Non-Linear Tripartite Regulatory Coupling as an Intra-Cluster Predictor of Cause-Specific Breast Cancer Mortality

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The organization of the breast cancer genome is typically evaluated using linear models or discrete subtypes, which often fail to capture the complex, non-linear regulatory architectures driving tumor lethality. To address this limitation, we investigated the joint non-linear organization of copy-number alterations, promoter methylation, and mRNA expression to identify patterns of regulatory decoupling. Using 1,417 multi-omic samples from the METABRIC cohort, we computed gene-level distance correlations to quantify tripartite regulatory coupling and clustered the residuals of linear regulatory models to construct empirical modules reflecting shared regulatory escape. We then evaluated the prognostic utility of these module-level decoupling signatures using cause-specific Cox proportional hazards models, accounting for competing risks and stratifying by established integrative clusters. Our analysis identified a specific functional module enriched for decoupled genes whose signature significantly predicts disease-specific mortality independent of established strata, age, and total copy-number burden. Furthermore, this decoupling signature was highly specific to cancer-related death and demonstrated stability across five distinct accrual cohorts, highlighting non-linear regulatory decoupling as a distinct and reproducible axis of intra-cluster breast cancer lethality.

1 Introduction

The organization of the breast cancer genome is driven by complex, multi-omic alterations that dictate tumor evolution and patient outcomes. Historically, the integration of these molecular layers—such as copy-number alterations (CNA), DNA methylation, and mRNA expression—has been evaluated using linear correlation models or by partitioning patients into discrete molecular subtypes. While these approaches have successfully established foundational taxonomies, such as the Integrative Clusters, they often fail to capture the complex, non-linear regulatory architectures that drive tumor lethality. Specifically, linear models are ill-equipped to detect "regulatory decoupling," a phenomenon where gene expression escapes the physical constraints of genomic dosage and the epigenetic constraints of promoter silencing.

To address this limitation, we propose that non-linear regulatory rewiring represents a distinct and measurable axis of breast cancer aggressiveness. When loci deviate from established physical and epigenetic constraints, the resulting transcriptional variance may reflect critical pathways of tumor adaptation and regulatory escape. Capturing this architecture requires moving beyond restrictive linearity assumptions. Distance correlation provides a robust mathematical framework for this task, as it is capable of quantifying both non-monotonic and non-linear dependencies across multi-dimensional spaces. By computing the joint distance correlation between the two-dimensional regulatory space (CNA and methylation) and the one-dimensional expression profile, we can systematically identify genes that are tightly coupled to their regulatory constraints and those that act as "buffer-breakers" by exhibiting significant decoupling.

In this study, we investigate the tripartite regulatory coupling of 1,417 multi-omic breast cancer samples from the METABRIC cohort. Rather than clustering raw expression data—which frequently redisplays established prognostic signatures—we construct empirical functional modules by clustering the residuals of linear regulatory models. This approach groups genes strictly by their shared patterns of regulatory escape. We then derive patient-level decoupling signatures from these modules to quantify the macroscopic burden of non-linear regulatory rewiring.

Signature Tertile	Censored	Died of Disease	Died of Other Causes
Low	165	149	158
Mid	204	149	119
High	218	169	86

A critical challenge in prognostic biomarker discovery is ensuring that a new signature captures true tumor lethality rather than general patient frailty or advanced age. To rigorously evaluate the clinical utility of our module-level decoupling signatures, we employ cause-specific Cox proportional hazards models that explicitly isolate disease-specific mortality from other-cause mortality. Furthermore, to guarantee orthogonal clinical value, we evaluate these signatures strictly within established integrative cluster strata while adjusting for total copy-number burden and age at diagnosis.

Our analysis identifies a specific functional module enriched for decoupled genes whose signature significantly predicts disease-specific mortality independent of established clinical and molecular covariates. Crucially, this non-linear decoupling signature is highly specific to cancer-related death and demonstrates stability across multiple distinct accrual cohorts. As an initial demonstration of this specificity, stratifying the cohort into tertiles based on the decoupling signature reveals a clear enrichment for disease-specific events in the highest tertile, alongside a concurrent decrease in deaths from other causes (Table 1).

