

# Preliminary Analysis of Within-Sample Co-methylation Patterns in Normal and Cancerous Breast Samples

Lillian Sun<sup>1</sup>, Surya Namboodiri<sup>2</sup>, Emily Chen<sup>3</sup> and Shuying Sun<sup>4</sup> 

<sup>1</sup>Stanford University, Stanford, CA, USA. <sup>2</sup>The University of Texas at Dallas, Richardson, TX, USA.

<sup>3</sup>Brown University, Providence, RI, USA. <sup>4</sup>Department of Mathematics, Texas State University, San Marcos, TX, USA.

Cancer Informatics  
Volume 18: 1–14  
© The Author(s) 2019  
Article reuse guidelines:  
[sagepub.com/journals-permissions](http://sagepub.com/journals-permissions)  
DOI: 10.1177/1176935119880516



**ABSTRACT:** DNA methylation plays a significant role in regulating the expression of certain genes in both cancerous and normal breast tissues. It is therefore important to study within-sample co-methylation, ie, methylation patterns between consecutive sites in a chromosome. In this article, we develop 2 new methods to compare co-methylation patterns between normal and cancerous breast samples. In particular, we investigate the co-methylation patterns of 4 different methylation states/levels separately. Using these 2 methods, we focus on addressing the following questions: How often does 1 methylation state change to other methylation states and how is this change dependent on chromosome distance? What co-methylation patterns do normal and cancerous breast samples have? Do genomic sites with different methylation states/levels have different co-methylation patterns? Our results show that cancerous and normal co-methylation patterns are significantly different. We find that this difference exists even when the physical distance of 2 sites are less than 50 bases. Breast cancer cell lines tend to remain in the same methylation state more often than normal samples, especially for the no/low or high/full methylation states. We also find that the co-methylation region lengths for various methylation states (no/low, partial, and high/full methylation states) are very different. For example, the co-methylation region lengths for partial methylation regions are shorter than the unmethylated or fully methylated regions. Our research may provide a deep understanding of co-methylation patterns. These co-methylation patterns will aid in discovering and understanding new methylation events that may be related to novel biomarkers.

**KEYWORDS:** Within-sample co-methylation, bioinformatics, breast cancer

**RECEIVED:** September 11, 2019. **ACCEPTED:** September 14, 2019.

**TYPE:** Translating Cancer Methylation Data Results from Bench to Bedside Using Informatics Tools - Original Research

**FUNDING:** The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This project was supported by Dr.

Sun's Start-Up Funds and the Texas State University Research Enhancement Program.

**DECLARATION OF CONFLICTING INTERESTS:** The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**CORRESPONDING AUTHOR:** Shuying Sun, Department of Mathematics, Texas State University, 601 University Drive, San Marcos, TX 78666, USA. Email: [ssun@txstate.edu](mailto:ssun@txstate.edu)

## Introduction

Cancer is the second leading cause of death in the United States. In 2019, there is an estimated 1 762 450 new cancer cases diagnosed and 606 880 cancer deaths in the United States. Even with advanced technology and medicine, 1 in 3 cancer cases are terminal.<sup>1</sup> However, cancer mortality can be reduced if the disease is detected and treated early; in fact, cancer can be reduced or controlled by implementing evidence-based strategies for cancer prevention, early detection of cancer, and proper treatment of cancer patients.<sup>2</sup> Thus, it is very important to identify genetic and epigenetic factors associated with cancer so that medical professionals will be able to more effectively identify and treat cancer cases. DNA methylation is an important epigenetic event that is associated with cancer. In this article, we will study and compare co-methylation patterns in cancerous and normal breast samples in an attempt to understand whether co-methylation of certain degrees and/or in certain genomic regions is associated with breast cancer. Next, we will review the definition and roles of DNA methylation.

In a mammalian genome, DNA methylation is the covalent bonding of a methyl group (CH<sub>3</sub>) to a cytosine base in a dinucleotide 5'-CG-3', or a CG site (also known as a CpG site, which is a cytosine base linked to a guanine base by a phosphate bond in the DNA sequence).<sup>3</sup> DNA methylation of some CG sites may be related to gene expression loss, especially for some key tumor suppressor genes (TSGs). It usually

restricts transcription factors from gaining access to the gene promoter and therefore turns off gene transcription. This function of DNA methylation often associates with chromatin structure changes. The chromatin becomes more condensed instead of remaining open and functional as is required for gene transcription.<sup>3</sup> DNA methylation has also been shown to potentially affect gene control via other means. It may stimulate transcription elongation, impact splicing, and alter the activities of enhancers, which are regions of DNA that can promote transcription.<sup>4,5</sup> In promoter regions, the methylation of CpG islands (regions rich in CG sites) can also be associated with long-term silencing of gene expression.<sup>6,7</sup> Although both methylation and mutation can silence TSGs, methylation is potentially reversible.<sup>3,8</sup> Thus, ongoing biomedical research is attempting to use demethylating agents to treat cancer cells, because these agents can activate TSGs and suppress tumor proliferation.<sup>9</sup>

Because DNA methylation plays a key role in regulating the expression of some genes, it is necessary to study methylation patterns thoroughly to fully understand cancerous DNA. Co-methylation, or spatial correlation, is an important type of methylation patterns.<sup>10-16</sup> Generally speaking, there are 2 main types of co-methylation<sup>12,16</sup>: between-sample (BS) co-methylation and within-sample (WS) co-methylation. The BS co-methylation describes the methylation similarity or correlation of CG sites or genes across a set of samples and in different



genomic locations (eg, genes can be on different chromosomes).<sup>17-21</sup> For this type of co-methylation, the WGCNA R package<sup>22</sup> is often used to find co-methylated CG sites, genes, or modules.<sup>23-25</sup> WS co-methylation describes the degree of methylation over distance, ie, similar methylation patterns in nearby CG sites located in the same chromosome of 1 single sample.<sup>11,14,26-28</sup> In addition to the above description, in the Introduction of Sun and Sun,<sup>16</sup> there is a clear explanation of the difference between WS and BS co-methylation patterns.

As mentioned above, consecutive CG sites in a chromosome region tend to have similar methylation, un-methylation, or partial methylation patterns. This WS co-methylation pattern decreases as the genomic distance increases.<sup>11-13,29,30</sup> The relationship between co-methylation decay and genomic distance can vary in different cell types and tissues. This relationship is also disputable due to the reports of different levels of co-methylation decay.<sup>12,15</sup> Although co-methylation has been studied in different cell types and tissues, to the best of our knowledge, it has not been well studied yet in normal breast tissues and breast cancer cell lines using the whole genome bisulfite sequencing (WGBS) data. In addition, previous studies investigate WS co-methylation patterns by considering CG sites with different methylation levels/states all together. However, the co-methylation patterns of unmethylated, partially methylated, and highly/fully methylated sites can have very different co-methylation patterns. Conducting WS co-methylation studies without separating CG sites with different methylation levels will lead to inaccurate results. Therefore, CG sites with different methylation levels/states should not be considered together when studying WS co-methylation. In this article, we will conduct WS co-methylation for breast tissues using WGBS data by considering CG sites with different methylation levels/states separately. Note that there is a publication on co-methylation for breast cancer by Akulenko and Helms.<sup>17</sup> However, their study is for BS co-methylation, not for WS co-methylation, and they use the Illumina 27K array data, not the WGBS data. An Illumina 27K data set can only have methylation signals for 0.1% of what WGBS data set can include.

We will focus on studying WS co-methylation patterns in breast tissues. Because there are not many WGBS data publicly available for normal breast tissues and breast cancer cell lines, we will conduct the analysis using 3 currently available samples: 1 normal breast sample and 2 breast cancer cell lines. In particular, we will conduct statistical analyses to address several questions. How different are the WS co-methylation patterns in cancerous and normal samples? Are the co-methylation patterns of various methylation levels/states (eg, no/low, partial, or high/full methylation) different? How often does a methylation state (low, high, or partial methylation state) remain the same in cancerous and normal samples? Are the length distributions of cancerous and normal co-methylation regions the same? If not, to what degree do they differ and what is the

significance of these differences? To answer these questions, we will study the relationship of WS co-methylation patterns of consecutive CG sites across a chromosome and compare co-methylation patterns in both cancerous and normal samples.

