Multidimensional Regulatory Discordance Profiles of Active Loci as Cause-Specific Predictors of Breast Cancer Lethality

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Integrating multi-omic data to identify localized regulatory failures often conflates specific biological rewiring with macroscopic genomic instability or established clinical taxonomies. To isolate these localized failures, we constructed a Multidimensional Regulatory Discordance Profile (RDP) across 1,417 primary breast tumors using matched copy number, methylation, and transcriptomic data. By employing a leave-one-cohort-out cross-validation framework, we identified regulatory-active genes independent of standard molecular subtypes and computed expression residuals via non-linear regression to capture deviations from expected copy-number and methylation states. We then evaluated the prognostic utility of these high-dimensional profiles using cause-specific Cox proportional hazards models and competing risks analysis, adjusting for age and global transcriptional noise. This approach identified a stable core signature of 44 genes, including BAP1 and SUZ12, whose regulatory discordance stratifies disease-specific lethality. Although the overall linear concordance index was modest, cumulative incidence analysis demonstrated that the RDP specifically stratified disease-related death without inadvertently stratifying mortality from other causes. Furthermore, we observed that dosage buffering failure—where transcript levels decouple from gene dosage—was significantly more prevalent in high-risk tumors, whereas epigenetic override was rare. These findings indicate that the localized decoupling of expression from genomic dosage provides an independent, albeit cohort-sensitive, axis of breast cancer lethality.

1 Introduction

The integration of multi-omic data has fundamentally advanced our understanding of breast cancer, enabling the classification of tumors into robust molecular taxonomies based on genomic and transcriptomic landscapes. However, standard approaches to multi-omic integration often capture macroscopic features of the tumor, such as global genomic instability, broad aneuploidy, or established clinical subtypes. Consequently, these methods can obscure localized regulatory failures—specific instances where the transcriptional output of a gene fundamentally decouples from its genomic dosage or epigenetic state. Isolating these localized instances of biological rewiring is critical, as they may represent independent drivers of tumor lethality that are currently masked by broader molecular classifications.

Gene expression is typically constrained by a combination of copy-number alterations (CNAs) and epigenetic modifications, such as promoter methylation. When these regulatory mechanisms fail, the resulting "regulatory discordance" can manifest in several ways, including dosage buffering failure—where transcript levels fail to scale with gene amplification or deletion—or epigenetic override. Quantifying this discordance requires moving beyond simple linear correlations to establish the expected expression of a gene given its specific multi-omic context, and subsequently measuring the deviation from this expectation. Furthermore, to ensure that such a metric captures true cancer-specific biological rewiring, it must be rigorously decoupled from confounding factors such as patient age, global transcriptional noise, and all-cause mortality.

To address this challenge, we introduce the Multidimensional Regulatory Discordance Profile (RDP). Utilizing a comprehensive dataset of 1,417 primary breast tumors with matched copy number, methylation, and transcriptomic data, we systematically identified "regulatory-active" loci whose expression is tightly linked to their genomic or epigenetic state. By employing non-linear regression within a leave-one-cohort-out cross-validation framework, we computed expression residuals that capture the magnitude of regulatory escape at these specific loci. Crucially, we explicitly excluded the primary drivers of established molecular

subtypes to ensure that the RDP provides an orthogonal axis of information rather than merely proxying known clinical taxonomies.

In this study, we evaluate the prognostic utility of the RDP using cause-specific Cox proportional hazards models and competing risks analysis. By explicitly separating disease-specific deaths from deaths due to other causes, we demonstrate that regulatory discordance specifically stratifies breast cancer lethality without inadvertently capturing general patient frailty. We identify a stable core signature of 44 genes, including key regulators such as *BAP1* and *SUZ12*, whose regulatory failure is highly indicative of poor disease-specific outcomes. Although the overall linear predictive power of the profile exhibits cohort sensitivity, our cumulative incidence and mechanistic analyses reveal that localized decoupling of expression from genomic dosage—specifically dosage buffering failure—constitutes a distinct and significant dimension of high-risk breast cancer biology.

