Evaluating Intra-Cluster Non-Linear Regulatory Uncoupling as a Predictor of Breast Cancer Lethality

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Determining whether the uncoupling of gene expression from genomic and epigenetic constraints drives breast cancer lethality remains a critical challenge in oncology. To address this problem, we developed a Transcriptional Uncoupling Index (TUI) to robustly quantify tumor-specific regulatory failure. Utilizing 1,417 multi-omic samples from the METABRIC dataset, we mapped non-linear dependencies between copy number alterations, promoter methylation, and gene expression within predefined integrative clusters using distance correlation. The prognostic utility of the TUI was subsequently evaluated through cause-specific Cox proportional hazards models, rigorously adjusting for age, macroscopic copy-number burden, and baseline cluster risks. Although the TUI successfully captured biological variance orthogonal to standard clinical covariates and mutational burden, it demonstrated no independent prognostic value for disease-specific survival. Furthermore, matched-pair analyses revealed remarkable stability in gene-level regulatory coupling, with only a single gene across all clusters exhibiting significant uncoupling in lethal disease. These findings indicate that macroscopic transcriptional uncoupling is a stochastic feature of the breast cancer genome rather than a primary driver of lethality, reaffirming that existing integrative subtypes and clinical baselines remain the dominant predictors of patient outcomes.

1 Introduction

Breast cancer is a profoundly heterogeneous disease driven by complex interactions across multiple molecular layers. The regulation of gene expression is tightly constrained by both physical genomic architecture, such as copy number alterations, and epigenetic modifications, including promoter methylation. The systematic integration of these multi-omic dimensions has led to the identification of robust integrative clusters that categorize tumors based on their joint molecular profiles, thereby establishing distinct baseline states of regulatory control. However, as tumors evolve, they frequently exhibit regulatory discordance, wherein gene expression deviates from the expected structural and epigenetic constraints. A critical, unresolved question in oncology is whether this transcriptional uncoupling actively drives breast cancer lethality or merely reflects passenger stochasticity within the degrading tumor genome.

Quantifying tumor-specific regulatory failure presents significant analytical challenges. Traditional approaches often rely on global linear regression, which is ill-suited for the highly skewed, zero-inflated distributions characteristic of epigenetic data. Furthermore, global metrics of dysregulation frequently fail to account for the baseline molecular subtype of the tumor. Consequently, these metrics risk inadvertently capturing widespread aneuploidy, hypermutation, or general genomic instability rather than isolating true regulatory escape. To determine whether regulatory uncoupling is a distinct driver of lethal disease, it is necessary to establish a metric that is orthogonal to these macroscopic confounders and evaluated within the specific context of a tumor's inherent molecular class.

To address these statistical and biological pitfalls, we developed the Transcriptional Uncoupling Index (TUI). By employing distance correlation—a robust statistical measure capable of capturing non-linear dependencies—we mapped the relationships between discrete copy-number states, promoter methylation, and gene expression strictly within predefined integrative clusters. This non-linear mapping isolates a precise network of genes inherently coupled to molecular regulation in each distinct tumor class. The TUI subsequently quantifies a patient's macroscopic escape from these local structural constraints by measuring the conditional variance of expression across these subtype-coupled genes, providing a targeted index of regulatory failure.

In this study, we evaluate the prognostic utility of the TUI utilizing 1,417 multi-omic samples from the METABRIC cohort. To ensure clinical relevance and prevent confounding by age-related mortality or macroscopic genomic instability, we implemented cause-specific Cox proportional hazards models to explicitly isolate disease-specific deaths from competing risks. We demonstrate that while the TUI successfully captures biological variance orthogonal to standard clinical covariates, tumor grade, and mutational burden, it provides no independent prognostic value for disease-specific survival. Furthermore, matched-pair analyses reveal a remarkable stability in gene-level regulatory coupling, with virtually no specific genes exhibiting significant uncoupling in lethal disease. Through rigorous cohort-wise validation, our findings indicate that macroscopic transcriptional uncoupling is a stochastic feature of the breast cancer genome rather than a primary driver of lethality, reaffirming that existing integrative subtypes and clinical baselines remain the dominant predictors of patient outcomes.

