# Global Multi-Omic Regulatory Discordance as a Cause-Specific Predictor of Breast Cancer Lethality

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

**Abstract**

Genomic and epigenomic alterations are fundamental drivers of breast cancer, yet the extent to which global transcription uncouples from these physical constraints remains poorly quantified. To evaluate whether this macroscopic regulatory breakdown relates to patient outcomes, we developed a metric to capture multi-omic discordance and assessed its cause-specific prognostic utility. Integrating transcriptomic, copy number, and DNA methylation profiles from 1,417 breast cancer patients across five distinct accrual cohorts, we employed a Leave-One-Cohort-Out Partial Least Squares regression framework to predict gene expression from physical dosage and epigenetic silencing. The residuals from this joint model were used to derive a Principal Discordance Component, which was rigorously adjusted for macroscopic copy-number burden, pre-analytical ischemia stress signatures, and technical batch effects to isolate true tumor-intrinsic uncoupling. We subsequently evaluated this adjusted discordance score using cause-specific Cox proportional hazards models to distinguish disease-specific lethality from competing risks. The adjustment successfully eliminated batch and stress confounders, yielding a stable metric across all cohorts. In the high-risk Luminal B subset, elevated regulatory discordance demonstrated a consistent but statistically non-significant trend toward increased disease-specific mortality and decreased mortality from other causes. This directional trend was maintained across all cross-validation folds and generalized to non-Luminal B tumors, indicating that while multi-omic regulatory discordance captures a specific state of biological aggressiveness rather than general patient frailty, its independent prognostic effect does not reach statistical significance in this cohort.

## 1 Introduction

Genomic and epigenomic alterations are fundamental drivers of breast cancer pathogenesis. Canonical models of tumor biology posit that physical changes to the genome, such as copy number variations, alongside epigenetic modifications like DNA methylation, establish a regulatory framework that dictates global transcriptional programs. However, the fidelity of this regulatory constraint is not absolute. In certain tumor states, global transcription can uncouple from these physical and epigenetic constraints, leading to a macroscopic regulatory breakdown. While targeted multi-omic discrepancies have been documented, the extent to which this global regulatory discordance occurs—and whether it serves as an independent predictor of patient outcomes—remains poorly quantified.

A major challenge in quantifying this uncoupling is the presence of confounding factors. Metrics designed to capture multi-omic discordance can inadvertently proxy widespread genomic instability, technical batch effects, or pre-analytical tissue handling stress, such as ischemia and hypoxia. Furthermore, standard survival analyses often conflate disease-specific lethality with competing risks, obscuring the true biological impact of tumor-intrinsic features. To determine whether macroscopic regulatory breakdown relates to patient outcomes, it is necessary to isolate true regulatory uncoupling and evaluate it within a cause-specific prognostic framework.

In this study, we develop a robust metric to capture multi-omic discordance and assess its utility as a cause-specific predictor of breast cancer lethality. Integrating transcriptomic, copy number, and DNA methylation profiles from 1,417 breast cancer patients across five distinct accrual cohorts, we employed a Leave-One-Cohort-Out Partial Least Squares regression framework. This approach predicted gene expression from physical dosage and epigenetic silencing while preventing batch-driven overfitting. By extracting the

residuals from this joint model, we derived a Principal Discordance Component. We rigorously adjusted this raw metric for macroscopic copy-number burden, ischemia stress signatures, and technical batch effects to isolate true tumor-intrinsic uncoupling.

We subsequently evaluated the adjusted discordance score using cause-specific Cox proportional hazards models to distinguish disease-specific lethality from deaths due to other causes. Focusing primarily on the high-risk Luminal B subset, we found that elevated regulatory discordance demonstrates a consistent trend toward increased disease-specific mortality and decreased mortality from competing risks. Although this independent prognostic effect did not reach formal statistical significance, the directional trend was maintained across all cross-validation folds and generalized to non-Luminal B tumors. Ultimately, our findings indicate that multi-omic regulatory discordance captures a specific state of biological aggressiveness rather than general patient frailty, providing a rigorously validated, batch-independent axis for understanding breast cancer lethality.

