An Analysis of DNA Methylation Data in the TCGA-BRCA Study Using the Illumina Infinium HumanMethylation27 BeadArray

Joe LaRocca
Columbia University
P8109: Statistical Genetic Modeling
Professor Shuang Wang
May 2, 2025

Introduction

Created in 2005, less than five years after the complete sequencing of the human genome, The Cancer Genome Atlas (TCGA) is a comprehensive database of multi-omic information synthesized for the purpose of studying the molecular-level biological mechanisms behind dozens of types of cancer. The Cancer Genome Atlas Breast Invasive Carcinoma (TCGA-BRCA) study focuses on breast invasive carcinoma, a subtype of breast cancer in which the cancer has spread from the breast ducts to the surrounding breast tissue. The study collected a wide variety of data on a total of 1,098 individuals, including but not limited to clinical data (i.e. patients' age, sex, and medical history), genomic data, transcriptomic data, and DNA methylation data.

Our scope is limited to DNA methylation data. DNA is methylated in CpG regions, or sections of DNA where a cytosine nucleotide is directly followed by a guanine nucleotide (the methyl group involved in methylation is attached directly to the cytosine nucleotide). CpG sites are often found in the promoter regions of genes in large clusters known as "CpG islands". One common way of quantifying DNA methylation across the genome is by using the beta value, which ranges from 0 to 1. A beta value of 0 represents no methylated DNA, while a beta value of 1 represents fully methylated DNA. Usually, CpG sites in promoter regions with lower beta values are associated with higher gene expression, and vice versa.

To conduct our analysis, we used the Illumina Infinium HumanMethylation27 BeadArray, which contains beta values for over 27,000 CpG sites – nearly 75% of which are on CpG islands – spanning over 14,000 genes. Although there were 343 samples for which DNA methylation data was available, there were only 27 patients for which both primary tumor tissue and normal tissue data was available. We therefore used a paired case-control design, in which each of the 27 participants had both a tumor sample and a normal tissue sample. All 27 of the tumor samples were taken from primary tumors (as opposed to metastatic tumors). All 27 of the participants were women, and 25 of them were White (the other 2 were Black). The women in our analysis were initially diagnosed with breast cancer between 1999 and 2010. The patients' age ranged from 35 to 88, with a mean of 54.1, a median of 50, and a middle 50% range of (43, 64). 19 and 21 of the women had progesterone- and estrogen-positive cancers, respectively.

Many past studies have examined the link between DNA methylation and breast cancer. Patients with breast cancer have been observed to have regional hypermethylation and global hypomethylation – that is, more methylation in specific genes that may be associated with cancer, but less methylation elsewhere.[1] More recently, the literature has shifted from simply identifying abnormal methylation patterns to identifying specific genes or groups of genes associated with breast cancer and its survival rate.[2] Some studies have isolated genes of interest, such as the BRCA-1 and BRCA-2 genes, for which hypermethylation has been linked with both downregulation and breast cancer.[3,4]

The aim of our project was to determine whether CpG sites have different methylation patterns in samples taken from tumor tissues compared to adjacent normal tissues, and to identify the CpG site(s) most significantly associated with breast cancer. We additionally sought to make conclusions about whether the most important CpG sites have similar location patterns – namely, whether they are on CpG islands, are found on the same chromosome, or are connected via a biological pathway.

Methods

Data Preparation

All of our data, including patient characteristics and CpG site beta values, came from the TCGA database. We obtained our data using the Bioconductor package in R, selecting only the patients for which 1) DNA methylation data was available and 2) data was present for both tumor tissue and normal tissue, and connected their clinical and DNA methylation characteristics using their unique patient barcodes. While the Illumina Human Methylation platform contained information on a total of 27,578 CpG sites, about 15% of the CpG sites had at least some missing data. Due to concerns about the impact of missing data on our analysis, we decided to keep only the 23,120 CpG sites for which beta values were available for all data. We also included several important covariates that we thought could be strongly associated with tissue type: age, race, the type of surgical procedure done (if any), progesterone receptor status, and estrogen receptor status.