To the best of our knowledge, this research work is the first study that focuses on investigating breast tissue WS co-methylation patterns by considering different methylation states separately using WGBS data. Analyses based on WGBS data can provide a more comprehensive understanding of WS co-methylation patterns. Studying cancerous and normal breast samples' co-methylation patterns will also allow us to further understand how specific co-methylation patterns are associated with breast cancer. Answers to the above questions can add to the understanding of the causes or hallmarks of breast cancer. Furthermore, an increased understanding of co-methylation patterns in different samples may also help researchers recover lost information in low-coverage methylation sequencing data<sup>31</sup> and even lead to more efficient methylation sequencing.<sup>12</sup> To simplify our writing for the rest of this article, when we use "co-methylation," we mean "WS co-methylation."

## Method

### *Data and methylation state definition*

We use publicly available DNA methylation sequencing data (GSE29127 and GSM3526804) generated for normal human mammary epithelial cells (HMECs) and 2 breast cancer cell lines, HCC1954<sup>32</sup> and MCF7.<sup>33</sup> The raw sequencing reads of these samples are generated using the WGBS technique and are saved in the FASTQ format. BRAT-bw,<sup>34</sup> a publicly available software package, is used to preprocess and align raw sequencing reads with the reference genome hg19. After processing the raw sequencing data, each methylation data set includes information on chromosome number, base position, sequencing coverage, and methylation ratio (MC ratio) (see Table 1). Sequencing coverage refers to the number of times a nucleotide is read or sequenced. A relatively high sequencing coverage may indicate more accurate sequencing results; thus, we choose to use the base positions or CG sites with at least  $3\times$  coverage for more accurate results.

The MC ratio for each base position in a chromosome ranges from 0 to 1, where 0 indicates no methylation and 1 indicates full methylation. To remove the impact of sequencing error and to simplify the analysis, we define 4 methylation states: "A" for no methylation or low methylation levels in the range of [0, 0.25), "B" for low partial methylation levels in the range of [0.25, 0.5), "C" for high partial methylation levels in the range of [0.5, 0.75), and "D" for full methylation or high methylation levels in the range of [0.75, 1]. "NA" is used for missing data. We define methylation states for all CG sites and then add them as an additional column to our data. We also calculate the distances between consecutive CG sites and add them to our data.

**Table 1.** Sample section of chromosome 1 breast cancer data.

CHR	POSITION	SEQUENCE COVERAGE	MC RATIO	METHYLATION STATE	DISTANCE
chr1	534314	6	0.666667	C	12
chr1	534326	3	1	D	3
chr1	534329	4	1	D	14
chr1	534343	4	0.25	B	17
chr1	534360	4	0.75	D	45
chr1	534405	3	0.666667	C	31
chr1	534436	0	NA	NA	108

Abbreviation: MC, methylation ratio.

As for which chromosome or region to study, our explorative analysis on a short chromosome has shown similar patterns and answers when comparing with the long chromosome analysis. In addition, because our focus is WS co-methylation for nearby CG sites, not BS co-methylation in a whole genome, using 1 chromosome is sufficient to address the questions of interest. Therefore, we only focus on chromosome 1, as it is the longest chromosome.

### Two analysis methods

We first determine the methylation state of each CG site as A, B, C, or D based on its methylation level as mentioned before. The methylation states A, B, C, and D represent low (or no), low partial, high partial, and high (or full) methylation, respectively, as explained above. We then study WS co-methylation patterns using 2 different analysis methods. First, we study how often or how frequently a methylation state (eg, A) remains the same (eg, AA) or changes to other methylation states (eg, AB, AC, or AD). Second, we investigate how long each region of a specific methylation state is (eg, AAAA . . .). We conduct analyses under several different sets of conditions, including different distance levels. Next, we use the chi-square test and the Wilcoxon rank sum test to determine whether the differences we observe between normal and cancerous DNA are statistically significant. The key novelty of our methods is that we study the WS co-methylation patterns by studying CG sites with different methylation states (A, B, C, and D) separately for breast tissues. The co-methylation patterns of CG sites that have different methylation levels can be different. When considering the CG sites that have different methylation levels together, the co-methylation analysis results will not be accurate. Therefore, we conduct our analysis for each methylation level/state separately to obtain accurate co-methylation analysis results for breast tissues.

*Method 1: analyze the relationship between methylation state changes and distance.* We first look at consecutive pairs of CG sites to see whether there is a pattern in how often 1

methylation state remains the same or changes to another methylation state. We divide our data based on the distance between consecutive CG sites to determine whether the observed patterns are only present among CG sites that are a short distance from each other or these patterns are also apparent over a longer distance. We first study the CG sites in the distance intervals [0, 200), [200, 500), and [500, Infinite [or Inf]). We then conduct further analysis using distance intervals incremented by 50: [0, 50), [50, 100), [100, 150), and so on, to [500, Inf). The distance is measured by the number of bases between consecutive CG sites. For example, when using the distance interval [0, 50), we select CG sites that are 0 to 50-base away from the next consecutive CG site. Our results for these distance intervals are shown in the “Results” section.

*Method 2: analyze the distribution of co-methylation region length.* We investigate the WS co-methylation patterns by studying the distribution of the co-methylation region length. We identify the co-methylation regions that have the same methylation state (eg, AAAAA or BBBBB). We then count the number of CG sites in each region and calculate the length of each region. Note that the co-methylation regions of different methylation states/levels can have different co-methylation patterns (eg, different co-methylation region lengths). We will study different methylation states/levels separately. In addition, the co-methylation regions of different methylation states/levels may have different numbers of CG sites, and most of these regions may consist of a small number of CG sites. We will conduct further analysis on numbers of CG sites in all co-methylation regions.

## Results

### Results of method 1

Our method 1 analysis results are summarized in Tables 2 to 5 and Figures 1 and 2. The first portion of our analysis is shown in Table 2 and Figures 1 and 2. Table 2 indicates how often 1 methylation state changes to another when the distance levels between 2 consecutive CG sites are 0 ~ 200 bases, 200 ~ 500

**Table 2.** Frequency of each methylation state change between consecutive CG Sites.

	DISTANCE	STATE (%)	A	B	C	D
Normal breast tissue (HMEC)	[0, 200)	A	74.023	6.737	6.401	12.839
		B	19.504	15.226	18.13	47.14
		C	11.33	10.798	18.947	58.925
		D	3.726	4.656	9.664	81.953
	[200, 500)	A	25.517	13.172	17.084	44.226
		B	17.459	13.454	18.274	50.813
		C	14.065	11.868	17.02	57.047
		D	8.4	7.476	12.754	71.37
	[500, Inf)	A	20.769	13.505	18.161	47.565
		B	17.769	14.098	18.916	49.218
		C	15.841	12.355	16.92	54.884
		D	9.554	7.78	12.704	69.962
Cancerous breast tissue (HCC1954)	[0, 200)	A	78.109	8.521	5.177	8.192
		B	29.762	21.303	18.234	30.701
		C	13.009	13.128	20.433	53.43
		D	2.802	2.941	7.187	87.07
	[200, 500)	A	61.845	12.098	8.824	17.233
		B	32.744	15.121	15.946	36.189
		C	18.333	11.779	16.348	53.54
		D	6.382	5.335	9.774	78.509
	[500, Inf)	A	63.188	12.141	8.801	15.869
		B	34.707	16.363	15.401	33.53
		C	20	13.03	16.145	50.825
		D	7.72	5.704	10.39	76.186
Cancerous breast tissue (MCF7)	[0, 200)	A	80.565	7.379	5.035	7.02
		B	25.821	22.294	20.165	31.719
		C	10.358	11.716	22.022	55.904
		D	1.71	2.219	6.713	89.357
	[200, 500)	A	53.536	13.723	12.088	20.654
		B	28.046	14.891	17.737	39.326
		C	16.117	11.487	17.55	54.845
		D	5.562	5.336	11.264	77.838
	[500, Inf)	A	57.04	13.038	11.614	18.308
		B	31.504	15.667	17.877	34.953
		C	18.535	12.939	18.586	49.941
		D	7.735	6.562	13.143	72.56

Abbreviation: HMEC, human mammary epithelial cell.

