Politecnico di Torino

Master's Degree in Mathematical Engineering

Material for Thesis

3– About GSE67919's Application

Elisabetta Roviera s328422

Contents

Note The papers summarized in this report represent the main references related to the GSE69914 dataset and to the preliminary filtering and quality control of CpG sites performed prior to the core analyses. These studies provide the methodological and technical background necessary to understand probe reliability, normalization strategies, and preprocessing pipelines applied to Illumina HumanMethylation450 data. Additional references may be integrated in the future to refine specific analytical steps. For a complete understanding of the concepts and results discussed, please refer to the original publications cited in the bibliography of this document.

1 Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray

Keywords DNA methylation, CpG sites, Illumina 450K array, cross-reactive probes, polymorphic CpGs, SNPs, microarray reliability, sex-associated methylation, probe specificity [1]

Cross-reactive probes Approximately 6% of the probes on the Illumina HumanMethylation450 array hybridize to multiple genomic regions with high sequence similarity ($\geq$94%). These cross-reactive probes can generate spurious methylation signals, particularly in autosomal sites that co-hybridize with sex chromosomes. As a result, apparent sex-associated methylation differences may arise as technical artifacts rather than biological effects. Probes showing $\geq$47 matched bases to unintended genomic loci were classified as cross-reactive and should be excluded from downstream analyses.

Polymorphic CpGs About 13.8% of probes overlap known single nucleotide polymorphisms (SNPs), affecting either the cytosine, guanine, or the base immediately preceding the CpG site. These polymorphic CpGs reflect underlying genetic variation instead of true methylation differences. Their inclusion can distort methylation quantitative trait loci (mQTL) analyses or group comparisons if genotype effects are not accounted for.

Implications for preprocessing Cross-reactive and polymorphic probes can introduce false biological associations and reduce reproducibility. For accurate methylation profiling, these probes must be filtered before normalization and statistical modeling. The lists provided by Chen et al. are now widely used as reference sets in preprocessing pipelines for datasets such as GSE69914.

Impact on Illumina 450K analysis This study established essential quality-control guidelines for HumanMethylation450 data. By identifying unreliable probes and defining sequence-based exclusion criteria, it laid the groundwork for standardized preprocessing workflows and for accurate biological interpretation of CpG methylation patterns.

2 DNA methylation outliers in normal breast tissue identify field defects that are enriched in cancer

Keywords Breast cancer, field defects, DNA methylation, Illumina 450K, epigenetic outliers, differential variability, iEVORA, DVMCs, adipose deconvolution, WNT signaling, GSE69914 [2].

Study design and cohorts Genome-wide DNA methylation (DNAm) was profiled using the Illumina HumanMethylation450 BeadChip in a total of 407 breast tissue samples. The core discovery set included: (i) 50 normal/benign breast tissue samples from cancer-free women, (ii) 42 normal breast tissue samples adjacent to invasive breast cancers from the same patients ("normal-adjacent"), and (iii) 305 breast cancer samples, including 42 that were matched to the normal-adjacent tissues. Additional independent cohorts were used for validation: normal breast from reduction mammoplasty, normal-adjacent tissue from other patients, ductal carcinoma in situ (DCIS), and TCGA breast cancers. The goal was to detect *epigenetic field defects*: focal methylation alterations already present in histologically normal tissue near a tumor, potentially representing early clonal expansions that precede malignancy.

Data preprocessing and quality control DNA was bisulfite converted and hybridized to the Illumina 450K array following standard protocol. Raw IDATs were processed with the `minfi` Bioconductor package. Probes with detection $p > 0.01$ were set to missing; CpG sites with $> 1\%$ missing values across all samples were removed. Remaining missing values were imputed using k-nearest neighbors ($k = 5$). Because Infinium I and Infinium II chemistries have different signal distributions, each sample was normalized with BMIQ (Beta MIxture Quantile normalization) to correct the type-II probe bias. After intra-sample normalization, the data matrix contained ∼485,000 probes across 397 primary discovery samples. Inter-sample effects were evaluated using singular value decomposition (SVD): the leading components of variation tracked biological factors such as normal vs. tumor state rather than technical batches, indicating no dominant batch confounding. The dataset was later deposited as GEO accession GSE69914, which is the same platform and pipeline used in this thesis.