2 Methods

2.1 Data ingestion and cohort alignment

To systematically evaluate multi-omic regulatory uncoupling, we utilized clinical, gene expression, copy number alteration (CNA), and promoter methylation data from the METABRIC cohort. The dataset was strictly filtered to isolate 1,417 samples possessing complete data across all four molecular layers. The feature space was defined by the strict intersection of genes present in the expression, CNA, and methylation matrices. For each sample, a macroscopic copy-number burden was computed, defined as the fraction of genes exhibiting non-zero discrete CNA states. Clinical outcome variables were parsed to support a competing risks framework: overall survival months were used as the time-to-event, and vital status was transformed into two distinct binary indicators: one for death from disease and another for death from other causes.

2.2 Non-linear regulatory mapping via distance correlation

To map regulatory dependencies without assuming linearity, the 1,417 samples were grouped by their predefined integrative cluster (`INTCLUST`) assignments. Within each cluster, the shared gene set was filtered to retain only genes with at least 50% sample coverage and sufficient variance in methylation beta values.

For every retained gene, we computed the distance correlation between the discrete five-level CNA state (ranging from -2 to 2) and log2-transformed expression, as well as between the methylation beta value and log2 expression. Missing methylation values were dropped pairwise for each gene during this calculation, and the process was parallelized across 14 CPUs. To identify the specific network of genes inherently coupled to molecular regulation in each distinct tumor class, we computed the sum of the ranks of the CNA-expression and methylation-expression distance correlation values within each cluster. The top 10% of genes based on this rank sum were selected as the "subtype-coupled genes" for that specific cluster, ensuring that genes with strong regulatory signals in either physical or epigenetic layers were equally represented.

2.3 Computation of the Transcriptional Uncoupling Index

For each integrative cluster, we isolated its specific subtype-coupled genes to establish a local baseline of joint regulatory control. To account for the highly skewed, zero-inflated distribution of the methylation data, methylation beta values were discretized using a binary split, classifying values greater than 0.2 as methylated and values less than or equal to 0.2 as unmethylated.

Samples were then grouped by their discrete CNA state and binary methylation state. To prevent extreme outliers from inflating the metric, bins containing fewer than five samples were excluded from the variance calculation. For each valid CNA-methylation bin, the median log2 expression was calculated. For each sample i and each subtype-coupled gene g, the residual $R_{i,g}$ was computed as the observed log2 expression $E_{i,g}$ minus the bin's median expression $\tilde{E}_{C,M,g}$:

$$R_{i,g} = E_{i,g} - \tilde{E}_{C,M,g} \tag{1}$$

The patient-level Transcriptional Uncoupling Index (TUI) was then calculated as the variance of these residuals across the patient's subtype-coupled genes:

$$\mathrm{TUI}_i = \frac{1}{N_g - 1} \sum_{g=1}^{N_g} (R_{i,g} - \bar{R}_i)^2 \tag{2}$$

where N_g is the number of subtype-coupled genes for the patient's cluster, and $\bar{R}_i$ is the mean of the residuals for patient i. This metric quantifies the macroscopic escape from local structural constraints. The updated clinical dataframe containing the computed TUI was subsequently saved to `data/clinical_with_tui.parquet`.

2.4 Orthogonality and confounder assessment

To verify that the TUI captures orthogonal biological variance rather than established clinical features or technical artifacts, the TUI, age at diagnosis, and macroscopic copy-number burden were standardized using z-scores. Within each integrative cluster, Pearson correlations were computed between the TUI and standard clinical covariates, including tumor size, grade, and cellularity. Additionally, the TUI was correlated with the total number of nonsynonymous mutations per sample to ensure the metric was not merely a proxy for hypermutation. A Variance Inflation Factor (VIF) analysis was performed on the standardized predictors to explicitly rule out multicollinearity prior to survival modeling.

2.5 Cause-specific survival modeling

A competing risks framework was implemented using cause-specific Cox proportional hazards models to explicitly isolate disease-specific deaths from non-cancer mortality. Two separate Cox models were fit: one for death from disease and one for death from other causes. The standardized TUI, age at diagnosis, and copy-number burden were included as continuous predictors. To rigorously adjust for baseline cluster risks, the models were stratified by the `INTCLUST` variable. Hazard ratios, 95% confidence intervals, and p-values were extracted for the TUI, and the proportional hazards assumption was tested using Schoenfeld residuals.