Distance is the number of base pairs between 2 consecutive CG sites. Cells show the percentages of CG sites with a specific methylation state remaining the same or changing to a different state in the next consecutive CG site.

**Table 3.** Methylation state changes in chromosome 1 for normal and cancerous data.

		A	B	C	D
Normal (HMEC)	A count	233 808	24 568	24 753	52 697
	A%	69.622	7.316	7.371	15.692
	B count	24 443	19 061	23 208	60 942
	B%	19.148	14.932	18.18	47.74
	C count	24 709	22 917	38 791	121 969
	C%	11.857	10.997	18.615	58.53
	D count	52 596	60 721	121 570	973 855
	D%	4.351	5.023	10.058	80.568
Cancerous (HCC1954)	A count	328 514	39 165	24 707	40 949
	A%	75.80%	9.00%	5.70%	9.50%
	B count	39 096	26 029	22 900	40 656
	B%	30.40%	20.20%	17.80%	31.60%
	C count	24 581	22 795	34 766	93 966
	C%	14.00%	12.90%	19.70%	53.40%
	D count	40 813	40 416	93 585	1 068 796
	D%	3.30%	3.30%	7.50%	85.90%
Cancerous (MCF7)	A count	236 253	25 277	18 400	27 103
	A%	76.947	8.233	5.993	8.827
	B count	25 372	19 960	18 880	31 782
	B%	26.431	20.793	19.668	33.108
	C count	18 502	18 807	34 024	89 049
	C%	11.537	11.726	21.214	55.523
	D count	26 903	31 898	89 217	1 079 114
	D%	2.192	2.599	7.27	87.938

Abbreviation: HMEC, human mammary epithelial cell.

The number in each cell represents the count or percentage of CG sites that display the specific methylation change.

**Table 4.** 0-Inf chromosome 1 chi-square test results.

VALUE	A	B	C	D
P-values	~0	~0	~0	~0
chi-square	11 756.38	9951.727	1570.923	30 424.83

The input of this chi-square test is shown in Table 3 with no restriction on the distance between CG sites. A, B, C, and D columns represent the results of tests conducted for each of the 4 methylation states.

bases, and larger than 500 bases respectively. For example, in Table 2, the cell in the A% row and the A column of the normal data [0, 200) is 74.023%, which means that the methylation state A remains as the state A 74.023% of the time, whereas the cell in the A% row and then B column of the normal data [0, 200) means that 6.737% times that the state A

changes to state B. The percentages are out of the row total. We summarize the patterns observed for the 0 ~ 200 base interval in Table 2.

First, for methylation states A and D (ie, the A% and D% rows), in both the normal (HMEC) and cancerous (HCC1954 and MCF7) data, the largest percentages occur in the AA and



**Table 5.** *P*-values of chr1 chi-square test with smaller distance restrictions.

DISTANCE INTERVAL	A	B	C	D
0-10	0	0	7.63E-291	0
10-20	5.41E-283	0	2.10E-155	0
20-30	1.46E-255	1.30E-244	5.16E-72	0
30-40	1.06E-282	4.12E-210	8.51E-42	0
40-50	4.78E-252	1.29E-122	4.21E-21	0
50-60	1.28E-240	4.22E-95	1.38E-13	0
60-70	2.47E-235	1.68E-84	4.72E-11	0
70-80	3.46E-239	5.96E-55	7.17E-15	3.45E-295
80-90	1.58E-222	2.42E-49	4.62E-08	6.62E-228
90-100	5.45E-211	1.48E-52	6.53E-07	3.84E-206
100-Inf	0	0	1.72E-99	0

Each “E-number” means “10<sup>-number</sup>.” For example, 5.41E-283 =  $5.41 \times 10^{-283}$ . Shown are test results using data with restrictions on distance between CG sites. Distance levels are shown in the first column. All distance intervals show a significant difference between normal and cancerous data.

DD cells, indicating that low or high methylation states tend to remain in the same state. However, these AA and DD percentages are higher in breast cancer cell lines (eg, 74.023% of AA for the normal breast sample HMEC, but 78.109% for HCC1954 and 80.565% for MCF7). This shows that co-methylation regions of A states and D states tend to be longer in cancerous data than in normal data.

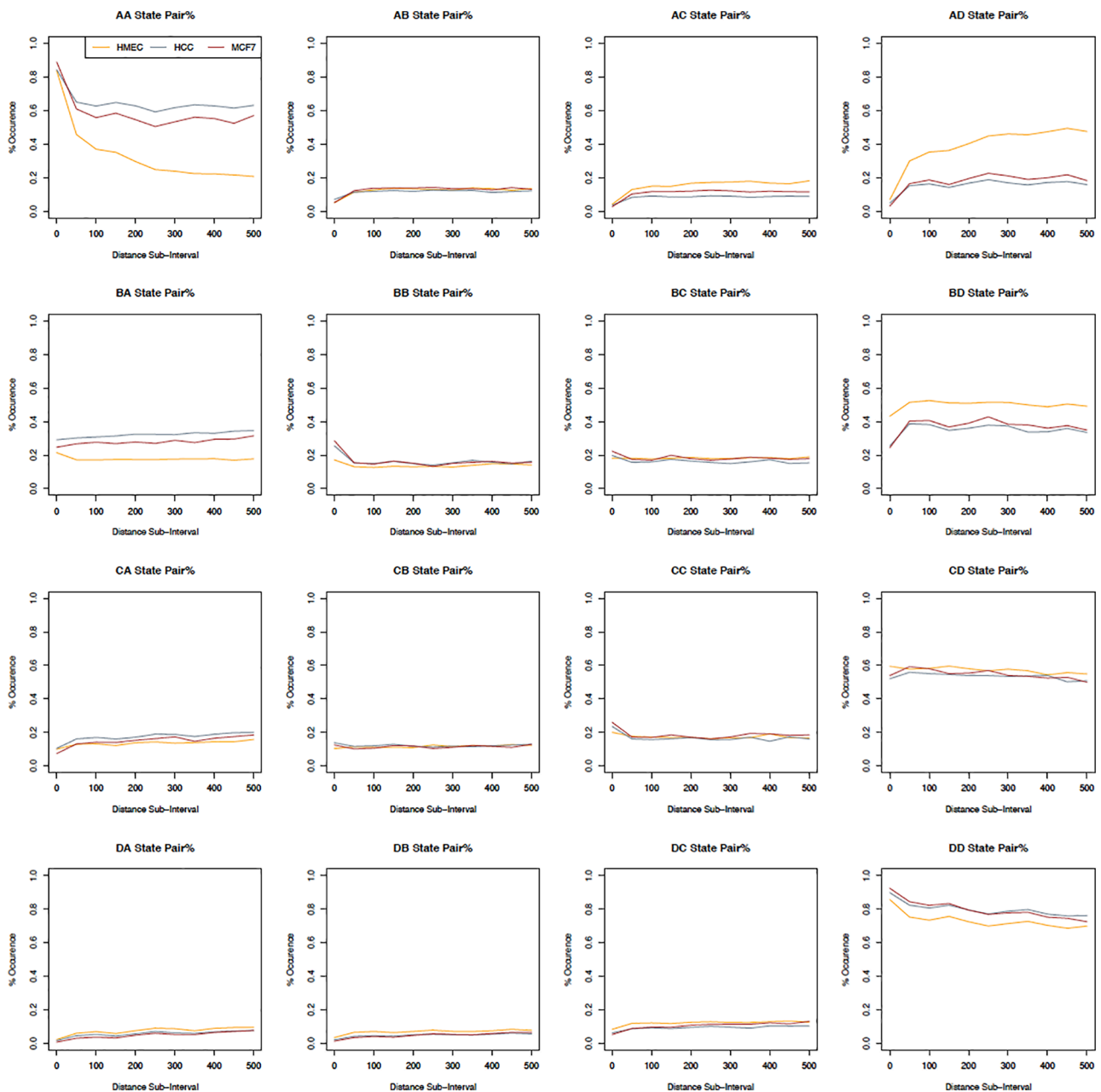
Second, for the partial methylation state B (ie, the B% row), it is more likely to become a D state in the normal HMEC sample (47.14%) than in the cancerous samples (30.701% for HCC1954 and 31.719% for MCF7). For the partial methylation state C, it has a similar pattern for the CD cells of the 3 samples. In fact, the CD% in the normal sample HMEC is higher than in 2 cancerous samples, 58.925% for the normal HMEC, 53.43% for the HCC1954, and 55.904% for the MCF7. For the B% row, the 4 transitions (BA, BB, BC, and BD) are more evenly distributed in 2 cancerous samples than in the normal sample.

