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†], Brock C. Christensen[1,5,6†**]

[1]Department of Epidemiology, Geisel School of Medicine at Dartmouth, Hanover, NH

[2]The Children's Environmental Health and Disease Prevention Research Center at Dartmouth, Hanover, NH

[3]Department of Veterinary & Animal Sciences, University of Massachusetts Amherst, Amherst, MA

[4] Department of Sociology, University of South Carolina, Columbus, SC

[5]Department of Molecular and Systems Biology, Geisel School of Medicine at Dartmouth, Hanover, NH

[6]Department of Community and Family Medicine, Geisel School of Medicine at Dartmouth, Hanover, NH

*Equal contributors as first authors

[†]Equal contributors as senior authors

**Corresponding author: BCC, Brock.C.Christensen@Dartmouth.Edu

Keywords: Human Milk, DNA methylation, Breast cancer, premenopause

ABSTRACT

Background: 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.

Methods: DNA methylation was assessed using Illumina HumanMethylation450k in 87 breast milk samples. After quality control, 368,171 autosomal CpG loci were analyzed. Cell type proportion estimates from RefFreeCellMix were calculated and adjusted for in this Epigenome Wide Association Study using linear mixed effects models adjusted for history of breast biopsy, age, time of delivery, cell type proportion estimates, array chip, and subject as random effect.

Results: Epigenome-wide analyses identified 58 differentially methylated 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 (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.

Conclusion: 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.

BACKGROUND

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, nontumor 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], [17]. 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 [18], suggesting that early measures of DNA methylation in the pathologically normal breast has value as a biomarker for future breast cancer risk [18]. Typically, mammary epithelial cells cannot be accessed without invasive procedures (breast biopsy), lavage, or other relatively impractical methods. However, exfoliated mammary epithelial cells are abundant in breast milk [19], 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 [20–22]. Given that 85% of 40 year-old women in the USA have given birth [23], breast milk is a viable noninvasive source of mammary epithelial cells [24]. We investigate the relation of early epigenetic alterations with breast cancer risk using cells obtained from breast milk in controls compared with prospectively collected milk specimens from subjects who were later diagnosed with breast cancer.

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 of age. They were either lactating or 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 participants characteristics has been described previously [25]. 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 [24], breast milk was processed within 24 hours of sample collection to obtain DNA. Briefly, DNA was extracted from 1 - 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, San Diego, CA) without normalization per the manufacturer's instructions. The methylation status for each individual CpG locus (β-value) was calculated as the ratio of fluorescent signals (β = Max(M,0)/ [Max(M,0) + 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 [26]. β-values were background corrected using methylumi-noob and normalized using functional normalization.[27] 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 [28,29]. 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,[30] 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 [31].

Cell mixture analysis

In order 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 $Y = M*\Omega^T$ [32]. 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)

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* [33]. 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 (5 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 [34], and we defined loci with q-value < 0.05 to be statistically significant. We focus on CpGs identified as differentially methylated in both prospectively diagnosed groups (ipsilateral and contralateral), and report individual group results in supplemental material. All analyses were carried out using the R statistical package, version 3.5.0 (Vienna, Austria; www.r-project.org/) [35].

Repetitive elements prediction and analysis

We use the package REMP, Repetitive Element Methylation Prediction [36], 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 [37] and the MSigDB v.6.2 curated database [38]. 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 [39] to calculate predicted methylation age ($mAge$) using the agep function from wateRmelon [40]. 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. *P*-values were calculated using the Kenward-Roger approach. Additionally, stem cell divisions were estimated using the epiTOC method [41], 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.

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**. 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.

193 Milk samples from subjects with any breast cancer diagnosis were classified according to
194 whether the cancer was in the ipsilateral or contralateral breast. Overall, about 70% of samples
195 from subjects with subsequent breast cancer were collected from the ipsilateral breast (n=14)
196 and 30% were from the contralateral breast (n=6).

197 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-diagnostic	N/A	20 (87.0)	
Post diagnosis	N/A	3 (13.0)	

198

199 We used a reference-free cell type estimation approach to identify the number of putative cell
200 types and the proportions of each cell type in each breast milk sample. The reference-free
201 method identified five putative cell types in human milk. In unadjusted models, we observed
202 differences in cell type proportions between breast milk samples from women who did not
203 developed breast cancer (henceforth named as “controls”) compared with those diagnosed with
204 breast cancer for three of the five putative cell types. The proportions of cell types 2 and 3 were
205 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) (**Figure 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 figure **Additional File 1**). 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.

