# POLITECNICO DI TORINO

**MASTER's Degree in Mathematical Engineering**


**MASTER's Degree Thesis**

# A Bayesian Cluster-Based Framework for Detecting Epigenetic Field Defects in Breast Cancer

**Supervisors**

**Prof. Alfredo BENSO**

**Prof. Roberta BARDINI**

**Dott. Sandro GAMBINO**

**Candidate**

**Eleonora GUARNASCHELLI**

**July 2026**

**Abstract**

Studies of early carcinogenesis suggest that epigenetic changes, in addition to genetic ones, contribute to tumor initiation. DNA methylation, defined as the covalent addition of a methyl group to cytosine residues within cytosine-phosphate-guanine (CpG) dinucleotides, is heavily involved in the regulation of gene activation and has been proposed as a potential biomarker for the early detection of breast cancer owing to the reversible nature of epigenetic modifications. The ultimate goal of this work was the identification of CpG sites that most strongly characterize the "early field defect", under the field cancerization hypothesis, according to which tissue adjacent to the tumor lesion, although apparently normal, may exhibit early epigenetic alterations. The analysis was conducted across three independent Gene Expression Omnibus (GEO) datasets comprising Normal, Adjacent, and Tumor tissue samples profiled on the Illumina 450K and EPIC BeadChip arrays. Standard epigenome-wide association studies (EWAS) test differences in mean methylation or variance between groups independently at each CpG site, thereby incurring a substantial multiple-testing burden and, most importantly, overlooking coordinated co-methylation patterns among neighboring CpGs. This raises the question of whether tumor development disrupts the local co-methylation patterns of healthy tissues, which this thesis aims to address. To account for such patterns, co-methylation clusters were identified in Normal tissue using SACOMA, which groups CpGs according to both genomic proximity and methylation similarity, providing a representation of the baseline epigenetic architecture of healthy breast tissue. Building on this cluster representation, the analysis proceeds through a Bayesian framework based on the Normal–Inverse-Wishart conjugate model for the identification and characterization of cluster-level methylation alterations. For each cluster, the co-methylation structure estimated from Normal samples was encoded as prior information and subsequently updated using methylation data from Adjacent and Tumor samples, separately. Cluster perturbation was quantified through a Kullback–Leibler-based decomposition, yielding two complementary and interpretable metrics for each comparison: a mean-shift component, measuring changes in average methylation levels, and a covariance-disruption component, capturing alterations in the underlying co-methylation architecture. These disruption metrics were integrated within a cluster-level feature-selection procedure designed to prioritize regions exhibiting consistent alterations across both the Normal–Adjacent and Normal–Tumor comparisons. The selected features were indirectly assessed through their ability to enable an XGBoost classifier to discriminate between Normal and Adjacent samples, and validated on an internal EPIC test set and a fully independent Illumina 450K external cohort. AUROC values of 0.86 and 0.75 were

achieved in the best configuration, despite the biological heterogeneity of adjacent tissue and cross-platform transferability challenges. Compared with established EWAS approaches such as iEVORA and limma, the proposed framework achieved superior external validation performance with a substantially smaller set of CpG sites, suggesting improved cross-dataset and cross-platform generalizability. Overall, these results demonstrate that a cluster-level co-methylation perspective yields an interpretable and generalizable framework for breast cancer epigenetic research.

*Alla mia famiglia.*

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

Breast cancer is the second most commonly diagnosed cancer worldwide and the leading cause of cancer-related death among women globally [1]. According to WHO estimates, in 2024, breast cancer caused approximately 694,000 deaths globally and represented the most common cancer among women in 164 out of 186 countries.

Breast cancer is characterized by the uncontrolled proliferation of abnormal cells in the breast, which can progressively develop into tumours and invade surrounding tissues. In its earliest stage, known as carcinoma in situ, the disease remains confined to its site of origin and is not considered life-threatening. However, when cancer cells acquire the ability to invade adjacent breast tissue, they can develop into invasive tumours that may manifest as palpable lumps or areas of thickening. If left undetected and untreated, these tumours can disseminate to distant organs and become potentially fatal. Therefore, effective strategies for early and timely diagnosis are essential to reduce breast cancer mortality.

According to the Field Cancerization Hypothesis, tumour development is a multistage process in which genetic and epigenetic alterations progressively accumulate over time. Thus, the transformation of a normal cell into a cancer cell generally requires the acquisition of multiple mutations and/or epigenetic changes affecting specific classes of genes. The progressive accumulation of these molecular alterations contributes to increasing tumour severity, ultimately leading to invasive disease and, in the absence of treatment, death.

Importantly, molecular alterations associated with carcinogenesis may occur before the emergence of a macroscopically evident tumour. In this context, apparently normal tissue adjacent to the tumour may already harbour early molecular changes related to the cancerization process. Within the framework of the epigenetic field defect hypothesis, it has been suggested that cells carrying an increased burden of epigenetic alterations may represent a pre-malignant field with an increased susceptibility to tumour development. Consequently, adjacent tissue should not be considered biologically equivalent to healthy normal tissue, but rather as a

potentially altered environment reflecting early events of cancer development. This interpretation is supported by previous studies investigating DNA methylation alterations in adjacent breast tissue [2].

The potential of tumour-adjacent tissue as a source of information for early detection is therefore substantial. However, while differences between normal and tumour tissue are often evident through histological examination, adjacent tissue generally appears indistinguishable from normal tissue at the morphological level. Therefore, molecular analyses are required to identify subtle alterations that may provide information about the earliest stages of carcinogenesis. In particular, differences between normal and adjacent tissue may not be detectable through conventional histological assessment but can emerge through the analysis of molecular changes, including epigenetic alterations such as DNA methylation patterns. Investigating these methylation changes represents the main focus of this thesis.

DNA is composed of four nucleotide bases: cytosine (C), guanine (G), adenine (A), and thymine (T). When a cytosine is followed by a guanine on the same DNA strand, the resulting dinucleotide is referred to as a CpG site. DNA methylation occurs when a methyl group is added to the cytosine within a CpG site, leading to the formation of a methylated CpG.

CpG sites located in close genomic proximity tend to display correlated methylation levels, a phenomenon known as co-methylation. This coordinated methylation pattern generally decreases as the genomic distance between CpG sites increases. Although DNA methylation does not modify the underlying DNA sequence, it plays a crucial role in regulating gene activity by influencing chromatin accessibility and transcriptional states. Consequently, alterations in DNA methylation patterns can disrupt normal gene regulation and have been widely associated with tumour development and progression.

The major biological question addressed in this thesis is whether alterations in DNA methylation patterns across disease groups can be exploited to identify potential molecular biomarkers for early detection of breast cancer. To address this question, this thesis develops a computational framework aimed at identifying a reduced subset of informative CpG regions characterized by epigenetic alterations associated with early stages of tumour development.

DNA methylation profiling generates high-dimensional datasets containing information from hundreds of thousands of CpG sites simultaneously. For example, widely used platforms such as the Illumina HumanMethylation450 BeadChip interrogate more than 450,000 methylation sites across the genome. The large scale and complexity of these datasets require computational approaches capable of identifying biologically meaningful patterns while reducing dimensionality.

Several statistical and computational strategies have been proposed for the analysis of DNA methylation data. Among these, region-based approaches aim to

overcome the limitations of single-CpG analyses by considering groups of correlated and genomically close CpG sites, relying on the comethylation property. Within this framework, this thesis focuses on an unsupervised region-based approach, in which CpG clusters are first defined according to a Normal reference methylation structure and subsequently investigated across disease groups through a spatially-aware clustering strategy.

The central biological hypothesis of this thesis is that tumour progression disrupts normal co-methylation patterns through two complementary mechanisms: alterations in the average methylation profile of CpG clusters and modifications of the dependency structure among neighbouring CpG sites. In other words, cancer-associated epigenetic alterations may affect not only the methylation level of individual CpGs, but also the coordinated relationships that define local methylation architectures.

To investigate this hypothesis, this thesis proposes a hierarchical Bayesian framework based on the Normal–Inverse–Wishart conjugate model. The co-methylation architecture of each CpG cluster, jointly represented by its mean vector and covariance matrix, is learned from normal tissue and used as prior information. This reference model is then updated using adjacent and tumour samples, allowing the quantification of cluster-level disruption through the divergence between the prior and posterior distributions. By decomposing this disruption into a mean-shift component and a covariance-structure component, the framework distinguishes between changes in average methylation levels and alterations of the underlying co-methylation organization.

Quantifying cluster-level disruption at each comparison, however, only partially addresses the problem of identifying candidate biomarkers. A second, equally important question must be answered: which type of disruption pattern is more likely to represent a genuine early biomarker signal?

Two competing selection hypotheses can be formulated in response to this question. Under Hypothesis A, the most informative CpG clusters are those exhibiting the strongest disruption between Normal and Adjacent tissue, considered in isolation. This strategy is consistent with the standard approach adopted in the field-cancerization literature, exemplified by the iEVORA framework [2], in which candidate loci are first identified based on their extreme behaviour in the Normal–Adjacent comparison, and only subsequently examined, as a post-hoc validation step, for consistent amplification in tumour tissue.

Under Hypothesis B, the most informative clusters are instead those exhibiting a coherent, progressive disruption trajectory across the N$\rightarrow$A$\rightarrow$T axis, rather than an extreme disruption confined to a single stage. This alternative formulation follows directly from the field cancerization hypothesis itself: if Adjacent tissue genuinely represents an intermediate stage of tumour progression, a credible early biomarker should display early, moderate disruption in Adjacent tissue and a

correspondingly amplified disruption in Tumour tissue — rather than being defined by extreme disruption at either stage alone. Hypothesis B therefore reverses the standard workflow adopted in the literature: rather than selecting clusters based on their extremeness in the Normal–Adjacent comparison and subsequently verifying amplification in tumour as a validation step, the progressive-trajectory pattern itself, evaluated jointly across both comparisons, is used directly as the selection criterion.

Because no ground truth exists regarding which CpGs are genuinely implicated in the early stages of breast cancer development, the biomarker potential of a candidate CpG subset cannot be established directly, but only indirectly, through two complementary computational criteria.

The first criterion is classification performance on the Normal versus Adjacent task, rather than Normal versus Tumor. This choice reflects the clinical relevance of each comparison: distinguishing Tumor from Normal tissue is a comparatively straightforward task, since tumour tissue is already histologically identifiable and molecularly divergent from normal tissue at this stage, so high classification performance on this comparison provides limited insight into the genuine biomarker potential of a candidate subset. Distinguishing Adjacent from Normal tissue is, by contrast, a substantially harder task, precisely because the two tissue types are histologically indistinguishable and the underlying molecular differences are highly stochastic and heterogeneous across samples. It is this task, rather than Normal versus Tumor, on which the clinical value of an early-detection biomarker ultimately depends.

The second criterion is cross-cohort — and, in this thesis, also cross-platform — generalization. Given the pronounced heterogeneity that characterizes the epigenetic signal of Adjacent tissue, a feature selection strategy that merely maximizes performance on the training cohort risks capturing cohort-specific artifacts rather than a genuine biological signal. The ability of a candidate subset to generalize to an independent cohort is a property more consistent with a robust and clinically transferable biomarker candidate, and is therefore adopted as a complementary, and arguably more stringent, validation criterion.

The remainder of this thesis is organized as follows. Chapter 2 introduces the biological background necessary to contextualize the analyses developed in this work. Chapter 3 reviews the existing literature on epigenome-wide association study approaches, at both the CpG and region level. Chapter 4 describes the DNA methylation profiling platforms used in this thesis and details the datasets employed, together with the initial preprocessing pipeline, which reduces the initial set of approximately 450,000 CpGs to the 149,147 CpGs retained for downstream analysis. Chapter 5 presents the proposed methodological framework, including cluster definition, the Normal–Inverse–Wishart Bayesian model, the operational use of the resulting disruption metrics, and the selection of the classification algorithm

adopted for external validation. Chapter 6 reports and discusses the results obtained. Chapter 7 summarizes the main findings of this thesis and outlines directions for future work.

# Chapter 2

# Biological Background

This chapter introduces the biological background necessary to interpret the methodological framework developed in this thesis. Section 2.1 presents the field cancerization hypothesis, which provides the conceptual and clinical rationale for using tumor-adjacent tissue as a surrogate for premalignant tissue. Building on this rationale, the chapter then turns to DNA methylation, describing its organization in the normal genome and its disruption in cancer. Finally, Section 2.3 introduces co-methylation, the spatial dependency between neighboring CpG sites that motivates the region-based, cluster-level approach adopted throughout this thesis.

## 2.1 Field Cancerization Hypothesis

The initiation of a solid tumor is widely believed to result from a multistep process driven by the sequential accumulation of genetic mutations or transmissible epigenetic alterations, known as epimutations. These events are thought to drive early tumorigenesis by conferring a proliferative advantage to specific cells relative to their surrounding counterparts within the tissue, thereby promoting clonal expansion and the progressive replacement of neighboring cell populations under the action of natural selection [3, 4]. Consequently, tissues harboring an increased burden of epigenetic alterations may represent premalignant states from which new cancers are likely to arise.

The identification and characterization of such alterations are therefore of considerable clinical interest: early-stage detection and treatment (stages I and II) is associated with a markedly higher 5-year survival rate (>93%) compared to late-stage diagnosis (stage IV: 22%) [5], underscoring the clinical value of identifying molecular signals capable of revealing the earliest stages of carcinogenesis before further malignant progression occurs.

However, the study of premalignant evolution in humans is intrinsically challenging, since longitudinal samples tracking the same tissue from a normal state through malignant transformation are rarely available. Moreover, access to the normal cells that originally occupied the sites from which tumors eventually arise is inherently impossible. Consequently, direct observation of the earliest stages of carcinogenesis is generally not feasible. These limitations, together with the growing recognition of the susceptibility of normal tissue to acquire early genetic and epigenetic alterations following exposure to carcinogens, motivated the formulation of the field cancerization hypothesis, first proposed by Slaughter and colleagues in 1953 [6]. According to this hypothesis, a "field defect" or "field of cancerization" is a region of tissue that appears histologically normal but harbors molecular alterations reflecting a succession of premalignant events due to its proximity to a tumor.

Field defects are of particular interest because they may provide a unique window into the earliest stages of carcinogenesis, which are otherwise inaccessible to direct observation in humans, and may ultimately serve as biomarkers of cancer risk and early detection. Consequently, the field cancerization hypothesis provides a rationale for identifying early molecular alterations through the comparison of tissues from healthy individuals and histologically normal tissues adjacent to tumors, the latter being commonly used as a surrogate for premalignant tissue [7], [8].

This thesis builds on this rationale, focusing on breast cancer tissue and working with Normal and Adjacent-to-Tumor samples, with the goal of characterizing altered epigenetic patterns that distinguish the two tissue types

## 2.2 The role of DNA methylation in Cancer

DNA methylation, i.e. the addition of a methyl group (-CH3) via a covalent bond to cytosine residues within the cytosine-phosphate-guanine (CpG) dinucleotides, represents one of the most extensively studied epigenetic mechanisms in cancer. In mammals, DNA methylation is essential for normal development and is associated with a number of key processes, including genomic imprinting, X-chromosome inactivation, repression of transposable elements, and tissue-specific gene regulation. Since it is a mechanism of epigenetic modification, it does not cause a change in the nucleotide sequence, unlike genetic alterations, but instead modulates gene expression by influencing chromatin accessibility and transcriptional activity.

While these methylation patterns are tightly regulated in healthy tissues, their dysregulation has been implicated in disease progression, including cancer [9] [10].

To appreciate how methylation patterns become disrupted in cancer, it is useful to first consider their organization in the healthy genome. CpG sites are not uniformly distributed across the genome: certain regions, known as CpG islands

(CGIs), are characterized by a locally elevated density of CpG dinucleotides and are typically found in an unmethylated state in normal tissue, particularly when located in the promoter regions of genes. These islands extend approximately 0.5–3 kb and occur on average every 100 kb throughout the genome [11]. Flanking each island are the so-called shores (regions within 2 kb of a CGI) and shelves (regions within 2–4 kb), while the remaining CpG sites dispersed throughout the genome are referred to as Open Sea.

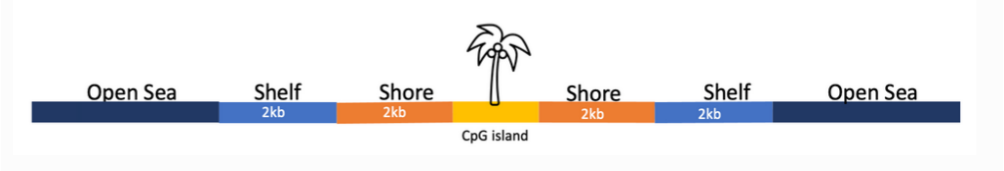


**Figure 2.1:** Genomic Regions along the genome

In Figure 2.1 a graphical representation of this regions is provided.

Outside of CGIs, CpG sites are generally methylated in normal tissue, a pattern that contributes to genome stability by suppressing the activity of transposable elements and maintaining heterochromatic regions.

In cancer, this orderly landscape is profoundly disrupted. Two hallmark alterations have been described: focal hypermethylation of CGIs located in the promoter regions of tumor suppressor genes, leading to transcriptional silencing and loss of growth regulatory control [5]; and global hypomethylation, initially underappreciated but subsequently demonstrated through high-throughput profiling to affect large portions of the genome. Crucially, a significant proportion of cancer-associated methylation changes do not occur at CGIs themselves, but rather at shores, shelves, and Open Sea regions, suggesting that the spatial relationship between a CpG site and its genomic context — whether located in a gene promoter, gene body, intergenic region, or in proximity to a CGI — is an important determinant of its susceptibility to aberrant methylation in cancer. Furthermore, it has been shown that many CGIs acquiring aberrant methylation in tumors are not associated with classical tumor suppressor genes, indicating that the landscape of cancer-related methylation alterations is considerably broader than initially recognized [11].

Crucially, since DNA methylation is an epigenetic modifications, it is also potentially reversible, making it a particularly attractive therapeutic target. In fact, it is in principle possible to reactivate genes that have been aberrantly silenced, thereby halting or reversing disease progression. This reversibility has motivated the development of epigenetic therapies aimed at restoring normal gene expression programs, and two demethylating agents — azacitidine (Vidaza) and decitabine — have subsequently received United States Food and Drug Administration (FDA) approval, highlighting the medical promise of mapping and understanding the role

of DNA methylation [12].

However, despite the growing evidence linking aberrant DNA methylation to cancer, the extent to which methylation alterations contribute to field defects and characterize premalignant tissue states remains poorly explored, providing the main motivation for the analyses presented in this thesis.

## 2.3 The role of co-methylation

DNA methylation does not occur independently at individual CpG sites. Instead, correlated methylation of neighboring CpG sites across the genome had been recognized and denominated as comethylation phenomena. Early studies reported that methylation levels of adjacent CpGs are positively correlated, such that methylation at one site increases the probability that nearby sites are methylated as well. This phenomenon, commonly referred to as *co-methylation*, reflects the spatial dependence of DNA methylation along the genome [13]. Generally speaking, there are 2 main types of co-methylation: between-sample (BS) co-methylation and within-sample (WS) co-methylation. The BS comethylation describes the methylation similarity or correlation of CG sites across a set of samples and in different genomic locations. WS co-methylation describes the degree of methylation over distance, ie, similar methylation patterns in nearby CG sites located in the same chromosome of 1 single sample.

The biological basis of co-methylation lies in the fact that nearby CpGs frequently share the same chromatin environment. CpGs located within promoters, enhancers, CpG islands (CGIs), or gene bodies are exposed to common regulatory influences, including chromatin accessibility, nucleosome positioning, histone modifications, and transcription factor binding. As a consequence, methylation changes tend to occur in a coordinated manner across short genomic intervals rather than at isolated CpG sites. Such coordinated patterns are biologically meaningful because they often reflect the functional state of the underlying genomic region and contribute to the regulation of gene expression and maintenance of genome stability.

Several studies have investigated how methylation correlation varies as a function of genomic distance. A consistent observation is that co-methylation extends over genomic regions spanning approximately 1-2 kb, with correlation generally decreasing as the distance between CpGs increases [14, 15]. However, the strength and persistence of this correlation strongly depend on genomic context.

In particular, [15] showed that CpGs located within CpG islands exhibit remarkably stable co-methylation patterns. Correlation remains high even for CpGs separated by distances approaching 1 kb, often exceeding 0.6. This behavior likely reflects the high CpG density and shared regulatory environment characteristic of CGIs, which promotes coordinated epigenetic regulation across extended genomic

segments. In contrast, CpGs located in Open Sea regions display a much more rapid decay of correlation with increasing distance. In these regions, methylation correlation is generally lower even at short distances and decreases substantially within a few hundred base pairs (approximately 200-400 bp). Similar trends were observed in other low-CpG-density genomic contexts such as intergenic and repetitive regions. These findings suggest that the spatial organization of co-methylation is strongly shaped by local genomic architecture.

Recent studies suggest that tumorigenesis is accompanied not only by alterations in methylation levels at individual CpGs but also by changes in the correlation structure linking neighboring sites [13, 16]. The breakdown, reorganization, or emergence of co-methylation domains may therefore provide additional information about epigenetic deregulation that cannot be captured by single-CpG analyses alone. For this reason, investigating how co-methylation patterns differ between normal and adjacent to tumor tissues may offer valuable insights into the mechanisms through which cancer reshapes the epigenetic landscape.

Taken together, these observations suggest that DNA methylation should not be viewed as a collection of independent CpG measurements, but rather as a coordinated regional process. This motivates the cluster-based modeling strategy adopted throughout this thesis.

# Chapter 3

# EWAS: Literature Overview

This chapter aims to present the existing approaches to address the search for relevant epigenetic alterations between conditions. Epigenome-wide association study (EWAS) has been applied to analyze DNA methylation variation in complex diseases for a decade. EWAS aims to use a variety of microarray-based or sequencing-based analysis techniques to obtain the association between epigenetic markers and phenotypes, which can ultimately explain the cause of the disease better and promote the development of new therapies and diagnostic method [17]. The commonly used EWAS analysis process usually starts with a reasonable hypothesis. Then a suitable population, tissue sample and corresponding DNA methylation dataset is selected. DNA methylation analyses can be broadly categorized according to the biological scale at which methylation alterations are investigated. Existing approaches can be grouped into two main families: single-CpG methods, which evaluate each CpG site independently, and region-based methods, which exploit the spatial dependency between neighboring CpG sites. The remainder of this chapter reviews the major methods within each of these two families.

## 3.1 CpG-level approaches

Within the CpG-level family, two complementary paradigms have emerged: Differential Methylation (DM), which focuses on differences in average methylation levels, and Differential Variability (DV), which instead investigates changes in methylation variability across samples. Therefore, while DM methods detect changes in the expected methylation level, DV approaches investigate whether disease progression alters the stability of methylation states across individuals.

### 3.1.1 DM methods

To date, the main focus when analysing DNA methylation data has been detecting CpG sites that are differentially methylated between groups in terms of mean levels. Several methods exist to conduct this kind of analysis, for example a non parametric Wilcoxon rank sum test, used in methyAnalysis package, or a simple t-test, used for example in IMA package. Other methods are reviewed in [18]. In this chapter the focus is on the one implemented in the `limma` package, provided by the `lmFit` and `eBayes` functions [19], which serves as the benchmark method used throughout this thesis for comparison. It works in the following way. For each locus $i$, a linear model is first fitted to estimate $\beta_i^*$, the vector of coefficients characterizing methylation differences between groups. Denoting by $y_i^T = (y_{i_1}, \ldots, y_{i_n})$ the methylation values at locus $i$ across both groups, with $n = n_1 + n_2$, the model takes the form

$$E(y_i) = X\beta_i^*, \tag{3.1}$$

where $X$ is a full-column-rank design matrix and $\beta_i^*$ is the coefficient vector. This vector is composed of $(\beta_{i0}^*, \beta_{i1}^*)$, where $\beta_{i0}^*$ represents the mean methylation level in the normal group and $\beta_{i1}^*$ represents the difference in mean methylation between the tumor and normal groups. The null hypothesis of interest is therefore $H_0 : \beta_{i1}^* = 0$ for each locus $i = 1, \ldots, m$. To test $H_0$, a moderated $t$-statistic is used, combining classical and Bayesian shrinkage of the variance estimates:

$$\tilde{t}_{ij} = \frac{\hat{\beta}_{ij}^*}{\tilde{s}_i \sqrt{v_{ij}}}. \tag{3.2}$$

The corresponding $p$-value for $H_0 : \beta_{ij}^* = 0$ is obtained from a $t$-distribution with $d_i + d_0$ degrees of freedom, where $d_0$ and the moderated residual variance $\tilde{s}_i$ arise from the empirical Bayes shrinkage step, and $v_{ij}$ is the corresponding scaling factor from the design matrix.

### 3.1.2 DV methods

DM methods have been conducted to compare Normal and Tumor samples, which are affected by more evident differences all over the genome, but when applied for a more difficult task, such as the comparison of Normal and Adjacent samples, they led to no clear biological conclusions [2]. To address this specific task, another approach was proposed, based on the following considerations:

1. Cancer is a heterogeneous disease. Within each type of cancer there is the potential for tumour growth to be driven by perturbations in many different molecular pathways, and these perturbations will vary from individual to individual. This heterogeneity also affects adjacent-to-tumor tissue.

