Politecnico di Torino

Politecnico di Torino

Master's Degree in Mathematical Engineering
Academic Year 2025/2026
Graduation Session March 2026

DNA Methylation Dynamics Along the Normal–Adjacent Axis in Breast Cancer

A Quantitative Statistical and Machine Learning Approach

Supervisors:
Prof. Alfredo BENSO
Dott. Sandro GAMBINO

Candidate:
Elisabetta ROVIERA

A mio nonno Franco,
sempre con me.

Abstract

Breast carcinogenesis is not a binary transition from normal to malignant tissue but a progressive deviation from epigenetic homeostasis. Field cancerisation theory posits that histologically normal tissue adjacent to a tumour harbours early epigenetic alterations — detectable as increased methylation variability rather than directional mean shifts — that precede overt neoplastic transformation. Their reproducible characterisation across independent cohorts under strict statistical control remains an open methodological challenge.

This thesis addresses whether DNA methylation alone can discriminate Normal from Normal-Adjacent breast tissue across three independent cohorts (GSE69914, GSE225845, GSE287331), spanning two Illumina array generations (HumanMethylation450K and EPIC), under a fully leakage-controlled framework. A platform-aware preprocessing pipeline addressed probe-design bias and statistical scale distortion. Exploratory analysis confirmed that Normal–Adjacent differences are globally subtle – mean $|\Delta\beta|$ from 0.009 to 0.031 – yet consistently focal, with outlier burden ratios up to 13.3 and silhouette coefficients up to 0.34.

A multi-stage stability-based feature selection framework integrated empirical variability filtering, biologically weighted resampling with directional stability constraints, correlation-based redundancy pruning, and constrained genomic diversification, yielding 5,000-CpG panels per cohort. Inter-dataset concordance analysis revealed that two cohorts share a concordant drift axis supporting bidirectional zero-shot transfer (AUC 0.776–0.794), while the third exhibits systematic polarity inversion with direct consequences for cross-cohort generalisation.

A Mixed-Integer Linear Programme inspired by the 0–1 knapsack formulation was introduced as the central methodological contribution, compressing each panel to 50 loci under joint optimisation of discriminative performance and biological diversity. Hard constraints enforce chromosomal balance, spatial non-redundancy, gene-level diversity, and enrichment in COSMIC breast cancer genes.

For multi-cohort integration, polarity heterogeneity was resolved via absolute deviation transformation from cohort-specific Normal references, followed by NeuroCombat batch harmonisation. The resulting 50-CpG panel achieves AUC = 0.902 and balanced accuracy = 0.799 on the held-out test set ($n = 177$), mapping to 18 COSMIC breast cancer genes — spanning tumour suppressors, chromatin regulators, developmental transcription factors, and signalling components — compared to 11 in the unconstrained baseline. These results demonstrate that biologically constrained combinatorial optimisation produces compact, interpretable CpG panels that preserve near-baseline predictive performance while embedding explicit prior knowledge.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Research Motivation

Breast cancer is the most commonly diagnosed malignancy worldwide and remains a leading cause of cancer-related mortality among women. Despite substantial advances in early detection and targeted therapy, the molecular mechanisms governing the earliest phases of tumour initiation remain incompletely understood. Carcinogenesis is not a binary transition from normal to malignant, but a progressive deviation from epigenetic homeostasis. The epigenetic hallmarks of frank tumour tissue, including genome-wide hypomethylation and focal hypermethylation of tumour-suppressor loci, have been extensively characterised [1, 2]. The methylation signatures of invasive breast carcinoma are therefore well established. While the existence of epigenetic field defects in Normal-Adjacent breast tissue has been established [3], their reproducible and transferable characterisation across independent cohorts — under strict statistical control — remains an open methodological challenge. This observation forms the basis of the *field cancerisation* hypothesis, originally proposed in oral cancer and subsequently extended to breast tissue through genome-wide methylation profiling [3]. The field effect posits that carcinogenesis does not arise from a single isolated cell but emerges within a broader epigenetically conditioned tissue environment — a molecularly altered field that remains invisible to the pathologist yet detectable at the epigenomic level. DNA methylation is particularly well suited for investigating such early alterations. Methylation marks are mitotically heritable, chemically stable, and measurable at single-locus resolution across hundreds of thousands of CpG sites using Illumina array platforms. Prior work has demonstrated that alterations accumulating in Normal-Adjacent breast tissue preferentially target Polycomb-regulated and bivalent chromatin loci, precisely the genomic regions implicated in cancer-associated silencing [3, 4]. Critically, these alterations manifest predominantly as increased

methylation *variability* rather than as large directional mean shifts, rendering classical differential methylation analyses insufficient for their detection [5]. This thesis addresses whether DNA methylation alone can reproducibly discriminate Normal from Normal-Adjacent breast tissue across independent cohorts under strict statistical control.

1.2 Unresolved Limitations and Open Problem

Despite the biological importance of the Normal–Adjacent state, four unresolved limitations motivate the present work. Standard approaches identify discriminative CpGs by comparing group means via linear models or t-tests; however, field-effect alterations are expected to manifest as dispersion increases rather than uniform shifts. Teschendorff and colleagues formalised the distinction between differentially methylated (DM) and differentially variable (DV) loci, showing that DV CpGs in Normal-Adjacent tissue are enriched for biologically meaningful genomic features [3, 5], so mean-based selection risks missing a substantial component of the pre-malignant signal. At the same time, field-cancerisation alterations are not homogeneous across individuals or loci: as documented in Chapters 5 and 6, the direction of epigenetic drift along the Normal–Adjacent axis is not consistent across independent cohorts, and such polarity heterogeneity has direct consequences for predictive transfer that have not been systematically characterised in prior work. A further structural challenge stems from dimensionality: Illumina 450K and EPIC arrays profile up to 8.6×10^5 CpG loci while breast tissue cohorts typically include fewer than $n = 300$ samples, and this HDLSS regime ($p \gg n$) induces instability in feature selection [6]; small perturbations in the training set may yield markedly different CpG rankings, and information leakage during preprocessing can further inflate performance estimates [7]. Finally, cross-cohort evaluations of Normal–Adjacent classifiers implicitly assume consistent drift orientation across studies, an assumption that is rarely tested explicitly. The present work shows it does not universally hold: one of the three cohorts exhibits a systematically inverted drift axis, producing directed misclassification in transfer settings, and ignoring this heterogeneity leads to misleading conclusions about signature transferability. These four limitations define the central open problem addressed in this thesis.

> *Is it possible to construct a robust, transferable CpG-based signature that discriminates Normal from Normal-Adjacent breast tissue, while controlling for high-dimensional instability and enforcing strict cross-cohort reproducibility?*

1.3 Central Hypothesis

The central hypothesis of this thesis is that Normal-Adjacent breast tissue occupies a quantitatively characterisable intermediate position along an epigenetic trajectory connecting histologically normal tissue to invasive tumour. This Normal–Adjacent–Tumour axis is not merely conceptual but measurable from high-dimensional methylation data under appropriate statistical constraints. This hypothesis yields four testable implications. First, intra-dataset discrimination of Normal from Normal-Adjacent tissue should be statistically feasible using a variance-sensitive, leakage-controlled CpG panel, yielding AUC substantially above chance. Second, cohorts exhibiting concordant drift orientation should support bidirectional cross-dataset transfer with non-trivial predictive performance. Third, polarity-inverted cohorts should produce systematic misclassification in transfer settings, leading to directed sub-random AUC values rather than random degradation. Fourth, a compact CpG panel derived within a single cohort should retain measurable discriminative power when transferred to an external cohort with aligned drift orientation. Failure of these predictions would weaken or refute the proposed model.

1.4 Objectives and Thesis Structure

The thesis pursues five methodological objectives, addressed across the following chapters.

Chapter 2 establishes the biological background, covering age-dependent methylation drift, field cancerisation, and the Normal–Adjacent–Tumour axis.

Chapter 3 introduces the three cohorts (GSE69914, GSE225845, GSE287331) and characterises their intra- and inter-dataset structure.

Chapter 4 constructs a leakage-controlled preprocessing pipeline enforcing strict separation between training and test data at all stages.

Chapter 5 develops a stability-aware feature selection framework integrating variance sensitivity, directional stability, correlation pruning, and biological priors, and reports inter-dataset concordance analyses.

Chapter 6 evaluates predictive performance under intra- and cross-dataset designs, derives a compact 50-CpG panel, and characterises the consequences of drift polarity heterogeneity for signature transferability.

Chapter 7 integrates findings across cohorts and discusses implications for epigenetic biomarker discovery.

Finally, Chapter 8 summarises the main contributions of the thesis, reflects on the methodological limitations encountered, and outlines directions for future work.

Chapter 2

Biological and Epigenetic Background

2.1 Breast Cancer: Epidemiology and Molecular Subtypes

Breast cancer is the most frequently diagnosed malignancy in women worldwide and represents a major public health burden across both high- and low-income countries [8]. Its clinical and biological heterogeneity is one of the most distinctive features of the disease: not all breast cancers behave alike, progress at the same rate, or respond to the same treatments. At the histological level, the vast majority of breast malignancies originate from the epithelial compartment of the mammary gland, most commonly from the luminal cells lining the ductal and lobular structures. The classical progression model describes a continuum from normal epithelium through non-atypical hyperplasia, atypical ductal hyperplasia (ADH), ductal carcinoma in situ (DCIS), and ultimately invasive carcinoma [9]:

$$\text{Normal epithelium} \rightarrow \text{Hyperplasia} \rightarrow \text{DCIS} \rightarrow \text{Invasive carcinoma.}$$

At the molecular level, tumors are stratified according to the expression status of estrogen receptor (ER), progesterone receptor (PR), and HER2, defining biologically distinct subtypes — luminal A, luminal B, HER2-enriched, and triple-negative — with different prognostic and therapeutic implications. This molecular classification reflects underlying differences in the cell of origin, epigenetic landscape, and selective pressures operative during tumor evolution. A subset of breast cancers arises on an inherited background of germline susceptibility. Pathogenic variants in *BRCA1* and *BRCA2*, genes encoding key components of the homologous recombination DNA repair machinery, substantially increase lifetime risk [10]. BRCA-associated

tumors tend to be triple-negative and exhibit distinct baseline epigenetic profiles that differ fundamentally from sporadic disease [11, 12]. For this reason, BRCA mutation carriers require separate analytical treatment in any study of epigenetic field effects, as discussed in Chapter 3.

2.2 The Hallmarks Framework and Epigenetic Reprogramming

The conceptual framework of cancer biology has been progressively organized around the notion of "hallmarks": a set of functional capabilities that normal cells must acquire to become malignant [8, 9]. These include sustained proliferative signaling, resistance to growth suppression, evasion of apoptosis, unlimited replicative potential, induction of angiogenesis, and the activation of invasion and metastasis. Importantly, the updated hallmarks framework explicitly recognizes epigenetic reprogramming as a core enabling mechanism [8]. Tumor cells do not merely accumulate genetic mutations; they undergo widespread reorganization of their chromatin architecture and DNA methylation landscape, which silences tumor-suppressive programs and derepresses oncogenic pathways. This repositions epigenetic alterations not as secondary consequences of malignant transformation but as primary and early drivers of the tumorigenic process.

2.3 DNA Methylation: Mechanism and Genomic Architecture

DNA methylation consists of the covalent addition of a methyl group to the 5-carbon position of cytosine residues within CpG dinucleotides, a reaction catalyzed by a family of DNA methyltransferases (DNMT3A, DNMT3B for *de novo* methylation; DNMT1 for maintenance after replication) [2]. CpG dinucleotides are non-randomly distributed in the genome: statistically expected at roughly one occurrence per 16 bp, they are in practice four- to fivefold depleted genome-wide — a depletion attributed to the historical mutability of methylated cytosines, which deaminate spontaneously to thymine. The exception to this depletion is found at *CpG islands*: dense, GC-rich clusters of CpG dinucleotides, typically 500–2000 bp in length, located at the promoters of approximately 60% of human protein-coding genes [2]. In normal somatic cells, CpG islands are kept constitutively unmethylated by active mechanisms, preserving transcriptional access.

2.3.1 CpG Islands, Shores, Shelves, and Open Sea

The regulatory significance of DNA methylation varies substantially depending on where in the genome a CpG locus resides. A widely adopted classification, used in Illumina array annotation, distinguishes four positional contexts relative to CpG islands [13].

- **CpG Islands**: regions with high CpG density and high GC content, operationally defined as stretches of ≥ 200 bp with CpG observed/expected ratio > 0.6 and GC content $> 50\%$. Methylation at island-overlapping promoters is the canonical mechanism of transcriptional silencing in both development and cancer.

- **Shores**: regions within 2 kb flanking a CpG island on either side. Although less CpG-dense than islands, shores are the sites of the most dynamic and biologically informative methylation variation. Landmark work by Irizarry et al. demonstrated that tissue-specific methylation differences between normal cell types are concentrated at shores rather than at islands, and that cancer-associated hypermethylation also preferentially targets shores [13]. This makes shores the primary locus of epigenetic switching between cell states.

- **Shelves**: regions 2–4 kb from the island boundary. Their methylation patterns are less variable than shores and less directly tied to promoter regulation, though they retain some contextual regulatory relevance.

- **Open Sea**: CpG loci not associated with any island, shore, or shelf. These constitute the majority of CpGs in the genome ($\sim$70%) and tend to be constitutively methylated in somatic tissues. Their methylation changes are more often associated with global hypomethylation and genomic instability than with specific gene regulation.

This spatial hierarchy has direct consequences for how methylation data should be interpreted and weighted in analytical frameworks. Loci overlapping islands and shores are enriched for functional regulatory activity — they are more likely to be associated with gene expression changes when their methylation state shifts. Shelf and open-sea loci, by contrast, tend to reflect structural genomic states rather than specific transcriptional events. When constructing feature selection or scoring systems based on biological relevance, this gradient from island to open sea provides a principled basis for differential weighting.

2.3.2 Genomic Position Determines the Functional Consequence of Methylation

A further distinction of biological relevance concerns the relationship between methylation and transcription depending on the position of the CpG locus relative to the gene structure. The classical rule — methylation silences expression — applies specifically to *promoter-associated* CpG islands. Within the gene body (introns and exons), the relationship is inverted: gene body methylation correlates *positively* with transcriptional activity [14]. This is thought to reflect the need to suppress spurious transcription initiation within highly transcribed genes. The functional interpretation of a methylation change therefore depends critically on whether the affected locus falls in a promoter or an intragenic context. In genome-wide methylation profiling studies, the Illumina array annotation provides positional information (TSS200, TSS1500, 5'UTR, 1stExon, Body, 3'UTR) that, combined with the island/shore/shelf classification, allows a nuanced characterization of each locus. CpGs annotated to TSS regions within islands or shores carry the highest prior probability of regulatory relevance and are therefore the most informative for studies of transcriptional silencing in cancer.

2.4 Epigenetic Alterations in Cancer: Too Much and Too Little

Cancer genomes are characterized by a paradoxical co-occurrence of global hypomethylation and focal hypermethylation [1], a duality confirmed across virtually all tumor types, including breast cancer. *Global hypomethylation* affects repetitive sequences (satellite DNA, LINE-1 elements) and promotes genomic instability through aberrant activation of transposable elements and loss of imprinting, resulting in transcriptional derepression at normally silenced genes [1]. *Focal hypermethylation*, by contrast, targets CpG islands at the promoters of specific genes, leading to heritable transcriptional silencing. Affected genes frequently include classical tumor suppressors (e.g., *CDKN2A*, *BRCA1*, *MLH1*), as well as genes involved in apoptosis, cell adhesion, and developmental signaling. This epigenetic silencing is functionally equivalent to mutation but potentially reversible, making it an attractive therapeutic target. Critically, neither phenomenon is random. Genes susceptible to hypermethylation in cancer are systematically enriched for targets of the Polycomb Repressive Complex 2 (PRC2) and display bivalent chromatin states in embryonic stem cells — a configuration marking developmental genes as both transcriptionally poised and repressed [4], suggesting that epigenetic alterations in malignancy partially recapitulate stem cell programs.

2.5 Epigenetic Drift, Aging, and Cancer Risk

Aging is the single strongest risk factor for breast cancer. The incidence of most breast malignancies rises sharply with age, and this epidemiological observation has a direct molecular correlate in the epigenome. Over the course of a lifetime, DNA methylation patterns undergo progressive and partially stochastic changes collectively referred to as *epigenetic drift* [15]. This phenomenon is characterized by gradual increases in methylation variability across both CpG loci and individuals. The drift is not uniform: it preferentially affects genes associated with Polycomb repression and bivalent chromatin — precisely the same loci that tend to become hypermethylated in cancer [4]. Horvath's landmark study demonstrated that DNA methylation patterns can be exploited to construct a highly accurate molecular clock: the "epigenetic age" of a tissue, estimated from a weighted combination of CpG methylation values, correlates strongly with chronological age across more than 50 tissue types [15]. Crucially, epigenetic age acceleration — a discordance between molecular and chronological age — has been associated with cancer risk, suggesting that the pace of epigenetic drift itself may be a biological marker of malignant susceptibility. From a mathematical perspective, aging-associated drift can be modeled as a diffusion-like process in epigenetic state space: methylation levels at susceptible loci undergo a gradual increase in inter-individual variance over time. This growing dispersion may create a permissive landscape for malignant transformation by producing a fraction of cells with aberrant epigenetic configurations from which selection can act. An important study by Teschendorff et al. in breast tissue documented precisely this mechanism: aging-related DNA methylation changes in normal breast tissue were found to target the same genomic regions altered in cancer, establishing a quantitative link between physiological aging and pre-malignant epigenetic reprogramming [16].

2.6 Field Cancerisation: The Concept and Its Epigenetic Basis

A central concept motivating the present thesis is that of *field cancerisation*. Originally proposed by Slaughter in 1953 from histopathological observations in oral cancer, the field effect hypothesis posits that carcinogenesis does not occur in isolation but within a "conditioned" tissue field broadly reprogrammed by molecular alterations preceding frank malignancy. In breast cancer, epigenetic field defects have been systematically characterized using genome-wide methylation profiling. Teschendorff et al. demonstrated that histologically normal breast tissue adjacent to tumors harbors a specific pattern of methylation alterations distinct from both distant normal tissue and tumor tissue [3], enriched at PRC2 target

genes, CTCF-binding regions, and loci involved in WNT and FGF developmental pathways — precisely the categories implicated in the aging-cancer methylation axis described above. Crucially, this field defect signature was detectable not through classical differential mean analysis but through *differential variability*, a framework quantifying the increase in methylation dispersion across samples rather than directional shifts in average levels [3, 5]. Under a classical framework, loci with increased variance but small mean differences are invisible to standard hypothesis tests and require variance-sensitive statistics such as the Levene test or its robust variants. The biological interpretation is that field defects reflect a partial and heterogeneous epigenetic reprogramming, where the stochastic nature of the alteration manifests as increased dispersion across individuals rather than a uniform shift. The progressive nature of this process is further supported by the observation that hypervariable loci in adjacent tissue tend to become hypermethylated in the matched invasive tumor, suggesting a deterministic convergence from a heterogeneous pre-malignant landscape to a consolidated malignant state [3]:

$$\text{Normal} \rightarrow \text{Epigenetically reprogrammed adjacent tissue} \rightarrow \text{Invasive carcinoma.}$$

This model implies that adjacent tissue represents an earlier stage on the same carcinogenic trajectory rather than mere contamination from the nearby tumor, making the detection and characterization of its methylation signature both biologically and clinically significant.

2.7 Differential Variability in Cancer

The discovery that field defects are better captured by differential variability than by differential means represents a genuine shift in the analytical paradigm for epigenetic cancer studies [5]. Formally, let X_j denote the methylation level at locus j across a cohort of samples. Classical differential methylation analysis focuses on the comparison of group means:

$$\Delta_j = \mathbb{E}[X_j \mid A] - \mathbb{E}[X_j \mid B]. \tag{2.1}$$

Differential variability analysis instead targets the comparison of group variances:

$$\delta\sigma_j^2 = \mathrm{Var}(X_j \mid A) - \mathrm{Var}(X_j \mid B). \tag{2.2}$$

A locus is said to be *differentially variable* (DV) if $\delta\sigma_j^2$ is significantly non-zero. Under a pre-malignant model, one expects $\delta\sigma_j^2 > 0$ in normal-adjacent tissue relative to distant normal tissue, reflecting increased epigenetic disorder in the field-affected region. Teschendorff et al. formalized this framework and developed statistical tools for its application to genome-wide methylation data [5]. Their

analysis demonstrated that DV CpGs in normal-adjacent breast tissue are substantially enriched for biologically meaningful genomic features (PRC2 targets, bivalent domains) and display stronger associations with cancer risk markers — including tumor proliferation as measured by KI67 — than differentially methylated CpGs identified by mean-based tests alone [3]. The implication for any analytical framework aimed at characterizing early epigenetic alterations is direct: a reliance on mean-based feature selection alone will systematically miss a biologically relevant component of the pre-malignant signal. This motivates the integration of variability-aware preprocessing and feature selection strategies, which are developed in the methodological chapters of this thesis.

2.8 The Normal–Adjacent–Tumor Axis as an Analytical Framework

Breast carcinogenesis is not a sudden event but a progressive deviation from epigenetic homeostasis, driven by the interplay of aging, stochastic drift, and selective pressures across a population of cells. The biological evidence reviewed in this chapter converges on a model in which this progression is both measurable and partially predictable from methylation data alone. Normal breast tissue accumulates methylation variability with age at loci associated with Polycomb regulation and bivalent chromatin, creating a heterogeneous epigenetic landscape from which pre-malignant field defects emerge. Adjacent tissue, histologically indistinguishable from normal, harbors a measurable epigenetic signature of this reprogramming — detectable predominantly as increased variability rather than a directional mean shift. As carcinogenesis proceeds, this heterogeneous pre-malignant state converges toward a consolidated malignant configuration characterized by focal hypermethylation at tumor-suppressive loci and global hypomethylation elsewhere. This progressive model has three direct consequences for the quantitative analyses developed in this thesis. First, the Normal–Adjacent–Tumor axis should be treated as a continuous biological spectrum: adjacent tissue occupies a genuinely intermediate epigenetic state, distinct from both normal and tumor, and warrants explicit characterization rather than collapsing into either extreme. Second, feature selection must be sensitive to variance differences and not only to mean shifts, as field defects are heterogeneous signals that may not survive standard mean-based filtering. Third, cross-cohort reproducibility is a necessary validation criterion: signatures replicating across independent datasets are more likely to reflect genuine biological structure than cohort-specific noise. These three principles — spectrum-aware design, variance sensitivity, and cross-cohort validation — directly motivate the construction, preprocessing, and comparative characterization of the three cohorts introduced in Chapter 3.

Chapter 3

Data Preparation and Exploratory Analysis

3.1 Methodological Objectives

This chapter establishes the methodological foundation for all subsequent analyses by addressing the construction, organization, and preliminary characterization of the DNA methylation datasets used in this thesis. Following the biological background and problem formulation introduced in Chapters 1 and 2, the present chapter acts as a bridge between domain-specific knowledge and quantitative modeling, ensuring that downstream analyses are grounded on a technically sound and well-understood data representation. Genome-wide DNA methylation data are characterized by extremely high dimensionality, heterogeneous signal-to-noise ratios, and the coexistence of subtle biological effects with multiple sources of technical variability. In this context, premature application of statistical models or feature-selection procedures may lead to biased conclusions driven by artefacts rather than genuine biological structure. A careful examination of data quality, distributional properties, and internal consistency is therefore a necessary prerequisite for any reliable inference [17]. The primary objective of this chapter is to construct a coherent and reproducible representation of the available datasets and to assess their structural and statistical properties prior to any form of preprocessing or modeling. Specifically, this chapter aims to: (i) describe the organization of methylation matrices and associated phenotype information; (ii) characterize the global and local variability of methylation levels within each dataset; (iii) evaluate the degree of separation between Normal, Adjacent, and Tumor tissues using exploratory tools; and (iv) compare datasets at an inter-cohort level to assess consistency, heterogeneity, and potential limitations in signal transferability. The scope of this chapter is deliberately restricted to exploratory and diagnostic analyses. No feature

selection, predictive modeling, or formal statistical inference is performed at this stage. All analyses are intended to be descriptive and comparative, serving to highlight structural properties of the data and to identify potential technical or biological issues that must be addressed before proceeding further. Methodologically, the chapter is organized around a unified exploratory framework applied consistently across all datasets. Intra-dataset analyses are used to investigate internal structure, variability patterns, and group-level behavior, while inter-dataset comparisons focus on assessing cross-cohort consistency and platform-related differences. Throughout the chapter, visualizations and summary metrics are employed as diagnostic tools rather than as decision-making instruments. The results of this exploratory phase reveal both shared characteristics and dataset-specific behaviors, as well as clear limitations of working directly with raw or minimally processed methylation data. These observations motivate the need for a structured and rigorous data preprocessing pipeline, which is developed and justified in the following chapter. The present chapter therefore provides the conceptual and technical basis for the preprocessing choices adopted in Chapter 4 and for all subsequent modeling steps.

3.2 Dataset Construction, Metadata Integration, and Data Organization

This section provides the methodological framework for the selection, construction of the dataset, and the organisation process adopted in this study. It contextualises the data sources, representation choices, metadata management, and computational design principles that underpin the detailed implementation described in the following sections.

3.2.1 Data Sources and Study Design

All datasets analyzed in this thesis were obtained from the NCBI Gene Expression Omnibus (GEO) [18], which was selected as a unified data source to minimize heterogeneity in data formats, metadata conventions, and access mechanisms. GEO provides curated series-level accessions, stable sample identifiers, and standardized links to processed methylation matrices and sample-level annotations, making it a natural reference repository for large-scale DNA methylation studies.

The study design adopts a multi-cohort comparative framework focused on genome-wide DNA methylation in breast tissue. Three independent GEO series were selected for analysis: **GSE69914** [19], based on the Illumina HumanMethylation450 platform, and two more recent EPIC-based cohorts, **GSE225845** [20] and **GSE287331** [21]. Together, these datasets span normal/healthy, adjacent-normal, and tumor tissue states across two generations of Illumina methylation arrays

(450K and EPIC), enabling both within-dataset characterization and cross-dataset comparisons. Dataset selection was driven by strict inclusion criteria aimed at ensuring technical compatibility, biological relevance, and analytical feasibility. In particular, only studies providing processed genome-wide methylation matrices in terms of β-values were considered, allowing a uniform data representation and avoiding the need for low-level signal reprocessing from raw `.idat` intensity files. Raw `.idat` files store probe-level fluorescence intensities measured in the methylated ($y^{(M)}$) and unmethylated ($y^{(U)}$) channels, from which β-values are derived as a normalized ratio [22]:

$$\beta := \frac{\max(y^{(M)},0)}{\max(y^{(M)},0) + \max(y^{(U)},0) + \alpha}, \qquad \beta \in [0,1]. \tag{3.1}$$

where α is a small offset (typically 100) used to stabilize the denominator. Additional requirements included the use of Illumina HumanMethylation450 (450K) or MethylationEPIC (EPIC) platforms, the availability of multiple tissue classes within each cohort, and sufficiently large sample sizes to support robust statistical analysis.

3.2.2 Methylation Data Representation and Storage

In several datasets, the original authors applied established normalization procedures, including methods specifically designed to correct probe-type design bias such as BMIQ [23]. Reprocessing raw IDAT files would therefore not add information relevant to the objectives of this thesis, while substantially increasing technical complexity and reducing reproducibility. For these reasons, all analyses were conducted on the processed β-value matrices as distributed via GEO. To support scalable analysis and efficient input/output operations, all working methylation matrices were converted to a standardized `sample` $\times$ `CpG` layout and stored in columnar Parquet format using single-precision floating-point encoding. This choice is motivated by both numerical and computational considerations. Since β-values are strictly bounded in the interval [0,1] [22], and biologically meaningful methylation differences typically occur at magnitudes on the order of 10^{-2}–10^{-3}, single-precision floating point (`float32`, machine $\varepsilon \approx 10^{-7}$) provides more than sufficient numerical accuracy. At the same time, it substantially reduces memory usage and I/O overhead compared to double precision. This representation is consistent with recent large-scale genomics frameworks that process molecular features, including DNA methylation data, entirely in `float32` precision for efficiency and scalability [24].

3.2.3 Phenotype Tables and Label Harmonization

For each dataset, a dedicated phenotype table was constructed by extracting and parsing GEO sample-level metadata. These tables encode tissue labels, sample identifiers, and relevant clinical or technical covariates when available. Given the heterogeneity of original annotations across studies, tissue classes were mapped to harmonized numeric encodings to ensure consistency across cohorts while preserving dataset-specific attributes in separate fields. A strict one-to-one correspondence between methylation matrices and phenotype tables was enforced. Samples lacking either methylation measurements or corresponding metadata were excluded to prevent inconsistencies and downstream misalignment. This design guarantees that all analyses operate on synchronized molecular and phenotypic representations and enables reliable stratification by tissue class in both intra- and inter-dataset comparisons.

3.2.4 Reproducibility and Computational Workflow

The overall data organization and computational design follow a multi-dataset analytical paradigm commonly adopted in recent large-scale DNA methylation studies, in which cohort-specific analyses are complemented by cross-dataset comparisons to explicitly evaluate robustness and generalizability, while simultaneously revealing cohort-dependent effects and limitations in signal transferability that motivate subsequent methodological refinement [25]. All dataset construction and organization steps were implemented using deterministic, script-based pipelines with explicit schema definitions and typed data representations. Particular care was taken to avoid unnecessary in-memory operations on high-dimensional methylation matrices, favoring streaming ingestion, chunked processing, and out-of-core transformations when required by dataset size. The resulting data organization yields a set of standardized, self-contained datasets, each consisting of a genome-wide methylation matrix and an aligned phenotype table stored in efficient, portable formats. This design ensures full reproducibility of the data preparation process and provides a robust foundation for the preprocessing, exploratory analysis, and modeling steps developed in subsequent chapters.

3.3 Intra-Dataset Exploration and Visualization

This section presents a structured intra-dataset exploratory analysis aimed at characterising the internal statistical and biological organisation of each methylation cohort prior to any preprocessing or modelling step. Each dataset is analysed independently in order to assess data integrity, quantify variability patterns, and evaluate whether Normal, Adjacent, and Tumor tissues exhibit distinguishable

Table 3.1: Overview of the GSE69914 dataset after initial cohort curation.

CpGs	Samples	Normal	Adjacent	Tumor	Normal (BRCA1)	Tumor (BRCA1)
485,512	407	50	42	305	3	7

epigenetic behaviour. In the context of breast cancer, epigenetic deregulation is known to manifest through both global and local phenomena, including genome-wide hypomethylation, focal hypermethylation of tumour-suppressor regions, and the presence of pre-neoplastic "field defects" in histologically normal tissues proximal to tumours [1, 2, 3]. These mechanisms motivate an exploratory investigation focused not only on overt tumour-associated changes, but also on more subtle alterations in Normal and Adjacent samples, which are central to the biological hypothesis of this thesis.

This section therefore applies a unified exploratory framework to each dataset – GSE69914 (Section 3.3.1), GSE225845 (Section 3.3.2), GSE287331 (Section 3.3.3)–, systematically examining data quality, global methylation distributions, sample-level summaries, CpG-wise variability and instability, correlation structure, dimensionality reduction, and genomic context coverage. The use of a consistent analytical protocol enables direct comparison of intra-dataset behaviour while preserving cohort-specific characteristics.

3.3.1 Dataset GSE69914

Dataset overview GSE69914 is a breast tissue DNA methylation cohort profiled using the Illumina Infinium HumanMethylation450 BeadChip, providing genome-wide coverage of approximately 4.8×10^5 CpG loci. The dataset comprises samples spanning the full Normal–Adjacent–Tumor spectrum that motivates the biological framework of this thesis. An overview of the sample composition and CpG coverage is reported in Table 3.1, including the distribution across tissue classes and hereditary-risk subgroups. A limited subset of samples is associated with germline *BRCA1* mutations. *BRCA1*, in particular, encodes a key DNA repair protein, and pathogenic variants are known to substantially increase lifetime breast and ovarian cancer risk [10]. Multiple studies have shown that BRCA-associated breast tissues exhibit distinct baseline epigenetic profiles, reflecting inherited genomic instability rather than sporadic tumorigenesis [11, 12]. Since the primary objective of this thesis is to characterise early, field defect DNA methylation alterations along the Normal–Adjacent axis in sporadic breast cancer, samples carrying *BRCA1* mutations were excluded from the core exploratory and preprocessing pipeline. Their inclusion would introduce a confounding hereditary epigenetic signal, potentially obscuring subtle methylation shifts in non-mutated normal and adjacent tissues.

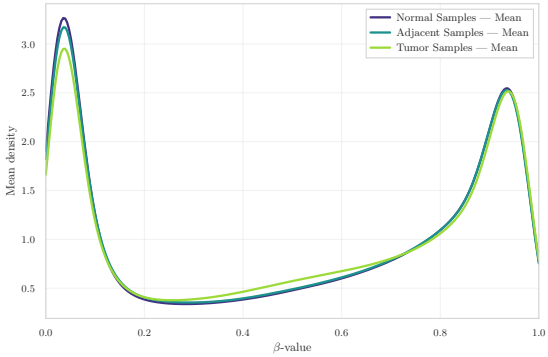
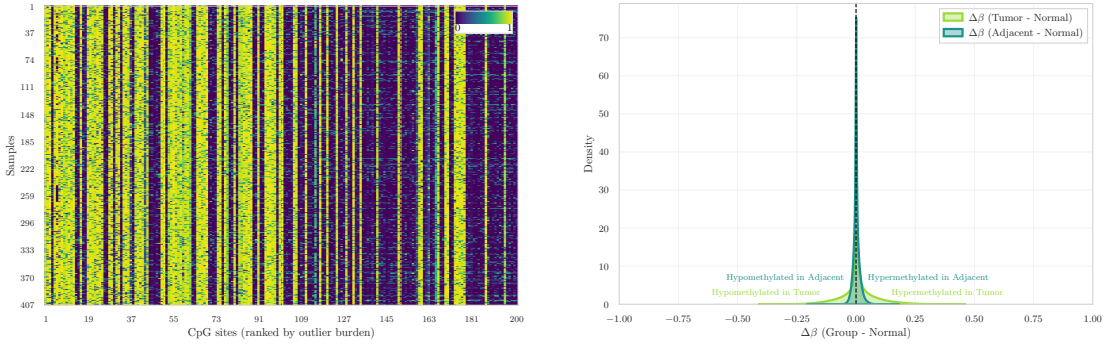


Figure 3.1: Group-wise mean β-value density in GSE69914.

Global methylation distributions To characterise the overall methylation landscape of the dataset, the distribution of β-values (fractional methylation in [0,1]) was first examined. The group-wise mean density curve (Figure 3.1) displays the characteristic *bimodal* profile of Illumina 450k arrays, with peaks near unmethylated ($\beta \approx 0$) and fully methylated ($\beta \approx 1$) CpG sites. This pattern reflects the underlying biology of CpG regulation, where many loci tend to be either transcriptionally active (hypomethylated) or repressed (hypermethylated) [2, 26]. When comparing tissue groups, a clear gradient emerges. Tumor samples show a slightly flatter high-β peak and a broader low-β tail, consistent with the well-described phenomenon of global hypomethylation and increased heterogeneity in cancer [1]. Adjacent samples lie between Normal and Tumor, suggesting early epigenetic drift and subtle field defects occurring in histologically non-neoplastic tissue [3].

CpG-level instability and recurrent outliers To characterise locus-specific instability, the 200 CpG sites with the highest outlier burden across samples were examined. The heatmap of the top outlier loci (Figure 3.2a) indicates that epigenetic disruption is *not uniform*: specific CpG sites and specific samples display extreme deviations, rather than a diffuse genome-wide shift. Moreover, both directions of alteration are present — focal hypermethylation and focal hypomethylation. In cancer biology, hypermethylation can silence tumor-suppressor regions, whereas hypomethylation can derepress oncogenic pathways and weaken genomic stability, reflecting the classic "too much and too little methylation" behaviour of tumor genomes [1]. Complementary $\Delta\beta$ distributions comparing Tumor and Adjacent tissues against Normal (Figure 3.2b) show a clear shift toward hypomethylation in Tumor samples and a subtler but detectable drift in Adjacent tissues. This provides evidence that epigenetic alterations emerge early in histologically normal tissue located near the tumor, supporting the field-cancerisation model [3].



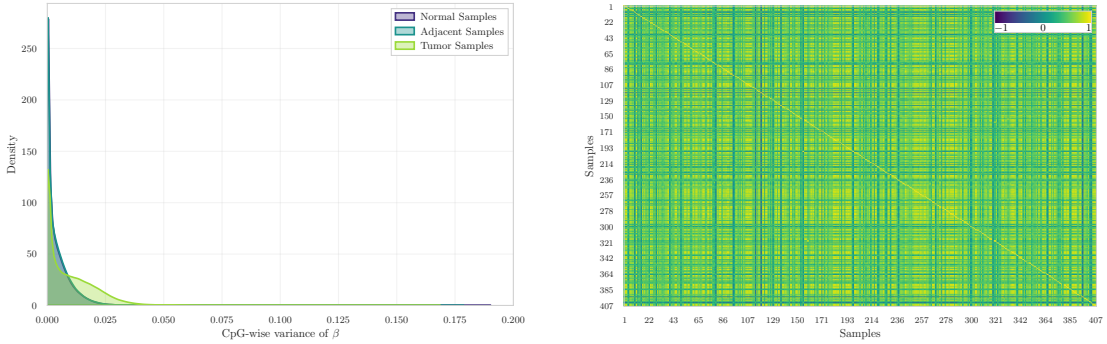
(a) Top outlier CpG loci.

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

Figure 3.2: Heatmap of the top outlier CpGs (left) and corresponding $\Delta\beta$ distributions (right), illustrating CpG-level instability in GSE69914.

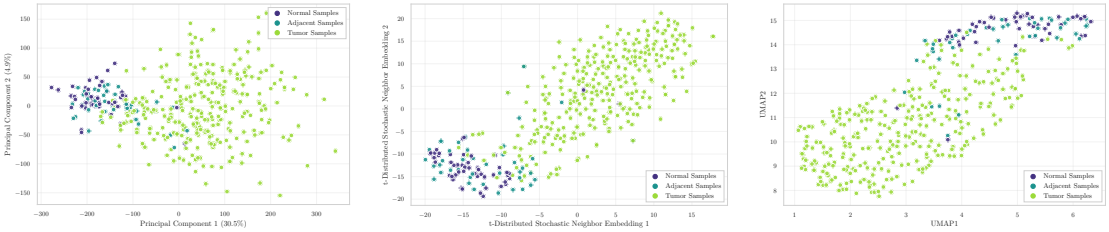
Variance and correlation structure To quantify within-group epigenetic variability, the distribution of CpG-wise variance across samples was examined. The density curves (Figure 3.3a) show that Tumor samples have markedly higher variance, reflecting increased epigenetic instability. In contrast, Normal and Adjacent samples display almost identical variance distributions that are tightly concentrated near zero. This indicates that, at the level of CpG-wise variability, Adjacent tissues do not yet exhibit detectable divergence from Normal samples. A complementary perspective is provided by the sample correlation heatmap (Figure 3.3b), which shows the expected high level of pairwise similarity across all samples. This behaviour is typical of genome-wide DNA methylation data, where a large fraction of CpG sites is stable across individuals, resulting in consistently strong correlations [23]. Importantly, the heatmap also reveals that all samples are highly mutually correlated, with no sharp blocks or boundaries separating the three tissue labels. This confirms that global methylation structure is largely conserved across Normal, Adjacent and Tumor tissues, and that group differences are too subtle to be detected through sample–sample correlation alone.

Low-dimensional embeddings Dimensionality reduction was applied using Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP) to visualise global similarities among samples (Figure 3.4). Across all three methods, Normal and Adjacent samples show substantial overlap, with no clear separation between these two groups. Tumor samples appear more dispersed, but only mildly shifted relative to the Normal/Adjacent cluster, and no sharp cluster boundaries



(a) CpG-wise variance distributions. **(b)** Sample correlation heatmap.

Figure 3.3: Variance and correlation structure across Normal, Adjacent, and Tumor samples in GSE69914.

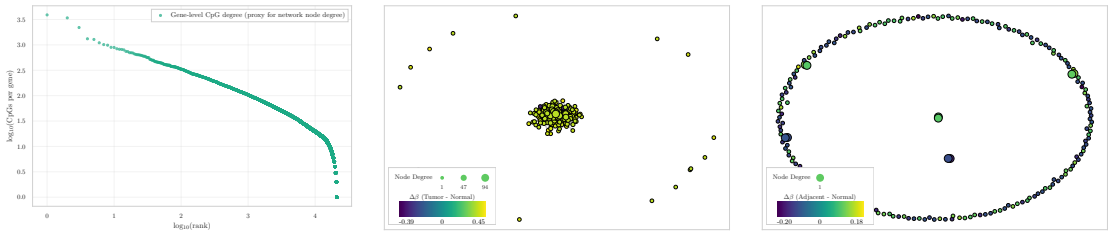


(a) PCA trustworthiness = 0.869. **(b)** t-SNE trustworthiness = 0.888. **(c)** UMAP trustworthiness = 0.874.

Figure 3.4: Low-dimensional embeddings samples in GSE69914.

emerge in any embedding. The trustworthiness scores are similar across embeddings (0.869–0.888), indicating that local neighbourhood relationships are largely preserved in the low-dimensional projections. These projections indicate that global methylation patterns alone do not provide strong discriminative structure among the three tissue states. This suggests that group differences are subtle at the whole-methylome level and are more effectively captured through locus-specific analyses rather than through global unsupervised embeddings.

Genome annotation coverage CpG probes were annotated using the official *Infinium MethylationEPIC v1.0 B5* [27] manifest file, the standard Illumina resource containing genomic coordinates, probe type, CpG island context (Island, Shore, Shelf, Open Sea), and gene-level annotations. Using this manifest, the dataset exhibits the expected non-uniform genomic distribution of CpG contexts: the largest fraction of probes maps to open-sea or non-island regions, followed by CpG islands, then north and south shores, and finally north and south shelves. This



(a) Gene-level CpG degree distribution.

(b) DNB-like network: Tumor vs Normal.

(c) DNB-like network: Adjacent vs Normal.

Figure 3.5: Dynamic-network proxy visualizations in GSE69914.

ordering mirrors the characteristic architecture of the HM450 arrays reported in the literature [28, 29]. The coverage observed is therefore consistent with the known genomic composition of the Illumina methylation arrays, supporting correct manifest alignment.

Dynamic-network proxies Although this analysis does not follow the full mathematical formalism of Dynamic Network Biomarkers (DNBs), the degree-based and network-level metrics (Figure 3.5) provide a qualitative view of how methylation perturbations may organize at the network level. DNB theory, originally formulated for gene-expression and regulatory networks, predicts that near a critical transition a subset of components displays increased internal fluctuations and coordinated behaviour [30, 31, 32]. This conceptual framework is here adapted to CpG-level methylation patterns by examining degree distributions and $\Delta\beta$-driven network layouts, enabling exploration of whether tumour-associated instability manifests as locally coordinated methylation shifts without implementing the complete DNB pipeline. The degree–rank plot (Figure 3.5a) displays a heavy-tailed distribution of CpG counts per gene, consistent with the hub–periphery structure typical of biological systems. In the Tumor vs Normal comparison (Figure 3.5b), a compact nucleus of high-degree nodes with larger $\Delta\beta$ deviations suggests a locally coordinated and perturbed subset of CpG/gene units. The Adjacent vs Normal network (Figure 3.5c), by contrast, shows no comparable hub concentration and smaller, more diffuse deviations, consistent with only early or weakly organised perturbations. Together, these observations are conceptually compatible with the interpretation that tumour tissue occupies a more unstable network regime while adjacent tissue remains closer to a stable, pre-transition configuration.

Overall interpretation The exploratory analysis of GSE69914 reveals a consistent but overall subtle separation between the three tissue states. Normal and Adjacent samples exhibit highly similar global methylation profiles, variance

Table 3.2: Overview of the GSE225845 dataset after initial cohort curation.

CpGs	Samples	Normal	Adjacent	Tumor
750,426	477	113	140	224

distributions, correlation structure, and low-dimensional embeddings, indicating that large-scale methylome organisation remains largely conserved in Adjacent tissue. Across these global summaries, the two groups are so intermixed that no clear or marked distinction emerges between Normal and Adjacent samples when using whole-methylome descriptors. Tumor samples display the expected increase in heterogeneity and a shift toward hypomethylation; however, these differences remain modest at the whole-methylome level and do not produce clear separation in unsupervised embeddings. Locus-specific analyses, however, highlight focused patterns of disruption. Outlier CpGs exhibit both hyper- and hypomethylation events, and $\Delta\beta$ distributions reveal a distinct hypomethylation bias in Tumor samples together with a milder but detectable shift in Adjacent tissue. Dynamic-network proxies further indicate a more compact and perturbed core in Tumor samples, whereas Adjacent tissue appears more diffuse and weakly structured.

Overall, GSE69914 constitutes a technically clean dataset with well-structured methylation patterns and biologically interpretable deviations. The results indicate that tumour-related alterations are primarily localized at specific CpG loci rather than reflected in global methylome reorganization. Within this framework, the Adjacent tissue exhibits subtle yet measurable deviations from the Normal baseline, consistent with previously described field-effect patterns in breast cancer epigenomics.

3.3.2 Dataset GSE225845

Dataset overview GSE225845 is a breast tissue DNA methylation cohort profiled using the Illumina Infinium MethylationEPIC BeadChip (850K). The dataset is part of the NCI-Maryland Breast Cancer Cohort and includes samples spanning the Normal–Adjacent–Tumor spectrum that motivates the biological framework of this thesis. An overview of the sample composition and CpG coverage after cohort harmonization is reported in Table 3.2, including the distribution across tissue classes. The dataset further provides detailed demographic and clinical metadata (e.g., age at surgery, race, sex, and tissue descriptors), enabling controlled downstream analyses when required.

