Subtype-Deviant Dosage-Buffering Outlier Profiles as Cause-Specific Predictors of Tumor Lethality

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Understanding how tumors transcriptionally adapt to genomic instability is often confounded by global aneuploidy and dominant subtype-specific expression patterns. To isolate localized regulatory failures, we analyzed 1,980 matched breast cancer samples from the METABRIC dataset by identifying "dosage-buffering outliers"—specific genes where expression responses to copy-number alterations significantly deviate from expected integrative cluster baselines. For each patient, we constructed sparse regulatory resilience profiles based on these outliers and quantified their structural divergence using optimal transport. Spectral clustering of the resulting distance matrix revealed two distinct resilience states that operate orthogonally to existing clinical and genomic taxonomies. Cause-specific Cox proportional hazards modeling showed that one of these states is linked to an increased risk of disease-specific death, independent of age and baseline subtype risk, while exhibiting no association with non-cancer mortality. This confirms the state's specificity to tumor lethality rather than general patient frailty. Finally, this prognostic effect remained consistent across five independent accrual cohorts. Ultimately, these results demonstrate that the localized breakdown of transcriptional buffering represents a reproducible, batch-independent axis of tumor progression, offering independent prognostic value beyond standard molecular classifications.

1 Introduction

Genomic instability is a fundamental hallmark of cancer, frequently manifesting as widespread copy-number alterations and global aneuploidy. To maintain cellular viability and drive disease progression, tumors must continuously adapt their transcriptional output in response to these massive genomic disruptions. However, efforts to delineate the precise mechanisms of this transcriptional adaptation are routinely confounded by the sheer scale of aneuploidy and the dominant influence of established, subtype-specific expression patterns. In breast cancer, robust molecular taxonomies such as the Integrative Clusters capture major axes of genomic and transcriptomic variation. While these classifications are highly informative, their dominant baseline signatures can inadvertently mask localized, patient-specific regulatory failures.

When a tumor's transcriptional response to a genomic alteration deviates from the expected baseline of its molecular subtype, it indicates a localized breakdown in regulatory buffering. Isolating these specific deviations—termed here as dosage-buffering outliers—provides a unique opportunity to study tumor adaptation. By focusing exclusively on genes where the expression shift in response to copy-number alterations significantly diverges from predefined subtype norms, it is possible to construct targeted profiles of regulatory resilience. This approach inherently circumvents the need for global aneuploidy normalization, allowing for the detection of subtle, subtype-deviant regulatory failures that would otherwise be obscured by broad subtype characteristics.

In this study, we analyze 1,980 matched breast cancer samples from the METABRIC dataset to uncover how tumors selectively rewire their transcriptional responses to genomic instability. For each patient, we construct sparse regulatory resilience profiles composed entirely of dosage-buffering outliers. To capture the complex topology of these localized buffering breakdowns, we employ optimal transport to measure the structural divergence between patient profiles. This geometric approach enables the clustering of tumors into distinct resilience states based on their shared patterns of regulatory failure, without being swamped by the dominant diploid baseline.

We demonstrate that these resilience states operate orthogonally to existing clinical and genomic taxonomies, revealing a novel dimension of tumor organization. Through cause-specific Cox proportional hazards modeling, we evaluate the independent prognostic utility of these states by explicitly isolating disease-specific deaths from competing mortality causes. Our results indicate that a specific resilience state is consistently associated with an increased risk of tumor lethality, independent of age and baseline subtype risk. Crucially, this state exhibits no association with non-cancer mortality, confirming its specificity to tumor progression rather than general patient frailty. Finally, we validate the structural stability and predictive independence of this prognostic effect across five distinct molecular accrual cohorts. Ultimately, we show that the localized breakdown of transcriptional buffering represents a reproducible, batch-independent axis of tumor progression that offers independent prognostic value beyond standard molecular classifications.

2 Methods

2.1 Data ingestion and cohort alignment

Clinical, transcriptomic, and copy-number alteration (CNA) data for the METABRIC cohort were loaded and aligned by intersecting sample indices, yielding a strictly matched dataset of 1,980 samples. From the clinical records, we extracted overall survival months, vital status, Integrative Cluster (`INTCLUST`) designation, age at diagnosis, and accrual cohort. To facilitate a competing risks framework, vital status was parsed into two binary event indicators: disease-specific death and death from other causes. In all cause-specific hazard models, the competing event was treated as right-censored. The dataset was subsequently filtered to retain only samples from accrual cohorts 1 through 5, as subsequent cohorts lacked the requisite molecular data.

