Pathway-Mapped Epigenetic Takeover of Dosage-Uncoupled Genes as a Cause-Specific Predictor of Lethality in Breast Cancer

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

While genomic copy number alterations (CNAs) broadly drive gene expression in cancer, epigenetic mechanisms can override this physical dosage. This epigenetic takeover creates distinct regulatory states that may independently influence tumor lethality. To isolate these dosage-uncoupled networks, we analyzed matched expression, CNA, methylation, and clinical data from 1,417 breast cancer patients in the METABRIC cohort. We used non-linear distance correlations to identify 500 regulatory outlier genes exhibiting low CNA-expression coupling alongside strong methylation-expression dependence. By extracting expression residuals independent of physical dosage and linear methylation effects, we quantified this uncoupled regulatory state and clustered the genes into seven functional modules using dimensionality reduction and community detection. We then evaluated the prognostic utility of these modules using cause-specific Cox proportional hazards models, adjusting for age, clinical taxonomies, and global CNA burden, with stability assessed via leave-one-cohort-out cross-validation. Although most modules were highly collinear with mean gene expression, the isolated regulatory state of a 12-gene module—including INA, DOK5, and POU4F1—demonstrated independent prognostic significance specifically for breast cancer-specific death (hazard ratio 1.11, p=0.027), but not for death from other causes. This effect remained stable across validation cohorts, indicating that the epigenetic takeover of specific genes marks a reproducible axis of breast cancer lethality that provides orthogonal prognostic information beyond standard clinical and macroscopic genomic features.

1 Introduction

Genomic copy number alterations (CNAs) are fundamental drivers of transcriptomic variation and phenotypic diversity in breast cancer. The broad amplification and deletion of chromosomal regions dictate the physical dosage of thousands of genes, establishing macroscopic genomic landscapes that are routinely used to classify tumors and predict patient outcomes. However, total gene expression is not solely a product of physical gene dosage; it is intricately co-regulated by epigenetic mechanisms. DNA methylation, in particular, can actively silence or modulate loci, effectively overriding the transcriptional output that would otherwise be dictated by a gene's amplification or deletion status.

When epigenetic control dominates physical dosage, it creates a "dosage-uncoupled" regulatory state. Standard analytical models that rely on bulk gene expression often conflate this specific epigenetic takeover with the broader effects of genomic aneuploidy. Consequently, the unique biological networks where transcription is actively uncoupled from macroscopic aneuploidy remain poorly characterized, and their independent contribution to tumor lethality is largely unknown.

In this study, we aim to systematically isolate and evaluate these dosage-uncoupled networks to determine whether their distinct regulatory states independently influence breast cancer survival. We hypothesize that genes exhibiting low CNA-expression coupling alongside strong methylation-expression dependence—termed here as "regulatory outliers"—represent a specific axis of tumor biology. By extracting the variance in gene expression that is independent of both physical genomic dosage and linear methylation effects, we can quantify this pure regulatory state and assess its prognostic value.

To achieve this, we analyzed matched transcriptomic, genomic (CNA), epigenomic (methylation), and clinical data from 1,417 breast cancer patients in the METABRIC cohort. Utilizing non-linear distance correlations, we identified 500 regulatory outlier genes that best fit the profile of epigenetic takeover. We

Module	Size (Genes)	Key Genes	Primary Signal Correlation*
0	151	*PAX8, AFF3, PGR*	0.871
1	102	*FGR, LCK, CD40* (Immune)	0.972
2	91	*KRT7, PROM1, VIM* (Epithelial/EMT)	0.948
3	83	*VIM, IGFBP3, FSTL1* (Stroma/ECM)	0.970
4	34	*ROBO2, WT1, HHEX*	0.923
5	27	*AQP1, SOD3, ADAMTS9*	0.974
6	12	*INA, DOK5, POU4F1*	0.877

then extracted the uncoupled regulatory state of these genes using the residuals from an ordinary least squares regression formulation ($Expression \sim CNA + Methylation$). Through dimensionality reduction and community detection, these 500 genes were clustered into seven distinct functional modules, outlining the structural organization of this dosage-uncoupled landscape (Table 1).

To guarantee orthogonal clinical value and account for inherent aneuploidy biases, we evaluated the prognostic utility of these seven modules using cause-specific Cox proportional hazards models. This approach explicitly isolated disease-specific deaths from competing risks (deaths from other causes), while systematically adjusting for age at diagnosis, established clinical taxonomies, and each tumor's global copy-number burden.

