Molecular and Cellular Pathobiology

Genome-wide Methylation Analysis Identifies Genes Specific to Breast Cancer Hormone Receptor Status and Risk of Recurrence

Mary Jo Fackler¹, Christopher B. Umbricht¹,², Danielle Williams¹, Pedram Argani³, Leigh-Ann Cruz¹, Vanessa F. Merino¹, Wei Wen Teo¹, Zhe Zhang¹, Peng Huang¹, Kala Visvananthan¹,⁴, Jeffrey Marks⁵, Stephen Ethier⁶, Joe W. Gray⁷, Antonio C. Wolff¹, Leslie M. Cope¹, and Saraswati Sukumar¹

Abstract

To better understand the biology of hormone receptor–positive and –negative breast cancer and to identify methylated gene markers of disease progression, we carried out a genome-wide methylation array analysis on 103 primary invasive breast cancers and 21 normal breast samples, using the Illumina Infinium HumanMethylation27 array that queried 27,578 CpG loci. Estrogen and/or progesterone receptor–positive tumors displayed more hypermethylated loci than estrogen receptor (ER)-negative tumors. However, the hypermethylated loci in ER-negative tumors were clustered closer to the transcriptional start site compared with ER-positive tumors. An ER-classifier set of CpG loci was identified, which independently partitioned primary tumors into ER subtypes. A total of 40 (32 novel and 8 previously known) CpG loci showed differential methylation specific to either ER-positive or ER-negative tumors. Each of the 40 ER subtype–specific loci was validated in silico, using an independent, publicly available methylome dataset from the Cancer Genome Atlas. In addition, we identified 100 methylated CpG loci that were significantly associated with disease progression; the majority of these loci were informative particularly in ER-negative breast cancer. Overall, the set was highly enriched in homeobox containing genes. This pilot study shows the robustness of the breast cancer methylome and illustrates its potential to stratify and reveal biological differences between ER subtypes of breast cancer. Furthermore, it defines candidate ER-specific markers and identifies potential markers predictive of outcome within ER subgroups. Cancer Res; 71(19); 6195–207. ©2011 AACR.

Introduction

Approximately 200,000 women are diagnosed each year in the United States with breast cancer, and nearly 50,000 die of their metastatic disease. Significant improvement made in both early detection and local/systemic therapy in the past few decades has significantly improved patient outcomes, especially survival. Breast cancers are characterized by their estrogen receptor (ER) and progesterone receptor (PR) status, and it is established that ER expression (ER-positive) identifies a tumor phenotype with improved near/midterm prognosis and likely benefits from adjuvant endocrine therapy when compared with ER-negative tumors. Yet, little is known about the genomic features within each ER subtype of breast cancer that could explain why some patients with the same ER status have a good outcome whereas others do poorly regardless of treatment.

Current decision algorithms based on standard clinico-pathologic factors (1) stratify ER-negative disease as having a high risk for recurrence (2–4). Although patients are now routinely offered adjuvant chemotherapy, most patients with node-negative, ER-negative disease remain disease free after local therapy alone, including approximately 80% of ER-negative patients with tumors of 1 cm or less (5) and up to 60% of all with stage 1 disease (6). Consequently, there are patients with ER-negative disease that might do well without adjuvant chemotherapy and could avoid its potential toxicities, whereas others with a high residual risk despite it might be offered trials of novel therapies. Unfortunately, existing markers routinely used in clinical practice are of limited or no use in ER-negative patients (7). For example, commonly used gene expression tests by reverse
transcriptase PCR have no clear prognostic/predictive utility in ER-negative disease (8, 9), and microarray assays developed so far seem to identify essentially all such patients as high risk (10, 11), whereas other markers are still in development. Consequently, there is a critical need to develop better prognostic factors to improve assessment of residual risk and better predictive markers to optimize patient selection for standard and investigational systemic therapies.

Established clinically annotated tissue banks from prospective randomized clinical trials and extensive databases on expression profiling in breast cancer in the past decade have allowed the prospective–retrospective development of the several prognostic and predictive tests. For instance, clinicians have come to accept foregoing adjuvant chemotherapy in patients with ER-positive, node-negative disease with a low risk of distant recurrence of less than 10% at 10 years according to the 21-gene expression profile Oncotype DX assay, while strongly recommending it in those with a residual risk of more than 20% despite 5 years of adjuvant tamoxifen (9). Similarly useful tests are urgently needed for ER-negative breast cancer.

Multiple published studies using candidate gene approaches have suggested the utility of analyzing genes that undergo tumor-specific and promoter-specific hypermethylation as biomarkers for early detection and for prediction of outcome in multiple types of cancer (reviewed in ref. 12). Methylated genes are particularly robust as biomarkers. In past studies, we developed a cancer detection panel using a quantitative cumulative methylation assay (quantitative multiplex methylation-specific PCR; QM-MSP) wherein the methylation status of multiple genes could be determined individually and cumulatively from picograms of input DNA, such as is retrieved from ductal carcinoma in situ (DCIS; refs. 16–18). Furthermore, histopathologically normal ducts in the vicinity of tumor display detectable hypermethylation of genes that are present in the adjacent DCIS or invasive cancer, whereas normal ducts present farther away do not (18–21). However, using the candidate marker approach, it has been difficult to identify markers informative of the biology specifically of ER-positive or -negative breast cancer or those that predict response to therapy, disease progression, and survival. Therefore, we tested whether a genome-wide discovery platform would identify gene loci in tumors that better predict clinical outcomes (22–26).

As the first step toward studies with clinical trial samples, we carried out methylation array analyses on a discovery set of 103 primary invasive tumors and 21 normal samples. We found that distinctly different gene CpG loci typify the methylome of ER-positive and ER-negative breast cancers. Forty gene loci were identified that stratified tumors according to ER status. We also identified a putative “prognostic signature” of 100 CpG loci that are individually and collectively associated with outcome in patients with breast cancer. This feasibility study shows that CpG locus methylation levels could reveal important biological differences in the epigenome between breast cancer subtypes and provide ancillary clinical diagnostic, prognostic, and predictive tools.

