Epigenome-wide Association Study of Breast Cancer Using Prospectively Collected Sister Study Samples

Zongli Xu, Sophia C.E. Bolick, Lisa A. DeRoo, Clarice R. Weinberg, Dale P. Sandler, Jack A. Taylor

Background
Previous studies have suggested DNA methylation in blood is a potential epigenetic marker of cancer risk, but this has not been evaluated on a genome-wide scale in prospective studies for breast cancer.

Methods
We measured DNA methylation at 27,578 CpGs in blood samples from 298 women who developed breast cancer 0 to 5 years after enrollment in the Sister Study cohort and compared them with a random sample of 612 cohort women who remained cancer free. We also genotyped women for nine common polymorphisms associated with breast cancer.

Results
We identified 250 differentially methylated CpGs (dmCpGs) between case subjects and noncase subjects (false discovery rate [FDR] Q < 0.05). Of these dmCpGs, 75.2% were undermethylated in case subjects relative to noncase subjects. Women diagnosed within 1 year of blood draw had small but consistently greater divergence from noncase subjects than did women diagnosed at more than 1 year. Gene set enrichment analysis identified Kyoto Encyclopedia of Genes and Genomes cancer pathways at the recommended FDR of Q less than 0.25. Receiver operating characteristic analysis estimated a prediction accuracy of 65.8% (95% confidence interval = 61.0% to 70.5%) for methylation, compared with 56.0% for the Gail model and 58.8% for genome-wide association study polymorphisms. The prediction accuracy of just five dmCpGs (64.1%) was almost as good as the larger panel and was similar (63.1%) when replicated in a small sample of 81 women with diverse ethnic backgrounds.

Conclusions
Methylation profiling of blood holds promise for breast cancer detection and risk prediction.


Breast cancer is the most common cancer in women and the leading cause of cancer mortality for women worldwide (1). According to estimates for 2011, there are an estimated 230,480 incident cases and 40,000 deaths from the disease each year in the United States (2). The Gail model, based on known risk factors including age and reproductive, medical, and family history, is the best breast cancer risk prediction method currently available for populations, but its predictive accuracy for individuals is only about 58.0% to 59.0% (3,4). Rare inherited mutations of the breast cancer susceptibility genes BRCA1 and BRCA2 are strongly associated with familial breast cancer, but together only account for 5% to 10% of breast cancers in the United States (5).

Recent genome-wide association studies (GWASs) have identified common polymorphisms associated with breast cancer risk. A panel of 10 such SNPs have a predictive accuracy of 59.7% (4). Thus the known environmental and common genetic risk factors for breast cancer have limited use in predicting a woman’s risk of disease (4–6).

Epigenetic modifications, including DNA methylation, are increasingly recognized as important determinants of gene transcriptional regulation that have both heritable and acquired characteristics (7). Aberrant DNA methylation patterns are among the earliest and most common events in carcinogenesis (8), and genome-wide methylation profiling has recently been extended to retrospectively collected blood samples in case–control studies of ovarian, bladder, and head and neck cancers (9–11). But unlike genotype, which remains constant, epigenetic modifications may differ from cell to cell, over time, and with exposure, making the results of case–control studies subject to a number of potential biases (12).

In this study we used genomic arrays of 27,578 CpGs and prospectively collected blood samples from women in the Sister Study cohort to compare women who subsequently developed breast cancer to those who remained cancer-free. We use these data to address two questions: Does the epigenetic profile differ between women who subsequently develop cancer vs those who do not? Within the case subject group, is there an effect of time to diagnosis—the interval between blood draw and clinical diagnosis of disease? In addition, we examined the predictive power of blood methylation compared with the Gail model and GWAS single nucleotide polymorphisms (SNPs).
Methods

Study Design

The National Institute of Environmental Health Sciences (NIEHS) Sister Study is a nationwide, prospective cohort study designed to explore genetic and environmental determinants of breast cancer (see Supplementary Methods, available online). To be eligible, women could not have had breast cancer themselves but must have had a biological sister with breast cancer. Participants provided extensive information at baseline interview, and informed consent and blood samples were obtained during a home visit.

