Investigating X Chromosome Co-methylation in Breast and Endometrial Cancer

Bridget Bangert and Madison Glenwinkel

Abstract

Introduction: Research focusing on the X chromosome and DNA methylation (DNAm) has shown that the X chromosome may play an important role in the cause of breast and endometrial cancers. In DNAm, a methyl group is added to the 5' end of a cytosine nucleotide that resides next to a guanine nucleotide, and is known as a CpG pair. In this study we will conduct CpG site level analyses using matched normal and tumor samples in alive and dead breast cancer (BRCA) and endometrial cancer (UCEC) data. **Methods:** We calculated the total number of highly correlated CpG sites, distinguishing between absolute high correlations, high positive correlations and high negative correlations. Then we calculated the distances between highly correlated CpG pairs. **Results**: BRCA and UCEC cancer co-methylate differently. UCEC samples had more highly correlating CpG sites than BRCA samples. In UCEC, tumor samples had more highly correlating CpG sites than normal samples though this pattern was not seen in BRCA. The majority of high correlations were positive, only the UCEC dead sample had a non-zero average of high negative correlations. Distances between highly correlated CpG sites were shorter in tumor samples than in normal samples, and this difference was more pronounced in BRCA than in UCEC. **Conclusion:** There are differences in the co-methylation patterns between BRCA and UCEC and between normal and tumor samples. Further research should be done on these patterns including conducting these same site level analyses on the autosomes.

Introduction

Epigenetics is the field of study concerned with modifications to the genome, deoxyribonucleic acid (DNA), that result in altered gene expression without changing the DNA sequence. One such modification is DNA methylation (DNAm) in which a methyl group is added to the 5' end of a cytosine nucleotide that resides next to a guanine nucleotide, known as a CpG pair. The methyl group added through this process blocks DNA transcription, and thus silences gene expression (Sun and Sun 2019). DNAm can lead to co-methylation, in which there is a high correlation between the methylation signals of multiple CpG sites (Sun et al. 2022). This methylation can be understood as occurring between CpG sites from different genomic regions, or between sample methylation, or between CpG sites from similar genomic regions, or within sample (WS) co-methylation. Studies have shown that WS co-methylation is associated with

various different cancers (Zhang and Huang 2017), including breast cancer (Sun et al. 2022) and endometrial cancer (Men et al. 2017).

Sun and colleagues (2022) studied WS co-methylation patterns assessing highly correlated CpG pairs (|r|>=0.8) in normal (non-tumor) versus tumor samples across all chromosomes. They found that normal samples had more negative highly correlated sites than tumor samples (6.6% vs 2.8%) and that chromosome X displayed a unique pattern with regard to the distances between highly correlated CpG pairs compared to the autosomes. In the autosomes, there was large variation in the distances between highly correlated CpG pairs, with normal samples having slightly larger distances between pairs than tumor samples. However, on the X chromosome this difference was more marked, with almost 45% of highly correlated CpG pairs in tumor samples being within 10 million base pairs of each other compared to only 17% in normal samples. They also found that chromosome X had a higher number of CpG sites that did not highly correlate with other CpG sites.

Research focusing on the X chromosome and DNA methylation has shown that the X chromosome may play an important role in the cause of breast and endometrial cancers. This pattern is interesting in the context of the unique expression of the X chromosome in females. Because the X chromosome carries a large number of genes, and females carry two X chromosomes, one is deactivated to prevent double gene expression. Both Kang and colleagues (2015) and Sun and colleagues (2015) found that the inactive X chromosome tended to be lost in breast and ovarian cancers. Both studies also found that the X chromosome in these instances tended to be hypomethylated, or have lower levels of DNAm relative to the autosomes.

As opposed to gene level analyses, studying DNAm at the CpG site level allows for a more accurate and less biased representation of the epigenetic changes occurring in the genome (Sun et al. 2022). Based on previous research finding aberrant patterns in the X chromosome, we propose that it is necessary to delve deeper into X chromosome co-methylation patterns between normal and tumor samples to identify significant, site level patterns. In addition, learning about the epigenetic changes on the X chromosome can inform detection, prevention and treatment methods for some of the most common female-only cancers in the world. In this study we compare WS co-methylation patterns on the X chromosome in normal versus tumor samples for both alive and dead individuals with breast cancer (BRCA) and endometrial cancer (UCEC). By comparing normal versus tumor samples in alive and dead BRCA and UCEC data, our objectives are as follows: 1) Compare the total number of highly correlated CpG sites 2) Compare high positively and high negatively correlated CpG sites, and 3) Compare distances between highly correlated CpG sites.

Materials and Methods

Sample Description

Both breast cancer data (BRCA) and endometrial cancer data (UCEC) show the methylation signals in a subset of CpG sites of the genomes of both dead and alive cancer patients. BRCA alive data consists of 53 samples, BRCA dead data consists of 32 samples, UCEC alive data

consists of 26 samples and UCEC dead data consists of 7 samples. Each sample consists of the methylation signal for 9498 CpG sites (table 1).

Table 1: Description of data used. N refers to the number of samples analyzed and nrow refers to the number of CpG sites in each sample.

Sample	N	nrow
BRCA Alive Normal	53	9498
BRCA Alive Tumor	53	9498
BRCA Dead Normal	32	9498
BRCA Dead Tumor	32	9498
UCEC Alive Normal	26	9498
UCEC Alive Tumor	26	9498
UCEC Dead Normal	7	9498
UCEC Dead Tumor	7	9498

Software

Analyses were performed by two independent researchers using RStudio Version 2023.06.1+524 (2023.06.1+524) and macOS Terminal Version 2.14 (453).

Identifying Highly Correlated CpG sites

To identify the number of highly correlated CpG sites for these data, analyses were conducted on each sample as follows:

1) Compute Correlation Matrices:
    a) Generate pairwise correlation matrices for CpG sites within each sample group.
2) Identify High Correlations:
    a) Select correlations with an absolute value >= 0.8, excluding self-correlations.
    b) For the UCEC dead data, the threshold was set to 0.9 due to the small sample size.
3) Sum High-Correlation Instances:
    a) For each CpG site (row), count how many correlations exceed the threshold.

b) Aggregate these counts for all rows in the matrix.

The outcome of this analysis is a single value representing the total number of high-correlation instances per group.

Identifying High Negative and High Positive Correlations

The same correlation matrices generated in step one of the previous step were then used to distinguish between high positive and highly negative correlated CpG sites.

1) Compute Correlation Matrices:
    a) Generate pairwise correlation matrices for CpG sites within each sample group.
2) Identify High Positive Correlations:
    a) Select correlations >= 0.8, excluding self correlations.
3) Identify High Negative Correlations:
    a) Select correlations <=(-0.8)
4) Sum High-Correlation Instances:
    a) For each CpG site (row), count how many correlations exceed the threshold for both the positive and negative correlations
    b) Aggregate these counts for all rows in the matrix.

The outcome of this analysis is a table that contains the total number of high positive correlations and the total number of high negative correlations for each sample group.

Comparing Distances Between Highly Correlated CpG Pairs

To calculate the distances between highly correlated CpG pairs, analyses were conducted as follows:

1) Calculate Pairwise Distances:
    a) Generate the top half of a pairwise distance matrix for all CpG site pairs in each sample group using the start value for each CpG site.
2) Filter by High Correlation:
    a) Using the previously generated pairwise correlation matrices, extract the indexes of the highly correlated CpG sites.
3) Extract Distances
    a) Extract the distances from the distance matrix using the indexes of highly correlated sites.
    b) Do this for high absolute, negative, and positive correlations.

The outcome of this analysis is a table that contains five columns: index 1, index 2, start position of index 1, start position of index 2, and distance.

Comparing Generated Data with Noise

To generate the original data with noise, the following procedure is as follows:

1. Generate a noise matrix for each sample.
    a. Generate a matrix of random numbers between a certain noise threshold ([-0.1,01] was used for this analysis) for each sample dataset.
    b. Add these matrices to its original sample dataset.
2. Conduct the same analysis as the previous steps listed above.
    a. Calculate the correlation matrix for each sample.
    b. Extract the positive, negative, and absolute high correlation counts and positions.
    c. Extract the new distances with the original pairwise distance matrix for each sample.
3. Compare these results to their corresponding original sample data.

The outcome of this analysis will allow us to determine if the patterns apparent from the original analyses are true or impacted by the inclusion of noise in the data.

Results

Identifying Highly Correlated CpG sites

High correlations between CpG sites methylation signals suggest that these sites are likely regulated together or share similar methylation patterns. When two or more CpG sites exhibit high correlation, it often indicates that their methylation status is co-regulated, meaning they may be involved in the same biological processes, pathways, or gene regulatory networks. In the context of cancer, high correlations between CpG sites could imply that certain genes or regions of the genome are undergoing similar epigenetic changes. Therefore, high correlations in comethylation data could provide valuable insights into potential biomarkers for cancer, or help identify key regulatory regions of the genome (in this analysis, specifically the X Chromosome) that are altered in disease.

BRCA

The analysis of highly correlated CpG sites presented challenges due to the significant clustering of values around zero. To address this, a fourth-root transformation was applied to the high-correlation counts, yielding a distribution that is much easier to interpret (this transformation will be applied to most of the values in this analysis). Despite the transformation, the breast cancer data for all samples remained clustered near zero; however, samples from dead samples exhibited fewer zero values compared to those from the living samples.

When comparing dead and living samples, the fourth-root-transformed counts for tumor data demonstrated a stronger clustering around zero than the corresponding counts for normal tissue. Additionally, the tumor data showed a reduced spike near a value of one, particularly in the living samples, with a similar pattern observed in the deceased data. Notably, the outlier values were smaller for tumor counts compared to normal counts. These trends were consistent across both the deceased and living breast cancer datasets, highlighting distinct distributional differences between the groups.