2 Methods

2.1 Data assembly and cohort alignment

To construct the Multidimensional Regulatory Discordance Profile (RDP), we assembled matched clinical, transcriptomic (log2 expression), copy-number alteration (CNA), and promoter methylation (beta values) data. The dataset was strictly subset to 1,417 primary breast tumor samples possessing complete data across all four molecular layers. All matrices were aligned by a unique sample identifier. For each sample, we extracted key clinical and technical variables: overall survival months, vital status, age at diagnosis, `INTCLUST` subtype, accrual cohort (1 through 5), and the `THREEGENE` proliferation index. To control for macroscopic aneuploidy in downstream analyses, we calculated the global CNA burden per sample, defined as the fraction of genes exhibiting non-zero CNA states. Finally, for cause-specific competing risks analysis, vital status was transformed into two binary event indicators: disease-specific death and death from other causes.

2.2 Refined active loci identification and subtype decoupling

We established a Leave-One-Cohort-Out (LOCO) cross-validation framework using the five distinct accrual partitions to ensure batch-independent reproducibility. In each fold, four cohorts comprised the training set and the remaining cohort served as the validation set. Genes were first filtered to meet a minimum baseline variance threshold across the cohort. To ensure the RDP captured localized regulatory failures rather than proxying established molecular subtypes, we performed a Kruskal-Wallis test on each gene's expression across the `INTCLUST` categories within the training set and explicitly excluded the top 20% most significant genes.

Strictly after this exclusion, we computed the Spearman correlation between log2 expression and CNA, and separately between log2 expression and methylation beta values, within each `INTCLUST` group. The plan required selecting the top active loci based on absolute correlation capped at 1,000 genes, but the executed code instead selected the top 10% of the remaining genes (capped at 1,000). The stability of these selected active loci across the five LOCO folds was evaluated using the Jaccard index.

2.3 Non-linear computation of the regulatory discordance profile

For each identified active gene, we modeled the expected expression given its multi-omic context. We fit a non-linear regression model (using XGBoost or LightGBM) on the training set to predict log2 expression from the categorical CNA state, continuous methylation beta value, and the sample's global CNA burden. These fitted models were applied to both the training and validation sets to compute the expected expression. The expression residual R_{ig} for gene g in sample i was calculated as the difference between the observed log2 expression Y_{ig} and the expected expression $\hat{Y}_{ig}$:

$$R_{ig} = Y_{ig} - \hat{Y}_{ig} \tag{1}$$

To ensure the resulting RDP captured cancer-specific lethality rather than general aging effects, we regressed out age at diagnosis by fitting a linear model on the training set that predicted each gene's residual from age,

and subsequently subtracted this age-predicted component from the residuals in all samples. Finally, we calculated the total variance of these adjusted residuals per sample to serve as a proxy for global transcriptional noise.

2.4 Supervised dimensionality reduction and stability analysis

To isolate discordance modules linked to survival, we fit a baseline Cox proportional hazards model on the training set using only age at diagnosis and one-hot encoded `INTCLUST` categories against disease-specific survival, extracting the martingale residuals. We then applied Partial Least Squares (PLS) regression to predict these martingale residuals from the high-dimensional RDP matrix. PLS models were fit using strictly nested internal cross-validation within the training folds to determine the optimal number of components (up to a maximum of 5). Both training and validation RDP matrices were transformed into these PLS components. To define a stable core signature of regulatory failure, we evaluated the absolute PLS loadings of the genes across the five folds and employed a consensus approach, retaining genes that appeared in the top features in at least 4 out of the 5 folds.

