Subtype-Specific Regulatory Rewiring Modules as Age-Stratified Competing Risk Predictors in Breast Cancer

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

While breast cancer is driven by complex multi-omic alterations, it remains unclear whether localized disruptions in joint epigenetic and genetic regulatory networks provide independent prognostic value beyond global genomic instability and established clinical taxonomies. To address this, we analyzed 1,417 primary breast tumors from the METABRIC cohort across transcriptomic, copy-number alteration (CNA), and promoter methylation layers. We computed gene-wise distance correlations to capture non-linear regulatory dependencies and clustered these profiles to identify distinct Regulatory Rewiring Modules. We then quantified patient-specific module disruption and CNA-versus-methylation dominance shifts, rigorously adjusting for global confounders such as total copy-number burden using Random Forest regression. The prognostic utility of these adjusted metrics was evaluated in an independent validation set using cause-specific Cox proportional hazards models, stratified by age to account for competing risks of disease-specific and other-cause mortality. We identified twelve distinct modules, including an ERBB2-associated network exhibiting strong copy-number-driven expression control in HER2-enriched subtypes. However, survival analysis revealed that neither the adjusted disruption nor the dominance shift of this module provided statistically significant independent prognostic value for disease-specific mortality. Furthermore, the effect sizes of these module metrics were comparable to those of a genomic-location-matched null module, indicating that while multi-omic regulatory rewiring forms distinct biological architectures, current patient-level quantifications of these localized disruptions do not yield prognostic information beyond established clinico-pathological variables.

1 Introduction

1.1 Multi-omic regulation in breast cancer

Breast cancer is a highly heterogeneous disease driven by complex multi-omic alterations, including copy-number alterations (CNAs) and epigenetic modifications such as promoter methylation. Integrative clustering approaches have successfully stratified tumors into distinct molecular subtypes with varying clinical trajectories. However, the interplay between these molecular layers—specifically how genetic and epigenetic mechanisms jointly regulate gene expression—remains a critical area of investigation.

1.2 Challenges in quantifying regulatory rewiring

While global genomic instability and broad epigenetic shifts are established hallmarks of cancer, it is less clear whether localized disruptions in these joint regulatory networks offer independent prognostic information. Tumor-intrinsic regulatory rewiring may manifest as a shift in dominance from methylation-mediated silencing to copy-number-driven expression control. Yet, quantifying the prognostic value of such localized multi-omic uncoupling is challenging. It requires rigorously isolating targeted regulatory failure from macroscopic confounders, such as total copy-number burden and global transcriptional noise. Furthermore, survival analyses in heterogeneous cohorts must account for clinical confounders like age and the competing risks of disease-specific versus other-cause mortality.

Stratum	Event	CIF (5-yr)	CIF (10-yr)	N
Global: RRM_11 Disruption High	Disease	0.209	0.272	155
Global: RRM_11 Disruption Low	Disease	0.130	0.179	155
IntClust 5: RRM_11 Disruption High	Disease	0.423	0.503	26
IntClust 5: RRM_11 Dominance Shift High	Disease	0.483	0.555	29

1.3 Study overview and findings

To address these challenges, we analyzed 1,417 primary breast tumors from the METABRIC cohort across transcriptomic, CNA, and promoter methylation layers. We computed gene-wise distance correlations to capture non-linear regulatory dependencies and clustered these profiles to identify distinct Regulatory Rewiring Modules (RRMs). By projecting individual patient profiles onto these modules, we quantified patient-specific module disruption and CNA-versus-methylation dominance shifts. We adjusted these metrics for global genomic and microenvironmental confounders using Random Forest regression and evaluated their prognostic utility in an independent validation set using cause-specific Cox proportional hazards models.

Our analysis identified twelve distinct RRMs, including an *ERBB2*-associated network (RRM_11) that exhibits strong copy-number-driven expression control, particularly in HER2-enriched subtypes such as Integrative Cluster 5 (IntClust 5). Unadjusted cumulative incidence estimates suggested that high disruption and dominance shifts in RRM_11 are associated with substantial disease-specific mortality in these subgroups (Table 1.3).

