# 76558 | CBIO | Hackathon

Guy Lutsker 207029448, Nivi Shenker 207227687, Eran Eben Chaime 308240597, Or Amar 311166169

## 1 Introduction

DNA methylation is an epigenetic mechanism in which a methyl group is added to cytosine. This mechanism causes silencing of gene expression and has an important role in multiple biological processes such as gene transcription regulation, aging, differentiation of stem cells, genomic imprinting and more. It has been shown that the patterns of DNA methylation are non-random, well regulated and tissue-specific **(link)**. These conclusions are consistent with the biological importance of DNA methylation. Given this notion, it is reasonable to think that the methylation status of close CpG sites is co-dependent. Indeed, several studies have demonstrated that there is a correlation in methylation status between neighboring CpG sites in healthy tissues **(link)**. In oncogenic tissues, a different scenario holds in the sense of tissue methylation sites status. On one hand, cancer silences tumor suppressor genes by methylating their promoter regions where on the other hand, hypomethylation has been recognized as a cause of oncogenesis **(link)**. This relationship raises a question concerning the co-methylation patterns in oncogenic tissues. It has been shown that the methylation status in these tissues was aberrant, at least for CpG islands **(link) (link2 needs to be checked)**. In our work, we wish to delve into the data and get a better understanding of the phenomena of co-methylation between neighboring sites. Moreover, we aimed to see if co-methylation holds in tumorigenic tissues.

## 2 Preprocessing

In our work we used data of breast tissue (BRCA) from The Cancer Genome Atlas (TCGA) **(link)**. Our data consists of 3 files of methylation patterns of healthy tissue, primary tumor and metastatic tumor. The first problem we encountered is the missing data in the files - many features (methylation sites) had nan entries. To handle this problem we chose to remove features with low percentage of data (mostly nan values) and to fill the remaining features with imputation methods. Firstly, to find the right threshold to remove entire features, we plotted the percentage of missing values over the number of features (~450K methylation sites) to make an educated decision:

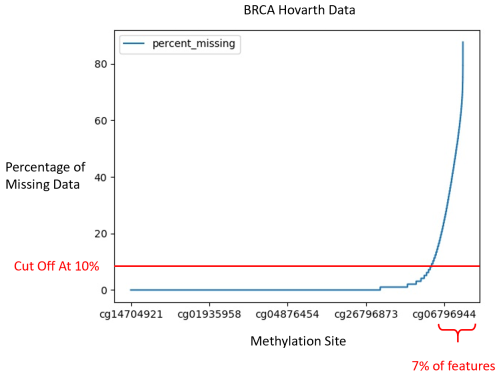

Figure 1: Example of missing data percentage in BRCA healthy file

As we can see, the features with most missing values do not take up much of the data itself and so we can intelligently remove some of the data and still keep most of it. In order to retain as much features as possible, while still getting rid of the "bad" features we chose to remove features with more than 10% nan values.

This method allowed us to retain 93% percent of the data, so we have data of better quality. After removing those features, the next step we preformed was filling the remaining nan values. There are many ways around this problem, and in our project we chose to replace the missing values with KNN imputation. Our intuition behind this step is that we assume that samples which are close in high dimensional space (~ 450K) are similar, and so we can use a weighted sum of the 3 most nearest neighbors to fill the missing values.

# 3 Distance-Co-methylation Analysis

## 3.1 Healthy Tissue Analysis

In this section, we wanted to analyze the healthy BRCA tissue data. We aimed to find the relation between the physical distance of methylation sites and their methylation status correlation. In order to do so, we used data from USCS **(link?)** containing the methylation sites positions. We partitioned the data by chromosomes, and then we sorted by the relative position in that chromosome. We then created a new table of all the sites pairs which their distance was less then 100nt. The reason we chose this filter was that from previous studies, it has been shown that co-methylation between neighboring sites exists, within sites that are no distant than 50nt apart. The next step was to calculate methylation correlation between all sites pairs. To do so, we calculated their correlation using the Pearson Correlation Coefficient. We graphed the co-methylation per distance to visualized whether the aforementioned relationship holds in our data:

