 2/24/05.
Grant support: The microarray studies were funded by the Hollings Cancer Center. The MIMS trial was funded by the Department of Defense N00014-99-1-0784.
The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.
Requests for reprints: Kaidi Mikhitarian, Department of Surgery, Medical University of South Carolina, 96 Jonathan Lucas Street, Suite 420, P.O. Box 250613, Charleston, SC 29425. Phone: 843-792-7789; Fax: 843-792-4813; E-mail: mikhitar@musc.edu.

molecular markers were informative for the detection of metastatic breast cancer, only a few were informative for the detection of micrometastatic breast cancer. Of particular interest, the molecular marker [*mammaglobin* (*mam*)] that was most highly expressed in pathology-positive ALN also had the highest apparent sensitivity for the detection of micrometastatic disease. Recent IHC studies have shown that *mam* is a valid surrogate of micrometastatic disease (9).

In this study, using quantitative real-time RT-PCR data from the MIMS Trial, we did a rigorous statistical analysis of the relationship between relative levels of gene expression and the ability of individual molecular markers to detect micrometastatic breast cancer. The result of this analysis is a statistical validation of the concept that the most informative markers for detection of micrometastatic disease are those that are most highly expressed in metastatic disease. To further test this hypothesis, we developed an innovative microarray strategy to identify genes that are most likely to be informative for the detection of micrometastatic disease.

Materials and Methods

Expression values used for frequency distribution analysis. The design, enrollment criteria, tissue acquisition protocols, and determination of gene expression values for patients enrolled in the MIMS trial are described in a separate publication (8). *Pathology-positive ALN*. Subjects (145 of the 489) enrolled in the MIMS Trial had at least one ALN that was positive by H&E staining for the presence of metastatic cancer. This population of subjects is referred to as the H&E (+) subjects. *Pathology-negative ALN*. Subjects (344 of the 489) had no evidence of metastatic breast cancer in the ALN by routine H&E staining done at the participating site. This population of subjects is called the H&E (−) subjects. For frequency distribution analyses of H&E (+) or H&E (−) ALN, the lymph node that had the highest combined level of relative gene expression from a given patient was selected (data not shown). Normal cervical lymph nodes were obtained from patients with no evidence of malignancy undergoing elective carotid endarterectomy at the Medical University of South Carolina.

Artificial neural network development and generation of frequency distribution analyses. Frequency distribution analyses were done using an artificial neural network (ANN) developed by one of the authors (J.S.A.) following published guidelines (10). The guidelines consist of using bootstrapped cross-validation for automatic identification of feedforward ANN models, including both regression early stopping and optimal topology selection, whereas avoiding model overfitting. The ANN model is detailed in Eq. A, where w_1, w_2, b_1, and b_2 are variable vectors with the length equal to the number of hidden nodes in the optimal topology of the ANN model.

$$p(\Delta C_t) = (1 + e^{-w_2 \cdot h + b_2})^{-1}$$

$$h = \tanh(w_1 \cdot \Delta C_t) + b_1 \qquad \text{(A)}$$

Because the ANN can be expressed as an algebraic expression, its symbolic derivative, $dp(\Delta C_t)/d\Delta C_t$, can also be used directly to generate the frequency distribution of a given molecular marker. The derivative of the ANN is described in Eq. B. The m-files (in MATLAB code) implementing Eq. A and its symbolic derivative can be obtained from J.S.A. at almeidaj@musc.edu.

$$A = \tanh(W_1 \times \Delta C_t + b_1)$$

$$B = \exp(-\Sigma(W_2 \cdot A) - b_2)$$

$$C = \Sigma W_2 \cdot (1 - A^2) \cdot W_1$$

$$\frac{d[p(\Delta C_t)]}{(d\Delta C_t)} = \left(\frac{1}{1} + B\right)^2 \cdot C \cdot B \qquad \text{(B)}$$

Statistical analyses. Regression analyses were done using SAS version 9.0 Software (SAS Institute, Inc., Cary, NC).

RNA isolation for dilutional microarray analysis. For microarray analysis, we used a metastatic ALN in which *mam* was overexpressed at a level 5.3×10^7-fold higher than the mean expression in normal lymph nodes. In addition, four normal lymph nodes were used. Quality and quantification of RNA was assessed by an Agilent 2100 Bioanalyzer System (Agilent Technologies, Inc., Palo Alto, CA). RNA from the metastatic lymph node was diluted into a pool of normal lymph node RNA at ratios of 1:50, 1:2,500, and 1:125,000. For all of these conditions, expression values were obtained for a total of 22,283 gene transcripts spotted on an Affymetrix U133A array. Total cellular RNA was isolated as follows: ≤0.15 g of lymph node tissue was homogenized in 1 mL of RNA STAT-60 (TEL-TEST, Friendswood, TX) using a model 395 type 5 polytron (Dremel, Racine, WI). Total RNA isolation was done as per manufacturer's instructions up to the aqueous phase separation. Aqueous phase containing RNA was removed from organic phase and mixed with an equal volume of 70% ethanol. The sample was then loaded into an Rneasy Mini column (Qiagen, Valencia, CA) and purified according to the manufacturer's protocol. The RNA pellet was dissolved in 50 μL of RNase-free water.

GeneChip microarray analysis. Expression levels of 22,283 gene transcripts were determined on oligonucleotide microarrays using: (a) pooled RNA from four normal lymph nodes, (b) RNA from an ALN with a large breast cancer metastasis, and (c) RNA from an ALN with a large breast cancer metastasis diluted into pooled normal lymph node RNA at dilutions of 1:50, 1:2,500, and 1:125,000. Eight microgram of total RNA per sample was used for microarray analysis. First- and second-strand cDNA synthesis, double-stranded cDNA cleanup, biotin-labeled cRNA synthesis, cleanup, and fragmentation were done according to protocols in the Affymetrix GeneChip Expression Analysis technical manual (Affymetrix, Santa Clara, CA). Microarray analysis was done by the DNA Microarray and Bioinformatics Core Facility at the Medical University of South Carolina using U133A GeneChips (Affymetrix). Fluorescent images of hybridized microarrays were obtained by using a HP GeneArray scanner (Affymetrix). For normalization, the microarray office suite was used such that all fluorescence values were multiplied by a factor that resulted in a mean fluorescent score for all genes equal to 150.

