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 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 [2].

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.

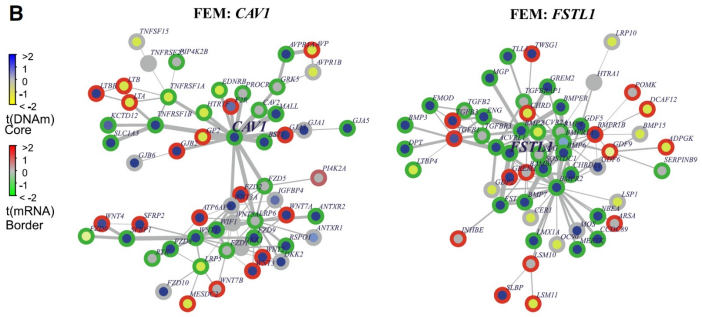

Figure 1: Examples of two FEMs centred around seed genes CAV1 and FSTL1 in ER+ breast cancer.

3 An integrative pan-cancer-wide analysis of epigenetic enzymes reveals universal patterns of epigenomic deregulation in cancer

Keywords DNA methylation, epigenetic enzymes, Illumina 450K, TCGA, HyperZ/HypoZ instability indices, UHRF1, WHSC1, CBX7, GSE69914, breast cancer subtypes, luminal B, promoter hypermethylation, open-sea hypomethylation, pan-cancer analysis [3].

Study aim and rationale The study performs a pan-cancer integrative analysis of DNA methylation and gene expression across ten tumor types from The Cancer Genome Atlas (TCGA), in order to identify epigenetic enzymes (EEs) that act as consistent regulators of cancer-associated methylation changes. Epigenetic enzymes are defined broadly as writers, erasers, readers and editors of chromatin and DNA methylation states (212 genes total). The goal is to (i) determine which EE genes are universally deregulated across cancers, (ii) quantify how their expression relates to genome-wide DNA hypermethylation and hypomethylation, and (iii) prioritize candidate master regulators of these methylation changes. A particular focus is on whether these processes are shared across different tissues and tumor types, and whether methylation instability detected in TCGA tumors is also observable in external breast tissue methylation data such as GSE69914.

Data sources and preprocessing Matched RNA-Seq (gene-level, $\log_2$-transformed RSEM values) and DNA methylation data (Illumina HumanMethylation450K β-values) were collected from TCGA for ten cancer types, including breast, bladder, colon, head and neck, kidney, liver, lung adenocarcinoma, lung squamous, thyroid, and endometrial carcinoma. Quality control relied on singular value decomposition to confirm that the dominant source of variation was biological (tumor vs. normal) rather than technical batch structure. For methylation arrays, probes with excessive missingness were removed, missing values were imputed using k-nearest neighbors, and type II probe bias was corrected using BMIQ. DNA methylation data from an independent breast tissue cohort (GSE69914; 30 healthy normals, 21 normal-adjacent to tumor, and 165 tumors profiled on Illumina 450K) were processed analogously (minfi, impute, BMIQ) and used to test whether "normal-adjacent" samples already carry field defects comparable to tumor, or whether they can still function as a quasi-normal reference for large-scale deviations.

Genome-wide methylation instability indices (HyperZ and HypoZ) To quantify global epigenetic disruption per tumor, two *instability indices* were defined for each cancer sample by comparing that sample to reference normal tissue from the same organ:

- **HyperZ**: measures aberrant promoter CpG-island *hypermethylation*. For each promoter CGI cluster, a Z-score is computed relative to normals. Only positive deviations (hypermethylation relative to normal) contribute, and the mean of these positive deviations defines HyperZ for that tumor.
- **HypoZ**: measures aberrant *hypomethylation* in low-CpG-density "open sea" regions (intergenic). For each open-sea cluster, the absolute negative deviation from normal methylation contributes to the index. The mean of these hypomethylation deviations defines HypoZ.

These indices summarize how far a given tumor departs from the normal epigenome along two distinct axes: focal promoter CGI hypermethylation vs. large-scale open-sea hypomethylation. When HyperZ and HypoZ were plotted against each other for tumors within each cancer type, they were only weakly correlated (per-cancer $R^2 \approx 0.1$ or less), indicating that promoter CGI hypermethylation and intergenic hypomethylation are at least partially independent processes. In breast cancer specifically, luminal B tumors showed the highest HyperZ and HypoZ scores, consistent with stronger epigenetic deregulation compared to luminal A, and HyperZ was also elevated in $HER2^+$ tumors. This supports that the instability indices capture clinically relevant subtype differences.

Universal differential expression of epigenetic enzyme genes Expression of the 212 curated EE genes was compared between tumor and matched/adjacent normal tissue for each TCGA cancer type using moderated t-tests. A gene was considered "consistently deregulated" if it showed the same direction of change (up or down) in at least 8 of the 10 cancers at nominal $p < 0.05$. This yielded 62 EE genes with pan-cancer deregulation: 35 recurrently overexpressed (putative epigenetic oncogenes) and 27 recurrently underexpressed (putative epigenetic tumor suppressors). The probability of observing this many consistently deregulated genes by chance, given 212 candidates, was estimated to be effectively zero (binomial model $P \ll 10^{-30}$). The overexpressed group included *EZH2* (a PRC2 histone methyltransferase), *DNMT1*, *DNMT3A*, *DNMT3B* (DNA methyltransferases), multiple histone deacetylases (e.g. *HDAC1*), *KDM1A* (histone demethylase), and *UHRF1* (a coordinator of maintenance methylation via DNMT1 recruitment). The underexpressed group included chromatin regulators such as *CBX7* (polycomb- and H3K36-associated reader linked to tumor suppression), and histone acetyltransferases (*KAT2B*, *KAT5*) and PRDM-family methyltransferases (*PRDM2*, *PRDM5*). This step narrows down which epigenetic modifiers are repeatedly altered at the transcriptional level across tissues.

Linking EE expression to global hypermethylation and hypomethylation Next, the expression level of each EE gene was correlated (Pearson correlation, converted to Fisher Z) with HyperZ and HypoZ across tumor samples, cancer type by cancer type. Genes with significant ($p < 0.05$) and directionally consistent correlations in at least 6 of the 10 cancers were retained. This analysis showed:

- Certain EE genes correlate positively with HyperZ (higher expression $\rightarrow$ stronger promoter CGI hypermethylation relative to normal) across many cancer types.
- Other EE genes correlate positively with HypoZ (higher expression $\rightarrow$ stronger open-sea hypomethylation).
- Generally, EE genes associated with HyperZ are *not* the same as those associated with HypoZ, reinforcing that promoter CGI hypermethylation and intergenic hypomethylation are controlled by partially distinct regulatory programs.

An important exception is that a few genes (*EZH2*, *PCNA*, etc.) show consistent association with both HyperZ and HypoZ, suggesting that proliferative/repair-linked programs can influence both focal hypermethylation and large-scale hypomethylation simultaneously.

Candidate master regulators of the cancer methylome By intersecting (i) pan-cancer differential expression status and (ii) pan-cancer correlation with HyperZ/HypoZ, the study prioritizes 18 EE genes as putative global methylome regulators. These split into:

- 11 overexpressed "epigenetic oncogenes" whose elevated expression is associated with higher HyperZ and/or HypoZ, including *UHRF1*, *WHSC1* (H3K36 methyltransferase), *EZH2*, *KDM1A*, *SUV39H2*, *HDAC1*, *PCNA*, *TTF2*, *RAD54L*, *TDG*, *TET3*.
- 7 underexpressed "epigenetic tumor suppressors" whose reduced expression is associated with higher HyperZ and/or HypoZ, including *CBX7*, *PRDM2*, *PRDM5*, *SETBP1*, *EYA4*, *DUSP1*, *KAT2B*.

A causal network modeling step (partial correlation / multivariate regression) was then used to distinguish direct from indirect associations: for each EE gene, the correlation with HyperZ/HypoZ was re-estimated while controlling for (a) promoter methylation of that EE gene itself and (b) the expression of the other EE genes. After this adjustment, three genes emerged as the most plausible *drivers* rather than passengers:

- **UHRF1**: overexpression associates with increased promoter CGI hypermethylation (higher HyperZ), suggesting a role in establishing or maintaining aberrant promoter hypermethylation in tumors. UHRF1 is known to recruit DNMT1 to hemimethylated DNA during replication.
- **WHSC1** (also known as *NSD2*): overexpression associates with increased promoter hypermethylation. WHSC1 writes H3K36me2/3, and altered H3K36 methylation has been linked to aberrant targeting of DNA methylation gains at CpG islands.
- **CBX7**: loss of expression associates with widespread intergenic hypomethylation (higher HypoZ), implying that CBX7 normally helps maintain DNA methylation in open-sea regions; its downregulation may destabilize large-scale methylation domains.

These predictions imply that tumors across many tissues converge on the same few epigenetic nodes (UHRF1, WHSC1, CBX7) to reshape their methylomes.

Shared genomic targets across cancer types For each of the three predicted master regulators (UHRF1, WHSC1, CBX7), genomic regions (clusters of CpGs) were ranked by how strongly their methylation correlated with that regulator's expression in one cancer type, then checked in other cancer types. The same promoter CpG islands and open-sea blocks tended to recur at the top of these rankings across distinct cancers. This indicates that the loci most sensitive to UHRF1- or WHSC1-associated hypermethylation and CBX7-associated hypomethylation are not tumor-type specific, but instead define a conserved "epigenetic response program" reused in multiple tissues.