Global methylation distributions The group-wise mean β-value densities (Figure 3.6) display the expected bimodal configuration of EPIC arrays. As

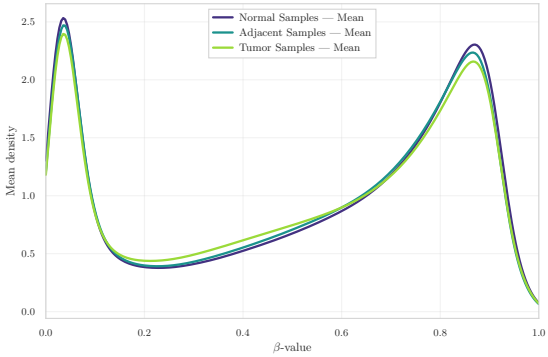
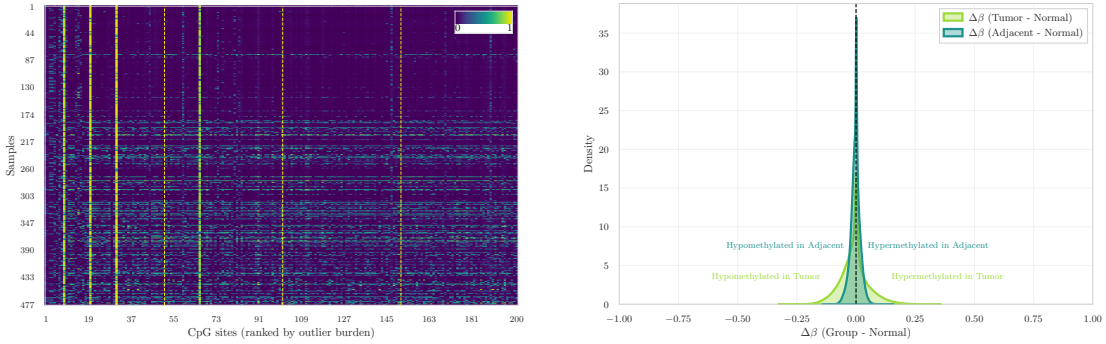


Figure 3.6: Group-wise mean β-value density in GSE225845.

observed in GSE69914, the three tissue groups show a high degree of overlap across the entire β range. Normal and Adjacent samples exhibit nearly indistinguishable density profiles. Tumor samples present only minimal deviations, characterised by a slight increase in density at low β values and a marginal attenuation of the high-β peak. At the whole-methylome level, no marked global redistribution of methylation levels is evident.

CpG-level instability and recurrent outliers The heatmap of the top outlier CpG loci (Figure 3.7a) indicates that instability within this subset is predominantly directional. Recurrent outlier events are largely associated with very low β values, consistent with focal hypomethylation affecting a restricted set of CpG sites. The perturbation pattern therefore appears concentrated rather than diffuse, suggesting localized instability rather than a genome-wide redistribution. The corresponding $\Delta\beta$ distributions comparing Tumor and Adjacent tissues against Normal (Figure 3.7b) are sharply centered around zero, indicating that most loci remain stable. However, both comparisons exhibit a heavier negative tail relative to the positive side. Tumor samples display the most extended negative tail, consistent with stronger hypomethylation shifts, whereas Adjacent tissue shows a milder but still detectable asymmetry toward $\Delta\beta < 0$.

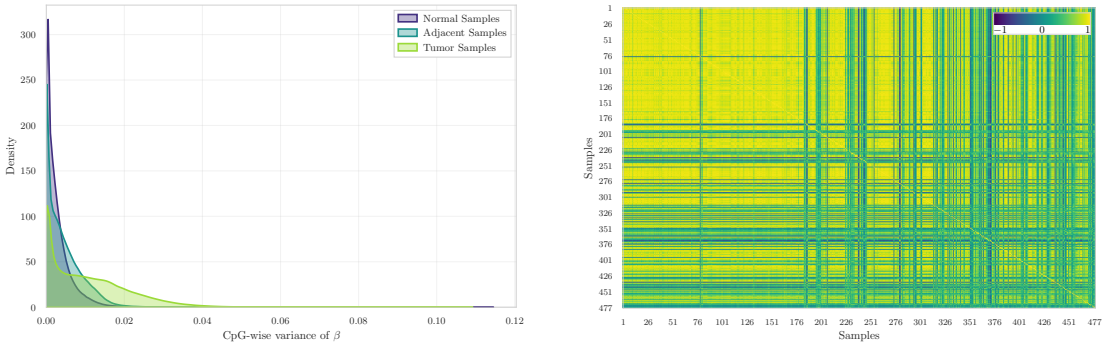
Variance and correlation structure The CpG-wise variance distributions (Figure 3.8a) indicate clear differences in variability across tissue groups. Normal samples exhibit highly concentrated variance values near zero, whereas Tumor samples display a substantially longer right tail, reflecting an increased proportion of CpG loci with elevated variability. Adjacent samples show a broader distribution than Normal tissues, with partial overlap with both Normal and Tumor profiles, but without forming a strictly intermediate pattern. The sample correlation heatmap (Figure 3.8b) shows uniformly positive correlations across the dataset. Most pairwise



(a) Top outlier CpG loci.

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

Figure 3.7: Heatmap of the top outlier CpGs (left) and corresponding $\Delta\beta$ distributions (right), illustrating CpG-level instability in GSE225845.



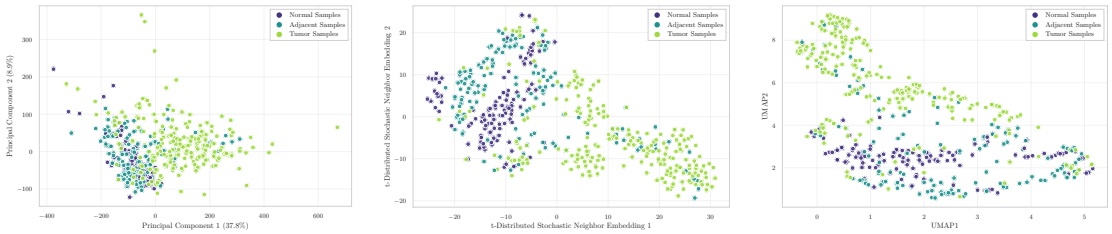
(a) CpG-wise variance distributions.

(b) Sample correlation heatmap.

Figure 3.8: Variance and correlation structure across Normal, Adjacent, and Tumor samples in GSE225845.

values lie within a narrow high-correlation range, and no distinct block structures separating tissue classes are observed. This pattern indicates preservation of global methylation structure, with group differences emerging primarily from locus-specific variability rather than from large-scale shifts in overall methylome organization.

Low-dimensional embeddings Dimensionality reduction via PCA, t-SNE, and UMAP (Figure 3.9) reveals limited but consistent group structure. In PCA space, the three tissue groups largely overlap, with Tumor samples showing slightly broader dispersion along the second principal component. Nonlinear embeddings reveal additional structure: Tumor samples occupy a more extended and peripheral region in both t-SNE and UMAP, while Normal and Adjacent samples form compact yet



(a) PCA
trustworthiness = 0.881.

(b) t-SNE
trustworthiness = 0.939.

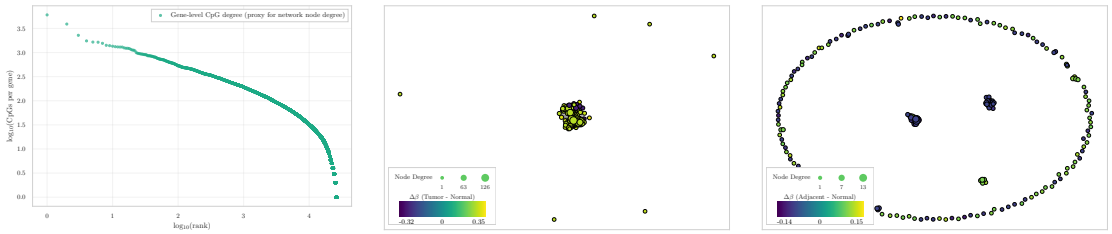
(c) UMAP
trustworthiness = 0.894.

Figure 3.9: Low-dimensional embeddings samples in GSE225845.

overlapping clusters. Trustworthiness scores range from 0.881 to 0.939, indicating good preservation of local neighbourhood structure. None of the methods yields sharply separated clusters.

Genome annotation coverage Consistent with the architecture of the EPIC platform, the genomic distribution of annotated CpG contexts is markedly unbalanced. The largest fraction of probes maps to open-sea or non-island regions, which constitute the dominant category. CpG islands represent the second most abundant class, followed by north and south shores in comparable proportions, whereas north and south shelves account for the smallest fractions. This annotation profile aligns with the expected genomic composition of the EPIC array and supports correct manifest integration.

Dynamic-network proxies The degree–rank distribution (Figure 3.10a) exhibits a heavy-tailed profile, with a limited number of genes associated with many CpGs and a long tail of low-degree nodes. Such right-skewed degree structures are characteristic of heterogeneous biological networks [33]. In the Tumor vs Normal layout (Figure 3.10b), nodes aggregate into a dense and highly compact central cluster, with only a small number of isolated peripheral points. This geometry indicates that the largest Tumor–Normal $\Delta\beta$ deviations are concentrated within a cohesive subset of CpG–gene units rather than broadly distributed across the network. In contrast, the Adjacent vs Normal layout (Figure 3.10c) displays a markedly different configuration. While a few small central aggregates are visible, most nodes are arranged along a wide annular structure, forming a pronounced ring-like embedding. This organisation reflects weaker and more spatially dispersed deviations, without the central concentration observed in the Tumor comparison. Altogether, the proxy visualisations indicate that Tumor–Normal contrasts generate a compact and locally coherent network signal, whereas Adjacent–Normal differences remain diffuse and geometrically fragmented.



(a) Gene-level CpG degree distribution. **(b)** DNB-like network: Tumor vs Normal. **(c)** DNB-like network: Adjacent vs Normal.

Figure 3.10: Dynamic-network proxy visualizations in GSE225845.

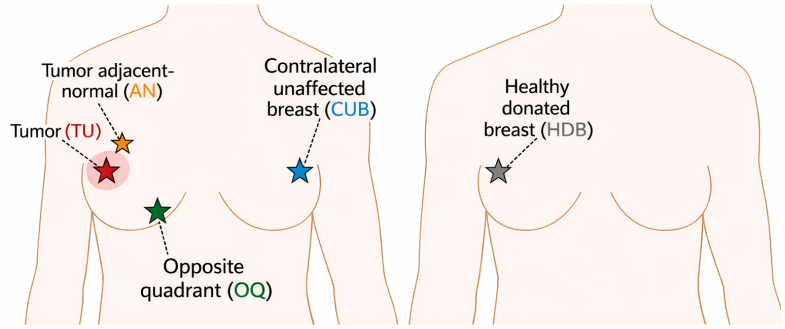


Figure 3.11: Schematic representation of the five tissue categories collected along the tumor proximity axis (TPxA), reproduced from [34].

Overall interpretation The exploratory analysis of GSE225845 indicates that, at the whole-methylome level, the three tissue states share highly similar global methylation profiles. Group-wise mean β-value densities are nearly superimposed, and both correlation structure and linear embeddings show substantial overlap. These observations indicate preservation of large-scale methylome organisation across Normal, Adjacent, and Tumor tissues, without evidence of pronounced genome-wide redistribution of methylation levels. Locus-focused analyses reveal more structured differences. Among the 200 CpGs with the highest outlier burden, instability is predominantly directional and characterised by focal hypomethylation. Variance distributions further show increased dispersion in Tumor samples, while Adjacent tissues exhibit broader variability than Normal without forming a strictly intermediate pattern. Nonlinear embeddings (t-SNE and UMAP) emphasise differences in spatial dispersion rather than discrete cluster separation. Normal and Adjacent samples occupy relatively compact regions, whereas Tumor samples extend over a wider area, reflecting increased heterogeneity. Dynamic-network proxies show that Tumor–Normal contrasts generate a dense central cluster of perturbed CpG–gene units, while Adjacent–Normal differences are distributed along a broader,

Table 3.3: Overview of the GSE287331 dataset after initial cohort curation.

CpGs	Samples	Normal	Adjacent	Tumor	OQ	CUB
866,552	446	182	60	69	67	68

ring-like structure without strong central concentration. Despite the larger sample size of this cohort, which provides stable estimates of global methylation structure, the separation between Normal and Adjacent tissues remains modest. Differences between these two groups are therefore primarily focal and do not manifest as large-scale structural shifts.

Overall, GSE225845 represents a globally homogeneous dataset in which tissue-state differences become apparent only when examined through locus-specific perturbations, variability patterns, and small-scale network organisation.

3.3.3 Dataset GSE287331

Dataset overview GSE287331 is a breast tissue DNA methylation cohort profiled using the Illumina Infinium MethylationEPIC v1.0 BeadChip. The dataset is organised along a tumor proximity axis (TPxA) spanning multiple anatomical and pathological contexts. Five tissue categories are represented: healthy donated breast (HDB), contralateral unaffected breast (CUB), ipsilateral opposite quadrant (OQ), adjacent normal tissue (AN), and tumor (TU). An overview of the full cohort composition is reported in Table 3.3. For cross-dataset consistency, only three biologically aligned classes are retained for downstream analyses: HDB (treated as Normal baseline), AN (Adjacent), and TU (Tumor). OQ and CUB samples are excluded, as they represent intermediate benign tissues along the proximity axis and are not directly comparable with the three-class framework adopted in the other cohorts.

Global methylation distributions The group-wise mean β-value density curves (Figure 3.12) display the expected bimodal configuration of Illumina methylation arrays. A dominant peak is observed at high methylation levels, accompanied by a smaller peak consistent with standard EPIC methylation profiles [35]. Although the overall shapes remain comparable across tissue groups, a clear ordering is visible at the high-β peak. Normal samples exhibit the highest density, Adjacent samples show a slightly attenuated peak, and Tumor samples present the lowest amplitude together with a broader distribution extending toward intermediate β values. This progressive reduction in the fully methylated peak is consistent with increasing methylation variability along the Normal–Adjacent–Tumor axis.

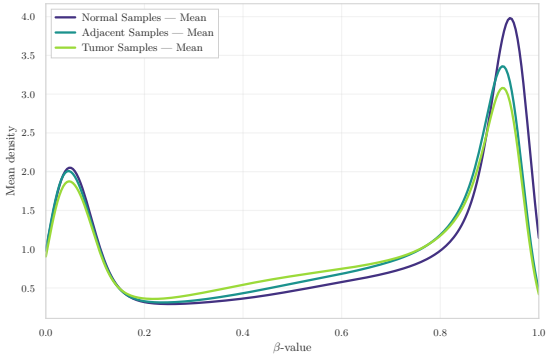
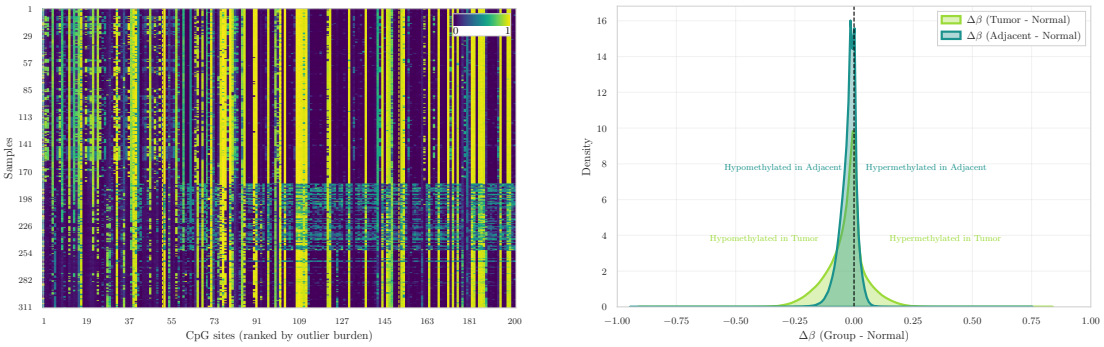


Figure 3.12: Group-wise mean β-value density in GSE287331.

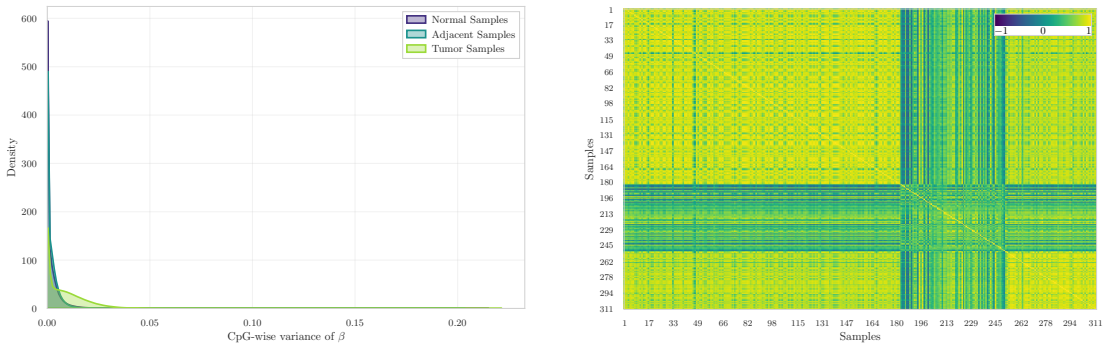


(a) Top outlier CpG loci.

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

Figure 3.13: Heatmap of the top outlier CpGs (left) and corresponding $\Delta\beta$ distributions (right), illustrating CpG-level instability in GSE287331.

CpG-level instability and recurrent outliers The heatmap of the top outlier CpG loci (Figure 3.13a) shows that epigenetic instability is not uniformly distributed. Alterations concentrate at specific CpG sites and within specific samples, forming vertical and horizontal bands rather than a diffuse genome-wide pattern. Both directions of deviation are observed, with focal hypermethylation and focal hypomethylation present within the selected loci [1]. The corresponding $\Delta\beta$ distributions (Figure 3.13b) highlight distinct profiles for the two contrasts. The Tumor–Normal curve is broader and displays a higher density of small negative shifts, indicating widespread but moderate hypomethylation in Tumor samples. In contrast, the Adjacent–Normal distribution is more sharply centered around zero but exhibits a comparatively more pronounced negative tail, reflecting fewer yet larger hypomethylated deviations. Overall, hypomethylation represents the dominant direction of change in both comparisons.

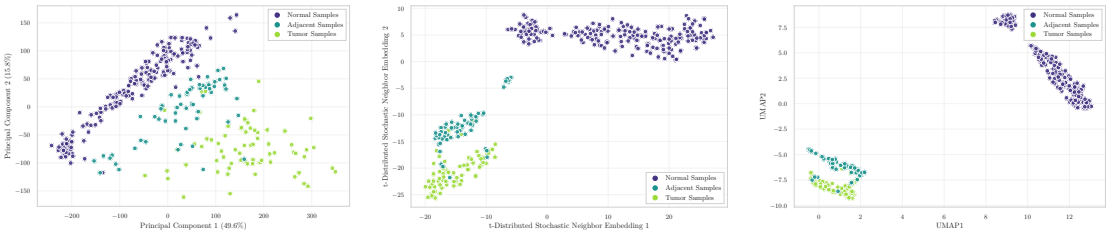


(a) CpG-wise variance distributions. **(b)** Sample correlation heatmap.

Figure 3.14: Variance and correlation structure across Normal, Adjacent, and Tumor samples in GSE287331.

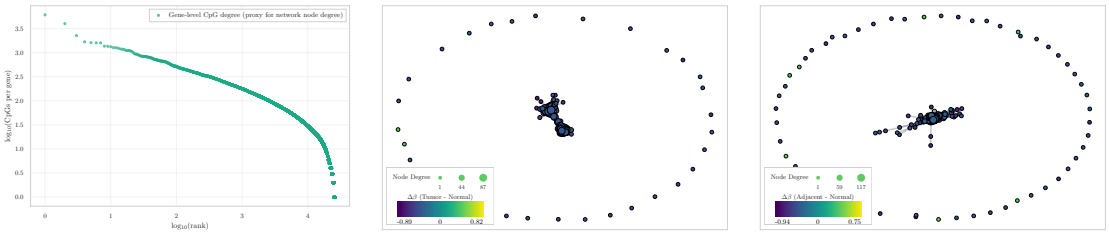
Variance and correlation structure The CpG-wise variance density curves (Figure 3.14a) show that Normal and Adjacent samples are nearly superimposed, with both groups exhibiting extremely low and tightly concentrated variance values. Tumor samples display a slightly broader right tail, indicating modestly increased genome-wide variability; however, the overall distribution remains sharply peaked near zero. At the level of CpG-wise variance, Adjacent tissue remains closely aligned with the Normal baseline, whereas Tumor samples show only a mild elevation in variability. A complementary perspective is provided by the sample correlation heatmap (Figure 3.14b). Correlations are uniformly high across the cohort, consistent with the predominance of stable CpG loci in genome-wide methylation data [23]. Although tissue labels are not overlaid on the matrix, localized blocks of higher similarity are visible, indicating structured relationships within subsets of samples rather than abrupt global separation.

Low-dimensional embeddings Dimensionality reduction using PCA, t-SNE, and UMAP was applied to visualise the global organisation of methylation profiles (Figure 3.15). Across all three embeddings, a consistent and well-defined structure is observed. Normal and Adjacent samples form two clearly separated clusters. The separation is visible in PCA space, becomes more pronounced in t-SNE, and appears as fully detached manifolds in the UMAP projection. Tumor samples occupy a distinct and more dispersed region of the embedding space, reflecting increased epigenetic heterogeneity. Trustworthiness scores are uniformly high (0.961–0.980), indicating excellent preservation of local neighbourhood structure in all low-dimensional projections. No substantial overlap between Tumor and the other two groups is observed in any of the projections, indicating strong intrinsic group structure in this cohort.



(a) PCA trustworthiness = 0.961. **(b)** t-SNE trustworthiness = 0.980. **(c)** UMAP trustworthiness = 0.973.

Figure 3.15: Low-dimensional embeddings of GSE287331 samples.



(a) Gene-level CpG degree distribution. **(b)** DNB-like network: Tumor vs Normal. **(c)** DNB-like network: Adjacent vs Normal.

Figure 3.16: Dynamic-network proxy visualizations for GSE287331.

Genome annotation coverage The genomic distribution of annotated CpG contexts is dominated by Open Sea regions, followed by CpG Islands, then Shores, with Shelves representing the smallest fractions. This ordering is consistent with the established architecture of EPIC arrays [28, 29].

Dynamic-network proxies The degree–rank curve (Figure 3.16a) shows a heavy-tailed distribution of CpG counts per gene, consistent with a hub–periphery architecture typical of biological networks. Both $\Delta\beta$-based network layouts — Tumor vs Normal and Adjacent vs Normal (Figure 3.16b, Figure 3.16c) — share a similar overall structure: a dense and compact central block of nodes connected by short-range edges, surrounded by a ring of peripheral, weakly connected nodes. This geometry reflects the underlying degree distribution and indicates that most CpG–gene units contribute only minimally to group-level differences. Despite this shared organisation, the Tumor vs Normal comparison exhibits a more pronounced central cluster with larger $\Delta\beta$ magnitudes, suggesting that a restricted subset of CpG–gene units undergoes locally coordinated perturbations. In contrast, the Adjacent vs Normal layout preserves the same global shape but shows weaker deviations, indicating only mild coordination of methylation changes.

Overall interpretation The exploratory analysis of GSE287331 indicates that, at the whole-methylome level, the three tissue states retain a broadly preserved global methylation architecture. Group-wise mean β-value densities follow the typical bimodal EPIC profile with only a subtle gradient (Normal > Adjacent > Tumor), and both CpG-wise variance distributions and sample correlation structure show high overall similarity across tissues, with only mild variability increases in Tumor samples. In contrast to the previous cohorts, low-dimensional embeddings reveal a clear and consistent separation of all three groups. Normal and Adjacent samples form distinct clusters in PCA, with the separation further strengthened in t-SNE and UMAP, where the groups appear as detached manifolds. Tumor samples occupy a separate and more diffuse region, reflecting increased epigenetic heterogeneity. These projections indicate that, despite similar first- and second-order global summaries, multivariate structure in this dataset captures strong group-level organisation. Locus-specific analyses reveal focused disruptions. Tumor samples exhibit widespread mild hypomethylation, whereas Adjacent tissue shows fewer but larger-magnitude deviations. Dynamic-network proxies further indicate a compact and perturbed central core in the Tumor comparison, while Adjacent–Normal differences display weaker coordination, consistent with early-stage methylation alterations.

Overall, GSE287331 emerges as a structurally coherent and biologically informative dataset in which Normal–Adjacent–Tumor separation is markedly more pronounced than in the other cohorts, while global methylation architecture remains broadly conserved.

3.4 Inter Dataset Exploration and Comparison

The objective of this section is not to merge the datasets—an approach that would be biologically and technically inappropriate due to differences in array platforms (HM450 vs. EPIC), preprocessing pipelines—but rather to compare their internal behaviours. Such a cross-dataset perspective enables the assessment of whether the early epigenetic alterations identified in the intra-dataset analyses are *reproducible across independent cohorts*, thereby providing support for the central biological premise of this thesis: the detection of early methylation drift in histologically normal tissues and the evaluation of its potential contribution to tumour initiation.

3.4.1 Analytical Framework

Across all three datasets, the intra-dataset analyses revealed three highly consistent phenomena.

– Normal and Adjacent tissues are globally similar, with nearly superimposed

Table 3.4: Cross-dataset comparison of Normal (N) vs Adjacent (A) contrasts using global and instability metrics.

Dataset	Mean $\lvert\Delta\beta\rvert$ (A–N)	Var Ratio (A/N)	Outlier Ratio (A/N)	Silhouette (A/N)
GSE69914	0.009	0.988	1.353	0.003
GSE225845	0.015	0.972	2.790	0.042
GSE287331	0.031	0.930	13.329	0.344

β-value distributions and comparable CpG-wise variance profiles.

– Tumour samples show increased heterogeneity and a clear bias toward hypomethylation, although the magnitude of this shift varies across datasets.

– Adjacent tissues exhibit early methylation drift, detectable through $\Delta\beta$ distributions and outlier profiles even when global metrics remain highly similar to Normal.

These patterns are consistent with the expected biological progression from Normal to Tumour, in which epigenetic deregulation emerges gradually and affects normal-appearing tissue surrounding the tumour. The inter-dataset analysis therefore evaluates the *directional consistency* of these effects rather than aiming at strict quantitative comparability.

3.4.2 Cross-Dataset Comparative Metric

To quantify the epigenetic deviation of Adjacent tissue from the Normal methylome across the three cohorts, a set of global and instability-oriented metrics was computed independently within each dataset for the Adjacent versus Normal contrast. All metrics reported in Table 3.4 were computed using the complete CpG set available in each dataset, rather than the cross-platform intersection. The resulting values therefore reflect cohort-specific dimensionality and internal variance structure. Collectively, these metrics indicate the presence of early epigenetic drift in histologically non-tumour tissue. Across cohorts, increasing mean $\lvert\Delta\beta\rvert$, progressive variance imbalance, elevated outlier ratios, and improved silhouette separation consistently support a directional shift from Normal to Adjacent states, in agreement with the field-defect framewor [3] and reinforced in other tissue contexts.

Mean absolute methylation difference For each CpG, the absolute mean difference between Adjacent and Normal was defined as:

$$|\Delta\beta_i| = \left|\beta_i^{(A)} - \beta_i^{(N)}\right|. \tag{3.2}$$

The cross-cohort pattern reveals progressively larger locus-specific shifts from GSE69914 to GSE225845 and GSE287331. Although the absolute magnitude of the effect remains modest, it is consistent in direction across cohorts. This monotonic increase reflects a strengthening early deviation in Adjacent tissues. Comparable locus-wise methylation drifts in histologically normal tissue have been reported in independent systems, such as colonic mucosa adjacent to adenomas [36], supporting the interpretation of these alterations as early manifestations of field cancerisation.

Global variance ratio For each sample s, the global methylation variance was computed across all CpG sites: $v_s = \mathrm{Var}(\beta_{s,1}, \beta_{s,2}, \ldots, \beta_{s,m})$, where m denotes the number of CpGs. Group-wise mean variances were then defined as:

$$\overline{v}^{(N)} = \frac{1}{|N|} \sum_{s \in N} v_s, \qquad \overline{v}^{(A)} = \frac{1}{|A|} \sum_{s \in A} v_s, \tag{3.3}$$

and the reported metric corresponds to their ratio:

$$\frac{\mathrm{Var(A)}}{\mathrm{Var(N)}} = \frac{\overline{v}^{(A)}}{\overline{v}^{(N)}}. \tag{3.4}$$

Across cohorts, this ratio remained consistently close to unity, indicating that Adjacent tissue does not exhibit diffuse genome-wide instability. Such behaviour suggests that early pre-neoplastic alterations are predominantly *focal* rather than global, with only specific CpG sites displaying abnormal dispersion while overall variance remains largely preserved [36]. This interpretation is consistent with contemporary models of early tumourigenesis, in which subtle and locally restricted perturbations accumulate prior to large-scale chromatin deregulation and overt neoplastic transformation [37].

Outlier burden ratio For each sample, the outlier burden is defined as the number of CpG sites whose methylation deviates from the Normal reference distribution by more than three times the Normal interquartile range:

$$\mathrm{outlier}_s = \#\left\{ i \ : \ \left|\beta_{i,s} - \tilde{\beta}_i^{(N)}\right| > 3 \cdot IQR_i^{(N)} \right\}, \tag{3.5}$$

where $\tilde{\beta}_i^{(N)}$ and $IQR_i^{(N)}$ denote the median and interquartile range of CpG i across Normal samples. The reported A/N ratio corresponds to the mean outlier burden in Adjacent tissue divided by the mean burden in Normal tissue. Across cohorts, this metric exhibits the most pronounced cross-dataset gradient, with progressively stronger enrichment of outlier events in Adjacent tissue. Such behaviour mirrors evidence that increased local methylation variability in histologically normal tissue represents a sensitive marker of early instability [36], and provides quantitative support for the presence of marked field effects in GSE287331.

Silhouette score For each sample s, let $a(s)$ denote the mean distance from s to all other samples within the same group (cohesion), and let $b(s)$ denote the smallest mean distance from s to samples in the alternative group (separation). The silhouette coefficient for sample s is defined as:

$$\mathrm{sil}(s) = \frac{b(s) - a(s)}{\max\{a(s),\, b(s)\}}, \tag{3.6}$$

and the reported score corresponds to the mean silhouette computed across all Normal and Adjacent samples. Higher values indicate clearer multivariate separation between Normal and Adjacent states, whereas values near zero reflect substantial overlap. Across cohorts, the silhouette analysis reveals limited separation in GSE69914 and GSE225845, consistent with minimal global displacement in multivariate space. In contrast, GSE287331 exhibits partial separation, indicating that the epigenetic landscape of Adjacent tissue is measurably shifted relative to Normal. This behaviour aligns with mechanistic models in which epigenetic dysregulation constitutes one of the earliest events in tumour initiation [37], preceding morphological transformation. Collectively, the set of metrics delineates a consistent Normal–Adjacent gradient across datasets. Early epigenetic drift is detectable in all cohorts, while its magnitude, focality, and multivariate visibility progressively intensify from GSE69914 to GSE287331. Importantly, the combination of stable global variance with pronounced outlier enrichment supports a model in which early pre-neoplastic changes arise through *localised* perturbations, a hallmark of field cancerisation [3, 36].

3.4.3 Visual Comparative Analysis

To complement the quantitative summary reported in Table 3.4, this section provides a visual characterisation of the Normal–Adjacent contrast across the three cohorts. All analyses are performed on the three-way intersection of 326,330 CpG sites, ensuring maximal comparability across platforms (HM450 and EPIC) and preprocessing pipelines [28]. The objective is not to achieve numerical concordance at the single-CpG level—a notoriously difficult task due to platform differences, probe-design effects, and cohort-specific processing [29, 38]—but rather to evaluate whether the *shape*, *direction*, and *global behaviour* of early methylation drift remain coherent across studies.

Global $\Delta\beta$ distributions The overlaid $\Delta\beta$ (Adjacent–Normal) density curves across the three cohorts are shown in Figure 3.17. All distributions are sharply centred around zero, indicating that the N–A methylation drift remains globally subtle in each dataset, consistent with the focal and low-amplitude nature of early epigenetic alterations described in pre-neoplastic tissues [3, 39]. Despite differences

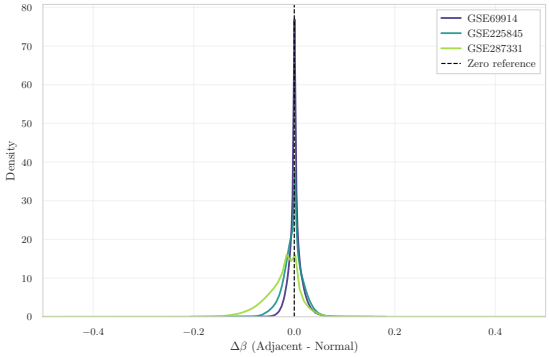
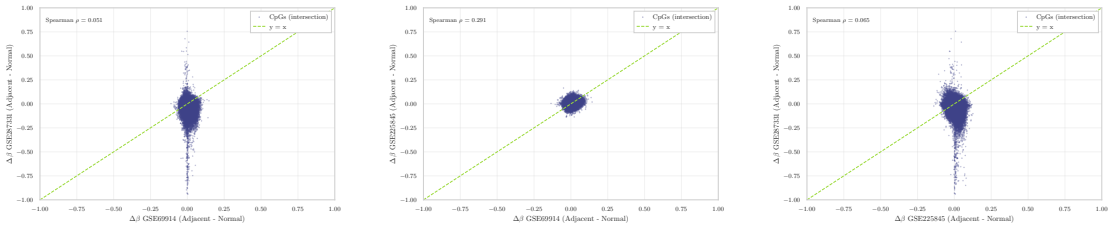


Figure 3.17: Overlay of $\Delta\beta$ (Adjacent–Normal) density curves across the three datasets, computed on the 326,330 CpG sites shared by all cohorts.

in dispersion—reflecting platform-dependent noise, probe-type composition, and study-specific variability [26, 29]—the three curves exhibit a highly similar shape. This pattern indicates that early epigenetic deviations in Adjacent tissue, although small in magnitude, exhibit reproducible global distributional features across independent cohorts, rather than exact concordance at the single-CpG level.

Pairwise $\Delta\beta$ concordance across datasets To assess whether locus-specific A–N shifts replicate across cohorts, pairwise scatterplots are shown for all dataset pairs (Figure 3.18a, Figure 3.18b, Figure 3.18c). Each point corresponds to a CpG in the three-way intersection; axes represent the mean $\Delta\beta$ (A–N) computed independently within each dataset. Across all pairs, Spearman correlations remain weak ($\rho = 0.05$–0.29), consistent with the well-established difficulty of achieving single-CpG reproducibility across independent methylation studies [28, 38, 40]. The comparison between GSE69914 and GSE225845 exhibits mild monotonic agreement, whereas pairs involving GSE287331 show near-zero correlation. These results indicate that while the *global* N–A drift remains consistent across cohorts (as demonstrated by the density curves), *fine-scale, CpG-specific* effects are strongly influenced by platform architecture, batch structure, and study-specific variability [29, 38].

t-SNE embeddings on the shared CpG intersection To investigate the multivariate structure of the N–A contrast across cohorts, t-SNE embeddings are computed on M-values restricted to the shared CpGs. Two complementary visualisations of the embedding are provided in Figure 3.19a (coloured by dataset) and Figure 3.19b (coloured by tissue state). When coloured by dataset, the embedding exhibits clear clustering by study, reflecting the dominant influence of batch and platform effects in high-dimensional methylation data [38, 40]. This

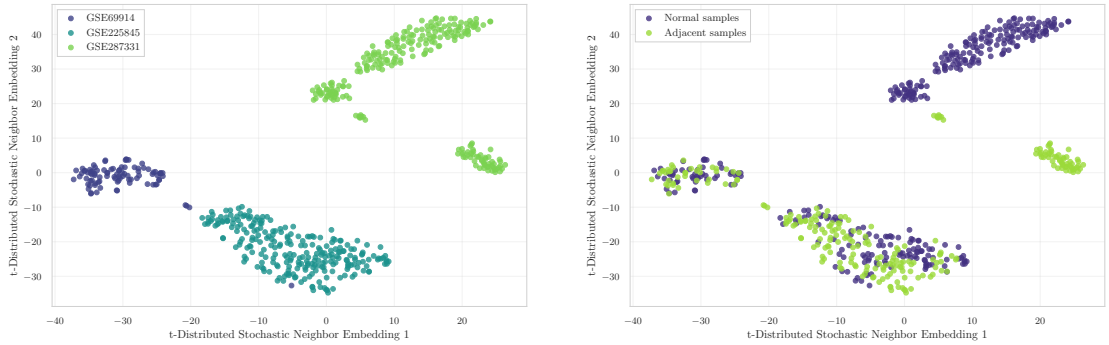


(a) GSE69914-GSE287331, Spearman $\rho \approx 0.051$

(b) GSE69914-GSE225845, Spearman $\rho \approx 0.291$

(c) GSE225845-GSE287331, Spearman $\rho \approx 0.007$

Figure 3.18: Pairwise $\Delta\beta$ (Adjacent–Normal) scatterplots across the three methylation datasets, computed on the 326,330 shared CpG sites. Spearman ρ values are reported in each panel.



(a) t-SNE coloured by dataset trustworthiness = 0.976

(b) t-SNE coloured by tissue state trustworthiness = 0.976

Figure 3.19: t-SNE embeddings computed on the 326,330 CpGs shared across GSE69914, GSE225845, and GSE287331.

pattern indicates that the global methylation structure is largely driven by inter-cohort variability, supporting the decision not to merge datasets at the raw feature level. When coloured by tissue state (Normal vs. Adjacent), no global separation across cohorts emerges, as the embedding remains primarily structured by dataset identity. However, within individual study-specific clusters, a state-dependent displacement becomes visible. This separation is particularly evident in GSE287331, where Normal and Adjacent samples form more clearly distinct substructures. This behaviour is consistent with early, localised methylation drift, mirroring the quantitative metrics reported in Table 3.4.

Overall, the visual analyses corroborate the quantitative metrics: early epigenetic drift is detectable in all cohorts, but its magnitude and visibility vary substantially, with GSE287331 exhibiting the clearest field-defect signature. The combination

of globally subtle yet directionally consistent $\Delta\beta$ patterns, weak CpG-level concordance, and dataset-specific multivariate structure reinforces the view that the Normal→Adjacent transition reflects genuine early epigenetic deregulation rather than dataset-specific artefacts.

3.5 Exploratory Findings and Pre-Processing Implications

The exploratory analyses presented in this chapter reveal a consistent picture across all three cohorts: the Normal–Adjacent transition does not manifest as a global reorganisation of the methylome, but as a subtle, focal drift that becomes detectable only through locus-specific perturbations and multivariate displacements. This characteristic makes it particularly sensitive to technical sources of variability — probe-design bias, batch effects, and platform differences — which operate at a comparable or larger scale than the biological signal of interest. This sensitivity has direct methodological consequences. Raw methylation data, as distributed via GEO, cannot be used directly for downstream modelling without first addressing these technical confounders: doing so would risk attributing platform-specific artefacts to genuine epigenetic alterations, or conversely, suppressing the subtle Normal–Adjacent differences under overly aggressive normalisation. The challenge is therefore not simply to clean the data, but to do so in a way that preserves the biological structure identified here.

Chapter 4 addresses this challenge by developing a structured, platform-aware preprocessing pipeline. Each step is motivated by the specific limitations identified in the present chapter.

Chapter 4

Data preprocessing

4.1 Preprocessing Rationale and Technical Constraints

DNA methylation arrays superimpose biological variability and platform-specific artefacts on the same high-dimensional signal. The exploratory analyses of Chapter 3 showed that Normal and Adjacent tissues exhibit nearly identical global distributions, variance profiles, and correlation structure across all three cohorts. Meaningful differences emerge exclusively at focal CpG loci — as quantified by locus-specific $\Delta\beta$ distributions and outlier burden ratios — and remain subtle even at that scale (mean $|\Delta\beta|$ from 0.009 in GSE69914 to 0.031 in GSE287331). This makes the Normal–Adjacent signal particularly vulnerable to technical confounders operating at a comparable or larger scale, so preprocessing must be precise: aggressive enough to remove noise, yet conservative enough to preserve the epigenetic drift that motivates this thesis. Three main classes of technical issues must be addressed before any statistical modelling. First, *probe-design bias* arises from the coexistence of Type I and Type II probes, which produce inherently different β-value distributions and distort cross-sample comparability if uncorrected [41, 42]. This issue is not uniform across datasets: GSE69914 and GSE225845 were released with normalization already applied, whereas GSE287331 shows clear diagnostic evidence of missing probe-type correction, requiring an additional decision with non-trivial consequences for downstream group separation (Section 4.3). Second, *unreliable or biologically confounded probes* — including cross-reactive loci, SNP-affected sites, non-CpG targets, and sex-chromosome probes — introduce systematic measurement error that can generate spurious associations or mask genuine methylation differences [43, 44]. Third, the *statistical scale of β-values*, with their bounded [0,1] support and mean-dependent heteroscedasticity, violates the assumptions of standard parametric models, motivating transformation to a

more statistically well-behaved representation [41].

This chapter addresses these issues through a structured, platform-aware pipeline applied consistently across all cohorts. Section 4.2 describes the general framework; Section 4.3 documents dataset-specific decisions, focusing on the handling of probe-type bias in GSE287331; Section 4.4 summarises the effect of preprocessing on CpG dimensionality and cross-dataset overlap, providing the basis for Chapter 5.

4.2 General Preprocessing Framework

This section describes the general preprocessing framework applied to all three methylation datasets. The pipeline comprises four sequential steps, each addressing a distinct class of technical issue identified in the overview above. Two of these steps—data integrity verification and Infinium probe-design bias correction—require dataset-specific decisions whose rationale and consequences are discussed in Section 4.3. The remaining two steps—technical probe filtering and β-to-M transformation—are applied uniformly across all cohorts following the same protocol.

4.2.1 Structural Integrity and Cohort Harmonization

Prior to any preprocessing operation, each dataset undergoes a structural verification stage to ensure the consistency and completeness of the input data. This step checks for correct alignment between the methylation matrix and its associated phenotype table, verifies sample identifier uniqueness, and confirms the absence of missing metadata fields. Only samples for which both a complete methylation profile and a valid tissue label are available are retained for downstream analysis. Probe-level and sample-level missingness in the β-value matrix is also assessed at this stage. For two of the three datasets (GSE69914 and GSE225845), the released matrices are fully observed and require no imputation. For GSE287331, a non-negligible fraction of CpG sites exhibits missing values across samples, necessitating a dedicated filtering and imputation strategy consistent with current recommendations for EPIC-array quality control [45, 46]. The dataset-specific treatment of missingness is detailed in Section 4.3.3.

Finally, cohort harmonization is applied where necessary to align tissue labels to the three-class framework (Normal, Adjacent, Tumor) adopted uniformly across datasets. Samples belonging to tissue categories outside this framework are excluded at this stage to ensure cross-dataset comparability.

4.2.2 Probe-Type Bias Diagnostics and Correction

A well-documented source of systematic bias in Illumina Infinium methylation arrays is the coexistence of two chemically distinct probe designs. Type I probes use two bead types per CpG (one for the methylated channel, one for the unmethylated channel), whereas Type II probes use a single bead type and rely on single-base extension with two-color detection. This design difference produces inherently distinct β-value distributions for the two probe types: Type I probes tend to yield more extreme values near 0 and 1, while Type II probes exhibit a broader, intermediate distribution. If uncorrected, this imbalance propagates into downstream statistical analyses and can generate probe-type-specific artefacts in differential methylation testing [23, 26]. The canonical approach to mitigating this bias is Beta-Mixture Quantile normalisation (BMIQ), introduced by Teschendorff et al. [23]. BMIQ explicitly models probe-type bias at the β-value level through a state-aware distributional alignment procedure. For each sample independently, a three-component Beta mixture model is first fitted to the β-value distribution of Type I probes, representing unmethylated, partially methylated, and fully methylated states. These fitted components define the reference state-specific distributions. Type II probes are then probabilistically assigned to one of the three methylation states based on their β-values, and within each state a quantile-mapping step is applied to align the Type II distribution to the corresponding Type I reference distribution. Because the procedure is performed independently for each sample, BMIQ does not introduce cross-sample information or global harmonisation across tissue classes; it strictly corrects within-sample probe-type bias. For each dataset, a systematic diagnostic is performed prior to any correction step [23, 47]: probes are stratified by design type using the official Illumina manifest, and two complementary assessments are carried out—(i) a density comparison of β-value distributions for Type I and Type II probes, and (ii) a quantile–quantile plot of Type II versus Type I quantiles, with the median absolute deviation (MAD) of the empirical Q–Q curve from the identity line as a scalar diagnostic metric. Datasets for which the original authors applied BMIQ or an equivalent normalisation are expected to exhibit near-overlapping distributions and a MAD close to zero; datasets lacking probe-type correction will display a pronounced distributional mismatch and elevated MAD.

The outcome of this diagnostic, and the corrective action taken, differs across cohorts and is reported in full in Section 4.3. Critically, the decision of whether to apply BMIQ is not made mechanically: as discussed in Section 4.3.3, BMIQ correction can attenuate genuine biological structure when group-level differences are expressed as broad distributional shifts, a phenomenon documented in systematic evaluations of EPIC-array normalisation methods [17, 48, 49]. The choice between correcting and preserving the raw data is therefore made on the basis of empirical evidence from the diagnostic plots and quantitative assessment of group separability

before and after correction.

4.2.3 Technical Probe Filtering

After addressing probe-design bias, unreliable CpG loci are removed through a systematic filtering procedure based on curated external annotation resources. The objective is to exclude probes whose measured β-values may reflect technical measurement error rather than genuine DNA methylation, thereby reducing noise and improving the reliability of downstream differential analyses. Four categories of technically problematic probes are systematically excluded.

Cross-reactive and multi-mapping probes. Probes that hybridise to multiple genomic loci produce β-values that cannot be unambiguously attributed to a single CpG site. These are identified using the catalogues of Naeem et al. [50], Chen et al. [51], Pidsley et al. [29], Zhou et al. [28], and McCartney et al. [52]. Full descriptions of each resource are provided in Appendix A.

SNP-affected probes. Probes whose interrogated CpG dinucleotide, single-base extension site, or probe body overlaps a common genetic variant may measure genotype rather than methylation status, producing population-stratified artefacts in group comparisons [28, 29]. These are excluded using the `MASK_snp5` and `MASK_extBase` fields from the Zhou annotation [28], supplemented by variant-aware lists from Naeem et al. [50].

Non-CpG probes. Probes targeting non-CpG cytosine contexts (identified by the `ch` prefix in Illumina probe identifiers) are excluded, as their interpretation differs from canonical CpG methylation and is outside the scope of this analysis [28, 29].