2. The resulting alterations arise stochastically: they do not necessarily affect the same CpGs or CpG regions across different individuals, at least in the early stages [2].

For these reasons, the differential variability approach was explored. Several methods have been proposed for detecting differential variability in DNA methylation data, including DiffVar [20], as well as approaches based directly on Bartlett's test or Levene's test for the homogeneity of variances. Among these, iEVORA [2] was selected as a benchmarking method for the results obtained in this thesis, as it was specifically developed to address the comparison between Normal and Adjacent samples. Notably, it was originally validated on GSE69914, which is also used as the external validation cohort in this work. iEVORA leads to the definition of differentially variable and differentially methylated CpGs (DVMCs). The method consists in conducting, at the single-CpG level, a Bartlett's test to identify CpGs whose DNA methylation patterns are differentially variable between normal samples and normal samples adjacent to tumours. CpGs that passed a stringent FDR $< 0.001$ threshold, as estimated with the q-value, were designated differentially variable CpGs (DVCs). Because Bartlett's test is overly sensitive to single outliers, the procedure was regularized by re-ranking DVCs according to their differential DNA methylation t-statistic, and selecting those with a t-test unadjusted P-value of less than 0.05. The resulting subset of DVCs is then termed DVMCs. Within this subset, CpGs are ranked according to their differential methylation $t$-statistic, so that those with the largest $t$-statistics – reflecting the strongest evidence of a consistent, non-outlier-driven difference – are ranked highest, a procedure that penalizes DVCs driven by only a few outliers.

## 3.2 Region-based approaches

While single-CpG methods such as those described above offer high resolution, they neglect the existence of co-methylation patterns between CpG loci. Current technology measures methylation at hundreds of thousands, or millions, of CpG sites throughout the genome, and neighboring CpG sites are often highly correlated with one another; current literature suggests that clusters of adjacent CpG sites are co-regulated. As a consequence, modest but coordinated methylation changes distributed across neighboring CpGs, which may nonetheless be biologically relevant, risk being overlooked or diluted under a strictly per-CpG testing framework.

Motivated by this limitation, the field has recently undergone a methodological shift toward region-based analysis, whereby information from contiguous CpGs is aggregated to identify differentially methylated regions (DMRs), rather than differentially methylated CpGs (DMCs) considered in isolation. Such region-level investigations not only mitigate the multiple testing burden, but also align more

closely with the biological reality that regulatory methylation changes frequently manifest across groups of neighboring CpGs rather than at isolated single sites.

Region-based methods can be broadly classified into two methodological frameworks: supervised and unsupervised.

Supervised frameworks first identify CpG sites that show significant associations with a phenotype or exposure, and subsequently combine adjacent significant sites into co-methylated regions – a phenotype-guided, or targeted, approach in which candidate regions are surveyed directly for association with the condition of interest [21]. In practice, a P-value or corresponding t-statistic is first computed at each CpG; because of the large number of probes on the array (almost half a million), multiple-comparison-adjusted P-values are typically computed for each probe first, and the genome is then scanned for regions enriched in significant adjusted P-values, using user-specified criteria such as the minimum number of CpGs required to define a region. The majority of these tools employ their own strategy to aggregate nearby CpGs, typically using genomic proximity together with p-value significance thresholds to define contiguous regions of coordinated change; examples include DMRcate [22] and Bump Hunting [23]. This phenotype-guided approach provides a practical and computationally efficient framework, particularly when strong site-level signals exist. However, its dependence on initial single-site testing limits its ability to detect subtle coordinated methylation changes that are distributed across several CpGs: such changes may not reach site-level significance individually, and therefore remain undetected during region formation.

Unsupervised methods, by contrast, first define co-methylated regions independently of any phenotype information, relying solely on intrinsic methylation correlation patterns and, in most cases, genomic distance among CpG sites. After regions have been defined an association analysis is performed. These models aim to uncover the natural structural organization of the methylome, reflecting the underlying chromatin and regulatory landscape, without bias from the outcome of interest. A particular use case of this family is the semi-targeted approach, in which regions are defined using one condition as a reference and subsequently investigated for consistency, or disruption, in a second condition [21]. Although unsupervised models avoid the limitations of supervised approaches, they can still fail to accurately capture the true co-methylation structure, either by grouping weakly correlated CpGs together or by fragmenting biologically coherent regions; the resulting regions can introduce noise, inflate false positives, or reduce statistical power in downstream region-based testing.

Within the unsupervised family, two further methodological currents can be distinguished. Some methods constrain region definition to a maximum genomic distance beyond which correlation is assumed to decay (e.g., Aclust2, SACOMA, coMethDMR, comeBack), while others rely purely on the correlation structure among CpGs, irrespective of genomic distance (e.g., WGCNA); under this latter

approach, modules of co-methylated CpGs can emerge even between sites that are genomically distant from one another. This thesis focuses on the former, spatially-constrained family of methods, as further discussed in Chapter 5.

### 3.2.1 Sacoma Algorithm

Having introduced the distinction between spatially-constrained and correlation-only unsupervised methods, a detailed description of SACOMA algorithm, the specific spatially-constrained algorithm adopted in this thesis for co-methylation cluster definition, is provided.

SACOMA (Spatially-Aware Clustering for Co-Methylation Analysis) is an unsupervised, data-driven framework for detecting co-methylated regions – genomic regions in which neighboring CpG sites show correlated methylation levels [24]. The algorithm frames region detection as a spatially constrained hierarchical clustering problem, in which candidate groupings of CpGs are learned by jointly accounting for how similar their methylation profiles are and how close they lie to one another along the genome, rather than relying on rigid, pre-fixed distance or correlation cutoffs as in earlier tools. This joint optimization is governed by a single, data-adaptive mixing parameter, estimated directly from the data rather than fixed a priori, which limits the number of hyperparameters that must be manually specified. Within the broader EWAS landscape, SACOMA therefore serves a dual role: it detects regions of co-regulated methylation, and it reduces the dimensionality of the original CpG-level data by aggregating sites into a smaller number of analytical units suitable for downstream region-based association testing.

SACOMA (Spatially-Aware Clustering for Co-Methylation Analysis) is a flexible, data-driven, and unsupervised framework designed to identify co-methylated regions, that is, genomic regions where adjacent CpG sites show correlated methylation levels, and in doing so capture biologically meaningful epigenetic patterns. The core idea is to formulate region detection in DNA methylation data as a hierarchical clustering problem that jointly minimizes within-cluster differences in terms of both methylation levels and genomic distance [24].

**Preprocessing and genomic binning** The algorithm starts from the Illumina array manifest file, which contains the genomic coordinates and probe annotations for each CpG site. CpG sites are first grouped by chromosome and ordered by genomic position, in order to preserve spatial continuity across loci. Sites located within a user-defined maximum distance are then merged into the same genomic bin $B_r$, forming contiguous CpG clusters that represent potential co-methylated regions. Bins containing fewer than a user-defined minimum number m of CpG sites are excluded to ensure sufficient CpG density and biological relevance.

**Construction of the dissimilarity matrices** For each bin containing n CpG sites with indices $I_r \in \{1, ..., n\}$, a n x n pairwise correlation matrix $C = [\rho_{ij}]$ is computed, where $\rho_{ij}$ denotes the correlation coefficient (Pearson or Spearman, as chosen by the user) between the methylation profiles of sites i and j across all samples. A user-defined threshold $r \in [0,1]$ is then applied to determine which CpG pairs show sufficient co-methylation to be considered part of the same region. This initial correlation matrix is then transformed into a binary adjacent matrix $A = [a_{ij}]$ where $a_{ij} = 1_{(|\rho|_{ij} \geq r)}$. To quantify the absence of co-methylation relationships between pairs CpG, the within cluster dissimilarity matrix in the methylation space $D_0 = [d_{ij}^{(0)}]$ is then computed by considering $d_{ij}^{(0)} = 1 - a_{ij}$. Simultaneously, the spatial dissimilarity matrix $D_1 = [d_{ij}^{(1)}]$ is constructed based on the genomic positions of the CpG sites, where each site is assigned a scalar coordinate given by the midpoint between start and end probe location, i.e. $d_{ij}^{(1)} = |g_i - g_j|$.

**The ClustGeo-based clustering criterion** These two matrices are the input of ClustGeo, an hierarchical clustering algorithm [25]. The ClustGeo algorithm relies on a convex combination of two dissimilarity matrices, one representing methylation similarity and another representing spatial proximity, thus enabling a soft, tunable integration of spatial information while preserving the interpretability and hierarchical nature of clustering. Building $D_0$ provides dissimilarities in the feature space, i.e. the methylation levels space, while $D_1$ provides dissimilarities in the constraint space, i.e. spatial contiguity space. In order to understand how ClustGeo works, it is important to define some fundamental quantities. Given the current partition in $K$ clusters $P_K$, the pseudo-inertia of a cluster $C_k$ generalizes the classical Ward inertia to the case of non-Euclidean dissimilarities:

$$I(C_k) = \sum_{i \in C_k} \sum_{j \in C_k} \frac{w_i w_j}{2\mu_k} d_{ij}^2 \tag{3.3}$$

where $\mu_k = \sum_{i \in C_K} w_i$ is the total weight of cluster $C_k$. The smaller $I(C_k)$ is, the more homogeneous are the observations belonging to the cluster $C_k$. This general definition can be then apply on both $D_0$ and $D_1$ in order to define $I_0(C_k)$, which is the within-cluser dispersion in the feature space, and $I_1(C_k)$, which is the within-cluster dispersion in the spatial dissimilarity space.

Starting from these concepts, we can generalize the formula 3.3 in order to take into accounts our two conditions, one on methylation similarity the other on geographical closeness, and define the mixed pseudo inertia of the cluster $C_K$ for a fixed mixing parameter $\alpha \in [0,1]$ as a convex combination of the pseudo inertia associated with $D_0$ and $D_1$:

$$I_\alpha(C_k) = (1 - \alpha) I_0(C_k) + \alpha I_1(C_k), \tag{3.4}$$

where the mixing parameter $\alpha$ controls the trade-off between methylation homogeneity and spatial compactness. When $\alpha$ increases, the homogeneity calculated with $D_0$ decreases whereas the homogeneity calculated with $D_1$ increases. The core idea of the algorithm is to determine an optimal value of $\alpha$ which increases the spatial-contiguity without deteriorating too much the quality of the solution on the variables of interest.

Another important quantity we have to define is the pseudo within-cluster inertia of the partition $P_k$:

$$W(P_k) = \sum_{k=1}^{K} I(C_k)$$

The smaller this pseudo within-inertia $W(P_K)$ is, the more homogenous is the partition in K clusters.

For the same reason as before, it is possible to generalize the previous quantity by defining the mixed pseudo within-cluster inertia of a partition $P_K^\alpha$ as follows:

$$W_\alpha(P_K^\alpha) = \sum_{k=1}^{K} I_\alpha(C_k)$$

**The agglomerative procedure** At each step of the algorithm, for a fixed $\alpha$, two clusters $A$ and $B$ from the current partition $P_{k+1}^\alpha$ are merged into a single cluster to obtain a new partition $P_k^\alpha$ in K clusters, such that $P_k^\alpha$ shows the minimun increase in the mixed within-cluster inertia (heterogeneity, variance). In fact, the optimization problem can be written in the most general form as:

$$argmin_{A,B \in P_{k+1}^\alpha} W_\alpha(P_k^\alpha)$$

where $P_k^\alpha = P_{k+1}^\alpha \setminus \{A, B\} U \{AUB\}$ and

$$W_\alpha(P_k^\alpha) = W_\alpha(P_{k+1}^\alpha) - I_\alpha(A) - I_\alpha(B) + I_\alpha(AUB)$$

If we define the aggregation measure between two cluster

$$\delta_\alpha(A, B) = I_\alpha(AUB) - I_\alpha(A) - I_\alpha(B)$$

this is the actual quantity that we want to minimize at each step of the hierarchical clustering algorithm, since $W_\alpha(P_{k+1}^\alpha)$ is fixed for a given partition $P_{k+1}^\alpha$. This quantity, which we want to minimize over the set of all possible clusters that can be merged in one, can be interpreted as the increase of the mixed within-cluster inertia (loss of homogeneity caused by the merger), since

$$\delta_\alpha(A, B) = W_\alpha(P_k^\alpha) - W_\alpha(P_{k+1}^\alpha)$$

The algorithm is initialized with $K = n$ singleton clusters and it assume as stopping criterion $K = 1$, meaning that we obtain partition $P_1$ containing n observations. The hierarchically-nested set of such partitions $\{P_n^\alpha, ..., P_K^\alpha, ..., P_1^\alpha\}$ can be represented graphically by dendogram, where the height of a cluster $C = AUB$ is $h_\alpha(C) = \delta_\alpha(A, B)$. The aggregation measures between the new cluster $AUB$ and any cluster $D$ of $P_{k+1}$ are calculated at each step thanks to the well-known Lance and Williams equation:

$$\begin{aligned}\delta_\alpha(A \cup B, D) = &\frac{\mu_A + \mu_D}{\mu_A + \mu_B + \mu_D}\,\delta_\alpha(A, D) \\ &+ \frac{\mu_B + \mu_D}{\mu_A + \mu_B + \mu_D}\,\delta_\alpha(B, D) \\ &- \frac{\mu_D}{\mu_A\mu_B + \mu_D}\,\delta_\alpha(A, B) \end{aligned} \tag{3.5}$$

**Selection of the optimal mixing parameter** SACOMA applies ClustGEO across a predefined grid $\{\alpha_1, ..., \alpha_J\}$. For each $\alpha_j$, the resulting dendogram is cut into $K = [|n|/3]$ clusters, in order to define the partition $P_K^\alpha$. The quality of each resulting partition is then assessed through the proportion of explained pseudo-inertia computed separately in the feature and spatial spaces:

$$Q_0(\alpha_j) = 1 - \frac{W_0(P_K^{\alpha_j})}{W_0(P_1^{\alpha_j})}, \quad Q_1(\alpha_j) = 1 - \frac{W_1(P_K^{\alpha_j})}{W_1(P_1^{\alpha_j})}$$

The optimal mixing parameter is then selected as

$$\alpha^* = argmin_{\alpha_j} |Q_0(P_K^{\alpha_j}) - Q_1(P_K^{\alpha_j})|$$

Using the selected $\alpha^*$, a final hierarchical clustering is performed to generate a dendrogram. This dendrogram is then cut at height zero, corresponding to the lowest possible aggregation level, to extract all subclusters formed during the agglomerative process. Among these, SACOMA retains only those clusters that contain $> 2$ CpG sites. If no such clusters are found, the bin is classified as noise/non-co-methylated; otherwise, all retained clusters are labeled as signal/co-methylated regions. These steps are repeated independently for each genomic bin.

In summary, SACOMA provides a principled, data-driven procedure for translating raw, high-dimensional CpG-level methylation data into a smaller set of spatially-contiguous, co-methylated clusters, without relying on the rigid, fixed distance or correlation cutoffs adopted by earlier region-definition methods. The rationale for adopting SACOMA, in particular relative to the other unsupervised approaches introduced above, is discussed in Chapter 5.

# Chapter 4

# Measuring Methylation

## 4.1 The Illumina BeadChip technology

Illumina BeadChip arrays are commonly used to generate DNA methylation data for large epidemiological studies, due to their high throughput, good accuracy, small sample requirement and relatively low cost.

Over the years, Illumina methylation arrays have evolved to interrogate an increasing number of CpG sites across the genome. Two platforms are particularly relevant for this work: the HumanMethylation450 BeadChip (450K), which measures approximately 485,000 CpG sites, and its successor, the HumanMethylationEPIC BeadChip (EPIC), introduced in 2016 and capable of interrogating more than 850,000 CpG sites.

As discussed in Chapter 2, DNA methylation modifies cytosine residues without altering the underlying DNA sequence. As a consequence, an initial step of bisulfite conversion needs to be performed, since standard sequencing platforms cannot distinguish a methylated cytosine from an unmethylated one by reading the DNA reads. During this process, unmethylated cytosines are chemically converted into uracils, whereas methylated cytosines remain unchanged. This differential conversion enables the subsequent discrimination between methylated and unmethylated alleles at individual CpG sites.

Infinium Type I probes employ two distinct 50-base oligonucleotide probes for each CpG site: one designed to hybridize with the methylated allele and the other with the unmethylated allele. Both signals are measured using the same fluorescence channel. In contrast, Infinium Type II probes use a single probe per CpG site and discriminate between methylated and unmethylated alleles through two different fluorescence channels. Following bisulfite conversion and hybridization, the methylated signal is measured in the green channel whereas the unmethylated signal is measured in the red channel.

The two probe designs were developed to balance measurement accuracy and genomic coverage. Type I probes are preferentially located in CpG-dense regions and generally provide a wider dynamic range of methylation measurements. Type II probes, on the other hand, require fewer array features and therefore allow a substantially larger number of CpG sites to be interrogated on a single chip. For this reason, they constitute the majority of probes on both the HumanMethylation450K and HumanMethylationEPIC arrays.

The DNA methylation level at CpG site i, called $\beta$-value, is estimated as the ratio of the methylated intensity $\gamma_{i,methy}$ to the total intensity across all cells of a given sample:

$$\beta_i = \frac{\max(\gamma_{i,methy},0)}{\max(\gamma_{i,unmethy},0) + \max(\gamma_{i,methy},0) + \alpha}$$

where $\alpha$ is a constant offset (by default equal to 100) added to regularize $\beta$-value when both methylated and unmethylated probe intensities are low [26, 27]. The $\beta$-value statistic results in a number between 0 and 1, representing the percentage of methylation. Under ideal conditions, a value of zero indicates that all copies of the CpG site in the sample were completely unmethylated (no methylated molecules were measured) and a value of one indicates that every copy of the site was methylated.

The IDAT data can be affected by several technical artifacts such as defective probes, background fluorescence, dye-bias from the use of two color channels, bias caused by type I/II probe design and batch effects, that may confound analyses. [28].

A first source of artifactual data comes from defective probes: the scanner can encounter difficulties in correctly reading the signal for some probes, due to their low intensities or to spatial artifacts on the array surface. This translates into a high detection p-value for the probes concerned, flagging them as unreliable during the pre-processing pipeline before obtaining the $\beta$-values.

Additional sources of variation originate from the fluorescence measurement process itself. Background fluorescence results from the presence of non-specific signal in the total signal present in the array. Dye bias instead arises from systematic differences between the red and green fluorescence channels used by the assay, leading to unequal signal distributions across channels. Together, these reduce the dynamic range of the $\beta$-value and skew Infinium Type II $\beta$-values differentially across samples.

A further challenge is represented by the different characteristics of Infinium Type I and Type II probes, resulting in systematic differences in the distribution of $\beta$-values generated by the two probe design, making a preprocessing step necessary to rescale and make them comparable. Specifically, the two probes differ in terms of cardinality and in terms of CpG density, with more CpGs mapping to CpG islands for type I probes (57%) as compared with type II probes (21%). Moreover,

compared with Infinium I probes, the range of -values obtained from the Infinium II probes is smaller; in addition, the Infinium II probes also appear to be less sensitive for the detection of extreme methylation values and display a greater variance between replicates [29].

It has to be observed that at least a part of the Infinium I/II-type bias is a combination of the background and dye-bias artifacts. It is worth noting that at least part of the Type I/II bias is itself a combination of background and dye bias artifacts [28]. For this reason, methods designed ad hoc to correct the Type I/II bias should be able to indirectly account also for the dye bias and the background artifacts.

To account for these three problems, it is necessary to include in the pre-processing pipeline the within-array normalization step. Several methods have been proposed: some of them account for background correction and dye-bias normalization as Noob [30], others just for the Type I/II type bias, such as SWAN on idat and BMIQ on $\beta$-values [31]. The existing procedures can either target a single issue or be designed to address multiple tecnhical noises through a combination of methods, and in the figure 4.1 there is a clear scheme of the most important ones.

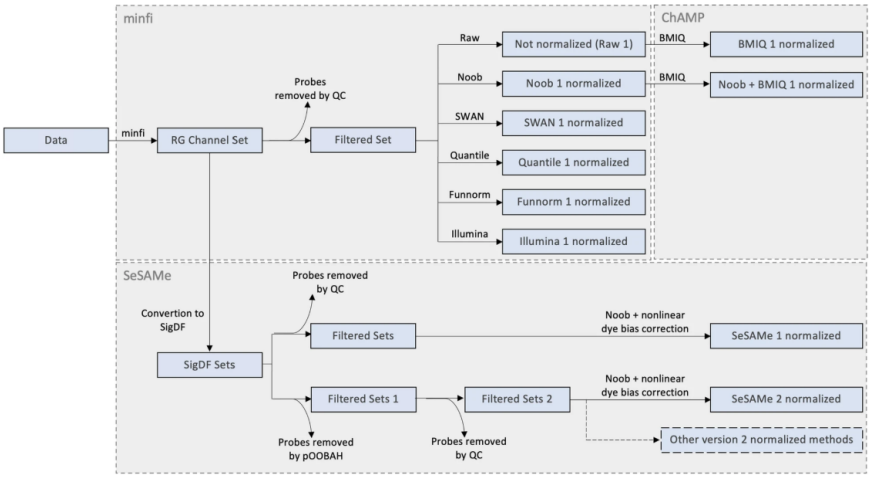


**Figure 4.1:** Normalization pipelines comprehensive scheme from [32]

The last artifact to consider during the pre-processing pipeline is represented by batch effects, which correspond to non-biological variations existing between subgroups of samples that, for instance, have not been processed the same day, on the same scanner, or by the same experimenter. The position of the array on the slide and the slide itself inside a same batch of samples can also generate non-biological variations. Batch effects can generate artifacts that only affect a subset of probes, so they cannot be eliminated by globally normalizing the data as

in the within-array normalization. Therefore, an additional correction step, called between-array normalization, should be applied to reduce as much as possible batch and slide effects. It is important to keep in mind that the best way to avoid problems linked to batch and slide effects is to have a good design of the experiment, meaning a good distribution of the samples labels on the slides and processing of all the samples on the same day by the same experimenter using the same scanner [29]. If this is not possible, one can rely on a posteriori methods applicable either directly to $\beta$-values or to raw data. Examples of the former include 'ComBat', 'sva', and 'RUVm'. The supervised nature of these tools allows them to remove unwanted variation aggressively while keeping variation associated with the labels of interest, requiring the user to provide either batch parameters or an outcome of interest. There exists a between array method, which is called Functional Normalization [33], that works on idat and attempts to remove unwanted variation by adjusting for covariates estimated from a control probe matrix. It has been noted that between array normalization of DNA methylation data must be handled with care. Most methods bring little or no benefit and actually decrease data quality by overcorrecting and removing biological signal. It has been suggested that correcting for known sources of technical variance (including background and dye bias) yields the safest and best implicit between-array normalization [34, 28].

Overall, several pre-processing methods have been implemented for DNA methylation dataset, but it was beyond the objective of this thesis to understand and compare them all, starting the analysis from the raw IDAT files. For this reason, pre-processed $\beta$-values already provided by the original authors were chosen as starting point, with the awareness of the need to manage the batch effect that arises from the integration of datasets coming from different platforms and previously subjected to different processing pipelines.

## 4.2 Data Description & Preparation

The datasets used in this study were obtained from the public Gene Expression Omnibus (GEO) repository. Dataset selection was restricted to studies providing DNA methylation measurements for all three tissue categories of interest: healthy breast tissue obtained from cancer-free female donors (*Normal*), tumor-adjacent normal tissue (*Adjacent*), defined as morphologically normal tissue located at least 3 cm from the tumor margin, and breast tumor tissue (*Tumor*).

Figure 4.2 provides a schematic representation of the three tissue categories considered throughout this work.

Based on these criteria, three breast cancer cohorts were selected: GSE69914, GSE287331, and GSE225845.

Table 4.1 summarizes the main characteristics of the selected cohorts, including

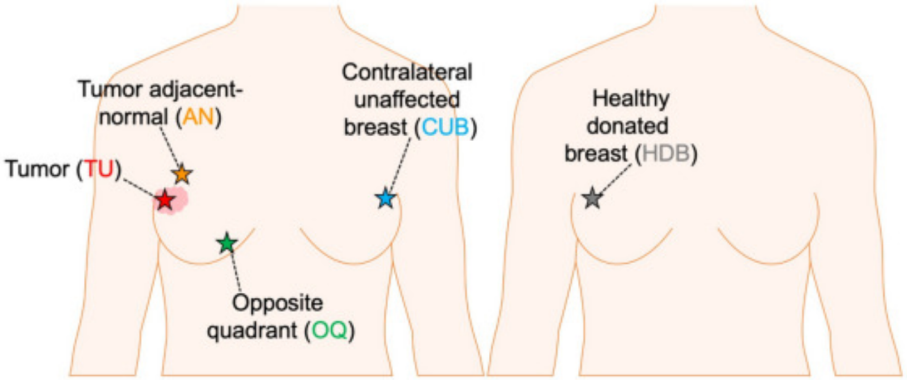


**Figure 4.2:** Visual representation of the three tissue categories considered in this study, reproduced from [35].

the microarray platform, sample distribution across tissue groups, GEO accession number, and preprocessing strategy adopted in the original studies.

