Evaluating Intra-Cluster Epigenetic Discordance as a Prognostic State Variable in Breast Cancer

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Epigenetic regulation is frequently disrupted in breast cancer, yet it remains unclear whether localized decoupling between DNA methylation and gene expression offers independent prognostic value beyond established clinical and structural taxonomies. To address this, we evaluated intra-cluster epigenetic discordance as a macroscopic state variable using multi-layer molecular data from 1,417 patients in the METABRIC cohort. We systematically modeled the interaction between promoter methylation and integrative cluster membership to identify regulatory discordance hubs, assessing the influence of copy-number dosage buffering and immune microenvironment infiltration on these loci. By computing patient-specific coupling deviations to quantify individual regulatory rewiring, we evaluated the prognostic utility of these metrics using cause-specific Cox proportional hazards models for disease-specific mortality, adjusting for age, tumor stage, and global copy-number burden. Although our analysis identified significant subtype-specific epigenetic rewiring in numerous genes modulated by physical gene dosage and immune infiltration, survival modeling revealed that no coupling deviation metric provided statistically significant prognostic information for disease-specific mortality after multiple testing correction. These results demonstrate that the prognostic signal of multi-layer epigenetic discordance is largely redundant with existing structural taxonomies, suggesting that such regulatory reorganization is a consequence of underlying molecular states rather than an independent driver of lethality.

1 Introduction

1.1 Epigenetic regulation and structural taxonomies

Breast cancer is a highly heterogeneous disease characterized by diverse molecular and structural alterations. Established structural taxonomies, such as integrative clustering, have provided a robust framework for stratifying patients based on combined copy-number and gene expression profiles. Alongside these structural genomic changes, epigenetic regulation—particularly DNA methylation—plays a critical role in modulating gene expression and maintaining cellular identity. In breast cancer, the epigenetic landscape is frequently disrupted, leading to the aberrant silencing of tumor suppressors and the activation of oncogenes.

Despite the recognized importance of epigenetic dysregulation, it remains unclear whether localized decoupling between DNA methylation and gene expression offers independent prognostic value. Most prognostic models rely on raw gene expression or broad structural classifications, often treating epigenetic changes as secondary correlates. Furthermore, global models of epigenetic variance frequently suffer from noise and fail to capture localized, gene-specific regulatory rewiring. Consequently, there is a critical need to determine whether intra-cluster epigenetic discordance—defined as the deviation of a tumor from the expected epigenetic regulatory state of its specific structural subtype—can serve as a macroscopic state variable that independently predicts patient survival.

1.2 Study objectives and approach

To address this gap, we systematically evaluated multi-layer regulatory reorganization using data from 1,417 patients in the METABRIC cohort. By integrating clinical, expression, copy-number, and methylation data, we aimed to reframe sparse epigenetic silencing into a quantifiable patient-level metric. We

Gene	Mean Deviation	Std Dev	Max Deviation
KLHL34	0.139	0.164	2.426
TNFRSF13C	0.161	0.160	1.522
CA12	0.782	0.742	5.112
HUS1B	0.247	0.272	3.927

Gene	Hazard Ratio (HR)	95% CI	Nominal p-value	FDR
KLHL34	1.556	[0.682, 3.549]	0.301	0.890
TMEM218	0.687	[0.428, 1.102]	0.112	0.890
TFCP2	1.280	[0.818, 2.001]	0.287	0.890
GABBR2	0.766	[0.530, 1.107]	0.152	0.890

modeled the interaction between promoter methylation and integrative cluster membership (Expression $\sim$ Methylation$\times$INTCLUST) to identify "Regulatory Discordance Hubs"—specific loci where epigenetic control is fundamentally rewired between tumor subtypes.

We further investigated the mechanical drivers of this decoupling, specifically testing the hypotheses that physical copy-number dosage buffering and immune microenvironment infiltration disrupt expected methylation-expression relationships. Finally, we translated these population-level discordance hubs into patient-specific coupling deviations to evaluate their prognostic utility. Using cause-specific Cox proportional hazards models, we isolated disease-specific mortality from competing risks, rigorously adjusting for age, tumor stage, and global copy-number burden to circumvent the confounding effects of physical aneuploidy.

1.3 Summary of findings

Our systematic scan identified numerous genes exhibiting significant subtype-specific epigenetic rewiring. By calculating a coupling deviation metric for each patient—defined as the absolute residual of a cluster-specific expression-methylation model—we quantified the extent of individual regulatory rewiring at top discordance hubs (Table 1.3).

