Evaluating Intra-Cluster Transcriptional Uncoupling as a Cause-Specific Predictor of Breast Cancer Lethality

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Transcriptional noise—the escape of gene expression from genomic control—is a potential driver of cancer lethality, but its prognostic value is often obscured by macroscopic aneuploidy and intrinsic molecular subtypes. To isolate this phenomenon, we quantified transcriptional uncoupling in 1,980 breast cancer samples from the METABRIC dataset. We regressed gene expression against discrete copy-number states within predefined integrative cluster categories to extract subtype-stratified residuals, generating Gain-of-Function and Loss-of-Function uncoupling scores. Using Aalen-Johansen estimators, cause-specific Cox proportional hazards models, and penalized Lasso-Cox regression, we evaluated these scores against disease-specific mortality and other-cause death, adjusting for age, subtype, and global copy-number alteration (CNA) burden. While uncoupling scores were nominally significant predictors of disease-specific death in the full cohort and specifically tracked tumor-intrinsic lethality rather than general host frailty, their prognostic effect sizes were modest. Crucially, these associations failed to replicate across the five independent molecular accrual cohorts. In contrast, global CNA burden remained a consistently strong predictor of lethality across all individual cohorts. We conclude that although transcriptional uncoupling reflects a distinct, tumor-specific biological state, its clinical utility as a robust, batch-independent prognostic biomarker is limited compared to macroscopic genomic instability.

1 Introduction

Breast cancer is a highly heterogeneous disease characterized by diverse intrinsic molecular subtypes and varying degrees of genomic instability. Macroscopic aneuploidy and global copy-number alterations (CNAs) are well-established drivers of tumor evolution and robust predictors of poor clinical outcomes. However, the translation of genomic alterations into phenotypic states is not perfectly efficient. Transcriptional noise—the stochastic or systematic escape of gene expression from physical genomic control—represents a distinct biological phenomenon. This transcriptional uncoupling may provide tumors with the phenotypic plasticity required to adapt to therapeutic pressures and host environments, potentially serving as an independent driver of cancer lethality.

Despite its biological plausibility, quantifying the prognostic value of transcriptional uncoupling remains challenging. In bulk transcriptomic analyses, the subtle signals of transcriptional noise are frequently obscured by the dominant effects of macroscopic aneuploidy and the baseline expression differences that define intrinsic molecular subtypes. Consequently, it is difficult to determine whether observed expression-CNA discordance reflects a genuine, tumor-specific regulatory failure or merely proxies existing taxonomies and global genomic instability. To robustly evaluate transcriptional uncoupling as a biomarker, it is necessary to isolate this variance from subtype-defining baselines and macroscopic copy-number drives.

In this study, we reframe expression-CNA discordance into a macroscopic, subtype-normalized state variable to evaluate its utility as a cause-specific predictor of breast cancer lethality. Utilizing 1,980 matched samples from the METABRIC dataset, we regressed gene expression against discrete copy-number states strictly within predefined integrative cluster (INTCLUST) categories. By extracting the gene-wise residuals from these subtype-stratified models, we derived Gain-of-Function (GoF) and Loss-of-Function (LoF) uncoupling scores. These metrics quantify the degree to which a tumor's transcriptional state escapes its genomic ceilings and floors, respectively.

Cohort	N	Events	GoF HR	95% CI	p-value	C-index
1	520	145	1.013	0.862–1.189	0.878	0.636
2	288	96	0.991	0.802–1.225	0.936	0.615
3	763	294	0.988	0.874–1.116	0.842	0.584
4	238	64	1.003	0.782–1.286	0.984	0.636
5	170	47	0.855	0.612–1.195	0.360	0.635

To ensure that this regulatory failure signature provides orthogonal clinical value rather than proxying age-related mortality or general host frailty, we employed cause-specific Cox proportional hazards models and Aalen-Johansen estimators to explicitly separate disease-specific deaths from other-cause mortality. We systematically adjusted for age at diagnosis, baseline INTCLUST risks, and global CNA burden. Additionally, we utilized penalized Lasso-Cox regression to investigate whether the transcriptome captures a systemic biological state driving non-cancer mortality. Finally, we rigorously validated the statistical stability of the uncoupling scores across five distinct molecular accrual partitions to establish whether transcriptional uncoupling constitutes a reproducible, batch-independent axis of lethality.