For our regression analysis, we transformed the beta values into M-values, where $M = \log_2(\beta/1 - \beta)$, for two key reasons: first, to stabilize variances, especially near the extremes (i.e. $\beta = 0$ or $\beta = 1$), and second, to make the predictor variables more Normally distributed, as the assumptions of regression are best satisfied when explanatory variables approximately follow a Normal distribution. For some visuals, such as Figure 1, we decided to present beta values for the sake of interpretability, while for others, such as Figure 2, we used M-values in order to explain other key characteristics of the visual (such as the biological significance of $\log_2$ fold change).

Exploratory Data Analysis

Before conducting our full analysis, we analyzed the distributions of beta values across all CpG sites for both tumor and normal tissues. We also created a volcano plot that examined whether, and to what degree, certain CpG sites were either hypomethylated or hypermethylated in tumor tissue compared to normal tissue, and classified points that had both a false discovery rate-adjusted P-value less than 0.05 and a $\log_2$ fold change either less than -2 (corresponding to hypomethylation) or greater than 2 (corresponding to hypermethylation). Though different $\log_2$ fold change thresholds are often used by bioinformaticians, we chose a cutoff of 2 (among the highest cutoffs frequently found in the literature) in order to isolate a small fraction of the CpG sites used for our analysis.[5] Finally, we used principal component analysis in order to determine whether there were clear separations between tumor tissue and normal tissue based on the first two principal components.

Regression Analysis

Since there were over 23,000 CpG sites for which data were available, we decided against using ordinary least squares regression, since the results would be difficult to interpret and the model would be at severe risk of overfitting. As a result, we opted to use elastic net regression, which combines the shrinkage effect of Ridge regression and the selection effect of LASSO regression to create a model that only uses CpG sites for which a strong association is present and shrinks coefficients that are included in the model but are not particularly important. Additionally, since our primary response variable (tissue type) was binary, we decided to use logistic regression instead of linear regression.

For our elastic net model, we selected $\alpha = 0.5$ to achieve a balance between the effects of Ridge and LASSO regression. We then selected the regularization parameter λ using cross-validation and used the λ_{1SE} coefficient instead of the λ_{min} coefficient for our final model since we aimed for sparsity (i.e. a model in which more predictors are removed through L1-regularization). We then analyzed the significant CpG sites from our regression analysis and determined whether there were any clear patterns in the location of these CpG sites within the human genome.

Results

Exploratory Data Analysis

We found that the distributions of beta values from tumor tissue and normal tissue among all 27 patients throughout all CpG sites were remarkably similar (Figure 1). Both distributions were heavily right-skewed, with a mode beta value of approximately 0.05 and a small increase in frequency for beta values between 0.9 and 0.95.