Sex-chromosome probes. Probes located on chromosomes X and Y are removed to prevent sex-driven confounding effects. Sex-chromosome CpGs exhibit substantially different methylation profiles between males and females, and including them in joint normalisation has been shown to introduce artificial sex-related bias into autosomal probes [53]. Since the datasets analysed here contain samples from both sexes and sex-specific methylation is not the object of investigation, removal therefore represents a conservative strategy to prevent confounding without compromising autosomal inference [26].

The filtering resources are applied under a union criterion: a probe is removed if it appears in at least one curated list. Platform-specific manifests (HumanMethylation450 v1.2 for GSE69914 [54]; MethylationEPIC v1.0 B5 for GSE225845 and

GSE287331 [55]) are used for manifest-based validation, chromosome annotation, and non-CpG probe identification. The number of probes flagged by each resource, and the redundancy of flags across resources, are reported per-dataset in Section 4.3.

4.2.4 Statistical Transformation: β-to-M Values

The final preprocessing step transforms the filtered β-value matrix into M-values prior to statistical modelling. As defined in Equation (3.1), β-values are bounded proportions in [0,1] derived from methylated and unmethylated probe intensities. While β-values are intuitive and directly interpretable as fractional methylation, their bounded support induces strong mean-dependent heteroscedasticity: variance is inflated near intermediate values ($\beta \approx 0.5$) and compressed near the boundaries, violating the homoscedasticity assumptions underlying standard linear modelling frameworks [22]. The logit transformation to M-values,

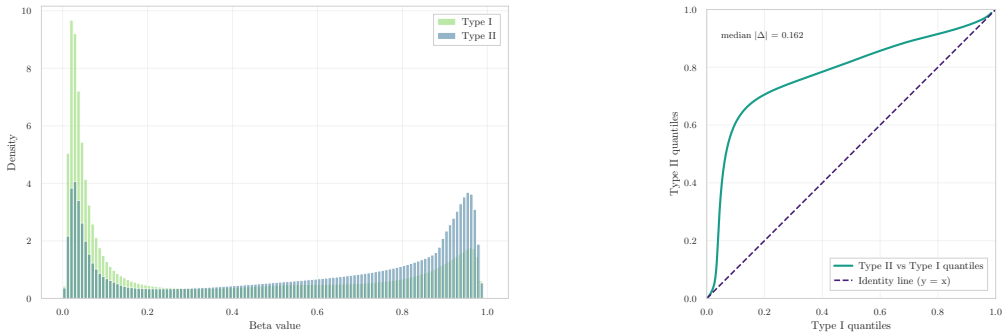
$$M = \log_2\left(\frac{\beta + \varepsilon}{1 - \beta + \varepsilon}\right), \quad \varepsilon = 10^{-6}, \tag{4.1}$$

maps $\beta \in (0,1)$ to an unbounded real-valued scale and substantially stabilises the mean–variance relationship [22]. As shown in their analysis, the relationship between β and M-values is a logit transformation yielding an approximately linear correspondence in the intermediate methylation range and nonlinear compression at the extremes. The small offset ε is introduced to handle boundary values numerically. The resulting M-value distributions are approximately symmetric and exhibit markedly flatter variance profiles across the full methylation range, making them the standard input for differential methylation analysis within linear modelling frameworks and related inferential procedures [41, 42, 56].

It is important to note that M-values are adopted exclusively for statistical modelling: β-values are retained in parallel for biological interpretation, effect-size quantification, and visualisation, following the dual-representation convention recommended by Du et al. [22] and widely adopted in the literature [56, 57]. The transformation is applied to the complete post-filtering matrix for each dataset. Distributional diagnostics confirming the expected variance-stabilising effect are reported in Section 4.3. This choice is consistent with the feature-selection and modelling procedures developed in the subsequent Chapter 5, which rely on variance-based ranking, linear statistics, and stability criteria that assume approximately homoscedastic behaviour.

4.3 Dataset-Specific Implementation

This section documents the dataset-specific application of the preprocessing pipeline described in Section 4.2. For each cohort, the outcome of structural verification,



(a) β-value density by probe type. **(b)** Q–Q plot (Type II vs. Type I).

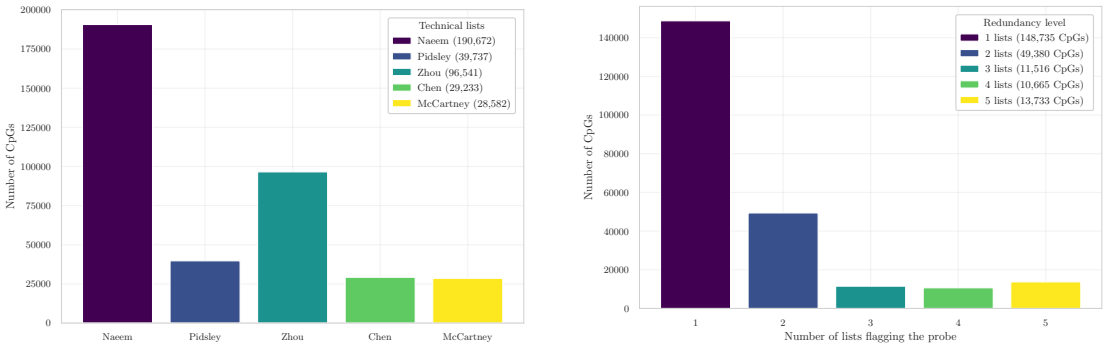
Figure 4.1: Infinium Type I/II bias diagnostics in GSE69914.

probe-type bias diagnostics, technical probe filtering, and β-to-M transformation is reported in detail. Although the general framework is applied consistently across all datasets, certain steps—most notably probe-type bias correction and handling of missing values—require cohort-specific decisions. These are explicitly justified within each subsection. The results are presented separately for GSE69914 (Section 4.3.1), GSE225845 (Section 4.3.2), and GSE287331 (Section 4.3.3).

4.3.1 Dataset GSE69914

Structural Integrity and Cohort Harmonization The dataset comprises 397 samples and 485,512 CpGs at load time. All sample identifiers are unique, and the phenotype table contains complete metadata for the required fields. Alignment between the phenotype annotation and the β-value matrix is exact. The methylation matrix is fully observed, with no missing CpG values and no incomplete samples. No exclusion or imputation was required.

Probe-Type Bias Diagnostics and Correction Raw intensity data were reported to have been processed by the original authors using the `minfi` pipeline with BMIQ normalization [19]. Diagnostic evaluation confirms that probe-type bias is effectively mitigated in the released matrix. The β-value density distributions of Type I and Type II probes exhibit substantial overlap, with no evidence of the characteristic multi-modal distortion observed in uncorrected data (Figure 4.1a). Quantile–quantile analysis further supports this result: the Type II versus Type I quantiles align closely to the identity line, with median absolute deviation median$|\Delta| = 0.162$, indicating balanced probe-type scaling (Figure 4.1b). Together, these diagnostics are consistent with effective BMIQ correction, in agreement with the normalization behaviour described in [23, 26, 47].



(a) CpGs flagged by each technical list. **(b)** Redundancy across filtering lists.

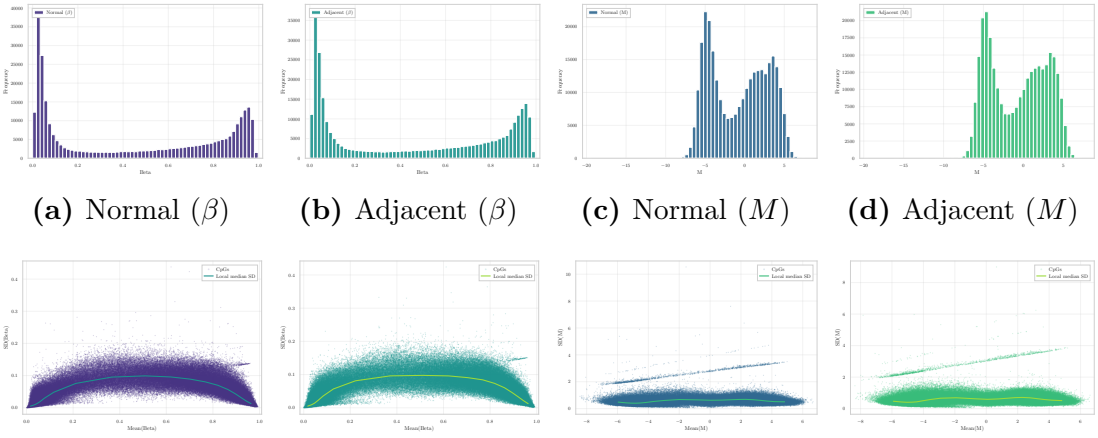
Figure 4.2: Technical filtering summary in GSE69914.

Table 4.1: CpG filtering summary for the GSE69914 dataset.

Initial CpGs	Naeem	Pidsley	Zhou	Chen	McCartney	Non-CpG	X/Y	Final CpGs
485,512	190,672	39,737	96,541	29,233	28,582	875	7,728	251,483

Technical Probe Filtering Intersection with external technical filtering resources resulted in substantial probe removal, as summarised in Table 4.1. The largest individual contributions derive from the Naeem and Zhou annotations, while the remaining resources flag smaller but non-negligible subsets. The redundancy structure across lists is shown in Figure 4.2b. Because these catalogues target partially distinct classes of technical artefacts—such as cross-reactivity, SNP interference, and repeat-associated probes—limited overlap across lists is expected. Most excluded CpGs are supported by a single resource, with progressively fewer loci flagged by multiple independent annotations. This pattern reflects complementary rather than redundant filtering criteria, while the subset of probes flagged by several lists indicates concordance for loci exhibiting multiple technical vulnerabilities. Annotation-based validation using the HumanMethylation450 v1.2 manifest [58] further excluded non-CpG probes and loci lacking reliable genomic annotation. Subsequent removal of sex-chromosome probes (X/Y), consistent with established preprocessing recommendations in mixed-sex cohorts [26, 53], ensured a strictly autosomal working set. A detailed per-probe exclusion summary is available in the project repository at `removed_cpgs_all_filters_summary_gse69914.csv`.

Statistical Transformation: β-to-M Values Following transformation, the empirical distributions confirm the expected variance-stabilising behaviour of M-values (Figure 4.3). In β-space, both Normal and Adjacent tissues exhibit the characteristic bounded, U-shaped distribution with accumulation near 0 and 1.



(a) Normal (β) **(b)** Adjacent (β) **(c)** Normal (M) **(d)** Adjacent (M)

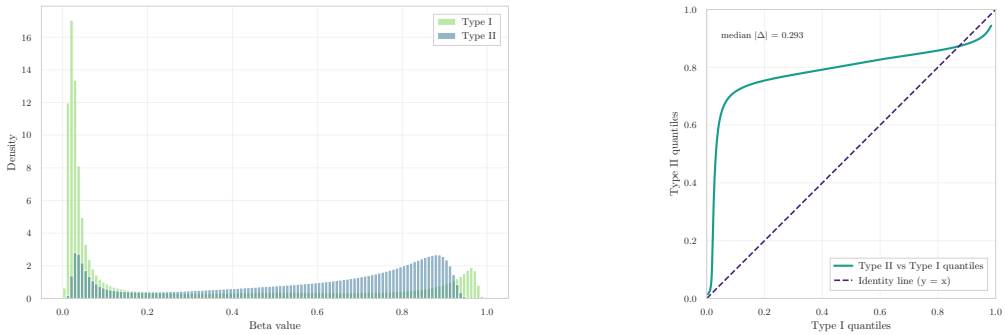
(e) Mean–SD Normal (β) **(f)** Mean–SD Adjacent (β) **(g)** Mean–SD Normal (M) **(h)** Mean–SD Adjacent (M)

Figure 4.3: Distributional comparison of β and M in Normal and Adjacent tissues in GSE69914.

After logit transformation, the corresponding M-value distributions become substantially more symmetric and centered, while preserving the relative structure between tissue groups. The mean–SD relationships further illustrate the scale effect. In β-space, variance is strongly mean-dependent and inflates near intermediate methylation levels, producing a pronounced curvature in the mean–SD plots. In contrast, M-values display markedly flatter dispersion profiles across the full range, indicating substantial reduction of heteroscedasticity. This behaviour is fully consistent with the theoretical and empirical findings of Du et al. [22], which motivate the use of M-values for inferential modelling.

4.3.2 Dataset GSE225845

Structural Integrity and Cohort Harmonization The dataset comprised 477 samples and 750,426 CpGs at load time, consistent with the dimensionality of the Illumina EPIC platform. The phenotype table contained 477 entries, and alignment between phenotype annotation and the β-value matrix was exact. No duplicated entries were detected, and all required metadata fields were present. Probe-level and sample-level missingness assessment confirmed that the methylation matrix was fully observed. No missing CpG values and no incomplete samples were detected. Consequently, no imputation or early NaN filtering was required prior to probe-type bias diagnostics and technical filtering.



(a) β-value density by probe type. **(b)** Q–Q plot (Type II vs. Type I).

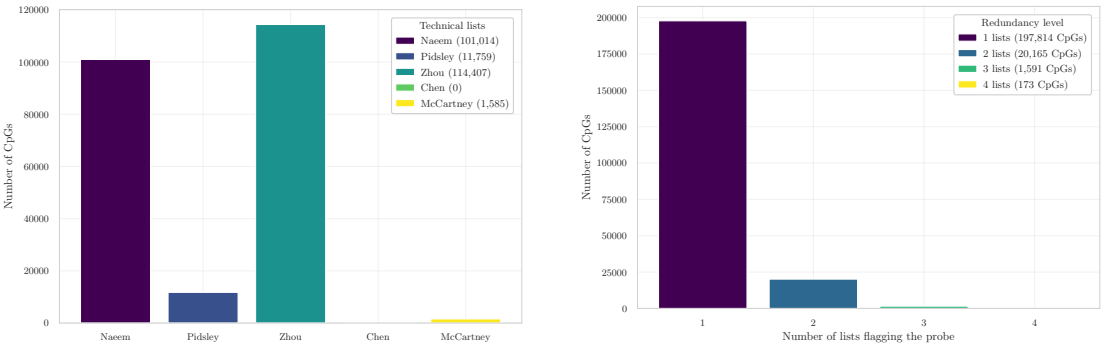
Figure 4.4: Infinium Type I/II bias diagnostics in GSE225845.

Table 4.2: CpG filtering summary for the GSE225845 dataset.

Initial CpGs	Naeem	Pidsley	Zhou	Chen	McCartney	Non-CpG	X/Y	Final CpGs
750,426	101,014	11,759	114,407	0	1,585	773	14,071	530,683

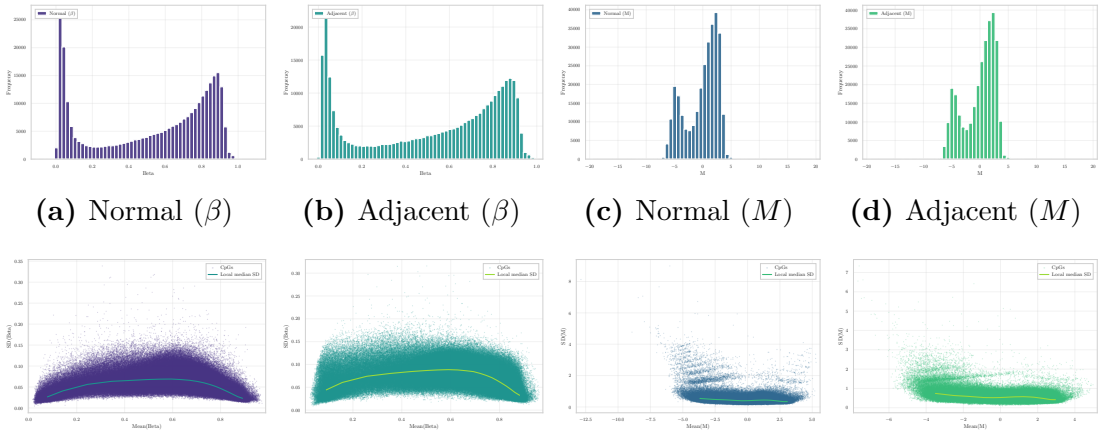
Probe-Type Bias Diagnostics and Correction Stratification by probe design yielded 119,760 Type I and 627,842 Type II probes. Raw intensity data were reported by the original authors to have been processed using the `minfi` pipeline with BMIQ normalization (v1.4) [20]. The β-value density distributions (Figure 4.4a) show substantial overlap between probe families. Type I probes retain the characteristic concentration near the unmethylated and fully methylated boundaries, while Type II probes display a broader distribution; however, no pronounced multimodal distortion or systematic separation is observed. Quantile–quantile analysis (Figure 4.4b) further supports this interpretation. Although minor deviations from the identity line ($y = x$) are visible, the overall alignment indicates largely balanced probe-type scaling. The median absolute deviation median$|\Delta| = 0.293$ suggests that probe-type differences are attenuated but not entirely eliminated, consistent with residual variation commonly observed after BMIQ normalization [23, 26, 47]. Overall, the diagnostic patterns are compatible with prior application of probe-type bias correction.

Technical Probe Filtering Intersection with external technical filtering resources resulted in substantial probe removal, as summarised in Table 4.2. The largest individual contributions derive from the Zhou and Naeem annotations, while the remaining resources flag smaller subsets. The redundancy structure across lists is shown in Figure 4.5b. A detailed per-probe exclusion summary is available in the project repository at `removed_cpgs_all_filters_summary_gse225845.csv`.



(a) CpGs flagged by each technical list. **(b)** Redundancy across filtering lists.

Figure 4.5: Technical filtering summary in GSE225845.



(a) Normal (β) **(b)** Adjacent (β) **(c)** Normal (M) **(d)** Adjacent (M)

(e) Mean–SD Normal (β) **(f)** Mean–SD Adjacent (β) **(g)** Mean–SD Normal (M) **(h)** Mean–SD Adjacent (M)

Figure 4.6: Distributional comparison of β and M in Normal and Adjacent tissues in GSE225845.

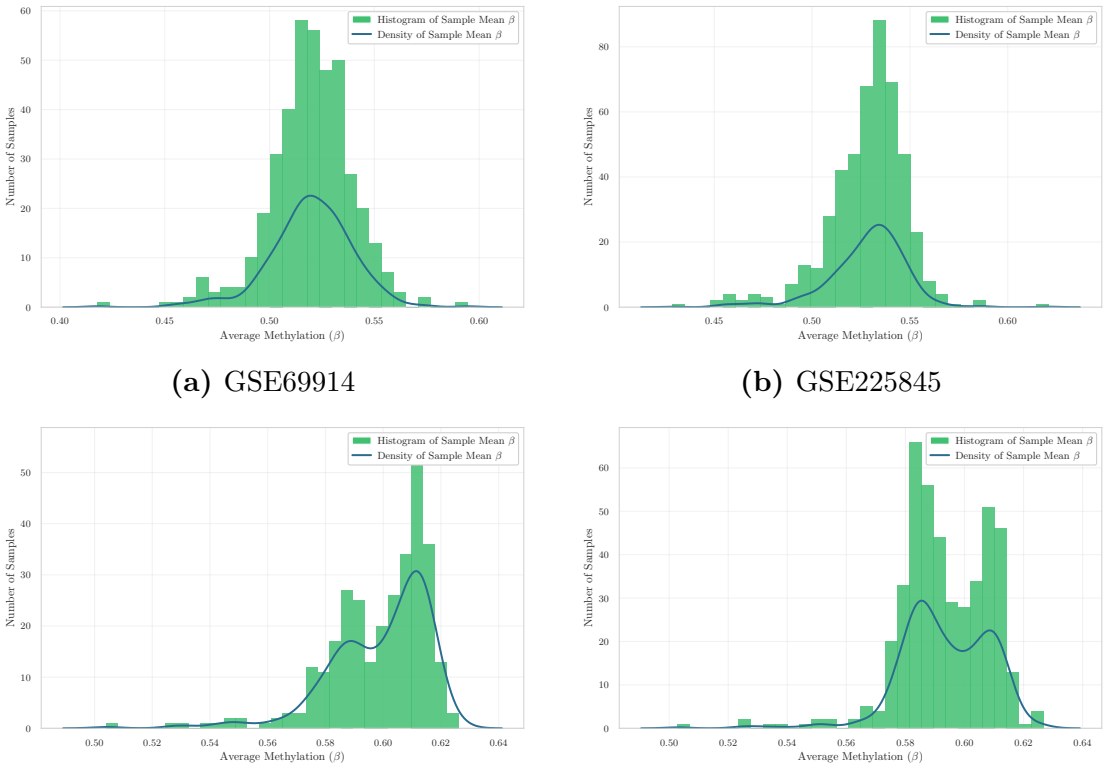
Statistical Transformation: β-to-M Values All filtered β-values were transformed to M-values using the logit relation defined in Equation (4.1). The empirical distributions (Figure 4.6) show the characteristic bounded, U-shaped profile in β-space for both Normal and Adjacent tissues. After transformation, the corresponding M-value distributions become substantially more symmetric and centered, while maintaining the relative structure between tissue classes. The mean–SD relationships further illustrate the scale effect. In β-space, variance exhibits pronounced mean dependence, with inflation at intermediate methylation levels and compression near the boundaries. In contrast, M-values display markedly flatter dispersion profiles across the full range, indicating substantial reduction of heteroscedasticity.

4.3.3 Dataset GSE287331

Structural Integrity and Cohort Harmonization The dataset was imported from its Parquet representation, yielding an initial β-value matrix of 446 samples and 866,552 CpGs. The phenotype table contained the same number of entries, with all required metadata fields present and alignment between phenotype annotation and the β-value matrix exact. Unlike GSE69914 and GSE225845, the methylation matrix was not fully observed. Probe-wise missingness is a well-known source of unreliability in Illumina methylation arrays: Islam *et al.* [45] excluded probes with missing β-values in more than 2% of samples, while Mansell *et al.* [46] applied a stricter 1% threshold for detection failure in their EPIC-array quality control pipeline. Consistent with these recommendations, all CpG sites missing in more than 1% of samples were removed, prioritising high-confidence measurements while minimising reliance on imputed values and leveraging the large dimensionality of the EPIC platform. After applying this criterion, the remaining sparsity was negligible: 58,561 NaN values, corresponding to 0.0187% of all entries. Given this extremely low level of sparsity, median-based imputation is both statistically robust and distribution-preserving; residual missing values were therefore imputed using the probe-wise median β-value across samples, a strategy that avoids introducing group-specific bias and is consistent with recent harmonisation workflows such as mLiftOver, where missing values are explicitly replaced by probe-wise median substitution to preserve biologically plausible levels [59].

After harmonising phenotype labels and applying dataset-specific selection rules, only samples belonging to the three core tissue states of interest—Normal, Adjacent, and Tumour—were retained. The final working matrix consisted of 311 samples and 702,916 CpGs, with no residual missing values.

Probe-Type Bias Diagnostics and Correction Using the official EPIC v1.0 manifest, 115,705 Type I and 585,514 Type II probes were identified. Unlike GSE69914 and GSE225845, for which the original authors reported prior application of BMIQ normalisation, no probe-design bias correction was documented in the processing pipeline of GSE287331. This absence was first suggested by the per-sample mean β-value distributions (Figure 4.8): whereas GSE69914 and GSE225845 display a single broad peak, GSE287331 shows a narrower, right-shifted distribution characterised by two distinct modes in the high-β range—a pattern largely absent in the other cohorts and strongly indicative of uncorrected Type I/II probe-design bias. Stratification of probes by design type confirmed this hypothesis. The β-value distributions of the two chemistries showed a marked mismatch (Figure 4.8a), with Type II probes displaying a strong right-shifted profile. The quantile–quantile comparison confirmed a substantial systematic offset, with a median absolute deviation of 0.380 (Figure 4.8c), fully incompatible with BMIQ-corrected data.



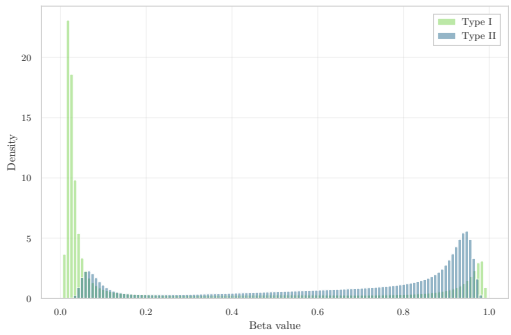
(a) GSE69914 **(b)** GSE225845

(c) GSE287331 pre BMIQ **(d)** GSE287331 post BMIQ

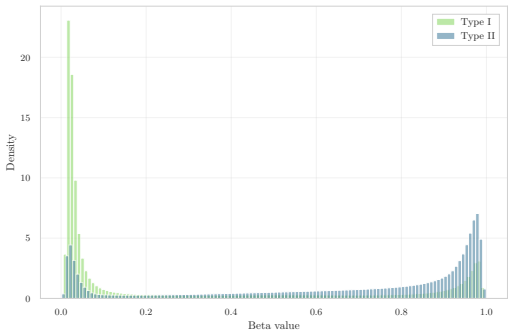
Figure 4.7: Per-sample mean β distributions for the three datasets.

Inspection of the original publication [34] further confirmed the absence of any mention of BMIQ, SWAN, noob, or equivalent probe-type correction methods.

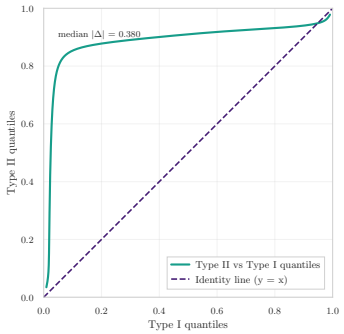
Given this evidence, the canonical BMIQ algorithm [23] was applied using the `wateRmelon` implementation [60], operating strictly on a per-sample basis: Type I probes define the reference mixture distribution within each sample, and Type II probes are rescaled accordingly, with no cross-sample or cross-group information introduced at any stage. Samples with insufficient numbers of probes per design type were left uncorrected to avoid unstable mixture estimation. Post-correction diagnostics showed only a modest reduction in probe-type mismatch ($|\Delta|$: $0.380 \rightarrow 0.370$), with the overall distributional discrepancy remaining substantial and the bimodality of per-sample mean β distributions persisting (Figure 4.8, Figure 4.8b). This limited harmonisation is consistent with documented caveats of quantile-based normalisation on EPIC arrays, which can artificially reduce biological variability, remove genuine group-level signal, or yield distorted distributions when differences are expressed as broad distributional shifts [17, 48, 49]. The impact on biological structure was assessed quantitatively via silhouette analysis: the



(a) GSE287331 pre BMIQ



(b) GSE287331 post BMIQ



(c) Q–Q comparison of Type II vs. Type I probes (pre BMIQ).

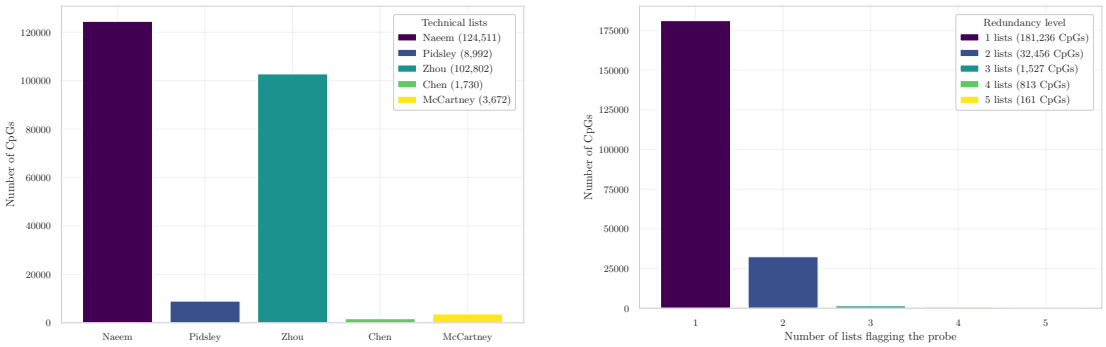
Figure 4.8: Infinium Type I/II bias diagnostics in GSE287331. Pre-BMIQ densities and Q–Q analysis indicate marked probe-type imbalance; post-BMIQ densities show partial alignment.

average silhouette coefficient declined from $\text{sil}_{\text{pre}} = 0.2876$ to $\text{sil}_{\text{post}} = 0.2537$ after correction, indicating that BMIQ reduced the natural clusterability of the data and collapsed genuine group-level separation—particularly between Normal and Adjacent samples. For this reason, all downstream analyses are carried out on the non-BMIQ version of the dataset, which preserves sharper cluster boundaries and avoids normalisation-induced shrinkage of biological signal.

Technical Probe Filtering Intersection with external technical filtering resources resulted in substantial probe removal, as summarised in Table 4.3. The largest individual contributions derive from the Naeem and Zhou annotations, while the remaining resources flag smaller but non-negligible subsets (Figure 4.9a). Importantly, in the unfiltered EPIC matrix the Zhou annotation alone would have excluded 192,856 CpGs. This number substantially exceeds the net contribution observed at this stage, indicating that a considerable fraction of Zhou-flagged

Table 4.3: CpG filtering summary for the GSE287331 dataset.

Initial CpGs	Naeem	Pidsley	Zhou	Chen	McCartney	Non-CpG	X/Y	Final CpGs
866,552	124,511	8,992	102,802	1,730	3,672	500	12,579	486,725

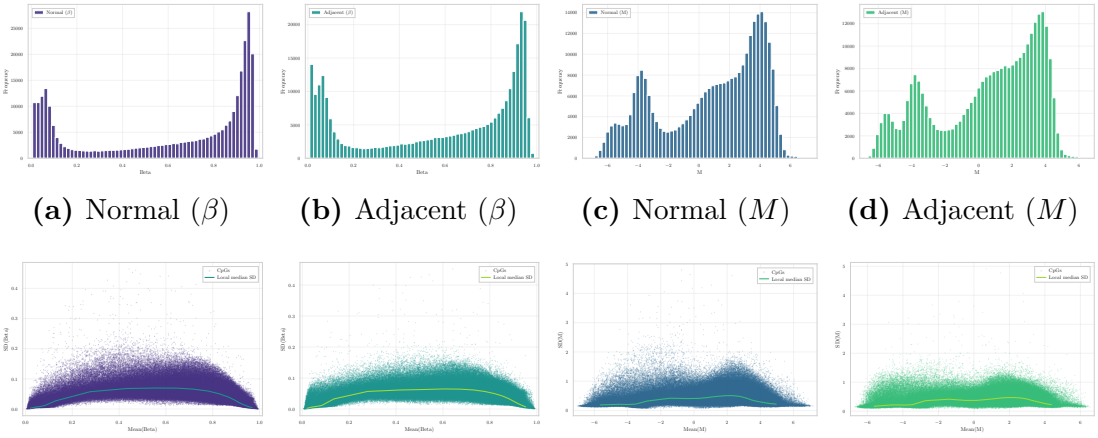


(a) CpGs flagged by each technical list. **(b)** Redundancy across filtering lists.

Figure 4.9: Technical filtering summary in GSE287331.

probes had already been removed during the earlier missingness-cleaning step. The missingness filter and the Zhou reliability masks therefore exhibit non-trivial overlap. This concordance supports the robustness of the preprocessing strategy: probes excluded due to empirical data sparsity largely coincide with loci independently annotated as technically unreliable, suggesting that early NaN-based exclusion was not arbitrary but aligned with established probe-quality criteria. The redundancy structure across lists (Figure 4.9b) further indicates that most excluded CpGs are supported by a single resource, with progressively fewer loci flagged by multiple independent annotations. Annotation-based validation using the MethylationEPIC v1.0 manifest [55] further excluded non-CpG probes and loci lacking reliable genomic annotation. Across all filters, the dimensionality reduction is detailed in Table 4.3. A probe-level exclusion summary is available in the project repository at `removed_cpgs_all_filters_summary_gse287331.csv`.

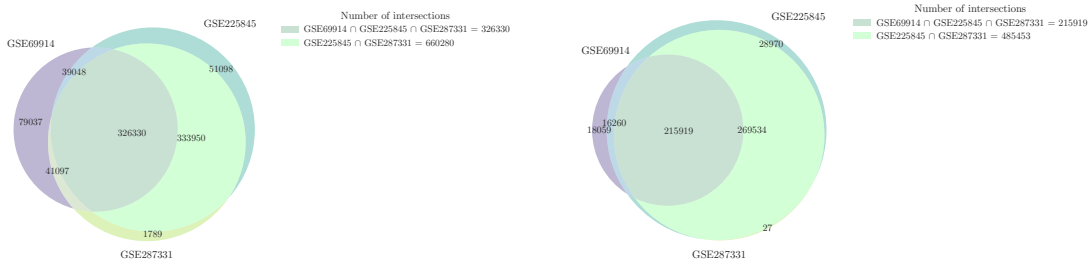
Statistical Transformation: β-to-M Values Following technical filtering, the β-value matrix was transformed into M-values using the logit relation defined in Equation (4.1), in accordance with established recommendations for variance stabilisation [22, 41, 42]. The empirical distributions (Figure 4.10) retain the characteristic bounded structure of β-values; however, in contrast to GSE69914 and GSE225845, the profiles appear less sharply bimodal and more broadly dispersed. Both Normal and Adjacent tissues exhibit reduced concentration at the methylation extremes and a wider intermediate range, consistent with the greater technical



(a) Normal (β) **(b)** Adjacent (β) **(c)** Normal (M) **(d)** Adjacent (M)

(e) Mean–SD (Normal, β) **(f)** Mean–SD (Adjacent, β) **(g)** Mean–SD (Normal, M) **(h)** Mean–SD (Adjacent, M)

Figure 4.10: Distributional comparison of β and M in Normal and Adjacent tissues in GSE287331.



(a) CpG overlap before preprocessing. **(b)** CpG overlap after preprocessing.

Figure 4.11: CpG overlap across datasets before and after preprocessing.

heterogeneity observed in this cohort. After transformation, the corresponding M-value distributions become more symmetric, as expected under the logit mapping, yet remain visibly broader than in the other datasets. The mean–SD relationships confirm that while variance stabilisation is achieved, dispersion remains comparatively elevated across the full methylation range. This behaviour is coherent with the stronger preprocessing burden documented in earlier steps and reflects the intrinsic variability of the released matrix.

Table 4.4: CpG dimensionality before and after preprocessing.

Dataset	Platform	CpGs *before*	CpGs *after*
GSE69914	450K	485,512	251,483
GSE225845	EPIC	750,426	530,683
GSE287331	EPIC	866,554	486,725

4.4 Inter-Dataset Consistency After Preprocessing

After completing the intra-dataset preprocessing for each cohort, a concise inter-dataset analysis is performed to summarise the effects of the pipeline and to motivate the subsequent feature selection step.

Consistency of the Preprocessing Framework All three datasets underwent a harmonised preprocessing workflow, including initial integrity checks, literature-based technical filtering, manifest-driven probe validation, removal of sex-chromosome loci, and transformation from β-values to M-values. While specific implementation details differ slightly depending on platform characteristics and data quality, the conceptual structure of the pipeline is consistent across datasets. Table 4.4 reports, for each dataset, the dimensionality of the methylation matrix before and after preprocessing. Despite substantial differences in initial platform coverage (450K vs. EPIC), technical filtering consistently removes a large fraction of probes, yielding cleaner and more reliable feature spaces.

CpG overlap across datasets CpG overlap was evaluated both before and after preprocessing to characterise the relationship between cohorts. Prior to filtering (Figure 4.11a), the intersection is dominated by probes shared between the two EPIC datasets, with a smaller but substantial core common to all three cohorts, reflecting platform design constraints and baseline annotation differences. After the complete preprocessing pipeline (Figure 4.11b), the total number of CpGs is markedly reduced, yet a large shared core is preserved: the three-way intersection comprises 215,919 CpGs, representing a stable autosomal backbone consistently retained across platforms after technical filtering.

4.5 Post-Preprocessing Feature Space and Dimensionality Implications

The preprocessing pipeline developed in this chapter has systematically addressed the three main classes of technical confounders identified in Chapter 3: probe-design bias, unreliable probe sets, and statistical scale distortion. The resulting matrices are technically clean, fully observed, and expressed on a variance-stabilised M-value scale suitable for inferential modelling. Across all three cohorts, the three-way CpG intersection comprises 215,919 autosomal loci retained after platform-aware filtering, constituting a stable and harmonised feature backbone for cross-dataset analyses. Yet the preprocessing outcome also sharpens the central analytical challenge. The post-processed datasets remain extremely high-dimensional — 251,483 CpGs in GSE69914, 530,683 in GSE225845, and 486,725 in GSE287331 — and the exploratory analyses of Chapter 3 established that the Normal–Adjacent signal is focal and low-amplitude, with mean $|\Delta\beta|$ ranging from 0.009 to 0.031 across cohorts. In this regime, the overwhelming majority of retained CpGs carry no discriminative information for the contrast of interest: including them in downstream modelling would dilute the signal, inflate variance estimates, and compromise both interpretability and generalisation. The limiting factor is therefore no longer data quality but *signal localisation*: identifying the restricted subset of CpG loci that capture biologically meaningful Normal–Adjacent variation, consistently across cohorts and independently of platform-specific noise. This requires a principled feature selection strategy that is sensitive to the subtle, focal nature of the epigenetic drift documented here, robust to the cohort-level heterogeneity observed in the inter-dataset comparisons, and compatible with the homoscedastic M-value representation adopted for modelling.

Chapter 5 addresses this challenge directly, developing and evaluating feature selection methodologies applied separately within each dataset-specific CpG space. In order to preserve the full informational richness of the EPIC platforms, feature selection is not restricted to the three-way CpG intersection but is performed on the complete post-processed feature set of each cohort, thereby avoiding artificial dimensional truncation driven by the lower coverage of the 450K array.

Chapter 5

Robust Feature Selection for Epigenetic Drift Characterisation

5.1 Statistical Challenges of High-Dimensional DNA Methylation Data

The preprocessing framework developed in Chapter 4.5 established a technically harmonised stabilised methylation space. However, preprocessing alone does not address the central statistical challenge of genome-wide DNA methylation data: extreme dimensionality relative to sample size. This challenge is particularly acute in the present study, where the objective is to characterise epigenetic differences along the Normal → Adjacent → Tumour axis in breast cancer — a setting in which biologically meaningful signal is expected to be subtle, spatially structured, and embedded in a feature space of several hundred thousand CpG loci. High-throughput methylation arrays, such as the Illumina EPIC platform, interrogate approximately 850 000 CpG sites per sample [29], while the number of available subjects in any given cohort remains comparatively limited. This high-dimensional, low-sample-size (HDLSS) regime is well characterised in the statistical literature as a setting in which standard estimators become geometrically pathological: as dimensionality grows relative to n, pairwise distances concentrate, covariance estimation becomes unstable, and classifiers overfit to noise [61, 62]. In genomics specifically, the consequences are well documented — unconstrained feature spaces yield inflated classification accuracy that fails to replicate in independent cohorts [7, 63]. Feature selection is therefore not a cosmetic reduction step but a structural necessity for any learning procedure applied to molecular profiling data.

5.1.1 Structural Properties Motivating Feature Selection

Three properties of DNA methylation data collectively motivate explicit dimensionality control.

High Dimensionality and Overfitting Risk In HDLSS settings, discriminative models may achieve near-perfect separation on training data while capturing noise rather than biological signal. The phenomenon has been formally characterised by [61], who showed that in very high dimensions, data from distinct classes become approximately equidistant unless genuine low-dimensional structure is present. Practically, this means that feature selection must precede, not follow, model fitting — and must itself be performed exclusively on training data to avoid optimistic bias [7].

Spatial Correlation and Redundancy CpG loci exhibit strong local correlation driven by genomic proximity, chromatin organisation, and shared regulatory context. This spatial autocorrelation has been well characterised in large-scale mapping studies: contiguous CpGs within co-methylated blocks ("methylation domains") can span hundreds to thousands of base pairs and behave as single epigenetic units [64, 65]. Without explicit redundancy control, learning algorithms may over-represent densely correlated genomic regions while underweighting distributed epigenetic signals. This is especially relevant here, as differentially methylated regions (DMRs) along the Normal–Adjacent axis are expected to be spatially clustered rather than randomly distributed [66].

Biological Sparsity Disease-associated methylation alterations are typically sparse and structured. In breast cancer, systematic analyses have demonstrated that tumour-associated hypermethylation preferentially targets specific regulatory programmes — including Polycomb-repressed developmental genes and oestrogen-regulated loci — while genome-wide methylation changes in histologically normal adjacent tissue are more subtle but nonetheless detectable [3, 67]. This sparsity motivates filtering strategies that prioritise stable, biologically coherent CpG subsets over exhaustive feature retention.

5.1.2 Methodological Challenges in Genomic Feature Selection

Feature selection in genomic data presents challenges that go beyond simple variable ranking.

Information Leakage The most consequential pitfall is information leakage: when feature selection is performed on the full dataset prior to cross-validation, the evaluation procedure has implicitly seen the test labels during variable ranking. [7] demonstrated empirically that this leakage can inflate estimated accuracy by 10–40 percentage points in gene expression data; analogous inflation has been documented for methylation-based classifiers [63]. Strict containment of all selection steps within the training fold is therefore essential and is enforced throughout the present framework.

Significance Versus Stability A second challenge concerns the relationship between statistical significance and effect stability. In large feature spaces, extremely small p-values may arise for negligible or unstable effect sizes — a manifestation of the multiple testing problem that is not fully resolved by correction alone. Conversely, biologically relevant features may show moderate but reproducible signal across subsamples. Stability selection [6] addresses this by quantifying the reproducibility of feature inclusion across repeated subsampling, providing selection frequency as a complementary criterion to statistical rank.

Predictive Optimisation Versus Interpretability A third challenge is the tension between predictive optimisation and interpretability. Aggressive feature selection driven solely by discriminative performance may yield compact but biologically opaque models, while overly conservative selection may retain sufficient redundancy to obscure the underlying regulatory architecture. In translational contexts, this tension has additional implications: in-silico classification accuracy does not automatically imply clinical utility [68], and the biological coherence of the selected feature set is therefore a criterion in its own right.

5.1.3 Design Objectives of the Selection Framework

In light of the statistical and biological considerations discussed above, the feature selection strategy developed in this chapter is designed to satisfy four interconnected objectives. First, strict leakage control ensures that all selection procedures are confined to training data, preventing inflation of evaluation metrics through inadvertent information exposure. Second, stability enforcement guarantees that CpGs are retained only if they demonstrate reproducible discriminative behaviour across repeated stratified resampling, ensuring robustness to training-set perturbations. Third, redundancy reduction through correlation-based pruning limits the over-representation of co-methylated genomic blocks and promotes independent signal components. Fourth, genomic diversification imposes structural constraints to prevent excessive concentration of selected features within specific chromosomes or regulatory contexts. The hyperparameters governing each stage of the pipeline

— including variability thresholds, stability cutoffs, correlation bounds, and diversification budgets — were fixed based on domain knowledge and methodological precedent rather than systematic optimisation. A formal sensitivity analysis lies outside the scope of the present work and is identified as a direction for future investigation in Chapter 8.

The objective is therefore not merely dimensionality reduction, but the identification of a stable, biologically coherent, and generalisable CpG subset capable of characterising the Normal–Adjacent epigenetic transition and supporting cross-dataset validation.

5.2 General Feature Selection Framework

Building upon the statistical considerations outlined above, this section formally describes the feature selection workflow adopted in the present study. The procedure operates on the post-preprocessing methylation matrices defined in Chapter 4 and is structured as a sequence of strictly train-only operations. These steps progressively reduce dimensionality while preserving statistical validity, stability, and biological coherence. From variability screening and scale transformation to stability-based ranking and redundancy control, each component addresses a specific structural property of DNA methylation data, culminating in a diversified CpG subset suitable for downstream cross-dataset evaluation.

5.2.1 Empirical Variability-Based Dimensionality Reduction

The first stage of the feature selection framework consists of an unsupervised dimensionality reduction step aimed at removing CpGs that exhibit negligible variability within the tissue under study. This strategy is grounded in the empirically driven data reduction method proposed by [69], who demonstrated that a substantial fraction of CpGs interrogated by the Illumina 450K array are effectively non-variable within a given tissue context, contributing to the multiple testing burden — by inflating the number of hypotheses tested — without increasing the probability of detecting true differential methylation signals.

Data restriction and structural cleaning Let$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$, where $x_i \in [0,1]^p$ represents the vector of β-values for p CpG loci and $y_i \in \{0,1\}$ encodes the binary Normal–Adjacent label, with $y_i = 0$ denoting histologically normal tissue and $y_i = 1$ denoting tumour-adjacent tissue. Samples belonging to other tissue states (e.g. Tumour) are excluded from $\mathcal{D}$ prior to feature selection. This restriction ensures that the selected CpGs specifically characterise early epigenetic alterations

occurring in histologically normal tissue adjacent to tumour, rather than late-stage tumour-specific methylation changes.

Train–test partition The dataset is partitioned into a training set $\mathcal{D}_{\text{train}}$ and a held-out test set $\mathcal{D}_{\text{test}}$ via stratified random sampling:

$$\mathcal{D} = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{test}}, \quad |\mathcal{D}_{\text{train}}| = 0.80\, n, \quad |\mathcal{D}_{\text{test}}| = 0.20\, n, \tag{5.1}$$

with class proportions preserved in both sets. Crucially, all feature selection statistics — including the variability filter defined below — are estimated exclusively on $\mathcal{D}_{\text{train}}$ and the resulting CpG subset is applied unchanged to $\mathcal{D}_{\text{test}}$. This design prevents information leakage and ensures that dimensionality reduction is statistically independent of evaluation [7].

Definition of the variability statistic For each CpG $j \in \{1, \ldots, p\}$, variability is quantified using the robust inter-quantile range statistic proposed by [69]:

$$r_{\beta,j} = Q_{0.90}(\beta_j \mid \mathcal{D}_{\text{train}}) - Q_{0.10}(\beta_j \mid \mathcal{D}_{\text{train}})\,, \tag{5.2}$$

where $Q_{0.90}$ and $Q_{0.10}$ denote the empirical 90th and 10th percentiles of the β-value distribution computed across all samples in $\mathcal{D}_{\text{train}}$, pooling both tissue classes. Note that $r_{\beta,j}$ measures population-level dispersion unconditional on class membership: it quantifies whether a CpG is variable across individuals at all, not whether it discriminates between tissue types. A CpG is retained if and only if: $r_{\beta,j} \geq \tau,\ \tau = 0.05$, yielding the retained index set $\mathcal{J}_{\text{Edgar}} = \{j \in \{1, \ldots, p\} : r_{\beta,j} \geq \tau\}$. The threshold $\tau = 0.05$ corresponds to the empirical criterion introduced in [69]: CpGs exhibiting a population-level methylation range below 5% were operationally defined as non-variable and excluded. In the original study, this filter removed approximately 20–30% of probes on the 450K array across diverse tissue types [69].