2.5 Cause-specific Cox modeling and competing risks analysis

For each training fold, we constructed cause-specific Cox proportional hazards models to explicitly separate the competing risks of mortality. Two separate models were fit: one for time to disease-specific death and another for time to death from other causes. The feature matrix for these models consisted of the PLS discordance components concatenated with age at diagnosis, one-hot encoded `INTCLUST`, and the total residual variance. Including total residual variance ensured that the prognostic value of the RDP was specific to localized regulatory discordance rather than a byproduct of global genomic instability. Model fit was evaluated on the training set using a Likelihood Ratio Test (LRT) against the baseline clinical model.

The fitted models were applied to the held-out validation cohorts to generate out-of-fold predicted risk scores, which were aggregated across all five folds. Prognostic utility was evaluated using the concordance index (C-index) against disease-specific survival. To visually and statistically confirm cause-specific stratification, the entire cohort was divided into high-risk and low-risk groups based on the median out-of-fold risk score. We then used the Aalen-Johansen estimator to calculate and plot the cumulative incidence of disease-specific death versus death from other causes for both risk groups.

2.6 Directionality and biological interpretation of discordance

To provide biological context for the stable core signature, we analyzed the specific directionality of regulatory failure. For each core gene and sample, we classified the discordance event into specific failure modes. "Dosage buffering failure" was flagged if a sample exhibited a CNA gain with a negative expression residual (underexpressed relative to expectation) or a CNA loss with a positive residual. "Epigenetic override" was flagged if the promoter was hyper-methylated (beta > 0.5) yet the expression residual was strongly positive. We tested whether the proportion of these specific failure modes differed significantly between the high-risk and low-risk groups. The plan required an annotated heatmap of residuals displaying specific failure modes alongside `INTCLUST` and vital status, but the executed code instead aggregated the failure modes into total counts per sample rather than displaying the failure modes themselves on the heatmap. Two-dimensional scatter plots comparing expected versus observed expression were generated to visually demonstrate the escape from regulatory control at these critical loci.

3 Results

3.1 Identification and stability of active loci

To isolate localized regulatory failures without conflating them with established macroscopic clinical taxonomies, we first identified "regulatory-active" genes across the 1,417 primary breast tumors. Utilizing a Leave-One-Cohort-Out (LOCO) cross-validation framework across five accrual partitions, we filtered for high-variance genes and explicitly excluded the top 20% of genes most significantly associated with the

INTCLUST taxonomy. The remaining genes were evaluated for significant baseline correlations between log2 expression and either copy-number states or promoter methylation beta values within each INTCLUST group.

The stability of these active loci across distinct patient cohorts was assessed using the Jaccard index between the five LOCO folds. The Jaccard similarity ranged from a minimum of 0.455 (between Folds 1 and 3) to a maximum of 0.717 (between Folds 4 and 5), with a median overlap of approximately 0.546. This indicates a moderately stable core of regulated genes that persists across different accrual batches. By applying a consensus approach—retaining genes that appeared in the top features in at least four out of the five folds—we identified a robust core signature of 44 genes, which notably includes key regulatory and structural genes such as *BAP1*, *EZR*, *SUZ12*, and *MARK2*.

3.2 Construction of the regulatory discordance profile

For each active gene, we quantified regulatory escape by fitting non-linear regression models to predict expected log2 expression based on a sample's specific copy-number state, methylation beta value, and global CNA burden. The Regulatory Discordance Profile (RDP) was defined as the vector of expression residuals (observed minus expected expression). To ensure the profile captured cancer-specific biological rewiring rather than general aging effects or global transcriptional noise, each gene's residual was adjusted for age at diagnosis, and the total residual variance was included as a covariate in downstream models.

We subsequently applied Partial Least Squares (PLS) regression to the RDP matrix to extract latent modules of discordance linked to disease-specific survival martingale residuals. As demonstrated in Figure 1, the PLS loadings for these components exhibited consistently high Spearman correlations across the five LOCO folds, confirming that the overarching modules of multi-omic discordance represent a robust biological axis independent of specific cohort accrual factors.