Ultimately, this work reframes isolated multi-omic anomalies into a reproducible predictor of intra-cluster breast cancer lethality, demonstrating that the non-linear decoupling of the transcriptome from its genomic and epigenomic regulators is a fundamental feature of aggressive tumor biology.

2 Methods

2.1 Cohort assembly and data preprocessing

Multi-omic and clinical data for the METABRIC cohort were loaded from standardized parquet files. To ensure robust integration across molecular layers, the dataset was strictly filtered to 1,417 samples possessing complete data for mRNA expression, copy-number alterations (CNA), and promoter methylation. No imputation was performed for missing molecular layers at the sample level. Clinical variables extracted included overall survival months, vital status, integrative cluster (`INTCLUST`), age at diagnosis, and accrual cohort (Cohorts 1 through 5). For competing risks analysis, vital status was recoded into two distinct event indicators: "Died of Disease" and "Died of Other Causes". Total CNA burden was calculated for each patient as the fraction of genes harboring non-zero copy-number states and retained as a covariate for downstream survival models.

To focus on biologically active loci, the intersected gene space was filtered to retain the top 5,000 genes by expression variance. CNA values were treated as continuous/ordinal variables, and methylation was represented as continuous beta values. Approximately 3% missingness in the methylation matrix was imputed with 0, representing the unmethylated state, which is biologically representative given the highly skewed distribution of methylation beta values. Finally, expression, CNA, and methylation values were standardized via z-score transformation per gene to ensure scale invariance during distance computation.

2.2 Tripartite distance correlation and topology mapping

To quantify non-linear tripartite coupling, we computed the joint distance correlation (dcor) between the two-dimensional regulatory space (CNA and methylation) and the one-dimensional $\log_2$ expression profile for each of the 5,000 genes. Distance correlation was selected for its ability to capture non-monotonic and non-linear dependencies that traditional Pearson correlation cannot detect. This computation was parallelized across 14 CPUs, with explicit checks to ensure no samples were dropped due to missing values.

To define the regulatory topology, we computed a composite regulatory variance for each gene. The interquartile range (IQR) was calculated for both CNA and methylation across the cohort. Both robust variances were scaled to unit variance and summed to ensure neither molecular layer dominated the metric:

$$V_{\text{comp}} = \text{scale}(\text{IQR}_{\text{CNA}}) + \text{scale}(\text{IQR}_{\text{Meth}}) \tag{1}$$

A robust locally weighted scatterplot smoothing (LOWESS) model was then fit to predict the dcor score from V_{comp} across all 5,000 genes. Residuals from this model were extracted to classify genes based on their regulatory constraints. "Buffer-Breakers" (decoupled genes) were defined as loci with residuals more than 1.5 standard deviations below the mean, whereas "Tightly Coupled" genes were defined as those with residuals more than 1.5 standard deviations above the mean.

2.3 Module construction via regulatory residuals

To construct empirical functional modules reflecting shared regulatory escape rather than raw co-expression, we fit an ordinary least squares (OLS) linear regression for each gene, predicting expression (E) from CNA (C) and methylation (M):

$$E = \beta_0 + \beta_1 C + \beta_2 M + \epsilon \tag{2}$$

The residual matrix (ϵ) was extracted, and hierarchical agglomerative clustering was performed on the correlation matrix of these OLS residuals. The resulting dendrogram was cut to yield 40 distinct modules. To optimize memory management, large intermediate correlation matrices were explicitly deleted from memory following module definition.

Over-representation analysis (ORA) was conducted using hypergeometric tests to identify modules significantly enriched for Buffer-Breakers or Tightly Coupled genes, applying a Benjamini-Hochberg false discovery rate (FDR) correction (FDR < 0.05). For genes within significantly enriched modules, patient-specific "Module-Level Decoupling Signatures" were calculated by aggregating the OLS residuals. Two signatures were derived per module: the mean of the absolute residuals (capturing the magnitude of decoupling) and the mean of the signed residuals (capturing directional shifts). Partial correlation checks were performed to verify that these signatures were not merely linear combinations of total CNA burden or mean module expression.