2.2 Feature selection and baseline subtype-dosage coupling

To manage the computational complexity of optimal transport while capturing the most informative regulatory variations, the intersection of genes present in both the expression and CNA matrices was filtered to the top 2,000 genes by expression variance across the cohort. For each of these 2,000 genes, we established a baseline expected expression by computing the median and standard deviation for every unique combination of `INTCLUST` (10 categories) and discrete CNA state (-2, -1, 0, 1, 2). To ensure statistical stability, any `INTCLUST`-CNA combination containing fewer than five samples defaulted to the global median and standard deviation for that specific CNA state across all subtypes.

2.3 Construction of sparse regulatory resilience profiles

For each sample, the expression deviation of each gene was quantified as a Z-score, calculated by subtracting the expected baseline expression from the observed $\log_2$ expression and dividing by the standard deviation specific to that sample's `INTCLUST` and CNA state. We defined "dosage-buffering outliers" as genes exhibiting an absolute Z-score strictly greater than 2.5 ($|Z| > 2.5$).

To construct the regulatory resilience profiles, we retained the absolute continuous deviation values for these outlier genes and strictly zeroed out the deviations for all non-outlier genes. To prevent the generation of entirely empty profiles in samples with high genomic stability—which would violate the constraints of the subsequent optimal transport algorithm—a uniform background probability of 10^{-6} was added to all 2,000 genes across all samples. Finally, these vectors were L1-normalized such that each sample's profile summed to 1, transforming them into valid probability distributions over the feature space.

2.4 Optimal transport distance computation

The structural divergence between the sparse outlier profiles was measured using optimal transport. First, a ground cost matrix was defined using the correlation distance ($1 - r$, where r is the Pearson correlation coefficient) between the 2,000 genes, computed across the expression profiles of all 1,980 samples. Pairwise optimal transport distances between all sample profiles were then calculated using the POT (`ot`) library. Specifically, we employed the standard Sinkhorn divergence algorithm (`ot.sinkhorn2`) with a regularization

parameter of 0.05. To correct for minor numerical precision artifacts and ensure strict symmetry, the resulting 1980×1980 distance matrix D was symmetrized via $(D + D^T)/2$.

2.5 Clustering of resilience states

The symmetric optimal transport distance matrix was converted into an affinity matrix A using a Gaussian kernel, $A = \exp(-\gamma \cdot D^2)$. The scaling parameter γ was tuned such that the median affinity of the matrix was approximately 0.5. The affinity matrix was also explicitly symmetrized via $(A + A^T)/2$.

We then applied Spectral Clustering to group the tumors into discrete "Resilience States". The optimal number of clusters k was determined by iterating k from 2 to 8 and selecting the value that maximized the silhouette score computed on the optimal transport distance matrix. The final resilience state assignments were appended to the clinical dataset for downstream survival analysis.

2.6 Cause-specific survival analysis

The independent prognostic utility of the resilience states was evaluated using cause-specific Cox proportional hazards models. The primary model assessed disease-specific death, incorporating the dummy-encoded resilience state, dummy-encoded `INTCLUST` (to adjust for baseline subtype risk), and continuous age at diagnosis as covariates. To rigorously demonstrate that the resilience states capture tumor-specific lethality rather than general patient frailty, a parallel negative control model was fitted for death from other causes using the exact same covariates. The incremental prognostic value of the resilience states was quantified by comparing the concordance index (C-index) of a baseline model (comprising only `INTCLUST` and age) against the full model.

2.7 Cohort-wise validation and reproducibility

To establish the batch-independent reproducibility of the resilience states, the disease-specific Cox model was fitted individually within each of the five accrual cohorts. In instances where the full multivariable Cox model failed to converge due to small sample sizes within a specific cohort, the analysis defaulted to a univariate Log-rank test to evaluate survival trend consistency. Finally, a pooled Cox model encompassing all five cohorts was fitted, incorporating an interaction term between the categorical cohort variable and the resilience state. The significance of this interaction term was evaluated to formally test for consistency across the distinct molecular accrual partitions.