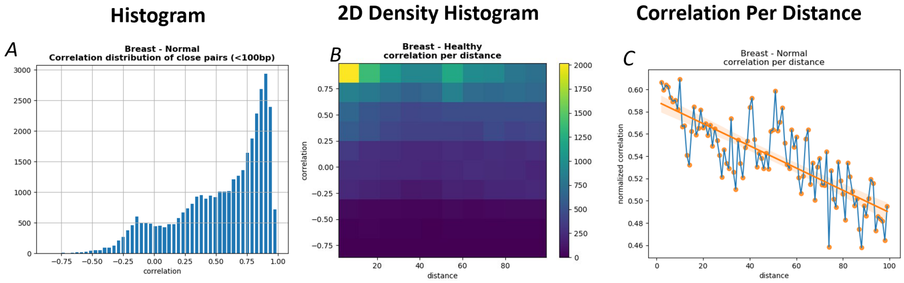

Figure 2: Healthy BRCA Tissue Analysis - (A) Histogram of correlations. (B) Density histogram of correlation per distance (C) Normalized correlation per distance with regression line, regression function: $y = -9.9 \cdot 10^{-4} \cdot x$

As expected, in Fig2.A we got mostly high correlations pairs, resulting from the fact we used pairs only at most 100nt apart. In Fig2.B we can see the 2D density of the data, and it shows us that physically close pairs have high correlation. Specifically, this result expands the phenomena which we observed in Fig2.A. Meaning, in distances that are less than 100nt apart, highly correlated methylation sites make up most of the data. In Fig2.C we normalized the correlations with relation to the number of samples we had in each distance. This results confirm the results shown in previous studies**(link)**, which means, co-methylation is high between neighboring sites and decreases as distances grow. To see whether our results are significant, we used permutation test with the statstic being $s = mean(f(d))\ s.t\ (d < N)$ where $d$ is the distance value, $f(d)$ is the correlation between pair of methylation sites and $N = 50$ is the distance threshold given in nt. The reason we chose to use $N = 50$ is because in several papers have demonstrated that two methylation sites which are less than 50nt apart, have high chance of being co-methylated **(link)**. $p_{val} = \frac{\#(s_i \geq s)}{R}$ where $R$ is our number of permutations and $s_i$ is the statistic for the $i'th$ run. In our test, $\#(s_i \geq s) = 0$ which results in $p_{val} < \frac{1}{100000}$. The null hypothesis is that there is no connection between physical distance and co-methylation. Our results suggest than we can reject the null hypothesis.

## 3.2 Comparing Healthy Tissues to Cancer Tissues

Next we wanted to analyze whether the patterns we saw in healthy tissues hold to tumorgenic tissues as well. We know from previous studies that a known property of cancer is that it may cause changes in methylation patterns **(LINK)**. To visualize if in our data the methylation sites between healthy and oncogenic tissues are indeed different we ran a UMAP analysis. UMAP is non-linear dimensionality reduction method which is capable of preserving both global & local structures in data. The result of this analysis is given in Fig.3.

