Evaluation of Cross-Layer Regulatory Coupling Variance as a Predictor of Breast Cancer Lethality

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Molecular regulatory mechanisms frequently break down in cancer, yet it remains unclear whether the variance in cross-layer molecular coupling provides independent prognostic information regarding tumor lethality. To address this, we investigated whether deviations in mRNA expression from levels predicted by copy-number and promoter methylation could serve as a replicable, cause-specific hazard predictor. Using 1,417 multi-omic breast cancer samples from the METABRIC cohort, we modeled baseline regulatory coupling via robust regression and isolated negative deviations representing targeted under-expression. Non-negative matrix factorization was then applied to extract four genomically independent regulatory signatures. We evaluated the prognostic utility of these signatures using cause-specific Cox proportional hazards models adjusted for competing risks and standard clinicopathological covariates, followed by a meta-analysis across five independent accrual cohorts. While biological characterization confirmed that these signatures capture a distinct layer of transcriptional variance orthogonal to established breast cancer gene panels, survival analyses revealed that none of the extracted signatures provided significant independent predictive value for disease-specific mortality. Furthermore, likelihood ratio tests indicated no incremental prognostic benefit within specific molecular subtypes, and the multi-cohort meta-analysis confirmed a highly consistent null effect. Ultimately, although multi-layer regulatory decoupling is a measurable feature of breast tumor biology, it does not constitute an independent axis of lethality beyond established clinicopathological markers and existing molecular taxonomies.

1 Introduction

Breast cancer is a highly heterogeneous disease driven by complex alterations across multiple molecular layers, including genomic copy number variations, epigenetic modifications, and transcriptomic dysregulation. Over the past decades, comprehensive multi-omic profiling has established robust molecular taxonomies that successfully stratify patients, predict disease progression, and guide targeted therapies. These established frameworks, however, predominantly rely on static molecular features—such as the presence of specific mutations, the amplification of oncogenes, or the absolute expression levels of predefined gene panels.

While these static markers are invaluable, tumorigenesis is fundamentally characterized by the breakdown of normal molecular control mechanisms. In a healthy cellular state, mRNA expression is tightly coupled to structural genomic dosage and epigenetic regulation. In cancer, this cross-layer regulatory coupling frequently deteriorates. Despite the widespread recognition of this biological phenomenon, it remains unclear whether the variance in this coupling—specifically, the extent to which mRNA expression deviates from the levels predicted by copy-number and promoter methylation—provides independent prognostic information regarding tumor lethality.

1.1 Quantifying cross-layer regulatory failure

To capture the biological breakdown of molecular control, it is necessary to move beyond absolute expression levels and instead quantify regulatory decoupling. This can be conceptualized as the residual transcriptional variance that remains after accounting for structural and epigenetic baselines. Of particular interest are negative deviations, where observed gene expression is substantially lower than what is predicted by the multi-omics baseline. These deviations represent targeted under-expression or silencing, a hallmark of regulatory failure that may drive aggressive tumor phenotypes.

However, quantifying this decoupling presents significant analytical challenges. Tumors exhibit vastly different levels of macroscopic copy-number burden, ranging from highly stable genomes to pervasive chromosomal instability. Any measure of regulatory failure must explicitly decouple localized transcriptional deviations from global genomic instability to ensure that the resulting metrics are not merely trivial proxies for known macro-features like tumor cellularity or broad chromosomal chaos.

1.2 Study objectives and summary of findings

In this study, we systematically investigate whether cross-layer regulatory coupling variance serves as a replicable, cause-specific hazard predictor of breast cancer lethality. We utilized a highly annotated cohort of 1,417 multi-omic breast cancer samples, modeling baseline regulatory coupling via robust regression to account for heteroscedasticity driven by genomic instability. By isolating negative transcriptional deviations and applying non-negative matrix factorization, we extracted genomically independent regulatory signatures that explicitly normalize for each tumor's macroscopic copy-number burden.

We rigorously evaluated the independent prognostic utility of these signatures using cause-specific Cox proportional hazards models. To ensure the extracted signal provided orthogonal clinical value rather than rediscovering known static taxonomies or general patient frailty, our survival framework systematically adjusted for established clinicopathological covariates and explicitly modeled disease-specific mortality against competing risks. Finally, we tested the statistical stability of these signatures across five distinct molecular accrual partitions to establish whether multi-layer regulatory failure constitutes a reproducible, batch-independent axis of lethality.