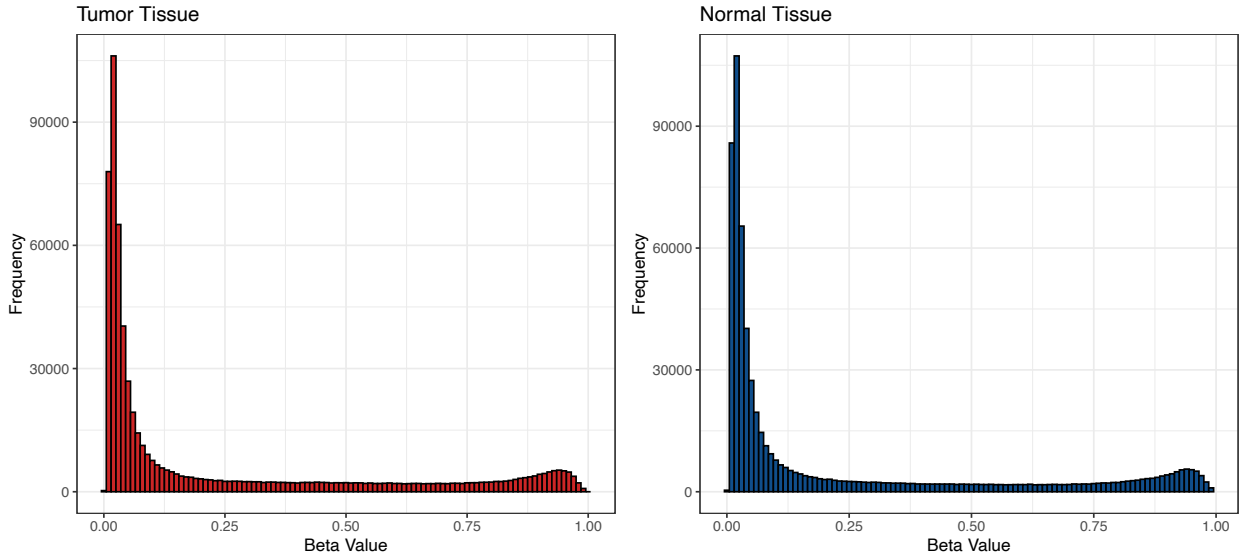

Figure 1: Histograms of Beta Values for Tumor Tissue and Normal Tissue. Methylation beta values are calculated for the 24,342 of the 27,578 CpG sites for which data was available; n = 27 subjects for both tumor tissue and normal tissue.

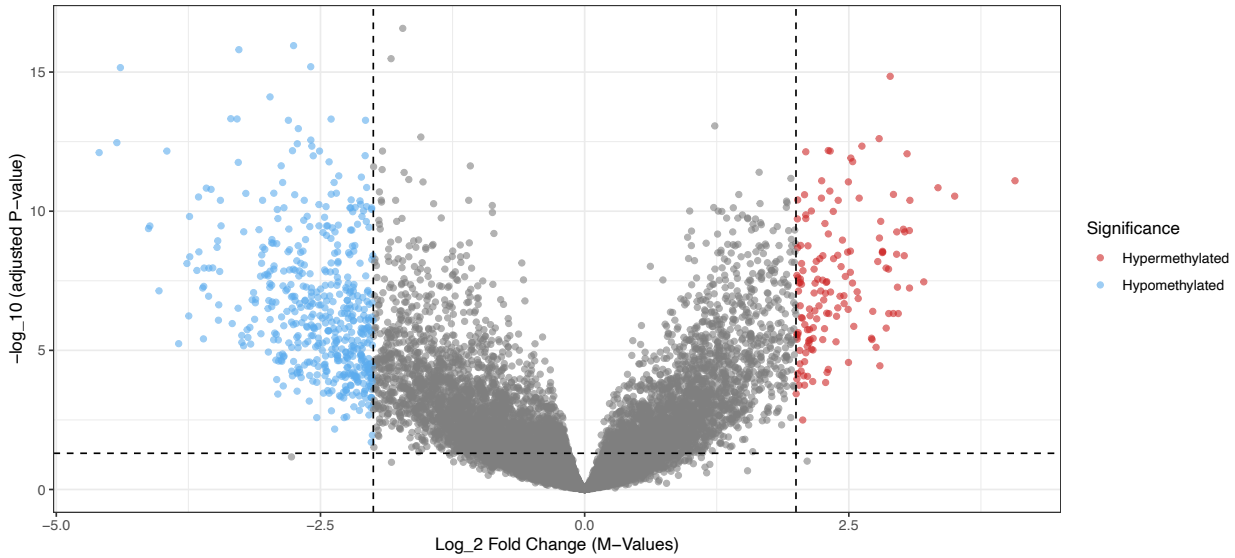

Figure 2: Volcano Plot of Log_{10} P-Value Against Log_2 Fold Change. M-values are transformed beta values, in which $M = \log_2(\beta/1 - \beta)$. Points that have both an FDR-adjusted P-value < 0.05 and Log_2 Fold Change with magnitude > 2 are considered significant. Blue points represent CpG sites hypomethylated in tumor tissue, while red points represent CpG sites hypermethylated in tumor tissue.