Interpretation and scope Because $r_{\beta,j}$ is computed without reference to class labels, this filtering step is strictly unsupervised: it cannot introduce label-dependent bias and is therefore safe to apply upstream of the stratified cross-validation framework described in Section 5.1.2. Its function is orthogonal to subsequent supervised ranking steps — it defines the candidate space over which discriminative selection will operate, not the final feature set. CpGs excluded by this criterion ($r_{\beta,j} < 0.05$) typically correspond to loci consistently hypo- or hypermethylated across individuals, frequently enriched in CpG islands and constitutively silenced promoter regions [69, 70]. Retaining such loci inflates the effective dimensionality of the feature space without contributing discriminative variance, exacerbating the HDLSS geometry. Low-variance independent filtering is a standard strategy in high-dimensional genomic analyses to reduce noise and improve statistical efficiency

without compromising type I error control [71]. From a statistical perspective, $|\mathcal{J}_{\text{Edgar}}|$ defines the reduced multiple testing space on which all downstream differential methylation analyses and stability selection procedures are conducted. The CpG subset $\mathcal{J}_{\text{Edgar}}$, estimated on $\mathcal{D}_{\text{train}}$, is applied identically to $\mathcal{D}_{\text{test}}$ prior to any evaluation (Equation 5.1), preserving strict train–test separation throughout.

5.2.2 Re-alignment to M-values and Residual Variance Regularisation

The empirical variability screening of Section 5.2.1 is carried out in β-space, preserving continuity with the original inter-quantile dispersion criterion of [69]. This is consistent with the dual-representation strategy of Section 4.2.4: β-values are retained for biological interpretability, while M-values are adopted for statistical modelling because their logit transformation substantially stabilises the mean–variance relationship [22]. Once $\mathcal{J}_{\text{Edgar}}$ has been determined exclusively on $\mathcal{D}_{\text{train}}$, the analysis transitions to the M-value representation defined in Equation (4.1). The transformation is applied deterministically to the post-Edgar β matrices of both training and test sets: no model parameters are estimated at this stage. For each retained CpG $j \in \mathcal{J}_{\text{Edgar}}$, using the same numerical stabilisation constant introduced during preprocessing. The CpG subset selected in β-space is transferred unchanged to M-space, preserving strict train–test separation throughout.

Residual low-variance trimming (train-only) Although the Edgar filter removes CpGs with negligible population-level dispersion in β-space, a residual fraction of near-constant loci may survive after logit transformation. This occurs because the logit map is nonlinear: loci with β-values concentrated near the boundaries of [0,1] can exhibit low inter-quantile range yet produce near-constant M-values once mapped to the real line. Retaining such loci inflates the effective dimensionality of the covariance structure without contributing discriminative signal, which is particularly detrimental in HDLSS settings. To address this, the pipeline systematically discards the lowest α-quantile of CpGs ranked by their empirical standard deviation in M-space on $\mathcal{D}_{\text{train}}$. Letting s_j denote this standard deviation, the retained index set is:

$$\mathcal{J}_{\text{var}} = \{j \in \mathcal{J}_{\text{Edgar}} : s_j \geq Q_\alpha(s)\}, \qquad \alpha = 0.02, \tag{5.3}$$

corresponding to the removal of the bottom 2% of CpGs by dispersion. The step is unsupervised — class labels play no role — and introduces no information leakage. The quantile threshold is estimated solely on $\mathcal{D}_{\text{train}}$; the resulting binary mask is then applied unchanged to $\mathcal{D}_{\text{test}}$, maintaining complete statistical independence of the evaluation set. The resulting set $\mathcal{J}_{\text{var}}$ constitutes the final candidate feature space entering the supervised discriminative ranking stage described in Section 5.1.2.

5.2.3 Biologically Weighted Stability Selection and Region-Level Consolidation

After structural dimensionality control (Sections 5.2.1 and 5.2.2), the feature space remains in the order of several hundred thousand CpGs. In HDLSS regimes, univariate ranking alone is insufficient: small perturbations in training composition can induce large instability in feature ordering, particularly when signal-to-noise ratios are moderate [61, 62]. To address this, we adopt a resampling framework that integrates statistical evidence, directional consistency, and biologically informed weighting across four sequential components.

Repeated Stratified Stability Framework Let $X \in \mathbb{R}^{n_{\text{train}} \times p}$ denote the post-filtered M-value matrix and $y \in \{0,1\}^{n_{\text{train}}}$ the binary Normal–Adjacent label. A Repeated Stratified K-Fold (RSKF) procedure is employed with $K = 5$ folds and $R = 10$ repetitions, yielding $S = 50$ train/validation splits. Repeated stratified resampling reduces the variance of estimated feature importance and mitigates selection instability arising from dependence on a single partition [7, 63]. For each split $s = 1, \ldots, S$ and each CpG j, three quantities are computed exclusively on the training fold.

1. The signed mean difference $\Delta M_j^{(s)} = \bar{M}_{1,j}^{(s)} - \bar{M}_{0,j}^{(s)}$ is computed for all CpGs as an effect-size measure.

2. To reduce computational burden in the HDLSS regime, Mann–Whitney U p-values are evaluated only for the top 50,000 CpGs ranked by $|\Delta M_j^{(s)}|$ within the training fold; CpGs outside this subset are assigned $p_j^{(s)} = 1$. This two-stage procedure introduces a structural dependency between prescreening and rank aggregation, since excluded CpGs receive a conservative rank by construction rather than by statistical evidence. The approximation is justified by the width of the prefilter: retaining 50,000 candidates corresponds to approximately 6% of the full feature space, substantially exceeding the final selection target of 5,000 CpGs and limiting the probability of excluding genuinely discriminative loci. The Mann–Whitney test is adopted as a rank-based nonparametric procedure to avoid distributional assumptions and ensure robustness to residual deviations from normality in M-space [72].

3. The same signed difference is computed independently on the validation fold to assess directional reproducibility across the partition.

Two rank statistics are then derived per split, $r_{p,j}^{(s)}$ (ascending rank of $p_j^{(s)}$), $r_{|\Delta M|,j}^{(s)}$

(descending rank of $|\Delta M_j^{(s)}|$), and averaged across splits:

$$\overline{r}_{p,j} = \frac{1}{S}\sum_{s=1}^{S} r_{p,j}^{(s)}, \qquad \overline{r}_{\Delta M,j} = \frac{1}{S}\sum_{s=1}^{S} r_{|\Delta M|,j}^{(s)}. \tag{5.4}$$

The combined statistical score is the equal-weight convex combination:

$$\text{score}_{\text{stat},j} = 0.5\,\overline{r}_{p,j} + 0.5\,\overline{r}_{\Delta M,j}. \tag{5.5}$$

This rank aggregation follows a Borda-count-like scheme [73], in which equal-weight combination of complementary ranking criteria has been shown to reduce selection variance relative to either criterion alone [74]. The resulting statistical score is subsequently min–max normalised to the unit interval:

$$\text{score}_{\text{stat},j}^{\text{norm}} = \frac{\text{score}_{\text{stat},j} - \min(\text{score}_{\text{stat}})}{\max(\text{score}_{\text{stat}}) - \min(\text{score}_{\text{stat}})}.$$

Relying solely on p-values risks prioritising negligible but stable effects under large sample sizes, whereas magnitude-only ranking may overweight noisy large deviations. Combining both criteria mitigates this trade-off and is consistent with resampling-based stability principles in high-dimensional feature selection [6].

Directional Stability Constraint A CpG whose estimated effect direction reverses across resampling splits is unlikely to reflect a genuine, reproducible biological signal. We therefore impose a directional reproducibility criterion. For each CpG j, let $\text{sign}_{\text{train}}^{(s)}$ and $\text{sign}_{\text{val}}^{(s)}$ denote the signs of $\Delta M_j^{(s)}$ on the training and validation folds of split s, respectively. The directional stability score is:

$$\text{stab}_j = \frac{\#\left\{s : \text{sign}_{\text{train}}^{(s)} = \text{sign}_{\text{val}}^{(s)} \neq 0\right\}}{\#\left\{s : \text{sign}_{\text{train}}^{(s)} \neq 0 \text{ and } \text{sign}_{\text{val}}^{(s)} \neq 0\right\}}. \tag{5.6}$$

The directional stability score is computed as the proportion of resampling splits in which both training and validation folds exhibit non-zero effect direction and agree in sign. CpGs for which the denominator equals zero — i.e. all splits yield $\Delta M_j^{(s)} = 0$ on at least one fold — are assigned $\text{stab}_j = 0$ by convention, rendering them ineligible for selection under the stability constraint. Only CpGs with $\text{stab}_j \geq 0.75$ are retained, requiring directional agreement in at least 75% of the splits for which a non-zero effect direction is observed in both folds. The threshold of 0.75 is chosen in analogy with the selection frequency cutoff recommended by [6], where a 75% inclusion rate across subsampling splits is identified as a conservative criterion for stable feature identification. The same numerical threshold is here applied to directional consistency rather than selection frequency, extending the stability principle from feature inclusion to effect reproducibility — a particularly relevant criterion when the target signal is moderate in magnitude, as expected for early epigenetic drift in histologically normal adjacent tissue.

Adaptive Effect-Size Filtering A global effect estimate is computed on the full training set, $\Delta M_j^{\text{full}} = \bar{M}_{1,j} - \bar{M}_{0,j}$, and a CpG is retained only if $|\Delta M_j^{\text{full}}| \geq Q_{0.60}(|\Delta M^{\text{full}}|)$. Rather than applying a fixed ΔM cutoff — which would be sensitive to inter-cohort differences in dispersion — this adaptive percentile-based criterion maintains scale invariance across datasets. A fallback to the 50th percentile is applied when the resulting candidate pool would otherwise be insufficient for downstream selection. Together with the directional stability constraint, this filter ensures that only CpGs with both reproducible direction and non-negligible magnitude reach the biological weighting stage. The fallback to the 50th percentile is triggered whenever the resulting candidate pool contains fewer than 15,000 CpGs, ensuring sufficient capacity for the downstream consolidation stage.

Integration of Biological Priors Purely statistical ranking may underweight loci in regulatory regions known to mediate transcriptional control. Promoter-associated CpGs and CpG islands are well-established sites of regulatory methylation changes in cancer and early field effects [64, 67]. Each CpG is assigned a discrete biological prior weight $w_j \in \{1.00, 0.70, 0.50, 0.20\}$ according to its island and promoter annotation:

$$w_j = \begin{cases} 1.00 & \text{Island + TSS} \\ 0.70 & \text{Island only} \\ 0.50 & \text{Promoter (TSS) only} \\ 0.20 & \text{Other loci} \end{cases} \tag{5.7}$$

The discrete weights are subsequently min–max normalised to the unit interval, yielding $w_j^{\text{norm}} \in [0,1]$. In order to prioritise biologically annotated loci while preserving a lower-is-better scoring convention, the biological contribution is defined as $1 - w_j^{\text{norm}}$. After min–max normalisation of the statistical score $\text{score}_{\text{stat},j}$, the final integrated score is:

$$\text{score}_j = (1-\lambda)\,\text{score}_{\text{stat},j}^{\text{norm}} + \lambda\,(1 - w_j^{\text{norm}}), \qquad \lambda = 0.20. \tag{5.8}$$

Min–max normalisation is applied independently to each component prior to combination, ensuring commensurability of scale. The procedure is sensitive to extreme values; however, given the large number of CpGs ($|\mathcal{J}_{\text{var}}| \gg 10^4$), the influence of individual outliers on the normalised distribution is expected to be negligible. The coefficient $\lambda = 0.20$ reflects a deliberate design choice rather than an empirically optimised parameter: statistical evidence contributes 80% of the integrated score, while the remaining 20% provides modest upweighting for loci with established regulatory function in promoter methylation and transcriptional silencing [14]. The top 15,000 CpGs satisfying both the stability and adaptive effect-size constraints define the candidate feature pool passed to the next stage.

Region-Level Anchoring on CpG Islands CpGs are not independent genomic entities: spatially proximal loci within co-methylated blocks co-vary as coordinated regulatory units [64]. Selecting isolated top-ranked probes therefore risks capturing stochastic single-probe fluctuations rather than coherent epigenetic events, which is particularly problematic when the signal of interest — early field-effect drift — is expected to manifest as concerted changes across regulatory domains. To enforce spatial coherence, candidate CpGs are grouped by annotated CpG island. For each island containing at least two candidate CpGs, a region-level score is defined as:

$$\text{region\_score} = \text{median}(\text{score}_j) \times \text{fraction}_{\text{stable}}, \tag{5.9}$$

where $\text{fraction}_{\text{stable}}$ denotes the proportion of CpGs within the island satisfying both $\text{stab}_j \geq 0.75$ and $|\Delta M_j^{\text{full}}| \geq \tau$, with τ denoting the adaptive effect-size threshold selected in the previous step (60th percentile or 50th percentile fallback). The multiplicative formulation jointly penalises islands in which either the median discriminative score is weak or the proportion of internally stable CpGs is low: an island driven by isolated high-scoring probes surrounded by unstable loci receives a substantially discounted region score, ensuring that anchored islands exhibit both statistical quality and internal coherence. Only islands with $\text{fraction}_{\text{stable}} \geq 0.60$ and at least two candidate CpGs are retained. From each such island, the top $m = 2$–3 CpGs are selected into an anchored pool; the remaining candidates form a complementary non-anchored set. This hierarchical consolidation reduces susceptibility to single-probe artefacts and aligns feature selection with the biological expectation that early epigenetic drift occurs in clustered regulatory domains rather than as isolated stochastic events.

5.2.4 Correlation-Based Redundancy Pruning and Graph Theoretic Clustering

Despite stability filtering and region-level anchoring, the candidate feature space retains strong local correlation. DNA methylation levels exhibit pronounced spatial autocorrelation driven by genomic proximity, shared chromatin context, and coordinated regulatory control [64, 65]: in high-density arrays such as EPIC, neighbouring CpGs within islands and promoter regions frequently co-vary as near-redundant blocks. Retaining multiple highly correlated loci inflates the effective dimensionality of the design matrix without increasing independent information content, inducing multicollinearity, unstable coefficient estimates, and variance inflation in downstream classifiers [62, 61]. Explicit correlation-based redundancy pruning is therefore applied to the candidate pool before passing it to the diversification stage.

Correlation Structure on Training Data Let $Z \in \mathbb{R}^{n_{\text{train}} \times q}$ denote the M-value matrix restricted to the q candidate CpGs, with each column standardised

using training-set mean μ_j and standard deviation s_j:

$$Z_{ij} = \frac{M_{ij} - \mu_j}{s_j}. \tag{5.10}$$

where s_j is the sample standard deviation computed with Bessel correction (ddof = 1) on the training set. A small numerical constant is added in practice to prevent division by zero in degenerate cases. The empirical Pearson correlation matrix is then:

$$C = \frac{1}{n_{\text{train}} - 1} Z^\top Z. \tag{5.11}$$

Because columns of Z are standardised to zero mean and unit sample variance, Equation (5.11) coincides with the empirical Pearson correlation matrix estimated exclusively on $\mathcal{D}_{\text{train}}$. All quantities are estimated exclusively on $\mathcal{D}_{\text{train}}$ to preserve strict evaluation independence.

Graph Construction and Connected Components An undirected graph $G = (V, E)$ is constructed with one vertex per candidate CpG; an edge connects j and k whenever their absolute correlation exceeds the threshold:

$$(j, k) \in E \iff |C_{jk}| \geq \tau_c, \qquad \tau_c = 0.85, \tag{5.12}$$

consistent with strong-correlation redundancy thresholds commonly adopted in high-dimensional omics analyses [75, 76]. By convention, each vertex is connected to itself in the adjacency matrix; this does not alter the connected component structure but simplifies implementation. Clusters are defined as the connected components of G: a cluster $\mathcal{C}$ is a maximal vertex subset in which every pair (j, k) is joined by at least one path in G. This graph-theoretic formulation is strictly preferable to pairwise sequential pruning because correlation propagates transitively: two CpGs with $|C_{jk}| < \tau_c$ may both be strongly correlated to a third, and would therefore represent the same co-methylated block without being linked directly. Connected components capture this transitive redundancy structure in a single, consistent pass.

Representative Selection per Cluster For each connected component $\mathcal{C}_m$, a single representative CpG is retained:

$$j^* = \arg\min_{j \in \mathcal{C}_m} \text{score}_j, \tag{5.13}$$

where score_j is the integrated biologically weighted stability score of Section 5.2.3, in which a lower value denotes stronger statistical and biological evidence. CpGs

lacking a valid stability score are assigned an infinite value and are therefore ineligible for selection as cluster representatives. Selecting the minimum thus retains the most discriminative and stable CpG within each correlated block, discarding all redundant co-varying loci. While region-level anchoring leverages genomic annotation to enforce spatial coherence based on CpG island structure, the present clustering step operates purely on empirical covariance structure estimated from the training data.

Separate Clustering of Anchored and Non-Anchored Pools To preserve the island-level spatial structure established by region-level anchoring before applying global redundancy reduction, clustering is performed separately on the region-anchored pool A (CpGs selected via island consolidation) and the non-anchored complementary pool B. Let $\mathcal{R}_A$ and $\mathcal{R}_B$ denote the representative sets obtained from each pool. Representatives from A are prioritised in the final union, and the redundancy-pruned feature set is: $\mathcal{R} = \mathcal{R}_A \cup \mathcal{R}_B$.

Statistical and Biological Rationale From a bias–variance standpoint, removing highly correlated predictors reduces estimator variance while preserving most of the predictive signal, since strongly correlated CpGs encode largely overlapping epigenetic information. In HDLSS settings this is particularly consequential: redundant features exacerbate the ill-conditioning of the design matrix, inflate effective model complexity, and destabilise decision boundaries [62]. Removing redundant predictors reduces effective dimensionality and improves the numerical conditioning of the design matrix without materially reducing predictive information content. The result is a structurally compact, non-redundant feature set that retains statistical strength, directional stability, biological coherence, and spatial structure — forming the input to the genomic diversification stage described in Section 5.2.5.

5.2.5 Genomic Diversification via Constrained Greedy Selection

After correlation-based redundancy pruning (Section 5.2.4), the feature set $\mathcal{R}$ remains enriched for highly ranked CpGs but may still exhibit structural imbalance. Genome-wide methylation arrays are inherently non-uniform — CpG density varies substantially across chromosomes, islands, shores, shelves, and open sea regions — so unconstrained greedy ranking tends to over-concentrate features in CpG-dense promoter islands, high-signal chromosomes, and local windows of coordinated drift. Such concentration reduces effective genomic coverage and increases the risk that downstream models capture locus-specific rather than system-level epigenetic patterns, undermining structural generalisability across cohorts. A diversification step is therefore introduced as a constrained subset selection problem.

Problem Formulation Let $\mathcal{R} = \{j_1, \ldots, j_m\}$ denote the redundancy-pruned CpG pool ordered by increasing integrated score score_j (Section 5.2.3). CpGs lacking valid manifest annotation for chromosome and genomic position are excluded prior to diversification, as spatial constraints require coordinate information. The goal is to select a subset $\mathcal{S} \subset \mathcal{R}$, $|\mathcal{S}| = K = 5000$, that maximises overall discriminative quality under structural diversity constraints. In practice, the effective target is defined as $K_{\text{eff}} = \min(K, |\mathcal{R}|)$ to accommodate cases in which fewer than K mapped CpGs are available; all constraint budgets are computed with respect to K_{eff}. The target cardinality $K = 5000$ represents approximately 1.0% of the probes interrogated by the Illumina 450K array and approximately 0.6% of those interrogated by the EPIC array. It is selected to provide sufficient genomic coverage for cross-cohort evaluation while maintaining a feature-to-sample ratio compatible with stable supervised learning in the available cohorts. The choice is acknowledged as a design parameter rather than an optimised quantity; its sensitivity is identified as a direction for future investigation in Chapter 8. Formally:

$$\begin{aligned} \min_{\mathcal{S} \subseteq \mathcal{R}} \quad & \sum_{j \in \mathcal{S}} \text{score}_j \\ \text{s.t.} \quad & |\mathcal{S}| = K_{\text{eff}}, \\ & |\{j \in \mathcal{S} : c(j) = \chi\}| \leq \lfloor \alpha_{\text{chr}} K_{\text{eff}} \rfloor \quad \forall \chi, \\ & |\{j \in \mathcal{S} : \gamma(j) = g\}| \leq \lfloor \alpha_{\text{ctx}} K_{\text{eff}} \rfloor \quad \forall g, \\ & |\{j \in \mathcal{S} : w(j) = \omega\}| \leq B_\omega \quad \forall \omega,\ B_\omega = 15. \end{aligned} \tag{5.14}$$

Each constraint targets a distinct axis of potential structural imbalance and is described in turn below.

Chromosomal Balance Constraint Let $c(j)$ denote the chromosome of CpG j. No chromosome may contribute more than a fixed fraction of the final set:

$$|\{j \in \mathcal{S} : c(j) = \chi\}| \leq \alpha_{\text{chr}} K, \qquad \alpha_{\text{chr}} = 0.08. \tag{5.15}$$

The budget per chromosome is computed as $\lfloor \alpha_{\text{chr}} K_{\text{eff}} \rfloor$, with a minimum of one CpG allowed per chromosome. The 8% cap prevents over-representation of chromosomes that harbour dense signal regions, a configuration known to affect classifier stability when disease-associated loci cluster in specific chromosomal domains [75].

Genomic Context Constraint Let $\gamma(j)$ denote the CpG island context of locus j (Island, Shore, Shelf, or OpenSea). The fraction of any single context class is bounded by:

$$|\{j \in \mathcal{S} : \gamma(j) = g\}| \leq \alpha_{\text{ctx}} K, \qquad \alpha_{\text{ctx}} = 0.65. \tag{5.16}$$

The context budget is computed as $\lfloor \alpha_{\text{ctx}} K_{\text{eff}} \rfloor$, with at least one CpG permitted per context class. Although promoter islands are biologically important, cancer-related methylation changes are not confined to islands and frequently involve shores and open sea regions [64, 67]. The soft cap ensures that non-island regulatory regions retain representation in the final panel.

Spatial Window Constraint To prevent local hyper-concentration, each chromosome is partitioned into non-overlapping windows of $W = 500\,\text{kb}$, and the number of selected CpGs per window is hard-capped: $|\{j \in \mathcal{S} : w(j) = \omega\}| \leq 15$, where $w(j)$ is the window index of CpG j. Formally, windows are defined via integer binning $w(j) = \lfloor \text{pos}(j)/W \rfloor$ within each chromosome. Since methylation domains frequently span tens to hundreds of kilobases [64], this constraint preserves regional representation while preventing dominance by a single hyper-variable block — a risk that correlation pruning alone cannot fully eliminate at the macro-scale.

Greedy Approximation with Progressive Relaxation The constrained optimisation problem (5.14) is combinatorial and NP-hard in general, as it resembles cardinality-constrained subset selection with multiple linear constraints. Rather than solving a full mixed-integer program, we adopt a deterministic greedy strategy: CpGs are processed in order of increasing score_j, ties are resolved deterministically by stable ordering, and each candidate is added to $\mathcal{S}$ if and only if all active constraints are simultaneously satisfied, until $|\mathcal{S}| = K$. If the strict constraint regime fails to reach K, constraints are relaxed by staged deactivation (rather than by increasing thresholds), first removing the spatial window constraint, then the genomic context constraint, and finally the chromosome balance constraint. At each stage, the specified constraint is fully deactivated while previously relaxed constraints remain inactive. Spatial over-concentration is relaxed first because local genomic crowding is most detrimental to coverage; chromosomal imbalance is relaxed last as it has the weakest direct effect on independent information content. The constrained greedy strategy does not admit formal approximation guarantees in this setting, as the objective is linear rather than submodular and the constraint structure involves multiple simultaneous bounds. The approach is adopted as a computationally tractable heuristic; its empirical adequacy is assessed indirectly through the constraint activation frequency reported in Section 5.3. The frequency with which progressive constraint relaxation is activated in practice is reported in Section 5.3 for each dataset, allowing empirical assessment of whether the final selection approximates the fully constrained solution or degrades toward unconstrained score-based ranking.

5.3 Dataset-Specific Implementation

This section documents the dataset-specific application of the feature selection framework described in Section 5.2. For each cohort, the outcome of empirical variability screening, M-value realignment and residual variance trimming, stability-based discriminative ranking with directional consistency constraints, redundancy reduction via correlation-graph clustering, and constrained genomic diversification is reported in detail.

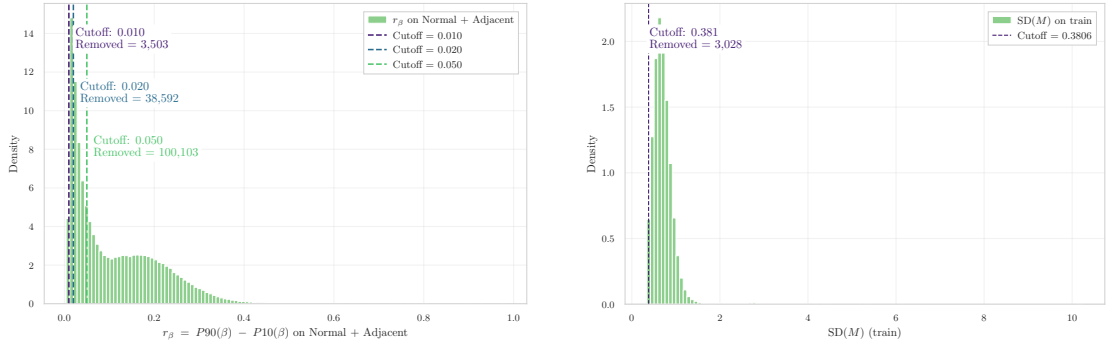
The results are presented separately for GSE69914 (Section 5.3.1), GSE225845 (Section 5.3.2), and GSE287331 (Section 5.3.3).

5.3.1 Dataset GSE69914

The feature selection framework described in Section 5.2 was applied to the pre-processed GSE69914 cohort, restricted to the Normal and Adjacent tissue classes, yielding a final panel of $K = 5000$ CpGs after variability filtering, stability ranking, redundancy pruning, and diversification. The selected loci are available in the project repository at `final_5000_gse69914.csv`.

Empirical Variability-Based Dimensionality Reduction Variability was quantified via the inter-quantile range $r_{\beta,j} = P90(\beta) - P10(\beta)$ computed on the training set. Figure 5.1(a) shows the empirical distribution of r_β across CpGs. The distribution is strongly right-skewed, with a substantial mass concentrated below $r_\beta = 0.05$. Applying the predefined threshold $\tau = 0.05$ removed 100,103 low-variability loci, defining the reduced candidate space for downstream supervised ranking. Lower candidate thresholds (0.01, 0.02) are shown for reference but were not adopted, as they would retain a substantially larger fraction of near-constant probes. To assess whether this filtering step induced structural genomic bias, the distribution of invariant loci across gene-feature categories was examined via fold-change enrichment analysis. As reported in Appendix B, invariant CpGs showed moderate enrichment in promoter-proximal regions (TSS200, first exon, 5'UTR), whereas gene bodies and 3'UTR regions were relatively under-represented. No extreme over-representation was observed, indicating that the variability-based reduction primarily eliminates low-dispersion loci without introducing pathological distortion of the genomic feature space.

Re-alignment to M-values and Residual Variance Regularisation The retained CpGs were deterministically transformed to M-values, after which a residual low-variance trimming was applied in M-space. Figure 5.1b displays the distribution of empirical standard deviations $SD(M)$ on the training set. The distribution is unimodal with limited heavy-tail behaviour. Removal of the bottom



(a) $r_\beta = P90(\beta) - P10(\beta)$ on the training set. Candidate thresholds (0.01, 0.02, 0.05) are indicated.

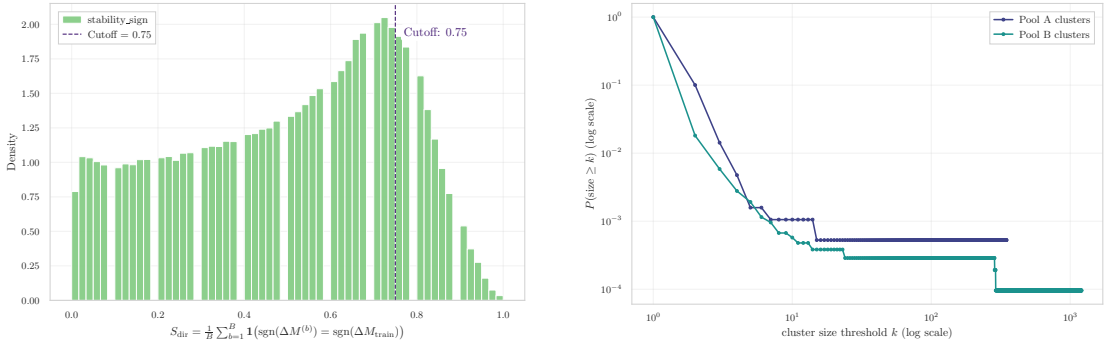
(b) Distribution of $SD(M)$ on the training set. The dashed line indicates the bottom 2% quantile used for residual variance trimming.

Figure 5.1: Variability-based filtering in β-space and residual variance trimming in M-space in GSE69914.

2% of CpGs by dispersion eliminated 3,028 loci. Although quantitatively modest, this step improves covariance conditioning in the HDLSS regime.

Biologically Weighted Stability Selection and Region-Level Consolidation Repeated stratified resampling was used to estimate stability of signed mean differences. Figure 5.2a shows the distribution of the directional stability score S_{dir}. The majority of CpGs exhibit high directional reproducibility, with substantial mass above 0.75. The threshold $\text{stab}_j \geq 0.75$ excludes loci whose effect direction reverses across resampling splits, enforcing robustness consistent with the stability-selection principle of [6]. After stability and adaptive effect-size filtering (60th percentile, corresponding to $|\Delta M| \approx 0.15$), the top 15,000 CpGs were retained by design, defining the candidate pool passed to correlation-based redundancy pruning.

Correlation-Based Redundancy Pruning and Graph-Theoretic Clustering After stability and effect-size filtering, correlation-based graph clustering was applied separately to the anchored and non-anchored pools using a threshold of $|r| \geq 0.85$ computed on the training set. Figure 5.2b shows the cluster size survival curves, whose rapid decay indicates that most connected components are of size 1–3, reflecting limited local redundancy. Only a small number of large correlated blocks are observed, corresponding to densely co-methylated regions. Selecting a single representative CpG per component reduced the candidate space to 12,313 loci, defining the redundancy-pruned set passed to the diversification stage.



(a) Distribution of directional stability scores S_{dir}. The dashed line indicates the cutoff $\text{stab}_j \geq 0.75$.

(b) Cluster size survival curves for anchored (Pool A) and non-anchored (Pool B) CpGs under $|r| \geq 0.85$.

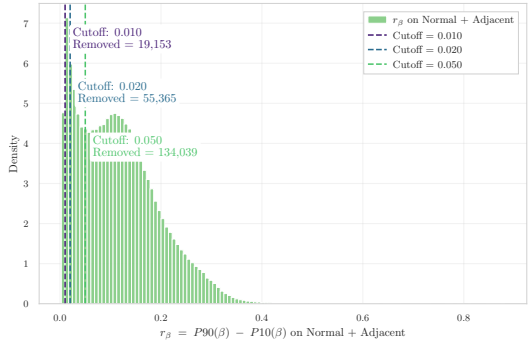
Figure 5.2: Stability enforcement and correlation-based redundancy structure in GSE69914.

Genomic Diversification via Constrained Greedy Selection The constrained greedy diversification procedure selected $K = 5000$ CpGs under chromosome, genomic context, and 500kb spatial-window constraints. The strict constraint regime was sufficient to reach K without requiring relaxation (maximum chromosome share 8%, maximum 15 CpGs per 500kb window). The final panel comprises 2,151 Island loci, 1,342 OpenSea loci, 712 N_Shore loci, 506 S_Shore loci, and 150 N_Shelf loci, indicating a diversified genomic-context composition rather than promoter-restricted enrichment. OpenSea loci were not excluded a priori, as distal regulatory regions may capture long-range epigenetic alterations relevant to early field effects.

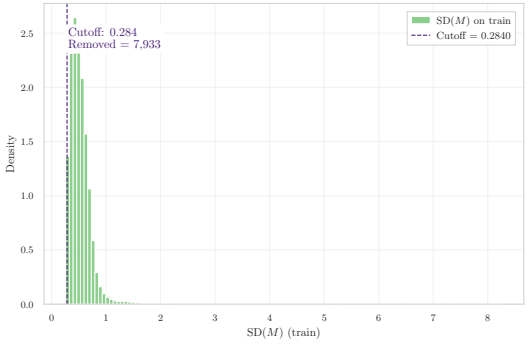
5.3.2 Dataset GSE225845

The feature selection framework described in Section 5.2 was applied to the preprocessed GSE225845 cohort, restricted to the Normal and Adjacent tissue classes, yielding a final panel of $K = 5000$ CpGs after variability filtering, stability ranking, redundancy pruning, and diversification. The selected loci are available in the project repository at `final_5000_gse225845.csv`.

Empirical Variability-Based Dimensionality Reduction Variability was quantified via the inter-quantile range $r_{\beta,j} = P90(\beta) - P10(\beta)$ computed on the training set. Figure 5.3a shows the empirical distribution of r_β across CpGs. Applying the predefined threshold $\tau = 0.05$ removed 134,039 low-variability loci, defining the reduced candidate space for downstream supervised ranking. Lower



(a) $r_\beta = P90(\beta) - P10(\beta)$ on the training set. Candidate thresholds (0.01, 0.02, 0.05) are indicated.



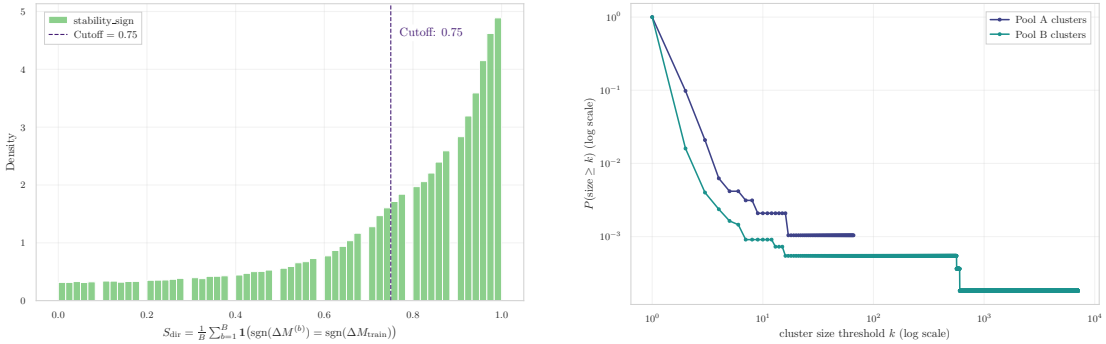
(b) Distribution of $SD(M)$ on the training set. The dashed line indicates the bottom 2% quantile used for residual variance trimming.

Figure 5.3: Variability-based filtering in β-space and residual variance trimming in M-space in GSE225845.

candidate thresholds (0.01, 0.02) are shown for reference but were not adopted. As observed in GSE69914, the variability-based reduction does not induce pathological genomic bias: invariant loci display moderate enrichment in promoter-proximal regions, while gene bodies and 3'UTR regions are relatively under-represented (see Appendix B).

Re-alignment to M-values and Residual Variance Regularisation The retained CpGs were deterministically transformed to M-values, after which a residual low-variance trimming was applied in M-space. Figure 5.3b displays the distribution of empirical standard deviations $SD(M)$ on the training set. Removal of the bottom 2% of CpGs by dispersion eliminated 7,933 loci (cutoff $SD(M) = 0.284$), defining the variance-regularised candidate space entering the stability-ranking stage.

Biologically Weighted Stability Selection and Region-Level Consolidation Repeated stratified resampling was used to estimate stability of signed mean differences. Figure 5.4a shows the distribution of the directional stability score S_{dir}. The distribution is strongly right-skewed, with a substantial concentration of CpGs exhibiting S_{dir} close to 1, indicating highly consistent effect directions across resampling splits. The threshold $\text{stab}_j \geq 0.75$ therefore removes only loci with unstable or sign-reversing effects, retaining the majority of directionally robust signals in this cohort. After stability and adaptive effect-size filtering (60th percentile), the top 15,000 CpGs were retained by design, defining the candidate



(a) Distribution of directional stability scores S_{dir}. The dashed line indicates the cutoff $\text{stab}_j \geq 0.75$.

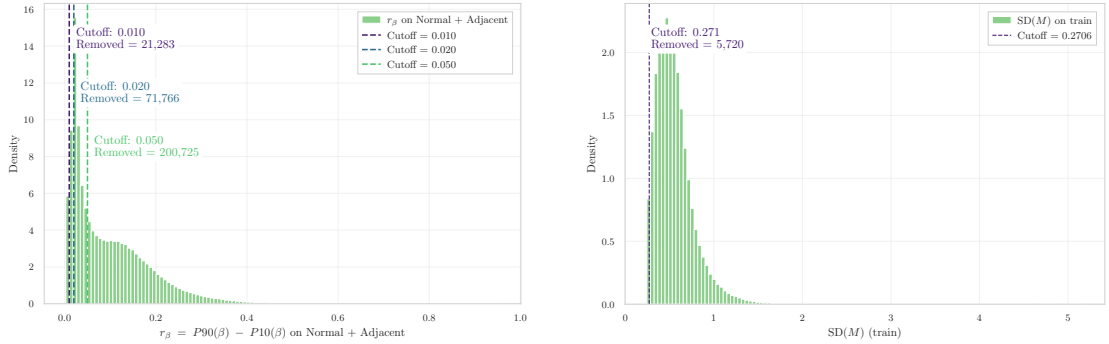
(b) Cluster size survival curves for anchored (Pool A) and non-anchored (Pool B) CpGs under $|r| \geq 0.85$.

Figure 5.4: Stability and redundancy diagnostics in GSE225845.

pool passed to correlation-based redundancy pruning.

Correlation-Based Redundancy Pruning and Graph-Theoretic Clustering After stability and effect-size filtering, correlation-based graph clustering was applied separately to the anchored and non-anchored pools using a threshold of $|r| \geq 0.85$ computed on the training set. Figure 5.4b shows the cluster size survival curves. The survival functions decay rapidly for small k, indicating that most connected components are of size 1–3. However, a pronounced heavy tail is observed, particularly in the non-anchored pool, with a limited number of large correlated blocks spanning tens to hundreds of CpGs. This redundancy pruning step substantially reduced the candidate space from 15,000 to 6,453 loci, indicating a markedly higher correlation structure in GSE225845 compared to GSE69914. The result suggests extended co-methylation domains consistent with the denser EPIC probe coverage. Selecting a single representative CpG per connected component defines the redundancy-pruned set passed to the diversification stage.

Genomic Diversification via Constrained Greedy Selection The constrained greedy diversification procedure was applied to the 6,453 redundancy-pruned CpGs under chromosome, genomic-context, and 500kb spatial-window constraints. The strict constraint regime was sufficient to reach $K = 5000$ without requiring relaxation (maximum chromosome share 8%, maximum context share 65%, maximum 15 CpGs per 500kb window). The final panel comprises 2,115 Island loci, 1,746 OpenSea loci, 534 N_Shore loci, 436 S_Shore loci, and 91 N_Shelf loci, indicating a diversified genomic-context composition. The chromosome cap was actively binding in the strict stage, ensuring balanced genomic representation



(a) $r_\beta = P90(\beta) - P10(\beta)$ on the training set. Candidate thresholds (0.01, 0.02, 0.05) are indicated.

(b) Distribution of $SD(M)$ on the training set. The dashed line indicates the bottom 2% quantile used for residual variance trimming.

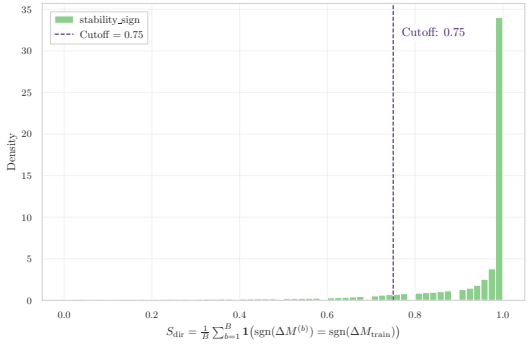
Figure 5.5: Variability-based filtering in β-space and residual variance trimming in M-space in GSE287331.

without overconcentration in high-density regions. The resulting panel therefore satisfies all diversification constraints by construction.

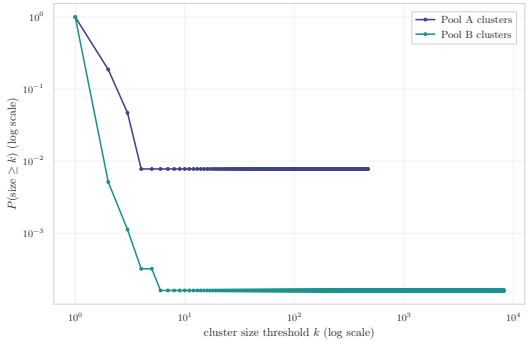
5.3.3 Dataset GSE287331

The feature selection framework described in Section 5.2 was applied to the preprocessed GSE287331 cohort, restricted to the Normal and Adjacent tissue classes, yielding a final panel of $K = 5000$ CpGs after variability filtering, stability ranking, redundancy pruning, and diversification. The selected loci are available in the project repository at `final_5000_gse287331.csv`.

Empirical Variability-Based Dimensionality Reduction Variability was quantified via the inter-quantile range $r_{\beta,j} = P90(\beta) - P10(\beta)$ computed on the training set. Figure 5.5a shows the empirical distribution of r_β across CpGs, which is heavily concentrated near zero, indicating a large fraction of weakly variable loci. Applying the predefined threshold $\tau = 0.05$ removed 200,725 low-variability CpGs, substantially reducing the dimensionality of the feature space prior to supervised ranking. Lower candidate thresholds are shown for reference but were not adopted, as they would retain a markedly larger proportion of near-constant probes. Consistent with the other cohorts, invariant CpGs exhibit moderate enrichment in promoter-proximal regions (notably TSS200 and first exon), while distal or non-promoter regions are relatively under-represented (see Appendix B). The magnitude of enrichment is less pronounced than in GSE69914 and GSE225845,



(a) Distribution of directional stability scores S_{dir}. The dashed line indicates the cutoff $\text{stab}_j \geq 0.75$.



(b) Cluster size survival curves for anchored (Pool A) and non-anchored (Pool B) CpGs under $|r| \geq 0.85$.

Figure 5.6: Stability and redundancy diagnostics in GSE287331.

suggesting that the variability structure of this dataset is more diffusely distributed across genomic contexts.

Re-alignment to M-values and Residual Variance Regularisation The retained CpGs were deterministically transformed to M-values, after which a residual low-variance trimming was applied in M-space. Figure 5.5b displays the distribution of empirical standard deviations $SD(M)$ on the training set, which is unimodal with moderate right-tail behaviour. Removal of the bottom 2% of CpGs by dispersion (cutoff $SD(M) = 0.271$) eliminated 5,720 loci, defining the variance-regularised candidate space entering the stability-ranking stage.

Biologically Weighted Stability Selection and Region-Level Consolidation Repeated stratified resampling was used to estimate stability of signed mean differences. Figure 5.6a shows the distribution of the directional stability score S_{dir}, which is sharply concentrated near 1. A pronounced peak at high stability values indicates that the vast majority of CpGs exhibit highly consistent effect directions across resampling splits. The threshold $\text{stab}_j \geq 0.75$ therefore removes only a small subset of unstable loci, enforcing directional robustness in line with the stability-selection principle of [6] while retaining most candidate signals. After stability and adaptive effect-size filtering (60th percentile), the top 15,000 CpGs were retained by design, defining the candidate pool passed to correlation-based redundancy pruning.

Correlation-Based Redundancy Pruning and Graph-Theoretic Clustering After stability and effect-size filtering, correlation-based graph clustering was applied

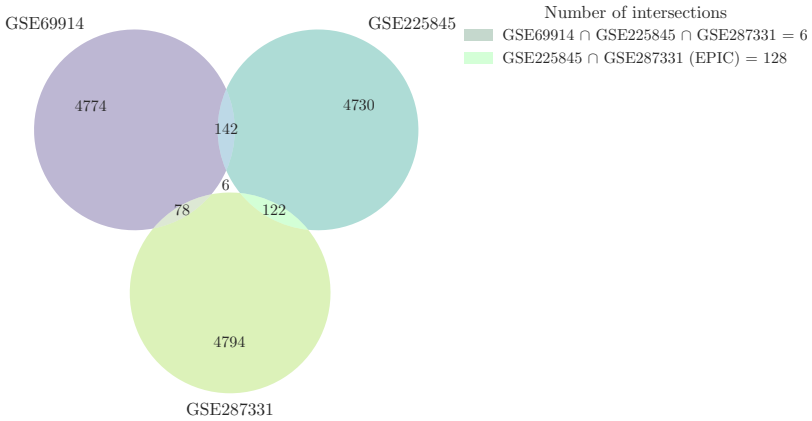


Figure 5.7: Three-way overlap of the final 5,000 CpG signatures independently selected in GSE69914, GSE225845 and GSE287331.

separately to the anchored and non-anchored pools using a threshold of $|r| \geq 0.85$ computed on the training set. Figure 5.6b shows the cluster size survival curves. The survival functions exhibit a rapid decay, indicating that most connected components are of size 1–3 and that large correlated blocks are rare. Compared to GSE225845, the absence of a pronounced heavy tail suggests a more locally structured correlation pattern without extended co-methylation domains. Selecting a single representative CpG per connected component reduced the candidate space to 6,300 loci, defining the redundancy-pruned set passed to the diversification stage.

Genomic Diversification via Constrained Greedy Selection The constrained greedy diversification procedure was applied to the 6,300 redundancy-pruned CpGs under chromosome, genomic-context, and 500kb spatial-window constraints. The strict constraint regime was sufficient to reach $K = 5000$ without requiring relaxation. The final panel comprises 1,992 OpenSea loci, 1,692 Island loci, 556 N_Shore loci, 507 S_Shore loci, and 135 S_Shelf loci, indicating a diversified genomic-context composition with a substantial contribution from distal regions.

