Global Regulatory Efficiency Index via Burden-Normalized Residual Variance as a Cause-Specific Predictor of Lethality in Breast Cancer

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Genomic instability profoundly alters the transcriptional landscape of tumors, but quantifying a tumor's ability to buffer these aneuploidy-induced changes remains challenging due to confounding by the sheer volume of copy number alterations. To address this, we developed the Regulatory Efficiency Index (REI), a metric designed to measure how effectively a tumor buffers its specific level of aneuploidy. Using 1,980 matched breast cancer samples from the METABRIC cohort, we regressed microarray gene expression against discrete copy number states to extract dosage-independent expression residuals. The per-sample variance of these residuals was then normalized by the tumor's global macroscopic copy number burden using non-parametric LOESS regression. We evaluated the independent prognostic utility of the index using cause-specific Cox proportional hazards models, adjusting for baseline clinicopathological covariates and integrative cluster subtypes, and validated its stability across five distinct molecular accrual partitions. The index demonstrated high internal consistency across diverse gene sets and exhibited statistical stability across accrual cohorts with zero significant heterogeneity. However, multivariable survival analysis revealed that the index was not a statistically significant independent predictor of disease-specific lethality (HR = 0.729, p = 0.4957). While the metric showed an association with non-cancer mortality (HR = 0.302, p = 0.0211), the formal difference between disease-specific and other-cause hazard coefficients was not statistically significant (p = 0.2051). Ultimately, while burden-normalized regulatory efficiency is a highly reproducible state variable in breast cancer, it does not independently drive disease-specific survival.

1 Introduction

Genomic instability is a fundamental hallmark of cancer, frequently manifesting as widespread copy number alterations (CNAs) and aneuploidy. These structural aberrations profoundly alter the transcriptional landscape of tumors, driving oncogenesis, tumor evolution, and therapeutic resistance. However, tumors exhibit varying capacities to buffer these aneuploidy-induced transcriptional changes. While some tumors tightly regulate their expression profiles despite massive genomic disruption, others experience widespread transcriptional dysregulation. Quantifying this regulatory efficiency—a tumor's intrinsic ability to maintain transcriptional homeostasis in the face of genomic instability—remains a significant challenge. Traditional approaches are often confounded by the sheer volume of copy number alterations, making it difficult to isolate the regulatory buffering capacity from the macroscopic burden of genomic instability itself.

To robustly quantify macroscopic transcriptional control without confounding from the sheer volume of genomic instability, we developed the Regulatory Efficiency Index (REI). This metric is designed to measure how effectively a tumor buffers its specific level of aneuploidy, reframing expression-CNA discordance into a biologically interpretable scalar value. By regressing microarray gene expression against discrete copy number states, we extract dosage-independent expression residuals. The per-sample variance of these residuals is then explicitly normalized by the tumor's total macroscopic copy number burden using non-parametric LOESS regression. This approach avoids the overfitting and circularity inherent in subgroup-stratified dependency metrics, providing a global baseline of expected transcriptional variance for any given level of genomic instability.

In this study, we apply the REI framework to 1,980 matched breast cancer samples from the METABRIC cohort to establish whether genome-wide regulatory efficiency constitutes a reproducible, batch-independent

axis of breast cancer lethality. To ensure the extracted signal provides orthogonal clinical value rather than merely proxying established structural taxonomies, we evaluate the independent prognostic utility of the REI using cause-specific Cox proportional hazards models. This framework explicitly isolates disease-specific deaths from deaths due to other causes, systematically adjusting for baseline clinicopathological covariates such as age at diagnosis, histological grade, tumor size, and precomputed integrative cluster (INTCLUST) categories. Furthermore, we rigorously validate the statistical stability of the REI across five distinct molecular accrual partitions to confirm its robustness as a biological state variable.

Our analysis reveals that burden-normalized regulatory efficiency is a highly reproducible and internally consistent metric across diverse gene sets. Furthermore, the REI exhibits remarkable statistical stability across different accrual cohorts, with zero significant heterogeneity observed in meta-analysis. However, multivariable survival analysis demonstrates that the index is not a statistically significant independent predictor of disease-specific lethality (HR = 0.729, p = 0.4957). Interestingly, while the metric shows an association with non-cancer mortality (HR = 0.302, p = 0.0211), the formal difference between disease-specific and other-cause hazard coefficients does not reach statistical significance (p = 0.2051). Ultimately, we conclude that while the Regulatory Efficiency Index successfully captures a robust and reproducible dimension of transcriptional buffering in breast cancer, it does not independently drive disease-specific survival.