Real-time reverse transcription-PCR validation of dilutional microarray analysis on frozen tissue samples. Twenty H&E (+) ALN, 40 control cervical lymph nodes, and 72 H&E (−)/PCR (+) ALN were used in this study. Frozen tissue specimens were obtained as part of the MIMS Trial, which was approved by the Institutional Review Board at the Medical University of South Carolina and all participating institutions. mRNA sequences of genes identified in this study were retrieved from the National Center for Biotechnology Information database. Intron-spanning primers were designed and tested in breast cancer cell lines MDA-MB-231 or SK-BR-3: *TFF1* forward 5′-AATGGCCACCATGGAGAACA-3′, reverse 5′-ACCACAATTCTGTCTTTCACGG-3′; *TFF3* forward 5′-TTTGACTCCAGGATCCCTGGAG-3′, reverse 5′-AGGTGCCTCAGAAGGTGCATTC-3′; *PRO1708* forward 5′-AAGAATGCCCTGTGCAGAAGAC-3′, reverse 5′-TTCTGTGCAGCATTTGGTGACT-3′; *Lipophilin B* forward 5′-ACGGATCAGATGTCCCTTCAG-3′, reverse 5′-TTGAAAGACAGTGGAAACCAGG-3′; *FBJ* forward 5′-CGTTGTGAAGACCATGACAGGA-3′, reverse 5′-TCCTTTCCCTTCGGATTCTCC-3′. Primers to the *mam* gene has been previously described (11, 12). cDNA was made from 5 μg of total RNA using 200 units of Moloney murine leukemia virus reverse transcriptase (Promega, Madison, WI) and 0.5 μg Oligo $(dT)_{12\text{-}16}$ in a reaction volume of 20 μL (10 minutes at 70°C, 50 minutes at 42°C, and 15 minutes at 70°C). Real-time RT-PCR analysis was done on a PE Biosystems Gene Amp 5700 Sequence Detection System (Foster City, CA). The standard reaction volume was 10 μL and contained

1× QuantiTect SYBR Green PCR Master Mix (Qiagen), 0.1 unit AmpErase UNG enzyme (PE Biosystems); 0.7 μL cDNA template; and 0.25 μmol/L of both forward and reverse primer. The initial step of PCR was 2 minutes at 50°C for AmpErase UNG activation, followed by a 15-minute hold at 95°C. Cycles (n = 40) consisted of a 15-second denaturation step at 95°C followed by a 1-minute annealing/extension step at 60°C. The final step was a 60°C incubation for 1 minute. All reactions were done in triplicate.

Real-time reverse transcription-PCR validation of dilutional microarray analysis on paraffin-embedded tissue samples. A 20- to 50-μm section was cut from nine H&E (+) ALN tissue blocks for mRNA extraction following the method of Specht et al. (13). An adjacent 5-μm section was cut for standard H&E staining and examined by a pathologist to confirm the presence or absence of metastatic breast cancer. Briefly, paraffin-embedded tissue sections were deparaffinized twice with 1 mL of xylene at 37°C or room temperature for 10 minutes. The pellet was subsequently washed with 1 mL of 100%, 90%, and 70% of ethanol and air-dried at room temperature for 2 hours. The pellet was resuspended in 200 μL of RNA lysis buffer [2% lauryl sulfate, 10 mmol/L Tris-HCl (pH 8.0), and 0.1 mmol/L EDTA] and 100 μg of proteinase K and incubated at 60°C for 16 hours. RNA was extracted using 1 mL of phenol/chloroform (5:1) solution (Sigma, St. Louis, MO). The aqueous layer containing RNA was transferred to a new 1.5-mL tube. Phenol/chloroform extraction was done a total of three times. RNA was precipitated with an equal volume of isopropanol, 0.1 volume of 3 mol/L sodium acetate, and 100 μg of glycogen at −20°C for 16 hours. After centrifugation at 12,000 rpm for 15 minutes (4°C), the RNA pellet was washed with 70% of ethanol and air-dried at room temperature for 2 hours. Finally, the pellet was dissolved in 12 μL of DEPC water. cDNA synthesis was done as described above with an exception that 500 ng of a panel of truncated gene-specific primers were used instead of oligo(dT)$_{12\text{-}16}$. Truncated gene-specific primers for reverse transcription were designed to correspond to the 5′-end of reverse primer designed for real-time PCR: *TFF1* 5′-ACCACAATTCTGTCTT-3′, *TFF3* 5′-AGGTGCCTCAGAAG-3′, *PRO1708* 5′-TTCTGTGCAGCAT-3′, *Lipophilin B* 5′-TTGAAAGACAGTGGAA-3′, *FBJ* 5′-TCCTTTCCCTTCGG-3′, and *mam* as described previously (14).

Results

Frequency distribution analyses and determination of mean gene expression in specific subsets of axillary lymph nodes. To understand the disparity in the ability of individual molecular markers to detect metastatic versus micrometastatic breast cancer, we first attempted to determine mean levels of gene expression in metastatic breast cancer tissue. For this purpose, we generated frequency distribution curves for the seven molecular markers in H&E (+) ALN used in the MIMS trial (8). We observed that for all seven molecular markers, the distribution of expression in the H&E (+) ALN was bimodal (Fig. 1, *solid trace*). In each instance, one population expressed a given marker at high levels, whereas the other population expressed the marker at low (or undetectable) levels. Because all seven genes were initially selected based on high expression levels in tissues containing breast cancer (11), we reasoned that the population expressing an individual molecular marker at high levels corresponds to ALN tissues containing metastatic breast cancer. Conversely, the population expressing an individual molecular marker at low levels corresponds to either (a) tissue containing no metastatic breast cancer (one of the limitations of the MIMS Trial study design is the potential for sampling error) or (b) tissue containing metastatic cancer that does not overexpress the particular molecular marker. Thus, we reasoned that for H&E (+) ALN, the best estimate for the true population mean corresponds to the peak of the population expressing the particular molecular marker at high levels (Fig. 1, *peak 1*).