5.4 Inter-Dataset Stability of the Final 5,000 CpG Signatures

The feature selection framework described in Section 5.3 was applied independently to each cohort, yielding three dataset-specific panels of $K = 5{,}000$ CpG loci optimised for within-dataset discrimination along the Normal–Adjacent axis. The present section quantifies the degree to which these panels converge on a shared

Table 5.1: Permutation-based enrichment of pairwise overlaps among the three final 5,000-CpG panels.

Pair	Shared CpGs	Expected Overlap	Enrichment Ratio	p-value
GSE69914–GSE225845	148	69.0	2.14	$< 10^{-4}$
GSE69914–GSE287331	84	68.3	1.23	0.024
GSE225845–GSE287331	128	37.9	3.38	$< 10^{-4}$

epigenetic drift signal, thereby assessing the cross-cohort replicability of the feature selection outcome. Consistency is evaluated at three complementary levels: (i) identity-level overlap of selected loci (set-level replicability), (ii) concordance of estimated effect sizes across cohort intersections (effect-level agreement), and (iii) transferability of the drift structure as a discriminative ranking in held-out datasets. This multi-resolution framework mirrors the analytical strategy adopted for the pre-selection inter-dataset analysis of Section 5.3, now applied to the output of the full selection pipeline rather than to the global methylome.

5.4.1 Set-Level Replicability

Pairwise intersections between the three final 5,000-CpG panels were computed and are visualised in Figure 5.7. The identity-level overlap is summarised by the Jaccard similarity coefficient, which for two selection sets A and B is defined as:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}, \tag{5.17}$$

where $|A \cap B|$ denotes the cardinality of the shared loci and $|A \cup B|$ the size of their union. This scale-invariant measure normalises the raw intersection count against the total selection universe, rendering it comparable across cohort pairs with different platform intersections. The observed values, $J(\text{GSE69914}, \text{GSE225845}) = 0.0150$, $J(\text{GSE69914}, \text{GSE287331}) = 0.0085$, $J(\text{GSE225845}, \text{GSE287331}) = 0.0130$, indicate that fewer than 1.5% of selected CpGs are shared between any two cohort-specific panels, confirming that identity-level replicability is limited across all pairs. This pattern is consistent with the weak locus-level reproducibility observed in the pre-selection inter-dataset analysis (Section 6.4). Given the high-dimensional setting, cohort-specific noise structure, and heterogeneous sampling conditions, only partial overlap between independently derived panels is expected.

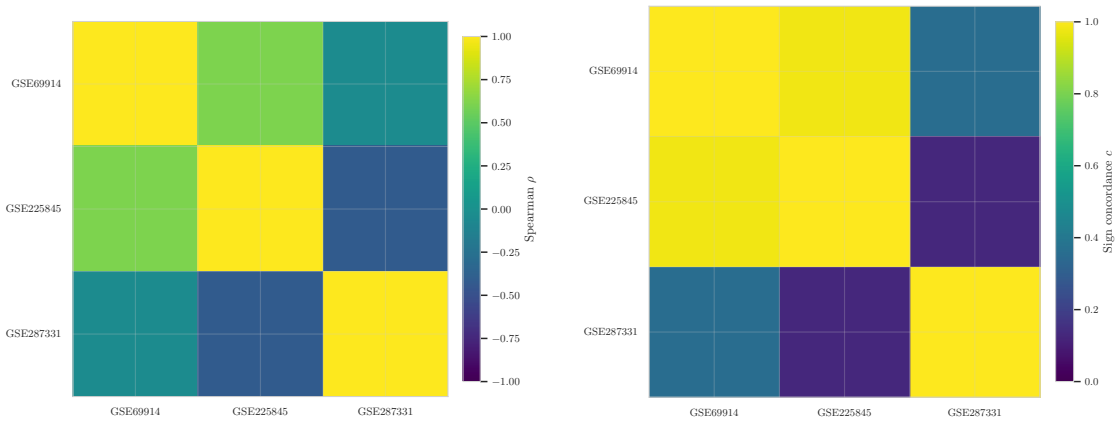
Permutation-based assessment of overlap significance A naive interpretation of the raw intersection size would be misleading without accounting for the baseline expected overlap under random selection. The appropriate null model must reflect the actual constraints of the selection procedure: each panel draws exactly

$K = 5{,}000$ CpGs from its own post-preprocessing feature space, and the two panels can only share loci that are present in *both* datasets after platform-specific filtering. The effective sampling universe for each pair is therefore the post-preprocessing pairwise intersection $\mathcal{U}_{AB} = \mathcal{C}_A \cap \mathcal{C}_B$, where $\mathcal{C}_A$ and $\mathcal{C}_B$ denote the full sets of CpGs retained after preprocessing in datasets A and B, respectively. Under this null, 10,000 independent replications were generated by drawing two sets of $K = 5{,}000$ CpGs uniformly at random from $\mathcal{U}_{AB}$, without replacement, and recording their intersection size. The empirical p-value is defined as the proportion of replications in which the random overlap equals or exceeds the observed value. This construction is conservative relative to null models based on the full genome or on the union of platform probes, since the restricted universe $\mathcal{U}_{AB}$ raises the baseline expected overlap; significance under this test therefore provides a more stringent criterion for enrichment. Results are reported in Table 5.1. All three pairwise overlaps exceed the permutation null at nominal significance. The most pronounced enrichment is observed for the EPIC–EPIC pair (GSE225845–GSE287331; enrichment $\times 3.4$), reflecting the larger shared platform intersection available to that pair. The HM450–EPIC pairs yield smaller but still significant enrichments, indicating that the selection procedure recovers a non-random shared subset even across platform boundaries.

Importantly, statistical significance here quantifies enrichment relative to the large-cardinality random baseline, and should not be conflated with biological equivalence of the selected panels. The restricted sampling universe $\mathcal{U}_{AB}$ is itself of order $\mathcal{O}(10^5)$, so even modest absolute overlaps can achieve nominal significance when drawn from such a large pool. Jaccard indices below 0.02 demonstrate that fewer than one in fifty selected CpGs is reproducibly identified across independent cohorts, and that identity-level concordance is insufficient to constitute a shared genomic signature. Cross-cohort replicability must therefore be evaluated at the level of effect structure rather than probe identity, as formalised in the following subsections.

5.4.2 Effect-Direction Concordance

The absence of strong identity-level overlap does not preclude the existence of a shared differential methylation structure: independently selected panels may still converge on the same underlying drift axis if the corresponding effect sizes are concordant across cohorts. To evaluate this possibility, signed mean methylation differences $\Delta\beta_j = \bar{\beta}_j^{\mathrm{Adj}} - \bar{\beta}_j^{\mathrm{Norm}}$ were computed for each CpG j in the pairwise intersection of the respective 5,000-CpG panels. Concordance was assessed through two complementary measures: the Spearman rank correlation ρ of signed effect sizes, which captures both magnitude and direction, and the directional concordance



(a) Spearman correlation ρ of $\Delta\beta$ **(b)** Sign concordance c

Figure 5.8: Inter-dataset effect concordance of the final 5,000 CpG signatures.

rate:

$$c = \widehat{\mathbb{P}}\Big(\text{sign}(\Delta\beta_j^{(A)}) = \text{sign}(\Delta\beta_j^{(B)})\Big)\,, \tag{5.18}$$

estimated with Wilson confidence intervals, which isolates the directional component independently of magnitude. The distinction between ρ and c is methodologically important: a low signed correlation combined with high directional concordance indicates attenuated but direction-preserving effects, whereas negative ρ accompanied by $c \ll 0.5$ is diagnostic of systematic polarity inversion rather than mere amplitude attenuation. Both quantities are visualised jointly in Figure 5.8.

GSE69914–GSE225845: shared drift axis The HM450–EPIC pair exhibits strong positive rank correlation ($\rho = 0.605$) and near-complete directional alignment ($c = 0.980$, Wilson 95% CI excluded from 0.5 by a wide margin). Together, these values indicate that the two datasets not only agree on the direction of each locus-specific shift but also preserve the relative ordering of effect magnitudes. This constitutes the clearest evidence of a shared epigenetic drift axis across independent cohorts, despite the known platform heterogeneity between the HM450 and EPIC arrays.

GSE287331: systematic polarity inversion Pairs involving GSE287331 display a qualitatively distinct pattern. The Spearman correlations are weakly negative ($\rho = -0.042$ with GSE69914; $\rho = -0.420$ with GSE225845), and directional concordance falls substantially below 0.5 ($c = 0.357$ and $c = 0.125$, respectively). The value $c = 0.125$ in particular implies that only one in eight selected CpGs shares the same Adjacent–Normal polarity between GSE225845 and GSE287331, a pattern incompatible with stochastic disagreement and indicative of a systematic reversal

of the dominant drift direction. Crucially, this inversion is not attributable to the absence of a discriminative signal in GSE287331: the intra-dataset analyses of Section 5.3 demonstrated that this cohort yields the strongest and most geometrically coherent Normal–Adjacent separation of the three (silhouette = 0.3442; outlier A/N ratio = 13.3). Rather, the inversion reflects a structural asymmetry in the orientation of the drift manifold, likely arising from a combination of platform-specific probe-type composition, the distinct tumour-proximity sampling design of GSE287331 (TPxA axis), and residual confounding by cohort-specific batch structure.

5.4.3 Magnitude-Stratified Concordance

The signed and directional analyses of Section 5.4.2 characterise the global behaviour of effect sizes across the pairwise intersection of selected panels. A complementary question is whether concordance is uniform across the effect magnitude spectrum or, rather, whether it is concentrated in loci exhibiting the strongest drift signals — a distinction with direct implications for the biological interpretability of cross-cohort consistency. To address this, CpGs in the pairwise intersection were stratified by decile of absolute effect magnitude $|\Delta\beta|$, computed independently in each dataset. For each decile bin, the fraction of loci whose effect direction agreed across cohorts was computed, yielding a magnitude-stratified directional concordance profile.

For the GSE69914–GSE225845 pair, directional concordance is high and stable across all deciles, with no systematic attenuation in lower-magnitude strata. Crucially, the top decile — comprising CpGs with the largest $|\Delta\beta|$ in both datasets — exhibits the highest concordance, indicating that the strongest signals are preferentially shared. This behaviour supports the interpretation that the cross-cohort agreement between these two datasets is not an artefact of magnitude-insensitive directional alignment, but reflects a genuinely shared structure concentrated at loci with the most pronounced Adjacent–Normal differential methylation.

Pairs involving GSE287331 display a markedly different pattern. Directional concordance remains depressed across all decile strata, with no recovery in the high-magnitude bins. This rules out the scenario in which directionality disagreement would be confined to low-amplitude, noisy loci — as would be expected under pure stochastic disagreement — and instead confirms that the inversion of effect orientation is structural and affects loci of all effect magnitudes, including those most strongly discriminative within GSE287331. Taken together with the negative signed correlations reported in Section 5.4.2, this evidence positions the GSE287331 discordance firmly in the *directional inversion* regime rather than the *structural heterogeneity* regime of Section 5.5: the selected loci are not uninformative, but their dominant drift axis is systematically rotated relative to that of the other two cohorts.

5.4.4 Cross-Dataset Transferability

Set-level overlap and effect-level concordance characterise the structure of the selected panels on their own terms. A more stringent criterion for cross-cohort replicability is *functional transferability*: whether the differential methylation signal inferred in one cohort retains discriminative power when applied to an independent dataset, without any retraining or adaptation. Formally, for each ordered pair of datasets (A, B), a linear scoring function $f^{(A)} : \mathbb{R}^{|S_{AB}|} \to \mathbb{R}$ was defined on the intersection S_{AB} of the respective 5,000-CpG panels, using the cohort-A effect sizes $\Delta\beta^{(A)}$ as fixed coefficients:

$$f^{(A)}(\mathbf{x}_i^{(B)}) = \sum_{j \in S_{AB}} \Delta\beta_j^{(A)} \cdot x_{ij}^{(B)}, \tag{5.19}$$

where $\mathbf{x}_i^{(B)}$ denotes the methylation profile of sample i in dataset B. This zero-shot transfer procedure isolates the intrinsic portability of the drift structure itself, independent of classifier fitting. Discriminative performance was evaluated via the area under the ROC curve (AUC), with values near 0.5 indicating absence of transfer, values substantially above 0.5 indicating directionally consistent transfer, and values substantially below 0.5 indicating systematic inversion — i.e. the scoring function trained on cohort A assigns *higher* scores to Normal than to Adjacent samples in cohort B.

GSE69914–GSE225845 Bidirectional transfer between the HM450 and EPIC cohorts was evaluated by training a univariate $\Delta\beta$ ranking rule in one dataset and applying it directly to the other. Specifically, given an effect-size vector $\Delta\beta$ estimated in the source cohort, each sample x in the target cohort was assigned the discriminative score:

$$s(x) = \sum_{j \in \mathcal{S}} \Delta\beta_j \, x_j, \tag{5.20}$$

where $\mathcal{S}$ denotes the selected CpG panel and x_j the methylation level of locus j. No re-estimation or adaptation to the target distribution was performed. Discriminative performance was quantified via the area under the receiver operating characteristic curve (AUC), defined as:

$$\mathrm{AUC} = \int_0^1 \mathrm{TPR}(t) \, d\mathrm{FPR}(t), \tag{5.21}$$

where TPR and FPR denote the true- and false-positive rates across classification thresholds. For the binary Normal–Adjacent problem, this quantity is equivalently expressed as:

$$\mathrm{AUC} = \mathbb{P}(s(X_A) > s(X_N)), \tag{5.22}$$

that is, the probability that a randomly selected Adjacent sample receives a higher score than a randomly selected Normal sample. The AUC takes values in [0,1], where 0.5 corresponds to chance-level discrimination, values above 0.5 indicate discriminative ability in the correct direction, and 1 corresponds to perfect separation. Values below 0.5 reflect systematic inversion of the ranking rule.

The resulting AUC was 0.776 when training on GSE69914 and testing on GSE225845, and 0.794 in the reverse direction. Both values are substantially above chance, indicating that the shared drift axis identified through concordance analysis translates into genuine discriminative transferability. These results provide the strongest cross-cohort evidence in the present study that the GSE69914–GSE225845 pair captures a reproducible and functionally portable epigenetic drift structure.

Transfers involving GSE287331 All four directed transfer pairs involving GSE287331 yield AUC values in the range [0.017, 0.226], substantially and consistently below chance level. Sub-random AUC is not a failure of discriminability per se, but rather a signature of systematic directional inversion: the scoring function derived from one cohort actively misranks samples in the other, assigning higher drift scores to the tissue class with lower methylation deviation. This outcome is the functional counterpart of the near-zero or negative sign concordance ($c = 0.125$–0.357) and negative Spearman correlations ($\rho = -0.042$ to -0.420) reported in Section 5.4.2, and provides definitive evidence that the polarity inversion observed at the descriptive level has concrete consequences for predictive transfer.

Interpretation The transferability results complete the multi-level concordance portrait of the three selected panels. The GSE69914–GSE225845 pair satisfies all criteria for the *shared drift structure* regime: significant set enrichment, high directional concordance, magnitude-stratified agreement concentrated at strong-effect loci, and bidirectional AUC well above chance. Pairs involving GSE287331 are consistent with the *directional inversion* regime: non-trivial set enrichment, systematically inverted effect polarity across all magnitude strata, and sub-random transfer performance. These asymmetries have direct methodological consequences for the downstream modelling strategy developed in Chapter 6: naive cross-cohort transfer from or to GSE287331 is not only non-informative but actively misleading, motivating the need for orientation-aware or cohort-adaptive learning approaches.

5.5 Feature Selection Outcomes and Modelling Implications

The feature selection framework developed in this chapter addresses the core challenge posed by genome-wide DNA methylation data in the HDLSS regime:

Table 5.2: Number of CpGs retained at each step of the feature selection pipeline (training set).

Pipeline step	CpGs in GSE69914	CpGs in GSE225845	CpGs in GSE287331
Post-preprocessing space	251,483	530,683	486,725
After Edgar variability filter	151,380	396,644	286,000
After M-value SD trimming	148,352	388,711	280,280
After stability + effect-size filter	15,000	15,000	15,000
After correlation pruning	12,313	6,453	6,300
After constrained diversification	5,000	5,000	5,000

identifying a compact, stable, and biologically coherent CpG panel capable of characterising the Normal–Adjacent transition across independent cohorts. The pipeline applies strictly train-only dimensionality reduction stages, each targeting a distinct source of noise or redundancy. Table 5.2 summarises the progressive reduction from post-preprocessing spaces of 251,483–530,683 CpGs to final panels of $K = 5{,}000$ loci. The resulting panels exhibit directional stability scores concentrated near unity, indicating reproducible Adjacent–Normal polarity across resampling splits; correlation pruning eliminates co-methylated blocks; and genomic diversification enforces broad chromosomal and regulatory-context coverage. The inter-dataset evaluation of Section 5.4 reveals partial and asymmetric cross-cohort replicability. GSE69914 and GSE225845 satisfy the *shared drift structure* regime: their panels are mutually enriched beyond chance, exhibit near-complete directional concordance ($c = 0.980$), and support bidirectional zero-shot transfer (AUC = 0.776 and 0.794). GSE287331, by contrast, occupies the *directional inversion* regime: its panel captures a robust but systematically inverted drift axis, yielding sub-random transfer performance (AUC $\in [0.017, 0.226]$) against both partner cohorts. This asymmetry implies that linear $\Delta\beta$ signatures, however stable within a cohort, do not constitute a universally portable representation of the Normal–Adjacent transition.

These structural properties and cross-cohort limitations jointly define the downstream modelling strategy. Within each cohort, the selected panels provide a stability-filtered input for supervised Normal–Adjacent classification, evaluated in Chapter 6. Where drift orientation is preserved — as between GSE69914 and GSE225845 — cross-dataset transfer is a principled objective; where systematic inversion is present, orientation-aware strategies are required. Chapter 7 addresses this through multi-cohort pooling and biological synthesis, investigating whether signal consolidation across heterogeneous cohorts supports a more robust characterisation of the Normal–Adjacent transition despite divergent drift orientations.

Chapter 6

Predictive Modelling, Constrained Optimisation and Biological Interpretation

6.1 From Robust Feature Selection to Predictive Modelling

Chapter 5 established a statistically stable and structurally diversified CpG panel of fixed cardinality ($K = 5{,}000$) for each dataset, integrating variability screening, biologically weighted stability selection, redundancy pruning, and genomic diversification. The inter-dataset evaluation of Section 5.4 further revealed that cross-cohort replicability is partial and asymmetric: GSE69914 and GSE225845 share a concordant drift orientation and support bidirectional zero-shot transfer, whereas GSE287331 exhibits a systematically inverted drift axis, yielding sub-random transfer performance against both partner cohorts. The objectives of this chapter are threefold. First, intra-dataset classification performance is quantified to verify whether the selected panels support accurate Normal–Adjacent discrimination within each cohort; a linear Support Vector Machine is adopted for GSE225845 and GSE287331, and a k-Nearest Neighbours classifier for GSE69914 to probe the degree of linear separability. Second, cross-dataset transferability is examined under a zero-leakage design — training on one cohort and evaluating directly on another — to quantify how well the inferred epigenetic drift axis generalises across independent cohorts, and whether performance degradation is systematically associated with the directional concordance structure of Chapter 5. Third, a constrained combinatorial compression step derives a compact 50-CpG subset via mixed-integer linear programming, investigating whether predictive

structure and biological diversity can be jointly preserved under severe dimensional reduction. The biological coherence of the compressed panels is then evaluated through CpG-to-gene mapping, cross-referenced against the COSMIC Cancer Gene Census (CGC, version 103, GRCh38) [77, 78], and pathway-level enrichment analysis — closing the loop between statistical discrimination and functional relevance. The limitations that emerge, in particular the failure of linear $\Delta\beta$ signatures to generalise across cohorts with divergent drift orientations, directly motivate the representation-learning approach of Chapter 7.

6.2 Supervised Learning Framework

Each dataset is restricted to its corresponding $K = 5{,}000$ CpG panel and two baseline classifiers are evaluated independently within each cohort: a linear Support Vector Machine for GSE225845 and GSE287331, and a k-Nearest Neighbours for GSE69914 to probe linear separability. The primary aim is not to maximise predictive performance but to quantify the discriminative content encoded in the selected panels.

6.2.1 Intra-Dataset Supervised Modelling

Data partitioning and leakage control For each dataset, samples are partitioned into stratified training and test sets under an 80/20 split, preserving the Normal–Adjacent class proportions. Feature selection was performed strictly within training folds in Chapter 5, so no information from the held-out set influences the choice of CpGs. Standardisation is fitted exclusively on the training partition and applied without re-estimation to the test set, ensuring complete leakage control throughout the pipeline.

Linear Support Vector Machine A linear Support Vector Machine (SVM) is adopted as the primary classifier for GSE225845 and GSE287331 [79]. This choice is motivated by two reasons. First, the feature space is high-dimensional relative to sample size, a regime where linear models exhibit strong empirical performance and avoid the curse of dimensionality affecting kernel methods. Second, the linear decision boundary admits a direct geometric interpretation along the epigenetic drift axis: the weight vector w identifies the methylation direction most separating Normal and Adjacent tissues, enabling hyperplane comparison across cohorts in the cross-dataset analysis of Section 6.2.2. Given training data $\{(x_i, y_i)\}_{i=1}^{n}$ with $x_i \in \mathbb{R}^p$ and $y_i \in \{-1, +1\}$, the soft-margin primal formulation is:

$$\min_{w,b,\xi} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \quad \text{s.t.} \quad y_i(w^\top x_i + b) \geq 1 - \xi_i,\ \xi_i \geq 0. \tag{6.1}$$

The decision function is $f(x) = w^\top x + b$, with class assignment given by $\text{sign}(f(x))$. All models are trained on standardised M-values with fixed hyperparameters ($C = 0.5$, `max_iter` = 10,000), deliberately avoiding data-driven tuning to keep the evaluation as a measure of the panel's intrinsic discriminative content rather than of classifier optimisation.

k-Nearest Neighbours (GSE69914 only) For GSE69914, a k-Nearest Neighbours classifier is introduced [80]. Its inclusion serves a specific diagnostic purpose: it is possible to assess whether the discriminative signal in GSE69914 is linearly structured or whether it resides in local manifold geometry. For a query x, let $\mathcal{N}_k(x)$ denote the set of the k closest training samples under Euclidean distance; the predicted label is:

$$\hat{y}(x) = \text{mode}\{y_i : \ x_i \in \mathcal{N}_k(x)\}. \tag{6.2}$$

The method is applied on standardised features with fixed $k = 21$.

Performance metrics Classification performance is evaluated through a complementary set of metrics designed to provide a complete picture under potential class imbalance. The Area Under the ROC Curve (AUC), defined in Chapter 5, serves as the primary threshold-free discriminability measure. To assess performance independently of class distribution, two summary scalar indices are adopted. *Balanced accuracy* (BAcc) is defined as the arithmetic mean of sensitivity and specificity:

$$\text{BAcc} = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP}}\right). \tag{6.3}$$

The *Matthews Correlation Coefficient* (MCC) provides a single scalar that accounts simultaneously for all four entries of the confusion matrix, making it particularly robust when class sizes differ substantially:

$$\text{MCC} = \frac{\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}. \tag{6.4}$$

At the operating threshold, performance is further characterised through precision p — the fraction of positive predictions that are correct — recall r — the fraction of true positives retrieved — and their harmonic mean, the F1-score, which provides a single balanced summary of the precision–recall trade-off:

$$p = \frac{\text{TP}}{\text{TP} + \text{FP}}, \qquad r = \frac{\text{TP}}{\text{TP} + \text{FN}}, \qquad \text{F1} = \frac{2\,p \cdot r}{p + r}. \tag{6.5}$$

6.2.2 Cross-Dataset Generalisation Protocol

The cross-cohort transferability analysis addresses a question that intra-dataset evaluation cannot answer: whether the epigenetic drift axis inferred within a single cohort encodes a portable biological signal or a cohort-specific artefact. The design is deliberately zero-leakage — no information from the target dataset enters the training procedure at any stage — so that performance on the external cohort reflects true out-of-distribution generalisation rather than any form of implicit adaptation.

Feature space alignment Cross-dataset evaluation requires strict alignment of the feature space between source and target cohorts. For each ordered pair (A, B), the CpG panel is restricted to the intersection of the effective 5,000-CpG sets retained after intra-dataset feature selection in cohort A: only CpGs present in both datasets and included in the source panel are considered. Column ordering is explicitly harmonised to ensure identical feature indexing across matrices, and all transformations — including the $\beta \to M$ mapping where required — are applied consistently prior to modelling. No re-selection, re-ranking, or any other data-driven operation is performed on the target cohort.

Train-on-A / Test-on-B design The classifier is trained exclusively on the training split of cohort A over the aligned CpG space. Feature standardisation is fitted on the source training partition and applied without re-estimation to cohort B. Probability calibration is performed via sigmoid fitting with three-fold internal cross-validation, entirely within the source training data, and the resulting calibrated decision function is evaluated directly on the target cohort. No hyperparameter tuning, retraining, or distributional adaptation is performed on B at any stage. This strict separation quantifies true zero-shot transferability of the learned separating hyperplane.

Evaluation criteria Performance is quantified using the same metrics as in the intra-dataset setting — AUC, balanced accuracy, and MCC — reported jointly for the internal test split of cohort A and the external cohort B, so that within-cohort and cross-cohort discrimination can be compared directly. Degradation in external AUC or balanced accuracy is interpreted as evidence that the learned separating hyperplane does not align with the drift structure of the target cohort. Of particular interest is the case of systematic performance inversion, examined against the directional concordance structure documented in Chapter 5: external performance below random expectation (AUC ≈ 0.5) is treated as a signature of reversed drift orientation rather than of weak shared signal, a distinction that

carries direct methodological implications for the multi-cohort strategy developed in Chapter 7.

6.2.3 Constrained Subset Optimisation via Knapsack Formulation

The 5,000-CpG panel retains substantial redundancy relative to the intrinsic dimensionality of the discriminative signal. A further compression step is therefore introduced with the dual aim of isolating the most informative loci and verifying whether predictive performance and biological diversity can be jointly preserved under severe cardinality reduction. The target panel size is fixed at $K = 50$, corresponding to a 100-fold compression relative to the input set. The selection problem is cast as a multi-constraint mixed-integer linear programme (MILP), inspired by the classical 0–1 knapsack problem of [81] and extended to incorporate structural and biological requirements analogous to those employed in marker panel design [82, 83]. Unlike the classical single-capacity formulation, the present model couples a composite statistical objective with eight hard constraints that jointly enforce cardinality, correlation independence, chromosomal balance, biological-region anchoring, gene-level linking, per-gene diversity, gene-count diversity, and methylation directionality.

Decision variables and composite score Let $\mathcal{J}$ denote the index set of $|\mathcal{J}| = 5{,}000$ candidate CpGs retained after feature selection. For each $j \in \mathcal{J}$, a binary decision variable $x_j \in \{0,1\}$ encodes inclusion in the compressed panel. Gene-level linking variables $y_g \in \{0,1\}$ indicate whether at least one CpG annotated to gene g is selected, and enter both the diversification objective (6.7) and constraint (6.12). The statistical utility of each locus is summarised by a composite score:

$$v_j \;=\; w_{\text{coef}}\,\tilde{c}_j + w_{\Delta M}\,\tilde{d}_j + w_{\sigma}\,\tilde{s}_j + w_{\text{COSMIC}}\,\mathbf{1}_{\{j\in\text{CGC}\}}, \tag{6.6}$$

where $\tilde{c}_j$, $\tilde{d}_j$, and $\tilde{s}_j$ are rank-normalised components corresponding, respectively, to the SVM coefficient magnitude, the absolute ΔM effect size, and the inverse cross-sample standard deviation of M-values on the training set, such that loci with low baseline variability in Normal tissue receive higher utility; the indicator $\mathbf{1}_{\{j\in\text{CGC}\}}$ confers an additional reward to loci annotated in the COSMIC Cancer Gene Census [77]. The normalised diversification functional is defined as:

$$\hat{\Phi}(x,y) \;=\; \frac{1}{K(2+\eta)}\left(\sum_{j:\,\text{prom}(j)} x_j + \sum_{j:\,\text{isl}(j)} x_j + \eta\sum_{g} y_g\right), \tag{6.7}$$

where $\text{prom}(j)$ indicates that locus j is promoter-proximal (TSS200, TSS1500, 1stExon, or 5'UTR), $\text{isl}(j)$ indicates membership in a CpG island, and $\eta > 0$

weights gene-level richness relative to regional coverage. The denominator $K(2+\eta)$ normalises $\hat{\Phi}$ to the unit interval, so that μ is directly interpretable as the maximum fractional penalty on the statistical objective.

Complete MILP formulation Combining (6.6) with (6.7), the full optimisation problem reads:

$$\max_{x} \quad \sum_{j \in \mathcal{J}} v_j \, x_j + \mu \, \hat{\Phi}(x, y)$$

s.t.

$$\sum_{j \in \mathcal{J}} x_j = K \tag{6.8}$$

$$x_i + x_j \leq 1 \qquad \forall \, (i,j) : \, |r_{ij}| \geq 0.85 \tag{6.9}$$

$$\sum_{j \in \mathcal{J}_c} x_j \leq \delta_{\text{chr}} \qquad \forall \, c \tag{6.10}$$

$$\sum_{j \in \mathcal{J}_A} x_j \geq K_A \qquad \text{if } \mathcal{J}_A \neq \emptyset \tag{6.11}$$

$$x_j \leq y_{g(j)} \qquad \forall \, j : \, g(j) \neq \emptyset \tag{6.12}$$

$$\sum_{j \colon g(j) = g} x_j \leq \delta_{\max} \qquad \forall \, g \tag{6.13}$$

$$\sum_{g} y_g \geq G_{\min} \tag{6.14}$$

$$\sum_{j \colon \Delta M_j > 0} x_j \geq h_{\min} \tag{6.15}$$

$$x_j \in \{0,1\}, \quad y_g \in \{0,1\}.$$

Each constraint encodes a distinct biological or structural requirement. Constraint (6.8) fixes the panel cardinality exactly at $K = 50$. Constraint (6.9) prevents the simultaneous selection of locus pairs whose Spearman correlation satisfies $|r_{ij}| \geq 0.85$, extending the redundancy-pruning principle of Chapter 5 to the compressed regime. Constraint (6.10) imposes an absolute ceiling $\delta_{\text{chr}} = 10$ on the number of loci selected from any single chromosome, preventing genomic concentration. Constraint (6.11) enforces a lower bound K_A on the number of loci drawn from Pool A — the biologically anchored subset of CpG-island-proximal sites identified during feature selection — and is active only when $\mathcal{J}_A \neq \emptyset$. Constraints (6.12) and (6.13) jointly govern gene-level diversity: (6.12) activates the linking variable $y_{g(j)}$ whenever CpG j is selected, while (6.13) independently caps

the number of selected loci per gene at $\delta_{\max} = 2$, preventing any single locus neighbourhood from dominating the panel. Constraint (6.14) imposes a lower bound $G_{\min}$ on the number of distinct genes represented, promoting broad regulatory coverage. Finally, constraint (6.15) enforces a minimum count $h_{\min}$ of loci with positive SVM coefficient, i.e. hypermethylated in Adjacent relative to Normal tissue, preventing polarity collapse towards a purely hypomethylated solution.

Optimisation strategy and parameter selection The scalar $\mu \geq 0$ governs the trade-off between statistical score and structural diversification. At $\mu = 0$ the problem reduces to a pure score-maximising knapsack; as μ increases, the objective increasingly rewards genomic balance at the expense of discriminative utility. To identify a stable operating point, μ is swept over the discrete grid $\{0, 0.05, 0.1, 0.25, 0.5, 1.0, 2.0\}$ and the resulting Pareto frontier between $f_1 = \sum_j v_j x_j$ and $f_2 = \hat{\Phi}(x, y)$ is examined. The final value is selected at the empirical knee of this curve, defined as the point of maximum perpendicular distance from the line connecting the two frontier endpoints, beyond which marginal gains in diversification no longer compensate for losses in statistical score. An independent sweep over $w_{\text{COSMIC}} \in \{0, 0.1, 0.2, 0.3, 0.5, 0.7, 1.0\}$ is subsequently performed at the knee μ, and the value maximising COSMIC enrichment subject to a tolerance of 0.1% on f_1 is retained. Solution stability across μ values is assessed by computing the Jaccard similarity between each solution and the knee-point panel, confirming that the selected subset is not sensitive to small perturbations of the diversification weight. Given the moderate input dimensionality ($|\mathcal{J}| = 5{,}000$), the MILP is solved to certified global optimality using the COIN-OR Branch-and-Cut (CBC) solver interfaced via the PuLP modelling library, with a wall-clock time limit of 120 seconds per instance.

6.2.4 Functional and Gene-Level Interpretation

Having established the discriminative capacity of the compressed panels, this subsection examines whether the loci selected through purely statistical and structural criteria converge on genes with documented roles in cancer biology. The analysis proceeds in two stages: CpG-to-gene mapping followed by over-representation testing against curated pathway databases.

CpG-to-gene mapping and methylation directionality Each of the 50 selected CpGs is annotated using the `UCSC_RefGene_Name` field of the Illumina array manifest (HumanMethylation450 for GSE69914; EPIC for GSE225845 and GSE287331), with genes identified via the GRCh38 assembly. When a single CpG maps to multiple gene symbols — a common occurrence for probes located in promoter-flanking or exon-boundary regions — all mapped symbols are retained.

Gene symbols are deduplicated, and unmapped CpGs (annotated as - in the manifest) are excluded from gene-level analyses. Let $\mathcal{G}$ denote the set of unique genes obtained from the 50-CpG panel after deduplication. Methylation directionality for each locus is assigned from the sign of the corresponding linear SVM coefficient: a positive coefficient ($w_j > 0$) indicates that the locus is hypermethylated in the Adjacent class and labelled HYPER; a negative coefficient ($w_j < 0$) indicates hypomethylation and is labelled HYPO. This directionality assignment is consistent with the decision function $f(x) = w^\top x + b$ derived in Section 6.2.1, where the positive class corresponds to Adjacent tissue.

COSMIC cross-reference Genes in $\mathcal{G}$ are cross-referenced against the COSMIC Cancer Gene Census (CGC, version 103, GRCh38) [77], filtered to entries with breast tissue listed under `TUMOUR_TYPES_SOMATIC`. For CpGs mapping to multiple genes, COSMIC membership is flagged if any mapped gene appears in the filtered CGC. The role in cancer (`ROLE_IN_CANCER`) and tumour type annotations are reported at the CpG level to characterise the functional profile of the compressed panel.

Over-representation analysis Functional enrichment is assessed via over-representation analysis using Enrichr [84], queried through the `gseapy` interface. Three complementary databases are interrogated: GO Biological Process 2023, KEGG 2021 Human, and MSigDB Hallmark 2020. The gene list submitted is $\mathcal{G}$; no custom background is specified, so enrichment is evaluated against the Enrichr reference universe for each library. Terms with Benjamini–Hochberg adjusted p-value below 0.05 are considered significant.

6.3 Intra-Dataset Results

The predictive framework described in Section 6.2.1 is applied independently to each of the three cohorts. For each dataset, results are organised around four questions: whether the 5,000-CpG panel supports accurate intra-cohort discrimination; whether this structure is preserved under compression to 50 CpGs; whether the learned separating axis transfers to external cohorts; and whether the compressed panel converges on biologically coherent gene sets.

6.3.1 Dataset GSE69914

GSE69914 presents the most challenging discrimination task among the three cohorts. As established in Chapter 5, the Normal–Adjacent $\Delta\beta$ distribution in this dataset is the narrowest and least directionally coherent of the three, suggesting that

Table 6.1: Performance metrics — KNN, 5,000 CpGs (GSE69914, internal test).

AUC	BAcc	Precision	Recall	F1	MCC
0.7778	0.6167	0.7500	0.3333	0.4615	0.2858

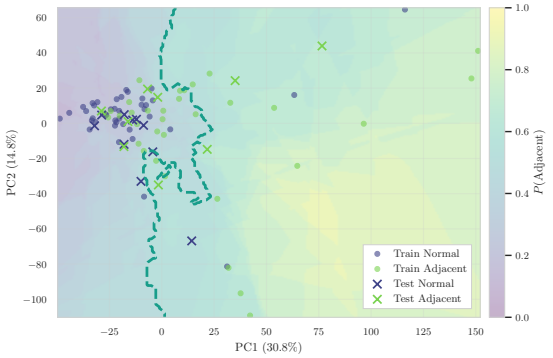


Figure 6.1: Projection of the KNN decision surface onto PC1–PC2 (explaining 30.8% and 14.8% of variance, respectively).

the field-cancerisation signal is either attenuated or partially confounded by residual technical structure. The analyses below quantify the downstream consequences of this property on classification performance and external transferability.

Intrinsic separability of the 5,000-CpG panel A k-Nearest Neighbours classifier is applied to the full 5,000-CpG panel under the stratified 80/20 split, with $k = 21$ selected by scanning odd values in $\{5, 11, 15, 21, 31\}$ on the training partition and retaining the value that maximised balanced accuracy. Test-set performance is reported in Table 6.1. The confusion matrix $\left[\begin{smallmatrix} 9 & 1 \\ 6 & 3 \end{smallmatrix}\right]$ reveals markedly asymmetric behaviour: Normal samples are identified with high reliability, but only one third of Adjacent samples are correctly classified (recall 0.33). The AUC of 0.78 confirms that some discriminative signal is present in the panel, yet the low MCC (0.29) and F1-score (0.46) indicate that no decision threshold achieves balanced separation. The PCA projection (Figure 6.1) and silhouette coefficient of 0.07 are consistent with overlapping class geometries rather than the compact separated clusters observed in the other cohorts.

Compression to 50 CpGs via MILP Following the intra-dataset KNN training, the learned permutation importance scores — alongside the absolute ΔM effect sizes and the inverse baseline variability scores — are fed as input to the MILP formulation of Section 6.2.3. The resulting 50-CpG panel, selected under $\mu = 1.0$ and $w_{\text{COSMIC}} = 1$, is shown in Figure 6.2 with its component-wise utility

Table 6.2: Performance metrics — KNN, 50 CpGs (GSE69914, internal test).

AUC	BAcc	Precision	Recall	F1-score	MCC
0.6500	0.6167	0.7500	0.3333	0.4615	0.2858

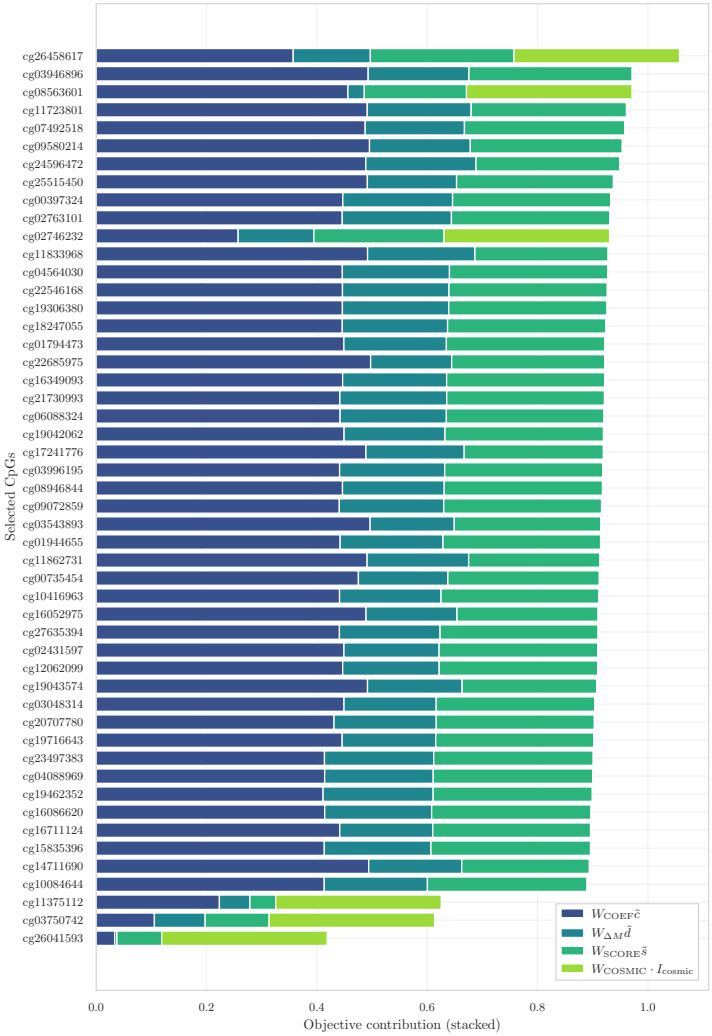
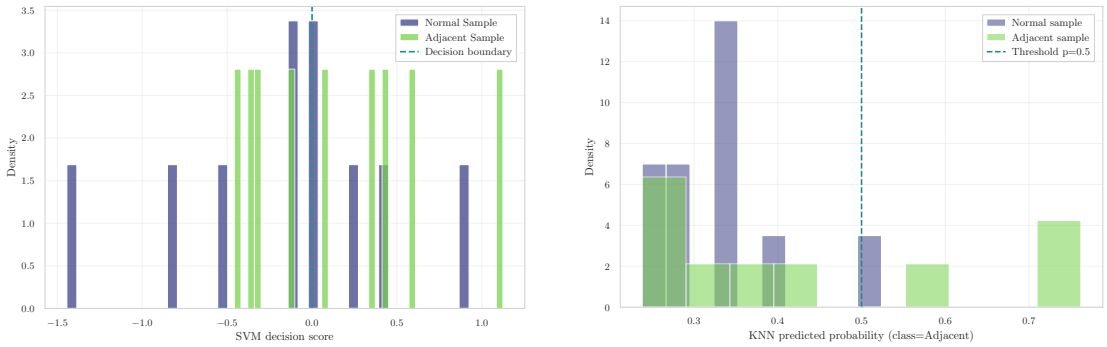


Figure 6.2: Component-wise decomposition of the MILP objective for the 50-CpG panel in GSE69914. Each bar is one selected locus, sorted by total utility; stacked segments show contributions from $w_{\text{coef}}\tilde{c}$, $w_{\Delta M}\tilde{d}$, $w_{\sigma}\tilde{s}$, and $w_{\text{COSMIC}} \cdot \mathbf{1}_{\text{cosmic}}$.

decomposition. A KNN classifier retrained exclusively on these 50 loci yields the metrics in Table 6.2. Threshold-level metrics are identical to the 5,000-CpG model, indicating that the compressed panel reproduces the same decision structure at



(a) Predicted probability distribution for the 5,000-CpG panel.

(b) Predicted probability distribution for the 50-CpG panel.

Figure 6.3: Score distributions for Normal and Adjacent samples in GSE69914.

100-fold lower dimensionality. The AUC reduction ($0.78 \rightarrow 0.65$) reflects the loss of loci with marginal discriminative weight that contributed to the ranking surface but not to the operating threshold. The score distributions (Figure 6.3) confirm that neither panel version achieves confident class separation.

Separation along the Tumor axis To verify that the Normal–Adjacent–Tumor biological axis is preserved in the feature space, the same KNN classifier trained to discriminate Normal from Adjacent tissue is evaluated on a Normal (test split) versus Tumor (all available samples) comparison, without any retraining. This zero-shot evaluation yields AUC = 1.000 and mean $|\Delta M| = 1.84$, confirming that the dataset harbours a strong tumour-driven methylation signal that projects clearly onto the learned discriminative axis. The comparatively subtle performance on the Normal–Adjacent task therefore reflects the biological attenuation of the field-cancerisation signal rather than a failure of the feature space, consistent with the model of progressive epigenetic drift discussed in Chapter 2.

Gene-level coherence and functional enrichment The 50-CpG panel maps to 40 unique genes. The component-wise decomposition of the MILP objective (Figure 6.2) shows that most selected loci carry contributions from all three statistical components ($w_{\text{coef}}\tilde{c}$, $w_{\Delta M}\tilde{d}$, $w_{\sigma}\tilde{s}$), with a subset receiving additional weight from the COSMIC indicator. The relationship between classifier importance and biological effect size is visualised in Figure 6.4: COSMIC-annotated loci (squares) span a wide range of $\Delta\beta$ values and are not systematically co-localised with the highest-importance region, consistent with the weak Spearman correlation ($\rho = -0.22$) across the full panel. Among the COSMIC-annotated loci, six genes warrant biological commentary.

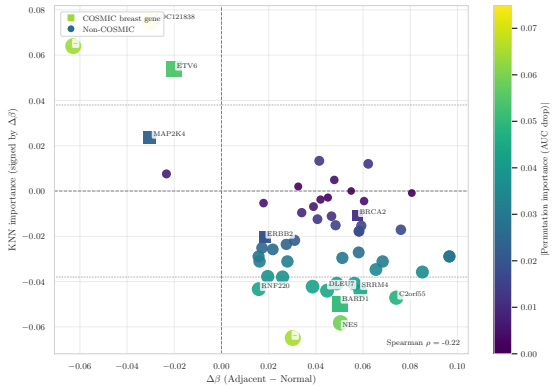


Figure 6.4: KNN permutation importance (signed by $\Delta\beta$) versus $\Delta\beta$ (Adjacent − Normal) for the 50 selected loci. COSMIC breast cancer genes (squares) are annotated by gene name. Spearman $\rho = -0.22$ in GSE69914.

BRCA2 (cg26458617, hypomethylated in Adjacent): a tumour suppressor central to homologous recombination repair of double-strand DNA breaks; germline mutations substantially increase lifetime risk of breast and ovarian cancer, and promoter methylation of *BRCA2* has been linked to the BRCAness phenotype in sporadic tumours, potentially conferring sensitivity to PARP inhibitor therapy [85].

BARD1 (cg08563601, hypermethylated in Adjacent): the obligate heterodimeric partner of BRCA1 in the RING-domain complex involved in DNA repair and apoptosis; loss-of-function variants in *BARD1* confer a moderate-penetrance risk for breast cancer (OR ≈ 2.90), with particularly elevated risk in triple-negative breast cancer [86].

MAP2K4 (cg02746232, hypermethylated in Adjacent): a MAPKK-family kinase that activates both the JNK and p38 stress-response axes; its role in breast cancer is context-dependent, with homozygous deletions and loss-of-function mutations supporting a tumour-suppressive function, while overexpression has also been reported to promote invasion via PI3K/AKT activation [87].

ERBB2 (cg26041593, hypermethylated in Adjacent): the *ERBB2* proto oncogene encodes the HER2 receptor tyrosine kinase at chromosome 17q12; amplification and overexpression are among the most established oncogenic drivers in breast cancer, conferring aggressive tumour behaviour and therapeutic susceptibility to targeted anti-HER2 agents [88].