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 identified 57 significantly differentially methylated CpG sites associated with milk from the ipsilateral breast after correction for multiple comparisons (*q-value* < 0.05). Among these 57 CpGs, one CpG in an island region and associated with both the *LRRC61* and *ACTR3C* genes was significantly hypermethylated in breast milk from subjects who were later diagnosed with breast cancer (**Figure 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 (**Figure 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 EWAS results are available as **Additional Files 2 and 3**.

230

231

232

233 Table 2. CpG loci that are hypomethylated in breast cancer

CpG ID	Gene	Enhancer	Genomic Context
cg22063056	CLCC1		Island
cg00954003	TMSB10		Island
cg01221484	ZNF584		Island
cg02014690	DGCR6		Island
cg02191044	MAP10	x	North Shore
cg04637598		x	Island
cg05698228	ENC1		Island
cg14399369	VRK2		Island
cg18453621	LMX1B		Island
cg19286631	TRIM27		Open Sea
cg21458073	SEPTIN7		Island
cg26421123	COMMD5		Island
cg01996304	ZNF668; ZNF646		Island
cg02236651	LIMD2		Island
cg03644271	LDHA		Island
cg06363887	UTP3		Island
cg06952862	NHEJ1		South Shore
cg08790491	PSMA3-AS1; ARID4A		Island
cg09422220	ELMOD2		Island
cg09523472	RAD21		Island
cg09974136	RAB34	x	Island
cg12276298	ECD; FAM149B1		Island
cg12538369	SERTAD1		Island
cg14500569	PTCH1		Island
cg14610853	EEF1A2		South Shelf
cg15698995	NAT14		Island

cg16914272	HIST1H2BN; HIST1H2AK		Island
cg19337593	DHPS		Island
cg19570943	MAGOHB		Island
cg20923184		x	Open Sea
cg24104616	ZNF311		Open Sea
cg24663984	UBE4A	x	Island

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 ρ=-0.03, *P*=0.5), *TMSB10* (ρ = -0.08, P=0.07) and *SEPTIN7* (ρ =-0.05, *P*=0.2), see **Additional File 4**. The range of DNA methylation level observed for each CpG tested in the TCGA tumors was comparable to that observed in our samples.

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 5, Table S5**. Mean Alu subfamily methylation was significantly lower in breast cancer cases compared to controls (β = -0.21, *p*-value = 2.9E-4), and mean LINE-1 subfamily methylation was also lower in cases than controls (β = -0.073, *p*-value = 0.10) (**Figure 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 7 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 [42]. 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 [43] and granulocyte differentiation in acute promyelocytic leukemia [44] , see **Additional File 5, Table S6**.

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

Breast Cancer Group[2]	**Island Regions**		**Enhancer Regions**	
	OR (95% CI)	***P***[1]	**OR (95% CI)**	***P***[1]
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

[1]*P* determined using the Cochran-Mantel-Haenszel test

[2]Reference level is controls with no breast cancer history

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 amongst breast cancer subjects (β = 0.013, *p*-value = 3.2E-04, **Figure 4a**). A marginally significant increase in age acceleration subjects with breast cancer compared to controls was also observed (β = 2.7, *p*-value = 0.071, **Figure 4b**). After adjusting for covariates, neither epiTOC estimates nor methylation age were related to

275 breast cancer status. In the adjusted analyses, putative cell type proportions were related to
276 epiTOC estimates and age acceleration.

277 **DISCUSSION**

278 We identified significant differences in DNA methylation after controlling for cell type and other
279 confounders in subjects with a subsequent diagnosis of breast cancer compared with controls.
280 In the subjects who were diagnosed with breast cancer after the milk collection, nearly all of the
281 significantly differentially methylated CpGs were hypomethylated. Several of the genes whose
282 CpG sites were differentially methylated in prospectively diagnosed cases have previously been
283 associated with breast cancer. For example, *TMSB10* is overexpressed in breast cancer cells,
284 has elevated protein expression in serum of breast cancer patients, and is elevated with
285 increasing breast cancer stage and distant metastasis [45]. Linking a systemic marker of breast
286 cancer risk to our tissue-specific approach, promoter CpG island hypomethylation of *ZNF584*
287 was associated with a breast cancer diagnosis both here and in peripheral blood DNA from
288 breast cancer patients [46]. Further, using TCGA breast tumor data, we showed the functional
289 relationship of *ZNF584* DNA methylation with gene expression. We also observed
290 hypomethylation at CpGs in *SEPTIN7*, *TRIM27*, *LIMD2*, and *LDHA,* which have been
291 associated with breast cancer metastasis, invasion, and proliferation, [47–50]. Apart from
292 *SEPTIN7*, all these genes showed negative correlation between gene expression and DNA
293 methylation in TCGA breast cancer samples, again demonstrating functional consequences of
294 altered DNA methylation to gene regulation. These results support our hypothesis that
295 epigenetic alterations in human milk have utility for noninvasive molecular assessment of breast
296 cancer risk.