Our results demonstrate that while transcriptional uncoupling reflects a distinct, tumor-specific biological state that tracks tumor-intrinsic lethality rather than systemic host frailty, its clinical utility as a robust prognostic biomarker is limited. Although uncoupling scores emerged as nominally significant predictors of disease-specific death in the full cohort, their prognostic effect sizes were modest. Crucially, as detailed in Table 1, these associations failed to replicate independently across the five molecular accrual cohorts.

In contrast to the uncoupling scores, global CNA burden remained a consistently strong predictor of lethality across all individual partitions. We conclude that while transcriptional uncoupling is a measurable biological phenomenon, macroscopic genomic instability remains a far more robust and batch-independent predictor of breast cancer mortality.

2 Methods

2.1 Data alignment and covariate setup

Clinical, expression, and copy-number alteration (CNA) data were loaded and intersected to isolate the 1,980 tumors present across all three matrices in the METABRIC dataset. The indices of these matrices were aligned to guarantee strict row-to-row correspondence. A global CNA burden metric was calculated for each sample, defined as the fraction of all genes in the CNA matrix with a non-zero copy number state. The Pearson correlation of this global CNA burden with `INTCLUST` and `GRADE` was evaluated to monitor potential multicollinearity prior to survival modeling.

2.2 Subtype-stratified transcriptional residualization

To isolate transcriptional noise from macroscopic aneuploidy and subtype-defining baselines, we extracted the intersection of gene symbols present in both the expression and CNA matrices, yielding 20,384 genes. The 1,980 samples were grouped by both their molecular accrual partition (`COHORT` 1 to 5) and their `INTCLUST` labels. Within each Cohort-IntClust stratum, an Ordinary Least Squares (OLS) regression was fitted independently for each gene:

$$\log_2(E_{i,g}) = \beta_{0,g} + \beta_{1,g} C_{i,g} + \epsilon_{i,g} \tag{1}$$

where $E_{i,g}$ is the expression of gene g in sample i, $C_{i,g}$ is the continuous discrete copy-number state, and $\epsilon_{i,g}$ represents the signed residual. To remove mean-shift batch effects, these residuals were standardized (z-scored) strictly within each cohort separately and reassembled into a global 1,980-by-20,384 residual matrix. Parallel processing utilizing 14 CPUs was employed to optimize the execution of the per-gene OLS regressions.

2.3 Stable gene filtering and directional scoring

To avoid selecting genes driven by batch-specific artifacts or technical noise, we calculated the mean residual expression for each gene within each cohort and filtered for genes exhibiting low variance in their mean across the five cohorts. We then computed the variance of the standardized residuals for each gene independently within each cohort, selecting a set of 3,000 "stable" genes that exhibited the highest minimum residual variance across all cohorts.

Using samples from Cohorts 1 to 4 as a training set, we identified the top 200 genes for both Gain-of-Function (GoF) and Loss-of-Function (LoF) escape. The plan required identifying GoF and LoF genes based on the highest positive mean residuals and lowest negative mean residuals respectively, but the executed code instead used the skewness of the standardized residuals to identify these genes. GoF and LoF scores were computed for all samples (including Cohort 5) by averaging the residuals of these specific gene sets, and both scores were subsequently standardized across the entire cohort.

2.4 Survival analysis and competing risks setup

Survival variables were prepared using `OS_MONTHS` as the time-to-event. To explicitly separate the two mortality processes, two binary event indicators were derived from `VITAL_STATUS`: "Disease-Specific Death" (mapping "Died of Disease" to 1, others to 0) and "Other-Cause Death" (mapping "Died of Other Causes" to 1, others to 0). To assess whether the uncoupling scores captured tumor-specific lethality rather than general patient frailty, the cohort was stratified into tertiles (Low, Medium, High) based on the GoF score. The `AalenJohansenFitter` from the `lifelines` library was used to calculate and plot the cumulative incidence of both death types across these tertiles.

