Breast Cancer RESEARCH

RESEARCH ARTICLE

Open Access

DNA methylation profiling in the Carolina Breast Cancer Study defines cancer subclasses differing in clinicopathologic characteristics and survival

Kathleen Conway[1,2*], Sharon N Edmiston[2], Ryan May[3], Pei Fen Kuan[4], Haitao Chu[5], Christopher Bryant[6], Chiu-Kit Tse[1], Theresa Swift-Scanlan[2,7], Joseph Geradts[8], Melissa A Troester[1,2] and Robert C Millikan^

Abstract

Introduction: Breast cancer is a heterogeneous disease, with several intrinsic subtypes differing by hormone receptor (HR) status, molecular profiles, and prognosis. However, the role of DNA methylation in breast cancer development and progression and its relationship with the intrinsic tumor subtypes are not fully understood.

Methods: A microarray targeting promoters of cancer-related genes was used to evaluate DNA methylation at 935 CpG sites in 517 breast tumors from the Carolina Breast Cancer Study, a population-based study of invasive breast cancer.

Results: Consensus clustering using methylation (β) values for the 167 most variant CpG loci defined four clusters differing most distinctly in HR status, intrinsic subtype (luminal versus basal-like), and p53 mutation status. Supervised analyses for HR status, subtype, and p53 status identified 266 differentially methylated CpG loci with considerable overlap. Genes relatively hypermethylated in HR$^+$, luminal A, or p53 wild-type breast cancers included *FABP3*, *FGF2*, *FZD9*, *GAS7*, *HDAC9*, *HOXA11*, *MME*, *PAX6*, *POMC*, *PTGS2*, *RASSF1*, *RBP1*, and *SCGB3A1*, whereas those more highly methylated in HR-, basal-like, or p53 mutant tumors included *BCR*, *C4B*, *DAB2IP*, *MEST*, *RARA*, *SEPT5*, *TFF1*, *THY1*, and *SERPINA5*. Clustering also defined a hypermethylated luminal-enriched tumor cluster 3 that gene ontology analysis revealed to be enriched for homeobox and other developmental genes (*ASCL2*, *DLK1*, *EYA4*, *GAS7*, *HOXA5*, *HOXA9*, *HOXB13*, *IHH*, *IPF1*, *ISL1*, *PAX6*, *TBX1*, *SOX1*, and *SOX17*). Although basal-enriched cluster 2 showed worse short-term survival, the luminal-enriched cluster 3 showed worse long-term survival but was not independently prognostic in multivariate Cox proportional hazard analysis, likely due to the mostly early stage cases in this dataset.

Conclusions: This study demonstrates that epigenetic patterns are strongly associated with HR status, subtype, and p53 mutation status and may show heterogeneity within tumor subclass. Among HR$^+$ breast tumors, a subset exhibiting a gene signature characterized by hypermethylation of developmental genes and poorer clinicopathologic features may have prognostic value and requires further study. Genes differentially methylated between clinically important tumor subsets have roles in differentiation, development, and tumor growth and may be critical to establishing and maintaining tumor phenotypes and clinical outcomes.

* Correspondence: kconway@med.unc.edu
^Deceased
[1]Department of Epidemiology, Campus Box 7435, School of Public Health, University of North Carolina, Chapel Hill, NC 27599, USA
[2]Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC 27599, USA
Full list of author information is available at the end of the article

BioMed Central

Introduction

Breast cancer is a complex and heterogeneous disease composed of several major subtypes with different molecular alterations, clinical behavior, and outcomes [1-3]. Microarray-based gene expression profiling of breast tumors has identified at least six major intrinsic subtypes—luminal A, luminal B, human epidermal growth factor receptor 2-positive/estrogen receptor-negative ($HER2^+/ER^-$), basal-like, claudin-low, and normal-like—that are thought to originate from different precursor cells and follow different progression pathways [4-6]. In addition, the genetic pathways leading to breast cancer vary by subtype. For example, basal-like tumors exhibit the highest and the luminal A tumors exhibit the lowest prevalence of p53 mutations [7]. These intrinsic subtypes differ in incidence by race and menopausal status [7] and show differences in risk factors [8], outcomes [7,9], and responsiveness to chemotherapy [10].

Although genetic alterations such as mutations, rearrangements, and copy number changes are established contributors to carcinogenesis, epigenetic alterations, including DNA methylation, also play an integral role. DNA methylation most commonly occurs when a methyl group is added to a cytosine preceding a guanosine (CpG). CpGs are often found at high densities in 'CpG islands', particularly within the promoter regions of genes; hypermethylation of CpG islands can result in the transcriptional silencing of tumor suppressor genes in cancer, whereas CpG hypomethylation may lead to gene activation [11-13]. Because alterations in DNA methylation often occur early in cancer development, candidate methylation markers may be valuable for early, specific cancer detection or for predicting clinical response to therapeutic agents or cancer prognosis.

In this study, we characterized DNA methylation profiles by using a microarray approach targeting CpG loci in the promoters of cancer-related genes in 517 breast tumors from the Carolina Breast Cancer Study (CBCS), a large, population-based study of mostly early-stage breast cancer in North Carolina. We hypothesized that DNA methylation events might be important determinants of tumor biology, could delineate tumor groups with distinct survival differences, and may help identify early etiologic events in breast carcinogenesis. In this report, we describe the results of this DNA methylation profiling analysis, focusing on the identification of tumor subclasses and the differentially methylated genes distinguishing them, survival differences among these tumor clusters, and characterization of hypermethylated breast tumors that may be manifestations of the CpG island methylator phenotype (CIMP) originally observed in colorectal cancer [14].

Materials and methods

Carolina Breast Cancer Study population

The CBCS is a population-based, case-control study of breast cancer. Participants include women, 20 to 74 years old, residing in 24 contiguous counties of central and eastern North Carolina [15]. Women with a first diagnosis of invasive breast cancer between 1993 and 1996 (phase 1 of the CBCS) were identified by the North Carolina Central Cancer Registry through a rapid case ascertainment system. Women diagnosed prior to age 50 and African-American women were over-sampled to ensure that they comprised roughly half the study sample. Race was self-reported; additional details are included in Table 1. Additional details of the study design are described elsewhere [15,16]. This study was approved by the Institutional Review Board at the University of North Carolina (UNC) School of Medicine. In total, 861 breast cancer cases were eligible for and consented to participate in the CBCS during phase 1. All CBCS patients provided written informed consent. Epidemiologic risk factor information was obtained from questionnaires that were administered to participants in their homes by trained nurse-interviewers. Clinical data and information on tumor characteristics were obtained from medical records or direct histopathologic review of tumor tissue. ER and progesterone receptor (PR) status of breast tumors was determined primarily through review of medical records (90% of cases) and by immunohistochemistry (IHC) staining in the remaining cases in the Tissue Procurement and Analysis Facility at UNC as described previously [7].

Tumor tissue preparation and histopathologic evaluation

Formalin-fixed paraffin-embedded (FFPE) tumor blocks were obtained from pathology departments at participating hospitals for 798 of the 861 breast cancer cases eligible for phase 1 of the CBCS. Of these, 684 had sufficient tumor tissue for molecular analyses. Tumors were sectioned as previously described [17] and underwent standardized histopathologic review by the study pathologist (JG) to confirm diagnosis, determine histologic subtype, and score standard histopathology features (grade, mitotic index, and so on). With the hematoxylin-and-eosin-stained slide used as a guide, the area of invasive tumor was selectively dissected away from other surrounding non-tumor tissue and then processed for DNA.

Breast tumor intrinsic subtypes

Subtypes were previously identified [7] by using a panel of IHC protein markers to assess expression of ER, PR, HER2, cytokeratins 5 and 6 (CK5 and CK6), and epidermal growth factor receptor (EGFR). Subtypes included luminal A (HR^+ (ER^+ or PR^+ or both) and $HER2^-$), luminal B ($HR^+/HER2^+$), basal-like ($HR^-/HER2^-/CK5^+$ or $CK6^+$

Table 1 Characteristics of Carolina Breast Cancer Study breast cancer cases or tumors evaluated or not evaluated for DNA methylation profile

Characteristic		Cases evaluated for methylation (n =517)		Cases not evaluated (n =163)		*P* value
		N	(%)	N	(%)	
Age	Mean ± SD, years	49.7 ± 11.9		52.2 ± 11.9		
	<50 years	318	(62)	84	(52)	0.02
	50+ years	199	(38)	78	(48)	
Race						
	White/Other[a]	301	(58)	106	(65)	0.12
	African-American	216	(42)	57	(35)	
Menopausal status						
	Premenopausal	275	(53)	76	(47)	0.14
	Postmenopausal	242	(47)	87	(53)	
Stage[b]						
	I	178	(37)	68	(45)	0.37
	II	245	(51)	67	(44)	
	III	45	(9)	12	(8)	
	IV	13	(3)	4	(3)	
Primary tumor size						
	≤2 cm	250	(50)	87	(57)	0.28
	>2-5 cm	205	(41)	52	(34)	
	>5 cm	42	(9)	13	(9)	
Lymph node status						
	Negative	291	(58)	101	(66)	0.11
	Positive	207	(42)	53	(34)	
Hormone receptor expression						
	ER^+/PR^+	250	(50)	62	(42)	0.38
	ER^+/PR^-	48	(10)	17	(12)	
	ER^-/PR^+	39	(8)	15	(10)	
	ER^-/PR^-	163	(33)	53	(36)	
Combined tumor grade[c]						
	I	126	(25)	47	(29)	0.48
	II	156	(30)	49	(30)	
	II	228	(45)	65	(40)	
Histologic type						
	Ductal[d]	388	(75)	122	(75)	0.23
	Ductal variants[e]	13	(3)	9	(5)	
	Poorly differentiated[f]	22	(4)	3	(2)	
	Lobular[g]	46	(9)	13	(8)	
	Mixed lobular/Ductal	48	(9)	16	(10)	

Table 1 Characteristics of Carolina Breast Cancer Study breast cancer cases or tumors evaluated or not evaluated for DNA methylation profile *(Continued)*

IHC intrinsic subtype[h]						
	Luminal A	212	(51)	42	(51)	0.76
	Luminal B	65	(16)	12	(15)	
	Basal-like	86	(21)	14	(17)	
	$HER2^+/HR^-$	26	(6)	7	(8)	
	Unclassified	24	(6)	7	(8)	
p53 mutation status						
	Positive	218	(42)	47	(34)	0.08
	Negative	297	(58)	91	(66)	

[a]The white/other cases evaluated included 291 Caucasians, 3 American Indians, 6 Asians, and 1 other. [b]According to the American Joint Committee on Cancer breast tumor staging guidelines. [c]Nottingham grade based on mitotic index, histologic grade, and nuclear grade. [d]Ductal not otherwise specified (n = 372), medullary (n = 3), neuroendocrine (n =2), apocrine (n = 2), and other mixed (n = 9). [e]Ductal variants include mucinous (n =8), papillary (n = 1), and cribriform (n = 4). [f]Poorly differentiated include metaplastic carcinoma (n = 6), anaplastic carcinoma (n =3), and undifferentiated high grade carcinoma (n = 13). [g]Lobular, classic, and/or variant (n = 46). [h]Intrinsic subtype was determined by estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) status determined by medical records or immunohistochemistry (IHC), and IHC staining for CK5, CK6, and epidermal growth factor receptor.
HR, hormone receptor; SD, standard deviation.

or both), $HER2^+/HR^-$, and unclassified (all markers negative). This IHC marker panel was previously validated against gene expression profiles [18] and was found to provide superior classification of basal-like tumors and outcome prediction over the triple-negative markers [19].