297 Amongst subjects with incident breast cancer, the group of hypomethylated CpGs found to be
298 significantly differentially methylated in milk samples from both contralateral and ipsilateral
299 breast compared to those from controls were enriched for CpG island regions. Methylation at

CpG island regions can reduce gene expression in associated genes [51]. 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* protooncogene which has been linked to growth suppression in breast cancer cells [52]. 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 [53]. 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 [54]. The remaining two pathways were related to tumor invasion and granulocyte differentiation.

We also observed differences in measures of methylation age including the epiTOC estimator and Horvath's methylation age between breast cancer subjects and control subjects, but these associations were not robust to adjustment for potential confounders. Notably, in our study population the subjects with a cancer diagnosis were slightly older than control subjects. Putative cell type proportions, however, were related to all measures of methylation age. Although CpG loci utilized in each algorithm are not supposed to be dependent on cell type, there have been consistent trends of accelerated age in breast tissues when using the Horvath methylation age approach. Accelerated biologic age inferred using DNA methylation has recently been associated with breast cancer risk in a very large prospective study using peripheral blood [55]. However, to date, unlike peripheral blood, there are no DNA methylation clocks for inference of biologic age that are calibrated 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 were potentially had clinically occult disease when providing a milk specimen, others were not diagnosed until years later. 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.

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 women with breast cancer. Importantly, because we identified both overlapping results with work that used peripheral blood as a surrogate biospecimen and results distinct 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 a biospecimen for understanding the molecular basis of disease risk and prospective risk assessment.

Conclusions

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.

List of abbreviations:

Epigenome-wide association studies (EWAS), epigenetic mitotic clock tick rate (epiTOC), molecular signatures database (MSigDB), University of Massachusetts Amherst (UMass), New Hampshire Birth Cohort study (NHBCS), methylation age ($mAge$).

Declarations

Ethics approval and consent to participate: 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.

Consent for publication: Not applicable

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.

Competing interests: The authors declare that they have no competing interests

Funding: This work was supported by funds of the COBRE Center for Molecular Epidemiology at Dartmouth P20GM104416, and R01CA216265 to BCC; NIEHS P01ES022832 and EPA RD83544201 to MRK; and R01CA230478-01A1 to KFA

Authors' contributions: LAS and SNL elaborated the analysis plan, analyzed the data and wrote the first draft of the manuscript, EPB, ECP, DLA, MRK provided technical and methodological feedback to the original analysis and final version, KFA and BCC generated the original idea and codirected the analyses. All authors approved the final version of this manuscript.

Acknowledgements: Not applicable.

Figures titles:

Figure 1. Percentage of reference-free cell estimates in subjects with and without breast cancer

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 1 hypermethylated. Gene names were added to those CpG sites that overlapped with CpGs differentially methylated in the contralateral breast.

Figure 3. Differences in repetitive element CpG methylation by breast cancer status

Figure 4. Measures of age inferred from methylation values

Additional files:

Additional File 1: Figure S1. Percentages of reference-free cell estimates by breast cancer group

Additional File 2: Table S1: EWAS results contralateral breast milk new cancer diagnosis after milk donation. **Table S2:** EWAS results ipsilateral breast milk new breast cancer diagnosis after milk donation. **Table S3:** EWAS results contralateral breast milk previous breast cancer. **Table S4:** EWAS results ipsilateral breast milk previous breast cancer.

Additional File 3: Figure S2. Volcano plots differentially methylated sites in breast milk according to breast source cancer status

Additional File 4: Figure S3. Correlation between gene expression (mRNA) and DNA methylation in breast tumor samples from TCGA

Additional File 5: Table S5. LINE-1 element methylation in realtion to breast cancer diagnosis (P-value < 0.01). **Table S6.** Molecular Signatures Database (MSigDB) pathways enriched using missMethyl

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA Cancer J Clin. 2019;69:7–34.

2. National Cancer Institute. Breast Cancer Risk Assessment Tool [Internet]. 2018 [cited 2018 Oct 12]. Available from: https://bcrisktool.cancer.gov/calculator.html

3. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst. 1989;81:1879–86.