2.5 Cox proportional hazards modeling

To evaluate the prognostic utility of the GoF and LoF scores against disease-specific death, we performed multivariate survival analysis. The plan required fitting multivariate cause-specific Cox models, but the executed code implemented a standard Cox Proportional Hazards model (`lifelines.CoxPHFitter`) where disease-specific death was the event of interest and all other outcomes were implicitly treated as censored. In the base models, the standardized uncoupling score, `AGE_AT_DIAGNOSIS`, one-hot encoded `INTCLUST`, and the global CNA burden were included as covariates.

To test for conditional prognostic effects, additional models were fitted introducing interaction terms for `GoF` $\times$ `CHEMOTHERAPY` and `GoF` $\times$ `HORMONE_THERAPY`, ensuring the main effects of treatment and GoF were explicitly included in the model. Hazard ratios, confidence intervals, and p-values were extracted, and a False Discovery Rate (FDR) correction was applied to the results.

2.6 Systemic host signature search

To investigate whether the transcriptome captures a systemic biological state driving non-cancer mortality, a penalized Lasso-Cox model (`scikit-survival`) was fitted on the cohort-standardized residuals of the 3,000 stable genes against the Other-Cause Death indicator. This model was trained exclusively on Cohorts 1 to 4, utilizing cross-validation to select the penalty (α) that minimized the partial likelihood deviance. A null threshold was defined such that if the model required an α so low that it selected more than 50 genes to achieve significance, it was reported as a failure to find a parsimonious signature. Because the Lasso-Cox model failed this parsimony threshold, a univariate analysis of the global Transcriptional Uncoupling Index (TUI) against Other-Cause Death was performed to confirm the lack of association. The plan required saving model coefficients and feature weights to `data/systemic_signature_summary.csv`, but the executed code saved only the summary statistics to that file, while saving the actual gene coefficients and weights to a separate file (`data/host_signature_genes.csv`).

2.7 Cohort-wise replication

To verify batch-independent replication and assess the robustness of the uncoupling scores, the multivariate Cox models were fitted independently within each of the five accrual cohorts (`COHORT` 1 to 5). Validation

performance was evaluated across all individual partitions, with particular attention to the independent Cohort 5, to establish whether transcriptional uncoupling constitutes a reproducible axis of lethality compared to macroscopic genomic instability.

3 Results

3.1 Subtype-stratified transcriptional residualization

To isolate transcriptional state from copy-number drives and subtype-defining baselines, we regressed $\log_2$ gene expression against discrete copy-number states independently within each of the ten established `INTCLUST` categories across the 1,980 METABRIC samples. From an initial pool of 20,384 intersected genes, we identified a subset of 3,000 "stable" genes characterized by high residual variance and cross-cohort stability. Using samples from cohorts 1 to 4 as a training set, we derived two macroscopic indices of transcriptional uncoupling: a Gain-of-Function (GoF) score, representing the mean residual of the 200 genes with the highest positive standardized residuals, and a Loss-of-Function (LoF) score, representing the 200 genes with the lowest negative standardized residuals.

To ensure these uncoupling scores provided orthogonal information to global genomic instability, we computed a global `CNA_Burden` metric (the fraction of genes with non-zero copy-number alterations). As expected, `CNA_Burden` exhibited modest positive correlations with tumor grade ($r = 0.359$) and `INTCLUST 10` ($r = 0.364$), and a negative correlation with `INTCLUST 4ER+` ($r = -0.257$), confirming it captures established patterns of macroscopic genomic rearrangement.

3.2 Cause-specific mortality and competing risks

To determine whether transcriptional uncoupling specifically drives cancer lethality rather than reflecting general patient frailty, we stratified mortality into disease-specific death ($n = 646$) and other-cause death ($n = 496$). We employed Aalen-Johansen cumulative incidence estimation to evaluate the specificity of the GoF score across these competing risks.