Materials and Methods

Tissues

Frozen breast cancer tissues that were excised from patients with stages 1 to 3 disease prior to treatment (n = 103) were retrieved from Surgical Pathology at Johns Hopkins Hospital (Baltimore, Maryland) and confirmed to contain more than 50% epithelial cells. Normal breast organoids were prepared by enzymatic digestion of reduction mammoplasty specimens (n = 15; median patient age = 52 years, range = 47–71). Normal ducts from breast tissue more than 2 cm away from the tumor (n = 6) were isolated from cryosections, using laser capture microdissection (PALM MicroBeam; Carl Zeiss Microimaging). The studies were done with Institutional Review Board approval. Tumor characteristics are provided in Table 1 and Supplemental Table S1.

Table 1. Characteristics of the ER+ and ER− primary breast cancer patients in the study

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>ER+ (N = 44)</th>
<th>ER− (N = 38)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recurrences</td>
<td>7</td>
<td>11</td>
</tr>
<tr>
<td>DFS at 5 y (estimated by Kaplan–Meier)</td>
<td>87%</td>
<td>71%</td>
</tr>
<tr>
<td>HER2+/no. of cases annotated</td>
<td>4/20</td>
<td>10/25</td>
</tr>
<tr>
<td>Median % Ki67</td>
<td>20</td>
<td>50</td>
</tr>
<tr>
<td>AJCC stage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>II</td>
<td>22</td>
<td>19</td>
</tr>
<tr>
<td>III</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>Median tumor size, mm</td>
<td>28</td>
<td>59</td>
</tr>
<tr>
<td>Having &lt;1 mm margin</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>Therapy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Locoregional therapy</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>Endocrine</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Hormone</td>
<td>34</td>
<td>0</td>
</tr>
<tr>
<td>Chemotherapy</td>
<td>21</td>
<td>27</td>
</tr>
</tbody>
</table>

NOTE: A total of 21 additional samples were arrayed and used for ER classification, but not for outcome analyses. These 21 cases were excluded from the outcome analysis for the following reasons: neoadjuvant treatment (n = 8), samples obtained 6 months after the initial diagnosis (n = 10), and progression within 6 months after diagnosis (n = 10). Abbreviations: DFS, disease-free survival or total follow-up in No Progression cases; AJCC, American Joint Commission on Cancer.

*Therapies add up to more than totals because of cases with both endocrine and chemotherapy; locoregional therapy only (surgery ± radiation).
Genomic DNA extraction, sodium bisulfite conversion, and quality assurance

DNA extraction and quality assurance were carried out as described previously (27, 28) and in Supplementary Materials and Methods (Supplementary Fig. S1A).

Methylome analysis

Bisulfite-converted DNA was analyzed by using Illumina Infinium Human Methylation27 BeadChip Kit (WG-311-I202) in the JHU DNA Microarray Core. Locus methylation was calculated as a β-value within GenomeStudio software, low to high ranging from 0 to 1, respectively.

Data analysis

Data were analyzed by GenomeStudio software (Illumina, Inc.) and Bioconductor in R (http://www.bioconductor.org). Unsupervised cluster analysis was used to visualize and characterize broad methylation patterns in the data. All tests were 2-tailed and \( P < 0.05 \) was considered significant. Cox regression and Kaplan–Meier plots were used to model associations between methylation levels and time to recurrence, with and without adjustment for relevant clinical covariates, and to identify potential predictive markers. Covariates used were patients’ age at diagnosis, tumor grade, pathologic T stage, lymph node status, ER, progesterone receptor, type of primary surgery (with or without radiotherapy), and adjuvant therapy (chemotherapy and/or endocrine therapy). To identify methylated genes associated with ER status and their biology, a different approach was taken, emphasizing genes in which methylation changed dramatically between ER-positive and ER-negative samples. To achieve this, the initial selection was based on large fold changes. To evaluate the predictive capability of a panel of loci associated with ER status, we used independent samples to carry out receiver operating characteristic (ROC) analysis of a summary score of methylation derived as follows: (i) methylation at each locus was standardized to have a common scale by subtracting the mean methylation level and dividing by the SD for that locus so that low methylation resulted in negative values, whereas high methylation gave positive values; (ii) high methylation was associated with ER-positive status at some loci, and with ER-negative status at others, so standardized methylation scores for these latter loci were multiplied by −1, such that a high score uniformly indicated ER-positive samples; and (iii) genes were combined by averaging the standardized methylation scores for each patient, and the average score used in ROC analyses. The same procedure was used to summarize multic locus homeobox panels associated with recurrence. Data can be accessed at the Gene Expression Omnibus (accession number GSE31979).

Validation in the Cancer Genome Atlas samples

To verify that patterns of methylation observed in association with ER status and risk of recurrence within the JHU cohort were characteristic of breast cancer, we downloaded and analyzed data publicly available from the Cancer Genome Atlas Project (TCGA; http://tcga-data.nci.nih.gov/).