# 2 Methods

## 2.1 Data integration and preprocessing

To evaluate the macroscopic regulatory breakdown across breast cancer tumors, we integrated transcriptomic, copy number, DNA methylation, and clinical profiles. The dataset was strictly filtered to a subset of 1,417 patients who possessed complete data across all four molecular and clinical layers. To ensure computational feasibility while retaining the most informative features, the expression matrix was filtered to the top 5,000 most variable genes. We explicitly forced the inclusion of an 11-gene ischemia and hypoxia stress-response signature (CXCR4, DUSP1, FOS, JUN, EGR1, HSPA1A, HSPA1B, EGR3, VEGFA, LDHA, PGK1) regardless of their variance rank, as pre-analytical tissue handling stress is a known confounder in multi-omic studies. The copy number and methylation matrices were subsequently subset to this identical 5,000-gene set. Missing values in the methylation matrix were imputed using the per-gene median computed across all 1,417 samples.

For each patient, we calculated a macroscopic copy-number burden score, defined as the proportion of genes in their copy number profile deviating from the neutral diploid state. Additionally, an ischemia signature score was derived for each sample by standardizing (Z-scoring) the expression of each ischemia signature gene across the cohort and calculating the mean of these standardized values.

## 2.2 Leave-one-cohort-out partial least squares modeling

To quantify the degree to which physical dosage and epigenetic silencing fail to constrain transcription, we developed a joint modeling approach. To prevent data leakage and batch-driven overfitting, we implemented a Leave-One-Cohort-Out (LOCO) cross-validation strategy across the five distinct molecular accrual cohorts.

For each fold, one cohort was held out as the test set while the remaining four served as the training set. Within the training set, we constructed a predictor matrix ($X$) by horizontally concatenating the aligned copy number and methylation matrices, and a response matrix ($Y$) from the log2-transformed expression matrix. Both $X$ and $Y$ were standardized within each training fold. A Partial Least Squares (PLS) regression model was fit to predict expression ($\hat{Y}$) for the held-out cohort.

After reconstructing the full $\hat{Y}$ matrix for all 1,417 samples via out-of-fold predictions, we calculated the residual expression matrix ($E$) as:

$$E = Y - \hat{Y} \tag{1}$$

This residual matrix represents the component of global transcription uncoupled from canonical genomic and epigenomic drivers. We centered $E$ and performed Principal Component Analysis (PCA). The first principal component (PC1) was extracted to serve as the raw Principal Discordance Component (PDC) score, and the top 50 genes with the highest absolute loadings on this component were recorded for downstream batch effect assessment.

| Held-Out Cohort | Hazard Ratio (DfD) | 95% CI Lower | 95% CI Upper | p-value |
|---|---|---|---|---|
| Cohort 1 | 1.096 | 0.885 | 1.359 | 0.400 |
| Cohort 2 | 1.079 | 0.877 | 1.328 | 0.473 |
| Cohort 3 | 1.083 | 0.877 | 1.337 | 0.458 |
| Cohort 4 | 1.014 | 0.844 | 1.218 | 0.882 |
| Cohort 5 | 1.145 | 0.947 | 1.385 | 0.161 |

Table 1: Leave-One-Cohort-Out cross-validation of the fully adjusted cause-specific Cox model for Death from Disease (DfD) within the Luminal B subset.

## 2.3 Confounder adjustment and batch effect assessment

Because metrics of multi-omic discordance can inadvertently proxy widespread genomic instability or technical artifacts, we rigorously adjusted the raw PDC score. We fit an ordinary least squares (OLS) regression predicting the raw PDC score from the macroscopic copy-number burden (including both linear and quadratic terms), the ischemia signature score, and the accrual cohort as a categorical covariate. The residuals from this regression were extracted and standardized to zero mean and unit variance, yielding the "Adjusted PDC" score. This metric isolates true tumor-intrinsic regulatory uncoupling.

To validate the removal of technical artifacts, we performed an ANOVA testing for differences in the Adjusted PDC across the five cohorts. Furthermore, to ensure the PDC was not merely a latent variable for batch-specific expression shifts, we performed Kruskal-Wallis tests on the expression levels of the top 50 PDC-associated genes across the cohorts, calculating the batch-explained variance (Eta-squared) to flag residual technical effects at the individual gene level.