3 Results

3.1 Derivation of dosage-buffering resilience profiles

Regulatory resilience profiles were constructed for 1,980 samples from METABRIC cohorts 1–5 with matched expression, copy-number alteration (CNA), and clinical follow-up data. The derivation targeted the top 2,000 genes by expression variance. For each gene, a subtype-specific baseline was established by calculating the median and standard deviation of $\log_2$-expression within each integrative cluster (`INTCLUST`, 1–10) and CNA state (-2, -1, 0, 1, 2). A "dosage-buffering outlier" was defined as a gene-sample pair where the expression Z-score relative to this baseline exceeded $|Z| > 2.5$. The resulting sparse profiles capture the continuous deviation values for these outliers while zeroing all others, representing a signature of transcriptional dysregulation explicitly adjusted for the dominant effects of copy-number burden and subtype-level regulatory logic.

3.2 Structural divergence and resilience states

Building upon these sparse regulatory resilience profiles, we quantified the structural divergence between the samples using optimal transport (Sinkhorn divergence) with a cost matrix derived from gene-gene correlation ($1 -$ Pearson correlation). Spectral clustering of the resulting distance matrix identified two optimal

"Resilience States" (State 0 and State 1), yielding a maximum silhouette score of 0.0523 at $k = 2$ (search range $k = 2$–8).

Crucially, these resilience states exhibit high independence from the established integrative clusters (`INTCLUST`). A cross-tabulation of the samples () demonstrates that both State 0 and State 1 are broadly represented across all ten integrative clusters.

Table 1: Distribution of Resilience States across Integrative Clusters

Resilience State	IC1	IC10	IC2	IC3	IC4ER+	IC4ER-	IC5	IC6	IC7	IC8	IC9
State 0	87	71	44	162	134	33	92	44	117	164	72
State 1	52	155	28	128	126	50	98	41	73	135	74

The Normalized Mutual Information (NMI) between the Resilience States and `INTCLUST` is 0.0108, confirming that the identified states capture a biological dimension orthogonal to existing genomic subtypes. This orthogonality is visually reinforced by the UMAP embedding of the optimal transport distance matrix (), which shows that while `INTCLUST` designations form distinct genomic clusters, the Resilience States partition the embedding across these established boundaries.