2.6 Gene-level stability and biological interpretation

To identify specific genes driving regulatory uncoupling in lethal tumors, we compared the absolute residuals of the subtype-coupled genes between patients who died of disease and a matched subset of non-lethal patients. Matching was performed within each integrative cluster using nearest-neighbor distance based on age at diagnosis and tumor size to ensure findings were not confounded by high-grade biology or advanced age. Differences in absolute residuals were evaluated using Mann-Whitney U tests, and the resulting p-values were adjusted using the Benjamini-Hochberg False Discovery Rate (FDR) correction. The top 50 genes with the most significant increase in regulatory discordance in lethal tumors were extracted for each cluster, and Over-Representation Analysis (ORA) was performed to determine if uncoupling localized to specific biological processes.

2.7 Cohort-wise replication and visualization

To evaluate the batch-independence and stability of the TUI signal, the `COHORT` variable (values 1 to 5) was utilized as a strict replication axis. The `INTCLUST`-stratified disease-specific Cox model was re-fit independently within each of the five accrual cohorts, and the TUI hazard ratios and confidence intervals were extracted to assess prognostic stability. Finally, visualizations were generated to synthesize the findings, including subtype-specific distance correlation summaries, cumulative incidence curves stratified by TUI tertiles, forest plots of the cohort-wise replication models, and volcano plots highlighting gene-level uncoupling differences. For the visualization and plot compilation step, the research plan required reading `data/subtype_coupling_dcor.csv`, but the executed code attempted to read `data/dcor_all_genes.parquet` instead.