**Table 4.1:** Summary of the methylation datasets included in this study.

| **Dataset** | **Platform** | **Sample distribution** | **Preprocessing (raw → beta)** |
|---|---|---|---|
| GSE69914 | Illumina 450K | Normal (50) / Tumor (305) / Adjacent (42) | `preprocessRaw` [36] + `impute.knn` ($k = 5$) + BMIQ |
| GSE287331 | Infinium EPIC | Normal (182) / Tumor (69) / Adjacent (60) | `SeSAMe` [35] |
| GSE225845 | Infinium EPIC | Normal (113) / Tumor (224) / Adjacent (140) | `Noob` + `Functional Normalization` [36] |

To ensure comparability across datasets, several sequential steps were implemented. First, probes affected by known technical artifacts were discarded. Since two cohorts were generated using the Illumina MethylationEPIC array and the last one belongs to Illumina HumanMethylation450K array, EPIC-specific probes were mapped to the 450K annotation through a lift-over procedure, and only CpG sites shared across all datasets were retained. The resulting datasets were then partitioned into training, test and external validation cohorts.

To focus the analysis on biologically informative methylation signals, tissue-specific non-variable CpG sites were removed according to the filtering strategy

proposed by Edgar et al [37]. Subsequently, methylation beta values were transformed into M-values, used as input for the Combat algorithm [38], used to correct batch effects arising from the integration of independent cohorts. A detailed description of each preprocessing step is provided in the following subsections.

### 4.2.1 Technical - based filtering

The first step of data preparation consists in excluding probes whose measured $\beta$-values may reflect technical artifacts rather than genuine DNA methylation, in order to reduce noise and improve the reliability of downstream analyses.

Three categories of technically unreliable probes are systematically removed based on curated blacklists derived from the literature. The first category comprises **cross-reactive probes**, i.e. probes that map to multiple locations in the genome and therefore measure a mixture of specific and non-specific signals rather than the intended target. Blacklists of cross-reactive probes for the 450K array were derived from [39], while those for the EPIC array were derived from [40] and [41]. The second category includes **probes overlapping SNPs or small insertions and deletions (INDELs)** at or near the target CpG site, whose hybridization signal may reflect underlying genetic variation rather than methylation status [39, 42, 41, 43]. The third category, identified by [42], encompasses **probes located in repetitive or low-complexity genomic regions**, where reduced sequence specificity compromises the reliability of hybridization.

A second filtering step was applied using the official Illumina Manifest file, removing two additional categories of probes. First, **non-CpG probes** (prefixed `ch*`), which target cytosines in CHH or CHG sequence contexts rather than canonical CpG dinucleotides and are therefore outside the scope of this analysis. Second, **probes mapping to sex chromosomes** (X and Y), which were excluded to avoid confounding effects related to differences in sex chromosome dosage between male and female samples.

The number of CpGs retained after these two filtering steps for each dataset is reported in Table 4.2.

| Dataset | Platform | CpGs retained |
|---|---|---|
| GSE225845 | EPIC | 530,683 |
| GSE287331 | EPIC | 486,973 |
| GSE69914 | 450K | 251,483 |

**Table 4.2:** Number of CpGs retained after technical filtering for each dataset.

### 4.2.2 CpG intersection and dataset splitting

Since the three cohorts were profiled on different array platforms and subjected to independent preprocessing pipelines by the original authors, the set of CpGs available after quality filtering differs across datasets. To ensure a common feature space for all downstream analyses, only the CpGs present in all three datasets were retained. This yielded an intersection of **215,919 CpGs**.

The intersection is smaller than the number of CpGs available in the 450K cohort, as shown in 4.2 for two reasons. First, the EPIC array covers approximately 90% of the 450K probes by design, but not all 450K probes are represented on the EPIC platform. Second, the independent preprocessing pipelines applied by the original authors — including probe-level filtering based on detection p-values — may have excluded different subsets of probes across cohorts, further reducing the overlap. The two EPIC cohorts (GSE225845 and GSE287331) were each split into a train and a test set with an 80–20 ratio, stratified by class label to preserve the original class composition in both partitions. The 450K cohort (GSE69914) was not split and was reserved entirely as an **external validation set**, serving to assess the generalization of the framework to an independent platform.

This asymmetric treatment of GSE69914 is motivated by its within-adjacent heterogeneity, documented by [7]. Reserving GSE69914 in its entirety as an external, held-out set provides a stronger generalization test, since the framework, trained exclusively on the pooled EPIC cohorts, is thus required to transfer both across platform (EPIC → 450K) and across an independently characterized, more heterogeneous adjacent-tissue population.

The resulting sample distributions are reported in Table 4.3.

**Train set**

| Cohort | Normal | Adjacent | Tumor |
|---|---|---|---|
| GSE225845 | 90 | 112 | 179 |
| GSE287331 | 146 | 48 | 55 |
| **Total** | **236** | **160** | **234** |

**Test set**

| Cohort | Normal | Adjacent | Tumor |
|---|---|---|---|
| GSE225845 | 23 | 28 | 45 |
| GSE287331 | 36 | 12 | 14 |
| **Total** | **59** | **40** | **59** |

**GSE69914**

| Normal | Adjacent | Tumor |
|---|---|---|
| 50 | 42 | 305 |

**Table 4.3:** Sample distribution across train, test and external validation sets for each cohort, after stratified 80–20 splitting of the EPIC datasets.

### 4.2.3 Tissue- specific non variable CpGs Filtering

Illumina series of DNAm arrays, while highly useful as tools for epigenetic studies, were not designed for any specific human tissue, and a large number of cytosine-guanine pairs (CpGs) lack variability within single tissue studies on the arrays. CpGs that are non-variable in a study of a specific disease or tissue may be variable in another context and therefore are still valuable on the 450K. However, these tissue specific nonvariable CpGs contribute to the high dimensionality of the data

and partially necessitate the need for severe multiple test correction. Removal of these independently verified nonvariable CpGs should then allow for a reduced multiple testing space and allow for more power to detect differential DNAm in the study of interest.

Edgar et al. [37] recognized this issue and proposed an empirically derived dimensionality reduction method for EWAS that was then reproduced in this thesis for breast cancer datasets, thanks to the availability of the code.

First of all, cancer samples were excluded, as cancer is associated with high DNAm variability. To designate a CpG as non-variable in a tissue, a threshold of 5% range in beta values (DNAm level ranging from 0 to 1) between the 10th and 90th percentile was used. A slightly more stringent definition of change in beta of 5% was used, since we are asking only that the population as a whole varies by at least 5% and are not testing an effect size between groups. CpGs with less than 5% reference range of beta values in a single tissue population were considered non-variable in that tissue. The definition of non-variable CpG list was made using only the Epic training set, to avoid data leakage, before correcting for batch effects between laboratories, leaving in variability in the data due to technical factors but generating a conservative list of non-variable CpGs, robust to study specific technical variability.

### 4.2.4 MliftOver

The Infinium DNA methylation BeadChip has evolved significantly since its inception, progressing from the HM450 to the EPIC and EPICv2 array. Since GSE69914 was profiled on the Illumina 450K platform while GSE225845 and GSE287331 were profiled on the Infinium EPIC array, direct integration of the three datasets for predictive modeling requires projecting the EPIC beta values onto the HM450 probe grid. This was accomplished using the mLiftOver function from the sesame package [34], which performs a principled cross-platform harmonization by mapping EPIC probes to their HM450 equivalents

mLiftOver, a feature in the SeSAME package, is a tool that harmonizes data from different Infinium platforms across three dimensions: probe names, $\beta$ methylation levels, and signal intensities. It is useful in this context because it manages platform-specific bias for accurate data conversion. mLiftOver can also translate data to and from new and previous Infinium platforms [44] mLiftOver preserves epigenetic identity across platforms. For GSE225845 (477 samples, 530,683 CpGs), the procedure reduced the probe space to 232,179 CpGs mapped onto the HM450 grid. An analogous reduction was obtained for GSE287331. After lifting both datasets independently, only previsously defined as tissue specific variable CpGs were mantained and the train-test split was rebuild, providing an EPIC feature space directly comparable to the probes in GSE69914.

### 4.2.5 Logit Transformation for statistical analysis

In high-throughput statistical data analyses, many of them, like canonical linear models or ANOVA, assume the data is normally distributed and homoscedastic, i.e. the variable variances are approximately constant. The violation of those assumptions imposes serious challenges when applying these analyses to high-throughput data. A common way to check the homoscedasticity of the data is by visualizing the relations between mean and standard deviation. The standard deviation of $\beta$-value is greatly compressed in the low (between 0 and 0.2) and high (between 0.8 and 1) ranges. This means $\beta$-value has significant heteroscedasticity in the low and high methylation range. The problem of heteroscedasticity is effectively resolved after transforming $\beta$-value to M-value, defined as:

$$M_i = log_2(\frac{\beta_i + \epsilon}{1 - \beta_i + \epsilon}) \quad \epsilon = 10^{-6}$$

A M-value close to 0 indicates a similar intensity between the methylated and unmethylated probes, which means the CpG site is about half-methylated. Positive M-values mean that more molecules are methylated than unmethylated, while negative M-values mean the opposite. Beta-values are severely compressed at the extremes when compared with M-values, since for very small intensity values (especially between 0 and 1), small changes of the methylated and unmethylated probe intensities can result in large changes in the M-value.

Its standard deviation is approximately constant across the entire methylation range for M-values. The M-value statistic is therefore much more appropriate for the homoscedastic assumptions of most statistical models used for microarray analysis. An M value transformation resulted attractive for statistical models derived for expression arrays also because are more compatible with the normality assumption. However, $\beta$-value method has a direct biological interpretation, since it corresponds roughly to the percentage of a site that is methylated. For this reason, in this thesis, while M-values are used for statistical modelling, $\beta$-values are retained in parallel for biological interpretation, as recommended by Du et al. [45].

### 4.2.6 Cohort Harmonization via ComBat

The final step of the data preparation pipeline consisted of cohort harmonization to remove batch effects from the training data. In this study, batch effects mainly arise because the two training cohorts were generated independently and underwent different preprocessing procedures before beta value extraction. These technical differences may introduce systematic variation unrelated to the biological signal of interest and therefore require correction before downstream analyses.

Several methods have been proposed to account for batch effects in DNA methylation studies, including RUVm and SVA. However, these methods estimate

latent technical factors to be included as covariates in downstream statistical analyses rather than producing a corrected methylation matrix. Since the objective of this work is to perform statistical modelling and machine learning analyses on DNA methylation data, a method capable of returning batch-corrected methylation matrices was required. For this reason, the ComBat algorithm was adopted.

Originally developed for gene expression microarray data, ComBat has become one of the most widely used approaches for harmonizing genomic datasets across different studies, laboratories, and experimental platforms. It models batch effects as systematic, non-biological shifts in the mean and variance of each genomic feature. Let $Y_{ijg}$ denote the M-value of CpG $g$ for sample $j$ belonging to batch $i$. The ComBat model is defined as

$$Y_{ijg} = \alpha_g + X_{ij}\beta_g + \gamma_{ig} + \delta_{ig}\epsilon_{ijg},$$

where $\alpha_g$ is the overall mean methylation level of CpG $g$ across all batches, $X_{ij}$ is the design matrix encoding the biological covariates, and $\beta_g$ is the corresponding vector of regression coefficients. The parameters $\gamma_{ig}$ and $\delta_{ig}$ represent the additive and multiplicative batch effects, respectively, while the residual errors $\epsilon_{ijg}$ are assumed to follow a normal distribution with mean zero and variance $\sigma_g^2$. Batch effect parameters are estimated through an empirical Bayes framework, which borrows information across CpGs to obtain stable estimates before adjusting the data. In this standard formulation, all batches are jointly re-estimated whenever the set of input batches changes, meaning that the correction of a given batch depends on the other batches supplied to the model.

Batch correction was performed using the `ComBat()` function from the `sva` Bioconductor package (version 3.26.0) on the M-values of the training set. The cohort identifier was specified as the batch variable, while tissue class (Normal, Adjacent, or Tumor) was included in the model matrix to preserve the biological variability of interest during harmonization. The standard ComBat procedure described above was applied exclusively to the two training cohorts (GSE225845 and GSE287331), producing a single harmonized training set of 630 samples on which all downstream analyses were performed.

To ensure a consistent preprocessing strategy for unseen data and to avoid data leakage, one of the training cohorts within the harmonized training matrix was subsequently designated as the reference batch. This reference was then used to harmonize both the internal test set and the independent external validation cohort through the `ref.batch` option, allowing these datasets to be adjusted according to the training reference without influencing the estimation of batch correction parameters.

In this reference-based formulation, the model becomes

$$Y_{ijg} = \alpha_{rg} + X_{ij}\beta_{rg} + \gamma_{rig} + \delta_{rig}\epsilon_{ijg},$$

where $\alpha_{rg}$ is the mean methylation level of CpG $g$ in the chosen reference batch $r$, and $\gamma_{rig}$ and $\delta_{rig}$ represent the additive and multiplicative differences between batch $i$ and the reference, rather than deviations from a global cross-batch mean. As long as the reference batch itself remains unchanged, the resulting correction is invariant to the additional batches subsequently harmonized with it.

Following cohort harmonization, the corrected M-value matrices were used as input for downstream feature selection and machine learning analyses.

# Chapter 5

# Methodology

This chapter presents the methodological contribution of this thesis, namely a feature selection pipeline designed to address some of the limitations of existing approaches.

The iEVORA method is based on the assumption that biologically informative CpGs may be identified through differences in variability between groups (differential variability, DV) [7]. Conversely, the moderated t-test used jointly with the empirical Bayes procedure proposed in limma is based on the hypothesis that the most relevant features are those that differ significantly in terms of mean level (differential mean, DM).

The hypothesis underlying this work is instead that the most relevant features are those groups of CpGs, referred to as CpG clusters, that differ significantly between groups in terms of co-methylation, a signal that is invisible to both DM- and DV-based approaches operating at the single-CpG level.

The biological question underlying this pipeline of feature selection is:

*"Does tumour development disrupt local co-methylation patterns observed in normal tissue?*
*If so, where and how?"*

The pipeline can be summarized in two main steps. The first consists in the identification of CpG clusters, i.e. groups of probes that are genomically close and show high pairwise correlation in normal samples, which are taken as the reference and thus provide a snapshot of the baseline co-methylation architecture.

The second step consists in characterizing, at the cluster level, how this baseline co-methylation architecture is perturbed during disease development.

To this end, Adjacent and Tumor samples are treated independently within a multivariate conjugate Bayesian framework, in which the co-methylation structure observed in Normal tissue is used to define prior distributions over the cluster-specific

parameters. Information from each target group is subsequently incorporated through Bayesian updating, yielding posterior distributions that reflect both the Normal-derived prior knowledge and the evidence provided by the observed samples. For each comparison, the extent to which a cluster departs from its Normal reference state is then quantified through a Kullback–Leibler divergence decomposition, which separates alterations in average methylation levels ($\delta_{mean}$) from changes affecting the covariance structure ($\delta_{cov}$), and therefore the underlying co-methylation relationships among the CpGs of the cluster.

Finally, clusters are selected as candidate features based on a composite score combining both comparison axes (Normal-Adjacent and Normal-Tumor), and used as features for the XGBoost binary classification task.

## 5.1 Cluster definition

The decision to work with clusters was guided by two main motivations. The first is the aim of capturing the true biological characteristics of methylation. As described in the 'Biological Context' chapter 2.3, CpGs are not independent; therefore, the objective was to develop a pipeline that was as faithful as possible to the biology of the phenomenon. An implicit advantage of this approach is the possibility of reducing dimensionality by aggregating sites into a smaller set of meaningful analytical units that can be used for downstream region-based association analysis.

Having established the need for a region-based unit of analysis, two further design choices had to be made: *where* clusters should be defined, and *which* clustering algorithm should be used to define them.

### 5.1.1 Defining clusters on the Normal reference condition.

As outlined in the State of the Art chapter 3, there exists various approaches to conducting a region-based analysis. The type of change in the co-methylation structure of the CpGs that this thesis aims to find is a group of CpGs strongly correlated in one condition but not in the other. This means using a clustering algorithm in a semi-supervised way, targeting the normal samples as the reference condition in cluster definition, consistently with other existing cluster-level analyses [8], where Adjacent and Tumor tissues are framed as progressive deviations from a Normal reference state.

Starting from Normal tissue implicitly assumes that co-methylation structures present in healthy tissue are progressively lost as disease develops. Consistent with this hypothesis, Sun et al. [46] report that normal breast tissue exhibits a higher degree of co-methylation than tumor tissue, particularly in terms of negatively correlated CpG pairs, and show that a subset of CpG sites undergoes pronounced

co-methylation changes between normal and tumor samples – changes that are only partially explained by differential methylation alone.

This choice, however, is not the only one that could a priori be justified. In fact, it cannot be excluded that co-methylation in diseased tissue is not simply destroyed, but rather reorganized into new, disease-specific structures that are not visible when Normal is used as the reference: a cluster absent in Normal could still be present and biologically meaningful in Tumor. Under this scenario, defining clusters on Tumor samples as an alternative reference condition would also be informative, and could reveal a complementary aspect of the phenomenon that this thesis, by construction, is not designed to capture. Investigating co-methylation structures anchored on a diseased-tissue reference is therefore left as a future direction of this work.

### 5.1.2 Choosing a spatially-aware clustering algorithm.

Since co-methylation has been shown to decay with genomic distance [14, 15], the natural choice was to use a clustering algorithm that explicitly accounts for this aspect of the phenomenon. A spatially-constrained clustering approach carries a strong default biological interpretation: physically neighboring CpGs that co-vary are likely to share the same local regulatory machinery (the same cis-regulatory element, nucleosome, or small-scale chromatin domain). An alternative approach would have been to consider co-methylation in general, without any constraint due to genomic distance, adapting a different clustering algorithm such as WGCNA [47] to the methylation context. A WGCNA module without spatial constraints can capture trans-correlations – distant CpGs, or CpGs on different chromosomes, that co-vary because they are regulated by a common upstream factor (e.g., a transcription factor acting genome-wide, or effects of cell composition, batch, or biological age). Using such a method would therefore answer an interesting biological question – *which groups of loci, wherever they may be, move together in tumor tissue?* – but one that is intrinsically different from the question motivating this thesis, i.e., *does the local co-methylation observed in normal tissue break down, and where?*

The choice of using spatially-aware clusters is also reinforced by a purely computational constraint. For the 149,147 CpGs retained after preprocessing, the full pairwise correlation matrix would contain approximately $149{,}147^2/2 \approx 11$ billion pairs. WGCNA is already computationally demanding at ∼20–30k features (400–900 million pairs); at 149k features it is not feasible without drastic blockwise approximations or aggressive preliminary dimensionality reduction, either of which would undermine the analysis. A distance-informed algorithm avoids this problem intrinsically, since it only ever evaluates correlations within genomically local windows.

### 5.1.3 Algorithm selection

Among the available spatially-aware clustering methods, SACOMA [24] was selected. Through extensive simulations, SACOMA has demonstrated superior sensitivity while maintaining effective false-positive control compared to existing methods, such as Aclust2 [48] and coMethDMR [49]. A further reason for this choice concerns how the two criteria of methylation correlation and genomic distance are combined. Aclust2 merges neighboring CpGs through a two-stage greedy agglomerative procedure, first restricting candidate neighbors to a genomic distance window and merging based on methylation correlation within that window, then further aggregating clusters subject to both a correlation threshold and a distance threshold enforced jointly. In both stages, distance and correlation act as independent, non-negotiable constraints, with no mechanism allowing one criterion to compensate for the other. SACOMA instead combines methylation correlation and spatial proximity into a single weighted objective through a mixing parameter $\alpha$, estimated independently within each genomic bin, building on ClustGeo [25], an established hierarchical clustering framework based on a convex combination of two dissimilarity matrices. This allows SACOMA to accommodate a partial trade-off between the two criteria, rather than enforcing independent cutoffs as in Aclust2. This local adaptivity reduces, though does not eliminate, the reliance on globally-fixed hyperparameters.

As explained in the State of the Art Chapter 3, SACOMA relies on three user-defined hyperparameters: the minimum number of CpGs required to retain a region (minCpGs), the maximum allowed genomic distance in base pairs between adjacent CpG sites for them to be grouped into the same region (maxGap), and the minimum correlation threshold among CpGs within a region (rthreshold). Since guidance on how to select these parameters is not provided in the original publication – where they are only explored in the context of simulation studies – and their effect on real data is not straightforward to anticipate, a sensitivity analysis was conducted on both maxGap and rthreshold.

**maxGap.** The choice of maxGap must reflect the spatial scale within which searching for co-methylation is biologically meaningful. Since co-methylation has been shown to decay within approximately 1000–2000 bp [14, 15], maxGap was explored over $\{1000, 2000\}$bp, anchoring the parameter choice to the empirically observed decay of correlation with genomic distance, rather than to an arbitrary cutoff. A limitation of fixing a single maxGap value, even when biologically motivated at the global level, is that the rate of co-methylation decay is not uniform across CpG neighborhood classes (Island, Shore, Shelf, OpenSea): the "best" maxGap identified through the sensitivity analysis therefore remains an average compromise across genomic contexts with heterogeneous local dynamics.

This is not a shortcoming specific to this choice, however, but a limitation inherent to any single global-parameter approach – the same applies, for instance, to Aclust2's fixed distance threshold (bp.thresh.dist/max.dist).

**rthreshold.** The rthreshold parameter was explored over $\{0.4, 0.5\}$, consistent with the values used in the original SACOMA publication [24]. This parameter poses a region-dependent trade-off: a threshold set too high risks missing weak but biologically relevant co-methylation in OpenSea regions, where correlation is generally lower; conversely, a threshold set too low risks introducing noise in CpG Islands, where co-methylation tends to be stronger and to decay more slowly with distance.

## 5.2 Bayesian Framework

This section presents the mathematical formulation of the Bayesian framework adopted to quantify cluster-level perturbations in co-methylation patterns. The proposed pipeline does not aim at identifying individual CpGs exhibiting differential methylation; rather, it seeks to characterize how the multivariate methylation structure of a co-methylated cluster changes when moving from healthy tissue to altered tissue states.

### 5.2.1 NIW model

The fundamental assumption underlying this framework is that CpG clusters constitute independent units of analysis. Consequently, the model described below is applied separately to each cluster $k \in \{1, \dots, K\}$. Furthermore, Adjacent and Tumor samples are analyzed independently, yielding two distinct assessments of cluster perturbation relative to the same Normal reference.

By construction, CpG clusters are defined based on co-methylation patterns observed in Normal samples and therefore represent genomic regions exhibiting a stable dependency structure in healthy tissue. This motivates the use of a Bayesian formulation in which the Normal-derived co-methylation architecture is treated as prior information and updated upon observing methylation data from altered tissues. The resulting posterior distribution can thus be interpreted as a perturbed version of the Normal reference state, allowing deviations from the baseline co-methylation structure to be quantified within a coherent probabilistic framework. We now formalize this construction in details. Let $g \in \{A, T\}$ denote the altered tissue group under consideration. For each co-methylation cluster $k \in \{1, \dots, K\}$), let $p_k$ denote the cluster cardinality, i.e. the number of CpG sites it contains. Then the multivariate M-values $\mathbf{X}_i^{(k,g)} \in \mathbb{R}^{p_k}$ associated with cluster $k$ and sample $i$

belonging to tissue group $g$ are assumed to arise from a Gaussian distribution:

$$\mathbf{X}_i^{(k,g)} \mid \boldsymbol{\mu}_k^g, \boldsymbol{\Sigma}_k^g \overset{\text{i.i.d.}}{\sim} \mathcal{N}_{p_k}(\boldsymbol{\mu}_k^g,\, \boldsymbol{\Sigma}_k^g)\,, \qquad i = 1, \ldots, n_g. \tag{5.1}$$

with a conjugate Normal-Inverse-Wishart prior:

$$\boldsymbol{\Sigma}_k^g \;\sim\; \mathcal{IW}(\boldsymbol{\Phi}_{0,k},\, \nu_{0,k})\,, \qquad \boldsymbol{\mu}_k^g \mid \boldsymbol{\Sigma}_k^g \;\sim\; \mathcal{N}_{p_k}\left(\boldsymbol{\mu}_{0,k},\, \frac{\boldsymbol{\Sigma}_k^g}{\kappa_0}\right). \tag{5.2}$$

This model can be also represented via a Direct Acyclic Graph (DAG), as shown in Figure 5.1.

| Symbol | Meaning | Range / Type |
|---|---|---|
| $k$ | CpG cluster index | $k = 1, \ldots, K$ |
| $p_k$ | number of CpGs in cluster $k$ | $p_k \geq 3$, integer |
| $g$ | biological group | $g \in \{N, A, T\}$ |
| $n_g$ | number of samples in group $g$ | $n_N,\, n_A,\, n_T \in \mathbb{N}^+$ |
| $i$ | sample index | $i = 1, \ldots, n_g$ |
| $\mathbf{X}_i^{(k,g)}$ | M-value vector for sample $i$ | $\mathbf{X}_i^{(k,g)} \in \mathbb{R}^{p_k}$ |
| $\bar{\mathbf{x}}_k^{(g)}$ | sample mean of cluster $k$ in group $g$ | $\bar{\mathbf{x}}_k^{(g)} \in \mathbb{R}^{p_k}$ |
| $\mathbf{S}_k^{(g)}$ | scatter matrix of cluster $k$ in group $g$ | $\mathbf{S}_k^{(g)} \in \mathbb{R}^{p_k \times p_k}$ |

**Table 5.1:** Summary of Indices and variables of the model.

The posterior distribution, due to conjugacy, is again a Normal–Inverse–Wishart distribution with updated parameters given by (full derivation reported in Appendix):

$$\kappa_1 = \kappa_0 + n_g, \tag{5.3}$$

$$\boldsymbol{\mu}_{1,k} = \frac{\kappa_0\, \boldsymbol{\mu}_{0,k} + n_g\, \bar{\mathbf{x}}_k^{(g)}}{\kappa_0 + n_g}, \tag{5.4}$$

$$\nu_{1,k} = \nu_{0,k} + n_g, \tag{5.5}$$

$$\boldsymbol{\Phi}_{1,k} = \boldsymbol{\Phi}_{0,k} + \mathbf{S}_k^{(g)} + \frac{\kappa_0\, n_g}{\kappa_0 + n_g} \left(\bar{\mathbf{x}}_k^{(g)} - \boldsymbol{\mu}_0^{(k)}\right) \left(\bar{\mathbf{x}}_k^{(g)} - \boldsymbol{\mu}_0^{(k)}\right)^\top, \tag{5.6}$$

where

$$\bar{\mathbf{x}}_k^{(g)} = \frac{1}{n_g} \sum_{i=1}^{n_g} \mathbf{X}_i^{(k,g)} \in \mathbb{R}^{p_k}, \qquad \mathbf{S}_k^{(g)} = \sum_{i=1}^{n_g} \left(\mathbf{X}_i^{(k,g)} - \bar{\mathbf{x}}_k^{(g)}\right) \left(\mathbf{X}_i^{(k,g)} - \bar{\mathbf{x}}_k^{(g)}\right)^\top \in \mathbb{R}^{p_k \times p_k}$$

are the sample mean and the scatter matrix of the altered-group samples, respectively.

