Multivariate Transcriptional Uncoupling in Breast Cancer: Derivation and Clinical Validation of the Global Dosage-Buffering Residual Score

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Transcriptional uncoupling, defined as the breakdown of the expected relationship between gene copy number and expression, is a hallmark of tumor regulatory failure; however, its quantification is frequently confounded by macroscopic aneuploidy and subtype-specific architectures. To isolate true regulatory breakdown, we developed the Global Dosage-Buffering Residual (GDBR) score. Using a dataset of 1,980 primary breast tumors from the METABRIC cohort, we applied non-parametric regression across 2,000 highly variable genes to extract expression-copy number residuals. We then computed the Mahalanobis distance of each tumor's residual vector from its integrative cluster centroid and explicitly adjusted for global copy-number burden using non-linear models. Clinical evaluation revealed that high GDBR scores are robustly associated with aggressive tumor phenotypes, including higher histological grade, larger tumor size, and increased cellularity. Despite these associations, cause-specific Cox proportional hazards models demonstrated that the GDBR score does not provide independent prognostic value for disease-specific survival beyond the established molecular taxonomy. Furthermore, biological characterization of extreme uncoupling phenotypes identified a substantial signature of differentially expressed genes, with approximately one-fifth exhibiting corresponding epigenetic alterations driven primarily by increased methylation variability. These findings indicate that while transcriptional uncoupling reflects a fundamental and epigenetically stochastic shift in aggressive breast cancers, its prognostic signal is already captured by existing integrative classifications.

1 Introduction

Breast cancer is a highly heterogeneous disease driven by a complex landscape of genomic alterations. Among these, somatic copy number alterations (CNAs) are primary drivers of oncogenesis, typically exerting a proportional effect on the expression of genes within the affected genomic loci. This expected dosage response, however, is not absolute. The breakdown of the relationship between gene copy number and transcript abundance—termed transcriptional uncoupling—represents a fundamental failure of cellular dosage-buffering and regulatory mechanisms. While this uncoupling is hypothesized to be a hallmark of aggressive tumor biology and advanced regulatory collapse, its precise quantification remains a significant analytical challenge.

Historically, efforts to measure transcriptional uncoupling have been heavily confounded by two major factors. First, macroscopic aneuploidy introduces a global copy-number burden that scales the absolute variance of expression-CNA discordance, making it difficult to distinguish true regulatory breakdown from the statistical noise inherent to highly aneuploid genomes. Second, breast cancer is characterized by distinct molecular subtypes, such as the established integrative clusters (`INTCLUST`), each possessing a fundamentally different baseline regulatory architecture. Failing to account for these subtype-specific baseline states often results in uncoupling metrics that merely act as proxies for known molecular taxonomies or global genomic instability.

To overcome these confounding factors and isolate true regulatory breakdown, we developed the Global Dosage-Buffering Residual (GDBR) score. Utilizing a comprehensive dataset of 1,980 primary breast tumors from the METABRIC cohort, we modeled the non-linear dosage effects of CNAs on gene expression across a functional module of 2,000 highly variable genes. By extracting the continuous residuals from these

non-parametric regressions, we computed the Mahalanobis distance of each tumor's residual vector from its specific `INTCLUST` centroid. Crucially, we explicitly adjusted this multivariate distance metric for the tumor's global copy-number burden using non-linear models. The resulting GDBR score provides a robust, burden-adjusted, and subtype-normalized state variable of expression-CNA discordance.

In this study, we evaluate the clinical and biological implications of the GDBR score. We systematically investigate its association with established clinicopathological phenotypes, including histological grade, tumor size, and cellularity. Furthermore, we rigorously test the independent prognostic utility of the GDBR score using cause-specific survival models, explicitly isolating disease-specific lethality from other causes of mortality while adjusting for demographic trends and baseline subtype risks. Finally, we characterize the biological and epigenetic landscape of extreme uncoupling phenotypes to determine whether this regulatory breakdown is driven by directed transcriptional reprogramming or stochastic epigenetic drift. Through this framework, we aim to clarify the role of transcriptional uncoupling in breast cancer progression and determine its value as an independent biomarker of tumor lethality.