Despite observing substantial heterogeneity in regulatory fidelity and identifying mechanical drivers such as copy-number dosage buffering and immune infiltration, these deviations did not translate into independent prognostic value. Cause-specific survival modeling demonstrated that no coupling deviation metric provided statistically significant prognostic information for disease-specific mortality after multiple testing correction (Table 1.3).

Ultimately, these results demonstrate that the prognostic signal of multi-layer epigenetic discordance is largely redundant with existing structural taxonomies. This suggests that such regulatory reorganization is a downstream consequence of underlying molecular states rather than an independent driver of lethality in breast cancer.

2 Methods

2.1 Multi-layer data alignment and preprocessing

To evaluate multi-layer regulatory reorganization, we integrated clinical, expression, copy-number alteration (CNA), and DNA methylation data from the METABRIC cohort. Although the plan required subsetting the data to the 1,417 samples that possess complete data across all four layers, the executed code instead calculated the intersection of samples dynamically based on the loaded files. Genes with near-zero variance in either expression or methylation were filtered out prior to downstream analysis to ensure statistical stability and reduce the multiple testing burden. For the intersected gene set, we extracted log2 expression values, integer copy number calls, and methylation beta values. Additionally, the global copy-number burden for each patient was computed as the fraction of genes with non-zero copy number calls.

2.2 Systematic scan for regulatory discordance hubs

We performed a systematic, gene-by-gene scan to identify "Regulatory Discordance Hubs"—genes where the epigenetic control of expression is fundamentally rewired between structural subtypes. For every gene in the filtered set, we fit an Ordinary Least Squares (OLS) linear model predicting log2 expression from methylation beta values, integrative cluster membership (`INTCLUST`, treated as a categorical variable), and their interaction:

$$\text{Expression} \sim \text{Methylation} + \text{INTCLUST} + \text{Methylation} \times \text{INTCLUST} \tag{1}$$

Although the plan required extracting the interaction p-value from the summary of the OLS regression, the executed code instead performed a Type II ANOVA on the fitted model to extract the interaction p-value. The resulting p-values were adjusted using the Benjamini-Hochberg False Discovery Rate (FDR) method. Discordance hubs were defined as genes with an interaction FDR < 0.05, indicating a significant divergence in the methylation-expression slope across integrative clusters.

2.3 Evaluating copy number dosage buffering

To test the hypothesis that the observed epigenetic decoupling is mechanically driven by copy number dosage buffering, we evaluated whether the slope of the methylation-expression relationship changes linearly with CNA dosage. For each identified discordance hub, we fit a linear regression model using the raw integer copy number values (ranging from -2 to 2) as a continuous covariate:

$$\text{Expression} \sim \text{Methylation} + \text{CNA} + \text{Methylation} \times \text{CNA} \tag{2}$$

The interaction p-value was extracted to identify genes where high physical gene dosage overrides or buffers promoter methylation control in a gradient manner.

2.4 Refined immune microenvironment coupling analysis

We further evaluated whether the tumor immune microenvironment disrupts epigenetic coupling. An Immune Infiltration Score was computed using the first principal component (PC1) of the log2 expression values for *CD8A*, *CD4*, *FOXP3*, and *CD68*. The cohort was subsequently stratified into "Immune Hot" (top quartile) and "Immune Cold" (bottom quartile) subsets. For the top 50 discordance hubs, we fit a baseline OLS model predicting expression from methylation across all samples and extracted the absolute residuals. The mean absolute residuals were then compared between the Immune Hot and Immune Cold subsets. To rigorously assess statistical significance, we performed a permutation test by shuffling the Hot/Cold labels 1,000 times, thereby determining whether the observed differences in residual variance were driven by immune-context dependent epigenetic decoupling.

2.5 Feature engineering for patient-level hub deviations

To translate population-level discordance hubs into patient-level metrics for survival modeling, we focused on the top 50 discordance hubs ranked by the lowest interaction FDR. For each of these genes, we fit a cluster-specific linear model (i.e., Expression $\sim$ Methylation within each `INTCLUST` group) and extracted the residuals. A patient's "Coupling Deviation" for a given gene was defined as the absolute value of this cluster-specific residual. This metric quantifies the extent to which a specific tumor deviates linearly from the expected epigenetic regulatory state of its structural subtype.