Despite these distinct biological architectures, rigorous survival analysis revealed that neither the adjusted disruption nor the dominance shift of RRM_11 provided statistically significant independent prognostic value for disease-specific mortality when accounting for established clinico-pathological variables. The effect sizes of these module metrics were comparable to those of a genomic-location-matched null module. These findings indicate that while multi-omic regulatory rewiring can be robustly quantified, current patient-level metrics of these localized disruptions do not yield prognostic information beyond established clinical taxonomies and global genomic profiles.

2 Methods

2.1 Data assembly and preprocessing

We utilized clinical, transcriptomic, copy-number alteration (CNA), and promoter methylation data from the METABRIC cohort. The analysis was restricted to 1,417 primary breast tumor samples possessing complete data across all four molecular and clinical layers. To ensure independent module definition and prevent statistical circularity, the cohort was partitioned into a Discovery Set comprising Cohorts 1, 2, and 3 ($n = 1,107$) and a Validation Set comprising Cohorts 4 and 5 ($n = 310$).

For all samples, we computed three global covariates to capture macroscopic tumor characteristics: macroscopic copy-number burden, defined as the fraction of genes with non-zero CNA states; global transcriptional noise, calculated as the variance across the top 1,000 highly variable genes; and a stromal/immune signature, derived as the first principal component of canonical marker genes (e.g., PTPRC, CD3D, COL1A1) via principal component analysis. Missing methylation beta values were imputed using a k-nearest neighbors approach based on expression profiles, followed by a logit transformation to convert beta values to M-values for downstream analysis.

2.2 Subtype-specific distance correlation profiling

To capture non-linear regulatory dependencies, we computed distance correlations (*dcor*) strictly within the Discovery Set. For the top 1,000 highly variable genes, we calculated the distance correlation between $\log_2$ expression and discrete CNA state, and separately between $\log_2$ expression and methylation M-value. These calculations were performed independently within each Integrative Cluster (IntClust) category. Because distance correlation is strictly non-negative, we assigned directionality by multiplying each *dcor* value by the sign of the corresponding standard Pearson correlation coefficient.

The signed *dcor* values were then subjected to a Fisher Z-transformation to place them on comparable scales. For each gene within each subtype, a Regulatory Dominance Score was computed as the difference between its transformed signed CNA *dcor* and its transformed signed methylation *dcor*. Finally, the transformed CNA *dcor*, methylation *dcor*, and Dominance Scores were standardized via Z-scoring across genes within each IntClust group.

2.3 Identification of regulatory rewiring modules

The standardized coupling profiles from the Discovery Set were concatenated into a feature matrix. We applied Uniform Manifold Approximation and Projection (UMAP) for dimensionality reduction, followed by Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) to cluster genes into distinct Regulatory Rewiring Modules (RRMs) based on their cross-layer uncoupling mechanisms. Hypergeometric enrichment analysis was performed on the resulting RRMs using standard gene sets to provide biological annotation.

To establish a rigorous background control for signal-to-noise quantification, we constructed a Null Module. This module consisted of genes matched to the genes of RRM_11 (an *ERBB2*-associated network prioritized for its biological relevance) based on baseline expression variance, mean expression, genomic location, and baseline CNA and methylation characteristics.

2.4 Patient-level dominance shift and confounder adjustment

For each patient in both the Discovery and Validation sets, we quantified the raw disruption score of each RRM and the Null Module by calculating the module-specific expression variance. Additionally, we calculated a patient-specific Dominance Shift by projecting individual expression, CNA, and methylation profiles onto the ideal gene-wise coupling weights derived from the Discovery Set, preserving the directionality of CNA versus methylation dominance. The Discovery cohort median for each module was subtracted to capture patient-specific rewiring.

To isolate targeted regulatory failure from global genomic and microenvironmental noise, we trained a Random Forest regressor on the Discovery Set to predict the raw disruption score and Dominance Shift using the global copy-number burden, stromal/immune signature, global transcriptional noise, tumor size, and cellularity. A parallel Random Forest model was trained for the Null Module using the same covariates. These models were applied to both sets to estimate the expected metrics. The residuals from these predictions were extracted as the Adjusted Disruption Score and Adjusted Dominance Shift, which were subsequently Z-scored within each cohort.