Normal breast tissues

Nine FFPE histologically normal breast tissues from women without cancer or other premalignant breast conditions were obtained from the Tissue Procurement Facility at UNC and processed for DNA. Patient consent was provided to the facility and tissues were dispersed to this study in anonymized form.

DNA extraction

FFPE tissues were processed for DNA lysates by using a Proteinase K extraction method as previously described [20].

p53 mutation screening

P53 mutation screening of 656 FFPE breast tumors in the CBCS was previously accomplished by using a combination of the Roche p53 AmpliChip (Roche Molecular Systems, Pleasanton, CA, USA), single-strand conformational polymorphism analysis, and direct radio-labeled DNA sequencing. Details of the p53 methods are provided in Additional file 1.

Bisulfite treatment of DNA

Sodium bisulfite modification of DNA obtained from FFPE tissue was performed by using the EZ DNA Methylation Gold kit (Zymo Research, Orange, CA, USA) as previously described [21].

Illumina GoldenGate Cancer Panel I methylation array analysis

Array-based DNA methylation profiling was accomplished by using the Illumina GoldenGate Cancer Panel I methylation bead array to simultaneously interrogate 1505 CpG loci associated with 807 cancer-related genes. We previously determined that this array showed high reproducibility; results obtained in FFPE tissues were highly correlated with those from matched non-FFPE samples (r =0.97), and published tumor-specific methylation profiles were detectable when DNA specimens contained at least 70% tumor cells [21].

Bead arrays were run in the Mammalian Genotyping Core laboratory at UNC at Chapel Hill. The Illumina GoldenGate methylation assay was performed as described previously [22] and imaged by using the BeadArray Scanner. Methylation status of the interrogated CpG sites was determined by comparing the ratio of the fluorescent signal from the methylated allele with the sum from the fluorescent signals of both methylated and unmethylated alleles. Controls for methylation status used on each bead array included the Zymo Universal Methylated DNA Standard as the positive, fully-methylated control, and a GenomePlex (Sigma-Aldrich, St. Louis, MO, USA) whole genome amplified DNA used as the negative, unmethylated control. Array data have been deposited in Gene Expression Omnibus under accession number GSE51557.

Array data filtering and quality control

Data were assembled by using GenomeStudio Methylation software from Illumina (San Diego, CA, USA). All array data points were represented by fluorescent signals from both methylated (Cy5) and unmethylated (Cy3) alleles. The methylation level of individual interrogated CpG sites was represented by the β value, defined as the ratio of fluorescent signal from the methylated allele to the sum of the fluorescent signals of both the methylated and unmethylated alleles and calculated as β = max (Cy5,0)/(|Cy5| + |Cy3| +100) [22]. β values ranged from 0 in the case of completely unmethylated to 1 in the case of fully methylated DNA.

Methylation array profiling was initially performed on 625 primary breast tumors by using the Illumina Cancer Panel I array that contained a total of 1,505 CpG probes. A series of filtering steps were then carried out as follows: (1) tumors with more than 25% unreliable detection *P* values of more than 10^{-5} were removed (n =14) [23]; (2) 411 CpG probes that were previously reported to overlap a single-nucleotide polymorphism (SNP) or repeat [24] were removed since these probes were potentially unreliable in some samples, especially in a racially diverse dataset such as CBCS; (3) CpG probes were removed with detection *P* value of more than 10^{-5} (n =19); (4) CpG probes with standard deviation of less than 0.06 (n =140) were removed according to Illumina's quality control algorithm [25]. Three tumors were removed because they became ineligible for the study. Finally, data from 90 tumors with replicate samples (89 with duplicates and 1 tumor with triplicates) were averaged. β values of replicate samples were highly correlated, having Pearson correlations of more than 0.900 for all but five tumors, with the remaining five tumor sets having correlations ranging from 0.899 to 0.720. The final data set consisted of 935 CpG loci (within 609 genes) in 517 breast tumors. All subsequent statistical analyses were carried out by using the R statistical programming language [26], with specific R functions noted below.

Consensus clustering and differential methylation analysis

Consensus clustering [27] was performed by using ConcensusClusterPlus [28] and CalclCL functions in R to determine subgroups of tumors on the basis of the most variable CpG sites; sites with standard deviation less than 0.2 were excluded, leaving 167 sites. This algorithm determines "consensus" clusters by measuring the stability of clustering results from the application of a given clustering method to random subsets of the data. In each iteration, 80% of the tumors were sampled, and the k-means algorithm, with the Euclidean squared distance metric, was used with k =2 to k =10 groups; these results were compiled over 100 iterations, and the stability of each clustering was determined. We chose the greatest number of clusters that had at least 90% cluster consensus. The consensus cluster heatmap was constructed by using the gplots and heatmap.2 functions in R.

Non-parametric Wilcoxon rank sum (for two class) and Kruskal-Wallis (for multiclass) tests were used to identify CpG sites that were significantly differentially methylated between tumor subgroups identified by consensus clustering. Multivariate analyses were conducted by using general linear regression models fitted to the logit transformed β methylation values to assess the association between methylation at each CpG locus and clinical or tumor covariates, adjusting for age, race, menopausal status, and stage as appropriate. *P* values were adjusted by using the Benjamini-Hochberg false discovery rate (FDR) [29] to adjust for multiple comparisons, and probes were selected at FDR of 0.05. Volcano plots were used to display global association patterns of differential methylation in which the estimated coefficients from multivariate analysis for grade, tumor size, or clinical stage were plotted against the negative logarithm of the *P* values obtained from the hypothesis test if the estimated coefficients are non-zero. The volcano plots were plotted by using raw *P* values (that is, not adjusted for multiple comparisons).

Survival analyses

Kaplan-Meier plots were used to illustrate disease-specific or overall survival among breast tumor clusters defined by methylation profiles. Survival analyses were carried out by using the survival package in R [30]. To identify methylation profiles associated with survival, multivariate Cox proportional hazard models [31] were fit with methylation cluster indicator by using R functions coxph [32] and cox.zph [33], with demographic and clinical attributes (age, race, menopausal status, stage, and other prognostic factors) as covariates. The *P* values for the Cox regression coefficients were adjusted by using Benjamini-Hochberg FDR for multiple comparisons [29].

Gene ontology term enrichment analysis for groups of differentially methylated genes

The DAVID (Database for Annotation, Visualization and Integrated Discovery) Bioinformatics Resources 6.7 Functional Annotation Tool [34] was used to perform gene-gene ontology (GO) term enrichment analysis to identify the most relevant GO terms associated with the genes found to be differentially methylated between breast tumor subsets defined by intrinsic subtype, hormone receptor (HR) status, p53 mutation status, or methylation cluster (for example, the hypermethylation cluster 3 versus other clusters). DAVID calculates an enrichment score and enrichment *P* value for each GO term to highlight the most relevant GO terms associated with the selected gene list. We used the Entrez gene IDs from each list and compared these with the background list of 609 genes evaluated from the Illumina Cancer Panel I array after filtering. Genes with more than one CpG site were listed only once in the analysis. We performed functional annotation clustering with default settings. Terms that were significantly enriched (FDR *P* <0.05) are listed.

Validation using The Cancer Genome Atlas data

Breast tumor methylation and gene expression data from 581 breast cancer patients in The Cancer Genome Atlas (TCGA) [35] were used to validate and test for relationships with gene expression at CpG probes that were among the top differentially methylated markers in the CBCS. TCGA breast cancer patients were older than CBCS patients (69% >50 years compared with 36% in CBCS), included few blacks (9% compared with 42% in CBCS), and had more later-stage 3 or 4 disease (26% compared with 12% in CBCS). Only 371 CpG probes from the GoldenGate array exactly matched probes on the Illumina 450 K array used in TCGA. Of the 935 CpG probes interrogated on the Illumina GoldenGate platform in CBCS, 21 were among our top differentially methylated probes and had exact matches for probes on the 450 K array. Based on these 21 matched 450 K CpG probes, correlations were determined with gene expression by using RNAseq expression data for all tumors (n =581) and separately for basal-like (n =102) and luminal A (n =321) tumors classified by PAM50. Pearson correlation coefficients were calculated on the basis of RNAseq (Illumina) log2 RSEM gene-normalized expression values with methylation β values for 450 K CpG probes, with significance set at *P* value of less than 0.05.

Results

Characteristics of breast cancer cases evaluated for promoter methylation

Demographic and clinical characteristics of the 517 breast cancer cases whose tumors were evaluated for Illumina promoter methylation are detailed in Table 1. The mean age of cases was 49.7 years, with 62% age 50 or younger, and 53% premenopausal. Breast cancer cases were mostly early-stage (88% stages 1 or 2), node-negative (58%), and HR^+ (ER^+ or PR^+ or both) (68%). Intrinsic tumor subtypes, defined by a panel of IHC markers (ER, PR, HER2, CK5, CK6, EGFR), identified 51% luminal A, 16% luminal B, 21% basal-like, 6% $HER2^+/HR^−$, and 6% unclassified. Histologic subtypes included approximately 75% ductal and 18% lobular or mixed lobular histologic types. Nearly all tumors were rigorously screened for p53 gene mutations, and 42% were mutation-positive. Compared with cases not evaluated (n =163), cases who were methylation profiled were younger (p =0.02) and were marginally more likely to have p53 mutation-positive tumors (*P* =0.08).

Consensus clustering of methylation β values of the most variant CpG loci in breast tumors identifies distinct molecular and subtype-related signatures

In total, 935 CpG probes (listed in Additional file 2: Table S1) were successfully screened for methylation in 517 primary breast tumors; because initial clustering indicated that many of these exhibited negligible variation in methylation level across tumors in the dataset, we performed consensus clustering by using the most variant 167 CpG probes (with standard deviation >0.2) (listed in Additional file 3: Table S2) in order to focus on CpG sites that were more likely to be useful for the subclassification of tumors. As illustrated in Figure 1A, four distinct clusters of breast tumors (numbered 1 to 4) were determined by consensus clustering when using 90% cluster consensus across all clusters as the criterion. Methylation profiles for nine normal breast tissues obtained from healthy/non-cancer patients from the UNC Tissue Procurement core are shown as a separate panel in Figure 1B, with probes similarly ordered. For the most variant probes, mean β values for each consensus cluster, together with those for normal breast tissues, are provided in Additional file 4: Table S3.

The four methylation-defined tumor clusters differed in their demographic, clinical, and molecular characteristics (Figure 1A and C and Table 2). In particular, cluster 2 was highly enriched for $HR^−$ tumors (80%) and contained the highest proportion of basal-like tumors (61%). Fewer basal-like tumors were found in clusters 1 (25%), 3 (4%), and 4 (3%). As expected, basal-enriched cluster 2 contained tumors that were of higher grade, larger size (>2 cm), stage 2 or higher, and more frequently p53 mutant. In contrast, clusters 3 and 4 were highly enriched for HR^+ luminal breast tumors, containing mainly mixtures of luminal A ($HER2^−$) and luminal B ($HER2^+$) subtypes. Cluster 4 contained the largest proportion of luminal tumors (91%), followed by clusters 3 (85%), 1 (61%), and 2 (20%). Luminal-enriched clusters 3 and 4 exhibited fewer p53 mutations compared with the basal-enriched cluster 2. Moreover, although cluster 3 consisted of mostly luminal breast cancers, these cases exhibited the highest lymph node positivity (50%) of the four tumor clusters and larger tumor size and higher stage (>stage 2) similar to tumors in basal-enriched cluster 2. Differences in age (*P* =0.006) and race (*P* =0.02) were noted among the four clusters, with the cases in basal-enriched cluster 2 being somewhat younger (<50 years) and more frequently African-American compared with the other clusters.