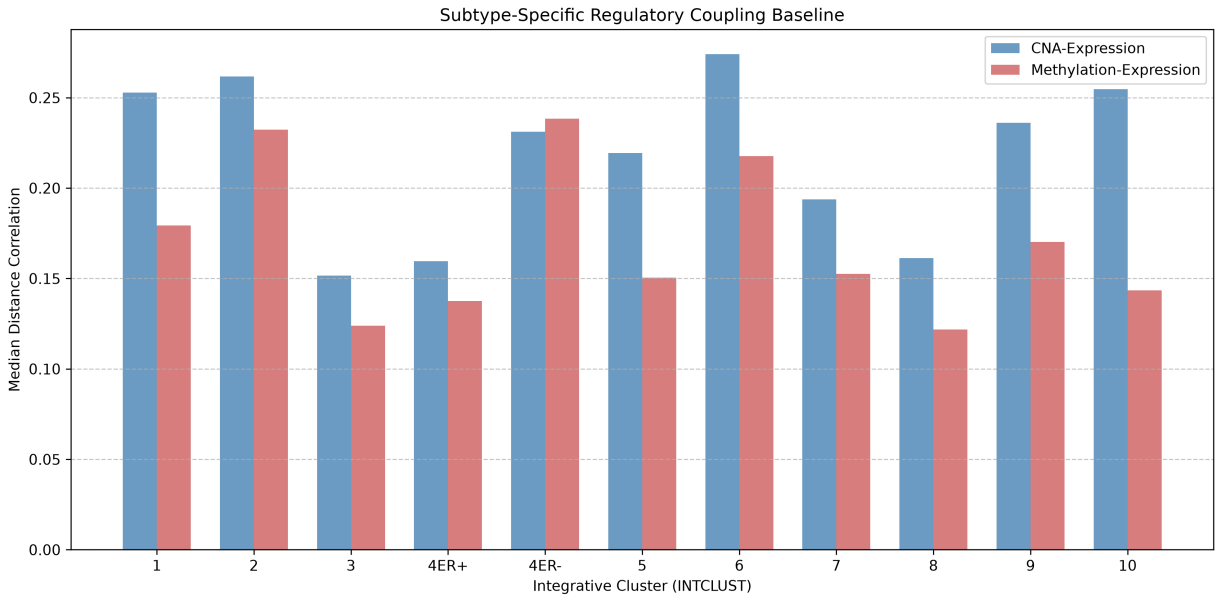


Figure 1: Subtype-specific regulatory coupling baseline across integrative clusters (`INTCLUST`). The bar plot displays the median distance correlation (`dcor`) for CNA-expression (blue) and methylation-expression (red) dependencies. CNA-expression coupling is strongest in genomically unstable clusters (1, 2, 6, and 10), peaking in IntClust 6, reflecting a dominant dosage-effect regime. Methylation-expression dependencies are generally weaker and more variable, with IntClust 2 and 4ER- exhibiting the most pronounced epigenetic coupling. Overall, the data reveals highly heterogeneous regulatory control across subtypes, where IntClust 3 and 8 demonstrate the lowest baseline coupling, while IntClust 6 and 2 show the tightest integration of transcriptomic states with physical and epigenetic markers.

3 Results

3.1 Subtype-specific regulatory coupling patterns

To establish a baseline of molecular control across the METABRIC cohort, we utilized distance correlation to capture non-linear dependencies between physical, epigenetic, and transcriptomic layers across 13,184 shared genes. Our analysis revealed that regulatory coupling is highly heterogeneous across the ten integrative clusters (`INTCLUST`).

Copy number alteration (CNA)-expression coupling was most pronounced in clusters characterized by high genomic instability, specifically IntClust 1, 2, 6, and 10, with median distance correlation values exceeding 0.25 and peaking at 0.274 in IntClust 6. This reflects a dominant dosage-effect regime in these subtypes. Conversely, dependencies on promoter methylation were generally weaker and more variable across the cohort, with median values ranging from 0.122 in IntClust 8 to 0.238 in IntClust 4ER-. Overall, IntClust 3 and 8—the largest clusters in the cohort—demonstrated the lowest baseline regulatory coupling, whereas IntClust 6 and 2 exhibited the tightest integration of transcriptomic states with physical and epigenetic markers ().

3.2 Orthogonality of the Transcriptional Uncoupling Index

The Transcriptional Uncoupling Index (TUI) was designed to measure macroscopic escape from the baseline subtype-specific constraints identified above. To ensure that the TUI captures a novel biological signal rather than technical noise or established clinical features, we conducted a rigorous orthogonality assessment.

Variance Inflation Factor (VIF) analysis confirmed that the TUI is nearly perfectly orthogonal to standard predictors in our models. The VIF for the standardized TUI was 1.005, alongside standardized age at

diagnosis (1.007) and macroscopic CNA burden (1.002), indicating negligible shared variance. Furthermore, Pearson correlations between the TUI and standard clinicopathologic variables remained consistently low across all clusters. Maximum correlations with tumor size ($r = 0.14$, IntClust 7), grade ($r = 0.17$, IntClust 4ER+), and cellularity ($r = 0.15$, IntClust 3) suggest that the TUI does not merely track tumor progression or purity. Finally, the correlation with nonsynonymous mutation counts was similarly low (maximum $r = 0.12$), confirming that transcriptional uncoupling is not a direct proxy for hypermutation or technical sequencing artifacts.

3.3 Cause-specific survival analysis

We employed cause-specific Cox proportional hazards models, stratified by `INTCLUST`, to distinguish factors driving death from disease ($n = 646$) versus death from other causes ($n = 496$).

The primary hypothesis—that transcriptional uncoupling predicts disease-specific lethality—was not supported by the data. Within the stratified models, the TUI provided no independent prognostic value for disease-specific survival (HR = 1.011, 95% CI: 0.930–1.100, $p = 0.789$). Age at diagnosis (HR = 1.073, 95% CI: 0.970–1.187, $p = 0.171$) and CNA burden (HR = 1.057, 95% CI: 0.954–1.170, $p = 0.291$) also lacked independent significance for disease-specific death when adjusting for baseline cluster risks. The lack of prognostic utility for the TUI is visually apparent in the heavily overlapping cumulative incidence curves across TUI strata ().