The updated parameters admit a clear statistical interpretation. In particular, the posterior mean $\boldsymbol{\mu}_{1,k}$ can be seen as a weighted combination of prior knowledge, encoded by $\boldsymbol{\mu}_{0,k}$ and derived from Normal samples, and the empirical mean of the altered tissue group. The relative contribution of these two sources of information is controlled by $\kappa_0$, which acts as an effective prior sample size. Similarly, the updated scale matrix $\boldsymbol{\Phi}_{1,k}$ can be decomposed into three distinct contributions: (i) the prior co-methylation structure inferred from Normal samples, (ii) the within-group variability captured by the scatter matrix $\mathbf{S}_k^{(g)}$, and (iii) a correction term accounting for discrepancies between prior and observed group means, which captures between-group shifts in the multivariate methylation profile.

The specification of the prior hyperparameters $(\boldsymbol{\mu}_{0,k}, \kappa_0, \boldsymbol{\Phi}_{0,k}, \nu_{0,k})$ plays a crucial role in ensuring consistency between cluster construction and subsequent statistical inference. In this work, these hyperparameters are not treated as non-informative quantities; instead, they are empirically estimated from Normal samples, in order to ensure consistency between cluster construction and statistical inference.

**Prior mean: $\boldsymbol{\mu}_{0,k}$.** The prior mean is set to the sample mean of the Normal group:

$$\boldsymbol{\mu}_{0,k} = \bar{\mathbf{x}}_k^{(N)}. \tag{5.7}$$

This encodes the biologically motivated belief that, in the absence of perturbation, the expected methylation pattern of cluster $k$ is the one observed in Normal tissue.

**Prior strength on the mean: $\kappa_0$.** The parameter $\kappa_0$ controls how strongly the prior on $\boldsymbol{\mu}_k$ resists updating. From Equation (5.4), $\boldsymbol{\mu}_{1,k}$ can be interpreted as a weighted average between the Normal-derived prior mean and the empirical mean of the altered tissue group, where $\kappa_0$ acts as an effective prior sample size. A natural choice would be to set $\kappa_0 = n_N$.

**Prior scale matrix: $\boldsymbol{\Phi}_{0,k}$.** The expectation of the Inverse-Wishart distribution $\mathcal{IW}(\boldsymbol{\Phi}_{0,k}, \nu_{0,k})$ is given by:

$$\hat{\boldsymbol{\Sigma}}_{0,k} = \mathbb{E}[\boldsymbol{\Sigma}_k] = \frac{\boldsymbol{\Phi}_{0,k}}{\nu_{0,k} - p_k - 1}.$$

By matching this quantity to the empirical covariance structure of the Normal group, $\frac{\mathbf{S}_k^{(N)}}{n_N}$, we obtain:

$$\boldsymbol{\Phi}_{0,k} = \frac{\mathbf{S}_k^{(N)}}{n_N}(\nu_{0,k} - p_k - 1), \tag{5.8}$$

where $\mathbf{S}_k^{(N)}$ denotes the scatter matrix of the Normal samples for cluster $k$.

**Prior degrees of freedom: $\nu_{0,k}$.** The parameter $\nu_{0,k}$ controls the strength of the prior on the covariance structure. To guarantee the existence of the posterior expectation of the covariance matrix, it is required that $\nu_{0,k} > p_k + 1$. The minimal admissible choice is therefore $\nu_{0,k} = p_k + 2$. To ensure coherence with the effective prior strength on the mean, we impose:

$$\nu_{0,k} - p_k - 1 = \kappa_0, \tag{5.9}$$

which yields:

$$\nu_{0,k} = \kappa_0 + p_k + 1. \tag{5.10}$$

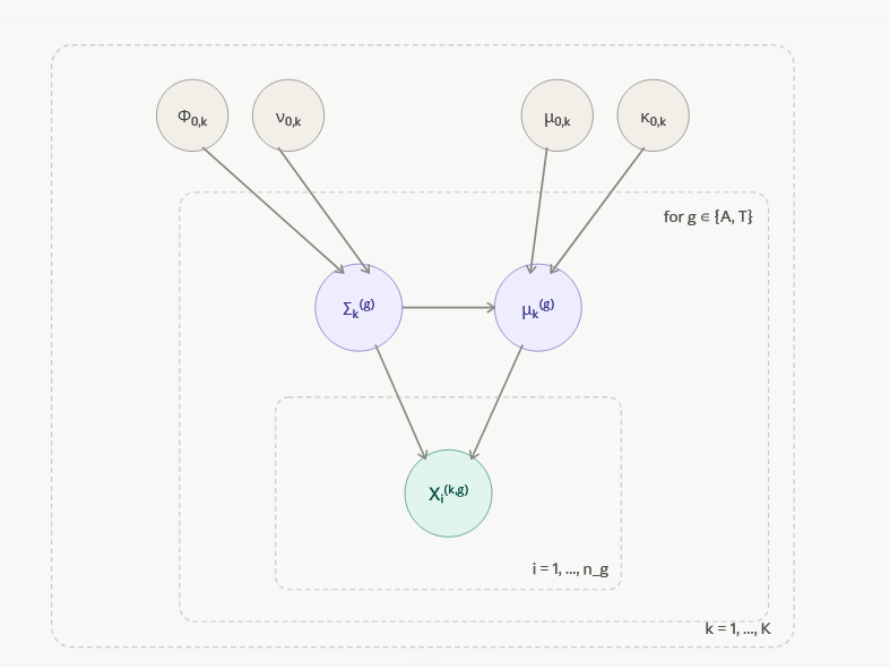


**Figure 5.1:** DAG representation of the Normal–Inverse–Wishart hierarchical model for cluster-level co-methylation inference.

### 5.2.2 Disruption Metrics Definition

The Bayesian model introduced in the previous section provides, for each cluster, a prior and a posterior distribution describing the multivariate methylation state of the cluster before and after observing samples from an altered tissue group.

A natural way to quantify the extent of such perturbation is through the Kullback–Leibler (KL) divergence between the prior and posterior distributions.

In a multivariate Gaussian setting, the KL divergence measures the amount of information required to move from a reference state to an updated state and therefore provides a principled measure of cluster disruption.

Since the Normal-Inverse-Wishart prior is conjugate, both prior and posterior belong to the same distributional family. Although a closed-form expression for the KL divergence between two NIW distributions exists, its interpretation is less straightforward from a biological perspective. For this reason, a plug-in approximation is adopted. The covariance matrices are replaced by their posterior expectations,

$$\hat{\boldsymbol{\Sigma}}_{0,k} = \frac{\boldsymbol{\Phi}_{0,k}}{\nu_{0,k} - p_k - 1}, \qquad \hat{\boldsymbol{\Sigma}}_{1,k} = \frac{\boldsymbol{\Phi}_{1,k}}{\nu_{1,k} - p_k - 1},$$

yielding the two Gaussian states;

$$f_{0,k} = \mathcal{N}_{p_k}\left(\boldsymbol{\mu}_{0,k}, \frac{\hat{\boldsymbol{\Sigma}}_{0,k}}{\kappa_0}\right), \qquad f_{1,k} = \mathcal{N}_{p_k}\left(\boldsymbol{\mu}_{1,k}, \frac{\hat{\boldsymbol{\Sigma}}_{1,k}}{\kappa_1}\right).$$

The Kullback–Leibler divergence between the posterior state $f_{1,k}$ and the prior state $f_{0,k}$ admits the following closed-form expression:

$$D_{KL}(f_{1,k} \,\|\, f_{0,k}) = \frac{1}{2}\Bigg[\underbrace{\kappa_0(\boldsymbol{\mu}_{1,k} - \boldsymbol{\mu}_{0,k})^\top \hat{\boldsymbol{\Sigma}}_{0,k}^{-1}(\boldsymbol{\mu}_{1,k} - \boldsymbol{\mu}_{0,k})}_{\text{Mean shift}} + \underbrace{\frac{\kappa_0}{\kappa_1}\operatorname{tr}\left(\hat{\boldsymbol{\Sigma}}_{0,k}^{-1}\hat{\boldsymbol{\Sigma}}_{1,k}\right) - p_k + \log\left(\frac{\det \hat{\boldsymbol{\Sigma}}_{0,k}}{\det \hat{\boldsymbol{\Sigma}}_{1,k}}\right) + \log(\frac{\kappa_1}{\kappa_0})}_{\text{Covariance structure change}}\Bigg]. \tag{5.11}$$

It naturally leads to define the following disruption measures:

$$\delta_{\text{mean}}^{(k,g)} = \frac{\kappa_0}{2\,p_k}\,(\boldsymbol{\mu}_{1,k} - \boldsymbol{\mu}_{0,k})^\top \hat{\boldsymbol{\Sigma}}_{0,k}^{-1}(\boldsymbol{\mu}_{1,k} - \boldsymbol{\mu}_{0,k}). \tag{5.12}$$

$$\delta_{\text{cov}}^{(k,g)} = \frac{1}{2\,p_k}\left[\frac{\kappa_0}{\kappa_1}\operatorname{tr}\left(\hat{\boldsymbol{\Sigma}}_{0,k}^{-1}\hat{\boldsymbol{\Sigma}}_{1,k}\right) - p_k + \log\left(\frac{\det \hat{\boldsymbol{\Sigma}}_{0,k}}{\det \hat{\boldsymbol{\Sigma}}_{1,k}}\right) + \log\left(\frac{\kappa_1}{\kappa_0}\right)\right]. \tag{5.13}$$

$\delta_{\text{mean}}$ is a Mahalanobis distance between the posterior and prior cluster means, scaled by the prior's uncertainty $\frac{\hat{\boldsymbol{\Sigma}}_{0,k}}{\kappa_0}$ and normalized by $p_k$. It quantifies shifts in the average methylation profile of the cluster when moving from Normal to the altered tissue group $g$.

$\delta_{\text{cov}}$ captures the change in the shape and volume of uncertainty between the prior and posterior, reflecting alterations in the co-methylation architecture of the cluster $k$ induced by the group $g$.

## 5.3 Exploratory Analysis and Adjusted Outlyingness Framework

After the Bayesian framework procedure, each cluster is associated with four disruption metrics. An independent exploratory analysis was first carried out, and an initial cluster selection based on the degree of outlyingness of each metric was considered. Subsequently, the joint analysis of the four metrics, together with a literature-derived biological hypothesis on the directionality of alterations along the $N \rightarrow A \rightarrow T$ axis, led to the development of composite cluster selection scores.

### 5.3.1 Independent EDA

The analysis begins with a qualitative description of the marginal distributions of $\delta_{mean}$ and $\delta_{cov}$, through density plots shown in Figure 5.2 and descriptive statistics, i.e. mean, variance, skewness, and excess kurtosis, reported in Table 5.2.

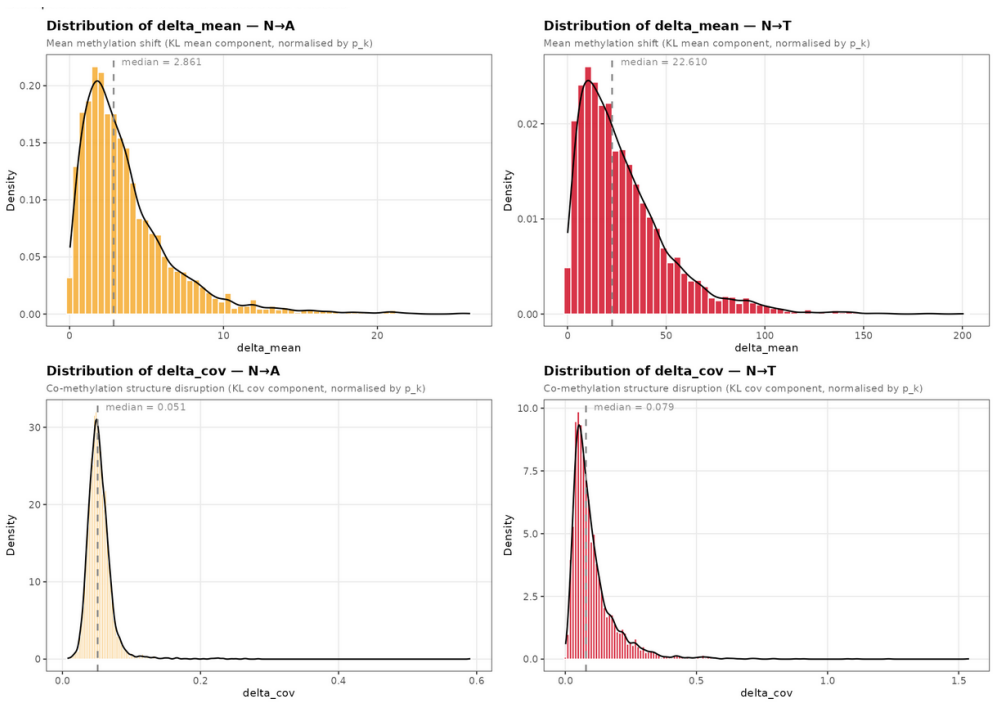


**Figure 5.2:** Disruption metrics distributions

The distributions of the mean disruption metric $\delta_{mean}$ are compared for both the $N \rightarrow A$ and $N \rightarrow T$ transitions. Two main aspects can be observed. First, regarding distributional shape, both distributions are strongly right-skewed, as confirmed by the positive skewness values reported in Table 5.2. Additionally, similar positive

**Table 5.2:** Descriptive statistics of disruption metrics across 5,033 clusters.

| Metric | Mean | Median | Variance | Skewness | Excess Kurtosis |
|---|---|---|---|---|---|
| $\delta_{mean}^{N\to A}$ | 3.708 | 2.861 | $1.02 \times 10^{1}$ | 2.18 | 6.98 |
| $\delta_{mean}^{N\to T}$ | 28.742 | 22.610 | $5.63 \times 10^{2}$ | 1.78 | 4.81 |
| $\delta_{cov}^{N\to A}$ | 0.054 | 0.051 | $4.83 \times 10^{-4}$ | 6.84 | 112.34 |
| $\delta_{cov}^{N\to T}$ | 0.107 | 0.079 | $9.45 \times 10^{-3}$ | 3.97 | 30.41 |

excess kurtosis values indicate that both distributions are leptokurtic, meaning that observations are more concentrated around the mean and extreme values are more frequent than in a Normal distribution. Second, despite the similarity in shape, the two distributions differ substantially in scale. The median of $\delta_{mean}^{N\to A}$ is 2.861, with a maximum value of approximately 25, while the median of $\delta_{mean}^{N\to T}$ is 22.610, with a maximum value of approximately 200. This suggests that cluster-level mean disruption has a larger magnitude in the Normal vs Tumor comparison than in the Normal vs Adjacent one, with the median approximately 8 times higher. This difference in magnitude is consistent with the biological hypothesis underlying this work: adjacent tissue, being by definition apparently normal, exhibits a less pronounced disruption of co-methylation structure relative to healthy tissue, compared to what is observed in tumor tissue. Furthermore, the variance of $\delta_{mean}^{N\to T}$ is considerably higher than that of $\delta_{mean}^{N\to A}$, indicating greater heterogeneity in cluster-level behaviour in Tumor tissue. However, since these observations are based on marginal distributions, whether they reflect a monotonic progression at the cluster level along the $N \to A \to T$ axis requires further per-cluster analysis.

Turning to the covariance disruption metric $\delta_{cov}$, both distributions are again right-skewed, as confirmed by the positive skewness values in Table 5.2. However, the excess kurtosis reveals substantial differences between the two. Both values are positive, implying that observations are more concentrated around the mean and extreme values are more frequent than in a Normal distribution. The strongly higher excess kurtosis of $\delta_{cov}^{N\to A}$ (112.34) compared to $\delta_{cov}^{N\to T}$ (30.41) indicates that the former deviates considerably more from normality relative to its own scale.

This is visually confirmed in Figure 5.2: the distribution of $\delta_{cov}^{N\to A}$ presents a very sharp and tall peak (density $\approx 30$ around 0.05), with the curve descending almost vertically on both sides and a very thin right tail vanishing before 0.6. Nearly all the mass is compressed within a narrow range around the median, consistent with an excess kurtosis of 112. In contrast, $\delta_{cov}^{N\to T}$ shows a much lower and broader peak (density $\approx 10$), with a heavier right tail extending up to approximately 1.5, indicating that a larger fraction of clusters reaches higher disruption values.

In terms of magnitude, the medians are comparable (0.051 vs 0.079, a ratio of approximately 1.5×), considerably lower than the 8× difference observed for $\delta_{mean}$. Nevertheless, the variance of $\delta_{cov}^{N\to T}$ is approximately one order of magnitude higher than that of $\delta_{cov}^{N\to A}$ ($9.45\times10^{-3}$ vs $4.83\times10^{-4}$), consistently with the greater spread visible in the distribution.

Finally, a common pattern can be observed across both metrics: the distribution in the $N\to T$ comparison is more spread out along the x-axis and lower in density than in the $N\to A$ comparison. However, the two metrics differ in the relative shift of their medians: while for $\delta_{cov}$ the medians remain comparable (1.5×), for $\delta_{mean}$ they differ by a factor of approximately 8×, reflecting a considerably more pronounced displacement along the x-axis in the latter.

### 5.3.2 The Adjusted Outlyingness Definition

In order to select the most relevant clusters, the focus can be initally be placed on the heavy tails of both the $N\to A$ and $N\to T$ distributions, i.e. the outliers, since these correspond to clusters whose disruption is higher and anomalous.

To correctly select the outliers of both distributions indipendently, the *Adjusted Outlyingness* (AO) measure proposed by Hubert et al. [50] is adopted as reference. Outlyingness quantifies how far an observation lies from the centre of its reference distribution. Unlike classical symmetric measures, the adjusted outlyingness explicitly accounts for skewness, making it particularly suitable for the strongly right-skewed disruption distributions considered in this work.

Firstly, it is necessary to introduce the Med Couple (MC) measure of skewness, defined as:

$$\mathrm{MC}(X_n) = \operatorname*{med}_{x_i<\mathrm{med}_n<x_j} h(x_i, x_j)$$

where $\mathrm{med}_n$ is the sample median, and

$$h(x_i, x_j) = \frac{(x_j - \mathrm{med}_n) - (\mathrm{med}_n - x_i)}{x_j - x_i}$$

Then, given $Q_1$ and $Q_3$, the first and third quartiles respectively, $\mathrm{IQR} = Q_3 - Q_1$, and MC, which in our case is positive due to the right-skeweness of the involved distributions, the adjusted boxplot whiskers are defined as:

$$w^- = Q_1 - 1.5\, e^{-4\,\mathrm{MC}}\,\mathrm{IQR} \qquad w^+ = Q_3 + 1.5\, e^{3\,\mathrm{MC}}\,\mathrm{IQR}$$

The original Adjusted Outlyingness score assigns positive values to all observations, measuring the distance from the median scaled by the half-width of the respective side of the adjusted boxplot. As a result, observations below the median are mapped to $(0, +\infty)$ symmetrically to those above it, making the two sides indistinguishable in sign.

To improve interpretability, this thesis adopts a sign-extended variant, defined as follows:

$$\mathrm{AO}_i = \begin{cases} \dfrac{x_i - \mathrm{med}(x)}{w^+ - \mathrm{med}(x)} & \text{if } x_i > \mathrm{med}(x) \\ -\dfrac{\mathrm{med}(x) - x_i}{\mathrm{med}(x) - w^-} & \text{if } x_i \leq \mathrm{med}(x) \end{cases} \tag{5.14}$$

With this formulation, $\mathrm{AO} = 0$ corresponds to the median, observations below the median receive negative values in $(-\infty, 0)$, and their relative ordering with respect to the median is preserved. At the same time, $\mathrm{AO} = 1$ corresponds exactly to the upper adjusted boxplot outlier threshold, while $0 < \mathrm{AO} < 1$ identifies observations above the median but below the outlier threshold.

By computing the AO score independently for each disruption distribution, an initial subset of clusters is selected for each metric. The results are shown in Figure 5.3

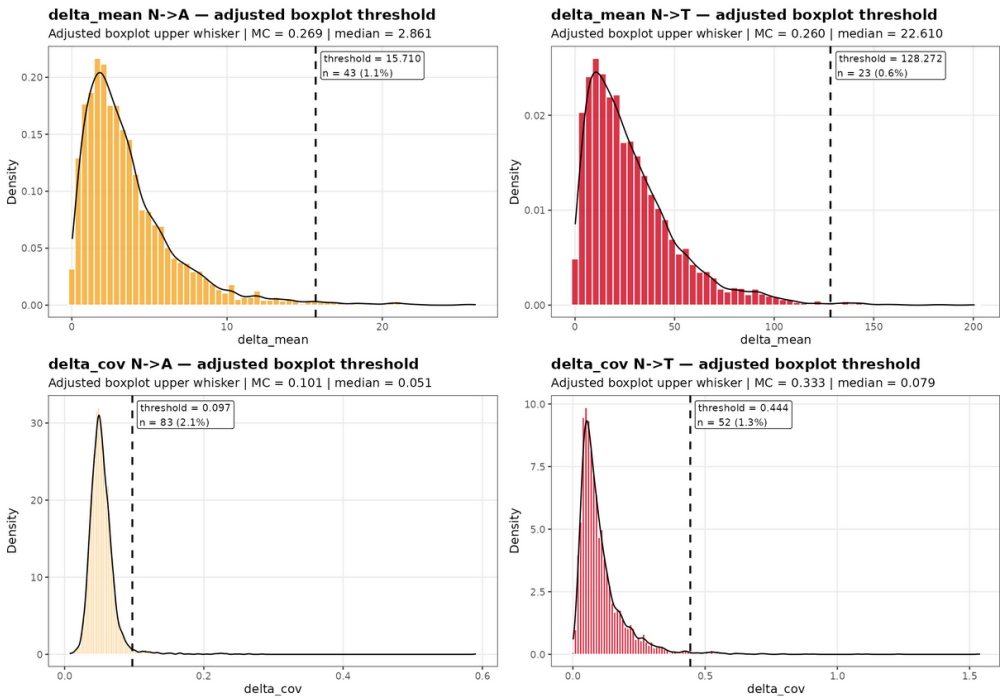


**Figure 5.3:** Outlier selections for all metrics

The fact that for both metrics the $N \to T$ comparison yields fewer selected outliers than the $N \to A$ one does not imply that tumour tissue is less disrupted, but is a consequence of the intrinsically higher variability of the $N \to T$ distributions, which causes the adjusted upper whisker to rise accordingly. The adjusted boxplot identifies clusters that are anomalous within their own distributional context —

that is, those exhibiting a disruption magnitude considerably higher than the rest — but this does not imply that clusters below the threshold are biologically uninformative.

This observation highlights a limitation of relying exclusively on an outlier-based criterion. Although the adjusted boxplot correctly identifies the most extreme clusters within each disruption distribution, biologically relevant changes may also occur among clusters that are not classified as statistical outliers. This naturally motivates a complementary perspective. Rather than considering each disruption metric separately, clusters can be evaluated according to their joint behaviour across the N→A and N→T comparisons. Under this view, a cluster is not considered biologically relevant solely because it exhibits an exceptionally large disruption in one comparison, but because its pattern of disruption is consistent with a specific hypothesis of change along the $N \rightarrow A \rightarrow T$ axis. The following Section aims in investigating the conjoint behaviour of the same clsuter along the two comparisons.

