Hemimethylation Patterns in Breast Cancer Cell Lines

Jazmine Castanon and Noah Ledbetter

Abstract:

DNA methylation is an epigenetic event that occurs in both cancerous and normal cell development. Hemimethylation occurs when an extra methyl group is added to a cytosine site. Analyzing CpG site hemimethylation could potentially reveal patterns within the genome that are undergoing tumor growth and deactivation of tumor suppressors. Using publicly available data on hg19, we identify hemimethylated CpG sites in the breast cancer cell lines using Wilcoxon signed rank tests on Chromosome 22 and X with the hope of detecting hemimethylation. Our results reveal less than 6% of our data among CHR22 and CHRX showed some level of significance at an alpha level of 0.05 and absolute value mean difference between the forward and reverse strands above 0.4, 0.6, and 0.8. A sliding window approach provided evidence to suggest that most distances between HM CpG sites is below 50 bp, while a clustering approach also revealed a cluster length of less than 50 bp. All 23 chromosomes were used in a novel cluster approach and helped create a cutoff of 0.13 for sites in clusters, versus the cutoff of 0.05 for singleton sites.

Introduction:

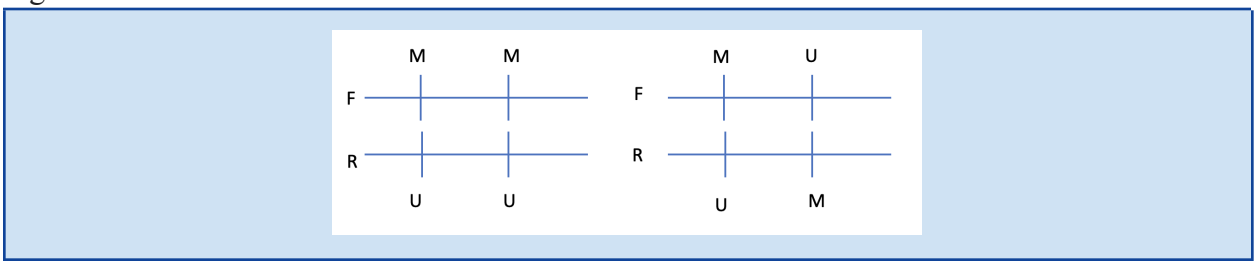
Breast cancer is the most commonly diagnosed cancer among American women. In fact, in 2021, it is expected that 43,600 women in the U.S. will die from breast cancer. Much research is currently geared towards early detection. Many women are encouraged to get checked during their yearly pap appointments, schedule mammograms when appropriate, and perform self examinations for early breast cancer detection. Currently, most women are also warned about harmful variants of BRCA1 and BRCA2 on chromosome 13 and 17 that have shown to be associated with a higher likelihood of developing breast and ovarian cancer.

Epigenetic markers and their ties to breast cancer have been a popular field of study in the last twenty years. DNA methylation, a significant epigenetic mechanism, has become a key player in the search for these epigenetic markers. DNA methylation occurs when a methyl group (CH3) is added to the fifth carbon of a cytosine-guanine dinucleotide (CG or CpG site) on a DNA molecule in a mammalian cell (1). Previous studies have suggested that hemimethylated CpG

sites are intermediaries during carcinogenesis, and something that should be studied further to help researchers trace cancer cell lines (2). If scientists can find which locations of hemimethylated CpG sites are unique to cancer, the search will shift to target these sites and reverse the effects of methylation.

The goal of our research is to use publicly available bisulfite-sequencing data to study hemimethylation in 7 breast cancer cell lines across the entire genome. Our Wilcoxon signed rank tests were performed on Chromosome 22 and X. We used publicly available data for reduced representation bisulfite-sequencing (RRBS) data of 7 breast cancer cell lines (BT20, BT474, MCF7, MDAMB231, MDAMB468, T47D, and ZR751). Other statistical analysis using R software revealed some patterns about the distance between HM CpG sites and the distance of HM CpG clusters.

Figure 1.