Adjustment for adipose content (cell-type heterogeneity) Normal breast tissue contains a large adipose component, so cellular composition can confound DNAm differences. A reference-based deconvolution strategy was implemented to estimate, per sample, the proportion of adipose vs. epithelial/stromal signal. Reference 450K methylation profiles were collected for human mammary epithelial cells (HMEC) and for adipose tissue, using ENCODE and independent fat tissue datasets. From these references, 1,320 CpGs were selected as markers: CpGs with absolute beta-value difference > 0.7 between HMEC and adipose, and located in DNase hypersensitive regions (cell-type-informative, regulatory sites). For each mixed sample, constrained projection (CP) was applied to infer the relative fat fraction $w(\text{FAT})$ and the complementary HMEC/stromal fraction $w(\text{HMEC}) = 1 - w(\text{FAT})$. This deconvolution step was validated on independent adipose data and confirmed that the top global source of DNAm variation across normal tissues correlates with fat content. Importantly, fat content did *not* differ significantly between normal and normal-adjacent samples, and adjusting for fat content did not yield genome-wide significant differentially methylated CpGs in standard mean-based testing. Therefore, large compositional shifts in adipose alone do not explain the epigenetic differences of interest.

Differential variability vs. differential mean Instead of assuming that early carcinogenic changes are uniform across all patients (which is the logic of standard differential methylation of the mean, DM), the analysis explicitly targeted *heterogeneous, stochastic* alterations that may appear only in a subset of at-risk cells. For each CpG, two complementary statistics were considered comparing normal vs. normal-adjacent tissue:

1. Differential variability (DV): Bartlett's test on variance of beta-values between the two groups.
2. Differential mean (DM): a standard two-sample t-test on group means.

Classical DM alone (mean shifts) did *not* detect any CpG at genome-wide significance after multiple testing (false discovery rate, FDR ≈ 0.3). In contrast, testing DV revealed widespread CpGs whose *variance* was significantly higher in normal-adjacent tissue, consistent with focal epigenetic hits present only in some cells or subclones.

iEVORA algorithm and definition of DVMCs To systematically extract these heterogeneous events, the study introduced **iEVORA** (improved Epigenetic Variable Outliers for Risk prediction Algorithm). The pipeline is:

1. For each CpG, run Bartlett's test comparing variance in normal vs. normal-adjacent tissue. CpGs passing a stringent threshold FDR < 0.001 are called differentially variable CpGs (DVCs).
2. Because a single extreme outlier can inflate variance, DVCs are then re-ranked using the t-statistic for differential mean methylation between groups. Only CpGs with unadjusted $p < 0.05$ for the mean shift are retained.
3. The retained set are called **DVMCs** (Differentially Variable and Differentially Methylated CpGs): they are both more variable (i.e. show outlier behavior) and directionally shifted.

Applying iEVORA to the 50 normal vs. 42 normal-adjacent samples yielded **7,318 DVMCs** ($\sim$1.5% of all interrogated CpGs), with the majority showing **increased variance** and **hypermethylation** in normal-adjacent tissue. Typical patterns at these loci are not subtle drifts: they show $\sim$20–30% jumps in beta-value in a subset of normal-adjacent samples, consistent with clonal epigenetic lesions. These lesions were often promoter-proximal (within 1.5 kb upstream of TSS) when hypermethylated, and enriched in regulatory regions controlling differentiation. The distribution of DVMC load per patient was highly uneven: most normal-adjacent samples had only a few altered CpGs, but some samples showed hundreds to thousands of altered loci, suggesting different levels of "field damage" around the tumor.

Probe reliability and exclusion of technical artifacts The analysis explicitly considered known Illumina 450K probe issues. Cross-reactive probes and polymorphic CpGs (as catalogued by Chen et al. 2013 for the 450K array) can create artificial methylation signals due to off-target hybridization or SNP overlap. Roughly 19% of 450K probes fall into these problematic categories globally. Among the 7,318 DVMCs identified by iEVORA, only 923 overlapped the Chen blacklist, far fewer than expected by chance, and the most biologically interesting class (hypervariable + hypermethylated DVMCs) was strongly *under*-enriched for problematic probes. All major downstream results remained valid after removing these potentially confounded probes. This supports that DVMCs represent true biological alterations, not array artifacts.