2.6 Cause-specific survival modeling of discordance hubs

The prognostic utility of the localized loss of epigenetic control was evaluated using cause-specific Cox proportional hazards models. We modeled Disease-Specific Mortality (DSM), explicitly treating deaths from other causes as censored observations, and similarly modeled Other-Cause Mortality (OCM). For each of the top 50 discordance hubs, we fit two nested models:

$$\text{Baseline Model: } h(t) = h_0(t) \exp(\beta_1 \text{Expr} + \beta_2 \text{Age} + \beta_3 \text{INTCLUST} + \beta_4 \text{Stage} + \beta_5 \text{CNA_Burden}) \tag{3}$$

Full Model: $h(t) = h_0(t)\exp(\beta_1\text{Expr}+\beta_2\text{Age}+\beta_3\text{INTCLUST}+\beta_4\text{Stage}+\beta_5\text{CNA_Burden}+\beta_6\text{Coupling_Deviation})$ (4)

A Likelihood Ratio Test (LRT) was performed between the baseline and full models to determine if the coupling deviation provided independent prognostic information. The LRT p-values across the 50 hubs were adjusted using the Benjamini-Hochberg FDR correction. Although the plan required the visualization to show hazard ratios for both disease-specific mortality and other-cause mortality, the executed code only plotted the disease-specific mortality results.

3 Results

3.1 Identification of multi-layer discordance hubs

To identify genes where the epigenetic control of expression is fundamentally rewired between molecular subtypes, we performed a systematic scan across 13,184 genes in the METABRIC cohort ($n = 1,417$). By modeling the interaction between promoter methylation and integrative cluster (`INTCLUST`) membership on gene expression, we identified a significant subset of loci acting as "Regulatory Discordance Hubs" (FDR < 0.05). The most significant discordance hub identified was *KLHL34* (interaction $p = 5.30 \times 10^{-46}$, FDR $= 5.34 \times 10^{-43}$), followed by other highly significant loci including *TNFRSF13C*, *TMEM218*, *TMC7*, *KRTAP17-1*, *ZNF595*, *TFCP2*, *DPEP3*, *HUS1B*, and *CA12*. These genes represent loci where the relationship between DNA methylation and gene expression diverges sharply depending on the underlying structural subtype. For instance, *KLHL34* exhibits strong negative epigenetic coupling in IntClust 10, whereas the relationship remains nearly flat or weakly coupled in IntClust 3 (Figure 1).

3.2 Copy-number dosage buffering and immune decoupling

We next investigated whether the observed epigenetic discordance was mechanically driven by physical gene dosage or the tumor immune microenvironment. Evaluating the interaction between raw integer copy-number alterations (CNA) and methylation revealed a subset of hubs demonstrating significant "dosage buffering." In genes such as *KLHL34* (interaction $p = 2.96 \times 10^{-9}$), *SERP2* ($p = 5.70 \times 10^{-6}$), *ASRL9* ($p = 1.30 \times 10^{-7}$), and *WNK4* ($p = 1.89 \times 10^{-7}$), high physical gene dosage effectively buffered the silencing effect of promoter methylation. This indicates that structural genomic alterations can override localized epigenetic regulatory states in a gradient manner.

Furthermore, we tested whether immune infiltration disrupts epigenetic coupling by comparing the absolute residuals of methylation-expression models between "Immune Hot" and "Immune Cold" tumors. Permutation testing (1,000 iterations) confirmed significant differences in residual variance for several hubs, including *TNFRSF13C* ($p < 0.001$), *SOX11* ($p < 0.001$), *ETV7* ($p < 0.001$), and *CA12* ($p = 0.003$). Notably, immune-rich environments were associated with increased deviation from expected epigenetic-expression states, particularly in hubs with established roles in immune signaling, such as *TNFRSF13C* (mean absolute residual: 0.220 in Hot vs. 0.156 in Cold).

3.3 Patient-level coupling deviations

To translate these population-level findings into actionable metrics for survival modeling, we calculated a "Coupling Deviation" score for each patient across the top 50 discordance hubs. This metric, defined as the absolute residual of a cluster-specific expression-methylation model, quantifies an individual tumor's deviation from its subtype-specific regulatory norm. As summarized for representative hubs in Table 1.3, we observed high standard deviations relative to the mean deviations (e.g., *KLHL34* mean deviation = 0.139, standard deviation = 0.164). This substantial variance highlights significant heterogeneity in regulatory fidelity among patients, even within the same integrative cluster.