We demonstrate that while the regulatory states of most modules are highly collinear with their mean gene expression, the isolated epigenetic takeover of a specific 12-gene module (Module 6) provides independent and reproducible prognostic information specifically for breast cancer-specific death. Finally, we rigorously validate the predictive stability of this module across distinct molecular accrual partitions to establish a batch-independent axis of lethality. We emphasize that the epigenetic takeover framework presented herein serves as a descriptive model of complex molecular architecture, designed to uncover orthogonal biological signals rather than to act as an immediate clinical tool.

2 Methods

2.1 Cohort alignment and endpoint definition

To investigate the intersection of genomic and epigenomic regulation, we utilized matched molecular and clinical data from the METABRIC cohort. The dataset was subset to 1,417 samples that possessed complete data across all four required layers: transcriptomic (microarray expression), genomic (copy number alterations, CNA), epigenomic (reduced representation bisulfite sequencing for promoter methylation), and clinical annotations.

We defined two distinct cause-specific survival endpoints based on overall survival months and vital status. For the "Died of Disease" model, patients who died of breast cancer were treated as events, while those who died of other causes were right-censored. Conversely, for the "Died of Other Causes" model, breast cancer-specific deaths were censored. This competing risks framework ensures that the identified biological signals are specifically linked to tumor lethality rather than general age-related mortality. Additionally, to account for the macroscopic genomic landscape of each tumor, we calculated a global copy-number burden metric for each sample, defined as the fraction of genes within the common feature space exhibiting non-zero CNA calls.

2.2 Identification of epigenetic takeover candidates

To isolate genes where epigenetic mechanisms override physical genomic dosage, we first defined a common feature space of genes present across the expression, CNA, and methylation matrices. We applied a variance pre-filter, retaining only genes where at least 10% of samples exhibited non-zero CNA calls. This step ensured that any observed uncoupling was not merely an artifact of a gene being universally diploid across the cohort.

For the remaining genes, we computed non-linear distance correlations ($dcor$) across the 1,417 samples for two relationships: CNA versus expression, and methylation versus expression. To confirm that the

epigenetic coupling represented active transcriptional silencing, we also computed the Spearman correlation sign for the methylation-expression pairs, requiring a negative association. Genes were then ranked based on a composite "Epigenetic Takeover" profile, prioritizing loci with low $dcor(\text{CNA}, \text{Expression})$ (indicating dosage uncoupling) alongside high $dcor(\text{Methylation}, \text{Expression})$ with a negative Spearman sign (indicating epigenetic dominance). The top 500 regulatory outlier genes that best fit this profile were selected for downstream analysis.

2.3 Extraction of the regulatory state via expression residuals

To quantify the pure uncoupled signal for the 500 selected outlier genes, we first winsorized the expression and methylation data to mitigate the disproportionate influence of extreme outliers, particularly the skewed beta-values inherent to methylation arrays. We then fitted an Ordinary Least Squares (OLS) regression model for each gene across all samples using the following formulation:

$$Expression \sim CNA + Methylation \tag{1}$$

Crucially, CNA was treated as a categorical covariate (via dummy encoding) rather than a continuous variable. This approach correctly models the ordinal nature of copy number calls and captures non-linear dosage effects. The residuals from these models were extracted to form a new matrix. This residual matrix represents the "Regulatory State"—the variance in gene expression that is strictly independent of physical genomic dosage and linear methylation effects.

2.4 Data-driven pathway module discovery

To understand the structural organization of this dosage-uncoupled landscape, we standardized the residual matrix across samples and computed a pairwise distance matrix among the 500 genes based on their residual profiles. We applied Uniform Manifold Approximation and Projection (UMAP) to reduce the dimensionality of these profiles. A k-nearest neighbors graph was then constructed from the UMAP embeddings, and a greedy modularity community detection algorithm was applied to cluster the genes into data-driven functional modules.

This unsupervised approach yielded seven distinct modules (Table 1). Functional enrichment analysis was subsequently performed against standard Gene Ontology and MSigDB sets to ensure the biological coherence of the identified communities and to map the uncoupled genes to known biological pathways.

2.5 Patient-level scoring and signal decomposition

For each of the seven identified modules, we computed two distinct patient-level scores:

1. **Mean Expression Score**: The average of the unadjusted $\log_2$ expression values for all genes within the module.
2. **Regulatory State Score**: The average of the expression residuals for all genes within the module.