Next, we summarize the patterns observed for the [200, 500) and [500, Inf) intervals. First, when the distance level increases, the AA% drops more dramatically in the normal HMEC sample than in the HCC1954 and MCF7. In fact, compared with the 0 ~ 200 base interval, in the HMEC, the AA% changes from 74.023% to 20.769%, ie, the difference is about 50%. In the HCC1954, the AA% changes from 78.109% to 63.188%, ie, the difference is just about 15%. In the MCF7, the AA% changes from 80.565% to 57.04%, ie, the difference is just about 23%. Although there is a large difference when comparing the cancerous with the normal samples based on the AA%, there is a small difference when comparing them using the DD%. Second, when the distance level increases, B% and C% rows do not change as dramatically as the A% row

does, especially in the normal HMEC sample. Finally, when comparing the percentages in all distance levels, we find that as the distance level increases, the percentages decrease in some cells but increase in others. At the last distance level, the 500 ~ Inf interval, the patterns differ greatly when compared with the [0, 200) interval, as there is a large distance between 2 consecutive sites (ie, >500 bases) in the last distance level.

After analyzing percentages of methylation-state-change using the distance levels of [0, 200), [200, 500), and [500, Inf), we find that between the distance intervals [0, 200) and [200, 500), there is a dramatic change in both normal and cancerous co-methylation patterns. It is unclear at what distance-level the co-methylation pattern starts to decay or change in breast tissues. We want to zoom in to pinpoint exactly where this drop occurs. We then use the distance intervals increased by 50 to get a closer look at the patterns (see Figure 1). These intervals are [0, 50), [50, 100), [100, 150), and so on, to [500, Inf). We further decrease the distance intervals to 10 to get an even closer look: [0, 10), [10, 20), [20, 30), and so on, to [90, 100). We compare the percentage methylation-state-change in 2 ways: percentage occurrence of CG pairs with the same first base (eg, AAs, ABs, ACs, and ADs; comparing graphs in 1 row of Figures 1 and 2), and percentage occurrence in cancerous vs normal data (comparing within each graph of Figures 1 and 2). We will explain the comparison results below.

In Figure 1, when comparing the breast cancer cell lines (gray and brown) with the normal sample (yellow), the AA and AD plots show that they have dramatic differences, but overall, the 2 cancer cell lines have similar co-methylation patterns. In the BA, BD, and DD plots, there are certain differences too, but not as much as the AA and AD plots. In addition, we observe similar trends as shown in Table 2. That is, the

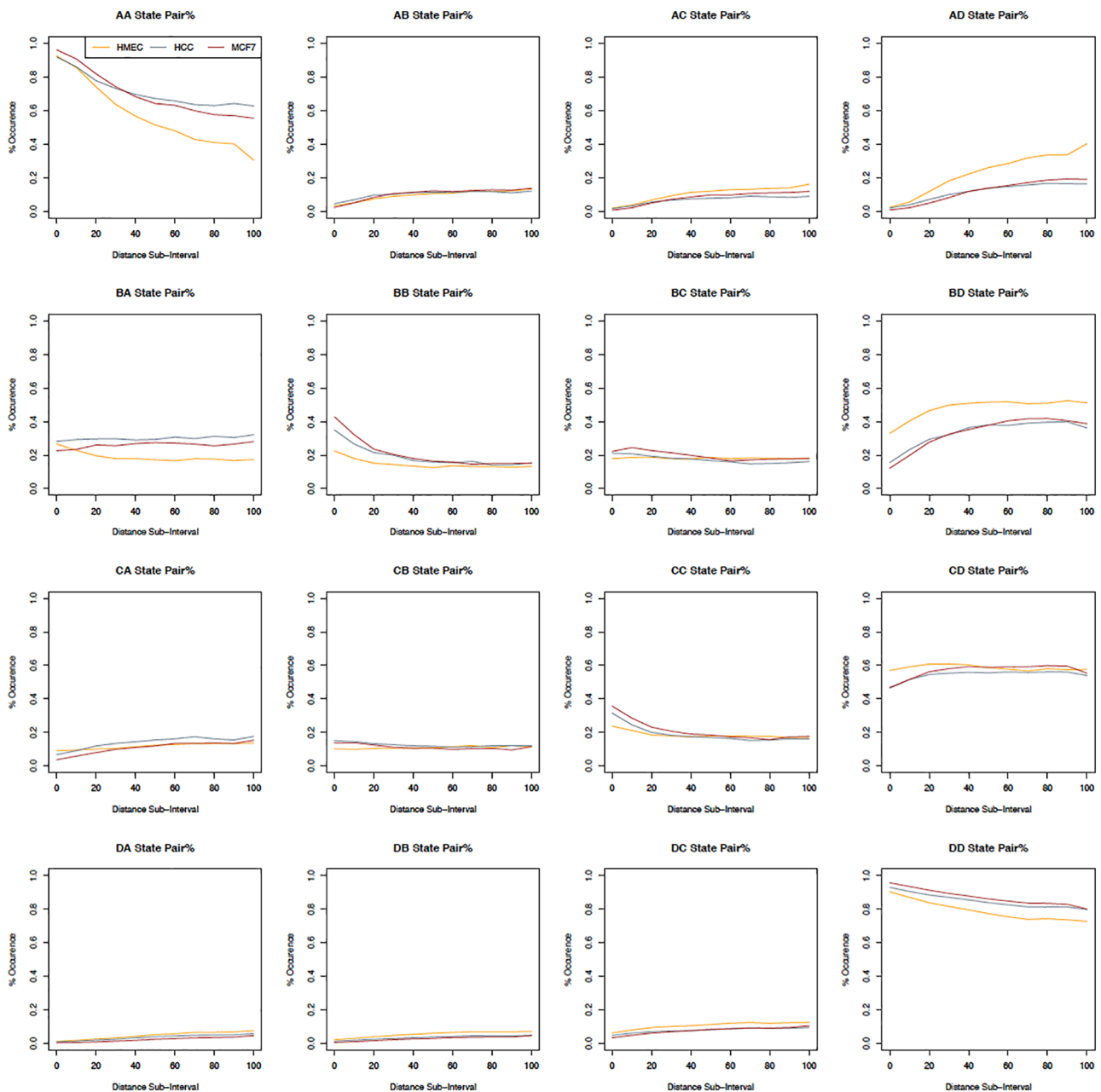


**Figure 1.** Methylation state changes between consecutive CG sites based on 50 base intervals. HMEC indicates human mammary epithelial cell. This figure displays the percentage of methylation state changes with an initial methylation state of A, B, C, and D. The horizontal axis is the distance between consecutive CG sites. The vertical axis is the percentage of occurrence on a 0 to 1 scale. The lines of 3 colors are for HMEC (yellow), HCC1954 (gray), and MCF7 (brown), respectively.

cancerous AA and DD state pairs and the normal DD state pairs have the highest percentages. For example, the 2 cancerous lines (gray for HCC1954 and brown for MCF7) in the AA state pair (or plot) are higher than the 2 cancerous lines in AB, AC, and AD graphs. In Figure 1, we also discover that the percentage of occurrence changes within the first 100 base pairs in each set of CG pairs. The most dramatic changes are shown in the AA and AD state pairs (the first and last plots of the top panel/row of Figure 1), where the likelihood of high and low methylation CG pairs changes significantly within the first 100 base pairs. Next, we zoom in to observe and pinpoint the patterns in the first 0 to 100 base regions by doing a similar

analysis for the intervals [0, 10), [10, 20), [20, 30), and so on, to [90, 100). We observe that the percentage of occurrence for the CG pairs AA and DD actually changes within 20 base pairs (eg, the AA, AD, BD, and DD plots of Figure 2).