2 Methods

2.1 Data acquisition and preprocessing

Clinical, microarray gene expression (log2 transformed), and discrete copy number alteration (CNA) data were obtained for the METABRIC cohort. Samples were strictly intersected across all three data modalities to ensure perfect alignment across the high-dimensional matrices, yielding a final cohort of 1,980 matched samples. The clinical dataset was subset to match these exact samples prior to downstream analysis.

2.2 Gene selection and raw residual variance

To ensure the derived metric captures a genome-wide regulatory efficiency signal that is robust to gene selection, four distinct sets of 2,000 genes were defined from the intersected gene space. Three sets were selected randomly using fixed seeds (42, 123, and 999), and a fourth set comprised the 2,000 genes exhibiting the highest expression variance across the cohort. For each gene set, an ordinary least squares (OLS) regression was fitted to predict each gene's log2 expression from its discrete CNA state across all samples. The residuals from these models represent dosage-independent expression. For each sample, the raw residual variance (RRV) was computed as the variance of these residuals across the 2,000 genes. The Pearson correlation of the RRVs across the four sets was calculated to confirm the internal consistency and stability of the transcriptional buffering signal.

2.3 Global burden calculation and regulatory efficiency index

The global macroscopic copy-number burden for each sample was calculated as the mean absolute CNA value across all intersected genes, providing a continuous measure of the magnitude of genomic instability. To establish a global baseline of expected transcriptional variance for any given level of genomic instability, a non-parametric LOESS (Locally Estimated Scatterplot Smoothing) regression was applied to predict the expected RRV from the global CNA burden across the entire cohort. The primary Regulatory Efficiency Index (REI) was defined as the residual from this global LOESS fit:

$$\mathrm{REI} = \mathrm{RRV}_{\mathrm{observed}} - \mathrm{RRV}_{\mathrm{expected}}$$

This LOESS normalization was repeated for the other three gene sets. Additionally, a secondary metric, $\mathrm{REI}_{\mathrm{stratified}}$, was calculated by performing the LOESS normalization separately within each precomputed integrative cluster (INTCLUST) category to test whether the regulatory signal was cluster-intrinsic or cluster-independent.

2.4 Cause-specific survival and competing risks analysis

Survival outcomes were parsed to handle competing risks, defining two distinct event types: disease-specific death and death from other causes, with overall survival months as the time variable. The cohort was filtered to include only samples with complete data for all planned covariates (age at diagnosis, INTCLUST, histological grade, and tumor size), leaving 1,875 samples for multivariable modeling. The cohort was dichotomized into high and low REI groups based on the median primary REI. Cumulative Incidence Functions (CIF) for both event types were estimated using the Aalen-Johansen estimator and stratified by REI group to visualize lethality and assess informative censoring.

Multivariable Cox proportional hazards models were fitted to predict disease-specific death, adjusting for the primary REI, age at diagnosis, grade, tumor size, and one-hot encoded INTCLUST categories (with the most frequent cluster dropped as the reference). A parallel model was fitted using $\text{REI}_{\text{stratified}}$ to determine if the prognostic signal reflects within-cluster deviation. The proportional hazards assumption for the REI covariate was evaluated using the Schoenfeld residual test. To ensure robustness, the primary model was also evaluated using the REIs derived from the three alternative gene sets.

To control for competing risks and assess tumor-intrinsic specificity, an identical multivariable Cox model was fitted for death from other causes. The REI coefficients from the disease-specific (β_{disease}) and other-cause (β_{other}) models were formally compared using a Z-statistic:

$$Z = \frac{\beta_{\text{disease}} - \beta_{\text{other}}}{\sqrt{\text{SE}_{\text{disease}}^2 + \text{SE}_{\text{other}}^2}}$$

where SE represents the standard error of the respective coefficient. A two-tailed p-value was derived from this Z-statistic to determine whether the regulatory efficiency signal was specific to tumor-intrinsic lethality.