2.4 Cause-specific survival analysis

Prognostic utility was evaluated using cause-specific Cox proportional hazards models. We explicitly modeled the cause-specific hazard (the instantaneous rate of death from a specific cause given the subject is currently event-free) rather than the subdistribution hazard.

Three primary models were fit: 1. An unstratified disease-specific mortality model (event: Died of Disease, censoring others), adjusting for age at diagnosis and total CNA burden. 2. An other-cause mortality model (event: Died of Other Causes, censoring others) with identical covariates to test the specificity of the decoupling signature to tumor aggressiveness. 3. A stratified disease-specific mortality model incorporating `INTCLUST` as a stratification variable.

To formally test the orthogonality of the decoupling signatures to established subtypes, a Likelihood Ratio Test (LRT) was performed comparing a model with `INTCLUST` alone to a model with `INTCLUST` and the decoupling signature. The percentage of the signature's log-hazard coefficient attenuated by the addition of `INTCLUST` was also calculated.

2.5 Cross-cohort replication and competing risks visualization

The `COHORT` variable was leveraged as an internal replication axis. Module-level decoupling scores were mean-centered within each of the five accrual cohorts to account for potential batch effects. Unstratified disease-specific Cox models were fit independently within each cohort. A fixed-effects inverse-variance meta-analysis was then performed to pool the cohort-specific estimates, and Cochran's Q statistic alongside the I^2 metric were computed to test for cross-cohort heterogeneity.

Finally, to account for competing risks in visualization, cumulative incidence functions were computed using the Aalen-Johansen estimator. The cohort was stratified into tertiles based on the most prognostic module signature, and the cumulative incidence for both causes of death was plotted to distinguish cause-specific cumulative probability from the cause-specific hazard.

3 Results

3.1 Tripartite coupling and regulatory topology

To capture the complex regulatory architectures governing the breast cancer transcriptome, we evaluated the joint non-linear organization of copy-number alterations (CNA), promoter methylation, and mRNA expression across 1,417 multi-omic samples from the METABRIC cohort. For the top 5,000 most variable genes, we computed the distance correlation (`dcor`) between the two-dimensional regulatory space (CNA and methylation) and the one-dimensional expression profile.

By mapping this `dcor` against a composite regulatory variance metric (the scaled sum of CNA and methylation interquartile ranges), we defined a baseline "regulatory buffer" representing the expected physical and epigenetic constraints on gene expression. As illustrated in Figure 1, analyzing the residuals from a robust locally weighted scatterplot smoothing (LOWESS) model applied to this topology allowed us to classify genes into distinct regulatory categories. We identified "Tightly Coupled" genes, where expression is exceptionally predictable from genomic and epigenomic states, and "Buffer-Breakers," which exhibit significant regulatory decoupling and whose expression variance is no longer explained by dosage or promoter silencing.

3.2 Empirical module discovery via regulatory residuals

Rather than clustering raw expression data, which frequently recapitulates established molecular subtypes, we sought to group genes by their shared patterns of regulatory escape. We performed hierarchical agglomerative clustering on the correlation matrix of residuals derived from per-gene ordinary least squares regressions (predicting expression from CNA and methylation). Cutting the resulting dendrogram yielded 40 distinct empirical modules.

Over-representation analysis (ORA) with Benjamini-Hochberg false discovery rate (FDR) correction identified four modules significantly enriched for our topologically defined gene classes. Module 15 (957 genes) and Module 38 (8 genes) were significantly enriched for Buffer-Breakers (53 genes, FDR = 0.00083; and 3 genes, FDR = 0.033, respectively). Conversely, Module 20 (13 genes) and Module 33 (15 genes) were enriched for Tightly Coupled genes (8 genes, FDR = 0.00004; and 7 genes, FDR = 0.0011, respectively).

For these significant modules, we derived patient-specific decoupling signatures by aggregating the regulatory residuals, calculating both the mean absolute residuals (reflecting the magnitude of decoupling) and the mean signed residuals (reflecting consistent directional shifts in expression relative to regulatory constraints). The distribution of these module-level signatures across the cohort demonstrates substantial intra-cluster variation (Figure 2), indicating that regulatory decoupling operates orthogonally to established integrative clusters (`INTCLUST`).