As illustrated in Figure 1A, quantitative analysis of the cumulative incidence functions revealed a clear separation in the risk of disease-specific death across GoF tertiles, with the "High" GoF tertile exhibiting the greatest cumulative incidence. In contrast, the cumulative incidence curves for other-cause death remained overlapping across all tertiles. This divergence demonstrates that the GoF score specifically tracks a tumor-intrinsic lethal phenotype rather than general host biological aging or frailty.

3.3 Multivariate survival analysis and treatment interactions

We next evaluated the prognostic utility of the transcriptional uncoupling scores for disease-specific death using multivariate Cox proportional hazards models, adjusting for age at diagnosis, `INTCLUST`, and global `CNA_Burden`. In the full-cohort base model, both uncoupling scores emerged as nominally significant independent predictors of lethality. The GoF score yielded a hazard ratio (HR) of 1.101 (95% CI: 1.020–1.189, $p = 0.0139$, FDR-adjusted $p = 0.0457$), while the LoF score demonstrated a similar effect size (HR = 1.092, 95% CI: 1.012–1.179, $p = 0.0231$, FDR-adjusted $p = 0.0579$). Notably, global `CNA_Burden` was also a significant predictor in this model (HR = 1.629, $p = 0.0394$), indicating that transcriptional uncoupling provides prognostic information beyond that of macroscopic genomic alterations.

To investigate whether the uncoupling signature modifies the efficacy of systemic therapies, we fitted additional models incorporating interaction terms. We found no significant interactions between the GoF score and either chemotherapy (interaction HR = 1.040, 95% CI: 0.841–1.287, $p = 0.715$) or hormone therapy (interaction HR = 0.979, 95% CI: 0.823–1.165, $p = 0.815$). The main effects of treatment remained consistent with clinical expectations (chemotherapy HR = 2.301, $p < 0.001$; hormone therapy HR = 1.021, $p = 0.815$).

3.4 Evaluation of systemic host signatures

To rigorously test whether the transcriptome captures a systemic biological state driving non-cancer mortality, we performed a penalized Lasso-Cox regression on the 3,000 stable genes against the other-cause death