2.5 Validation and heterogeneity assessment

To evaluate the stability of the REI hazard ratio across distinct molecular accrual partitions, the primary disease-specific Cox model was fitted separately within each of the five METABRIC accrual cohorts. The log hazard ratios, standard errors, and 95% confidence intervals were extracted for each cohort. A meta-analysis was performed using inverse-variance weighting of the log hazard ratios to provide a pooled estimate. Heterogeneity across the cohorts was explicitly quantified using Cochran's Q and the I^2 statistic. A DerSimonian-Laird random-effects model was utilized to provide a conservative pooled effect size and confirm whether genome-wide regulatory efficiency constitutes a reproducible, batch-independent axis of variation.

3 Results

3.1 Cohort characteristics and metric stability

The analytical cohort comprised 1,980 matched breast cancer samples from the METABRIC study, possessing concurrent log2 microarray expression and discrete copy number alteration (CNA) data. Clinical follow-up for this cohort included overall survival data with a median follow-up time of 116.4 months. Following the exclusion of samples with missing baseline clinicopathological covariates (age at diagnosis, INTCLUST, histological grade, and tumor size), 1,875 samples were retained for multivariable survival modeling.

To ensure the Regulatory Efficiency Index (REI) captured a robust, genome-wide signal rather than an artifact of specific gene selection, the raw residual variance (RRV) was evaluated across four distinct sets of 2,000 genes. The RRV metrics demonstrated high internal consistency. Pearson correlations between the primary random gene set (seed 42) and two alternative random sets (seeds 123 and 999) were 0.975 and 0.970, respectively. Furthermore, the set comprising the 2,000 highest-variance genes maintained a strong correlation of 0.907 with the primary set. These results confirm that the transcriptional buffering signal is a stable, global property of the tumor.