Validation in independent cohorts and progression to cancer The same DVMCs were tested in an *independent* cohort containing normal breast tissue from cancer-free women and normal-adjacent tissue from other patients. Over 60% of hypervariable DVMCs showed higher alteration frequency in normal-adjacent vs. normal also in this second dataset, confirming reproducibility. Next, progression was assessed by comparing DNAm in invasive breast cancers to healthy normals, and also within matched normal-adjacent vs. tumor pairs. DVMCs (especially those hypervariable + hypermethylated in normal-adjacent tissue) showed markedly stronger methylation deviations in the cancers, and in many cases the same CpG sites became more uniformly hypermethylated across tumors. Up to $\sim$32% of these hypervariable/hypermethylated DVMCs gained *further* methylation in the tumor, while only $\sim$2% reversed direction. This indicates that the outlier methylation changes seen in histologically normal tissue are not random noise: they expand and consolidate in the tumor, behaving like early epigenetic field defects that get clonally fixed during transformation.

Pathway-level and regulatory context The DVMCs are not randomly scattered. They are significantly enriched in binding sites of transcription factors (TFs) linked to chromatin architecture and Polycomb repression, including EZH2 and SUZ12 (PRC2 complex), as well as CTCF and RAD21. Regions bound by these factors tended to gain DNA methylation first in normal-adjacent tissue and then even more in cancer. Many DVMCs localize near promoters of genes involved in developmental and differentiation programs. Network-level enrichment analysis (FEM / EpiMod) showed coordinated hypermethylation in members of canonical pathways such as WNT and FGF signaling. Within a given patient, multiple genes in the same pathway (e.g. WNT ligands, FZD receptors, pathway modulators like *SFRP1*, *WIF1*) often showed concurrent promoter hypermethylation in the normal-adjacent sample, suggesting early, pathway-level epigenetic repression of differentiation signals.

Clinical correlations For each normal-adjacent sample and tumor, the study defined a progression score (a Z-score measuring how far the methylation profile at DVMCs deviates from the healthy normal baseline). Tumors with higher progression scores showed:

- higher proliferation index (KI67),
- larger tumor size,
- poorer overall survival.

These associations were strongest for the DVMC class that is hypervariable and hypermethylated in normal-adjacent tissue. The same progression score replicated in an independent, untreated breast cancer cohort (TCGA), indicating prognostic relevance. In matched pairs, tumors with higher deviation from their own adjacent normal tissue were more likely to be HER2-positive, linking these field defects to aggressive subtypes.

Practical takeaway for replication To reproduce this analysis on a new 450K dataset (e.g. GSE69914):

1. Import raw IDATs, drop failed probes (detection $p > 0.01$) and samples with poor control metrics.
2. Impute missing values (kNN), apply BMIQ per sample to correct type-II bias.
3. (Optional but recommended) Estimate adipose fraction using constrained projection on cell-type-informative CpGs; record fat content as a covariate.
4. Remove probes known to be cross-reactive / SNP-overlapping if desired, or at minimum flag them.
5. Split samples into biologically defined groups: e.g. healthy normal vs. "at-risk" normal (adjacent) vs. tumor.
6. Run iEVORA:
   - Bartlett's test for variance (normal vs. adjacent), control FDR < 0.001 to define DVCs.
   - Re-rank DVCs by t-test on mean difference; keep CpGs with unadjusted $p < 0.05$.
   - The retained CpGs are DVMCs (candidate field defects).
7. For each DVMC and each sample, compute deviation from the healthy-normal mean in units of standard deviations (z-score). Count significant outliers.
8. Test whether those deviations increase in matched tumors, and whether they concentrate in known regulatory elements or pathways (e.g. WNT, FGF).
9. Build per-sample progression scores and correlate with phenotype (stage, KI67, HER2, survival).