Our findings demonstrate that while the extracted signatures successfully capture a distinct layer of transcriptional variance orthogonal to established breast cancer gene panels, they consistently yield a null effect for predicting disease-specific mortality. Extensive survival analyses, likelihood ratio tests within specific molecular subtypes, and a multi-cohort meta-analysis all confirm that these regulatory deviations do not provide significant independent predictive value. Ultimately, this study rigorously establishes that although multi-layer regulatory decoupling is a measurable and distinct feature of breast tumor biology, it does not constitute an independent axis of lethality beyond established clinicopathological markers and existing molecular taxonomies.

2 Methods

2.1 Data assembly and cohort alignment

To systematically evaluate cross-layer regulatory coupling, we utilized multi-omic and clinical data from the METABRIC cohort. The analysis was strictly restricted to the intersection of samples possessing complete data across four molecular layers: mRNA expression, genomic copy number alterations (CNA), and promoter methylation, alongside comprehensive clinical follow-up. This intersection yielded a final analytical cohort of 1,417 primary breast tumor samples. All subsequent analyses and models rest entirely on this 1,417-sample four-layer intersection.

To evaluate cause-specific hazards, the vital status indicator was parsed to define competing risks. Two binary event indicators were created: disease-specific death (1 if the patient died of breast cancer, 0 otherwise) and death from other causes (1 if the patient died of any other cause, 0 otherwise). Overall survival in months was extracted as the time-to-event variable. Additionally, to account for the macroscopic genomic instability of each tumor, a copy-number burden metric (CN_BURDEN) was calculated for each sample as the fraction of all measured genes harboring a non-zero copy number call.

2.2 Subtype-specific baseline coupling estimation

To quantify the baseline regulatory coupling between molecular layers, we modeled mRNA expression as a function of copy number and promoter methylation. Samples were first grouped by their precomputed integrative cluster (`INTCLUST`) labels to establish subtype-specific baselines, with very small clusters pooled into their most biologically similar counterparts to ensure sufficient statistical power.

For each shared gene g and within each `INTCLUST` group, a robust linear regression model (Huber regression) was fitted to predict expression (E) from copy number state (C) and methylation beta value (M). Copy number was treated as a dummy-encoded categorical variable, with the neutral diploid state ($C = 0$) explicitly set as the reference category:

$$E_g = \beta_{0,g} + \sum_{k \in \{-2,-1,1,2\}} \beta_{C,k,g} I(C_g = k) + \beta_{M,g} M_g + \epsilon_g \tag{1}$$

where $I(\cdot)$ is the indicator function and ϵ_g represents the residual transcriptional variance.

To assess whether the variance of these residuals scaled with macroscopic genomic instability, a Breusch-Pagan test was performed for each gene against CN_BURDEN. Significant heteroscedasticity was observed in a substantial fraction of genes, justifying the subsequent use of robust standard errors in all survival models.

2.3 Isolation of negative regulatory deviations and signature extraction

To make the residuals comparable across genes and subtypes, they were standardized into z-scores (Z_g) within each `INTCLUST` group. Because our primary biological interest was targeted regulatory failure—specifically, under-expression relative to the expected multi-omics baseline—we isolated the negative residuals. A non-negative matrix of negative deviations (D) was constructed by taking the absolute value of negative z-scores and setting all positive residuals to zero:

$$D_g = \begin{cases} |Z_g| & \text{if } Z_g < 0 \\ 0 & \text{if } Z_g \geq 0 \end{cases} \tag{2}$$

Standard Non-negative Matrix Factorization (NMF) with Nonnegative Double Singular Value Decomposition (`nndsvd`) initialization was applied to this continuous, direction-preserving matrix. The optimal number of components (K) was determined empirically by evaluating the reconstruction error across $K \in [2, 10]$. The elbow method identified $K = 4$ as the optimal solution.

The resulting sample-wise NMF transformation weights represented the raw module scores. To ensure these scores captured localized regulatory decoupling rather than acting as trivial proxies for global genomic instability, each raw score (W_k) was orthogonalized against CN_BURDEN using robust linear regression:

$$W_k = \gamma_{0,k} + \gamma_{1,k} \text{CN_BURDEN} + S_k \tag{3}$$

The residuals from these models (S_k) were extracted to form four Genomically Independent Regulatory Signatures.