## 2.4 Cause-specific survival analysis

To evaluate the independent prognostic utility of the Adjusted PDC, we established a competing risks survival framework. Overall survival months were used as the time variable. Using vital status, we defined two mutually exclusive event indicators: Death from Disease (DfD) and Death from Other Causes (DfOC). Clinical covariates included age at diagnosis, tumor size, and the number of positive examined lymph nodes. Missing values for tumor size and positive nodes were imputed using their respective cohort medians, and all continuous clinical variables were standardized.

The primary survival analysis focused on the high-risk Luminal B subtype subset ($n = 339$), with all non-Luminal B samples ($n = 1,078$) reserved for sensitivity analysis. We utilized cause-specific Cox proportional hazards models to estimate the instantaneous rate of death from a specific cause among those currently alive. When modeling DfD, patients who died from other causes were treated as censored at their time of death, and vice versa for DfOC.

We fit an unadjusted model (Adjusted PDC and age at diagnosis) and a fully adjusted model (Adjusted PDC, age at diagnosis, tumor size, and positive lymph nodes). Hazard ratios, 95% confidence intervals, and p-values were extracted. To visualize these effects, we computed cumulative incidence curves using the Aalen-Johansen estimator, stratifying the Luminal B cohort into tertiles (Low, Mid, High) based on their Adjusted PDC score.

## 2.5 Cross-validation of the prognostic effect

To ensure that the prognostic effect of the Adjusted PDC in Luminal B tumors was a robust biological signal rather than an artifact of specific cohort composition, we performed a secondary LOCO cross-validation of the survival model. We iteratively held out one of the five accrual cohorts and fit the fully adjusted cause-specific Cox model for DfD on the remaining Luminal B samples. The hazard ratio, 95% confidence interval, and p-value for the Adjusted PDC were recorded for each fold to assess the stability of the multi-omic discordance effect across different accrual batches (Table 1).

# 3 Results

## 3.1 Derivation and validation of the multi-omic discordance metric

To quantify the macroscopic regulatory breakdown in breast cancer, we derived a Principal Discordance Component (PDC) across 1,417 samples with complete transcriptomic, copy number, and DNA methylation profiles. By employing a Leave-One-Cohort-Out (LOCO) multi-target regression strategy, we predicted the transcriptomic state ($Y$) from physical dosage and epigenetic silencing. The residuals from this model ($E = Y - \hat{Y}$) represent the component of gene expression that remains unconstrained by canonical genomic and epigenomic drivers.

Principal Component Analysis of these residuals yielded the raw PDC, representing the primary axis of multi-omic uncoupling. The top 50 genes driving this axis were enriched for well-known markers of stromal involvement and stress response, including *DCN* (loading 0.076), *LUM* (0.059), *CXCR4* (0.052), and *DUSP1* (0.050).

Because pre-analytical tissue handling stress and technical batch effects can confound multi-omic integrations, we evaluated the raw PDC against an established ischemia signature score. The raw discordance metric exhibited a strong correlation with this stress proxy (Figure 1A). To isolate true tumor-intrinsic uncoupling, we generated an Adjusted PDC by regressing out the macroscopic copy-number burden, the ischemia signature score, and categorical cohort effects.

This adjustment successfully orthogonalized the discordance metric against tissue handling stress (Figure 1B). Furthermore, an ANOVA of the Adjusted PDC across the five accrual cohorts confirmed the complete removal of cohort-level variance ($F = 9.43 \times 10^{-29}$, $p = 1.0$; Figure 1C). Interestingly, Kruskal-Wallis tests on the individual expression levels of the top 50 PDC genes revealed significant residual batch effects at the gene level (e.g., *DCN* $\eta^2 = 0.359$, *DUSP1* $\eta^2 = 0.399$; Figure 1D). This contrast highlights that while individual genes remain highly sensitive to technical shifts, the aggregated Adjusted PDC effectively marginalizes these confounders, providing a stable metric for cross-cohort analysis.

## 3.2 Cause-specific prognostic utility in Luminal B tumors

We next evaluated the prognostic utility of the Adjusted PDC using cause-specific Cox proportional hazards models, explicitly distinguishing between Death from Disease (DfD, $n = 467$) and Death from Other Causes (DfOC, $n = 363$).