We selected TCGA to carry out this analysis because Illumina Meth27K was used, enabling direct comparisons for the same 50 bp CpG locus probes. In total, 185 samples were available on the Illumina 27k Human Methylation platform, and 465 samples were available on the Agilent G4502A Expression Array. Time to recurrence was not available at the time of download, but time to death was obtained for 342 of the samples queried on expression array and 182 samples queried on methylation array. Probe level data (TCGA level 2) were obtained for the methylation platform, whereas gene-level summaries (TCGA level 3) were used for RNA expression. Rank-based Spearman correlations were calculated between methylation and expression, using the 182 samples. Each methylation probe was mapped to the nearest gene by using the open source Illumina methylation platform annotation package available from Bioconductor (http://www.bioconductor.org/packages/2.6/data/annotation/html/IlluminaHumanMethylation27k.db.html), and correlations calculated for probes mapping to genes found on the expression array. Benjamini–Hochberg adjusted \( P \) values are reported for each probe, alongside the correlation coefficient. Association between overall survival and methylation or expression was evaluated by Cox regression. The ability of molecular markers to predict ER status was measured by carrying out an ROC analysis by using the methylation and expression levels of individual genes as predictors and reporting the area under the ROC curve. For expression, the ROC analysis was based on the expectation of an inverse relationship between methylation and expression, so that in some cases, where a significant, positive association is observed between the 2 platforms, the area under the ROC may be substantially less than 0.5.

Quantitative multiplex methylation-specific PCR

The 2-step multiplexed methylation-specific PCR method was previously described (15, 27, 28). For AKR1B1 primers/probes, see Supplementary Materials and Methods.

Results

Methylation profiling of primary invasive breast cancer tumors

Whole-genome methylation array analysis was carried out by using the Illumina Infinium HumanMethylation27 BeadChip with primary invasive carcinoma samples (\( n = 103 \)), samples from microdissected normal breast tissue distant from the primary tumor (\( n = 6 \)), and epithelium-enriched organoids isolated from normal breast (\( n = 15 \)). The array quantifies the proportion of methylated cytosines (\( 3^nM \)) to total cytosines at each of the 27,578 different CpG dinucleotides. The steps followed for our analysis is shown as a flowchart in Figure 1.

To characterize the overall methylation profile of primary invasive breast tumors, unsupervised hierarchical cluster analysis using the Manhattan distance was carried out on the most varied probes across tumors (1,578 gene loci, SD > 1.60; Fig. 2A). Two distinct clusters of tumors were
observed. Cluster 1 was enriched for ER-positive breast cancer (21 of 28, 75%), whereas cluster 2 contained 85% of the ER-negative tumors (41 of 75, 55% of total). Given the importance of ER in breast cancer, it is not surprising to observe a strong association between predominant methylation patterns and ER status ($OR = 3.57; 95\% CI, 1.27–11.20; P = 0.082$), but the result also highlights the importance of gene methylation in the disease process. The data also suggested additional subgroups within clusters 1 and 2 with distinct methylation profiles such as cluster 2B, which contains all ER−PRþ tumor samples.

**Distinct groups of genes are specifically and reciprocally hypermethylated in ER-positive versus ER-negative breast cancer**

Very little is known about the genomic features within each ER subtype of breast cancer that could explain why some patients have a good outcome, whereas others will do poorly regardless of treatment. To determine the differences in breast cancer biology/behavior between ER subtypes, we characterized methylation patterns at 8,376 selected CpG loci according to ER status. These loci met 2 criteria: (i) showed the most variation across primary tumors (SD > 0.100) and (ii) had

---

**Figure 1.** Schema outlining study design for analysis of association of methylation with ER status (A) and disease outcome (B). FDR, false discovery rate.
Figure 2. A, unsupervised hierarchical cluster analysis of the most varied 5% of CpG loci probes among tumors (1,378 loci, SD > 0.160). 2D hierarchical cluster analysis was carried out by using the Manhattan distance on 103 tumors, and 1,378 loci showed 2 distinct clusters. Cluster 1 is enriched for ER-positive (ER⁺) tumors (pink bars); regions of cluster 2 appear enriched for ER-negative (ER⁻) tumors (purple bars). Normal organoids cluster together (yellow bars). Examples of breast cancer-related gene loci are indicated at left. B, histogram plot showing the frequency of loci differentially methylated in ER-positive and ER-negative tumors. Plotted on the x-axis is the fold difference in the mean locus methylation of ER-positive/ER-negative tumors, among 8,376 loci with (SD > 0.1 across all tumors). The 200 CpG loci (left and right shaded boxes) are most differentially methylated between ER-positive and ER-negative tumors. C, distance to the TSS of the top CpG loci identified in ER-positive versus ER-negative tumors. Location of 100 ER-positive (pink line), 100-ER negative (purple line), and all 8,376 (black line) CpG loci was plotted relative to the TSS indicated as 0 on the x-axis. D, validation of ER subtype markers. An ROC plot shows an AUC of 0.961 for detection of ER subtype in 50 independent tumors within TCGA, using the ER-specific 40 loci marker set (data in Fig. 3, Supplementary Table S2, and Supplementary Fig. S2).

Probe detection \( P < 0.0001 \) (indicating that DNA from that locus was present above background levels and that probe intensities were consistently measured across replicate beads; the distribution of methylation among these loci is shown in Supplementary Fig. S1B). A substantial number of loci were observed with median methylation levels of 0.15 or more in both groups of tumor and normal breast organoids. However, the majority of loci were more highly methylated in tumor than in normal organoid samples; 1,744 loci in tumors had median methylation more than 2-fold higher than normal organoids (Supplementary Fig. S1C).

ER-positive tumors were found to have a higher frequency of hypermethylated gene loci than ER-negative tumors (Fig. 2B). Methylation at 5,264 loci was higher (ratio > 1) in ER-positive tumors samples than methylation of 3,112 loci (ratio < 1) in ER-negative tumors. The top 100 hypermethylated CpG loci in each group of ER-positive and ER-negative tumors were selected (Fig. 2B; ER-negative loci, ratio: 0.52–0.15 and ER-positive loci, ratio: 3.98–2.23; Supplementary Table S2). Interestingly, ER-negative tumors had a higher number of hypermethylated loci located closer to the transcriptional start site (TSS) than ER-positive tumors or the 8,376 array loci as a whole (Fig. 2C). This finding suggested a more rigorous suppression of gene expression by methylation in the ER-negative subtype, because methylated regions overlapping the TSS have been shown to most tightly negatively regulate transcription.