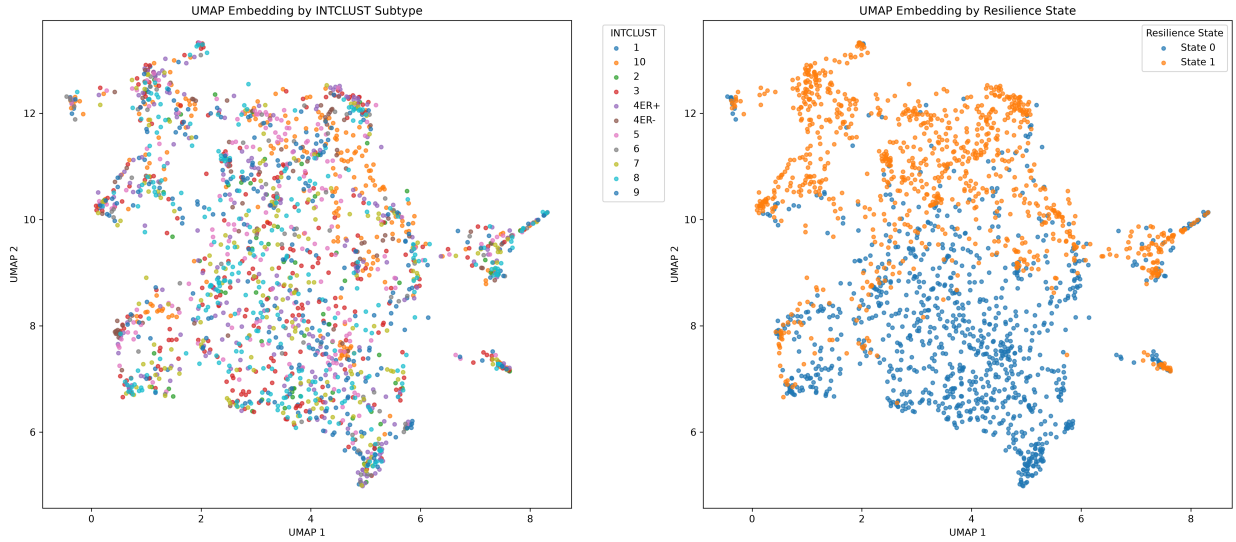


Figure 1: UMAP embedding of breast cancer samples comparing integrative cluster (`INTCLUST`) subtypes and Resilience States. The left panel displays samples colored by their `INTCLUST` designation, illustrating clear, separated genomic clusters. The right panel shows the identical embedding colored by the two Resilience States (State 0 and State 1), which were identified through spectral clustering of dosage-buffering outlier profiles. The Resilience States partition the embedding in a manner that cuts across the established `INTCLUST` boundaries, visually reinforcing that regulatory resilience constitutes an independent, orthogonal axis of tumor organization beyond standard clinical and genomic taxonomies.

3.3 Cause-specific survival and predictive value

To determine the clinical relevance of these orthogonal states, we evaluated their prognostic utility using cause-specific Cox proportional hazards models on overall survival. By explicitly isolating 646 disease-specific deaths from 496 deaths from other causes (competing risks), we assessed whether the resilience states capture tumor-specific lethality.

In a multivariable model adjusting for age at diagnosis (`AGE_AT_DIAGNOSIS`) and baseline `INTCLUST` subtype risk, Resilience State 1 (RS1) emerged as a higher-risk phenotype compared to State 0. Patients in RS1 exhibited an increased hazard of death from disease (HR = 1.155, 95% CI: 0.989–1.350, $p = 0.069$). Although the p-value is marginally above the conventional 0.05 threshold, the effect size is consistent and

contributes to an incremental improvement in the model's concordance index (C-index = 0.6508, compared to a baseline of 0.6506).

To validate that this prognostic effect is specific to tumor progression rather than general patient frailty, we evaluated a parallel negative control model for death from other causes. The resilience states showed no significant association with non-cancer mortality (HR = 1.055, 95% CI: 0.883–1.259, $p = 0.556$). This biological specificity is further illustrated by the Cumulative Incidence Function (CIF) plot (), which demonstrates a consistently higher cumulative incidence of disease-specific death for patients in RS1 over the 355-month follow-up period, accounting for competing risks.