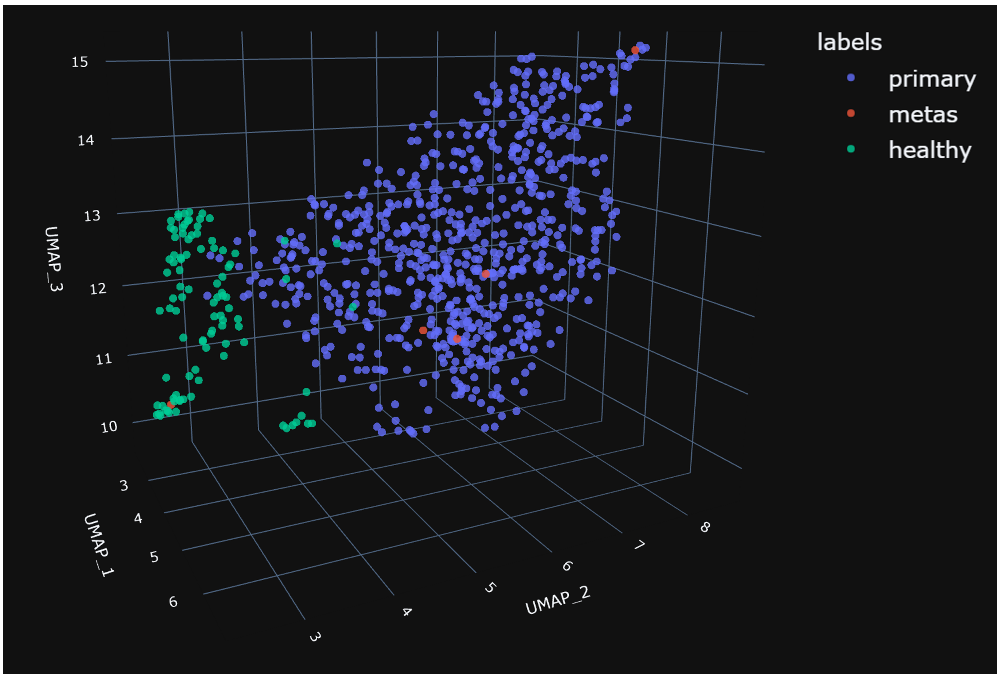

Figure 3: Methylation Patterns Comparison Between Healthy And Oncogenic Tissues - using UMAP analysis to distinct between samples from healthy tissues (green) and samples from oncogenic tissues (blue and red)

Using this dimensionality reduction method conforms with previous studies. Due to this notion, one would expect to find that cancer produces different methylation pattern also in neighboring methylation sites. However, our analysis shows that this thought is incorrect. In our work, we ran the same analysis mentioned in section 3.1 for both primary tumor tissues and metastatic tissues. The results are as follows:

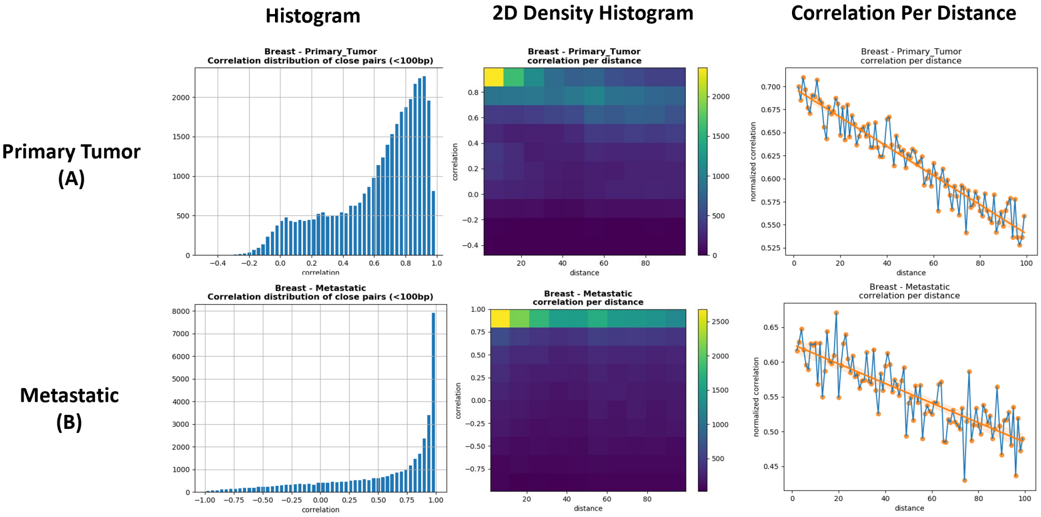

Figure 4: Oncogenic BRCA Tissue Analysis - (A) Primary tumor analysis, normalized correlation per distance with regression line, regression function: $y = -1.58 \cdot 10^{-3} \cdot x$. (B) metastatic tumor analysis, normalized correlation per distance with regression line, regression function: $y = -1.41 \cdot 10^{-3} \cdot x$.

