An Innovative Microarray Strategy Identifies 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 Mitas1

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 micrometastatic breast cancer (P < 0.05). These data provide 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) confirms 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 breast 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 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
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.

\[
p(\Delta C_i) = \left(1 + e^{-w_1 \cdot \Delta C_i + b_1}\right)^{-1}
\]

\[h = \tanh(w_1 \cdot \Delta C_i + b_1)
\]

Because the ANN can be expressed as an algebraic expression, its symbolic derivative, \( d[p(\Delta C_i)]/d\Delta C_i \), 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 almeida@musc.edu.

\[
A = \tanh(W_1 \times \Delta C_i + b_1)
\]

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

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

**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 \( \times 10^{-7} \)-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: \( \pm 0.15 \) 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 \( \mu \)L 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. Intronspanning primers were designed and tested in breast cancer cell lines MDA-MB-231 or SK-BR-3: TFF1 forward 5′-AATGGCCACCATGGA-CAACA-3′, reverse 5′-ACCAATTCGGTCTTACCAGG-3′; TFF3 forward 5′-TGGATCCACCATGGA-CAACA-3′, reverse 5′-ACCAATTCGGTCTTACCAGG-3′; reverse 5′-AGGTGGCTCA-GAAGGTCACTACGTCACTG-3′; PRO1708 forward 5′-AAGAATCCGCTGTCGTA-GAGAGC-3′, reverse 5′-TTCATGCACTTATGTCGTA-GAGAGC-3′, reverse 5′-TTCATGCACTTATGTCGTA-GAGAGC-3′, reverse 5′-TTCATGCACTTATGTCGTA-GAGAGC-3′, reverse 5′-TTCATGCACTTATGTCGTA-GAGAGC-3′, reverse 5′-TTCATGCACTTATGTCGTA-GAGAGC-3′, reverse 5′-TTCATGCACTTATGTCGTA-GAGAGC-3′. Primers to the mam gene has been previously described (11, 12). cDNA was made from 5 \( \mu \)g of total RNA using 200 units of Moloney murine leukemia virus reverse transcriptase (Promega, Madison, WI) and 0.5 \( \mu \)g Oligo (dT)21,16 in a reaction volume of 20 \( \mu \)L (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 \( \mu \)L and contained
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. All reactions were done in triplicate.

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 in the MIMS trial. 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 the peak of the population expressing the particular molecular marker at high levels (Fig. 1, peak 1).

Statistical Analysis of Breast Cancer–Associated Genes

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 breast cancer in the H&E (+) population was good (R² = 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

![Figure 1](https://www.aacrjournals.org/clin-cancer-research/article-pdf/11/10/3700/3367103/clin-cancer-research-05015-3700.pdf)
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, ΔCt 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 ΔCt 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 × 10^7 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.

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

<table>
<thead>
<tr>
<th>Gene</th>
<th>∆Ct, of peak 1</th>
<th>∆Ct, of mean</th>
<th>∆ΔCt, i</th>
<th>AE</th>
<th>RLGE</th>
<th>Log [RLGE]</th>
</tr>
</thead>
<tbody>
<tr>
<td>mam</td>
<td>0.02</td>
<td>21.50</td>
<td>21.48</td>
<td>0.96</td>
<td>1.89e+06</td>
<td>6.28</td>
</tr>
<tr>
<td>PIP</td>
<td>2.09</td>
<td>19.19</td>
<td>17.10</td>
<td>1.00</td>
<td>1.40e+05</td>
<td>5.15</td>
</tr>
<tr>
<td>mamB</td>
<td>9.90</td>
<td>22.49</td>
<td>12.59</td>
<td>0.99</td>
<td>5.79e+03</td>
<td>3.76</td>
</tr>
<tr>
<td>CEA</td>
<td>12.09</td>
<td>22.23</td>
<td>10.14</td>
<td>0.94</td>
<td>8.27e+02</td>
<td>2.92</td>
</tr>
<tr>
<td>PDEF</td>
<td>4.57</td>
<td>17.64</td>
<td>13.07</td>
<td>1.00</td>
<td>8.60e+03</td>
<td>3.94</td>
</tr>
<tr>
<td>CK19</td>
<td>1.94</td>
<td>17.25</td>
<td>15.31</td>
<td>0.66</td>
<td>2.35e+03</td>
<td>3.37</td>
</tr>
<tr>
<td>muc1</td>
<td>0.85</td>
<td>9.35</td>
<td>8.51</td>
<td>1.00</td>
<td>3.64e+02</td>
<td>2.56</td>
</tr>
</tbody>
</table>

*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 peak 1 and mean in normal control lymph nodes.

§Amplification efficiency of respective gene.