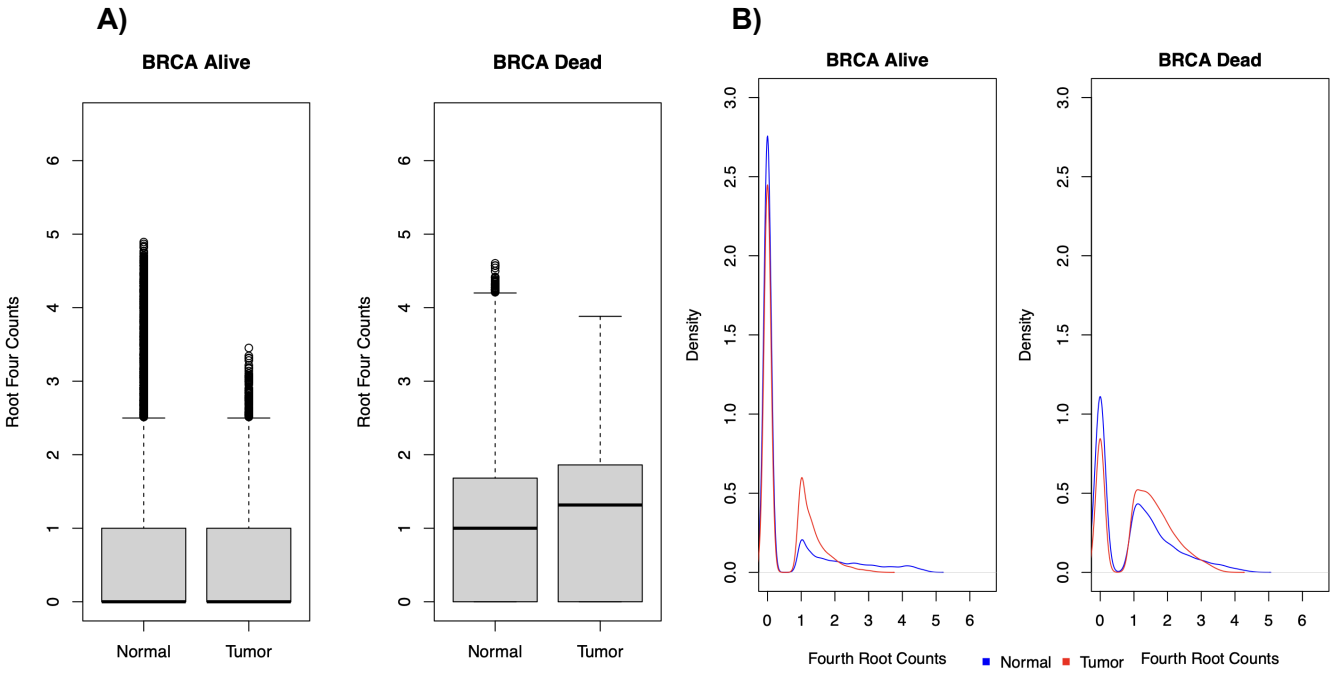


Figure 1: **A)** Boxplots showing the fourth root of high correlation counts between BRCA alive and dead sample groups. **B)** Density plots showing the distribution of high correlation counts between BRCA alive and dead samples.

Madison

A) File: /home/vbz21/math5376D/final.project/Plots/BRCA.highcorr.boxplots.pdf

B) File: /home/vbz21/math5376D/final.project/Plots/BRCA.highcorr.density.pdf

Code: /home/vbz21/math5376D/final.project/Code/11.25.highcorr.plots.R

Bridget

A) File: /home/bmb191/math5376D/Final.Project/Output.Files/Plots/BRCA.highly.correlated.counts.boxplot.pdf

B) File: /home/bmb191/math5376D/Final.Project/Output.Files/Plots/BRCA.highly.correlated.counts.density.plot.pdf

Code: /home/bmb191/math5376D/Final.Project/Scripts/generate.final.plots.R

UCEC

In the UCEC alive sample, the tumor group has a higher mean number of high correlation counts than the normal group. This pattern is also seen in the UCEC dead data, with the tumor group having a higher mean number of highly correlated CpG sites than the normal group. Interestingly with the UCEC dead sample, there were no CpG sites that were not highly correlated (figure 2A)

When looking at the density of high correlation counts, it is clear that in the UCEC alive data, the high correlation counts are clustering around zero in both normal and tumor samples. The density of high correlations in the higher count values are greater in tumor samples suggesting that the tumor samples have more CpG sites that highly correlate with a greater number of other sites. This same general pattern is seen in the UCEC dead data with the normal sample having a higher density of CpG sites that highly correlate with fewer total CpG sites than in the tumor sample (figure 2B).

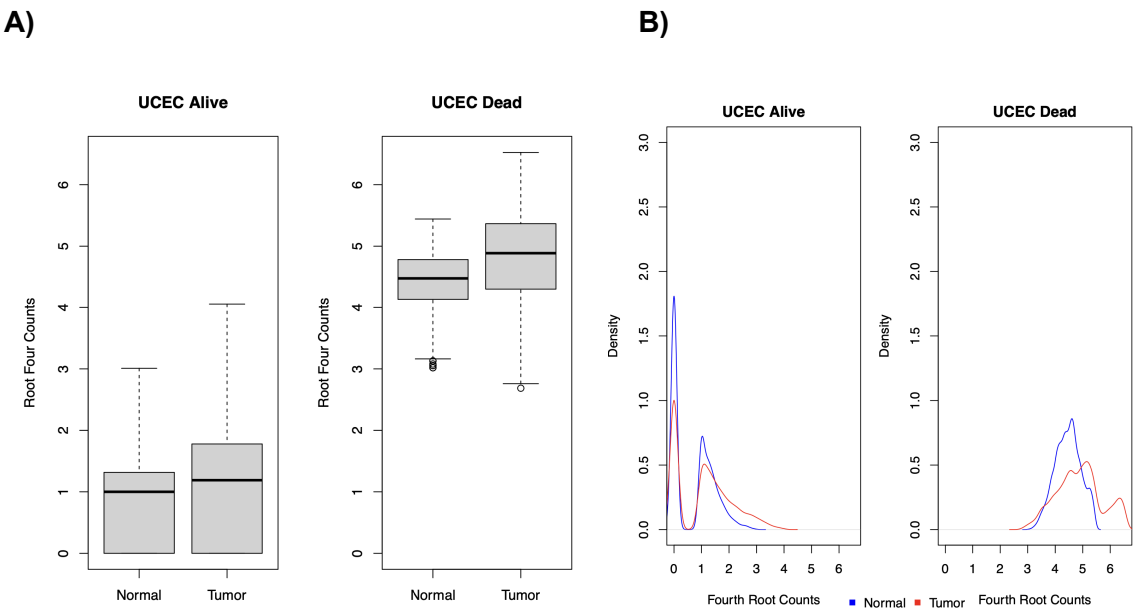


Figure 2: **A)** Boxplots showing the fourth root of high correlation counts between UCEC alive and dead sample groups. **B)** Density plots showing the distribution of high correlation counts between UCEC alive and dead samples.

Madison
- A) File: /home/vbz21/math5376D/final.project/Plots/UCEC.highcor.boxplots.pdf
- B) File: /home/vbz21/math5376D/final.project/Plots/UCEC.highcorr.density.pdf

Code: /home/vbz21/math5376D/final.project/Code/11.25.highcorr.plots.R

Bridget
- A) File: /home/bmb191/math5376D/Final.Project/Output.Files/Plots/UCEC.highly.correlated.counts.boxplot.pdf
- B) File: /home/bmb191/math5376D/Final.Project/Output.Files/Plots/UCEC.highly.correlated.counts.density.plot.pdf

Code: /home/bmb191/math5376D/Final.Project/Scripts/generate.final.plots.R

UCEC vs BRCA

There are slight differences between the pattern of high correlation counts in UCEC and BRCA data. The UCEC data shows more of a difference between the high correlation counts of normal and tumor samples than the BRCA data does. In BRCA data, the normal samples have higher variance while in the UCEC data, the tumor sample has higher variance (figure 3A).

The overall shape of the density distributions seen in figure 3B are the same across all samples, even when considering the UCEC dead data. In all samples, the normal samples have a higher

density of CpG sites that highly correlate with fewer other CpG sites than the tumor samples. When looking at the density distributions of BRCA alive and UCEC alive data, there are opposite peaks around the 1-2 fourth root count rates. In the BRCA data, the tumor sample has a much higher density of CpG sites that highly correlate with 1-2 (fourth rooted) other CpG sites. However in the UCEC alive data, the normal sample and tumor sample have almost the same distribution with the normal sample having a slightly higher density of counts within this value range.

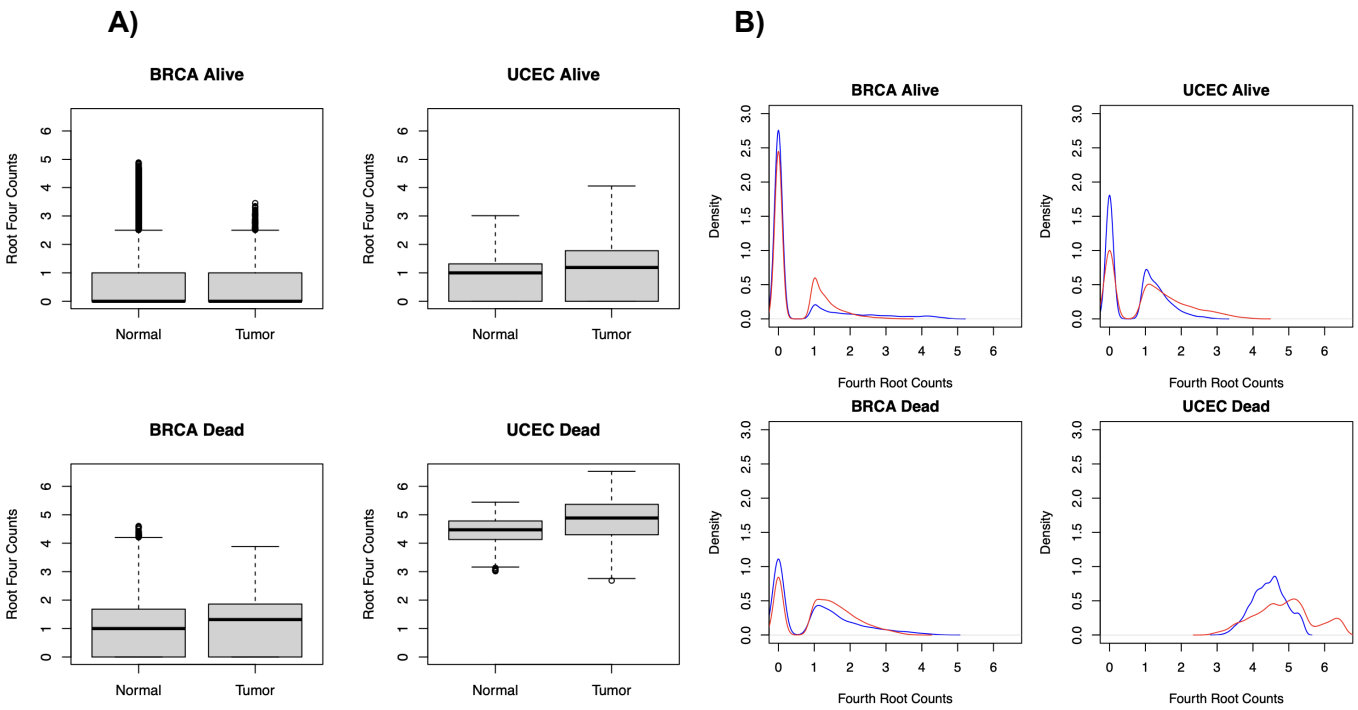


Figure 3: **A)** Boxplots showing the fourth root of high correlation counts for all samples **B)** Density plots showing the distribution of high correlation counts for all samples