As we can see, we get the similar patterns for both primary tumor tissues and metastatic tissues in histogram, density and correlation per distance analysis. **Moreover we tried to reject the null hypothesis using the Kruskal Wallis test a nonparametric statistical test for multiple comparisons. Our results show no significant difference between the tissues and we cannot reject the null hypothesis.** $p_{value} = XXXXX$

# 4 Highly Co-Methylated Sites Across Different Tissues

Next, we wanted to analyze if sites that were highly co-methylated in one tissue are also highly co-methylated across other tissues. It has been shown that different tissues produce distinctive methylation patterns **(link)**, however, we wanted to examine if highly co-methylated sites preserve over different tissues despite that notion. Prior to this step, we chose to visualize if different tissues indeed exhibit different methylation signal. Note that we ran the same preprocessing steps described under the "preprocessing" section before approaching this question.

## 4.1 Principle Component Analysis (PCA)

The first technique we chose to apply to our data is a linear dimensionality reduction named PCA. PCA is an unsupervised learning method which tries to preserve linear relationships in the data and is commonly used as a first step in analysis. Here we tried to see if PCA will be able to distinguish between tissues in all healthy tissues combined file:

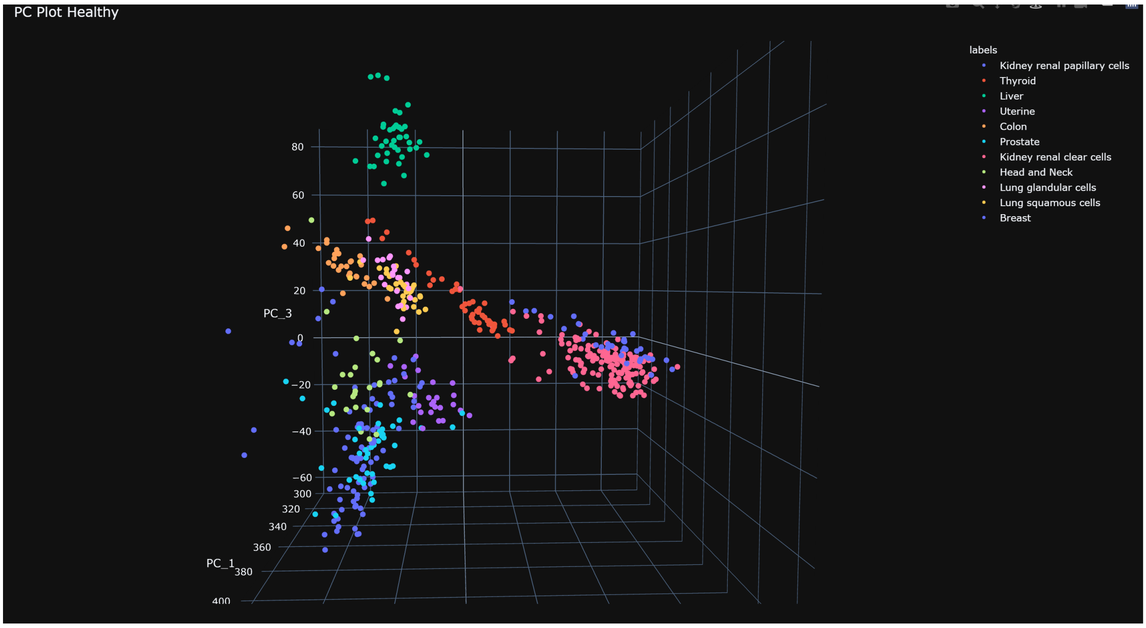

Figure 5: PCA of Healthy Tissue Combined

As we can see this unsupervised method was capable of distinguishing different tissues! Side note: the separation is even more pronounced in 3D : https://www.cs.huji.ac.il/~guy_lutsker/PCA_Plot_Healthy.html . This is very reasuring as it means that even though the data is very sparse, there is yet a strong signal in the data.