2.4 Biological characterization and subtype-specific evaluation

To verify that the extracted signatures did not merely rediscover known histological differentiations or low-cellularity artifacts, they were correlated with tumor cellularity, stage, grade, age at diagnosis, and the normal-like intrinsic subtype. Kruskal-Wallis tests were also performed across histological subtypes.

For functional characterization, the top 100 genes with the highest NMF component weights for each signature were extracted. These gene lists were evaluated for overlap with established breast cancer gene panels (e.g., PAM50) and subjected to functional enrichment analysis, explicitly focusing on pathways that were silenced or downregulated relative to expected dosage.

To quantify the incremental prognostic value of the signatures within specific biological contexts, the cohort was stratified by major intrinsic subtypes (e.g., Luminal A, Luminal B). Within each subtype, Likelihood Ratio Tests were used to formally compare a baseline clinical survival model against a full model incorporating the regulatory signatures.

2.5 Competing risks and cause-specific survival modeling

The independent prognostic utility of the regulatory signatures was evaluated using a competing risks framework. The cohort was stratified into "high" and "low" groups based on the median score of each signature.

The Cumulative Incidence Function (CIF) for disease-specific death was estimated and plotted using the Aalen-Johansen estimator, explicitly treating death from other causes as a competing risk.

Cause-specific Cox Proportional Hazards models were fitted on the continuous signature scores. The disease-specific model included the signatures alongside a comprehensive set of clinicopathological covariates: age at diagnosis, tumor size, grade, tumor stage, cellularity, and dummy-encoded histological subtype. To account for the heteroscedasticity identified during baseline coupling estimation, robust standard errors were employed. A parallel Cox model was fitted for death from other causes using the exact same covariates. The proportional hazards assumption was verified for all models using Schoenfeld residuals.

2.6 Multi-cohort meta-analysis

To rigorously test the statistical stability of the signatures as a reproducible, batch-independent axis of lethality, a meta-analysis was conducted across five distinct molecular accrual partitions (Cohorts 1 through 5). Within each cohort, the fully adjusted disease-specific Cox model was refitted independently. The log-hazard ratios and their standard errors were extracted for each signature and combined using an inverse-variance weighted meta-analysis. The I^2 statistic was calculated to quantify inter-cohort heterogeneity, with random-effects models employed where appropriate.

3 Results

3.1 Baseline regulatory coupling and extraction of negative deviations

Following the assembly of the 1,417-sample analytical cohort with complete multi-omic profiles, we first established subtype-specific baseline models of molecular control. By modeling mRNA expression as a function of discrete copy-number states and promoter methylation, we quantified the residual transcriptional variance for each gene. To validate our statistical framework, we performed Breusch-Pagan tests to assess whether the variance of these residuals scaled with macroscopic genomic instability. We observed significant heteroscedasticity ($p < 0.05$) relative to the copy-number burden in approximately 24% of the evaluated genes (including *RASD2*, *RBM17*, and *SUV39H1*). This finding empirically justified the use of robust standard errors in all subsequent survival models to account for the increased transcriptional variance inherent to more genomically unstable tumors.

To isolate targeted regulatory failure, we extracted the absolute values of the negative residuals—representing genes whose expression was substantially lower than predicted by the multi-layer baseline. Non-negative Matrix Factorization (NMF) was applied to this standardized matrix of negative deviations. By evaluating the reconstruction error across a range of components ($K \in [2, 10]$), the elbow method identified an optimal inflection point at $K = 4$, yielding a reconstruction error of 1967.5 (Figure 1).

The raw sample-wise NMF weights for these four components were subsequently orthogonalized against the macroscopic copy-number burden via robust linear regression. The residuals from these models constituted the four Genomically Independent Regulatory Signatures (Sig 1, Sig 2, Sig 3, and Sig 4) utilized in downstream analyses.

3.2 Biological characterization of regulatory signatures

To ensure that the extracted signatures captured a novel layer of transcriptional variance rather than rediscovering established molecular taxonomies, we evaluated their overlap with canonical breast cancer gene sets. The top 100 genes most strongly associated with each signature exhibited zero overlap with both the PAM50 panel and a standard ER/HER2/Proliferation pathway panel. Overlap with a 173-gene targeted breast cancer panel was also negligible, with Sig 1, Sig 2, and Sig 3 each sharing only a single gene, and Sig 4 sharing two genes.