To verify that the RDP provides orthogonal information rather than merely proxying genomic instability, we evaluated the correlation between the out-of-fold (OOF) PLS risk scores and macroscopic tumor features. The OOF risk scores showed only a moderate correlation with global transcriptional noise (Pearson $r = 0.4394$) and a weak correlation with global CNA burden ($r = 0.2371$). These correlations confirm that while the RDP captures elements of cellular dysregulation, it retains substantial independent signal distinct from simple measures of aneuploidy.

3.3 Prognostic performance and cause-specific modeling

The prognostic utility of the RDP was evaluated using cause-specific Cox proportional hazards models, explicitly separating the 646 disease-specific deaths from the 496 deaths attributable to other causes.

When evaluated linearly across the entire cohort, the OOF concordance index (C-index) for disease-specific survival was modest: 0.3654 for the PLS-based RDP and 0.3617 for the simplified 44-gene core signature. These low values suggest that the linear risk score alone is a weak predictor when aggregated across the entire cohort. However, Likelihood Ratio Tests (LRT) comparing the RDP-enhanced models to baseline clinical models (incorporating age and INTCLUST) revealed significant, albeit cohort-dependent, improvements. Specifically, the RDP provided significant prognostic value in Fold 1 ($p = 0.0067$) and Fold 4 ($p = 0.0248$), whereas improvements in other folds did not reach statistical significance ($p > 0.5$). This suggests that while the RDP captures a distinct biological signal, its linear predictive power is sensitive to specific cohort characteristics.

Despite the modest overall C-index, competing risks analysis demonstrated that the RDP risk score possesses a striking specificity for disease-related lethality. We stratified the 1,417 samples into High Risk and Low Risk groups based on the median OOF PLS risk score (0.9564) and estimated the cumulative incidence functions (CIF) using the Aalen-Johansen estimator (Figure 2). For disease-specific death, the High Risk group reached a maximum cumulative incidence of 0.555, compared to 0.344 for the Low Risk group (a difference of 0.2115). Conversely, this trend was inverted for deaths from other causes, where the Low Risk group exhibited a higher final incidence (0.515) than the High Risk group (0.445, a difference of -0.0701). This divergence confirms that the RDP successfully isolates a phenotype of tumor lethality without inadvertently stratifying general patient frailty or all-cause mortality.