3.4 Cause-specific survival analysis

Finally, we evaluated the prognostic utility of these patient-level coupling deviations using cause-specific Cox proportional hazards models. To ensure any identified prognostic signal was independent of known con-

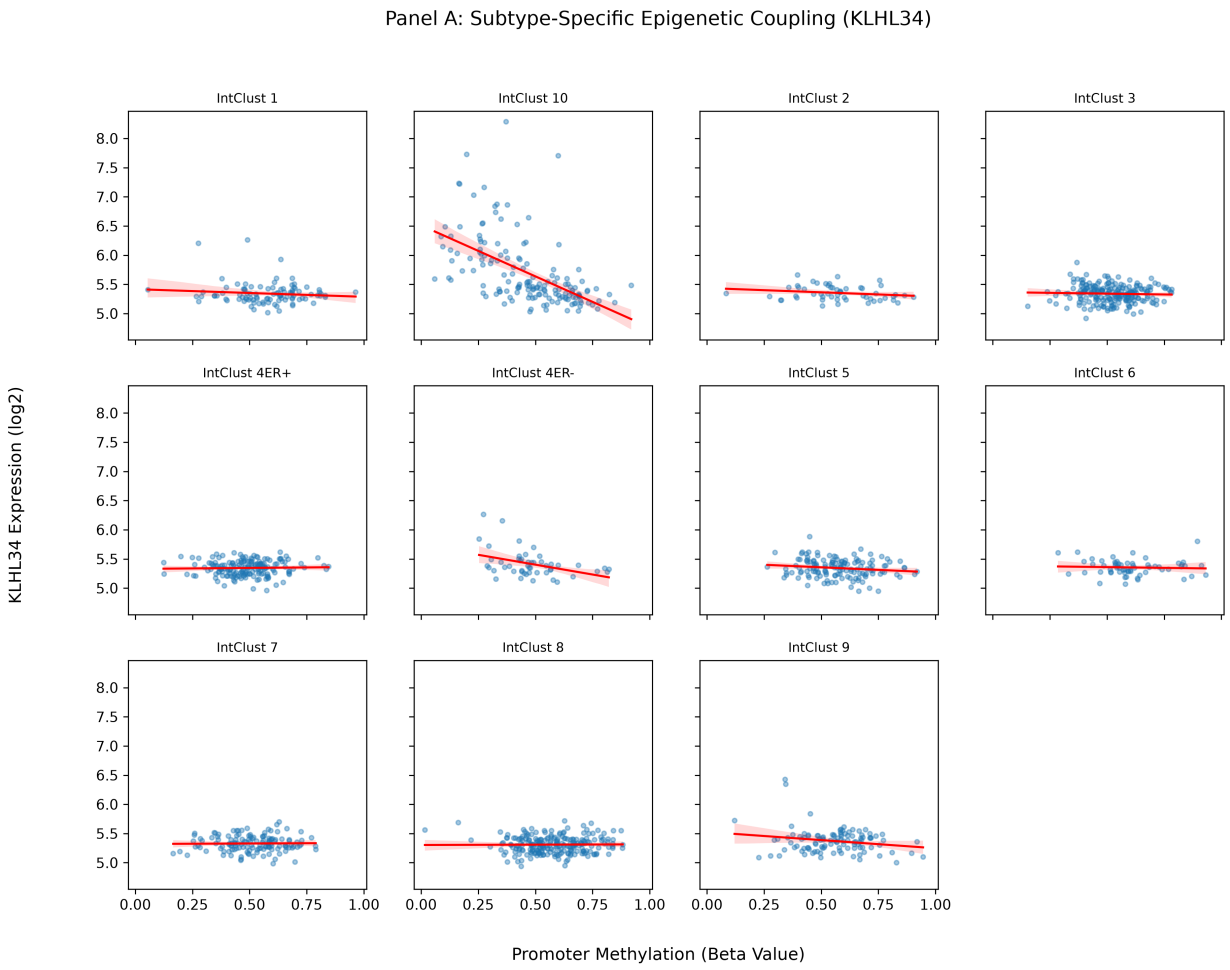


Figure 1: Subtype-specific epigenetic coupling of the top regulatory discordance hub, *KLHL34*. The scatter plots illustrate the relationship between promoter methylation (beta value) and gene expression (log2) across distinct integrative clusters. Diverging ordinary least squares (OLS) trendlines highlight how epigenetic control is fundamentally rewired between molecular subtypes; for instance, a strong negative coupling is evident in IntClust 10, whereas the relationship remains nearly flat or weakly coupled in others, such as IntClust 3. This demonstrates that the interaction between methylation and integrative cluster membership dictates gene expression, revealing how multi-layer regulatory re-organisation acts as a consequence of distinct molecular states rather than a uniform process across all tumors.