Relevance to GSE69914 and field effects in normal-adjacent tissue Because TCGA "normal" tissues are often sampled adjacent to tumor, there is a risk that early field defects could bias the normal reference. To address this, Illumina 450K data from an external breast cohort (GSE69914) were analyzed: 30 normal tissues from healthy women, 21 histologically normal tissues adjacent to breast tumors, and 165 breast cancers. After identical preprocessing (minfi, BMIQ), methylation patterns in normal-adjacent breast tissue were compared to healthy normal tissue. Large-scale field defects were found to be rare, and the HyperZ/HypoZ instability indices were highly similar regardless of whether the reference normals came from healthy breast or adjacent-normal breast. This supports that the HyperZ/HypoZ framework is robust, and that the genome-wide deviations captured by these indices (promoter CGI hypermethylation and open-sea hypomethylation) reflect true tumor-associated remodeling rather than trivial contamination of the "normal" baseline.

Implications This work indicates that:

- Global promoter CGI hypermethylation and intergenic hypomethylation are partially decoupled processes within tumors and likely driven by distinct epigenetic programs.
- A relatively small set of epigenetic regulators (notably *UHRF1*, *WHSC1*, and *CBX7*) appears to control pan-cancer methylation instability, affecting the same genomic loci in multiple tissues.
- Breast cancer subtypes with worse prognosis (e.g. luminal B) show stronger global methylation instability (higher HyperZ/HypoZ), linking these methylome deviations to clinical aggressiveness.
- External breast methylation data (GSE69914) confirm that these instability patterns are reproducible and not an artifact of using normal-adjacent tissue as baseline.

Overall, the study proposes a generalizable systems-level framework: quantify genome-wide methylation deviation (HyperZ/HypoZ), link it to expression of epigenetic enzyme genes, and nominate direct master regulators of tumor methylome remodeling. This framework directly involves datasets such as GSE69914 and supports the idea that a limited set of conserved epigenetic mechanisms underlies methylome deregulation across cancers.

4 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 [4].

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** (∼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 ∼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 ∼32% of these hypervariable/hypermethylated DVMCs gained *further* methylation in the tumor, while only ∼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).

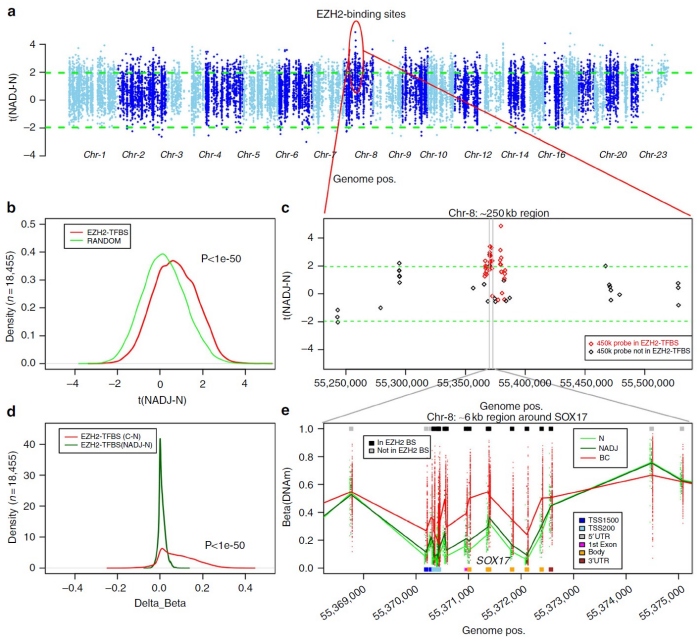

Figure 2: EZH2-associated DNA methylation patterns CpG probes located within EZH2-binding sites show consistent hypermethylation in normal-adjacent breast tissue compared with normal tissue, indicating early Polycomb-associated epigenetic changes. The density and regional plots reveal a gradual methylation 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.

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.

6 Large-scale analysis of DFNA5 methylation reveals its potential as biomarker for breast cancer

Keywords DFNA5, breast cancer, DNA methylation, Illumina HumanMethylation450K, CpG biomarkers, logistic regression, AUC, survival, TCGA, ER status, ductal vs. lobular, prognostic markers [6].

Study design and datasets This study used The Cancer Genome Atlas (TCGA) to perform a large-scale, locus-specific analysis of DNA methylation in the *DFNA5* (also known as *GSDME*) gene in breast cancer. Only female, untreated, ductal or lobular breast adenocarcinoma samples were included. Methylation data were available for 668 primary breast adenocarcinomas and 85 histologically normal breast tissues sampled at a distance from the tumor; 79 of these patients had matched tumor/normal pairs. Expression data were available for 476 tumors and 56 normals (Agilent microarray) and 666 tumors and 71 normals (RNA-seq). Clinical parameters were collected for each patient: estrogen receptor (ER) status, progesterone receptor (PR) status, HER2 status, histological type (ductal vs. lobular), pathological tumor stage (I–IV), age at diagnosis, and overall survival (OS) up to 5 years after diagnosis. Three independent public datasets (GEO: GSE52865, GSE69914, GSE60185) were later used to externally validate classifier performance.

Methylation and expression profiling Genome-wide DNA methylation profiles were obtained from TCGA level-3 Illumina HumanMethylation450K BeadChip data. For *DFNA5*, 22 distinct CpG loci on chromosome 7 were available. Methylation at each CpG was represented as a β value, defined as the ratio of methylated probe intensity to the total probe intensity (methylated + unmethylated). Gene expression of *DFNA5* was quantified using Agilent microarray and RNA-seq data, normalized as $\log_2$ fold-changes or abundance. Only 5 of 570 sequenced breast adenocarcinomas carried *DFNA5* variants (3 missense, 2 silent), indicating that *DFNA5* is rarely mutated; therefore, analysis focused on epigenetic regulation.

Statistical framework All statistical analyses were performed in `R`. Linear mixed models with batch as random effect and age as covariate were used for methylation/expression associations. For paired tumor/normal comparisons, paired t-tests were used. Stepwise logistic regression with tenfold cross-validation identified CpGs that best discriminated tumor from normal tissues, optimizing the area under the ROC curve (AUC). Cox proportional hazards models assessed whether CpGs contributed to 5-year OS prediction beyond age and tumor stage.

DFNA5 methylation landscape The 22 CpGs were distributed as follows:

- **Gene body (6 CpGs):** CpG17790129, CpG14205998, CpG04317854, CpG12922093, CpG17569154, CpG19260663.
- **Promoter region (14 CpGs):** CpG09333471, CpG00473134, CpG03995857, CpG07320646, CpG07293520, CpG04770504, CpG24805239, CpG01733570, CpG25723149, CpG22804000, CpG07504598, CpG15037663, CpG19706795, CpG20764575.
- **Upstream region (2 CpGs):** CpG06301139, CpG26712096.

Breast cancers displayed clear promoter hypermethylation and gene-body hypomethylation. In the promoter, tumor β-values ranged 0.6–0.75 vs. 0.3–0.4 in normal tissue, while gene-body CpGs showed the opposite pattern (hypomethylation in tumor). These results indicate focal hypermethylation of regulatory regions concurrent with hypomethylation of coding regions.

Relationship between methylation and expression Tumors exhibited both promoter hypermethylation and lower *DFNA5* expression compared with normal tissue, but direct CpG–expression correlations were modest ($|\rho| < 0.35$). Multivariate models explained about 20% of expression variance in tumors, suggesting methylation partially controls gene repression. This partial correlation may reflect tumor heterogeneity and the complex regulation of *DFNA5* transcription.

Diagnostic classifier Stepwise logistic regression identified two CpGs that maximized tumor/normal discrimination:

- **CpG12922093** (gene body, hypomethylated in tumor),
- **CpG07504598** (promoter, hypermethylated in tumor).

The logistic model:

$$\Pr(\text{tumor}) = \frac{e^{7.49 - 10.77\beta(\text{CpG12922093}) + 6.33\beta(\text{CpG07504598})}}{1 + e^{7.49 - 10.77\beta(\text{CpG12922093}) + 6.33\beta(\text{CpG07504598})}}$$

yielded tenfold cross-validated AUC = 0.93 (95% CI: 0.92–0.95). Using a cutoff of 0.87 achieved 85.3% sensitivity and 100% specificity (accuracy 87%). The model replicated well in GSE52865, GSE69914, and GSE60185 datasets. In contrast, *DFNA5* expression alone provided lower AUCs (0.82–0.88), confirming methylation as a superior diagnostic marker.

Clinicopathological correlations

- **ER status:** Promoter CpGs were more methylated in ER^+ tumors, gene-body CpGs less methylated. *DFNA5* expression was lower in ER^+ cancers.
- **PR status:** 15 CpGs correlated with PR status, but expression differences were not significant.
- **HER2 status:** Only CpG04317854 associated with HER2; expression unaffected.
- **Tumor stage:** Five CpGs showed significant stage correlation.
- **Histology:** Lobular carcinomas had higher promoter methylation (10 CpGs) and higher *DFNA5* expression than ductal carcinomas, suggesting subtype-specific regulation.

Prognostic relevance Cox models revealed that five **gene-body CpGs**—CpG17790129, CpG14205998, CpG12922093, CpG17569154, and CpG19260663—were significantly associated with 5-year OS ($p < 0.05$). Higher methylation at these loci predicted worse prognosis independently of stage and age. Promoter CpGs, while diagnostic, were not prognostic. Hence, promoter methylation primarily distinguishes tumors, while gene-body methylation predicts aggressiveness.

