Politecnico di Torino

Master's Degree in Mathematical Engineering

Material for Thesis

b– Dataset Exploration and Visualization
Intra and Inter Analysis

Elisabetta Roviera s328422

Contents

Abstract

SCRIVERE ABSTRACT + METTERE LINK NOTEBOOK

1 Intra-Dataset Exploration and Visualization

Exploratory Data Analysis (EDA) represents a fundamental step in DNA methylation studies, particularly in cancer epigenomics, where alterations in global methylation levels, sample-level variability, and locus-specific instability often precede detectable phenotypic changes. Before applying any preprocessing, filtering, bias correction, or machine-learning models, each dataset must be evaluated independently in order to assess its internal structure, detect technical anomalies, and identify biologically meaningful patterns.

In the context of breast cancer, epigenetic deregulation manifests through both global and local phenomena: genome-wide hypomethylation, focal hypermethylation of tumor-suppressor regions, and the presence of pre-neoplastic "field defects" in histologically normal tissues close to tumors [1], [2], [3]. The primary biological goal of this thesis—*to identify DNA methylation alterations in normal tissues and evaluate whether they may contribute to the transition toward tumorigenesis*—motivates this detailed intra-dataset investigation.

This section analyzes each dataset (GSE69914 [1.1], GSE225845, GSE287331) separately and systematically, focusing on: missingness and quality control, global methylation distributions, sample-level summaries, CpG-level instability, correlation structure, dimensionality reduction, and genome-context coverage.

1.1 Dataset GSE69914

Dataset overview The GSE69914 dataset provides genome-wide DNA methylation profiles generated using the Illumina Infinium HumanMethylation450 BeadChip on bisulfite-converted DNA, covering more than 480,000 CpG sites. The series, contributed by Teschendorff and Widschwendter and publicly released on June 18, 2015, is available on GEO under accession GSE69914.

In total, the dataset includes **407 breast tissue samples** spanning the three tissue states central to this thesis:

- 50 Normal samples, representing the physiological methylation baseline;

- 42 Tumor-adjacent samples, reflecting early epigenetic alterations in proximity to the tumor;

- 305 Tumor samples, providing the malignant endpoint of the Normal → Adjacent → Tumor axis.

A small subset consists of BRCA1-related samples (7 normal carriers and 3 tumors). BRCA1 encodes a key DNA-repair protein, and pathogenic variants substantially increase lifetime breast and ovarian cancer risk [4], adding a hereditary-risk dimension to the cohort.

The final working matrix therefore contains 407 tissue samples and 485,512 CpG probes, offering a structured and biologically coherent dataset for intra-dataset exploratory analyses.

Missingness and data quality An assessment of missing values across all samples and CpG probes confirmed that the GSE69914 methylation matrix contains *no missing entries* of any kind. The dataset is therefore complete, technically clean, and well suited for direct exploratory analysis and visualization without requiring imputation or preliminary data recovery steps.

Global methylation distributions To characterise the overall methylation landscape of the dataset, I first examined the distribution of β-values (fractional methylation in $[0, 1]$). The group-wise mean density curve (Fig. 1) displays the characteristic *bimodal* profile of Illumina 450k arrays, with peaks near unmethylated ($\beta \approx 0$) and fully methylated ($\beta \approx 1$) CpG sites. This pattern reflects the underlying biology of CpG regulation, where many loci tend to be either transcriptionally active (hypomethylated) or repressed (hypermethylated) [2], [5].

When comparing tissue groups, a clear gradient emerges. Tumor samples show a slightly flatter high-β peak and a broader low-β tail, consistent with the well-described phenomenon of **global hypomethylation** and increased heterogeneity in cancer [1]. Tumor-adjacent samples lie between Normal and Tumor, suggesting early epigenetic drift and subtle field effects occurring in histologically non-neoplastic tissue [3]. Normal samples exhibit the sharpest and most stable bimodality, representing the expected physiological baseline.