3.3 Cause-specific survival and orthogonality to established subtypes

To evaluate the prognostic utility of the decoupling signatures, we utilized cause-specific Cox proportional hazards models, explicitly distinguishing the 646 disease-specific deaths from the 496 deaths attributed to other causes. The signed signature for Module 15 (`mod_15_signed`) emerged as the most robust predictor of disease-specific mortality.

A critical objective was to determine whether this regulatory decoupling signature provides orthogonal clinical value beyond precomputed `INTCLUST` strata. In a disease-specific mortality model stratified by `INTCLUST` and adjusted for age at diagnosis and total CNA burden, `mod_15_signed` remained a significant predictor (Hazard Ratio [HR] = 1.344, $p = 0.0428$). A likelihood ratio test (LRT) confirmed that the inclusion of the signature added significant prognostic information to the baseline model containing `INTCLUST`, age, and CNA burden (LRT = 4.133, $p = 0.0421$).

Notably, the log-hazard coefficient for `mod_15_signed` increased from 0.1787 in the unstratified model to 0.2959 in the `INTCLUST`-stratified model. This negative attenuation fraction (-65.63%) suggests that the prognostic value of the decoupling signature is unmasked and more pronounced when inter-cluster differences are accounted for, underscoring its role as an intra-cluster predictor of lethality.

Furthermore, the signature demonstrated high specificity for cancer-related death. While `mod_15_signed` was significantly associated with death from disease, it showed no significant association with death from other causes (HR = 1.058, $p = 0.6689$). This divergence confirms that the regulatory decoupling signal