2.5 Competing risks and survival analysis

Clinical data were processed to establish a competing risks framework, defining two distinct event indicators: disease-specific death and death from other causes. Age at diagnosis was discretized into quartiles to serve as a strict categorical stratification variable, addressing the severe age confounding between the competing risks.

In the Validation Set, we employed the Aalen-Johansen estimator to compute the cumulative incidence of disease-specific death, explicitly accounting for the competing risk of death from other causes. To visually assess the prognostic signal, these cumulative incidence curves were stratified by high versus low Adjusted Disruption Score and Adjusted Dominance Shift specifically for the RRM_11 feature within key IntClust subgroups.

To rigorously evaluate prognostic utility, we fit cause-specific Cox proportional hazards models exclusively on the Validation Set ($n = 293$ complete cases). A Base model was first fitted using IntClust, histological grade, tumor size, and the number of positive lymph nodes, stratified by age quartiles. Full models were then constructed by adding the Adjusted Disruption Score and its interaction with IntClust, and separately, the Adjusted Dominance Shift and its interaction with IntClust. These models were evaluated for both disease-specific mortality and other-cause mortality, with the latter serving as a negative control to ensure the metrics did not merely capture general patient frailty.

Finally, to quantify the signal-to-noise ratio, parallel Full cause-specific Cox models were fitted using the Adjusted Disruption Score and Adjusted Dominance Shift of the Null Module. The prognostic independence of the RRM_11 metrics was assessed by directly comparing their hazard ratios and confidence intervals against those of the Null Module.

3 Results

3.1 Identification of multi-omic regulatory rewiring modules

To systematically map the joint molecular organization of breast cancer, we analyzed the distance correlation profiles between gene expression and both copy-number alterations (CNA) and promoter methylation for the top 1,000 highly variable genes within each integrative cluster (IntClust) in the Discovery Set ($n = 1,107$). Dimensionality reduction using UMAP followed by HDBSCAN clustering revealed that genes do not vary independently but organize into distinct biological architectures within the multi-omic space. This approach identified 12 distinct Regulatory Rewiring Modules (RRMs) characterized by coordinated patterns of transcriptomic, CNA, and promoter methylation coupling (Figure 1).

3.2 Subtype-specific coupling architectures in RRM_11

Among the identified modules, we prioritized RRM_11, which contains the *ERBB2* driver gene, to contrast targeted regulatory failure with global genomic instability. Analysis of the multi-omic regulatory coupling profiles demonstrated that RRM_11 genes exhibit a characteristic dominance of copy-number-driven expression control over methylation-mediated silencing (Figure 2). This CNA dominance was particularly pronounced in HER2-enriched subtypes, such as IntClust 5, where the Dominance Score (CNA-coupling minus methylation-coupling) was consistently elevated. In contrast, a genomic-location-matched Null module displayed lower overall coupling magnitudes and higher variance in regulatory drivers across the same subtypes, confirming that RRM_11 captures a specific, localized regulatory architecture.

3.3 Cumulative incidence of disease-specific mortality

We next evaluated the prognostic implications of RRM_11 disruption and dominance shift in the independent Validation Set ($n = 310$). Using the Aalen-Johansen estimator to account for the competing risk of death from other causes, we observed visible stratification in the cumulative incidence of disease-specific mortality based on patient-level module metrics (Figure 3).

As detailed in Table 1.3, high RRM_11 metrics specifically identified a high-risk subgroup within IntClust 5. Patients in this subgroup with a high RRM_11 Dominance Shift experienced a 10-year disease-specific mortality rate exceeding 50% (CIF = 0.555), compared to the global cohort baseline.