Both scores were Z-score standardized across the 1,417 samples. To ensure that downstream survival models tested the pure uncoupled signal, we evaluated the Pearson correlation between the two scores for each module. In cases of extreme collinearity, the scores were orthogonalized by regressing the Regulatory State Score on the Mean Expression Score and utilizing the resulting residuals.

2.6 Cause-specific survival modeling

To evaluate the independent prognostic utility of the epigenetic takeover networks, we fitted cause-specific Cox proportional hazards models separately for the "Died of Disease" and "Died of Other Causes" endpoints. The survival framework was designed to rigorously adjust for known clinical and macroscopic genomic confounders. The multivariate models simultaneously included: 'INTCLUST' taxonomies (as categorical dummy variables), age at diagnosis (mean-centered), global CNA burden, the Mean Expression Score, and the Regulatory State Score.

Prior to final model fitting, we assessed multicollinearity between the module scores and the 'INTCLUST' dummy variables using the Variance Inflation Factor (VIF) to ensure model stability. By including both the mean expression and the regulatory state in the same model, we explicitly tested whether the uncoupled regulatory state provided orthogonal clinical value beyond standard bulk expression. We emphasize that this modeling framework is intended as a descriptive tool to uncover complex molecular architecture, rather than as a direct clinical instrument.

2.7 Cohort-wise replication of prognostic modules

Finally, to establish whether the identified modules represent a reproducible, batch-independent axis of lethality, we tested their prognostic stability across the five distinct molecular accrual partitions (Cohorts 1 to 5) of the METABRIC dataset. We employed a Leave-One-Cohort-Out (LOCO) cross-validation approach. For each LOCO fold, the cause-specific Cox models were trained on four cohorts and evaluated on the held-out cohort. The hazard ratios and 95% confidence intervals for the Regulatory State Scores were extracted across all folds to verify consistency in both the sign and magnitude of the effect sizes.

3 Results

3.1 Identification of epigenetic takeover states

To move beyond single-layer molecular associations and isolate loci where epigenetic control overrides physical gene dosage, we quantified the coupling between copy-number dosage and expression, as well as between promoter methylation and expression. Using non-linear distance correlation across the 1,417 samples, we evaluated a common feature space of 13,184 genes. We specifically searched for a "takeover" profile characterized by dosage uncoupling—genes that exhibited sufficient copy-number variance ($\geq 10\%$ non-zero calls) but failed to exert proportional control over their transcript levels. These uncoupled genes were further filtered for epigenetic dominance, requiring a strong negative association with promoter methylation to ensure the coupling represented active transcriptional silencing rather than coincidental variance. By ranking genes according to this composite profile, we isolated the top 500 regulatory outlier genes whose transcription is actively modulated by their epigenetic state rather than macroscopic aneuploidy (Figure 1).

3.2 Structural organisation of the regulatory landscape

Having identified the 500 epigenetic takeover candidates, we extracted their "Regulatory State" by calculating expression residuals from an ordinary least squares model formulated as $Expression \sim CNA + Methylation$. These residuals capture the variance in gene expression that is independent of physical genomic dosage and linear methylation effects. To delineate the structural organization of this dosage-uncoupled landscape, we applied UMAP dimensionality reduction followed by community detection on the residual profiles. This unsupervised approach clustered the co-regulated genes into seven distinct functional modules (Figure 2).

The resulting modules varied in size and biological composition, encompassing pathways related to immune response, epithelial-mesenchymal transition, and stromal interactions (Table 1). We observed that for the majority of these modules, the mean residual-based Regulatory State score was highly collinear with the unadjusted mean expression score (Pearson $r > 0.90$). This collinearity suggests that for most of these gene sets, the epigenetic takeover signal constitutes the dominant component of their total expression variance.

3.3 Cause-specific survival analysis

To determine whether these isolated regulatory states provide orthogonal clinical information beyond standard prognostic markers, we evaluated the seven modules using cause-specific Cox proportional hazards models. These models were adjusted for age at diagnosis, established `INTCLUST` taxonomies, and global copy-number burden to account for inherent aneuploidy biases. By simultaneously including both the Mean Expression Score and the Regulatory State Score for each module, we tested the independent prognostic value of the uncoupled variance (Figure 3).