ASPM (cg03750742, hypermethylated in Adjacent): a centrosomal spindle-assembly protein whose elevated expression in breast cancer is associated with advanced tumour grade, poor prognosis, and reduced relapse-free and overall survival [89].

ETV6 (cg11375112, hypomethylated in Adjacent): a transcriptional repressor

Table 6.3: Performance metrics — Linear SVM, 5,000 CpGs (GSE225845, internal test).

AUC	BAcc	Precision	Recall	F1	MCC
0.9565	0.9030	0.9259	0.8929	0.9091	0.8034

of the ETS family; in breast cancer, *ETV6* is rearranged in secretory carcinoma through the *ETV6–NTRK3* fusion, and loss-of-function alterations have been implicated in tumour progression across multiple tissue types [90].

Cross-dataset transferability External evaluation yields near-random discrimination on both target cohorts: AUC = 0.517 on GSE225845 and AUC = 0.498 on GSE287331, with balanced accuracy collapsing to 0.5 and zero recall for the Adjacent class in both cases. The failure is symmetric across targets, ruling out a directional mismatch specific to one partner cohort and suggesting instead that the discriminative direction identified in GSE69914 does not project onto a shared epigenetic axis — a finding consistent with the low pairwise $\Delta\beta$ concordance documented in Chapter 5.

6.3.2 Dataset GSE225845

GSE225845 presents strong intra-cohort linear separability, with the 5,000-CpG panel achieving high AUC and balanced class discrimination. The 50-CpG compressed panel largely preserves this structure, and the cohort yields informative cross-dataset transferability results given its intermediate position along the field-cancerisation signal axis.

Linear separability of the 5,000-CpG panel A linear SVM is applied to the full 5,000-CpG panel under the stratified 80/20 split. Test-set performance is reported in Table 6.3. The confusion matrix $\left[\begin{smallmatrix} 21 & 2 \\ 3 & 25 \end{smallmatrix}\right]$ confirms balanced discrimination across both classes; the high MCC (0.80) rules out class-imbalance artefacts. The PCA projection (Figure 6.5) and silhouette coefficient of 0.09 show a well-separated class geometry, in sharp contrast to the overlapping structure observed in GSE69914.

Compression to 50 CpGs via MILP Following the intra-dataset SVM training, the learned coefficient magnitudes — alongside the absolute ΔM effect sizes and the inverse baseline variability scores — are fed as input to the MILP formulation of Section 6.2.3. The resulting 50-CpG panel, selected under $\mu = 0.5$ and $w_{\text{COSMIC}} = 1$, is shown in Figure 6.6 with its component-wise utility decomposition. A linear

Table 6.4: Performance metrics — Linear SVM, 50 CpGs (GSE225845, internal test).

AUC	BAcc	Precision	Recall	F1-score	MCC
0.9363	0.8851	0.9231	0.8571	0.8889	0.7666

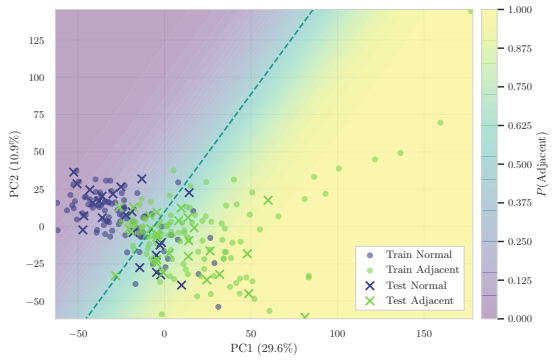


Figure 6.5: Projection of the calibrated SVM decision surface onto PC1–PC2 (29.6% and 10.9% variance explained).

SVM retrained exclusively on these 50 loci yields the metrics in Table 6.4. The AUC decreases modestly while MCC remains high (0.77), with the confusion matrix $\begin{bmatrix} 21 & 2 \\ 4 & 24 \end{bmatrix}$ showing only one additional false negative. Unlike GSE69914, the stacked decomposition reveals that the COSMIC component dominates the objective for the majority of loci, with statistical components providing secondary contributions. The score distributions (Figure 6.7) confirm that both panel versions achieve confident class separation.

Separation along the Tumor axis The same 50-CpG SVM is evaluated on a Normal versus Tumor comparison without retraining, yielding AUC = 0.9781 (five-fold CV). The silhouette coefficient rises from 0.09 to 0.13. This gradient confirms that the selected loci lie on the biological continuum of progressive epigenetic drift.

Gene-level coherence and functional enrichment The 50-CpG panel maps to 42 unique genes. The component-wise decomposition (Figure 6.6) shows the COSMIC indicator dominating the objective for most loci. The relationship between classifier importance and biological effect size is visualised in Figure 6.8: a positive Spearman correlation ($\rho = 0.38$) indicates partial concordance between univariate effect size and multivariate discriminative weight, in contrast to the misalignment observed in GSE69914. Among the COSMIC-annotated loci, four genes warrant biological commentary.

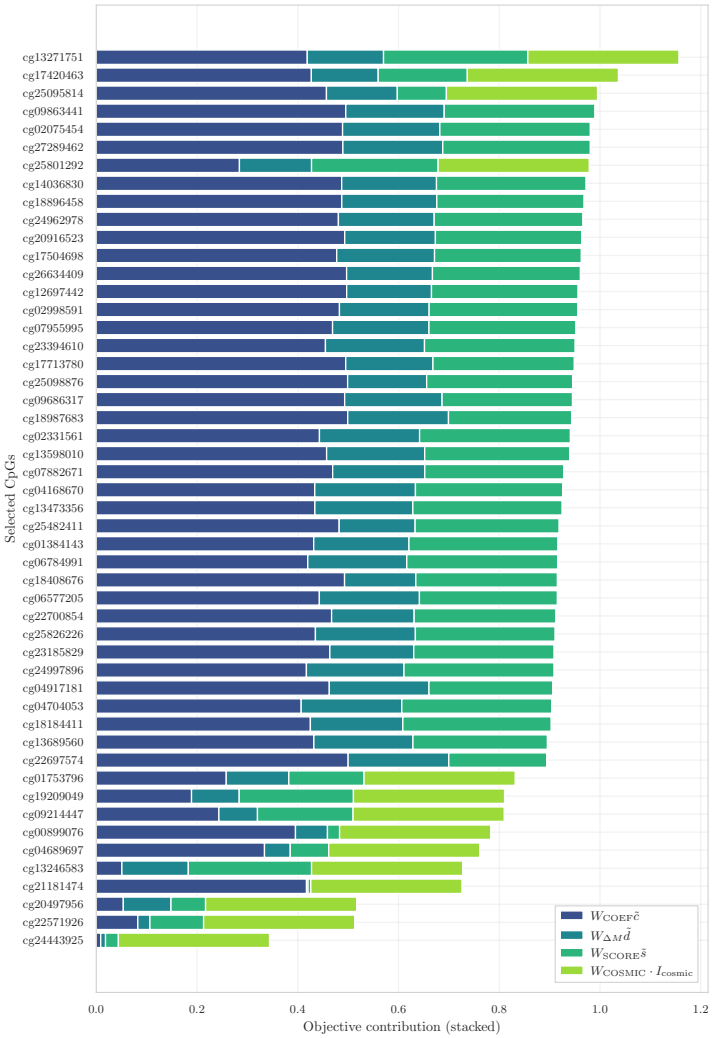
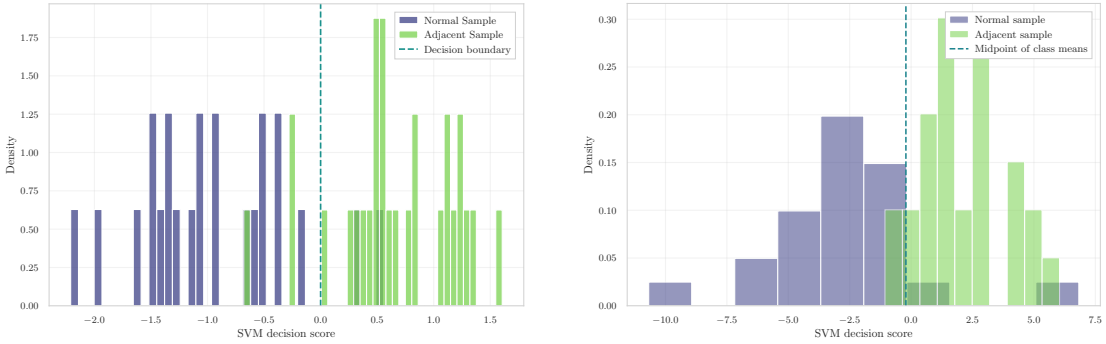


Figure 6.6: Component-wise decomposition of the MILP objective for the 50-CpG panel in GSE225845. Each bar is one selected locus, sorted by total utility; stacked segments show contributions from $w_{\mathrm{coef}}\tilde{c}$, $w_{\Delta M}\tilde{d}$, $w_{\sigma}\tilde{s}$, and $w_{\mathrm{COSMIC}} \cdot \mathbf{1}_{\mathrm{cosmic}}$.

CASP8 (cg25095814, hypomethylated in Adjacent): an initiator caspase and critical effector of the extrinsic apoptosis pathway; promoter methylation of *CASP8* silences its expression in breast cancer, and demethylating agents restore apoptotic sensitivity [91].

MAP3K13 (cg17420463 hypomethylated, cg20497956 hypermethylated): a leucine-zipper kinase activating NF-κB and JNK signalling; in breast cancer, *MAP3K13* overexpression stabilises Myc transcriptional activity, with high expression correlating with poor survival in ER-negative, Myc-high tumours [92].



(a) SVM decision score distribution for the 5,000-CpG panel.

(b) SVM decision score distribution for the 50-CpG panel.

Figure 6.7: Score distributions for Normal and Adjacent samples in GSE225845.

Figure 6.8: SVM coefficient (signed) versus $\Delta\beta$ (Adjacent − Normal) for the 50 selected loci. COSMIC breast cancer genes (squares) are annotated by gene name. Spearman $\rho = 0.38$ in GSE225845.

CCND1 (cg13271751 hypermethylated, cg19209049 hypomethylated): the gene encoding Cyclin D1, core regulator of the G1/S cell-cycle transition; amplified at 11q13 in approximately 15–20% of breast cancers and among the most recurrently altered oncogenes in luminal subtypes [93].

MAP3K1 (cg22571926, hypermethylated): a MAPKKK-family kinase whose loss-of-function mutations are among the most frequent somatic alterations in luminal A breast cancer, associated with improved overall survival in hormone-receptor-positive disease [94].

The remaining COSMIC-annotated loci — *NOTCH1* (cg21181474, hyper), *FOXA1* (cg00899076, hypo), *NCOR1* (cg04689697 hypo, cg24443925 hyper), *KEAP1* (cg25801292, hypo), *PIK3CA* (cg01753796, hyper), *ARID1A* (cg09214447,

Table 6.5: Performance metrics — Linear SVM, 5,000 CpGs (GSE287331, internal test).

AUC	BAcc	Precision	Recall	F1	MCC
1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

hypo), and *SALL4* (cg13246583, hyper) — are established breast cancer genes whose roles are well characterised in the literature [95].

Cross-dataset transferability External evaluation yields near-random or inverted discrimination on both target cohorts: AUC = 0.578 on GSE69914 and AUC = 0.029 on GSE287331, with balanced accuracy collapsing to 0.5 and zero positive predictions in both cases. The collapse on GSE287331 to AUC ≈ 0 suggests systematic score inversion rather than mere signal attenuation.

6.3.3 Dataset GSE287331

GSE287331 occupies the opposite end of the separability spectrum relative to GSE69914. The Normal–Adjacent contrast is characterised by a large mean $|\Delta M|$ and a high silhouette coefficient, yielding perfect internal classification at both 5,000 and 50 CpGs. Cross-dataset transfer fails completely, confirming that the discriminative axis learned in this cohort is structurally inverted relative to the shared drift orientation of the other two datasets.

Linear separability of the 5,000-CpG panel A linear SVM is applied to the full 5,000-CpG panel under the stratified 80/20 split. Test-set performance is reported in Table 6.5. The confusion matrix shows zero misclassifications across both classes. The test partition comprises 49 samples (39 Normal, 10 Adjacent), reflecting the class imbalance of the full cohort. Perfect classification of all 10 held-out Adjacent samples is consistent with the large effect size documented above; the limited cardinality of the minority class means, however, that this result should be interpreted as a lower bound on generalisation error rather than a precise estimate of out-of-sample performance. Perfect internal performance here is not indicative of overfitting: the silhouette coefficient of 0.40 and mean $|\Delta M| = 2.40$ document a genuinely large effect size. The PCA projection (Figure 6.9) confirms a two-cluster geometry with a wide, uncontested margin along PC1 (50.7%, PC2 5.7%).

Compression to 50 CpGs via MILP Following the intra-dataset SVM training, the learned coefficient magnitudes — alongside the absolute ΔM effect sizes

Table 6.6: Performance metrics — Linear SVM, 50 CpGs (GSE287331, internal test).

AUC	BAcc	Precision	Recall	F1	MCC
1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

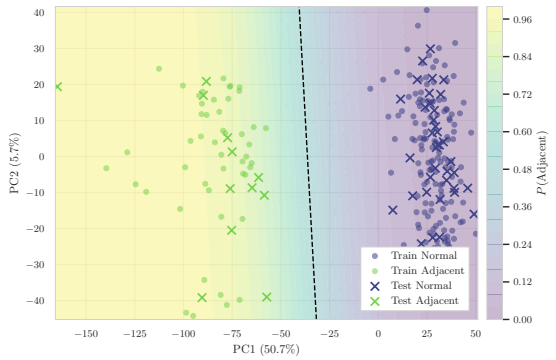


Figure 6.9: Projection of the calibrated SVM decision surface onto PC1–PC2 (50.7% and 5.7% variance explained).

and the inverse baseline variability scores — are fed as input to the MILP formulation of Section 6.2.3. The resulting 50-CpG panel, selected under $\mu = 1.0$ and $w_{\text{COSMIC}} = 0.7$, is shown in Figure 6.10 with its component-wise utility decomposition. A linear SVM retrained exclusively on these 50 loci yields the metrics in Table 6.6. Complete separability is preserved at 100-fold compression, demonstrating that the discriminative signal is intrinsically low-dimensional. Unlike GSE225845, the stacked decomposition shows that $w_\sigma \tilde{s}$ (inverse baseline variability) dominates the objective for most loci, with the COSMIC component providing secondary contributions. The score distributions (Figure 6.11) confirm complete class separation with no density overlap in both panel sizes.

Separation along the Tumor axis The same 50-CpG SVM is evaluated on a Normal (test split) versus Tumor (all available samples) comparison without retraining, yielding AUC = 1.000, silhouette = 0.44, and mean $|\Delta M| = 1.98$.

Gene-level coherence and functional enrichment The 50-CpG panel maps to 42 unique genes. The component-wise decomposition (Figure 6.10) shows $w_\sigma \tilde{s}$ dominating for most loci, consistent with the strong low-variability signal in this cohort. The relationship between classifier importance and biological effect size is visualised in Figure 6.12: a weak negative Spearman correlation ($\rho = -0.19$) mirrors the pattern observed in GSE69914, indicating that discriminative and

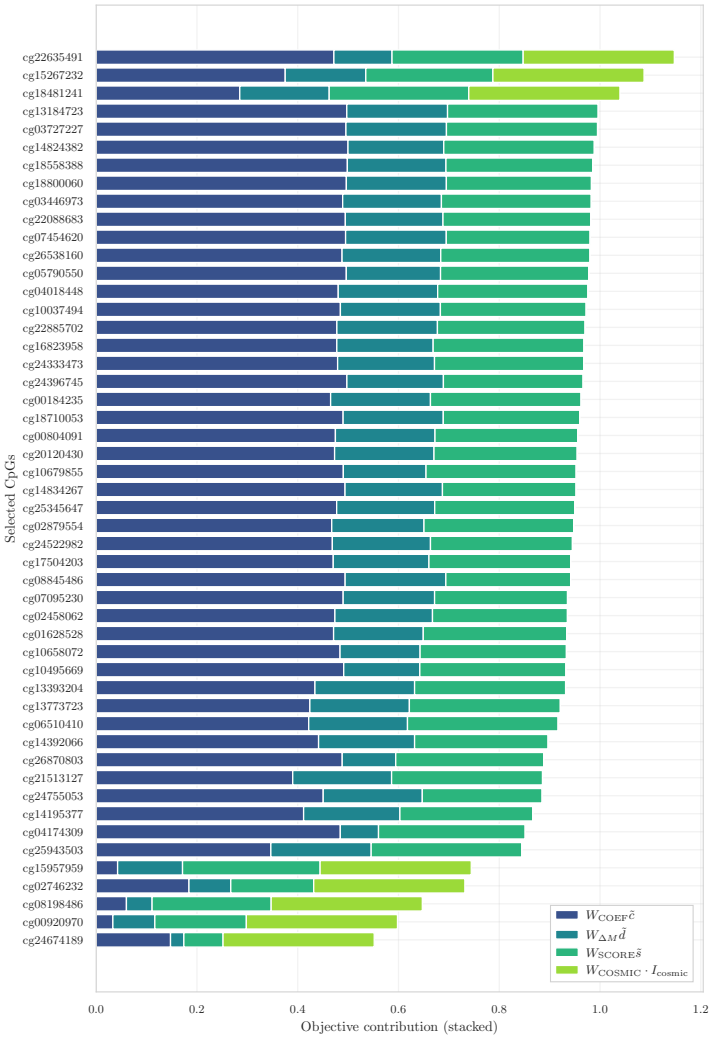
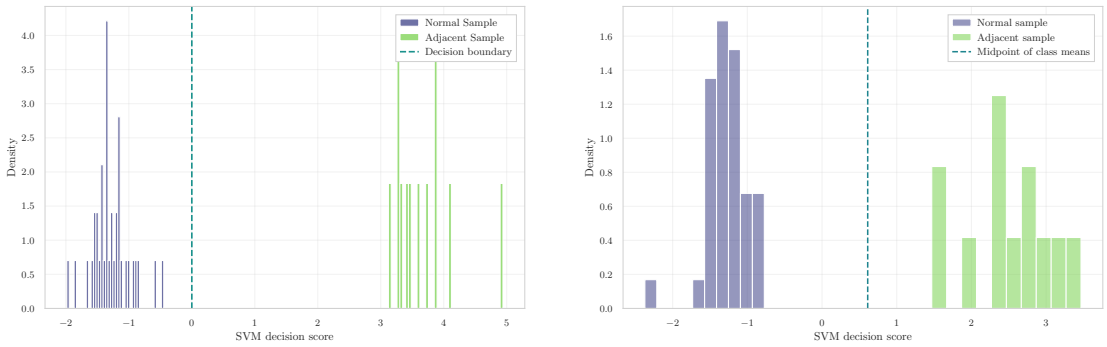


Figure 6.10: Component-wise decomposition of the MILP objective for the 50-CpG panel in GSE287331. Each bar is one selected locus, sorted by total utility; stacked segments show contributions from $w_{\mathrm{coef}}\tilde{c}$, $w_{\Delta M}\tilde{d}$, $w_{\sigma}\tilde{s}$, and $w_{\mathrm{COSMIC}} \cdot \mathbf{1}_{\mathrm{cosmic}}$.

biologically drifted loci are not well aligned. Among the COSMIC-annotated loci, four genes warrant biological commentary.

TBX3 (cg22635491, hypomethylated in Adjacent): a T-box transcriptional repressor overexpressed in breast cancer, where it promotes proliferation, invasion, and cancer stem cell expansion by bypassing senescence through repression of $\mathrm{p14}^{\mathrm{ARF}}$ and $\mathrm{p21}^{\mathrm{WAF1}}$; high *TBX3* expression correlates with shorter metastasis-free survival [96].

GATA3 (cg15267232, hypomethylated in Adjacent): the master transcriptional



(a) SVM decision score distribution for the 5,000-CpG panel.

(b) SVM decision score distribution for the 50-CpG panel.

Figure 6.11: Score distributions for Normal and Adjacent samples in GSE287331.

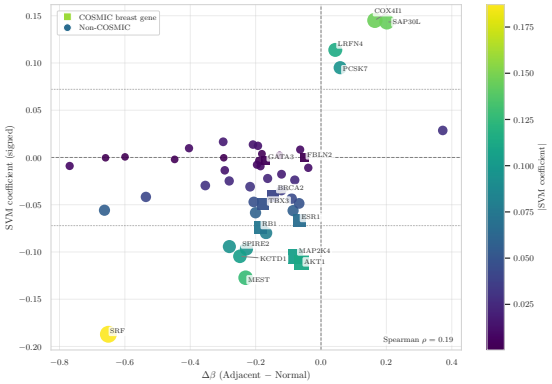


Figure 6.12: SVM coefficient (signed) versus $\Delta\beta$ (Adjacent − Normal) for the 50 selected loci. COSMIC breast cancer genes (squares) are annotated by gene name. Spearman $\rho = -0.19$ in GSE287331.

regulator of luminal cell differentiation in the mammary gland; loss of *GATA3* expression marks progression from well-differentiated to poorly differentiated, highly metastatic breast tumours, and *GATA3* is among the most frequently mutated genes in luminal breast cancer [97].

RB1 (cg18481241, hypomethylated in Adjacent): the retinoblastoma tumour suppressor, a master regulator of the G1/S cell-cycle checkpoint; loss of *RB1* function through deletion, mutation, or silencing is a recurrent event in breast cancer, particularly in triple-negative and basal-like subtypes [95].

MAP2K4 (cg02746232, hypomethylated in Adjacent): previously discussed in the context of GSE69914 (Section 6.3.1); its selection in this independent cohort with concordant hypomethylation reinforces the cross-cohort relevance of this locus.

FBLN2 (cg08198486, hypomethylated in Adjacent): fibulin-2, a secreted extracellular matrix glycoprotein required for basement membrane integrity in mammary epithelium; its loss has been associated with breast cancer invasion, and its expression varies systematically across molecular subtypes [98]. The hypomethylation observed here is consistent with ECM remodelling in the tumour-adjacent microenvironment [99].

The remaining COSMIC-annotated loci — *BRCA2* (cg24674189, hyper), *AKT1* (cg15957959, hypo), and *ESR1* (cg00920970, hypo) — are established breast cancer genes whose roles are well characterised in the literature [95].

Cross-dataset transferability Despite perfect internal classification, external evaluation reveals complete loss of discriminative ability: AUC = 0.436 on GSE69914 and AUC = 0.510 on GSE225845, with balanced accuracy collapsing to 0.5 and zero recall for the Adjacent class in both cases. Performance on GSE69914 falls below random expectation (AUC < 0.5), consistent with directional inversion of the learned separating hyperplane. The dichotomy between perfect internal separability and zero external transfer reflects the coexistence of a strong but cohort-specific methylation signal with a structural misalignment that prevents cross-cohort generalisation.

6.4 Inter-Dataset Structure and Batch–Biology Diagnostics

The intra-dataset results of Sections 6.3.1–6.3.3 establish that each cohort supports robust within-cohort discrimination, but that no trained classifier transfers successfully to an external dataset. This section quantifies the relative contribution of batch and biological variance to the joint feature space, and characterises the directional structure of the Normal–Adjacent drift across cohorts.

PC1 as a Diagnostic Axis The discriminative power of PC1, computed on the union panel of all three cohorts, is evaluated for within-dataset phenotype separation and between-dataset separation (Figure 6.13). In all pairwise comparisons, batch-separation AUC meets or exceeds the corresponding phenotype-separation AUC, confirming that dataset origin dominates the leading variance axis. Any cross-cohort modelling strategy operating directly in this space will absorb cohort effects before biological signal.

Directional Concordance of Methylation Shifts Pairwise concordance of ΔM (Adjacent – Normal) is assessed via Spearman correlation and sign agreement on the shared CpG intersection (Table 6.7, Figures 6.14–6.15). GSE69914 and

Table 6.7: Pairwise ΔM concordance across cohorts.

Pair	Spearman ρ	Sign concordance
GSE69914–GSE225845	0.08	0.76
GSE69914–GSE287331	0.49	0.50
GSE225845–GSE287331	0.38	0.42

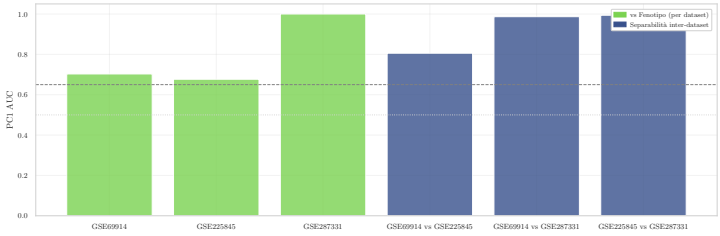
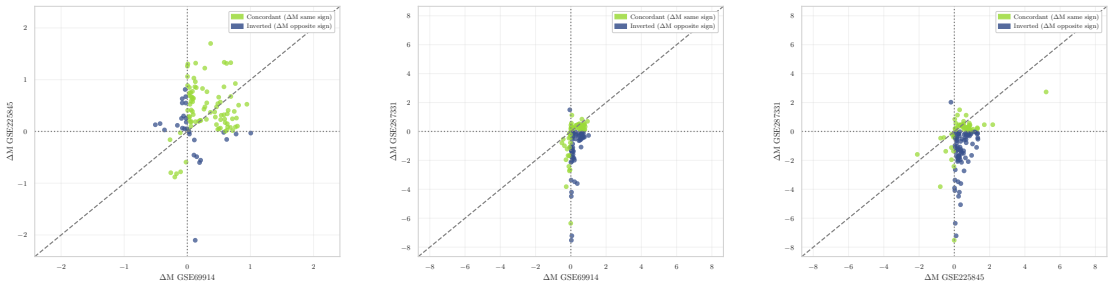


Figure 6.13: PC1 discriminative power for phenotype separation (Normal vs. Adjacent) and batch separation (cohort identity), reported as AUC.

GSE225845 share a consistent drift orientation despite differing in effect magnitude, consistent with a common biological axis modulated by cohort-specific noise. Pairs involving GSE287331 show sign concordance at or below 50%, indicating systematic polarity inversion — a structural property that cannot be resolved by scaling or harmonisation alone.

6.5 Synthesis and Implications

The results of this chapter establish three converging findings. Within each cohort, the Normal–Adjacent contrast is learnable: AUC ranges from 0.78 (GSE69914) to 1.00 (GSE287331) on the 5,000-CpG panel, and compression to 50 CpGs via MILP preserves the majority of discriminative structure while concentrating loci on biologically annotated genes. Across cohorts, however, no trained classifier transfers successfully: AUC collapses to near-random or inverted values in all six directed transfer experiments, with the 50-CpG panels sharing zero loci across datasets. This failure is not attributable to overfitting or weak signal. The PC1 diagnostics show that batch variance dominates the joint feature space, consistently exceeding phenotype variance along the leading principal direction. The directional concordance analysis further reveals that GSE287331 exhibits systematic polarity inversion of the ΔM axis relative to the other two cohorts — a structural property that cannot be resolved by scaling or simple harmonisation. These two obstacles together define the *domain shift* problem in this setting: models trained on one methylation cohort learn a separating hyperplane that is cohort-specific, and the learned axis neither aligns with the biological drift



(a) GSE69914 vs. GSE225845 ($\rho = 0.08$, $c = 0.76$).

(b) GSE69914 vs. GSE287331 ($\rho = 0.49$, $c = 0.50$).

(c) GSE225845 vs. GSE287331 ($\rho = 0.38$, $c = 0.42$).

Figure 6.14: Pairwise ΔM concordance scatter plots on the shared CpG intersection. Green = concordant sign, blue = inverted sign.

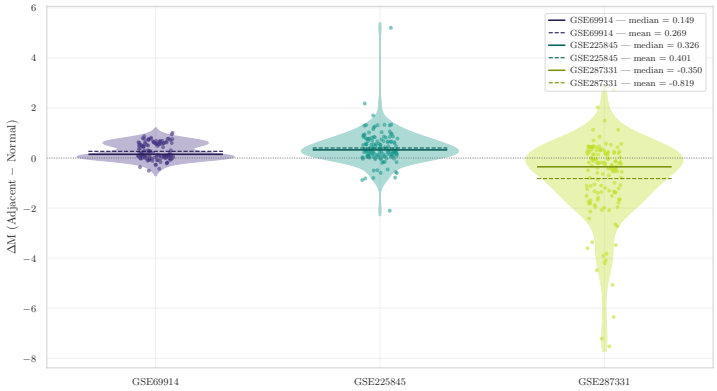


Figure 6.15: Distribution of ΔM (Adjacent – Normal) across the three cohorts, stratified by sign-concordance group.

direction of external cohorts nor survives the dominant batch component of the joint variance structure [25]. Addressing both obstacles simultaneously requires an approach that explicitly disentangles cohort variance from phenotype variance while remaining agnostic to the polarity of the drift direction. The NeuroCombat harmonisation framework [100] provides a principled starting point by removing additive and multiplicative batch effects while preserving biological covariance, operating directly on DNA methylation β-values without requiring a reference batch. Chapter 7 develops this strategy, evaluating whether a harmonised joint representation recovers a shared discriminative axis and whether cross-cohort transfer improves from the near-random baseline documented here.

Chapter 7

Multi-Cohort Integration and Joint Modelling

7.1 Rationale for Multi-Cohort Integration

The single-cohort analyses presented in Chapters 5 and 6 established that Normal–Adjacent methylation differences are detectable within each dataset independently, yet the cross-cohort transfer experiments of Section 6.2.2 revealed a fundamental obstacle: the discriminative axes identified in isolation are not mutually transferable, and in one direction performance inverts below chance level. This failure is attributable to two concurrent factors. First, the three cohorts originate from different Illumina platforms (HumanMethylation450K for GSE69914, MethylationEPIC for GSE225845 and GSE287331), introducing systematic probe-design and coverage differences. Second, and more critically, the uncorrected pool is dominated by inter-study variance that absorbs the comparatively subtle biological signal of early epigenetic drift. The present chapter addresses both obstacles simultaneously by constructing a unified analytical framework in which all three cohorts are harmonised, jointly preprocessed, and modelled as a single integrated dataset. The objective is not external validation — which would require a fully held-out study — but the assessment of signal stability under controlled multi-cohort pooling, where biological heterogeneity and platform variability coexist within a single analytical setting. A discriminative panel that retains predictive ability in this pooled regime can be considered cohort-invariant to a meaningful degree, encoding shared biological processes rather than study-specific artefacts.

7.2 Multi-Cohort Integration Framework

This section describes the sequence of operations applied to construct the harmonised multi-cohort representation. The pipeline proceeds in three ordered steps — feature-space alignment, absolute deviation transformation, and batch harmonisation — each designed to address a specific obstacle to cross-cohort integration while maintaining strict separation between training and test data.

7.2.1 Construction of a Shared CpG Feature Space

Integration begins with the construction of a shared feature space. After applying the per-dataset preprocessing pipelines described in Chapter 4 — including technical probe filtering, sex-chromosome exclusion, and manifest-based annotation validation — the three cleaned CpG sets are intersected. The resulting common space comprises 215,919 autosomal CpGs consistently retained across all three platforms after filtering, forming the backbone for all downstream operations. Before any transformation or modelling step, a single stratified train–test split is performed on the aggregated pool of 587 samples. Stratification is carried out jointly on dataset identity and tissue label, yielding a training set of 410 samples and a test set of 177 samples, with class and cohort proportions preserved in both partitions. This split is fixed for the entire chapter; the test set is never consulted during preprocessing, transformation, feature selection, or model training, guaranteeing a strictly leakage-free evaluation.

7.2.2 Absolute Deviation Transformation

A key challenge identified in Chapter 6 is the directional inconsistency of methylation drift across cohorts: the sign of ΔM (Adjacent − Normal) is not uniform, with GSE287331 exhibiting a predominantly opposite direction relative to the two EPIC cohorts (Figure 6.14). Training a standard linear classifier on signed M-values under these conditions forces the model to learn contradictory decision boundaries, suppressing cohort-invariant signal. To resolve this, each sample is transformed into its absolute deviation from the within-cohort Normal mean before pooling. Formally, for each dataset d and each CpG j, the transformed feature is:

$$\tilde{x}_{ij} = \left| x_{ij} - \bar{\mu}^{(d)}_{\text{Normal},j} \right|, \tag{7.1}$$

where $\bar{\mu}^{(d)}_{\text{Normal},j}$ is the mean β-value of Normal samples in the training partition of cohort d. This mean is estimated exclusively on the training split of each dataset and applied without re-estimation to the corresponding test samples, preserving the leakage-free design. Under the field cancerisation hypothesis, Adjacent tissue

is expected to deviate more from the Normal reference than Normal tissue itself, regardless of drift direction; the absolute deviation representation makes this deviation comparable across cohorts with opposite drift orientations.

7.2.3 Batch Harmonisation via NeuroCombat

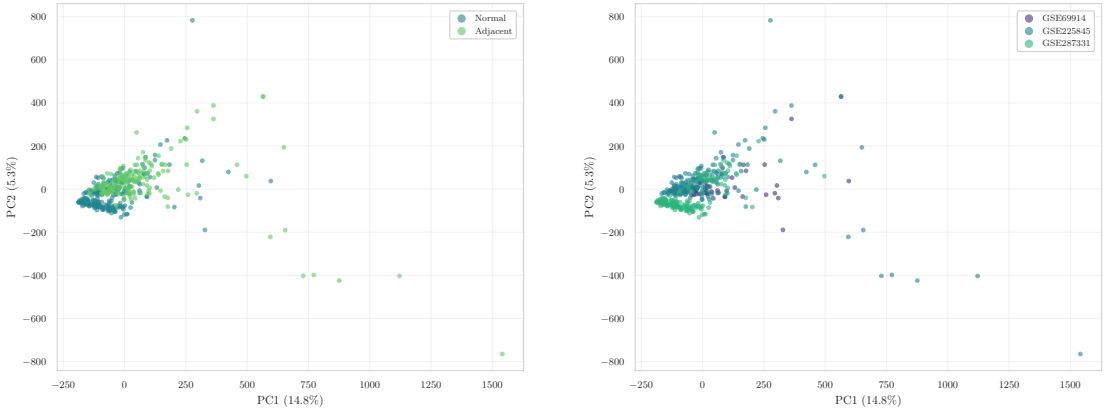
Even after the absolute deviation transformation, residual inter-study variance persists in the pooled training matrix, reflecting differences in sample preparation, technical processing, and cohort-specific biological composition. To remove these additive and multiplicative batch effects while preserving biologically relevant covariance, the NeuroCombat harmonisation framework [100] is applied to the pooled training set of 410 samples. NeuroCombat models each feature as:

$$Y_{ijd} = \alpha_j + X\beta_j + \gamma_{jd} + \delta_{jd}\varepsilon_{ijd}, \tag{7.2}$$

where γ_{jd} and δ_{jd} are the additive and multiplicative batch parameters for cohort d, estimated via empirical Bayes shrinkage. The biological covariate matrix X includes tissue label (Normal vs. Adjacent) as a categorical variable and chronological age as a continuous covariate. Age is included because DNA methylation varies with age in a systematic, locus-specific manner [15]; including it as a protected covariate prevents ComBat from absorbing age-related methylation variance into the batch correction. The chronological age used here is derived from the metadata of each cohort; its estimation and quality assessment are detailed in Appendix C. ComBat parameters are estimated exclusively on the training pool. The resulting estimates are then applied to the test set via `neuroCombatFromTraining`, which applies the pre-fitted transformation without re-estimating any parameters — a strict zero-leakage design.

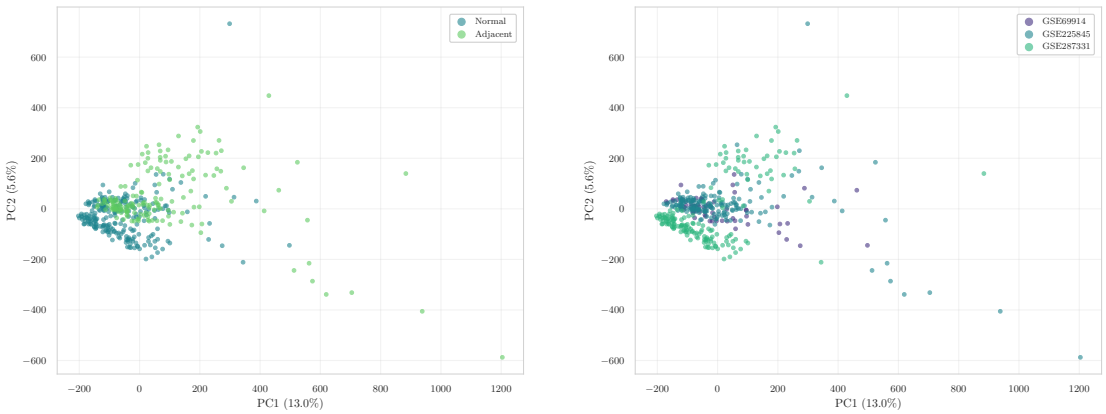
7.3 Batch Harmonisation Diagnostics

The effectiveness of the harmonisation pipeline is assessed through principal component analysis of the pooled training matrix, computed before and after ComBat correction. Before harmonisation (Figure 7.1), the first principal component captures 14.8% of variance and is dominated by cohort identity: GSE287331 samples form an elongated scatter extending along PC1, consistent with the strong platform-specific signal documented in earlier chapters. When the same projection is coloured by tissue label, no Normal–Adjacent separation is discernible, confirming that the biological axis is masked by batch-driven variance. After NeuroCombat correction (Figure 7.2), the cohort-driven elongation is substantially reduced. The three datasets mix more uniformly in the PC1–PC2 plane, and a partial separation by tissue label begins to emerge along PC1 (13.0%). The modest but consistent



(a) Coloured by tissue label. **(b)** Coloured by cohort of origin.

Figure 7.1: PCA of the pooled training set *prior* to NeuroCombat (PC1 = 14.8%, PC2 = 5.3%). Batch variance dominates PC1; no tissue-label separation is visible.



(a) Coloured by tissue label. **(b)** Coloured by cohort of origin.

Figure 7.2: PCA of the pooled training set *after* NeuroCombat (PC1 = 13.0%, PC2 = 5.6%). Cohort-driven elongation is substantially reduced; a partial Normal–Adjacent separation emerges along PC1.

restructuring of the variance landscape confirms that ComBat has attenuated the dominant inter-study component while exposing the underlying biological signal.

7.4 Feature Selection and Predictive Modelling

This section describes the feature selection and classification pipeline applied to the harmonised pooled matrix, and reports the resulting predictive performance on the

held-out test set. The pipeline mirrors the intra-dataset procedure of Chapter 5 and Chapter 6, here adapted to the multi-cohort setting.

7.4.1 Feature Selection on the Harmonised Pool

Following harmonisation, the stability-driven feature selection procedure introduced in Chapter 5 is applied to the pooled training matrix. Edgar β-variability filtering reduces the initial 215,919 CpGs to 106,289, and subsequent $\beta \to M$ transformation with variance-based pruning (bottom 2% SD removed) yields a working set of 104,163 features. Biological weighting assigns higher priority to CpG island and promoter-proximal loci, and repeated stratified K-fold stability ranking ($K = 5$, $R = 10$, 50 splits total) with $S_{\text{dir}} \geq 0.75$ selects 15,000 candidate CpGs, with Spearman $\rho(\text{score}, |\Delta M|) = -0.78$ confirming strong alignment between ranking and effect size. Region anchoring on CpG islands identifies 1,165 stable regions and produces Pool A of 2,740 anchored CpGs. Correlation-based redundancy pruning ($|r| \geq 0.85$) and genomic diversification with chromosomal (max 8% per chromosome) and spatial constraints (max 15 CpGs per 500 kb window, zero violations) yield the final panel of 5,000 CpGs.

7.4.2 Linear SVM Training and Panel Compression

A linear SVM with StandardScaler preprocessing is trained on the 5,000-CpG. Five-fold cross-validation over $C \in \{0.001, 0.01, 0.1, 1.0, 10.0\}$ yields a stable CV AUC of 0.973 ± 0.021 across all regularisation values; $C = 0.1$ is selected. The learned coefficient vector is fed into the MILP knapsack formulation (Section 6.2.3), which compresses the panel to 50 biologically constrained CpGs under joint optimisation of discriminative performance and COSMIC breast-cancer gene enrichment ($\mu = 1.0$, $w_{\text{COSMIC}} = 0.7$, selected via knee-point analysis on the AUC–enrichment sweep). The knapsack increases the number of COSMIC breast-cancer genes from 11 (at $\mu = 0$, pure performance) to 18, the number of CpG island loci from 39 to 42, and the number of promoter-proximal loci from 29 to 32, while mapping to 42 unique genes.

7.4.3 Predictive Performance

Performance on the pooled test set (177 samples: 104 Normal, 73 Adjacent) is reported in Table 7.1. The per-cohort breakdown reveals a heterogeneous picture. GSE287331 achieves perfect separation at both panel sizes, consistent with the strong intra-dataset signal documented in Chapter 6. GSE225845 retains good discriminative ability at 50 CpGs, representing a moderate but acceptable degradation from the 5,000-CpG upper bound. GSE69914, however, remains near

Table 7.1: Performance of the 5,000-CpG SVM and the 50-CpG knapsack panel on the pooled test set, with per-cohort AUC breakdown.

Model	AUC	BACC	AUC GSE69914	AUC GSE225845	AUC GSE287331
SVM 5,000 CpG	0.944	0.849	0.528	0.951	1.000
SVM 50 CpG knapsack	0.902	0.799	0.487	0.827	1.000

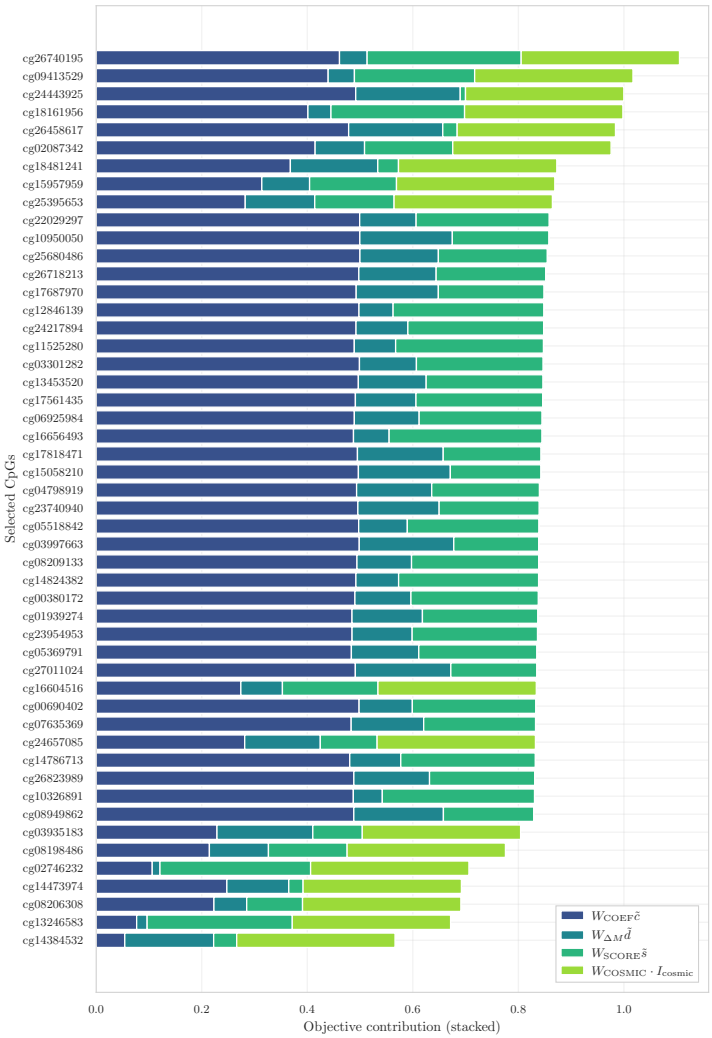


Figure 7.3: Component-wise decomposition of the MILP objective for the 50-CpG multi-cohort panel.

chance level even with the full panel, mirroring the low intra-dataset signal observed in isolation and suggesting that the HM450K cohort contributes structural diversity to the panel without contributing proportional discriminative power.

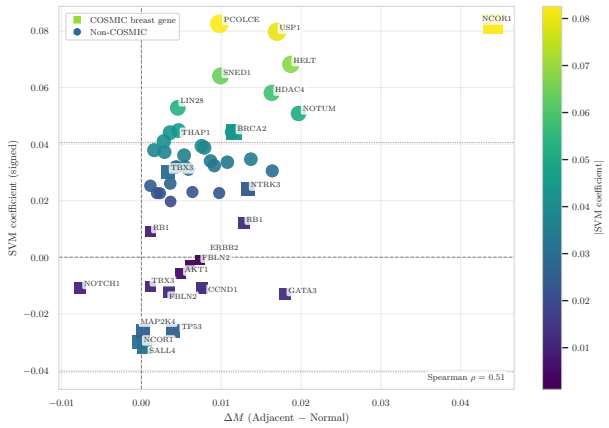


Figure 7.4: SVM coefficient (signed) versus ΔM (Adjacent − Normal, absolute-deviation space) for the 50 selected loci. Spearman $\rho = 0.51$ indicates moderate concordance between discriminative weight and deviation magnitude.

7.5 Biological Interpretation of the 50-CpG Panel

The 50 selected CpGs map to 49 unique genes via Illumina manifest annotation (50/50 CpGs annotated). Of these, 18 overlap with the COSMIC Cancer Gene Census breast-cancer subset (44 genes total), compared to 11 in the unconstrained baseline ($\mu = 0$), demonstrating the effectiveness of the knapsack biological penalty.

