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[1], aging[2,3], differentiation of stem cells[4,5], genomic imprinting[6] and more. It has been shown that the patterns of DNA methylation are non-random, well regulated and tissue-specific[7]. 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[8]. In oncogenic tissues, a different scenario holds in the sense of tissue methylation sites status. On one hand, during cancerous processes silencing of tumor suppressor genes occurs by methylating their promoter regions where on the other hand, hypomethylation has been recognized as a cause of oncogenesis[9]. 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[10]. 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 aim to see if co-methylation holds in tumorigenic tissues. Additionally, it has been shown that different tissues exhibit different methylation patterns[11]. Given this notion, we wish to investigate whether co-methylation of neighboring sites show to the same trend.

2 Data

2.1 Data Description

Firstly, we used data of breast tissue, from the breast cancer (BRCA) data base, extracted from The Cancer Genome Atlas (TCGA)[12]. TCGA consist of thousands of cancerous and healthy samples using Illumina 450K array. Illumina is a methylation profiling platform providing quantitative methylation measurement at the single-CpG–site level. The 450K array covers only about 1.5% of CpGs in the human genome.. Our data consists of three data-bases of methylation patterns of healthy tissue, primary tumor and metastatic tumor.

2.2 Pre-processing

The first issue we encountered is the missing data in the files - many features (methylation sites) had missing entries. To handle this issue we chose to remove features with low percentage of data (mostly missing 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, as mentioned) to make an educated decision:

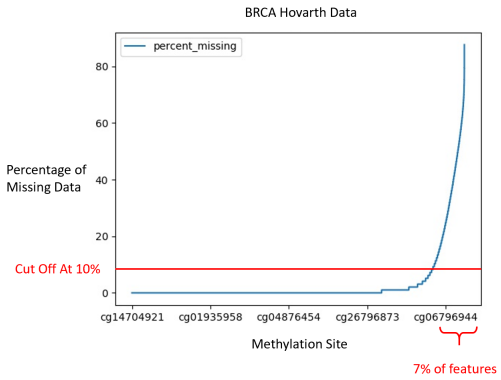

Figure 1: Example of missing data percentage in BRCA healthy file

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% missing values. This method allowed us to retain 93% of the data, so we have data of better quality. After removing those features, the next step we preformed was filling the remaining missing values. There are many ways around this problem, and in our project we chose to replace the missing values with KNN imputation, which has been shown to provide generally effective results[13]. KNN imputation is not robust to mostly missing values, but because we removed highly sparsed features, KNN imputation provides good fit. 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. We chose $k = 3$ for the imputation due to the spacious nature of our data.

3 Distance-Co-methylation Analysis

3.1 Healthy Tissue Analysis

In this section, we wanted to analyze healthy tissue from the BRCA data base. We aimed to find the relation between the physical distance of methylation sites and their methylation correlation. In order to do so, we used data from USCS[14] containing the methylation sites positions. We partitioned the data by chromosomes, sorted by the relative position in that chromosome and created a new table of all the sites pairs which their distance was less then 100nt. Previous studies have shown that co-methylation between neighboring sites exists within sites that are no distant than 50nt apart[15], and therefore we chose to use this filter. The next step was to calculate methylation correlation between all sites pairs. In this analysis we assumed that the relationship of methylation between sites is linear and therefore we chose to use Pearson Correlation Coefficient. We graphed the co-methylation per distance to visualized whether the aforementioned relationship holds in our data:

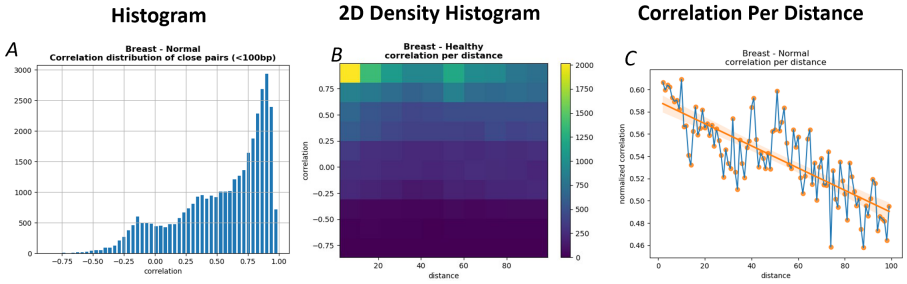

Figure 2: Healthy BRCA Tissue Analysis - (A) Histogram of correlations. This graph describes the correlation distribution of close methylation sites pairs. Distance between two methylation sites is at most 100bp. We can infer that close sites have as expected mostly high correlation (B) 2D Density histogram of correlation per distance. Here we present the count of number of entries into a correlated section as a function of distance. For example, in the upper left cell there were ~2000 rows in which the distance was between 2 to 10 bases, and the correlation in distance was higher than 0.75. In overall, in short distance between different sites (including those with at most 100 bases) most entries are in higher correlation. (C) Normalized correlation per distance with regression line, regression function: $y = -9.9 \cdot 10^{-4} \cdot x$. In this graph we mapped the average correlation (axis y) for a value of distance (axis x). We normalized the correlations with relation to the number of samples we had in each distance. This results is consistent with previous studies[16,17,18], 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 statistic 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[8,15]. $p_{val} = \frac{\#(s_i \geq s)}{R}$ 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 that 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[19,20]. 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.

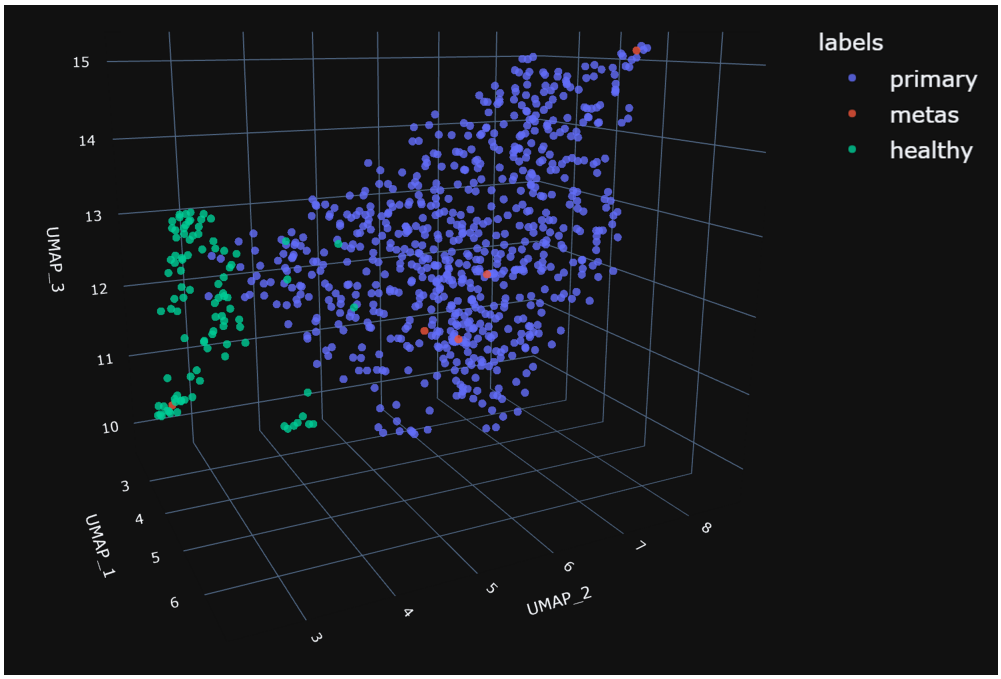

Figure 3: Methylation Patterns Comparison Between Healthy And Oncogenic Tissues - using UMAP analysis to distinct between samples from healthy tissues (green) and samples from oncogenic tissues (blue and red)

Indeed, the data plotted by this dimensionality reduction method conforms with previous studies, as we can observe that there is a good separation between healthy and oncogenic tissues. Due to this notion, one might 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:

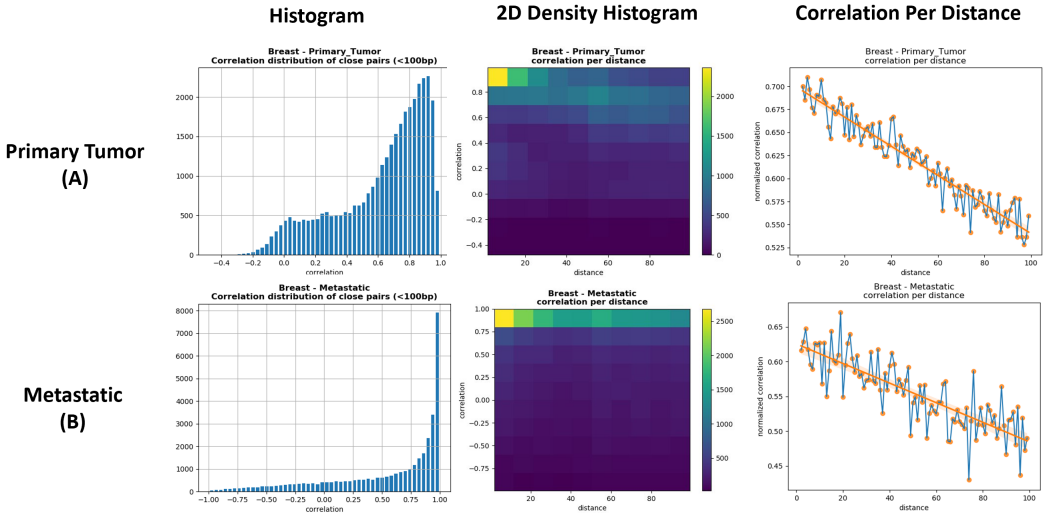

Figure 4: Oncogenic BRCA Tissue Analysis - (A) Primary tumor analysis, normalized correlation per distance with regression line, regression function: $y = -1.58 \cdot 10^{-3} \cdot x$. (B) metastatic tumor analysis, normalized correlation per distance with regression line, regression function: $y = -1.41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

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 non-parametric statistical test for multiple comparisons. When Running the test we got $p_{value} =$

6.6×10^{-24} . And so our results show a significant difference between the tissues and we reject the null hypothesis. The affect the test found can be visualized if we overlay the slopes of the graphs to see the difference:

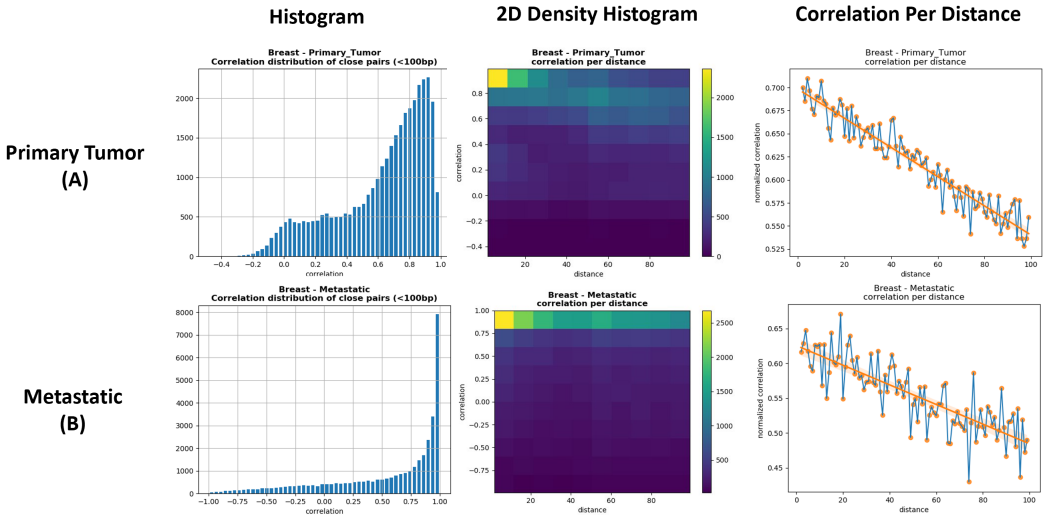

Figure 5: Nivi overlay

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:

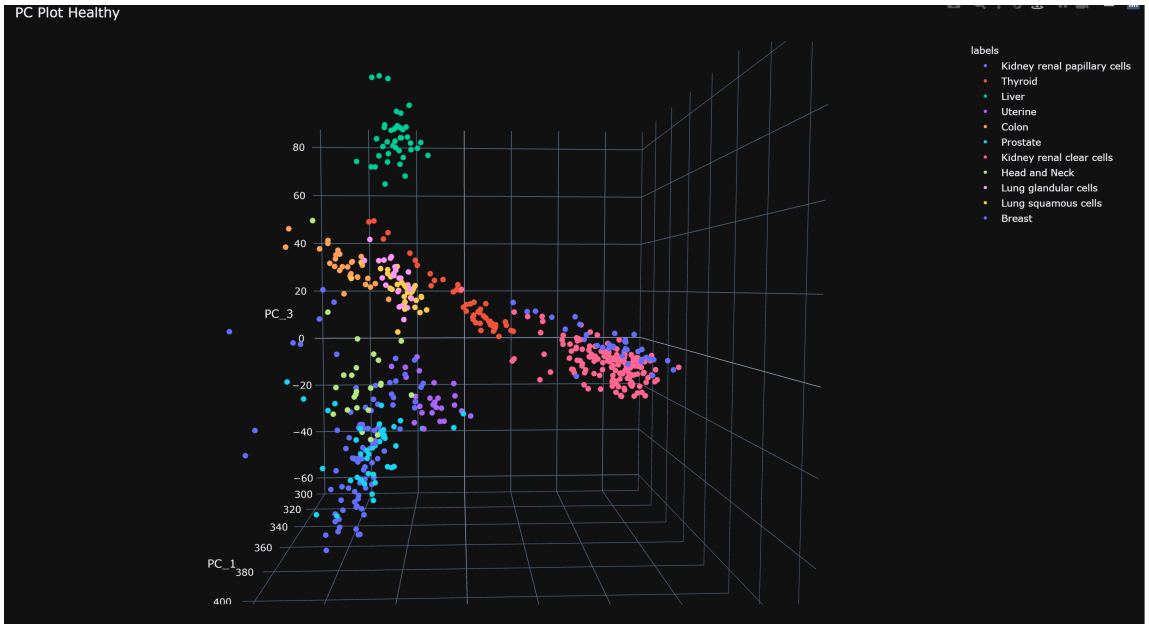

Figure 6: PCA of Healthy Tissue Combined

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:

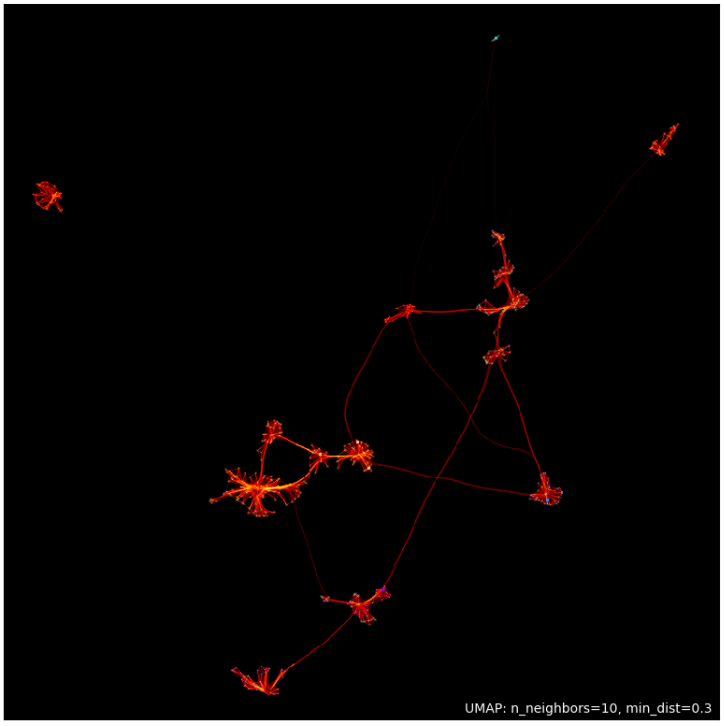

Figure 7: Neighbors graph on UMAP on Healthy Tissue Combined

And the UMAP graph itself:

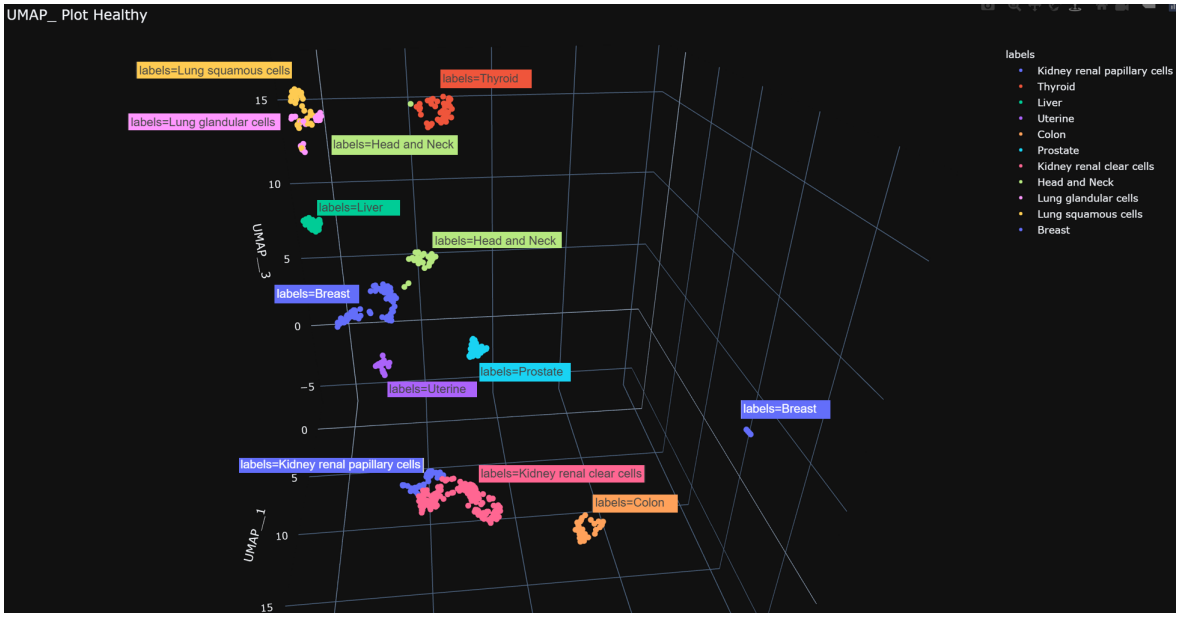

Figure 8: UMAP of Healthy Tissue Combined

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

4.3 Statistical Analysis

Now, that we have seen that these tissues indeed have a differentiating signal, we can move on to conduct some numerical (and not just visual) statistical tests. Our first idea was to run a Kruskal-Wallis test to see whether it is capable of finding that these samples come from different groups. And unsurprisingly so, the results show $p_{value} = 1.1 \times 10^{-171}$. Which means that we can say with statistical significant that these indeed exhibit an unsupervised differentiating signal. In addition we wanted to calculate the distance to correlation analysis for all tissues, and got the following plots:

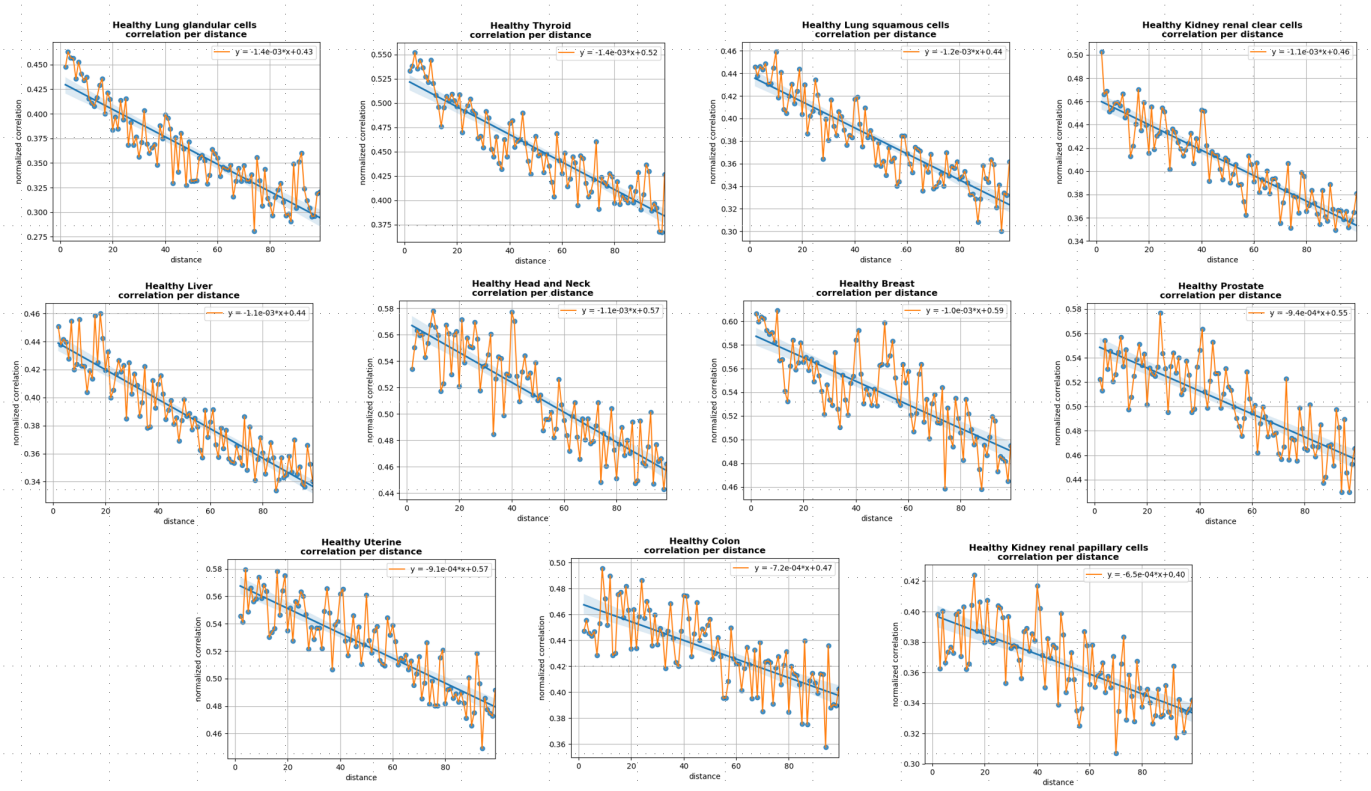

Figure 9: Oncogenic BRCA Tissue Analysis - (A) Primary tumor analysis, normalized correlation per distance with regression line, regression function: $y = -1.58 \cdot 10^{-3} \cdot x$. (B) metastatic tumor analysis, normalized correlation per distance with regression line, regression function: $y = -1.41 \cdot 10^{-3} \cdot x$.

5 References

1. Chan, M. F., Liang, G., & Jones, P. A. (2000). Relationship between transcription and DNA methylation. Current topics in microbiology and immunology, 249, 75-86.