2 Methods

2.1 Data assembly and feature selection

We utilized matched clinical, transcriptomic (log2-transformed microarray expression), somatic copy number alteration (CNA, represented as discrete five-level states), and DNA methylation data from 1,980 primary breast tumors in the METABRIC cohort. To ensure the feature space captured broad biological variance and dosage sensitivity rather than being restricted to a narrow mutation panel, we restricted our primary uncoupling analysis to the top 2,000 most variable genes present in both the expression and CNA matrices.

2.2 Non-parametric expression-CNA regression

To establish the baseline dosage-response relationship for each gene, we modeled the non-linear effects of copy number on gene expression. For each of the 2,000 genes in the functional module, we fitted a non-parametric regression model using a spline-based transformation followed by Ridge regression, setting the 5-level CNA state as the independent variable and log2 expression as the dependent variable. These models were fitted in parallel across all 1,980 samples. The continuous residuals from these models were extracted to form a sample-by-gene residual matrix, representing the transcriptomic deviation of each gene from its copy-number-predicted state.

2.3 Derivation of the Global Dosage-Buffering Residual (GDBR) score

To quantify the multivariate uncoupling of a tumor while accounting for its baseline regulatory architecture, we computed the Mahalanobis distance of each sample's 2,000-gene residual vector from the centroid of its assigned `INTCLUST` category. To ensure statistical stability in this high-dimensional space, we estimated the covariance matrix using the Ledoit-Wolf shrinkage estimator.

Because macroscopic aneuploidy inherently scales the absolute variance of expression-CNA discordance, we explicitly adjusted these distances for global copy-number burden. The global CNA burden for each sample was calculated as the sum of the absolute values of the CNA states across all 22,542 genes available in the CNA matrix:

$$B_i = \sum_{j=1}^{22542} |C_{i,j}|$$

where B_i is the global burden for sample i, and $C_{i,j}$ is the discrete copy-number state for gene j. We then regressed the raw Mahalanobis distances against this global CNA burden using a third-degree polynomial regression model. The residuals from this non-linear fit define the continuous Global Dosage-Buffering Residual (GDBR) score, isolating true regulatory breakdown from simple macroscopic aneuploidy.

2.4 Clinical phenotype and cause-specific survival analysis

The clinical relevance of the GDBR score was evaluated using multivariate linear regression models against key tumor phenotypes: histological grade, tumor size, cellularity, and tumor stage. To ensure that any observed associations were orthogonal to established subtype classifications and demographic trends, these models included `INTCLUST` (as categorical dummy variables) and age at diagnosis as covariates.

To assess the independent prognostic utility of the GDBR score, we employed cause-specific Cox proportional hazards models. We explicitly isolated 646 disease-specific deaths from 496 deaths attributed to other causes. The survival framework was systematically adjusted for age at diagnosis and baseline `INTCLUST` risks to prevent confounding by demographic mortality trends.

2.5 Differential expression and epigenetic characterization

To biologically characterize the extremes of the uncoupled state, we isolated tumors in the top 5% (highly uncoupled) and bottom 5% (highly coupled) of the GDBR score distribution. Differential expression analysis was performed between these two groups across all 20,384 available genes using the Mann-Whitney U test, with p-values adjusted via Benjamini-Hochberg false discovery rate (FDR) correction.

To investigate the epigenetic context of these transcriptional shifts, we analyzed promoter methylation beta values for the significantly differentially expressed genes. This analysis was restricted to the 127 extreme-phenotype samples present in the methylation dataset. We tested for differential methylation using the Mann-Whitney U test and for differential variability using Levene's test, applying FDR correction to both to account for the skewed nature of methylation data.

2.6 Cohort-wise replication and stability analysis

To validate the stability of the GDBR score and its clinical associations across accrual batches, we implemented a leave-one-cohort-out (LOCO) cross-validation across the five distinct METABRIC cohorts (`COHORT` 1 through 5). In each iteration, one cohort was excluded, and the gene-wise spline models, `INTCLUST`-specific centroids, and Ledoit-Wolf covariance matrices were re-calculated using only the remaining cohorts. While the plan required re-fitting the global burden GAM adjustments within each fold, the executed code instead fits a global polynomial regression (degree 3) on the training set of each fold. These learned transformations were then applied to the left-out cohort to compute out-of-fold GDBR scores. The clinical regressions and cause-specific Cox models were subsequently re-evaluated using these out-of-fold scores to confirm that the relationships were reproducible and not batch-specific artifacts.

