Politecnico di Torino

Master's Degree in Mathematical Engineering

Material for Thesis

a– Exploratory Data Visualization of DNA Methylation in GSE69914

Elisabetta Roviera s328422

Contents

Abstract

This report summarizes the exploratory visual analysis performed on the breast cancer methylation dataset GSE69914 [1]. I first assess global distributions and quality, then quantify methylation outlier burden, contrast Tumor vs. Normal at the CpG level, and finally embed samples via dimensionality reduction. The goal is to establish a clear visual baseline before preprocessing the data.

All supporting materials are available in the project repository: the full repo THESIS, the curated medical/biological papers with summaries (Medical and Biological information), the data visualization papers and summaries (Data Exploration & Visualization), and the complete notebook with all code that this document summarizes (01-data-exploration.ipynb).

1 Introduction

Breast cancer represents one of the most extensively studied models of tumorigenesis and remains a leading cause of cancer-related mortality worldwide. Beyond genetic mutations, its initiation and progression are strongly influenced by epigenetic deregulation, particularly through alterations in DNA methylation patterns at cytosine–phosphate–guanine (CpG) sites [2] [3]. Such aberrant methylation contributes to transcriptional silencing of tumor-suppressor genes, genomic instability, and the establishment of pre-neoplastic "field defects" that precede visible malignancy [4].

The notebook, **01-data-exploration.ipynb**, constitutes the first computational step of my thesis project. Its goal is to perform an in-depth exploratory analysis and visualization of DNA methylation data from the **GSE69914** dataset [1], which profiles breast tissue samples from Normal, Tumor-adjacent, and Tumor contexts using the Illumina HumanMethylation450 BeadChip platform.

The analyses presented here focus on establishing a clear and interpretable overview of the dataset's global structure before applying any normalization or filtering. Specifically, the notebook explores:

- Global methylation distributions and sample-level variability;

- Methylation outlier burden and its progression across tissue groups;

- Group-level contrasts between Tumor and Normal samples at the CpG level;

- Low-dimensional embeddings (t-SNE) capturing the overall organization of the methylome.

This exploratory phase provides the visual and quantitative foundation required for subsequent steps—namely, data preprocessing.

2 Dataset

The **GSE69914** dataset [1] provides DNA methylation profiles from **407 breast tissue samples** obtained using the **Illumina Infinium HumanMethylation450 BeadChip**, covering approximately **485,512 CpG sites**. It includes **56 normal tissues**, **49 adjacent tissues**, **302 tumors**, and a small subset of **BRCA1-related samples** (8 normal carriers and 3 tumors). After transposition, each row corresponds to a sample and each column to a CpG probe, with β-values in $[0, 1]$ representing methylation levels. The array distinguishes CpG (`cg`) and non-CpG (`ch`) contexts [5]. This dataset is an established reference for studying breast cancer epigenetics, and exhibits complete data coverage with no missing values, confirming its suitability for visualization and subsequent normalization.

3 Visualization Pipeline and Key Findings

The exploratory analysis of *GSE69914* was designed as a stepwise visual pipeline, moving from global quality assessment to group comparison and, finally, to low-dimensional structure. The main results are summarized in Fig. 1.

3.1 Global methylation patterns and data quality

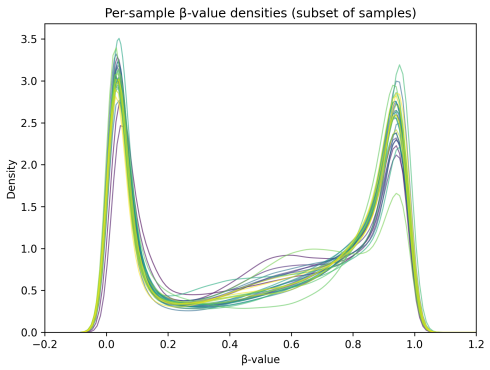
I first examined the distribution of β-values (fractional DNA methylation in $[0, 1]$) across individual samples. The per-sample density curves (Fig. 1a) show the characteristic **bimodal profile** of Illumina 450k arrays, with one peak near unmethylated CpGs ($\beta \approx 0$) and one near fully methylated CpGs ($\beta \approx 1$). This behaviour reflects the biology of CpG regulation, where many loci are either transcriptionally active (hypomethylated) or repressed (hypermethylated), rather than occupying stable intermediate states [3], [6]. When these distributions are stratified by group (Normal, Tumor-adjacent, Tumor; Fig. 1b), I observe that Tumor samples show a flatter high-β peak and a broader tail, consistent with emerging **global hypomethylation** and higher heterogeneity. Tumor-adjacent samples tend to fall in between Normal and Tumor, suggesting early epigenetic drift in histologically non-tumor tissue [4].

