Hemimethylation in Breast Cancer Cell Lines

Jazmine Castanon and Noah Ledbetter
November 17, 2021

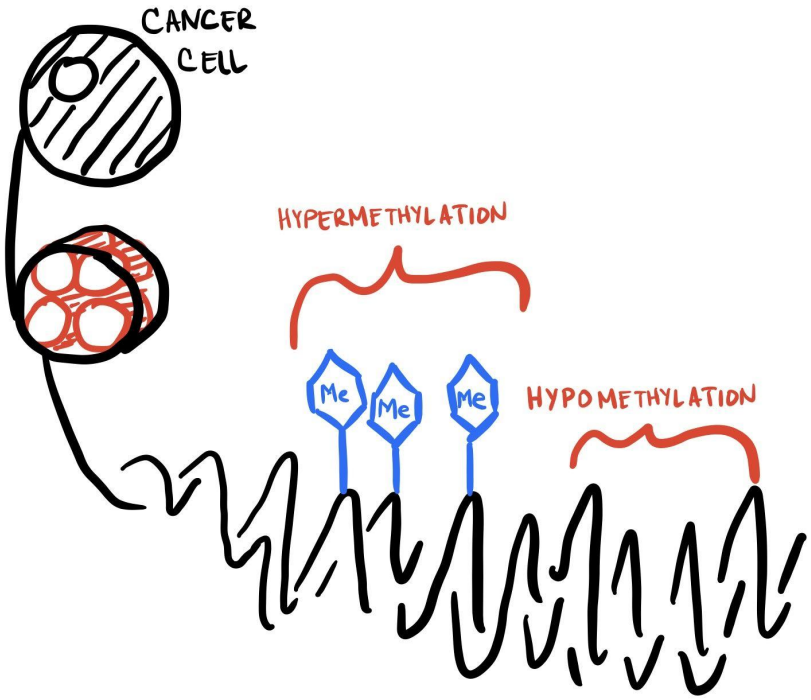
Motivation

43,600 women in the U.S. are expected to die in 2021 from breast cancer.

1 in 8 U.S. women (about 13%) will develop invasive breast cancer over the course of her lifetime.

Breast cancer is the most commonly diagnosed cancer among American women. In 2021, it's estimated that about 30% of newly diagnosed cancers in women will be breast cancers.

Previous Research Conclusions



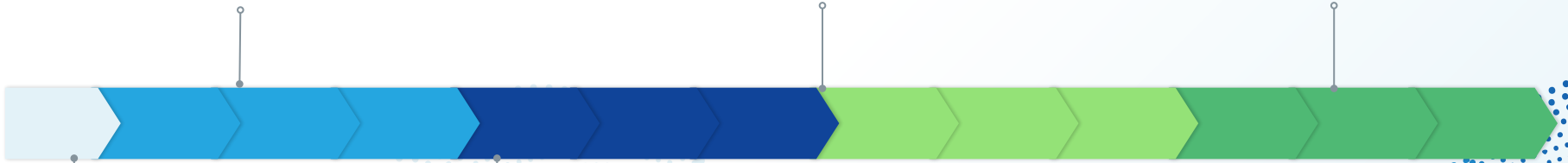
- Methylation occurring on only one DNA strand of a CpG site but not the other.
- Hemimethylation may be closely related to hypermethylation and hypomethylation patterns found in a cancer genome.

Our Process

We ran Wilcoxon tests on Forward and Reverse strands to detect methylation signals.

Grouping into Clusters, creating Manhattan plot, and applying a Sliding Window Approach.

Refining Sliding Window, looked at more chromosomes



We cleaned and filtered our data.

Filtered data out by |mean diff| and p value.