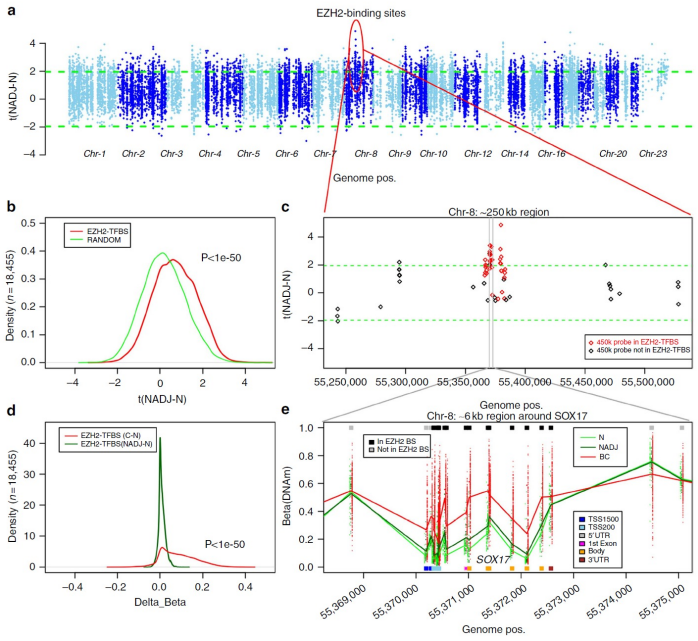

Figure 1: EZH2-associated DNA methylation patterns CpG probes located within EZH2-binding sites show consistent hypermethylation in normal-adjacent breast tissue compared with normal tissue, indicating early Polycomb-associated epigenetic changes. The density and regional plots reveal a gradual methylation increase along the normal → adjacent → cancer sequence, exemplified by the *SOX17* locus, where EZH2-bound CpGs gain methylation progressively. These patterns suggest that PRC2 targets undergo early and coordinated epigenetic activation preceding tumor development.

3 The integrative epigenomic-transcriptomic landscape of ER-positive breast cancer

Keywords ER-positive breast cancer, DNA methylation, RNA-Seq, Illumina 450K, TCGA, FEM algorithm, iCluster, luminal-A/B subtypes, WNT signaling, TGF-beta pathway, differential methylation, network analysis [3].

Study design and rationale The study integrated DNA methylation (Illumina 450K) and RNA-Seq data from 724 estrogen receptor–positive (ER+) breast cancers and 111 normal adjacent tissues from the TCGA dataset to define functional epigenetic alterations driving tumor subtypes. Using the **Functional Epigenetic Modules (FEM)** algorithm, the authors identified network-level hotspots of coordinated DNA methylation and gene-expression changes, focusing on how epigenetic regulation contributes to luminal-A and luminal-B classification.

Data preprocessing and integration Methylation data (395,775 CpGs) and RNA-Seq expression data (20,531 genes) were preprocessed using standard Illumina and `limma` normalization procedures. Genes were assigned promoter methylation values by averaging probes mapping to TSS200 or first exon regions; if unavailable, TSS1500 was used. Probes mapping to gene bodies were excluded to focus on regulatory methylation. Empirical Bayes statistics (`limma`) were computed for differential methylation and expression between normal and ER+ samples. These gene-level statistics were integrated into a protein–protein interaction (PPI) network derived from high-confidence interactomes.

The FEM algorithm Each PPI edge connecting genes g, h was weighted using an anti-correlation rule combining differential methylation (t_g^D) and expression (t_g^R):

$$w_{gh} = \frac{1}{2}\left(t_g^I + t_h^I\right), \quad t_g^I = H(t_g^D)H(-t_g^R) + H(-t_g^D)H(t_g^R),$$

where $H(x)$ is the Heaviside step function. The network was scanned for dense subnetworks ("modules") maximizing average edge weight (local modularity) using a spin-glass community detection algorithm. Significant modules were validated via permutation tests and comparison to null networks.

Identified FEM modules Nine significant FEMs were detected, encompassing 257 unique genes (146 both differentially methylated and expressed, 99 showing anti-correlation). Each FEM corresponded to a biologically meaningful hotspot. Representative examples include:

- **CAV1 module** – enriched in WNT signaling genes (e.g., *WIF1*, *WNT3A*, *SFRP1*), showing promoter hypermethylation and transcriptional repression.
- **FSTL1 module** – enriched in TGF-β/BMP signaling members (*BMP2*, *BMP6*, *BMP7*, *TGFB2*), with widespread hypermethylation of tumor suppressor components.
- **CCL11 and LEP modules** – involved in chemokine and GPCR signaling, with hypomethylation and overexpression consistent with enhanced metastatic potential.
- **PROC and MME modules** – linked to coagulation and endothelin pathways, known to support tumor cell migration and proliferation.

Validation and reproducibility Independent datasets (Germany: 254 ER+ and 49 normal tissues; Yu: 110 ER+ and 13 normals) confirmed the FEM hotspots by network modularity and directional consistency of methylation and expression t-statistics. Four of nine FEMs achieved significant validation ($p < 0.05$), and all showed concordant differential patterns across cohorts. Methylation datasets included GEO accession GSE69914, confirming overlap with the platform used in the present thesis.

Integrative clustering (iCluster analysis) Joint latent variable modeling (`iCluster`) of 463 ER+ TCGA tumors with matched DNAm and mRNA data (FEM genes only) identified exactly two integrative clusters ($k = 2$), strongly corresponding to luminal-A and luminal-B subtypes (Fisher test $p < 10^{-10}$). Luminal-B tumors exhibited significantly higher deviation scores from the normal reference in both methylation and expression, indicating stronger epigenetic deregulation rather than distinct pathway activation. A similar two-cluster structure persisted when extending the input to 4311 anti-correlated genes genome-wide, confirming the homogeneity of ER+ epigenetic architecture.

