Human Molecular Genetics, 2020, Vol. 29, No. 4 662–673

doi: 10.1093/hmg/ddz301
Advance Access Publication Date: 16 January 2020
General Article One

GENERAL ARTICLE ONE

Prediagnostic breast milk DNA methylation alterations in women who develop breast cancer

Lucas A. Salas[1,2], Sara N. Lundgren[1,2], Eva P. Browne[3], Elizabeth C. Punska[3], Douglas L. Anderton[4], Margaret R. Karagas[1,2], Kathleen F. Arcaro[3] and Brock C. Christensen[1,5,6,*]

[1]Department of Epidemiology, Geisel School of Medicine at Dartmouth, Hanover, NH 03766, USA, [2]The Children's Environmental Health and Disease Prevention Research Center at Dartmouth, Hanover, NH 03766, USA, [3]Department of Veterinary & Animal Sciences, University of Massachusetts Amherst, Amherst, MA 01003, USA, [4]Department of Sociology, University of South Carolina, Columbus, SC 29208, USA, [5]Department of Molecular and Systems Biology, Geisel School of Medicine at Dartmouth, Hanover, NH 03766, USA and [6]Department of Community and Family Medicine, Geisel School of Medicine at Dartmouth, Hanover, NH 03766, USA

*To whom correspondence should be addressed at: Department of Epidemiology, Geisel School of Medicine at Dartmouth College, 1 Medical Center Drive, Williamson Level 6, HB7650, Lebanon, NH 03766, USA. Tel: +1 6036501828; Fax: +1 6036501840; Email: brock.c.christensen@dartmouth.edu

Abstract

Prior candidate gene studies have shown tumor suppressor DNA methylation in breast milk related with history of breast biopsy, an established risk factor for breast cancer. To further establish the utility of breast milk as a tissue-specific biospecimen for investigations of breast carcinogenesis, we measured genome-wide DNA methylation in breast milk from women with and without a diagnosis of breast cancer in two independent cohorts. DNA methylation was assessed using Illumina HumanMethylation450k in 87 breast milk samples. Through an epigenome-wide association study we explored CpG sites associated with a breast cancer diagnosis in the prospectively collected milk samples from the breast that would develop cancer compared with women without a diagnosis of breast cancer using linear mixed effects models adjusted for history of breast biopsy, age, RefFreeCellMix cell estimates, time of delivery, array chip and subject as random effect. We identified 58 differentially methylated CpG sites associated with a subsequent breast cancer diagnosis (q-value <0.05). Nearly all CpG sites associated with a breast cancer diagnosis were hypomethylated in cases compared with controls and were enriched for CpG islands. In addition, inferred repeat element methylation was lower in breast milk DNA from cases compared to controls, and cases exhibited increased estimated epigenetic mitotic tick rate as well as DNA methylation age compared with controls. Breast milk has utility as a biospecimen for prospective assessment of disease risk, for understanding the underlying molecular basis of breast cancer risk factors and improving primary and secondary prevention of breast cancer.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors, and the last two authors should be regarded as joint senior authors

Received: August 7, 2019. **Revised:** November 30, 2019. **Accepted:** December 6, 2019

Introduction

Breast cancer is the most common non-keratinocyte cancer in women in the USA, with over 270 000 new cases each year (1). Established risk factors for breast cancer include age, reproductive history and family history of disease and can be used to estimate disease risk (2,3). Additionally, and beyond the recognized role of inherited *BRCA* mutation, individual germline genetic variants and even polygenic risk scores from genome-wide association studies have also contributed to breast cancer risk assessment (4–6). Nonetheless, a large gap in the capacity to predict breast cancer risk remains, and the molecular basis of breast cancer risk and carcinogenesis has largely not been studied using target-organ biospecimens from premenopausal women.

Epigenome-wide association studies (EWAS), using surrogate tissues such as peripheral blood DNA, have also had some success testing the relation of DNA methylation with cancer risk (7–9). However, unlike genetic variation and germline alterations that confer cancer risk, cytosine modifications that contribute to cancer risk as disease initiating and promoting events are overwhelmingly tissue-specific. Defining and leveraging knowledge of tissue-specific early DNA methylation alterations for screening or risk models in normal, non-tumor human tissues is challenging for most common tumor types. Yet, use of breast-specific substrate to investigate breast cancer risk has shown promise in early studies measuring cell composition, cytology and candidate gene DNA methylation from nipple aspirate fluid, though as a substrate, nipple aspirate fluid can be challenging to obtain and typically yields very low volume (10–14). Recently, the utility of altered DNA methylation in cancer screening and risk assessment was established in colon cancer as part of the Cologuard multi-target assay where a tissue-specific biospecimen (stool) is obtained and measured without using an invasive procedure (15).

The majority of extensive DNA methylation alterations observed in invasive breast cancer compared with normal breast tissue are already present in pre-invasive disease (16–18). In addition, age-related variation in normal breast tissue DNA methylation has been shown to occur at CpG sites that are more likely to be altered in breast tumors (16), suggesting that early measures of DNA methylation in the pathologically normal breast have value as a biomarker for future breast cancer risk (16). Typically, mammary epithelial cells cannot be accessed without invasive procedures (breast biopsy), lavage or other relatively impractical methods. However, exfoliated mammary epithelial cells (lactocytes, myoepithelial and progenitor/stem cells) are abundant in mature breast milk (as high as 98% of cells) (19,20), providing a tissue-specific substrate obtained without invasive procedure. These cells are an excellent target for biomarker development, and prior candidate gene studies have shown that methylation-induced silencing of tumor suppressor genes in breast milk is related with history of breast biopsy, an established risk factor for breast cancer (21–23). Also as we learn more about human milk as a potential research target, we are now aware of potential sources of variability for these biospecimens (24). Given that 85% of 40-year-old women in the USA have given birth (25), breast milk is a viable non-invasive source of mammary epithelial cells (26). We investigate the relation of early epigenetic alterations with breast cancer risk using cells obtained from breast milk of women who did not develop breast cancer (hereafter named as 'controls') compared with prospectively collected milk specimens from subjects who were later diagnosed with breast cancer.

Table 1. Subject characteristics

Variable	N (%) or mean [range]		P
	Controls ($n = 64$)	Breast cancer ($n = 23$)	
Age (years)	33.2 [23–44]	36.3 [29–45]	0.01
BMI	26.5 [18.2–43.6]	25.2 [18.4–38.7]	0.40
BMI category			0.20
Normal/underweight	28 (43.8)	8 (34.8)	
Overweight/obesity	27 (42.2)	13 (56.5)	
Missing	9 (14.1)	2 (8.7)	
Breast biopsy			<0.001
No	50 (78.1)	0 (0.0)	
Yes	14 (21.9)	23 (100.0)	
Time since delivery (months)	2.2 [0–10]	10.8 [0.2–20]	<0.001
Parity	2 [1–5]	2 [1–4]	<0.001
Milk sample			N/A
Ipsilateral	N/A	16 (69.6)	
Contralateral	N/A	7 (30.4)	
Milk collection			N/A
Pre-diagnosis	N/A	20 (87.0)	
Post-diagnosis	N/A	3 (13.0)	

Results

Genome-scale DNA methylation was measured in breast milk samples from 87 subjects using the Illumina HumanMethylation450 beadchip. Subject demographic and sample details are provided in Table 1. A total of 64 (73%) samples were from cancer-free subjects and 23 were from subjects who had a breast cancer diagnosis of which 20 (87%) were collected prior to diagnosis. Milk samples from subjects with any breast cancer diagnosis were classified according to whether the cancer was in the ipsilateral or contralateral breast. Overall, about 70% of samples from subjects with subsequent breast cancer were collected from the ipsilateral breast (n = 14) and 30% were from the contralateral breast ($n = 6$). Unless stated otherwise, here we summarize the comparison between cancer-free subjects and those with subsequent breast cancer focusing on the ipsilateral breast and exploring the potential effects in the contralateral breast. Results of comparisons between cancer-free subjects with the other three groups are deposited in zenodo (27).