Our (Initial) Data

- CHR information
- CpG Site Location
- Position
- 7 breast cancer cell lines

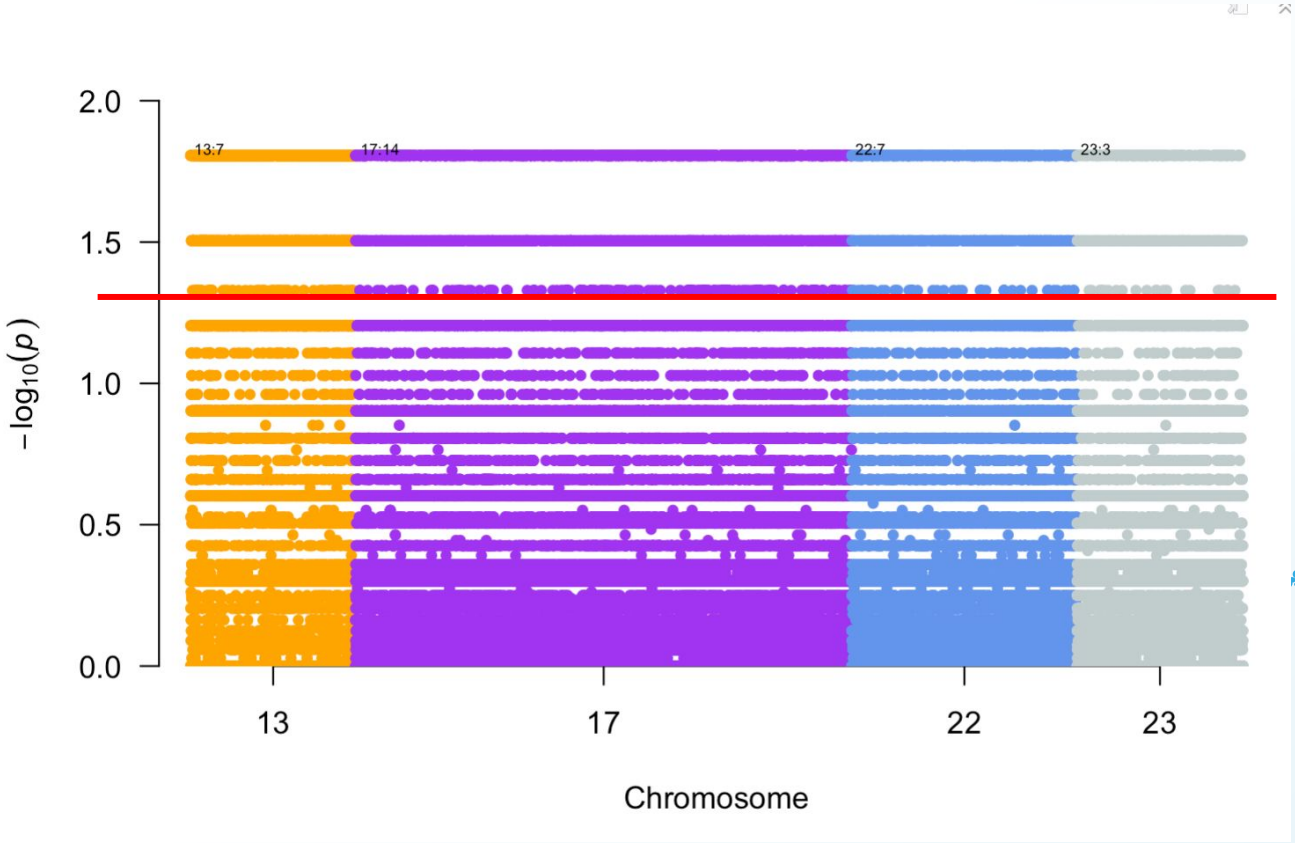
	*Length	CPG Sites	Density	*Reduced Length (excluding sites with 4 or more NAs)
Chromosome 22	50,818,468	578,097	Length/CPG ≈ 87.907	20218
Chromosome X	156,040,895	1,246,401	Length/CPG ≈ 125.193	14875

Wilcoxon Results

	CHR X		CHR 22	
	No p-value filter	p-value < 0.05	No p-value filter	p-value < 0.05
|Mean difference| ≥ 0	14875	640	20218	1190
|Mean difference| ≥ 0.4	5666	384	8661	711
|Mean difference| ≥ 0.6	4407	319	7113	612
|Mean difference| ≥ 0.8	2873	226	5127	471

Figure NUMBER.