3.2 Group-wise summary at the sample level

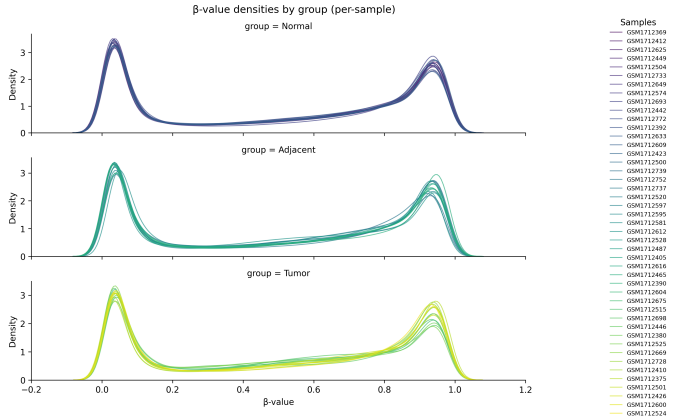
I then summarized each sample by its global methylation level and variability. The group-wise distributions of per-sample mean β and dispersion (IQR) in Fig. 1c show two key trends: (i) **Tumor samples have slightly lower global methylation** than Normal, consistent with genome-wide hypomethylation being a hallmark of cancer and associated with genomic instability [3]; (ii) **Tumor samples are more variable**, indicating increased epigenetic instability and stochastic deregulation [7]. This agrees with the outlier-burden analysis performed in the notebook, where Tumor carries a substantially higher number of CpG outliers than either Normal or Tumor-adjacent, mirroring the "instability load" phenotype reported in ageing and cancer epigenomes [7].

3.3 Low-dimensional organization of the methylome

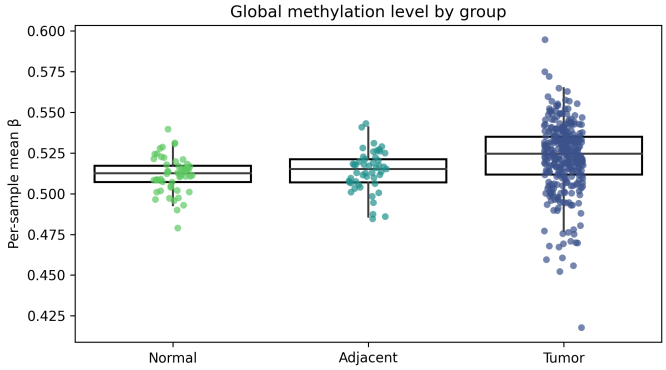
To test whether these differences appear at a global, genome-wide level, I embedded all samples in two dimensions using a PCA→t-SNE pipeline (Fig. 1d). Briefly, methylation profiles (expressed as M-values) were first compressed with PCA to the top 50 components, then projected into 2D using t-SNE. PCA captures major axes of variance, while t-SNE preserves local neighborhood structure in the reduced space [8], [9]. The resulting map shows **clear separation between Normal and Tumor samples**, with Tumor-adjacent samples occupying an intermediate region rather than overlapping completely with either group. This pattern supports a **continuous epigenetic gradient** from Normal → Tumor-adjacent → Tumor, instead of an abrupt "all-or-nothing" transition [4].



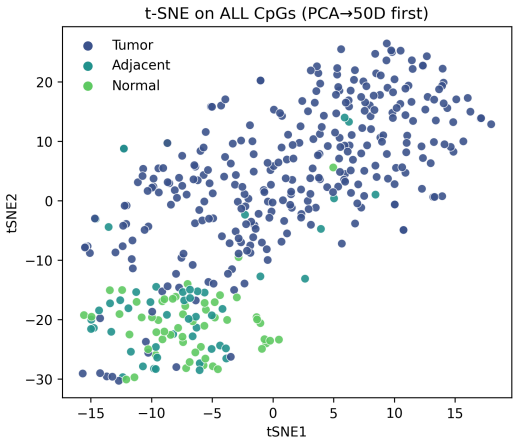
(a) Global β-value density per sample (bimodality check).



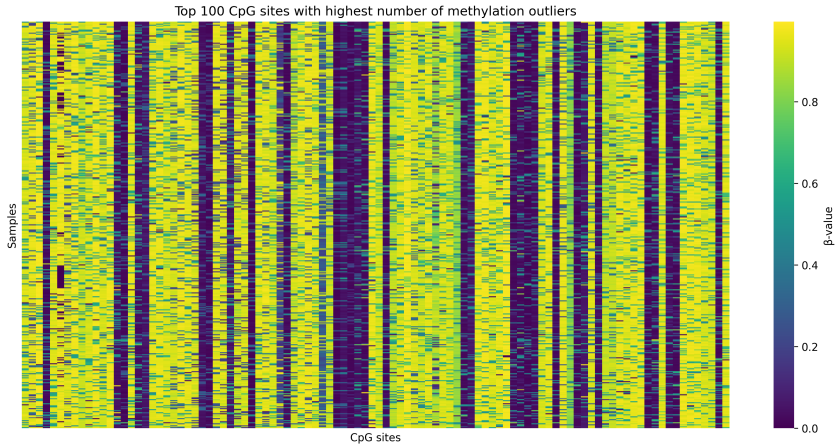
(b) Ridge plot by group.