3.4 Cause-specific prognostic evaluation and signal-to-noise quantification

To rigorously determine whether these localized regulatory rewiring signals provide independent prognostic value beyond established clinical taxonomies, we fitted cause-specific Cox proportional hazards models on the complete cases of the Validation Set ($n = 293$). These models were adjusted for baseline clinico-pathological variables (histological grade, tumor size, number of positive lymph nodes, and IntClust) and strictly stratified by age quartiles to isolate tumor-intrinsic lethality from age-related confounding (Figure 4).

Despite the visible stratification in the unadjusted cumulative incidence curves, multivariate modeling revealed that neither the Adjusted Disruption nor the Adjusted Dominance Shift of RRM_11 reached nominal statistical significance for disease-specific mortality (Disruption: HR 0.95, 95% CI 0.77–1.18, $p = 0.655$; Dominance Shift: HR 1.13, 95% CI 0.91–1.41, $p = 0.259$). Furthermore, evaluation of other-cause mortality as a negative control confirmed no significant association (Disruption: HR 0.95, 95% CI 0.77–1.17, $p = 0.614$; Dominance Shift: HR 1.04, 95% CI 0.83–1.32, $p = 0.715$), indicating that these metrics do not merely reflect general patient frailty.

Crucially, the effect sizes observed for RRM_11 were comparable to those of the genomic-location-matched Null module (Null Disruption: HR 1.10, 95% CI 0.89–1.35, $p = 0.366$; Null Dominance Shift: HR