### 5.3.3 Joint EDA

An initial joint exploratory analysis of the two disruption distributions is conducted to address the following question:

> *Does the degree of extremeness observed in the $N \rightarrow A$ comparison become amplified, attenuated, or unchanged in the $N \rightarrow T$ comparison?*

This question is biologically motivated: if a cluster already displays anomalous co-methylation disruption in adjacent tissue — a site that is histologically normal yet epigenetically altered — it is natural to ask whether this early signal anticipates the degree of disruption observed in tumor. Answering this question requires comparing the two disruption axes jointly, at the cluster level.

A direct comparison of raw $\delta$ values across the two comparisons would however be misleading, since the two empirical distributions differ substantially in location, spread, and skewness. Equal numerical values of $\delta$ do not correspond to the same degree of extremeness: for instance, a cluster with $\delta_{\text{mean}} = 20$ constitutes an extreme outlier in the $N \rightarrow A$ distribution — where the outlier threshold is 15.71 — while in the $N \rightarrow T$ distribution the same value falls below the median of 22.61, indicating a below-average degree of disruption.

A joint normalization of the two distributions to a common scale would not resolve this issue without introducing a different problem: the difference in spread and shape between the two distributions is itself biologically informative, reflecting the fact that tumor tissue exhibits a qualitatively different and more extreme remodeling of co-methylation structure than adjacent tissue. Collapsing the two distributions onto a common scale would erase this information.

The appropriate object of comparison is therefore not the absolute magnitude of disruption, but the *relative position* of each cluster within its own disruption distribution — that is, how extreme a cluster's behaviour is with respect to the typical behaviour observed on that axis. The Adjusted Outlyingness score introduced in Section 5.3.2 is adopted precisely for this purpose: it expresses each cluster's disruption relative to the centre and tails of its own empirical distribution, while explicitly accounting for skewness via the MedCouple correction. As a result, AO values share a common interpretation across the two comparisons: a cluster with $\mathrm{AO}_{\mathrm{ext}} = 1.2$ on both axes is unambiguously an extreme outlier in both settings, while a cluster with $\mathrm{AO}_{\mathrm{ext}} = -0.2$ is stable in both — regardless of the absolute scale of the underlying $\delta$ values. This enables a meaningful joint analysis of cluster behaviour across the $N \to A$ and $N \to T$ disruption axes, without imposing any assumption of distributional equivalence between them.

To explore the grade of outlyiness of each cluster in comparison with respect to both metrics a scatter plot was used. Each point represents a cluster, and its position in the space quantifies its degree of outlyingness with respect to both disruption metrics. Clusters shown in grey exhibit low $\mathrm{AO}_{\mathrm{ext}}$ scores on both axes, indicating stable co-methylation structure across both the $N \to A$ and $N \to T$ transitions, and are therefore of limited biological relevance. Quadrant $Q_1$ contains clusters that are outlying on both axes simultaneously, meaning their mean methylation pattern is disrupted in both adjacent and tumor tissue relative to normal. Quadrant $Q_2$ groups clusters that are outlying exclusively along the $N \to T$ axis, while Quadrant $Q_4$ collects those disrupted only in the $N \to A$ comparison. As shown in Figure 5.4, 1216 out of 3891 clusters exhibit stable behaviour on both axes, meaning that, despite being defined as co-methylated on normal samples only, they preserve their co-methylation signal in both adjacent and tumor tissue, showing no detectable early or late epigenetic alterations.

In contrast, Quadrant Q1, which should in principle capture clusters consistently displaying strong disruption of the normal co-methylation landscape across both comparisons, only contains 4 clusters. The sparsity of Quadrant $Q_1$ indicates that simultaneous strong disruption along both axes is not the prevailing biological mechanism linking the $N \to A$ and $N \to T$ comparisons. This suggests that, if a joint structure connecting the two metrics exists, it manifests through subtler patterns across the full distribution of clusters, rather than through the co-occurrence of extreme outlyingness on both axes.

Nonetheless, a closer inspection of Quadrant $Q_2$ (19 clusters) reveals that a portion of them already display moderate AO score in the $N \to A$ comparison, suggesting that their disruption may not arise abruptly in tumor but rather amplifies a weaker signal already present in adjacent tissue. Conversely, clusters in Quadrant $Q_4$ (39) outliers exclusively along the $N \to A$ axis — seem to less display a corresponding moderate signal in the $N \to T$ comparison. This may suggest that

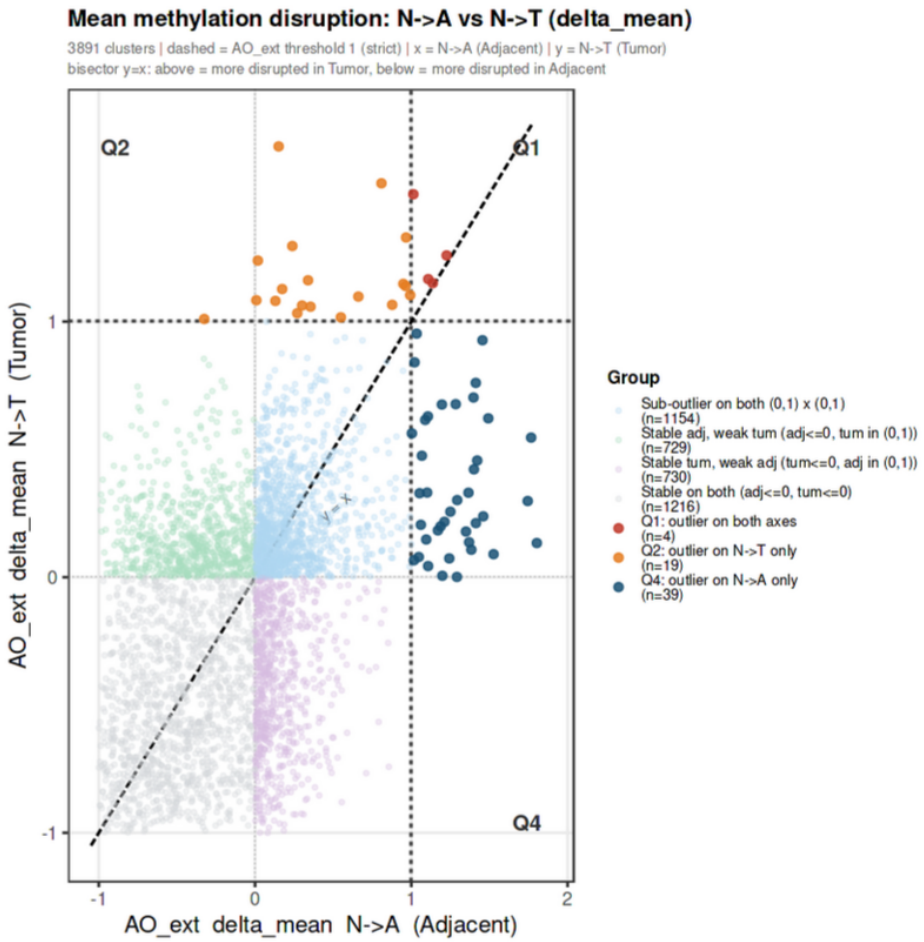


**Figure 5.4:** Mean Methylation Disruption behaviour across clusters from SACOMA configuration [maxGap=1000bp, r=0.5]

their disruption is specific to the adjacent field and does not propagate or amplify in tumor tissue.

This pattern becomes more evident when examining Figure 5.5, which displays clusters ranked in increasing order of outlyingness according to one metric (shown in green), and coloured according to their AO score on the other axis, ranging from blue (stable) to red (strongly disrupted). An adjacent binary encoding is also provided, marking clusters as outliers (red), sub-outliers (light blue), or stable (grey) in the other comparison.

In the left heatmap, where $\delta^{NA}_{\text{mean}}$ defines the ranking, the colour scale of the $N \to T$ axis reveals predominantly light orange and white tones among the $N \to A$ outliers: among the 43 $N \to A$ outliers, only approximately 9 display intensely

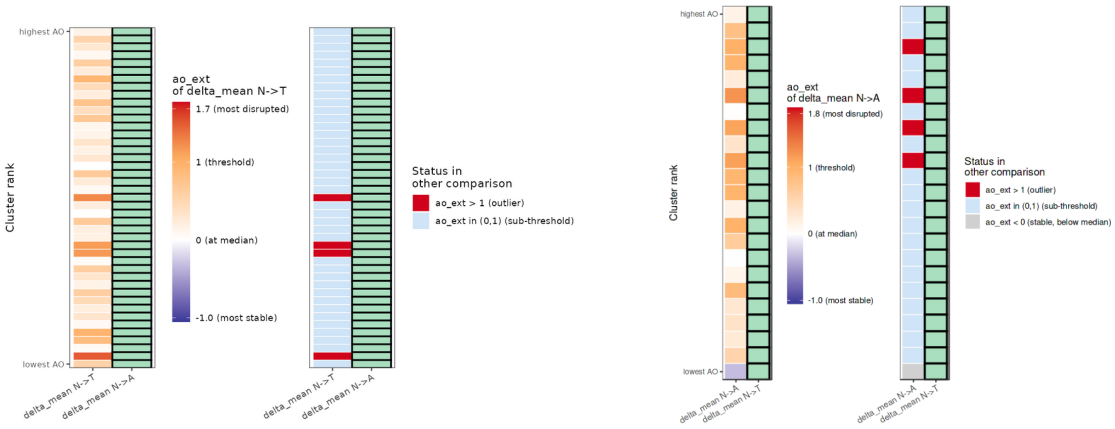


**Figure 5.5:** Outliers comparison

orange colouring on the $N \to T$ axis. This may suggest that their disruption in adjacent tissue is either an early, not yet consolidated alteration that does not propagate to tumor, or a stochastic and heterogeneous signal more likely to be cohort-specific.

The right heatmap, where $\delta^{NT}_{\text{mean}}$ defines the ranking, reveals a qualitatively different picture. Among the 23 $N \to T$ outliers, approximately 11 already display moderate $\text{AO}_{\text{ext}}$ scores in the $N \to A$ comparison, appearing as sub-outliers on the adjacent axis — consistent with the progressive amplification hypothesis suggested by the scatter plot. A smaller subset of clusters shows $N \to A$ scores close to the median, with one case falling below it, indicating strong adjacent stability; these may represent late-stage, tumor-specific disruption events. Notably, the clusters belonging to $Q_1$ — outliers on both axes — are those with the strongest $N \to T$ disruption, yet their $N \to A$ scores remain only moderately above threshold, further supporting the asymmetric and progressive nature of co-methylation breakdown.

Taken together, these observations suggest that a binary outlier classification based on a single axis discards potentially relevant information. Among the 1154 clusters classified as sub-outliers on both axes, a subset may in fact follow the progressive disruption trajectory — exhibiting moderate adjacent signal and stronger tumor signal — but remain below the strict threshold. Since $\text{AO}_{\text{ext}} \in (0, 1)$ already implies a disruption above the median of the reference distribution, sub-outlier clusters may still carry biologically relevant signal that a strict binary classification would discard. This motivates a joint, continuous use of both metrics as a selection criterion, complementary to the univariate outlier detection on either axis alone.

## 5.4 Hypothesis-Driven Composite Score for Cluster Selection

As discussed in Section 2.1, field defects arise from epigenetic alterations in normal stem cells that confer a reproductive advantage, leading to the clonal expansion of epigenetically altered cell patches that progressively accumulate further alterations along the path toward tumorigenesis [4]. As access to the normal cells at the exact locations where cancers arise is naturally impossible, the only approach in humans to identify early molecular alterations is by comparison of normal tissue from healthy individuals to the normal tissue found adjacent to tumours [7]. Recent research indicates that cells within a field defect characteristically exhibit an increased frequency of epigenetic alterations, and these may be fundamentally important as underlying factors in progression to cancer [7].

In this section, mean composite score is defined to jointly account for both the $N \to A$ and $N \to T$ comparisons. The underlying intuition is that the

two metrics are not biologically independent: a score that considers only one comparison at a time cannot capture the directional dependency assumed under the field cancerization hypothesis. The goal of the composite scores defined here is therefore to identify clusters whose disruption pattern along the $N \rightarrow A \rightarrow T$ axis is consistent with a plausible evolutionary trajectory toward tumorigenesis.

### 5.4.1 Previous Biological Evidence

The biological rationale underlying the proposed score is motivated by the observations reported by Teschendorff et al. [7]. In their study, performed on the same external validation cohort considered in this work, the authors showed that a conventional differential methylation analysis between normal and normal-adjacent tissues failed to identify genome-wide significant field defects. They argued that early DNA methylation alterations are highly heterogeneous and stochastic across individuals, making mean-based differential methylation approaches insufficient for detecting these early events.

To account for the heterogeneity and stochasticity of epigenetic alterations in normal-adjacent tissue, Teschendorff et al proposed the iEVORA algorithm, previously presented in Chapter 3. During their analysis, Teschendorff et al. [7] explored the hypothesis that the identified epigenetic alterations were not confined to normal-adjacent tissue but represent early manifestations of a progressive biological process. Specifically, they argue that if the DNAm defects identified in normal-adjacent tissue mark cells that are important for the development of breast cancer, one would expect these loci to exhibit larger differences in DNAm when comparing breast cancers to normal samples from healthy subjects. To test this, they compute differential methylation statistics between breast cancers and normal tissue for all previously identified DVMCs, observing that the great majority exhibit significant differential DNAm changes in breast cancer, with much stronger associations than a randomly selected set of CpGs.

These results provide the conceptual basis for the progression hypothesis adopted in this thesis: the field defects captured by DVMCs in normal-adjacent tissue, despite their inherent stochasticity, may already constitute early manifestations of the same disruption process that becomes fully established, with substantially larger magnitude, in tumour tissue.

### 5.4.2 Mean Disruption Score Definition

The score proposed in this thesis builds on this progression hypothesis, but reverses the logic used by Teschendorff et al. to establish it: rather than starting from alterations identified in normal-adjacent tissue and proposing a post-hoc validation criterion via their amplification in tumour, the progressive change hypothesis is

used directly as a selection rule, jointly combining $AO_k^{NA}$ and $AO_k^{NT}$ within a single composite score.

It is important to note that, while the score is motivated by a progression hypothesis, the underlying analysis remains cross-sectional: samples from the Normal, Adjacent, and Tumor compartments are not derived from the same patients over time, and the proposed score cannot support direct inference of temporal causality.

Additionally, the use of AO scores rather than raw values for the definition of the combined score is motivated by the fact that both axes are thus expressed in units of their respective adjusted boxplot half-widths, making the score invariant to the absolute scale of each metric.

We now formalize both the considerations and the associated modeling choices for score definition.

**Eligibility domain.** First, the search for early field defects is restricted to a biologically motivated eligibility domain $E = \{k : \mathrm{AO}_k^{N\to A} > 0,\ \mathrm{AO}_k^{N\to T} > 0\}$. The rationale for this joint condition is the following. Clusters with $\mathrm{AO}_k^{N\to A}$ near or above the outlier threshold but with $\mathrm{AO}_k^{N\to T} \leq 0$ do not show any amplified signal in Tumour and therefore cannot be interpreted as field defects under the progressive change hypothesis – their disruption, if present, is confined to the Adjacent stage and does not progress. Symmetrically, clusters with $\mathrm{AO}_k^{N\to T} > 0$ but $\mathrm{AO}_k^{N\to A} \leq 0$ exhibit a tumour-associated signal without any early trace in normal-adjacent tissue: these may represent late-emerging, tumour-specific alterations rather than field defects, and fall outside the biological hypothesis being tested here. Clusters with both scores below zero are stable in co-methylation across both comparisons and should also be excluded. This filtering step defines a hypothesis-driven search space rather than a global ranking, thereby constraining the analysis to a subset of candidates consistent with the progression hypothesis.

**Multiplicative combination.** Second, the two signals are combined multiplicatively rather than additively. An additive combination would assign a high score even when only one of the two signals is strong; a multiplicative combination, by contrast, encodes a necessary coexistence condition between the two signals.

**Asymmetric role of the two axes.** Third, the two signals are not treated symmetrically within the score. Tumor-associated outlyingness defines the core of the signal: the score requires, as a necessary condition, a high degree of outlyingness in Tumor, since this is the target signal the score aims to capture – the higher the degree of outlyingness associated with a cluster in the Normal-Tumor comparison, the stronger its evidence of disruption. This translates into a linear involvement of $\mathrm{AO}_k^{N\to T}$ in the score.

The role of Adjacent-tissue outlyingness is instead more subtle. As shown by the exploratory analysis presented in Section 5.3.3, clusters showing the highest degree of outlyingness in the $N \to A$ comparison tend not to be outliers in the $N \to T$ comparison, reinforcing the view that the strongest Adjacent outliers are not necessarily carrying the progressive signal under investigation. Consequently, the Adjacent-tissue signal is not included linearly but smoothed via a bounded modulating function – here the hyperbolic tangent, although other saturating functions could in principle be investigated – so as to actively prevent $\mathrm{AO}_k^{N\to A}$ from dominating the selection when the degree of outlyingness in Tumour is comparatively weak. This limits the impact of extreme normal-adjacent outliers on the final score.

In particular, for each cluster $k$ in the eligibility domain $E$, the Tumor-associated outlyingness $\mathrm{AO}_k^{N\to T}$ is multiplied by the following amplification coefficient:

$$\alpha_k = 1 + \tanh\Big(\mathrm{AO}_k^{N\to A}\Big) \in (1,2).$$

The additive shift ensures that $\alpha_k$ starts at exactly 1 when $\mathrm{AO}_k^{N\to A} = 0$ and increases monotonically toward a bounded maximum of 2. Its interpretation is therefore that of a genuine reward: any positive degree of Adjacent-tissue outlyingness amplifies the Tumor-associated signal, even below the threshold of $\mathrm{AO} = 1$ conventionally used to define a strong outlier on its own scale. The saturation of tanh guarantees, at the same time, that this reward remains bounded, preventing extreme values of $\mathrm{AO}_k^{N\to A}$ from disproportionately dominating the score.

To further illustrate the necessity of the asymmetric formulation, consider two clusters with the following scores:

| | $\mathrm{AO}^{N\to T}$ | $\mathrm{AO}^{N\to A}$ |
|---|---|---|
| Cluster A | 0.6 | 2.0 |
| Cluster B | 1.2 | 0.7 |

**Table 5.3:** Illustrative example of two clusters with contrasting outlyingness profiles across the two axes.

Under a symmetric multiplicative score $S^{\times}(k) = \mathrm{AO}_k^{N\to T} \cdot \mathrm{AO}_k^{N\to A}$:

$$S^{\times}(\mathrm{A}) = 0.6 \times 2.0 = 1.200$$
$$S^{\times}(\mathrm{B}) = 1.2 \times 0.7 = 0.840$$

Cluster A is preferred. However, this ranking is biologically questionable: $\mathrm{AO}^{N\to T} = 0.6$ places Cluster A only marginally above the median of the $N \to T$ distribution, indicating a weak signal of co-methylation disruption in the tumour state, while

the high $\mathrm{AO}^{N\to A} = 2.0$ may instead reflect cohort-specific epigenetic variability in normal-adjacent tissue rather than a biologically meaningful early field defect. Cluster B, by contrast, exhibits a robust outlier signal in Tumour ($\mathrm{AO}^{N\to T} = 1.2 > 1$) accompanied by a moderate but present disruption in Adjacent ($\mathrm{AO}^{N\to A} = 0.7 < 1$, i.e., below the strong-outlier threshold, yet clearly positive) – consistent with the progressive change pattern in which alterations initiated in normal-adjacent tissue become amplified in breast cancer. Under $S_{\text{mean}}$ as defined in Equation 5.15:

$$S^{\tanh}(\mathrm{A}) = 0.6 \cdot (1 + \tanh(2.0)) = 0.6 \cdot 1.964 = 1.178$$
$$S^{\tanh}(\mathrm{B}) = 1.2 \cdot (1 + \tanh(0.7)) = 1.2 \cdot 1.604 = 1.925$$

Cluster B is now correctly preferred. The modulator $(1 + \tanh(\mathrm{AO}^{N\to A}))$, bounded in $(1, 2)$, amplifies the primary signal proportionally to the degree of disruption in Adjacent, but cannot compensate for a weak $\mathrm{AO}^{N\to T}$: even with $\mathrm{AO}^{N\to A} = 2.0$ fully saturating the modulator at $\approx 1.964$, the low primary factor $\mathrm{AO}^{N\to T} = 0.6$ keeps Cluster A's score below that of Cluster B.

**Prevent late effect driven signals.** A potential limitation of $S^{\tanh}$ is that clusters with very high $\mathrm{AO}_k^{N\to T}$ and low, yet strictly positive, $\mathrm{AO}_k^{N\to A}$ – which may represent late tumour-specific alterations rather than early field defects – could still receive high scores, despite the initial restriction to the eligibility domain $E$. A further refinement is possible to account for late-effect-driven signals. The idea is to introduce a hyperparameter $\lambda$ that shifts the modulating function to the left, updating the score definition into the following final formulation:

$$S^{\tanh}(k) = \begin{cases} \mathrm{AO}_k^{N\to T} \cdot \Big(1 + \tanh\Big(\mathrm{AO}_k^{N\to A} - \lambda\Big)\Big) & \text{if } k \in E \\ 0 & \text{otherwise} \end{cases} \tag{5.15}$$

The rationale behind the design of $S^{\tanh}$ can be understood by considering its two regimes, separated by the threshold $\lambda$. For $0 < \mathrm{AO}_k^{N\to A} < \lambda$, the multiplicative factor $\Big(1 + \tanh(\mathrm{AO}_k^{N\to A} - \lambda)\Big)$ lies strictly below 1, acting as a penalizing factor for clusters that have not yet entered the moderate-outlyingness phase in Adjacent tissue. This design specifically penalizes clusters that appear strongly disrupted in Tumor tissue without a correspondingly pronounced early signal in Adjacent tissue: such clusters are treated as inconsistent with a progressive N→A→T trajectory, regardless of how extreme their Tumor-stage disruption is.

Beyond the threshold ($\mathrm{AO}_k^{N\to A} > \lambda$), the model enters a regime in which Adjacent outlyingness is considered sufficiently pronounced: the multiplicative factor exceeds 1 (approaching an asymptotic value of 2), and $S^{\tanh}$ effectively becomes an amplifier of $\mathrm{AO}_k^{N\to T}$. As $\mathrm{AO}_k^{N\to A}$ increases beyond $\lambda$, the contribution

of Adjacent disruption to the composite score increases up to a bounded maximum, reflecting the idea that early co-methylation instability already visible in Adjacent tissue strengthens the biological relevance of the corresponding disruption observed in Tumor.

A visual representation of the behaviour of $\alpha_k$ as a function of $\mathrm{AO}_k^{N\to A}$ is provided in Figure 5.6.

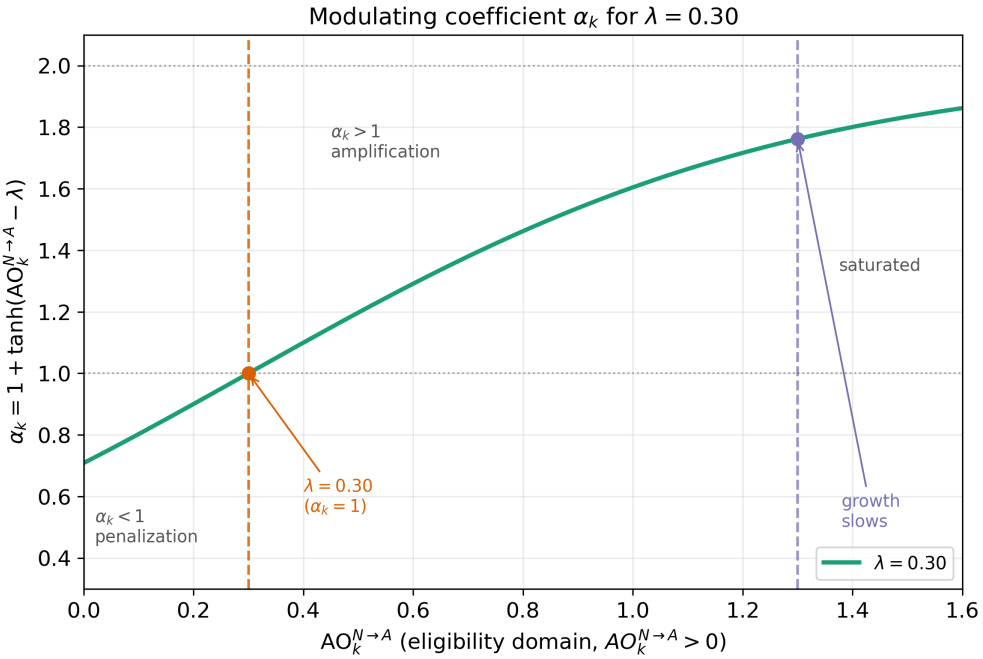


**Figure 5.6:** Behaviour of the modulating coefficient $\alpha_k = 1 + \tanh(\mathrm{AO}_k^{N\to A} - \lambda)$ as a function of $\mathrm{AO}_k^{N\to A}$, illustrated for $\lambda = 0.30$. Below the threshold $\lambda$ (orange dashed line), $\alpha_k < 1$ and the Tumor-associated signal is penalized. Above the threshold, $\alpha_k > 1$ and the signal is amplified, with the growth rate progressively slowing beyond $\mathrm{AO}_k^{N\to A} = \lambda + 1$ (purple dashed line) as the coefficient approaches its bounded maximum of 2.

A final scheme for the methodological rationate behind the mean distruption composite score definition is provided in Figure 5.7.