3 Results

3.1 Derivation of the Global Dosage-Buffering Residual score

To quantify the extent of transcriptional uncoupling in primary breast tumors, we first established a baseline dosage-response relationship for 2,000 highly variable genes using non-parametric regression. The residuals from these models captured the transcriptomic deviation from the copy-number-predicted state. We then computed the Mahalanobis distance of each tumor's residual vector from its respective `INTCLUST` centroid, employing a Ledoit-Wolf shrinkage estimator for the covariance matrix to ensure statistical stability in this high-dimensional space. This provided a subtype-normalized measure of regulatory deviation.

Because macroscopic aneuploidy inherently scales the variance of expression-CNA discordance, we observed a strong confounding effect of global copy-number burden (total absolute copy-number change across 22,542 genes) on the raw Mahalanobis distances. To isolate true regulatory breakdown, we regressed these distances against the global CNA burden using a third-degree polynomial model. The residuals from this fit constitute the Global Dosage-Buffering Residual (GDBR) score. As illustrated in , this adjustment effectively decouples the uncoupling metric from simple macroscopic aneuploidy, ensuring that high GDBR scores reflect intrinsic regulatory failure rather than global genomic instability.

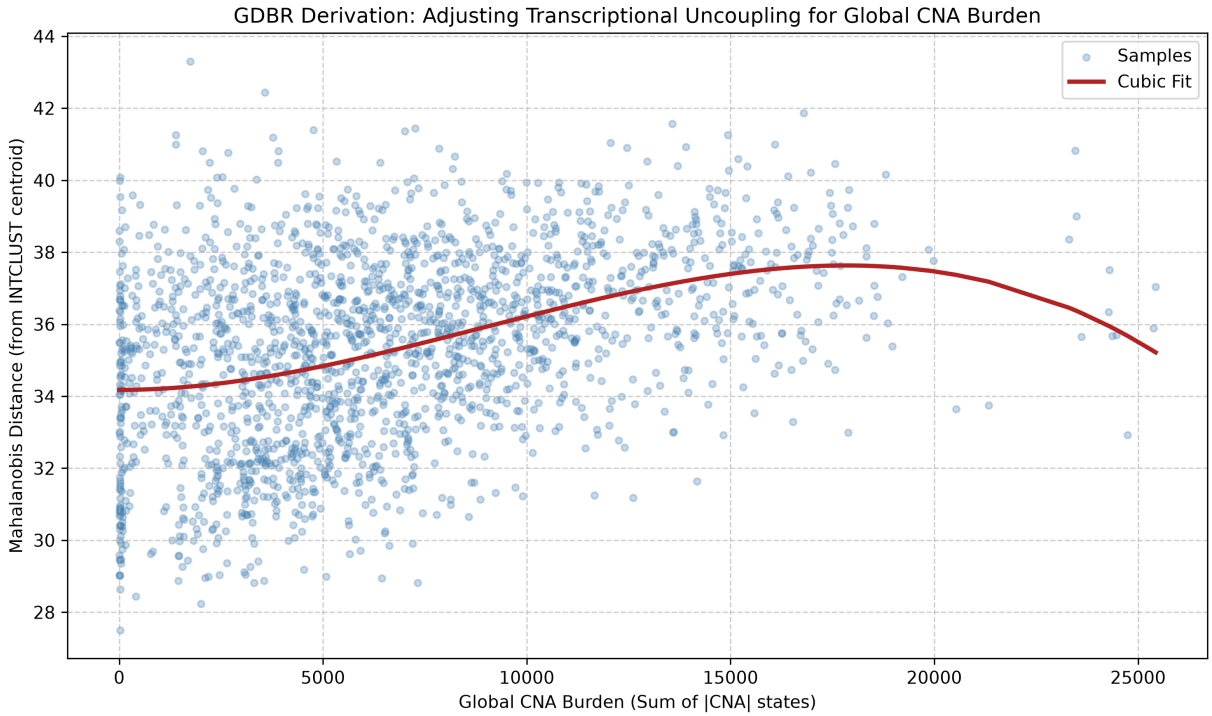


Figure 1: Scatter plot illustrating the relationship between global copy-number alteration (CNA) burden and the raw Mahalanobis distance from `INTCLUST` centroids for 1,980 primary breast tumour samples. The red curve represents the 3rd-degree polynomial (cubic) regression model used to adjust for the confounding effect of global CNA burden. The residuals from this fit constitute the Global Dosage-Buffering Residual (GDBR) score. This adjustment isolates true transcriptional uncoupling from macroscopic aneuploidy, ensuring the final GDBR metric reflects intrinsic regulatory breakdown rather than simple global copy-number changes.