Replication protocol To reproduce these findings:

1. Extract β-values for the 22 CpGs from 450K data (TCGA or GEO).
2. Perform paired t-tests for tumor vs. normal and visualize mean β by genomic position.
3. Fit logistic regression with all 22 CpGs; select the 2-CpG model above for classification.
4. Validate AUC through cross-validation or on external datasets (expected AUC ≈ 0.93).
5. Associate CpG methylation with ER, PR, HER2, histology using linear mixed models.
6. Fit Cox models for OS; expect significant gene-body CpG effects on prognosis.

Conclusion Croes et al. demonstrated that *DFNA5* promoter hypermethylation and gene-body hypomethylation constitute a reproducible epigenetic signature in breast cancer. The 22 CpGs identified delineate functional regions with diagnostic and prognostic value. The 2-CpG logistic model (CpG12922093, CpG07504598) offers a compact, high-specificity classifier (AUC = 0.93). Gene-body methylation correlates with poor prognosis, suggesting that *DFNA5* methylation captures both tumor presence and aggressiveness. These results position *DFNA5* as a promising epigenetic biomarker for breast cancer detection and risk stratification.

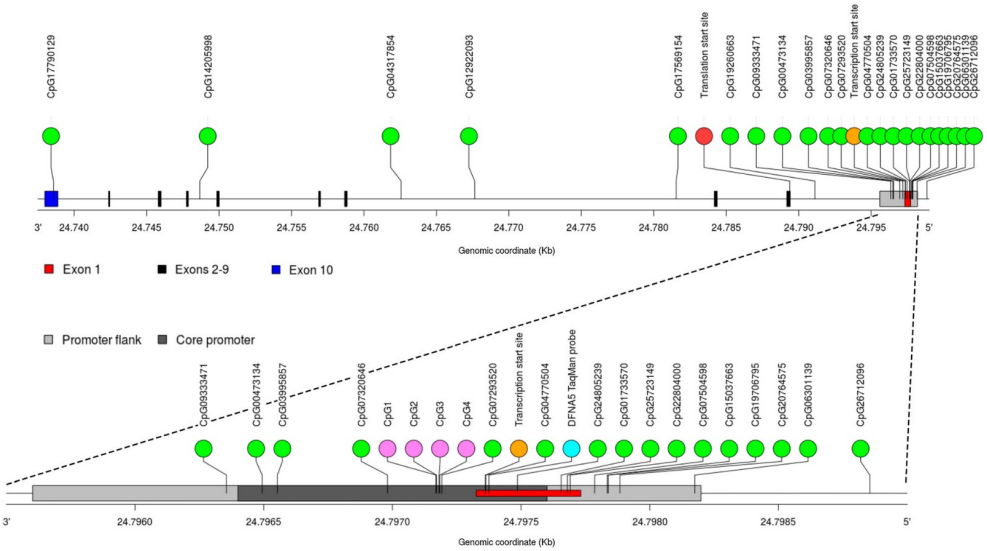

Figure 3: The DFNA5 gene with annotation of the 22 CpGs. The 10 exons and the promoter and gene body region of the DFNA5 gene are indicated.

7 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 [7].

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.

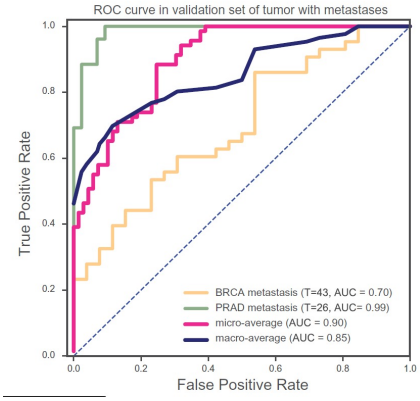

Figure 4: Validation of tumor specific classifier in tumors with metastases. ROC curve of multiclass tumor specific classifier in metastatic breast cancer (GSE58999) and metastatic prostate cancer (GSE73549 and GSE38240).

8 EPISCORE: cell type deconvolution of bulk tissue DNA methylomes from single-cell RNA-Seq data

Keywords DNA methylation, deconvolution, cell-type composition, breast tissue, Illumina 450K, single-cell RNA-Seq, EPISCORE, TCGA, GSE69914, cell-type-specific differential methylation [8].

Aim The work introduces **EPISCORE**, a computational pipeline that estimates cell-type proportions and detects cell-type-specific DNA methylation changes in bulk tissue samples. The core motivation is that most DNA methylation datasets (including breast cancer datasets on the Illumina HumanMethylation450K array such as GSE69914) are generated from heterogeneous tissues, where epithelial, stromal, endothelial, adipose, and immune cells are mixed. Without correcting for this heterogeneity, it is not possible to tell whether a methylation difference reflects a true epithelial (or stromal, etc.) change or just a shift in cellular composition. EPISCORE overcomes the lack of genome-wide single-cell methylation atlases by *inferring* cell-type-specific DNA methylation reference profiles from single-cell *RNA*-Seq atlases.

Core idea Instead of requiring purified cell-type DNA methylomes, EPISCORE learns which genes have promoter methylation that is predictably (usually anti-) correlated with their expression across many purified cell/tissue types. Those genes are then used as "anchors" to *translate* single-cell RNA-Seq cell-type signatures into approximate, cell-type-specific methylation signatures. These inferred methylation signatures act as a reference to decompose any bulk 450K (or WGBS) sample into proportions of its constituent cell types and to localize differential methylation to specific cell types.

EPISCORE workflow EPISCORE proceeds in four main steps:

1. **Build a tissue-specific single-cell expression reference.** Single-cell RNA-Seq atlases (e.g. mammary epithelium, stroma, endothelium, immune infiltrate) are clustered into major cell types (for breast: basal epithelial, luminal epithelial, stromal/fibroblast, endothelial, lymphocytes, macrophages). For each cell type, highly specific marker genes are selected (high expression in that type, near-zero in others). The result is a cell-type-by-gene matrix of median expression.

2. **Identify "imputable" genes with expression–methylation coupling.** Using *independent* matched DNA methylation and bulk expression data from many purified primary cell types (Epigenomics Roadmap, Stem Cell Matrix Compendium), genes are scanned genome-wide. A gene is kept if its promoter (or enhancer) methylation and its expression are strongly anti-correlated across purified cell/tissue samples, after requiring sufficient dynamic range. This yields a core set (∼2–3k) of genes for which promoter methylation can be predicted from expression.

3. **Impute a DNA methylation reference for the tissue.** For each marker gene that is both (i) cell-type-specific in single-cell RNA-Seq and (ii) "imputable", a probabilistic model (logistic / Bayesian regression) is used to convert its median expression (per cell type) into an estimated promoter methylation level for that same cell type. Aggregating over all such genes produces a **synthetic, tissue-specific DNA methylation reference matrix**: rows = marker loci (CpGs near TSS/enhancers of those genes), columns = cell types (e.g. luminal epithelium, basal epithelium, fibroblast, endothelial, immune). Each locus also receives a weight reflecting confidence in the imputation.

4. **Deconvolution and cell-type-specific differential methylation.** Given a bulk Illumina 450K profile (for instance, normal breast, normal-adjacent breast, tumor breast from GSE69914), EPISCORE solves a robust weighted regression to estimate the fraction of each cell type in that sample using the imputed methylation reference. These estimated fractions are then passed to CellDMC, which fits interaction models to call **differentially methylated cytosines by cell type** (DMCTs): i.e. CpGs whose methylation shifts in the epithelial compartment vs. the stromal compartment, etc., between biological groups (e.g. ER^+ vs. triple-negative tumors; tumor vs. normal-adjacent tissue).

Validation in lung and breast tissue EPISCORE is validated on (i) lung squamous cell carcinoma and lung adenocarcinoma from TCGA, and (ii) breast tissue data including the Erlangen cohort/GSE69914 (50 normal breast, 42 normal-adjacent, 305 cancers profiled on the 450K array). Key observations:

- In lung tumors, EPISCORE detects strong epithelial fractions (as expected) and reduced endothelial fractions. CellDMC then localizes thousands of cancer-associated differentially methylated CpGs to the epithelial compartment and, importantly, also to endothelial cells. Endothelial-specific DMCTs are enriched for TGF-β/SMAD2/3 signaling and endothelial-to-mesenchymal transition (EndoMT) programs, suggesting vascular remodeling that may facilitate invasion and metastasis.

- In breast cancer, EPISCORE-derived deconvolution distinguishes basal-like (triple-negative) vs. luminal ER^+ tumors at the *epithelial* level, not just at bulk level. Most CpGs that differ between ER^+ and triple-negative tumors are assigned specifically to the epithelial compartment, and these same CpGs reproduce differences seen directly in luminal vs. basal breast cancer cell lines.

- When applied to breast Illumina 450K data (including normal vs. normal-adjacent vs. tumor profiles from GSE69914), the method confirms that methylation deviations in normal-adjacent tissue are already detectable in the epithelium, consistent with "field defects" near tumors. This directly supports the interpretation that GSE69914 captures early, spatially localized epithelial epigenetic damage rather than mere shifts in stromal content.

Relevance for the GSE69914 analysis For GSE69914-like data (paired normal breast, normal-adjacent breast, and tumor breast on the 450K array), EPISCORE provides:

- An estimated cellular composition per sample (luminal epithelium, basal epithelium, stroma/fibroblast, endothelium, immune).

- A way to test whether methylation differences between normal and normal-adjacent tissue are epithelial-intrinsic (true early field defects) or driven by microenvironmental admixture.

- A way to map which CpGs are specifically altered in epithelial cells of aggressive subtypes (e.g. triple-negative) versus luminal ER^+ cases, reproducing subtype biology.