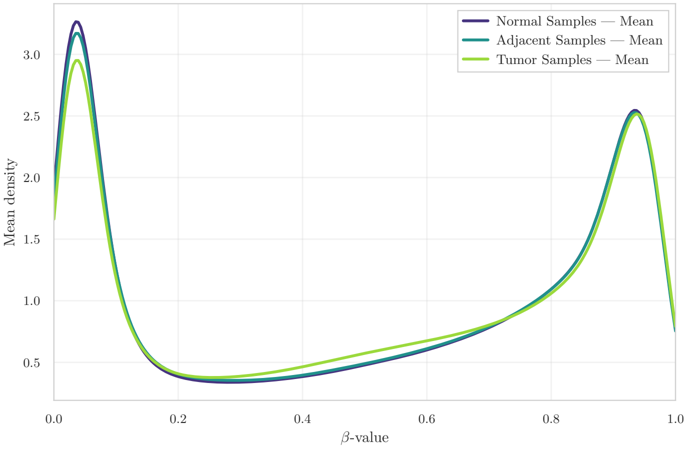
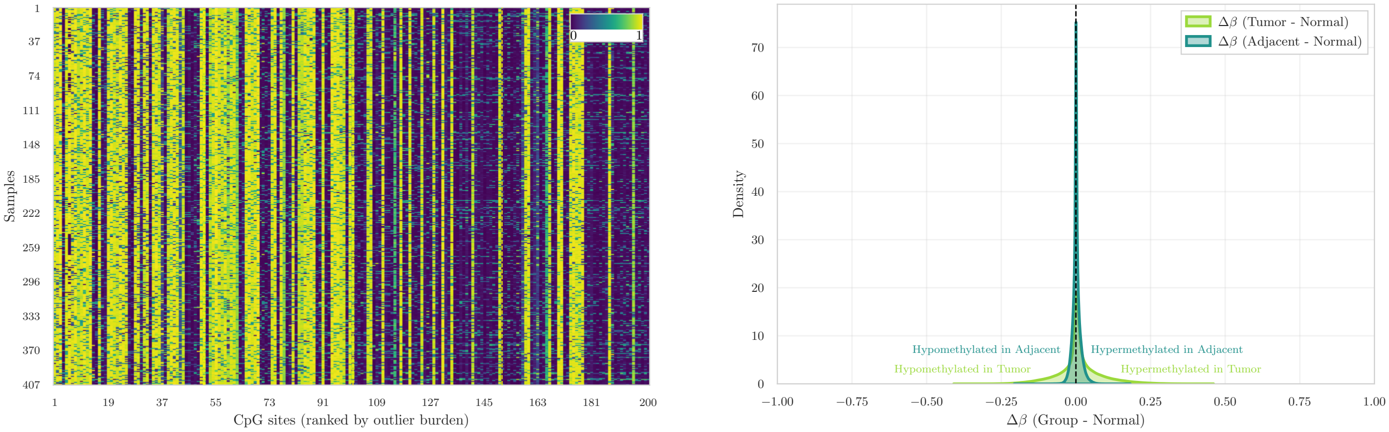


Figure 1: Group-wise mean β-value density.

CpG-level instability and recurrent outliers To characterise locus-specific instability, I examined CpG sites that show the strongest deviations across samples. The heatmap of the top outlier loci (Fig. 2a) indicates that epigenetic disruption is *not uniform*: specific CpG sites and specific samples display extreme deviations, rather than a diffuse genome-wide shift. Moreover, both directions of alteration are present — focal **hypermethylation** (very high β) and focal **hypomethylation** (very low β). In cancer biology, hypermethylation can silence tumor-suppressor regions, whereas hypomethylation can derepress oncogenic pathways and weaken genomic stability, reflecting the classic "too much and too little methylation" behaviour of tumor genomes [1].

Complementary $\Delta\beta$ distributions comparing Tumor and Adjacent tissues against Normal (Fig. 2b) show a clear shift toward hypomethylation in Tumor samples and a subtler but detectable drift in Adjacent tissues. This provides evidence that epigenetic alterations emerge early in histologically normal tissue located near the tumor, supporting the field-cancerization model [3].



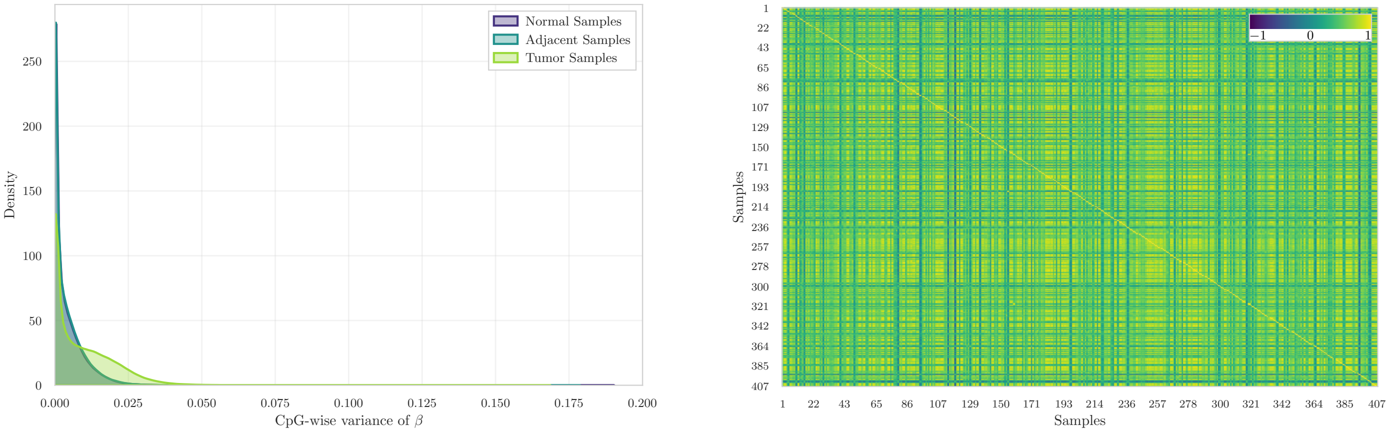
(a) Top outlier CpG loci.