Methodology:

Data was obtained for each chromosome containing the chromosome name, CpG site index, position in the chromosome, and methylation signal on the forward and reverse strand for seven breast cancer cell lines. The data was then cleaned, stripping out CpG sites that contained less than three methylation signals for either the forward or reverse strand for the various cell lines. Note that although data was removed from the data set, the CpG site indices were kept, as well as the positions on the genome for each CpG site. This information would aid us in our approach when trying to group CpG Sites based on adjacency and when applying our sliding window approach on adjacent sites. The resulting sizes can be seen in **Table 1.**

Table 1

	*Length	CPG Sites	Density	*Reduced Length (excluding sites with 4 or more NAs)
Chromosome 22	**50,818,468**	**578,097**	**Length/CPG ≈ 87.907**	**20218**
Chromosome X	**156,040,895**	**1,246,401**	**Length/CPG ≈ 125.193**	**14875**

Table 1. Length of Chromosome 22 and Chromosome X, as well as the number of CpG sites both before and after cleaning.

For each chromosome, a Wilcoxon signed rank test was performed between the methylation signal of the seven cell lines on the forward and reverse strands. A Wilcoxon signed rank test is a non-parametric paired test that can be used over the paired student T test when the data analyzed is not normally distributed and the sample size is small. A CpG site with a calculated p-value less than 0.05 has a statistically significant difference of methylation signals between the forward and reverse strand, which is evidence indicating the presence of hemimethylation.

Once the Wilcoxon signed rank test was performed and a p-value was calculated for each site in our cleaned data set, a mean difference was calculated by comparing methylation signals on each strand. The mean difference calculation can help supplement our identification of hemimethylated sites, since just because there is a statistically significant difference between the forward and reverse strands, it doesn't mean that there is a biological difference. The mean difference helps capture and quantify the biological difference. The resulting data set size can be seen below in **Table 2** for various filtering criteria. For example, when looking at CpG sites in chromosome X, when filtering for p-values less than 0.05 and mean difference greater than 0.4, there are 384 significant sites.

Table 2

	CHR X		CHR 22	
	No p-value filter	**p-value < 0.05**	**No p-value filter**	**p-value < 0.05**
|Mean difference| ≥ 0	14875 - 100%	640 - 4.3%	20218 - 100%	1190 - 5.9%
|Mean difference| ≥ 0.4	5666 - 3.8%	384 -2.6%	8661 - 4.3%	711 - 3.5%
|Mean difference| ≥ 0.6	4407 - 3%	319 -2.2%	7113 - 3.5%	612 - 3%
|Mean difference| ≥ 0.8	2873 - 1.9%	226 - 1.5%	5127 - 2.5%	471 - 2.3%

Table 2. Size of the data set with various filtering criteria applied. Percentage for each cell is calculated by cell value divided by the cleaned data size.

Results:

Previous research has suggested that although Hemimethylation occurs within CpG clusters and singletons, it is significantly more likely to occur in clusters. Clusters that are hemimethylated are also signs of particular methylation events, and can possibly have a larger impact on gene function (2). These findings lead to more interest in a sliding window approach across Chromosomes 22 and X. Our sliding window approach rolled down our dataset of methylated CpG sites and looked at consecutive sites with a window of size 2. Similar sliding window approaches were used in a 2017 research project looking to decode methylation patterns in the honeybee genome. Welsh, Maleszka, and Foret had a slightly more complex operation applied to their 1200 bp windows (typically including 8 methylated CpG sites) as they performed a paired t-test on the methylation levels of each CpG site on each strand(3).

Our sliding windows were of size 2 (consecutive CpG sites) and consisted of two main calculations: a rolling mean on the absolute mean difference in methylation signals between two

sites and the rolling mean p-value between two sites. A forward and reverse rolling function was calculated for both types of averages to ensure all possible groups of consecutive sites/pairings occured. Four new rows are added to our data set for Chromosome 22 and X and named as follows: “Forward Average P Value, “Reverse Average P Value,” “Forward Average Mean Difference,” and “Reverse Average Mean Difference.”

Significant CpG clusters and singletons were extracted and a visualization displaying the difference between CpG sites in base pairs(units) was created. The results are shown in Figure 1.

Figure 1.