We used a reference-free cell type estimation approach to identify the number of putative cell types and the proportions of each cell type in each breast milk sample. The reference-free method identified five putative cell types in human milk. In unadjusted models, we observed differences in cell type proportions between breast milk samples from women who did not develop breast cancer (controls) compared with those diagnosed with a new breast cancer (ipsilateral or contralateral) for three of the five putative cell types. The proportions of cell types 2 and 3 were higher in subjects with a prospective diagnosis of breast cancer than controls ($P = 5.2E-06$ and 7.1E-04), and the proportion of cell type 4 was lower in milk from subjects with breast cancer compared to controls ($P = 1.2E-05$) (Fig. 1). In these models, differential abundance of putative cell types in controls versus cases was similar irrespective of whether the samples were from the ipsilateral or contralateral breast, or whether the breast cancer diagnosis occurred prior or subsequent to breast milk sample collection (see Additional File, Supplementary Material, Fig. S1). After adjusting for maternal age (years), time since delivery (months) and BeadArray slide number, cell

type proportions were no longer associated with breast cancer diagnosis. We also explored the cell composition of the samples using a modified reference-based hierarchical EpiDISH approach (28,29). The predominant cells were epithelial cells and 'neutrophils' (granulocytes and other related cells) (Additional File, Supplementary Material, Fig. S2).

DNA methylation was compared using linear mixed effect models adjusted for time since delivery in months, maternal age in years, estimated cell type proportions and array chip with subject as a random effect. We explored potential sources of variability between the ipsilateral new breast cancer and our model covariates, the major sources of variation were time since delivery ($r = 0.73$) and age at donation ($r = 0.22$) and putative cell types 2–4 (see Additional File, Supplementary Material, Table S1). We identified 58 significantly differentially methylated CpG sites associated with milk from the ipsilateral breast after correction for multiple comparisons (q-value <0.05). Among these 58 CpGs, two CpGs in island regions and associated one with both the *LRRC61* and *ACTR3C* genes and the other with *SLC18B1* (previously *C6orf192*) were significantly hypermethylated in breast milk from subjects who were later diagnosed with breast cancer (Fig. 2). The remaining 56 CpG sites were significantly hypomethylated in prospectively collected breast milk from the ipsilateral breast of subjects who developed cancer compared with controls (Fig. 2). The most statistically significantly hypomethylated CpG site related to breast cancer diagnosis was located in the island region of the *CLCC1* gene. Additional genes with hypomethylated loci included *TMSB10*, *ZNF584*, *MAP10* (previously *KIAA1383*), *TRIM27* and *SEPTIN7* (previously *SEPT7*). A total of 32 of these CpGs also were hypomethylated in prospectively collected milk from women who developed cancer in the contralateral breast compared to controls (Table 2). The full set of the unadjusted and adjusted EWAS results are deposited in zenodo (27); an overview of the results is available as Additional File, Supplementary Material, Figure S3.

We accessed TCGA breast tumor data using cBioportal to determine whether genes we identified as having hypomethylated CpGs related to breast cancer were associated with gene regulation. We found negative correlations between DNA methylation with mRNA expression z-scores (RNA seq) for many of these genes including *ZNF584* ($P = 2.41E-17$), *MAP10* ($P = 1.61E-76$), *TRIM27* ($P = 6.01E-14$), *LIMD2* ($P = 1.14E-59$) and *LDHA* ($P = 6.06E-06$). In contrast, there was little to no correlation between DNA methylation and expression of *CLCC1* (Spearman $\rho = -0.03$, $P = 0.5$), *TMSB10* ($\rho = -0.08$, $P = 0.07$) and *SEPTIN7* ($\rho = -0.05$, $P = 0.2$), see Additional File, Supplementary Material, Figure S4. The range of DNA methylation level observed for each CpG tested in the TCGA tumors was comparable to that observed in our samples. When we compared the DNA methylation level change between normal adjacent and breast cancer in TCGA, 22 CpGs from our results followed the same direction in TCGA samples including *TRIM27* and *MAP10* (see Additional File, Supplementary Material, Table S2). Of those, seven CpGs were also hypomethylated when comparing normal breast tissue versus breast cancer tissue in the dataset by Teschendorff *et al.* (18).

Given the preponderance of CpG-specific breast milk DNA hypomethylation associated with breast cancer, and that repeat element hypomethylation is well established in cancer, we further assessed repetitive element methylation. To do so, we inferred Alu (37 subfamilies) and LINE-1 (115 subfamilies) DNA methylation using array data and the repetitive element methylation prediction (REMP), as detailed in the Methods section. None of the individual repetitive elements reached statistical significance after multiple comparison correction. The nominally significant are summarized in Additional File, Supplementary Material, Table S3. Mean Alu subfamily methylation was significantly lower in breast cancer cases compared to controls ($\beta = -0.21$, P-value = 2.9E-4), and mean LINE-1 subfamily methylation was also lower in cases than controls ($\beta = -0.073$, P-value = 0.10) (Fig. 3).

To evaluate the location in the genome where breast cancer-related DNA methylation alterations in breast milk were occurring, we performed enrichment analyses for both genomic context and gene sets. Differentially methylated CpGs (q-value <0.05) associated with a subsequent diagnosis of breast cancer were enriched for CpG island regions in milk from both the ipsilateral and contralateral breast (Table 3). Among CpGs whose methylation was significantly related with cancer diagnosis we also tested for enrichment of gene sets using the molecular signatures database (MSigDB) v. 6.2, and identified seven gene sets enriched for the 32 CpG sites that were differentially methylated in both ipsilateral and contralateral samples. The top two pathways are related to highly conserved motif clusters matching transcription factor binding sites (30). Three pathways are related to upregulation of genes in CD8(+) T lymphocytes, T regulatory cells and dendritic cells. Finally, two gene sets are associated to tumor invasion (31) and granulocyte differentiation in acute promyelocytic leukemia (32), see Additional File, Supplementary Material, Table S4.

In univariate linear mixed effect analyses we also tested for DNA methylation age acceleration and elevated epigenetic mitotic clock tick rate (epiTOC) in association with breast cancer status. The epiTOC estimates were significantly higher among breast cancer subjects ($\beta = 0.013$, P-value = 3.2E-04, Fig. 4A); when restricting to the ipsilateral samples the association was still significant ($\beta = 0.017$, P-value = 5.4E-05). A marginal non-statistically significant increase in age acceleration subjects with breast cancer compared to controls was also observed in new cancer diagnoses using samples from any breast ($\beta = 2.7$, P-value = 0.071, Fig. 4B), and when restricting to the ipsilateral samples ($\beta = 2.5$, P-value = 0.1).

Discussion

We identified significant differences in DNA methylation after controlling for cell type and other confounders in subjects with a subsequent diagnosis of breast cancer compared with controls. In the subjects who were diagnosed with breast cancer after the milk collection, nearly all of the significantly differentially methylated CpGs were hypomethylated. Several of the genes whose CpG sites were differentially methylated in prospectively diagnosed cases have previously been associated with breast cancer. For example, *TMSB10* is overexpressed in breast cancer cells, has elevated protein expression in serum of breast cancer patients and is elevated with increasing breast cancer stage and distant metastasis (33). Linking a systemic marker of breast cancer risk to our tissue-specific approach, promoter CpG island hypomethylation of *ZNF584* was associated with a breast cancer diagnosis both here and in peripheral blood DNA from breast cancer patients (34). Further, using TCGA breast tumor data, we showed the functional relationship of *ZNF584* DNA methylation with gene expression. We also observed hypomethylation at CpGs in *SEPTIN7*, *TRIM27*, *LIMD2* and *LDHA*, which have been associated with breast cancer metastasis, invasion and proliferation, (35–38). Apart from *SEPTIN7*, all these genes showed negative correlation between gene expression and DNA methylation