The study was approved by the institutional review boards of the NIEHS, National Institutes of Health, and the Copernicus Group. Our nested case–cohort design included 329 incident case subjects who were diagnosed with breast cancer during the interval between their blood draw and May 2008, and a random sample of 709 women drawn from the 29026 participants enrolled by June 2007. Methylation profiling was performed on DNA extracted from 1016 whole blood samples and 22 blood clot samples. Details of extraction, bisulfite conversion, array hybridization, genotyping, and quality control measures are provided in the Supplementary Materials (available online). After excluding 27 low-quality samples, we had high-quality measures on 1011 samples. The majority of the women were non-Hispanic whites (n = 928) with the remaining women being black (n = 37), Hispanic (n = 23), or from other racial groups (n = 23). Because of the possibility that effects would differ by race/ethnicity, we restricted our analysis to the 910 non-Hispanic whites with DNA extracted from whole blood samples and used the 81 women from other ethnic groups (25 case subjects and 56 noncase subjects) with DNA from whole blood samples as a separate validation set. Of the 910 women, 620 women belonged to the subcohort; 8 subcohort members subsequently became case subjects. Details of gene set enrichment analysis and receiver operator characteristics analysis are provided in the Supplementary Methods (available online).

Statistical Analysis

An association test between CpG methylation profiles (β value) and binary phenotype case subject/noncase subject status was carried out using a case–cohort proportional hazard model (13,14), where age was treated as the primary time factor, age at blood draw was the left truncation time, and right censoring time was May 15, 2008. For each CpG site, the methylation β value was the predictor variable for the case–cohort analysis. There was no indication that the proportional hazards assumption was violated: The interaction term of age by methylation was not significant at a P value threshold of .05 for any CpG. In a separate set of analyses restricted to case subjects, a linear mixed model was employed to test the association between methylation β value and time interval from blood draw to diagnosis plate was modeled as contributing a random effect. Both case–cohort and case-only analyses were performed for each CpG separately, adjusting for plate and bisulfite conversion efficiency. To minimize the effect of outliers, for each CpG probe we excluded β values that were more than 3 standard deviations from the mean based on the combined data from case subjects and noncase subjects. To correct for multiple testing, we estimated the false discovery rate (FDR) using the Q value framework (15). A Q value of less than 0.05 was considered statistically significant. All statistical tests were two-sided. The sample size provided more than 80% power to detect a case subject vs noncase subject methylation β value difference of 0.6% with a variance of 0.9% (the mean variance of all CpGs on the array in our data) at type 1 error rate of 0.05.

Results

Breast Cancer–Associated CpG Sites

The mean age at enrollment for noncase subjects (n = 612) was 55 years (range = 35–75), whereas case subjects (n = 298) had a mean age of 57 years (range = 36–75). Noncase subjects had an average follow-up time (time between blood draw and analysis cutoff date in May 2008) of 908 days (range = 351–1681 days), whereas case subjects had an average time to diagnosis of 465 days (range = 5–1389 days). There was increased risk associated with number of first-degree relatives with breast cancer (P = .008) and with having a mother with breast cancer (P = .03), but there was no association with age at menarche, ever having had a full-term pregnancy, or age of menopause (Table 1). Tumor characteristics for case subjects are provided in Table 2.

Using case–cohort proportional hazard regression we identified 250 differentially methylated CpG (dmCpG) markers at a FDR threshold of Q equal to 0.05 (Figure 1; Supplementary Figure 1A, available online; Supplementary Table 1, available online). A total of 188 (75.2%) of the 250 dmCpGs were undermethylated in case subjects relative to noncase subjects (Figure 2A). This frequency of undermethylated dmCpGs in case subjects was statistically significantly different (ρ2 = 61; P < .001) from that seen on the array as a whole, where 50.5% (13 922 of 27 578) of CpGs were undermethylated in case subjects relative to noncase subjects. In addition, 222 (88.8%) of the 250 dmCpGs were located in CpG islands. This frequency differed statistically significantly (ρ2 = 32;
HER-2/NEU Overexpression 23 (7.72)
Location Ductal 225 (75.50)
Estrogen receptor assay Positive 217 (72.82)
Breast cancer stage* 0 55 (18.46)
Metastasis classification* M0 270 (90.60)
Lymph nodes classification* pN0 210 (70.47)