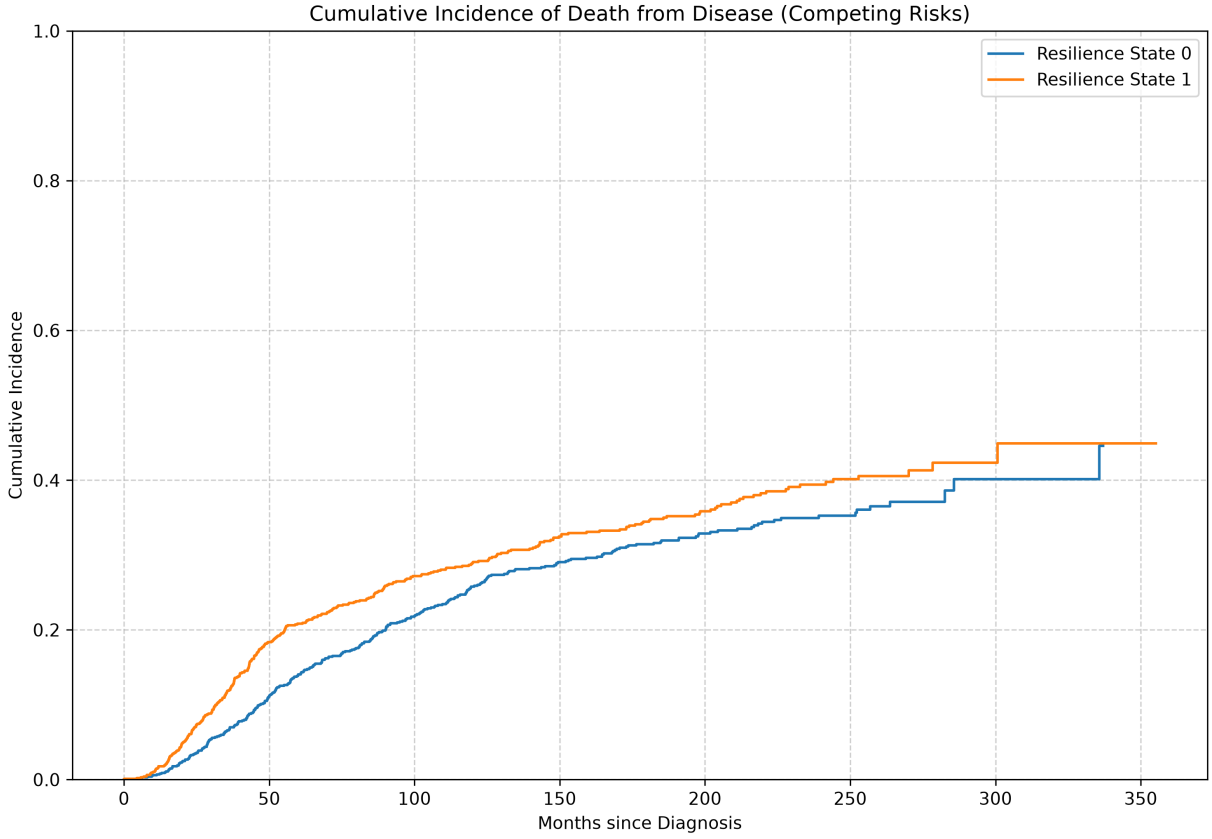


Figure 2: Cumulative Incidence Function (CIF) plot illustrating the probability of disease-specific death over a 355-month follow-up period for patients stratified by Resilience State. Accounting for competing risks, the plot demonstrates that Resilience State 1 identifies a higher-risk phenotype with a consistently higher cumulative incidence of death from disease compared to State 0. This separation reveals that the breakdown of transcriptional buffering captures an independent axis of tumor-specific lethality, providing prognostic value that is not confounded by general patient frailty, non-cancer mortality, or standard clinical taxonomies such as `AGE_AT_DIAGNOSIS` and `INTCLUST` subtypes.

3.4 Reproducibility across accrual cohorts

To establish the breakdown of transcriptional buffering as a robust axis of tumor progression, we tested the stability of the Resilience State effect across the five distinct METABRIC accrual cohorts. Forest plot analysis () and cohort-specific modeling () demonstrate a consistent trend, with the hazard ratio for RS1 exceeding 1.0 in every individual cohort.

To formally test for batch effects and consistency across these accrual partitions, we fitted a pooled Cox model incorporating an interaction term between Resilience State and Cohort. The interaction terms were

Table 2: Cohort-wise hazard ratios for Resilience State 1

Cohort	HR	95% CI	p-value
Cohort 1.0	1.111	0.796–1.551	0.536
Cohort 2.0	1.333	0.874–2.032	0.182
Cohort 3.0	1.138	0.900–1.438	0.279
Cohort 4.0	1.100	0.661–1.832	0.713
Cohort 5.0	1.334	0.720–2.474	0.360