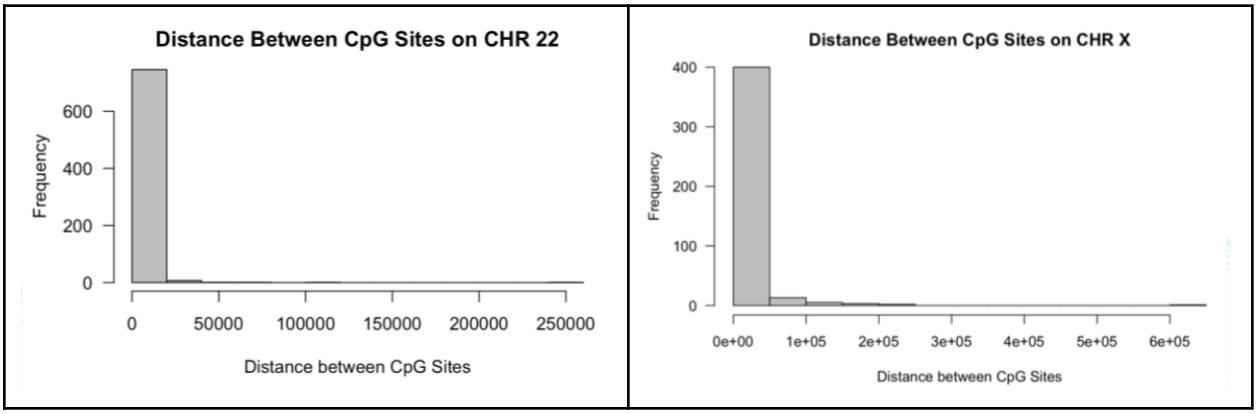


Figure 1. Since our main interest was in clusters or singletons that had a significant p value and higher mean difference, a filter was applied to extract significant sites, with the goal of mapping the length between these significant CpG sites. A filter was applied to singletons to extract those with a p value below 0.05 and a mean difference above 0.40. Another filter was applied to adjacent CpG sites to extract those that had a forward sliding average p value below 0.05 and a forward sliding average mean difference above 0.40 or adjacent CpG sites that had a reverse sliding average p value below 0.05 and a reverse sliding average mean difference above 0.40. **Figure 1** above shows a significant peak and at lower values indicating the distance between hemimethylated CpG sites is fairly small.

Figure 2.

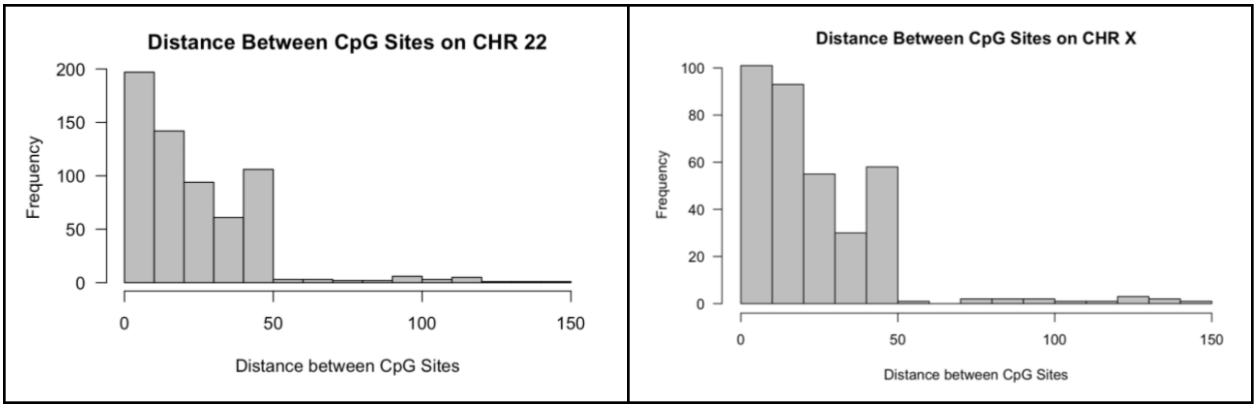


Figure 2. The histogram shows a steep decrease in frequency of CpG sites around 50 base pairs on both chromosomes. This suggests there may be higher occurrences of hemimethylated CpG sites with distances of 50 base pairs or less between adjacent hemimethylated sites.