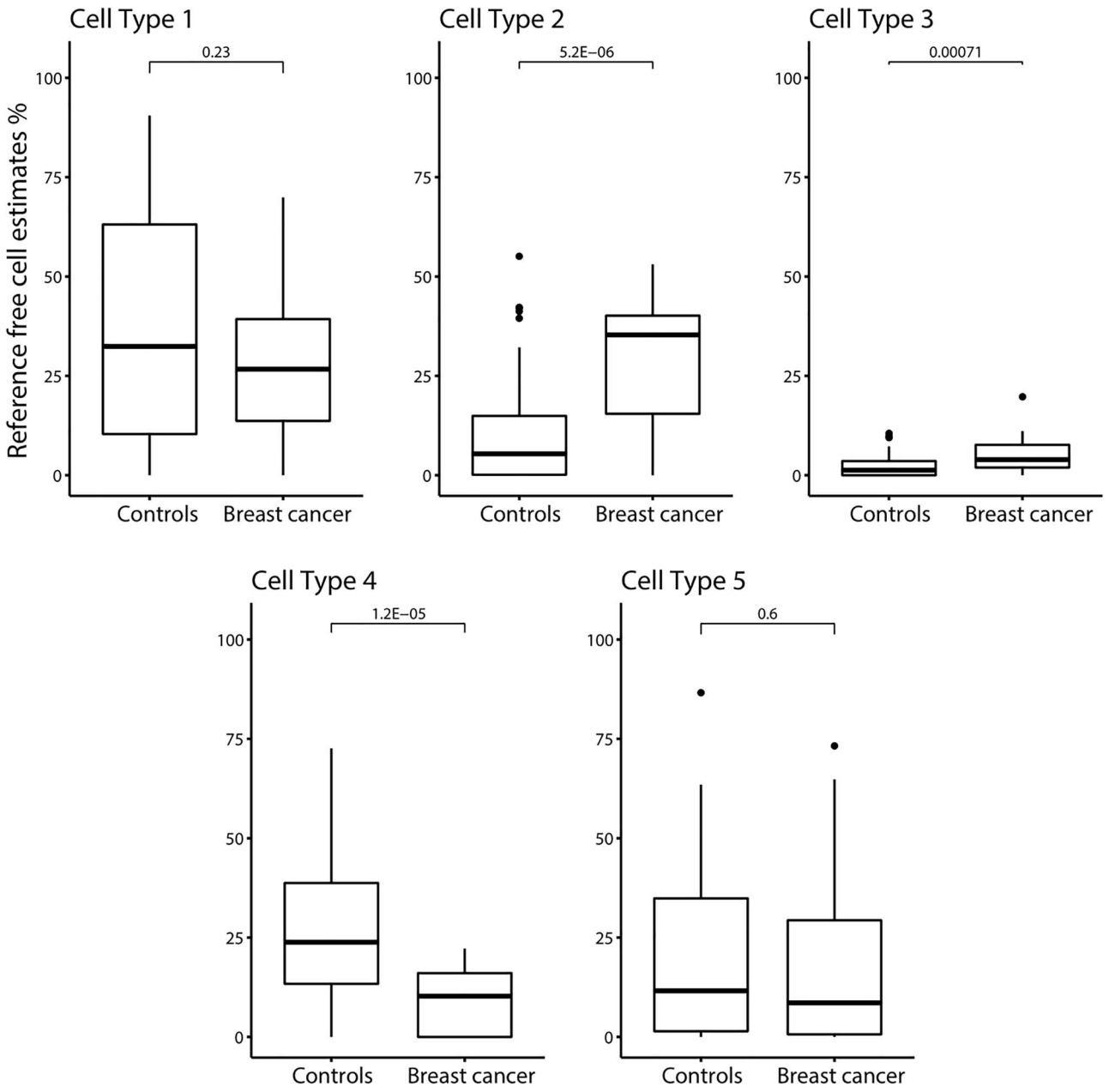

Figure 1. Percentage of reference-free cell estimates in subjects with and without breast cancer. Note: here all the samples from both contralateral ($n = 6$) and ipsilateral breast ($n = 20$) are shown in the graph.

in TCGA breast cancer samples, again demonstrating functional consequences of altered DNA methylation to gene regulation. These results support our hypothesis that epigenetic alterations in human milk have utility for non-invasive molecular assessment of breast cancer risk.

Among subjects with incident breast cancer, the group of hypomethylated CpGs found to be significantly differentially methylated in milk samples from both contralateral and ipsilateral breast compared to those from controls was enriched for CpG island regions. Methylation at CpG island regions can reduce gene expression in associated genes (39). Since the majority of differentially methylated CpGs were hypomethylated, this may correspond to increased expression of genes with promoters in these regions, and consistent with our observations of local and potentially systemic effects, our pathway enrichment analyses identified both proto-oncogene signatures and immune dysregulation signatures. One pathway with strong enrichment is associated with a motif for the *ELK-1* a regulator of the *c-Fos* proto-oncogene which has been linked to growth suppression in breast cancer cells (40). The second pathway includes CpGs related to a motif for *SP-1*, a part of the Kruppel-like family that also has been associated as a prognostic factor in breast cancer (41). Three more pathways pointed to genes upregulated in CD8(+) T lymphocytes, activated T-regulatory cells and dendritic cells, cornerstones of tumor immune response in breast cancer murine models (42). The remaining two pathways were related to tumor invasion and granulocyte differentiation.

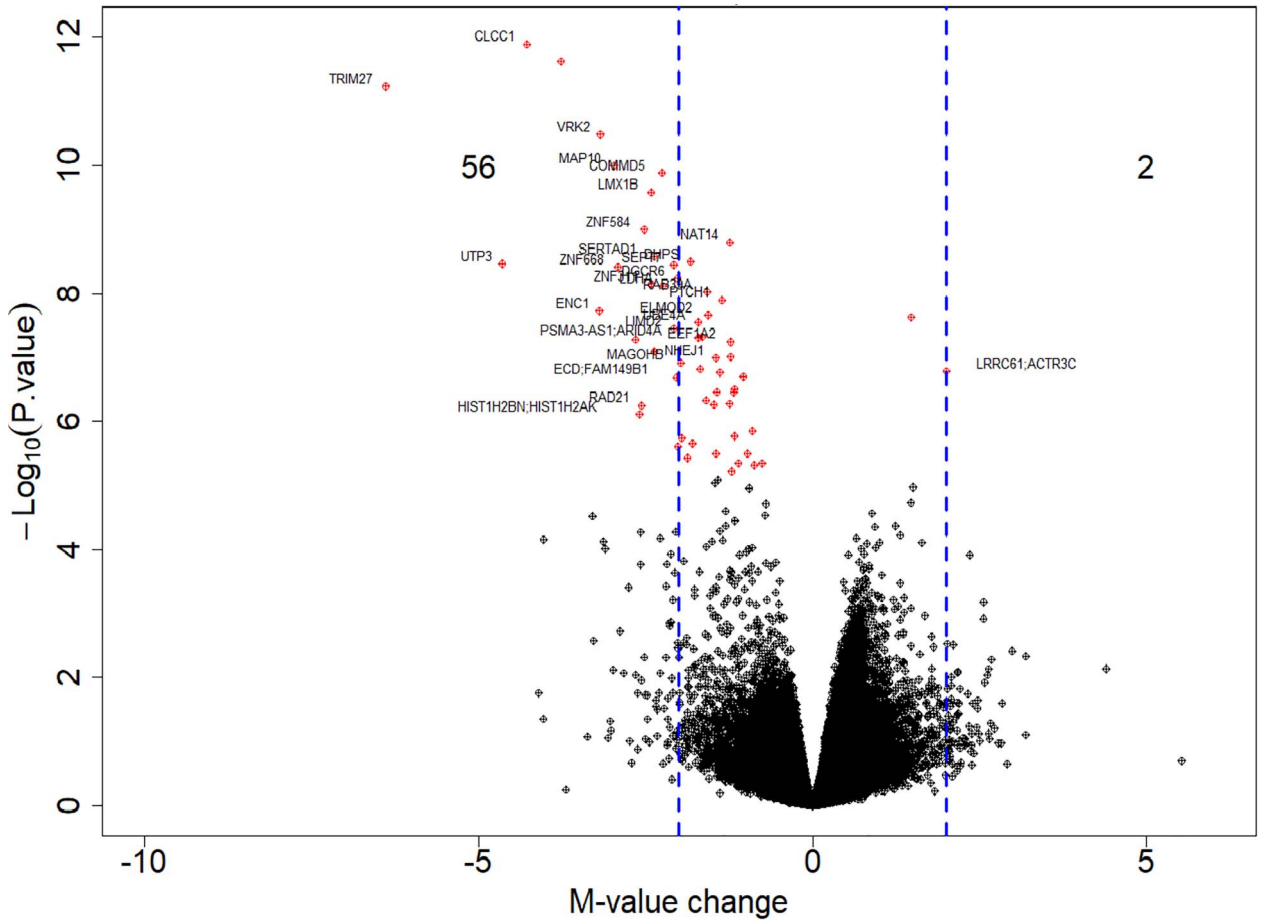

Figure 2. Volcano plots differentially methylated sites in milk from the ipsilateral breast in prospectively diagnosed cancer patients. Note: In red those CpGs that were differentially methylated (q-value <0.05), 56 hypomethylated and 2 hypermethylated. Gene names were added to those CpG sites that overlapped with CpGs differentially methylated in the contralateral breast.