(c) Per-sample summary: mean β and IQR β by group.



(d) t-SNE using ALL ∼480k CpGs.



(e) Heatmap of top 100 CpG outliers.

Figure 1: Summary of the exploratory visual analysis of DNA methylation in *GSE69914*. Panels 1a–1b show the expected bimodal distribution of β-values across samples, with systematic differences between Normal, Tumor-adjacent, and Tumor tissues. Panels 1c–1d highlight that Tumor samples display lower average methylation and markedly higher variability, and that a PCA→t-SNE projection separates Normal and Tumor while placing Tumor-adjacent samples in between, suggesting a progressive epigenetic continuum rather than a binary switch [4], [7], [8], [9]. Panel 1e shows that recurrent CpG outliers arise in both directions (focal hypermethylation and focal hypomethylation), consistent with the known coexistence of promoter silencing and global/segmental demethylation in cancer [3], [6].

3.4 Locus-level instability and recurrent CpG outliers

Finally, I investigated which individual CpG sites show the most extreme deviations. The heatmap in Fig. 1e displays the top 100 CpG loci with the highest number of outlier events across samples. Two observations emerge: (i) epigenetic disruption is **not uniform** — instability occurs at specific CpG sites and in specific samples, rather than as a smooth genome-wide shift; (ii) both directions are present — I observe focal **hypermethylation** (very high β) and focal **hypomethylation** (very low β). In cancer biology, focal hypermethylation can silence tumor-suppressor regions, while focal hypomethylation can derepress oncogenic programs and weaken genomic stability [3]. The coexistence of both extremes in the same tissue context reflects the classic "too much and too little methylation" behavior of tumor genomes [3].

4 Conclusion

These results are consistent with what I expect biologically and technically for HM450 data. Across multiple visualization strategies, results were consistent with known biological patterns:

- **Bimodal β-value distributions** typical of Illumina HM450 data;
- **Systematic group shifts**, with Tumor samples exhibiting lower global methylation and greater variability than Normal;
- **Intermediate profiles** for Tumor-adjacent tissues, suggesting early epigenetic drift rather than abrupt transitions [4], [7].

At the CpG level, heatmaps and boxplots highlighted **recurrently variable loci** where both focal hypermethylation and hypomethylation coexist—processes known to silence tumor suppressors and activate oncogenic pathways [3]. Dimensionality-reduction results (PCA $\rightarrow$ t-SNE) further supported a **continuous epigenetic gradient** from Normal to Tumor, confirming structured rather than random methylation differences.

Overall, these findings validate the dataset's quality and provide a strong rationale for the upcoming *data pre-processing* phase.

References

[1] National Center for Biotechnology Information, *Gse69914 on geo datasets*, Gene Expression Omnibus (GEO), [Online]. Available: `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE69914`, 2015.

[2] A. Bird, "Dna methylation patterns and epigenetic memory," *Genes & Development*, vol. 16, no. 1, pp. 6–21, Jan. 2002.

[3] M. Ehrlich, "Dna methylation in cancer: Too much, but also too little," *Oncogene*, vol. 21, no. 35, pp. 5400–5413, Aug. 2002.

[4] A. E. Teschendorff, Y. Gao, A. Jones, et al., "Dna methylation outliers in normal breast tissue identify field defects that are enriched in cancer," *Nature Communications*, vol. 7, p. 10 478, 2016.

[5] W. Zhou, P. W. Laird, and H. Shen, "Comprehensive characterization, annotation and innovative use of infinium dna methylation beadchip probes," *Nucleic Acids Research*, vol. 45, no. 4, e22, 2016, [Online]. Available: `https://doi.org/10.1093/nar/gkw967`. DOI: `10.1093/nar/gkw967`

[6] J. Maksimovic, L. Gordon, and A. Oshlack, "Swan: Subset-quantile within array normalization for illumina infinium humanmethylation450 beadchips," *Genome Biology*, vol. 13, no. 6, R44, 2012.

[7] F. Seeboth et al., "Dna methylation outlier burden, health and ageing in generation scotland and lothian birth cohorts," *Clinical Epigenetics*, vol. 12, no. 1, p. 103, 2020.

[8] F. Pedregosa et al., "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[9] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.