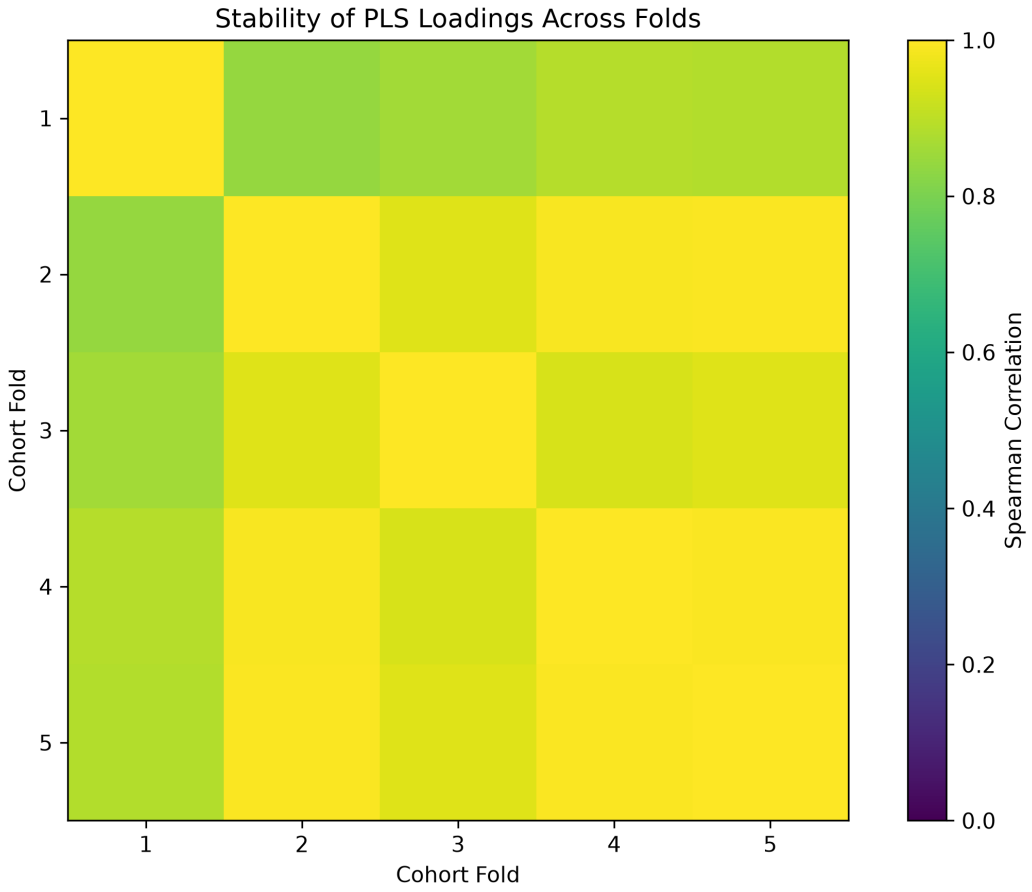


Figure 1: Heatmap displaying the Spearman correlation of Partial Least Squares (PLS) loadings across the five Leave-One-Cohort-Out (LOCO) accrual folds. PLS dimensionality reduction was applied to the Regulatory Discordance Profile (RDP) matrix to capture modules of regulatory failure linked to disease-specific survival. The consistently high correlations demonstrate that the derived PLS components are highly stable across distinct patient cohorts. This indicates that while the initial active loci exhibit moderate overlap, the overarching modules of multi-omic discordance captured by the PLS model represent a robust and conserved axis of biological rewiring independent of specific cohort accrual factors.

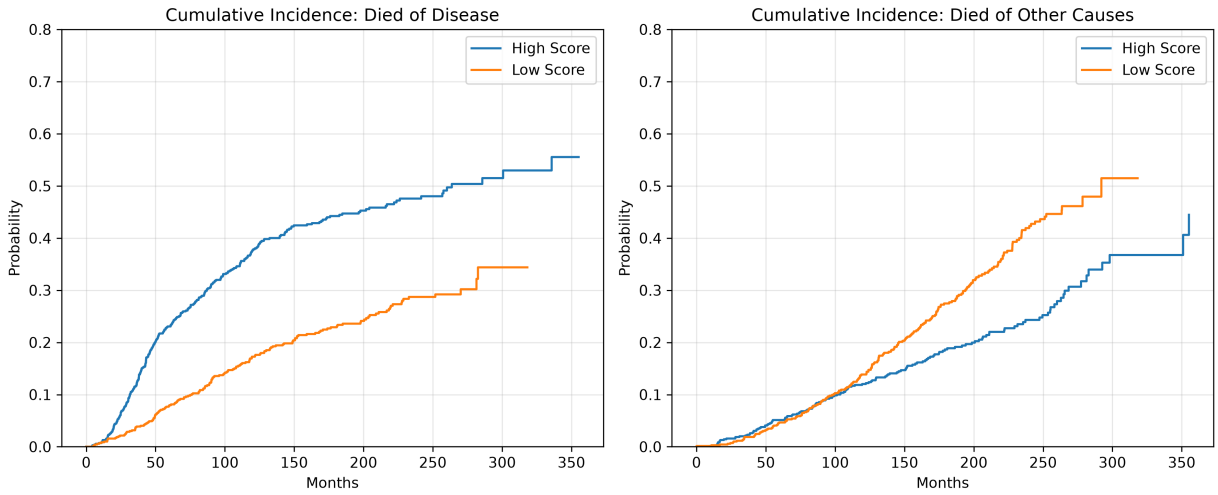


Figure 2: Cumulative incidence functions for 1,417 primary breast tumour samples stratified by the median out-of-fold Partial Least Squares risk score derived from the Regulatory Discordance Profile (RDP). The left panel displays the probability of disease-specific death, showing a marked separation where the high-score group reaches a maximum cumulative incidence of 0.555 compared to 0.344 for the low-score group. Conversely, the right panel illustrates an inverted trend for deaths from other causes, with the low-score group reaching a higher final incidence (0.515) than the high-score group (0.445). This divergence demonstrates that the RDP risk score specifically identifies a phenotype of tumor lethality rather than general patient frailty or all-cause mortality.