We also observed differences in measures of methylation age including the epiTOC estimator and non-significant differences using the Horvath's methylation age between breast cancer subjects and control subjects. Notably, in our study population, the subjects with a cancer diagnosis were slightly older than control subjects. There have been inconsistent trends of accelerated age in normal breast tissues when using the Horvath methylation age approach (16). Accelerated biologic age inferred using DNA methylation has recently been associated with breast cancer risk in a very large prospective study using peripheral blood (43). However, to date, unlike peripheral blood, there are no DNA methylation clocks for inference of biologic age that are calibrated specifically to biospecimens from the breast. In the future, larger breast-tissue-specific studies are needed to advance our understanding and opportunity to leverage biologic age estimates for breast cancer risk assessment and primary prevention.

This study has several strengths and limitations. Strengths of our study include the use of prospectively collected specimens, tissue-specific measures of DNA methylation and two independent cohorts. Although some subjects potentially had clinically occult disease when providing a milk specimen, others were not diagnosed until years later. Based on a cohort study of women exposed to breast radiation due to benign disease, the latency between the initiating event and the development of a clinical stage ranged between 10 and 30 years (44). In our cohort, the women who later developed breast cancer could have already had an asymptomatic tumor that could be detected in the exfoliated cells in the breast milk. One limitation of this study is sample size, though investigating genome-scale DNA methylation measures in breast milk is novel. One potential limitation is that we pooled controls from two different cohorts processed in different technical batches. Although we controlled for technical differences in our models and used a conservative approach that adjusted for cell estimates which also captures technical differences, we cannot completely exclude some residual technical noise between cohorts affecting our results. In the future, we expect to gather additional information about the cell composition of breast milk samples that will allow the application of improved reference-based approaches (including additional cell subtypes not available in our exploratory approach). This will allow discerning between closely related cell types which currently are collapsed into wider categories (subtypes of epithelial cells and immature/progenitor myeloid cells).

We identified early DNA methylation alterations in breast milk associated with subsequent breast cancer occurrence. These loci were either in genes expressed in breast cancers, related to breast cancer progression or found in peripheral blood samples of women with breast cancer. Importantly, because we identified both overlapping results with work that used peripheral blood as a surrogate biospecimen and results distinct

Table 2. CpG loci that are hypomethylated in breast cancer

CpG ID	Gene	Enhancer	Genomic context	Both breasts[a]
cg00954003	*TMSB10*		Island	Yes
cg22063056	*CLCC1*		Island	Yes
cg04637598		x	Island	Yes
cg19286631	*TRIM27*		Open sea	Yes
cg14399369	*VRK2*		Island	Yes
cg02191044	*MAP10*	x	N-Shore	Yes
cg26421123	*COMMD5*		Island	Yes
cg18453621	*LMX1B*		Island	Yes
cg01221484	*ZNF584*		Island	Yes
cg15698995	*NAT14*		Island	Yes
cg12538369	*SERTAD1*		Island	Yes
cg06363887	*UTP3*		Island	Yes
cg19337593	*DHPS*		Island	Yes
cg21458073	*SEPTIN7*		Island	Yes
cg01996304	*ZNF668*		Island	Yes
cg02014690	*DGCR6*		Island	Yes
cg24104616	*ZNF311*		Open sea	Yes
cg03644271	*LDHA*		Island	Yes
cg09974136	*RAB34*	x	Island	Yes
cg14500569	*PTCH1*		Island	Yes
cg05698228	*ENC1*		Island	Yes
cg09422220	*ELMOD2*		Island	Yes
cg24663984	*UBE4A*	x	Island	Yes
cg02236651	*LIMD2*		Island	Yes
cg08790491	*PSMA3-AS1; ARID4A*		Island	Yes
cg20923184		x	Open sea	Yes
cg20605045	*SFXN4*	x	Island	No
cg14610853	*EEF1A2*		S-Shelf	Yes
cg24471039	*RAB3GAP1*		Open sea	No
cg09827701	*USP19*		Island	No
cg06952862	*NHEJ1*		S-Shore	Yes
cg19570943	*MAGOHB*		Island	Yes
cg16400434	*PPME1; C2CD3*		Island	No
cg01228243	*GPAT4*		Island	No
cg12276298	*SEM1; FAM149B1*		Island	Yes
cg26973266	*TRAF4*		Island	No
cg00496455	*TUBA4A*		Island	No
cg18522266	*SMARCA4*		Island	No
cg19584875	*KCNK13*		Island	No
cg20287461	*TMEM102*		N-Shore	No
cg06094142		x	Open sea	No
cg26292521	*GATA3-AS1; GATA3*	x	Island	No
cg09523472	*RAD21*		Island	Yes
cg16914272	*H2BC15; H2AC15*		Island	Yes
cg04422896	*C12orf43*		Island	No
cg04193422	*PON2*		Island	No
cg19483159	*DYNLT1*		S-Shore	No
cg05677943		x	Open sea	No
cg14328761			Open sea	No
cg10291648	*TIRAP*		Island	No
cg25977304	*POU2F1*		Island	No
cg21077559	*TMEM155; PP12613*		Island	No
cg03243700	*WDR5*		Island	No
cg04799218	*LPCAT3*		Island	No
cg07496106	*GPT2*		Island	No
cg24717401	*CCM2*		Island	No

[a]Hypomethylation statistically significant on both ipsilateral and contralateral breast milk samples compared to non-cancer controls or not. For those marked as no, the hypomethylation was significant for the ipsilateral sample only.

to breast milk, we expect that our tissue-specific approach has high potential for follow-up work. We expect that future investigations of DNA methylation changes present in cells from breast milk from disease-free women will have value for risk assessment and primary prevention of breast cancer, perhaps with specific strength in application to premenopausal disease. However, larger studies are needed to validate our findings and to further establish the utility of breast milk as