From the volcano plot, we can see that while most CpG sites did not have significant under- or over-methylation in cancer tissue, there were 503 CpG sites with a $\log_2$ fold change value under -2, indicating hypomethylation, and 146 CpG sites with a $\log_2$ fold change value over 2, indicating hypermethylation (Figure 2). Specifically, given the base 2 log structure, a $\log_2$ fold change of +2 represents $2^2 = 4$ times as much methylation in tumor tissue, while a $\log_2$ fold change of -2 represents $2^{-2} = ¼$ as much methylation in tumor tissue (i.e. 4 times as much methylation in normal tissue). We chose a threshold Our results are potentially consistent with the observation made by Szyf et al. (2004) that while some regions may be hypermethylated in tumor tissue, global hypomethylation can also occur (we would need more information about the specific location of the CpG sites in each category to confirm this observation).

Our principal component plot shows a clear separation between tumor tissue and normal tissue (Figure 3). The first two principal components explain 13.5% and 8.9% of the variance, respectively. Normal tissue had lower values for PC1 than tumor tissue, and the variance for tumor tissue was much larger than that of normal tissue, indicating potential differences in the DNA methylation structure of tumors between the 27 patients in the study.

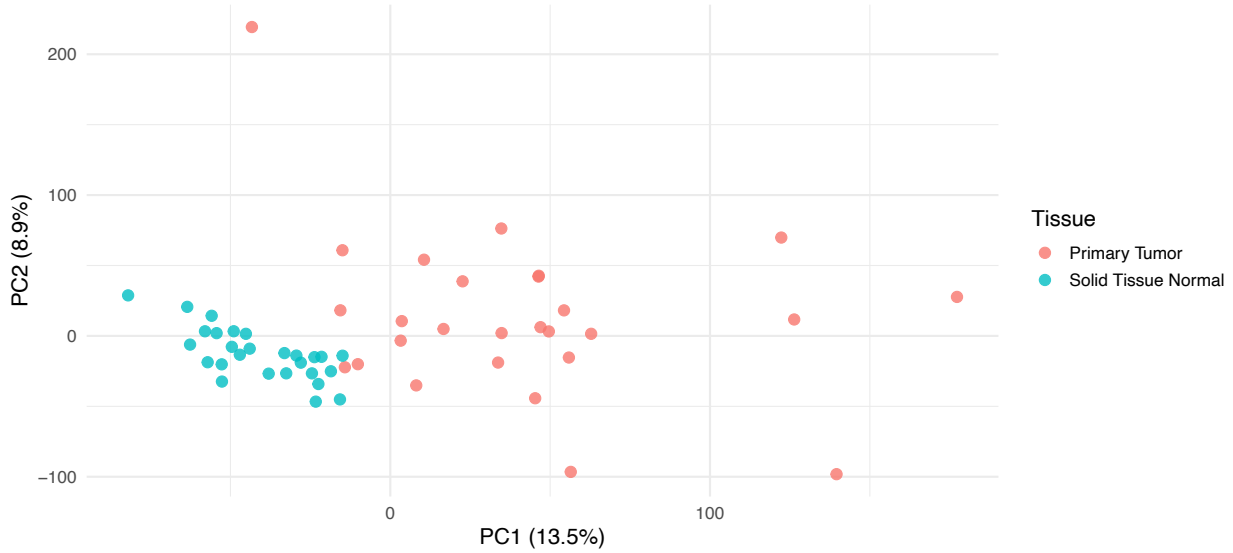

Figure 3: PC1 vs. PC2 plot. Principal components were created using all predictor variables, including CpG sites and relevant covariates (age, race, surgical procedure, estrogen/progesterone receptor status).