4. Zhang X, Rice M, Tworoger SS, Rosner BA, Eliassen AH, Tamimi RM, et al. 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. 2018;15:e1002644.

5. Wang S, Qian F, Zheng Y, Ogundiran T, Ojengbede O, Zheng W, et al. Genetic variants demonstrating flip-flop phenomenon and breast cancer risk prediction among women of African ancestry. Breast Cancer Res Treat. Springer US; 2018;168:703–12.

6. Khera A V., Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet. Springer US; 2018;50:1219–24.

7. Li L, Zheng H, Huang Y, Huang C, Zhang S, Tian J, et al. DNA methylation signatures and coagulation factors in the peripheral blood leucocytes of epithelial ovarian cancer. Carcinogenesis. 2017;38:797–805.

8. Tang Q, Holland-Letz T, Slynko A, Cuk K, Marme F, Schott S, et al. DNA methylation array analysis identifies breast cancer associated RPTOR, MGRN1 and RAPSN hypomethylation in peripheral blood DNA. Oncotarget. 2016;7:64191–202.

9. Baglietto L, Ponzi E, Haycock P, Hodge A, Bianca Assumma M, Jung C-H, et al. DNA methylation changes measured in pre-diagnostic peripheral blood samples are associated with smoking and lung cancer risk. Int J cancer. 2017;140:50–61.

10. King EB, Barrett D, Petrakis NL. Cellular composition of the nipple aspirate specimen of breast fluid. II. Abnormal findings. Am J Clin Pathol. 1975;64:739–48.

11. Krassenstein R, Sauter E, Dulaimi E, Battagli C, Ehya H, Klein-Szanto A, et al. Detection of breast cancer in nipple aspirate fluid by CpG island hypermethylation. Clin Cancer Res. 2004;10:28–32.

12. Tice JA, Miike R, Adduci K, Petrakis NL, King E, Wrensch MR. Nipple aspirate fluid cytology and the Gail model for breast cancer risk assessment in a screening population. Cancer Epidemiol Biomarkers Prev. 2005;14:324–8.

13. Wrensch MR, Petrakis NL, Gruenke LD, Ernster VL, Miike R, King EB, et al. Factors associated with obtaining nipple aspirate fluid: analysis of 1428 women and literature review. Breast Cancer Res Treat. 1990;15:39–51.

14. Zhu W, Qin W, Hewett JE, Sauter ER. Quantitative evaluation of DNA hypermethylation in malignant and benign breast tissue and fluids. Int J cancer. 2010;126:474–82.

15. Imperiale TF, Ransohoff DF, Itzkowitz SH, Levin TR, Lavin P, Lidgard GP, et al. Multitarget

439 stool DNA testing for colorectal-cancer screening. N Engl J Med. 2014;370:1287–97.

440 16. Johnson KC, Koestler DC, Fleischer T, Chen P, Jenson EG, Marotti JD, et al. DNA
441 methylation in ductal carcinoma in situ related with future development of invasive breast
442 cancer. Clin Epigenetics. 2015;7:75.

443 17. Fleischer T, Frigessi A, Johnson KC, Edvardsen H, Touleimat N, Klajic J, et al. Genome-
444 wide DNA methylation profiles in progression to in situ and invasive carcinoma of the breast with
445 impact on gene transcription and prognosis. Genome Biol. 2014;15:435.

446 18. Johnson KC, Houseman EA, King JE, Christensen BC. Normal breast tissue DNA
447 methylation differences at regulatory elements are associated with the cancer risk factor age.
448 Breast Cancer Res. 2017;19:81.

449 19. Witkowska-Zimny M, Kaminska-El-Hassan E. Cells of human breast milk. Cell Mol Biol Lett.
450 2017;22:11.

451 20. Wong CM, Anderton DL, Smith-Schneider S, Wing MA, Greven MC, Arcaro KF. Quantitative
452 analysis of promoter methylation in exfoliated epithelial cells isolated from breast milk of healthy
453 women. Epigenetics. 2010;5:645–55.

454 21. Browne EP, Dinc SE, Punska EC, Agus S, Vitrinel A, Erdag GC, et al. Promoter methylation
455 in epithelial-enriched and epithelial-depleted cell populations isolated from breast milk. J Hum
456 Lact. 2014;30:450–7.

457 22. Browne EP, Punska EC, Lenington S, Otis CN, Anderton DL, Arcaro KF. Increased
458 promoter methylation in exfoliated breast epithelial cells in women with a previous breast
459 biopsy. Epigenetics. 2011;6:1425–35.