Much of our findings suggest that hemimethylated CpG sites in our cancer cell lines in chromosome 22 and X have a distance of about 50 base pairs or less. To further explore such patterns, we readjusted our subset to alter the clusters and singletons we were selecting. The p value, forward average p value, and reverse average p value all kept a cutoff of .05, but the cutoff for average mean difference, forward average mean difference and reverse average mean difference changed from 0.4, to 0.6, to 0.8. **Table 3** shows there is a significant increase in sites included from the distance of 25 or less base pairs to 50 base pairs on both chromosomes. There is also a significant drop off after 50 base pairs.

Table 3

a) CpG sites selected based on Mean Difference of 0.4 or higher							
CHR	**d ≤ 5**	**d ≤ 10**	**d ≤ 15**	**d ≤25**	**d ≤ 50**	**d ≤ 100**	**No filter on Position distance(d)**
X	41	101	154	223	337	344	424
22	75	197	282	395	600	616	757
(b) CpG sites selected based on Mean Difference of 0.6 or higher							
CHR	**d ≤ 5**	**d ≤ 10**	**d ≤15**	**d ≤ 25**	**d ≤ 50**	**d ≤ 100**	**No filter on Position distance(d)**
X	35	83	130	182	275	281	347
22	65	170	238	337	513	526	641
(b) CpG sites selected based on Mean Difference of 0.8 or higher							
CHR	**d ≤ 5**	**d ≤ 10**	**d ≤ 15**	**d ≤ 25**	**d ≤ 50**	**d ≤100**	**No filter on Position distance(d)**
X	25	62	93	133	203	206	247
22	51	131	187	256	395	405	492

Table 3. It is important to note that there are more CpG Sites captured in Table 3 than in Table 2 because our criteria changes to include not only significant singleton sites with a p value less than 0.05 but to also include HM sites adjacent to other HM sites with forward and reverse Average p values less than 0.05 and forward and reverse average mean differences greater than 0.4, 0.6, and 0.8.

Additionally, we explored a novel approach to clusters where, prior to filtering by mean difference and p-value, each site with sufficient data was grouped with all of their immediately adjacent sites. The main difference between this clustering approach and the sliding window approach was that all adjacent CpG sites were considered part of a cluster, rather than just adjacent pairs, as is the case in the sliding window approach. Histograms were created for each chromosome of the CpG cluster length as can be seen in **Figures 3 and 4.**

Figure 3.

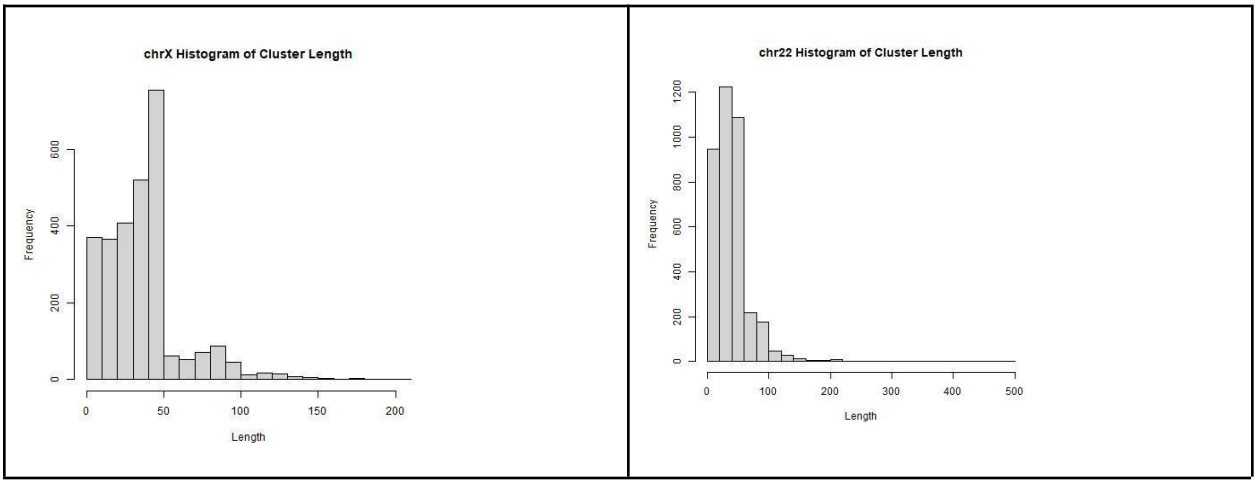


Figure 3. The histograms are both right skewed, with a sharp decline before a length of 100 base pairs.