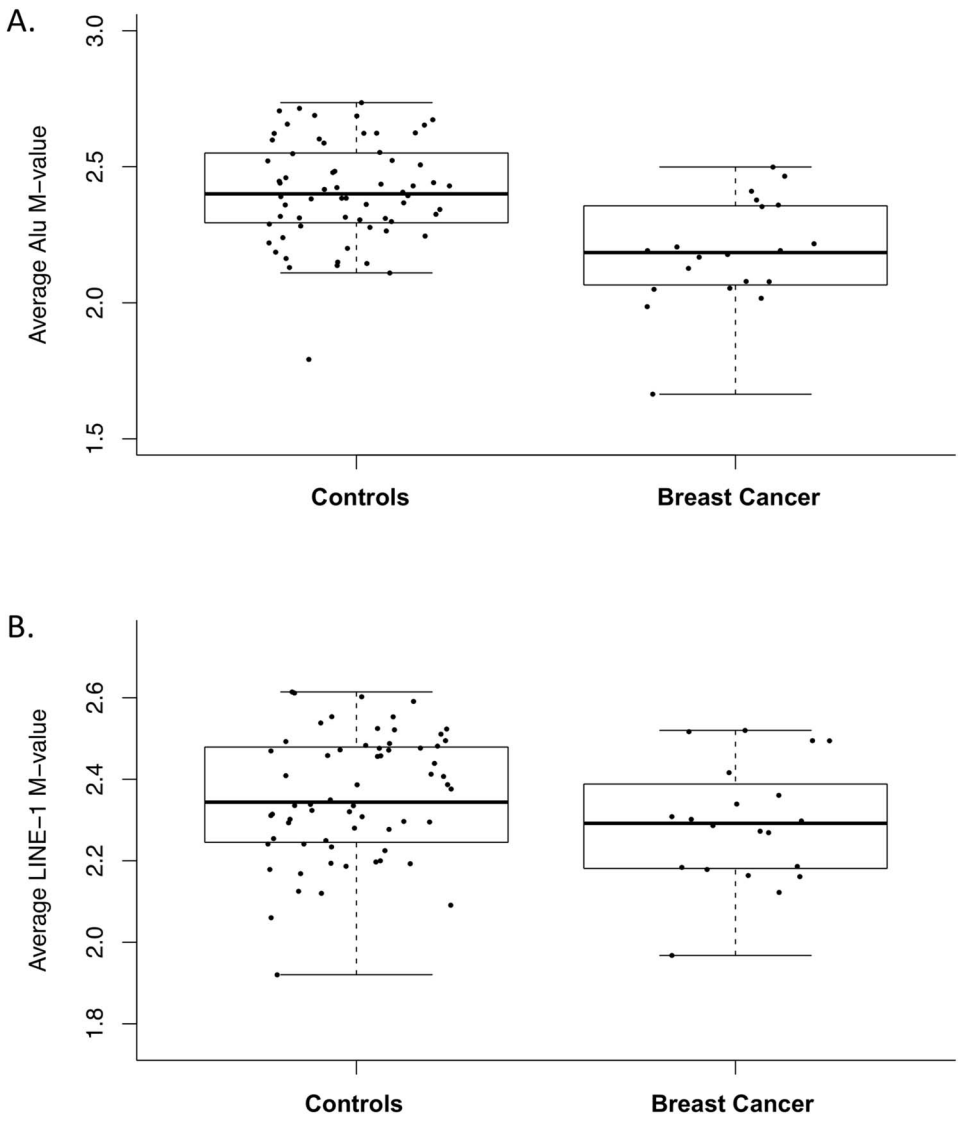

Figure 3. Differences in repetitive element CpG methylation by breast cancer status. Note: here all the samples from both contralateral ($n = 6$) and ipsilateral breast ($n = 20$) are shown in the graph.

Table 3. Enrichment for genomic context in CpGs with $q < 0.05$

Breast cancer group[a]	Island regions		Enhancer regions	
	OR (95% CI)	P[b]	OR (95% CI)	P[b]
Ipsilateral	3.48 (1.75, 7.45)	9.3E-05	1.05 (0.45, 2.18)	8.5E-01
Contralateral	4.28 (1.64, 13.30)	8.6E-04	1.01 (0.30, 2.67)	1.0E+00

[a]Reference level is controls with no breast cancer history.
[b]P determined using the Cochran–Mantel–Haenszel test.

a biospecimen for understanding the molecular basis of disease risk and prospective risk assessment.

In conclusion, we assessed genome-wide DNA methylation in breast milk from subjects with and without breast cancer; specific loci were hypomethylated in breast cancer subjects compared to control subjects. These differentially methylated regions were more likely to occur in island regions of the genome. Our results suggest that breast milk has utility for prospective assessment of breast cancer risk.

Materials and Methods

Study population

Two different study populations were included in this study: (1) women from the 'Molecular Biomarkers for Assessing Breast-Cancer Risk' project at the University of Massachusetts Amherst (UMass) and (2) participants of the New Hampshire Birth Cohort Study (NHBCS) at Dartmouth College. UMass subjects were women older than 18 years. They were either lactating or

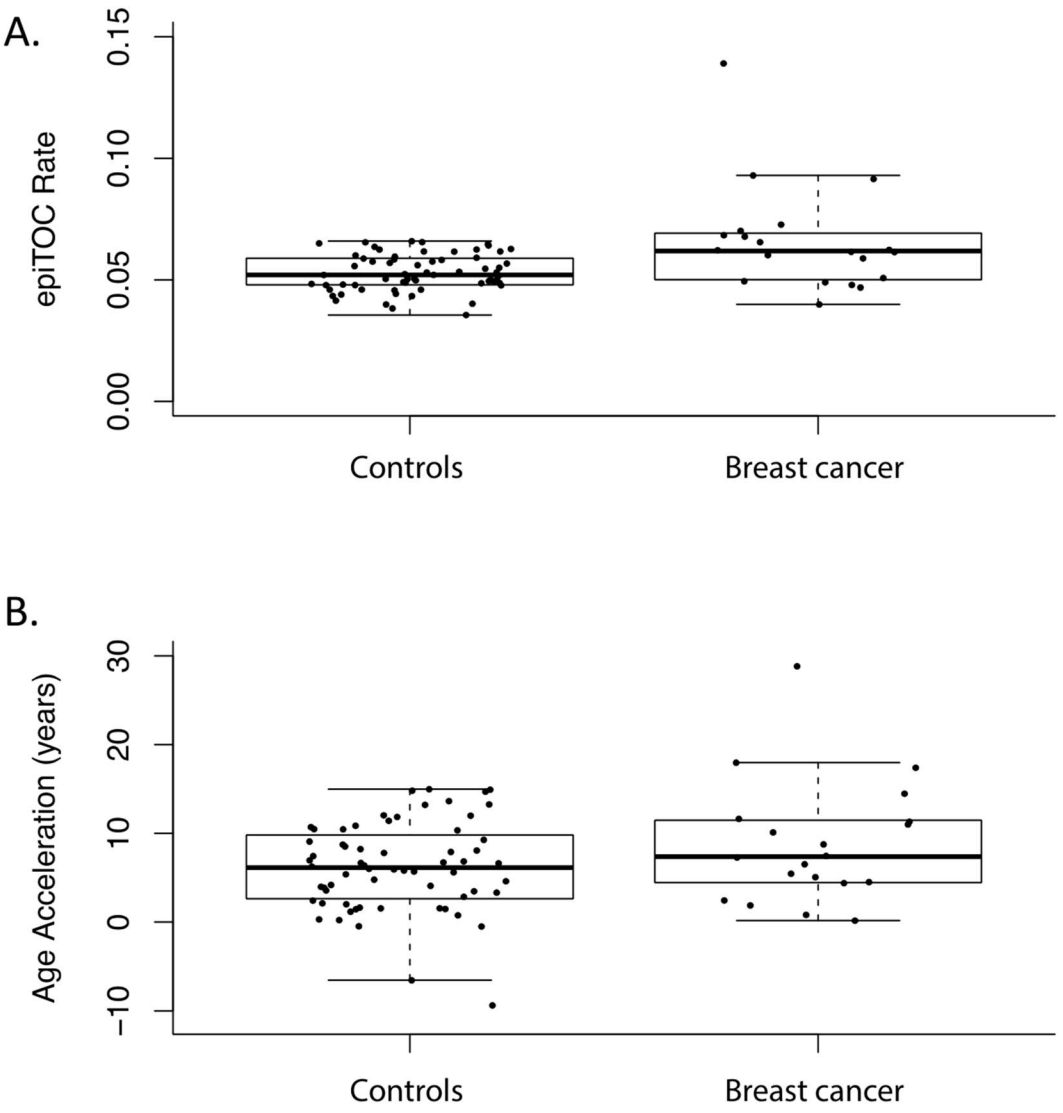

Figure 4. Measures of age inferred from methylation values. (A) Epigenetic mitotic clock tick rate (epiTOC) between controls and subjects who later developed breast cancer. (B) Age acceleration (difference between chronological and DNA methylation age) between controls and subjects who later developed breast cancer. Note: Here all the samples from both contralateral ($n = 6$) and ipsilateral breast ($n = 20$) are shown in the graph.

have recently given birth, and they had a history of either breast biopsy or breast cancer. UMass subjects were asked to provide one or two breast milk samples expressed in a single pumping session. NHBCS participant characteristics have been described previously (45). Briefly, NHBCS eligibility criteria included: English speaking, literate and mentally competent women carrying a singleton pregnancy, 18–45 years of age and whose primary source of residential water was a private well. Women who planned to move during pregnancy were excluded from this study. NHBCS participants were asked to bring bilateral breast milk samples to the postpartum follow-up appointment. All study participants provided written informed consent prior to the study according to the guidelines of Institutional Review Board of the University of Massachusetts Amherst and the Committee for the Protection of Human Subjects at Dartmouth. Women in both studies were asked to complete a questionnaire about general health, reproductive health and personal breast biopsy and breast cancer history. Each woman's samples were classified into five different groups: (1) no breast cancer history, (2) healthy breast, contralateral breast cancer before donation, (3) ipsilateral breast cancer diagnosis before donation, (4) healthy breast contralateral cancer diagnosis after donation and (5) sample from the ipsilateral breast with cancer after donation. For this analysis, we report the results of model milk samples from control subjects and from subjects with a subsequent diagnosis of breast cancer.