The average methylation content of each cluster, estimated by using the mean methylation (β) values across the 167 most variant CpG sites, differed significantly between consensus clusters (*P* <0.0001), with the luminal-enriched cluster 3 exhibiting the highest mean β overall, cluster 1 containing the lowest level of methylation, and basal-enriched cluster 2 and the highly luminal-enriched cluster 4 exhibiting intermediate levels (Figure 1D).

Cluster 3 hypermethylation gene signature

To identify the hypermethylated CpG loci that defined luminal-enriched cluster 3, the Wilcoxon rank sum test

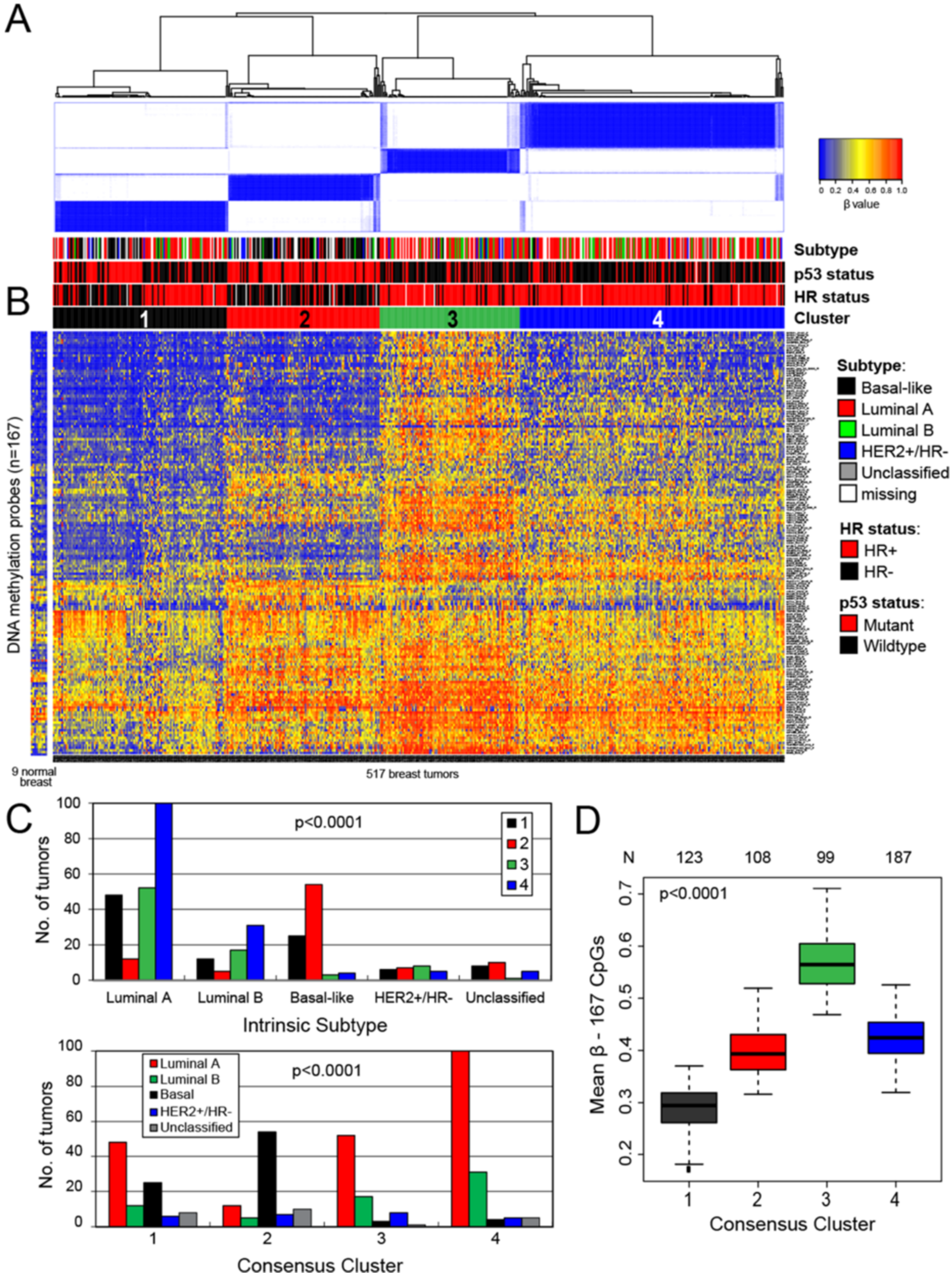

Figure 1 (See legend on next page.)

(See figure on previous page.)

Figure 1 Consensus clustering of methylation β values in breast tumors using the GoldenGate Cancer Panel I array. DNA methylation profiles in 517 breast tumors and 9 normal breast tissues are shown. Columns represent tissue samples; rows represent CpG (cytosine preceding a guanosine) loci. Beta (β) value, indicating the fraction of DNA methylated, varies from 0 (blue, unmethylated) to 1 (red, highly methylated), with intermediate values shown in yellow. **(A)** Unsupervised clustering of the 167 most variable CpG sites having standard deviation of methylation β values of more than 0.2 from among the 935 CpG sites evaluated after filtering (see Materials and methods). The four tumor clusters are numbered 1 (n =123), 2 (n =108), 3 (n =99), and 4 (n =187). Primary tumor characteristics are indicated at the top of the heatmap as intrinsic subtype (defined by immunohistochemistry, or IHC): luminal A (red), luminal B (green), basal-like (black), HER2$^+$/ER$^-$ (blue), unclassified (gray), missing values (white); p53 mutation status: mutant (red), wild-type (black); and hormone receptor (HR) status: HR$^+$ (red), HR$^-$ (black). **(B)** Methylation in the same 167 CpG sites in 9 normal breast tissues, with probes ordered as in the consensus clustered heatmap. **(C)** Relationship between methylation cluster and intrinsic tumor subtype, shown according to intrinsic subtype (top panel) or according to methylation cluster (bottom panel). **(D)** Box-and-whisker plot showing differences (*P* <0.0001) in methylation (β) of the four consensus clusters, with numbers of tumors within each cluster shown along the top of the boxplot. Luminal-enriched tumor cluster 3 exhibits distinctly higher methylation than other clusters. ER, estrogen receptor; HER2, human epidermal growth factor receptor 2.

was used to compare mean β at each of the 167 CpG sites in cluster 3 versus all other breast tumors. Cluster 3 showed significant differential methylation at 149 CpGs in 116 genes compared with all other breast tumors after accounting for multiple comparisons (Additional file 5: Table S4 and Additional file 6: Figure S1); the great majority of loci, though not all, were relatively hypermethylated in cluster 3 to varying degrees, with such genes as *ASCL2*, *GFI1*, *IPF1* (or *PDX1*), *IRAK3*, *ISL1*, *JAK3*, *KIT*, *MME*, *PENK*, *RARA*, *RASSF1*, *SEPT9*, *VIM*, and *WT1* showing the largest differential methylation. Of the 149 cluster 3-defining CpGs, 92 were unmethylated or poorly methylated (mean β <0.2) in normal breast tissues. The cluster 3-defining gene set was enriched in homeobox genes and other developmental transcription factors: *HOXB13*, *HOXA5*, *HOXA9*, *ISL1*, *EYA4*, *ASCL2*, *IHH*, *IPF1*, *ONECUT2*, *PAX6*, *SOX1*, *SOX17*, *TBX1*, and *GAS7*. To assess the functions of the 116 genes in the cluster 3 signature, a GO search performed via DAVID Bioinformatics Resources 6.7 identified 49 significant terms (FDR *P* <0.05) related to various aspects of cellular, tissue, and organ development; cell differentiation; hormonal response; cell communication; and cell motility (Additional file 7: Table S5). Additionally, the CBCS cluster 3 hypermethylation signature showed substantial overlap with the 'methyl-deviator' signature at the CpG probe (n =64) or gene (n =60) level described in the study of Killian *et al.* [36] that also used the Illumina Cancer Panel I array (Additional file 5: Table S4 and Additional file 8: Figure S2A). Comparing the significant hypermethylated probes from CBCS cluster 3 with those from the 'methyl deviator' signature, each identified from the 1505 CpG-probe GoldenGate background, we observed a highly significant correlation (*P* <0.0001, Fisher's exact test), even though different algorithms were employed to derive each of these signatures. Moreover, despite the enormous difference in the numbers of CpGs interrogated between the Illumina GoldenGate and 450 K array platforms used in the CBCS versus TCGA, respectively, genes in our hypermethylated cluster 3 were also found within the hypermethylated tumor cluster reported in TCGA (Additional file 5: Table S4) [35].

Identification of genes differentially methylated according to clinicopathologic characteristics

In addition to unsupervised analysis, we looked for patterns that varied as a function of specific clinical characteristics across all 935 CpG loci. Multivariate linear regression analysis controlling for age, race, menopausal status, stage, and multiple comparisons using FDR identified 467 CpG sites in 350 genes that were significantly (*P* <0.05) differentially methylated according to HR status (ER$^+$ or PR$^+$ or both versus ER$^-$/PR$^-$), 341 CpG sites in 264 genes that were significantly differentially methylated between basal-like and luminal A breast tumor subtypes, and 402 CpG sites in 296 genes that were significantly differentially methylated between p53 mutant and wild-type breast tumors. Complete lists of differentially methylated CpG loci are provided in Additional files 9, 10, and 11: Tables S6-S8. After controlling for intrinsic subtype in the regression model for p53 mutation, 164 significantly differentially methylated CpGs persisted, suggesting that some p53-related methylation events are independent of subtype. There was considerable overlap in the CpG loci (n =266) differentially methylated by HR status, intrinsic subtype, and p53 status (Figure 2A), with only 68 CpGs, 9 CpGs, and 61 CpGs being uniquely differentially methylated, respectively. Similar numbers of differentially methylated CpG loci were relatively hypermethylated or hypomethylated in association with HR status, subtype, or p53 mutation status (Figure 2B). Genes more highly methylated in HR$^+$, luminal, or p53 wild-type breast cancers included *FABP3*, *FGF2*, *FZD9*, *GAS7*, *HDAC9*, *HOXA11*, *MME*, *PAX6*, *POMC*, *PTGS2*, *RASSF1*, *RBP1*, and *SCGB3A1*; among the genes more highly methylated in HR$^-$, basal-like, or p53 mutant tumors were *BCR*, *C4B*, *CDH17*, *DAB2IP*, *MEST*, *RARA*, *SEPT5*, *SERPINA5*, *TFF1*, and *THY1*. Among the p53-related genes, 34 were also associated with p53 mutation status in the study of Ronneberg *et al.* [37] (Additional file 8: Figure S2B).