P < .001) from the representation on the chip, where 20006 (72.5%) of the 27578 CpGs were located in CpG islands. Among the 28 dmCpGs located in nonisland regions, 23 (82.1%) were overmethylated, which did not differ statistically (χ² = 2.9; P = .09) from the chip as a whole, where 4919 (65.0%) of 7572 nonisland CpGs were overmethylated in case subjects. The overall distribution of methylation beta value variability for the 250 dmCpGs was slightly greater than that for the array as a whole (Supplementary Figure 1B, available online). A total of 69 CpG probes on the HumanMethylation27 BeadChip (Illumina Inc, San Diego, CA) were located near known breast cancer susceptibility genes (Supplementary Table 1, available online) with a total of eight of the 14 KEGG cancer pathways (Supplementary Figure 2A, available online). Similarly, analysis using both the geneSetTest (18) method and the hypergeometric test method confirmed the enrichment for cancer pathways at a two-sided P less than .05 (Supplementary Figure 2, B and C, available online) with a total of eight of the 14 KEGG cancer pathways enriched in two or more analyses.

### Table 2. Breast tumor characteristics among case subjects

<table>
<thead>
<tr>
<th>Tumor characteristics</th>
<th>Category</th>
<th>No.</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tumor classification*</td>
<td>Tis</td>
<td>63</td>
<td>21.14</td>
</tr>
<tr>
<td></td>
<td>T1</td>
<td>146</td>
<td>48.99</td>
</tr>
<tr>
<td></td>
<td>T2</td>
<td>44</td>
<td>14.77</td>
</tr>
<tr>
<td></td>
<td>T3</td>
<td>6</td>
<td>2.01</td>
</tr>
<tr>
<td></td>
<td>TX</td>
<td>14</td>
<td>4.70</td>
</tr>
<tr>
<td></td>
<td>Missing</td>
<td>25</td>
<td>8.59</td>
</tr>
<tr>
<td>Lymph nodes classification*</td>
<td>pN0</td>
<td>210</td>
<td>70.47</td>
</tr>
<tr>
<td></td>
<td>pN1</td>
<td>48</td>
<td>16.11</td>
</tr>
<tr>
<td></td>
<td>pN2</td>
<td>8</td>
<td>2.68</td>
</tr>
<tr>
<td></td>
<td>pN3</td>
<td>3</td>
<td>1.01</td>
</tr>
<tr>
<td></td>
<td>pNX</td>
<td>4</td>
<td>1.34</td>
</tr>
<tr>
<td></td>
<td>Missing</td>
<td>25</td>
<td>8.59</td>
</tr>
<tr>
<td>Metastasis classification*</td>
<td>M0</td>
<td>270</td>
<td>90.60</td>
</tr>
<tr>
<td></td>
<td>MX</td>
<td>3</td>
<td>1.01</td>
</tr>
<tr>
<td></td>
<td>Missing</td>
<td>25</td>
<td>8.59</td>
</tr>
<tr>
<td>Breast cancer stage*</td>
<td>0</td>
<td>55</td>
<td>18.46</td>
</tr>
<tr>
<td></td>
<td>I</td>
<td>118</td>
<td>39.60</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>63</td>
<td>21.14</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>13</td>
<td>4.36</td>
</tr>
<tr>
<td></td>
<td>Cannot stage</td>
<td>24</td>
<td>8.05</td>
</tr>
<tr>
<td></td>
<td>Missing</td>
<td>25</td>
<td>8.59</td>
</tr>
<tr>
<td>Estrogen receptor assay</td>
<td>Positive</td>
<td>217</td>
<td>72.82</td>
</tr>
<tr>
<td></td>
<td>Negative</td>
<td>43</td>
<td>14.43</td>
</tr>
<tr>
<td></td>
<td>Missing</td>
<td>38</td>
<td>12.75</td>
</tr>
<tr>
<td>Progestosterone receptor assay</td>
<td>Positive</td>
<td>168</td>
<td>56.38</td>
</tr>
<tr>
<td></td>
<td>Negative</td>
<td>75</td>
<td>25.17</td>
</tr>
<tr>
<td></td>
<td>Missing</td>
<td>55</td>
<td>18.46</td>
</tr>
<tr>
<td>Location</td>
<td>Ductal</td>
<td>225</td>
<td>75.50</td>
</tr>
<tr>
<td></td>
<td>Lobular</td>
<td>25</td>
<td>8.39</td>
</tr>
<tr>
<td></td>
<td>Both</td>
<td>16</td>
<td>5.37</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>2</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>Missing</td>
<td>30</td>
<td>10.07</td>
</tr>
<tr>
<td>HER-2/NEU</td>
<td>Overexpression</td>
<td>23</td>
<td>7.72</td>
</tr>
<tr>
<td></td>
<td>Not overexpression</td>
<td>150</td>
<td>50.34</td>
</tr>
<tr>
<td></td>
<td>Missing</td>
<td>125</td>
<td>41.95</td>
</tr>
</tbody>
</table>