Sample collection

Using a previously described method (26), breast milk was processed within 24 h of sample collection to obtain DNA. Briefly, DNA was extracted from 1 to 10 ml of milk from each breast and stored at −20°C until DNA extraction.

DNA extraction and genome-wide DNA methylation array

DNA was isolated using the Qiagen DNeasy Blood and Tissue Kit (Qiagen, Valencia, CA) and bisulfite converted using the EZ DNA Methylation kit (Zymo, Irvine, CA). Samples were

randomized across several plates and subsequently subjected to epigenome-wide DNA methylation assessment using Illumina Infinium HumanMethylation450 BeadChip, which measured ~485 000 CpG sites genome-wide (Illumina, San Diego, CA). Microarrays were processed at USC core facility following standard protocols. The data were assembled using GenomeStudio methylation software (Illumina) without normalization as per the manufacturer's instructions. The methylation status for each individual CpG locus (β-value) was calculated as the ratio of fluorescent signals ($\beta = \text{Max(M,0)}/[\text{Max(M,0)} + \text{Max(U,0)} + 100]$), ranging from 0 (no methylation) to 1 (complete methylation) using the average probe intensity for the methylated (M) and unmethylated (U) alleles. We read the idat files using the minfi R package (46). β-values were background corrected using methyluminoob and normalized using functional normalization.(47) Our pipeline included array control probes to assess sample quality and evaluate potential problems such as poor bisulfite conversion or color-specific issues for each array as described previously (48,49). All CpG loci on X and Y chromosomes, CpH and loci with potential problems of cross-reactivity, tracking to polymorphisms with minor allele frequencies over 5% for the general population, or common copy number alterations,(50) were excluded from the analysis, leaving 368 171 autosomal CpG loci in 92 samples. Principal components analysis and multiple dimension scaling were used to identify potential technical batches. Additionally, we used a principal component regression analysis to investigate the top eight principal components in relation to potential batch-associated differences. Subjects with missing covariate data were excluded from modeling, resulting in 87 samples. DNA methylation β-values were logit_2 transformed to M-values for the analyses (51).

Cell mixture analysis

Given the lack of cell-specific DNA methylation references for most of the breast epithelial cell subtypes being interrogated (lactocytes, myoepithelial, progenitor cells, among others) (20), we were unable to apply a reference-based approach to cell type deconvolution in the EWAS models. Instead, to identify and adjust for potential cell type heterogeneity in the breast milk samples, we used a reference-free decomposition (RefFreeCellMix) of the DNA methylation matrix into cell-type distributions and cell-type methylomes, using the expression $Y = M * \Omega^T$ (52). We explored a range of k cell types from 2 to 10. Note that the decomposition will be based on Y, but Yfinal (=Y by default) was used to determine the final value of M based on the last iterated value of Ω. We explored the global cell composition of the samples using a modified hierarchical EpiDISH approach. As references we used the epithelial, fibroblast and immune cell matrix included in EpiDISH and for the immune cell composition we used the legacy 450 k library included in FlowSorted.Blood.EPIC (28,29), cell proportions were estimated through robust partial correlations.

Locus-by-locus analysis for detecting differentially methylated CpG loci

We implemented a locus-by-locus analysis to identify differentially methylated CpG sites between samples obtained from control subjects without breast cancer diagnosis and those from healthy and diseased breasts before or after the cancer development using the R package *limma* (53). Five groups were compared: (1) Controls with no breast cancer history, (2) Contralateral Prior Diagnosis (sample from healthy breast of a woman previously diagnosed breast cancer), (3) Ipsilateral Prior Diagnosis (sample from affected breast of a woman previously diagnosed breast cancer), (4) Contralateral New Diagnosis (sample from healthy breast of a woman with incident breast cancer) and (5) Ipsilateral New Diagnosis (sample from affected breast of a woman with incident breast cancer). Briefly, linear mixed effects models were fit to each CpG site separately, with the CpG β-value as the response against the five groups. A random effect for subject was included to control for within subject correlation in subjects with bilateral samples (30 subjects). The models were adjusted for time from delivery (in months), maternal age (in years), RefFreeCellMix proportion estimates (five putative cell types), and the microarray slide to control residual batch confounding. *P*-values were adjusted for multiple comparisons by computing the Benjamini–Hochberg *q*-values (54), and we defined loci with *q*-value < 0.05 to be statistically significant. For this analysis, we focus on CpGs identified as differentially methylated in both prospectively diagnosed groups (ipsilateral and contralateral) and report individual group results in supplemental material (27). All analyses were carried out using the R statistical package, version 3.5.0 (Vienna, Austria; www.r-project.org/) (55). We accessed TCGA breast tumor data using cBioportal to determine whether genes we identified as having hypomethylated CpGs related to breast cancer were also associated with gene regulation. Finally, we used public data from TCGA (774 breast cancer samples and 97 normal adjacent breast tissue) and from Teschendorff *et al.* (305 breast cancer samples and 50 normal breast tissue, deposited in GEO, GSE69914) (18) to explore whether the CpGs observed in our analyses were following the same directionality when comparing normal breast tissue (or normal-adjacent tissue for TCGA) versus breast cancer tissue.

Repetitive element prediction and analysis

We use the package REMP (56) to estimate the DNA methylation levels on both *LINE-1* and *Alu* transposons using the information from the DNA methylation microarray. This random forest approach covers 37 *Alu* subfamilies and 115 *LINE-1* subfamilies. We computed the average *Alu* and *LINE-1* methylation levels for each sample, and tested the association with prospectively diagnosed breast cancer, excluding the three samples from subjects with a prior diagnosis of breast cancer. *P*-values were computed using the Kenward–Roger approach.

Enrichment analyses

The probes that were differentially methylated were tested for pathway and gene set enrichment using missMethyl (57) and the MSigDB v.6.2 curated database (58). A minimum of two genes were required for further exploring the specific pathway. We also tested for over- or underrepresentation of differentially methylated CpGs identified in the locus-by-locus analysis in (1) enhancer regions and (2) CpG island regions. Loci with a *q*-value <0.05 were considered to be statistically significant. Odds ratios, 95% confidence intervals and *P*-values were computed with the Cochran–Mantel–Haenszel test and were adjusted for probe type.

Predicted methylation age and stem cell divisions

We used Horvath's DNA methylation age estimation algorithm (59) to calculate predicted methylation age (mAge) using the agep function from wateRmelon (60). Using those estimates, age acceleration was defined as: Age acceleration = mAge − Age. We tested for differences in age acceleration between control

subjects and subjects with breast cancer using a linear mixed effects model to control for within subject correlations. *P*-values were calculated using the Kenward–Roger approach. Additionally, stem cell divisions were estimated using the epiTOC method (61), but only 334 of 385 CpGs were available to calculate estimates. epiTOC estimates were compared between cases and controls using unadjusted linear mixed effect models analogously to the age acceleration models.

Supplementary Material

Supplementary material is available at *HMG* online.

Availability of data and material

The datasets generated and analyzed during the current study are available in the GEO (https://www.ncbi.nlm.nih.gov/geo/) under the accession number GSE133918. We also used the public dataset by Teschendorff *et al.* available under the accession number GSE69914, and the data from TCGA available through https://gdac.broadinstitute.org/. The EWAS results (unadjusted and adjusted for confounding covariates) are deposited in Zenodo (https://doi.org/10.5281/zenodo.3362478).

Funding

National Institutes of General Medical Sciences [Centers of Biomedical Research Excellence (COBRE) Center for Molecular Epidemiology at Dartmouth P20GM104416 to M.R.K. and B.C.C.]; National Cancer Institute [R01CA216265 to B.C.C., R01CA230478-01A1 to K.F.A.]; National Institute of Environmental Health Sciences [P01ES022832 to M.R.K.]; Environmental Protection Agency [RD-83544201-1 to M.R.K.]; Congressionally Directed Medical Research Program [W81XWH-08-1-0721 to K.F.A.].

Authors' contributions

L.A.S. and S.N.L. elaborated the analysis plan, analyzed the data and wrote the first draft of the manuscript. E.P.B., E.C.P., D.L.A. and M.R.K. provided technical and methodological feedback to the original analysis and final version and K.F.A. and B.C.C. generated the original idea and codirected the analyses. All authors approved the final version of this manuscript.