Madison

A) File: /home/vbz21/math5376D/final.project/Plots/BRCA_UCEC_highcorrcounts_boxplots_pdf

B) File: /home/vbz21/math5376D/final.project/Plots/BRCA_UCEC_highcorrcounts_densityplots_legends.pdf

Code: /home/vbz21/math5376D/final.project/Code/11.25.highcorr.plots.R

Bridget

A) File: /home/bmb191/math5376D/Final.Project/Output.Files/Plots/counts.boxplots.pdf

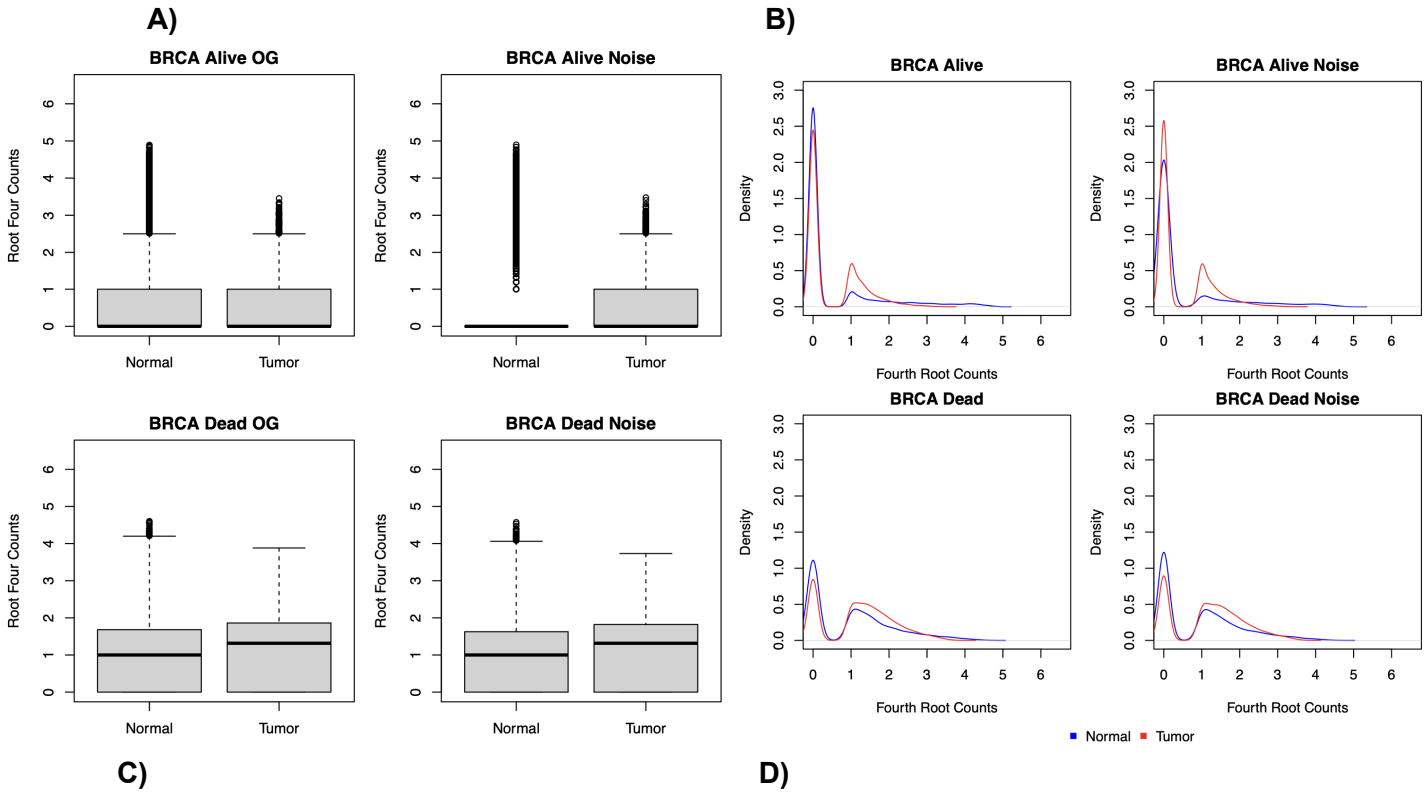
B) File: /home/bmb191/math5376D/Final.Project/Output.Files/Plots/counts.density.plots.pdf

Code: /home/bmb191/math5376D/Final.Project/Scripts/plot.generator.Nov.18.R

Adding Noise

To evaluate whether the patterns of high correlation counts in BRCA and UCEC data were impacted by the noise in the data, randomly generated noise was added to the original data sets and analyses of high correlation counts. By checking if the patterns are maintained even with the addition of noise, we can confirm that these patterns actually exist.

All samples maintained the same pattern even after the addition of noise aside from the BRCA alive normal data (Figure 5A, B). This sample clustered towards zero in the noise added data, with the density of high correlation in the tumor sample having a higher density of CpG sites with 0 correlations in the noise data whereas the opposite was true in the original data (Figure 5B). In the UCEC data, all patterns were consistent even after the addition of noise (figure 5C, D).



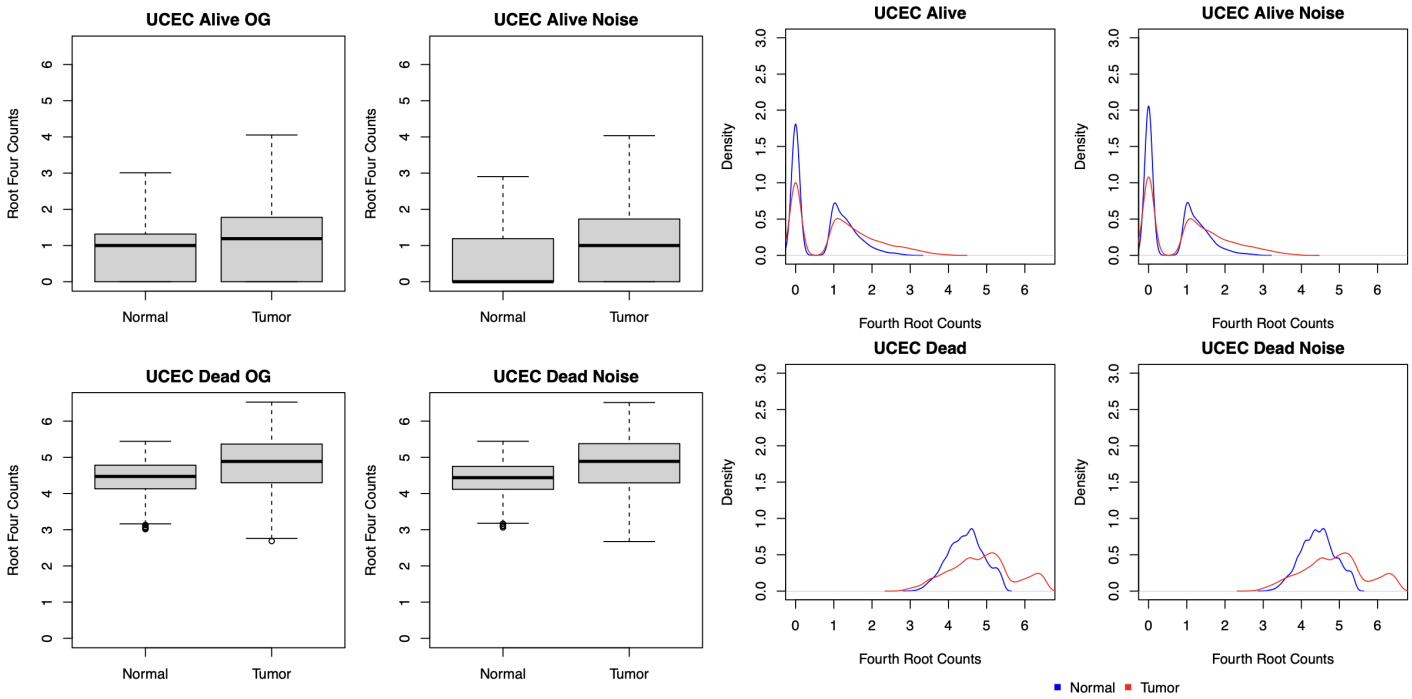


Figure 5: **A)** Boxplots showing the fourth root of high correlation counts in original and noise-added BRCA samples. **B)** Density plots showing the distribution of high correlation counts for original and noise-added BRCA samples. **C)** Boxplots showing the fourth root of high correlation counts in original and noise-added UCEC samples. **D)** Density plots showing the distribution of high correlation counts for original and noise-added UCEC samples.