Deviation scoring and prognosis A per-sample FEM deviation score quantified the combined distance (Z-normalized) of each gene's methylation and expression from the normal baseline:

$$FEM_s = \frac{1}{m}\sum_{g=1}^{m} |Z_{gs}^D - \alpha Z_{gs}^R|,$$

with $\alpha = \sigma_Z^D/\sigma_Z^R$ scaling data-type variance. Luminal-B samples displayed higher deviation scores than luminal-A in all modules ($p < 10^{-5}$). Prognostic modeling (Cox regression across TCGA, METABRIC, and Fleischer datasets) showed that higher FEM scores and cluster membership were associated with worse survival (meta-analysis $p = 0.013$ for DNAm-based classifier).

Coordination of methylation–expression changes Unlike copy number alterations, methylation changes across FEM genes were found to be **coordinated**, not mutually exclusive, within tumors. Binary matrices of gene activation (1) vs normal-like (0) states revealed significantly smaller Manhattan distances between FEM genes than expected under random permutation ($p < 0.001$), demonstrating intra-tumor coherence of epigenetic deregulation.

Biological interpretation ER+ luminal-A and luminal-B cancers share the same deregulated epigenetic pathways, dominated by silencing of WNT and BMP/TGF-*beta* signaling antagonists. Luminal-B tumors exhibit larger magnitude deviations—reflecting stronger pathway repression, higher proliferation (correlation of FEM score with PCNA $r \approx 0.38$), and worse clinical outcomes. Epigenetic silencing of *WIF1*, *SFRP1*, and *FSTL1* likely enhances WNT/TGF-β signaling activity, promoting cell self-renewal and EMT. Additional deregulation in chemokine and endothelin pathways underscores a coordinated shift toward proliferative and migratory phenotypes.

Summary of methodology for replication To reproduce the FEM-based integration in a 450K dataset (e.g., GSE69914):

1. Preprocess IDATs (filter detection $p > 0.01$; impute missing values $k = 5$; exclude body probes).
2. Compute per-gene differential methylation and expression using empirical Bayes (`limma`).
3. Build a weighted PPI using the Heaviside anti-correlation rule.
4. Apply spin-glass modularity optimization to extract FEMs; validate with 1000 permutations.
5. Perform pathway enrichment (MSigDB) and compute per-sample FEM deviation scores.
6. Integrate FEM DNAm and mRNA matrices via `iCluster` to identify subtypes ($k = 2$).
7. Optionally, compare FEM-derived subtypes to luminal-A/B and assess prognostic value.

Conclusion The integrative analysis reveals that ER+ breast cancers form two principal epigenetic clusters corresponding to luminal-A and luminal-B phenotypes. Both subtypes share deregulated WNT and TGF-β/BMP networks, but luminal-B tumors exhibit greater magnitude of methylation-driven transcriptional repression. The study establishes a reproducible, network-based framework for identifying epigenetic driver modules and for quantifying coordinated methylation–expression shifts in large-scale 450K datasets.

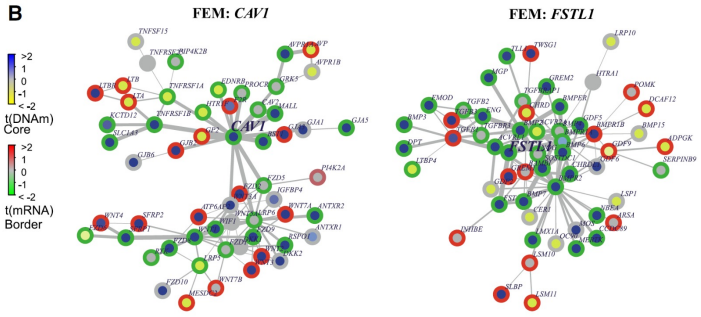

Figure 2: Examples of two FEMs centred around seed genes CAV1 and FSTL1 in ER+ breast cancer.

4 Integrative analysis identifies potential DNA methylation biomarkers for pan-cancer diagnosis and prognosis

Keywords DNA methylation, pan-cancer analysis, CpG biomarkers, TCGA, XGBoost, logistic regression, diagnostic classifier, prognostic model, differential methylation, feature selection, GSE69914 [4].