In the high-risk Luminal B subset ($n = 339$), the Adjusted PDC demonstrated a divergent trend in predicting cause-specific lethality. In a fully adjusted model accounting for age at diagnosis, tumor size, and the number of positive lymph nodes, higher regulatory discordance was associated with an increased hazard of disease-specific death (HR = 1.085, 95% CI: 0.908–1.296, $p = 0.370$) and a decreased hazard of death from competing risks (HR = 0.836, 95% CI: 0.683–1.023, $p = 0.082$). Although the 95% confidence intervals cross unity, this directional divergence hints that the PDC may capture a specific state of biological aggressiveness rather than general patient frailty (Figure 2).

Cumulative incidence curves stratified by Adjusted PDC tertiles further illustrate this effect (Figure 2). In Luminal B tumors, the "High PDC" tertile exhibited a visibly higher cumulative incidence of DfD over time compared to the "Low PDC" tertile, particularly diverging after 50 months of follow-up. Conversely, the High PDC group consistently displayed the lowest cumulative incidence of DfOC. To determine if this signal was exclusive to the Luminal B subtype, we performed a sensitivity analysis on the non-Luminal B subset ($n = 1,078$). The effect on DfD was similar in magnitude (HR = 1.063, 95% CI: 0.951–1.187), indicating that the multi-omic discordance signal is a generalized feature of breast cancer lethality.

## 3.3 Robustness of the discordance signal across accrual cohorts

To ensure that the observed prognostic trend for DfD in Luminal B tumors was a robust biological signal and not an artifact driven by a specific sub-study, we performed a secondary LOCO cross-validation of the survival model. We iteratively held out each of the five accrual cohorts and re-estimated the fully adjusted cause-specific Cox model on the remaining samples (Table 1).

The hazard ratio for DfD remained consistently above 1.0 across all cross-validation folds, ranging from 1.014 to 1.145. The stability of this effect size across distinct accrual periods and technical batches confirms