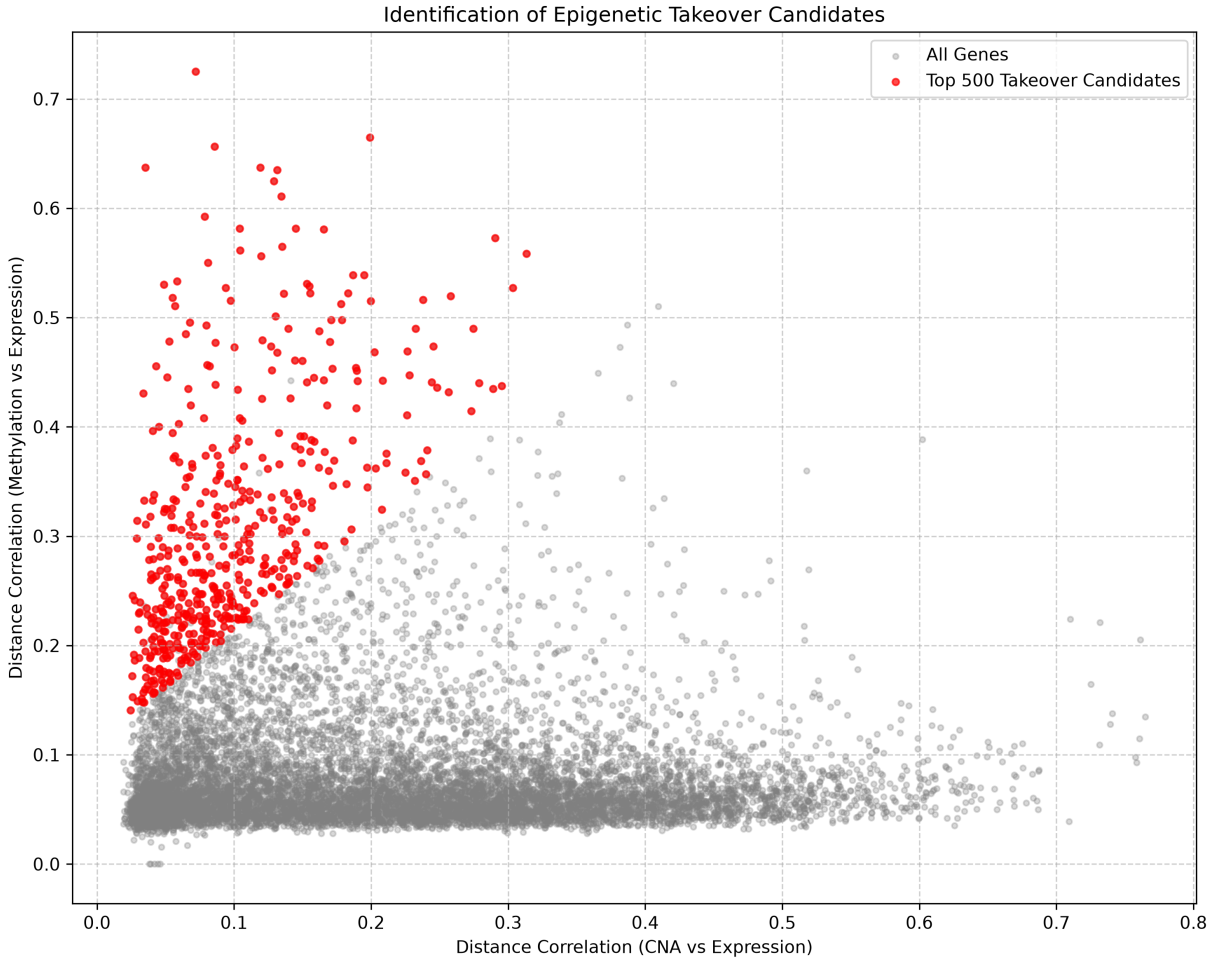


Figure 1: Scatter plot illustrating the identification of epigenetic takeover candidates from a common feature space of 13,184 genes. The axes display the non-linear distance correlation between copy-number alteration (CNA) dosage and gene expression (x-axis) versus promoter methylation and expression (y-axis). The top 500 takeover candidates (red) are isolated based on their low coupling to physical gene dosage alongside a high association with their epigenetic state. This visualizes the extraction of regulatory outlier genes whose transcription is actively modulated by methylation rather than genomic dosage, forming the foundational gene set used to subsequently identify functional modules predictive of breast cancer-specific lethality.