Study design and objectives This large-scale integrative study aimed to identify CpG-based DNA methylation biomarkers with diagnostic and prognostic potential across multiple cancer types. The authors systematically analyzed genome-wide methylation data from 26 cancer types in The Cancer Genome Atlas (TCGA), covering 9,685 tumor samples and 729 matched normal tissues. The central goal was to develop a minimal CpG signature able to accurately (i) distinguish tumors from normal samples across cancers, (ii) identify the tissue of origin, and (iii) predict patient prognosis.

Data preprocessing and feature selection DNA methylation β-values (Illumina HumanMethylation450K array) were retrieved from TCGA. CpG probes with missing values in more than 20% of samples or low variance ($\sigma^2 < 0.01$) were removed, leaving approximately 350,000 high-quality CpGs. Batch effects were minimized through quantile normalization and cross-cancer scaling of β-values to the [0,1] range. For each cancer type, differentially methylated CpGs (DMCs) were identified using a two-sided t-test comparing tumor vs. normal samples, controlling false discovery rate (FDR < 0.05) and $|\Delta\beta| > 0.2$. These DMCs were pooled across cancers to obtain a global candidate set of $\sim$3,500 CpGs, which served as input for model training.

Diagnostic model construction (XGBoost + logistic regression) An extreme gradient boosting (**XGBoost**) model was used for feature selection to capture non-linear relationships among CpG methylation levels. The dataset was split into training (70%) and testing (30%) sets while maintaining cancer type stratification. Each CpG feature was assigned an importance score based on gain and coverage metrics from XGBoost iterations. The top 30 CpGs with the highest feature importance were retained, and a multivariate **logistic regression classifier** was trained on their β-values. Recursive feature elimination was then applied to minimize redundancy, resulting in a final diagnostic signature of **seven CpGs**. These seven CpGs achieved an average area under the ROC curve (AUC) of 0.982 in 10-fold cross-validation, and above 0.95 in nine independent cancer cohorts. Validation was also conducted on external GEO datasets, including breast tissue datasets such as GSE69914 and GSE76938, confirming the robustness of the selected CpGs across platforms.

Functional and genomic annotation of the 7 CpGs These 7 CpGs function as a universal "epigenetic fingerprint" that clearly separates the methylation profiles of healthy tissues from those of tumors. The 7 CpG sites are distributed across genes involved in tumorigenesis and transcriptional regulation:

- **cg08244313 (ANKRD11)** – located in the promoter of a chromatin remodeling gene frequently mutated in breast and lung cancer; hypermethylation leads to transcriptional silencing and impaired cell differentiation.
- **cg17735539 (ZNF582)** – located in a zinc-finger transcription factor promoter; hypermethylation is recurrent in cervical, colon, and breast cancers, acting as a universal tumor suppressor marker.
- **cg21361244 (TRIM15)** – associated with ubiquitin-mediated protein degradation; its hypomethylation correlates with higher expression in invasive tumors.
- **cg26157345 (CCDC181)** – localized in the gene body of a microtubule-associated protein; consistent hypermethylation across epithelial cancers suggests a pan-epithelial marker.
- **cg11510243 (PDLIM4)** – involved in cytoskeletal anchoring and tumor suppression; promoter hypermethylation represses expression and promotes migration and metastasis.
- **cg12542207 (SPG20)** – regulates cell cycle and WNT signaling; hypermethylated in multiple carcinomas, including breast, liver, and colon.
- **cg18081940 (ZSCAN18)** – zinc-finger transcription regulator; hypermethylated in breast and endometrial cancers, associated with chromatin repression and proliferation.

Collectively, these CpGs represent **pan-cancer epigenetic switches** targeting genes involved in chromatin regulation, cytoskeletal integrity, and transcriptional control—biological processes frequently altered in early tumorigenesis. Their hypermethylation consistently marks the transition from normal to malignant epigenetic states.

Model validation and cross-cancer generalization To evaluate generalizability, the seven-CpG classifier was tested on unseen TCGA cancers (e.g., prostate, ovarian, kidney, brain). All achieved diagnostic AUC > 0.97, confirming that the same CpG panel discriminates tumors from normals regardless of tissue origin. Moreover, an extended classifier including 12 additional CpGs was trained to predict cancer type (i.e., tissue of origin). This model achieved a macro-AUC of 0.95 across 26 TCGA cancers, correctly classifying both primary and metastatic samples in over 90% of cases. The consistency of CpG methylation patterns across datasets and tissues underscores the universality of these epigenetic changes.