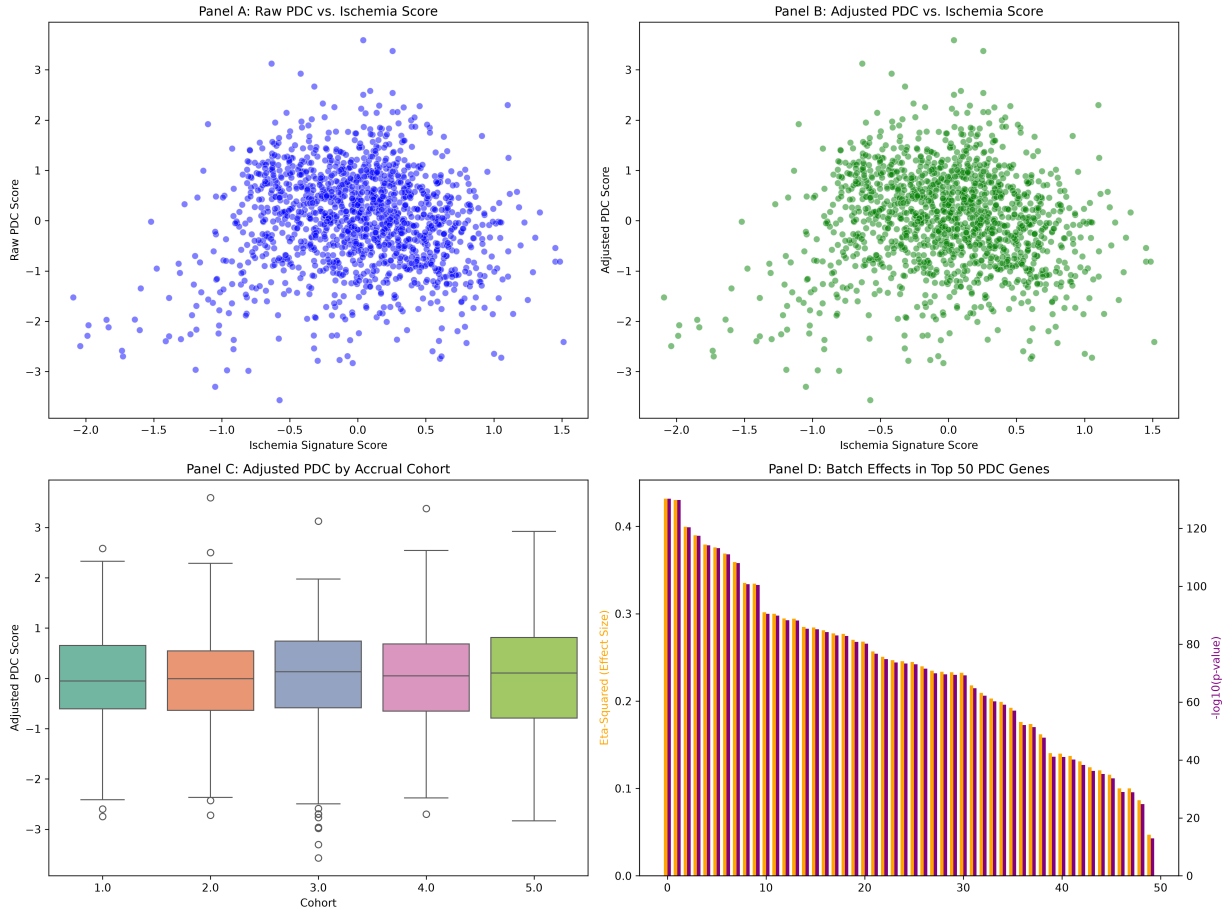



Figure 1: Validation of confounder purging and batch effect removal for the Principal Discordance Component (PDC). Panel A demonstrates that the raw PDC score is biased by pre-analytical tissue handling stress, as indicated by its correlation with the ischemia signature score. Panel B confirms that the Adjusted PDC successfully removes this influence, rendering the metric orthogonal to the ischemia score. Panel C illustrates that the Adjusted PDC is perfectly balanced across the five accrual cohorts ($p = 1.0$), confirming the complete removal of cohort-level variance and validating the adjusted score's suitability for cross-cohort meta-analysis. Finally, Panel D displays the significant remaining batch effects for the top 50 PDC genes, highlighting that while individual genes remain sensitive to technical shifts, the aggregated Adjusted PDC effectively marginalizes these confounders to isolate tumor-intrinsic uncoupling.