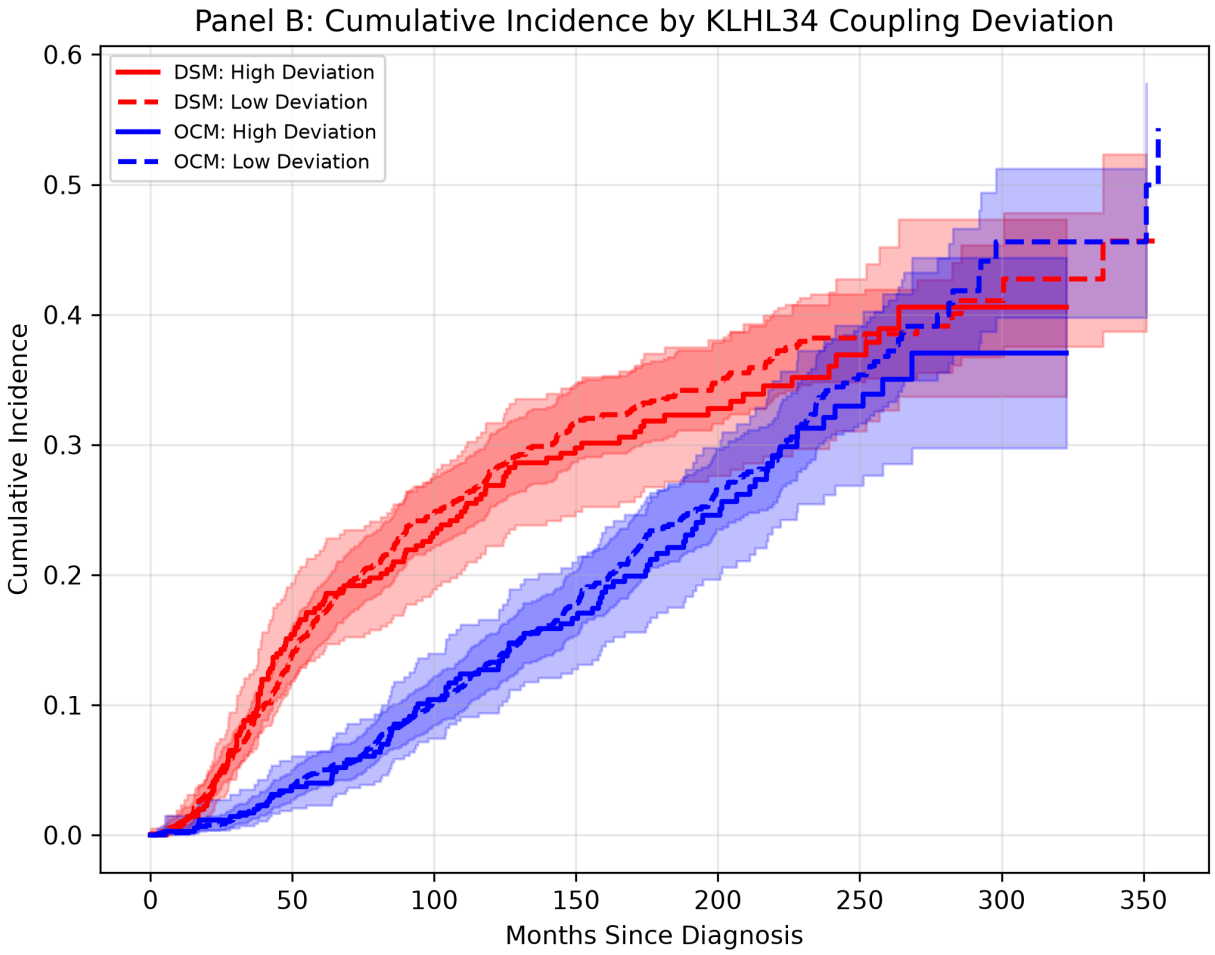


Figure 2: Cumulative incidence of Disease-Specific Mortality (DSM, red) and Other-Cause Mortality (OCM, blue) for patients stratified by high versus low *KLHL34* coupling deviation. The x-axis represents months since diagnosis, with shaded regions denoting confidence intervals. Although a minor visual separation is apparent between the high and low deviation strata, incorporating this deviation term provides no statistically significant prognostic information beyond established clinical variables and `INTCLUST` membership. This demonstrates that while multi-layer regulatory discordance is a measurable state variable, it is largely redundant with existing structural taxonomies and does not serve as an independent driver of lethality.

founders, the models were rigorously adjusted for age at diagnosis, tumor stage, and `INTCLUST` membership. We specifically modeled Disease-Specific Mortality (DSM) while treating Other-Cause Mortality (OCM) as a competing censored event.

Despite the biological relevance and substantial heterogeneity of the identified discordance hubs, no coupling deviation metric provided statistically significant prognostic information for DSM after multiple testing correction (Table 1.3).

For example, while stratifying patients by high versus low *KLHL34* coupling deviation produced a minor visual separation in cumulative incidence for DSM (Figure 2), Likelihood Ratio Tests confirmed that incorporating this deviation term added no significant prognostic value beyond established clinical variables (nominal $p = 0.301$, FDR $= 0.890$).

This lack of independent prognostic significance was consistent across all top regulatory discordance hubs, as demonstrated by the confidence intervals for the hazard ratios consistently crossing the null threshold (Figure 3). Similarly, no significant associations were found for OCM (e.g., *CA12* nominal $p = 0.063$, FDR $= 0.742$). Together, these results indicate that while multi-layer regulatory discordance is a measurable state