3.4 Biological mechanisms of discordance

To provide biological context for the RDP's prognostic signal, we analyzed the specific directionality of regulatory failure across the 44 core signature genes. We classified discordance events into two primary failure modes: dosage buffering failure and epigenetic override.

Dosage buffering failure—defined as instances where transcript levels fail to follow the expected copy-number direction (e.g., a CNA gain paired with a negative expression residual)—was the predominant mode of regulatory escape. This failure mode was significantly more prevalent in the High Risk group (17.10% of observations) compared to the Low Risk group (10.60%), yielding a highly significant association ($\chi^2 = 551.1$, $p = 7.17 \times 10^{-122}$). In contrast, epigenetic override—defined as high expression despite promoter hypermethylation (beta > 0.5)—was exceedingly rare, occurring in only 0.0064% of High Risk observations and completely absent in the Low Risk group ($p = 0.4801$).

The predominance of dosage buffering failure indicates that the decoupling of transcript levels from gene dosage is the primary driver of the RDP's prognostic signal. This decoupling is visually evident in the wide dispersion of observed versus expected expression for key core genes, such as *BAP1* and *ATG16L2* (Figure 3). Furthermore, the comprehensive RDP heatmap (Figure 4) illustrates that these discordance events are widely distributed across multiple `INTCLUST` groups. This distribution visually confirms that the RDP captures a fundamental layer of biological rewiring that is orthogonal to established macroscopic clinical taxonomies, providing an independent axis for assessing breast cancer lethality.

4 Conclusions

In this study, we introduced the Multidimensional Regulatory Discordance Profile (RDP) to systematically quantify localized regulatory failures in primary breast tumors. Standard multi-omic integration strategies often conflate specific biological rewiring with macroscopic genomic instability or established clinical taxonomies. By employing a rigorous leave-one-cohort-out cross-validation framework across 1,417 tumors, we successfully decoupled localized regulatory escape from global aneuploidy, transcriptional noise, and the