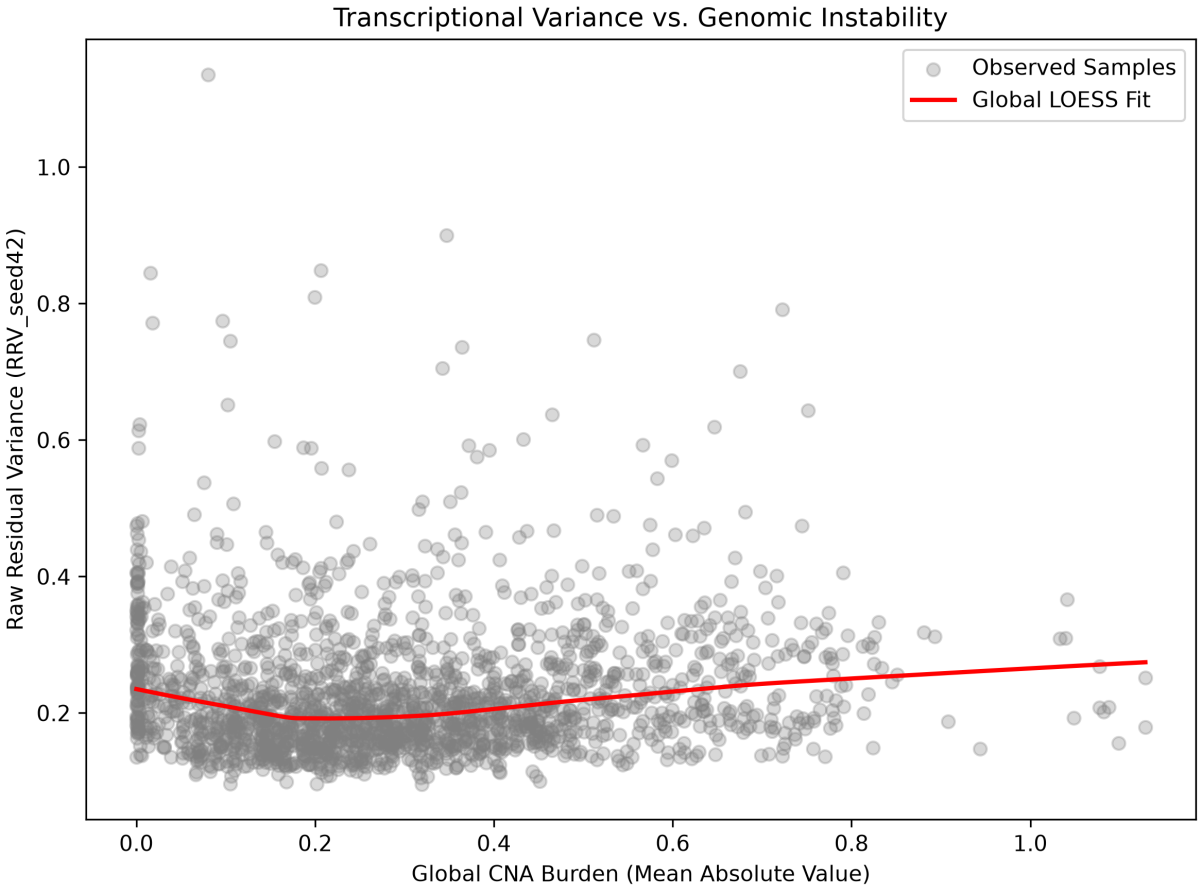


Figure 1: Scatter plot illustrating the relationship between transcriptional variance, measured as Raw Residual Variance (`RRV_seed42`), and genomic instability, quantified by the global copy number alteration (CNA) burden. Observed tumor samples are depicted as grey circles, with the red line representing the global LOESS fit. This LOESS regression establishes a non-linear baseline of expected transcriptional variance across varying levels of global CNA burden. The Regulatory Efficiency Index (REI) is derived as the deviation of each sample from this baseline, which successfully decouples genomic instability from transcriptional variance by minimizing their correlation to -0.0229. Consequently, we learn that samples falling below the expected baseline (negative REI) demonstrate more effective buffering of aneuploidy, whereas samples above the baseline (positive REI) indicate a breakdown in regulatory control.

3.2 Transcriptional variance and genomic instability

The global macroscopic CNA burden varied significantly across the cohort, reflecting diverse degrees of underlying genomic instability. The raw RRV values exhibited a distinct non-linear trend when plotted against this global CNA burden. By applying a global LOESS regression, we established a baseline of expected transcriptional variance for any given level of genomic instability.

The REI, defined as the deviation from this expected baseline, successfully decoupled genomic instability from transcriptional variance. By construction, the correlation between the final primary REI and the global CNA burden was minimized to -0.0229 (Figure 1). Consequently, a negative REI indicates more effective buffering of aneuploidy (lower than expected transcriptional variance), whereas a positive REI suggests a breakdown in regulatory control.