Madison

A) File: /home/vbz21/math5376D/final.project/Plots/BRCA.highcorrcounts.noisevsOG.boxplot.pdf
B) File: /home/vbz21/math5376D/final.project/Plots/BRCA.highcorrcounts.noisevsOG.density.pdf
C) File: /home/vbz21/math5376D/final.project/Plots/UCEC.highcorrcounts.noisevsOG.boxplot.pdf
D) File: /home/vbz21/math5376D/final.project/Plots/UCEC.highcorrcounts.noisevsOG.density.pdf

Code: /home/vbz21/math5376D/final.project/Code/11.27.noise.matrix.R

Bridget

A) File: /home/bmb191/math5376D/Final.Project/Output.Files/Plots/BRCA.highly.correlated.counts.boxplot.with.noisepdf
B) File: /home/bmb191/math5376D/Final.Project/Output.Files/Plots/BRCA.highly.correlated.counts.density.plot.with.noise.pdf

Code: /home/bmb191/math5376D/Final.Project/Scripts/generate.final.plots.R

To further investigate the effect of the addition of noise, scatter plots were created comparing noise added and original data for each sample. In the BRCA samples, all scatter plots showed a relatively linear 1:1 ratio of original to noise data, meaning that noise did not have an impact on the distribution of high correlation counts (Figure 6A). Interestingly in the UCEC data which appeared unaffected in the box and density plots (Figure 5C, D) shows less of a linear distribution in the scatter plots. The UCEC dead normal sample shows a very wide distribution of points around the 1:1 line suggesting that noise does have an impact on the general pattern of results. Interestingly, only the dead normal sample appears to be greatly affected (figure 6B). UCEC alive normal also shows some deviations from a linear relationship suggesting that noise may have a slight impact (figure 6B). Based on the box and density plots of the UCEC dead data (Figure 3B) that look very different from the plots for the rest of the samples, these scatter plots hint that noise may be impacting the distributions of the low sample UCEC dead data.

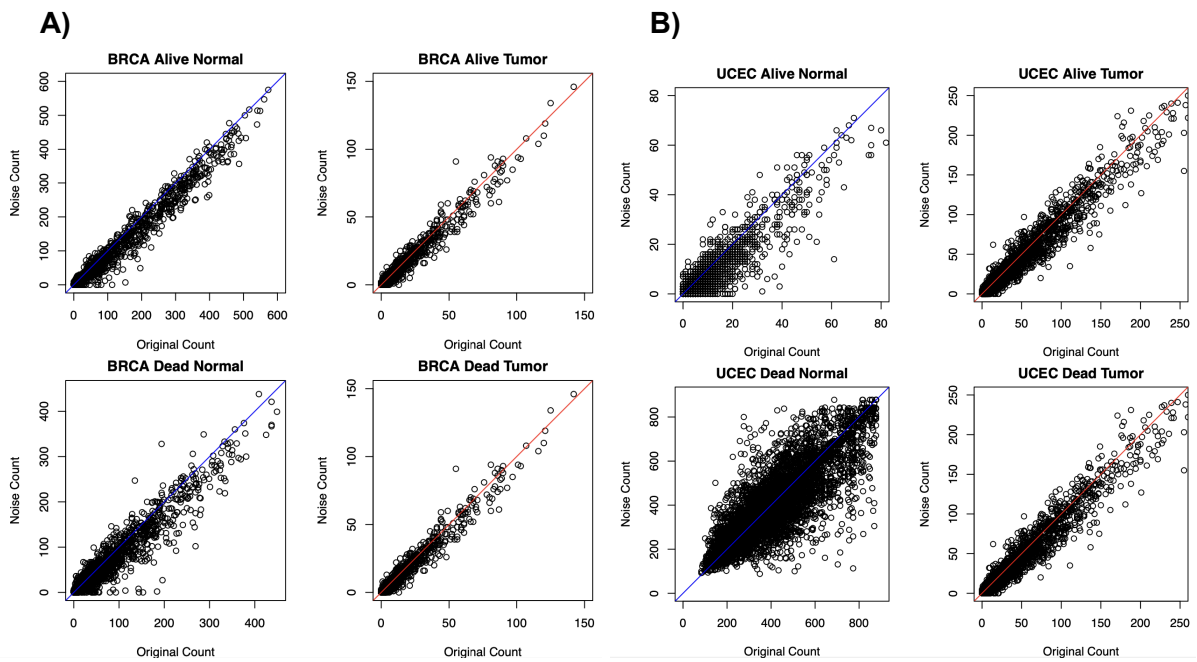


Figure 6: Scatter plots with original versus noise-added high correlation counts. The x axis consists of high correlation counts from the original data. The y axis consists of high correlation counts from the noise-added data. A line with a slope of 1 was added to each plot, the line being blue in normal samples and red in tumor samples. A) BRCA samples. B) UCEC samples.

Madison

A) File: /home/vbz21/math5376D/final.project/Plots/BRCA.original.noise.scatter.pdf
B) File: /home/vbz21/math5376D/final.project/Plots/UCEC.original.noise.scatter.pdf

Code: /home/vbz21/math5376D/final.project/Code/11.27.noise.matrix.R

Bridget

A) File: /home/bmb191/math5376D/Final.Project/Output.Files/Plots/BRCA.alive.normal.highly.corr.cnts.scatterplot.with.noise.pdf
B) File: /home/bmb191/math5376D/Final.Project/Output.Files/Plots/BRCA.alive.tumor.highly.corr.cnts.scatterplot.with.noise.pdf
C) File: /home/bmb191/math5376D/Final.Project/Output.Files/Plots/BRCA.dead.normal.highly.corr.cnts.scatterplot.with.noise.pdf
D) File: /home/bmb191/math5376D/Final.Project/Output.Files/Plots/BRCA.dead.tumor.highly.corr.cnts.scatterplot.with.noise.pdf

Code: /home/bmb191/math5376D/Final.Project/Scripts/generate.noise.plots.R

Identifying Negative and Positive Correlations

In this part of the analysis, we independently analyze the high negative and high positive correlation counts of the samples to gain insights into their distinct distributional properties. By treating these counts as a separate analysis, we aim to explore potential biological or statistical significance within the samples. This approach allows for a more detailed examination of the correlation structures and their implications across different sample groups.

BRCA Positive

Both groups exhibit similar overall distributions; however, the deceased group demonstrates a broader range. Additionally, the distribution for the deceased group appears to be centered at a median around 1, compared to the median of 0 in the living group (figure 7A). The living group also contains a greater number of outliers with higher counts. Based on the density plot, there is minimal difference within the distribution of the deceased group. In contrast, the living group's

tumor data shows a slightly higher spike at 1 compared to the normal group (figure 7B). These distributions are very similar to the absolute high correlation distribution since the majority of them are positive in the BRCA samples.

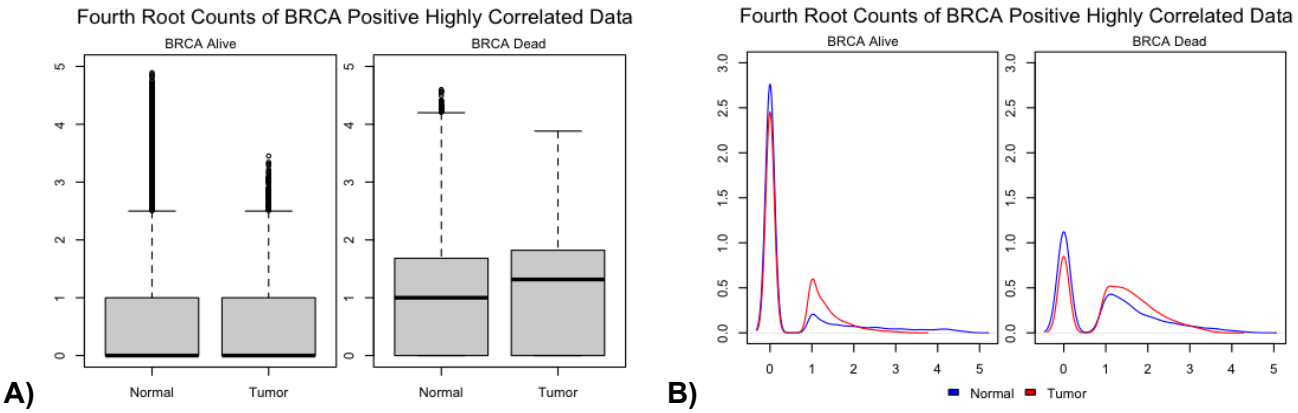


Figure 7: **A)** Boxplots showing the fourth root of positive high correlation counts between BRCA alive and dead sample groups. **B)** Density plots showing the distribution of positive high correlation counts between BRCA alive and dead samples.

Madison

A) File: /home/vbz21/math5376D/final.project/Plots/BRCA.UCEC.positive.highcorr.boxplots.pdf

B) File: /home/vbz21/math5376D/final.project/Plots/BRCA.UCEC.positive.highcorr.density.pdf

Code: /home/vbz21/math5376D/final.project/Code/Pos.Neg.Corrs.R

Bridget

A) File: /home/bmb191/math5376D/Final.Project/Output.Files/Plots/BRCA.positive.highly.correlated.distances.boxplot.pdf

B) File: /home/bmb191/math5376D/Final.Project/Output.Files/Plots/BRCA.positive.highly.correlated.distances.density.plot.pdf

Code: /home/bmb191/math5376D/Final.Project/Scripts/generate.final.plots.R

BRCA Negative

The majority of CpG sites exhibit counts at 0, particularly within the tumor group. This observation aligns with the limited presence of high negative correlations in the BRCA dataset. Notably, the clustering at 0 is more pronounced in the tumor samples from living samples, whereas the normal tissue data maintains a consistent distribution across groups (Figure 8)

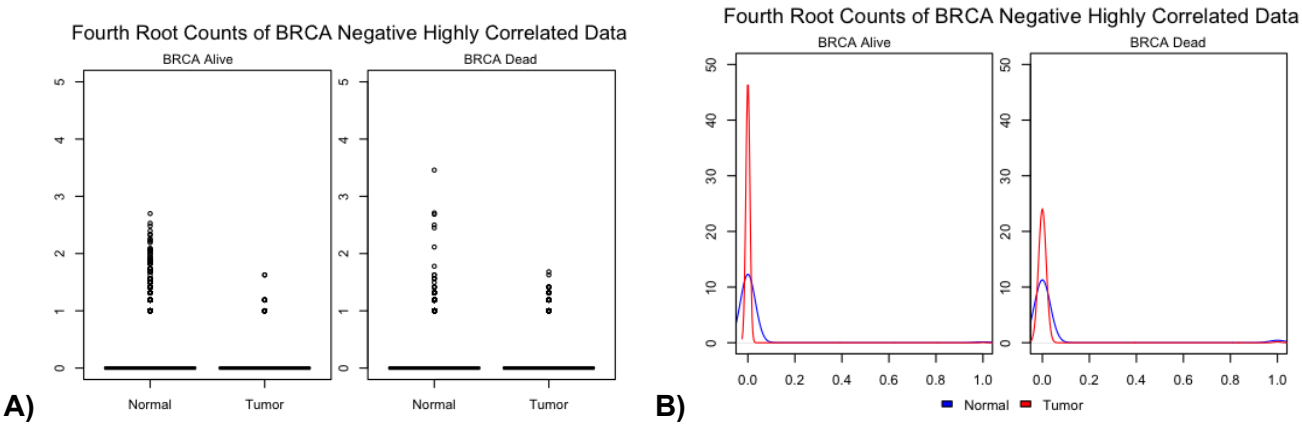


Figure 8: **A)** Boxplots showing the fourth root of negative high correlation counts between BRCA alive and dead sample groups. **B)** Density plots showing the distribution of negative high correlation counts between BRCA alive and dead samples.