(b) $\Delta\beta$ distributions (Tumor/Adjacent − Normal).

Figure 2: CpG-level instability: heatmap of top outlier CpGs (left) and corresponding $\Delta\beta$ distributions (right).

Variance and correlation structure To quantify within-group epigenetic variability, I examined the distribution of CpG-wise variance across samples. The density curves (Fig. 3a) show that Tumor samples have markedly higher variance, reflecting increased epigenetic instability. In contrast, Normal and Tumor-adjacent samples display almost identical variance distributions that are tightly concentrated near zero. This indicates that, at the level of CpG-wise variability, Adjacent tissues do not yet exhibit detectable divergence from Normal samples, despite being histologically proximate to the tumor.

A complementary perspective is provided by the sample correlation heatmap (Fig. 3b), which shows the expected high level of pairwise similarity across all samples. This behaviour is typical of genome-wide DNA methylation data, where a large fraction of CpG sites is stable across individuals, resulting in consistently strong correlations [6].



(a) CpG-wise variance distributions.

(b) Sample correlation heatmap.

Figure 3: Variance and correlation structure across Normal, Adjacent, and Tumor samples.

Low-dimensional embeddings

Dimensionality reduction (Figures 4–6) reveals a clear separation between Normal and Tumor samples. Adjacent samples consistently populate a transitional manifold between these two groups, supporting the hypothesis of a continuous epigenetic gradient rather than an abrupt transition [3].

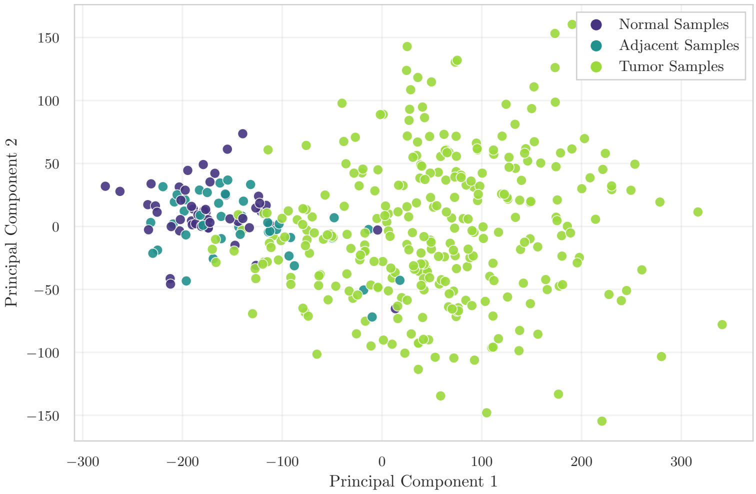


Figure 4: PCA projection.

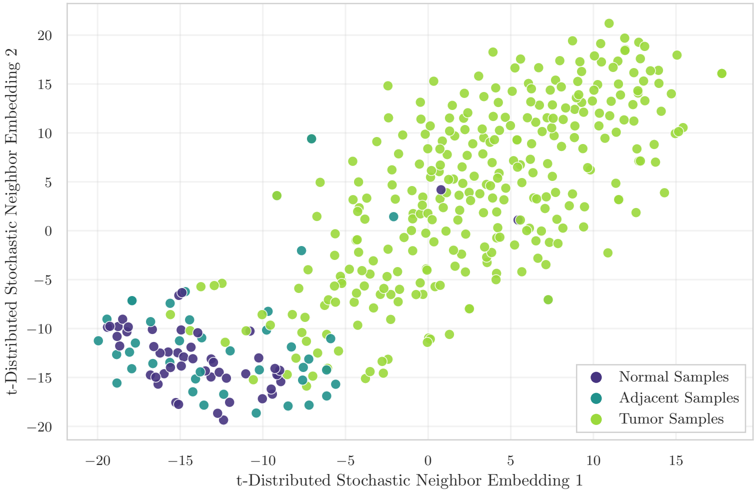


Figure 5: t-SNE embedding.