In summary, four distinct selection strategies are investigated with respect to the $\delta_{\text{mean}}$ metric: clusters selected as outliers according to $\delta_{\text{mean}}^{N\to T}$ alone, clusters selected according to $\delta_{\text{mean}}^{N\to A}$ alone, and clusters selected via the two composite scores, $S^{\tanh}$ and $S^{\times}$, to empirically demonstrate the necessity of rewarding a moderate, rather than extreme, degree of outlyingness in Adjacent. Additionally, for $S^{\tanh}$ a sensitivity analysis was conducted over $\lambda = \{0.20, 0.30, 0.40\}$, the lower bound beyond which $AO_k^{N\to A}$ is considered to enter a moderate-outlyingness regime. For

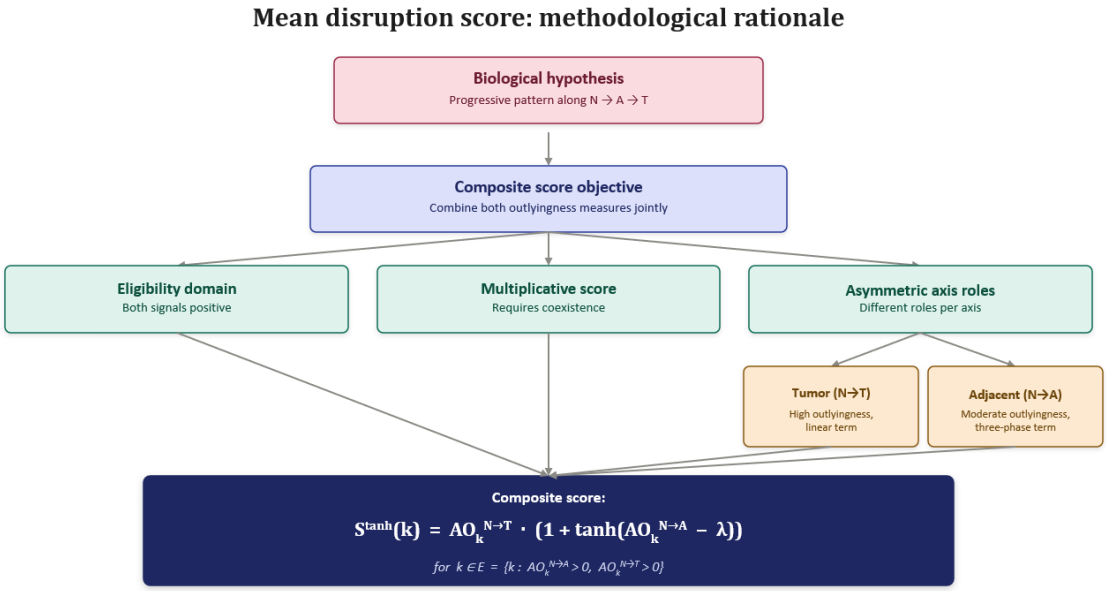


**Figure 5.7:** Methodological Rationale for the Composite Score Definition

every type of multiplicative score, the final selection consists of the clusters whose score strictly exceeds the 95th and 99th percentile of the empirical score distribution over the eligibility domain $E$, corresponding to the top 5% and top 1% thresholds, respectively.

Which of these four subsets is most biologically meaningful for characterizing early field defects is assessed via indirect validation through XGBoost classification.

## 5.5 Classification via XGBoost

This section presents the final classification method chosen to indirectly validate the feature selection pipeline developed in this thesis: XGBoost. XGBoost (eXtreme Gradient Boosting) [51] is a scalable implementation of gradient-boosted decision trees, an ensemble learning method in which weak learners (shallow regression trees) are added sequentially, each one trained to correct the residual errors of the ensemble built so far. At each boosting iteration $t$, the model updates its prediction as

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i), \tag{5.16}$$

where $f_t$ is the new tree added at iteration $t$, selected to greedily minimize a regularized objective function combining a differentiable loss term (here, logistic loss, given the binary classification tasks addressed in this thesis) and a penalty on tree complexity, which controls overfitting by discouraging overly deep or unbalanced

trees. This additive, regularized formulation makes XGBoost particularly suited to high-dimensional, moderate-sample-size settings such as the CpG feature sets considered here, and its widespread adoption in methylation-based classification tasks motivated its selection as the downstream classifier for this validation step.

Given the high dimensionality and comparatively small sample size typical of these tasks, careful hyperparameter tuning is required to avoid overfitting. The hyperparameter tuning procedure adopted in this thesis was adapted from [52], via randomized search (`RandomizedSearchCV`, `scikit-learn`) over the following grid:

- `max_depth` $\in \{2, 3\}$
- `n_estimators` $\in \{500, 800, 1200\}$
- `subsample` $\in \{0.5, 0.25\}$
- `colsample_bytree` $\in \{0.5, 0.25, 0.125\}$
- `learning_rate` $= 0.189$ (fixed)

The search sampled $n_{\text{iter}} = 5$ hyperparameter combinations, each evaluated via 3-fold cross-validation ($K_{\text{CV}} = 3$) on the EPIC training set, optimizing for AUC-ROC (`scoring = "roc_auc"`). Both the base model and the search procedure used a fixed random seed (`random_state = 42`) to ensure reproducibility. The hyperparameter combination achieving the highest mean cross-validated AUC-ROC was selected, and a final model was refit on the entire EPIC training set using these best-performing hyperparameters.

**Evaluation.** The final model was evaluated on two independent sets: the EPIC test set and the external 450k validation cohort (GSE69914). For each evaluation set, model performance was summarized using two metrics: the area under the ROC curve (AUC-ROC) and classification accuracy.

The AUC-ROC quantifies the model's ability to discriminate between the two classes across all possible probability thresholds. Given predicted probabilities $\hat{p}_i \in [0,1]$ and true labels $y_i \in \{0,1\}$, the ROC curve plots the true positive rate against the false positive rate as the classification threshold varies, and the AUC-ROC is defined as the probability that a randomly chosen positive sample is assigned a higher predicted probability than a randomly chosen negative sample:

$$\text{AUC-ROC} = P\left(\hat{p}_i > \hat{p}_j \mid y_i = 1,\ y_j = 0\right). \tag{5.17}$$

Unlike accuracy, this metric is threshold-independent and therefore provides a global measure of separability between classes.

Accuracy, computed at the default classification threshold of 0.5, is defined as the proportion of correctly classified samples:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{5.18}$$

where $TP$, $TN$, $FP$ and $FN$ denote true positives, true negatives, false positives and false negatives, respectively. Since the Normal and Adjacent classes were approximately balanced in both the EPIC test set and the external 450k validation cohort, accuracy remains an informative and interpretable complement to AUC-ROC in this setting, without being distorted by class imbalance effects.

# Chapter 6

# Results

This chapter compares the CpG clusters selected under several strategies in terms of their actual ability to enable XGBoost classifier to distinguish Normal versus Adjacent breast tissue, evaluated on both the EPIC test set and the external 450k cohort. The analysis is conducted on clusters obtained after a sensitivity analysis from the most promising SACOMA configuration, i.e. maxGap = 2000bp and rthreshold = 0.4; results for the remaining explored configurations are reported in Appendix A for completeness.

First, results from the different multiplicative scores are compared in terms of classification performance; a visual representation of the selected clusters is then provided to characterize how the different scores affect the composition of the selected subsets. Second, a comparison between the best-performing configuration and the independent, single-axis selection via Adjbox-$\delta$ is provided; the best-performing configuration is also benchmarked against two established methods from the literature, iEVORA and limma. Since iEVORA was originally developed using GSE69914 as discovery set – the same dataset used here as external validation cohort – particular attention is given to verifying whether the CpGs driving the best-performing configuration overlap with those identified by iEVORA: the absence of substantial overlap would suggest that the proposed pipeline captures a signal that is, at least in part, orthogonal to that recovered by existing single-CpG differential methods. Third, to assess whether the observed classification performance could plausibly arise by chance, the best-performing subset is compared against a null distribution of AUROC values obtained by repeatedly sampling random clusters of matched cardinality (e.g., 1,000 random selections of 53 CpGs), positioning the observed performance relative to this empirical baseline and thereby demonstrating the necessity of an informed selection strategy over random subsetting. Finally, the best-performing configuration is mapped to genes and genomic regions, and its biological relevance is assessed through gene set and pathway enrichment analysis, together with a comparison against the COSMIC cancer gene census, providing

a biological interpretation of the disruption patterns captured by the proposed composite score.

## 6.1 Comparison between selection strategies

This section first compare the multiplicative scores in terms of performance. On the internal test set, all methods and $\lambda$ values yield comparably strong AUROC (0.86–0.95), since the test set shares the same underlying batch and platform characteristics as the training data. Since cross-dataset generalization is the most challenging task to achieve, it is more informative to compare the different scores in terms of performance on GSE69914.

**Table 6.1:** Generalization trend on the external 450k cohort, stratified by selection cardinality (top 1% vs. top 5%). AUROC values below 0.5 are marked in bold to highlight generalization failure.

| | | **Top 1%** | | **Top 5%** | |
|---|---|---|---|---|---|
| **Method** | $\lambda$ | $n_{\text{CpG}}$ | **AUROC** | $n_{\text{CpG}}$ | **AUROC** |
| $S^{\times}$ (no modulator) | – | 53 | 0.7014 | 275 | **0.3748** |
| $S^{\tanh}$ | 0 | 53 | 0.7133 | 272 | 0.6943 |
| $S^{\tanh}$ | 0.20 | 53 | 0.7248 | 271 | 0.6848 |
| $S^{\tanh}$ | 0.30 | 53 | 0.7505 | 269 | 0.6424 |
| $S^{\tanh}$ | 0.40 | 53 | 0.7510 | 270 | **0.4457** |

Table 6.1 reveals substantial and divergent differences in generalization capability depending on both $\lambda$ and the cardinality of the selected cluster set.

At the top 1% selection level ($n \approx 53$ clusters), increasing $\lambda$ monotonically improves generalization to the external cohort, from 0.7133 ($\lambda = 0$) to 0.7510 ($\lambda = 0.40$).

At the top 5% selection level ($n \approx 270$ clusters), however, the trend reverses sharply: AUROC drops from 0.6943 ($\lambda = 0$) to 0.4457 ($\lambda = 0.40$) – a value below chance level. This suggests that, once the selection is relaxed to include a substantially larger number of clusters, an aggressive $\lambda$ increasingly favors clusters whose apparent Adjacent-stage outlyingness reflects EPIC-specific noise or batch artifacts rather than a genuine, cross-platform field-defect signal – a textbook case of overfitting to platform-specific structure, masked by strong within-cohort performance.

This effect is most evident when comparing against $S^{\times}$, the purely multiplicative score without the tanh modulator. At top 5% selection, $S^{\times}$ collapses to an AUROC

of 0.3748 on the external cohort, the worst generalization performance observed across all configurations. This provides strong empirical evidence in support of the modulating role of the tanh term in $S^{\tanh}$: without it, the score is more prone to selecting clusters that generalize poorly once the selection is broadened beyond the most extreme outliers.

To explore the effect of different multiplicative scores on the final selection, a scatter plot on the eligible domain is provided at the 5% threshold, which does not yield the best performance but includes a larger number of clusters, making it more informative for illustrating how the scores differ. In particular, the clusters selected with $\lambda = 0$, $\lambda = 0.30$, $\lambda = 0.40$ and $S^{\times}$ (no modulator) are shown in Figure 6.1. The

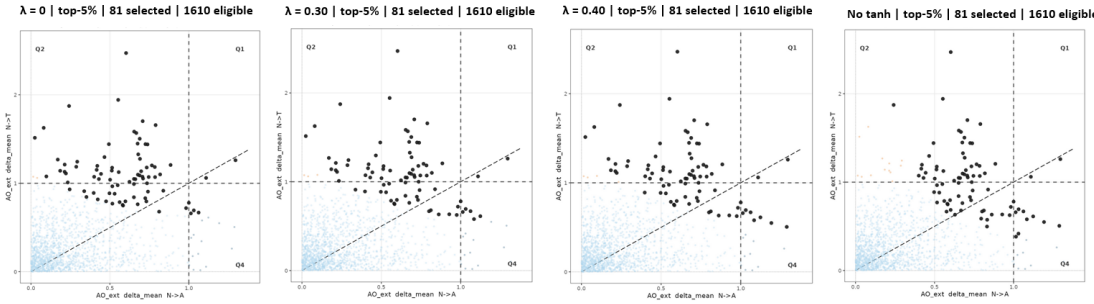


**Figure 6.1:** Visual comparison of different multiplicative score selection

general trend that appears to emerge is the following. The score without $\lambda$, which therefore exhibits only two regimes – a strong amplification of the Tumor signal as the degree of Adjacent disruption increases, and a saturation regime, in which excessively high degrees of Adjacent outlyingness no longer contribute to the score proportionally to their magnitude but in a saturated fashion – effectively selects relatively few clusters from Q4. The selection instead shifts toward clusters in the left and central region, concentrated around the Tumor outlyingness threshold. It therefore appears to favor clusters that are moderately disrupted in Tumor, with a moderate-to-low signal in Adjacent. Increasing $\lambda$ produces a shift toward selecting an increasing number of clusters belonging to the Q4 region, of which the simple multiplicative score represents the most pronounced example.

This can be explained by noting that, as $\lambda$ increases, the threshold required to qualify for amplification rises: among the clusters surviving the top-5% selection, those with a moderate-to-low $AO^{N\to A}$ are more easily excluded, falling into the now wider penalization region, whereas those with an already extreme $AO^{N\to A}$ remain competitive because they still exceed the higher threshold.

Finally, a comparison with independent, single-axis selection criteria is performed.

**Table 6.2:** Single-axis selection (Adjbox-$\delta$, no score composition), across the EPIC test set and the external 450k cohort.

| | | **EPIC Test** | | **External 450k** | |
|---|---|---|---|---|---|
| **Axis** | $n_{\text{CpG}}$ | **AUROC** | **Accuracy** | **AUROC** | **Accuracy** |
| N→A | 68 | 0.9369 | 0.80 | **0.3448** | 0.42 |
| N→T | 184 | 0.9229 | 0.77 | 0.7276 | 0.70 |

In Table 6.2 emerges that clusters selected as outliers on the Adjacent axis alone fail to generalize to the external cohort, consistent with the exploratory analysis discussed above: the strongest Adjacent outliers do not reliably carry the progressive signal under investigation. Clusters selected as outliers on the Tumor axis alone, by contrast, generalize almost as well as the best $S^{\text{tanh}}$ configuration – marginally better on the EPIC test set, marginally worse on the external 450k cohort. The key difference lies in panel size: the Tumor-only selection requires 184 CpGs to achieve this performance, whereas $S^{\text{tanh}}$ (top-1%, $\lambda = 0.30$) reaches comparable external generalization with only 53 CpGs – roughly a threefold reduction.

Figure 6.2 visualizes the clusters selected under the best-performing configuration. As expected from the score's design, the selected clusters form a subset of the Tumor-axis outliers, positioned high on the $N \to T$ axis and at a moderate degree of outlyingness on the $N \to A$ axis – the exact pattern the composite score was designed to reward.

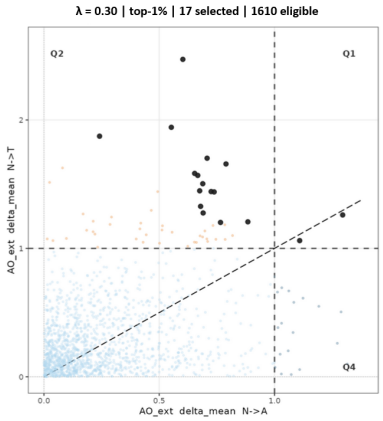


**Figure 6.2:** Clusters selected under the best-performing configuration ($S^{\text{tanh}}$, top-1%, $\lambda = 0.30$), shown against the full eligible domain.

This comparable-or-better generalization achieved with roughly a third of the CpGs raises a natural question: is the smaller $S^{\tanh}$ panel itself a particularly informative subset, or does the Tumor-axis selection as a whole carry the discriminative signal, with $S^{\tanh}$ merely picking out CpGs that would have performed just as well on their own? The latter would undermine the rationale for the composite score altogether, since the same panel could then be obtained by simply taking the top Tumor outliers directly. To disentangle these two explanations, an ablation experiment was performed: the 53 CpGs selected by the best $S^{\tanh}$ configuration were removed from the 184-CpG Adjbox-$\delta$ Tumor-axis selection, leaving a remainder of 131 CpGs. This remainder was evaluated under the same classification pipeline and hyperparameter search used throughout this chapter.

**Table 6.3:** Ablation experiment: ablated CpGs, isolated from the 184-CpG N→T Adjbox-$\delta$ selection, from which the CpGs shared with the best $S^{\tanh}$ configuration were removed.

| | | **EPIC Test** | | **External 450k** | |
|---|---|---|---|---|---|
| **Selection** | $n_{\text{CpG}}$ | **AUROC** | **Accuracy** | **AUROC** | **Accuracy** |
| $S^{\tanh}$, best | 53 | 0.8597 | 0.73 | 0.7505 | 0.70 |
| Adjbox-$\delta$, N→T | 184 | 0.9229 | 0.77 | 0.7276 | 0.70 |
| N→T \ best | 131 | 0.9284 | 0.82 | **0.5814** | 0.54 |

The remainder set fits the internal EPIC test set even better (AUROC = 0.9284) but its performance on the external 450k cohort collapses by roughly 17 percentage points (from 0.7505 to 0.5814), approaching the level expected under random classification. This result indicates that the 53 CpGs selected by $S^{\tanh}$ are not simply a representative sample of the Tumor-axis outlier pool, but a specifically informative subset: removing them from the larger Tumor-axis selection is sufficient to erase most of its external generalization capacity, even though the remaining 131 CpGs still outperform it internally. This supports the conclusion that the asymmetric treatment of Adjacent-tissue outlyingness introduced by $S^{\tanh}$ is not redundant with a Tumor-only selection criterion, and that the resulting panel captures signal that is disproportionately relevant to the underlying progression hypothesis.

## 6.2 Benchmarking against iEVORA and limma

This section presents a comparison between the proposed pipeline and two established methods from the literature: limma and iEVORA. Starting from the 149147

CpGs obtained after the preliminary technical filtering steps, both methods were applied to the same EPIC train set used throughout this thesis, yielding very large CpG subsets: 25684 for iEVORA and 110142 for limma. For iEVORA, the code was derived directly from the associated paper; limma, on the other hand, is a package installed via Bioconductor. The resulting CpG sets from each method were then used for classification, replicating the same training procedure presented in 5.5.

Results are reported in Table A.3.

**Table 6.4:** Benchmarking against iEVORA and limma methods

| | | | | **EPIC Test** | | **External 450k** | |
|---|---|---|---|---|---|---|---|
| **Selection** | **Method** | $\lambda$ | $n_{\text{CpG}}$ | **AUROC** | **Accuracy** | **AUROC** | **Accuracy** |
| Top 1% | $S^{\tanh}$ | 0.30 | 53 | 0.8597 | 0.73 | 0.7505 | 0.70 |
| $p < 0.05$ | iEVORA | – | 25684 | 0.9648 | 0.87 | 0.3819 | 0.43 |
| $p_{\text{BH}} < 0.05$ | limma | – | 110142 | 0.9814 | 0.91 | 0.3810 | 0.43 |

On the internal EPIC test set, the CpGs selected by the other two methods lead to higher classification performance than the CpG subset identified by the method presented in this thesis, which nonetheless remains high (AUROC 0.8597, accuracy 0.73). Cross-dataset, cross-platform generalization on GSE69914 is instead good for the proposed subset (AUROC 0.7505, accuracy 0.70), whereas the other two methods show a marked drop of approximately 37% in AUROC and approximately 30% in terms of accuracy.

The intersection between the CpG sets obtained by the three methods was also examined. 25470 CpGs are shared between limma and iEVORA alone. Of the 53 CpGs selected by the proposed method, 52 are also found by limma and 40 are also found by iEVORA.

Finally, it was verified whether the CpGs identified by the proposed method also rank among the CpGs most strongly supported by limma and iEVORA, respectively, within their own significant subsets. For limma, CpGs are ranked directly by the moderated t-test p-value, with lower values indicating stronger evidence of differential methylation. For iEVORA, CpGs are ranked according to the unadjusted p-value of the t-test, with a significance threshold of 0.05, computed on the CpGs previously selected as differentially variable via Bartlett's test (FDR $< 0.001$).

For both methods, rank deciles were computed on the corresponding p-value distribution and a histogram was used to visualize where the proposed method's CpGs fall within that ranking. Results are reported in Figure 6.3.

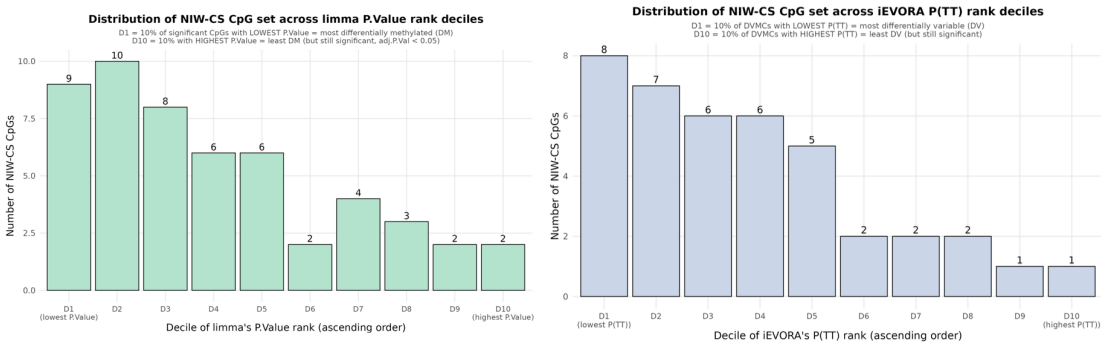


**Figure 6.3:** Distribution of the CpGs selected by the proposed pipeline across the p-value rank deciles of limma and iEVORA.

A clear concentration in the lowest (most significant) deciles is observed for both methods:

- limma: D1+D2 = 19/52 (36.5%) vs. D9+D10 = 4/52 (7.7%)
- iEVORA: D1+D2 = 15/40 (37.5%) vs. D9+D10 = 2/40 (5%)

This indicates that, on the one hand, the established methods are able to capture the signal identified by the proposed pipeline, but they also capture a large amount of additional signal that appears to be cohort-specific. At the same time, the best selected subset captures a consistent share of CpGs that limma independently ranks among its most differentially methylated probes, and that iEVORA independently ranks among its most differentially methylated probes within the subset of CpGs it has already identified as differentially variable.

## 6.3 Random Feature Baseline

A final experiment was conducted to further assess the robustness of the proposed feature selection pipeline. Identifying predictive features from high-dimensional datasets is a major task in biomedical research; however, it is difficult to determine the robustness of selected features. Developing and optimizing predictive models as combinations of algorithms, features, and other design choices requires testing baselines to demonstrate the value of these choices [53]. In this section, the performance of randomly chosen features, referred to as "random feature baselines" (RFBs), is investigated in the context of Normal - Adjacent prediction. For machine learning workflows that include feature selection, RFBs represent an important benchmark for confirming the value of the selected features; for biomarker development, they provide a test of the predictive performance of the candidate

biomarker. This type of analysis is particularly relevant given findings such as those of Venet et al. [54], who demonstrated that in breast cancer any set of 100 or more randomly selected genes has a 90% chance of being significantly associated with outcome.

Consistently with Venet et al., the experiment was conducted by randomly selecting 1000 subsets of CpGs of the same cardinality as the best configuration of the proposed pipeline, drawn from the 149147 CpGs obtained after the initial filtering steps, prior to cluster definition via SACOMA and the feature selection procedure proposed in this thesis. For each subset, an XGBoost classifier was then trained independently, preserving the same implementation details described in Section 5.5 for the best configuration, and AUROC performance was recorded. The resulting empirical distributions of AUROC were then visualized for both the EPIC test set and the external Illumina 450k validation set.

Results are reported in Figure 6.4.

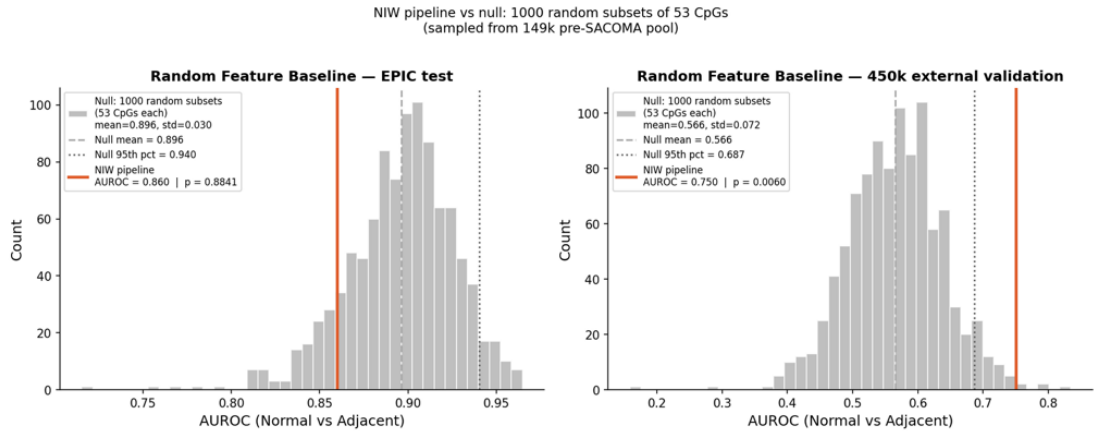


**Figure 6.4:** Random Feature Baseline experiment results in terms of AUROC null distributions over the EPIC test set and the Illumina 450k external validation set.