Manhattan Plot

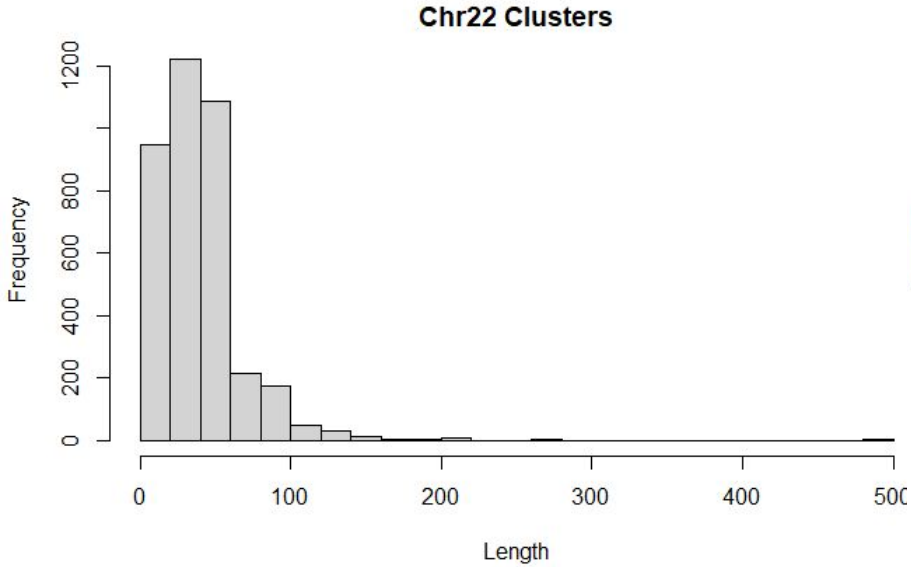


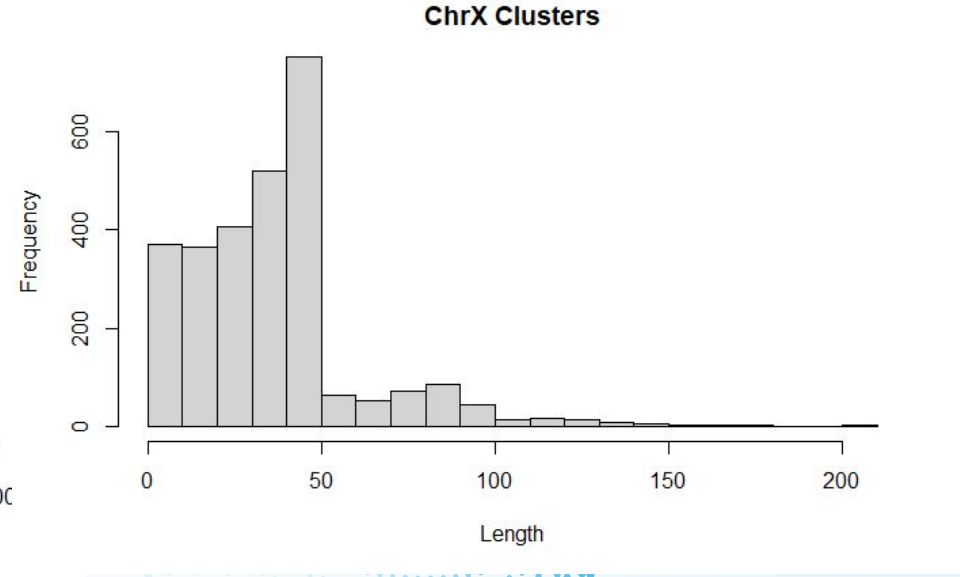
A Clustering Approach

- Each adjacent cg site is put in a cluster with solo sites put into the label 0
- Filtering is done by lowering the threshold of significance for all cg sites in a cluster

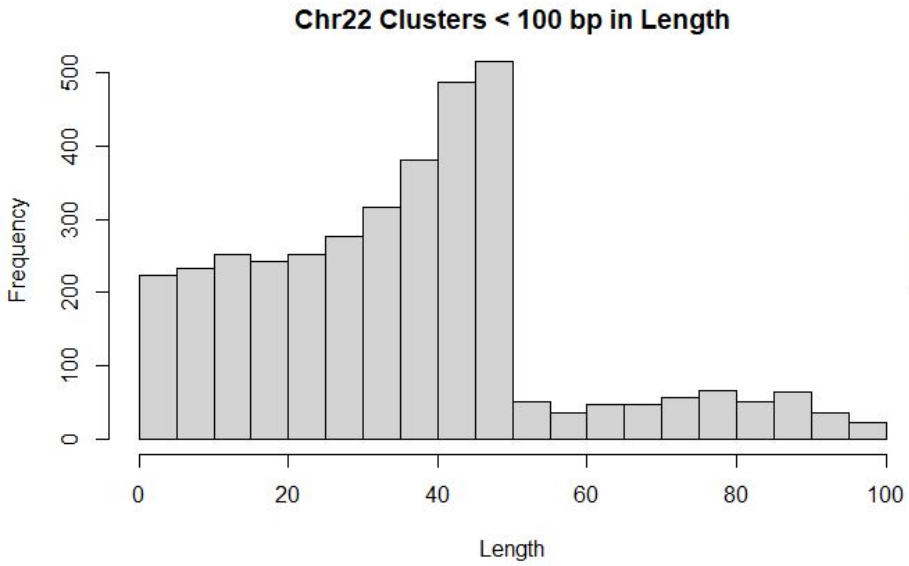
```
grouping <- function(df){
  df$adj <- rownames(df)
  groupNumber = 1
  df$groupNumber = df$adj
  for (group in (split(rownames(df), cumsum(c(1,diff(as.integer(rownames(df))) != 1))))
    if (length(group) > 1) {
      df[df$adj %in% group,'groupNumber'] = groupNumber
      groupNumber = groupNumber + 1
    } else{
      df[df$adj %in% group,'groupNumber'] = 0
    }
  }
  return(df)
}
```