Furthermore, the signatures demonstrated weak or non-significant associations with macroscopic tumor features. Sig 4 exhibited the strongest, albeit weak, correlations with age at diagnosis (Spearman $\rho = 0.187$, $p < 0.001$) and tumor cellularity ($\rho = 0.204$, $p < 0.001$). Kruskal-Wallis tests indicated that while Sig 2 and Sig 4 varied slightly across histological subtypes ($p = 0.018$ and $p = 0.019$, respectively), Sig 1 and Sig 3 did

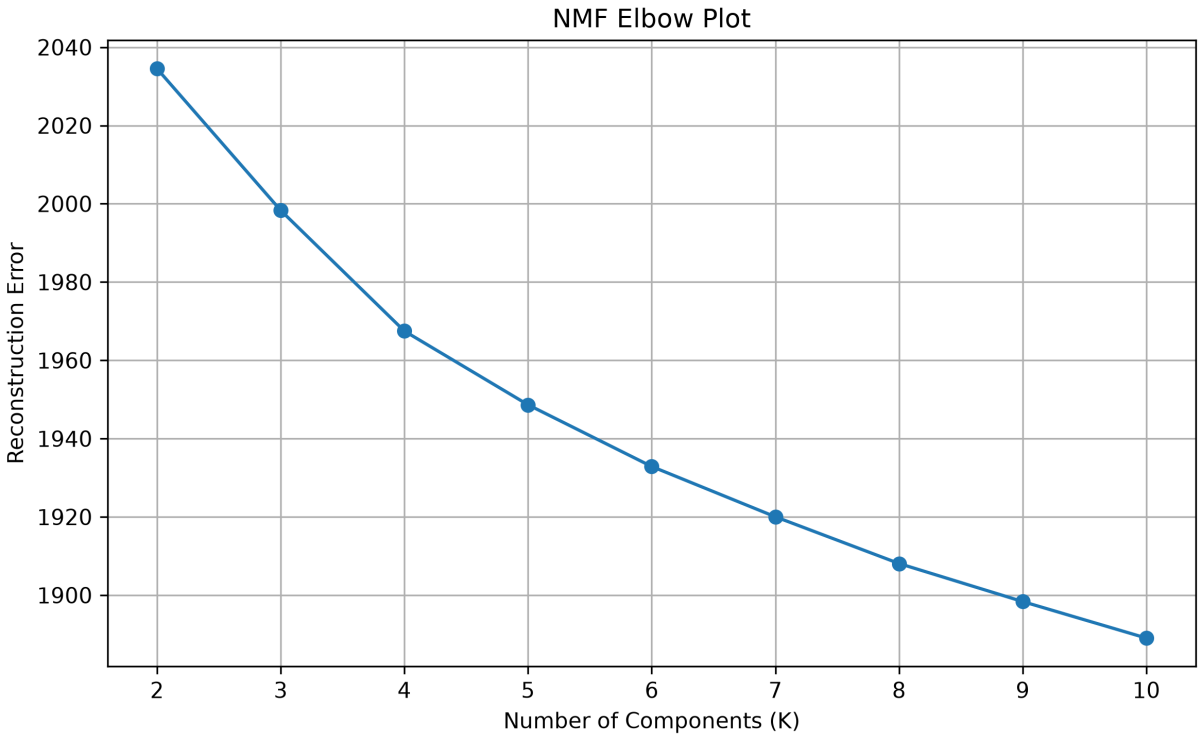


Figure 1: Non-negative Matrix Factorization (NMF) elbow plot illustrating the reconstruction error across a range of components ($K \in [2, 10]$) applied to the standardized matrix of negative regulatory deviations. The plot demonstrates a continuous decrease in reconstruction error as the number of components increases. Based on the elbow method, an inflection point is observed at $K = 4$ (reconstruction error = 1967.5), which justifies the selection of a 4-component solution to extract the genomically independent regulatory signatures representing localized regulatory failure for subsequent prognostic evaluation.

not distinguish between histological groups ($p > 0.49$). Together, these results confirm that the regulatory signatures represent a distinct biological signal orthogonal to established static markers.