Conclusion EPISCORE shows that cell-type-resolved methylation analysis can be recovered *without* having physical purified methylomes from each breast cell type. This is crucial for heterogeneous cohorts such as GSE69914, where early epigenetic lesions in morphologically "normal-adjacent" epithelium are subtle and would otherwise be obscured by stromal/immune admixture. The pipeline therefore justifies interpreting GSE69914 not only at bulk level but explicitly at the epithelial compartment level, using inferred per-sample cell-type fractions and epithelial-specific differentially methylated CpGs.

9 Prediction of Disease Genes Based on Stage-Specific Gene Regulatory Networks in Breast Cancer

Keywords Breast cancer, DNA methylation, gene expression, stage-specific analysis, gene regulatory networks, WGCNA, hub genes, TCGA, GSE69914, GSE15852, candidate driver genes [9].

Study aim and rationale This study proposes a computational framework to predict *stage-specific* candidate disease genes in breast cancer by integrating transcriptomic and DNA methylation data with clinical staging information. Unlike many previous approaches that ignore stage and analyze tumors as a single group, this method builds three distinct regulatory models for Stage I, Stage II, and Stage III breast cancer. The central idea is that genes driving cancer initiation may differ from those driving progression, and therefore network structure and key regulators should be inferred separately at each tumor stage. The pipeline identifies stage-specific gene modules, characterizes their biological functions, and prioritizes their core genes as putative disease genes for that stage. The approach is then validated on external datasets (GSE15852 and GSE69914) to assess reproducibility.

Datasets and preprocessing Gene expression (RNA-Seq) and DNA methylation data were downloaded from The Cancer Genome Atlas (TCGA), together with phenotype and staging data. The gene expression matrix contained 60,484 genes across 1,217 samples; the DNA methylation matrix contained 485,578 CpG sites across 890 samples. Only paired samples (each tumor sample matched to its own adjacent normal tissue) were retained. Samples were then stratified by clinical stage, yielding: 29 tumor/normal pairs for Stage I, 94 pairs for Stage II, and 32 pairs for Stage III. Stage IV had only two usable pairs and was excluded as not statistically convincing.

For methylation, multiple CpG probes can map to the same gene. The authors collapsed CpG-level data to gene-level methylation by averaging the β values of all CpGs annotated to that gene. The β-value represents the methylation level; values > 0.8 defined hypermethylated genes and values < 0.2 defined hypomethylated genes.

For gene expression, normalized FPKM values were used. Genes with missing values in more than 15% of samples were filtered out. Only those samples with both gene expression and matched methylation data (tumor and normal from the same patient) were kept for downstream paired analyses.

Differential expression and methylation filtering Within each stage separately (Stage I, Stage II, Stage III), tumor vs. matched normal samples were compared to identify both (i) differentially expressed genes and (ii) aberrantly methylated genes.

Differential expression analysis was performed using the `limma` R package, selecting genes with $p < 0.05$ and $|\log \mathrm{FC}| < 0.5$ as reported thresholds. In parallel, DNA methylation status was used to label genes as hypermethylated ($\beta > 0.8$) or hypomethylated ($\beta < 0.2$). The intersection of these two criteria — genes that were both differentially expressed and abnormally methylated — produced the gene sets carried forward for network modeling. This yielded 1,027 such genes in Stage I, 1,012 in Stage II, and 1,220 in Stage III.

Across all stages, the authors confirmed the expected inverse relationship between DNA methylation and gene expression: higher methylation corresponded to lower expression, consistent with promoter silencing effects.

Stage-specific gene regulatory network construction To model regulatory control at each stage, the study built three transcriptional regulatory networks (one per stage). Transcription factor (TF) $\rightarrow$ target gene pairs were obtained from GRNdb, a curated TF–target interaction resource. These TF–target pairs were filtered so that the target gene was among the stage-specific differentially expressed and hyper/hypomethylated genes defined above.

For each retained TF–target pair, the Pearson correlation coefficient (PCC) between TF expression and target expression was computed across the paired tumor/normal samples of that stage. Edges with $|\text{PCC}| \geq 0.5$ were kept. This procedure produced three **stage-specific gene regulatory networks:**

- Stage I network: 1,129 nodes and 4,429 edges.
- Stage II network: 1,066 nodes and 4,879 edges.
- Stage III network: 1,339 nodes and 6,461 edges.

These per-stage networks reflect transcriptional wiring specific to that disease stage, integrating both expression deregulation and methylation dysregulation relative to normal tissue.

Module detection via WGCNA Within each stage-specific regulatory network, modules (i.e., co-regulated and co-expressed gene communities) were identified using Weighted Gene Co-expression Network Analysis (WGCNA). The steps were: (i) hierarchical clustering of genes in each network; (ii) Dynamic Tree Cut to segment the dendrogram into discrete modules, enforcing a minimum module size of 30 genes.

Results:

- Stage I network was partitioned into 11 modules (e.g., S1_turquoise, S1_brown, S1_blue), with the largest (S1_turquoise) containing 270 genes.
- Stage II network was partitioned into 10 modules, with the largest (S2_turquoise) containing 337 genes.
- Stage III network was partitioned into 13 modules, with the largest (S3_turquoise) again containing 337 genes.

Genes that were differentially expressed *only* in one stage (not in the others) were labeled as "stage-specific genes": 92 genes for Stage I, 60 for Stage II, and 187 for Stage III. The authors then mapped these uniquely stage-specific genes back to the WGCNA modules to identify which modules were most enriched for stage-unique signals. They found:

- Stage I-specific genes mainly cluster in S1_brown, S1_turquoise, and S1_blue.
- Stage II-specific genes mainly cluster in S2_turquoise.
- Stage III-specific genes mainly cluster in S3_turquoise, S3_brown, and S3_green.

These seven modules (S1_brown, S1_turquoise, S1_blue, S2_turquoise, S3_turquoise, S3_brown, S3_green) were therefore defined as the **stage-specific core modules**.

Topological analysis of the core modules For each of the seven core modules, the study characterized network topology using Cytoscape. Metrics included node degree, betweenness centrality, and closeness centrality.

- Degree distributions in S1_turquoise, S2_turquoise, and S3_turquoise were broad and extended to high connectivity: most node degrees lay between ∼100 and 400, implying dense regulatory connectivity in these "turquoise" modules.
- In S1_brown, S1_blue, S3_brown, and S3_green, degree values were generally lower (mostly 50–100) but still showed scale-free-like (power-law) behavior.
- Betweenness centrality was high for many nodes across all seven modules, and closeness centrality values for most nodes ranged from 0.5 to 0.9.

These metrics indicate that each stage-specific module is internally well connected, suggesting that its high-centrality nodes function as regulatory hubs. Such hubs are natural candidates for stage-relevant driver genes.

Functional enrichment of stage-specific modules Gene set enrichment analysis for each of the seven core modules was performed with Metascape, using a significance cutoff of $p < 0.01$. The most significantly enriched biological themes were:

- **Cell cycle control and mitosis**: "cell cycle phase transition," "chromosome segregation," "DNA replication," "spindle organization," and "cell division." These functions dominated the large turquoise modules of Stage I, Stage II, and Stage III (S1_turquoise, S2_turquoise, S3_turquoise).
- **Transcriptional regulation and chromatin state**: modules such as S1_brown, S1_blue, S3_brown, and S3_green were enriched for transcriptional activator/repressor activity, chromatin binding, histone modification, nuclear receptor activity, telomerase complex, and macromolecule methylation.

Importantly, regulatory complex assembly at promoters and chromatin binding functions were found across *all* seven modules, suggesting that stage-specific dysregulation in breast cancer converges on control of transcription, chromatin structure, and cell cycle checkpoints.

Candidate disease gene prioritization Within each of the seven core modules, the authors defined "core genes" using two complementary strategies and then took their intersection:

1. **Correlation structure within the module**: for each module, they computed a gene–gene correlation matrix and retained genes with correlation ≥ 0.8 and $p < 0.05$, marking them as internally coherent regulators.
2. **Network centrality**: they ranked genes by degree, betweenness centrality, and closeness centrality within that module and selected the top 5% as hub-like nodes.

Genes that satisfied both criteria were labeled **candidate disease genes** for that stage.
Results by stage:

- **Stage I** (modules S1_brown, S1_turquoise, S1_blue): 20 candidate genes, including *E2F2*, *E2F8*, *TPX2*, *BUB1*, *CKAP2L*, *CBX3*, *KPNA2*, *NEK2*, *TTK*, *LMNB1*, *SLC25A36*, *SLC39A1*, *MRPS12* (PCNA-related), and additional transcriptional/transport regulators like *CASC5*, *CREBRF*, *PAN2*, *BTAF1*, *ZC3H6*, *DDX49*.
- **Stage II** (module S2_turquoise): 12 candidate genes, including *E2F2*, *E2F8*, *TPX2*, *KPNA2*, *CKAP2L*, *CBX3*, *BUB1*, *CCNE2*, *CASC5*, *SPDL1*, *TOP2A*, and *DDIAS*.
- **Stage III** (modules S3_turquoise, S3_brown, S3_green): 22 candidate genes, including *E2F2*, *RAD21*, *FBXO5*, *CCNE2*, *CBX3*, *STIL*, *CKAP2L*, *PCNA*, *NEK2*, *TTK*, *CSE1L*, *H2AFZ*, *NR2F6*, *TRAPPC6A*, *IGSF8*, *FDXR*, *SLC39A1*, *EXOSC5*, *RBBP5*, *KDM5B*, *H3F3A*, and *CDC42SE1*.

Three genes — *E2F2*, *CKAP2L*, and *CBX3* — were shared across all three stages, suggesting persistent dysregulation from early to late disease.