To further refine this set to identify ER subtype–specific biological/molecular functions most driven by the epigenome in breast cancer, we selected a subgroup of 40 hypermethylated loci of the 200 CpG locus set that individually showed the highest subtype specificity in individual tumor samples. Each individual locus was selected whose methylation profile showed (i) robust reciprocal methylation between the 2 ER subtypes, (ii) an incidence of more than 20% of methylation within the breast cancer subtype, and (iii) low methylation in normal breast epithelium/stroma and leukocytes (\( \beta < 0.15; \) Fig. 3). Using these selection criteria, in the discovery set, we identified 27 loci/probes aberrantly and reciprocally hypermethylated in ER-positive tumors and 13 loci/probes aberrantly hypermethylated in ER-negative tumors. The majority of these were at loci newly identified as hypermethylated in...
Figure 3. CpG methylation biomarkers of ER-positive versus ER-negative subtypes in primary tumors. Left (Discovery set, JH), CpG loci were evaluated by using 103 primary tumors. y-axis, β-methylation; x-axis, breast cancer samples; and N, normal breast organoids. The ER status of the tumors is denoted on top of each box. Right (Validation set, TCGA), validation on an independent dataset (n = 50) from TCGA is shown. Representatives of novel (A) and previously known (B) CpG loci specifically hypermethylated in ER+ and ER− breast cancers are shown.
breast cancer, and some never observed before as hypermethylated in cancer (Tables 2 and 3). As shown in Figure 3A, Supplementary Figure S2, and Supplementary Table S3, ACADL, ADAMTS1L, ARFGAP3, B3GAT1, CDCA7, FAM78A, FAM89A, FLJ31951 (RNF145), FLJ34922 (SLFN11), GAS6, HAAO, HEY2, HOXB9, ITGA11, NETO, PROX1, PSAT1, RECK, SMOC1,

### Table 2. Hypermethylated loci newly identified in breast cancers

<table>
<thead>
<tr>
<th>Gene symbol</th>
<th>Identified in this study</th>
<th>Hypermethylated in other cancers</th>
<th>Known aberrant expression</th>
<th>Gene Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACADL</td>
<td>ER⁺</td>
<td>No</td>
<td>No</td>
<td>Acyl-CoA dehydrogenase, long chain</td>
</tr>
<tr>
<td>ADAMTS1L</td>
<td>ER⁺</td>
<td>No</td>
<td>No</td>
<td>ADAMTS-like 1</td>
</tr>
<tr>
<td>ARFGAP3</td>
<td>ER⁺</td>
<td>No</td>
<td>No</td>
<td>ADP-ribosylation factor GTPase activating protein 3</td>
</tr>
<tr>
<td>B3GAT1</td>
<td>ER⁺</td>
<td>No</td>
<td>No</td>
<td>β-1,3-Glucuronitransferase 1</td>
</tr>
<tr>
<td>CDCA7</td>
<td>ER⁺</td>
<td>No</td>
<td>No</td>
<td>Cell division cycle associated 7</td>
</tr>
<tr>
<td>FAM78A</td>
<td>ER⁺</td>
<td>No</td>
<td>No</td>
<td>Family with sequence similarity 78, member A</td>
</tr>
<tr>
<td>FAM89A</td>
<td>ER⁺</td>
<td>No</td>
<td>No</td>
<td>Family with sequence similarity 89, member A</td>
</tr>
<tr>
<td>FLJ31951 (RNF145)</td>
<td>ER⁺</td>
<td>No</td>
<td>No</td>
<td>Unknown</td>
</tr>
<tr>
<td>FLJ34922 (SLFN11)</td>
<td>ER⁺</td>
<td>No</td>
<td>No</td>
<td>Schlafen family member 11</td>
</tr>
<tr>
<td>GAS6</td>
<td>ER⁺</td>
<td>No</td>
<td>No</td>
<td>Growth arrest-specific 6</td>
</tr>
<tr>
<td>HAAO</td>
<td>ER⁺</td>
<td>No</td>
<td>No</td>
<td>3-Hydroxyanthranilate 3,4-dioxygenase</td>
</tr>
<tr>
<td>HEY2</td>
<td>ER⁺</td>
<td>No</td>
<td>No</td>
<td>Hairy/enhancer-of-split related with YRPW motif 2</td>
</tr>
<tr>
<td>HOXB9</td>
<td>ER⁺</td>
<td>No</td>
<td>No</td>
<td>Homeobox B9</td>
</tr>
<tr>
<td>ITGA11</td>
<td>ER⁺</td>
<td>No</td>
<td>No</td>
<td>Integrin, alpha 11</td>
</tr>
<tr>
<td>NETO</td>
<td>ER⁺</td>
<td>No</td>
<td>No</td>
<td>Neuropilin (NRP) and tolloid (TLL)-like 2</td>
</tr>
<tr>
<td>PROX1</td>
<td>ER⁺</td>
<td>No</td>
<td>No</td>
<td>Prospero homeobox 1</td>
</tr>
<tr>
<td>PSAT1</td>
<td>ER⁺</td>
<td>No</td>
<td>No</td>
<td>Phosphoserine aminotransferase 1</td>
</tr>
<tr>
<td>RECK</td>
<td>ER⁺</td>
<td>Yes</td>
<td>No</td>
<td>Reversion-inducing cysteine-rich protein with kazal motifs</td>
</tr>
<tr>
<td>SMOC1</td>
<td>ER⁺</td>
<td>No</td>
<td>No</td>
<td>SPARC related modular calcium binding 1</td>
</tr>
<tr>
<td>SND1</td>
<td>ER⁺</td>
<td>No</td>
<td>No</td>
<td>Staphylococcal nuclease and tudor domain containing 1</td>
</tr>
<tr>
<td>TNFSF9</td>
<td>ER⁺</td>
<td>Yes</td>
<td>No</td>
<td>TNF (ligand) superfamily, member 9</td>
</tr>
<tr>
<td>ADHFE1</td>
<td>ER⁺</td>
<td>Yes</td>
<td>No</td>
<td>Alcohol dehydrogenase, iron containing, 1</td>
</tr>
<tr>
<td>DYNLRB2</td>
<td>ER⁺</td>
<td>No</td>
<td>No</td>
<td>Dynein, light chain, roadblock-type 2</td>
</tr>
<tr>
<td>HSD17B8</td>
<td>ER⁺</td>
<td>No</td>
<td>No</td>
<td>Hydroxysteroid (17-beta) dehydrogenase 8</td>
</tr>
<tr>
<td>PISD</td>
<td>ER⁺</td>
<td>No</td>
<td>No</td>
<td>Phosphatidylserine decarboxylase</td>
</tr>
<tr>
<td>PDOXK (C21orf124)</td>
<td>ER⁻</td>
<td>No</td>
<td>No</td>
<td>Pyridoxal kinase (vitamin B6 kinase)</td>
</tr>
<tr>
<td>WNK4</td>
<td>ER⁻</td>
<td>No</td>
<td>No</td>
<td>WNK lysine deficient protein kinase 4</td>
</tr>
</tbody>
</table>
yielded AUCs higher than 0.563, a range in which we expect under the curve (AUC) of 0.95. In all, 156 of 200 ER probes (Supplementary Table S2A). The median area under the ROC curve exceeded the same 5% threshold. This is consistent with the high degree of correlation observed between expression and methylation measurements of these genes in the TCGA data. At a false discovery rate of 0.05, 142 of 200 CpG loci are significantly inversely correlated with expression. And, as seen in Figure 3, the TCGA data provided support for the existence of ER subtype–specific methylation in breast cancer. To evaluate the predictive performance of the 40 locus panel (Fig. 2D), we derived an average methylation score for the entire set as described in Materials and Methods. Using this score, ROC analysis showed a high classification accuracy for the ER subtype in TCGA data with an area under the ROC curve of 0.961, with a specificity of 89% at a sensitivity of 90% for the ER subtype. Here, 121 of the 175 unique genes from our ER panel were discovered that showed tumor-specific and ER subtype–specific hypermethylation. Existing literature provided further validation to our current observations.