Regression Analysis

We chose a regularized logistic regression model with $\alpha = 0.5$, using the M-values for each of the CpG sites as well as a selected group of covariates (age, race, surgical procedure, and progesterone/estrogen receptor status) as predictor variables. Through cross-validation, we found the value of λ_{1SE} to be approximately 0.0304, with a standard error of about 0.0533. Out of the 23,120 possible CpG sites, our final elastic net model kept 109 CpG sites as predictors (using an alternative model chosen by $\lambda = \lambda_{min}$ would have yielded 128 predictors instead). All of the predictors the model selected were CpG sites, meaning that when adjusting for CpG site, none of the relevant covariates were significantly associated with tissue type according to our model. Using a probability threshold of 0.5, the model correctly classified all 54 samples as tumor tissue or normal tissue; all of the probabilities for tumor tissue were greater than 0.9, while all of the probabilities for normal tissue were less than 0.1.

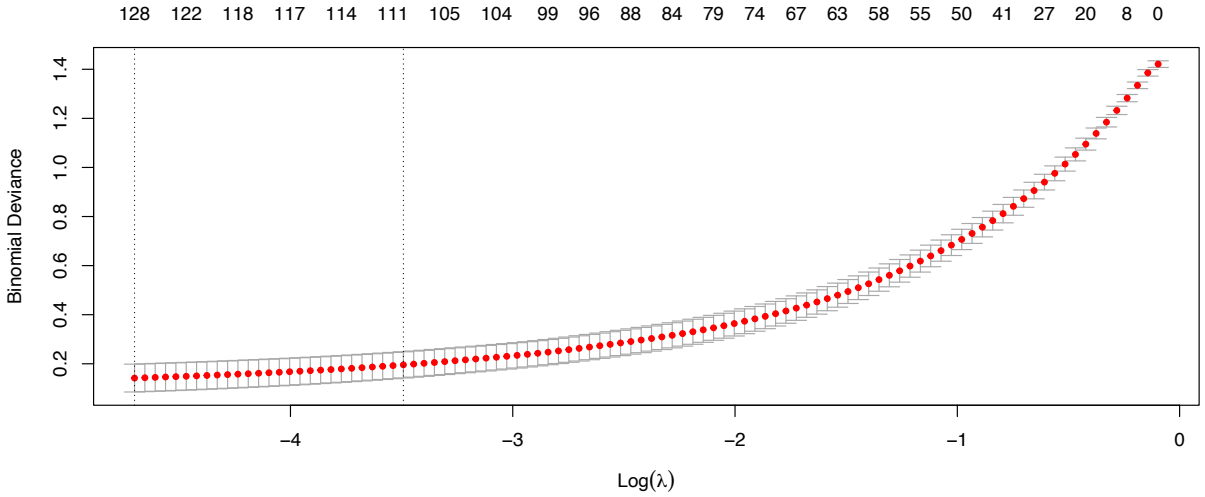

Figure 4: Plot of Optimal λ Values for Elastic Net Regression Model. Lower values of λ generally have lower residual deviance. For the purpose of sparsity, λ_{1SE} was used to fit the final model instead of λ_{min}.

