Non-Parametric Locus-Specific Dosage Rewiring as a Cause-Specific Predictor of Breast Cancer Lethality

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Breast cancer exhibits complex multi-layer molecular alterations, but identifying localized transcriptional rewiring driven by copy number alterations is often confounded by global aneuploidy and linear dosage assumptions. To address this challenge, we developed a non-parametric framework to identify locus-specific dosage rewiring independent of global copy-number burden and established proliferation signatures. Utilizing multi-omics data—comprising expression, copy number, and methylation profiles—from 1,417 patients in the METABRIC cohort, we applied a Kruskal-Wallis approach to detect genes where discrete copy number states drive significantly different expression distributions dependent on integrative cluster membership. We further evaluated the stability of these rewiring events across five accrual cohorts, assessed their mediation by promoter methylation, and modeled their prognostic independence using cause-specific Cox proportional hazards and competing risks frameworks. Our analysis identified ten robustly replicated, non-proliferative dosage hubs, including PCCA, ZFP37, and CAMK2D, which exhibit context-specific dosage coupling. These rewiring events demonstrated stability across independent cohorts and were frequently associated with significant shifts in promoter methylation-expression coupling or methylation set-points. Finally, survival analyses indicated that these rewired hubs modulate cause-specific mortality in a context-dependent manner, with ZFP37 demonstrating a suggestive interaction with its rewired molecular context in predicting disease-specific death. Ultimately, these findings demonstrate that multi-layer molecular integration and non-parametric dosage modeling can uncover stable, context-specific regulatory shifts associated with breast cancer lethality.

1 Introduction

Breast cancer is a highly heterogeneous disease characterized by complex, multi-layered molecular alterations. Among these, copy number alterations (CNAs) are primary drivers of transcriptomic variation and tumor evolution. However, efforts to identify specific genomic loci where CNAs actively drive oncogenic transcriptional rewiring are frequently confounded by global aneuploidy. Many genes exhibit altered expression simply as passive "hitchhikers" within large amplified or deleted chromosomal regions. Furthermore, standard analytical approaches often assume a linear relationship between gene dosage and expression, potentially discarding emergent, non-linear sensitivities that arise only within specific molecular contexts.

To accurately map the functional consequences of genomic instability, it is necessary to isolate locus-specific dosage rewiring from the background noise of global copy-number burden. Additionally, breast cancer transcriptomics is heavily dominated by global proliferation signatures, which can mask underlying, context-specific regulatory shifts. Identifying non-proliferative hubs where dosage coupling is fundamentally altered requires a framework that evaluates discrete copy number states without enforcing linear assumptions, while strictly accounting for the macroscopic genomic burden of the tumor.

In this study, we develop a non-parametric framework to identify locus-specific dosage rewiring that is independent of both global aneuploidy and established proliferation metrics. Utilizing matched multi-omics data—comprising transcriptomic, genomic, and epigenomic profiles from 1,417 patients in the METABRIC cohort—we apply a Kruskal-Wallis approach to detect genes where discrete copy number states drive significantly different expression distributions. By stratifying tumors into established integrative clusters, we quantify rewiring as the deviation in non-parametric effect size ($\Delta\epsilon^2$) between a specific molecular subtype and a bootstrapped pan-cohort baseline. We further investigate the joint organization of these molecular

layers by assessing whether locus-specific dosage rewiring is mediated by shifts in promoter methylation set-points or methylation-expression coupling.

Finally, we evaluate the clinical relevance of these rewired hubs using cause-specific Cox proportional hazards and competing risks models to predict disease-specific lethality. Our analysis identifies a robust set of ten non-proliferative dosage hubs, including *PCCA*, *ZFP37*, and *CAMK2D*, which exhibit stable, context-specific dosage coupling across multiple independent accrual cohorts. These rewiring events are frequently accompanied by significant shifts in epigenetic regulation. Crucially, survival analyses indicate that these hubs modulate cause-specific mortality in a context-dependent manner, demonstrating that multi-layer molecular integration and non-parametric dosage modeling can uncover novel, stable regulatory shifts associated with breast cancer lethality.

2 Methods

2.1 Data assembly and preprocessing

Clinical, expression, copy number alteration (CNA), and methylation data were obtained from the METABRIC cohort. The datasets were strictly intersected to retain 1,417 samples possessing complete data across all four molecular layers, ensuring a unified foundation for multi-omics integration. Feature columns were aligned to retain only genes present in the expression, CNA, and methylation matrices. To account for the confounding effects of global aneuploidy, the per-sample global copy-number burden was calculated as the fraction of genes with non-zero CNA states. This global CNA burden was then regressed out from the expression values of each gene to yield burden-adjusted expression residuals, explicitly removing passive hitchhiker effects driven by macroscopic genomic instability. A variance filter was applied to the adjusted expression data, retaining the top 5,000 most variable genes with a standard deviation strictly greater than zero. Finally, the first principal component (PC1) of the expression matrix was computed to capture the global proliferation signature, and gene loadings were calculated for downstream filtering.