Figure 2 is a zoomed-in analysis of the 100-base pattern shown in Figure 1. In Figure 2, we find that when the distance level is less than 20 base pairs, the AA plot shows that the 2 cancer cell lines and the normal (HMEC) sample are similar. After 20 base pairs, the normal AA state pair frequency (yellow line) dips below the 2 cancerous AA lines (gray for HCC1954 and brown for MCF7) (see the AA plot of Figure 2). We also find that the AD state pairs in normal and cancerous



**Figure 2.** Methylation state changes between consecutive CG sites based on 10 base intervals. HMEC indicates human mammary epithelial cell. The figure displays the percentage of methylation state changes with an initial methylation state of A, B, C, and D. The horizontal axis is the distance between consecutive CG sites. The vertical axis is the percentage of occurrence on a 0 to 1 scale. The lines of 3 colors are for HMEC (yellow), HCC1954 (gray), and MCF7 (brown), respectively.

data have similar percentage of occurrences when the distance sub-interval is less than roughly 10 base pairs (see the AD plot of Figure 2), but when the distance increases, there is a dramatic difference between the breast cancer lines (gray and brown) and the normal sample (yellow). The normal and cancerous frequencies for the DD state pair show that the 2 cancer cell lines (gray and brown) are very similar, and they are different from the normal sample (yellow), see the DD plot of Figure 2. In addition, when looking at the partial methylation state B row (ie, the second row of the Figure 2), the zoomed-in analyses of the BA and BD plots also show that cancer cell lines are very different from the normal sample. The

methylation state C row (ie, the third row of the Figure 2) does not show that obvious difference between cancerous and normal samples.

Figures 1 and 2 show that co-methylation patterns in normal and cancerous samples are different. However, it is unclear whether these differences are statistically significant. Therefore, we use the chi-square test to investigate these differences. Table 3 shows the count and percentage of all CG sites in chromosome 1 (not just CG sites selected based on specific intervals). We run the chi-square test on the count data (in Table 3) from our method 1 to compare cancerous and normal samples. In Table 4, we display the chi-square test performed on count data



**Table 6.** Co-methylation region length summary.

	STATE	MINIMUM	FIRST QUARTER	MEDIAN	MEAN	THIRD QUARTER	MAXIMUM
Normal data (HMEC)	A	2	42	116	223.6	288	5493
	B	2	22	61	132.7	161	2576
	C	2	24	63	133.2	162	3278
	D	2	55	189	394.3	499	13 700
Cancerous data (HCC1954)	A	2	62	206	503.4	573	18 750
	B	2	16	46	125.5	147	2805
	C	2	17	51	121.9	147	2013
	D	2	66	247	584.3	711	100 300
Cancerous data (MCF7)	A	2	67	207	394.3	504	11 780
	B	2	20	56	136.4	168	3347
	C	2	22	62	138.7	175	2473
	D	2	75	275	565.5	728	18 410

Abbreviation: HMEC, human mammary epithelial cell.

Shown are co-methylation region length summaries for normal and cancerous data separately.

with no distance restrictions, which is for all the CG sites in chromosome 1. The test is conducted for all 3 samples, and it shows significant results.

We use our method 1 analysis data to further investigate if there is also a significant difference between CG sites that are within a certain distance level. That is, we will conduct this test using the  $[0, 50)$ ,  $[50, 100)$ ,  $\dots$ ,  $[450, 500)$  50-base distance intervals. The results of these tests are that  $P$ -values in all intervals A, B, C, and D are extremely small in the chromosome 1 analysis ( $P$ -values are not shown here). Therefore, we conclude that the co-methylation patterns in the cancerous and normal samples are significantly different even for CG sites that are just 50 bases away. We additionally test our data using the chi-square test to see whether co-methylation patterns within much smaller 10-base distance-levels continue to show the significant difference between cancerous and normal samples. In Table 5, we show our test results on the 10-base intervals (ie,  $[0, 10)$ ,  $[10, 20)$ ,  $\dots$ ). These results show that the normal and cancerous data sets are statistically different.

### Results of method 2

As mentioned before, we will conduct the method 2 analysis. That is, we investigate the WS co-methylation patterns by studying the distribution of the co-methylation region length. In particular, we will study the WS co-methylation region length for 4 methylation states (A, B, C, and D) separately. We identify the co-methylation regions that have the same methylation state (eg, AAAAA or BBBBB), and then count the number of CG sites in each region and calculate the length of each region. For example, we may report a DDDD region of

length 100 base pairs, which represents a region that consists of 4 consecutive CG sites with methylation state “D” and its length is 100 bases. We identify all possible A, B, C, and D methylation-state regions. We also calculate the length and count of each region (“count” means the number of CG sites within a region). All method 2 analysis results are shown in Table 6. This table shows that the medians of cancer co-methylation region length is larger than the median of the normal sample by more than a 100 base pairs for both the A and D methylation states. In fact, for the methylation state A region, the median length is 116 bases for the normal HMEC sample, but it is 206 and 207 bases for 2 cancer cell lines HCC1954 and MCF7, respectively. For the methylation state D region, the median length is 189 base pairs for the normal HMEC sample, but it is 247 and 275 base pairs for the 2 cancer cell lines. This indicates that lengths of cancerous co-methylation A or D regions are consistently greater than the lengths of normal co-methylation regions, with the difference growing larger in higher quantiles (eg, 75% or third quarter). For the partial methylation state B and C regions, Table 6 shows that the partial methylation region length ranges from 46 to 63 base pairs (see the fifth or the median column of the B and C states). There is not an obvious difference for the partial methylation region length.

Table 6 also shows that the majority (about 75%) of co-methylation regions are at most 288 (for state A), 161 (for state B), 162 (for state C), and 499 (for state D) base pairs in the normal HMEC sample. For the breast cancer cell line HCC1954, the majority (about 75%) of co-methylation regions are at most 573 (for state A), 147 (for state B), 147 (for state C), and 711 (for state D) base pairs. For the breast cancer cell line

**Table 7.** Wilcoxon rank sum test results.

	A	B	C	D
Co-methylation region length				
P-values	0	1.42E-40	9.05E-69	0
Chi-square	3583.585	183.5097	313.3517	4959.549
Number of CG sites in each co-methylation region				
P-values	1.07E-157	5.80E-71	2.74E-21	0
Chi-square	722.87	323.4529	94.69338	6575.451

MCF7, the majority (about 75%) of co-methylation regions are at most 504 (for state A), 168 (for state B), 175 (for state C), and 728 (for state D) base pairs. That is, most of the co-methylation regions are very short. However, for both the normal and cancerous samples, a small proportion of the co-methylation regions are very long, which can be as long as several thousand base pairs (see the last column of the Table 6). For example, even the partial methylation regions can be 2000 to 3000 base pairs; for the no/low (state A) and high/full (state D) co-methylation regions, they can be as long as 5000 and even more than 10 000 base pairs.

Our method 2 analysis results in Table 6 show that co-methylation patterns, especially the length of co-methylation regions of methylation states A and D, are different in normal and cancerous samples. Next, we use the Wilcoxon rank sum test to investigate whether these differences are statistically significant. We conduct the test twice: first comparing cancer with normal using the co-methylation region length data and then comparing cancer with normal using the count data (see Table 7). Each count is the number of CG sites per region. Table 7 shows that there is a significant difference for each methylation-state region.