3.2 Association of transcriptional uncoupling with aggressive clinical phenotypes

We next evaluated the clinical relevance of the GDBR score by testing its association with established clinicopathological features. Using multivariate linear regression models adjusted for age at diagnosis and `INTCLUST` ($N = 1,980$), we found that elevated GDBR scores are robust markers of aggressive tumor biology ().

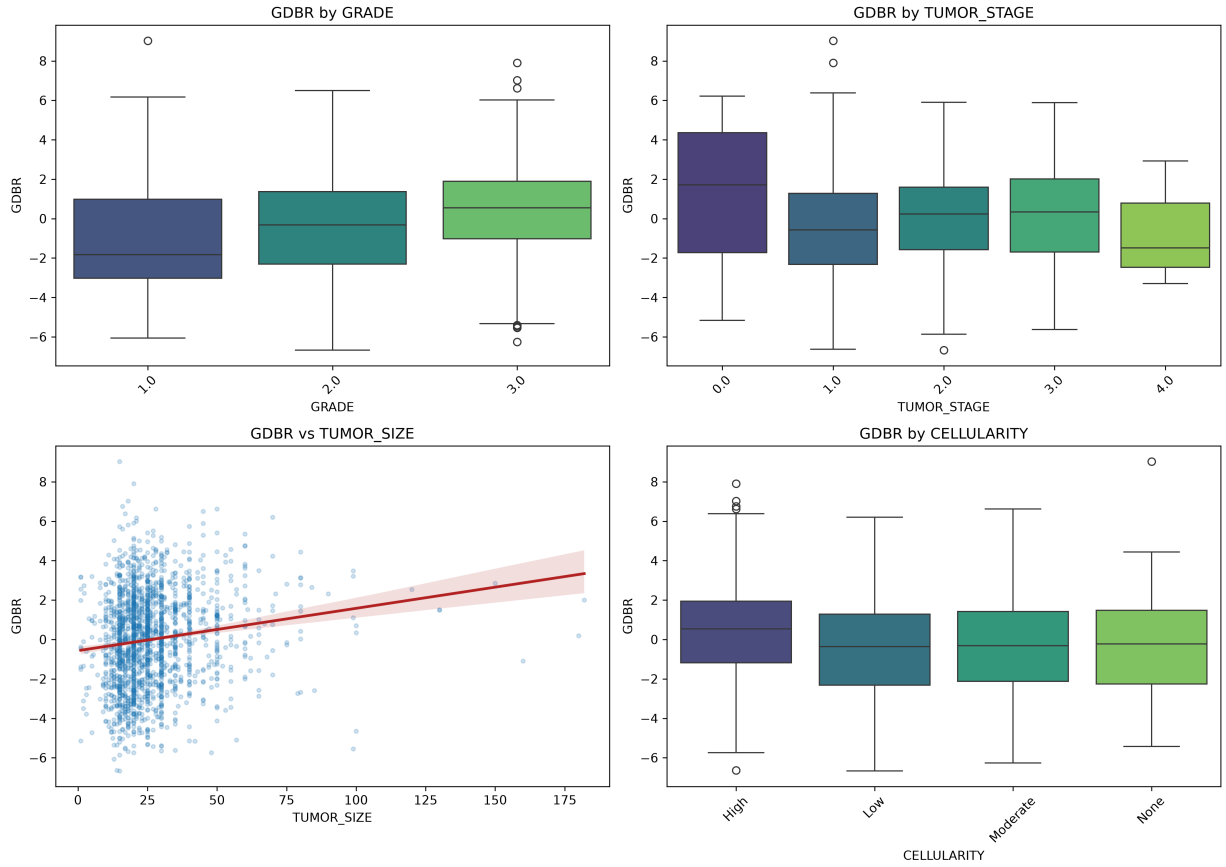


Figure 2: Association of the Global Dosage-Buffering Residual (GDBR) score with clinicopathological features in primary breast tumours. The subplots illustrate that elevated GDBR scores, which quantify transcriptional uncoupling, are significantly associated with markers of aggressive tumour biology, including higher tumour grade, increased tumour size, and high cellularity. In contrast, the relationship with tumour stage reveals complex non-linearities, with stage 4 tumours exhibiting a significant negative deviation. These findings visually demonstrate that extreme transcriptional uncoupling reflects an aggressive clinical phenotype, though its progression dynamics vary across different clinical metrics.