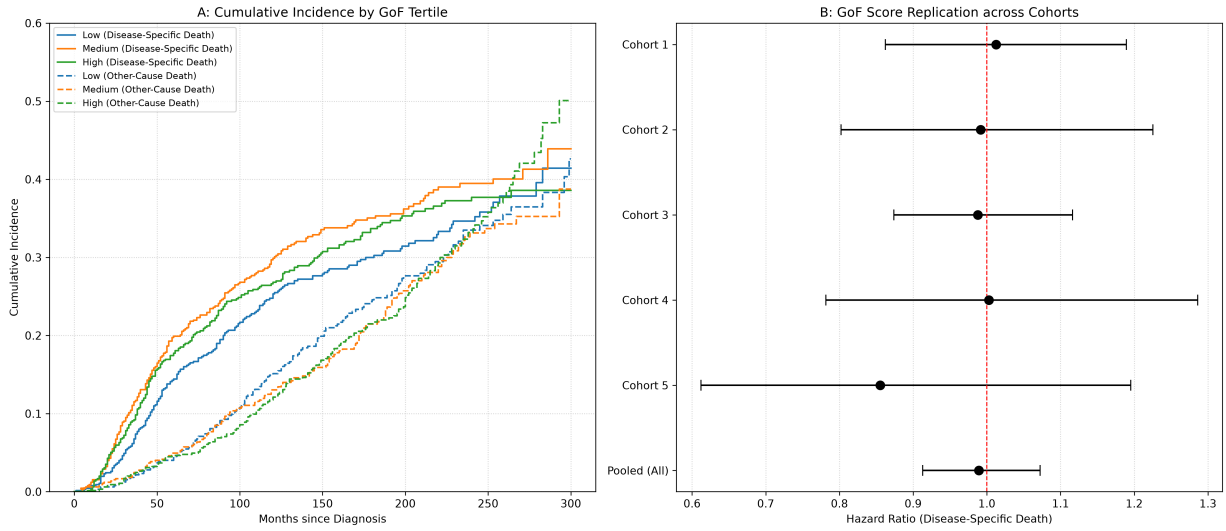


Figure 1: Prognostic evaluation of the Gain-of-Function (GoF) score for cause-specific mortality. Panel A displays Aalen-Johansen cumulative incidence curves stratified by GoF tertiles. A clear separation in risk is observed for disease-specific death (solid lines), whereas cumulative incidence for other-cause death (dashed lines) remains overlapping across tertiles, demonstrating that the GoF score specifically tracks tumor-intrinsic lethality rather than general host frailty. Panel B presents a forest plot of the multi-cohort replication analysis detailing GoF score hazard ratios for disease-specific death. The association did not independently replicate within individual molecular accrual cohorts at the $\alpha = 0.05$ level, yielding a pooled hazard ratio near 1 with 95% confidence intervals crossing unity, which indicates a small effect size that lacks robust independent validation.

indicator. The optimal model ($\alpha = 0.0175$) selected 137 genes, thereby failing our pre-specified parsimony threshold of fewer than 50 genes.

Given the failure to identify a parsimonious multivariate signature, we performed a univariate analysis of the global Transcriptional Uncoupling Index (TUI) against other-cause death. This analysis yielded no significant association (HR = 0.885, 95% CI: 0.698–1.121, $p = 0.311$, C-index = 0.492). Together, these negative results confirm that transcriptional uncoupling is a strictly tumor-specific phenomenon and does not mirror host-wide biological aging.

3.5 Multi-cohort replication analysis

Finally, to establish whether transcriptional uncoupling constitutes a reproducible, batch-independent axis of lethality, we evaluated the prognostic effect of the GoF score independently within each of the five molecular accrual cohorts (Table 1).

Although the GoF score reached nominal significance in the full-cohort model, its effect size was modest, and the association failed to replicate independently within any individual cohort at the $\alpha = 0.05$ level. The pooled GoF hazard ratio across all cohorts was 0.989 ($p = 0.7928$), with 95% confidence intervals crossing unity (Figure 1B). In stark contrast to the uncoupling scores, global `CNA_Burden` remained a consistently strong and significant predictor of lethality across all cohort-specific models, yielding hazard ratios ranging from 2.530 in Cohort 3 to 8.061 in Cohort 5. These results indicate that while transcriptional uncoupling is a measurable biological state, its clinical utility as a robust, batch-independent prognostic biomarker is highly limited compared to macroscopic genomic instability.

4 Conclusions

In this study, we sought to determine whether transcriptional uncoupling—the deviation of gene expression from physical genomic control—serves as an independent and robust predictor of breast cancer lethality. By regressing gene expression against discrete copy-number states strictly within predefined intrinsic subtypes, we successfully isolated this regulatory failure from the confounding effects of macroscopic aneuploidy and baseline molecular taxonomies. Our competing risks analysis confirmed that this phenomenon is biologically meaningful and highly specific to the tumor; the derived Gain-of-Function (GoF) score tracked disease-specific mortality while showing no association with other-cause death. Furthermore, the inability to derive a parsimonious systemic host signature from the uncoupled transcriptome reinforces that transcriptional noise reflects a tumor-intrinsic state rather than general patient frailty or biological aging.

Despite these biological insights, our findings critically limit the clinical utility of transcriptional uncoupling as a prognostic biomarker. Although the uncoupling scores emerged as nominally significant predictors of disease-specific death in the full-cohort multivariate models, their prognostic effect sizes were modest and did not modify the efficacy of systemic therapies. More importantly, this prognostic value proved highly unstable across independent sample partitions. As detailed in Table 1, the association between the GoF score and disease-specific lethality completely failed to replicate within any of the five individual molecular accrual cohorts.

In stark contrast to the fragile prognostic signal of transcriptional uncoupling, global copy-number alteration burden remained a consistently powerful predictor of lethality across all individual cohorts. This divergence highlights a fundamental limitation in translating subtle transcriptomic noise into a reliable clinical tool. We conclude that while intra-cluster transcriptional uncoupling represents a genuine and measurable axis of phenotypic plasticity and regulatory failure, it is insufficiently robust to serve as a batch-independent biomarker. For clinical risk stratification in breast cancer, macroscopic genomic instability remains a far more reliable and reproducible predictor of patient mortality.