As for the co-methylation region length, our median result of methylation state A of the normal HMEC sample listed in Table 6 is 116. This result is close to the co-methylation region length reported in Guo et al,<sup>29</sup> which is an average of 95 base pairs and is calculated based on a correlation method. However, if we look at the co-methylation region length of the methylation state B, C, and D, they are not the same no matter if we use the median or the mean. This difference may be due to the reasons shown below.

First, Guo et al did the analysis by considering the no/low methylation (ie, the methylation state A) regions, 2 types of partial methylation (methylation states B and C) regions, and high/full methylation (ie, methylation state D) regions together. That is, their result of the “average 95 base pairs” is a “pooled” analysis of all types of co-methylation regions. However, our analysis considers these 4 types of regions separately. As shown in Table 6, co-methylation regions of different methylation states (A, B, C, and D) have different lengths. To obtain accurate analysis results, it is better to consider these 4 different methylation states separately.

Second, Guo et al calculated methylation haplotype blocks based on regions with at least 3 CG sites, but we calculate the length for co-methylation regions that have a minimum of 2 CG sites. Note that, in our original analysis, we calculate the co-methylation region length calculation for regions with at least 4 CG sites. Later, we change it to analyze co-methylation region of at least 2 CG sites. We make this change to be consistent with the co-methylation analysis we conducted in a recent publication.<sup>16</sup> Another reason for this change is that we find that there are a large proportion of co-methylation regions with just 2 CG sites; including them or not can affect the co-methylation region length summary. We will discuss this in detail below (see Table 8).

When we study co-methylation region length and the number of CG sites belonging to each region, we find that many co-methylation regions have only 2 or 3 CG sites (see Table 8). Therefore, when we calculate the co-methylation region length for each methylation state, the selection of a minimum number of CG sites can affect the results. For example, Table 8 shows that, for the co-methylation regions of the methylation state A, 46.3% of them consist of just 2 CG sites (ie, AA) for the HMEC sample. For the methylation states B, C, and D of the HMEC samples (ie, BB, CC, and DD), they are 83.4%, 80.5%, and 28.2% (see the third column of the HMEC sample in Table 8). It is important to include the co-methylation region with only 2 CG sites in the study. Therefore, we use the minimum of 2 CG sites in this study and in our recent publication.<sup>16</sup> Without including the 2-CG site co-methylation region, the length could be longer than the reported one.

## Discussion

We have conducted analyses to study WS co-methylation patterns by comparing 1 normal breast sample and 2 breast cancer cell lines. Our analysis results will allow researchers to better understand co-methylation, which can aid in future discovery and understanding of how co-methylation patterns may be related to the onset of cancer. However, our study has some limitations. The first limitation is that the current study is only based on 3 samples: 1 normal breast sample and 2 breast cancer cell lines. Therefore, the results of this study may not be generalized or applied to the population level due to the lack of replicates and samples. However, our results are still useful for the

**Table 8.** Summary of number of CG sites in different co-methylation regions.

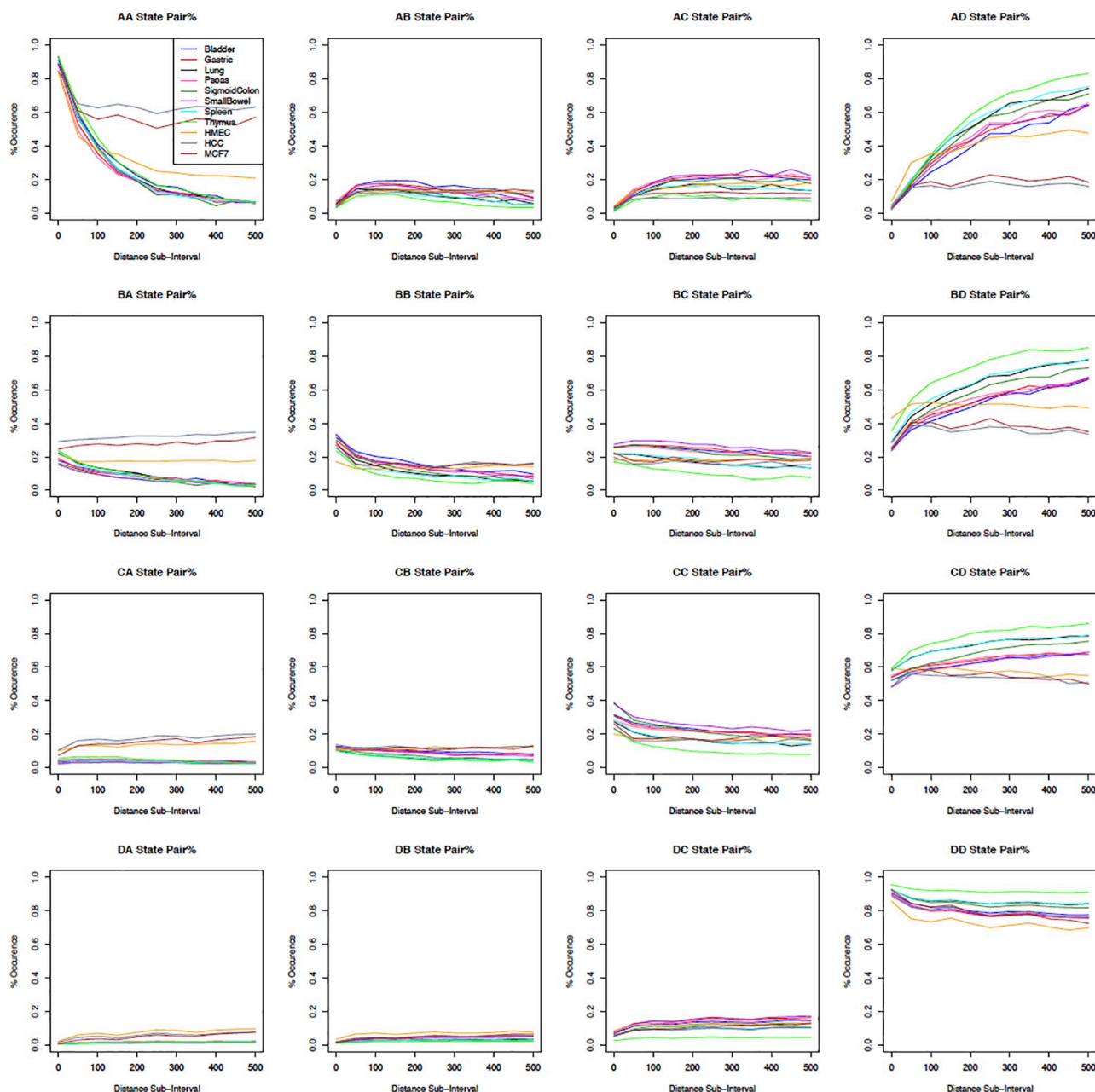
		2 CGS	3 CGS	4 CGS	5 CGS	6 CGS	>6 CGS
HMEC	A count	16 680	5964	2922	1680	1153	7620
	A%	46.3	16.6	8.1	4.7	3.2	21.2
	B count	13 147	2115	378	92	16	14
	B%	83.4	13.4	2.4	0.6	0.1	0.1
	C count	24 954	4767	968	225	55	32
	C%	80.5	15.4	3.1	0.7	0.2	0.1
HCC1954	D count	58 092	37 016	24 969	17 915	13 088	54 715
	D%	28.2	18	12.1	8.7	6.4	26.6
	A count	19 413	9316	5487	3551	2640	13 048
	A%	36.3	17.4	10.3	6.6	4.9	24.4
	B count	15 005	3301	844	255	90	60
	B%	76.7	16.9	4.3	1.3	0.5	0.3
MCF7	C count	20 892	4354	1045	263	99	65
	C%	78.2	16.3	3.9	1	0.4	0.2
	D count	36 875	23 435	16 334	12 360	9402	54 911
	D%	24.1	15.3	10.7	8.1	6.1	35.8
	A count	16 173	7922	4562	2799	1888	7995
	A%	39.1	19.2	11	6.8	4.6	19.3
MCF7	B count	11 204	2437	619	211	68	92
	B%	76.6	16.7	4.2	1.4	0.5	0.6
	C count	20 012	4283	1032	290	92	87
	C%	77.6	16.6	4	1.1	0.4	0.3
	D count	34 363	21 931	15 550	11 632	9066	56 394
	D%	23.1	14.7	10.4	7.8	6.1	37.9

Abbreviation: HMEC, human mammary epithelial cell.