Madison
- A) File: /home/vbz21/math5376D/final.project/Plots/BRCA.UCEC.negative.highcorr.boxplots.pdf
- B) File: /home/vbz21/math5376D/final.project/Plots/BRCA.UCEC.negative.highcorr.density.pdf

Code: /home/vbz21/math5376D/final.project/Code/Pos.Neg.Corrs.R

Bridget
- A) File: /home/bmb191/math5376D/Final.Project/Output.Files/Plots/UCEC.negative.highly.correlated.counts.boxplot.pdf
- B) File: /home/bmb191/math5376D/Final.Project/Output.Files/Plots/UCEC.negative.highly.correlated.counts.density.plot.pdf

Code: /home/bmb191/math5376D/Final.Project/Scripts/generate.final.plots.R

UCEC Positive

The plots below (Figure 9) reveal a noticeable difference between the alive and deceased datasets. However, it is important to interpret these findings cautiously due to the small sample size of the deceased group (n=7). In the alive group, the tumor data exhibits a noticeably larger range and third quartile compared to the normal group. For the deceased group, the counts are centered at significantly higher values, with similar central tendencies for both tumor and normal groups. Nonetheless, consistent with the alive group, the range is larger in the tumor group than in the normal group.

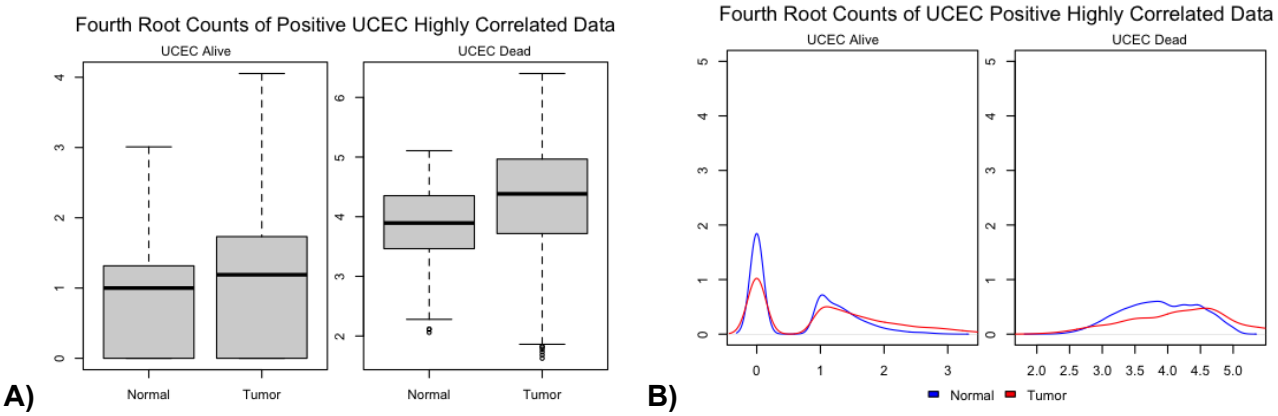


Figure 9. **A)** Boxplots showing the fourth root of positive high correlation counts between UCEC alive and dead sample groups. **B)** Density plots showing the distribution of positive high correlation counts between UCEC alive and dead samples.

Madison
- A) File: /home/vbz21/math5376D/final.project/Plots/BRCA.UCEC.positive.highcorr.boxplots.pdf
- B) File: /home/vbz21/math5376D/final.project/Plots/BRCA.UCEC.positive.highcorr.density.pdf

Code: /home/vbz21/math5376D/final.project/Code/Pos.Neg.Corrs.R

Bridget
- A) File: /home/bmb191/math5376D/Final.Project/Output.Files/Plots/UCEC.positive.highly.correlated.counts.boxplot.pdf
- B) File: /home/bmb191/math5376D/Final.Project/Output.Files/Plots/UCEC.positive.highly.correlated.counts.density.plot.pdf

Code: /home/bmb191/math5376D/Final.Project/Scripts/generate.final.plots.R

UCEC Negative

The alive group closely resembles the overall BRCA data, with a significant cluster around 0. However, the normal group exhibits a higher density at 0 compared to the tumor group. In the deceased group, the first through third quartiles are relatively similar between the tumor and normal groups, but the tumor group demonstrates a larger range (figure 10A). As evident from both density plots, the distributions show no notable differences apart from the slightly higher concentration at 0 observed in the alive normal group (figure 10B)

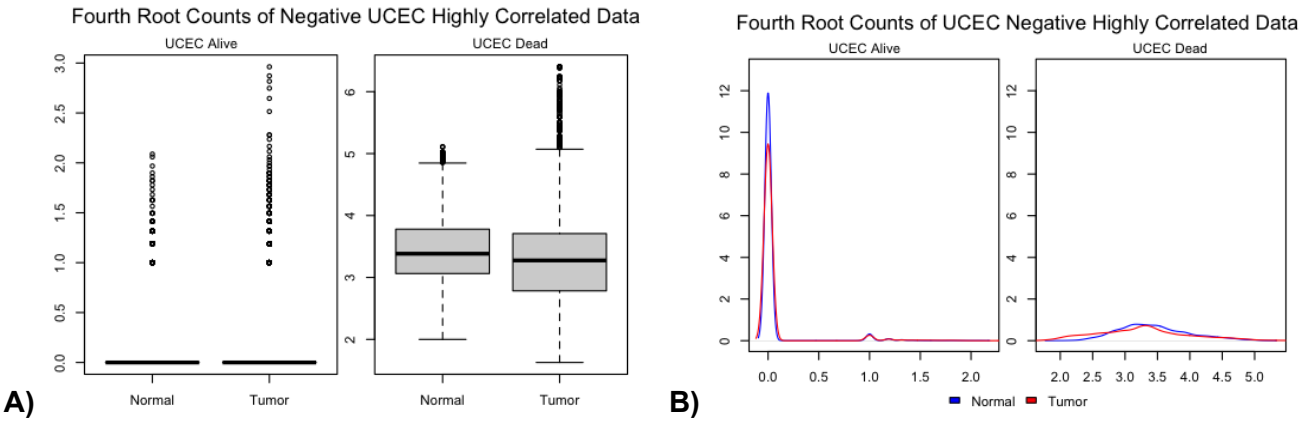


Figure 10. **A)** Boxplots showing the fourth root of negative high correlation counts between UCEC alive and dead sample groups. **B)** Density plots showing the distribution of negative high correlation counts between UCEC alive and dead samples.

Madison

A) File: /home/vbz21/math5376D/final.project/Plots/BRCA.UCEC.negative.highcorr.boxplots.pdf

B) File: /home/vbz21/math5376D/final.project/Plots/BRCA.UCEC.negative.highcorr.density.pdf

Code: /home/vbz21/math5376D/final.project/Code/Pos.Neg.Corrs.R

Bridget

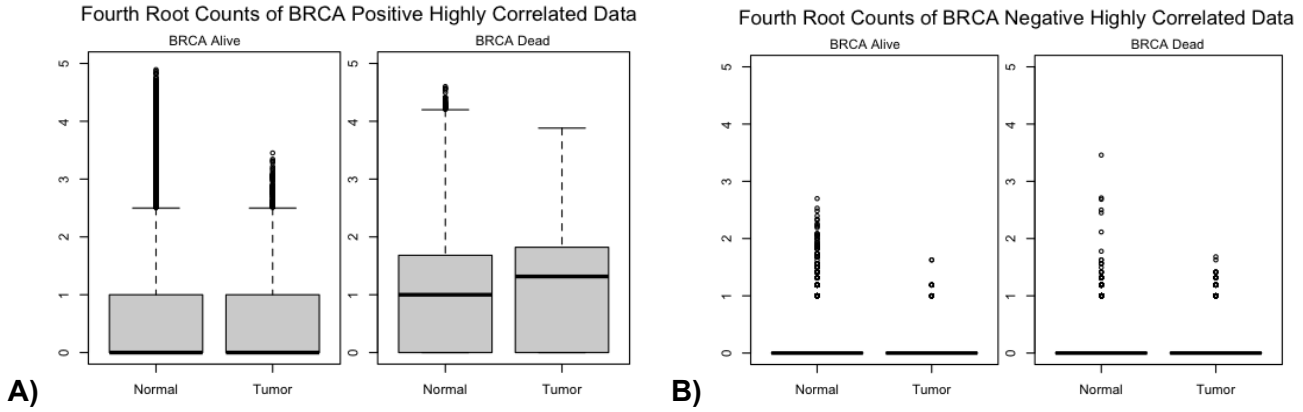
A) File: /home/bmb191/math5376D/Final.Project/Output.Files/Plots/UCEC.negative.highly.correlated.counts.boxplot.pdf

B) File: /home/bmb191/math5376D/Final.Project/Output.Files/Plots/UCEC.negative.highly.correlated.counts.density.plot.pdf

Code: /home/bmb191/math5376D/Final.Project/Scripts/generate.final.plots.R

Positive vs Negative

When we compare positive and negative correlations both to each other and across samples, differences become more apparent. There are more high positive correlations than high negative correlations for both BRCA and UCEC (figure 11). The mean number of negative correlations was 0 for BRCA alive normal and tumor, BRCA dead normal and tumor, and UCEC alive normal and tumor (figure 11B, D). Only in the UCEC dead was there a significant number of highly negatively correlated CpG sites (figure 11D).