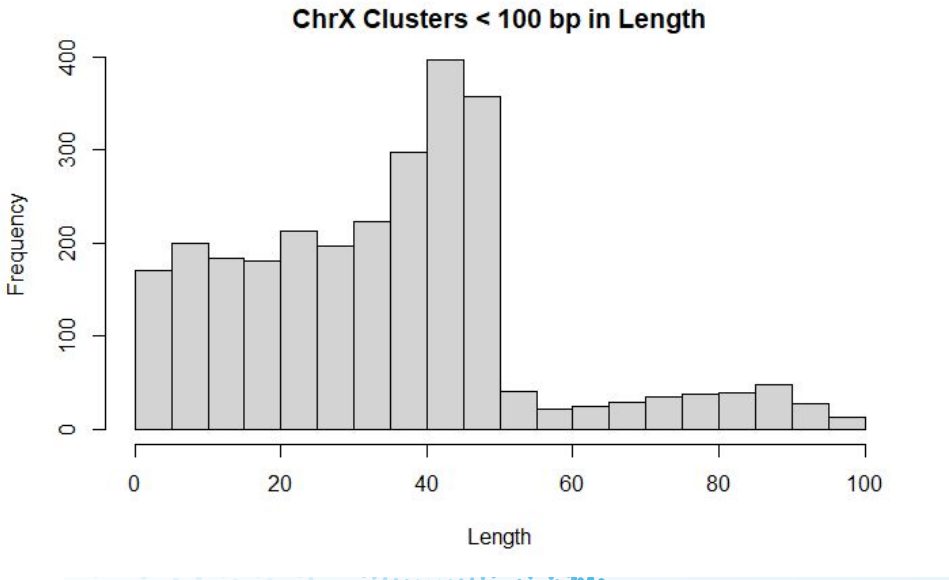
Cluster Length





Cluster Length < 100 BP





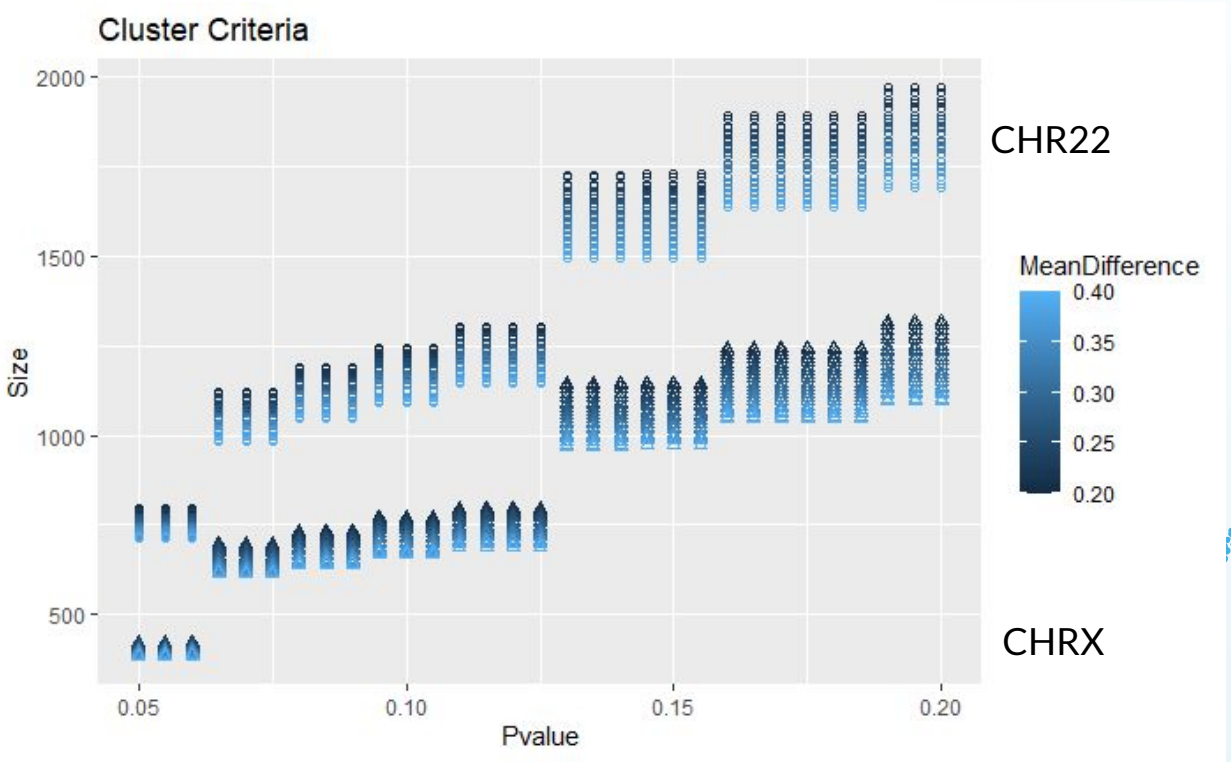
Supplemental

Supplemental Table

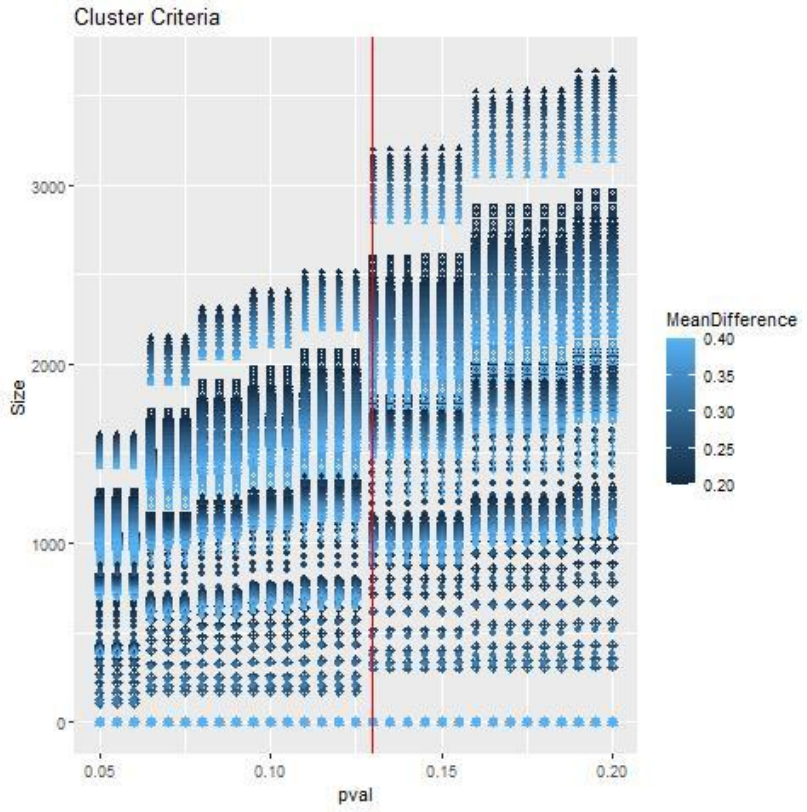
```
for (i in 1:22){
  print(i)
  fileFor <- read.table(paste0("cancer.hg19.for.chr", i,".txt"), header=TRUE, sep = "")
  fileRev <- read.table(paste0("cancer.hg19.rev.chr", i,".txt"), header=TRUE, sep = "")
  merged <- merge(fileFor, fileRev, by = "POSITION")
  merged <- composite(merged)
```

```
composite <- function(df){
  print("Removing missing values")
  df <- delete.na(df)
  print("calculating mean dif")
  df <- pval_meandif_calculation(df)
  print("Creating Groups")
  df <- grouping(df)
  print("Making Hist")
  cluster_hist(df)
  cluster_hist_zoomed(df)
  return(df)
}
```