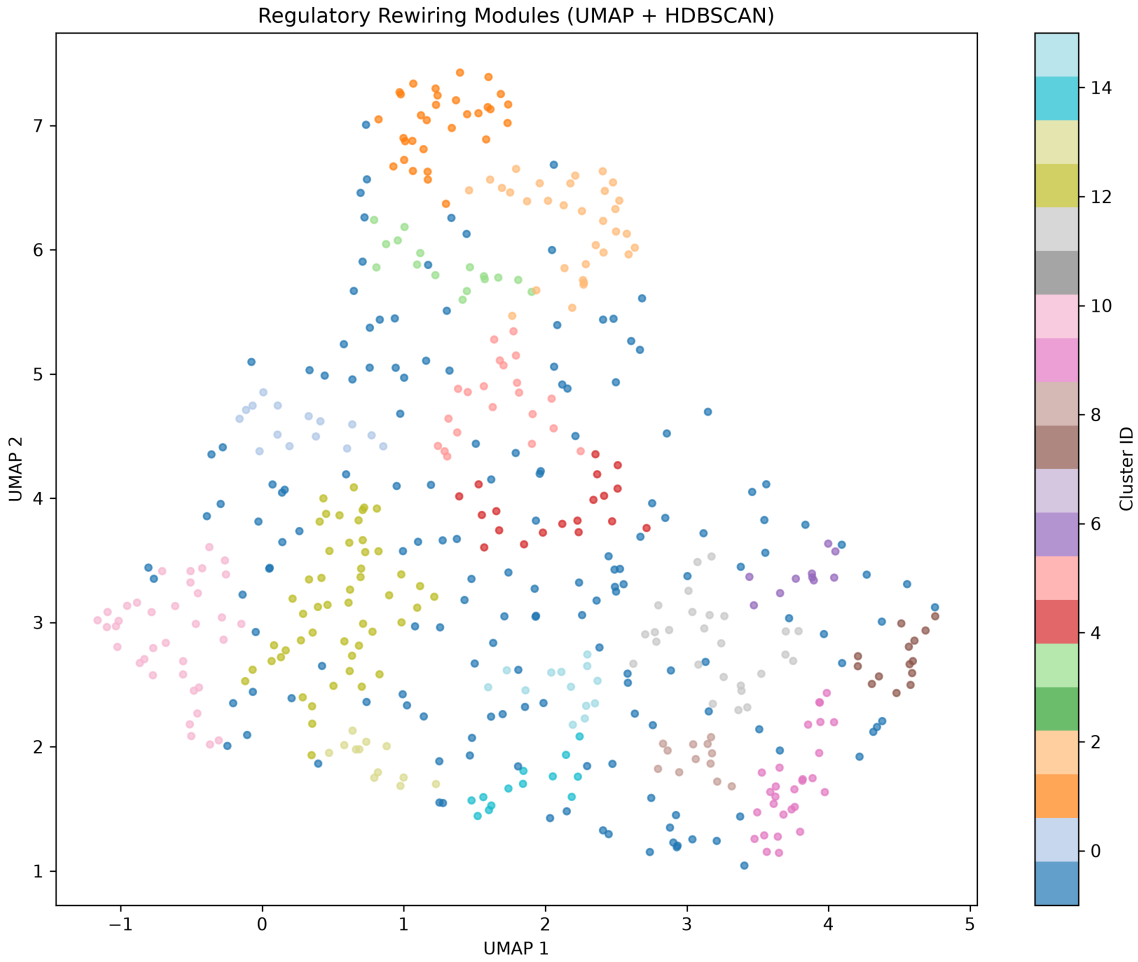


Figure 1: UMAP projection and HDBSCAN clustering of highly variable genes based on their multi-omic coupling profiles. Points represent individual genes mapped into a two-dimensional space and colored by their assigned Cluster ID, revealing distinct Regulatory Rewiring Modules (RRMs) characterized by coordinated patterns of transcriptomic, copy-number alteration (CNA), and promoter methylation coupling. The distinct grouping demonstrates that genes organize into specific biological architectures within the multi-omic space rather than varying independently. This dimensionality reduction identified 12 RRMs, including the *ERBB2*-containing RRM_11, which was prioritized for subsequent cause-specific survival evaluation to contrast targeted regulatory failure with global genomic instability.