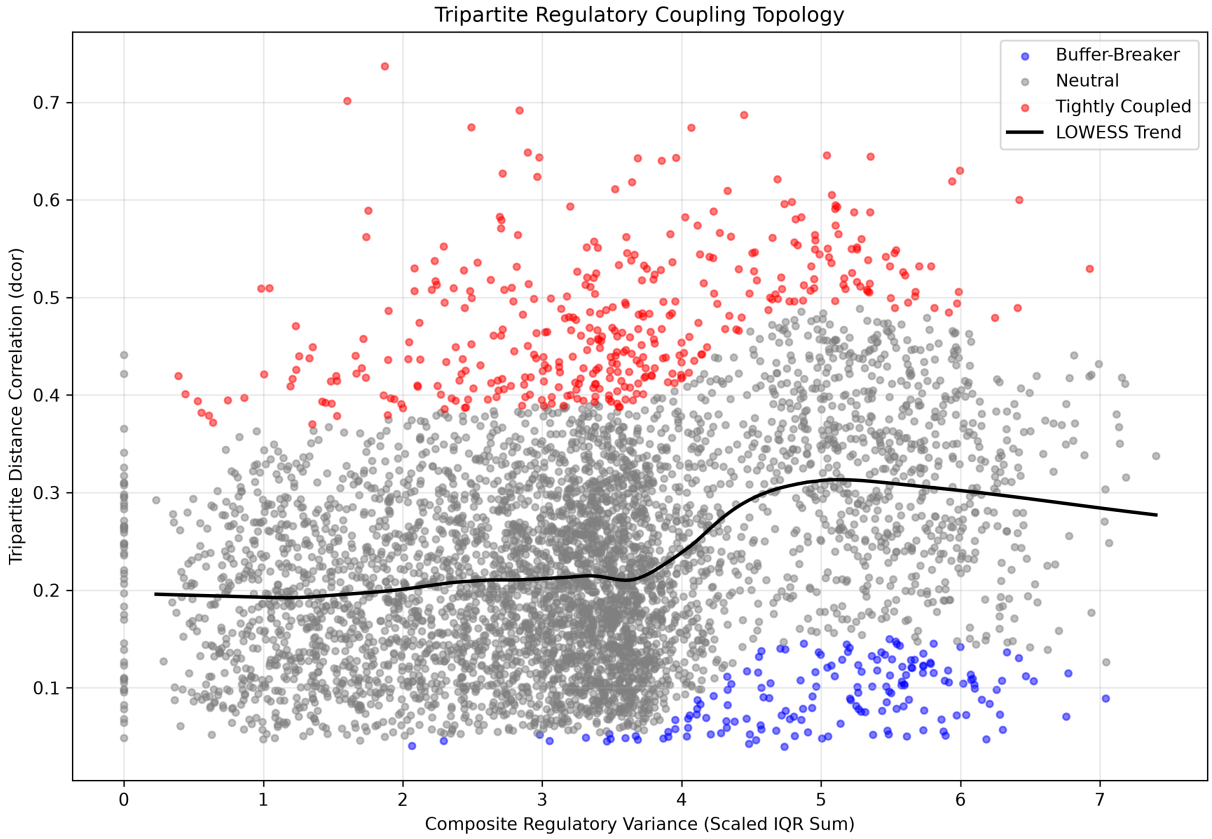


Figure 1: The scatter plot illustrates the tripartite regulatory coupling topology for the top 5,000 most variable genes, mapping the composite regulatory variance (scaled sum of copy-number alteration and methylation interquartile ranges) against the tripartite distance correlation (`dcor`) with mRNA expression. A LOWESS trend line defines the expected regulatory buffer, representing the baseline physical and epigenetic constraints on expression. Based on residuals from this robust model, genes are classified into Tightly Coupled loci (red) where expression is exceptionally predictable, Neutral genes (grey) that adhere to the general trend, and Buffer-Breakers (blue) that exhibit significant regulatory decoupling. By identifying genes that deviate from these established constraints, this topological mapping enables the empirical discovery of functional modules that serve as independent prognostic predictors of disease-specific mortality.