To obtain an estimate of the mean gene expression levels in normal tissue, we analyzed 51 normal control lymph nodes. For all of the molecular markers, the distribution of expression in normal lymph nodes approximated a normal distribution (data not shown). The mean gene expression levels for each molecular marker in normal lymph nodes are indicated in Fig. 1 (*arrowhead*). Of note, the mean gene expression levels in normal lymph nodes closely approximated the mean levels of expression observed in the low expressing population of both the H&E (+), and H&E (−) samples. These results confirm that normal lymph nodes obtained from patients with no history of malignancy are appropriate controls for this study.

Relative levels of gene expression for individual molecular markers are correlated with apparent sensitivity for the detection of micrometastatic breast cancer. Using the mean expression values for metastatic breast cancer (Fig. 1, *peak 1*) and normal lymph nodes (Fig. 1, *arrowhead*), we were able to calculate the relative levels of gene expression (RLGE) for all seven markers using the $2^{\Delta\Delta Ct}$ method (ref. 15; Table 1). *Muc1* had the lowest RLGE value (3.6×10^2), whereas *mam* had the highest (1.9×10^6).

To determine whether RLGE values are correlated with the ability of a given marker to detect micrometastatic disease, we did a linear regression analysis. We observed that the correlation coefficient between log (RLGE) values and detection of micrometastatic disease in the H&E (−) population was good ($R^2 = 0.69$, $P = 0.0211$, F test; Fig. 2). This result provides statistical validation of the concept that the most informative markers for detection of micrometastatic disease are those that are most highly expressed in metastatic disease.

Innovative microarray strategy improves the ability to identify molecular markers that are informative for the detection of micrometastatic breast cancer. Based on the results outlined above, we hypothesized that informative molecular markers for detection of micrometastatic breast cancer could be rapidly identified by an innovative microarray strategy based on RNA dilution.

To test this hypothesis, we did a microarray analysis whereby RNA isolated from a highly metastatic (breast cancer) ALN was diluted into normal lymph node RNA as described in Materials and Methods. Candidate breast cancer – associated genes from this analysis were then selected based on the following criteria: (a) absence of expression in the pooled normal lymph nodes, (b) a fluorescence signal that was above 500 relative units for the undiluted breast cancer sample, and (c) a fluorescence signal that was present in the 1:50 dilution. The per cent of genes that met each respective criterion were 52%, 8.1%, and 52%. Median relative fluorescent value for all genes was 74. Seventy-one genes were identified by criteria a and b, whereas 34 genes were identified by criteria a, b, and c. The 34 genes were sorted by relative intensity of metastatic signal and the top 15 are listed in Table 2, along with genes that we have used in the past for molecular detection of micrometastatic disease (indicated in bold). Of note, of the 34 genes identified by criteria a, b, and c, only *mam* and *trefoil factor 1* (*TFF1*) had fluorescence signals above 1,000 fluorescent units in the 1:50 dilution. These results suggest that both the *mam* and *TFF1* genes may be informative molecular markers for the

detection of micrometastatic breast cancer. The gene with the highest relative intensity was *mam*, a result that is consistent with results from the MIMS Trial, where *mam* was noted to be the molecular marker that was most highly expressed in ALN containing metastatic breast cancer, as well as being the most informative marker for the detection of micrometastatic breast cancer (8).

A closer examination of the results of the microarray analyses in this study confirms limitations of a standard (undiluted) microarray approach to gene identification. The fluorescence signal for *mam* was 6,348 in the undiluted sample, 1,335 in the 1:50 dilution, 38 at the 1:2,500 dilution, and at background levels in the 1:125,000 dilution. However, based on real-time RT-PCR measurements, we determined that