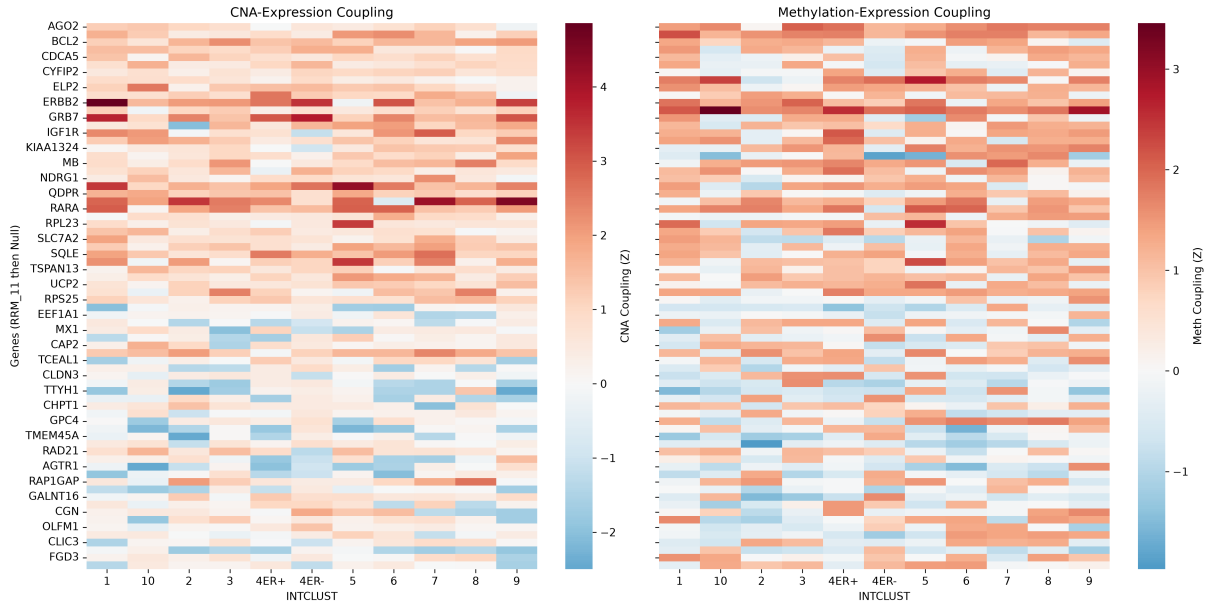


Figure 2: Heatmaps displaying the multi-omic regulatory coupling profiles for genes in RRM_11 and a genomic-location-matched Null module across `INTCLUST` subtypes. The left and right panels show copy-number alteration (CNA)-expression and methylation-expression coupling, respectively, measured as Z-scored distance correlations. RRM_11 genes (top rows, including the driver *ERBB2*) exhibit strong CNA-driven expression control, particularly in `INTCLUST` 5, alongside minimal coupling with promoter methylation. Conversely, Null module genes (bottom rows) display lower overall coupling magnitudes and higher variance in their regulatory drivers. This visualizes the characteristic dominance of copy-number over methylation-mediated regulation in RRM_11, although these distinct multi-omic architectures ultimately yielded prognostic effect sizes comparable to the Null module in downstream cause-specific survival evaluations, suggesting the localized regulatory rewiring signal resides within the measurement noise floor.