A notable phenomenon can be observed: the two null AUROC distributions differ substantially in scale. On the EPIC test set, the AUROC achieved by randomly selected feature sets ranges approximately between 0.75 and 0.95, suggesting that it is nearly impossible to obtain a poor classification performance in this cohort. This may reflect the fact that, in these cohorts, adjacent tissue is not strictly "normal" in a biological sense but is instead closer to tumor tissue, so that differences from normal tissue are more diffuse and more pronounced throughout the genome. The external validation cohort, by contrast, is recognized in the literature as being considerably less informative in terms of differences between normal and adjacent

tissue, and this is reflected in the corresponding null distribution, which is strongly shifted to the left, with a minimum AUROC of 0.20 and a maximum of 0.8. The proposed subset, achieving an AUROC of 0.75 on this cohort, falls beyond the 95th percentile of the null distribution, supporting the validity of the proposed feature selection pipeline.

These results indicate that the performance achieved is non-trivial: cross-platform generalization is inherently difficult, cross-cohort generalization even more so, and this is further compounded in the case of the external cohort considered here, which has been explicitly reported by Teschendorff et al. [2] to be biologically less informative. Generalization is therefore particularly challenging in this setting, and although the proposed method does not achieve very high performance in absolute terms, it nonetheless yields consistently reasonable results, suggesting that the captured signal is unlikely to be pure background noise and may instead reflect biologically relevant information. While these computational results provide a promising starting point, biological validation — including ad hoc laboratory data collection — is necessary before any definitive conclusion can be drawn. The thesis therefore concludes with the following section, which presents preliminary biological analyses, following established practice, and outlines directions for future work.

## 6.4 Biological Interpretation

Finally, an initial attempt at biological interpretation of the results was carried out. First, the 53 CpGs belonging to the best configurations were mapped to their corresponding genomic regions with respect to CpG islands. Each CpG was annotated using the Illumina HumanMethylation450 manifest, with respect to both its genomic context (Island, Shore, Shelf, or Open Sea, based on the `Relation_to_UCSC_CpG_Island` field) and its gene association, derived from the `UCSC_RefGene_Name` and `UCSC_RefGene_Group` fields.

The majority of the selected CpGs (58.5%) map to Open Sea regions, i.e. genomic regions located far from annotated CpG islands, followed by Shore (24.5%), Island (11.3%), and Shelf (5.7%). This enrichment in Open Sea regions may suggest a preferential involvement of distal regulatory elements, such as enhancers, rather than classical promoter-associated CpG islands. Secondly, the CpGs were mapped to their corresponding genic regions. Of the 53 CpGs, 38 showed a correspondence with a gene, yielding a final list of 13 genes. Table 6.6 reports the result, together with the corresponding genic region.

The majority of CpGs map to the gene Body (i.e., the region spanning from the end of the 5' UTR to the beginning of the 3' UTR), followed by TSS200 (23.7%) and TSS1500 (21.1%), the regions located respectively within 200 bp and 1500 bp

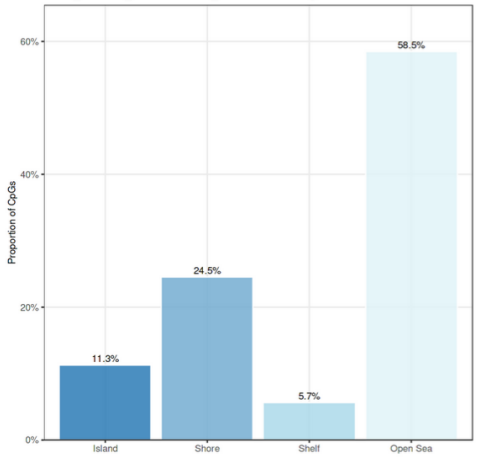


**Figure 6.5:** Genomic Regions mapping

| Gene | n_CpG | CGI_relation |
|---|---|---|
| MYOM2 | 4 | Open Sea |
| CMKLR1 | 3 | Open Sea |
| CRYBG3 | 3 | Open Sea |
| LOC100128554 | 3 | Open Sea |
| LOC283070 | 3 | Open Sea |
| ODZ2 | 3 | Open Sea |
| PTPRN2 | 3 | Open Sea |
| SFRS5 | 3 | Shore |
| SLIT2 | 3 | Island |
| ZFAT | 3 | Shore |
| ZNF423 | 3 | Shore |
| ZNF578 | 3 | Island |
| ADCYAP1R1 | 1 | Open Sea |

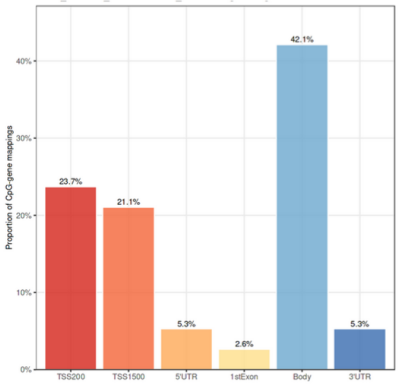


**Figure 6.6:** Gene & Genic Regions mapping

upstream of the transcription start site (TSS).

### 6.4.1 Gene Involvement in Cancer

Gene involvement in cancer was then explored. The 13 genes were first queried on the Human Protein Atlas platform [55] to assess whether their involvement in cancer, in general, is already recognized. Table 6.5 provides a comprehensive overview.

It is important to notice that, although annotated as "not found" in The Human Protein Atlas, LOC283070 has been functionally characterized as a long non-coding RNA involved in the development of androgen independence in prostate cancer

[56].

Moreover, the major consideration emerging from this first screening is that none of the identified genes shows a HPA-annotated association specifically with breast cancer; most prognostic associations instead concern other tumor types, particularly Kidney Renal Clear Cell Carcinoma (KIRC). This observation was further corroborated by cross-referencing the gene list with the COSMIC Cancer Gene Census restricted to genes strictly associated with breast cancer, which yielded zero overlapping genes. This absence may indicate that the selected CpGs capture more general cancer-related or regulatory processes rather than breast cancer-specific drivers, although it may also reflect the identification of epigenetic signals not yet annotated in current breast cancer-specific databases and therefore worthy of further investigation. Given this absence of a direct annotated link, a more in-depth literature search was conducted for each gene, to assess whether more recent or more specific studies had reported an association with breast cancer beyond what is captured by HPA and COSMIC.

**MYOM2** Although MYOM2 is annotated as "not detected" in cancer tissues by The Human Protein Atlas, additional evidence from the literature points to a relevant role in breast and prostate cancer. Kitz et al. [57] reported downregulation of MYOM2 in a clinical assessment of breast cancer patients, consistent with earlier multiplex RT-PCR data, and observed increased DNA methylation at the MYOM2 locus, suggesting promoter hypermethylation as a plausible mechanism underlying its silencing. Furthermore, cBioPortal analysis showed that decreased MYOM2 expression is associated with significantly worse progression-free survival in prostate cancer patients compared to those with high expression, indicating that MYOM2 silencing may have broader prognostic relevance beyond breast cancer.

**CMKLR1** CMKLR1, encoding the resolution-of-inflammation receptor ChemR23, is not expressed by tumor cells themselves but is instead a marker of tumor-associated macrophages (TAMs) infiltrating the tumor stroma; its expression correlates with TAM markers in breast cancer and its pharmacological activation was shown to reduce metastasis in a triple-negative breast cancer model, suggesting that the co-methylation disruption observed at this cluster may reflect immune microenvironment remodeling rather than a cell-intrinsic tumor alteration [58].

**ZNF423** Beyond the "prognostic marker in KIRC" annotation reported by HPA, ZNF423 has a well-characterized and mechanistically detailed role specifically in hormone-responsive breast cancer. ZNF423, epigenetically regulated via promoter CpG island hypermethylation, is induced by estrogen in breast cancer cells and trans-activates BRCA1, linking estrogen signaling to DNA damage response modulation;

SNPs in ZNF423 were also associated with reduced breast cancer risk during SERM-based preventive therapy [59].

**PTPRN2** PTPRN2 has been directly implicated in breast cancer metastasis through a specific molecular mechanism: it enzymatically reduces plasma membrane PI(4,5)P2 levels in metastatic breast cancer cells via two independent pathways, promoting actin remodeling and enhancing cell migration. Both genes are upregulated in highly metastatic breast cancer cells, and their increased expression is associated with metastatic relapse in patients, positioning PTPRN2 as a functionally validated driver of breast cancer metastatic progression.[60]

**SRSF5** SRSF5 has been mechanistically implicated in tamoxifen resistance in ER-positive breast cancer, which represents approximately 70% of breast cancer patients treated with adjuvant endocrine therapy. Low SRSF5 expression promotes production of the alternative splice variant BQ323636.1 (BQ) of NCOR2, a known driver of tamoxifen resistance, while SRSF5 overexpression reduces BQ production and restores drug sensitivity — a relationship confirmed in vitro, in vivo, and via tissue microarray analysis showing a significant inverse correlation between SRSF5 and BQ expression. Clinically, low SRSF5 expression was associated with tamoxifen resistance, local recurrence, metastasis, and poorer overall survival, positioning SRSF5 as both a prognostic marker and a potential therapeutic target in ER-positive breast cancer[61].

**SLIT2** SLIT2 is associated with epigenetic silencing.Dallol et al [62] showed that the SLIT2 promoter was methylated in 59% of breast cancer cell lines and 43% of primary breast tumors. Treating these cell lines with a demethylating agent restored SLIT2 expression, showing that methylation — not mutation — was directly responsible for silencing the gene. Most methylated tumors also showed allelic loss at the SLIT2 locus (4p15.2), and methylated tumors had lower SLIT2 expression than normal tissue (measured by qRT-PCR). When SLIT2 was re-expressed in breast cancer cell lines that normally lacked it, colony growth dropped by more than 70%. Together, these results identify SLIT2 as a strong tumor suppressor candidate, silenced in breast (and lung) cancer mainly through promoter hypermethylation combined with allelic loss.

**TENM2** Across multiple cancer types, TENM2 downregulation has been consistently reported: in prostate cancer, expression progressively decreases from locally advanced to metastatic stages; in early-stage cervical cancer, it is downregulated in tumors with lymph node metastases; and in breast hyperplastic lobular units — a premalignant precursor lesion — expression is significantly reduced compared

to normal tissue. In ovarian cancer, lower TENM2 expression correlates with poorly differentiated, more aggressive tumors, suggesting a tumor-suppressive role. Notably, however, TENM2 downregulation in both ovarian and breast cancer cells has been experimentally shown to be independent of DNA methylation — demethylating agents failed to restore expression in ovarian cancer cells despite the presence of predicted CpG islands, and downregulation induced by prolactin/estradiol treatment in breast cancer cells followed the same epigenetics-independent pattern — pointing instead to regulation via FGF8/FGF receptor signaling [63]

Overall, while none of the 13 genes identified shows an established, breast cancer-specific association in the Human Protein Atlas or COSMIC databases, a deeper literature search revealed the majority have been reported to be involved in breast cancer-related processes, through mechanisms ranging from direct promoter hypermethylation (*SLIT2*, *ZNF423*, *MYOM2*) to metastasis-promoting enzymatic activity (*PTPRN2*), immune microenvironment remodeling (*CMKLR1*), and therapy resistance (*SFRS5*). This methylation-based analysis should ideally be complemented by matched mRNA expression data, to determine whether the observed methylation changes at CpGs mapped to these genes translate into actual changes in gene activation — particularly given evidence, such as that for *TENM2*, that gene downregulation is not always methylation-dependent. This represents a natural direction for future work: matched transcriptomic data are available for at least one of the two EPIC training cohorts (GSE225845), which would allow a direct integration of DNA methylation and gene expression changes across the N→A→T progression.

**Table 6.5:** Cancer relevance of the genes identified in the analysis. Prognostic Summary and Cancer Specificity values are retrieved from The Human Protein Atlas [55].

| Gene | CpGs | Involvement in cancer | |
|---|---|---|---|
| | | **Prognostic Summary** | **Cancer Specificity** |
| MYOM2 | 4 | Not prognostic. | Not detected. |
| CMKLR1 | 3 | Not prognostic. | Low cancer specificity. |
| CRYBG3 | 3 | Prognostic marker in Kidney Chromophobe (KICH) and Kidney Renal Clear Cell Carcinoma (KIRC). | Low cancer specificity. |
| LOC100128554 | 3 | Not found | Not found |
| LOC283070 | 3 | Not found | Not found |
| ODZ2 (TENM2) | 3 | Not prognostic. | Group-enriched expression in Head and Neck Squamous Cell Carcinoma (HNSC) and Lung Squamous Cell Carcinoma (LUSC). |
| PTPRN2 | 3 | Prognostic marker in Glioblastoma Multiforme (GBM), Kidney Renal Clear Cell Carcinoma (KIRC), and Liver Hepatocellular Carcinoma (LIHC). | Cancer enhanced in Prostate Adenocarcinoma |
| SFRS5 | 3 | Prognostic marker in Kidney Renal Clear Cell Carcinoma (KIRC). | Low cancer specificity |
| SLIT2 | 3 | Not prognostic. | Cancer-enriched in Kidney Chromophobe (KICH). |
| ZFAT | 3 | Prognostic marker in Bladder Urothelial Carcinoma (BLCA), Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma (CESC), Glioblastoma Multiforme (GBM), Kidney Chromophobe (KICH), Kidney Renal Clear Cell Carcinoma (KIRC), and Pancreatic Adenocarcinoma (PAAD). | Low cancer specificity |
| ZNF423 | 3 | Prognostic marker in Kidney Renal Clear Cell Carcinoma (KIRC). | Cancer enhanced in Testicular Germ Cell Tumor |
| ZNF578 | 3 | Prognostic marker in Kidney Renal Clear Cell Carcinoma (KIRC). | Cancer enriched in Testicular Germ Cell Tumor |
| ADCYAP1R1 | 1 | Not prognostic. | Cancer-enriched in Glioblastoma Multiforme (GBM). |

*Note:* Prognostic Summary and Cancer Specificity values are reported as provided by The Human Protein Atlas [55]. Specificity of RNA expression across 17 cancer types is categorized as cancer enriched, group enriched, cancer enhanced, low cancer specificity, or not detected.

# Chapter 7

# Conclusions

The analytical framework presented in this thesis adopts a semi-targeted approach [21], in which CpG-clusters are identified on normal breast tissue samples from healthy women, and the NIW-based KL divergence decomposition is then applied to quantify how much Adjacent and Tumor samples deviate from the normal co-methylation structure. In particular, Normal tissue provides an unambiguous epigenetic baseline, and the two decomposed components, mean shift ($\delta_\mu$) and co-methylation structure disruption ($\delta_\Sigma$), acquire a clear interpretation precisely because the reference distribution encodes what coordinated methylation looks like in the absence of disease. Both components were found to increase in magnitude when moving from the Normal-Adjacent to the Normal-Tumor comparison, providing initial support for the hypothesis that co-methylation clusters are increasingly disrupted, in both location and structure, along the N→A→T axis. However, only the mean-shift component was ultimately operationalized into a cluster-selection criterion in this work, for reasons discussed in Section 7.2 The exploration of the mean disruption metric, both in terms of maximizing the difference between Normal and Adjacent tissue and in terms of exploiting the additional information contained in Tumor samples to identify clusters exhibiting a progressive degree of disruption, revealed that the latter strategy is more promising for identifying candidate biomarkers within the validation framework adopted in this thesis. Under a specific hyperparameter configuration, this strategy achieved both strong internal AUROC on the EPIC test set and non-trivial AUROC on the external 450k cohort. This result is reinforced by the direct comparison with the single-axis alternative: selecting clusters based on outlyingness in the Normal-Adjacent comparison alone yielded external generalization at chance level (AUROC $\approx 0.34$), whereas the progressive selection strategy reached an external AUROC of 0.75, providing direct evidence, within the same validation framework, in favor of the progression hypothesis over the single-comparison alternative. Notably, although the resulting composite score was not designed ad hoc to optimize Normal-versus-Adjacent discrimination, it

outperformed established methods from the literature, such as iEVORA and limma, on the external cohort, while relying on a panel of only 53 CpGs, several orders of magnitude smaller than the panels selected by these methods.

## 7.1 Limitations

While the biological questions posed at the outset of this thesis were addressed by the proposed framework, its limitations must also be acknowledged.

Cluster definition depends on the choice of the SACOMA hyperparameters. Although the ranges explored in the sensitivity analysis are biologically motivated by the existing literature, no principled a priori method currently exists for determining these hyperparameters, and establishing one would require a dedicated, larger-scale study specifically devoted to this question. As a consequence, this choice affects the validation patterns observed across the different configurations. Addressing this limitation properly would require access to a larger number of independent datasets, or dedicated simulation studies, in order to systematically characterize how the resulting clusters and downstream disruption metrics depend on these choices. This thesis limited itself to conducting a sensitivity analysis and to selecting, among the explored configurations, the one achieving the best performance under the validation criteria defined a priori; consequently, the reported external validation performance for that configuration should be interpreted as an upper bound on generalization performance rather than as a fully unbiased estimate, since the external cohort itself contributed to the selection of the configuration.

Related to this point, it would also be worth verifying whether the pipeline remains consistent when the spatially-aware clustering algorithm itself is changed, replacing SACOMA with alternative methods that share the same spatial-proximity constraint.

Another limitation, related more in general to semi-targeted approaches, discussed explicitly in the differential co-expression literature [21], is that they are blind to co-methylation structures that do not exist in the reference condition. If a group of CpG sites exhibits no coordinated methylation in normal tissue, it does not emerge as a cluster under SACOMA, but it is biologically possible that they can acquire a coherent co-methylation pattern in tumor or adjacent tissue, that the current framework is not able to capture by construction.

A further limitation concerns the Bayesian model itself: in deriving the posterior distribution of the mean, the covariance matrix is plugged in at its posterior expectation, so that a degree of distributional detail is inevitably lost in the process. An alternative would have been to work directly with the Kullback-Leibler divergence of the full joint posterior distribution; however, this would have come at the cost of the interpretability of the resulting disruption metrics, which was

prioritized in this work.

The procedure used to assign the prior hyperparameters could also be conducted in a more sophisticated manner, for instance by incorporating additional, independent datasets containing methylation information from normal tissue samples only, so as to obtain a more robust and better-powered estimate of the baseline co-methylation architecture.

The analyses should also be repeated including, at the preprocessing stage, an explicit correction for cellular heterogeneity, and including, at the classification stage, relevant clinical covariates known to be associated with the disease, such as age, which is provided for only one cohort but can be estimated for the others.

Finally, regarding the composite score, the hyperbolic tangent adopted as the modulating function could in principle be replaced with alternative modulating functions, and the sensitivity of the results to this specific choice was not explored in this work.

## 7.2 Future Directions

The most immediate extension of this work is the operational use of the covariance disruption metric. In the present study, only the mean-shift component was exploited for cluster selection, while the covariance disruption measure was derived but not incorporated into the selection criterion. Future work could investigate its use either as an alternative to the mean disruption metric or in combination with it, requiring clusters to exhibit both changes in average methylation levels and alterations in their co-methylation architecture. In this context, the work of Teschendorff et al. [8] provides an interesting biological perspective, suggesting that covariance disruption may follow a non-linear trajectory, reaching its maximum in pre-cancerous tissue before partially stabilizing in established tumours. Exploring this hypothesis within the proposed Bayesian framework could provide additional insight into the dynamics of epigenetic disruption during carcinogenesis.

Another natural and methodologically coherent extension concerns the semi-targeted nature of the proposed pipeline. Since clusters are defined exclusively on the reference condition, the current framework is unable to identify co-methylation structures that emerge de novo in disease. This limitation could be addressed by repeating the analysis using tumour tissue as the reference condition, thereby reversing the analytical perspective. In this setting, SACOMA would first identify co-methylation clusters characteristic of the tumour epigenome, and the NIW model would use tumour-derived prior hyperparameters to quantify how normal and adjacent tissues deviate from these disease-specific structures. Such an analysis would complement the current study by characterizing co-methylation architectures that arise specifically during tumour development.

An addictional direct extension of this work concerns the biological validation of the identified signal. Matched RNA-seq data, available for at least one of the EPIC cohorts used in this thesis (GSE225845), could be used to verify whether the methylation alterations detected at the selected CpG clusters translate into corresponding changes in gene expression, thereby providing orthogonal support for their biological relevance beyond the methylation layer alone.

Finally, the proposed framework could be naturally extended to other cancer types for which matched normal, adjacent, and tumour methylation datasets are available. Since neither SACOMA nor the Bayesian NIW model relies on assumptions specific to breast tissue, the methodology is, in principle, directly transferable to other solid tumours, such as colorectal, lung, or prostate cancer. Applying the same analytical pipeline across multiple cancer types would not only provide further methodological validation but also allow the investigation of whether progressive disruption of local co-methylation represents a general feature of carcinogenesis.

In conclusion, this thesis contributes to the field of early breast cancer detection by proposing a probabilistic framework for identifying potential methylation biomarkers based on alterations in local co-methylation patterns across disease states. Beyond the methodological contribution, this work challenges the conventional biomarker discovery paradigm by investigating whether early biomarkers are better characterized by pairwise differences between normal and adjacent tissue or by coherent molecular trajectories spanning the Normal–Adjacent–Tumour continuum. Within the proposed validation framework, the latter hypothesis proved more effective, suggesting that progressive disruption of local co-methylation may provide a more informative signature of early carcinogenesis.

Although the computational findings require dedicated biological validation before any clinical translation, they provide quantitative support for the Field Cancerization Hypothesis and demonstrate how probabilistic modeling, statistical inference, and machine learning can be integrated with biological knowledge to investigate the earliest epigenetic alterations associated with breast cancer development.

# Appendix A

# Results from other SACOMA hyperparameters configurations

**Table A.1:** SACOMA parameters: `maxGap` = 1000 bp, $\rho = 0.5$.

| | | | | **EPIC Test** | | **External 450k** | |
|---|---|---|---|---|---|---|---|
| **Selection** | **Method** | $\lambda$ | $n_{\text{CpG}}$ | **AUROC** | **Accuracy** | **AUROC** | **Accuracy** |
| Top 1% | $S^{\tanh}$ | 0 | 40 | 0.8415 | 0.78 | 0.7224 | 0.64 |
| | | 0.20 | 40 | 0.8890 | 0.78 | 0.6881 | 0.63 |
| | $S^{\tanh}$ | 0.30 | 40 | 0.8263 | 0.77 | 0.6676 | 0.59 |
| | | 0.40 | 40 | 0.8263 | 0.77 | 0.6676 | 0.59 |
| | $S^{\times}$ | – | 39 | 0.8542 | 0.77 | 0.6929 | 0.63 |
| Top 5% | $S^{\tanh}$ | 0 | 200 | 0.9492 | 0.81 | 0.5424 | 0.54 |
| | | 0.20 | 201 | 0.9530 | 0.81 | 0.5248 | 0.48 |
| | $S^{\tanh}$ | 0.30 | 201 | 0.9682 | 0.79 | 0.6405 | 0.63 |
| | | 0.40 | 199 | 0.9589 | 0.78 | 0.5814 | 0.54 |
| | $S^{\times}$ | – | 204 | 0.9356 | 0.82 | 0.3886 | 0.42 |
| Adjbox-$\delta_{\text{mean}}$ | N→A axis | - | 136 | 0.9538 | 0.84 | 0.3681 | 0.45 |
| | N→T axis | - | 77 | 0.8970 | 0.77 | 0.7248 | 0.66 |

**Table A.2:** SACOMA parameters: `maxGap` = 2000 bp, $\rho = 0.5$.

| | | | | **EPIC Test** | | **External 450k** | |
|---|---|---|---|---|---|---|---|
| **Selection** | **Method** | $\lambda$ | $n_{\text{CpG}}$ | **AUROC** | **Accuracy** | **AUROC** | **Accuracy** |
| Top 1% | $S^{\tanh}$ | 0 | 50 | 0.9119 | 0.77 | 0.4714 | 0.53 |
| | | 0.20 | 50 | 0.8966 | 0.77 | 0.4533 | 0.46 |
| | $S^{\tanh}$ | 0.30 | 49 | 0.9225 | 0.78 | 0.5319 | 0.54 |
| | | 0.40 | 49 | 0.9225 | 0.78 | 0.5319 | 0.54 |
| | $S^{\times}$ | – | 49 | 0.8805 | 0.76 | 0.6543 | 0.64 |
| Top 5% | $S^{\tanh}$ | 0 | 249 | 0.9428 | 0.81 | 0.6748 | 0.64 |
| | | 0.20 | 252 | 0.9593 | 0.80 | 0.6348 | 0.62 |
| | $S^{\tanh}$ | 0.30 | 252 | 0.9538 | 0.78 | 0.6171 | 0.61 |
| | | 0.40 | 252 | 0.9538 | 0.78 | 0.6171 | 0.61 |
| | $S^{\times}$ | – | 255 | 0.9648 | 0.82 | 0.4990 | 0.47 |
| Adjbox-$\delta_{\text{mean}}$ | N→A axis | – | 127 | 0.9169 | 0.77 | 0.4548 | 0.45 |
| | N→T axis | – | 80 | 0.9169 | 0.77 | 0.7152 | 0.72 |

**Table A.3:** SACOMA parameters: `maxGap` = 1000 bp, $\rho = 0.4$.

| | | | | **EPIC Test** | | **External 450k** | |
|---|---|---|---|---|---|---|---|
| **Selection** | **Method** | $\lambda$ | $n_{\text{CpG}}$ | **AUROC** | **Accuracy** | **AUROC** | **Accuracy** |
| Top 1% | $S^{\tanh}$ | 0 | 49 | 0.8966 | 0.79 | 0.5433 | 0.58 |
| | | 0.20 | 49 | 0.9157 | 0.76 | 0.5519 | 0.61 |
| | $S^{\tanh}$ | 0.30 | 49 | 0.9157 | 0.76 | 0.5519 | 0.61 |
| | | 0.40 | 49 | 0.9157 | 0.76 | 0.5519 | 0.61 |
| | $S^{\times}$ | – | 48 | 0.8949 | 0.79 | 0.5286 | 0.53 |
| Top 5% | $S^{\tanh}$ | 0 | 258 | 0.9475 | 0.81 | 0.5486 | 0.53 |
| | | 0.20 | 259 | 0.9708 | 0.81 | 0.5895 | 0.54 |
| | $S^{\tanh}$ | 0.30 | 261 | 0.9496 | 0.79 | 0.5395 | 0.52 |
| | | 0.40 | 261 | 0.9496 | 0.79 | 0.5395 | 0.52 |
| | $S^{\times}$ | – | 262 | 0.9453 | 0.87 | 0.3595 | 0.43 |
| Adjbox-$\delta_{\text{mean}}$ | N→A axis | – | 122 | 0.9619 | 0.84 | 0.4143 | 0.46 |
| | N→T axis | – | 156 | 0.9475 | 0.83 | 0.6643 | 0.64 |

# Bibliography


[1] Freddie Bray, Mathieu Laversanne, Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Isabelle Soerjomataram, and Ahmedin Jemal. «Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries». In: *CA: A Cancer Journal for Clinicians* 74.3 (2024), pp. 229–263. DOI: https://doi.org/10.3322/caac.21834. eprint: https://acsjournals.onlinelibrary.wiley.com/doi/pdf/10.3322/caac.21834. URL: https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.3322/caac.21834 (cit. on p. 1).