Biological/clinical interpretation and validation The candidate genes were then checked against known cancer gene resources (OMIM, COSMIC, DAVID) and literature in PubMed to test whether they are already associated with breast cancer biology.

Many top hits have clear functional relevance:

- *E2F2*, *E2F8*: transcription factors controlling cell cycle entry and proliferation; linked to poor prognosis and recurrence-free survival in breast cancer.
- *TPX2*, *BUB1*, *NEK2*, *TTK*, *TOP2A*, *PCNA*: mitotic spindle assembly, checkpoint control, DNA topology, and replication; these genes support uncontrolled proliferation and chromosomal instability.
- *CKAP2L*, *CASC5*, *FBXO5*, *STIL*: regulators of mitosis, centrosome function, or chromosomal segregation, repeatedly linked to aggressive tumor behavior and poor prognosis in breast cancer.
- *RAD21*: DNA repair and chromosomal cohesion, implicated in therapy response.
- *KDM5B*, *CBX3*, *H2AFZ*: chromatin and epigenetic regulators whose overexpression correlates with metastatic potential, proliferation, and reduced survival.
- *CSE1L*, *CCNE2*: associated with metastasis and poor clinical outcome in breast tumors.

Overall, 55% of Stage I candidates, 83% of Stage II candidates, and 64% of Stage III candidates were already supported by curated cancer resources or published studies as breast cancer–related. This high validation rate supports the predictive value of the stage-specific module approach.

External validation using independent datasets (GSE15852 and GSE69914) To further test generalizability, the authors reapplied their framework to two public GEO datasets:

- GSE15852: gene expression data from 43 primary breast cancers and their matched normal tissues.
- GSE69914: Illumina 450K DNA methylation data including paired normal-adjacent and tumor breast tissue samples.

From these datasets, they identified 79 genes that were both differentially expressed and aberrantly methylated. Using TF–target information and PCC ≥ 0.5, they built a combined breast cancer regulatory network with 195 nodes and 313 edges, then used WGCNA to divide it into modules (turquoise, blue, brown; gray genes were unassigned and discarded). Applying the same two-step prioritization (correlation ≥ 0.8 with $p < 0.05$ and top 5% centrality), they obtained a focused set of 10 candidate genes: *H2AFZ*, *NPM1*, *MAF*, *NR3C1*, *PTGER3*, *TCF4*, *IRF1*, *RARB*, *CHD2*, and *SMAD4*. All but *PTGER3* and *CHD2* were previously linked to breast cancer biology or prognosis. This reproduced the core logic of the pipeline (methylation + expression + regulatory structure + module centrality) on independent data, including GSE69914, and again recovered biologically meaningful breast cancer genes.

Replication workflow for a new dataset The procedure described in this work can be replicated on any breast cancer dataset with (i) gene expression, (ii) DNA methylation, and (iii) clinical staging or at least tumor vs. matched normal pairs:

1. **Stratify by stage**: split tumor/normal pairs into Stage I, Stage II, Stage III (or other clinically defined strata). Exclude strata with too few matched pairs.

2. **Call altered genes**: within each stage, find differentially expressed genes using `limma` with $p < 0.05$ and $|\log \mathrm{FC}| < 0.5$ as stated; classify genes as hypermethylated ($\beta > 0.8$) or hypomethylated ($\beta < 0.2$). Take the intersection of these two sets to focus on genes that are both transcriptionally deregulated and epigenetically abnormal.

3. **Build stage-specific TF→target networks**: obtain TF–target pairs (e.g., from GRNdb), keep those whose targets are in the intersected gene list, and compute Pearson correlation between TF and target across that stage's samples. Retain edges with $|\mathrm{PCC}| \geq 0.5$.

4. **Detect co-regulated modules**: run WGCNA on each stage-specific network. Use hierarchical clustering + Dynamic Tree Cut, enforcing a minimum of ∼30 genes per module.

5. **Identify "core" stage modules**: determine which modules are enriched for genes unique to that stage (i.e., differentially expressed only in Stage I, only in Stage II, etc.). These modules are considered the disease-relevant stage-specific modules.

6. **Prioritize candidate disease genes**: within each stage-specific module, (a) keep genes with strong within-module correlation (≥ 0.8, $p < 0.05$), and (b) keep the top 5% highest-ranked genes by degree, betweenness centrality, and closeness centrality. The intersection are the candidate stage-specific disease genes.

7. **Biological interpretation**: run functional enrichment (e.g., Metascape) to reveal dominant pathways (cell cycle, chromatin remodeling, spindle checkpoint, transcriptional control). Cross-reference the prioritized genes with OMIM, COSMIC, DAVID, and PubMed to evaluate known breast cancer relevance and prognostic value.

Conclusion By integrating DNA methylation, gene expression, TF–target regulatory structure, and explicit stage information, this framework reveals modules of tightly connected, transcriptionally dysregulated, and epigenetically altered genes that are specific to Stage I, Stage II, or Stage III breast cancer. These modules are enriched for cell cycle, chromatin regulation, mitotic checkpoint control, and transcriptional programs known to underlie tumor progression. The most central genes in these modules — including *E2F2*, *E2F8*, *TPX2*, *BUB1*, *CKAP2L*, *CBX3*, *RAD21*, *CCNE2*, *STIL*, *KDM5B*, *TOP2A*, *PCNA* — are strongly supported by literature as drivers of proliferation, chromosomal instability, epigenetic reprogramming, and metastatic potential in breast cancer. The method therefore provides a reproducible, stage-aware strategy to nominate disease genes and potential therapeutic targets, and it was shown to generalize to independent datasets including GSE69914.

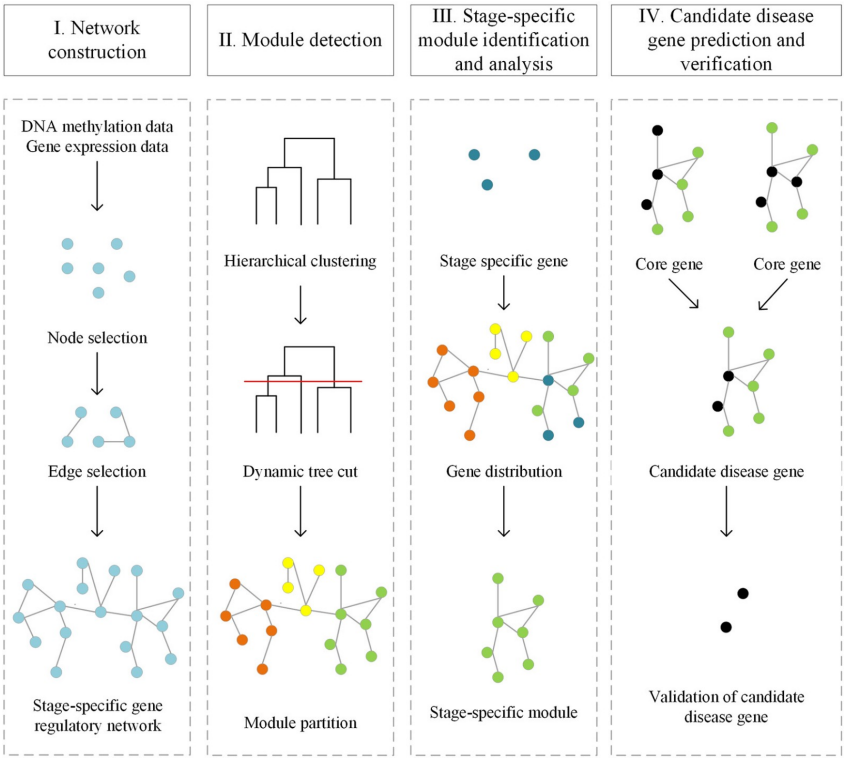

Figure 5: Workflow of the computational framework for predicting disease genes based on stage-specific gene regulatory network.

10 Inference of tissue relative proportions of the breast epithelial cell types luminal progenitor, basal, and luminal mature

Keywords Breast epithelium, luminal progenitor cells, basal cells, luminal mature cells, DNA methylation deconvolution, RNA-seq deconvolution, TNBC, BRCA1 carriers, Fanconi anemia pathway, GSE69914, TCGA, cell-type proportions, RLM-pseudoinverse[10].

Goal of the study The paper develops a computational framework to estimate the relative abundance of key breast epithelial subtypes — luminal progenitor, luminal mature, and basal cells — using bulk genomic data (DNA methylation arrays and RNA-seq). The motivation is that aggressive hormone receptor (HR)-negative, triple-negative breast cancers (TNBCs) are believed to originate from hormone-receptor-negative luminal progenitor cells. Therefore, being able to quantify the proportion of luminal progenitor-like cells in bulk tissue provides information about cancer risk and tumor phenotype. The method is designed for use in realistic cohort-scale data, including Illumina HumanMethylation450K datasets such as GSE69914 and TCGA breast cancer samples, where single-cell sequencing is not available for all individuals.

Two-step deconvolution model Cell-type proportions are inferred using a linear mixing model

$$X = AW + \epsilon,$$

where $X \in \mathbb{R}^{p\times n}$ is the observed bulk data (DNA methylation β-values or RNA-seq expression) across p features and n samples, $A \in \mathbb{R}^{p\times k}$ is a reference matrix containing cell-type-specific profiles for k epithelial subtypes, and $W \in [0,1]^{k\times n}$ is the unknown matrix of mixing proportions (cell-type fractions per sample).
The pipeline proceeds hierarchically:

1. First, estimate broad tissue compartments (epithelial, adipose, fibroblast, immune) using DNA methylation reference profiles from prior studies.