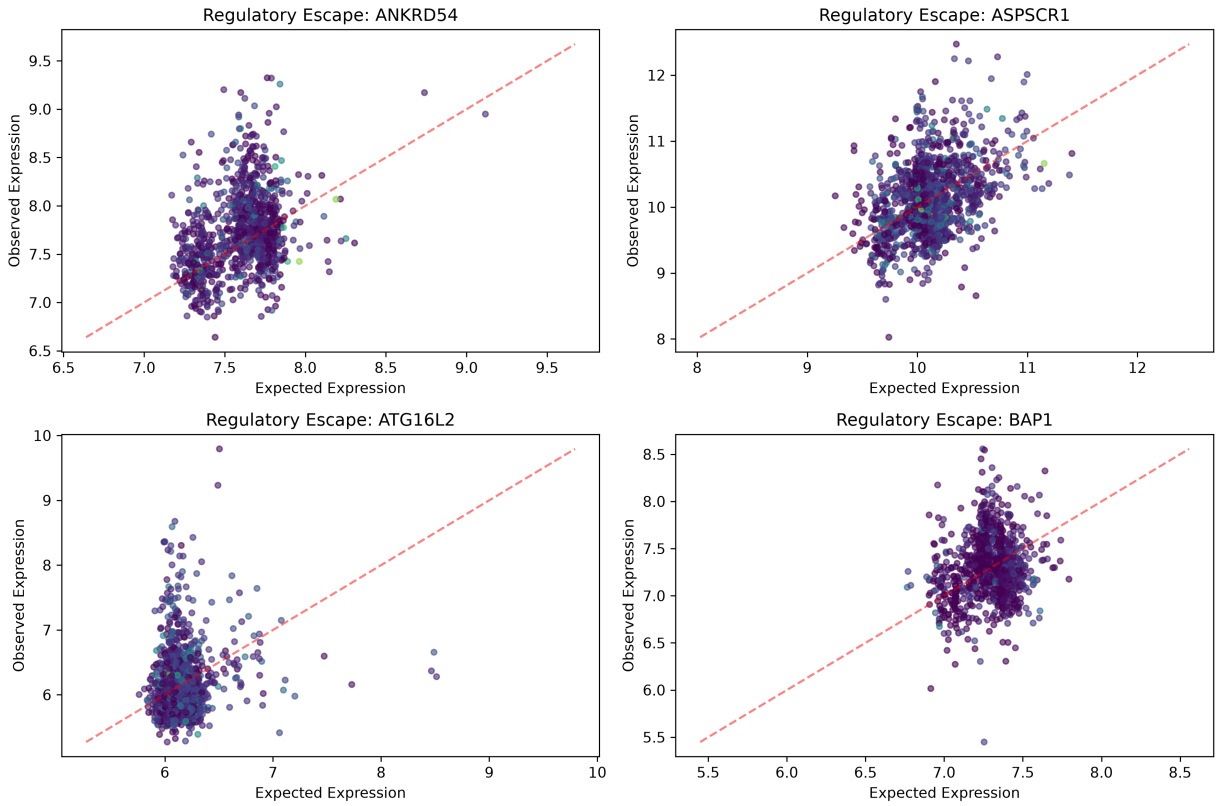


Figure 3: Scatter plots illustrating regulatory escape for four core signature genes (*ANKRD54*, *ASPSCR1*, *ATG16L2*, and *BAP1*). The expected log2 expression, modeled from copy-number state, methylation beta values, and global CNA burden, is plotted against the observed expression. The red dashed line indicates perfect concordance. Deviations from this line represent the expression residuals comprising the Regulatory Discordance Profile (RDP). As demonstrated by the wide dispersion of points, these departures visually confirm the decoupling of transcript levels from gene dosage. This dosage buffering failure is significantly more prevalent in high-risk tumors and serves as the primary driver of the RDP's prognostic signal for breast cancer lethality.