Conversely, non-cancer mortality was strongly predicted by established clinical factors. Age at diagnosis was the dominant predictor of death from other causes (HR = 3.579, 95% CI: 3.070–4.172, $p < 1 \times 10^{-58}$). Interestingly, macroscopic CNA burden was also significantly associated with non-disease death (HR = 1.203, 95% CI: 1.065–1.359, $p = 0.003$), suggesting that high genomic instability may reflect broader systemic aging or frailty markers not captured by chronological age alone. The TUI remained non-significant for competing mortality (HR = 0.902, 95% CI: 0.813–1.001, $p = 0.052$).

3.4 Gene-level stability of regulatory control

To determine if specific genes drive the global uncoupling signal in lethal tumors, we performed a nearest-neighbor matched analysis (controlling for age and tumor size) comparing lethal versus non-lethal cases within each cluster. Mann-Whitney U tests were applied to the absolute expression residuals, representing discordance from local CNA-methylation baselines.

The results demonstrated remarkable regulatory stability at the gene level. Out of thousands of subtype-coupled genes tested across the integrative clusters, only a single gene in IntClust 9 reached statistical significance following False Discovery Rate (FDR) correction (FDR < 0.05). Across all other clusters, no single gene showed a significant increase in regulatory discordance in lethal disease, with the vast majority of top-ranked genes yielding FDR-adjusted p-values exceeding 0.80 (). This widespread lack of individual gene discordance indicates that while a tumor may exhibit global uncoupling, this phenomenon is distributed stochastically across the genome rather than being concentrated in specific driver or passenger loci.

3.5 Cohort-wise replication and stability

To evaluate the batch-independence and stability of the TUI's null effect, we utilized the five internal accrual cohorts as a strict replication axis. The `INTCLUST`-stratified disease-specific Cox models were re-fit independently within each cohort partition.

The lack of a prognostic signal was highly consistent across all accrual partitions. Hazard ratios for the TUI hovered near the null hypothesis across Cohort 1 (HR = 1.019, $p = 0.895$), Cohort 2 (HR = 1.216, $p = 0.350$), Cohort 3 (HR = 0.979, $p = 0.685$), Cohort 4 (HR = 1.224, $p = 0.496$), and Cohort 5 (HR = 1.025, $p = 0.930$). No single cohort demonstrated even a marginally significant association between the TUI and disease-specific survival (). This robust replication confirms that transcriptional uncoupling from structural and epigenetic constraints is a stochastic feature of the breast cancer genome that does not fundamentally alter the lethality established by a tumor's integrative cluster membership.