2. Second, within the epithelial compartment, resolve the three epithelial subtypes of interest: luminal progenitor, luminal mature, and basal.

Estimation of W uses a robust linear model to compute a pseudo-inverse of A (RLM-pseudoinverse, implemented via `rlm()` in `MASS` in R). Negative fitted weights (which are interpreted as noise) are truncated to zero, and column weights are renormalized to sum to one. This allows estimation of biologically interpretable cell-type proportions from standard 450K methylation data and from bulk RNA-seq.

Reference matrix construction (A) The reference matrix A encodes the DNA methylation or RNA expression profiles of the three epithelial subtypes.
For DNA methylation:

- Purified luminal progenitor, luminal mature, and basal epithelial cells were profiled by bisulfite sequencing.
- CpG loci were retained only if (i) they were measured both in bisulfite-seq and in bulk Illumina 450K arrays (for downstream compatibility with datasets such as GSE69914), (ii) they showed low variance across non-epithelial lineages, (iii) they had at least a 0.5 difference in mean methylation between the epithelial subtype of interest and non-epithelial lineages, and (iv) they maximally discriminated the epithelial subtypes from each other.
- This yielded a compact DNA methylation reference matrix $A \in [0,1]^{58\times 3}$, where entries are subtype-specific methylation rates $\beta \in [0,1]$ at informative CpGs.

For RNA-seq:

- Single-cell RNA-seq data from 13,909 human breast epithelial cells were used to define luminal progenitor, luminal mature, and basal clusters.
- Clusters were identified with an unsupervised pipeline: spectral graph Laplacian embedding followed by Gaussian Mixture Modeling (GMM-LE), and then validated with known subtype marker genes.
- To enable deconvolution in bulk RNA-seq (which also contains adipose and stromal signal), synthetic reference profiles for stromal and adipose were generated by adapting bulk stromal/adipose libraries so that they were comparable in scale to epithelial single-cell profiles.
- Informative RNA features for each subtype were ranked using either (1) a Mahalanobis-distance-based statistic that accounts for covariance between features, or (2) classical differential expression statistics (edgeR) contrasting each subtype against all others. The top $\sim$250 discriminating genes per subtype were used to assemble A.

This systematic construction of A is crucial, because the accuracy of W depends on having subtype-specific markers that are robust across platforms.

Simulation robustness The method was stress-tested using simulated bulk DNA methylation data created from the known reference A and randomly generated mixing weights W. Increasing fractions of entries in X were then randomly replaced by uniform noise. Even after replacing up to 40% of the data with noise, the reconstructed subtype proportions remained accurate, with low mean squared error and with only a small fraction of samples failing to return valid estimates. This shows that the RLM-based inversion of A is numerically stable and tolerant to noise/missingness typical of 450K arrays.

DNA methylation vs RNA-seq agreement To evaluate biological validity in real data (rather than only simulations), the method was applied to $n = 175$ low-stage (Stage I–II) breast tumor biopsies from TCGA, for which both bulk DNA methylation and bulk RNA-seq data are available for the same samples. For each sample:

- Luminal progenitor, luminal mature, and basal proportions were inferred from DNA methylation using the hierarchical RLM-pseudoinverse approach.
- The same proportions were inferred from RNA-seq using several published deconvolution tools (URSM, Bisque, MuSiC, CIBERSORTx) as well as the same RLM-pseudoinverse strategy.

Correlations between DNAme-based and RNA-based estimates were then computed. When A was defined using the covariance-aware Mahalanobis-distance ranking and W was inferred with the robust RLM-pseudoinverse, Pearson correlations between DNAme-based and RNA-based estimates of luminal progenitor and luminal mature fractions were very high ($r \approx 0.85$–0.9 with narrow 95% CIs). Basal proportions showed moderate correlations ($r \approx 0.5$). These correlation values are notably higher than the typical correlation between mRNA and protein abundance (~ 0.4), indicating good cross-platform consistency in subtype proportion estimates.
This supports using Illumina 450K methylation datasets (including GSE69914) to infer epithelial subtype composition, even in the absence of matched RNA-seq.

Application to breast cancer biopsies and normal tissue The inferred proportions of luminal progenitor-like, luminal mature-like, and basal-like cells were examined in:

- Hormone receptor positive (HR+) vs. hormone receptor negative / triple-negative (TNBC) breast cancers.
- Healthy breast tissue from BRCA1/FANCS mutation carriers.
- Breast tumors from carriers of heterozygous mutations in non-BRCA Fanconi anemia (FA) pathway genes.

Key findings:

- TNBC / HR− tumors showed elevated proportions of luminal progenitor-like cells. This is consistent with the model in which TNBC originates from an HR− luminal progenitor lineage.
- HR+ tumors showed higher luminal mature-like proportions, coherent with the fact that luminal mature cells are typically hormone receptor positive.
- Healthy breast tissue from BRCA1 (FANCS) mutation carriers showed elevated luminal progenitor cell proportions compared to controls, in agreement with prior observations that BRCA1 carriers accumulate at-risk HR− progenitor-like populations.
- Tumors from carriers of heterozygous mutations in non-BRCA FA pathway genes (e.g. FANCD2, FANCG, FANCM) were enriched for luminal progenitor-like and basal-like profiles, and were more often HR− compared to tumors from FA-pathway wild-type individuals. This suggests that defects in DNA damage repair pathways are linked to expansion of progenitor-like epithelial compartments that are predisposed to form HR− / TNBC-like disease.

These results indicate that breast epithelial composition, as estimated from DNA methylation arrays such as GSE69914, reflects clinically relevant biology: specifically, the enrichment of luminal progenitor-like cells in tissues at risk for aggressive, HR− phenotypes.

Relevance to GSE69914 / replication notes The study explicitly uses Illumina HumanMethylation450K data from TCGA and GEO (including GSE69914) to estimate epithelial subtype proportions in bulk breast tissue. The pipeline is directly reproducible on GSE69914 data:

1. Preprocess 450K methylation data (background correction, removal of probes with $< 95\%$ coverage, kNN imputation with $k = 5$ for loci with detection $p > 0.05$).
2. Extract β-values at the 58 informative CpGs that define the epithelial subtype reference matrix A.
3. Estimate broad compartments (epithelial, adipose, fibroblast, immune).
4. Within the epithelial compartment, solve $X \approx AW$ using the robust linear-model pseudoinverse (RLM-PI) to obtain luminal progenitor / luminal mature / basal proportions for each sample.
5. Optionally, model subtype proportions against clinical covariates such as hormone receptor status, BRCA1/FA-pathway mutation status, age, and tumor stage using standard linear regression.

In short, this method turns bulk 450K methylation data into an approximate "cell-type mixture profile" for luminal progenitor, luminal mature, and basal epithelium. These proportions can then be correlated with tumor subtype, risk group, or mutation carrier status.

Conclusion The paper shows that DNA methylation arrays (including GSE69914) can be used not only for CpG-level differential methylation analysis, but also to infer biologically meaningful cell-type composition of the breast epithelium at scale. The deconvolution framework links high luminal progenitor content to HR− / TNBC-like biology, BRCA1/FANCS carrier status, and FA-pathway defects, suggesting that subtype proportion estimates derived from bulk methylation are informative early indicators of aggressive tumor potential and may support risk stratification.

11 An improved epigenetic counter to track mitotic age in normal and precancerous tissues

Keywords DNA methylation, Illumina 450K/EPIC, mitotic age, stem cell divisions, field defects, normal-adjacent tissue, TCGA, risk prediction, GSE69914-like breast cohorts [11].

Aim and conceptual framework The study develops and validates an epigenetic mitotic-age estimator called **stemTOC** (Stochastic Epigenetic Mitotic Timer of Cancer). stemTOC is designed to measure the cumulative number of stem-cell divisions ("mitotic age") in a tissue sample, which is considered a key determinant of cancer risk and early malignant progression. The work focuses on using DNA methylation (DNAm) rather than somatic mutations to quantify mitotic age in normal tissues, precancerous lesions, and tumors across many organs, including breast.

Data sources and link to breast cancer / GSE69914-type data The method is trained and evaluated on large Illumina HumanMethylation450K / EPIC BeadChip datasets, including:

- normal breast tissue from cancer-free donors (*"normal-healthy"*) vs. histologically normal breast tissue sampled adjacent to an invasive breast tumor (*"normal-adjacent"*), using the same type of paired design as in GSE69914 and related Teschendorff breast cohorts (50 normal-healthy vs. 42 normal-adjacent samples),
- TCGA normal-adjacent tissues and tumors across ∼15 cancer types,
- precancerous lesions (e.g. ductal carcinoma in situ in breast, colon adenomas, Barrett's esophagus),
- additional exposure/risk datasets (e.g. buccal swabs from smokers, liver tissue in obese individuals with NAFLD).

This makes the approach directly reusable for breast 450K datasets where "normal-healthy", "normal-adjacent", and tumor tissue are available, such as GSE69914-like designs.