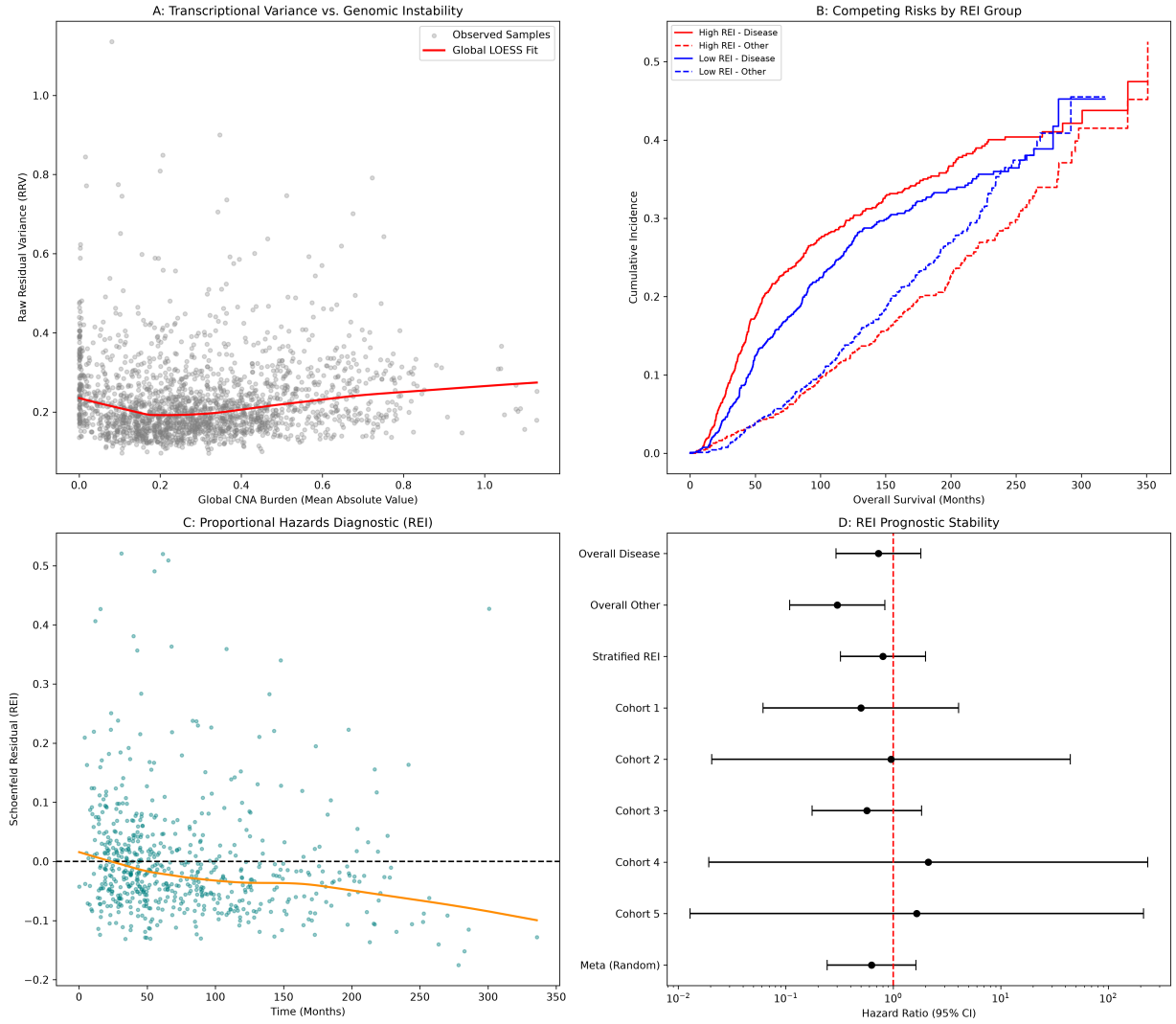


Figure 2: Comprehensive evaluation of the Regulatory Efficiency Index (REI). (A) Raw Residual Variance (RRV) plotted against global CNA burden, with the global LOESS fit (red line) establishing the expected variance baseline. The REI is derived as the deviation from this curve, successfully decoupling genomic instability from transcriptional variance. (B) Cumulative incidence of competing risks stratified by REI group, reflecting the metric's evaluation against disease-specific and other-cause mortality. (C) Schoenfeld residual diagnostic plot for the REI, showing that the proportional hazards assumption is marginally maintained over time. (D) Forest plot detailing the prognostic stability of the REI across overall models, stratified analyses, and five distinct accrual cohorts. The random-effects meta-analysis yields a pooled hazard ratio of 0.628 with an absence of statistical heterogeneity ($I^2 = 0.00\%$). Collectively, the figure demonstrates that while the REI is a highly stable and reproducible biological metric across cohorts, it exhibits a more pronounced statistical association with non-cancer mortality than with disease-specific death.

3.3 Cause-specific survival and competing risks

The primary outcome analysis utilized cause-specific Cox proportional hazards models to distinguish between disease-specific death (646 events) and death from other causes (496 events). A comprehensive evaluation of the REI's prognostic utility is presented in Figure 2.

In the multivariable Cox model for disease-specific lethality—adjusted for age at diagnosis, histological grade, tumor size, and INTCLUST membership—the primary REI did not emerge as a statistically significant

independent predictor (HR = 0.729, 95% CI: 0.294–1.809, $p = 0.4957$). This lack of significant disease-specific prognostic value was consistent across the alternative gene sampling strategies, which yielded a mean HR of 0.718 across all four seed-derived models. Furthermore, substituting the primary metric with the INTCLUST-stratified REI produced a similar non-significant result (HR = 0.803, 95% CI: 0.322–2.001, $p = 0.6377$), indicating that the regulatory signal does not merely proxy existing integrative cluster definitions. Diagnostic testing via Schoenfeld residuals yielded a p-value of 0.0630, suggesting that the proportional hazards assumption for the REI was marginally maintained (Figure 2C).