### External validation of methylation array findings in an independent test set of primary tumors

Next, we validated these findings in publicly available data on the breast cancer samples in TCGA (http://tcga-data.nci.nih.gov/), using an ROC analysis to evaluate predictive ability (Supplementary Table S2A). The median area under the ROC curve for the 200 loci was 0.7. 1 gene, SERPINA12, had an area under the curve (AUC) of 0.95. In all, 156 of 200 ER probes yielded AUCs higher than 0.563, a range in which we expect only 5% of CpG loci by chance alone. Interestingly, expression of most of these same genes is also a very strong predictor of ER status. Here, 121 of the 175 unique genes from our ER panel and available on the expression array had areas under the curve exceeding the same 5% threshold. This is consistent with the high degree of correlation observed between expression and methylation measurements of these genes in the TCGA data. At a false discovery rate of 0.05, 142 of 200 CpG loci are significantly inversely correlated with expression. And, as seen in Figure 3, the TCGA data provided support for the existence of ER subtype–specific methylation in breast cancer. To evaluate the predictive performance of the 40 locus panel (Fig. 2D), we derived an average methylation score for the entire set as described in Materials and Methods. Using this score, ROC analysis showed a high classification accuracy for the ER subtype in TCGA data with an area under the ROC curve of 0.961, with a specificity of 89% at a sensitivity of 90% (Fig. 2D; details in Supplementary Materials and Methods). A similar composite score derived from expression probes for the same genes showed some discriminatory ability in the TCGA data, albeit reduced, with an area under the ROC of 0.667 (data not shown).

### CpG loci associated with disease progression in patients with newly diagnosed invasive breast cancer

To develop an epigenomic signature that predicts outcome in patients with breast cancer, we conducted differential methylation analysis on primary tumors from recurrent versus nonrecurrent breast cancers. We used a subgroup of 82 well-annotated, invasive breast tumors derived from the discovery set of 103 tumors that included 44 ER-positive (7 recurrences) and 38 ER-negative (11 recurrences) breast cancers and