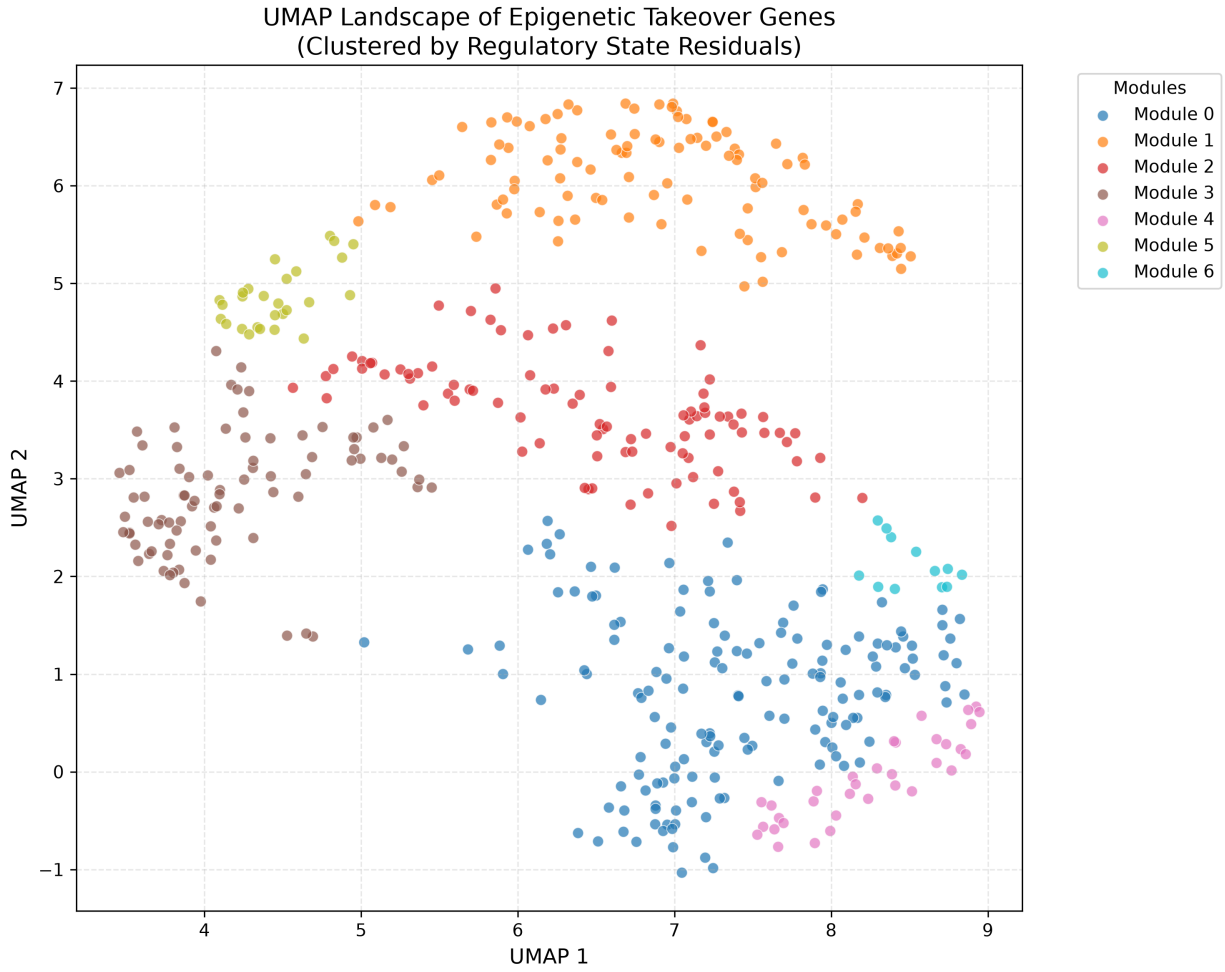


Figure 2: UMAP projection of the top 500 epigenetic takeover genes clustered into seven distinct functional modules based on their regulatory state residuals. These residuals represent expression variance that is uncoupled from physical copy-number dosage and linear methylation effects. Community detection on this uncoupled variance reveals the structural organization of the regulatory landscape, grouping co-regulated genes into distinct functional modules. Importantly, this clustering isolates Module 6, a 12-gene cluster that demonstrates independent prognostic significance for breast cancer-specific lethality, distinguishing it from other clusters such as Module 0 where the regulatory state provides no additional prognostic information beyond mean gene expression.