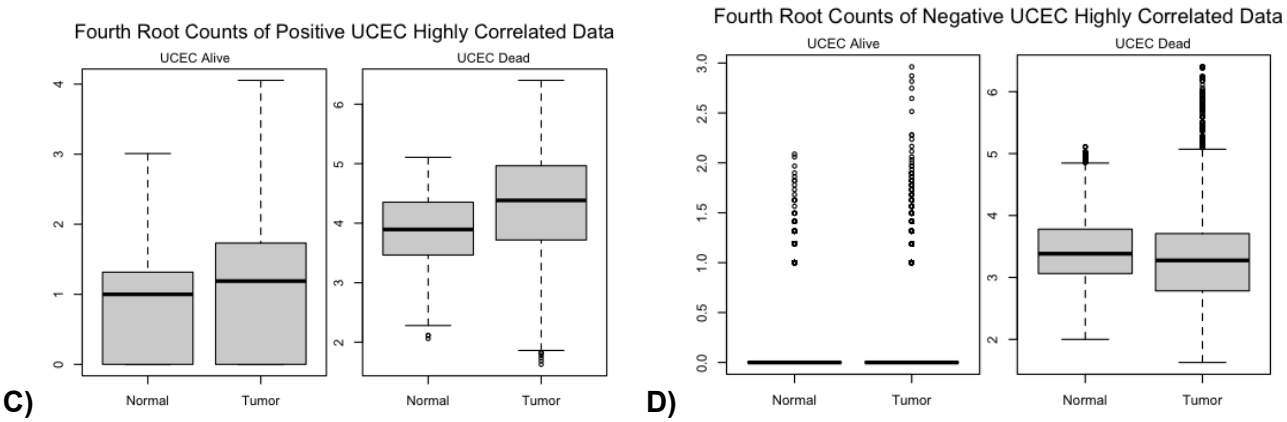


Figure 11. Fourth root of positive and negative high correlation counts. **A)** Positive high correlation counts between BRCA alive and dead sample groups **B)** Negative high correlation counts between BRCA alive and dead sample groups **C)** Positive high correlation counts between UCEC alive and dead sample groups. **D)** Negative high correlation counts between UCEC alive and dead sample groups.

Comparing Distances Between Highly Correlated CpG Sites

Comparing the pairwise distances of highly correlated CpG sites provides insights to their spatial and functional relationships within the genome. Highly correlated sites may share similar regulatory processes or epigenetic modifications, and analyzing their distances reveals whether they are located near each other or spread across the genome.

BRCA Absolute

The alive and deceased data appear to have similar overall distributions (Figure 12A). However, the tumor alive group exhibits a more uniform density distribution compared to the other groups with a larger range of central tendencies. In contrast, the other datasets show a steady increase in density to a sharp spike near the highest values in the density plot, followed by a rapid decline. Interestingly, both tumor groups display a spike around the smallest values, suggesting the distances between the CpG sites are smaller (Figure 12B).

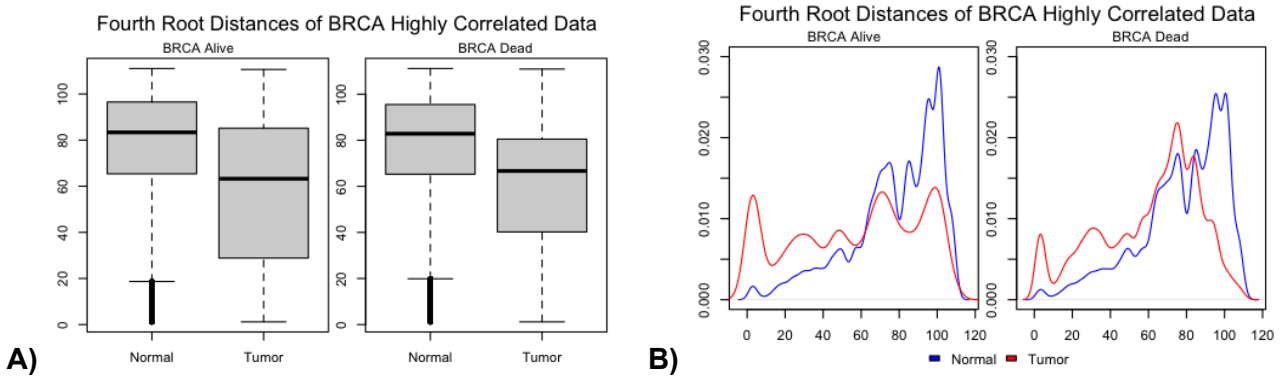


Figure 12. **A)** Boxplots showing the distribution of distances (in base pairs) between all highly correlated CpG sites in BRCA data. **B)** Density plots showing the distribution of distances between all highly correlated CpG sites in BRCA data.

Madison

A) File: /home/vbz21/math5376D/final.project/Plots/BRCA.distance.boxplots.pdf

B) File: /home/vbz21/math5376D/final.project/Plots/BRCA.density.distances.pdf

Code: /home/vbz21/math5376D/final.project/Code/11.25.distance.plots.R

Bridget

A) File: /home/bmb191/math5376D/Final.Project/Output.Files/Plots/distances.boxplots.pdf

B) File: /home/bmb191/math5376D/Final.Project/Output.Files/Plots/distances.density.plots.pdf

Code: /home/bmb191/math5376D/Final.Project/Scripts/generate.final.plots.R

UCEC Absolute

The distributions across the different groups for the UCEC data are almost identical, with the exception of a small spike observed at the smallest and largest values for the tumor distances in the alive group compared to the normal distances (Figure 12).

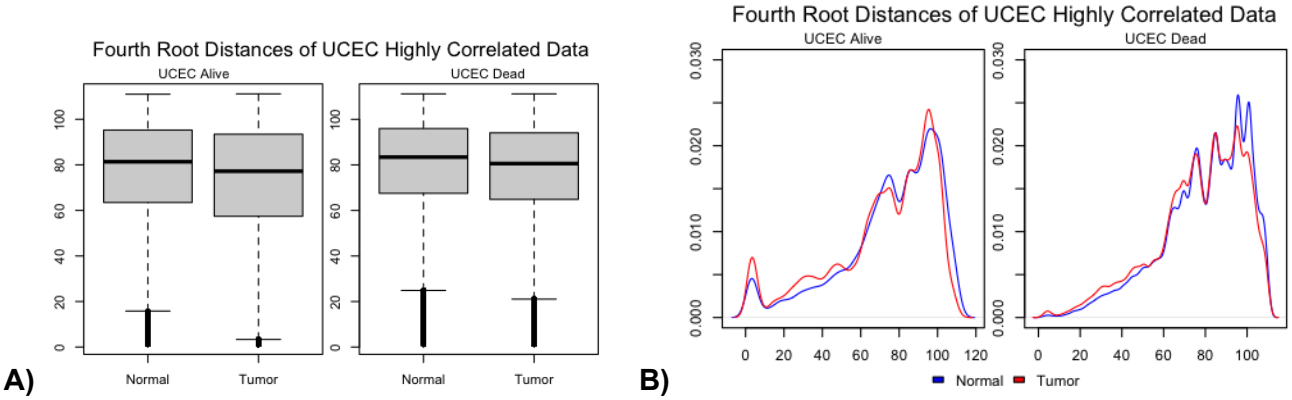


Figure 13. **A)** Boxplots showing the distribution of distances (in base pairs) between all highly correlated CpG sites in UCEC data. **B)** Density plots showing the distribution of distances between all highly correlated CpG sites in UCEC data.

Madison

A) File: /home/vbz21/math5376D/final.project/Plots/UCEC.distance.boxplots.pdf

B) File: /home/vbz21/math5376D/final.project/Plots/UCEC.density.distances.pdf

Code: /home/vbz21/math5376D/final.project/Code/11.25.distance.plots.R

Bridget

A) File: /home/bmb191/math5376D/Final.Project/Output.Files/Plots/distances.boxplots.pdf

B) File: /home/bmb191/math5376D/Final.Project/Output.Files/Plots/distances.density.plots.pdf

Code: /home/bmb191/math5376D/Final.Project/Scripts/generate.final.plots.R

BRCA Positive

The distributions for absolute and positive distances in the BRCA dataset are essentially the same (Figure 13A, 14A), with the tumor alive group exhibiting a more uniform density distribution compared to the other groups with a larger range of central tendencies. The other samples show a steady increase in density to a sharp spike near the highest values in the density plot, followed by a rapid decline. Again, both tumor groups display a spike around the smallest values, suggesting the distances between the CpG sites are smaller (Figure 14B).

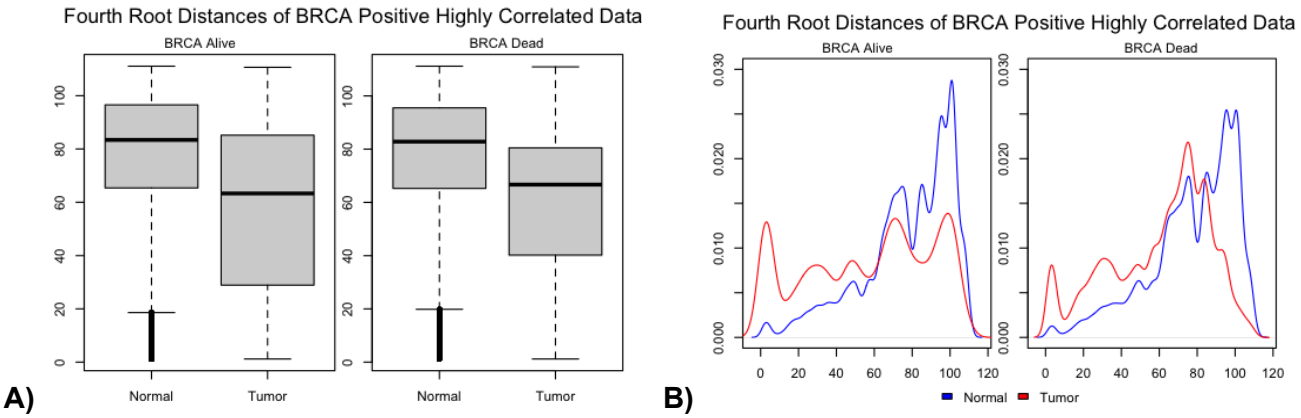


Figure 14. **A)** Boxplots showing the distribution of distances (in base pairs) between positive highly correlated CpG sites in BRCA data. **B)** Density plots showing the distribution of distances between positive highly correlated CpG sites in BRCA data.