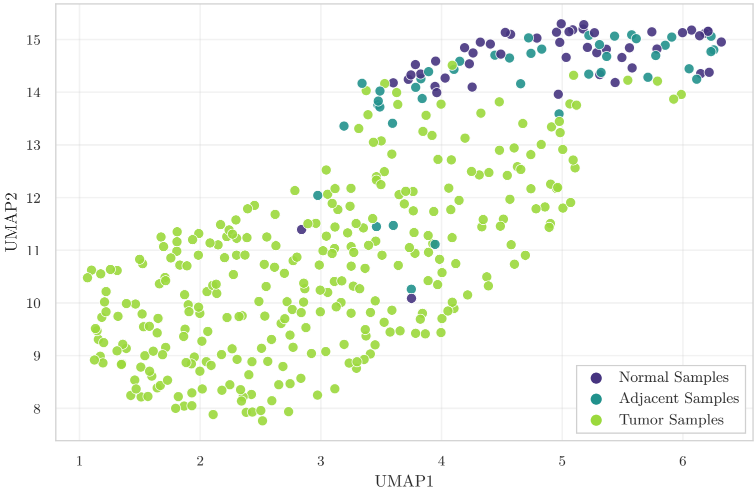


Figure 6: UMAP embedding.

Genome annotation coverage

Coverage across genomic contexts (islands, shores, shelves) matches the expected distribution for the 450K array, confirming correct manifest matching.

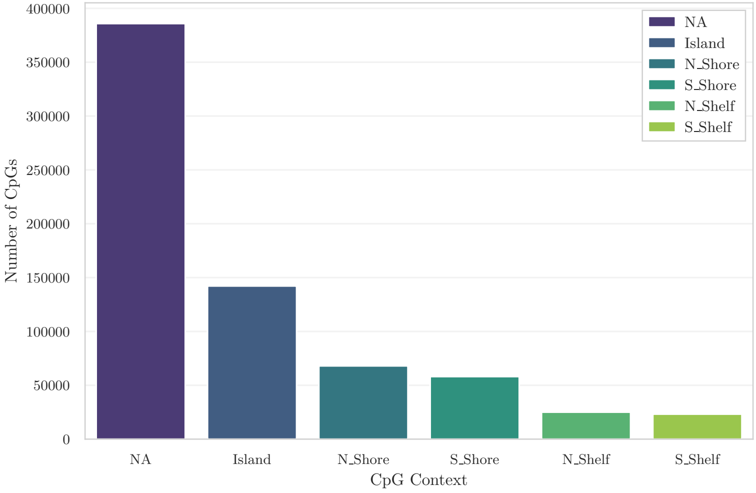


Figure 7: CpG context coverage.

Dynamic-network proxies

Although not a full Dynamic Network Biomarker (DNB) analysis, the degree-based proxies (Figures 8–10) show enhanced network instability in Tumors and early deviations in Adjacent samples.

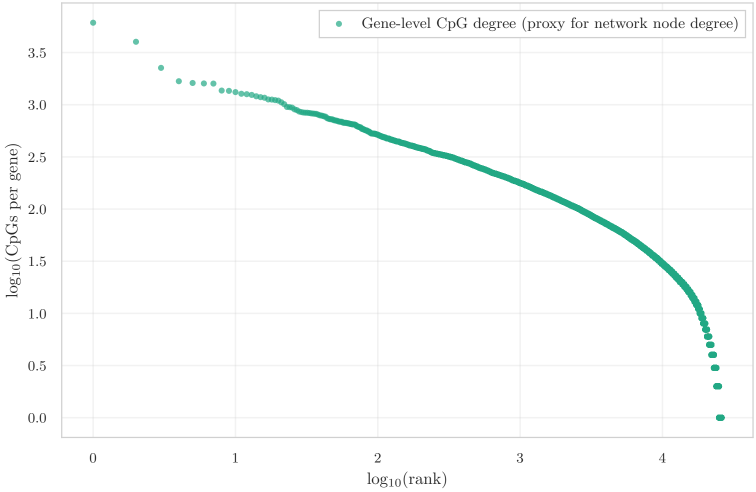


Figure 8: Gene-level CpG degree (proxy).