Figure 4.

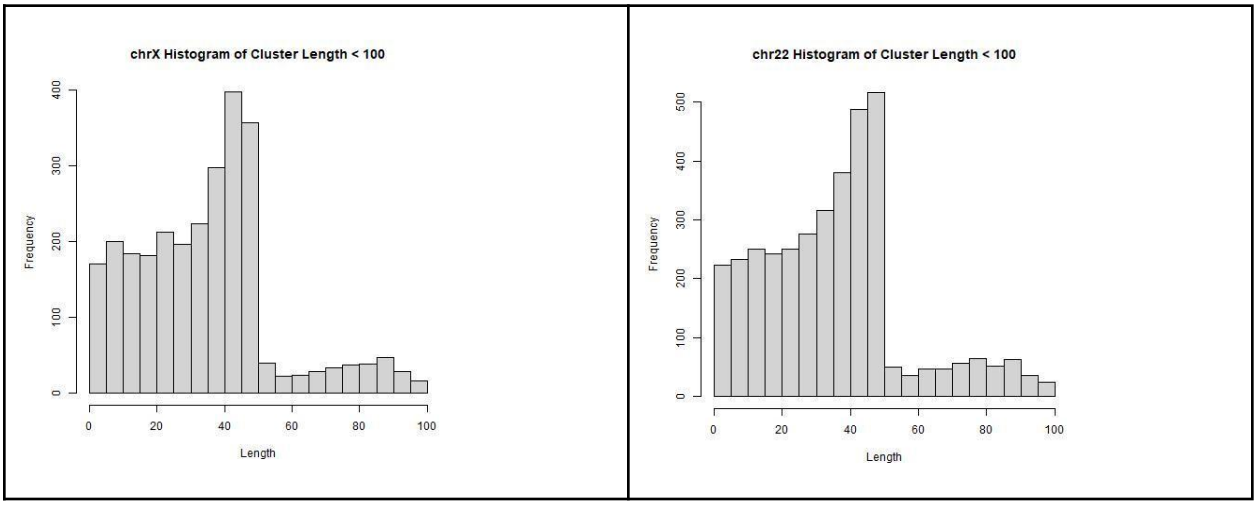


Figure 4. The histogram shows a steep decrease in frequency of CpG clusters at 50 base pairs in length.

Looking only at CpG clusters with a size less than 100, we can immediately see a sharp drop off at 50, with the majority of clusters containing less than 50 base pairs. This agrees with the sliding window approach, and previous results found by other investigators (4).

The primary motivation for clustering is to incorporate the additional information that CpG sites tend to occur in clusters into our results. Our goal was to decrease the requirement for significance for sites that occur in clusters. In pursuit of this, we examined the size of our data

sets under different filtering criteria for those sites which occur in clusters. Singletons with a calculated p-value less than 0.05 and mean difference greater than 0.4 were still included, but CpG sites in clusters were filtered out according to a reduced criteria. The results can be seen on the left of figure 5. We found a large leap when sites not included in clusters were filtered with a p-value less than 0.05 and mean difference greater than 0.4, and sites included in clusters had the reduced criteria of p-value less than .13 and mean difference greater than 0.4. Expanding this to the entire genome, the cut off continues to hold true as seen on the right of Figure 5 .

Using this cutoff for CpG sites in groups captures the previous research that CpG sites tend to occur in clusters, while not unnecessarily reducing the number of significant sites. Filtering based on this approach yielded the histograms found in Figure 6 This aligns with the previous conclusion from unfiltered data that CpG clusters tend to be less than 50 base pairs in length. This visualization was performed on every chromosome and the results can be found in SUPPLEMENTAL TABLE 1 and 2. The trend holds true across chromosomes.

Figure 5.