---

**Table 3.** Hypermethylated loci previously identified in breast cancers

<table>
<thead>
<tr>
<th>Gene symbol</th>
<th>Published</th>
<th>Hypermethylation in other cancers</th>
<th>Known aberrant expression</th>
<th>Gene</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>EVI1 (MECOM)</td>
<td>ER⁻ (25, 39)</td>
<td>No</td>
<td>Basal (40)</td>
<td>MDS1 and EVI1 complex locus</td>
<td>Nucleus</td>
</tr>
<tr>
<td>ETS1</td>
<td>ER⁻ (25)</td>
<td></td>
<td>Basal (38)</td>
<td>v-ets erythroblastosis virus E26 oncogene homologue 1</td>
<td>Nucleus</td>
</tr>
<tr>
<td>IRF7</td>
<td>ER⁻ (25)</td>
<td>Yes</td>
<td>Basal (38)</td>
<td>Interferon regulatory factor 7</td>
<td>Nucleus</td>
</tr>
<tr>
<td>LYN</td>
<td>ER⁻ (25)</td>
<td>Yes</td>
<td>Basal (38)</td>
<td>v-yes-1 Yamaguchi sarcoma viral related oncogene homolog</td>
<td>Cytoplasm</td>
</tr>
<tr>
<td>PDXK</td>
<td>ER⁻ (29)</td>
<td>No</td>
<td>Basal</td>
<td>Pyridoxal kinase (vitamin B6 kinase)</td>
<td>Cytoplasm</td>
</tr>
<tr>
<td>PTGS2 (COX2)</td>
<td>ER⁻ (25)</td>
<td>Yes</td>
<td>Basal (38)</td>
<td>Prostaglandin-endoperoxide synthase 2</td>
<td>Cytoplasm</td>
</tr>
<tr>
<td>RUNX3</td>
<td>ER⁻ (43)</td>
<td>Yes</td>
<td>Basal (38)</td>
<td>Runt-related transcription factor 3</td>
<td>Nucleus</td>
</tr>
<tr>
<td>VIM</td>
<td>ER⁻</td>
<td>Yes</td>
<td>Basal (41)</td>
<td>Vimentin</td>
<td>Cytoplasm</td>
</tr>
<tr>
<td>DAB2IP (4 loci)</td>
<td>ER⁻ (25)</td>
<td>Yes</td>
<td>Basal</td>
<td>DAB2 interacting protein</td>
<td>Plasma Membrane</td>
</tr>
<tr>
<td>HSD17B4</td>
<td>ER⁻ (44)</td>
<td>Yes</td>
<td>ER⁻ (44)</td>
<td>Hydroxysteroid (17β)-dehydrogenase 4</td>
<td>Cytoplasm</td>
</tr>
<tr>
<td>PER1</td>
<td>ER⁻ (23)</td>
<td></td>
<td></td>
<td>Period homologue 1 (Drosophila)</td>
<td>Nucleus</td>
</tr>
</tbody>
</table>

**SND1**, and **TNFSF9** were found hypermethylated almost exclusively in ER-positive tumors, whereas **ADHFE1, DYNLRB2, HSD17B8, PDXK, PISD**, and **WNK4** were found hypermethylated in ER-negative tumors. A number of genes previously reported as having subtype-specific methylation were also identified. **EVII, ETS1, IRF7, LYN, PTGS2 (COX2), RUNX3**, and **VIM** were found to be hypermethylated in ER-positive tumors, whereas **DAB2IP, HSD17B4**, and **PER1** were reported to be hypermethylated in ER-negative breast cancers (detailed information and references in Table 3). A second distinct CpG locus of **PDXK** was previously found hypermethylated in ER-positive breast cancers (29). We did not find any gene that was preferentially methylated in ER-positive or ER-negative tumors where the literature conflicted with our data. The concordance between current and published data is shown for 2 of these gene CpG loci, **EVII** (23–25) and **DAB2IP** (23), in Figure 3B. Thus, many novel and some published gene loci were discovered that showed tumor-specific and ER subtype–specific hypermethylation. Existing literature provided further validity to our current observations.
independently queried the ER-positive and ER-negative tumor groups (Table 1; Supplementary Table S1) as follows. Differential methylation analysis was carried out in GenomeStudio, using the DiffScore algorithm to compare tumors which later recurred with those that did not. The analysis was carried out separately on the ER-positive and ER-negative tumor groups. Candidate loci (50 per ER subtype) that met the following 3 criteria were selected: (i) more highly methylated in recurrent tumors than in nonrecurrent tumors, (ii) relatively unmethylated in normal samples ($\beta < 0.15$), and (iii) significantly differentially methylated above the false discovery rate cutoff (5%). Next, we carried out a multivariate Cox regression analysis for each of these candidate loci and generated Kaplan–Meier plots, showing the interrelationships between ER status and methylation and depicted in these plots as high/low with respect to the median methylation level for each CpG locus. From these 100 candidate CpG loci, a set of 32, selected for high Cox coefficients (Supplementary Table S4A) and visually striking Kaplan–Meier plots (Fig. 4; Supplementary Fig. S3) were followed up most closely, including with an extensive literature search to identify previous associations with outcome in breast cancer (Supplementary Table S4C). Novel associations with poor outcome were identified for (i) TMEM179, CRMP1, and SCN1B in ER-positive breast cancer; (ii) ALX1, COL1A1, EP405, EYA4, FLRT2, GPX7, KCNB2, LAMA1, LHX1, NEUROG1, POU3F2, and STMN3 in ER-negative breast cancer; and (iii) AKR1B1, COL6A2, EYA4, GPX7, HOXA13, HOXB13, NX6-2, NRIP2, POU4F2, REM1, and SLTRK2, in both ER-positive and ER-negative tumors. Cox regression $P$-values for each member of the 100 CpG loci set is presented in Supplementary Table S4A. Because the differential methylation analysis was designed in such a way to find loci most highly methylated in recurrent tumors, we did not observe hypomethylated loci associating with recurrence.

To verify array data using an independent assay, and to ensure future technical translation of the HumanMethylation27 array data to laboratory assays, we tested several methylated genes, such as EVI1, DAB2IP, and AKR1B1 by carrying out QM-MSP. In each case, we observed an excellent correlation between the levels of gene methylation assessed by both assays. A comparison of level of methylation in AKR1B1 assessed by the array and by QM-MSP in individual primary tumors and both data plotted as Kaplan–Meier plots as shown in Figure 4C.