2. Zhang, Z., Deng, C., Lu, Q., & Richardson, B. (2002). Age-dependent DNA methylation changes in the ITGAL (CD11a) promoter. Mechanisms of ageing and development, 123(9), 1257-1268.

3. Ahuja, N., & Issa, J. P. (2000). Aging, methylation and cancer.

4. Li, E. (2002). Chromatin modification and epigenetic reprogramming in mammalian development. Nature Reviews Genetics, 3(9), 662-673.

5. Reik, W., Dean, W., & Walter, J. (2001). Epigenetic reprogramming in mammalian development. Science, 293(5532), 1089-1093.

6. Reik, W., & Walter, J. (1998). Imprinting mechanisms in mammals. Current opinion in genetics & development, 8(2), 154-164.

7. Chen, Z. X., & Riggs, A. D. (2011). DNA methylation and demethylation in mammals. Journal of Biological Chemistry, 286(21), 18347-18353.

8. Affinito, O., Palumbo, D., Fierro, A., Cuomo, M., De Riso, G., Monticelli, A., ... & Cocozza, S. (2020). Nucleotide distance influences co-methylation between nearby CpG sites. Genomics, 112(1), 144-150.

9. Das, P. M., & Singal, R. (2004). DNA methylation and cancer. Journal of clinical oncology, 22(22), 4632-4642.

10. Costello, J. F., Frühwald, M. C., Smiraglia, D. J., Rush, L. J., Robertson, G. P., Gao, X., ... & Plass, C. (2000). Aberrant CpG-island methylation has non-random and tumour-type–specific patterns. Nature genetics, 24(2), 132-138.

11. Lokk, K., Modhukur, V., Rajashekar, B., Märtens, K., Mägi, R., Kolde, R., ... & Tõnisson, N. (2014). DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns. Genome biology, 15(4), 1-14.

12. The results presented in this work are in whole or part based upon data generated by the TCGA Research Network：https://www.cancer.gov/tcga.

13. Beretta, L., & Santaniello, A. (2016). Nearest neighbor imputation algorithms：a critical evaluation. BMC medical informatics and decision making, 16 Suppl 3(Suppl 3), 74.

14. Navarro Gonzalez J, Zweig AS, Speir ML, Schmelter D, Rosenbloom KR, Raney BJ, Powell CC, Nassar LR, Maulding ND, Lee CM, Lee BT, Hinrichs AS, Fyfe AC, Fernandes JD, Diekhans M, Clawson H, Casper J, Benet-Pagès A, Barber GP, Haussler D, Kuhn RM, Haeussler M, Kent WJ. The UCSC Genome Browser database：2021 update. Nucleic Acids Res. 2021 Jan 8;49(D1):D1046-D1057.

15. Witzmann, S. R., Turner, J. D., Mériaux, S. B., Meijer, O. C., & Muller, C. P. (2012). Epigenetic regulation of the glucocorticoid receptor promoter 17 in adult rats. Epigenetics, 7(11), 1290-1301.

16. Sun, L., Namboodiri, S., Chen, E., & Sun, S. (2019). Preliminary Analysis of Within-Sample Co-methylation Patterns in Normal and Cancerous Breast Samples. Cancer informatics, 18, 1176935119880516.

17. Li, Y., Zhu, J., Tian, G., Li, N., Li, Q., Ye, M., ... & Zhang, X. (2010). The DNA methylome of human peripheral blood mononuclear cells. PLoS biol, 8(11), e1000533.

18. Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V. K., Attwood, J., Burger, M., ... & Beck, S. (2006). DNA methylation profiling of human chromosomes 6, 20 and 22. Nature genetics, 38(12), 1378-1385.

19. Ghavifekr Fakhr, M., Farshdousti Hagh, M., Shanehbandi, D., & Baradaran, B. (2013). DNA methylation pattern as important epigenetic criterion in cancer. Genetics research international, 2013, 317569

20. Baylin, S. B., & Jones, P. A. (2016). Epigenetic Determinants of Cancer. Cold Spring Harbor perspectives in biology, 8(9), a019505