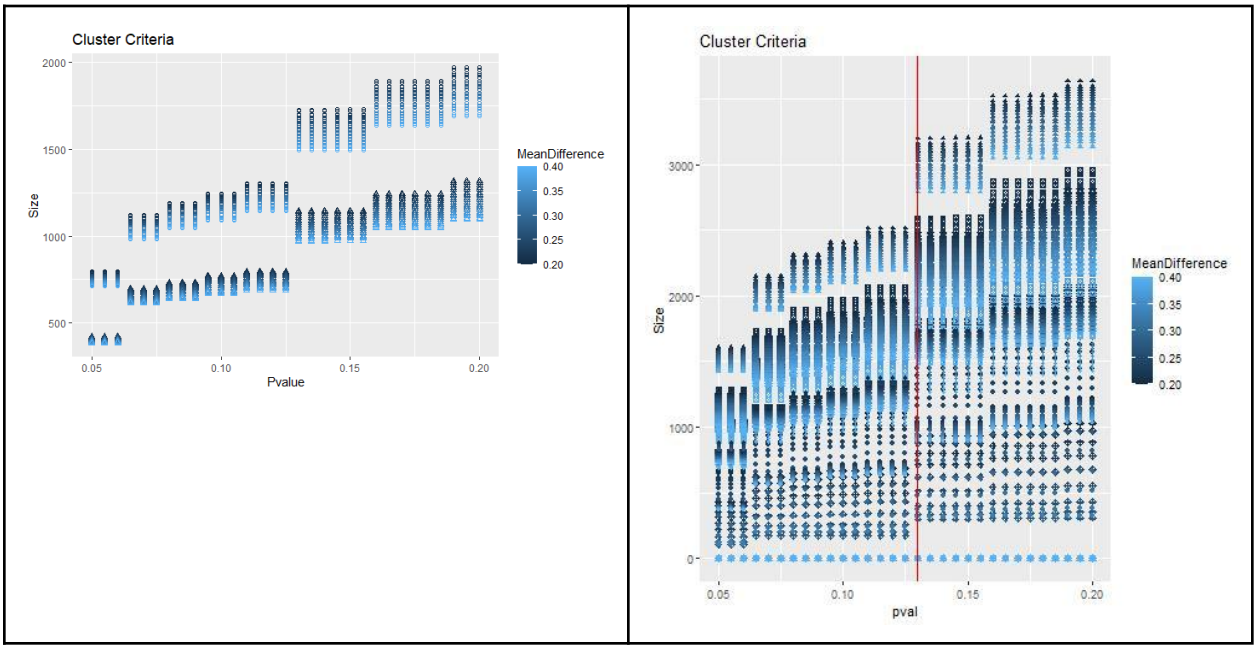


Figure 5. Left: Scatter plot of dataset size when p-value and mean difference cut offs vary for grouped CpG sites for chromosome 22 (circles) and chromosome X (triangles). Right: Scatter plot of dataset size when p-value and mean difference cut offs vary for grouped CpG sites for each chromosome

Figure 6.