3.3 Cause-specific survival and predictive utility

The primary objective of this study was to determine whether these regulatory deviations provide independent prognostic value for tumor lethality. We evaluated this using cause-specific Cox Proportional Hazard models, explicitly isolating the 646 disease-specific deaths from the 496 deaths attributed to competing risks.

After adjusting for age at diagnosis, tumor size, grade, tumor stage, cellularity, and histological subtype, none of the four signatures emerged as statistically significant independent predictors of disease-specific mortality. The adjusted Hazard Ratios (HR) remained close to unity: Sig 1 HR = 0.93 (95% CI: 0.82–1.05, $p = 0.222$); Sig 2 HR = 0.96 (95% CI: 0.87–1.07, $p = 0.482$); Sig 3 HR = 1.00 (95% CI: 0.89–1.13, $p = 0.940$); and Sig 4 HR = 0.92 (95% CI: 0.83–1.02, $p = 0.126$).

Consistent with the Cox models, cumulative incidence curves based on Aalen-Johansen estimates demonstrated no meaningful separation between high and low signature strata for disease-specific mortality when accounting for competing risks (Figure 2).

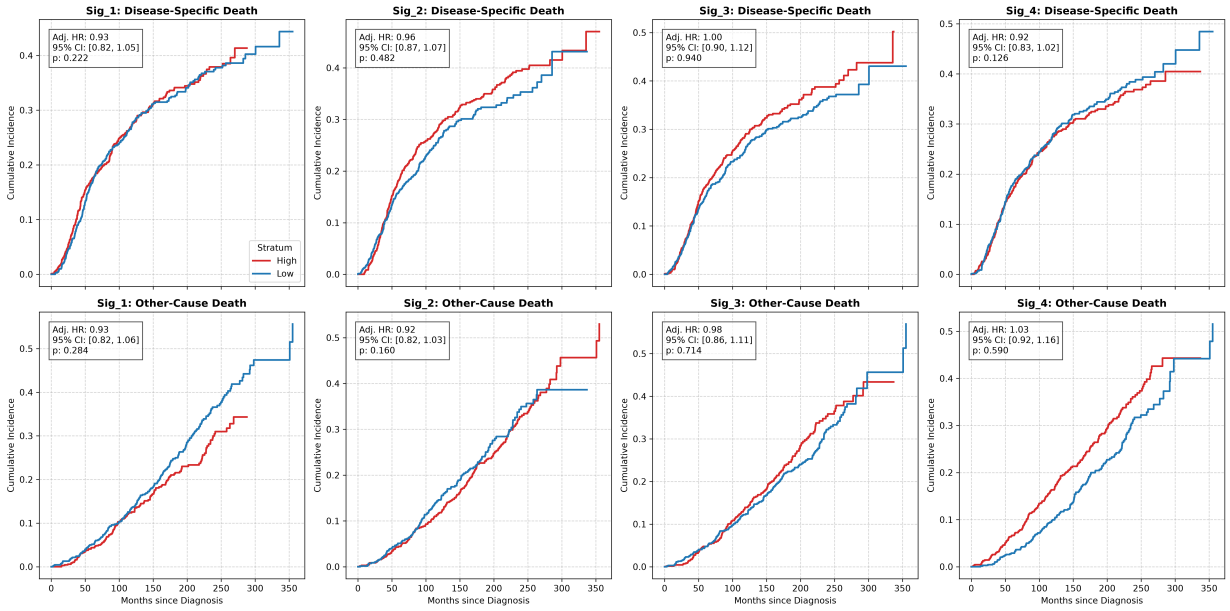


Figure 2: Cumulative incidence curves (Aalen-Johansen estimates) illustrating disease-specific mortality (top row) and death from other causes (bottom row), stratified by high and low levels of the four Genomically Independent Regulatory Signatures (Sig 1, Sig 2, Sig 3, and Sig 4). Cause-specific Cox Proportional Hazard models, adjusted for established clinicopathological markers and competing risks, reveal that none of the signatures are statistically significant independent predictors of mortality. The lack of consistent separation between the strata, alongside adjusted Hazard Ratios remaining close to unity (with 95% confidence intervals crossing 1 and $p > 0.05$), demonstrates that these multi-layer regulatory deviations do not add prognostic value for tumor lethality beyond existing clinical taxonomies.