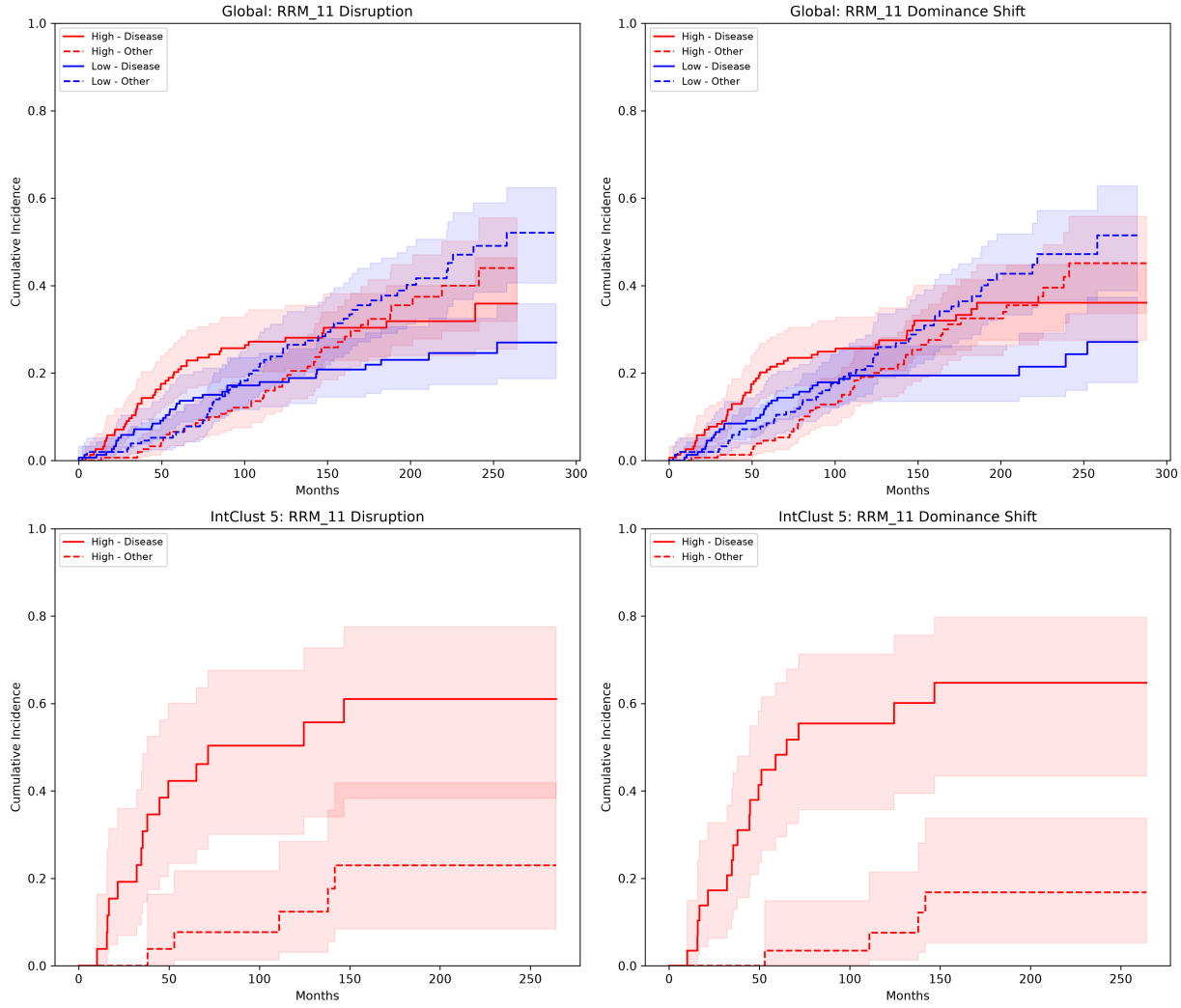


Figure 3: Cumulative incidence of disease-specific (solid lines) and other-cause (dashed lines) mortality stratified by RRM_11 metrics in the Validation Set. The top panels display the global cohort, comparing high (red) and low (blue) patient-level scores for RRM_11 Adjusted Disruption (left) and Dominance Shift (right) using the Aalen-Johansen estimator. The bottom panels focus on the `INTCLUST` 5 subgroup, revealing that high RRM_11 metrics identify a specific population with a 10-year disease-specific mortality exceeding 50%. The lack of association with other-cause death confirms these scores do not merely reflect general age-related frailty. However, despite the visible stratification in these cumulative incidence curves, multivariate Cox modeling demonstrates that these localized regulatory rewiring signals do not provide significant independent prognostic value beyond existing clinical taxonomies and baseline clinico-pathological variables.