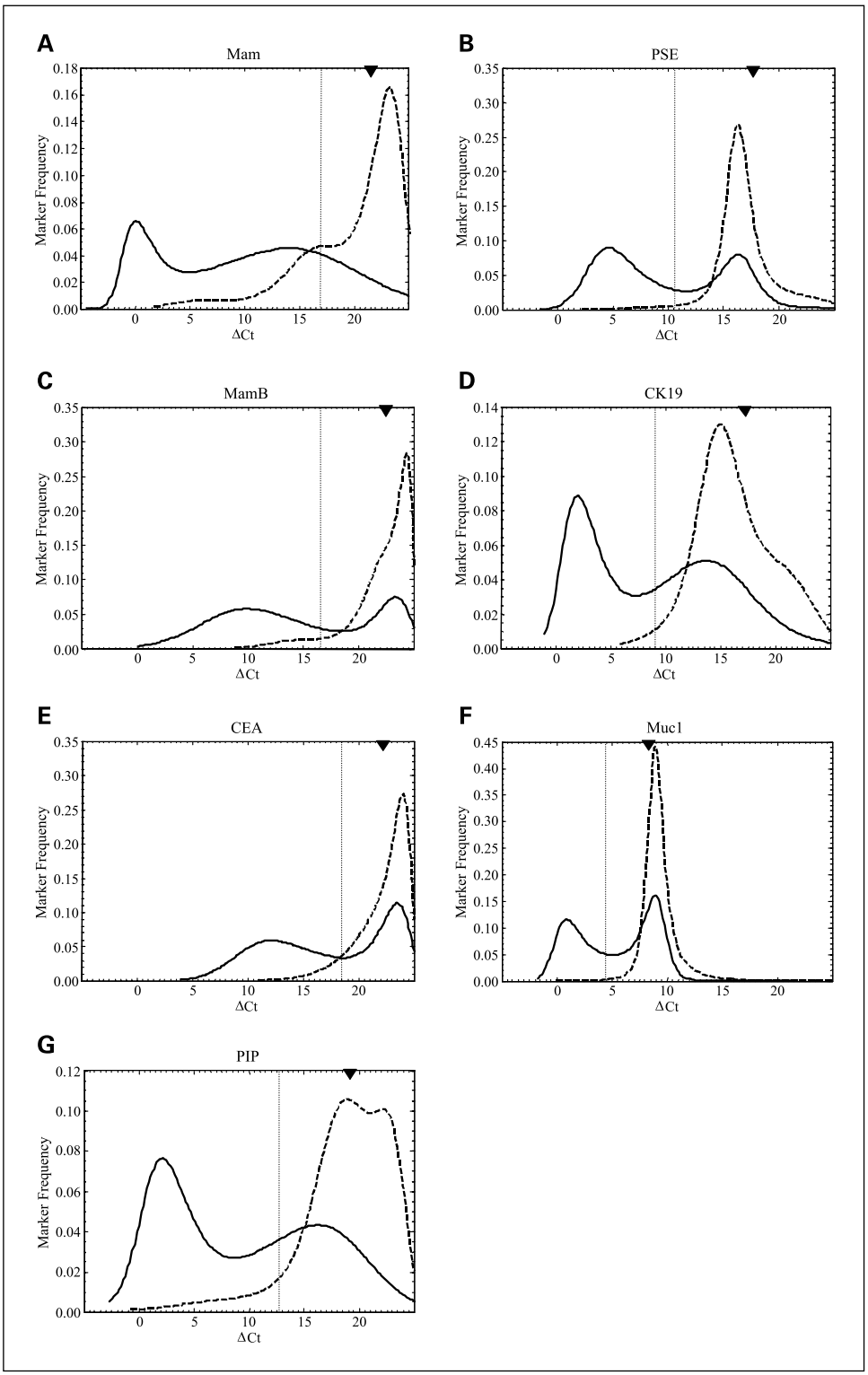

Fig. 1. Frequency distribution analyses of molecular markers expressed in axillary nodes of breast cancer patients. Frequency (or density) distribution analyses of the indicated genes were generated as described in Materials and Methods for H&E (+) (n = 145; *solid trace*) and H&E (−) (n = 344; *dashed trace*) samples. Vertical line corresponds to the threshold used in the MIMS Trial for marker positivity. Filled arrowhead (*top*), mean of normal control lymph nodes. *A, Mam*, mammaglobin. *B, PDEF* (*or PSE*), prostate-derived Ets transcriptional factor. *C, mamB*, mammaglobin B. *D, CK19*, cytokeratin 19. *E, CEA*, carcinoembryoinc antigen. *F, muc1*, mucin 1. *G, PIP*, prolactin-inducible protein. The comparison between experimental data and the ANN model for the cumulative distribution of values for all markers is provided in a supplement to this article.

Table 1. Population-based statistical analysis of real-time RT-PCR data generated from the MIMS Trial

Gene	ΔC_t of peak 1*	ΔC_t of mean normal†	$\Delta\Delta C_t$‡	AE§	RLGE‖	Log [RLGE]
mam	0.02	21.50	21.48	0.96	1.89e+06	6.28
PIP	2.09	19.19	17.10	1.00	1.40e+05	5.15
mamB	9.90	22.49	12.59	0.99	5.79e+03	3.76
CEA	12.09	22.23	10.14	0.94	8.27e+02	2.92
PDEF	4.57	17.64	13.07	1.00	8.60e+03	3.94
CK19	1.94	17.25	15.31	0.66	2.35e+03	3.37
muc1	0.85	9.35	8.51	1.00	3.64e+02	2.56

*Corresponds to high expressing peak observed in H&E (+) population.
† Corresponds to mean expression level observed in normal cervical control population reported from ref. (8).
‡ Difference between peak1 and mean in normal control lymph nodes.
§Amplification efficiency of respective gene.
‖Determined from the equation $(1 + AE)^{-\Delta\Delta C_t}$.

mam was overexpressed in this particular ALN at a level 5.3×10^7-fold higher than the mean expression in normal lymph nodes. We can conclude, therefore, that without dilution, *mam* is in the saturated range, whereas at the 1:50 and 1:2,500 dilutions, *mam* is at the upper and lower end of the linear detection range, respectively. Based on these findings, we conclude that for highly expressed genes, the hybridization signal at the undiluted level is likely to become saturated and is unlikely to be proportional to gene copy number.

Real-time reverse transcription-PCR confirms that TFF1 is highly expressed in axillary lymph nodes containing metastatic breast cancer. To determine whether *TFF1* and/or other markers identified by dilutional microarray analysis were potentially useful for the detection of metastatic and/or micrometastatic breast cancer, we selected the five most highly expressed genes (*TFF1*, *TFF3*, *PRO1708*, *Lipophilin B*, and *FBJ*) for further analyses. Primers (see Materials and Methods) were designed and validated using cDNA prepared from the breast cancer cell lines MDA-MB-361 and/or MDA-MB-231. Gene expression levels were determined in ALNs containing metastatic breast cancer ($n = 20$), as well as in control lymph nodes ($n = 8$; $n = 40$ for TFF1; Fig. 3). In control lymph nodes, ΔC_t values for *TFF1* ranged from 16.3 to 25.2 (mean, 22.4 ± 2.1) providing evidence that this gene is expressed poorly in normal tissue. In contrast, the ΔC_t values for *TFF1* in pathology-positive lymph nodes ranged from −2.6 to 23.7 (mean, 12.7 ± 8.1). Using a threshold of three SDs beyond the mean of control lymph nodes, we observed that at least one marker was overexpressed in 17 of 20 (85%) metastatic lymph nodes. *TFF1* was overexpressed in 10 of 20 (50%) metastatic ALN, providing evidence that this gene may be an informative marker for detection of metastatic disease. Consistent with previous studies (8, 11), *mam* was overexpressed in 14 of 20 (70%) samples. Of note, of the three samples that were marker positive but negative for *mam*, one was positive for *TFF1* and two were positive for *FBJ*. Of the remaining candidate genes, *lipophilin B* seemed a potentially informative marker and was overexpressed in 9 of 20 (45%) specimens. The sensitivities of the other markers tested were as follows: *FBJ* 6 of 20 (30%), *TFF3* 4 of 20 (20%), and *PRO1708* 1 of 20 (5%). Although *TFF1* was not as sensitive as *mam*, the level of overexpression of *TFF1* was comparable to the level of overexpression observed with *mam* (overexpression in individual ALN of up to 1.1×10^5 and 1.0×10^6, respectively). Because *TFF1* was the only gene whose level of detection for metastatic disease was not significantly different from *mam* at a $P < 0.05$ (χ^2 test; data not shown), we chose to analyze this gene in further detail.