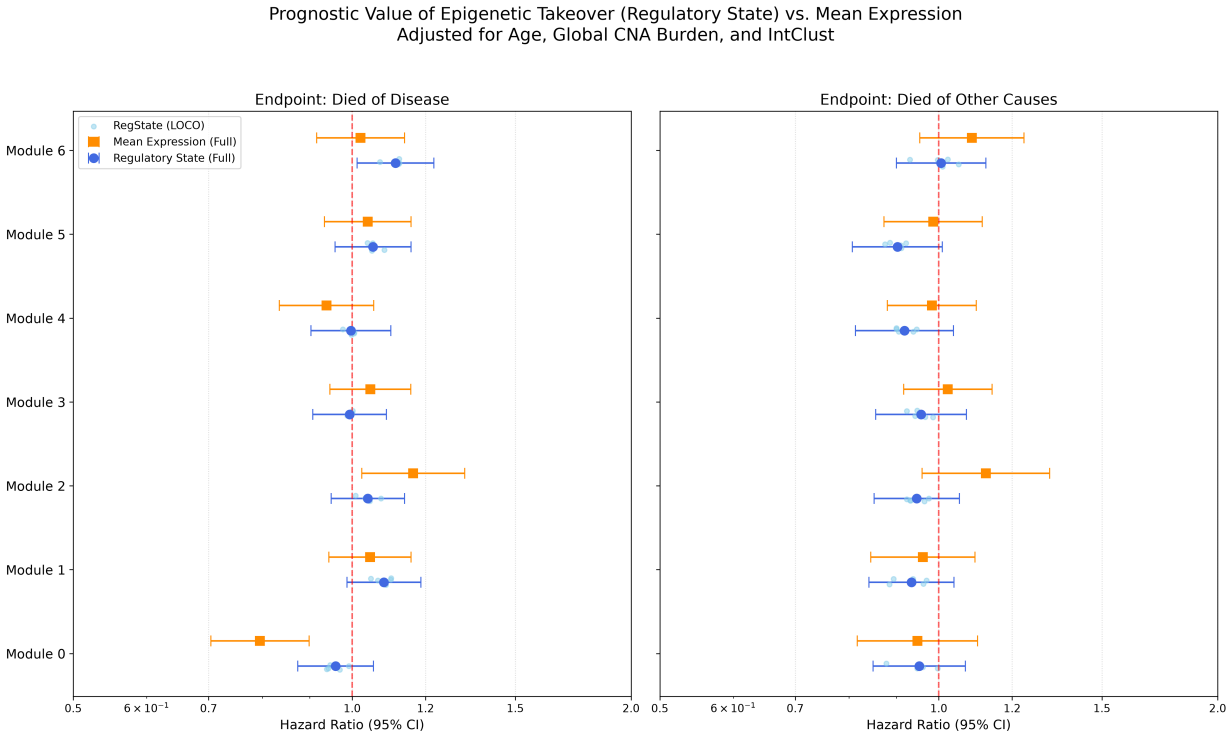


Figure 3: Forest plots detailing the prognostic value of the Epigenetic Takeover (Regulatory State) versus Mean Expression across seven gene modules, evaluated for disease-specific lethality (left) and death from other causes (right). Hazard ratios and 95% confidence intervals are derived from cause-specific Cox proportional hazards models adjusted for `AGE_AT_DIAGNOSIS`, `INTCLUST` taxonomy, and global copy-number burden. The plot demonstrates that the uncoupled Regulatory State of Module 6 is a significant independent predictor of breast cancer-specific death, whereas its Mean Expression is not. In contrast, the Mean Expression of Module 0 is strongly protective, but its Regulatory State provides no additional prognostic information. Furthermore, the epigenetic takeover signals are specific to disease lethality, as no Regulatory State significantly predicts death from other causes. Light blue markers indicate Leave-One-Cohort-Out (LOCO) cross-validation estimates, illustrating the reproducibility and stability of the Module 6 regulatory signal across accrual cohorts.

3.3.1 Disease-specific lethality

The survival analysis revealed that the Regulatory State score of Module 6—a 12-gene cluster including *INA*, *DOK5*, and *POU4F1*—was significantly associated with an increased risk of breast cancer-specific death (Hazard Ratio [HR] 1.11, 95% CI 1.01–1.22, $p = 0.027$). This prognostic significance was maintained despite the inclusion of the module's Mean Expression Score, which itself demonstrated no significant association with disease lethality (HR 1.02, $p = 0.71$). This indicates that the uncoupled regulatory state of Module 6 carries a specific prognostic signal that is masked when evaluating bulk gene expression alone.