460 23. Martinez G, Daniels K, Chandra A. Fertility of men and women aged 15-44 years in the
461 United States: National Survey of Family Growth, 2006-2010. Natl Health Stat Report.

462 2012;51:1–28.

463 24. Murphy J, Sherman ME, Browne EP, Caballero AI, Punska EC, Pfeiffer RM, et al. Potential
464 of breastmilk analysis to inform early events in breast carcinogenesis: rationale and
465 considerations. Breast Cancer Res Treat. Springer US; 2016;157:13–22.

466 25. Gilbert-Diamond D, Cottingham KL, Gruber JF, Punshon T, Sayarath V, Gandolfi AJ, et al.
467 Rice consumption contributes to arsenic exposure in US women. Proc Natl Acad Sci U S A.
468 2011;108:20656–60.

469 26. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al.
470 Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA
471 methylation microarrays. Bioinformatics. 2014;30:1363–9.

472 27. Fortin J, Labbe A, Lemire M, Zanke BW, Hudson TJ, Fertig EJ, et al. Functional
473 normalization of 450k methylation array data improves replication in large cancer studies.
474 Genome Biol. 2014;15:503.

475 28. Cardenas A, Koestler DC, Houseman EA, Jackson BP, Kile ML, Karagas MR, et al.
476 Differential DNA methylation in umbilical cord blood of infants exposed to mercury and arsenic
477 in utero. Epigenetics. 2015;10:508–15.

478 29. Koestler DC, Avissar-Whiting M, Houseman EA, Karagas MR, Marsit CJ. Differential DNA
479 methylation in umbilical cord blood of infants exposed to low levels of arsenic in utero. Environ
480 Health Perspect. 2013;121:971–7.

481 30. Zhou W, Laird PW, Shen H. Comprehensive characterization, annotation and innovative use
482 of Infinium DNA methylation BeadChip probes. Nucleic Acids Res. 2017;45:e22.

483 31. Du P, Zhang X, Huang C-C, Jafari N, Kibbe WA, Hou L, et al. Comparison of Beta-value and
484 M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics.

2010;11:587.

32. Houseman EA, Kile ML, Christiani DC, Ince TA, Kelsey KT, Marsit CJ. Reference-free deconvolution of DNA methylation data and mediation by cell composition effects. BMC Bioinformatics. 2016;17:259.

33. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:e47.

34. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B. 1995;57:289–300.

35. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2017. Available from: http://www.r-project.org/

36. Zheng Y, Joyce BT, Liu L, Zhang Z, Kibbe WA, Zhang W, et al. Prediction of genome-wide DNA methylation in repetitive elements. Nucleic Acids Res. Oxford University Press; 2017;45:8697–711.

37. Phipson B, Maksimovic J, Oshlack A. missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. Bioinformatics. 2016;32:286–8.

38. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011;27:1739–40.

39. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol. 2013;14:R115.

40. Pidsley R, Y Wong CC, Volta M, Lunnon K, Mill J, Schalkwyk LC. A data-driven approach to preprocessing Illumina 450K methylation array data. BMC Genomics. 2013;14:293.

507 41. Yang Z, Wong A, Kuh D, Paul DS, Rakyan VK, Leslie RD, et al. Correlation of an epigenetic
508 mitotic clock with cancer risk. Genome Biol. 2016;17:205.

509 42. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, et al. Systematic discovery
510 of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals.
511 Nature. 2005;434:338–45.

512 43. Teramoto H, Castellone MD, Malek RL, Letwin N, Frank B, Gutkind JS, et al. Autocrine
513 activation of an osteopontin-CD44-Rac pathway enhances invasion and transformation by H-
514 RasV12. Oncogene. 2005;24:489–501.

515 44. Park DJ, Vuong PT, de Vos S, Douer D, Koeffler HP. Comparative analysis of genes
516 regulated by PML/RAR alpha and PLZF/RAR alpha in response to retinoic acid using
517 oligonucleotide arrays. Blood. 2003;102:3727–36.

518 45. Zhang X, Ren D, Guo L, Wang L, Wu S, Lin C, et al. Thymosin beta 10 is a key regulator of
519 tumorigenesis and metastasis and a novel serum marker in breast cancer. Breast Cancer Res.
520 2017;19:15.

521 46. Khakpour G, Noruzinia M, Izadi P, Karami F, Ahmadvand M, Heshmat R, et al. Methylomics
522 of breast cancer: Seeking epimarkers in peripheral blood of young subjects. Tumour Biol.
523 2017;39:1010428317695040.