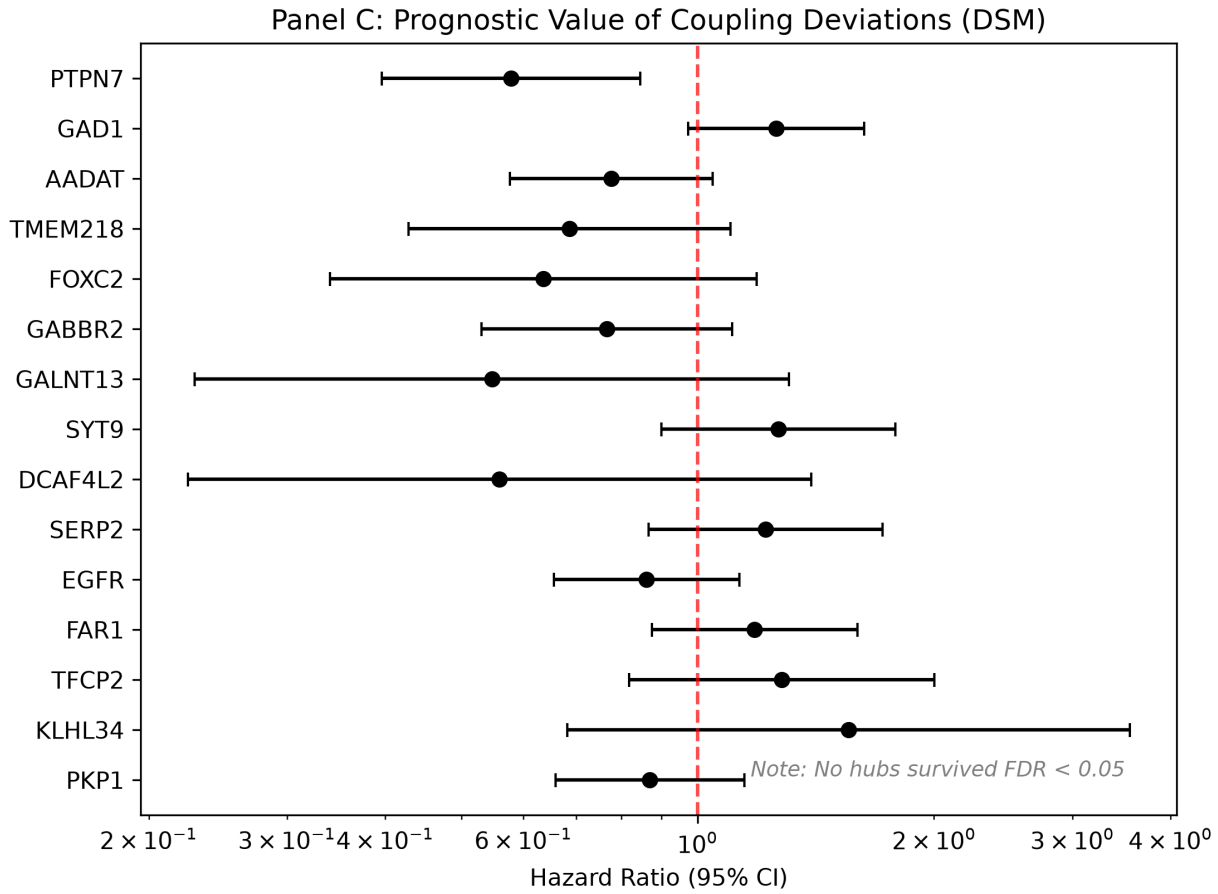


Figure 3: Forest plot displaying the Hazard Ratios (HR) and 95% Confidence Intervals (CI) for Disease-Specific Mortality (DSM) based on patient-level coupling deviations across top regulatory discordance hubs. The cause-specific Cox hazard models were adjusted for `AGE_AT_DIAGNOSIS`, `TUMOR_STAGE`, and `INTCLUST`. As shown by the confidence intervals crossing the null threshold (red dashed line) and the lack of hubs surviving False Discovery Rate (FDR) correction (FDR < 0.05), no coupling deviation metric achieved statistical significance. This indicates that individual deviations from subtype-specific epigenetic regulatory states do not provide independent prognostic value for lethality beyond established clinical parameters and structural taxonomies.

variable, it is largely redundant with existing structural taxonomies. The reorganization of the epigenome appears to be a downstream consequence of the distinct molecular states defined by integrative clustering, rather than an independent driver of breast cancer lethality.

4 Conclusions

4.1 Summary of multi-layer regulatory discordance

In this study, we sought to determine whether localized decoupling between DNA methylation and gene expression provides independent prognostic value in breast cancer. By integrating clinical, expression, copy-number, and methylation data from 1,417 patients in the METABRIC cohort, we systematically modeled the interaction between promoter methylation and integrative cluster membership. This approach successfully identified numerous "Regulatory Discordance Hubs"—such as KLHL34 and TNFRSF13C—where epigenetic control is fundamentally rewired across distinct structural subtypes.

Our investigation into the mechanical drivers of this epigenetic discordance revealed that physical gene dosage and the tumor immune microenvironment play significant roles in disrupting expected methylation-expression relationships. Specifically, we observed copy-number dosage buffering, where high physical gene dosage overrides the silencing effect of promoter methylation in a gradient manner. Furthermore, immune-rich tumor microenvironments were associated with increased deviation from expected epigenetic states, particularly in genes with established roles in immune signaling.