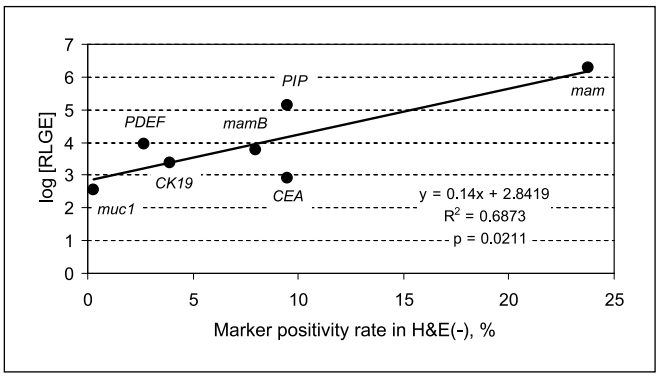

Fig. 2. RLGE are correlated with apparent sensitivity for the detection of micrometastatic disease in H&E (−) population. Marker positivity rates in H&E (−) patients ($n = 344$) were determined using the thresholds described in the text. A particular lymph node was determined to be positive by molecular marker analysis if the ΔC_t value of one or more marker(s) was above the respective threshold. A patient was determined to be marker positive if one or more lymph nodes were marker positive. Marker positivity rates (*x-axis*) reflect the percent of total. Data points correspond to the indicated gene. Regression line through the data points was generated using Microsoft Excel software. Relative overexpression values were calculated as described in Table 2. Correlation coefficients and *P*s were obtained using SAS version 9.0 Software.

TFF1 is highly expressed in paraffin-embedded lymph nodes containing metastatic breast cancer. One of the design limitations of the analysis outlined above is the potential for sampling error. To minimize this potential, we determined expression levels of *TFF1* gene in nine archived paraffin-embedded tissue samples containing metastatic breast cancer. Sections adjacent to those used for molecular analysis were analyzed by H&E to confirm the presence of metastatic breast cancer. In addition to *TFF1*, we also determined expression levels of four genes that have very high diagnostic accuracy for breast cancer: *mam*, *PIP*, *PDEF*, and *EpCAM* (refs. 11, 16;

Table 2. Gene expression results from dilutional microarray analysis

Rank*	Descriptions	Metastasis	1:50	1:2,500	1:125,000	N-mix
1	**mammaglobin 1 (*MGB1*)**	**6,348**	**1,335**	**38**	**11**	**9**
2	trefoil factor 1, breast cancer, estrogen-inducible sequence expressed in (*TFF1*)	5,443	1,013	90	51	47
3	trefoil factor 3, intestinal (*TFF3*)	4,243	195	92	87	90
4	**keratin 19 (*KRT19*)**	**2,991**	**150**	**9**	**2**	**2**
5	*PRO1708*	2,941	237	11	12	12
6	lipophilin B, uteroglobin family member, prostatein-like (*LPHB*)	2,819	132	8	4	6
7	S100 calcium-binding protein P (*S100P*)	1,852	88	4	17	3
8	v-fos FBJ murine osteosarcoma viral oncogene homologue, clone *MGC:11074*	1,783	220	41	34	27
9	serine (or cysteine) proteinase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 3 (*SERPINA3*)	1,514	167	92	119	82
10	Estrogen receptor 1 (*ESR1*)	1,483	191	82	99	111
*	**prolactin-induced protein (*PIP*)**	**1,376**	**70**	**17**	**41**	**49**
11	Human gastrointestinal tumor-associated antigen GA733-1 protein, clone *05516*	1,292	88	30	23	28
12	v-jun avian sarcoma virus 17 oncogene homologue	1,213	222	135	185	163
13	*Prostaglandin D synthase gene*	1,118	717	587	903	845
14	*cytokeratin 8*	917	55	67	54	45
15	alpha-fetoprotein (*AFP*)	904	55	28	24	31
*	**prostate epithelium – specific Ets transcription factor (*PDEF*)**	**424**	**45**	**17**	**19**	**13**
*	**tumor-associated calcium signal transducer 1 (*TACSTD1*; *EpCAM*)**	**346**	**18**	**14**	**16**	**22**
*	**carcinoembryonic antigen – related cell adhesion molecule 5 (*CEACAM5*), mRNA**	**328**	**88**	**54**	**88**	**116**
*	**Mammaglobin 2 (*MGB2*), mRNA**	**251**	**2**	**2**	**1**	**1**

NOTE: Bolded entries correspond to genes used in the MIMS Trial. Genes that lack a numeric value (*) in the rank category failed to meet the following selection criteria: signal present in 1:50 dilution (*PIP, PDEF, EpCam, CEA,* and *mamB*); signal >500 relative units in metastatic tissue (*PDEF*, *EpCam*, *CEA*, and *mamB*).
*Rank value was based on relative intensity of undiluted metastatic signal.

Fig. 4A). We observed that at least one marker was overexpressed in nine of nine (100%) of the samples. *TFF1* and *mam* were each overexpressed in six of nine samples (67%) at levels up to 1.0×10^4- and 2.9×10^5-fold above their respective thresholds. For *PDEF* and *EpCAM*, overexpression was detected in eight of nine samples (89%), a rate that is higher than the rate observed for *TFF1* or *mam*. However, although the *rates* of overexpression of *PDEF* and *EpCAM* were higher than *TFF1*, their *levels* of overexpression were not. For example, the mean value of gene overexpression for both *mam* and *TFF1* were greater than *EpCAM* (not shown, but see Fig. 4).