524 47. Dong T, Liu Z, Xuan Q, Wang Z, Ma W, Zhang Q. Tumor LDH-A expression and serum LDH
525 status are two metabolic predictors for triple negative breast cancer brain metastasis. Sci Rep.
526 Springer US; 2017;7:6069.

527 48. Peng H, Talebzadeh-Farrooji M, Osborne MJ, Prokop JW, McDonald PC, Karar J, et al.
528 LIMD2 is a small LIM-only protein overexpressed in metastatic lesions that regulates cell motility
529 and tumor progression by directly binding to and activating the integrin-linked kinase. Cancer

530 Res. 2014;74:1390–403.

531 49. Zoumpoulidou G, Broceño C, Li H, Bird D, Thomas G, Mittnacht S. Role of the tripartite motif
532 protein 27 in cancer development. J Natl Cancer Inst. 2012;104:941–52.

533 50. Zhang N, Liu L, Fan N, Zhang Q, Wang W, Zheng M, et al. The requirement of SEPT2 and
534 SEPT7 for migration and invasion in human breast cancer via MEK/ERK activation. Oncotarget.
535 2016;7:61587–600.

536 51. Deaton AM, Bird A. CpG islands and the regulation of transcription. Genes Dev.
537 2011;25:1010–22.

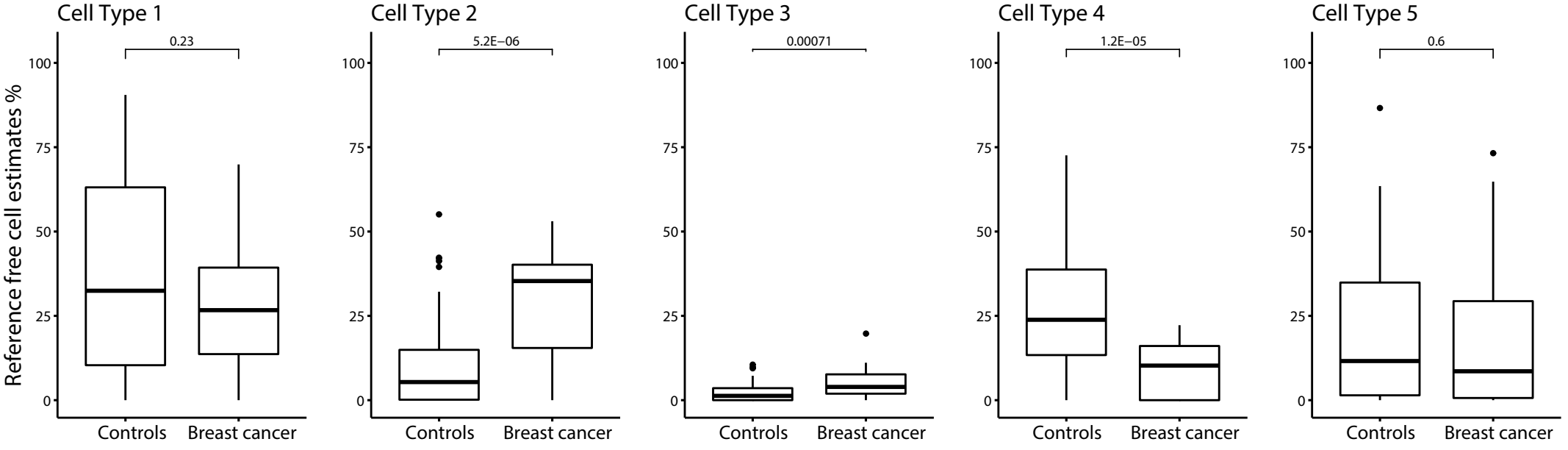
538 52. Chai YL, Chipitsyna G, Cui J, Liao B, Liu S, Aysola K, et al. c-Fos oncogene regulator Elk-1
539 interacts with BRCA1 splice variants BRCA1a/1b and enhances BRCA1a/1b-mediated growth
540 suppression in breast cancer cells. Oncogene. 2001;20:1357–67.