To assess whether the signatures offered incremental prognostic value within specific biological contexts, we performed Likelihood Ratio Tests comparing the baseline clinical models to models incorporating the signatures within major intrinsic subtypes. No subtype demonstrated a significant improvement in model fit at the $\alpha = 0.05$ level (Luminal A: $p = 0.421$; Luminal B: $p = 0.074$; Basal: $p = 0.928$; Her2: $p = 0.352$; Claudin-low: $p = 0.979$).

3.4 Multi-cohort meta-analysis and robustness

To rigorously verify the stability of these null findings, we evaluated the prognostic utility of the signatures independently across the five distinct molecular accrual partitions (Cohorts 1 through 5). A meta-analysis of the cohort-specific hazard ratios confirmed a highly consistent lack of prognostic signal across the accrual axis ().

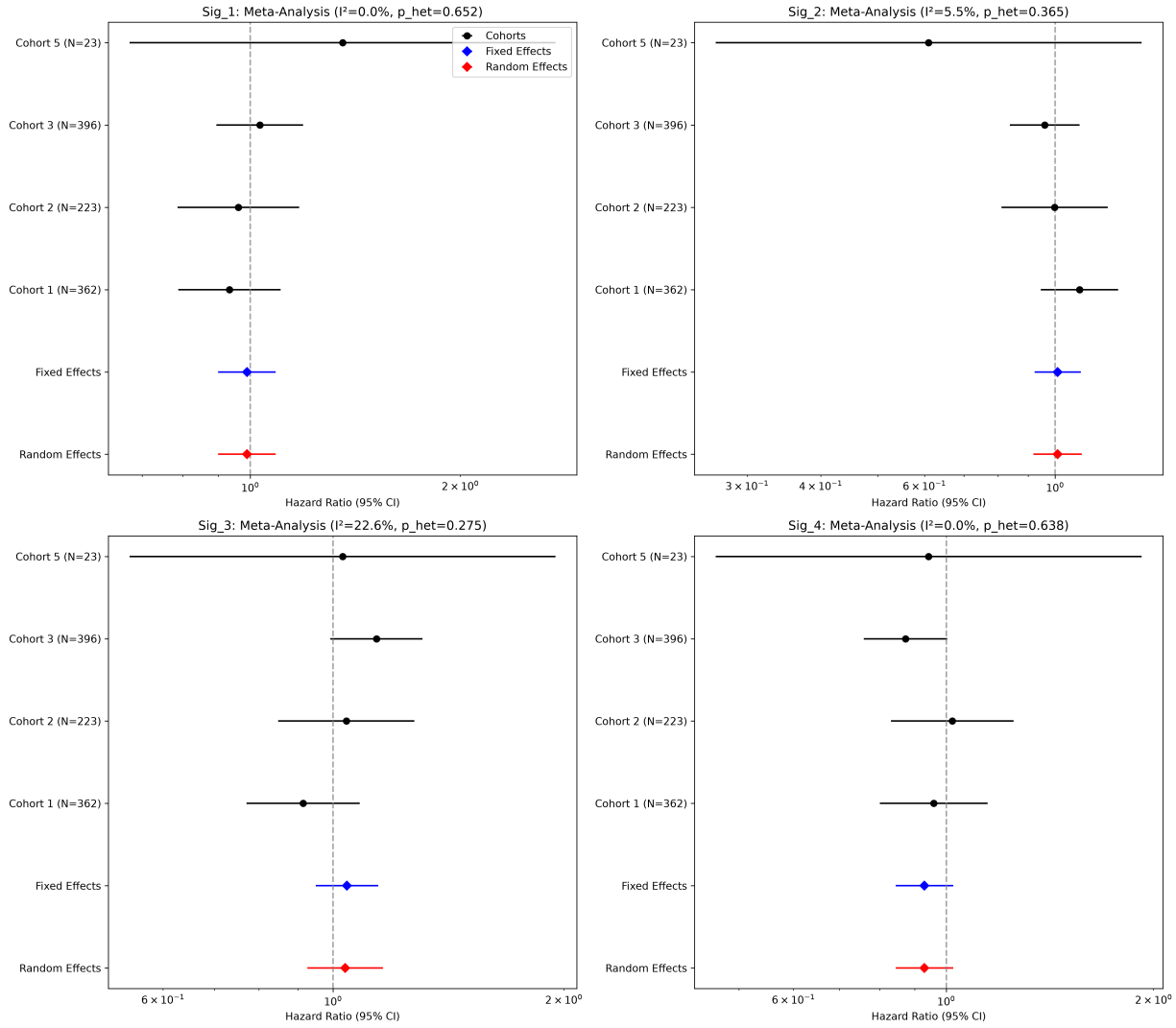


Figure 3: Forest plots presenting the multi-cohort meta-analysis of disease-specific mortality hazard ratios for the four genomically independent regulatory signatures (Sig 1, Sig 2, Sig 3, and Sig 4). Each panel displays the cohort-specific hazard ratios with 95% confidence intervals, alongside pooled estimates derived from Fixed Effects (blue diamonds) and Random Effects (red diamonds) models. The analysis reveals a consistent lack of prognostic signal across the accrual cohorts, as all pooled estimates cross the null threshold of unity. Furthermore, the heterogeneity between cohorts is low to moderate, with I^2 values ranging from 0% for Sig 1 and Sig 4 to a maximum of 22.6% for Sig 3 ($p_{\text{het}} > 0.27$). These low-heterogeneity null results demonstrate that while multi-layer regulatory deviations can be quantified, they do not provide independent predictive utility for tumor lethality beyond established clinical taxonomies.

Heterogeneity between the cohorts was low to moderate. Both Sig 1 and Sig 4 exhibited an I^2 of 0%, indicating perfect consistency of the null effect across all partitions. Sig 3 displayed the highest inter-cohort heterogeneity ($I^2 = 22.6\%$, $p_{\text{het}} = 0.275$), yet the pooled random-effects estimate remained entirely non-

significant (HR = 1.04, 95% CI: 0.93–1.16, $p = 0.531$). Similarly, Sig 2 showed minimal heterogeneity ($I^2 = 5.5\%$, $p_{\text{het}} = 0.365$) with a pooled HR of 1.01 (95% CI: 0.92–1.11, $p = 0.843$). These robust, low-heterogeneity results conclusively demonstrate that while cross-layer regulatory decoupling is a quantifiable feature of breast tumor biology, it does not constitute a reproducible, independent axis of lethality.

4 Conclusions

4.1 Summary of findings