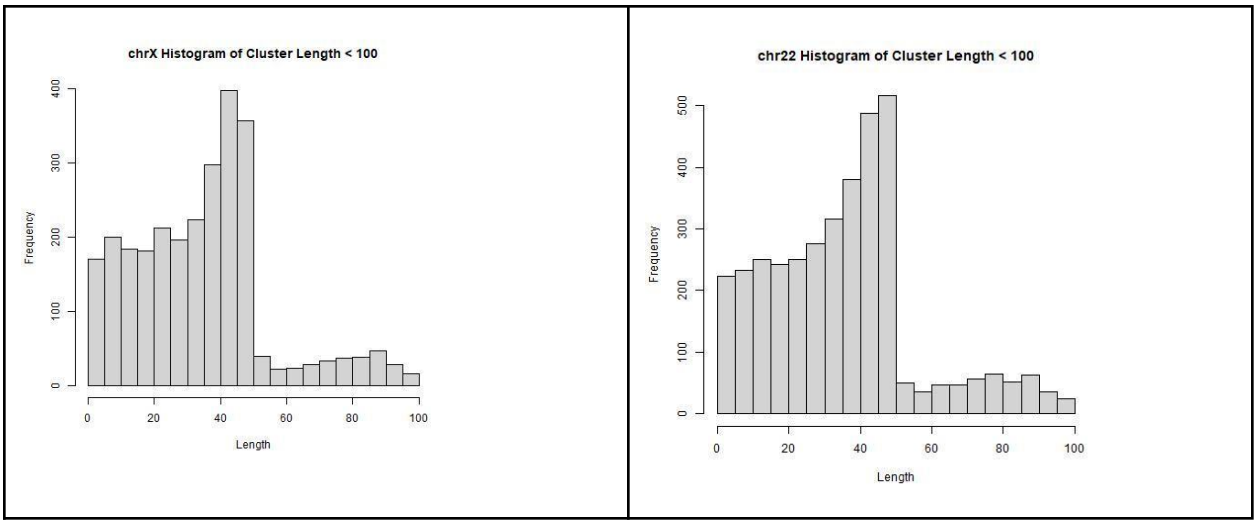


Figure 6. The histograms show a steep decrease in frequency of CpG clusters at 50 base pairs in length while non-significant CpG sites are filtered out.

Conclusion:

In this paper, we have conducted Wilcoxon tests to identify HM in 7 breast cancer cell lines. Through our analysis of CpG clusters that were hemimethylated, we found that the typical cluster length is less than 50 bp. When not just looking for significant singleton sites across the genome, one can lower the threshold of significance to a cutoff p value of 0.13 for clusters of CpG sites. This adjustment in cutoffs can help us include more data in our analysis. When using a non-clustered approach, a sliding window shows adjacent CpG sites typically have less than 50 bp lengths between them. Thorough analysis of hemimethylation across the genome is essential to help map the footprints of hemimethylation that are unique to breast cancer. Some genes containing HM CpG sites may even be linked to tumor growth and suppression(1). Knowing where to target reversal of methylation may help with treatment and early diagnosis of breast cancer.

There continues to be hopeful progress in the field of epigenetics as it relates to cancer detection. DNA Methylation Inhibitors 5-azacytidine (5-aza) and 5-aza-deoxycytidine have been studied and researched as a way to reverse the effects of methylation in Pediatric T-Cell Acute Lymphoblastic Leukemia(5)(6). There are also other studies as recent as 2021 that have shown an effort to isolate hypomethylation unique to breast cancer (5). Continued study of HM in breast cancer cells and normal cells can serve as a method to help detect, treat, and increase survival rates among women diagnosed with breast cancer.

References

1. Sun, S., Lee, Y. R., & Enfield, B. (2019). Hemimethylation patterns in breast cancer cell lines. *Cancer Informatics*, *18*, 1. https://doi.org/10.1177/1176935119872959

2. Ehrlich M, Lacey M. DNA hypomethylation and hemimethylation in cancer. *Adv Exp Med Biol*. 2013;754:31-56.https://pubmed.ncbi.nlm.nih.gov/22956495/

3. Welsh, L., Maleszka, R., & Foret, S. (2017). Detecting rare asymmetrically methylated cytosines and decoding methylation patterns in the Honeybee genome. *Royal Society Open Science*, *4*(9), 170248. https://doi.org/10.1098/rsos.170248

4. Sun, S., Lee, Y. R., & Enfield, B. (2019). Hemimethylation patterns in breast cancer cell lines. *Cancer Informatics*, *18*, 5. https://doi.org/10.1177/1176935119872959

5. Christman, J. 5-Azacytidine and 5-aza-2′-deoxycytidine as inhibitors of DNA methylation: mechanistic studies and their implications for cancer therapy. *Oncogene* 21, 5483–5495 (2002). https://doi.org/10.1038/sj.onc.1205699

6. Bensberg, M., Rundquist, O., Selimović, A., Lagerwall, C., Benson, M., Gustafsson, M., Vogt, H., Lentini, A., & Nestor, C. E. (2021, August 24). *TET2 as a tumor suppressor and therapeutic target in T-cell acute lymphoblastic leukemia*. PNAS. Retrieved December 4, 2021, from https://www.pnas.org/content/118/34/e2110758118.