**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 ∆Ct, 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 corresponding 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 Ps were obtained using SAS version 9.0 Software.
addition to at least one of seven molecular markers was overexpressed. In prior analysis of these 72 ALNs (8) indicated that for each node, pathology-negative ALN from the MIMS trial (Fig. 4B). Our hypothesis, we did real-time RT-PCR analyses on 72 frozen, detection of micrometastatic breast cancer. To test this rate observed for in eight of nine samples (89%), a rate that is higher than the PDEF thresholds. For their levels value of gene overexpression for both mam TFF1 and EpCAM were each overexpressed in six of nine samples (67%) at levels were over 1.0 x 10^5- and 2.9 x 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

![Imaging, Diagnosis, Prognosis](https://www.aacrjournals.org)

**Table 2. Gene expression results from dilutional microarray analysis**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Descriptions</th>
<th>Metastasis</th>
<th>1:50</th>
<th>1:2,500</th>
<th>1:125,000</th>
<th>N-mix</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>mammaglobin 1 (MGB1)</td>
<td>6,348</td>
<td>1,335</td>
<td>38</td>
<td>11</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>trefoil factor 1, breast cancer, estrogen-inducible sequence expressed in (TFF1)</td>
<td>5,443</td>
<td>1,013</td>
<td>90</td>
<td>51</td>
<td>47</td>
</tr>
<tr>
<td>3</td>
<td>trefoil factor 3, intestinal (TFF3)</td>
<td>4,243</td>
<td>195</td>
<td>92</td>
<td>87</td>
<td>90</td>
</tr>
<tr>
<td>4</td>
<td>keratin 19 (KRT19)</td>
<td>2,991</td>
<td>150</td>
<td>9</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>PRO1708</td>
<td>2,941</td>
<td>237</td>
<td>11</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>6</td>
<td>lipophilin B, uteroglobin family member, prostate-like (LPHB)</td>
<td>2,819</td>
<td>132</td>
<td>8</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>S100 calcium-binding protein P (S100P)</td>
<td>1,852</td>
<td>88</td>
<td>4</td>
<td>17</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>v-fos FBJ murine osteosarcoma viral oncogene homologue, clone MGC:11074</td>
<td>1,783</td>
<td>220</td>
<td>41</td>
<td>34</td>
<td>27</td>
</tr>
<tr>
<td>9</td>
<td>senne (or cysteine) proteinase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 3 (SERPINA3)</td>
<td>1,514</td>
<td>167</td>
<td>92</td>
<td>119</td>
<td>82</td>
</tr>
<tr>
<td>10</td>
<td>Estrogen receptor 1 (ESR1)</td>
<td>1,483</td>
<td>191</td>
<td>82</td>
<td>99</td>
<td>111</td>
</tr>
<tr>
<td>11</td>
<td>mammaglobin 2 (MGB2), mRNA</td>
<td>1,376</td>
<td>70</td>
<td>17</td>
<td>41</td>
<td>49</td>
</tr>
<tr>
<td>12</td>
<td>v-jun avian sarcoma virus 17 oncogene homologue</td>
<td>1,292</td>
<td>88</td>
<td>30</td>
<td>23</td>
<td>28</td>
</tr>
<tr>
<td>13</td>
<td>Prostaglandin D synthase gene</td>
<td>1,118</td>
<td>717</td>
<td>587</td>
<td>903</td>
<td>845</td>
</tr>
<tr>
<td>14</td>
<td>cytokeratin 8</td>
<td>917</td>
<td>55</td>
<td>67</td>
<td>54</td>
<td>45</td>
</tr>
<tr>
<td>15</td>
<td>alpha-fetoprotein (AFP)</td>
<td>904</td>
<td>55</td>
<td>28</td>
<td>24</td>
<td>31</td>
</tr>
<tr>
<td>16</td>
<td>prostate epithelium–specific Ets transcription factor (PDEF)</td>
<td>1,196</td>
<td>45</td>
<td>17</td>
<td>19</td>
<td>13</td>
</tr>
<tr>
<td>17</td>
<td>tumor-associated calcium signal transducer 1 (TACSTD1, EpCAM)</td>
<td>346</td>
<td>45</td>
<td>14</td>
<td>16</td>
<td>22</td>
</tr>
<tr>
<td>18</td>
<td>carcinoembryonic antigen–related cell adhesion molecule 5 (CEACAM5), mRNA</td>
<td>328</td>
<td>88</td>
<td>54</td>
<td>88</td>
<td>116</td>
</tr>
<tr>
<td>19</td>
<td>Mammaglobin 2 (MGB2), mRNA</td>
<td>251</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

**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 x 10^5- and 2.9 x 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).
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.

5 K. Mikhailian 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 MLMS 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 p18) 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


3704 www.aacrjournals.org Clin Cancer Res 2005;11(10) May 15, 2005
Downloaded from clincanceres.aacrjournals.org on April 14, 2017. © 2005 American Association for Cancer Research.
An Innovative Microarray Strategy Identities Informative Molecular Markers for the Detection of Micrometastatic Breast Cancer


Updated version
Access the most recent version of this article at:
http://clincancerres.aacrjournals.org/content/11/10/3697

Supplementary Material
Access the most recent supplemental material at:
http://clincancerres.aacrjournals.org/content/suppl/2005/06/24/11.10.3697.DC1

Cited articles
This article cites 22 articles, 4 of which you can access for free at:
http://clincancerres.aacrjournals.org/content/11/10/3697.full.html#ref-list-1

Citing articles
This article has been cited by 3 HighWire-hosted articles. Access the articles at:
/content/11/10/3697.full.html#related-urls

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.