* Breast cancer staging in the Sister Study used a modification of Tumor, Node, Metastasis (TNM) (31).

Functional Pathway Analysis

Gene set enrichment analysis (17) showed enrichment (at the recommended FDR threshold of 0.25) for six of 14 cancer pathways in the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Supplementary Figure 2A, available online). Similarly, analysis using both the geneSetTest (18) method and the hypergeometric test method confirmed the enrichment for cancer pathways at a two-sided P less than .05 (Supplementary Figure 2, B and C, available online) with a total of eight of the 14 KEGG cancer pathways enriched in two or more analyses.

### Time to Diagnosis

The average time to diagnosis (TTD) for case subjects, defined as the time interval between date of blood draw and date of diagnosis of breast cancer, was 465 days (range = 5–1389). In single CpG case subject–only analysis, only 10 (4.0%) of the 250 dmCpGs were associated with TTD at an unadjusted type 1 error threshold of 0.05, and none reached an FDR Q value threshold of 0.05.

However when we compared the rank ordering of means of the 250 dmCpG methylation values for noncase subjects, case subjects with TTD greater than 1 year (n = 129), and case subjects with TTD within 1 year (n = 169), there was some evidence of effect for TTD: for 179 (71.6%) of the 250 dmCpG sites, case subjects with TTD greater than 1 year had mean methylation values located between that of noncase subjects and case subjects with TTD within 1 year (Figure 2B). This was much higher than the 34% observed for the array as a whole (Fisher exact test, P < .001).

Receiver Operating Characteristic Analysis

Gail model score and GWAS SNPs are established a priori predictors of risk and therefore were evaluated using the entire dataset. The area under the curve (AUC) for Gail score was 56.0%, the AUC for nine GWAS SNPs was 58.8% (Figure 3), and the AUC for Gail plus SNPs was 61.2%.

We examined the performance of the methylation signature in blood DNA by a double-loop cross-validation approach (19), using some of the women as a learning set to identify markers and then evaluating those markers in an independent test set of women. An average of 57 CpGs applied in the independent test sets resulted in an AUC of 65.8% (95% confidence interval [CI] = 61.0% to 70.5%) (Figure 3).

Adding the Gail score or both the Gail score and the nine GWAS SNPs to methylation provided only modest improvement in AUC estimates above those for methylation alone: methylation + Gail score AUC = 65.9% (95% CI = 60.9% to 70.9%) and 5 CpGs + Gail score AUC = 64.7% (95% CI = 61.0% to 71.3%). Using only five CpGs produced slightly lower average AUC (64.1%, 95% CI = 59.7% to 68.0%). Again, there was small improvement in adding to those five CpGs either the Gail score or the SNP data: 5 CpGs + Gail score AUC = 65.6% (95% CI = 61.1% to 69.4%).