Table 2 Characteristics of the four methylation-based consensus clusters

Characteristic	Cluster 1 (n =123)		Cluster 2 (n =108)		Cluster 3 (n =99)		Cluster 4 (n =187)		*P* value
	N	(%)	N	(%)	N	(%)	N	(%)	
Age, years									
<50	73	(59)	82	(76)	54	(55)	109	(58)	0.006
50+	50	(41)	26	(24)	45	(45)	78	(42)	
Race									
White/Other	66	(54)	53	(49)	58	(59)	124	(66)	0.02
African-American	57	(46)	55	(51)	41	(41)	63	(34)	
Menopausal status									
Postmenopausal	61	(50)	39	(36)	51	(52)	91	(49)	0.09
Premenopausal	62	(50)	69	(64)	48	(48)	96	(51)	
Stage									
I	58	(49)	26	(27)	24	(25)	70	(41)	0.02
II	48	(40)	58	(60)	56	(60)	83	(48)	
III	12	(10)	9	(9)	11	(12)	13	(8)	
IV	1	(1)	4	(4)	3	(3)	5	(3)	
Missing	4		11		5		16		
Primary tumor size									
≤2 cm	72	(60)	35	(35)	36	(38)	107	(59)	0.0003
>2-5 cm	39	(33)	53	(54)	49	(51)	64	(35)	
>5 cm	9	(7)	11	(11)	11	(11)	11	(6)	
Missing	3		9		3		5		
Lymph node status									
Positive	44	(37)	40	(39)	48	(50)	75	(42)	0.24
Negative	76	(63)	62	(61)	48	(50)	105	(58)	
Missing	3		6		3		7		
Tumor grade									
I	29	(24)	3	(3)	17	(17)	77	(42)	<0.0001
II	34	(28)	16	(15)	40	(40)	66	(36)	
III	58	(48)	87	(82)	42	(43)	41	(22)	
Missing	2		2				3		
Estrogen receptor status									
Positive	56	(47)	11	(11)	79	(81)	153	(83)	<0.0001
Negative	64	(53)	90	(89)	18	(19)	31	(17)	
Missing	3		7		2		3		
Hormone receptor status									
Positive	69	(58)	20	(20)	83	(86)	166	(90)	<0.0001
Negative	51	(42)	81	(80)	13	(14)	18	(10)	
Missing	3		7		3		3		
Histologic type									
Ductal NOS	98	(80)	89	(82)	78	(79)	123	(66)	<0.0001
Ductal variants	3	(2)	0	(0)	4	(4)	6	(3)	
Poorly differentiated	7	(6)	11	(10)	2	(2)	2	(1)	

Table 2 Characteristics of the four methylation-based consensus clusters *(Continued)*

Lobular	8	(6)	1	(1)	7	(7)	30	(16)	
Mixed lobular	7	(6)	7	(7)	8	(8)	26	(14)	
Intrinsic subtype (IHC)									
Luminal A	48	(49)	12	(14)	52	(64)	100	(69)	<0.0001
Luminal B	12	(12)	5	(6)	17	(21)	31	(22)	
Basal-like	25	(25)	54	(61)	3	(4)	4	(3)	
HER2$^+$/ER$^-$	6	(6)	7	(8)	8	(10)	5	(3)	
Unclassified	8	(8)	10	(11)	1	(1)	5	(3)	
Missing	24		20		18		42		
p53 mutation status									
Mutant	59	(48)	90	(83)	25	(25)	44	(24)	<0.0001
Wild-type	64	(52)	18	(17)	74	(75)	141	(76)	
Missing							2		
EGFR status									
Positive	34	(32)	75	(75)	9	(10)	11	(7)	<0.0001
Negative	71	(68)	25	(25)	78	(90)	149	(93)	
Missing	18		8		12		27		
HER2 status									
Positive	25	(20)	16	(15)	31	(31)	50	(27)	0.02
Negative	97	(80)	92	(85)	68	(69)	135	(73)	
Missing	1						2		

Hormone receptor (HR) status: positive: estrogen receptor-positive (ER$^+$) or progesterone receptor-positive (PR$^+$) or both; negative: ER$^-$ and PR$^-$. Consensus methylation clusters 1 to 4 based on the most variant 167 CpG (cytosine preceding a guanosine) sites. Intrinsic subtypes: luminal A (ER$^+$ and/or PR$^+$, HER2$^-$), luminal B (ER$^+$ and/or PR$^+$, HER2$^+$), basal-like (ER$^-$, PR$^-$, HER2$^-$, CK5$^+$ and/or CK6$^+$ or EGFR$^+$), HER2$^+$/HR$^-$ (ER$^-$/PR$^-$/HER2$^+$), and unclassified (all markers negative). EGFR, epidermal growth factor receptor; HER2, human epidermal growth factor receptor 2; IHC, immunohistochemistry; NOS, not otherwise specified.

Volcano plots were used to visualize patterns of differential methylation across all CpGs according to grade, tumor size, or clinical stage, showing the coefficients from multivariate analyses and associated log^{-10} *P* values (Figure 2C). Higher tumor grade (II/III versus I), larger tumor size (<2 cm versus >2 cm), and increasing clinical stage (comparing across stages 1 to 4) were associated primarily with CpG hypermethylation; however, compared with HR status, subtype, or p53 status, fewer differentially methylated CpG loci were detected in association with these characteristics (177 CpGs for tumor grade, 24 CpGs for tumor size, and 14 for stage) (Figure 2B). Higher methylation at one or more CpG sites in the upstream regulatory region of *GSTM2* was correlated with increasing stage, larger tumor size, and higher grade. Similarly, higher methylation in the *MT1A* gene was correlated with larger tumor size. No CpG probes were found to be differentially methylated in relation to lymph node status. Boxplots showing the distribution of β values for the top differentially methylated CpG loci (at FDR *P* <0.05 level) in breast tumors according to clinicopathologic factors are given in Figure 2D. We also tested for differences in tumor methylation with age among premenopausal or postmenopausal cases, but no significant differences were detected after adjustment for multiple comparisons.

GO analysis of 350 genes that were differentially methylated between HR$^+$ and HR$^-$ breast tumors identified 71 GO terms that remained highly significant after adjusting for multiple comparisons (FDR adjusted *P* <0.05); similarly, GO analysis for 264 genes differentially methylated between luminal A and basal tumor subtypes identified 36 terms, and analysis of 296 genes differentially methylated between p53 mutant and wild-type tumors identified 44 GO terms. As expected, there was considerable overlap in the terms identified in these three analyses (Additional file 12: Table S9), which were related to signal transduction, anatomic development, cell differentiation and cell proliferation, and response to steroid hormone stimulus. Additional GO terms related to HR status included regulation of cell death, apoptosis, and programmed cell death.

Survival differences among methylation-based consensus clusters

Kaplan-Meier curves showing breast cancer-specific survival of CBCS cases revealed some differences between the

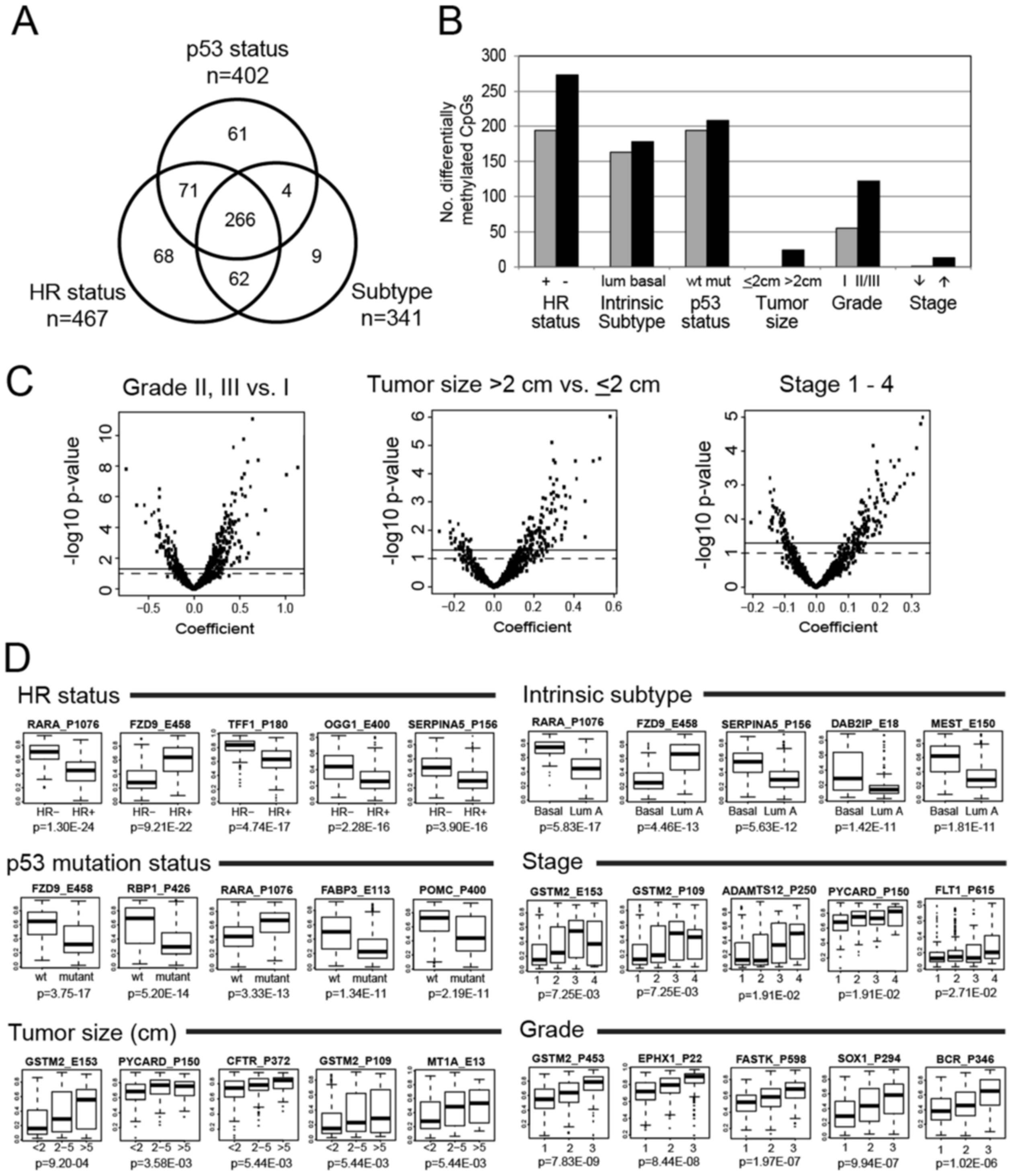

Figure 2 (See legend on next page.)