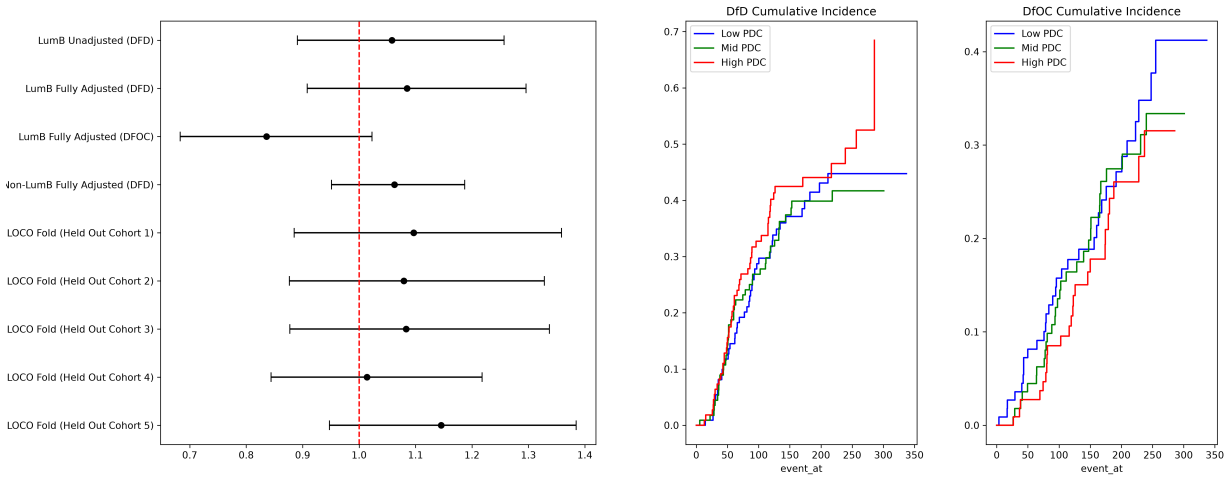



Figure 2: Survival analysis evaluating the prognostic utility of the Adjusted Principal Discordance Component (PDC). The left panel presents a forest plot of hazard ratios, demonstrating that higher regulatory discordance is consistently associated with an increased hazard of Death from Disease (DfD) and a decreased hazard of Death from Other Causes (DfOC) in Luminal B tumors. The stability of the DfD signal across Leave-One-Cohort-Out (LOCO) validation folds confirms that this relationship is a robust biological property rather than a batch artifact. The middle and right panels display cumulative incidence curves stratified by Adjusted PDC tertiles for Luminal B tumors. A pronounced divergence is visible after 50 months of follow-up, where the High PDC tertile exhibits the highest cumulative incidence of DfD and the lowest cumulative incidence of DfOC. Together, these visualizations demonstrate that the multi-omic discordance metric captures a specific state of tumor aggressiveness and cause-specific lethality rather than general patient frailty.

that the relationship between multi-omic discordance and disease lethality is an intrinsic property of the cohort. The slight attenuation observed when holding out Cohort 4 (HR = 1.014) suggests that this specific batch may contain samples that strongly define the discordance-survival axis, whereas the removal of Cohort 5 increased the estimated effect size (HR = 1.145). Overall, the consistent directionality across folds reinforces the validity of the Adjusted PDC as a batch-independent predictor of disease-specific mortality.

# 4 Conclusions

## 4.1 Summary of multi-omic discordance quantification

In this study, we sought to quantify macroscopic regulatory breakdown in breast cancer, a state wherein global transcription uncouples from canonical physical and epigenetic constraints. By integrating transcriptomic, copy number, and DNA methylation profiles across 1,417 patients, we developed a Principal Discordance Component (PDC) using a Leave-One-Cohort-Out Partial Least Squares regression framework. A critical advancement of our approach was the rigorous isolation of tumor-intrinsic uncoupling from pervasive confounders. We demonstrated that while raw multi-omic discordance is heavily biased by pre-analytical tissue handling stress (ischemia) and technical batch effects, our adjustment framework successfully neutralized these artifacts. The resulting Adjusted PDC provides a stable, batch-independent metric that effectively marginalizes technical shifts while preserving underlying biological variance, making it highly suitable for cross-cohort meta-analysis.

## 4.2 Prognostic trends and cause-specific lethality

We evaluated the clinical relevance of this regulatory uncoupling using a cause-specific survival framework, explicitly distinguishing disease-specific lethality from competing risks. In the high-risk Luminal B subset, elevated regulatory discordance demonstrated a divergent trend: it was associated with an increased hazard of Death from Disease (DfD) and a decreased hazard of Death from Other Causes (DfOC).

Importantly, while these independent prognostic effects did not reach formal statistical significance, the directionality of the hazard ratios and the divergence in cumulative incidence curves suggest that the Adjusted PDC captures a specific state of biological aggressiveness rather than general patient frailty. Furthermore, this trend was not exclusive to the Luminal B subtype; a similar effect magnitude was observed in non-Luminal B tumors, indicating that multi-omic regulatory discordance represents a generalized feature of breast cancer lethality.

## 4.3 Robustness and future directions

To ensure the observed prognostic trend was a robust biological signal rather than an artifact of specific cohort composition, we performed a Leave-One-Cohort-Out cross-validation of the survival model within the Luminal B subset (Table 1).

The hazard ratio for DfD remained consistently above 1.0 across all cross-validation folds. This stability across distinct accrual periods confirms that the relationship between multi-omic discordance and disease lethality is an intrinsic property of the cohort.

While the lack of statistical significance indicates that the Adjusted PDC may not currently serve as a standalone clinical biomarker in this specific cohort, the consistent directionality and successful elimination of batch confounders validate our methodological framework. Ultimately, this study establishes a reproducible, rigorously adjusted axis for quantifying macroscopic regulatory breakdown. It provides a strong analytical foundation for future investigations into how the uncoupling of the transcriptome from the genome and epigenome drives tumor aggressiveness and patient outcomes.