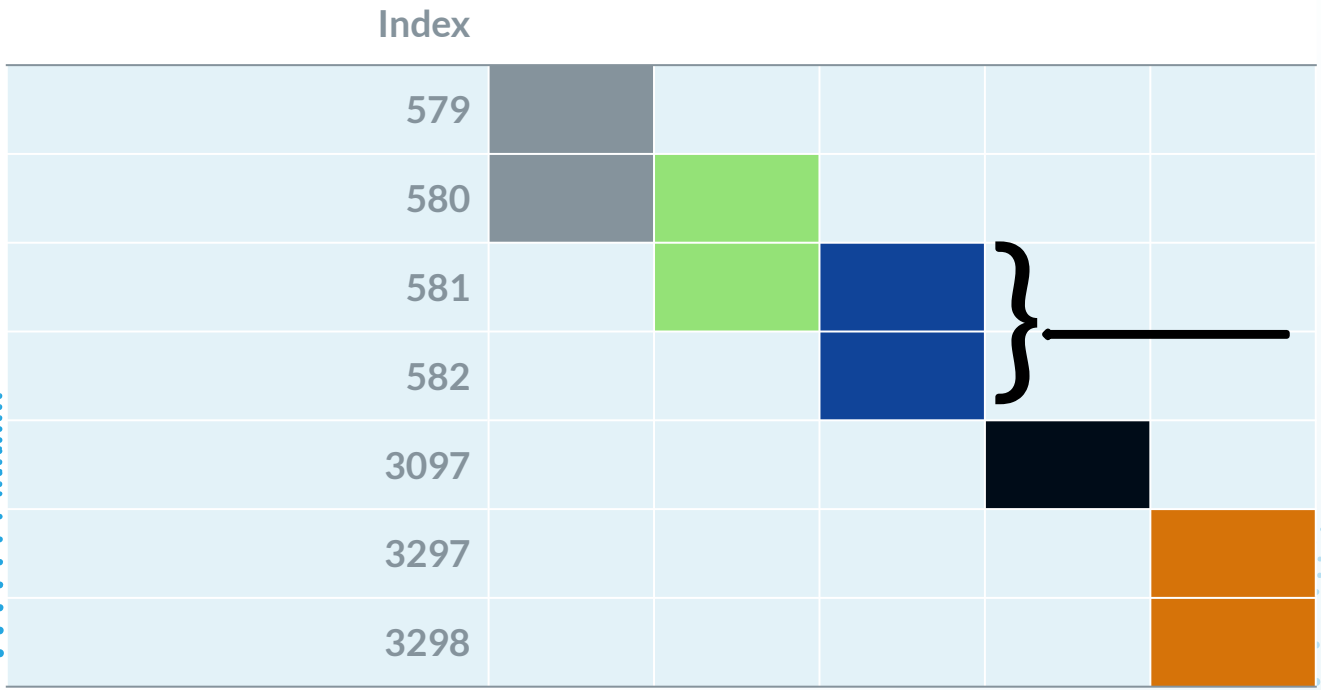
Old Cluster Criteria



Cluster Criteria with all Chr



Sliding Window Approach



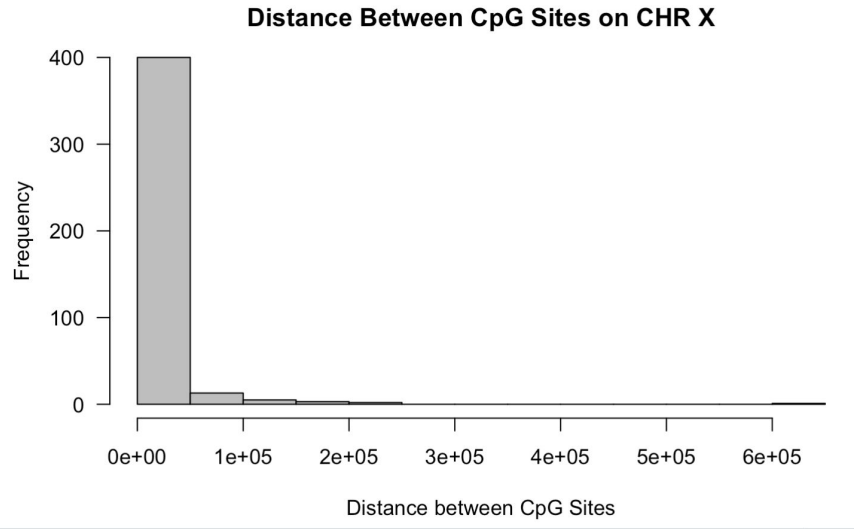
Each adjacent CpG site will be checked for two cutoff conditions (mean difference & p value).

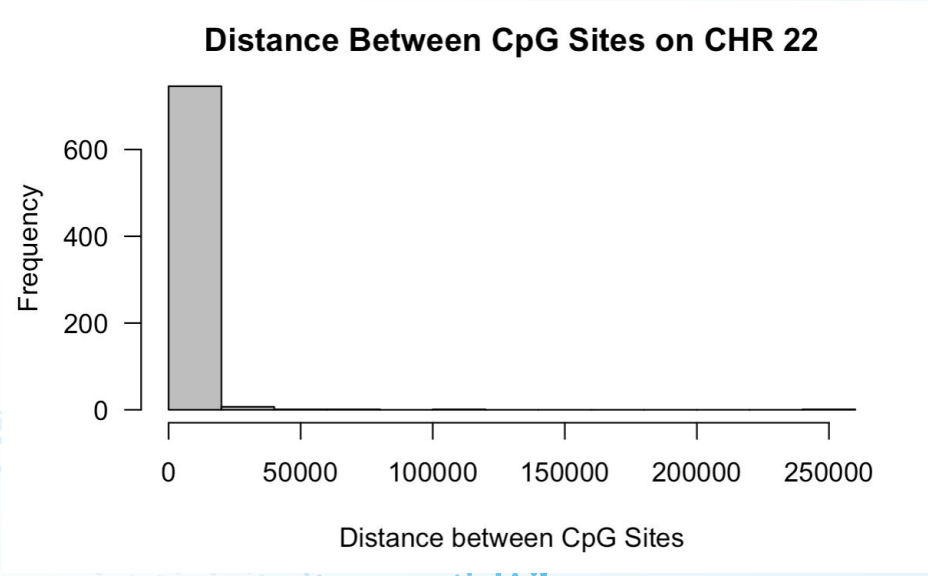
Exploring The CHRX Data Frame-Sliding Window