Conflict of interest statement

None declared.

References

1. Siegel, R.L., Miller, K.D. and Jemal, A. (2019) Cancer statistics, 2019. *CA. Cancer J. Clin.*, **69**, 7–34.
2. National Cancer Institute (2018) Breast Cancer Risk Assessment Tool. https://bcrisktool.cancer.gov/calculator.html (accessed October 12, 2018).
3. Gail, M.H., Brinton, L.A., Byar, D.P., Corle, D.K., Green, S.B., Schairer, C. and Mulvihill, J.J. (1989) Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J. Natl. Cancer Inst.*, **81**, 1879–1886.
4. Zhang, X., Rice, M., Tworoger, S.S., Rosner, B.A., Eliassen, A.H., Tamimi, R.M., Joshi, A.D., Lindstrom, S., Qian, J., Colditz, G.A. *et al.* (2018) Addition of a polygenic risk score, mammographic density, and endogenous hormones to existing breast cancer risk prediction models: a nested case-control study. *PLoS Med.*, **15**, e1002644.
5. Wang, S., Qian, F., Zheng, Y., Ogundiran, T., Ojengbede, O., Zheng, W., Blot, W., Nathanson, K.L., Hennis, A., Nemesure, B. *et al.* (2018) Genetic variants demonstrating flip-flop phenomenon and breast cancer risk prediction among women of African ancestry. *Breast Cancer Res. Treat.*, **168**, 703–712.
6. Khera, A.V., Chaffin, M., Aragam, K.G., Haas, M.E., Roselli, C., Choi, S.H., Natarajan, P., Lander, E.S., Lubitz, S.A., Ellinor, P.T. *et al.* (2018) Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.*, **50**, 1219–1224.
7. Li, L., Zheng, H., Huang, Y., Huang, C., Zhang, S., Tian, J., Li, P., Sood, A.K., Zhang, W. and Chen, K. (2017) DNA methylation signatures and coagulation factors in the peripheral blood leucocytes of epithelial ovarian cancer. *Carcinogenesis*, **38**, 797–805.
8. Tang, Q., Holland-Letz, T., Slynko, A., Cuk, K., Marme, F., Schott, S., Heil, J., Qu, B., Golatta, M., Bewerunge-Hudler, M. *et al.* (2016) DNA methylation array analysis identifies breast cancer associated RPTOR, MGRN1 and RAPSN hypomethylation in peripheral blood DNA. *Oncotarget*, **7**, 64191–64202.
9. Baglietto, L., Ponzi, E., Haycock, P., Hodge, A., Bianca Assumma, M., Jung, C.-H., Chung, J., Fasanelli, F., Guida, F., Campanella, G. *et al.* (2017) DNA methylation changes measured in pre-diagnostic peripheral blood samples are associated with smoking and lung cancer risk. *Int. J. cancer*, **140**, 50–61.
10. King, E.B., Barrett, D. and Petrakis, N.L. (1975) Cellular composition of the nipple aspirate specimen of breast fluid. II. Abnormal findings. *Am. J. Clin. Pathol.*, **64**, 739–748.
11. Krassenstein, R., Sauter, E., Dulaimi, E., Battagli, C., Ehya, H., Klein-Szanto, A. and Cairns, P. (2004) Detection of breast cancer in nipple aspirate fluid by CpG island hypermethylation. *Clin. Cancer Res.*, **10**, 28–32.
12. Tice, J.A., Miike, R., Adduci, K., Petrakis, N.L., King, E. and Wrensch, M.R. (2005) Nipple aspirate fluid cytology and the Gail model for breast cancer risk assessment in a screening population. *Cancer Epidemiol. Biomarkers Prev.*, **14**, 324–328.
13. Wrensch, M.R., Petrakis, N.L., Gruenke, L.D., Ernster, V.L., Miike, R., King, E.B. and Hauck, W.W. (1990) Factors associated with obtaining nipple aspirate fluid: analysis of 1428 women and literature review. *Breast Cancer Res. Treat.*, **15**, 39–51.
14. Zhu, W., Qin, W., Hewett, J.E. and Sauter, E.R. (2010) Quantitative evaluation of DNA hypermethylation in malignant and benign breast tissue and fluids. *Int. J. Cancer*, **126**, 474–482.
15. Imperiale, T.F., Ransohoff, D.F., Itzkowitz, S.H., Levin, T.R., Lavin, P., Lidgard, G.P., Ahlquist, D.A. and Berger, B.M. (2014) Multitarget stool DNA testing for colorectal-cancer screening. *N. Engl. J. Med.*, **370**, 1287–1297.
16. Johnson, K.C., Houseman, E.A., King, J.E. and Christensen, B.C. (2017) Normal breast tissue DNA methylation differences at regulatory elements are associated with the cancer risk factor age. *Breast Cancer Res.*, **19**, 81.
17. Fleischer, T., Frigessi, A., Johnson, K.C., Edvardsen, H., Touleimat, N., Klajic, J., Riis, M.L., Haakensen, V.D., Wärnberg, F., Naume, B. *et al.* (2014) Genome-wide DNA methylation profiles in progression to in situ and invasive carcinoma of the breast with impact on gene transcription and prognosis. *Genome Biol.*, **15**, 435.
18. Teschendorff, A.E., Gao, Y., Jones, A., Ruebner, M., Beckmann, M.W., Wachter, D.L., Fasching, P.A. and Widschwendter, M. (2016) DNA methylation outliers in normal breast tissue