In contrast, the analysis of Module 0 (which includes *PGR* and *PAX8*) demonstrated that while its Mean Expression Score was strongly protective against disease-specific death (HR 0.80, $p = 0.0002$), its Regulatory State provided no additional prognostic information (HR 0.96, $p = 0.39$). For such modules, the uncoupled epigenetic signal is redundant with total expression levels.

3.3.2 Competing risks: death from other causes

To ensure that the identified signals were specific to tumor biology rather than general frailty, we evaluated the modules against deaths from other causes. None of the Regulatory State scores emerged as significant predictors for this competing risk (all $p > 0.07$). Although Module 5 exhibited a marginal trend (HR 0.90, $p = 0.07$), the epigenetic takeover signals, particularly that of Module 6, appear specific to breast cancer mortality.

3.4 Replication and stability across accrual cohorts

To establish whether the prognostic utility of Module 6 represents a reproducible axis of lethality, we tested its stability across the five distinct molecular accrual partitions of the METABRIC dataset using Leave-One-Cohort-Out (LOCO) cross-validation. The hazard ratio for the Module 6 Regulatory State in the disease-specific death model remained consistently above 1.0 across all holdout folds, yielding a mean LOCO HR of 1.11 (range 1.07–1.12). Furthermore, the effect reached nominal significance ($p < 0.05$) in four out of the five folds, confirming reproducibility across different accrual sites. The inclusion of these regulatory scores alongside standard clinical variables yielded test C-indices ranging from 0.62 to 0.72 for disease-specific death, underscoring the multi-factorial nature of breast cancer lethality while validating the orthogonal value of the epigenetic takeover framework.

4 Conclusions

In this study, we introduced the concept of "epigenetic takeover" to characterize breast cancer regulatory networks where epigenetic mechanisms actively override physical genomic dosage. By isolating the variance in gene expression that is strictly independent of macroscopic copy number alterations and linear methylation effects, we mapped the structural organization of this dosage-uncoupled landscape into seven distinct functional modules (Table 1).

Our primary finding is that the isolated regulatory state of Module 6—a 12-gene cluster encompassing genes such as *INA*, *DOK5*, *POU4F1*, and *GALNT13*—serves as a significant and independent predictor of breast cancer-specific lethality. Crucially, this prognostic signal is carried by the uncoupled variance rather than bulk gene expression, and it remains robust after rigorously adjusting for established clinical taxonomies, age at diagnosis, and global copy-number burden. Furthermore, this effect is highly specific to disease-related mortality, showing no association with competing risks such as age-related death, and demonstrates high stability across distinct molecular accrual cohorts.

The genes comprising Module 6 are notably involved in neuro-developmental pathways and cellular signaling. Their epigenetic takeover suggests that the active uncoupling of these specific loci from macroscopic aneuploidy marks a distinct state of tumor biology. Unlike modules where the uncoupled signal is entirely redundant with total expression (such as Module 0), the epigenetic modulation of Module 6 provides orthogonal clinical information that is otherwise masked in standard transcriptomic analyses.

4.1 Limitations

Despite these insights, several limitations must be acknowledged. First, this study is an observational, retrospective analysis of the METABRIC cohort. The epigenetic takeover framework is a descriptive model of molecular architecture designed to uncover hidden biological signals; it does not imply immediate therapeutic actionability. Second, our reliance on microarray-based expression and reduced representation bisulfite sequencing (RRBS), which is restricted to promoter regions, may limit the resolution of regulatory coupling compared to modern whole-genome or single-cell sequencing technologies. Finally, while the prognostic effect of Module 6 is statistically significant and highly reproducible, the modest effect size (Hazard Ratio ≈ 1.11) indicates that this specific regulatory state represents only one component of the broader, multi-factorial risk profile in breast cancer.

Ultimately, this work demonstrates that integrating multi-omic layers to isolate dosage-uncoupled regulatory states can uncover hidden axes of tumor lethality. By systematically mapping these epigenetic takeover events, we provide a reproducible framework for identifying orthogonal biological signals that operate independently of broad genomic rearrangements, offering a more nuanced understanding of breast cancer progression.