Madison

A) File: /home/vbz21/math5376D/final.project/Plots/BRCA.distance.boxplots.pdf

B) File: /home/vbz21/math5376D/final.project/Plots/BRCA.density.distances.pdf

Code: /home/vbz21/math5376D/final.project/Code/11.25.distance.plots.R

Bridget

A) File: /home/bmb191/math5376D/Final.Project/Output.Files/Plots/distances.boxplots.pdf

B) File: /home/bmb191/math5376D/Final.Project/Output.Files/Plots/distances.density.plots.pdf

Code: /home/bmb191/math5376D/Final.Project/Scripts/generate.final.plots.R

UCEC Positive

Like what was seen in the BRCA data, analysis of the strictly positive high correlation for the UCEC distances yields the same conclusion as the UCEC absolute high correlation distances (Figure 12A, 13B, 15A, 14B), showing comparable patterns across the groups with no significant deviations.

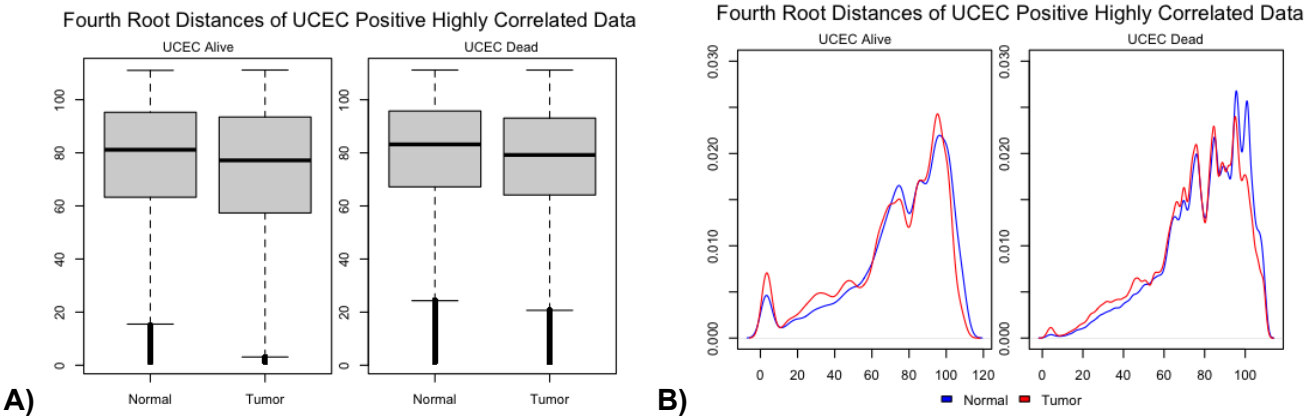


Figure 15. **A)** Boxplots showing the distribution of distances (in base pairs) between positive highly correlated CpG sites in UCEC data. **B)** Density plots showing the distribution of distances between positive highly correlated CpG sites in UCEC data.

Madison

A) File: /home/vbz21/math5376D/final.project/Plots/UCEC.distance.boxplots.pdf

B) File: /home/vbz21/math5376D/final.project/Plots/UCEC.density.distances.pdf

Code: /home/vbz21/math5376D/final.project/Code/11.25.distance.plots.R

Bridget

A) File: /home/bmb191/math5376D/Final.Project/Output.Files/Plots/distances.boxplots.pdf
B) File: /home/bmb191/math5376D/Final.Project/Output.Files/Plots/distances.density.plots.pdf
Code: /home/bmb191/math5376D/Final.Project/Scripts/generate.final.plots.R

BRCA Negative

The alive tumor group exhibits significantly smaller distances, with a central tendency around very small values (figure 16A, B). In contrast, the alive normal group, as well as the dead normal and tumor groups, display relatively similar distributions. However, the tumor dead group shows a slightly lower central tendency and overall distribution, consistent with other observed trends.

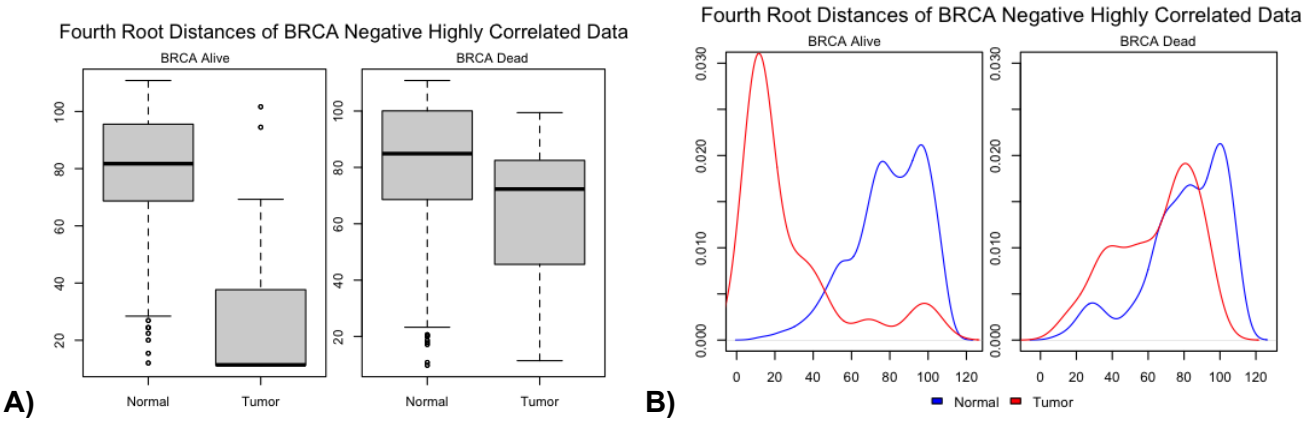


Figure 16. **A)** Boxplots showing the distribution of distances (in base pairs) between negative highly correlated CpG sites in BRCA data. **B)** Density plots showing the distribution of distances between negative highly correlated CpG sites in BRCA data.

Madison
C) File: /home/vbz21/math5376D/final.project/Plots/BRCA.distance.boxplots.pdf
D) File: /home/vbz21/math5376D/final.project/Plots/BRCA.density.distances.pdf
Code: /home/vbz21/math5376D/final.project/Code/11.25.distance.plots.R
Bridget
C) File: /home/bmb191/math5376D/Final.Project/Output.Files/Plots/distances.boxplots.pdf
D) File: /home/bmb191/math5376D/Final.Project/Output.Files/Plots/distances.density.plots.pdf
Code: /home/bmb191/math5376D/Final.Project/Scripts/generate.final.plots.R

UCEC Negative

The analysis of the strictly negative high correlation distances yields the same conclusion as the absolute and positive high correlation distances (figure 12, 15, 17), showing comparable patterns across the groups with no significant deviations.

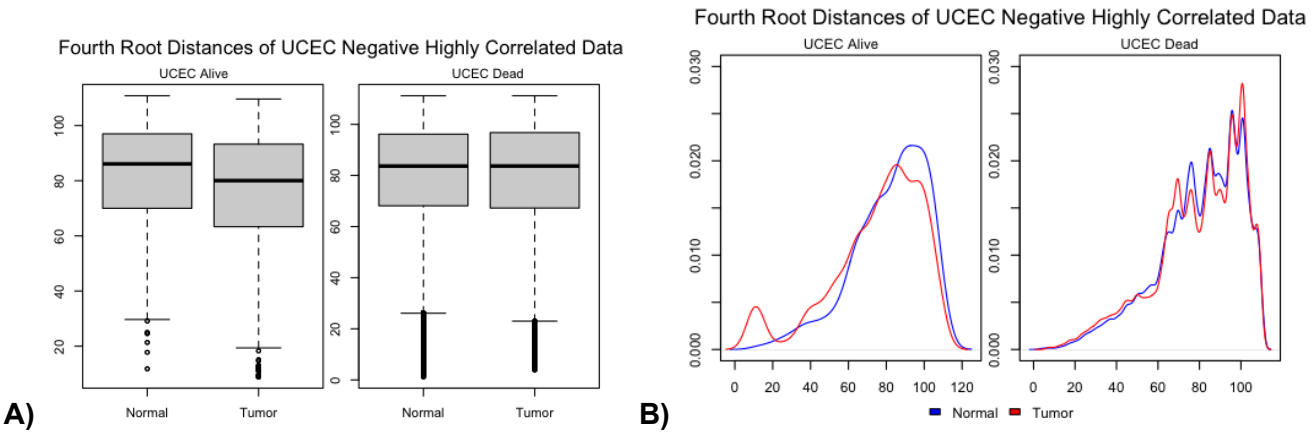


Figure 17. **A)** Boxplots showing the distribution of distances (in base pairs) between negative highly correlated CpG sites in UCEC data. **B)** Density plots showing the distribution of distances between negative highly correlated CpG sites in UCEC data.