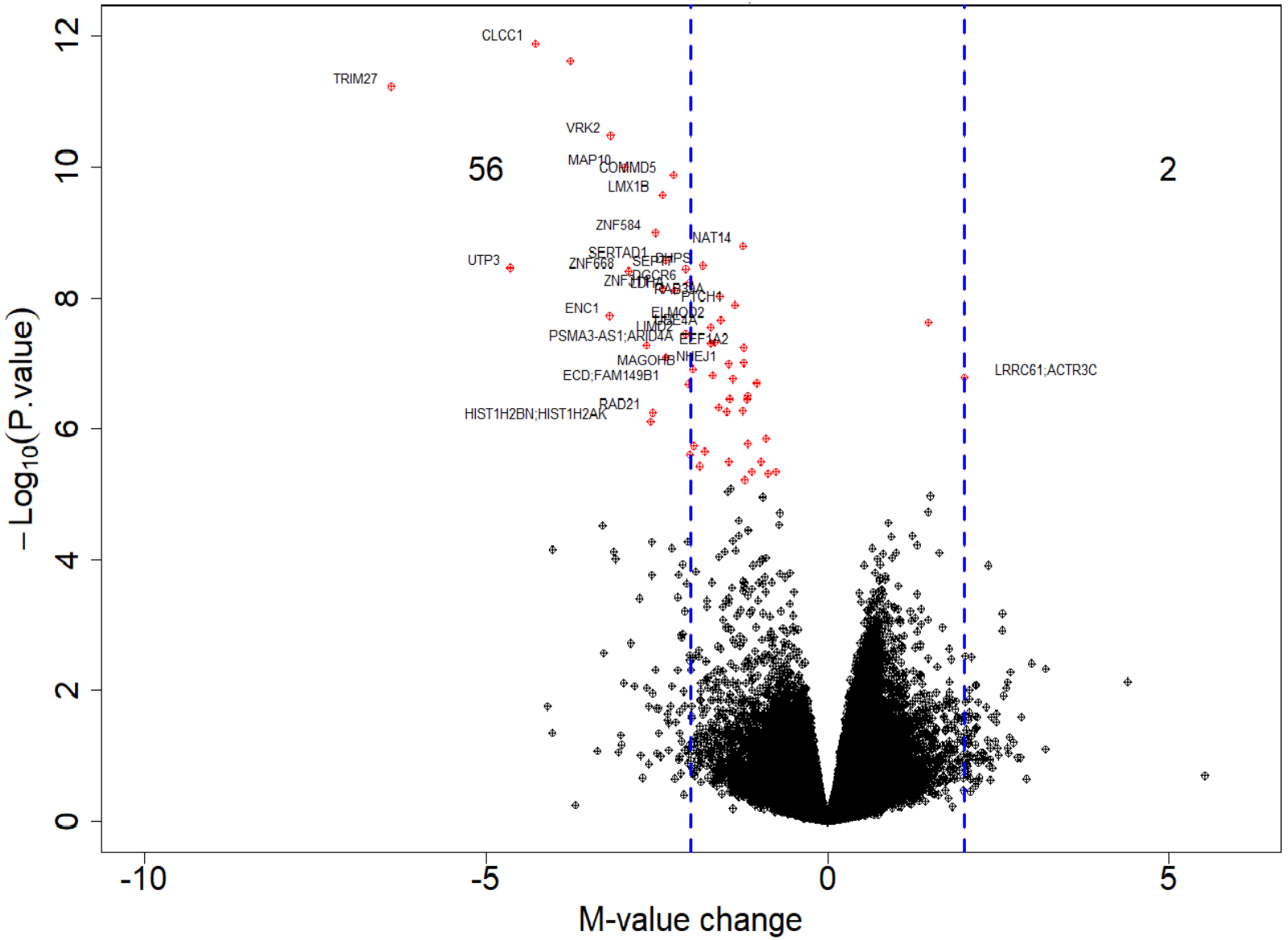
541 53. Hedrick E, Cheng Y, Jin U-H, Kim K, Safe S. Specificity protein (Sp) transcription factors
542 Sp1, Sp3 and Sp4 are non-oncogene addiction genes in cancer cells. Oncotarget.
543 2016;7:22245–56.

544 54. Goudin N, Chappert P, Mégret J, Gross D-A, Rocha B, Azogui O. Depletion of Regulatory T
545 Cells Induces High Numbers of Dendritic Cells and Unmasks a Subset of Anti-Tumour
546 CD8+CD11c+ PD-1lo Effector T Cells. PLoS One. 2016;11:e0157822.