identify field defects that are enriched in cancer. *Nat. Commun.*, **7**, 10478.
19. Hassiotou, F., Hepworth, A.R., Metzger, P., Tat Lai, C., Trengove, N., Hartmann, P.E. and Filgueira, L. (2013) Maternal and infant infections stimulate a rapid leukocyte response in breastmilk. *Clin. Transl. Immunol.*, **2**, e3.
20. Witkowska-Zimny, M. and Kaminska-El-Hassan, E. (2017) Cells of human breast milk. *Cell. Mol. Biol. Lett.*, **22**, 11.
21. Wong, C.M., Anderton, D.L., Smith-Schneider, S., Wing, M.A., Greven, M.C. and Arcaro, K.F. (2010) Quantitative analysis of promoter methylation in exfoliated epithelial cells isolated from breast milk of healthy women. *Epigenetics*, **5**, 645–655.
22. Browne, E.P., Dinc, S.E., Punska, E.C., Agus, S., Vitrinel, A., Erdag, G.C., Anderton, D.L., Arcaro, K.F. and Yilmaz, B. (2014) Promoter methylation in epithelial-enriched and epithelial-depleted cell populations isolated from breast milk. *J. Hum. Lact.*, **30**, 450–457.
23. Browne, E.P., Punska, E.C., Lenington, S., Otis, C.N., Anderton, D.L. and Arcaro, K.F. (2011) Increased promoter methylation in exfoliated breast epithelial cells in women with a previous breast biopsy. *Epigenetics*, **6**, 1425–1435.
24. Davis Lynn, B.C., Bodelon, C., Pfeiffer, R.M., Yang, H.P., Yang, H.H., Lee, M., Laird, P.W., Campan, M., Weisenberger, D.J., Murphy, J. *et al.* (2019) Differences in genome-wide DNA methylation profiles in breast milk by race and lactation duration. *Cancer Prev. Res.*, **12**, 781–790.
25. Martinez, G., Daniels, K. and Chandra, A. (2012) Fertility of men and women aged 15-44 years in the United States: National Survey of family growth, 2006-2010. *Natl. Health Stat. Report.*, **51**, 1–28.
26. Murphy, J., Sherman, M.E., Browne, E.P., Caballero, A.I., Punska, E.C., Pfeiffer, R.M., Yang, H.P., Lee, M., Yang, H., Gierach, G.L. *et al.* (2016) Potential of breastmilk analysis to inform early events in breast carcinogenesis: rationale and considerations. *Breast Cancer Res. Treat.*, **157**, 13–22.
27. Salas, L.A., Lundgren, S.N., Browne, E.P., Punska, E.C., Anderton, D.L., Karagas, M.R., Arcaro, K.F. and Christensen, B.C. (2019) EWAS results "Prediagnostic breast milk DNA methylation alterations in women who develop breast cancer". [Data set]. *Zenodo*. http://doi.org/10.5281/zenodo.3362478.
28. Zheng, S.C., Webster, A.P., Dong, D., Feber, A., Graham, D.G., Sullivan, R., Jevons, S., Lovat, L.B., Beck, S., Widschwendter, M. *et al.* (2018) A novel cell-type deconvolution algorithm reveals substantial contamination by immune cells in saliva, buccal and cervix. *Epigenomics*, **10**, 925–940.
29. Salas, L.A., Koestler, D.C., Butler, R.A., Hansen, H.M., Wiencke, J.K., Kelsey, K.T. and Christensen, B.C. (2018) An optimized library for reference-based deconvolution of whole-blood biospecimens assayed using the Illumina HumanMethylationEPIC BeadArray. *Genome Biol.*, **19**, 64.
30. Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S. and Kellis, M. (2005) Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, **434**, 338–345.
31. Teramoto, H., Castellone, M.D., Malek, R.L., Letwin, N., Frank, B., Gutkind, J.S. and Lee, N.H. (2005) Autocrine activation of an osteopontin-CD44-Rac pathway enhances invasion and transformation by H-RasV12. *Oncogene*, **24**, 489–501.
32. Park, D.J., Vuong, P.T., de Vos, S., Douer, D. and Koeffler, H.P. (2003) Comparative analysis of genes regulated by PML/RAR alpha and PLZF/RAR alpha in response to retinoic acid using oligonucleotide arrays. *Blood*, **102**, 3727–3736.
33. Zhang, X., Ren, D., Guo, L., Wang, L., Wu, S., Lin, C., Ye, L., Zhu, J., Li, J., Song, L. *et al.* (2017) Thymosin beta 10 is a key regulator of tumorigenesis and metastasis and a novel serum marker in breast cancer. *Breast Cancer Res.*, **19**, 15.
34. Khakpour, G., Noruzinia, M., Izadi, P., Karami, F., Ahmadvand, M., Heshmat, R., Amoli, M.M. and Tavakkoly-Bazzaz, J. (2017) Methylomics of breast cancer: seeking epimarkers in peripheral blood of young subjects. *Tumour Biol.*, **39**, 1010428317695040.
35. Dong, T., Liu, Z., Xuan, Q., Wang, Z., Ma, W. and Zhang, Q. (2017) Tumor LDH-A expression and serum LDH status are two metabolic predictors for triple negative breast cancer brain metastasis. *Sci. Rep.*, **7**, 6069.
36. Peng, H., Talebzadeh-Farrooji, M., Osborne, M.J., Prokop, J.W., McDonald, P.C., Karar, J., Hou, Z., He, M., Kebebew, E., Orntoft, T. *et al.* (2014) LIMD2 is a small LIM-only protein overexpressed in metastatic lesions that regulates cell motility and tumor progression by directly binding to and activating the integrin-linked kinase. *Cancer Res.*, **74**, 1390–1403.
37. Zoumpoulidou, G., Broceño, C., Li, H., Bird, D., Thomas, G. and Mittnacht, S. (2012) Role of the tripartite motif protein 27 in cancer development. *J. Natl. Cancer Inst.*, **104**, 941–952.
38. Zhang, N., Liu, L., Fan, N., Zhang, Q., Wang, W., Zheng, M., Ma, L., Li, Y. and Shi, L. (2016) The requirement of SEPT2 and SEPT7 for migration and invasion in human breast cancer via MEK/ERK activation. *Oncotarget*, **7**, 61587–61600.
39. Deaton, A.M. and Bird, A. (2011) CpG islands and the regulation of transcription. *Genes Dev.*, **25**, 1010–1022.
40. Chai, Y., Chipitsyna, G., Cui, J., Liao, B., Liu, S., Aysola, K., Yezdani, M., Reddy, E.S. and Rao, V.N. (2001) C-Fos oncogene regulator Elk-1 interacts with BRCA1 splice variants BRCA1a/1b and enhances BRCA1a/1b-mediated growth suppression in breast cancer cells. *Oncogene*, **20**, 1357–1367.
41. Hedrick, E., Cheng, Y., Jin, U.-H., Kim, K. and Safe, S. (2016) Specificity protein (Sp) transcription factors Sp1, Sp3 and Sp4 are non-oncogene addiction genes in cancer cells. *Oncotarget*, **7**, 22245–22256.
42. Goudin, N., Chappert, P., Mégret, J., Gross, D.-A., Rocha, B. and Azogui, O. (2016) Depletion of regulatory T cells induces high numbers of dendritic cells and unmasks a subset of anti-tumour CD8+CD11c+ PD-1lo effector T cells. *PLoS One*, **11**, e0157822.
43. Kresovich, J. K., Xu, Z., O'Brien, K. M., Weinberg, C. R., Sandler, D. P. and Taylor, J. A. (2019) Methylation-based biological age and breast cancer risk. *J. Natl. Cancer Inst.*, **111**, 1051–1058.
44. Olsson, H., Baldetorp, B., Fernö, M. and Perfekt, R. (2003) Relation between the rate of tumour cell proliferation and latency time in radiation associated breast cancer. *BMC Cancer*, **3**, 11.
45. Gilbert-Diamond, D., Cottingham, K.L., Gruber, J.F., Punshon, T., Sayarath, V., Gandolfi, A.J., Baker, E.R., Jackson, B.P., Folt, C.L. and Karagas, M.R. (2011) Rice consumption contributes to arsenic exposure in US women. *Proc. Natl. Acad. Sci. U. S. A.*, **108**, 20656–20660.
46. Aryee, M.J., Jaffe, A.E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A.P., Hansen, K.D. and Irizarry, R.A. (2014) Minfi: a flexible and comprehensive bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*, **30**, 1363–1369.
47. Fortin, J., Labbe, A., Lemire, M., Zanke, B.W., Hudson, T.J., Fertig, E.J., Greenwood, C.M. and Hansen, K.D. (2014) Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biol.*, **15**, 503.

48. Cardenas, A., Koestler, D.C., Houseman, E.A., Jackson, B.P., Kile, M.L., Karagas, M.R. and Marsit, C.J. (2015) Differential DNA methylation in umbilical cord blood of infants exposed to mercury and arsenic in utero. *Epigenetics*, **10**, 508–515.
49. Koestler, D.C., Avissar-Whiting, M., Houseman, E.A., Karagas, M.R. and Marsit, C.J. (2013) Differential DNA methylation in umbilical cord blood of infants exposed to low levels of arsenic in utero. *Environ. Health Perspect.*, **121**, 971–977.
50. Zhou, W., Laird, P.W. and Shen, H. (2017) Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Res.*, **45**, e22.
51. Du, P., Zhang, X., Huang, C.-C., Jafari, N., Kibbe, W.A., Hou, L. and Lin, S.M. (2010) Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, **11**, 587.
52. Houseman, E.A., Kile, M.L., Christiani, D.C., Ince, T.A., Kelsey, K.T. and Marsit, C.J. (2016) Reference-free deconvolution of DNA methylation data and mediation by cell composition effects. *BMC Bioinformatics*, **17**, 259.
53. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W. and Smyth, G.K. (2015) Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47.
54. Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B*, **57**, 289–300.
55. R Core Team (2019) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
56. Zheng, Y., Joyce, B.T., Liu, L., Zhang, Z., Kibbe, W.A., Zhang, W. and Hou, L. (2017) Prediction of genome-wide DNA methylation in repetitive elements. *Nucleic Acids Res.*, **45**, 8697–8711.
57. Phipson, B., Maksimovic, J. and Oshlack, A. (2016) missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. *Bioinformatics*, **32**, 286–288.
58. Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P. and Mesirov, J.P. (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**, 1739–1740.
59. Horvath, S. (2013) DNA methylation age of human tissues and cell types. *Genome Biol.*, **14**, –R115.
60. Pidsley, R., Wong, Y., C, C., Volta, M., Lunnon, K., Mill, J. and Schalkwyk, L.C. (2013) A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics*, **14**, 293.
61. Yang, Z., Wong, A., Kuh, D., Paul, D.S., Rakyan, V.K., Leslie, R.D., Zheng, S.C., Widschwendter, M., Beck, S. and Teschendorff, A.E. (2016) Correlation of an epigenetic mitotic clock with cancer risk. *Genome Biol.*, **17**, 205.