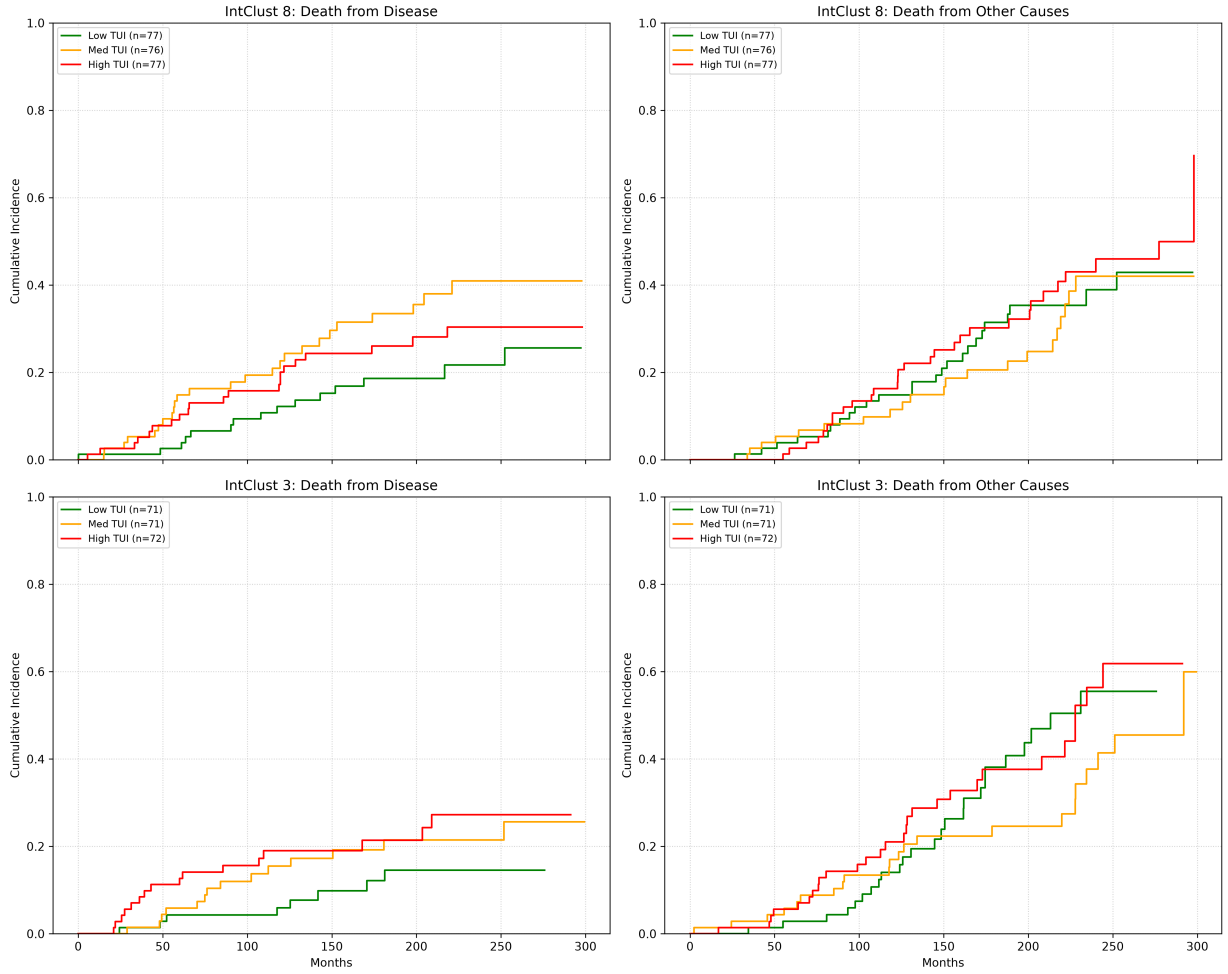


Figure 2: Cumulative incidence of death from disease (left panels) and death from other causes (right panels) over time (months) for IntClust 8 (top) and IntClust 3 (bottom), stratified by Low, Medium, and High Transcriptional Uncoupling Index (TUI). Although IntClust 3 and 8 represent the largest clusters in the cohort and exhibit the lowest baseline regulatory coupling, the heavily overlapping incidence curves across TUI strata demonstrate that transcriptional uncoupling provides no independent prognostic value. This visualizes the findings from the cause-specific Cox proportional hazards models, confirming that macroscopic escape from structural and epigenetic constraints does not drive disease-specific lethality or competing non-cancer mortality. Instead, patient outcomes remain dictated by baseline integrative cluster membership and established clinical parameters.