We also calculated prediction accuracy using the five most statistically significant dmCpGs (as listed in Supplementary Table 1, available online) with an average of 57 CpGs applied in the independent test sets resulted in an AUC of 65.8% (95% confidence interval [CI] = 61.0% to 70.5%) (Figure 3).

DNA Methylation and Leukocyte Subtypes

We examined a group of 50 CpGs that have been previously reported to be differentially methylated in leukocyte subtypes (16) but found that none were included in our 250 dmCpGs with Q less than or equal to 0.05 and none met the less stringent measure of unadjusted P less than or equal to .05.
available online) in a validation set of 81 women (25 case subjects, 56 noncase subjects) of different ethnicities who were not part of our original analysis. Although none of the individual effect estimates for these five dmCpGs reached statistical significance in the small validation set at \( P \) value threshold of .05, the mean methylation changes were similar to those observed in the main study population (Supplementary Figure 3, available online), and the prediction performance of the validation set (AUC = 63.1%) was very similar to that obtained in the main study population of non-Hispanic whites (AUC = 64.1%).

**Discussion**

We studied DNA methylation at 27578 CpGs in 910 women in a prospective case–cohort analysis of breast cancer using blood collected at enrollment into the Sister Study. We found evidence of differential methylation in blood DNA between women who went on to develop breast cancer compared with women who remained cancer-free. This difference in methylation was evident even among case subject women whose blood sample was collected more than 1 year before diagnosis and became more pronounced in women who provided blood in the year before their diagnosis. The use of prospectively collected samples avoids a number of potential biases that can arise in case–control studies, including case–control differences in sample collection, processing, and storage or changes in case blood samples due to diagnostic, surgical, treatment, or lifestyle effects of having breast cancer. These case–cohort data provide the strongest evidence to date that blood DNA methylation differences exist and are detectable months to years before the clinical diagnosis of breast cancer.

Several case–control studies of breast cancer have investigated global methylation, the average level of methylation in the genome, with mixed results (20–23). A selective survey of 49 genes found six that were differentially methylated, all of which were undermethylated in case subjects (24). A growing number of case–control studies have looked at known breast cancer susceptibility genes, showing differences in BRCA1 promoter methylation (25–27) and differences in ATM methylation, although not at the ATM promoter (23,28).

We found some evidence that known breast cancer susceptibility genes are differentially methylated in our study. CpGs near 18 known breast cancer susceptibility genes were more than twice as likely as other CpGs on the array to have case–cohort differences with unadjusted \( P \) values less than .05 and included CpGs near nine genes: *ATM, BRCA1, CHEK2, FAM84B, FGFR2, MLH1, MSH2, PTEN*, and *TNP1*. However only one of these CpGs, located in *FGFR2*, was statistically significant at FDR \( Q \) value threshold of 0.05. Using genetic pathway analysis, we found evidence that methylation differences were associated with cancer genes: although the KEGG database does not include a specific pathway for breast cancer, we did find that eight of the 14 KEGG cancer pathways were enriched in pathway analyses. It is possible that the methylation differences we observed were due to circulating tumor DNA, but this seems unlikely. The proportion of tumor DNA in circulation (either free tumor DNA or as tumor cells) would be miniscule compared with blood cell DNA and thus, if admixed, would be unlikely to change overall blood methylation values. Secondly, were admixture to be the cause of the methylation change, then the CpGs most likely to be affected would be those with the greatest difference between blood and tumor. However of the 2000 CpGs with the greatest methylation difference between noncase subject blood and breast tumor DNA (from the Cancer Genome Atlas), only six were in the set of 250 dmCpGs.

Differences in methylation profiles might also reflect differences in the proportions of the leukocyte subpopulations that make up whole blood. Recent analyses using methylation arrays

---

**Figure 1.** Manhattan plot of methylation case–cohort association \( P \) values. **Lower dashed horizontal black line** indicates unadjusted \( P \) equal to .05; **upper red dashed line** indicates false discover rate \( Q \) value threshold of 0.05 with 250 CpGs falling above this threshold.
have reported a set of 50 CpGs that differ in methylation level between leukocyte subpopulations (16,29). Many of these leukocyte CpGs were differentially methylated in reanalysis of three published epigenome-wide cancer case–control studies, suggesting that shifts in leukocyte subpopulations might be responsible for the observed differences between case subjects and control subjects (16). However none of these 50 leukocyte markers were differentially methylated in our case–cohort study at FDR $Q < 0.05$ or even at the less stringent threshold of $P < 0.05$.