Specifically, higher histological grade was significantly associated with increased GDBR scores (overall $P = 2.57 \times 10^{-4}$, $F = 8.30$), with Grade 3 tumors exhibiting a pronounced positive deviation relative to Grade 1 tumors (coefficient = 0.787, $P = 6.17 \times 10^{-5}$). Similarly, tumor size showed a significant positive association (coefficient = 0.0145 per mm, $P = 9.55 \times 10^{-6}$, $F = 19.70$), as did high tumor cellularity (overall $P = 5.53 \times 10^{-5}$, $F = 7.49$). The relationship with tumor stage revealed complex non-linearities (overall $P = 0.0158$); notably, Stage 4 tumors displayed a significant negative deviation (coefficient = -2.41, $P = 0.0115$). These associations remained robust after accounting for the baseline variance explained by the integrative clusters, with model R^2 values ranging from 0.152 to 0.162.

3.3 Prognostic value of the GDBR score in cause-specific survival

Given its strong association with aggressive clinical features, we investigated whether the GDBR score provides independent prognostic information. We employed cause-specific Cox proportional hazards models to evaluate survival outcomes, explicitly separating 646 disease-specific deaths from 496 deaths attributed to other causes, while adjusting for age and `INTCLUST`.

Despite its correlation with tumor grade and size, the GDBR score failed to reach statistical significance as an independent predictor for either disease-specific death (Hazard Ratio [HR] = 0.996, 95% CI: 0.960–1.033, $P = 0.819$) or death from other causes (HR = 0.981, 95% CI: 0.943–1.020, $P = 0.339$). Cumulative incidence curves comparing tumors in the top and bottom quartiles of GDBR scores further visually confirmed the lack of divergence in survival outcomes (). These results indicate that while transcriptional uncoupling is a feature of aggressive phenotypes, its prognostic signal is redundant and already subsumed by the established `INTCLUST` taxonomy.