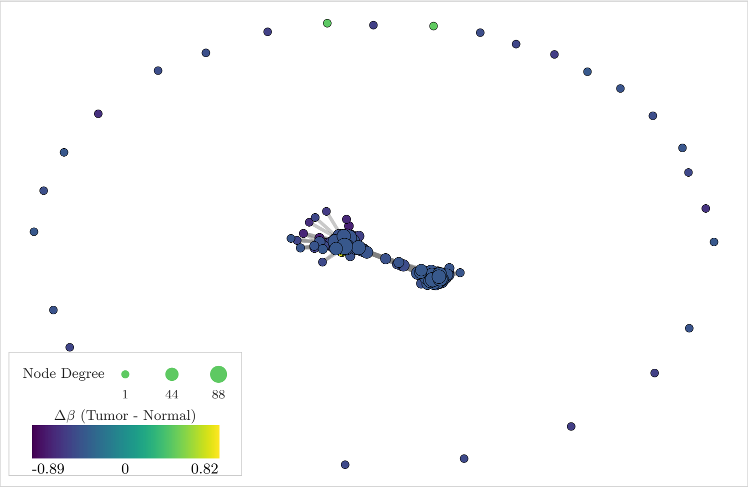


Figure 9: DNB-like network — Tumor vs Normal.

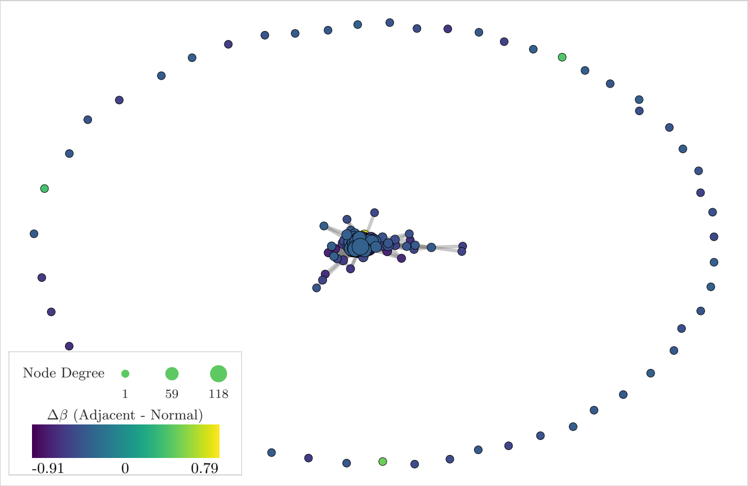


Figure 10: DNB-like network — Adjacent vs Normal.

Summary

The GSE69914 dataset exhibits a highly coherent and biologically interpretable structure. Across all analyses—global distributions, variability, outlier patterns, and embeddings— Tumor samples consistently show global hypomethylation, increased heterogeneity, and focal instability. Adjacent samples systematically occupy an intermediate position, consistent with early epigenetic alterations and the field-defect model. No technical issues were detected, making GSE69914 an excellent baseline dataset for subsequent comparative and modeling analyses.

1.2 GSE225845

1.3 GSE287331

2 Inter Dataset Analysis

References

[1] M. Ehrlich, "Dna methylation in cancer: Too much, but also too little," *Oncogene*, vol. 21, no. 35, pp. 5400–5413, 2002. DOI: 10.1038/sj.onc.1205651 [Online]. Available: https://doi.org/10.1038/sj.onc.1205651

[2] A. Bird, "Dna methylation patterns and epigenetic memory," *Genes & Development*, vol. 16, no. 1, pp. 6–21, 2002. DOI: 10.1101/gad.947102 [Online]. Available: https://doi.org/10.1101/gad.947102

[3] 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. DOI: 10.1038/ncomms10478 [Online]. Available: https://doi.org/10.1038/ncomms10478

[4] National Cancer Institute, *Brca gene changes: Cancer risk and genetic testing*, https://www.cancer.gov/about-cancer/causes-prevention/genetics/brca-fact-sheet, Accessed October 2025, 2024.

[5] 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. DOI: 10.1186/gb-2012-13-6-r44 [Online]. Available: https://doi.org/10.1186/gb-2012-13-6-r44

[6] A. E. Teschendorff, T. W. Zhuang, J. W. Breeze, and S. Beck, "A correlation-based network analysis approach for the identification of differential dna methylation interactions in cancer," *Genome Biology*, vol. 14, no. 10, R120, 2013. DOI: 10.1186/gb-2013-14-10-r120 [Online]. Available: https://doi.org/10.1186/gb-2013-14-10-R120