The volcano plot (Figure 7.4) shows the relationship between SVM discriminative weight and deviation magnitude for each of the 50 loci. COSMIC breast-cancer genes span the full range of coefficient magnitudes and both directions of the signed axis, consistent with the heterogeneous roles of these genes in epigenetic field reprogramming. Among the previously characterised COSMIC loci, notable examples include TBX3 (hypomethylated, promotes invasion by bypassing senescence [96]), GATA3 (master regulator of luminal identity [97]), RB1 (G1/S checkpoint tumour suppressor [95]), NOTCH1 (developmental signalling, discussed in Section 6.3.2), NCOR1 (transcriptional co-repressor, discussed in Section 6.3.2), BRCA2 (DNA repair, discussed in Sections 6.3.1–6.3.3), TP53, GATA3, AKT1, and SALL4 (discussed in respective single-cohort sections). Three COSMIC loci appear in the multi-cohort panel for the first time and warrant brief commentary. HDAC4 (class IIa histone deacetylase): a transcriptional co-repressor that suppresses CDK inhibitors including p21 and promotes cell-cycle progression; its overexpression has been linked to invasion and angiogenesis in multiple solid tumours through VEGF–HIF-1α signalling [101]. NTRK3 (neurotrophic tyrosine kinase receptor 3): the kinase domain partner in the ETV6–NTRK3 fusion oncogene that is the defining genetic event of secretory breast carcinoma, constitutively activating

RAS–MAPK and PI3K–AKT pathways [90]. LIN28B (RNA-binding protein): an oncogenic suppressor of the let-7 microRNA family that derepresses MYC, RAS, and HMGA2; overexpressed in triple-negative breast cancer and associated with poor differentiation and advanced disease stage [102, 103]. The Spearman $\rho = 0.51$ between signed coefficient and ΔM indicates moderate but imperfect concordance, reflecting the contribution of the knapsack biological penalty terms in selecting loci beyond pure effect size. The panel reflects the multifactorial nature of epigenetic field defects: it spans tumour suppressors (RB1, TP53), chromatin regulators (HDAC4, NCOR1), developmental transcription factors (TBX3, SALL4), signalling components (NOTCH1, NTRK3, AKT1), and an RNA-binding oncogene (LIN28B). The panel reflects the multifactorial nature of epigenetic field defects: it spans tumour suppressors (RB1, TP53), chromatin regulators (HDAC4, NCOR1), developmental transcription factors (TBX3, SALL4), signalling components (NOTCH1, NTRK3, AKT1), and an RNA-binding oncogene (LIN28B). The inclusion of 18 COSMIC breast-cancer genes — compared to 11 in the unconstrained $\mu = 0$ baseline — demonstrates that the knapsack constraints are effective in directing the panel towards biologically interpretable loci without a commensurate loss of discriminative performance.

7.6 Cross-Cohort Epigenetic Drift

The multi-cohort integration framework presented in this chapter demonstrates that early epigenetic drift in breast tissue encodes a cohort-invariant signal, accessible once inter-study batch effects and directional inconsistencies are addressed. The absolute deviation transformation resolves the polarity problem identified in Chapter 6 by recasting the classification problem in terms of deviation magnitude from the Normal reference. NeuroCombat harmonisation then removes residual platform-specific variance, as confirmed by the restructuring of the PCA landscape (Figures 7.1–7.2). Stability-based feature selection on the harmonised pool identifies loci that are consistently informative across heterogeneous study conditions, and the MILP knapsack produces a compact, biologically grounded 50-CpG panel (AUC = 0.902, BACC = 0.799 on the pooled test set) mapping to 18 COSMIC breast-cancer genes at 100-fold compression. The residual weakness on GSE69914 reflects a genuine biological signal limitation rather than a methodological failure: the Normal–Adjacent contrast in this HM450K cohort is inherently subtler, as documented in the intra-dataset analyses, and the harmonised multi-cohort framework cannot recover signal that is absent at source. The strong performance on GSE287331 and GSE225845, spanning two independent EPIC studies, provides meaningful evidence that the selected panel captures a shared epigenetic signature of field cancerisation in breast tissue.

Chapter 8

Conclusion and Future work

8.1 Conclusions

Field cancerisation represents one of the most compelling and, at the same time, methodologically challenging phenomena in cancer epigenomics. Histologically indistinguishable from healthy tissue, tumour-adjacent breast parenchyma harbours early epigenetic alterations that precede overt neoplastic transformation and whose systematic characterisation may ultimately inform both risk stratification and early detection strategies. The central biological question motivating this thesis — whether DNA methylation differences between Normal and Normal-Adjacent tissue are detectable, reproducible, and transferable across independent cohorts under strict statistical control — was addressed through an end-to-end computational pipeline integrating multi-cohort data harmonisation, biologically informed feature selection, and constrained combinatorial optimisation.

The study was built on three publicly available DNA methylation datasets (`GSE69914`, `GSE225845`, `GSE287331`) spanning the Normal – Adjacent – Tumour axis in breast tissue across two Illumina array generations (450K and EPIC). Exploratory analysis confirmed that, at the whole-methylome level, Normal and Adjacent tissues are globally similar, with nearly superimposed β-value distributions and comparable variance profiles. Yet locus-specific analyses consistently revealed a focal early drift in Adjacent samples: outlier burden ratios ranging from 1.35 in `GSE69914` to 13.33 in `GSE287331`, and silhouette scores up to 0.34, provided quantitative evidence for the presence of field effects across all three independent cohorts. This cross-dataset consistency confirms that the Normal – Adjacent signal is not an artefact of a single study design but a reproducible biological phenomenon, coherent with contemporary models of tumour initiation in which locally restricted epigenetic perturbations accumulate in histologically normal tissue prior to overt malignant transformation.

Building on this exploratory foundation, the preprocessing pipeline addressed the principal sources of confounding in multi-cohort methylation studies: probe-design bias, batch effects between the datasets (corrected via NeuroCombat with age as a continuous covariate, estimated strictly on the training set to prevent leakage), and age-dependent methylation structure (quantified through the Horvath multi-tissue epigenetic clock and subsequently controlled via age binning and CpG blacklisting). The transformation of raw β-values into absolute deviations from the Normal reference $|\beta - \mu_{\text{Normal}}|$ provided a biologically grounded input space that resolves the drift polarity heterogeneity identified across cohorts while aligning naturally with the field cancerisation hypothesis.

Feature selection proceeded through a multi-stage stability-based pipeline: empirical β-filtering to remove constitutively non-variable CpGs, variance filtering in M-space, repeated stratified k-fold stability ranking with directional weighting (RSKF, Spearman $\rho = -0.848$ between score and $|\Delta M|$), region anchoring to CpG islands, correlation-based redundancy pruning ($|r| \geq 0.85$), and genomic diversification with chromosomal and spatial constraints. The output of this pipeline — 5 000 stable, non-redundant, genomically diverse CpG sites — constituted the input space for a LinearSVC classifier, whose coefficients were then passed as relevance scores to the main methodological contribution of this work.

The MILP knapsack formulation represents the core novelty of this thesis. Rather than selecting CpG sites by greedily maximising predictive performance, the optimiser explicitly encodes biological prior knowledge as hard and soft constraints: enrichment in COSMIC breast cancer genes (W_{COSMIC}), upper bounds on chromosomal representation (CHR_MAX), spatial diversity (max per 500 kb window), redundancy control (CORR_THR), gene-level diversity (MAX_PER_GENE), and a minimum fraction of hypermethylated sites (MIN_HYPER). Applied to the multi-cohort pooled setting of Chapter 7, the resulting panel of 50 CpG sites achieves AUC = 0.902 on the pooled held-out test set ($n = 177$: 104 Normal, 73 Adjacent), with per-cohort breakdown $\text{AUC}_{\texttt{GSE287331}} = 1.000$, $\text{AUC}_{\texttt{GSE225845}} = 0.827$, $\text{AUC}_{\texttt{GSE69914}} = 0.487$, against the 5 000-CpG upper bound of AUC = 0.944. Of the 50 selected CpGs, 18 overlap with the COSMIC Cancer Gene Census breast-cancer subset, compared to 11 in the unconstrained baseline ($\mu = 0$), demonstrating the effectiveness of the knapsack biological penalty.

The biological interpretation of the panel corroborated the coherence of the constrained selection. The 50 CpGs map to 49 unique genes spanning tumour suppressors (*RB1*, *TP53*), chromatin regulators (*HDAC4*, *NCOR1*), developmental transcription factors (*TBX3*, *SALL4*, *GATA3*), signalling components (*NOTCH1*, *NTRK3*, *AKT1*), DNA repair (*BRCA2*), and an RNA-binding oncogene (*LIN28B*) — a multifactorial portrait of epigenetic field reprogramming that could not have emerged from a purely performance-driven selection. This is the central result of the thesis: *interpretability is not a by-product of performance — it must be*

structurally imposed.

Taken together, these results demonstrate that it is possible to construct a compact, clinically motivated CpG panel that preserves near-baseline classification performance while embedding explicit biological knowledge. The MILP knapsack framework is fully parametrisable: the constraint structure can be adapted to any tumour type or epigenetic phenotype for which a reference cancer gene database exists, making the approach general beyond the breast cancer context explored here.

8.2 Future Work

Several directions naturally extend the present work, ranging from immediate experimental validations to broader methodological generalisations.

External validation on independent cohorts The most pressing limitation of this study is the absence of evaluation on a dataset not involved in any step of the pipeline design. The three cohorts employed here were all used in training and feature selection, albeit with strict leakage controls. Identifying an additional dataset with Normal and Adjacent breast tissue samples — a non-trivial task given the scarcity of publicly available adjacent-normal methylation data with consistent phenotype annotation — would provide a genuine out-of-distribution test of the 50-CpG panel. Cross-platform validation, extending the panel to the EPIC v2 array ($\sim$930k CpG, broader promoter coverage), would further establish its robustness across array generations.

Sensitivity analysis on pipeline and MILP hyperparameters A systematic sensitivity analysis of the feature selection pipeline hyperparameters — including the directional stability threshold, the biological prior weight λ, the correlation redundancy cutoff τ_c, and the diversification target K — would provide empirical grounding for the design choices adopted in Chapter 5. Given the computational cost of re-running stability selection across the full hyperparameter grid, this analysis would benefit from a reduced resampling scheme or surrogate evaluation on a representative subset of CpGs. Analogously, varying each MILP constraint independently (μ, W_{COSMIC}, CHR_MAX, CORR_THR, MAX_PER_GENE, MIN_HYPER) while holding the others fixed would quantify which constraints drive biological enrichment and which are partially redundant, providing practical guidance for adapting the framework to other cancer types.

Replacement of SVM coefficients with transformer attention weights (CpGPT) In the present formulation, the relevance scores fed to the knapsack are

the linear coefficients of a LinearSVC trained on the 5 000 candidate CpGs. While interpretable and computationally efficient, linear coefficients capture only additive, univariate contributions and may miss higher-order co-methylation interactions. A natural extension is to replace the SVM with a transformer-based model pre-trained on large-scale methylation data, such as CpGPT [104], and to use the resulting attention weights as the relevance scores for the MILP objective. The self-attention mechanism [105] learns context-dependent representations where the contribution of each CpG is modulated by its co-methylation context across the genome; feeding these richer relevance estimates into the knapsack would preserve the full biological constraint structure while potentially improving both discriminative power and biological coherence of the selected panel.

Multi-omic integration DNA methylation captures only one layer of the epigenetic regulation underlying field cancerisation. The CpG sites identified by the knapsack panel represent natural candidates for functional validation through complementary omics modalities. Integration with RNA-seq data from matched Normal – Adjacent – Tumour samples would allow assessment of whether hypermethylation at the selected CpGs is associated with transcriptional silencing of the corresponding genes, directly linking the epigenetic signal to downstream expression changes. Similarly, ATAC-seq or ChIP-seq data for histone marks (H3K27me3, H3K4me3) at the selected loci would clarify whether the observed methylation alterations co-occur with chromatin accessibility changes, providing a more complete picture of the epigenetic state of adjacent tissue.

Generalisation to other tumour types and phenotypes The MILP constraint structure is fully parametrisable. Applying the same pipeline to colorectal, lung, or ovarian cancer — tumour types for which large EPIC methylation datasets with normal and adjacent tissue are increasingly available — would test whether biologically constrained CpG panel selection is a generally effective strategy or whether its advantages are specific to the breast cancer context. More broadly, the framework could be adapted to phenotypes other than tissue state classification: age-acceleration groups, molecular subtypes, or treatment response endpoints could serve as the target variable, with the COSMIC constraint replaced by phenotype-relevant gene sets from MSigDB or other curated databases.

8.3 Code and Data Availability

All code developed for this thesis, including preprocessing pipelines, feature selection notebooks, MILP formulation, and evaluation scripts, is publicly available at:

`https://github.com/elisabettaroviera/THESIS`.

Appendix A

Filtering lists

A.1 Probe Filtering Resources

This appendix documents the five external annotation resources adopted for technical probe filtering across all three datasets. Each resource targets distinct classes of measurement artefacts; together they define the union-based exclusion criterion applied in Section 4.2.3. Table A.1 summarises the categories covered by each resource.

A.1.1 Naeem *et al.* (2014)

Naeem *et al.* [50] proposed a structured filtering framework for the HumanMethylation450 array that excludes probes multi-mapping or cross-hybridising to additional loci, probes overlapping repetitive elements (LINE, SINE, Alu), and probes overlapping insertion–deletion polymorphisms. For SNP-affected probes, variants at the interrogated CpG or single-base extension site — which directly compromise the methylation measurement — are excluded, while nearby variants not interfering in bisulfite space are retained.

A.1.2 Chen *et al.* (2013)

Chen *et al.* [51] empirically identified 450K probes exhibiting off-target hybridisation or overlap with common SNPs. Only the cross-reactive probe list is used here, enumerating approximately 29,000 multi-mapping probes whose methylation signal cannot be unambiguously attributed to a single genomic locus.

Table A.1: Technical artefact categories covered by each filtering resource.

Category	Naeem	Chen	Pidsley	Zhou	McCartney
Cross-hybridisation	Yes	Yes	Yes	`MASK_mapping`	Yes
SNP at CpG	Yes	—	Yes	`MASK_snp5`, `MASK_extBase`	—
Nearby SNP	Conditional	—	—	—	—
Structural variant	Yes	—	—	—	—
Non-CpG probes	—	—	Yes	`MASK_nonCG`	Yes

A.1.3 Pidsley *et al.* (2016)

Pidsley *et al.* [29] provide annotated probe lists for the EPIC array flagging three relevant categories: CpG-targeting probes with sequence homology to additional loci; non-CpG-targeting probes with analogous off-target issues; and probes overlapping common genetic variation at the interrogated CpG, the single-base extension site of Type I probes, or within the probe body.

A.1.4 Zhou *et al.* (2016)

Zhou *et al.* [28] released a unified probe annotation for 450K and EPIC arrays using binary mask columns. The masks applied here flag probes with low mapping quality (`MASK_mapping`); Type I probes with a SNP at the extension base causing a colour-channel switch (`MASK_typeINextBaseSwitch`); probes with an inconsistent extension base (`MASK_extBase`); probes with non-unique 30-bp 3′ subsequences (`MASK_sub30.copy`); and probes overlapping SNPs within ±5 bp of the interrogated CpG (`MASK_snp5.common`, `MASK_snp5.GMAF1p`). The composite `MASK_general` field is used as the primary filter; RepeatMasker probes (`MASK_rmsk15`) are retained, consistent with standard practice.

A.1.5 McCartney *et al.* (2016)

McCartney *et al.* [52] assessed probe design artefacts on the EPIC array, publishing lists of cross-hybridising CpG- and non-CpG-targeting probes that complement the Pidsley and Zhou resources with an independent characterisation of off-target hybridisation.

The combined application of these resources ensures that downstream analyses operate on a probe set with maximised hybridisation specificity and minimal confounding by common genetic variation.

Appendix B

Extension of an Empirically Driven Variability-Based Filtering Framework to Breast Tissue

B.1 Motivation and Scope

The variability-based filtering strategy embedded in the preprocessing pipeline of this thesis builds directly on the method proposed by [69], who demonstrated that a substantial fraction of CpG sites interrogated by the Illumina HumanMethylation450K array are effectively non-variable within a given tissue context and can therefore be removed prior to downstream analysis without loss of discriminative information. The original study was conducted on three healthy reference tissues — blood, buccal epithelial cells, and placenta — and the resulting non-variable CpG lists were shown to be robust across preprocessing pipelines and normalisation strategies.

The present appendix documents the extension of this framework to breast tissue. The extension pursues two objectives: (i) to assess whether the non-variability criterion, as originally defined, generalises to a cancer-relevant tissue under increased biological heterogeneity; and (ii) to provide a general, reproducible protocol for incorporating additional tissues into the reference framework in future studies. The analysis operates exclusively on histologically normal samples to preserve comparability with the original reference tissues and to estimate baseline methylation stability in the absence of disease-related confounding. Importantly, this appendix does not modify the non-variability definition of [69]: a CpG is

classified as non-variable if its inter-quantile range $r_{\beta,j} = Q_{0.90}(\beta_j) - Q_{0.10}(\beta_j) < 0.05$. The contribution is therefore evaluative and integrative rather than definitional.

B.2 General Framework for Cross-Tissue Extension

Let $\mathcal{T} = \{T_1, \ldots, T_k\}$ denote a set of reference tissues, each represented by one or more independent methylation datasets from normal samples. For each tissue T_ℓ, the tissue-specific non-variable CpG set is defined as:

$$\mathcal{N}_\ell(\tau) = \left\{ j \in \{1, \ldots, p\} : Q_{0.90}\left(\beta_j^{(\ell)}\right) - Q_{0.10}\left(\beta_j^{(\ell)}\right) < \tau \right\}, \tag{B.1}$$

where $\tau = 0.05$ is the threshold inherited from [69]. The cross-tissue invariant core — CpGs non-variable across all reference tissues simultaneously — is given by the intersection:

$$\mathcal{N}_{\text{core}}(\tau) = \bigcap_{\ell=1}^{k} \mathcal{N}_\ell(\tau). \tag{B.2}$$

When a new tissue T_{k+1} is incorporated, the invariant core is updated as:

$$\mathcal{N}'_{\text{core}}(\tau) = \mathcal{N}_{\text{core}}(\tau) \cap \mathcal{N}_{k+1}(\tau), \tag{B.3}$$

which is strictly more conservative than the original core: $\mathcal{N}'_{\text{core}}(\tau) \subseteq \mathcal{N}_{\text{core}}(\tau)$. The validation of any extension requires verifying three properties before the updated core can be considered methodologically sound.

1. **Selection stability**: the non-variable set $\mathcal{N}_{k+1}(\tau)$ must be robust to dataset composition, i.e. not driven by any single cohort (Step A, Section B.3.1).

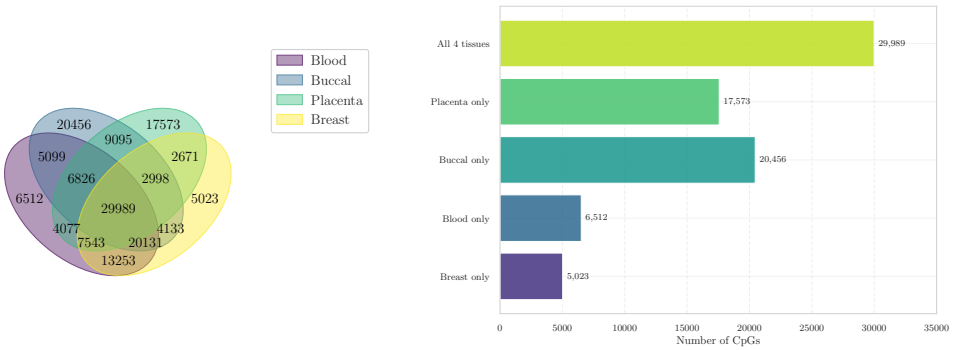
2. **Threshold optimality**: the chosen τ must lie at a favourable stability–dimensionality trade-off point in the new tissue context (Step B, Section B.3.2).

3. **Empirical coherence**: the selected CpGs must exhibit methylation and variability profiles consistent with the non-variability interpretation, ruling out selection artefacts (Steps C–D, Sections B.3.3–B.3.4).

This three-step validation protocol is general and applies identically to any future tissue extension.

B.3 Cross-Tissue Overlap Structure

Breast tissue was introduced alongside the three original reference tissues of [69]. Figure B.1 reports the overlap structure of tissue-specific non-variable CpG sets



(a) Pairwise and higher-order overlaps of tissue-specific non-variable CpG sets.

(b) Quantitative intersection sizes for key subsets.

Figure B.1: Global overlap structure of non-variable CpGs across blood, buccal epithelial cells, placenta, and breast tissue.

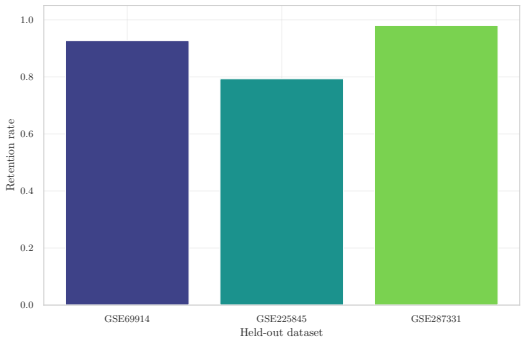
across all four tissues. The four-way invariant core $\mathcal{N}'_{\text{core}}(0.05)$ comprises 29,989 CpGs — a substantial shared subset whose persistence despite the inclusion of a cancer-relevant tissue suggests that at least a component of CpG non-variability reflects stable, context-independent properties of the methylation landscape rather than purely tissue-specific regulatory programmes. The breast-exclusive non-variable set (5,023 CpGs) is the smallest tissue-specific component, consistent with the expectation that breast tissue introduces additional heterogeneity relative to the original healthy tissue panel.

B.3.1 Step A: Stability of Non-Variable CpG Selection

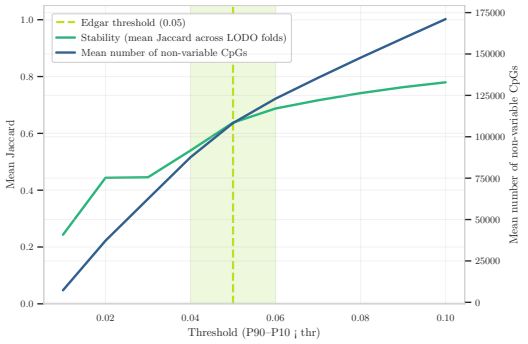
The first validation step assesses whether the breast-tissue non-variable set $\mathcal{N}_{\text{breast}}(\tau)$ is robust to dataset composition. A leave-one-dataset-out (LODO) strategy is adopted: for each held-out dataset $\mathcal{D}^{(-s)}$, the non-variable set is recomputed on the remaining datasets and compared to the full-reference set via the retention rate:

$$\rho^{(-s)} = \frac{\left|\mathcal{N}^{(-s)}_{\text{breast}}(\tau) \cap \mathcal{N}_{\text{breast}}(\tau)\right|}{|\mathcal{N}_{\text{breast}}(\tau)|}, \tag{B.4}$$

where $\mathcal{N}^{(-s)}_{\text{breast}}(\tau)$ denotes the non-variable set estimated without dataset s. The retention rates reported in Figure B.2a are consistently high across all three held-out datasets, demonstrating that $\mathcal{N}_{\text{breast}}(\tau)$ is not an artefact of any individual cohort but reflects a stable property of the underlying methylation variability structure in breast tissue. This result satisfies the first validation criterion stated in Section B.2 and is a necessary condition for interpreting non-variability as a genuine empirical property rather than a dataset-specific anomaly.



(a) Retention rates under the LODO protocol across held-out datasets (GSE69914, GSE225845, GSE287331).



(b) Mean Jaccard similarity $J(\tau)$ and mean non-variable CpG count as functions of the threshold.

Figure B.2: Assessment of robustness and trade-off between stability and dimensionality of non-variable CpG sets.

B.3.2 Step B: Stability–Dimensionality Trade-Off

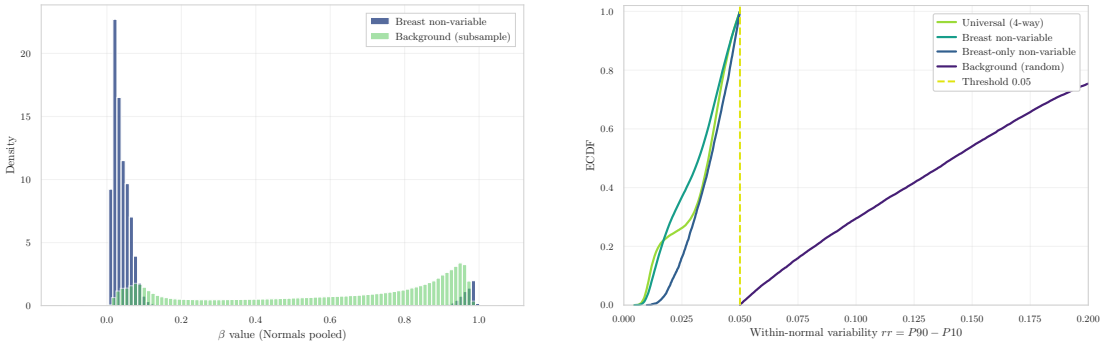
Having established stability at $\tau = 0.05$, Step B evaluates whether this threshold represents an optimal compromise between selection reproducibility and dimensionality reduction in the breast tissue setting. The mean Jaccard similarity across LODO folds and the mean number of non-variable CpGs are computed as functions of $\tau \in [0.01, 0.10]$:

$$J(\tau) = \frac{1}{S} \sum_{s=1}^{S} \frac{\left| \mathcal{N}_{\text{breast}}^{(-s)}(\tau) \cap \mathcal{N}_{\text{breast}}(\tau) \right|}{\left| \mathcal{N}_{\text{breast}}^{(-s)}(\tau) \cup \mathcal{N}_{\text{breast}}(\tau) \right|}. \tag{B.5}$$

As shown in Figure B.2b, $J(\tau)$ increases rapidly for small τ and saturates before the Edgar threshold, while the CpG count continues to grow approximately linearly. At $\tau = 0.05$, stability is already at its near-maximum value and the marginal gain from increasing the threshold further is negligible relative to the dimensionality cost incurred. This empirically validates the original threshold of [69] in the breast tissue context and provides a data-driven justification for its retention within the extended framework.

B.3.3 Step C: Methylation Distribution of Non-Variable CpGs

Step C characterises the methylation profiles of the selected non-variable CpGs by comparing their β-value distribution to a random background sample of array CpGs, computed on normal breast samples pooled across datasets. Non-variable



(a) Beta-value density for non-variable CpGs (breast tissue) versus genomic background.

(b) ECDF of within-normal variability $r_{\beta,j} = Q_{0.90} - Q_{0.10}$ across CpG subsets.

Figure B.3: Characterization of the selected non-variable CpG set in terms of methylation distribution and baseline risk behavior.

CpGs in breast tissue are strongly concentrated at extreme β values, consistent with constitutive hypo- or hypermethylated states. This pattern replicates the findings of [69] in healthy tissues and extends them to a cancer-relevant context, supporting the interpretation that non-variability reflects stable regulatory configurations rather than artefacts of tissue composition or disease-related heterogeneity. The enrichment of non-variable CpGs in CpG islands and constitutively silenced promoter regions, documented in the original study, is consistent with this bimodal profile [69, 70].

B.3.4 Step D: Baseline Variability Profile

Step D provides quantitative validation of the filtering criterion by examining the empirical cumulative distribution function (ECDF) of $r_{\beta,j}$ across four CpG subsets: the four-way shared non-variable core $\mathcal{N}'_{\text{core}}(0.05)$, the breast-specific non-variable set $\mathcal{N}_{\text{breast}}(0.05)$, the breast-exclusive component, and a random background set. The ECDFs in Figure B.3b show a clear separation: the non-variable subsets are sharply concentrated at low $r_{\beta,j}$ values, with the majority of CpGs well below $\tau = 0.05$, whereas the background distribution spans a substantially wider range. This confirms that the selected CpGs are empirically distinct in their variability profile and not merely defined as non-variable by construction. The consistency of this pattern across all non-variable subsets — including the breast-exclusive component — further supports the robustness of the extended framework.

B.4 Cohort-Resolved Characterisation of Breast-Tissue Non-Variability

The four-step validation protocol described in Section B.2 establishes the methodological soundness of the non-variability criterion at the level of pooled breast data. The present section extends this characterisation to the individual-cohort level, pursuing two complementary objectives: (i) to assess whether the filtering behaviour is consistent across independent datasets, each with distinct sample compositions and phenotypic configurations; and (ii) to characterise the biological identity of the invariant CpG set through genomic contextual enrichment analysis and cross-phenotype variability concordance. Together, these analyses situate the breast-tissue extension within a broader biological framework and provide the empirical grounding required for its use as a preprocessing step in Chapter 5.

B.4.1 Empirical Variability Spectrum Across Independent Breast Cohorts

To evaluate the consistency of the non-variability criterion across datasets, the inter-quantile range statistic $r_{\beta,j} = Q_{0.90}(\beta_j) - Q_{0.10}(\beta_j)$ was computed independently for each of the three breast cohorts (GSE69914, GSE225845, GSE287331). For each dataset, the analysis was carried out under two phenotypic configurations: (i) Normal + Adjacent (N+A), which includes only histologically normal and adjacent-normal samples, and (ii) Normal + Tumor (N+T), which additionally incorporates tumor samples. This design allows direct assessment of whether malignant tissue alters the global variability spectrum of CpG loci and, consequently, the proportion of sites classified as non-variable under the Edgar threshold $\tau = 0.05$.

Figure B.4 reports the empirical density of r_β under the N+T configuration for each dataset, with vertical markers indicating the three candidate thresholds $\tau \in \{0.01, 0.02, 0.05\}$. Several structural features are common to all three cohorts. The distribution is strongly right-skewed, with a pronounced mass concentrated near $r_\beta \approx 0$, indicating that a substantial fraction of CpG sites exhibit minimal within-group dispersion regardless of sample composition. The strict thresholds $\tau = 0.01$ and $\tau = 0.02$ remove only a small minority of loci, whereas the canonical $\tau = 0.05$ eliminates a sizeable but controlled proportion — 77,687 CpGs in GSE69914, 103,622 in GSE225845, and 168,663 in GSE287331 — consistent with the dimensionality reduction reported in the pooled analysis. A biologically informative pattern emerges from the comparison between the two phenotypic configurations. Under N+T, the Edgar filter removes fewer CpGs than under N+A: the greater methylation contrast between normal and tumor samples elevates r_β across a larger fraction of loci, reducing the proportion that falls below $\tau = 0.05$. By contrast, the N+A

configuration — whose analogous plots are reported in Chapter 5 — yields a higher count of non-variable CpGs, reflecting the subtler methylation differences between histologically normal tissue and adjacent-normal samples. This asymmetry is consistent with the concept of field cancerisation, whereby tissue adjacent to a tumour undergoes epigenetic alterations that, while not yet producing overt histological change, partially converge towards the tumour methylation profile. Under this interpretation, the reduced variability observed in N+A relative to N+T is not a statistical artefact but reflects a genuine biological phenomenon: adjacent tissue shares a greater proportion of its methylation landscape with normal tissue than tumor tissue does, resulting in a larger set of CpG sites that appear non-variable within the combined sample.

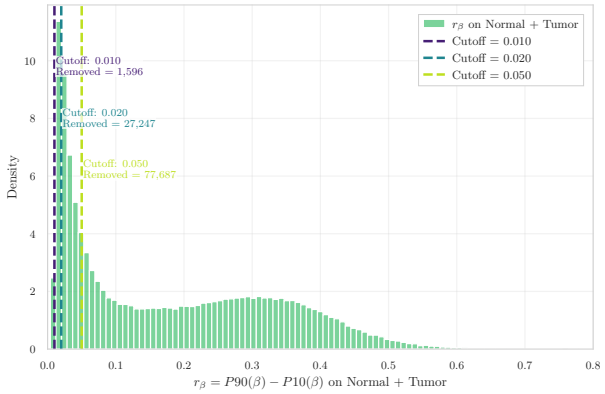
B.4.2 Contextual Enrichment Profile of Invariant CpGs

To characterise the genomic architecture of CpGs classified as non-variable under the Edgar threshold, a contextual enrichment analysis was performed using Illumina manifest annotations. Each probe was mapped to its corresponding gene-feature category (e.g., TSS200, TSS1500, 5'UTR, 1stExon, Body, 3'UTR, Intergenic), and enrichment was quantified as fold-change relative to the full set of interrogated CpGs within the dataset. Let $\mathcal{U}$ denote the universe of CpGs and $\mathcal{N}_{\text{breast}}(0.05)$ the invariant subset identified in the N+A configuration. For each annotation category g, enrichment was computed as:

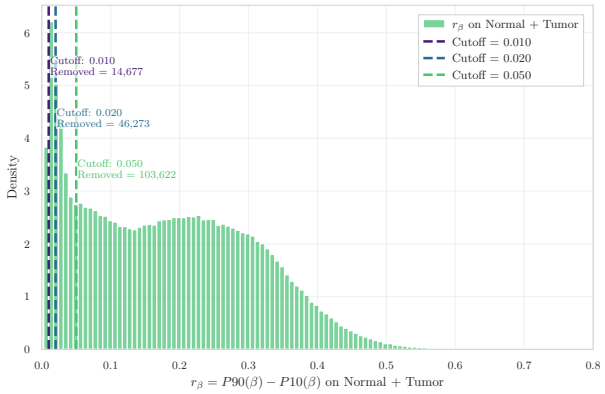
$$\text{FC}_g = \frac{\frac{|\mathcal{N}_{\text{breast}}(0.05) \cap g|}{|\mathcal{N}_{\text{breast}}(0.05)|}}{\frac{|\mathcal{U} \cap g|}{|\mathcal{U}|}}. \tag{B.6}$$

Figure B.5 reports the fold-change profile for the dominant gene-feature categories. Across datasets, invariant CpGs show consistent enrichment in proximal promoter regions (particularly TSS200 and 1stExon) and relative depletion in gene bodies and distal regions. The baseline reference (fold-change = 1) indicates the expected frequency under random sampling from the array; deviations above this threshold reflect preferential localisation of low-variability CpGs within regulatory domains. This pattern is coherent with the bimodal methylation behaviour described in Section B.3.3, where invariant CpGs were shown to concentrate at constitutive hypo- or hypermethylated states. Promoter-associated CpG islands are known to exhibit tight regulatory control and reduced dispersion in normal tissue, providing a biologically plausible explanation for the observed enrichment.

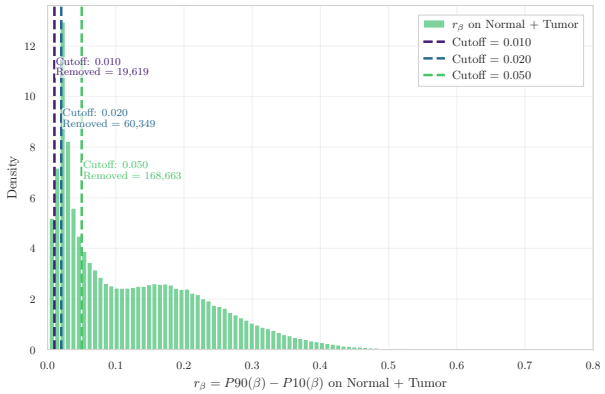
Importantly, the enrichment structure is stable across cohorts, indicating that genomic localisation of invariant CpGs reflects intrinsic regulatory constraints rather than dataset-specific artefacts.



(a) GSE69914

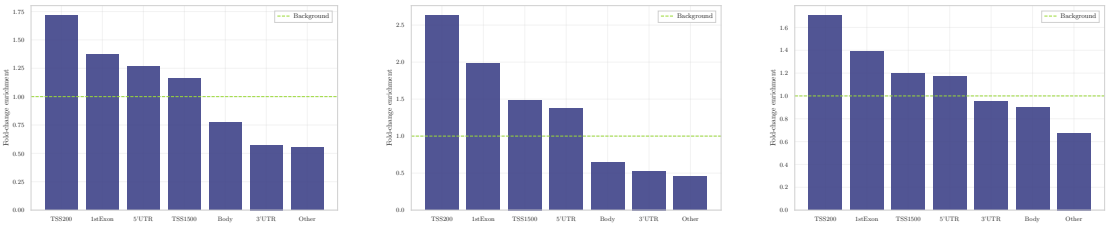


(b) GSE225845



(c) GSE287331

Figure B.4: Distribution of r_β computed on the Normal+Tumor comparison for each dataset. Vertical dashed lines indicate the selected threshold.

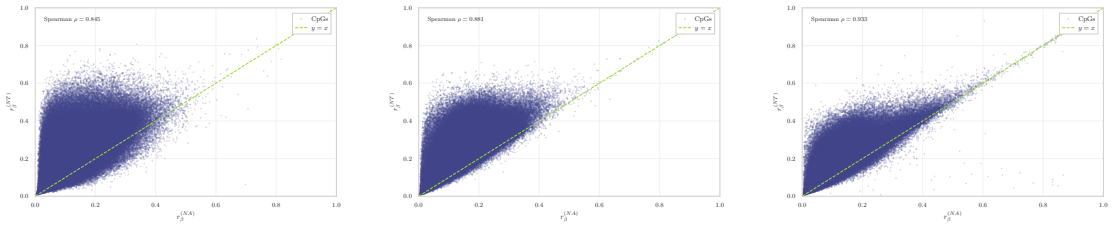


(a) GSE69914 **(b)** GSE225845 **(c)** GSE287331

Figure B.5: Fold-enrichment of genomic context categories among invariant CpGs relative to the background distribution in each dataset. The dashed line indicates enrichment equal to background (fold = 1).

B.4.3 Phenotype-Resolved Variability Concordance Analysis

Section B.4.1 established that the aggregate variability spectrum is qualitatively preserved when tumor samples are incorporated. The present section quantifies this concordance at the level of individual CpG sites by directly comparing r_β values computed under the N+A and N+T configurations. For each CpG j, the pair $\left(r_{\beta,j}^{(NA)}, r_{\beta,j}^{(NT)}\right)$ was examined and rank concordance was quantified via the Spearman correlation coefficient. Figure B.6 displays the resulting scatter plots together with the identity line $y = x$ for each dataset. The Spearman coefficients are $\rho \approx 0.845$ (GSE69914), $\rho \approx 0.881$ (GSE225845), and $\rho \approx 0.933$ (GSE287331), indicating strong to very strong rank preservation of CpG-level variability across phenotypic configurations. Several structural features emerge from the scatter plots. The bulk of CpG loci cluster tightly around the identity line, confirming that relative variability ordering is largely maintained when tumor samples are included. A minority of loci deviate upward from the diagonal, corresponding to CpGs whose dispersion increases in the presence of tumor tissue; by contrast, very few loci exhibit a systematic decrease in variability under N+T. This asymmetry is consistent with the expectation that malignant transformation introduces epigenetic heterogeneity rather than convergence, and corroborates the finding in Section B.4.1 that tumor inclusion primarily affects loci in the intermediate variability range. The high rank concordance implies that the non-variability criterion is largely phenotype-robust: CpGs classified as low-variability under the N+A configuration remain predominantly low-variability when tumor samples are incorporated. This result supports the interpretation that breast-tissue non-variability reflects stable structural properties of methylation regulation, with tumor-associated heterogeneity manifesting as localised deviations superimposed on an otherwise preserved variability backbone.



(a) GSE69914
Spearman $\rho \approx 0.845$

(b) GSE225845
Spearman $\rho \approx 0.881$

(c) GSE287331
Spearman $\rho \approx 0.933$

Figure B.6: Spearman correlation between CpG-level $\Delta\beta$ values computed in the Normal+Adjacent and Normal+Tumor contrasts for each dataset. The diagonal line represents identity.

B.4.4 Invariant-Set Partitioning Across Phenotypic Combinations

The analyses in Sections B.4.1– B.4.3 establish that non-variability in breast tissue is broadly stable across both cohorts and phenotypic configurations. The present section integrates the breast-tissue invariance structure with the reference framework of [69] by performing a set-theoretic decomposition that explicitly accounts for the effect of tumor inclusion. Let $\mathcal{B} = \mathcal{N}_{\text{Blood}} \cap \mathcal{N}_{\text{Buccal}} \cap \mathcal{N}_{\text{Placenta}}$ denote the universal invariant set across the three reference tissues of [69]. The breast-tissue non-variable sets $\mathcal{N}_{\text{breast}}(0.05)$ were estimated separately under the N+A and N+T configurations for each cohort. The intersection of $\mathcal{B}$ with these two sets defines three biologically interpretable invariant classes.

Core invariants $\mathcal{B} \cap \mathcal{N}^{(NA)}_{\text{breast}}(0.05) \cap \mathcal{N}^{(NT)}_{\text{breast}}(0.05)$ CpGs that are non-variable across all reference tissues and remain invariant in breast tissue irrespective of tumor inclusion. These loci constitute the most conserved component of the cross-tissue epigenetic backbone.

Conditionally variable universal CpGs $\mathcal{B} \cap \mathcal{N}^{(NA)}_{\text{breast}}(0.05) \setminus \mathcal{N}^{(NT)}_{\text{breast}}(0.05)$ Loci that are invariant across all reference tissues and in normal breast tissue but become variable when tumor samples are incorporated. This subset identifies CpGs whose stability is sensitive to malignant transformation and that may be of interest for cancer-specific methylation studies.

Breast-specific invariants $\left(\mathcal{N}^{(NA)}_{\text{breast}}(0.05) \cap \mathcal{N}^{(NT)}_{\text{breast}}(0.05)\right) \setminus \mathcal{B}$ CpGs that are consistently non-variable within breast tissue under both phenotypic configurations

but are not part of the universal cross-tissue invariant core. These loci reflect tissue-contextual stability that is specific to breast epithelium.

Figure B.7 visualises this set-theoretic decomposition for each dataset as a three-way Venn diagram. The quantitative breakdown is consistent across cohorts. The core invariant class comprises approximately 27,000–28,000 CpGs per dataset (GSE69914: 28,241; GSE225845: 26,542; GSE287331: 27,577), representing a highly conserved epigenetic backbone that is robust to both tissue context and phenotypic variation. The conditionally variable universal class is markedly smaller (GSE69914: 2,010; GSE225845: 1,353; GSE287331: 496), confirming that only a minor fraction of cross-tissue invariants is susceptible to tumor-induced destabilisation. The breast-specific invariant class is the largest component in datasets with denser sampling (GSE69914: 138,256; GSE225845: 76,223; GSE287331: 46,788), reflecting the greater CpG coverage of the corresponding platforms and the broader set of tissue-contextual constraints captured at higher array density. Several biological observations emerge from this decomposition. First, the invariant core is not globally eroded by cancer: tumor-associated variability acts locally on a bounded subset of CpGs, while the structural invariance backbone is preserved. Second, the breast-specific invariant class is predominantly shared between the N+A and N+T configurations, confirming that tissue-contextual stability is itself phenotype-robust. Third, the hierarchical organisation of the invariant partition — a context-independent core layer, a tissue-specific invariant layer, and a phenotype-sensitive component — mirrors the layered regulatory architecture of the methylome, in which CpG islands at constitutively regulated promoters are progressively supplemented by tissue-specific and stimulus-responsive methylation programmes [69, 70].

Taken together, the analyses in Section B.4 provide a cohort-resolved and phenotype-stratified characterisation of breast-tissue non-variability that complements the pooled validation in Section B.2. The convergence of results across independent datasets, genomic annotation categories, and phenotypic configurations establishes that the Edgar-based filtering strategy is a biologically grounded dimensionality reduction tool whose behaviour in breast tissue is consistent, interpretable, and robust to the heterogeneity introduced by malignant transformation.

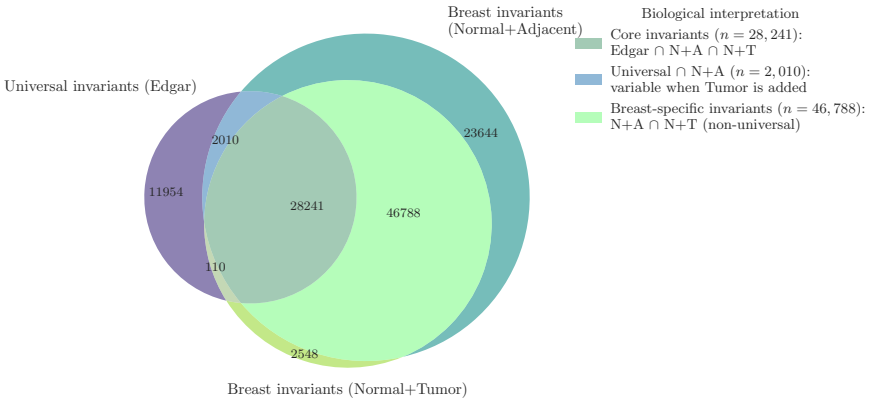
B.5 Positioning Within the Thesis and Generalisability

The extended framework described in this appendix provides the empirical grounding for the variability filter applied in Chapter 5. The four validation steps collectively establish that the non-variability criterion of [69] is stable, threshold-optimal, and biologically coherent when applied to breast tissue — necessary conditions for

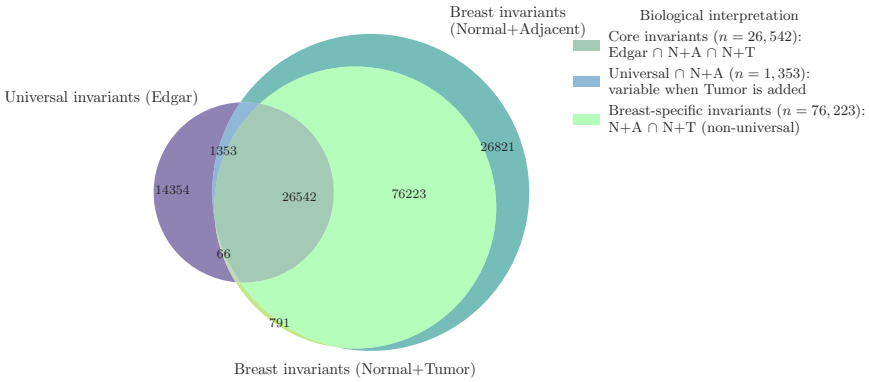
its use as a preprocessing step in the present study. The cohort-resolved characterisation presented in Section 5.3 further demonstrates that these properties hold independently within each dataset and are robust to the phenotypic heterogeneity introduced by tumor inclusion, consolidating the methodological foundations of the filter across the full range of analytical configurations considered in Chapter 5.

In the current analysis, datasets are processed independently and the non-variability threshold is estimated within each cohort separately, rather than enforcing a single shared CpG list across datasets. This design is motivated by two complementary considerations. First, it preserves dataset-specific methylation structure and avoids premature cross-cohort harmonisation that could mask genuine biological heterogeneity between sample populations. Second, the cohort-level validation reported in Section B.4 — in particular the consistency of the variability spectrum, the enrichment profile, and the phenotype-resolved concordance across GSE69914, GSE225845, and GSE287331 — ensures that applying the threshold independently within each cohort does not introduce arbitrary inter-dataset discordance, but rather instantiates a shared population-level filtering principle in each data context separately.