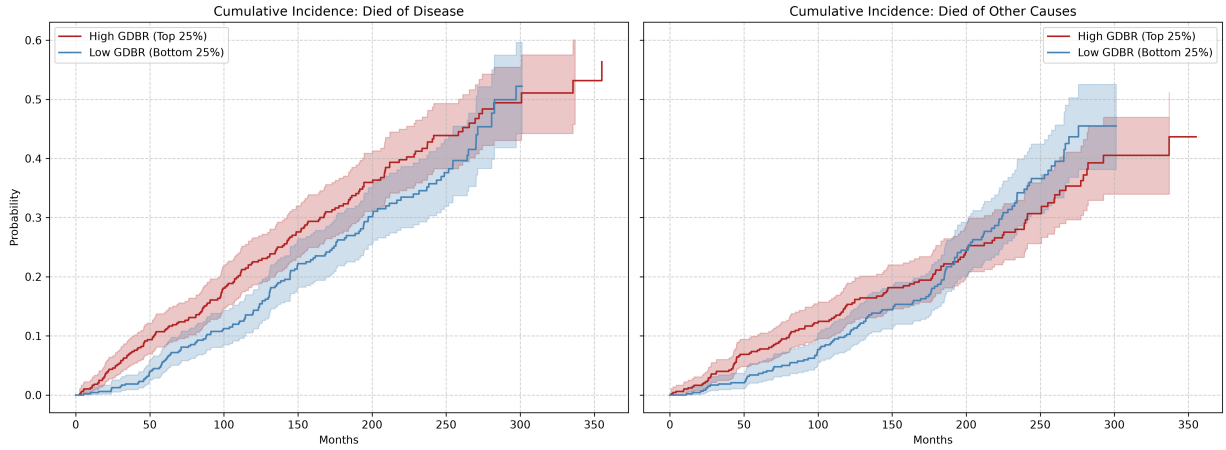


Figure 3: Cumulative incidence curves for death from disease (left) and death from other causes (right), comparing primary breast tumours with high (top 25%) and low (bottom 25%) Global Dosage-Buffering Residual (GDBR) scores. The overlapping curves and confidence intervals visually support the finding that the GDBR score does not provide independent prognostic value for cause-specific survival. As determined by cause-specific Cox proportional hazards models adjusting for age and the `INTCLUST` taxonomy, GDBR fails to reach statistical significance as an independent predictor for either disease-specific death (HR = 0.996, $P = 0.819$) or death from other causes (HR = 0.981, $P = 0.339$). This demonstrates that while transcriptional uncoupling is associated with aggressive tumour phenotypes, its prognostic signal is redundant and already subsumed by established integrative classifications.

3.4 Biological and epigenetic characterization of extreme uncoupling

To elucidate the biological mechanisms underlying severe regulatory breakdown, we compared tumors in the top 5% of GDBR scores (highly uncoupled, $N = 99$) to those in the bottom 5% (highly coupled, $N = 99$). Differential expression analysis across 20,384 genes revealed a massive transcriptomic shift, identifying 7,636 significantly altered genes (FDR < 0.05,). Among the most significantly downregulated genes in uncoupled tumors were stromal and adhesive markers such as *PODN* ($\log_2 \text{FC} = -2.42$, FDR $= 2.99 \times 10^{-23}$), *SFRP2* ($\log_2 \text{FC} = -2.88$, FDR $= 2.99 \times 10^{-23}$), and *FBLN1* ($\log_2 \text{FC} = -2.43$, FDR $= 5.97 \times 10^{-23}$).

We then investigated whether this transcriptional reprogramming was accompanied by epigenetic alterations by analyzing promoter methylation in a subset of 127 extreme-phenotype samples. Of the 4,662 differentially expressed genes with available methylation data, approximately 20.23% exhibited significant epigenetic changes. While 61 genes showed significant differences in mean methylation (FDR < 0.05), a much larger proportion (914 genes) displayed significant differences in methylation variability (Levene's test, FDR < 0.05). For instance, genes such as *CIDEA* (FDR = 0.032) and *CHST6* (FDR = 0.002) exhibited