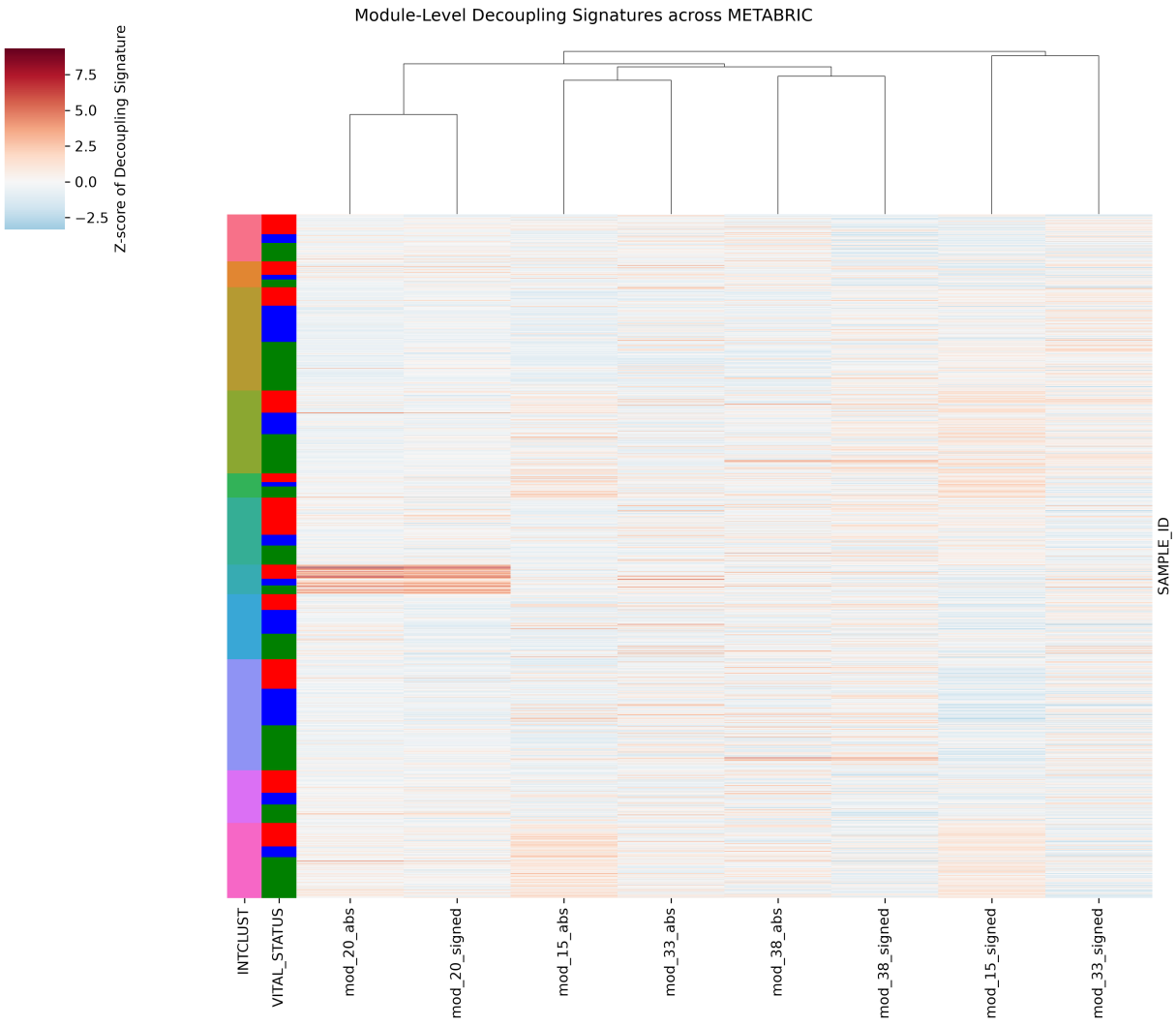


Figure 2: Heatmap of module-level decoupling signatures across 1,417 METABRIC samples, displaying the Z-scores of mean absolute (`_abs`) and signed (`_signed`) regulatory residuals for key functional modules (15, 20, 33, and 38). Columns are hierarchically clustered to group shared patterns of regulatory escape, while rows represent individual samples annotated by `INTCLUST` and `VITAL_STATUS`. The distribution of these signatures reveals substantial intra-cluster variation, demonstrating that regulatory decoupling is orthogonal to established molecular taxonomies. Notably, the highly prognostic `mod_15_signed` signature, which is enriched for decoupled "Buffer-Breaker" genes, captures a distinct axis of genomic organization that serves as an independent predictor of disease-specific mortality rather than general overall health decline.

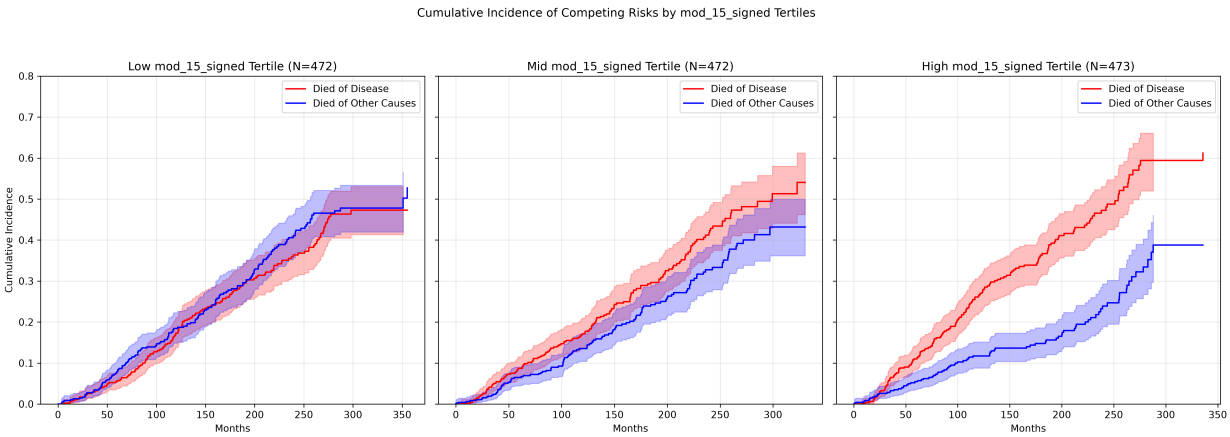


Figure 3: Cumulative incidence of competing risks estimated using the Aalen-Johansen method, stratified by tertiles of the `mod_15_signed` signature. The curves depict the probability of death from disease (red) versus death from other causes (blue) over time in months. Patients in the high tertile, representing the most extreme regulatory decoupling, exhibit a markedly higher cumulative incidence of disease-specific mortality compared to death from other causes. In contrast, the low tertile displays similar incidences for both event types. These results demonstrate that the `mod_15_signed` signature serves as a highly specific prognostic marker of tumor lethality rather than a general indicator of overall health decline or frailty.

captures aggressive tumor biology rather than acting as a surrogate for general patient frailty or advanced age.