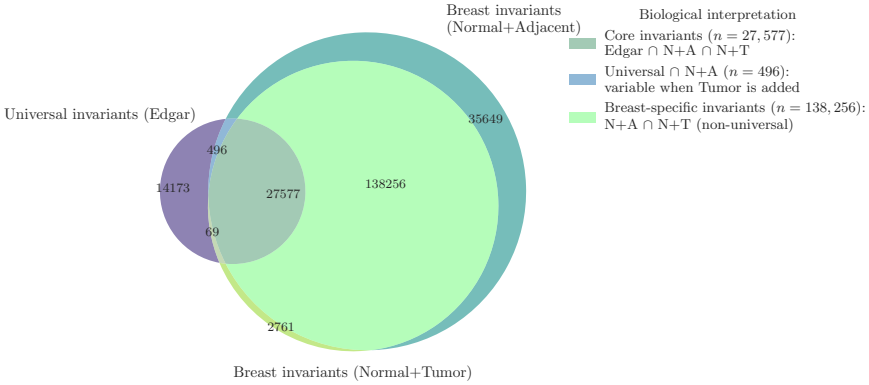
The general framework formalised in Section B.2 — update rule (Equation B.3), LODO stability assessment (Equation B.4), and Jaccard-based threshold analysis (Equation B.5) — is structurally independent of tissue type and platform density. It therefore remains applicable to future tissue extensions and to higher-density arrays such as the Illumina EPIC platform, whose expanded CpG coverage approximately doubles that of the 450K array [29]. In this respect, the breast-tissue extension documented here serves a dual purpose: it validates the criterion in a cancer-relevant context, and it demonstrates the operability of the general protocol for incorporating new tissues into the reference framework — a property of direct relevance for future studies seeking to apply variability-based filtering beyond the tissue panel considered here.



(a) GSE69914



(b) GSE225845



(c) GSE287331

Figure B.7: Overlap between universal invariants and breast-specific invariant sets derived from the Normal+Adjacent and Normal+Tumor contrasts in each dataset.

Appendix C

Epigenetic Age as a Covariate and Stratification Framework

This appendix describes the epigenetic age modelling step whose output is used in Chapter 7, where Horvath DNAmAge is employed as a covariate in ComBat batch correction to control for age-driven confounding across cohorts, and whose age-bin assignments provide a unified stratification framework across all three datasets. The clock was applied to GSE69914, GSE225845, and GSE287331; however, only GSE225845 provides chronological age at surgery, enabling a quantitative evaluation of prediction accuracy, whereas for the remaining two cohorts DNAmAge estimates were used solely for age-bin assignment without performance assessment.

C.1 The Horvath Multi-Tissue Clock

Chronological age provides only a coarse approximation of biological ageing: individuals of the same age may exhibit markedly different physiological states and disease susceptibility. DNA methylation has emerged as one of the most robust molecular correlates of ageing, with age-associated changes accumulating reproducibly at specific CpG sites in a largely monotonic manner. Epigenetic clocks exploit this property by combining methylation levels at selected loci into a predictive model of *biological age* (DNAmAge), correlating strongly with chronological age while capturing deviations associated with environmental exposures, disease states, and tissue-specific stressors [15, 106]. The difference between DNAmAge and chronological age — *age acceleration* — indicates altered biological ageing and has been associated with cancer risk and mortality [15, 106]. Among epigenetic

Table C.1: Global performance of the Horvath epigenetic clock on GSE225845.

Samples	MAE (years)	RMSE (years)	Pearson R	R^2	MAPE (%)
253	9.989	11.691	0.770	0.593	22.735

Table C.2: Performance of the Horvath clock stratified by tissue group.

Tissue	MAE (yr)	RMSE (yr)	Pearson R	R^2	MAPE (%)
Normal	11.6	12.6	0.703	0.494	29.1
Adjacent	8.71	10.9	0.715	0.511	17.6

clocks, multi-tissue models capture ageing-related patterns conserved across diverse tissues at the cost of some tissue-specific accuracy [15, 106], a critical property when working with solid and pathological samples.

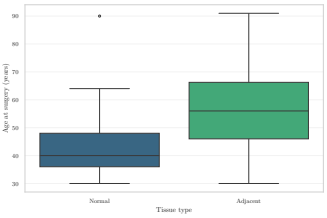
The Horvath clock [15] is the first large-scale multi-tissue estimator, trained on approximately 8,000 samples from 82 datasets spanning 51 tissue types. It applies elastic net regression to select 353 CpG sites from 21,369 loci shared across Illumina 27k and 450k platforms:

$$\text{DNAmAge} = \sum_{i=1}^{353} w_i \, \beta_i + b, \tag{C.1}$$

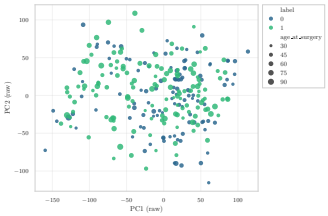
where β_i is the methylation level of the i-th CpG and w_i its regression coefficient. Although designed as a pan-tissue predictor, the clock reports reduced accuracy for breast tissue due to cellular heterogeneity, hormonal regulation, and epigenetic remodelling associated with mammary differentiation [15]. Malignant transformation further decouples methylation from normal ageing trajectories through global hypomethylation and focal hypermethylation, rendering DNAmAge unreliable in tumour tissue [4, 106]. DNAmAge estimation was therefore restricted to Normal and Adjacent samples, excluding tumours for both biological and methodological coherence.

C.2 Empirical Evaluation on GSE225845

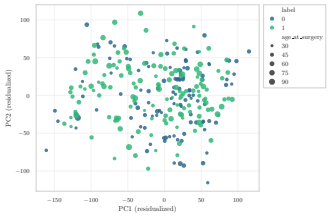
Epigenetic age estimation was performed on the Normal and Adjacent samples of GSE225845 using beta values from the preprocessing pipeline of Chapter 7, comprising probe-level technical filtering and BMIQ normalisation as recommended in the original Horvath implementation; tumour samples were excluded as detailed in Section C.1. Of the 353 required CpGs, 316 were directly available on the EPIC array; the remaining 37 were imputed using platform-specific reference values (`goldstandard/goldstandard2`) or $\beta = 0.5$ otherwise [15]. Performance



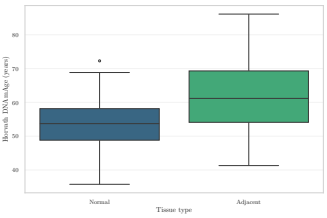
(a) Chronological age at surgery by tissue group.



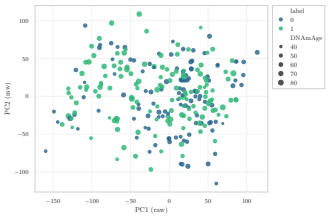
(b) Raw methylation PCA coloured by chronological age.



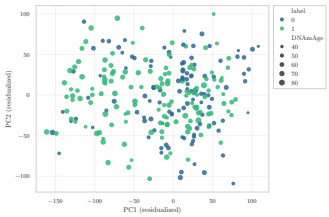
(c) Residualised PCA after removal of age effects.



(d) Horvath DNAmAge by tissue group.



(e) Raw methylation PCA coloured by DNAmAge.

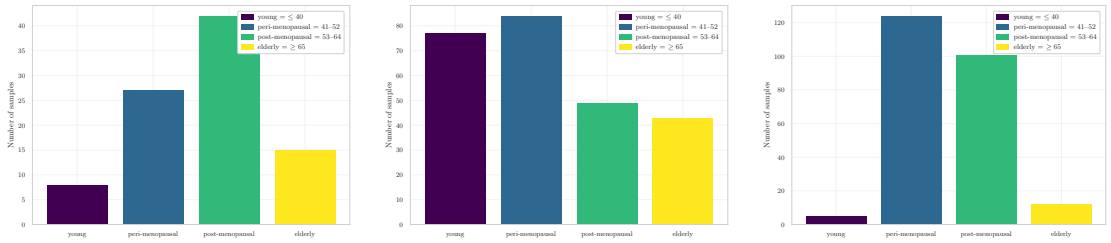


(f) Residualised PCA after removal of age-related structure.

Figure C.1: Age-related methylation structure (GSE225845). Age distributions and PCA before and after age residualisation confirm that DNAmAge captures structured methylation variance, motivating its use as a covariate in the downstream analyses of Chapter 7.

was evaluated via MAE, RMSE, Pearson R, R^2, and MAPE; age acceleration was defined as DNAmAge − chronological age.

DNAmAge exhibited a moderate-to-strong linear association with chronological age (Table C.1), consistent with prior applications of the Horvath clock to tissues outside its training optimum [15]. Stratification by tissue type (Table C.2) revealed that Normal samples exhibit higher estimation error than Adjacent samples while maintaining comparable correlation with chronological age. The lower error in Adjacent tissue may reflect more homogeneous epigenetic alterations associated with early field effects. Globally, samples displayed positive age acceleration (mean = +7.57 years, median = +9.10 years). Normal samples showed stronger acceleration (mean ≈ +10.9 years) than Adjacent samples (mean ≈ +4.9 years, more heterogeneous), suggesting that a subset of clinically normal tissues already harbours substantial epigenetic ageing. PCA on raw methylation data (Figure C.1) reveals that DNAmAge contributes more structured variance to the methylome than chronological age ($\text{corr}(\text{PC1}, \text{DNAmAge}) = 0.17$ vs. $\text{corr}(\text{PC1}, \text{age}) = -0.03$). After residualisation, both correlations drop to zero, confirming that age acts as a



(a) GSE69914 **(b)** GSE225845 **(c)** GSE287331

Figure C.2: Epigenetic age-bin distribution. Sample counts by age group (young, peri-, post-menopausal, elderly).

confounding factor and that its removal is effective. These results are consistent with the framework of Teschendorff et al. [4], whereby age-dependent methylation changes at Polycomb-regulated loci constitute a hallmark of cancer susceptibility.

C.3 Clinically Motivated Age Stratification

Beyond its continuous effect on DNA methylation, age marks distinct biological phases in breast tissue associated with major hormonal transitions. The menopausal transition represents a well-defined breakpoint characterised by profound changes in oestrogen exposure, tissue composition, and cancer risk: large-scale epidemiological studies place the median age at natural menopause in Western populations between 50 and 52 years [107], and clinical evidence confirms that breast cancers arising before and after menopause differ in incidence, molecular characteristics, and prognosis. Four bins were therefore defined to capture distinct hormonal phases: young (≤ 40 years), peri-menopausal (41–52), post-menopausal (53–64), and elderly (≥ 65). These intervals align with established clinical definitions of reproductive ageing and mitigate estimation error inherent in continuous epigenetic age prediction: assigning samples to clinically motivated bins reduces sensitivity to small prediction errors near decision boundaries, shifting the focus from exact age reconstruction to correct assignment to ranges associated with distinct breast cancer biology. The clock was applied to all three cohorts to derive DNAmAge estimates and bin assignments, establishing a unified stratification framework. The distribution across bins (Figure C.2) shows a broadly consistent pattern: peri- and post-menopausal samples predominate in GSE69914 and GSE287331, while GSE225845 exhibits a more balanced distribution with a substantial proportion of younger and elderly samples. Elderly samples are comparatively underrepresented in GSE287331, while very few young samples appear in that cohort.

Bibliography

[1] M. Ehrlich. «DNA methylation in cancer: Too much, but also too little». In: *Oncogene* 21 (2002), pp. 5400–5413. DOI: 10.1038/sj.onc.1205651 (cit. on pp. 1, 7, 15, 16, 26).

[2] A. Bird. «DNA methylation patterns and epigenetic memory». In: *Genes & Development* 16.1 (2002), pp. 6–21. DOI: 10.1101/gad.947102 (cit. on pp. 1, 5, 15, 16).

[3] A. E. Teschendorff, Y. Gao, A. Jones, M. Ruebner, M. W. Beckmann, D. L. Wachter, P. A. Fasching, and M. Widschwendter. «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. 1, 2, 8–10, 15, 16, 30, 32, 54).

[4] A. E. Teschendorff, U. Menon, A. Gentry-Maharaj, S. J. Ramus, D. J. Weisenberger, H. Shen, M. Campan, H. Noushmehr, C. G. Bell, A. P. Maxwell, D. A. Savage, E. Mueller-Holzner, C. Marth, G. Kocjan, S. A. Gayther, A. Jones, S. Beck, W. Wagner, P. W. Laird, I. J. Jacobs, and M. Widschwendter. «Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer». In: *Genome Research* 20.4 (2010), pp. 440–446. DOI: 10.1101/gr.103606.109 (cit. on pp. 1, 7, 8, 133, 135).

[5] A. E. Teschendorff, J. Zhuang, and M. Widschwendter. «Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies». In: *Bioinformatics* 27.11 (2011), pp. 1496–1505. DOI: 10.1093/bioinformatics/btr171 (cit. on pp. 2, 9).

[6] N. Meinshausen and P. Bühlmann. «Stability selection». In: *Journal of the Royal Statistical Society: Series B* 72.4 (2010), pp. 417–473. DOI: 10.1111/j.1467-9868.2010.00740.x (cit. on pp. 2, 55, 60, 68, 73).

[7] C. Ambroise and G. J. McLachlan. «Selection bias in gene extraction on the basis of microarray gene-expression data». In: *Proceedings of the National Academy of Sciences* 99.10 (2002), pp. 6562–6566. DOI: 10.1073/pnas.102102699 (cit. on pp. 2, 53–55, 57, 59).

[8] D. Hanahan and R. A. Weinberg. «Hallmarks of Cancer: The Next Generation». In: *Cell* 144.5 (2011), pp. 646–674. DOI: 10.1016/j.cell.2011.02.013 (cit. on pp. 4, 5).

[9] D. Hanahan and R. A. Weinberg. «The Hallmarks of Cancer». In: *Cell* 100.1 (2000), pp. 57–70. DOI: 10.1016/S0092-8674(00)81683-9 (cit. on pp. 4, 5).

[10] National Cancer Institute. *BRCA Gene Changes: Cancer Risk and Genetic Testing*. https://www.cancer.gov/about-cancer/causes-prevention/genetics/brca-fact-sheet. Accessed October 2025. 2024 (cit. on pp. 4, 15).

[11] K. P. M. Suijkerbuijk, M. J. Fackler, S. Sukumar, C. H. van Gils, T. van Laar, E. van der Wall, M. Vooijs, and P. J. van Diest. «Methylation is less abundant in BRCA1-associated compared with sporadic breast cancer». In: *Annals of Oncology* 19.11 (2008), pp. 1870–1874. DOI: 10.1093/annonc/mdn409 (cit. on pp. 5, 15).

[12] T. E. Bartlett, I. Evans, A. Jones, J. E. Barrett, S. Haran, D. Reisel, K. Papaikonomou, L. Jones, C. Herzog, N. Pashayan, B. M. Simões, R. B. Clarke, D. G. Evans, T. S. Ghezelayagh, S. Ponandai-Srinivasan, N. R. Boggavarapu, P. G. Lalitkumar, S. J. Howell, R. A. Risques, A. Flöter Rådestad, L. Dubeau, K. Gemzell-Danielsson, and M. Widschwendter. «Antiprogestins reduce epigenetic field cancerization in breast tissue of young healthy women». In: *Genome Medicine* 14.1 (2022), p. 64. DOI: 10.1186/s13073-022-01063-5 (cit. on pp. 5, 15).

[13] R. A. Irizarry, C. Ladd-Acosta, B. Wen, Z. Wu, C. Montano, P. Onyango, H. Cui, K. Gabo, M. Rongione, M. Webster, H. Ji, J. B. Potash, S. Sabunciyan, and A. P. Feinberg. «The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores». In: *Nature Genetics* 41.2 (2009), pp. 178–186. DOI: 10.1038/ng.298 (cit. on p. 6).

[14] P. A. Jones. «Functions of DNA methylation: islands, start sites, gene bodies and beyond». In: *Nature Reviews Genetics* 13.7 (2012), pp. 484–492. DOI: 10.1038/nrg3230 (cit. on pp. 7, 61).

[15] S. Horvath. «DNA methylation age of human tissues and cell types». In: *Genome Biology* 14.10 (2013), R115. DOI: 10.1186/gb-2013-14-10-r115 (cit. on pp. 8, 107, 132–134).

[16] A. E. Teschendorff, A. Jones, H. Fiegl, A. Sargent, J. J. Zhuang, H. C. Kitchener, and M. Widschwendter. «Epigenetic variability in cells of normal cytology is associated with the risk of future morphological transformation». In: *Genome Medicine* 4.3 (2012), p. 24. DOI: 10.1186/gm323 (cit. on p. 8).

[17] P. Yousefi, K. Huen, R. Aguilar Schall, A. Decker, E. Elboudwarej, H. Quach, L. Barcellos, and N. Holland. «Considerations for Normalization of DNA Methylation Data by Illumina 450K BeadChip Assay in Population Studies». In: *Epigenetics* 8.11 (Aug. 2013), pp. 1141–1152. DOI: 10.4161/epi.26037 (cit. on pp. 11, 38, 47).

[18] National Center for Biotechnology Information (NCBI). *Gene Expression Omnibus (GEO)*. https://www.ncbi.nlm.nih.gov/geo/. Public functional genomics data repository supporting MIAME-compliant submissions. Accessed 2025 (cit. on p. 12).

[19] National Center for Biotechnology Information (NCBI). *GSE69914: Genome wide DNA methylation profiling of normal breast, normal adjacent and breast cancer tissue*. Gene Expression Omnibus (GEO). 2015. URL: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE69914 (cit. on pp. 12, 41).

[20] National Center for Biotechnology Information (NCBI). *GSE225845: High neighborhood deprivation impacts DNA methylation and gene expression in cancer-related genes*. Gene Expression Omnibus (GEO). 2023. URL: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE225845 (cit. on pp. 12, 44).

[21] National Center for Biotechnology Information (NCBI). *GSE287331: DNA methylation patterns in breast cancer, paired benign tissue from ipsilateral and contralateral breast, and healthy controls*. Gene Expression Omnibus (GEO). 2025. URL: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE287331 (cit. on p. 12).

[22] P. Du, X. Zhang, C.-C. Huang, N. Jafari, W. A. Kibbe, L. Hou, and S. M. Lin. «Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis». In: *BMC Bioinformatics* 11.1 (2010), p. 587. DOI: 10.1186/1471-2105-11-587 (cit. on pp. 13, 40, 43, 49, 58).

[23] Andrew E. Teschendorff, Francesco Marabita, Matthias Lechner, Thomas Bartlett, David Tegner, Jonathan 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 (cit. on pp. 13, 17, 27, 38, 41, 44, 47).

[24] B. P. de Almeida, G. Richard, H. Dalla-Torre, C. Blum, L. Hexemer, P. Pandey, S. Laurent, C. Rajesh, M. Lopez, A. Laterre, M. Lang, U. Sahin, K. Beguir, and T. Pierrot. «A Multimodal Conversational Agent for DNA, RNA and Protein Tasks». In: *Nature Machine Intelligence* (2025). DOI: 10.1038/s42256-025-01047-1 (cit. on p. 13).

[25] A. V. Sokolov and H. B. Schiöth. «Decoding Depression: A Comprehensive Multi-cohort Exploration of Blood DNA Methylation Using Machine Learning and Deep Learning Approaches». In: *Translational Psychiatry* 14.1 (2024), p. 326. DOI: 10.1038/s41398-024-02992-y (cit. on pp. 14, 104).

[26] J. Maksimovic, L. Gordon, and A. Oshlack. «SWAN: Subset-Quantile Within Array Normalization for Illumina Infinium HumanMethylation450 BeadChips». In: *Genome Biology* 13.6 (2012), R44. DOI: 10.1186/gb-2012-13-6-r44 (cit. on pp. 16, 33, 38, 39, 41, 42, 44).

[27] Illumina Inc. *Infinium MethylationEPIC v1.0 B5 Manifest File*. https://support.illumina.com/array/array_kits/infinium-methylationepic-beadchip-kit/downloads.html. Accessed 2025 (cit. on p. 18).

[28] W. Zhou, P. W. Laird, and H. 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 pp. 19, 28, 32, 33, 39, 118).

[29] R. Pidsley, E. Zotenko, T. J. Peters, M. G. Lawrence, G. P. Risbridger, P. Molloy, S. Van Dijk, B. Muhlhausler, C. Stirzaker, and S. J. Clark. «Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling». In: *Genome Biology* 17 (2016), p. 208. DOI: 10.1186/s13059-016-1066-1 (cit. on pp. 19, 28, 32, 33, 39, 53, 118, 130).

[30] L. Chen, R. Liu, Z.-P. Liu, M. Li, and K. Aihara. «Detecting early-warning signals for sudden deterioration of complex diseases by dynamical network biomarkers». In: *Scientific Reports* 2 (2012), p. 342. DOI: 10.1038/srep00342 (cit. on p. 19).

[31] J. West, G. Bianconi, S. Severini, and A. E. Teschendorff. «Differential network entropy reveals cancer system hallmarks». In: *Scientific Reports* 2 (2012), p. 802. DOI: 10.1038/srep00802 (cit. on p. 19).

[32] J. Liu, D. Ding, J. Zhong, and R. Liu. «Identifying the critical states and dynamic network biomarkers of cancers based on network entropy». In: *Journal of Translational Medicine* 20.254 (2022). DOI: 10.1186/s12967-022-03445-0 (cit. on p. 19).

[33] P. Wang. «Network biology: Recent advances and challenges». In: *Gene & Protein in Disease* 1.2 (2022), p. 101. DOI: 10.36922/gpd.v1i2.101 (cit. on p. 23).

[34] S. R. Dennis, T. Tsukioki, G. Cottone, W. Zhou, P. A. Ganz, M. E. Sehl, Y. Luo, S. A. Khan, and S. E. Clare. «DNA methylation patterns in breast cancer, paired benign tissue from ipsilateral and contralateral breast, and healthy controls». In: *Breast Cancer Research* 27.1 (2025), p. 103. DOI: 10.1186/s13058-025-02057-y (cit. on pp. 24, 47).

[35] J. Maksimovic, B. Phipson, and A. Oshlack. «A cross-package Bioconductor workflow for analysing methylation array data». In: *F1000Research* 5 (2017), p. 1281. DOI: 10.12688/f1000research.8839.3 (cit. on p. 25).

[36] J. Yates, H. Schaufelberger, R. Steinacher, P. Schär, K. Truninger, and V. Boeva. «DNA-methylation variability in normal mucosa: A field cancerization marker in patients with adenomatous polyps». In: *Journal of the National Cancer Institute* 116.6 (2024), pp. 974–982. DOI: 10.1093/jnci/djae016 (cit. on pp. 31, 32).

[37] S. Zhang, X. Xiao, Y. Yi, X. Wang, L. Zhu, Y. Shen, D. Lin, and C. Wu. «Tumor initiation and early tumorigenesis: molecular mechanisms and interventional targets». In: *Signal Transduction and Targeted Therapy* 9 (2024), p. 149. DOI: 10.1038/s41392-024-01848-7 (cit. on pp. 31, 32).

[38] J.-P. Fortin, A. Labbé, M. Lemire, B. W. Zanke, T. J. Hudson, E. J. Fertig, C. M. T. Greenwood, and K. 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 pp. 32, 33).

[39] W. Timp and A. P. Feinberg. «Cancer as a dysregulated epigenome allowing cellular growth advantage at the expense of the host». In: *Nature Reviews Cancer* 13.7 (2013), pp. 497–510. DOI: 10.1038/nrc3486 (cit. on p. 32).

[40] J. T. Leek, R. B. Scharpf, H. Corrada Bravo, D. Simcha, B. Langmead, W. E. Johnson, D. Geman, K. Baggerly, and R. A. Irizarry. «Tackling the widespread and critical impact of batch effects in high-throughput data». In: *Nature Reviews Genetics* 11.10 (2010), pp. 733–739. DOI: 10.1038/nrg2825 (cit. on p. 33).

[41] T. J. Triche Jr, D. J. Weisenberger, D. Van Den Berg, P. W. Laird, and K. D. Siegmund. «Low-level processing of Illumina Infinium DNA Methylation BeadArrays». In: *Nucleic Acids Research* 41.7 (2013), e90. DOI: 10.1093/nar/gkt090 (cit. on pp. 36, 37, 40, 49).

[42] J.-P. Fortin, T. J. Triche Jr, and K. D. Hansen. «Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi». In: *Bioinformatics* 33.4 (2017), pp. 558–560. DOI: 10.1093/bioinformatics/btw691 (cit. on pp. 36, 40, 49).

[43] C. S. Wilhelm-Benartzi, D. C. Koestler, M. R. Karagas, J. M. Flanagan, B. C. Christensen, K. T. Kelsey, C. J. Marsit, E. A. Houseman, and R. Brown. «Review of processing and analysis methods for DNA methylation array data». In: *British Journal of Cancer* 109.6 (2013), pp. 1394–1402. DOI: 10.1038/bjc.2013.496 (cit. on p. 36).

[44] Z. Xu, L. Niu, and J. A. Taylor. «The ENmix DNA methylation analysis pipeline for Illumina BeadChip and comparisons with seven other preprocessing pipelines». In: *Clinical Epigenetics* 13 (2021), p. 216. DOI: 10.1186/s13148-021-01207-1 (cit. on p. 36).

[45] S. A. Islam, S. J. Goodman, J. L. MacIsaac, J. Obradović, R. G. Barr, W. T. Boyce, and M. S. Kobor. «Integration of DNA methylation patterns and genetic variation in human pediatric tissues helps inform EWAS design and interpretation». In: *Epigenetics & Chromatin* 12.1 (2019), p. 1. DOI: 10.1186/s13072-018-0245-6 (cit. on pp. 37, 46).

[46] G. Mansell, T. J. Gorrie-Stone, Y. Bao, M. Kumari, L. S. Schalkwyk, J. Mill, and E. Hannon. «Guidance for DNA methylation studies: Statistical insights from the Illumina EPIC array». In: *BMC Genomics* 20.1 (2019), p. 366. DOI: 10.1186/s12864-019-5761-7 (cit. on pp. 37, 46).

[47] T. Wang, W. Guan, J. Lin, N. Boutaoui, G. Canino, J. Luo, J. C. Celedón, and W. Chen. «A systematic study of normalization methods for Infinium 450K methylation data using whole-genome bisulfite sequencing data». In: *Epigenetics* 10.6 (2015), pp. 536–545. DOI: 10.1080/15592294.2015.1057384 (cit. on pp. 38, 41, 44).

[48] 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. 59. DOI: 10.1186/s13148-023-01459-z (cit. on pp. 38, 47).

[49] J. Liu and K. D. Siegmund. «An evaluation of processing methods for HumanMethylation450 BeadChip data». In: *BMC Genomics* 17.1 (2016), p. 469. DOI: 10.1186/s12864-016-2819-7 (cit. on pp. 38, 47).

[50] H. Naeem, N. C. Wong, Z. Chatterton, M. K. H. Hong, J. S. Pedersen, N. M. Corcoran, C. M. Hovens, and G. 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 (cit. on pp. 39, 117).

[51] Y. A. Chen, M. Lemire, S. Choufani, D. T. Butcher, D. Grafodatskaya, B. W. Zanke, S. Gallinger, T. J. Hudson, and R. 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 (cit. on pp. 39, 117).

[52] D. L. McCartney, R. M. Walker, S. W. Morris, A. M. McIntosh, D. J. Porteous, and K. 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 pp. 39, 118).

[53] Y. Wang, T. J. Gorrie-Stone, O. A. Grant, A. D. Andrayas, X. Zhai, K. D. McDonald-Maier, and L. C. Schalkwyk. «InterpolatedXY: a two-step strategy to normalize DNA methylation microarray data avoiding sex bias». In: *Bioinformatics* 38.16 (2022), pp. 3950–3957. DOI: 10.1093/bioinformatics/btac436 (cit. on pp. 39, 42).

[54] K. D. Hansen. *IlluminaHumanMethylation450kanno.ilmn12.hg19: Annotation for Illumina's 450k methylation arrays*. R package version 0.6.0. Bioconductor. URL: https://bioconductor.org/packages/IlluminaHumanMethylation450kanno.ilmn12.hg19 (cit. on p. 39).

[55] Illumina Inc. *Infinium MethylationEPIC BeadChip Manifest File v1.0 B5*. https://support.illumina.com/. Accessed November 2025. 2016 (cit. on pp. 40, 49).

[56] B. Thienpont, J. Steinbacher, H. Zhao, F. D'Anna, A. Kuchnio, A. Ploumakis, B. Ghesquière, L. Van Dyck, B. Boeckx, L. Schoonjans, E. Hermans, F. Amant, V. N. Kristensen, K. P. Koh, M. Mazzone, M. L. Coleman, T. Carell, P. Carmeliet, and D. Lambrechts. «Tumour hypoxia causes DNA hypermethylation by reducing TET activity». In: *Nature* 537.7618 (2016), pp. 63–68. DOI: 10.1038/nature19081 (cit. on p. 40).

[57] L. E. Reinius, N. Acevedo, M. Joerink, G. Pershagen, S. E. Dahlén, D. Greco, C. Söderhäll, A. Scheynius, and J. Kere. «Differential DNA methylation in purified human blood cells: Implications for cell lineage and studies on disease susceptibility». In: *PLOS ONE* 7.7 (2012), e41361. DOI: 10.1371/journal.pone.0041361 (cit. on p. 40).

[58] Illumina Inc. *Infinium HumanMethylation450 BeadChip: Product Files and Manifest (v1.2)*. https://support.illumina.com/downloads/infinium_humanmethylation450_product_files.html. Accessed December 2025. 2011 (cit. on p. 42).

[59] B. H. Chen and W. Zhou. «mLiftOver: harmonizing data across Infinium DNA methylation platforms». In: *Bioinformatics* 40.7 (2024). DOI: 10.1093/bioinformatics/btae423 (cit. on p. 46).

[60] R. Pidsley, C. C. Y. Wong, M. Volta, K. Lunnon, J. Mill, and L. C. Schalkwyk. «A data-driven approach to preprocessing Illumina 450K methylation array data». In: *BMC Genomics* 14.1 (2013), p. 293. DOI: 10.1186/1471-2164-14-293 (cit. on p. 47).

[61] P. Hall, J. S. Marron, and A. Neeman. «Geometric representation of high dimension, low sample size data». In: *Journal of the Royal Statistical Society: Series B* 67.3 (2005), pp. 427–444. DOI: 10.1111/j.1467-9868.2005.00510.x (cit. on pp. 53, 54, 59, 62).

[62] D. L. Donoho. *High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality*. Tech. rep. Lecture manuscript, AMS Conference on Math Challenges of the 21st Century. Stanford University, 2000 (cit. on pp. 53, 59, 62, 64).

[63] P. Smialowski, D. Frishman, and S. Kosol. «Pitfalls of supervised feature selection». In: *Bioinformatics* 26.3 (2010), pp. 440–443. DOI: 10.1093/bioinformatics/btp621 (cit. on pp. 53, 55, 59).

[64] K. D. Hansen, W. Timp, H. C. Bravo, S. Sabunciyan, B. Langmead, O. G. McDonald, B. Wen, H. Wu, Y. Liu, D. Diep, E. Briem, K. Zhang, R. A. Irizarry, and A. P. Feinberg. «Increased methylation variation in epigenetic domains across cancer types». In: *Nature Genetics* 43.8 (2011), pp. 768–775. DOI: 10.1038/ng.865 (cit. on pp. 54, 61, 62, 66).

[65] M. J. Aryee, A. E. Jaffe, H. Corrada-Bravo, C. Ladd-Acosta, A. P. Feinberg, K. D. Hansen, and R. A. Irizarry. «Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays». In: *Bioinformatics* 30.10 (2014), pp. 1363–1369. DOI: 10.1093/bioinformatics/btu049 (cit. on pp. 54, 62).

[66] C. Luo, C. L. Keown, L. Kurihara, J. Zhou, Y. He, J. Li, R. Castanon, J. Lucero, J. R. Nery, J. P. Sandoval, B. Bui, T. J. Sejnowski, T. T. Harkins, E. A. Mukamel, M. M. Behrens, and J. R. Ecker. «Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex». In: *Science* 357.6351 (2017), pp. 600–604. DOI: 10.1126/science.aan3351 (cit. on p. 54).

[67] T. Fleischer, X. Tekpli, A. Mathelier, S. Wang, D. Nebdal, H. P. Dhakal, K. K. Sahlberg, E. Schlichting, Oslo Breast Cancer Research Consortium (OSBREAC), A. L. Børresen-Dale, E. Borgen, B. Naume, R. Eskeland, A. Frigessi, J. Tost, A. Hurtado, and V. N. Kristensen. «DNA methylation at

enhancers identifies distinct breast cancer lineages». In: *Nature Communications* 8 (2017), p. 1379. DOI: 10.1038/s41467-017-00510-x (cit. on pp. 54, 61, 66).

[68] G. Poste. «Bring on the biomarkers». In: *Nature* 469.7329 (2011), pp. 156–157. DOI: 10.1038/469156a (cit. on p. 55).

[69] R. D. Edgar, M. J. Jones, W. P. Robinson, and M. 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. DOI: 10.1186/s13148-017-0320-z (cit. on pp. 56–58, 119, 120, 122, 123, 128, 129).

[70] C. Bock. «Analysing and interpreting DNA methylation data». In: *Nature Reviews Genetics* 13.10 (2012), pp. 705–719. DOI: 10.1038/nrg3273 (cit. on pp. 57, 123, 129).

[71] R. Bourgon, R. Gentleman, and W. Huber. «Independent filtering increases detection power for high-throughput experiments». In: *Proceedings of the National Academy of Sciences* 107.21 (2010), pp. 9546–9551. DOI: 10.1073/pnas.0914005107 (cit. on p. 58).

[72] H. B. Mann and D. R. Whitney. «On a Test of Whether One of Two Random Variables Is Stochastically Larger Than the Other». In: *The Annals of Mathematical Statistics* 18.1 (1947), pp. 50–60. DOI: 10.1214/aoms/1177730491 (cit. on p. 59).

[73] R. Fagin, R. Kumar, and D. Sivakumar. «Comparing Top-k Lists». In: *SIAM Journal on Discrete Mathematics* 17.1 (2003), pp. 134–160. DOI: 10.1137/S0895480102412856 (cit. on p. 60).

[74] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. «Rank Aggregation Methods for the Web». In: *Proceedings of the 10th International Conference on World Wide Web (WWW 2001)*. ACM, 2001, pp. 613–622. DOI: 10.1145/371920.372165 (cit. on p. 60).

[75] D. M. Witten and R. Tibshirani. «A Framework for Feature Selection in Clustering». In: *Journal of the American Statistical Association* 105.490 (2012), pp. 713–726. DOI: 10.1198/jasa.2010.tm09415 (cit. on pp. 63, 65).

[76] J. H. Friedman, T. Hastie, and R. Tibshirani. «Regularization Paths for Generalized Linear Models via Coordinate Descent». In: *Journal of Statistical Software* 33.1 (2010), pp. 1–22. DOI: 10.18637/jss.v033.i01 (cit. on p. 63).

[77] Z. Sondka, S. Bamford, C. G. Cole, S. A. Ward, I. Dunham, and S. A. Forbes. «The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers». In: *Nature Reviews Cancer* 18.11 (2018), pp. 696–705. DOI: 10.1038/s41568-018-0060-1 (cit. on pp. 83, 86, 89).

[78] COSMIC Cancer Gene Census. *COSMIC Cancer Gene Census, version 103 (GRCh38)*. https://cancer.sanger.ac.uk/cosmic/download/cosmic/v103/cancergenecensus. Wellcome Sanger Institute, accessed February 2026. 2023 (cit. on p. 83).

[79] C. Cortes and V. Vapnik. «Support-vector networks». In: *Machine Learning* 20.3 (1995), pp. 273–297. DOI: 10.1007/BF00994018 (cit. on p. 83).

[80] T. M. Cover and P. E. Hart. «Nearest neighbor pattern classification». In: *IEEE Transactions on Information Theory* 13.1 (1967), pp. 21–27. DOI: 10.1109/TIT.1967.1053964 (cit. on p. 84).

[81] G. B. Dantzig. «Discrete-variable extremum problems». In: *Operations Research* 5.2 (1957), pp. 266–288. DOI: 10.1287/opre.5.2.266 (cit. on p. 86).

[82] X.-L. Wu, J. Xu, G. Feng, G. R. Wiggans, J. F. Taylor, J. He, C. Qian, J. Qiu, B. Simpson, J. Walker, and S. Bauck. «Optimal Design of Low-Density SNP Arrays for Genomic Prediction: Algorithm and Applications». In: *PLOS ONE* 11.9 (2016), e0161719. DOI: 10.1371/journal.pone.0161719 (cit. on p. 86).

[83] S. Nishiyama, K. Sato, and R. Tao. «Integer programming for selecting set of informative markers in paternity inference». In: *BMC Bioinformatics* 23 (2022), p. 265. DOI: 10.1186/s12859-022-04801-z (cit. on p. 86).

[84] M. V. Kuleshov, M. R. Jones, A. D. Rouillard, N. F. Fernandez, Q. Duan, Z. Wang, S. Koplev, S. L. Jenkins, K. M. Jagodnik, A. Lachmann, M. G. McDermott, C. D. Monteiro, G. W. Gundersen, and A. Ma'ayan. «Enrichr: a comprehensive gene set enrichment analysis web server 2016 update». In: *Nucleic Acids Research* 44.W1 (2016), W90–W97. DOI: 10.1093/nar/gkw377 (cit. on p. 89).

[85] M. Panagopoulou, T. Panou, A. Gkountakos, G. Tarapatzi, M. Karaglani, I. Tsamardinos, and E. Chatzaki. «BRCA1 & BRCA2 methylation as a prognostic and predictive biomarker in cancer: Implementation in liquid biopsy in the era of precision medicine». In: *Clinical Epigenetics* 16 (2024), p. 178. DOI: 10.1186/s13148-024-01787-8 (cit. on p. 93).

[86] R. Graffeo, H. Q. Rana, F. Conforti, B. Bonanni, M. J. Cardoso, S. Paluch-Shimon, O. Pagani, A. Goldhirsch, A. H. Partridge, M. Lambertini, and J. E. Garber. «Moderate penetrance genes complicate genetic testing for breast cancer diagnosis: ATM, CHEK2, BARD1 and RAD51D». In: *The Breast* 65 (2022), pp. 32–40. DOI: 10.1016/j.breast.2022.06.003 (cit. on p. 93).

[87] S. Liu, J. Huang, Y. Zhang, Y. Liu, S. Zuo, and R. Li. «MAP2K4 interacts with Vimentin to activate the PI3K/AKT pathway and promotes breast cancer pathogenesis». In: *Aging* 11.22 (2019), pp. 10697–10710. DOI: 10.18632/aging.102485 (cit. on p. 93).

[88] M. M. Moasser. «The oncogene HER2: its signaling and transforming functions and its role in human cancer pathogenesis». In: *Oncogene* 26.45 (2007), pp. 6469–6487. DOI: 10.1038/sj.onc.1210477 (cit. on p. 93).

[89] J. Tang, M. Lu, Q. Cui, D. Zhang, D. Kong, X. Liao, J. Ren, Y. Gong, and G. Wu. «Overexpression of ASPM, CDC20, and TTK Confer a Poorer Prognosis in Breast Cancer Identified by Gene Co-expression Network Analysis». In: *Frontiers in Oncology* 9 (2019), p. 310. DOI: 10.3389/fonc.2019.00310 (cit. on p. 93).

[90] C. Tognon, S. R. Knezevich, D. Huntsman, C. D. Roskelley, N. Melnyk, J. A. Mathers, L. Becker, F. Carneiro, N. MacPherson, D. Horsman, and P. H. B. Sorensen. «Expression of the ETV6–NTRK3 gene fusion as a primary event in human secretory breast carcinoma». In: *Cancer Cell* 2.5 (2002), pp. 367–376. DOI: 10.1016/S1535-6108(02)00180-0 (cit. on pp. 94, 112).

[91] Y. Wu, M. Alvarez, D. J. Slamon, P. Koeffler, and J. V. Vadgama. «Caspase 8 and maspin are downregulated in breast cancer cells due to CpG site promoter methylation». In: *BMC Cancer* 10 (2010), p. 32. DOI: 10.1186/1471-2407-10-32 (cit. on p. 96).

[92] H. Han, Y. Chen, L. Cheng, E. V. Prochownik, and Y. Li. «microRNA-206 impairs c-Myc-driven cancer in a synthetic lethal manner by directly inhibiting MAP3K13». In: *Oncotarget* 7.13 (2016), pp. 16409–16419. DOI: 10.18632/oncotarget.7653 (cit. on p. 96).

[93] C. J. Ormandy, E. A. Musgrove, R. Hui, R. J. Daly, and R. L. Sutherland. «Cyclin D1, EMS1 and 11q13 amplification in breast cancer». In: *Breast Cancer Research and Treatment* 78.3 (2003), pp. 323–335. DOI: 10.1023/A:1023033708204 (cit. on p. 97).

[94] C. Li, G. Zhang, Y. Wang, B. Chen, K. Li, L. Cao, C. Ren, L. Wen, M. Jia, H. Mok, J. Lai, W. Xiao, X. Li, and N. Liao. «Spectrum of MAP3K1 mutations in breast cancer is luminal subtype-predominant and related to prognosis». In: *Oncology Letters* 23.2 (2022), p. 68. DOI: 10.3892/ol.2022.13187 (cit. on p. 97).

[95] Cancer Genome Atlas Network. «Comprehensive molecular portraits of human breast tumours». In: *Nature* 490.7418 (2012), pp. 61–70. DOI: 10.1038/nature11412 (cit. on pp. 98, 101, 102, 111).

[96] C. M. Fillmore, P. B. Gupta, J. A. Rudnick, S. Caballero, P. J. Keller, E. S. Lander, and C. Kuperwasser. «Estrogen expands breast cancer stem-like cells through paracrine FGF/Tbx3 signaling». In: *Proceedings of the National Academy of Sciences* 107.50 (2010), pp. 21737–21742. DOI: 10.1073/pnas.1007863107 (cit. on pp. 100, 111).

[97] M.-L. Asselin-Labat, K. D. Sutherland, H. Barker, R. Thomas, M. Shackleton, N. C. Forrest, L. Hartley, L. Robb, F. G. Grosveld, J. van der Wees, G. J. Lindeman, and J. E. Visvader. «Gata-3 is an essential regulator of mammary-gland morphogenesis and luminal-cell differentiation». In: *Nature Cell Biology* 9.2 (2007), pp. 201–209. DOI: 10.1038/ncb1530 (cit. on pp. 101, 111).

[98] A. A. WalyEldeen, S. Sabet, S. E. Anis, T. Stein, and A. M. Ibrahim. «FBLN2 is associated with basal cell markers Krt14 and ITGB1 in mouse mammary epithelial cells and has a preferential expression in molecular subtypes of human breast cancer». In: *Breast Cancer Research and Treatment* 208 (2024), pp. 673–686. DOI: 10.1007/s10549-024-07447-y (cit. on p. 102).

[99] A. Naba, K. R. Clauser, J. M. Lamar, S. A. Carr, and R. O. Hynes. «Extracellular matrix signatures of human mammary carcinoma identify novel metastasis promoters». In: *eLife* 3 (2014), e01308. DOI: 10.7554/eLife.01308 (cit. on p. 102).

[100] J.-P. Fortin, N. Cullen, Y. I. Sheline, W. D. Taylor, A. Aselcioglu, P. A. Cook, P. Adams, C. Cooper, M. Fava, P. J. McGrath, M. McInnis, M. L. Phillips, M. H. Trivedi, M. M. Weissman, and R. T. Shinohara. «Harmonization of cortical thickness measurements across scanners and sites». In: *NeuroImage* 167 (2018), pp. 104–120. DOI: 10.1016/j.neuroimage.2017.11.024 (cit. on pp. 104, 107).

[101] E. Cuttini, C. Goi, E. Pellarin, R. Vida, and C. Brancolini. «HDAC4 in cancer: a multitasking platform to drive not only epigenetic modifications». In: *Frontiers in Molecular Biosciences* 10 (2023), p. 1116660. DOI: 10.3389/fmolb.2023.1116660 (cit. on p. 111).

[102] E. Piskounova, C. Polytarchou, J. E. Thornton, R. J. LaPierre, C. Pothoulakis, J. P. Hagan, D. Iliopoulos, and R. I. Gregory. «Lin28A and Lin28B Inhibit let-7 MicroRNA Biogenesis by Distinct Mechanisms». In: *Cell* 147.5 (2011), pp. 1066–1079. DOI: 10.1016/j.cell.2011.10.039 (cit. on p. 112).

[103] R. Xie, Y. Wang, W. Nie, W. Huang, W. Song, Z. Wang, and X. Guan. «Lin28B expression correlates with aggressive clinicopathological characteristics in breast invasive ductal carcinoma». In: *Cancer Biotherapy and Radiopharmaceuticals* 29.5 (2014), pp. 215–220. DOI: 10.1089/cbr.2014.1610 (cit. on p. 112).

[104] L. P. de Lima Camillo, R. Sehgal, J. Armstrong, H. E. Miller, J. A. Lasky-Su, A. T. Higgins-Chen, S. Horvath, and B. Wang. «CpGPT: A Foundation Model for DNA Methylation». In: *bioRxiv* (2024). DOI: 10.1101/2024.10.24.619766 (cit. on p. 116).

[105] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. «Attention Is All You Need». In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017. DOI: 10.48550/arXiv.1706.03762 (cit. on p. 116).

[106] Q. Zhang, C. L. Vallerga, R. M. Walker, T. Lin, A. K. Henders, G. W. Montgomery, J. He, D. Fan, J. Fowdar, M. Kennedy, T. Pitcher, J. Pearson, G. Halliday, J. B. Kwok, I. Hickie, S. Lewis, T. Anderson, P. A. Silburn, G. D. Mellick, S. E. Harris, P. Redmond, A. D. Murray, D. J. Porteous, C. S. Haley, K. L. Evans, A. M. McIntosh, J. Yang, J. Gratten, R. E. Marioni, N. R. Wray, I. J. Deary, A. F. McRae, and P. M. Visscher. «Improved precision of epigenetic clock estimates across tissues and its implication for biological ageing». In: *Genome Medicine* 11.1 (2019), p. 54. DOI: 10.1186/s13073-019-0667-1 (cit. on pp. 132, 133).

[107] E. B. Gold. «The timing of the age at which natural menopause occurs». In: *Obstetrics and Gynecology Clinics of North America* 38.3 (2011), pp. 425–440. DOI: 10.1016/j.ogc.2011.05.002 (cit. on p. 135).