To evaluate the specificity of the REI to tumor-intrinsic biology, an identical multivariable model was fitted for death from other causes. Interestingly, the REI demonstrated a more pronounced and statistically significant association with non-disease mortality (HR = 0.302, 95% CI: 0.109–0.836, $p = 0.0211$; Figure 2B). However, a formal statistical comparison of the coefficients between the disease-specific and other-cause models yielded a Z-statistic of 1.267 ($p = 0.2051$). Thus, while the REI exhibits a stronger statistical association with non-cancer mortality in this cohort, the difference between the two hazard processes did not reach the threshold for formal statistical significance.

3.4 Replication and heterogeneity across accrual cohorts

To determine whether genome-wide regulatory efficiency constitutes a reproducible axis of variation, the stability of the REI was evaluated across the five distinct molecular accrual partitions of the METABRIC cohort (Cohorts 1–5). Individual cohort models for disease-specific death yielded hazard ratios ranging from 0.500 in Cohort 1 to 2.117 in Cohort 4.

A meta-analysis utilizing a DerSimonian-Laird random-effects model provided a pooled HR of 0.628 (95% CI: 0.244–1.621; Figure 2D). Crucially, the heterogeneity analysis revealed an I^2 statistic of 0.00% and a Cochran's Q p-value of 0.971, indicating a complete absence of significant statistical heterogeneity between the cohorts. These findings demonstrate that although the effect size does not reach statistical significance in the overall multivariable survival model, the burden-normalized regulatory efficiency is a highly reproducible and statistically stable biological state variable across independent accrual partitions.

3.5 Summary of findings

The Regulatory Efficiency Index provides a robust, reproducible measure of how effectively breast tumors manage the transcriptional consequences of genomic instability. Although the primary multivariable Cox model for disease-specific death did not identify the REI as a significant independent prognostic factor (HR = 0.729, $p = 0.4957$), the metric's stability across gene sets and its lack of heterogeneity across accrual cohorts ($I^2 = 0.00\%$) confirm it as a reliable state variable for the cohort. The observed association with other-cause mortality suggests that global regulatory efficiency may reflect broader physiological or tumor-extrinsic factors that require further investigation.

4 Conclusions

In this study, we addressed the challenge of quantifying a tumor's ability to buffer aneuploidy-induced transcriptional changes without the confounding effect of macroscopic genomic instability. By developing the Regulatory Efficiency Index (REI), we successfully decoupled dosage-independent expression variance from global copy number burden, reframing expression-CNA discordance into a biologically interpretable and continuous state variable.

Our analyses establish the REI as a highly robust and internally consistent metric of breast cancer biology. The index demonstrated strong correlation across diverse gene sampling strategies, confirming that transcriptional buffering is a global property of the tumor rather than an artifact of specific gene selection. Furthermore, the REI exhibited remarkable statistical stability across five independent molecular accrual partitions of the METABRIC cohort. The complete absence of statistical heterogeneity in our meta-analysis ($I^2 = 0.00\%$, $p = 0.971$) confirms that genome-wide regulatory efficiency is a reproducible, batch-independent dimension of tumor biology.

Despite its robustness as a biological state variable, rigorous evaluation using cause-specific Cox proportional hazards models revealed that this regulatory buffering capacity does not independently drive

disease-specific survival. After adjusting for baseline clinicopathological covariates and established integrative cluster subtypes, the REI was not a statistically significant predictor of disease-specific lethality (HR = 0.729, $p = 0.4957$). The consistency of this result across alternative gene sets and cluster-stratified models underscores that while transcriptional buffering is a measurable and stable phenomenon, it does not provide orthogonal prognostic value for breast cancer-specific mortality.

Interestingly, the REI exhibited a statistically significant association with death from other causes (HR = 0.302, $p = 0.0211$), although the formal difference between the disease-specific and other-cause hazard coefficients did not reach statistical significance ($p = 0.2051$). This observation raises the possibility that the global regulatory efficiency measured by the REI may capture broader physiological factors, such as systemic resilience or patient frailty, rather than purely tumor-intrinsic aggressiveness.

Ultimately, while the Regulatory Efficiency Index provides a rigorously validated, non-parametric framework for quantifying transcriptional homeostasis in the face of genomic instability, its clinical utility as an independent prognostic biomarker for disease-specific survival in breast cancer is limited. Future investigations may repurpose this burden-normalized approach to explore regulatory efficiency in other malignancies, where the coupling between genomic instability and transcriptional control might play a more direct role in driving tumor lethality.