## 4.2 Uniform Manifold Approximation and Projection (UMAP)

The next step we took is to analyze the data using a non-linear approach. UMAP is non-linear dimensionality reduction method which is capable of preserving both global & local structures in data. This method could prove useful to us since the manifold of our data lies in a very high dimension and so a method as robust as this could help us under stand the data better. UMAP analysis heavily relies on a neighbors graph generation, and we can learn quite a bit about the manifolds structure from the neighbors connection:

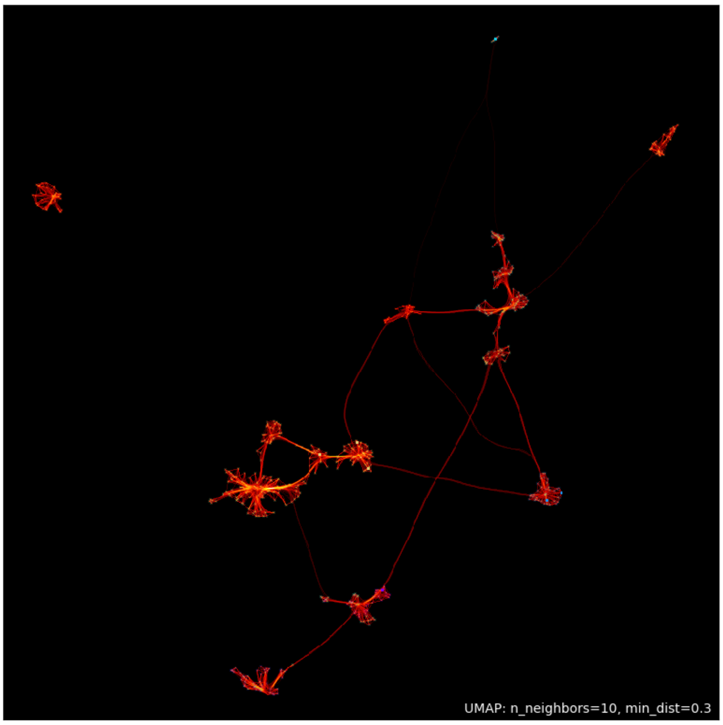

Figure 6: Neighbors graph on UMAP on Healthy Tissue Combined

And the UMAP graph itself:

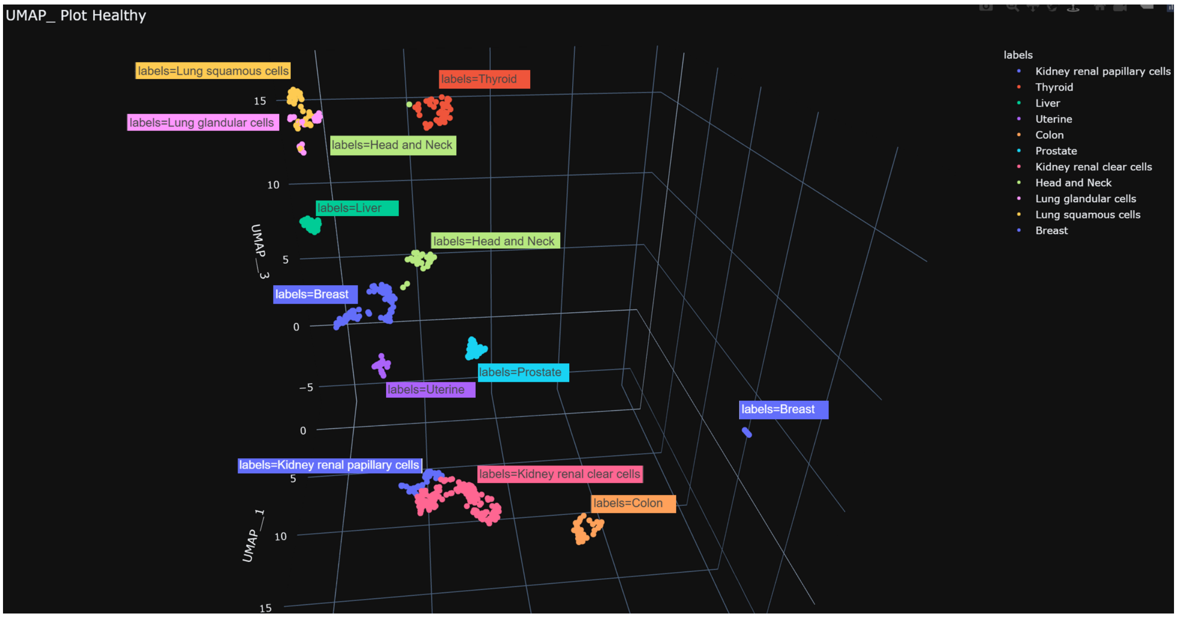

Figure 7: UMAP of Healthy Tissue Combined

Here the results are really good! UMAP with its local & global manifold preserving magic was able to separate the tissues almost perfectly. In addition, it has also been able to capture the similarity of tissues, for example we can that both Kidney renal paplilary cells & Kidney renal are 2 clusters “smashed” together, while still being seperated into 2 distinct clusters. The same is true for Lung data, and all of these phenomena are more pleasing in 3D: https://www.cs.huji.ac.il/~guy_lutsker/UMAP_Plot_Healthy.html

### 4.2.1 Within Tissue

- Removing features with high percentage of missing data
- Use a data completion method (such as KNN)

### 4.2.2 Between Tissues

- Merging the data from several tissues
- Removing features with high percentage of missing data
- Use a data completion method (such as KNN)

## 4.3 Visualization

- Visualize the data to have a better interpretation of what we are working with, using dimensionality reduction techniques such as PCA, UMAP, and diffusion maps
- Cluster the data to see if we can recognize some intrinsic properties in the manifold of the high dimensional data, with techniques such as K-means, HDBSCAN or soft clustering methods such as NMF, LDA

## 4.4 Connection Between ”Correlative Distance” & Physical Distance

- Calculating physical distance between methylation sites (in base pairs)
- Calculating ”correlative distance” based on Pearson correlation
- Finding if we have a connection between ”correlative distance” and physical distance

## 4.5 Connection Between Healthy Tissue & Cancer Tissue

- Same pipeline in section 3.3 will serve us for this section as well
- Find if the same correlations preserve in cancer tissues compared to healthy tissues

# 5 Model

Denoting $k$ as the number of samples (people) and $n$ as the number of features (methylation sites), we can write our data as follows:

$\{x\}_{i=1}^{k} \in \mathbb{R}^n$ e.g $x_5^2$ is the second methylation site of the fifth person in our data. In order to find the relation between physical distance of two sites in the genome (given in bp) and “correlative distance” we need to normalized the physical distance. To do so, we chose to use “Min-Max Normalization”, this normalization preserves linear relationships in the data. We denote the normalized value of the physical distance with $l_n$. In order to calculate the “correlative distance” we use the Pearson Correlation coefficient denoted with $Corr$ and is given by the formula:

$Corr(x^i, x^j) = \frac{\sum_t (x_t^i - \bar{x^i}) \cdot (x_t^j - \bar{x^j})}{\sqrt{\sum_t (x_t^i - \bar{x^i})^2 \cdot (x_t^j - \bar{x^j})^2}}$. Notice that both $l_n$ and $Corr$ have unit-less outputs and therefore we are allowed to compare them. We expect to find that two closely located sites will have similar methylation patterns and vice-versa. Therefore, we have constructed the following formula:

$$f(x^i, x^j) = \frac{1}{|Corr(x^i, x^j, )| - l_n(x^i, x^j)}$$

Our first intuition was to subtract the physical distance from the correlation in order to infer where the genome is “compacted” or “spreaded”. However, we understand that both high positive and high negative correlation might have some important biological insight (e.g both might suggest regulatory regions). Therefore, we have decided to take the absolute value of the correlation. Our last step was to invert the result because we expect to find the most interesting findings in this way. In order to preserve the information about the sign of the correlation, in the final step of the analysis we will use the following formula which incorporates the sign of the correlation:

$$g(x^i, x^j) = sign(Corr(x^i, x^j)) \cdot f(x^i, x^j)$$

# 6 Computational Tools

- KNN for missing data
- PCA for data visualization
- UMAP for data visualization
- Clustering algorithms such as K-means, HDBSCAN or soft clustering methods such as NMF, LDA
- Permutation Tests