(See figure on previous page.)
Figure 2 Differential CpG methylation in breast tumors according to clinical or tumor factors. Generalized linear regression models were used to compare methylation at each of 935 CpG sites in breast tumors according to clinical or prognostic factors while controlling for age, race, menopausal status, and stage (except in analyses of tumor size or nodal status; tumor size was adjusted for in the analysis of nodal status, and vice versa). **(A)** Venn diagram showing overlap of significantly differentially methylated sites (false discovery rate (FDR) P <0.05) according to hormone receptor (HR) status, intrinsic subtype (basal-like versus luminal A), and p53 status. Full lists of differentially methylated CpG loci are given in Additional files 9, 10, and 11: Tables S6-S8. **(B)** Bar graph summarizing the numbers of differentially methylated CpG loci that were relatively hypermethylated or hypomethylated in association with clinical or tumor characteristics. For analysis of stage, methylation varied between stages 1 to 4. **(C)** Volcano plots showing global patterns of differential methylation across all 935 CpGs. All multivariate models were adjusted for age, race, menopausal status, stage, except for stage (adjusted for age, race, and menopausal status only), and tumor size (adjusted for age, race, menopausal status, and lymph node status). Probes significantly differentially methylated at the P <0.05 level in multivariate analysis fall above the solid line and at P <0.1 above the broken line. **(D)** Box-and-whisker plots showing the top five CpGs exhibiting significant differential methylation according to clinical staging or tumor characteristics. Each box plot shows the median β-value (dark bar within box) and the interquartile range (IQR = Q3-Q1) (outer boundaries of box). The whiskers (broken line) cover (Q1 – 1.5IQR, Q3 + 1.5IQR). Multivariate and FDR-adjusted P values are shown for each boxplot. No CpGs were differentially methylated according to lymph node status.

four tumor subgroups defined by methylation signature (log-rank P =0.02) (Figure 3A). The luminal-enriched hypermethylated cluster 3 and the basal-enriched cluster 2 showed poorer survival compared with clusters 1 and 4. Although basal-enriched cluster 2 showed worse early survival, cluster 3 showed similar survival to the basal-enriched cluster by the end of the follow-up period. Overall survival was worse than breast cancer-specific survival for all clusters but generally reflected the relative differences noted between methylation-based clusters, with hypermethylated cluster 3 showing somewhat worse long-term survival than the other three clusters (Figure 3B). Figures 3C and D show survival plots for disease-specific and overall survival based among breast tumors distinguished by intrinsic subtype. Compared with the intrinsic subtypes, methylation-based clustering provided somewhat better distinction of patients differing in outcome based on log-rank P values.

To determine whether methylation profile would provide superior prognostic value for breast cancer compared with the known clinical or prognostic factors, each clinical attribute as well as methylation-based tumor cluster was tested in univariate Cox proportional hazard analysis; factors identified as significant in univariate analysis were then included in multivariate models to determine whether methylation signature was an independent predictor of survival (Table 3). In univariate analyses for patient or clinical variables, age, race, menopausal status, stage, HR status, tumor size, lymph node status, and tumor grade all showed significant hazard ratios (at P <0.05). Relative to luminal-enriched cluster 4 reference group, basal-enriched cluster 2 and hypermethylated cluster 3 showed significantly worse outcome in univariate analyses (hazard ratio =1.91, 95% confidence interval (CI) = 1.19 to 3.08 and hazard ratio = 1.71, 95% CI = 1.04 to 2.79, respectively) but were not independently associated with survival in multivariate Cox proportional hazard models fully adjusted for all significant covariates (age, race, stage, HR status, grade, tumor size, and lymph node status). Univariate Cox proportional hazards analysis of intrinsic subtypes identified only the HER2$^+$/HR$^-$ subtype as differing in breast cancer-specific survival compared with the luminal A subtype (hazard ratio =2.41, 95% CI = 1.24 to 4.70), but there were no significant differences among subtypes in multivariate analyses after controlling for other clinical or tumor characteristics.

Correlations between methylation and gene expression for GoldenGate-matched 450 K Illumina probes in TCGA

Because no gene expression data were available for our CBCS tumors, we sought to infer locus-specific CpG methylation correlations with gene expression in publically available TCGA breast tumor data. In total, only 371 probes from the 1,505 probes in GoldenGate array exactly match those on the 450K. Of the 935 Illumina GoldenGate probes interrogated in our study after filtering, 21 were both direct matches to the CpG probes on the 450K array and were found to be differentially methylated and of interest in our study. For these 21 matched 450K probes, Pearson correlation coefficients were calculated in all breast tumors (n = 581) and for basal-like (n = 102) and luminal A (n = 321) tumors comparing RNAseq (Illumina) log2 RSEM gene-normalized expression values with methylation β values for 450K CpG probes (Additional file 13: Table S10). For the TCGA breast samples, approximately half the probes with exact matches to the GoldenGate platform and showing differential methylation in our study exhibited significant inverse correlations with gene expression; among these are *CCND2*, *DBC1*, *FGF2*, *JAK3*, *KIT*, and *SERPINA5*.

Discussion

In this study, we describe the results of an array-based promoter methylation analysis of 935 CpG sites in cancer-related genes in a large, population-based study of mostly early-stage breast cancer. Consensus clustering of methylation levels for the 167 most variant CpG loci in 517 tumors identified four methylation-based tumor

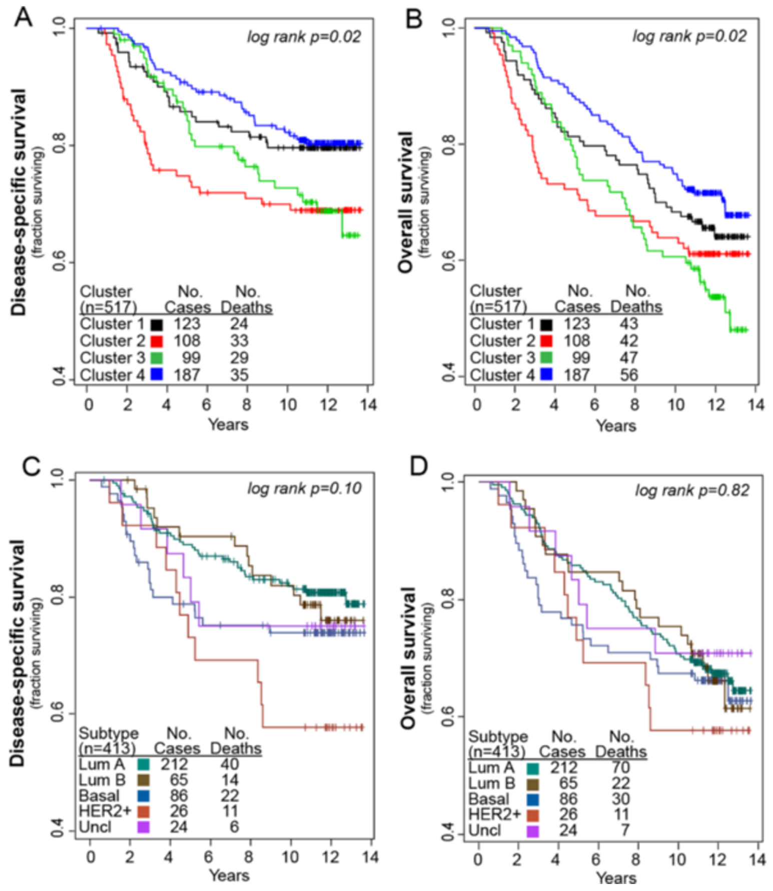

Figure 3 Kaplan-Meier plot showing survival of breast cancer case subsets defined by methylation-based consensus clustering or intrinsic subtyping. Consensus clustering of methylation β values for the 167 most variant CpG (cytosine preceding a guanosine) sites in 517 breast tumors defined four clusters. Kaplan-Meier plots show **(A)** breast cancer-specific survival or **(B)** overall survival for methylation-based clusters 1 through 4. Log-rank *P* values (*P* =0.02) indicate significant differences in survival among the methylation clusters. Intrinsic subtype information was available on 413 of the 517 tumors with methylation data. Subtypes defined by immunohistochemistry (IHC), as described in the Materials and methods section, were luminal A, luminal B, basal-like, human epidermal growth factor receptor 2-positive/hormone receptor-negative ($HER2^+/HR^-$), and unclassified. Kaplan-Meier plots for the five intrinsic subtypes show **(C)** breast cancer-specific survival and **(D)** overall survival. Numbers of cases and events within each group are noted in each plot.

subgroups that were associated with HR status or specific intrinsic subtypes (basal-like versus luminal A), thus confirming that intrinsic subtype may be an important determinant of some epigenetic markers. However, there are also important methylation phenotypes that are heterogeneously expressed within tumor subclass. For example, although clusters 3 and 4 were both composed of mostly luminal tumors (85% and 91%, respectively), methylation profiling distinguished cluster 3 as a hypermethylated subclass with poorer clinicopathologic characteristics (larger tumor size, higher grade, and more frequently lymph node-positive) and possibly worse outcomes.

Most HR^+ or luminal-enriched tumor clusters exhibited higher methylation across the most variant CpGs compared with HR^- or basal-like tumors. Genes previously observed to differ in methylation between luminal and basal-like subtypes and also noted in this study included *RASSF1*, *FZD9*, *PTGS2*, *MME*, *HOXA9*, *PAX6*, and *SCGB3A1*, which were more highly methylated in HR^+ and luminal tumors [37-39], and *CDH17*, *EPHX1*, *TFF1*, *RARA*, and *MEST*, which showed higher methylation in basal-like tumors. These methylation-based clusters also differed in the prevalence of p53 mutation, which is strongly correlated with intrinsic subtype, occurring with high prevalence among basal-like tumors in the CBCS [7]. However, even after intrinsic subtype differences were controlled for, 164 significant p53-related CpG methylation differences persisted, suggesting that at least some of these methylation events are independent of tumor subtype. Methylation also varied according to clinicopathologic characteristics, with higher tumor grade being

Table 3 Cox multivariate regression analysis for breast cancer survival according to CpG methylation profile, clinical factors, or intrinsic subtypes

	Univariate			Multivariate[a]		
Prognostic variable	**Hazard ratio**	**95% CI**	***P* value**	**Hazard ratio**	**95% CI**	***P* value**
Methylation cluster (n =517)						
4 (luminal-enriched) (n =187) (reference)	1.00	-		1.00	-	
1 (mixed) (n =123)	1.11	0.66-1.86	0.70	1.09	0.61-1.95	0.78
2 (basal-enriched) (n =108)	1.91	1.19-3.08	0.0075	1.41	0.76-2.64	0.28
3 (luminal-enriched) (n =99)	1.71	1.04-2.79	0.033	1.27	0.75-2.17	0.37
Clinical factor (n =517)						
Age at diagnosis (continuous)	0.97	0.96-0.99	0.004	0.99	0.96-1.01	0.38
Premenopausal (versus post)	1.76	1.20-2.57	0.004	1.27	0.72-2.22	0.41
African-American (versus white/other)	1.65	1.15-2.35	0.006	1.60	1.08-2.39	0.02
HR$^-$ (versus HR$^+$)	1.58	1.09-2.30	0.02	1.07	0.65-1.76	0.80
HER2$^+$ (versus HER2$^-$)	1.32	0.89-1.96	0.16	-	-	-
Stage (1, 2, 3, 4)	2.76	2.22-3.43	<0.0001	1.74	1.22-2.49	0.002
Grade 2/3 (versus 1)	2.51	1.46-4.31	0.0009	1.21	0.66-2.22	0.54
Lymph node-positive (versus negative)	5.25	3.44-8.00	<0.0001	3.40	2.03-5.71	<0.0001
Tumor size 2-5 cm (versus ≤2 cm)	2.32	1.52-3.55	0.0001	1.24	0.77-1.99	0.38
Tumor size >5 cm (versus ≤2 cm)	5.01	2.91-8.63	<0.0001	1.44	0.73-2.83	0.28
Intrinsic subtype (n =393)[b]						
Luminal A (n =212) (reference)	1.00	-		1.00	-	
Luminal B (n =65)	1.14	0.62-2.09	0.68	0.97	0.51-1.84	0.93
Basal-like (n =86)	1.49	0.88-2.50	0.13	1.10	0.63-1.91	0.74
HER2$^+$/HR$^-$ (n =26)	2.41	1.24-4.70	0.01	1.06	0.48-2.34	0.88
Unclassified (n =24)	1.34	0.57-3.17	0.50	1.19	0.49-2.87	0.70

[a] Multivariate Cox proportional hazards regression models for methylation-based clusters were adjusted for age (continuous), menopausal status (pre/post), race, stage (1, 2, 3, 4), hormone receptor (HR) status, grade (1 versus 2 + 3), lymph node status, and tumor size. Multivariate Cox proportional hazards regression models for intrinsic subtypes were adjusted for age (continuous), menopausal status (pre/post), race, stage (1, 2, 3, 4), grade (1 versus 2 + 3), lymph node status, and tumor size. [b] The reduced number of tumors included in models for intrinsic subtypes reflects missing data for subtype or other covariates. CI, confidence interval; HER2, human epidermal growth factor receptor 2.

strongly correlated with hypermethylation of such genes as *GSTM2*, *EPHX1*, and *BCR*, and larger primary tumor size correlated with hypermethylation of *GSTM2*, *PYCARD*, *MYCL2*, and *MT1A*. Methylation of several of these genes has been noted previously in breast cancer [37,40-43]. Moreover, methylation was significantly inversely correlated with gene expression for several of these genes in TCGA. Importantly, our findings are consistent with prior reports of heavier methylation among HR$^+$ breast tumors, less methylation in basal-like tumors [35,38,39,44], and significant correlation of breast tumor DNA methylation patterns with HR subtype [36,45], gene expression-based subtype [35,37,39,44,46,47], or p53 mutational status [37].