Prognostic analysis To assess survival relevance, patients were divided into high- and low-risk groups based on their mean methylation at the seven diagnostic CpGs. Kaplan–Meier and Cox proportional hazards analyses revealed that higher methylation of these CpGs correlated with shorter overall survival in seven cancer types, notably in breast, colon, and lung cancers (log-rank $p < 0.001$). The prognostic classifier demonstrated stable performance across cohorts, indicating that CpG methylation at these loci reflects tumor aggressiveness and progression dynamics.

Interpretation and reproducibility The seven identified CpGs form a minimal yet highly predictive signature of cancer-specific methylation. Their diagnostic capacity stems from consistent hypermethylation of regulatory promoters and transcription factor binding regions across epithelial cancers. Importantly, these CpGs can be measured on the Illumina 450K or EPIC arrays, allowing direct replication in datasets such as GSE69914. The pipeline—DMC selection, XGBoost feature ranking, logistic regression training, and ROC-based validation—offers a reproducible framework for building multi-cancer methylation classifiers.

Conclusion This integrative analysis demonstrates that DNA methylation profiling can yield robust, universal biomarkers for cancer detection and prognosis. The seven CpGs identified by Ding et al. serve as a compact and biologically interpretable panel capturing core epigenetic alterations across multiple tumor types. Their consistent hypermethylation patterns make them ideal candidates for clinical diagnostic assays and for cross-dataset validation in studies such as the present thesis.

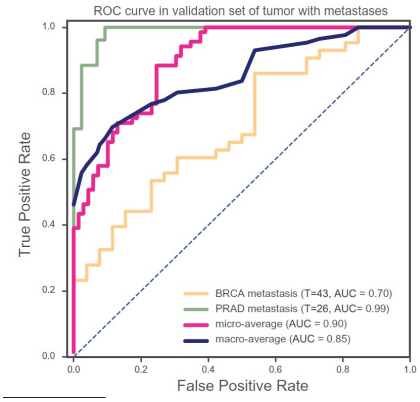

Figure 3: Validation of tumor specific classifier in tumors with metastases. ROC curve of multiclass tumor specific classifier in metastatic breast cancer (GSE58999) and metastatic prostate cancer (GSE73549 and GSE38240).

5 DNA Methylation Patterns in Normal Tissue Correlate more Strongly with Breast Cancer Status than Copy-Number Variants

Keywords DNA methylation, copy-number variants (CNVs), breast cancer, epigenetic field defects, iEVORA algorithm, EpiDISH, cell-type deconvolution, GSE67919, GSE69914, risk prediction [5].

Study objectives Gao, Widschwendter and Teschendorff (2018) investigated whether epigenetic or genetic alterations in normal breast tissue are more predictive of cancer risk. Specifically, they compared the ability of DNA methylation (DNAm) and copy-number variation (CNV) profiles—obtained from the same Illumina HumanMethylation450K arrays—to discriminate between **normal-healthy**, **normal-adjacent to tumor**, and **cancerous** breast samples. The analysis aimed to test whether DNAm changes in the normal epithelium better reflect cancer field defects than CNV alterations.

Datasets and preprocessing Two independent breast tissue cohorts were analyzed:

- The **Erlangen cohort**: 50 normal-healthy, 42 normal-adjacent, and 305 breast cancers, profiled on the Illumina 450K array.
- The **Validation cohort (GSE67919)**: 18 normal-healthy (from reduction mammoplasty) and 70 normal-adjacent samples.

Data normalization and background correction were performed using the `minfi` R package. CNV inference was derived directly from methylation signal intensities using the `conumee` package, ensuring both methylation and copy-number profiles were obtained from identical assays, avoiding batch effects.

Cell-type deconvolution and reference construction Given the cellular heterogeneity of breast tissue (epithelial, adipose, immune), the authors built a reference DNAm database of 349 CpGs discriminating nine major cell types (epithelial, adipocytes, and seven immune subtypes). The reference was validated using: (i) independent datasets (ENCODE, Blueprint), (ii) in-silico mixed cell populations, and (iii) purified cell samples. Using this reference, the **EpiDISH algorithm** estimated epithelial, adipose, and immune-cell fractions for each sample. This correction allowed methylation changes to be attributed to true epithelial alterations rather than cell composition shifts.