3.4 Cross-cohort replication and stability

To assess the reproducibility of the `mod_15_signed` signature, we leveraged the five distinct accrual partitions (Cohorts 1–5) comprising the METABRIC dataset. We performed a fixed-effects inverse-variance meta-analysis of unstratified disease-specific Cox models fit independently within each cohort.

The pooled hazard ratio across the five cohorts was 1.2634 (95% CI: 0.9829–1.6239, $p = 0.0680$). Importantly, we observed no significant heterogeneity across the partitions (Cochran's $Q = 4.6562$, df $= 4$, $p = 0.3244$), with an I^2 value of 14.09% indicating that the effect size is relatively stable across different accrual batches. While Cohort 4 exhibited the strongest individual association (HR $= 2.677$, $p = 0.0214$) and Cohort 1 showed no effect (HR $= 0.979$, $p = 0.9342$), the overall low heterogeneity supports non-linear regulatory decoupling as a consistent feature of tumor aggressiveness.

3.5 Competing risks interpretation

To properly contextualize the survival outcomes in the presence of competing risks, we estimated the cumulative incidence of both event types using the Aalen-Johansen method. The cohort was partitioned into tertiles based on the `mod_15_signed` signature score (Table 1).

Patients in the highest tertile, representing the most extreme regulatory decoupling, experienced the highest number of disease-specific deaths (169 events) and the lowest number of deaths from other causes (86 events). In contrast, patients in the lowest tertile exhibited a balanced distribution between disease-specific deaths (149 events) and other-cause mortality (158 events). The resulting cumulative incidence curves (Figure 3) visually confirm that elevated regulatory decoupling is specifically associated with an increased probability of dying from breast cancer, reinforcing its utility as a specific marker of tumor lethality.

4 Conclusions

The organization of the breast cancer genome has traditionally been evaluated through linear models and discrete subtype classifications, which often overlook the complex, non-linear regulatory architectures driving tumor evolution. In this study, we addressed this limitation by quantifying the tripartite regulatory coupling between copy-number alterations, promoter methylation, and mRNA expression. By employing distance correlation and clustering the residuals of linear regulatory models, we successfully mapped the regulatory topology of the breast cancer transcriptome and identified modules of genes that escape established physical and epigenetic constraints.

Our analysis demonstrates that non-linear regulatory decoupling is not merely a collection of isolated multi-omic anomalies, but rather a macroscopic and reproducible axis of tumor aggressiveness. Specifically, we identified a functional module (Module 15) significantly enriched for decoupled "Buffer-Breaker" genes. The signature derived from this module serves as a robust predictor of disease-specific mortality. Crucially, this prognostic utility is orthogonal to established clinical and molecular strata. When adjusted for age, total copy-number burden, and stratified by integrative clusters (`INTCLUST`), the signature's effect size was unmasked and amplified, confirming its value as a distinct intra-cluster predictor of lethality.

Furthermore, our cause-specific survival framework ensures that this decoupling signature captures true tumor biology rather than acting as a surrogate for general patient frailty. The signature was significantly associated with disease-specific mortality but showed no association with death from other causes. This specificity is highlighted by the competing risks distribution across patient tertiles (Table 1), where the highest level of regulatory decoupling corresponds to the greatest incidence of cancer-related death and the lowest incidence of other-cause mortality.

Finally, the stability of the decoupling signature across five distinct accrual cohorts, characterized by low cross-cohort heterogeneity, reinforces its reproducibility and potential for broad clinical application. By moving beyond restrictive linearity assumptions and raw co-expression clustering, this work establishes non-linear regulatory rewiring as a fundamental feature of aggressive breast cancer. These findings provide a novel framework for understanding intra-cluster tumor lethality and offer a robust, independent biomarker for refining patient risk stratification.