A striking observation was that nearly 20% of the recurrence loci (18 of 100 loci, 15 of 91 unique genes) were from homeobox-containing genes including the HOX, LHX, POU, ALX, and NK6 gene families (Fig. 5; Supplementary Table S4A). With only 375 homeobox loci (189 genes) present in the 27,578 loci (14,495 genes) array, this represented a dramatic enrichment of homeobox genes in our 100 recurrence related set (OR = 16.17, $P = 6.515e-13$). These data clearly implicate methylated homeobox genes as key factors in tumor progression. To determine if the other homeobox loci on the array exhibited similar methylation patterns, we extended our analysis to 60 homeobox loci which showed high variance (SD above the 95th percentile for the array) among the tumors, excluding the 18 loci represented in the recurrence sets. Two-dimensional (2D) hierarchical cluster analysis (using the Manhattan distance) was carried out to characterize these loci. As shown in the heatmap in Figure 5A, the 18 homeobox gene loci derived from the 100 recurrence locus set have distinctive methylation patterns, showing significant comethylation within the first cluster, with highly methylated samples tending to be methylated for all the loci. Interestingly, a similar clustering profile was observed with the 60 homeobox loci (Fig. 5B), suggesting that the homeobox genes as a group have a common methylation signature. To evaluate correlation with recurrence, we derived an average methylation score for the panel as described in Materials and Methods. In a multivariate analysis that included age, stage, treatment, and ER status, there was clear evidence of a significant additional and independent contribution to the model where the Cox coefficient was 1.74, with a $P$-value of 0.0042. Kaplan–Meier plots for the 18 and the 60 homeobox loci (but not for all 1,378 CpG loci that showed differential methylation across all the tumors) illustrate their predictive value. These results support the notion that highly methylated homeobox loci and loss of their expression may likely contribute to poor outcome in breast cancer.

**External validation of associations with outcome in an independent test set of primary tumors**

Next, we sought to validate these findings in publicly available TCGA breast cancer samples (http://tcga-data.nci.nih.gov/), using Cox regression to evaluate association between methylation and overall survival; progression-free survival was not available for these samples at the time of download. In total, survival information was available for 342 of the samples available on expression array, of which 182 were also available on methylation array. Despite the change of outcome variable and moderate sample sizes, results in TCGA data as a set confirmed the findings that these genes are significantly associated with outcome. An overwhelming majority of our recurrence marker loci (78 of 100) have positive Cox regression coefficients, indicating that hypermethylation of these loci is associated with a worse outcome in these samples as well. In comparison, we would expect only half of these loci to have positive Cox coefficients by chance alone, giving a composite $P$ value of 2.2e-09 in support of the association. Additional confirmation for the panel is provided by the fact that for more than two third of these genes, Cox regression analysis of TCGA expression data show that low expression correlates with worse outcome. This result is wholly consistent with the observed methylation results, and statistically significant in its own right, with a $P$-value of 0.00022. This is consistent with the high degree of correlation observed between expression and methylation measurements of these genes in the TCGA data. At a false discovery rate of 0.05, 43 of 100 CpG loci are significantly inversely correlated with expression. We also carried out a multivariate Cox regression analysis for each of these candidate loci and generated Kaplan–Meier plots for the sets of 18 loci (log-rank test, $P = 0.00027$) and 60 loci (log-rank test, $P = 0.00036$), compared with the top 5% of varied probes (1,378 probes, $P = 0.112$), showing significant interrelationships between homeobox gene methylation and survival (data not shown).
Figure 4. Methylated CpG loci associated with disease progression. Kaplan–Meier plots show CpG loci with strong associations between hypermethylation and high rate of disease progression among 82 tumors. Recurrence is shown for ER+ (dashed line) and ER- tumors (solid line). High (red line) and low (blue line) methylation was defined relative to the median methylation level for a given CpG loci within ER subtype. CpG loci associated with a high rate of recurrence in (A) ER+ and ER- breast tumors. B, ER-negative breast tumors. Cox regression P values for each of the 100 CpG loci sets are presented in Supplementary Table 4A. C, verification of HumanMethylation27 array data (top, blue bars) by QM-MSP analysis (bottom) for AKR1B1. x-axis, tumors; y-axis, methylation levels (β-value for array, % methylation for QM-MSP). Kaplan–Meier curves using array and QM-MSP values.
Discussion

In this study, we report the results of a genome-wide array analysis of primary invasive breast cancers of 27,578 CpG loci. This screen identified hypermethylated genes that specifically segregate with ER-positive or ER-negative tumor subtypes, which were then validated in silico by using the newly populated TCGA breast cancer database. The array analysis also identified 100 gene loci that were enriched for homeobox-containing genes and predicted recurrence in breast cancers. Many novel hypermethylated loci were identified. In summary, we show that the methylome is a rich source of genes whose hypermethylation has the potential to significantly contribute to the understanding of ER subgroups of breast cancer and predict recurrence in ER-positive and in ER-negative breast cancers.

We observed a significantly higher frequency of hypermethylation in ER-positive than ER-negative tumors \((P < 0.0001)\). The reason for the hypermethylated phenotype of ER-positive tumors is not yet clear. The simplest explanation could be that the ER-positive markers are drivers that enhance the expression of the DNA methyltransferases or inhibit the repair processes that remove methyl groups from DNA. On the contrary, reduced methylation in ER-negative tumors might offer an explanation for their relative aggressiveness because the uncontrolled expression of growth factors and their receptors may be facilitated by removing the protective imposition of methylation-mediated silencing. The observation of a higher frequency of hypermethylated genes in ER-positive tumors is substantiated by 5 recent studies describing the breast cancer methylome (22, 24–26, 29). In a study examining 44 primary tumors, Hill and colleagues (22) confirmed a highly significant difference \((\text{ANOVA, } P = 0.001)\) in hypermethylation depending on hormone receptor status. Similarly, in the Fang study (26), ER/PR-positive tumors displayed high level of methylation across the top 5% variant loci in the 27K Illumina array. Our study of 103 primary breast cancers identified many novel loci that have the ability to impressively segregate ER-positive and ER-negative tumors (Tables 2 and 3; Fig. 3A and B), shedding light on many novel...
paths and constituent genes that may be involved in the
genesis of these subgroups. We tested the strength of these
observations by using an independent dataset from TCGA
for both methylation and expression. All 40 CpG loci showed
reproducible associations with ER subtype and these
markers classified essentially all tumors into the correct ER
subtype (AUC, 0.961). Interestingly, expression of the same
genes was also found to be a very strong predictor of ER
status. Thus, independent TCGA data strongly validated our
findings. With the caveat that both the discovery and
validation cohorts represent small sample sets, the repro-
ducibility of the findings supports the strength of this plat-
form to reveal differences that can now be studied in detail.