How stemTOC is constructed The pipeline to build stemTOC is explicitly designed to avoid common confounders like cell-type heterogeneity and pure chronological age:

1) **CpG preselection:** start from ∼30,000 CpG sites located near promoters (TSS200) that are *consistently unmethylated* (DNAm $\beta < 0.2$) across 86 fetal/neonatal tissue samples covering 13 tissue types. These CpGs represent an "epigenetic ground state" shared across cell types.
2) **In vitro mitotic filter:** keep only CpGs that *gain* methylation (hypermethylate) as normal cells divide in culture (i.e. increasing population doublings), but *do not* gain methylation under growth arrest (mitomycin or low serum). This yields 629 "vitro-mitCpGs" that seem to respond to cell division itself, not just time.
3) **In vivo aging filter:** further restrict to CpGs that also gain methylation with chronological age *in vivo* in large whole-blood datasets, after adjusting for detailed immune cell composition. This step removes culture artefacts and keeps CpGs whose methylation realistically accumulates with stem/progenitor divisions in living tissue. This results in a final panel of 371 CpGs, called "stemTOC CpGs"
4) **Handling stochasticity / field defects:** in real tissue, early carcinogenic changes are mosaic and heterogeneous: only a subset of cells in a "normal-adjacent" sample may carry epigenetic damage. Instead of averaging methylation across all 371 CpGs, stemTOC defines mitotic age for a sample as the **95th percentile (upper quantile)** of the β-values across those CpGs. This emphasizes the most advanced subclone(s) in that tissue. The optimal 95% cut-off is calibrated by maximizing separation between normal-healthy vs. normal-adjacent breast tissue (50 vs. 42 samples from the breast dataset described above).

Core results (biological validation) Once defined, stemTOC is tested in multiple ways:

- **Chronological age vs. mitotic age in normal tissues:** in normal-adjacent tissues from TCGA, stemTOC correlates strongly with chronological age in high–self-renewal tissues such as colon/rectum, but shows weaker or no correlation in slowly cycling, hormone-sensitive tissues such as breast and endometrium. This matches the idea that tissues with faster stem-cell turnover accumulate mitotic divisions more linearly with age.
- **Discriminating early-risk tissue:** stemTOC is consistently higher in *normal-adjacent* breast tissue than in matched normal-healthy breast tissue, and similarly higher in other pre-invasive states (ductal carcinoma in situ in breast; colon adenoma; Barrett's esophagus; gastric intestinal metaplasia; NAFLD-associated premalignant liver). Thus, the DNAm-based mitotic age is already elevated before full malignancy.
- **Tumor purity / cell-of-origin:** in TCGA tumors across many cancer types (including breast ER+ and basal-like, colon, liver, pancreas, prostate, lung), stemTOC correlates with the inferred fraction of the tumor cell-of-origin (for example, luminal epithelial fraction in luminal breast cancer, hepatocyte fraction in liver cancer, basal epithelial fraction in lung squamous carcinoma). This indicates that higher mitotic age tracks expansion of the transformed lineage within the sample.

- **Exposure to risk factors:** stemTOC increases with known carcinogenic exposures:
  - higher mitotic age in buccal epithelium of smokers vs. never-smokers (all same chronological age),
  - higher mitotic age in normal lung tissue from smokers vs. non-smokers after adjusting for epithelial content,
  - higher mitotic age in obese liver tissue with advanced NAFLD/fibrosis compared to non-fibrotic obese liver.

  This supports the interpretation that carcinogenic stressors accelerate mitotic age in the relevant epithelial compartment.
- **Comparison to other epigenetic/replicative clocks:** stemTOC is benchmarked against previously published DNAm mitotic clocks (epiTOC, epiTOC2, EpiCMIT, HypoClock, RepliTali). It generally shows:
  - stronger association with tissue stem-cell division rate,
  - better discrimination between normal-healthy vs. normal-adjacent / precancerous tissue,
  - tighter coupling to tumor cell-of-origin fraction,
  - reduced sensitivity to shifts in cell-type composition.

  This suggests that stemTOC is more specific to mitotic history than earlier clocks.
- **Relation to mutational "clock-like" signatures:** stemTOC aligns with the burden of SBS1-type (MS1) somatic mutations, which are known to arise from deamination at methylated CpGs and scale with stem-cell division history, but not with SBS5-type (MS5) signatures. This supports that stemTOC is genuinely mitotic in nature.

Relevance for re-analysis of Illumina 450K breast data For a breast 450K dataset containing (i) normal-healthy, (ii) normal-adjacent epithelium, and (iii) tumor, the study shows a practical recipe:

- quantify mitotic age per sample using the 371 stemTOC CpGs and the 95th-percentile rule,
- confirm that normal-adjacent tissue already shows elevated mitotic age (i.e. early "field defect"),
- test whether mitotic age increases further in tumor,
- relate mitotic age to inferred epithelial/tumor cell fraction and to clinical covariates (grade, proliferation, etc.).

This is exactly the scenario of GSE69914-like breast cohorts, where matched normal-adjacent tissue is available and field defects have already been described. The work therefore suggests that mitotic-age scoring (stemTOC) can be directly layered on top of that dataset to quantify early risk and clonal expansion in histologically normal breast tissue.

Conclusion In summary, stemTOC provides an Illumina 450K/EPIC-compatible DNAm score that captures mitotic age in both normal and precancerous tissue. It separates truly healthy tissue from "normal-adjacent at risk", scales with stem/progenitor expansion, correlates with tumor cell-of-origin content across many cancers (including breast), and increases with exposure to carcinogenic stressors such as smoking and obesity. This makes it a directly applicable framework to interpret early epigenetic damage and field defects in datasets like GSE69914.

12 Genome-wide discovery of circulating cell-free DNA methylation signatures for the differential diagnosis of triple-negative breast cancer

Keywords Triple-negative breast cancer (TNBC), DNA methylation, circulating cell-free DNA (cfDNA), Illumina 450K / EPIC arrays, differential methylation, LASSO feature selection, logistic regression score, droplet digital PCR, subtype discrimination, GSE69914, TCGA. [12].

Aim and clinical motivation Triple-negative breast cancer (TNBC) is clinically aggressive and lacks targeted endocrine/HER2 therapies, so deciding early (pre-surgery) whether a lesion is TNBC or a different subtype (non-TNBC) directly affects neoadjuvant strategy, prognosis counseling, and surgical planning. Current subtype calls rely on immunohistochemistry on tissue biopsy, which is invasive and slow. The study by Gao *et al.* proposes a minimally invasive alternative: use DNA methylation patterns, measured either in tumor tissue or in plasma cfDNA, to **distinguish TNBC from non-TNBC**.

Study design overview The workflow has four main steps:

1. **Discovery in tumor tissue**: call differentially methylated CpG sites (DMCs) between TNBC and non-TNBC tumors from genome-wide arrays.
2. **Cross-cohort filtering / validation**: keep only CpGs that also separate TNBC vs. non-TNBC in a large external cohort, and remove CpGs methylated in blood leukocytes (to make them usable in cfDNA).
3. **Marker selection**: compress the CpGs to a minimal diagnostic panel using LASSO and build a methylation diagnostic score (MDS) with logistic regression.
4. **Liquid biopsy implementation**: translate the best markers into a multiplex droplet digital PCR (mddPCR) assay on plasma cfDNA, and test diagnostic power.

A crucial point is that one of the independent validation datasets is **GSE69914**, which is the breast tissue 450K dataset we are studying. This means GSE69914 has already been used to benchmark subtype-separating CpG signatures and to test whether TNBC methylation is distinguishable from other breast cancer subtypes.

Cohorts and preprocessing

- **In-house discovery set**: 5 TNBC vs. 9 non-TNBC primary breast tumor tissues. DNA methylation was profiled on the Illumina HumanMethylationEPIC (850K) BeadChip. After QC (filtering, correction, normalization), CpGs with $|\Delta\beta| > 0.10$ and $p < 0.05$ were called differentially methylated between TNBC and non-TNBC. This lenient cutoff was chosen because the discovery set was small. This yielded 32,787 DMCs: about 28% hypermethylated in TNBC and about 72% hypomethylated in TNBC. Pathway enrichment of these DMCs highlighted signaling programs such as WNT, PI3K–AKT, MAPK, Ras, Rap1, focal adhesion, etc., which are relevant to invasion, survival, and metastasis.
- **TCGA validation set**: 83 TNBC vs. 691 non-TNBC tumors from TCGA (Illumina 450K). Here they applied a *stricter* cutoff $|\Delta\beta| > 0.25$ and $p < 0.05$ to confirm which CpGs from discovery still show strong subtype-specific differential methylation in a large cohort. This reduced the list to 1,130 robust TNBC-vs-non-TNBC CpGs.
- **Blood background filter**: Because the final goal is a *plasma cfDNA* test, they removed CpGs that are heavily methylated or unmethylated in white blood cells (WBC), to avoid background noise in liquid biopsy. Using WBC methylomes (GSE50132, $n = 233$), they *kept only* CpGs whose average $\beta < 0.10$ or $\beta > 0.90$ in WBC (i.e. nearly fully unmethylated or fully methylated in leukocytes). After this leukocyte filter, 113 CpGs remained as high-quality TNBC-vs-non-TNBC candidates.
- **Cross-cohort external datasets**: Besides TCGA, they used GEO breast cancer methylation datasets, including **GSE69914**, as an *independent validation* cohort (342 tumors total are mentioned across GEO sets; GSE69914 provides Illumina 450K β-values for paired normal-adjacent and tumor samples and is widely used in breast methylation field-defect studies).

Feature reduction and final marker panel Even 113 CpGs are too many for a clinical test. Therefore:

1. They examined pairwise correlations among the 113 CpGs and then ran LASSO (least absolute shrinkage and selection operator) with 10-fold cross-validation on the TCGA cohort to deal with multicollinearity and pick the minimal discriminative subset.
2. LASSO retained **8 CpG sites** with non-zero coefficients:
    - cg19758859 (in *SASH1*, chr6, body)
    - cg01095157 (*GORASP2*, chr2, TSS1500)
    - cg14534279 (no gene symbol assigned, chr10, intergenic region)
    - cg17588293 (*ZBTB7B*, chr1, 5′UTR)
    - cg06268921 (intergenic on chr1)
    - cg04016621 (*GRK7*, chr3, TSS1500)
    - cg23247845 (intergenic on chr10)
    - cg02096552 (*DISP1*, chr1, body)

    Six of these CpGs were **hypomethylated** in TNBC vs. non-TNBC, and two were **hypermethylated** in TNBC.