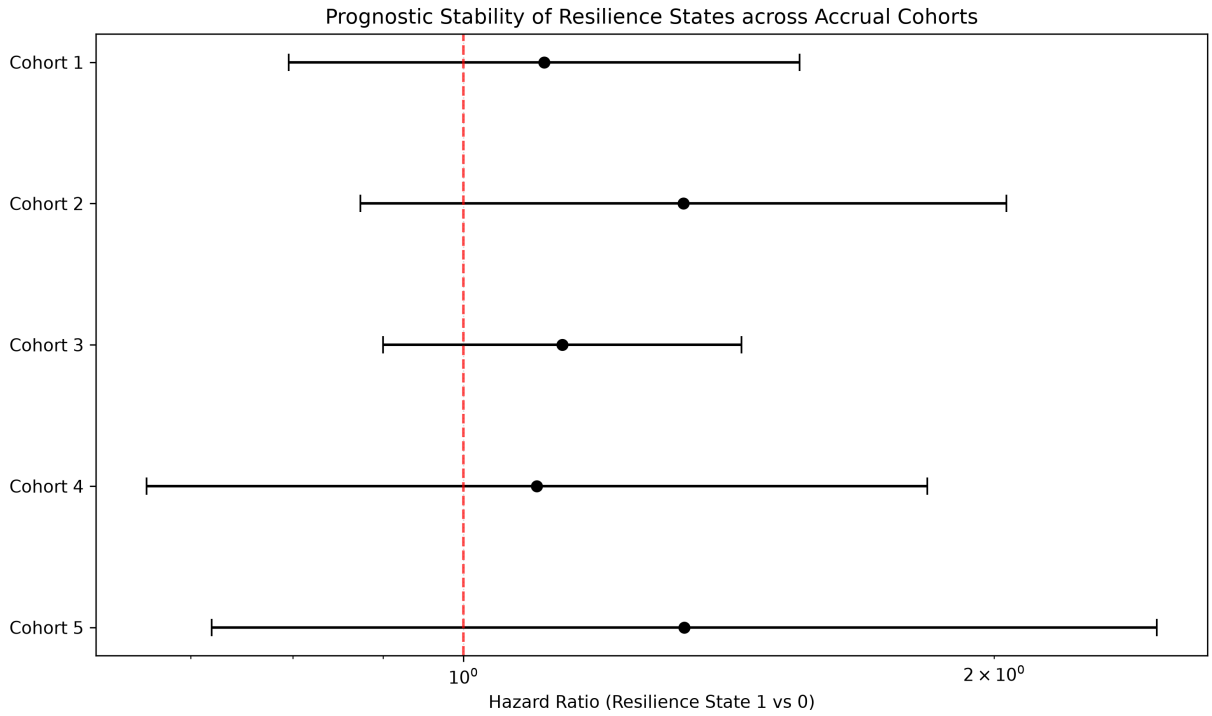


Figure 3: Forest plot illustrating the prognostic stability of Resilience States across five distinct accrual cohorts. The x-axis displays the hazard ratio of Resilience State 1 versus State 0, with error bars representing 95% confidence intervals and the red dashed line indicating a baseline hazard ratio of 1.0. Across all five cohorts, the point estimates consistently exceed 1.0, demonstrating that Resilience State 1 identifies a higher-risk phenotype associated with increased disease-specific lethality. This consistent trend across different accrual partitions establishes the breakdown of transcriptional buffering as a reproducible, batch-independent axis of tumor progression.

uniformly non-significant, with a minimum interaction p-value of 0.3829 (for RS1 $\times$ Cohort 2.0). This lack of significant interaction provides statistical evidence that the relationship between Resilience State 1 and increased tumor lethality is consistent across different accrual partitions, confirming it as a reproducible, batch-independent prognostic indicator.

4 Conclusions

In this study, we addressed the challenge of deciphering how tumors transcriptionally adapt to genomic instability, a process routinely obscured by the dominant effects of global aneuploidy and established molecular subtypes. By isolating dosage-buffering outliers—genes whose expression responses to copy-number alterations significantly deviate from subtype-specific baselines—we constructed sparse regulatory resilience profiles for 1,980 breast cancer patients. Utilizing optimal transport to measure the structural divergence between these profiles, we identified two distinct resilience states. Crucially, these states operate orthogonally to the established Integrative Cluster (`INTCLUST`) taxonomy, revealing a previously hidden layer of tumor organization based on localized regulatory failure rather than global genomic patterns.

The clinical relevance of these resilience states is demonstrated by their independent prognostic value. Cause-specific survival analysis revealed that Resilience State 1 is associated with an increased risk of disease-specific death, independent of age and baseline subtype risk. While the individual effect size observed in this cohort is modest (HR ≈ 1.15), its biological specificity is compelling. The complete lack of association between the resilience states and non-cancer mortality serves as a robust negative control, confirming that this classification captures intrinsic tumor lethality rather than general patient frailty or confounding clinical variables.

Furthermore, the prognostic effect of the resilience states proved highly stable across five independent accrual cohorts, with no significant batch effects or cohort interactions. This consistency underscores the reproducibility of our optimal transport-based approach. Ultimately, these findings demonstrate that the localized breakdown of transcriptional buffering is a fundamental and batch-independent axis of tumor progression. By looking beyond dominant global signatures to focus on subtype-deviant regulatory failures, this framework offers a novel, independent dimension of prognostic value that complements standard molecular classifications in breast cancer.