Rows 2 to 9 are for the normal sample HMEC. Rows 10 to 17 are for the cancer cell line HCC1954. Rows 18 to 25 are for the cancer cell line MCF7. For each sample, the top row is number of CG sites in each type of co-methylation region, ie, 2 CG sites, 3 CG sites, and so on. For each sample, "A count" and "A%" are the total number and percentage of co-methylation regions of methylation states "A." For example, for the HMEC sample, in the "2 CGs" column, "A count" is 16 680, and "A%" is 46.3. These 2 numbers mean that, among all the AA...A type co-methylation region, 16 680, ie, 46.3%, of them have only 2 CGs.

reasons and benefits listed below. Although analysis results on WS co-methylation over distance are reported,<sup>11,14,26-28</sup> they are disputable.<sup>12,15</sup> In addition, WS co-methylation is not well studied for breast tissues yet. In particular, previous studies do not conduct WS co-methylation analysis by considering different methylation states (no/low, partial, and high/full methylation) separately. Therefore, it is necessary and important to conduct this preliminary analysis, which allows for the initial focus to be on a specific sample before working with multiple samples. This preliminary analysis can provide insights for us to build more accurate models to identify WS co-methylation patterns of multiple samples. Our analysis results can also provide helpful insight and information for methylation data

analyses based on hidden Markov models.<sup>35-41</sup> However, the distribution of co-methylation region lengths is often unknown. Our WS co-methylation analysis results can provide useful information on this aspect. Meanwhile, we do plan to study the WS co-methylation patterns using multiple samples in each of the 2 groups (normal vs cancerous) in the near future. In fact, we recently published a paper on the analysis of co-methylation patterns of multiple normal samples/tissues.<sup>16</sup> The second limitation is that WS co-methylation patterns may also be related to genomic context.<sup>11-13,30</sup> To simplify our analysis, we choose not to consider this relationship when comparing cancerous data with normal data. We plan to consider the genomic context in another co-methylation project.



**Figure 3.** Comparison of co-methylation patterns of 11 samples. HMEC indicates human mammary epithelial cell.

In each plot, the 11 lines represent the 11 samples/tissues (HMEC, HCC1954, MCF7, and the 8 tissues of the STL0001). Each plot is for a state pair, ie, AA, AB, AC, AD, BA, BB, and so on.

The main reason for the first limitation mentioned above is that we could not find additional publicly available WGBS data of normal or cancerous breast samples. For example, although there are 12 359 methylation data generated and posted on The Cancer Genome Atlas Program (TCGA) website, they are all Illumina array data: 9756 Illumina 450K and 2603 Illumina 27K array data sets. When we search the GEO data sets using “breast bisulfite WGBS,” we can only find 6 items, which are mainly the data used in this study. However, to address this limitation, we have conducted some further analysis by showing the co-methylation patterns of the 3 breast samples with the other 8 different normal tissues (see Figure 3). This figure is similar to Figure 1, to which the additional 8

different tissues are added. These 8 tissues are bladder, gastric, lung, psoas, sigmoid colon, small bowel, spleen, and thymus of the sample STL0001.<sup>42</sup> The co-methylation patterns of these 8 tissues are reported in our recent publication.<sup>16</sup> We conduct this further analysis as an indirect way to show our analysis results. In Figure 3, the overall patterns of the 11 samples have some similarities and differences as well. The AA and AD plots show that normal breast sample HMEC (yellow line) has a more similar pattern with the other 8 normal tissues than with the 2 cancer cell lines (gray and brown lines). In the BA, CA, BD, and CD plots, we find that the breast samples are different from the other 8 tissues of the STL0001. This difference may be due to some sample or sequencing differences.



As for the coverage level in our analysis, we use all CG sites with at least  $3\times$  coverage. If the data sets that we use have very low coverages, eg, with an average of  $3\times$  or  $4\times$  coverage, then using a minimum of read depth of  $3\times$  will not be meaningful, and the results will not be reliable. However, the coverage of our 3 data sets are 27-fold (HMEC), 20-fold (HCC1954), and 36-fold (MCF7). Using the cutoff value of  $3\times$  coverage will not affect our analysis results much. In fact, we have conducted our analysis for 2 different coverage levels:  $\geq 3\times$  and  $\geq 6\times$ . The analysis results regarding co-methylation patterns are almost the same except that the lengths of co-methylation regions of the  $3\times$  and  $6\times$  levels are different. This difference is related to part of our method 2 conclusion. That is, the co-methylation region length of  $\geq 6\times$  data is shorter than the co-methylation region length  $\geq 3\times$  data. This is expected because less CG sites are selected when using  $\geq 6\times$  data. As for the method 1 conclusion, they are the same. We have this consistency because our method 1 conclusion is based on “percentage,” ie, how frequently 1 methylation state changes to another state. When the coverage level is increased, the “count” may become smaller, but the overall percentage is the same or similar in a whole chromosome.

In our WS co-methylation analysis, we consider CG sites with different methylation states/levels (A, B, C, and D) separately. That is, we group CG sites based on their methylation levels first before any further analysis. This type of method can be called a binning approach. For both WS and BS co-methylation analyses, the Pearson correlation-based method can be used when the sample size is relatively large. Next, we explain the benefits/advantages and downsides/disadvantages of using the binning approach against the Pearson correlation method on modeling data. First, when there are a large number of samples in each group of cancerous or normal samples, the Pearson correlation method can be used. For example, using the correlation-based method, Guo et al<sup>29</sup> studied WS correlation and Mallona et al<sup>15</sup> investigated BS co-methylation. When analyzing a small number of samples in each group (cancerous or normal), the binning approach can be used, but the Pearson correlation method cannot be used because not enough data can be used to calculate the correlation. The binning approach of studying WS co-methylation can provide helpful input for methylation analyses based on hidden Markov models.<sup>37-41</sup> Second, the binning approach can help to remove the impact of noise because DNA sequencing data can be very noisy; the Pearson correlation estimates can be easily affected by outliers. Third, the binning approach can help to investigate WS co-methylation patterns, especially co-methylation region of different methylation states, ie, A, B, C, and D. We have shown that the co-methylation patterns of these 4 different methylation states are different (see Table 6). Therefore, our method is useful and important for studying the co-methylation patterns accurately and thoroughly. Furthermore, it is better to study the co-methylation patterns for partial methylation sites separately

because changes in partially methylated domains are hallmarks of cell differentiation.<sup>43</sup>

## Conclusions

In this article, we conduct analyses to study breast tissue WS co-methylation patterns using WGBS data. We analyze normal and cancerous breast tissue methylation data in an attempt to determine whether normal and cancerous breast samples have significantly different WS co-methylation patterns. To do so, we assign each CG site a methylation state (A, B, C, and D) based on methylation signal levels. We find that WS co-methylation patterns change even within a short 50-base distance. We also show that the co-methylation patterns of 4 methylation levels/states (A, B, C, D) are different both within a breast sample and between different samples (normal vs cancerous breast samples). Using our methods, not only do we show that co-methylation patterns of normal and cancerous breast samples are significantly different, but also pinpoint the specific range of distances between CG sites, in which different co-methylation patterns occur. We also show that the co-methylation lengths of 4 different methylation states are different. To the best of our knowledge, this study is the first one that conducts a “zoomed-in” analysis for breast tissue co-methylation patterns by considering different methylation states (A, B, C, and D) separately. Our research may provide a deep understanding of co-methylation patterns. The presence of these specific co-methylation patterns may help tumor biologists more easily locate genes associated with cancer, which may contribute to more efficient and effective cancer diagnoses.