[2] Andrew E. Teschendorff, Yuan Gao, Ashley Jones, et al. «DNA methylation outliers in normal breast tissue identify field defects that are enriched in cancer». In: *Nature Communications* 7 (2016), p. 10478. DOI: 10.1038/ncomms10478 (cit. on pp. 2, 3, 12, 13, 64).

[3] Carol Bernstein, Harris Bernstein, C. Melville Payne, Karel Dvorak, and Hardeep Garewal. «Field Defects in Progression to Gastrointestinal Tract Cancers». In: *Cancer Letters* 260.1–2 (2008), pp. 1–10. DOI: 10.1016/j.canlet.2007.11.027. URL: https://doi.org/10.1016/j.canlet.2007.11.027 (cit. on p. 6).

[4] Carol Bernstein, Valery Nfonsam, Anil R. Prasad, and Harris Bernstein. «Epigenetic field defects in progression to cancer». In: *World Journal of Gastrointestinal Oncology* 5.3 (2013), pp. 43–49. DOI: 10.4251/wjgo.v5.i3.43 (cit. on pp. 6, 47).

[5] Marie-Claude Pouliot, Yvan Labrie, Caroline Diorio, and Francine Durocher. «The Role of Methylation in Breast Cancer Susceptibility and Treatment». In: *Anticancer Research* 35.9 (Sept. 2015), pp. 4569–4574 (cit. on pp. 6, 8).

[6] D. P. Slaughter, H. W. Southwick, and W. Smejkal. «Field cancerization in oral stratified squamous epithelium; clinical implications of multicentric origin». In: *Cancer* 6.5 (Sept. 1953), pp. 963–968. DOI: 10.1002/1097-0142(195309)6:5<963::AID-CNCR2820060515>3.0.CO;2-Q (cit. on p. 7).

[7] AE Teschendorff, Y Gao, A Jones, M Ruebner, MW Beckmann, DL Wachter, PA Fasching, and M Widschwendter. «DNA methylation outliers in normal breast tissue identify field defects that are enriched in cancer». In: *Nature Communications* 7 (Jan. 29, 2016), p. 10478. DOI: 10.1038/ncomms10478 (cit. on pp. 7, 25, 30, 47, 48).

[8] Andrew E. Teschendorff, Xiaoping Liu, Helena Carén, Steven M. Pollard, Stephan Beck, Martin Widschwendter, and Luonan Chen. «The Dynamics of DNA Methylation Covariation Patterns in Carcinogenesis». In: *PLoS Computational Biology* 10.7 (2014), e1003709. DOI: 10.1371/journal.pcbi.1003709. URL: https://doi.org/10.1371/journal.pcbi.1003709 (cit. on pp. 7, 31, 72).

[9] Hafiz Ahmed, Aziz Ullah, Abdul Malik, and Babar Islam. «Review; Role of epigenetics in cancer». In: *World Journal of Biology and Biotechnology* 5 (July 2020), pp. 51–54. DOI: 10.33865/wjb.005.02.0304 (cit. on p. 7).

[10] Sudhir Sharma, Timothy K. Kelly, and Peter A. Jones. «Epigenetics in Cancer». In: *Carcinogenesis* 31.1 (2010), pp. 27–36. DOI: 10.1093/carcin/bgp220. URL: https://doi.org/10.1093/carcin/bgp220 (cit. on p. 7).

[11] A. M. Deaton and A. Bird. «CpG islands and the regulation of transcription». In: *Genes & Development* 25.10 (2011), pp. 1010–1022. DOI: 10.1101/gad.2037511 (cit. on p. 8).

[12] Edvardas Kaminskas et al. «Approval Summary: Azacitidine for Treatment of Myelodysplastic Syndrome Subtypes». In: *Clinical Cancer Research* 11.10 (2005), pp. 3604–3608. DOI: 10.1158/1078-0432.CCR-04-2135. URL: https://doi.org/10.1158/1078-0432.CCR-04-2135 (cit. on p. 9).

[13] Li Sun, Sreekanth Namboodiri, Enyi Chen, and Shuyu Sun. «Preliminary Analysis of Within-Sample Co-methylation Patterns in Normal and Cancerous Breast Samples». In: *Cancer Informatics* 18 (2019), p. 1176935119880516. DOI: 10.1177/1176935119880516 (cit. on pp. 9, 10).

[14] J. T. Bell, A. A. Pai, J. K. Pickrell, D. J. Gaffney, R. Pique-Regi, J. F. Degner, Y. Gilad, and J. K. Pritchard. «DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines». In: *Genome Biology* 12.1 (2011), R10. DOI: 10.1186/gb-2011-12-1-r10 (cit. on pp. 9, 32, 33).

[15] Hongyu Ren, Robert B. Taylor, T. Luke Downing, and Elizabeth L. Read. «Locally Correlated Kinetics of Post-Replication DNA Methylation Reveals Processivity and Region Specificity in DNA Methylation Maintenance». In: *Journal of the Royal Society Interface* 19.195 (2022), p. 20220415. DOI: 10.1098/rsif.2022.0415. URL: https://doi.org/10.1098/rsif.2022.0415 (cit. on pp. 9, 32, 33).

[16] Olga Taryma-Leśniak, Jakub Bińkowski, Piotr K. Przybylowicz, et al. «Methylation patterns at the adjacent CpG sites within enhancers are a part of cell identity». In: *Epigenetics & Chromatin* 17.1 (2024), p. 30. DOI: 10.1186/s13072-024-00555-5 (cit. on p. 10).

[17] Shuai Wei et al. «Ten Years of EWAS». In: *Advanced Science* 8.20 (2021), e2100727. DOI: 10.1002/advs.202100727. URL: https://doi.org/10.1002/advs.202100727 (cit. on p. 11).

[18] Dongmei Li, Zongli Xie, Marjorie L. Pape, and Timothy Dye. «An Evaluation of Statistical Methods for DNA Methylation Microarray Data Analysis». In: *BMC Bioinformatics* 16 (2015), p. 217. ISSN: 1471-2105. DOI: 10.1186/s12859-015-0641-x. URL: https://doi.org/10.1186/s12859-015-0641-x (cit. on p. 12).

[19] Gordon K. Smyth. «Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments». In: *Statistical Applications in Genetics and Molecular Biology* 3.1 (2004), Article 3. ISSN: 1544-6115. DOI: 10.2202/1544-6115.1027. URL: https://doi.org/10.2202/1544-6115.1027 (cit. on p. 12).

[20] Belinda Phipson and Alicia Oshlack. «DiffVar: a new method for detecting differential variability with application to methylation in cancer and aging». In: *Genome Biology* 15.9 (2014), p. 465. DOI: 10.1186/s13059-014-0465-4 (cit. on p. 13).

[21] Bruno Tesson, Rainer Breitling, and Ritsert Jansen. «DiffCoEx: A simple and sensitive method to find differentially coexpressed gene modules». In: *BMC bioinformatics* 11 (Oct. 2010), p. 497. DOI: 10.1186/1471-2105-11-497 (cit. on pp. 14, 70, 71).

[22] Timothy J. Peters, Matthew J. Buckley, Aaron L. Statham, Ruth Pidsley, Katherine Samaras, Richard V. Lord, Susan J. Clark, and Peter L. Molloy. «De Novo Identification of Differentially Methylated Regions in the Human Genome». In: *Epigenetics & Chromatin* 8 (2015), p. 6. ISSN: 1756-8935. DOI: 10.1186/1756-8935-8-6. URL: https://doi.org/10.1186/1756-8935-8-6 (cit. on p. 14).

[23] Andrew E. Jaffe, Peter Murakami, Heather Lee, Jeffrey T. Leek, M. Daniele Fallin, Andrew P. Feinberg, and Rafael A. Irizarry. «Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies». In: *International Journal of Epidemiology* 41.1 (2012), pp. 200–209. DOI: 10.1093/ije/dyr238 (cit. on p. 14).

[24] Siddhant Meshram, Arindam Fadikar, Ganesan Arunkumar, and Suvo Chatterjee. «A spatially-aware unsupervised pipeline to identify co-methylation regions in DNA methylation data». In: *bioRxiv* (2025). DOI: 10.1101/2025.11.13.688356. eprint: https://www.biorxiv.org/content/early/2025/11/14/2025.11.13.688356.full.pdf. URL: https://www.biorxiv.org/content/early/2025/11/14/2025.11.13.688356 (cit. on pp. 15, 33, 34).

[25] Marie Chavent, Véronique Kuentz-Simonet, Alexis Labenne, and Jérôme Saracco. «ClustGeo: an R package for hierarchical clustering with spatial constraints». In: *Computational Statistics* 33.4 (2018), pp. 1799–1822. DOI: 10.1007/s00180-018-0791-1 (cit. on pp. 16, 33).

[26] Zhenxing Wang, XiaoLiang Wu, and Yadong Wang. «A framework for analyzing DNA methylation data from Illumina Infinium HumanMethylation450 BeadChip». In: *BMC Bioinformatics* 19.5 (2018), p. 115. ISSN: 1471-2105. DOI: 10.1186/s12859-018-2096-3. URL: https://doi.org/10.1186/s12859-018-2096-3 (cit. on p. 20).

[27] Daniel J. Weisenberger, David Van Den Berg, Fei Pan, Benjamin P. Berman, and Peter W. Laird. *Comprehensive DNA Methylation Analysis on the Illumina Infinium Assay Platform*. Illumina Application Note. Accessed: 2026-05-20. 2008. URL: https://www.illumina.com/content/dam/illumina-marketing/documents/products/appnotes/appnote_dna_methylation_analysis_infinium.pdf (cit. on p. 20).

[28] Jing Liu and Kimberly D. Siegmund. «An evaluation of processing methods for HumanMethylation450 BeadChip data». In: *BMC Genomics* 17 (2016), p. 469. DOI: 10.1186/s12864-016-2819-7 (cit. on pp. 20–22).

[29] Sarah Dedeurwaerder, Matthieu Defrance, Martin Bizet, Emilie Calonne, Gianluca Bontempi, and François Fuks. «A comprehensive overview of Infinium HumanMethylation450 data processing». In: *Briefings in Bioinformatics* 15.6 (Nov. 2014), pp. 929–941. DOI: 10.1093/bib/bbt054. URL: https://doi.org/10.1093/bib/bbt054 (cit. on pp. 21, 22).

[30] TJ Jr Triche, DJ Weisenberger, D Van Den Berg, PW Laird, and KD Siegmund. «Low-level processing of Illumina Infinium DNA Methylation BeadArrays». In: *Nucleic Acids Research* 41.7 (Apr. 2013), e90. DOI: 10.1093/nar/gkt090 (cit. on p. 21).

[31] Andrew E. Teschendorff, Francesco Marabita, Matthias Lechner, Thomas Bartlett, Jesper Tegner, David Gomez-Cabrero, and Stephan Beck. «A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450k DNA methylation data». In: *Bioinformatics* 29.2 (2013), pp. 189–196. DOI: 10.1093/bioinformatics/bts680. URL: https://doi.org/10.1093/bioinformatics/bts680 (cit. on p. 21).

[32] H. Welsh, C. M. P. F. Batalha, W. Li, K. L. Mpye, N. C. Souza-Pinto, M. S. Naslavsky, and E. J. Parra. «A systematic evaluation of normalization methods and probe replicability using Infinium EPIC methylation data». In: *Clinical Epigenetics* 15.1 (2023), p. 41. DOI: 10.1186/s13148-023-01459-z. URL: https://doi.org/10.1186/s13148-023-01459-z (cit. on p. 21).

[33] Jean-Philippe Fortin, Audrey Labbe, Mathieu Lemire, Brent W. Zanke, Thomas J. Hudson, Elana J. Fertig, Celia M. T. Greenwood, and Kasper D. Hansen. «Functional normalization of 450k methylation array data improves replication in large cancer studies». In: *Genome Biology* 15.12 (2014), p. 503. DOI: 10.1186/s13059-014-0503-2 (cit. on p. 22).

[34] Wanding Zhou, Jr Triche Timothy J, Peter W Laird, and Hui Shen. «SeSAMe: reducing artifactual detection of DNA methylation by Infinium BeadChips in genomic deletions». In: *Nucleic Acids Research* 46.20 (Nov. 2018), e123–e123. ISSN: 0305-1048. DOI: 10.1093/nar/gky691. eprint: https://academic.oup.com/nar/article-pdf/46/20/e123/26578142/gky691.pdf. URL: https://doi.org/10.1093/nar/gky691 (cit. on pp. 22, 26).

[35] SR Dennis, T Tsukioki, G Cottone, W Zhou, PA Ganz, ME Sehl, Y Luo, SA Khan, and S Clare. «DNA methylation patterns in breast cancer, paired benign tissue from ipsilateral and contralateral breast, and healthy controls». In: *Breast Cancer Research* 27.1 (June 11, 2025), p. 103. DOI: 10.1186/s13058-025-02057-y (cit. on p. 23).

[36] MJ Aryee, AE Jaffe, H Corrada-Bravo, C Ladd-Acosta, AP Feinberg, KD Hansen, and RA Irizarry. «Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays». In: *Bioinformatics* 30.10 (May 15, 2014), pp. 1363–1369. DOI: 10.1093/bioinformatics/btu049 (cit. on p. 23).

[37] Rachel D. Edgar, Meaghan J. Jones, Wendy P. Robinson, and Michael S. Kobor. «An empirically driven data reduction method on the human 450K methylation array to remove tissue specific non-variable CpGs». In: *Clinical Epigenetics* 9.1 (2017), p. 11. ISSN: 1868-7083. DOI: 10.1186/s13148-017-0320-z. URL: https://doi.org/10.1186/s13148-017-0320-z (cit. on pp. 24, 26).

[38] Yuqing Zhang, David F. Jenkins, Solaiappan Manimaran, and W. Evan Johnson. «Alternative empirical Bayes models for adjusting for batch effects in genomic studies». In: *BMC Bioinformatics* 19.1 (July 2018), p. 262. ISSN: 1471-2105. DOI: 10.1186/s12859-018-2263-6. URL: https://doi.org/10.1186/s12859-018-2263-6 (cit. on p. 24).

[39] Yi-An Chen, Melanie Lemire, Sylvia Choufani, Daniel T. Butcher, Daria Grafodatskaya, Brent W. Zanke, Steven Gallinger, Thomas J. Hudson, and Rosanna Weksberg. «Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray». In: *Epigenetics* 8.2 (2013), pp. 203–209. DOI: 10.4161/epi.23470. URL: https://doi.org/10.4161/epi.23470 (cit. on p. 24).

[40] Ruth Pidsley et al. «Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling». In: *Genome Biology* 17.1 (2016), p. 208. DOI: 10.1186/s13059-016-1066-1. URL: https://doi.org/10.1186/s13059-016-1066-1 (cit. on p. 24).

[41] Daniel L. McCartney, Rosie M. Walker, Stewart W. Morris, Andrew M. McIntosh, David J. Porteous, and Kathryn L. Evans. «Identification of polymorphic and off-target probe binding sites on the Illumina Infinium MethylationEPIC BeadChip». In: *Genomics Data* 9 (2016), pp. 22–24. DOI: 10.1016/j.gdata.2016.05.012 (cit. on p. 24).

[42] Hassan Naeem, Nicholas C. Wong, Zoe Chatterton, Mei K. H. Hong, John S. Pedersen, Nicola M. Corcoran, Christopher M. Hovens, and Geoff Macintyre. «Reducing the risk of false discovery enabling identification of biologically significant genome-wide methylation status using the HumanMethylation450 array». In: *BMC Genomics* 15.1 (2014), p. 51. DOI: 10.1186/1471-2164-15-51. URL: https://doi.org/10.1186/1471-2164-15-51 (cit. on p. 24).

[43] Wanding Zhou, Peter W. Laird, and Hui Shen. «Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes». In: *Nucleic Acids Research* 45.4 (2017), e22. DOI: 10.1093/nar/gkw967 (cit. on p. 24).

[44] Brian H Chen and Wanding Zhou. «mLiftOver: harmonizing data across Infinium DNA methylation platforms». In: *Bioinformatics* 40.7 (July 2024), btae423. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btae423. eprint: https://academic.oup.com/bioinformatics/article-pdf/40/7/btae423/58485851/btae423.pdf. URL: https://doi.org/10.1093/bioinformatics/btae423 (cit. on p. 26).

[45] Pan Du, Xiaowei Zhang, C. C. Huang, Nader Jafari, Warren A. Kibbe, Lifang Hou, and Simon M. Lin. «Comparison of Beta-Value and M-Value Methods for Quantifying Methylation Levels by Microarray Analysis». In: *BMC Bioinformatics* 11 (2010), p. 587. DOI: 10.1186/1471-2105-11-587. URL: https://doi.org/10.1186/1471-2105-11-587 (cit. on p. 27).

[46] Shuying Sun, Jael Dammann, Pierce Lai, and Christine Tian. «Thorough Statistical Analyses of Breast Cancer Co-Methylation Patterns». In: *BMC Genomic Data* 23.1 (2022), p. 29. DOI: 10.1186/s12863-022-01046-w (cit. on p. 31).

[47] Peter Langfelder and Steve Horvath. «WGCNA: an R package for weighted correlation network analysis». In: *BMC Bioinformatics* 9 (2008), p. 559. DOI: 10.1186/1471-2105-9-559 (cit. on p. 32).

[48] Tamar Sofer, Elizabeth D. Schifano, Jane A. Hoppin, Lifang Hou, and Andrea A. Baccarelli. «A-clustering: a novel method for the detection of co-regulated methylation regions, and regions associated with exposure». In: *Bioinformatics* 29.22 (Nov. 2013), pp. 2884–2891. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btt498. eprint: https://academic.oup.com/bioinformatics/article-pdf/29/22/2884/48898975/bioinformatics_29_22_2884.pdf. URL: https://doi.org/10.1093/bioinformatics/btt498 (cit. on p. 33).

[49] Liliana Gomez et al. «coMethDMR: accurate identification of co-methylated and differentially methylated regions in epigenome-wide association studies with continuous phenotypes». In: *Nucleic Acids Research* 47.17 (2019), e98. DOI: 10.1093/nar/gkz590 (cit. on p. 33).

[50] Mia Hubert and Stephan Van der Veeken. «Outlier detection for skewed data». In: *Journal of Chemometrics* 22.3-4 (2008), pp. 235–246. DOI: https://doi.org/10.1002/cem.1123. eprint: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/pdf/10.1002/cem.1123. URL: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/cem.1123 (cit. on p. 41).

[51] Tianqi Chen and Carlos Guestrin. «XGBoost: A Scalable Tree Boosting System». In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016). URL: https://api.semanticscholar.org/CorpusID:4650265 (cit. on p. 53).

[52] Ian Newsham, Martin Sendera, Sandeep G. Jammula, and Samaneh A. Samarajiwa. «Early detection and diagnosis of cancer with interpretable machine learning to uncover cancer-specific DNA methylation patterns». In: *Biology Methods and Protocols* 9.1 (2024), bpae028. DOI: 10.1093/biomethods/bpae028 (cit. on p. 54).

[53] Randall J. Ellis, Audrey Airaud, and Chirag J. Patel. *Random feature baselines provide distributional performance and feature selection benchmarks for clinical and 'omic machine learning*. 2024. arXiv: 2411.10574 [q-bio.QM]. URL: https://arxiv.org/abs/2411.10574 (cit. on p. 62).

[54] D Venet, JE Dumont, and V Detours. «Most random gene expression signatures are significantly associated with breast cancer outcome». In: *PLoS Computational Biology* 7.10 (2011), e1002240. DOI: 10.1371/journal.pcbi.1002240 (cit. on p. 63).

[55] Peter J. Thul and Cecilia Lindskog. «The Human Protein Atlas: A Spatial Map of the Human Proteome». In: *Protein Science* 27.1 (2018), pp. 233–244. DOI: 10.1002/pro.3307 (cit. on pp. 65, 69).

[56] Lina Wang et al. «Long non-coding RNA LOC283070 mediates the transition of LNCaP cells into androgen-independent cells possibly via CAMK1D». In: *American journal of translational research* 8 (Dec. 2016), pp. 5219–5234 (cit. on p. 66).

[57] J Kitz, C Lefebvre, J Carlos, LE Lowes, and AL Allan. «Reduced Zeb1 expression in prostate cancer cells leads to an aggressive partial-EMT phenotype associated with altered global methylation patterns». In: *International Journal of Molecular Sciences* 22.23 (2021), p. 12840. DOI: 10.3390/ijms222312840 (cit. on p. 66).

[58] Margot Lavy et al. «ChemR23 activation reprograms macrophages toward a less inflammatory phenotype and dampens carcinoma progression». In: *Frontiers in Immunology* 14 (July 2023). DOI: 10.3389/fimmu.2023.1196731 (cit. on p. 66).

[59] Heather M. Bond, Stefania Scicchitano, Emanuela Chiarella, Nicola Amodio, Valeria Lucchino, Annamaria Aloisio, Ylenia Montalcini, Maria Mesuraca, and Giovanni Morrone. «ZNF423: A New Player in Estrogen Receptor-Positive Breast Cancer». In: *Frontiers in Endocrinology* Volume 9 - 2018 (2018). ISSN: 1664-2392. DOI: 10.3389/fendo.2018.00255. URL: https://www.frontiersin.org/journals/endocrinology/articles/10.3389/fendo.2018.00255 (cit. on p. 67).

[60] CA Sengelaub, K Navrazhina, JB Ross, N Halberg, and SF Tavazoie. «PTPRN2 and PLC$\beta$1 promote metastatic breast cancer cell migration through PI(4,5)P2-dependent actin remodeling». In: *The EMBO Journal* 35.1 (2016), pp. 62–76. DOI: 10.15252/embj.201591973 (cit. on p. 67).

[61] H Tsoi, NN Fung, EPS Man, MH Leung, CP You, WL Chan, SY Chan, and US Khoo. «SRSF5 regulates the expression of BQ323636.1 to modulate tamoxifen resistance in ER-positive breast cancer». In: *Cancers (Basel)* 15.8 (2023), p. 2271. DOI: 10.3390/cancers15082271 (cit. on p. 67).

[62] A Dallol, NF Da Silva, P Viacava, JD Minna, I Bieche, ER Maher, and F Latif. «SLIT2, a human homologue of the Drosophila Slit2 gene, has tumor suppressor activity and is frequently inactivated in lung and breast cancers». In: *Cancer Research* 62.20 (2002), pp. 5874–5880 (cit. on p. 67).

[63] Giulia Peppino, Roberto Ruiu, Maddalena Arigoni, Federica Riccardo, Antonella Iacoviello, Giuseppina Barutello, and Elena Quaglino. «Teneurins: Role in Cancer and Potential Role as Diagnostic Biomarkers and Targets for Therapy». In: *International Journal of Molecular Sciences* 22.5 (2021). ISSN: 1422-0067. DOI: 10.3390/ijms22052321. URL: https://www.mdpi.com/1422-0067/22/5/2321 (cit. on p. 68).