TFF1 is an informative marker for the detection of micrometastatic breast cancer. The results described above confirm that *TFF1* is an informative marker for the detection of metastatic breast cancer. Furthermore, the degree of overexpression in ALN containing metastatic breast cancer suggests that *TFF1* is also a potentially informative marker for the detection of micrometastatic breast cancer. To test this hypothesis, we did real-time RT-PCR analyses on 72 frozen, pathology-negative ALN from the MIMS trial (Fig. 4B). Our prior analysis of these 72 ALNs (8) indicated that for each node, at least one of seven molecular markers was overexpressed. In addition to *TFF1* and *EpCAM*, we reanalyzed samples for *mam*, *PIP*, and *PDEF*. We observed that 46 of 72 samples were positive for one ($n = 32$) or more molecular markers. Out of all the 14 samples that were positive for two or more markers, *mam* was positive in 14 and *TFF1* in 11 cases. Overall, *mam* was overexpressed in 39 samples (54%), *TFF1* in 15 samples (21%), *PIP* in nine samples (13%), *PDEF* in one sample (1.4%), and *EpCAM* in one sample (1.4%). Of the seven samples that were molecular marker positive but *mam* negative, four were positive for *TFF1*, whereas four were positive for *PIP*. These results suggest that *TFF1* is an informative marker for the detection of micrometastatic breast cancer.

Discussion

We have recently completed an interim analysis of the MIMS, a multi-institutional prospective cohort study designed to determine the clinical relevance of micrometastatic breast cancer in sentinel and ALNs of breast cancer patients (8). The interim results reveal that overexpression of breast cancer–associated genes in breast cancer subjects with pathology-negative ALNs is correlated with traditional predictors of poor prognosis in breast cancer, providing strong evidence that molecular markers are valid surrogates for micrometastatic

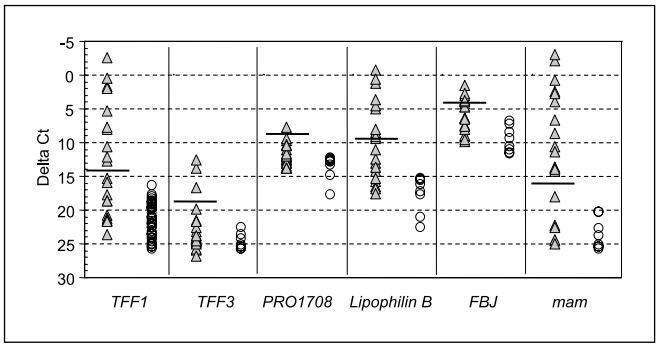

Fig. 3. Real-time RT-PCR analysis of pathology-positive ALNs and normal control lymph nodes. Real-time RT-PCR analysis of pathology-positive lymph nodes (n = 20; *filled triangles*) and normal control lymph nodes (n = 8; n = 40 for TFF1; *empty circles*) was done as described in Materials and Methods using primer pairs for the indicated genes. C_t values for each gene were determined from triplicate reactions. ΔC_t values were obtained by subtracting the mean C_t value of β_2-microglobin from the mean C_t value of each respective gene. Horizontal lines indicate ΔC_t threshold values (3 SD below average). Average ΔC_t (±SD) and ΔC_t threshold for each gene are as follows: *TFF1* 21.80 (±2.53), 14.22; *TFF3* 24.74 (±1.14), 21.32; *PRO1708* 13.6 (±8.01), 8.01; *Lipophilin B* 17.64 (±2.76), 9.37; *FBJ* 9.6 (±1.92), 3.84; *mam* 21.53 (±1.53), 16.9.

breast cancer. The validity of such a molecular approach is also underscored by recent IHC studies done by our laboratory and others. Ouellette et al. showed a 79% concordance between mammaglobin (*mam* and *mamB*) expression and IHC (9). We have done a similar side-by-side study whereby H&E-negative lymph nodes (n = 9) were subjected to both IHC (AE1-AE3 cytokeratin-based stain) and molecular analysis.[5] The concordance between IHC and molecular analysis was six of nine (67%) [3(IHC−/PCR−), 3(IHC+/PCR+), 3(IHC−/PCR+)]. Of note, for IHC, a 5-μm section was used, whereas 20- to 50-μm sections were used for PCR. Thus, based on the size of the paraffin sections, one would expect the rate of positive results to be higher for molecular analyses compared with IHC (decreased sampling error). Overall, these IHC analyses support the concept that molecular markers are a valid surrogate for micrometastatic disease.

Given this as a background, we did a statistical analysis of the molecular data from the MIMS Trial to determine predictors of informative markers for the detection of micrometastatic disease. Specifically, the data from the MIMS Trial provides a unique opportunity to explore the relationship between relative levels of gene expression in metastatic breast cancer and the ability to detect micrometastatic breast cancer. The large sample size (n = 489 breast cancer subjects) and quantitative data generated in the MIMS Trial have made it possible for the first time to develop artificial neural networks capable of generating statistically meaningful, accurate frequency distribution analyses of molecular markers known to be associated with breast cancer. In this article, we generated frequency distribution analyses from three different populations: subjects with pathology-positive ALN, subjects with pathology-negative ALN, and control subjects with no evidence of malignancy. The frequency distribution analyses seem internally consistent; the pathology-positive and pathology-negative populations are bimodal, and the peaks representing the low-expressing populations correspond to the mean of the control populations (Fig. 1). Furthermore, based on these frequency distribution analyses, we were able to calculate RLGE values for each molecular marker, a surrogate for the degree of gene overexpression observed in metastatic breast cancer. Finally, we did logistic regression analyses that seem to confirm the hypothesis that relative levels of gene overexpression in metastatic tissue are associated with apparent sensitivity for detection of micrometastatic disease.