Identification of differentially variable CpGs (DVMCs) To detect epigenetic field defects, the authors applied their **iEVORA algorithm**, designed to identify CpGs with differential variance rather than mean-level differences. Between normal-healthy and normal-adjacent tissues:

- CpGs with significant differential variance (FDR < 0.001, Bartlett's test) and mean difference ($p < 0.05$, t-test) were selected as **differentially variable and methylated CpGs (DVMCs)**.
- The majority showed increased variance ("hyperV DVMCs") in normal-adjacent samples, suggesting stochastic epigenetic deregulation in tissue at risk.

HyperV DVMCs were confirmed to be independent of cell-type fraction changes and to localize preferentially within epithelial genomic regions, indicating that they reflect true epigenetic instability rather than compositional artifacts.

CNV calling and comparative analysis CNV profiles were inferred using `conumee`, followed by segmentation via circular binary segmentation (CBS). Copy-number states (gain/loss) were determined with adaptive, sample-specific thresholds accounting for stromal contamination. Differential CN analysis between normal and normal-adjacent tissue revealed no genome-wide significant differences; only marginal gains were detected. While 2,845 genes exhibited CN changes exclusively in normal-adjacent samples, these alterations were weaker and less consistent than methylation changes. Both CN and DNAm alterations were enriched in the matched cancers, but only DNAm patterns were strong enough to separate tissue classes.

Risk prediction and model validation Cancer risk predictors were trained independently on DNAm and CNV features using five-fold cross-validation and an adaptive-index algorithm:

- **DNAm-based classifier**: achieved AUC = 0.94 (95% CI: 0.88–1.0) in the discovery set and AUC = 0.84 (0.74–0.94) in the validation cohort (GSE67919).
- **CNV-based classifier**: failed to discriminate normal vs. normal-adjacent tissue (AUC = 0.60 and 0.50, respectively).

Alternative CNV-calling pipelines (`cnAnalysis450k`, bin-level analysis, probe-level Elastic Net classifiers) confirmed the poor predictive performance of CNVs, indicating that the observed differences were not due to technical segmentation bias but reflect a genuine biological contrast.

Interpretation and implications DNAm alterations in normal-adjacent tissue mirror early "field defects" that are propagated and intensified in corresponding cancers. In contrast, CNVs—though enriched in tumors—show no significant discriminative power at the pre-cancer stage. The authors conclude that **epigenetic variability in normal cells better captures early carcinogenic processes** and thus provides a more sensitive predictor of breast cancer risk. The identified hyperV DVMCs often overlap with Polycomb (PRC2) target regions and developmental transcription factor binding sites, suggesting that aberrant methylation at these loci represents an early, reversible step toward neoplastic transformation.

Conclusion This work provides quantitative evidence that DNA methylation changes in histologically normal tissue are more predictive of cancer risk than copy-number variations. The methodological framework—EpiDISH correction, iEVORA feature selection, and risk-score modeling—demonstrates a reproducible way to extract early epigenetic markers from array data (e.g., GSE67919, GSE69914). The findings support a model where epigenetic instability precedes and possibly drives genetic alterations during tumor initiation.

References

[1] Y.-a. Chen et al., "Discovery of cross-reactive probes and polymorphic cpgs in the illumina infinium humanmethylation450 microarray," *Epigenetics*, vol. 8, no. 2, pp. 203–209, 2013. DOI: 10.4161/epi.23470 [Online]. Available: https://doi.org/10.4161/epi.23470

[2] A. E. Teschendorff et al., "Dna methylation outliers in normal breast tissue identify field defects that are enriched in cancer," *Nature Communications*, vol. 7, p. 10 478, 2016. DOI: 10.1038/ncomms10478 [Online]. Available: https://doi.org/10.1038/ncomms10478

[3] Y. Gao et al., "The integrative epigenomic-transcriptomic landscape of er positive breast cancer," *Clinical Epigenetics*, vol. 7, p. 126, 2015. DOI: 10.1186/s13148-015-0159-0 [Online]. Available: https://doi.org/10.1186/s13148-015-0159-0

[4] W. Ding, G. Chen, and T. Shi, "Integrative analysis identifies potential dna methylation biomarkers for pan-cancer diagnosis and prognosis," *Epigenetics*, vol. 14, no. 1, pp. 67–80, 2019. DOI: 10.1080/15592294.2019.1568178

[5] Y. Gao, M. Widschwendter, and A. E. Teschendorff, "Dna methylation patterns in normal tissue correlate more strongly with breast cancer status than copy-number variants," *EBioMedicine*, vol. 31, pp. 243–252, 2018. DOI: 10.1016/j.ebiom.2018.04.025 [Online]. Available: https://doi.org/10.1016/j.ebiom.2018.04.025