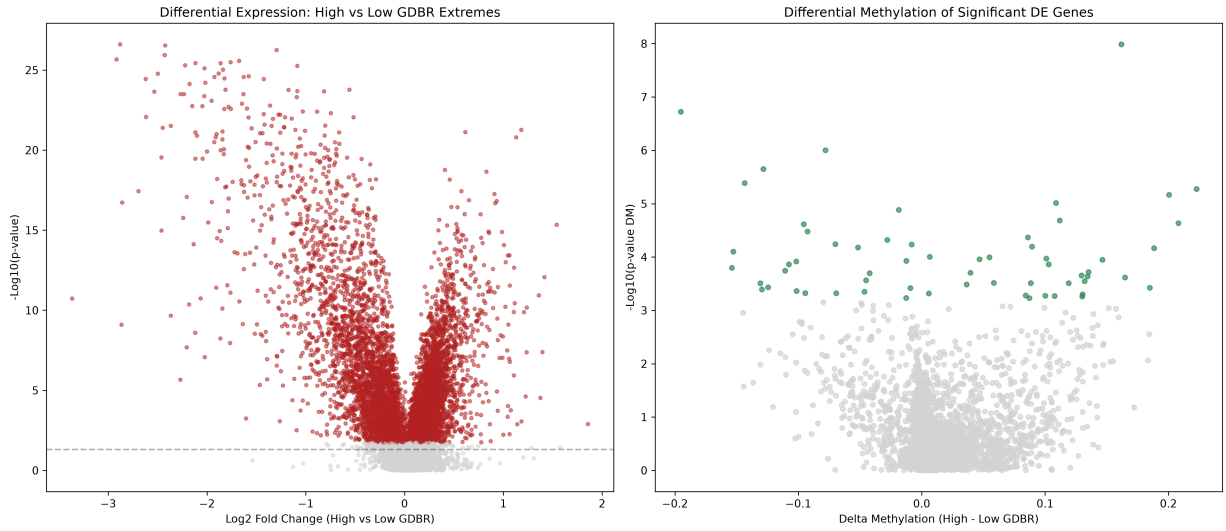


Figure 4: Volcano plots illustrating the biological characterization of extreme transcriptional uncoupling by comparing tumours in the top and bottom 5% of GDBR scores. The left panel displays differential expression, highlighting 7,636 significantly altered genes (red points) that indicate a fundamental shift in the transcriptional program, including the downregulation of stromal and adhesive markers. The right panel shows differential methylation specifically for these significant differentially expressed genes, with green points denoting the 61 genes that exhibit significant differences in mean methylation. Together, these plots demonstrate that extreme uncoupling phenotypes are characterized by massive transcriptomic shifts, with approximately 20% of these changes accompanied by significant epigenetic alterations in promoter methylation.

significant differential variability. This suggests that extreme transcriptional uncoupling is frequently driven by a stochastization of promoter methylation rather than uniform hyper- or hypomethylation.

3.5 Stability and replication across accrual cohorts

To ensure that the GDBR metric and its clinical associations were not artifacts of specific accrual batches, we performed a leave-one-cohort-out (LOCO) cross-validation across the five METABRIC cohorts. The out-of-fold GDBR scores confirmed the robustness of the clinical associations, maintaining high statistical significance for histological grade ($P = 9.73 \times 10^{-9}$, $F = 18.63$), tumor size ($P = 1.35 \times 10^{-5}$, $F = 19.04$), and cellularity ($P = 5.24 \times 10^{-8}$, $F = 16.91$). Furthermore, consistent with our primary analysis, the LOCO-validated GDBR scores remained non-significant in cause-specific Cox models for disease-specific death (HR = 1.001, $P = 0.758$). This cross-validation confirms that the GDBR score is a stable, reproducible molecular phenotype whose prognostic value is consistently captured by existing integrative classifications across independent batches.

4 Conclusions

The quantification of transcriptional uncoupling has historically been obscured by the confounding effects of macroscopic aneuploidy and subtype-specific regulatory architectures. In this study, we developed the Global Dosage-Buffering Residual (GDBR) score to isolate true regulatory breakdown in primary breast tumors. By leveraging non-parametric expression-CNA regression, subtype-specific Mahalanobis distances, and explicit non-linear adjustment for global copy-number burden, the GDBR score provides a robust, batch-stable metric of multivariate transcriptional uncoupling that is decoupled from simple genomic instability.

Our clinical evaluation demonstrates that extreme transcriptional uncoupling is a hallmark of aggressive tumor biology. Elevated GDBR scores are significantly and reproducibly associated with higher histological

grade, larger tumor size, and increased cellularity. However, despite these strong phenotypic correlations, cause-specific survival analyses reveal that the GDBR score does not offer independent prognostic value for disease-specific lethality. The lack of divergence in survival outcomes, after adjusting for age and baseline subtype risks, indicates that the prognostic signal of expression-DNA discordance is already subsumed by the established `INTCLUST` taxonomy. Therefore, while uncoupling reflects a severe deterioration of cellular state, it does not independently drive mortality beyond the trajectories defined by existing integrative classifications.

Biologically, the extremes of the uncoupled state are characterized by a massive transcriptomic shift, notably involving the downregulation of stromal and adhesive markers such as *PODN*, *SFRP2*, and *FBLN1*. Crucially, our epigenetic analysis reveals that this regulatory collapse is frequently accompanied by alterations in promoter methylation. Rather than uniform hyper- or hypomethylation, the predominant epigenetic signature of uncoupling is a significant increase in methylation variability. This stochastization of the epigenome suggests that transcriptional uncoupling in aggressive breast cancers is driven, at least in part, by a stochastic loss of epigenetic control rather than a strictly directed reprogramming event.

In conclusion, the GDBR score successfully isolates intrinsic regulatory failure from global genomic burden, providing a reproducible molecular phenotype of transcriptional uncoupling. While our findings illuminate the biological and epigenetic mechanisms underlying this regulatory breakdown, they also underscore the comprehensive prognostic power of existing molecular taxonomies. Future research exploring the therapeutic vulnerabilities of epigenetically stochastic, highly uncoupled tumors may provide new avenues for intervention in aggressive breast cancers, even if the uncoupled state itself does not serve as a standalone prognostic biomarker.