Based on this hypothesis, we proceeded with the development of a novel microarray strategy for the rapid identification of molecular markers that are informative for the detection of micrometastic breast cancer. Microarray analysis has proven to be a powerful tool for studying the mRNA expression profiles of normal and neoplastic tissues. However, the ability of this technology to identify informative molecular markers for the detection of micrometastatic disease has been limited. One major limitation of microarray analysis is that it is only semiquantitative. Thus, it is often difficult to determine which of several hundred candidate genes are likely to be most informative for detection of micrometastatic disease. We reasoned that the identification of informative markers for the detection of micrometastatic disease can be simplified by dilution of metastatic tissue (or RNA) into an excess of normal tissue (or RNA).

For our analyses, RNA from a metastatic lymph node was extracted and serially diluted into a pool of normal lymph node RNA at ratios of 1:50, 1:2,500, and 1:125,000. By virtue of this dilution strategy, we were able to rapidly identify those genes

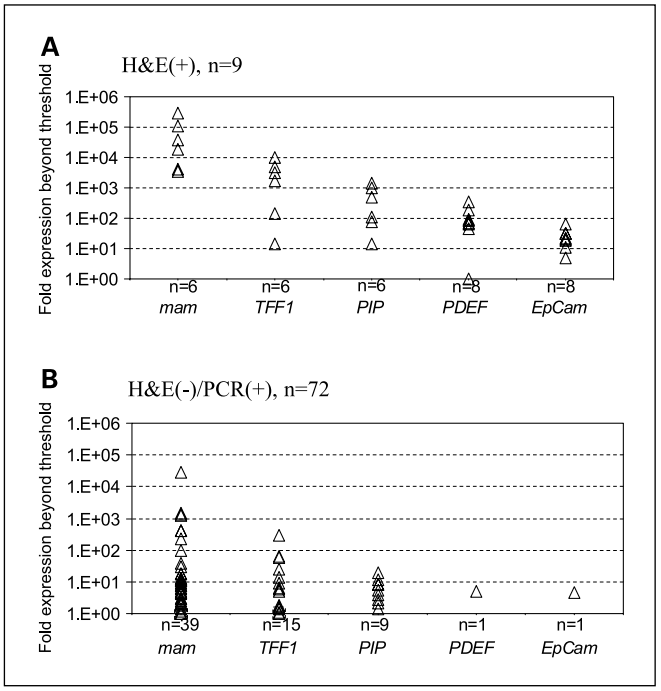

Fig. 4. Real-time RT-PCR analysis of archived paraffin-embedded pathology-positive and frozen pathology-negative marker-positive ALNs from breast cancer patients. Real-time RT-PCR analysis was done on (*A*) nine archived paraffin embedded pathology-positive [H&E(+)] and (*B*) 72 frozen pathology-negative marker-positive [H&E(−)/PCR(+)] ALNs from breast cancer patients. H&E(−)/PCR(+) samples were chosen based on the MIMS Trial results with *mam*, *PIP*, *mamB*, *CEA*, *CK19*, *muc1*, and *PDEF* genes (8). RNA extraction and RT-PCR were done as described in Materials and Methods. Gene overexpression is presented as fold expression beyond threshold. The number of samples positive for the particular gene in each data set is indicated as "*n*."

[5] K. Mikhitarian et al., unpublished results.

that were overexpressed at the highest level in metastatic tissue. Candidate marker genes were chosen based on three selection criteria and validated by real-time RT-PCR using experimental and control lymph nodes. Of the 22,283 genes contained on the Affymetrix U133-A chip, 34 genes met the three selection criteria. The most highly overexpressed gene was *mam*, a result that was consistent with our previous studies. Besides *mam*, only *trefoil factor 1* (*TFF1*) had a signal in the 1:50 dilution that was above 1,000 relative fluorescent units. In fact, the intensity signals of *TFF1* were similar to those of *mam*. The results of our analyses suggest that of the dilutions used, the 1:50 was the most informative for identification of genes useful for detection of micrometastatic disease. Real-time RT-PCR analyses of pathology-negative ALN nodes that had shown cancer-associated gene overexpression in the MIMS study ($n = 72$) confirm that of all the markers tested, *mam* and *TFF1* have the highest apparent sensitivity for detection of micrometastatic breast cancer (Fig. 4B).

TFF1, also known as *gastrointestinal trefoil protein pS2*, *breast cancer estrogen-inducible sequence* (*BCEI*), *pNR-2*, and *Md2*, is a secretory polypeptide encoded by a gene in chromosome 21q22.3. *TFF1* is involved in the formation of mucus and is highly expressed in stomach epithelium. Interestingly, *TFF1* expression has also been found in the regeneration stage of ulcerative and inflammatory gastrointestinal disorders and in various human carcinomas including breast carcinoma (17). In breast cancer, *TFF1* is regulated by estrogen and there is evidence that it can be used as a surrogate indicator for the response to anti–hormonal therapy and more favorable outcome. For example, Gillesby et al. showed that *TFF1* mRNA levels in breast cancer were positively associated with both estrogen receptor and progesterone receptor status and that *TFF1* was primarily expressed in small ($T = 2.0$ cm) but well-differentiated tumors (grade 1 and 2; ref. 18). In support of a prognostic role of *TFF1* in breast cancer, Thompson et al. reported that the combination of lymph node status and *TFF1* expression (analyzed by Northern blot) discriminated patients with good prognosis (node negative, *TFF1* positive) and patients with poor prognosis (node positive, *TFF1* negative; ref. 19). However, other studies suggest that high levels of *TFF1* expression may promote cancer cell invasion (particularly in interval cancers; ref. 20) and may be involved in establishing distant metastasis (21). To our knowledge, only one research group (van't Veer et al.) has studied *TFF-1* and *TFF-3* (also called *p1B*) as diagnostic markers for detection of metastatic breast cancer in axillary nodes (22) and in peripheral blood (23). Contrary to our results, their data indicate that *TFF-3* is superior to *TFF-1* for detection of metastatic disease. Although we suspect that the discrepancy is due to selection of a limited set of control samples and/or threshold values that were too high or too low, we cannot rule out the possibility that selection of primer sequences may play a role.