CpGSite <dbl>	POSITION <int>	pval <dbl>	meandiff <dbl>	Forward_Sliding_Mean_Pval <dbl>	Forward_Sliding_Mean_MeanDiff <dbl>	Reverse_Sliding_Mean_MeanDiff <dbl>	Reverse_Sliding_Mean_Pval <dbl>	diffCpGX <dbl>	diffPositionX <int>
52053	2723063	0.562500	0.851570000	0.4375000	0.8854334286	*NA*	*NA*	1	16
52054	2723079	0.312500	0.919296857	0.1640625	0.8877843571	0.8854334286	0.4375000	1	4
52055	2723083	0.015625	0.856271857	0.5078125	0.7988842143	0.8877843571	0.1640625	223	18913
52278	2741996	1.000000	0.741496571	1.0000000	0.4540816190	0.7988842143	0.5078125	110	4947
52388	2746943	1.000000	0.166666667	1.0000000	0.1666666667	0.4540816190	1.0000000	1	15
52389	2746958	1.000000	0.166666667	1.0000000	0.1666666667	0.1666666667	1.0000000	1	9
52390	2746967	1.000000	0.166666667	1.0000000	0.1833333333	0.1666666667	1.0000000	1	12
52391	2746979	1.000000	0.200000000	1.0000000	0.1460638236	0.1833333333	1.0000000	1	3
52392	2746982	1.000000	0.092127647	1.0000000	0.1710638236	0.1460638236	1.0000000	5	66
52397	2747048	1.000000	0.250000000	1.0000000	0.1939484143	0.1710638236	1.0000000	4	37

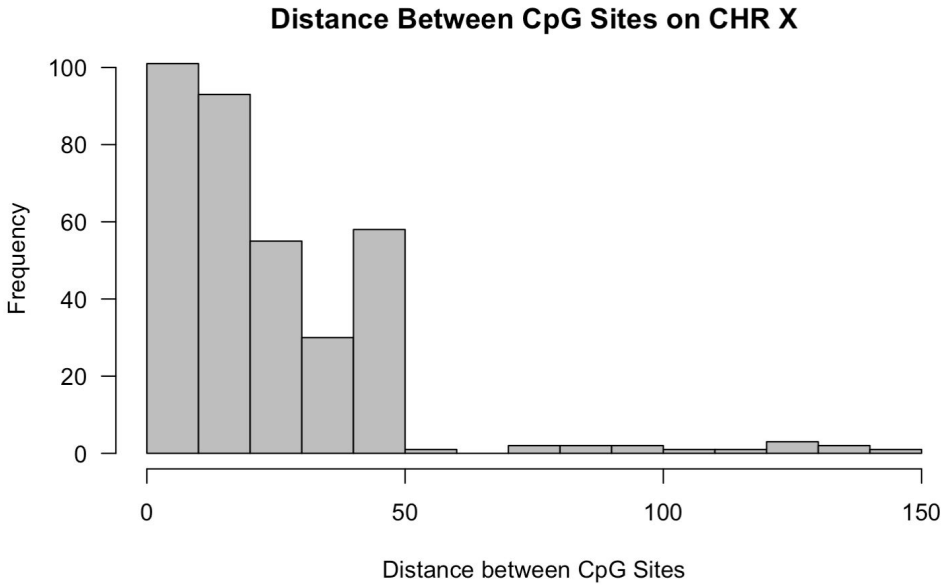
Rolling Window & Position Differences

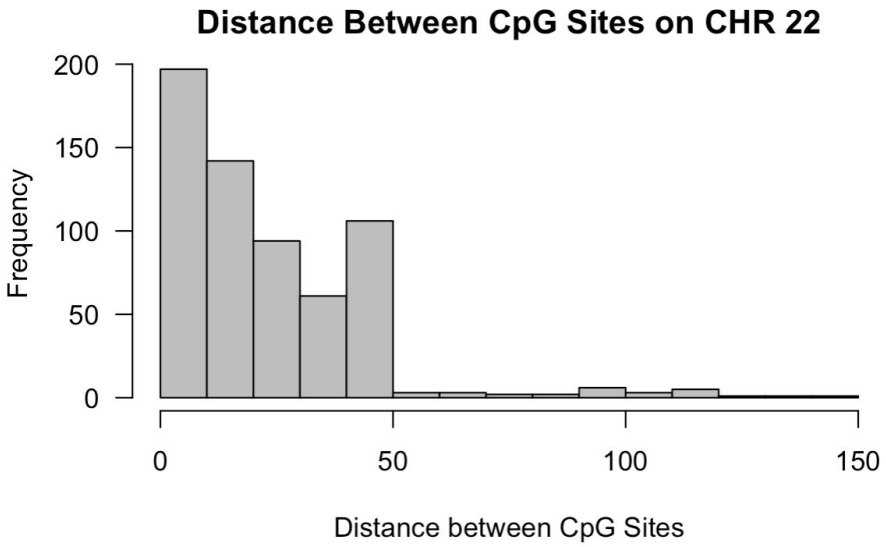
➔ We filtered on significant p values and mean differences.