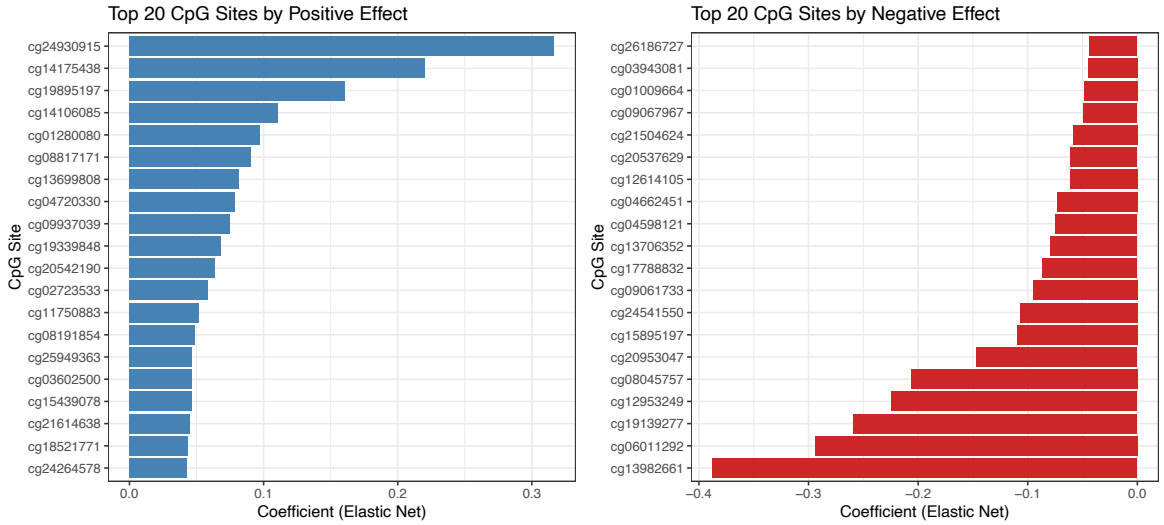

Figure 5: Plots of Top 20 CpG Sites by Positive and Negative Effects. Our final elastic net model had a total of 109 predictors.

We then ordered the coefficients by largest positive and largest negative effect size in order to see which CpG sites were most strongly associated with breast cancer (Figure 5). The cg24930915 site, which is on a CpG island on chromosome 16, had the largest positive effect size, while the cg13982661 site, which is located on a CpG island on chromosome 11, had the largest negative effect size. Given that both the CpG sites with the strongest positive and negative effect sizes were on CpG islands, it seemed plausible that a large portion of the CpG sites selected by our model would reside on CpG islands, therefore suggesting a link between methylation of CpG sites on CpG islands and breast cancer.

However, using the Bioconductor package's annotation feature, we were able to determine the genomic locations of the CpG sites selected by our regression model, and by examining the locations of the CpG sites that represented the top ten predictors by positive effect size, we found no clear association between chromosomal location, whether the CpG was located on a CpG island, and effect size (Table 1). Each of the top ten predictors by positive effect size was located on a different chromosome, six were on the "–" strand while four were on the "+" strand, and only three out of the ten were categorized as being on a CpG island. Three were on a "south shore", meaning less than 2 kilobases (kb) downstream of the nearest CpG island, two were on a "north shore", meaning less than 2 kb upstream of the nearest island, and the remaining two were in the "open sea", meaning more than 4 kilobases (kb) away from the nearest island.[6]

While there was a slightly more noticeable pattern for the top ten predictors by negative effect size, there still were no clear and conclusive associations. Three of these CpG sites were located on chromosome 11 and two were located on chromosome 8, with the other five being located on distinct chromosomes. Five CpG sites were on each of the "-" and "+" strands. Four of the CpG sites were located on CpG islands, one was located upstream of an island, and the remaining five were in the open sea.

Out of the 108 CpG sites selected by the model for which annotation data was available, 41 were located on CpG islands, followed by 32 in the open sea, 20 located on a south shore, and 12 on a north shore. The remaining three CpG sites were on the "shelves" of a CpG island, meaning that they were between 2 kb and 4 kb from the nearest CpG island.

Location	Overall	Top 10 by Positive Effect Size	Top 10 by Negative Effect Size
CpG Island	41	3	4
North Shore	12	2	1
South Shore	20	3	0
Island Shelf	3	0	0
Open Sea	32	2	5

Table 1. List of Locations of CpG Sites included in Elastic Net Regression Model (n = 108). "Shores" are within 2 kilobases (kb) of the nearest CpG island, "shelves" are between 2 kb and 4 kb from the nearest CpG island, and the "open sea" is more than 4 kb away from the nearest CpG island.