3. Using these 8 CpGs, they built a linear **Methylation Diagnostic Score (MDS)** by logistic regression:

$$\text{MDS} = \sum_i (\beta_i \times \text{coefficient}_i) + \text{intercept},$$

where β_i is the methylation level (β-value) of CpG i.

Diagnostic performance in tissues (TCGA and GSE69914) The 8-CpG MDS cleanly separated TNBC from non-TNBC tumors:

- **In TCGA**: area under the ROC curve (AUC) = 0.922 (95% CI 0.895–0.950). With a decision cutoff of -2.51, sensitivity was $\sim$ 94% in stage I/II TNBC and $\sim$ 82% in stage III/IV TNBC, at specificity $\sim$ 85% against non-TNBC. Importantly, the score also worked specifically in early-stage disease (stage I/II TNBC vs. non-TNBC), where AUC remained above 0.93.
- **In GSE69914 (external validation)**: applying *the same 8-CpG formula and the same cutoff*, the MDS still showed a clear TNBC vs. non-TNBC difference. The AUC was 0.875 (95% CI 0.789–0.961), with sensitivity $\sim$ 87% and specificity $\sim$ 90%.

This is very important for me: **GSE69914 was explicitly used as an independent validation cohort for the subtype-classifier.** In other words, GSE69914 is not just "tumor vs. normal"—it was mined to show that specific CpG patterns reliably distinguish *TNBC tumors* from *non-TNBC tumors*, and that this can be summarized into a portable 8-CpG score. That score generalized across platforms (EPIC 850K discovery $\rightarrow$ 450K validation) and across cohorts.

Prognostic signal in TNBC tissue They next asked if any of these CpGs also stratify outcome *within* TNBC. In TCGA TNBC cases:

- CpG **cg06268921** was significantly associated with both overall survival (OS) and disease-free survival (DFS). Patients with *lower* methylation at cg06268921 had *worse* survival.
- In multivariable Cox models adjusting for age and clinical stage, cg06268921 remained an independent prognostic marker for TNBC, with hazard ratio (HR) $\approx$ 0.25 for OS and $\approx$ 0.19 for DFS when comparing high- vs. low-methylation groups (cutoff $\beta = 0.50$). Lower methylation was linked to poorer outcome.

An external TNBC cohort (GSE72251) did not replicate this survival signal at strong significance, likely due to small size and heterogeneity, but the trend motivated trying to detect cg06268921 in plasma.

Translation to plasma: cfDNA assay and analytical sensitivity To move toward a non-invasive clinical test, the authors developed a **multiplex droplet digital PCR (mddPCR)** assay targeting two of the 8 CpGs (cg06268921 and cg23247845) plus an internal control (*ACTB*). Key technical points:

- Plasma cfDNA was isolated from 1–2 mL of blood from 33 TNBC and 80 non-TNBC patients.
- Primers/probes were designed after bisulfite conversion. The ddPCR platform partitions DNA into droplets and counts methylated copies directly.
- The assay limit of quantification (LOQ) for cg06268921 methylation was $\sim$ 0.01% (i.e. it can detect 1 methylated molecule among $\sim$ 10,000 unmethylated), which is vastly more sensitive than standard multiplex qMSP ($\sim$ 5% LOQ).

cfDNA-based classification performance In plasma cfDNA:

- **cg06268921** alone was significantly hypermethylated in TNBC vs. non-TNBC cfDNA (AUC $\approx$ 0.72).
- cg23247845 alone was weaker.
- Combining them into a cfDNA methylation diagnostic score (**cf-MDS**) still significantly separated TNBC vs. non-TNBC in blood, with AUC $\approx$ 0.73, specificity $\sim$ 82%, and sensitivity $\sim$ 55%. Importantly, this still worked in early-stage (I/II) TNBC.

So they effectively *exported* tumor-tissue methylation logic into a liquid biopsy signature for subtype discrimination, using only milliliters of blood and without sequencing.

What this means for my dataset (GSE69914) For my thesis context:

- **GSE69914 is already treated as a benchmark cohort** for differential DNA methylation in breast cancer, not only normal vs. tumor, but also TNBC vs. non-TNBC subtypes.
- The authors trained their 8-CpG TNBC classifier on TCGA and successfully validated it on GSE69914 using the same coefficients and cutoff. This means:
  1. Preprocessing of GSE69914 in this paper assumes high-quality Illumina 450K β-values, comparable to TCGA after normalization.
  2. At least some samples in GSE69914 are labeled by subtype (TNBC vs. non-TNBC) in a way consistent enough to reproduce high AUC (> 0.85).
  3. The signal is *subtype-specific*, not just "tumor vs. normal": CpGs such as cg06268921, cg19758859 (*SASH1*), cg17588293 (*ZBTB7B*), etc., capture biology unique to TNBC compared to other breast cancers.
- Biologically, many of these CpGs sit in or near genes involved in cell signaling, transcriptional control, adhesion, or chromatin state; pathway enrichment of the broader DMC set implicates WNT, PI3K–AKT, MAPK, focal adhesion, and cytoskeleton regulation — pathways known to drive invasiveness and poor prognosis in TNBC.
- Clinically, one of the CpGs (cg06268921) also carries prognostic information for TNBC survival in TCGA, suggesting that subtype-defining methylation may also stratify risk within TNBC.

In short, GSE69914 has already been mined to (i) define CpG markers that distinguish TNBC from non-TNBC with high accuracy, (ii) validate that this signal generalizes across platforms, and (iii) motivate liquid biopsy translation.

Take-home for replication If I want to replicate or extend this for my own analysis of GSE69914 (or integrate with cfDNA):

1. Separate samples by molecular subtype (TNBC vs. other) *within tumor tissue*.
2. Compute per-CpG $\Delta\beta$ between subtypes and keep only CpGs with strong, consistent shifts.
3. Optionally filter out CpGs that are noisy in blood (for liquid biopsy purposes) using leukocyte methylation references.
4. Use penalized regression (LASSO) to down-select to a minimal CpG panel.
5. Build a logistic regression score (MDS) and test its AUC within GSE69914 as an "external" cohort, exactly as Gao *et al.* did.
6. For translational relevance: design targeted ddPCR assays for the top CpGs (such as cg06268921) and measure them in cfDNA.

Methodologically, this is a clean blueprint for turning subtype-specific tumor methylation into a blood-based assay — and GSE69914 is already proven useful for that validation step.

Conclusion Gao *et al.* showed that:

- TNBC vs. non-TNBC tumors have reproducible subtype-specific DNA methylation differences that persist across platforms and cohorts, including GSE69914.
- A compact 8-CpG tissue panel (MDS) distinguishes TNBC from non-TNBC with AUC ≈ 0.9 in TCGA and ≈ 0.88 in GSE69914.
- One CpG (cg06268921) is also prognostic within TNBC.
- A two-CpG cfDNA ddPCR panel retains discriminatory power in plasma (AUC ≈ 0.73), even for early-stage disease, suggesting a feasible non-invasive test.

The key message is that **my dataset (GSE69914) has already been leveraged to validate a subtype-specific methylation classifier for TNBC**, and specific CpGs from that classifier (notably cg06268921) are already being pushed toward a blood test for pre-operative subtype calling.

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] 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

[3] Z. Yang, A. Jones, M. Widschwendter, and A. E. Teschendorff, "An integrative pan-cancer-wide analysis of epigenetic enzymes reveals universal patterns of epigenomic deregulation in cancer," *Genome Biology*, vol. 16, no. 1, p. 140, 2015. DOI: 10.1186/s13059-015-0699-9

[4] 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

[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

[6] J. Croes et al., "Large-scale analysis of dfna5 methylation reveals its potential as biomarker for breast cancer," *Clinical Epigenetics*, vol. 10, no. 1, pp. 51–64, 2018. DOI: 10.1186/s13148-018-0490-9

[7] 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

[8] A. E. Teschendorff, T. Zhu, C. E. Breeze, and S. Beck, "EPISCORE: Cell type deconvolution of bulk tissue DNA methylomes from single-cell RNA-Seq data," *Genome Biology*, vol. 21, no. 221, pp. 1–33, 2020. DOI: 10.1186/s13059-020-02126-9

[9] L. Fan, J. Hou, and G. Qin, "Prediction of disease genes based on stage-specific gene regulatory networks in breast cancer," *Frontiers in Genetics*, vol. 12, p. 717 557, 2021. DOI: 10.3389/fgene.2021.717557

[10] T. E. Bartlett, P. Jia, S. Chandna, and S. Roy, "Inference of tissue relative proportions of the breast epithelial cell types luminal progenitor, basal, and luminal mature," *Scientific Reports*, vol. 11, p. 23 702, 2021. DOI: 10.1038/s41598-021-03161-7

[11] T. Zhu, H. Tong, Z. Du, S. Beck, and A. E. Teschendorff, "An improved epigenetic counter to track mitotic age in normal and precancerous tissues," *Nature Communications*, vol. 15, no. 4211, 2024. DOI: 10.1038/s41467-024-48649-8

[12] L. Gao et al., "Genome-wide discovery of circulating cell-free dna methylation signatures for the differential diagnosis of triple-negative breast cancer," *PeerJ*, vol. 13, e19888, 2025. DOI: 10.7717/peerj.19888