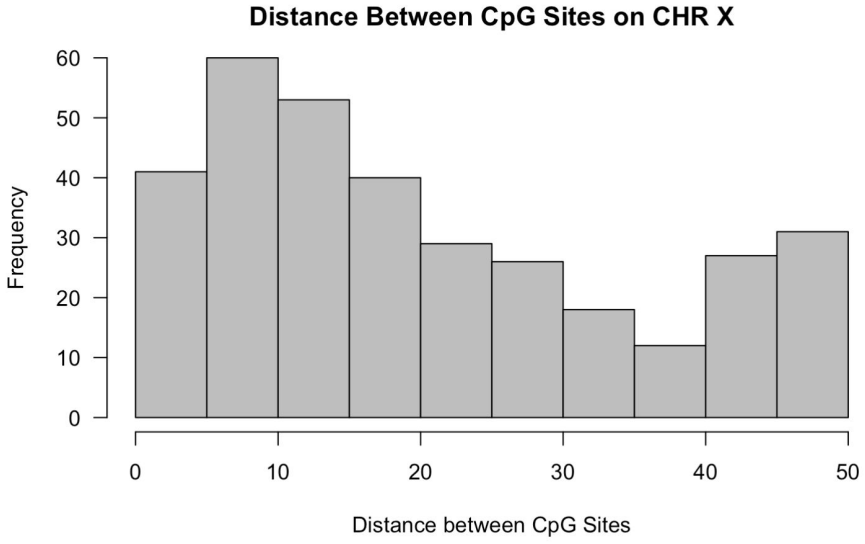


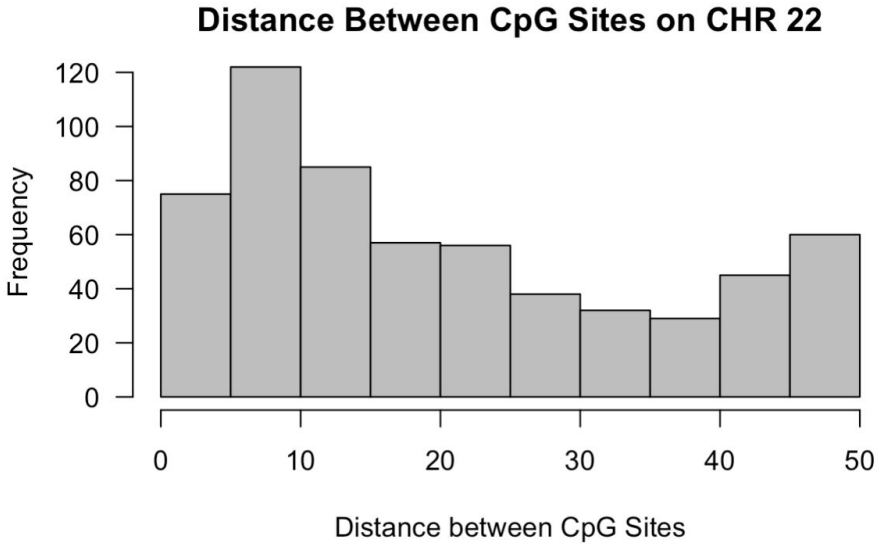
CpG Site Length <150





CpG Site Length <50





Significant CpG Sites Based on Different Criteria

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

Conclusion

- Cluster length falls off drastically after 50 base pairs.
- Distance between CG Sites is usually less than 50 base pairs.
- For clusters, a cut off p value of .13 and a cut off mean difference of .4 drastically improve sample size without lowering significance by much.
- Continued research in the field:
  - (2002)DNA Methylation Inhibitors **5-azacytidine (5-aza)** and **5-aza-deoxycytidine** https://www.nature.com/articles/1205699
  - (2010) Breast Cancer Epigenetics research effort to isolate what gene hyper or hypomethylation are unique to breast cancer. https://www.sciencedirect.com/science/article/pii/S1574789110000244
  - (2021) Pediatric T-cell acute lymphoblastic leukemia (T-ALL) research treatment with the DNA demethylating agent, **5-azacytidine (5-aza)** and T-ALL is correlated with hypermethylation of the *TET2* promoter https://www.pnas.org/content/118/34/e2110758118
  - (2021)Tumor biomarkers are usually proteins measured either in serum, plasma or tumor tissue, and other noninvasive screening methods.
    - Liquid biopsieshttps://www.aacr.org/blog/2021/04/14/aacr-annual-meeting-2021-facilitating-early-cancer-detection-with-liquid-biopsy/
  - Overall methylation method has **higher sensitivity** due to the presence of multiple methylation sites within a single gene. Still very promising approach.