Recent evidence suggests that the distinct differences in methylation observed according to intrinsic breast tumor subtype may reflect the methylation patterns of different cells of origin. Lineage-specific differentiation changes might lock tumors into certain growth programs that subsequently help to drive the tumor phenotype and clinical outcome. Kamalkaran *et al.* [46] found that methylation patterns in basal tumors are similar to breast progenitor cells but that the patterns in luminal A tumors are similar to those identified in the more differentiated CD24$^+$ luminal epithelial cells. Similarly, *in vitro* work by Bloushtain-Qimron *et al.* [48] reported that CD44$^+$ progenitor-like cells of normal mammary epithelium were hypomethylated compared with luminal epithelial (CD24$^+$ and MUC1$^+$) and myoepithelial (CD10$^+$) cells and that cell type-specific methylation patterns were conserved in breast cancer subtypes. Additionally, we observed differences in methylation of several genes that mediate or are markers for epithelial-to-mesenchymal transition (EMT) (for example, *NOTCH* or *VIM*) or signaling pathways (TGFβ, WNT/β-catenin, and FGF) linked to EMT [49]. Although differential methylation of some cadherins (*CDH17* and *PCDH1*) varied by subtype, HR status, and p53 mutational state, the EMT marker, CDH1, was not among them. Recently, Cohen *et al.* [50] mapped patterns of epigenetic pathway activation in breast and other tumor types and identified a gene expression pattern of EZH2 activation in luminal breast tumors, and HDAC4 pathway activation was seen in basal breast tumors. These two distinct activated pathways were mutually exclusive, supporting the idea that fundamentally different epigenetic programs characterize these tumor subtypes.

A growing number of studies have investigated the existence and possible clinical relevance of a CIMP in breast tumors, which has been described in other tumor types, most notably colorectal cancers [14,51-57]. Putative CIMP or gene hypermethylation signatures have been identified in subsets of HR^+ breast tumors that were independently associated with poorer clinical outcomes in multivariate Cox models [36,45,46] or with gene expression signatures indicative of poor prognosis [58]. Conversely, Fang *et al.* [59] found CIMP to be associated with HR^+ status, reduced metastatic potential, and better survival, suggesting the possibility that the hypermethylated CIMP signature primarily distinguished intrinsic subtypes which are known to differ in survival. The CIMP hypermethylation profile described among HR^+ tumors appears to manifest as a coordinated hypermethylation of a set of genes highly enriched for developmental transcription factors, polycomb repressor complex 2 gene targets, as well as genes involved in EMT and Wnt signaling [36,46,59]. A recent report from TCGA identified a hypermethylated, HR^+ breast tumor subset with lower Wnt-pathway gene expression and fewer PIK3CA and MAP3K1 mutations [35].

It is unclear whether CIMP-associated gene hypermethylation in breast tumors reflects the degree of lineage-specific differentiation or is a biologically distinct entity occurring through another mechanism. It has been proposed that CIMP may signify an underlying global derangement in epigenetic regulation [59], possibly mediated by overexpression of DNA methyltransferase 3b [58,60,61]. Moreover, it is important to note that gene hypermethylation independent of CIMP may also have prognostic value in breast tumors [45,62-65]. Notably, hypermethylation signatures predicted poorer outcomes in ER^- breast cancers [45,62], with one study identifying a prognostic signature highly enriched in homeobox genes [45].

In the CBCS, consensus clustering of the 167 most variant CpG loci revealed a CIMP-like hypermethylated cluster 3. Although this cluster was composed of predominantly HR^+/luminal tumors, it was associated with poorer clinicopathologic features and possibly worse prognosis, similar to basal-like breast cancers. In fact, methylation-based clustering provided similar discrimination of prognostically different subgroups as intrinsic subtyping based on IHC. The finding that DNA methylation profiling may identify breast cancer cases with worse outcomes irrespective of subtype, together with its particular suitability for FFPE tissues, suggests that methylation analysis could be useful for breast cancer prognosis. Cluster 3 tumors exhibited hypermethylated gene signatures enriched in homeobox domain and transcription factors important in development and differentiation, consistent with prior studies [36,37,45,46,59]. In particular, the cluster 3 CpG signature was similar to the ‘methyl deviator’ signature identified by Killian *et al.* [36] that independently predicted poor prognosis among HR^+ tumors. Moreover, the hypermethylated breast tumor gene signature identified in TCGA [35] overlapped with our cluster 3, showing both hypermethylation and reduced expression of genes such as *ASCL2*, *CCND2*, *COL1A2*, *EPHB1*, *FABP3*, *GAS7*, *IFNGR2*, *IRAK3*, *KLK10*, *POMC*, *SCGB3A1*, *SFRP1*, *SMO*, and *VCAN* (*CSPG2*).

Our findings from CBCS suggest that methylation patterns defined by the most variant CpG loci largely reflect cell lineage, as evidenced by the distinct differences in methylation patterns between HR^+ and HR^- or basal-like and luminal A breast tumors, and the extreme hypermethylation of genes important in development and differentiation in a subset of mostly HR^+ tumors. This is consistent with the idea that aberrant methylation occurs early in cancer development [66], suggesting that these methylation events may be important in carcinogenesis and could be linked with exposures that modulate risk of tumor subtypes. Our results also suggest that certain methylation events are associated with more aggressive tumor phenotypes irrespective of subtype and have the potential to provide prognostic information, consistent with other studies [36,37,45,46]. Owing to high representation of incident, early-stage breast cancer cases in the CBCS dataset (with relatively few deaths), our power to detect significant and independent survival differences may have been limited, particularly among the better-prognostic HR^+ cases. However, our results are derived from a population-based sample and therefore represent the distribution of incident breast cancer cases. Over time, extended follow-up of CBCS cases may allow more definitive ascertainment of the relationship between CpG methylation and breast cancer survival.

Major strengths of this study include the large size and population-based nature of the CBCS, inclusion of breast tumors with relatively complete histopathologic, subtyping, and outcome data. The sample size was large, allowing well-powered analysis of methylation signatures across a diverse spectrum of breast tumors. We used a stringent approach in methylation profiling by filtering out CpG probes that overlapped repeats or known SNPs, which might have produced unreliable results. A few limitations are also noted. The CBCS collected only FFPE tumor tissues that have been stored as cut sections for nearly 20 years. The difficulty in obtaining RNA of sufficient quality for gene expression array analysis from such tissues has precluded the direct comparison of promoter methylation and gene expression. Intrinsic tumor subtypes were defined by a panel of IHC protein expression markers, which may be less accurate than subtyping based on expression of 50 or more genes [67], and therefore likely resulted in some misclassification. This misclassification is most likely to occur among luminal breast cancers; however, given that the most prominent methylation differences were between luminal and basal-like breast cancers, this misclassification is unlikely to substantially alter the conclusions of the study. The data were collected on a first-generation methylation array which oversampled genes in cancer-related pathways; however, many genes on the platform had strong coverage for the best-studied methylation sites in breast cancer research. Additionally, information on treatment and breast cancer recurrence was not available in CBCS, and thus their impact on the relationship between methylation profile and survival could not be addressed.

Conclusions

In summary, we found evidence for a strong association of DNA methylation with HR status and breast tumor subtype as well as with p53 mutation status, which is inextricably linked to subtype. Moreover, epigenetic heterogeneity within tumor subclass is supported by identification of a hypermethylated tumor cluster enriched in developmental genes among primarily HR$^+$ luminal tumors. This hypermethylated signature may be related to more aggressive tumor growth features and, potentially, outcome. These findings provide proof-of-principle that epigenetic profiles may offer important information beyond expression-based subtyping for clinically or epidemiologically meaningful breast tumor classification.

Additional files

Additional file 1: p53 mutation screening methods.

Additional file 2: Table S1. Nine hundred thirty-five CpG sites/probes evaluated for methylation.

Additional file 3: Table S2. One hundred sixty-seven most variant CpG sites.

Additional file 4: Table S3. Mean beta values for 167 probes distinguishing four consensus clusters.

Additional file 5: Table S4. One hundred forty-nine CpG sites distinguishing hypermethylated cluster 3 from other breast tumors.

Additional file 6: Figure S1. Boxplots showing distribution of β methylation values for the top CpG markers defining the hypermethylated cluster 3. β values are shown for the four consensus clusters. CpG sites or genes that overlap the 'methyl deviator' signature (○) described by Killian *et al.* [36] or the hypermethylated cluster 3 described in breast tumors profiled within The Cancer Genome Atlas (♦) [35] are indicated.

Additional file 7: Table S5. Gene ontology (GO) terms for genes that define the cluster 3 hypermethylation signature.

Additional file 8: Figure S2. Venn diagrams showing overlap of differentially methylated CpGs/genes between CBCS and other published studies. **(A)** Overlap of the hypermethylated signature from cluster 3 in CBCS (149 CpGs, 116 genes) with the methyl-deviator signature (109 CpGs, 85 genes) identified in the study of Killian *et al.* [36], which also used the Illumina Cancer Panel I methylation platform. **(B)** Overlap of genes differentially methylated according to p53 mutation status in CBCS (402 CpGs in 296 genes) with genes included in the p53 signature (84 genes) reported in Ronneberg *et al.* [37]. CBCS, Carolina Breast Cancer Study.

Additional file 9: Table S6. Four hundred sixty-seven CpGs differentially methylated by hormone receptor
(HR) status.

Additional file 10: Table S7. Three hundred forty-one CpGs differentially methylated by subtype.

Additional file 11: Table S8. Four hundred two CpGs differentially methylated by p53 status.

Additional file 12: Table S9. Gene ontology (GO) terms for genes differentially methylated according to hormone receptor (HR) status, subtype, or p53 status.

Additional file 13: Table S10. The Cancer Genome Atlas (TCGA) correlations between GoldenGate-matched 450 K Illumina probe methylation and gene expression.