In conclusion, we have been able to provide a meaningful statistical evaluation of the concept that relative levels of gene expression/overexpression are correlated with the ability to detect micrometastatic disease. Furthermore, we have used this information to design an innovative microarray strategy for the rapid identification of marker genes that can be used for the molecular detection of micrometastatic cancer. The microarray analyses did confirm that *mam* is one of the most valuable molecular markers for the detection of micrometastatic breast cancer and have identified *TFF1* as another highly informative marker. These results underscore the importance of relative gene expression levels in evaluation of candidate molecular markers for molecular detection of cancer.

Acknowledgments

We thank Victor Fresco and the DNA Microarray and Bioinformatics Core Facility at the Medical University of South Carolina. We thank Margaret Romano and the Hollings Cancer Center Tissue Procurement/Tumor Bank.

References

1. Goldhirsch A, Glick JH, Gelber RD, Coates AS, Senn HJ. Meeting highlights: International Consensus Panel on the Treatment of Primary Breast Cancer. Seventh International Conference on Adjuvant Therapy of Primary Breast Cancer. J Clin Oncol 2001;19:3817–27.
2. Woo CS, Silberman H, Nakamura SK, et al. Lymph node status combined with lymphovascular invasion creates a more powerful tool for predicting outcome in patients with invasive breast cancer. Am J Surg 2002;184:337–40.
3. Cummings MC, Walsh MD, Hohn BG, et al. Occult axillary lymph node metastases in breast cancer do matter: results of 10-year survival analysis. Am J Surg Pathol 2002;26:1286–95.
4. de Mascarel I, Bonichon F, Coindre JM, Trojani M. Prognostic significance of breast cancer axillary lymph node micrometastases assessed by two special techniques: reevaluation with longer follow-up. Br J Cancer 1992;66:523–7.
5. McGuckin MA, Cummings MC, Walsh MD, et al. Occult axillary node metastases in breast cancer: their detection and prognostic significance. Br J Cancer 1996;73:88–95.
6. Sedmak DD, Meineke TA, Knechtges DS. Detection of metastatic breast carcinoma with monoclonal antibodies to cytokeratins. Arch Pathol Lab Med 1989;113:786–9.
7. Gardner BFJ. Are positive axillary nodes in breast cancer markers for incurable disease? Ann Surg 1993;218:270–7.
8. Gillanders WE, Mikhitarian K, Hebert R, et al. Molecular detection of micrometastatic breast cancer in histopathology-negative axillary lymph nodes correlates with traditional predictors of prognosis: an interim analysis of a prospective multi-institutional cohort study. Ann Surg 2004;239:828–37; discussion 837–40.
9. Ouellette RJ, Richard D, Maicas E. RT-PCR for mammaglobin genes, MGB1 and MGB2, identifies breast cancer micrometastases in sentinel lymph nodes. Am J Clin Pathol 2004;121:637–43.
10. Almeida JS. Predictive non-linear modeling of complex data by artificial neural networks. Curr Opin Biotechnol 2002;13:72–6.
11. Mitas M, Mikhitarian K, Walters C, et al. Quantitative real-time RT-PCR detection of breast cancer micrometastasis using a multigene marker panel. Int J Cancer 2001;93:162–71.
12. Mitas M, Mikhitarian K, Hoover L, et al. Prostate-specific Ets (PSE) factor: a novel marker for detection of metastatic breast cancer in axillary lymph nodes. Br J Cancer 2002;86:899–904.
13. Specht K, Richter T, Muller U, et al. Quantitative gene expression analysis in microdissected archival formalin-fixed and paraffin-embedded tumor tissue. Am J Pathol 2001;158:419–29.
14. Mikhitarian K, Reott S, Hoover L, et al. Enhanced detection of RNA from paraffin-embedded tissue using a panel of truncated gene-specific primers for reverse transcription. BioTechniques 2004;36:474–8.
15. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the $2(-\Delta\Delta C(T))$ method. Methods 2001;25:402–8.
16. Osta WA, Chen Y, Mikhitarian K, et al. EpCAM is overexpressed in breast cancer and is a potential target for breast cancer gene therapy. Cancer Res 2004;64:5818–24.
17. Ribieras S, Tomasetto C, Rio MC. The pS2/TFF1 trefoil factor, from basic research to clinical applications. Biochim Biophys Acta 1998;1378:F61–77.
18. Gillesby BE, Zacharewski TR. pS2 (TFF1) levels in human breast cancer tumor samples: correlation with clinical and histological prognostic markers. Breast Cancer Res Treat 1999;56:253–65.
19. Thompson AM, Elton RA, Hawkins RA, Chetty U, Steel CM. PS2 mRNA expression adds prognostic information to node status for 6-year survival in breast cancer. Br J Cancer 1998;77:492–6.
20. Crosier M, Scott D, Wilson RG, et al. High expression of the trefoil protein TFF1 in interval breast cancers. Am J Pathol 2001;159:215–21.
21. Prest SJ, May FE, Westley BR. The estrogen-regulated protein, TFF1, stimulates migration of human breast cancer cells. FASEB J 2002;16:592–4.
22. Weigelt B, Verduijn P, Bosma AJ, et al. Detection of metastases in sentinel lymph nodes of breast cancer patients by multiple mRNA markers. Br J Cancer 2004;90:1531–7.
23. Bosma AJ, Weigelt B, Lambrechts AC, et al. Detection of circulating breast tumor cells by differential expression of marker genes. Clin Cancer Res 2002;8:1871–7.

Clinical Cancer Research

An Innovative Microarray Strategy Identities Informative Molecular Markers for the Detection of Micrometastatic Breast Cancer

Kaidi Mikhitarian, William E. Gillanders, Jonas S. Almeida, et al.

Clin Cancer Res 2005;11:3697-3704.