Conclusion

Findings

The objective of our analysis was to determine DNA methylation patterns between tumor tissue and normal tissue in a cohort of 27 patients from the TCGA-BRCA study, which used a paired case-control design that allowed us to compare the two types of tissue in the same individuals. From both our exploratory analysis and regression model, we found a clear association between DNA methylation and tissue type; while the volcano plot showed that there are hundreds of CpG sites that were either hypo- or hypermethylated in tumor tissue, the principal component plot showed that the first two principal components effectively separated tumor tissue and normal tissue. With respect to the associations found in the volcano plot, over 600 CpG sites had a $\log_2$ fold change score with a magnitude greater than 2, even after adjusting for multiple comparisons using false discovery rate.

Our regularized logistic regression model gave us key insight into which CpG sites were most strongly associated with breast cancer, but we did not see any clear patterns among the CpG sites selected in our model with respect to their location in the human genome. The model also eliminated any clinical covariates such as age and race, which makes sense given the paired format of our data.

Limitations and Next Steps

Even though we worked with over 20,000 CpG sites for our analysis, there are estimated to be between 28-30 *million* CpG sites in the human genome, meaning that we used less than 0.1% of human CpG sites for our analysis, limiting our scope. From a model fit standpoint, given that our elastic net model predicted all 54 tissue types correctly on the training data, future analysis could determine whether the model generalizes well to novel testing data, as the model could potentially be overfit. Since we did not have much information on interactions between CpG sites, such as close proximity on the same chromosome, location on the same CpG island, or relatedness in biological pathways related to cancer, we could conduct a revised analysis with these variables being incorporated into the final model. Additionally, we could use information on interactions between CpG sites to consolidate groups of CpG sites into sets, reducing the impact of multiple comparisons and therefore allowing us to identify more CpG sites that may have statistically significant associations with breast cancer.

References

1. **Szyf, Moshe, Pouya Pakneshan, and Shafaat A. Rabbani.** "DNA Methylation and Breast Cancer." *Biochemical Pharmacology* 68, no. 6 (September 15, 2004): 1187–1197. https://doi.org/10.1016/j.bcp.2004.04.030.
2. **Zhang, Ming, Yilin Wang, Yan Wang, Longyang Jiang, Xueping Li, Hua Gao, Minjie Wei, and Lin Zhao.** "Integrative Analysis of DNA Methylation and Gene Expression to Determine Specific Diagnostic Biomarkers and Prognostic Biomarkers of Breast Cancer." *Frontiers in Cell and Developmental Biology* 8 (December 6, 2020). https://doi.org/10.3389/fcell.2020.529386.
3. **Tang, Qiuqiong, Jie Cheng, Xue Cao, Harald Surowy, and Barbara Burwinkel.** "Blood-Based DNA Methylation as Biomarker for Breast Cancer: A Systematic Review." *Clinical Epigenetics* 8, no. 1 (2016): 115. https://doi.org/10.1186/s13148-016-0282-6.
4. **Tarapara, Bhoomi, and Franky Shah.** "BRCA1/2 Methylation and Expression Dynamics in Hereditary Breast and Ovarian Cancer: Insights from Gene, Protein, and TCGA Analysis." *Clinical and Translational Oncology*, published online April 30, 2025. https://doi.org/10.1007/s12094-025-03934-w.
5. **Dalman, Mark R., Anthony Deeter, Gayathri Nimishakavi, and Zhong-Hui Duan.** "Fold Change and P-Value Cutoffs Significantly Alter Microarray Interpretations." *BMC Bioinformatics* 13, Suppl 2 (2012): S11. https://doi.org/10.1186/1471-2105-13-S2-S11.
6. **Visone, Rosa, Maria Giulia Bacalini, Simone Di Franco, Manuela Ferracin, Maria Luisa Colorito, Sara Pagotto, et al.** "DNA Methylation of Shelf, Shore and Open Sea CpG Positions Distinguish High Microsatellite Instability from Low or Stable Microsatellite Status Colon Cancer Stem Cells." *Epigenomics* 11, no. 6 (2019): 587–604. https://doi.org/10.2217/epi-2018-0153.