In this study, we sought to determine whether the breakdown of cross-layer molecular control—specifically, the targeted under-expression of mRNA relative to genomic and epigenetic baselines—could serve as an independent predictor of breast cancer lethality. By integrating copy number, DNA methylation, and transcriptomic data across 1,417 primary tumors, we successfully isolated four genomically independent regulatory signatures. Biological characterization confirmed that these signatures capture a distinct layer of transcriptional variance, showing negligible overlap with canonical breast cancer gene panels such as PAM50 and remaining largely orthogonal to macroscopic tumor features.

Despite representing a measurable and distinct biological phenomenon, our rigorous survival framework conclusively demonstrated that these regulatory deviations do not provide independent prognostic value. Cause-specific Cox proportional hazards models, which explicitly adjusted for competing risks and established clinicopathological covariates, yielded hazard ratios consistently centered around unity. Furthermore, likelihood ratio tests confirmed that the integration of these signatures offered no incremental predictive benefit within specific intrinsic molecular subtypes, such as Luminal A or Basal-like tumors.

4.2 Robustness and clinical implications

The strength of these findings lies in their statistical stability. A multi-cohort meta-analysis across five independent accrual partitions revealed a highly consistent null effect, characterized by low to moderate inter-cohort heterogeneity. This robust replication confirms that the lack of prognostic signal is not an artifact of a specific sub-cohort, limited sample size, or statistical anomaly, but rather a genuine reflection of the underlying tumor biology.

Ultimately, these results suggest that while the decoupling of mRNA expression from structural and epigenetic regulation is a fundamental feature of tumorigenesis, the variance of this decoupling does not constitute an independent axis of lethality. The aggressive potential of breast tumors appears to be sufficiently captured by established macroscopic clinical markers—such as tumor stage, grade, and cellularity—alongside existing static molecular taxonomies.

Future investigations into regulatory failure may benefit from focusing on highly localized, gene-specific decoupling events at critical oncogenic loci, or by exploring alternative post-transcriptional and translational mechanisms, rather than extracting broad, multi-gene regulatory signatures. We conclude that cross-layer regulatory coupling variance, while biologically quantifiable and reflective of molecular disorganization, does not warrant integration into clinical prognostic models for breast cancer disease-specific survival.