## Acknowledgements

This project was completed with the use of Texas State University facilities and resources and was made possible by the Mathworks summer program. The authors are grateful for the 2 reviewers, whose questions and comments help us improve this manuscript greatly.

## Author Contributions


SS initiated the project, suggested all key original ideas, and oversaw the whole process. All authors contributed to early coding. LS conducted the main analysis and interpretation of data with contribution from SN, who conducted the method 2 analysis of data. LS conducted all the analyses when we revised the manuscript to address all reviewers' questions. EC and SN drafted early portions of the manuscript. LS and SN drafted latter portions of the manuscript. SS did all the revision before the resubmission. SS gave suggestions over the course of the project and extensively reviewed and revised the final paper. All authors contributed expertise and edits. All authors have read and approved the final manuscript.

## Availability of Data and Material

Data sets used in this paper are publicly available (GSE29127 and GSM3526804). R code files are available on request.



## ORCID iD

Shuying Sun  <https://orcid.org/0000-0003-3974-6996>

## REFERENCES

- Siegel R, Ma J, Zou Z, Jemal A. Cancer statistics, 2014. *CA Cancer J Clin.* 2014;64:9-29.
- Houssami N, Given-Wilson R, Ciatto S. Early detection of breast cancer: overview of the evidence on computer-aided detection in mammography screening. *J Med Imaging Radiat Oncol.* 2009;53:171-176.
- Lim DH, Maher E. DNA methylation: a form of epigenetic control of gene expression. *Obstetric Gynaecol.* 2010;12:6.
- Chatterjee R, Vinson C. CpG methylation recruits sequence specific transcription factors essential for tissue specific gene expression. *Biochim Biophys Acta.* 2012;1819:763-770.
- Ehrlich M, Lacey M. DNA methylation and differentiation: silencing, upregulation and modulation of gene expression. *Epigenomics.* 2013;5:553-568.
- Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet.* 2012;13:484-492.
- Jones PA, Baylin SB. The epigenomics of cancer. *Cell.* 2007;128:683-692.
- Yang X, Yan L, Davidson NE. DNA methylation in breast cancer. *Endocr Relat Cancer.* 2001;8:115-127.
- Griffiths EA, Gore SD. DNA methyltransferase and histone deacetylase inhibitors in the treatment of myelodysplastic syndromes. *Semin Hematol.* 2008;45:23-30.
- Barrera V, Peinado MA. Evaluation of single CpG sites as proxies of CpG island methylation states at the genome scale. *Nucleic Acids Res.* 2012;40:11490-11498.
- Eckhardt F, Lewin J, Cortese R, et al. DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet.* 2006;38:1378-1385.
- Hickey PF. *The Statistical Analysis of High-Throughput Assays for Studying DNA Methylation* [doctoral thesis]. Melbourne, VIC, Australia: The University of Melbourne; 2015.
- Li YR, Zhu JD, Tian G, et al. The DNA methylome of human peripheral blood mononuclear cells. *PLoS Biol.* 2010;8:e1000533.
- Lister R, Ecker JR. Finding the fifth base: genome-wide sequencing of cytosine methylation. *Genome Res.* 2009;19:959-966.
- Mallona I, Aussó S, Díez-Villanueva A, Moreno V, Peinado MA. Modular dynamics of DNA co-methylation networks exposes the functional organization of colon cancer cells' genome. <https://www.biorxiv.org/content/biorxiv/early/2018/09/27/428730.full.pdf>. Update 2018.
- Sun L, Sun S. Within-sample co-methylation patterns in normal tissues. *Biodata Min.* 2019;12:9.
- Akulenko R, Helms V. DNA co-methylation analysis suggests novel functional associations between gene pairs in breast cancer samples. *Hum Mol Genet.* 2013;22:3016-3022.
- Busch R, Qiu W, Lasky-Su J, Morrow J, Criner G, DeMeo D. Differential DNA methylation marks and gene comethylation of COPD in African-Americans with COPD exacerbations. *Respir Res.* 2016;17:143.
- Martin TC, Yet I, Tsai PC, Bell JT. coMET: visualisation of regional epigenome-wide association scan results and DNA co-methylation patterns. *BMC Bioinformatics.* 2015;16:131.
- Wang F, Xu H, Zhao H, Gelernter J, Zhang H. DNA co-methylation modules in postmortem prefrontal cortex tissues of European Australians with alcohol use disorders. *Sci Rep.* 2016;6:19430.
- Zhang J, Huang K. Pan-cancer analysis of frequent DNA co-methylation patterns reveals consistent epigenetic landscape changes in multiple cancers. *BMC Genomics.* 2017;18:1045.
- Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics.* 2008;9:559.
- Horvath S, Zhang Y, Langfelder P, et al. Aging effects on DNA methylation modules in human brain and blood tissue. *Genome Biol.* 2012;13:R97.
- Rickabaugh TM, Baxter RM, Sehl M, et al. Acceleration of age-associated methylation patterns in HIV-1-infected adults. *PLoS ONE.* 2015;10:e0119201.
- Van Eijk KR, de Jong S, Boks MP, et al. Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects. *BMC Genomics.* 2012;13:636.
- Cokus SJ, Feng S, Zhang X, et al. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature.* 2008;452:215-219.
- Lacey MR, Ehrlich M. Modeling dependence in methylation patterns with application to ovarian carcinomas. *Stat Appl Genet Mol Biol.* 2009;8:e40.
- Landan G, Cohen NM, Mukamel Z, et al. Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. *Nat Genet.* 2012;44:1207-1214.
- Guo S, Diep D, Plongthongkum N, Fung HL, Zhang K, Zhang K. Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA. *Nat Genet.* 2017;49:635-642.
- Lister R, Pelizzola M, Dowen RH, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature.* 2009;462:315-322.
- Libertini E, Heath SC, Hamoudi RA, et al. Information recovery from low coverage whole-genome bisulfite sequencing. *Nat Commun.* 2016;7:11306.
- Hon GC, Hawkins RD, Caballero OL, et al. Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. *Genome Res.* 2012;22:246-258.
- Du Q, Bert SA, Armstrong NJ, et al. Replication timing and epigenome remodeling are associated with the nature of chromosomal rearrangements in cancer. *Nat Commun.* 2019;10:416.
- Harris EY, Ponts N, Le Roch KG, Lonardi S. BRAT-BW: efficient and accurate mapping of bisulfite-treated reads. *Bioinformatics.* 2012;28:1795-1796.
- Shafi A, Mitrea C, Nguyen T, Draghici S. A survey of the approaches for identifying differential methylation using bisulfite sequencing data. *Brief Bioinform.* 2017;19:737-753.
- Yu X, Sun S. Comparing five statistical methods of differential methylation identification using bisulfite sequencing data. *Stat Appl Genet Mol Biol.* 2016;15:173-191.
- Ernst J, Kellis M. Chromatin-state discovery and genome annotation with ChromHMM. *Nat Protoc.* 2017;12:2478-2492.
- Saito Y, Mituyama T. Detection of differentially methylated regions from bisulfite-seq data by hidden Markov models incorporating genome-wide methylation level distributions. *BMC Genomics.* 2015;16:S3.
- Shokoohi F, Stephens DA, Bourque G, Pastinen T, Greenwood CMT, Labbe A. A hidden Markov model for identifying differentially methylated sites in bisulfite sequencing data. *Biometrics.* 2019;75:210-221.
- Sun S, Yu X. HMM-Fisher: identifying differential methylation using a hidden Markov model and Fisher's exact test. *Stat Appl Genet Mol Biol.* 2016;15:55-67.
- Yu X, Sun S. HMM-DM: identifying differentially methylated regions using a hidden Markov model. *Stat Appl Genet Mol Biol.* 2016;15:69-81.
- NIH common fund epigenomics program. <http://commonfund.nih.gov/epigenomics>.
- Salhab A, Nordstrom K, Gasparoni G, et al. A comprehensive analysis of 195 DNA methylomes reveals shared and cell-specific features of partially methylated domains. *Genome Biol.* 2018;19:150.