Madison

E) File: /home/vbz21/math5376D/final.project/Plots/UCEC.distance.boxplots.pdf

F) File: /home/vbz21/math5376D/final.project/Plots/UCEC.density.distances.pdf

Code: /home/vbz21/math5376D/final.project/Code/11.25.distance.plots.R

Bridget

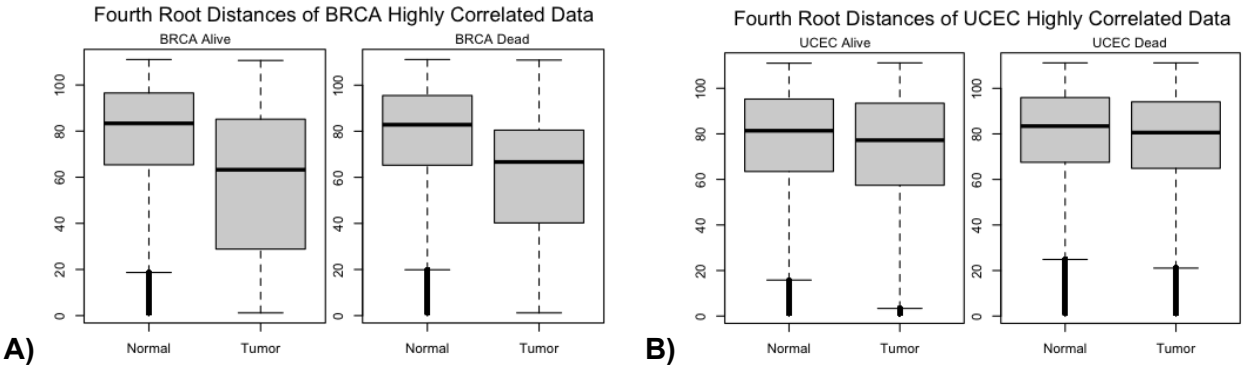
E) File: /home/bmb191/math5376D/Final.Project/Output.Files/Plots/distances.boxplots.pdf

F) File: /home/bmb191/math5376D/Final.Project/Output.Files/Plots/distances.density.plots.pdf

Code: /home/bmb191/math5376D/Final.Project/Scripts/generate.final.plots.R

All Sample Comparison

When comparing all samples it becomes apparent that BRCA data shows a stronger pattern of tumor samples having shorter distances between highly correlated CpG sites than seen in the normal samples (Figure 18). Normal and tumor distances in the UCEC data appear quite similar (figure 18B, D, F)



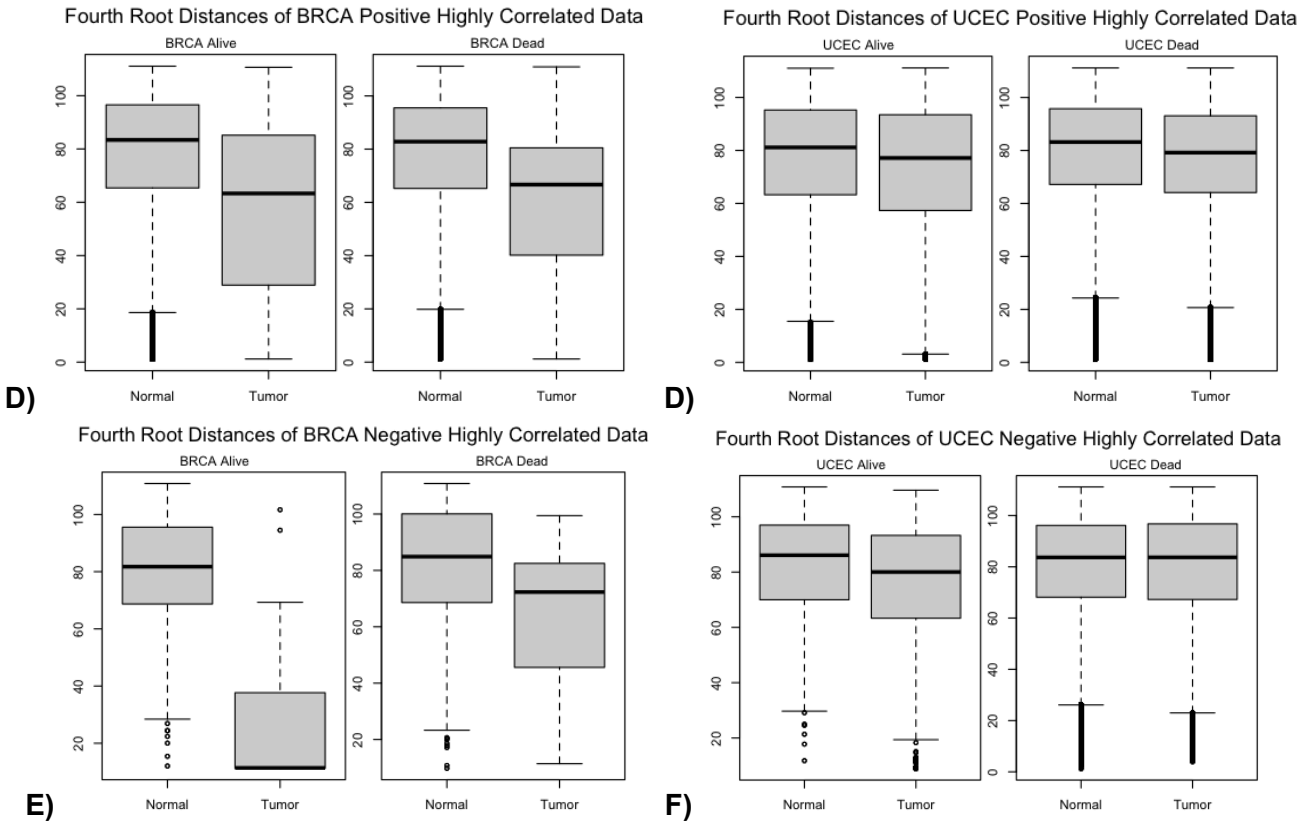


Figure 18. Boxplots of distances between absolute, positive, and negative highly correlated CpG sites in all samples. **A)** BRCA absolute high correlations. **B)** UCEC high correlations. **C)** BRCA high positive correlations **D)** UCEC high positive correlations. **E)** BRCA high negative correlations **F)** UCEC high negative correlations.

Discussion

Our study analyses CpG site level analyses of co-methylation on the X chromosome for breast and endometrial cancer using matched alive and dead, normal and tumor samples. To investigate patterns of methylation more precisely, we looked at the total number of high correlations per sample, the patterns of high positive versus high negative correlations, and the distance between these highly correlated CpG sites. These analyses have revealed important insights into the epigenetics of breast cancer (BRCA).

There are some interesting findings from our data, most notably that breast and endometrial cancer appear to co-methylate differently when it comes to the X chromosome. Both the UCEC alive and dead samples had more overall highly correlating CG sites than seen in the BRCA data (figure 3A). These differences in co-methylation were not unexpected, as methylation patterns are known to vary significantly across different tissues (Burris & Baccarelli 2013). However, there was a disparity in sample sizes between the BRCA and UCEC samples, with UCEC samples being much smaller than those for the BRCA data (53 alive BRCA, 26 alive UCEC, 32 dead BRCA, 7 dead UCEC).

In terms of the matched normal and tumor samples, tumor samples exhibited increased variability and disrupted co-methylation patterns compared to normal tissue. In contrast, normal tissue showed more stable methylation patterns, which aligns with known characteristics of healthy cells.

With regards to negative and positive high correlation counts, the analysis revealed notable co-methylation patterns between CpG sites in tumor and normal samples, as well as between living and deceased groups. Deceased samples show higher central values and increased variability, potentially reflecting advanced disease stages or post-mortem effects. However, no definitive conclusions can be drawn from the deceased data, as the cause of death is unknown for these samples.

Another significant observation is that the negatively correlated BRCA distances are much closer together, indicating a stronger consistency in the negative correlation patterns across tumor samples. These findings suggest that the alive tumor group may exhibit a more uniform or constrained methylation pattern, potentially reflecting distinct epigenetic regulation in tumor tissue. The lower central tendency in the tumor dead group could imply a further alteration in methylation patterns associated with disease progression. Overall, these differences in distribution highlight the dynamic nature of methylation alterations in both living and deceased cancer samples.

One limitation of this study is the lack of demographic information available for these samples. As such, we are not able to control for the effects of age, sex (for BRCA only), ethnicity, cancer stage, or other lifestyle factors. As previous studies have shown that such environmental or population based factors have the potential to impact DNA methylation (Burris & Baccarelli 2013).

While this analysis provides insights into the epigenetic alterations in breast and endometrial cancer, the findings must be interpreted with caution due to the limitations listed above. Further research with larger, more diverse samples, additional tissue types, and detailed demographic information will be crucial for a more comprehensive understanding of the role of methylation in cancer.

Conclusion

In this study we analyzed co-methylation on the X chromosome in breast and endometrial cancer. From these analyses we find that breast and endometrial cancer co-methylate differently. In breast cancer, dead tumor samples show a higher mean number of highly correlating CpG sites than the dead normal sample (figure 1A). In endometrial cancer, tumor samples show a higher mean number of highly correlated CpG sites in both the alive and dead groups.

Most high correlations between CpG sites are positive correlations. When looking at distances between highly correlated CpG sites in breast cancer, tumor samples have shorter distances

than normal samples suggesting that the highly correlated CpG sites in tumor samples are located more closely together. In endometrial cancer, this pattern is very weak and distances between highly correlated CpG sites for both normal and tumor samples are very similar, though tumor may be slightly lower.

Our findings support the concept that breast cancer is associated with DNA co-methylation and that tumor progression may be linked to further disruption in co-methylation patterns.

Future research should aim to extend these findings by comparing other chromosomal regions, such as chromosome 22, to assess whether similar patterns of negative correlation are present. Such investigations could offer more comprehensive insights into the role of epigenetic modifications in cancer and potentially uncover biomarkers or therapeutic targets for treatment.

References

Burris, H. H., & Baccarelli, A. A. (2014). Environmental epigenetics: from novelty to scientific discipline. Journal of Applied Toxicology, 34(2), 113-116.

Kang, J., Lee, H. J., Jun, S.-Y., Park, E. S., & Maeng, L. (2015). Cancer-Testis Antigen Expression in Serous Endometrial Cancer with Loss of X Chromosome Inactivation. *PLOS ONE*, *10*(9), e0137476. https://doi.org/10.1371/journal.pone.0137476

Men, C., Chai, H., Song, X., Li, Y., Du, H., & Ren, Q. (2017). Identification of DNA methylation associated gene signatures in endometrial cancer via integrated analysis of DNA methylation and gene expression systematically. *Journal of Gynecologic Oncology*, *28*(6), e83. https://doi.org/10.3802/jgo.2017.28.e83

Sun, L., & Sun, S. (2019). Within-sample co-methylation patterns in normal tissues. *BioData Mining*, *12*(1), 9. https://doi.org/10.1186/s13040-019-0198-8

Sun, S., Dammann, J., Lai, P., & Tian, C. (2022). Thorough statistical analyses of breast cancer co-methylation patterns. *BMC Genomic Data*, *23*(1), 29. https://doi.org/10.1186/s12863-022-01046-w

Zhang, J., & Huang, K. (2017). Pan-cancer analysis of frequent DNA co-methylation patterns reveals consistent epigenetic landscape changes in multiple cancers. *BMC Genomics*, *18*(1), 1045. https://doi.org/10.1186/s12864-016-3259-0