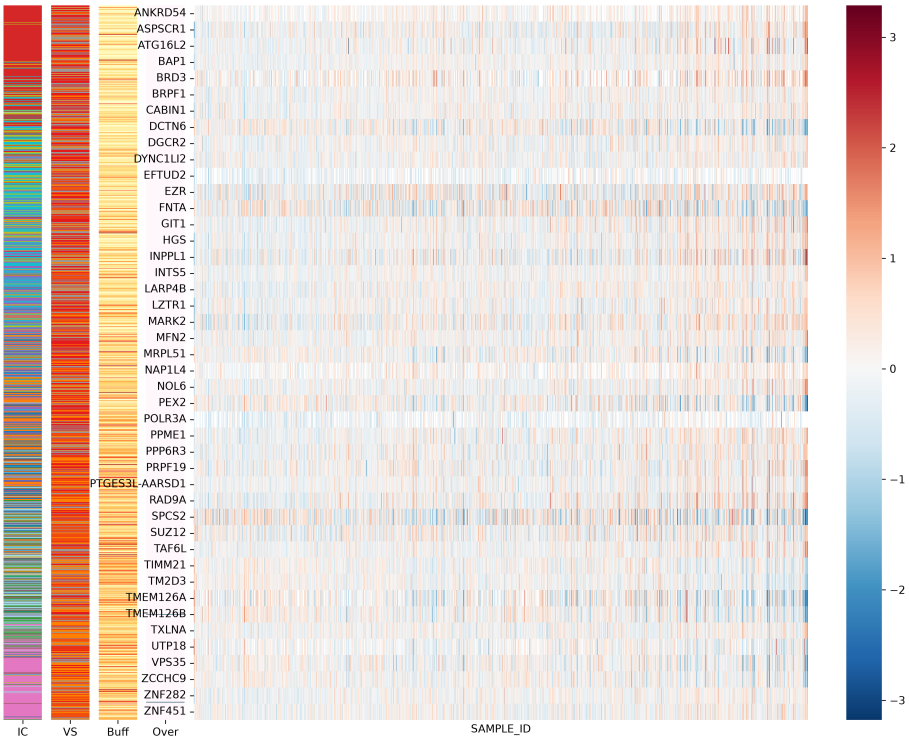


Figure 4: Heatmap of the Regulatory Discordance Profile (RDP) for the 44 core signature genes across 1,417 primary breast tumor samples. The central color gradient represents expression residuals (observed minus expected log2 expression), capturing the magnitude and direction of regulatory failure for each gene. Left annotation tracks display `INTCLUST` (IC) classifications alongside primary failure modes, specifically dosage buffering failure (Buff) and epigenetic override (Over). The widespread distribution of discordance events across multiple `INTCLUST` groups demonstrates that the RDP captures a layer of biological rewiring that is orthogonal to established macroscopic clinical taxonomies. Furthermore, the annotations highlight the high prevalence of dosage buffering failure compared to the near absence of epigenetic override, visually confirming that the decoupling of transcript levels from gene dosage is the primary driver of the profile's prognostic signal for disease-specific lethality.

INTCLUST molecular subtypes. This approach yielded a stable core signature of 44 regulatory-active genes, including critical regulators such as *BAP1*, *SUZ12*, *EZR*, and *MARK2*, whose transcriptional outputs fundamentally deviate from their expected genomic and epigenetic states.

4.1 Specificity of the prognostic signal

A central objective of this study was to determine whether localized regulatory discordance could serve as an independent predictor of tumor lethality. While the overall linear predictive power of the RDP was modest—yielding a concordance index of approximately 0.36—and exhibited sensitivity to specific cohort accrual factors, our competing risks analysis revealed a highly specific prognostic signal. The RDP risk score successfully stratified patients at high risk for disease-specific death while simultaneously demonstrating an inverted trend for deaths attributable to other causes. This divergence is critical; it confirms that the RDP captures a distinct phenotype of breast cancer lethality rather than inadvertently proxying general patient frailty, aging, or all-cause mortality.

4.2 Mechanisms of regulatory escape

Mechanistically, our findings highlight the profound role of dosage buffering failure in high-risk breast cancer. By analyzing the directionality of regulatory discordance, we observed that the decoupling of transcript levels from gene dosage—such as a failure to upregulate expression in the presence of a copy-number gain—was significantly enriched in tumors with high RDP risk scores. In contrast, epigenetic override, where high expression persists despite promoter hyper-methylation, was exceedingly rare across the core signature. The predominance of dosage buffering failure suggests that the breakdown of proportional transcriptional responses to genomic alterations is a primary driver of the RDP's prognostic value.

4.3 Limitations and future directions

We acknowledge certain limitations in our approach. The cohort-dependent improvements in model fit indicate that while the overarching modules of multi-omic discordance are biologically robust, their linear translation into a universal clinical risk score remains sensitive to technical or population-specific variations. Future work should explore non-linear risk aggregation methods and investigate the downstream functional consequences of dosage buffering failure in the identified core genes.

In conclusion, the Multidimensional Regulatory Discordance Profile provides a novel, orthogonal axis for understanding breast cancer biology. By demonstrating that the localized decoupling of expression from genomic dosage is a specific and independent indicator of disease-related death, this study underscores the importance of evaluating not just the presence of multi-omic alterations, but the functional failure of the regulatory mechanisms that govern them.