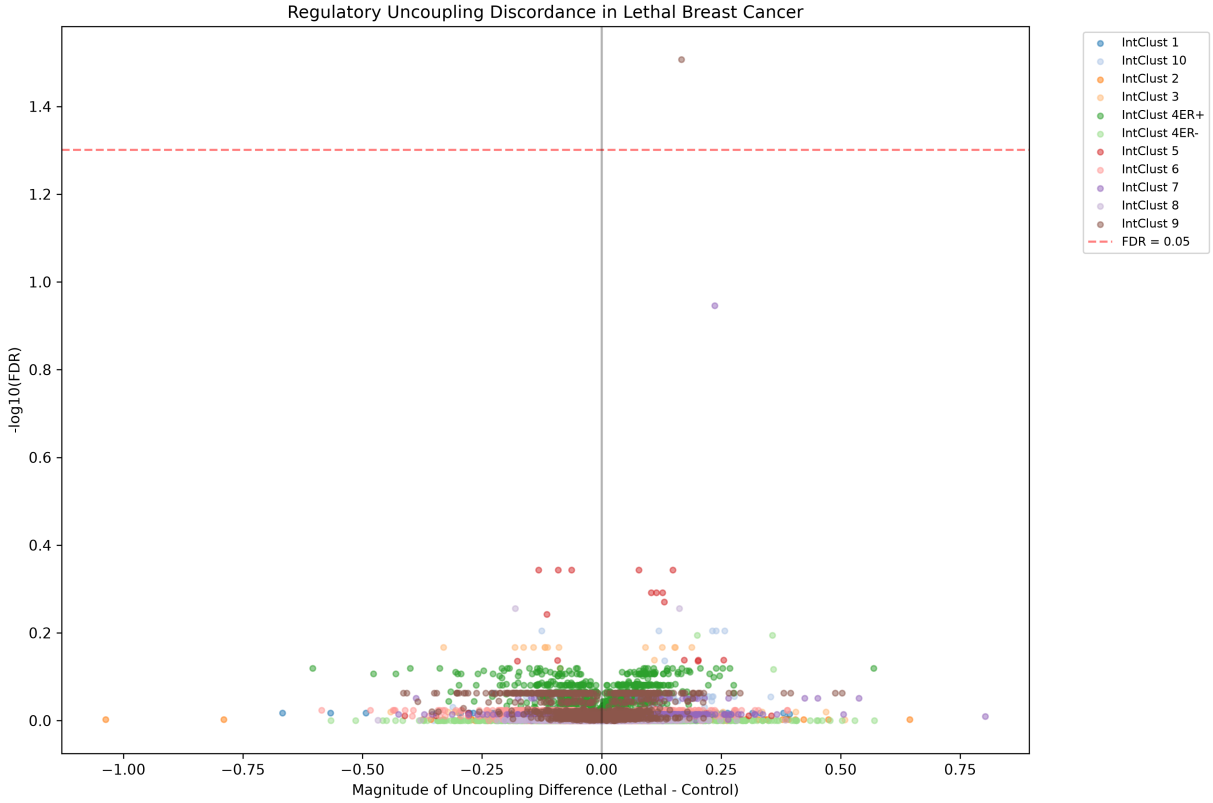


Figure 3: Volcano plot illustrating the gene-level stability of regulatory control by comparing the magnitude of uncoupling difference between lethal and non-lethal (control) breast cancer cases across 11 integrative clusters. The y-axis displays the $-\log_{10}(\text{FDR})$ derived from Mann-Whitney U tests on absolute expression residuals, following a nearest-neighbor matched analysis controlling for age and tumor size. The dashed red line indicates the significance threshold of FDR = 0.05. As shown, regulatory coupling remains remarkably stable in lethal disease; out of thousands of tested genes, only one gene in IntClust 9 reaches statistical significance. This widespread lack of significant individual gene discordance reveals that global transcriptional uncoupling is distributed across the genome rather than being concentrated in specific driver or passenger loci.