Abbreviations

CBCS: Carolina Breast Cancer Study; CIMP: CpG island methylator phenotype; CpG: cytosine preceding a guanosine; DAVID: Database for Annotation, Visualization and Integrated Discovery; EMT: epithelial to mesenchymal transition; ER: estrogen receptor; FDR: false discovery rate; FFPE: formalin-fixed paraffin-embedded; GO: gene ontology; HER2: human epidermal growth factor receptor 2; HR: hormone receptor; IHC: immunohistochemistry; PR: progesterone receptor; SNP: single-nucleotide polymorphism; TCGA: The Cancer Genome Atlas.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

KC conceived, designed, and implemented the study; analyzed and interpreted the data; drafted the manuscript; and obtained funding for the study. HC, PFK, RM, CB, TS-S, and C-KT analyzed the data and contributed to writing the manuscript. SE participated in study design, processed tissue samples, generated the methylation data, and contributed to analysis and interpretation of data and writing of the manuscript. RM (deceased), principal investigator of the CBCS, and MT contributed to data analysis and interpretation and provided a critical review of the manuscript. As the study pathologist, JG performed histopathologic review and scoring of tumors and their grading criteria and provided a critical review of the manuscript. All authors read and approved the final version of this manuscript.

Acknowledgments

This work was supported by a grant to KC from the Susan G. Komen for the Cure (grant #KG081397). The CBCS was funded by the University Cancer Research Fund of North Carolina and the National Cancer Institute Specialized Program of Research Excellence (SPORE) in Breast Cancer (NIH/NCI P50-CA58223). We acknowledge support of Jason Luo and staff of the Mammalian Genotyping Core Facility in performance of the Illumina methylation array assays and of the Tissue Procurement Core for making available normal breast tissue specimens. We thank Lin Wu and Nancy Patten of Roche Molecular Systems for providing p53 AmpliChip arrays. We are grateful to the participants of the CBCS for their continuing support of the study. This work is dedicated to the memory of our colleague, the late Robert Millikan, principal investigator of the CBCS.

Author details

[1]Department of Epidemiology, Campus Box 7435, School of Public Health, University of North Carolina, Chapel Hill, NC 27599, USA. [2]Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC 27599, USA. [3]The EMMES Corporation, Rockville, MD 20850, USA.

[4]Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, NY 11794-3600, USA. [5]Division of Biostatistics, MMC 303 School of Public Health, University of Minnesota, Minneapolis, MN 55455, USA. [6]Department of Biostatistics, School of Public Health, University of North Carolina, Campus Box 7420, Chapel Hill, NC 27599, USA. [7]School of Nursing, Campus Box 7460, University of North Carolina, Chapel Hill, NC 27599, USA. [8]Department of Pathology, School of Medicine, Duke University Medical Center DUMC 3712, Durham, NC 27710, USA.

Received: 21 November 2013 Accepted: 24 September 2014
Published online: 07 October 2014

References

1. Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lønning PE, Børresen-Dale AL, Brown PO, Botstein D: **Molecular portraits of human breast tumours.** *Nature* 2000, **406:**747–752.
2. Van't Veer LJ, Dai H, van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrish M, Atsma D, Witteveen A, Glas A, Delahaye L, van der Velde T, Bartelink H, Rodenhuis S, Rutgers ET, Friend SH, Bernards R: **Gene expression profiling predicts clinical outcome of breast cancer.** *Nature* 2002, **415:**530–536.
3. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Eystein Lonning P, Borresen-Dale AL: **Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.** *Proc Natl Acad Sci U S A* 2001, **98:**10869–10874.
4. Polyak K, Shipitsin M, Campbell-Marrotta L, Bloushtain-Qimron N, Park SY: **Breast tumor heterogeneity: causes and consequences.** *Breast Cancer Res* 2009, **11:**S18.
5. Prat A, Perou CM: **Deconstructing the molecular portraits of breast cancer.** *Mol Oncol* 2011, **5:**5–23.
6. Keller PJ, Arendt LM, Skibinski A, Logvinenko T, Klebba I, Dong S, Smith AE, Prat A, Perou CM, Gilmore H, Schnitt S, Naber SP, Garlick JA, Kuperwasser C: **Defining the cellular precursors to human breast cancer.** *Proc Natl Acad Sci U S A* 2012, **109:**2772–2777.
7. Carey LA, Perou CM, Livasy CA, Dressler LG, Cowan D, Conway K, Karaca G, Troester MA, Tse CK, Edmiston S, Deming SL, Geradts J, Cheang MC, Nielsen TO, Moorman PG, Earp HS, Millikan RC: **Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study.** *JAMA* 2006, **295:**2492–2502.
8. Millikan RC, Newman B, Tse CK, Moorman PG, Conway K, Dressler LG, Smith LV, Labbok MH, Geradts J, Bensen JT, Jackson S, Nyante S, Livasy C, Carey L, Earp HS, Perou CM: **Epidemiology of basal-like breast cancer.** *Breast Cancer Res Treat* 2008, 109:123–139.
9. O'Brien KM, Cole SR, Tse CK, Perou CM, Carey LA, Foulkes WD, Dressler LG, Geradts J, Millikan RC: **Intrinsic breast tumor subtypes, race, and long-term survival in the Carolina Breast Cancer Study.** *Clin Cancer Res* 2010, **16:**6100–6110.
10. Carey LA, Dees EC, Sawyer L, Gatti L, Moore DT, Collichio F, Ollila DW, Sartor CI, Graham ML, Perou CM: **The triple negative paradox: primary tumor chemosensitivity of breast cancer subtypes.** *Clin Cancer Res* 2007, **13:**2329–2334.
11. Plass C: **Cancer epigenomics.** *Hum Mol Genet* 2000, **11:**2479–2488.
12. Antequera F, Bird A: **Number of CpG islands and genes in human and mouse.** *Proc Natl Acad Sci U S A* 1993, **90:**11995–11999.
13. Bird A, Taggart M, Frommer M, Miller OJ, Macleod D: **A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA.** *Cell* 1985, **40:**91–99.
14. Ogino S, Nosho K, Kirkner GJ, Kawasaki T, Meyerhardt JA, Loda M, Giovannucci EL, Fuchs CS: **CpG island methylator phenotype, microsatellite instability, BRAF mutation and clinical outcome in colon cancer.** *Gut* 2009, **58:**90–96.
15. Newman B, Moorman PG, Millikan R, Qaqish BF, Geradts J, Aldrich TE, Liu ET: **The Carolina Breast Cancer Study: integrating population-based epidemiology and molecular biology.** *Breast Cancer Res Treat* 1995, **35:**51–60.
16. Moorman PG, Newman B, Millikan RC, Tse CK, Sandler DP: **Participation rates in a case–control study: the impact of age, race, and race of interviewer.** *Ann Epidemiol* 1999, **9:**188–195.
17. Dressler LG, Geradts J, Burroughs M, Cowan D, Millikan RC, Newman B: **Policy guidelines for the utilization of formalin-fixed, paraffin-embedded tissue sections: the UNC SPORE experience. University of North Carolina Specialized Program of Research Excellence.** *Breast Cancer Res Treat* 1999, **58:**31–39.
18. Nielsen TO, Hsu FD, Jensen K, Cheang M, Karaca G, Hu Z, Hernandez-Boussard T, Livasy C, Cowan D, Dressler L, Akslen LA, Ragaz J, Gown AM, Gilks CB, van de Rijn M, Perou CM: **Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma.** *Clin Cancer Res* 2004, **10:**5367–5374.
19. Cheang MC, Voduc D, Bajdik C, Leung S, McKinney S, Chia SK, Perou CM, Nielsen TO: **Basal-like breast cancer defined by five biomarkers has superior prognostic value than triple-negative phenotype.** *Clin Cancer Res* 2008, **14:**1368–1376.
20. Li Y, Millikan RC, Carozza S, Newman B, Liu E, Davis R, Miike R, Wrensch M: **p53 mutations in malignant gliomas.** *Cancer Epidemiol Biomarkers Prev* 1998, **7:**303–308.
21. Conway K, Edmiston SN, Khondker ZS, Groben PA, Zhou X, Chu H, Kuan PF, Hao H, Carson C, Berwick M, Olilla DW, Thomas NE: **DNA-methylation profiling distinguishes malignant melanomas from benign nevi.** *Pigment Cell Melanoma Res* 2011, **24:**352–360.
22. Bibikova M, Lin Z, Zhou L, Chudin E, Garcia EW, Wu B, Doucet D, Thomas NJ, Wang Y, Vollmer E, Goldmann T, Seifart C, Jiang W, Barker DL, Chee MS, Floros J, Fan JB: **High-throughput DNA methylation profiling using universal bead arrays.** *Genome Res* 2006, **16:**383–393.
23. Marsit CJ, Christensen BC, Houseman EA, Karagas MR, Wrensch MR, Yeh RF, Nelson HH, Wiemels JL, Zheng S, Posner MR, McClean MD, Wiencke JK, Kelsey KT: **Epigenetic profiling reveals etiologically distinct patterns of DNA methylation in head and neck squamous cell carcinoma.** *Carcinogenesis* 2009, **30:**416–422.
24. Byun HM, Siegmund KD, Pan F, Weisenberger DJ, Kanel G, Laird PW, Yang AS: **Epigenetic profiling of somatic tissues from human autopsy specimens identifies tissue- and individual-specific DNA methylation patterns.** *Hum Mol Genet* 2009, **18:**4808–4817.
25. Lynch AG, Dunning MJ, Iddawela M, Barbosa-Morais NL, Ritchie ME: **Considerations for the processing and analysis of GoldenGate-based two-colour Illumina platforms.** *Stat Methods Med Res* 2009, **18:**437–452.
26. **The R package.** [http://www.r-project.org/]
27. Monti S, Tamayo P, Mesirov J, Golub T: **Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data.** *Mach Learn* 2009, **52:**91–118.
28. Wilkerson M, Waltman P: *ConsensusClusterPlus: ConsensusClusterPlus*; R package version 1.16.0; 2013.
29. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *J R Statist Soc* 1995, **57:**289–300. Series B.
30. Therneau T: *A Package for Survival Analysis in S_. R package version 2.37-7*; [http://CRAN.R-project.org/package=survival]
31. Therneau T, Grambsch P: *Modeling Survival Data: Extending the Cox Model.* Statistics for Biology and Health. Springer-Verlag; 2000.
32. Andersen P, Gill R: **Cox's regression model for counting processes, a large sample study.** *Ann Stat* 1982, 10:1100–1120.
33. Grambsch P, Therneau T: **Proportional hazards tests and diagnostics based on weighted residuals.** *Biometrika* 1994, 81:515–526.
34. **DAVID Bioinformatics Resources 6.7 Functional Annotation Tool.** [http://david.abcc.ncifcrf.gov/home.jsp]
35. The Cancer Genome Atlas Network: **Comprehensive molecular portraits of human breast tumours.** *Nature* 2012, **490:**61–70.
36. Killian JK, Bilke S, Davis S, Walker RL, Jaeger E, Killian MS, Waterfall JJ, Bibikova M, Fan JB, Smith WI Jr, Meltzer PS: **A methyl-deviator epigenotype of estrogen receptor-positive breast carcinoma is associated with malignant biology.** *Am J Pathol* 2011, **179:**55–65.
37. Rønneberg JA, Fleischer T, Solvang HK, Nordgard SH, Edvardsen H, Potapenko I, Nebdal D, Daviaud C, Gut I, Bukholm I, Naume B, Børresen-Dale AL, Tost J, Kristensen V: **Methylation profiling with a panel of cancer related genes: Association with estrogen receptor, TP53 mutation status and expression subtypes in sporadic breast cancer.** *Mol Oncol* 2011, **5:**61–76.
38. Park SY, Kwon HJ, Choi Y, Lee HE, Kim SW, Kim JH, Kim IA, Jung N, Cho NY, Kang GH: **Distinct patterns of promoter CpG island methylation of breast cancer subtypes are associated with stem cell phenotypes.** *Mod Pathol* 2012, **25:**185–196.
39. Bediaga NG, Acha-Sagredo A, Guerra I, Viguri A, Albaina C, Ruiz Diaz I, Rezola R, Alberdi MJ, Dopazo J, Montaner D, de Renobales M, Fernández AF,

Field JK, Fraga MF, Liloglou T, de Pancorbo MM: **DNA methylation epigenotypes in breast cancer molecular subtypes.** *Breast Cancer Res* 2010, **12:**R77.