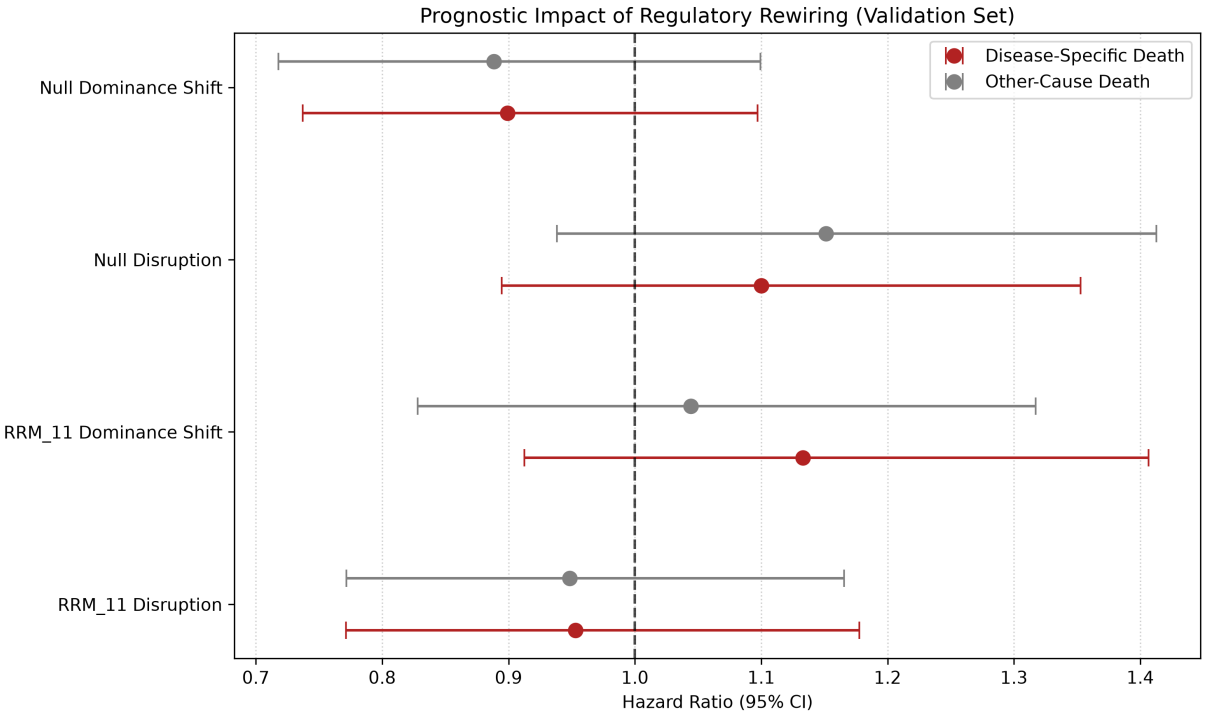


Figure 4: Forest plot illustrating the prognostic impact of regulatory rewiring metrics in the validation set (n=293). Cause-specific Cox proportional hazards models were used to estimate hazard ratios and 95% confidence intervals for RRM_11 and a genomic-location-matched Null module across two patient-level metrics: Adjusted Disruption and Dominance Shift. Results are shown for disease-specific death (red) and other-cause death (grey, serving as a negative control). All confidence intervals cross 1.0, indicating that the metrics do not reach nominal significance. The comparable effect sizes between RRM_11 and the Null module demonstrate that these localized regulatory rewiring signals reside within the measurement noise floor and do not provide independent prognostic value beyond standard clinico-pathological variables.

0.90, 95% CI 0.74–1.10, $p = 0.295$). The overlapping confidence intervals between the true biological module and the background control demonstrate that the localized regulatory rewiring signals detected by our patient-level quantifications reside within the measurement noise floor. Consequently, while RRMs represent distinct and quantifiable biological architectures, their current implementation does not yield prognostic information beyond what is already captured by global genomic profiles and standard clinical variables.

4 Conclusions

4.1 Biological architectures versus prognostic utility

This study systematically investigated whether localized disruptions in joint epigenetic and genetic regulatory networks provide independent prognostic value in breast cancer. By integrating transcriptomic, copy-number alteration (CNA), and promoter methylation data from 1,417 primary tumors in the METABRIC cohort, we successfully mapped distinct Regulatory Rewiring Modules (RRMs). Our multi-omic coupling analysis confirmed that genes do not uncouple independently; rather, they organize into specific biological architectures. Most notably, RRM_11—an *ERBB2*-associated network—demonstrated a clear, subtype-specific dominance of copy-number-driven expression control over methylation-mediated silencing, particularly within HER2-enriched tumors (IntClust 5).

At the unadjusted patient level, these localized regulatory shifts appeared to stratify patient outcomes. High disruption and dominance shifts in RRM_11 identified a high-risk subgroup within IntClust 5, where the 10-year cumulative incidence of disease-specific mortality exceeded 50% (Table 1.3).

However, the apparent prognostic utility of these metrics dissolved under rigorous multivariate evaluation. When adjusted for global genomic confounders—such as macroscopic copy-number burden and global transcriptional noise—and evaluated within a cause-specific Cox proportional hazards framework stratified by age, neither the adjusted disruption nor the dominance shift of RRM_11 provided statistically significant independent prognostic value for disease-specific mortality.

4.2 The necessity of rigorous multi-omic controls

A central strength of this study is the implementation of strict methodological controls, which contextualize the negative prognostic findings. By evaluating other-cause mortality as a negative control, we confirmed that the RRM metrics were not merely acting as surrogates for general patient frailty or age-related decline. More importantly, the inclusion of a genomic-location-matched Null module provided a critical baseline for signal-to-noise quantification. The hazard ratios associated with RRM_11 were statistically indistinguishable from those of the Null module. This overlap indicates that while the biological architecture of RRM_11 is distinct in the multi-omic feature space, the patient-level quantification of its disruption resides entirely within the measurement noise floor.

4.3 Concluding remarks

Our findings highlight a critical divergence between biological detectability and clinical utility in multi-omic biomarker discovery. While non-linear regulatory dependencies can be robustly quantified to reveal how tumor-intrinsic networks are locally rewired, current patient-level metrics of these disruptions do not exceed the predictive power of established clinical taxonomies and global genomic profiles. Future efforts to translate multi-omic regulatory rewiring into clinical biomarkers must prioritize high-resolution, single-cell multi-omics to overcome bulk-tissue noise floors, and must continue to employ rigorous background controls to prevent the misinterpretation of multi-omic variance as independent prognostic signal.