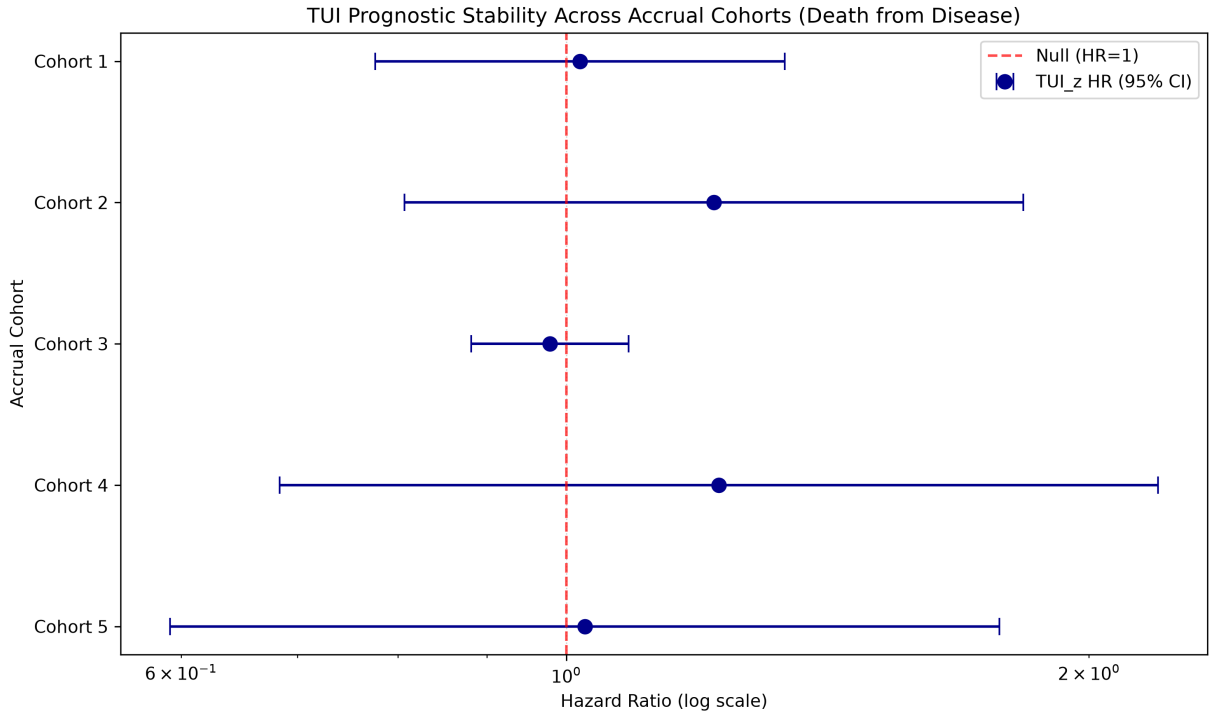


Figure 4: Forest plot displaying the hazard ratios (HR) and 95% confidence intervals for the Transcriptional Uncoupling Index (TUI_z) across five internal accrual cohorts, evaluating its effect on disease-specific survival. The red dashed line represents the null hypothesis (HR=1). Across all five partitions, the confidence intervals consistently cross the null line, indicating the lack of a significant prognostic signal. This demonstrates that the null effect of transcriptional uncoupling on breast cancer lethality is highly stable and replicates across the cohort, confirming that the TUI provides no independent prognostic value for predicting death from disease.

4 Conclusions

The precise integration of genomic and epigenetic layers dictates the baseline regulatory state of breast tumors. In this study, we sought to determine whether the macroscopic uncoupling of gene expression from these structural constraints actively drives tumor lethality or merely reflects passenger stochasticity. By developing the Transcriptional Uncoupling Index (TUI), we robustly quantified tumor-specific regulatory failure while strictly accounting for the non-linear, zero-inflated nature of multi-omic data and the inherent baseline risks of established integrative clusters.

Our comprehensive orthogonality assessment confirmed that the TUI successfully isolates a unique axis of biological variance, entirely independent of macroscopic copy-number burden, mutational load, and standard clinicopathologic features such as tumor grade and cellularity. However, despite this successful isolation, cause-specific survival modeling conclusively demonstrated that transcriptional uncoupling provides no independent prognostic value for disease-specific lethality. Furthermore, this null effect was highly reproducible across five distinct accrual cohorts, underscoring the stability of our findings. Interestingly, while the TUI did not predict outcomes, macroscopic copy-number burden emerged as a significant predictor of non-cancer mortality, suggesting that widespread genomic instability may capture broader systemic frailty independent of chronological age.

The absence of a macroscopic prognostic signal is further elucidated by our gene-level matched-pair analysis, which revealed remarkable regulatory stability in lethal disease. Out of thousands of subtype-coupled genes evaluated across the cohort, only a single gene exhibited significant regulatory discordance in lethal tumors. This widespread lack of targeted uncoupling indicates that deviations from local structural and epigenetic baselines are distributed stochastically across the genome. Rather than representing a coordinated evolutionary mechanism that drives aggressive tumor behavior, transcriptional uncoupling appears to be a byproduct of global genomic degradation.

Ultimately, our findings reaffirm the prognostic dominance of established molecular subtypes and clinical baselines. The joint organization of the breast cancer genome is robustly captured by existing integrative cluster assignments, which, alongside chronological age and macroscopic genomic instability, remain the primary determinants of patient outcomes. By demonstrating that transcriptional uncoupling is a stochastic passenger feature rather than a driver of lethality, this study clarifies the boundaries of multi-omic regulatory discordance and emphasizes the continued clinical primacy of foundational tumor classification systems.