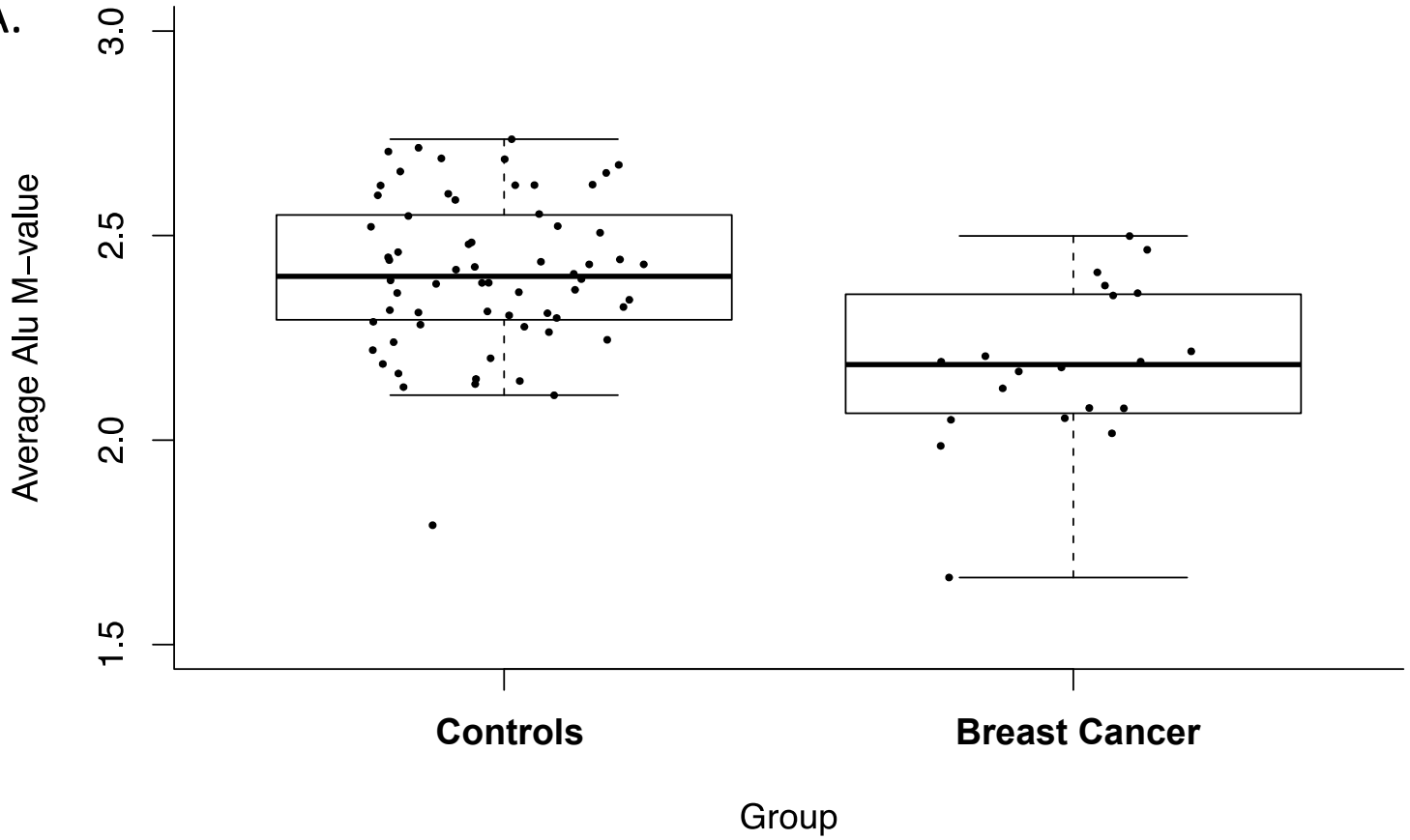
547 55. Kresovich JK, Xu Z, O'Brien KM, Weinberg CR, Sandler DP, Taylor JA. Methylation-based
548 biological age and breast cancer risk. J Natl Cancer Inst. 2019;111:1–8.

549


Cell Type 1
0.23
Cell Type 2
5.2E−06
Cell Type 3
0.00071
Cell Type 4
1.2E−05
Cell Type 5
0.6
Reference free cell estimates %
100
75
50
25
0
Controls
Breast cancer


CLCC1
TRIM27
VRK2
MAP10
COMMD5
LMX1B
56
2
ZNF584
NAT14
UTP3
ZNF668
SERTAD1
DHPS
EDC3
ZNF
PTCH1
ENC1
ELMOD2
PSMA3-AS1;ARID4A
EEF1A2
MAGOHB
NHEJ1
ECD;FAM149B1
LRRC61;ACTR3C
RAD21
HIST1H2BN;HIST1H2AK
0
2
4
6
8
10
12
-Log10(P.value)
-10
-5
0
5
M-value change

A.



B.