The differences in methylation that we observed might still be caused by subtle shifts in other circulating cell populations or by micronutrient or DNA methyltransferase activity in response to a growing tumor. Alternatively, the differences in methylation might be caused by prior diet, lifestyle, or environmental factors or heritable traits (30) that in turn reflect risk factors for cancer development. We cannot yet distinguish between these alternatives.

The range of TTD intervals among case subjects allowed us to explore whether methylation was a marker of predisposition to disease or a marker of preclinical disease. A predisposition marker (whether inherited or acquired from exposure) might be independent of TTD, whereas a marker of preclinical disease might strengthen with shorter TTD. Our results were equivocal. Only a few of the 250 dmCpGs individually show statistically significant association with TTD, which would support a susceptibility marker hypothesis, but the weak association evidence could also reflect inadequate sample size and low power to detect small differences. Conversely, the vast majority of the 250 dmCpGs did show increasing difference from control subjects with shorter TTD, providing some support for the early-detection marker hypothesis.

Despite the prevalence of breast cancer, risk prediction remains challenging. The Gail model remains the current standard for population-level prediction but has only modest predictive power for individuals. Family history of breast cancer is an important component of the model, and all women in our study by design have at least one sister with breast cancer. We might expect some consequent loss of predictive accuracy in our cohort because women were more similar to one another for family history than would be found in the general population. But we find that the Gail model...
had an AUC of 56.0%, only slightly lower than the 58.0% reported by Wacholder et al. (4) in the general population. In our study, common GWAS SNPs for breast cancer have modest performance either alone (AUC = 58.8%) or in combination with the Gail model (AUC = 61.2%). Here again our results are in close agreement with general population estimates for GWAS SNPs (AUC = 59.7%) and for GWAS SNPs plus the Gail model (AUC = 61.8%) (4).

Methylation data alone appear to generate better predictions of risk than the Gail model or GWAS SNPs alone, and if verified, this finding is a positive development for breast cancer prediction. The addition of Gail score and SNPs information to methylation data provided little further improvement to predictive precision, perhaps suggesting that the information that Gail score and SNPs provided about risk was already captured within methylation data. It is particularly encouraging that much of the predictive risk information could be ascertained with only five CpGs and that these predictions could be validated in a small set of women with varied ethnicities. Further validation of methylation markers requires that methylation findings be replicated in other studies of both non-Hispanic whites and other ethnic groups.

A particular strength of this study is the prospective collection of samples, with careful uniform processing and storage, which minimizes the possibility that the observed differences between case subjects and noncase subjects were due to diagnostic, treatment, or processing effects.

Our study was not without limitations. It was done within a cohort of women who each had a sister with breast cancer, so the estimated magnitude of association may not be generalizable to women without a similar family history. Comparisons in women without a family history of breast cancer could reveal other, potentially larger differences. The 27578 CpG array used in our study assesses a relatively small number of CpGs near gene transcription start sites. Newer arrays and direct sequencing technologies provide expanded genomic coverage that could reveal additional sites related to breast cancer risk.

References


Funding
This research was supported by the Intramural Research Program of the National Institutes of Health, National Institute of Environmental Health Sciences (Z01 ES044005, Z01 ES044032, and Z01 ES049033). This research was supported by the Intramural Research Program of the National Institutes of Health, National Institute of Environmental Health Sciences or the National Institutes of Health. We express sincere appreciation to all Sister Study participants and the study management group. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Environmental Health Sciences or the National Institutes of Health. We express sincere appreciation to all Sister Study participants and the study management group.

Affiliations of authors: Epidemiology Branch (ZX, LAD, DPS, JAT), Laboratory of Molecular Carcinogenesis (SCEB, JAT), and Biostatistics Branch (CRW), National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, NC.