40. Piotrowski A, Benetkiewicz M, Menzel U, Díaz de Ståhl T, Mantripragada K, Grigelionis G, Buckley PG, Jankowski M, Hoffman J, Bała D, Srutek E, Laskowski R, Zegarski W, Dumanski JP: **Microarray-based survey of CpG islands identifies concurrent hyper- and hypomethylation patterns in tissues derived from patients with breast cancer.** *Genes Chromosomes Cancer* 2006, **45:**656–667.
41. Moelans CB, Verschuur-Maes AH, van Diest PJ: **Frequent promoter hypermethylation of BRCA2, CDH13, MSH6, PAX5, PAX6 and WT1 in ductal carcinoma in situ and invasive breast cancer.** *J Pathol* 2011, **225:**222–231.
42. Virmani A, Rathi A, Sugio K, Sathyanarayana UG, Toyooka S, Kischel FC, Tonk V, Padar A, Takahashi T, Roth JA, Euhus DM, Minna JD, Gazdar AF: **Aberrant methylation of TMS1 in small cell, non-small cell lung cancer and breast cancer.** *Int J Cancer* 2003, **106:**198–204.
43. Mirza S, Sharma G, Prasad CP, Parshad R, Srivastava A, Gupta SD, Ralhan R: **Promoter hypermethylation of TMS1, BRCA1, ERalpha and PRB in serum and tumor DNA of invasive ductal breast carcinoma patients.** *Life Sci* 2007, **81:**280–287.
44. Holm K, Hegardt C, Staaf J, Vallon-Christersson J, Jonsson G, Olsson H, Borg A, Ringner M: **Molecular subtypes of breast cancer are associated with characteristic DNA methylation patterns.** *Breast Cancer Res* 2010, **12:**R36.
45. Fackler MJ, Umbricht CB, Williams D, Argani P, Cruz LA, Merino VF, Teo WW, Zhang Z, Huang P, Visvananthan K, Marks J, Ethier S, Gray JW, Wolff AC, Cope LM, Sukumar S: **Genome-wide methylation analysis identifies genes specific to breast cancer hormone receptor status and risk of recurrence.** *Cancer Res* 2011, **71:**6195–6207.
46. Kamalakaran S, Varadan V, Giercksky Russnes HE, Levy D, Kendall J, Janevski A, Riggs M, Banerjee N, Synnestvedt M, Schlichting E, Kåresen R, Shama Prasada K, Rotti H, Rao R, Rao L, Eric Tang MH, Satyamoorthy K, Lucito R, Wigler M, Dimitrova N, Naume B, Borresen-Dale AL, Hicks JB: **DNA methylation patterns in luminal breast cancers differ from non-luminal subtypes and can identify relapse risk independent of other clinical variables.** *Mol Oncol* 2011, **5:**77–92.
47. Christensen BC, Kelsey KT, Zheng S, Houseman EA, Marsit CJ, Wrensch MR, Wiemels JL, Nelson HH, Karagas MR, Kushi LH, Kwan ML, Wiencke JK: **Breast cancer DNA methylation profiles are associated with tumor size and alcohol and folate intake.** *PLoS Genet* 2010, **6:**e1001043.
48. Bloushtain-Qimron N, Yao J, Snyder EL, Shipitsin M, Campbell LL, Mani SA, Hu M, Chen H, Ustyansky V, Antosiewicz JE, Argani P, Halushka MK, Thomson JA, Pharoah P, Porgador A, Sukumar S, Parsons R, Richardson AL, Stampfer MR, Gelman RS, Nikolskaya T, Nikolsky Y, Polyak K: **Cell type-specific DNA methylation patterns in the human breast.** *Proc Natl Acad Sci U S A* 2008, **105:**14076–14081.
49. Tomaskovic-Crook E, Thompson EW, Thiery JP: **Epithelial to mesenchymal transition and breast cancer.** *Breast Cancer Res* 2009, **11:**213.
50. Cohen AL, Piccolo SR, Cheng L, Soldi R, Han B, Johnson WE, Bild AH: **Genomic pathway analysis reveals that EZH2 and HDAC4 represent mutually exclusive epigenetic pathways across human cancers.** *BMC Med Genomics* 2013, **6:**35.
51. Noushmehr H, Weisenberger DJ, Diefes K, Phillips HS, Pujara K, Berman BP, Pan F, Pelloski CE, Sulman EP, Bhat KP, Verhaak RG, Hoadley KA, Hayes DN, Perou CM, Schmidt HK, Ding L, Wilson RK, Van Den Berg D, Shen H, Bengtsson H, Neuvial P, Cope LM, Buckley J, Herman JG, Baylin SB, Laird PW, Aldape K, Cancer Genome Atlas Research Network: **Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma.** *Cancer Cell* 2010, **17:**510–522.
52. Shinjo K, Okamoto Y, An B, Yokoyama T, Takeuchi I, Fujii M, Osada H, Usami N, Hasegawa Y, Ito H, Hida T, Fujimoto N, Kishimoto T, Sekido Y, Kondo Y: **Integrated analysis of genetic and epigenetic alterations reveals CpG island methylator phenotype associated with distinct clinical characters of lung adenocarcinoma.** *Carcinogenesis* 2012, **33:**1277–1285.
53. Zhang QY, Yi DQ, Zhou L, Zhang DH, Zhou TM: **Status and significance of CpG island methylator phenotype in endometrial cancer.** *Gynecol Obstet Invest* 2011, **72:**183–191.
54. Nosho K, Irahara N, Shima K, Kure S, Kirkner GJ, Schernhammer ES, Hazra A, Hunter DJ, Quackenbush J, Spiegelman D, Giovannucci EL, Fuchs CS, Ogino S: **Comprehensive biostatistical analysis of CpG island methylator phenotype in colorectal cancer using a large population-based sample.** *PLoS One* 2008, **3:**e3698.
55. Strathdee G, Appleton K, Illand M, Millan DW, Sargent J, Paul J, Brown R: **Primary ovarian carcinomas display multiple methylator phenotypes involving known tumor suppressor genes.** *Am J Pathol* 2001, **158:**1121–1127.
56. Toyota M, Ahuja N, Suzuki H, Itoh F, Ohe-Toyota M, Imai K, Baylin SB, Issa JP: **Aberrant methylation in gastric cancer associated with the CpG island methylator phenotype.** *Cancer Res* 1999, **59:**5438–5442.
57. Jithesh PV, Risk JM, Schache AG, Dhanda J, Lane B, Liloglou T, Shaw RJ: **The epigenetic landscape of oral squamous cell carcinoma.** *Br J Cancer* 2013, **108:**370–379.
58. Van der Auwera I, Yu W, Suo L, Van Neste L, van Dam P, Van Marck EA, Pauwels P, Vermeulen PB, Dirix LY, Van Laere SJ: **Array-based DNA methylation profiling for breast cancer subtype discrimination.** *PLoS One* 2010, **5:**e12616.
59. Fang F, Turcan S, Rimner A, Kaufman A, Giri D, Morris LG, Shen R, Seshan V, Mo Q, Heguy A, Baylin SB, Ahuja N, Viale A, Massague J, Norton L, Vahdat LT, Moynahan ME, Chan TA: **Breast cancer methylomes establish an epigenomic foundation for metastasis.** *Sci Transl Med* 2011, **3:**75ra25.
60. Roll JD, Rivenbark AG, Jones WD, Coleman WB: **DNMT3b overexpression contributes to a hypermethylator phenotype in human breast cancer cell lines.** *Mol Cancer* 2008, **7:**15.
61. Nosho K, Shima K, Irahara N, Kure S, Baba Y, Kirkner GJ, Chen L, Gokhale S, Hazra A, Spiegelman D, Giovannucci EL, Jaenisch R, Fuchs CS, Ogino S: **DNMT3B expression might contribute to CpG island methylator phenotype in colorectal cancer.** *Clin Cancer Res* 2009, **15:**3663–3671.
62. van Hoesel AQ, van de Velde CJ, Kuppen PJ, Putter H, de Kruijf EM, van Nes JG, Giuliano AE, Hoon DS: **Primary tumor classification according to methylation pattern is prognostic in patients with early stage ER-negative breast cancer.** *Breast Cancer Res Treat* 2011, **131:**859–869.
63. Ulirsch J, Fan C, Knafl G, Wu MJ, Coleman B, Perou CM, Swift-Scanlan T: **Vimentin DNA methylation predicts survival in breast cancer.** *Breast Cancer Res Treat* 2013, **137:**383–396.
64. Xu J, Shetty PB, Feng W, Chenault C, Bast RC Jr, Issa JP, Hilsenbeck SG, Yu Y: **Methylation of HIN-1, RASSF1A, RIL and CDH13 in breast cancer is associated with clinical characteristics, but only RASSF1A methylation is associated with outcome.** *BMC Cancer* 2012, **12:**243.
65. Maier S, Nimmrich I, Koenig T, Eppenberger-Castori S, Bohlmann I, Paradiso A, Spyratos F, Thomssen C, Mueller V, Nährig J, Schittulli F, Kates R, Lesche R, Schwope I, Kluth A, Marx A, Martens JW, Foekens JA, Schmitt M, Harbeck N, European Organisation for Research and Treatment of Cancer (EORTC) PathoBiology group: **DNA-methylation of the homeodomain transcription factor PITX2 reliably predicts risk of distant disease recurrence in tamoxifen-treated, node-negative breast cancer patients-Technical and clinical validation in a multi-centre setting in collaboration with the European Organisation for Research and Treatment of Cancer (EORTC) PathoBiology group.** *Eur J Cancer* 2007, **43:**1679–1686.
66. Tommasi S, Karm DL, Wu X, Yen Y, Pfeifer GP: **Methylation of homeobox genes is a frequent and early epigenetic event in breast cancer.** *Breast Cancer Res* 2009, **11:**R14.
67. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z, Quackenbush JF, Stijleman IJ, Palazzo J, Marron JS, Nobel AB, Mardis E, Nielsen TO, Ellis MJ, Perou CM, Bernard PS: **Supervised risk predictor of breast cancer based on intrinsic subtypes.** *J Clin Oncol* 2009, **27:**1160–1167.

doi:10.1186/s13058-014-0450-6
Cite this article as: Conway *et al.*: **DNA methylation profiling in the Carolina Breast Cancer Study defines cancer subclasses differing in clinicopathologic characteristics and survival.** *Breast Cancer Research* 2014 **16:**450.