A major goal of our work is to find markers that can
prognosticate recurrence and predict benefit from therapy.
Expression array-based analyses have proven to be useful for
ER-positive breast cancers (30, 31). Their utility, however,
has been limited in ER-negative breast cancer. Also, DNA
mutation and copy number studies have been found less
useful in breast cancer compared with other cancers (e.g.,
lung and colon), probably reflecting the greater diversity
of breast cancer subtypes (32–34). Epigenetically mediated
gene silencing through DNA methylation occurs extremely
frequently and has now been accepted as a major driver in
neoplastic transformation, especially in the breast (12).

Genome-wide methylation analysis allowed us to identify
a tumor recurrence marker set of 100 gene loci; a few specific
to ER-positive tumors, many loci specific to ER-negative
tumors, and many common to both (Fig. 4; Supplementary
Table S4C). The emergence of a homeobox gene methylation
signature predictive of recurrence among the 100 recurrence
loci and also among all homeobox loci on the array is
notable. Substantial inverse correlation was seen between
methylation and expression of the genes, both of the ER-
stratifying set and the recurrence set of CpG loci, suggesting
functional relevance to the effects of methylated genes
observed in our study. These genes play critical roles in
differentiation and development, growth factor receptor
signaling, angiogenesis, and more recently an unequivocal
role in stem cell function (reviewed in ref. 35). At the same
time, particularly within the recurrence panel, expression
alone could not duplicate the level of performance achieved
with methylation probes.

The tissues used for the current analysis were from
an institutional cohort of frozen specimens and are there-
fore samples of convenience with their inherent drawbacks.
Additional studies will need to address the question of the
precise role of methylation signatures in prognosticating
outcome and predicting response to therapy. More discov-
ergy and validation will need to be carried out with annot-
tated samples from controlled studies, with more uniform
standards of sample collection, such as in the context of
large mature randomized clinical trials. To allow investiga-
tion on archival specimens, the rapid development of
methods to retrieve high-quality DNA from paraffin-embed-
ded tissues is imperative. Our recent success in standard-
izing restoration of DNA retrieved from formalin-fixed,
paraffin-embedded tissues, in collaboration with Illumina
(our unpublished data; ref. 36), bodes well for the future of
these investigations.

In summary, this study has shown the feasibility of distin-
guishing ER subtype in breast cancers and possibly predicting
outcome based on CpG DNA methylation. The study suggests
pathways that may explain distinctive behaviors among ER-
positive and ER-negative tumors. In conclusion, the data
strongly support upcoming planned studies that will use
existing clinically annotated tissues from previously con-
ducted prospective randomized trials to examine the prog-
nostic outcome and predictive therapeutic information
offered by methylation markers in a prospective–retrospective
fashion.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

We thank Dr. Wayne Yu for carrying out the methylation microarray, Drs.
Gedge Rosen and Michelle Manahan for providing the reduction mammaplasty
tissues, and Ms. Areli Lopez for excellent technical assistance.

Grant Support

This work was supported by grants from the Rubenstein and Cohen families
and the Breast Cancer SPORE: P50-CA-88843.

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.

Received May 11, 2011; revised August 3, 2011; accepted August 4, 2011;
published OnlineFirst August 8, 2011.

References

1. Carlson RW, Alfred DC, Anderson BO, Burstein HJ, Carter WB, Edge

2. Hammond ME, Hayes DF, Dowsett M, Alfred DC, Hagerty KL, Badve
S, et al. American Society of Clinical Oncology/Clinical Oncology/College Of American
Pathologists guideline recommendations for immunohistochemical
testing of estrogen and progesterone receptors in breast cancer. J

American Society of Clinical Oncology 2007 update of recommenda-
tions for the use of tumor markers in breast cancer. J Clin Oncol

4. Wolff AC, Hammond ME, Schwartz JN, Hagerty KL, Alfred DC, Cote RJ,
et al. American Society of Clinical Oncology/College of American
Pathologists guideline recommendations for human epidermal growth

5. Fisher B, Dignam J, Tan-Chiu E, Anderson S, Fisher ER, Wittliff JL,
et al. Prognosis and treatment of patients with breast tumors of one
centimeter or less and negative axillary lymph nodes. J Natl Cancer

DL, et al. Findings from recent National Surgical Adjuvant Breast and
Bowel Project adjuvant studies in stage I breast cancer. J Natl Cancer
Inst Monogr 2001;62–6.
Methylome Classifiers of Cancer Subtype and Prognosis


Genome-wide Methylation Analysis Identifies Genes Specific to Breast Cancer Hormone Receptor Status and Risk of Recurrence

Mary Jo Fackler, Christopher B. Umbricht, Danielle Williams, et al.

Cancer Res 2011;71:6195-6207. Published OnlineFirst August 8, 2011.

Updated version
Access the most recent version of this article at:
doi:10.1158/0008-5472.CAN-11-1630

Supplementary Material
Access the most recent supplemental material at:
http://cancerres.aacrjournals.org/content/suppl/2011/08/08/0008-5472.CAN-11-1630.DC1

Cited articles
This article cites 42 articles, 15 of which you can access for free at:
http://cancerres.aacrjournals.org/content/71/19/6195.full#ref-list-1

Citing articles
This article has been cited by 21 HighWire-hosted articles. Access the articles at:
http://cancerres.aacrjournals.org/content/71/19/6195.full#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, use this link
http://cancerres.aacrjournals.org/content/71/19/6195.
Click on "Request Permissions" which will take you to the Copyright Clearance Center's (CCC) Rightslink site.