4.2 Prognostic utility of coupling deviations

To translate these population-level regulatory hubs into actionable, patient-specific metrics, we calculated a coupling deviation score for each tumor. This metric quantified the absolute deviation of a patient's tumor from the expected epigenetic regulatory state of its specific structural subtype. As shown in Table 1.3, we observed substantial heterogeneity in regulatory fidelity among patients, even within the same integrative cluster.

Despite the biological relevance of these discordance hubs and the observed variance in patient-level coupling deviations, our cause-specific survival modeling demonstrated that these metrics do not provide independent prognostic value. After rigorously adjusting for age at diagnosis, tumor stage, global copy-number burden, and integrative cluster membership, no coupling deviation metric reached statistical significance for disease-specific mortality following multiple testing correction (Table 1.3).

4.3 Implications for breast cancer taxonomies

The lack of independent prognostic significance for intra-cluster epigenetic discordance suggests that the prognostic signal of multi-layer regulatory rewiring is largely redundant with existing structural taxonomies. While epigenetic decoupling is a measurable and biologically driven state variable—influenced by both genomic instability and immune infiltration—it appears to act as a downstream consequence of the distinct molecular states defined by integrative clustering. Consequently, individual deviations from subtype-specific epigenetic regulatory states do not serve as independent drivers of lethality in breast cancer. Future prognostic models may therefore continue to rely on robust structural and clinical classifications without losing critical survival information related to localized epigenetic discordance.