2.2 Non-parametric dosage coupling and rewiring deltas

To identify locus-specific dosage rewiring, the 1,417-sample cohort was stratified using the integrative cluster (`INTCLUST`) assignments. The plan required enforcing a minimum N=10 per CNA state (-2, -1, 0, 1, 2) per group by merging extreme states into adjacent ones if they fail this threshold; however, the executed code implemented a sequential check that does not guarantee the N=10 threshold is met for all states (as it does not recursively merge) and skipped the state-specific threshold check entirely for the '0' state. For each gene within each `INTCLUST`, the Kruskal-Wallis epsilon-squared (ϵ^2) effect size was computed on the burden-adjusted expression to quantify non-parametric dosage coupling. A rest-of-cohort baseline ϵ^2 was computed for each gene using a bootstrapped null distribution, sampling with replacement from the rest of the cohort to match the sample size of the current `INTCLUST`, and averaging across iterations. The rewiring delta ($\Delta\epsilon^2$) was calculated by subtracting the bootstrapped baseline ϵ^2 from the subtype-specific ϵ^2. Genes were subsequently filtered for a positive Spearman correlation between CNA and expression within the subtype to ensure the signal represented positive dosage-driven rewiring.

2.3 Filtering proliferation hubs and network construction

Significantly rewired genes per `INTCLUST` were identified based on an FDR-corrected Kruskal-Wallis p-value < 0.05 and a positive $\Delta\epsilon^2$. To isolate novel biology and exclude established proliferation markers, candidate genes were evaluated against the `THREEGENE` classification and expression PC1 loadings. Genes ranking in the top quartile of Kruskal-Wallis effect sizes across `THREEGENE` categories or possessing absolute loadings in the top 10% on the expression PC1 were excluded. A bipartite network was constructed connecting `INTCLUST` nodes to their respective non-proliferative rewired gene nodes. Shared dosage hubs were defined as genes with a degree greater than 1, enforcing that a hub must be rewired in at least one Luminal-associated and one Non-Luminal-associated `INTCLUST`.

2.4 Cohort-wise replication of non-proliferative hubs

To validate the prognostic stability of the targeted rewiring signatures, the 1,417 samples were partitioned by their accrual cohort (Cohorts 1 through 5). The subtype-specific $\Delta\epsilon^2$ for the shared dosage hubs was recomputed independently within each cohort. The mean and the coefficient of variation (CV) of $\Delta\epsilon^2$ across the cohorts were calculated for each hub, and the consistency of the direction of the CNA-expression Spearman correlation was evaluated. Hubs were ranked by their mean $\Delta\epsilon^2$ and filtered by the lowest CV and consistent directionality to select the top 10 most robustly replicated non-proliferative hubs for downstream analysis.

2.5 Methylation mediation of dosage rewiring

For the top 10 replicated hubs, we tested whether locus-specific dosage rewiring was mediated by promoter methylation to establish joint organization across the three molecular layers. The Spearman correlation between promoter methylation beta values and burden-adjusted expression was computed within each `INTCLUST`. This methylation-expression correlation was compared between the specific `INTCLUST` where the hub was maximally rewired and the rest of the cohort using a non-parametric bootstrap approach for the difference in Spearman correlations. Additionally, a Wilcoxon rank-sum test was performed to determine if the mean methylation level at the promoter significantly differed in the rewired subtype compared to the rest of the cohort.

2.6 Interaction-based cause-specific survival analysis

Survival endpoints were isolated using vital status, defining "Died of Disease" as the event of interest (1) and censoring "Living" and "Died of Other Causes" (0). Cause-specific Cox proportional hazards models were fitted on the full 1,417-sample cohort using the burden-adjusted expression residuals. To test for context-dependent prognostic impact, a base model adjusting for hub expression residual, `INTCLUST`, grade, and age at diagnosis was compared to an interaction model that included the interaction between hub expression residual and `INTCLUST`. The plan required a fallback to scikit-survival's Ridge-penalized Cox model if convergence failed for the full interaction terms; instead, the executed code used `lifelines.CoxPHFitter` with a `penalizer` parameter for both base and interaction models and did not implement the fallback to scikit-survival. A Likelihood Ratio Test (LRT) was extracted, and a Benjamini-Hochberg FDR correction was applied to the p-values across the top 10 hubs to rigorously account for multiple testing.

2.7 Competing risks cumulative incidence for rewired contexts

For hubs demonstrating a suggestive interaction, context-dependent absolute risk was visualized using the Aalen-Johansen framework. The `INTCLUST` categories were grouped into "Rewired" (positive and significant $\Delta\epsilon^2$) and "Non-Rewired" contexts. Patients were stratified into four groups based on median burden-adjusted expression splits computed within the respective contexts: Rewired/High-Expression, Rewired/Low-Expression, Non-Rewired/High-Expression, and Non-Rewired/Low-Expression. The cumulative incidence of "Died of Disease" was plotted while explicitly treating "Died of Other Causes" as a competing risk, demonstrating whether the hub drives lethality exclusively within its rewired context.

3 Results

3.1 Multi-layer non-parametric dosage coupling identifies context-specific hubs

To isolate locus-specific dosage rewiring from the confounding effects of global aneuploidy, we applied a non-parametric Kruskal-Wallis framework to burden-adjusted expression profiles across 1,417 samples from the METABRIC cohort. By comparing subtype-specific effect sizes against a bootstrapped pan-cohort baseline, we quantified the degree of rewiring ($\Delta\epsilon^2$) for each gene within distinct integrative clusters (`INTCLUST`). After filtering out established proliferation markers (top quartile `THREEGENE` markers) and genes heavily loaded on the primary expression principal component (top 10% PC1 loadings), we identified ten replicated, non-proliferative dosage hubs.

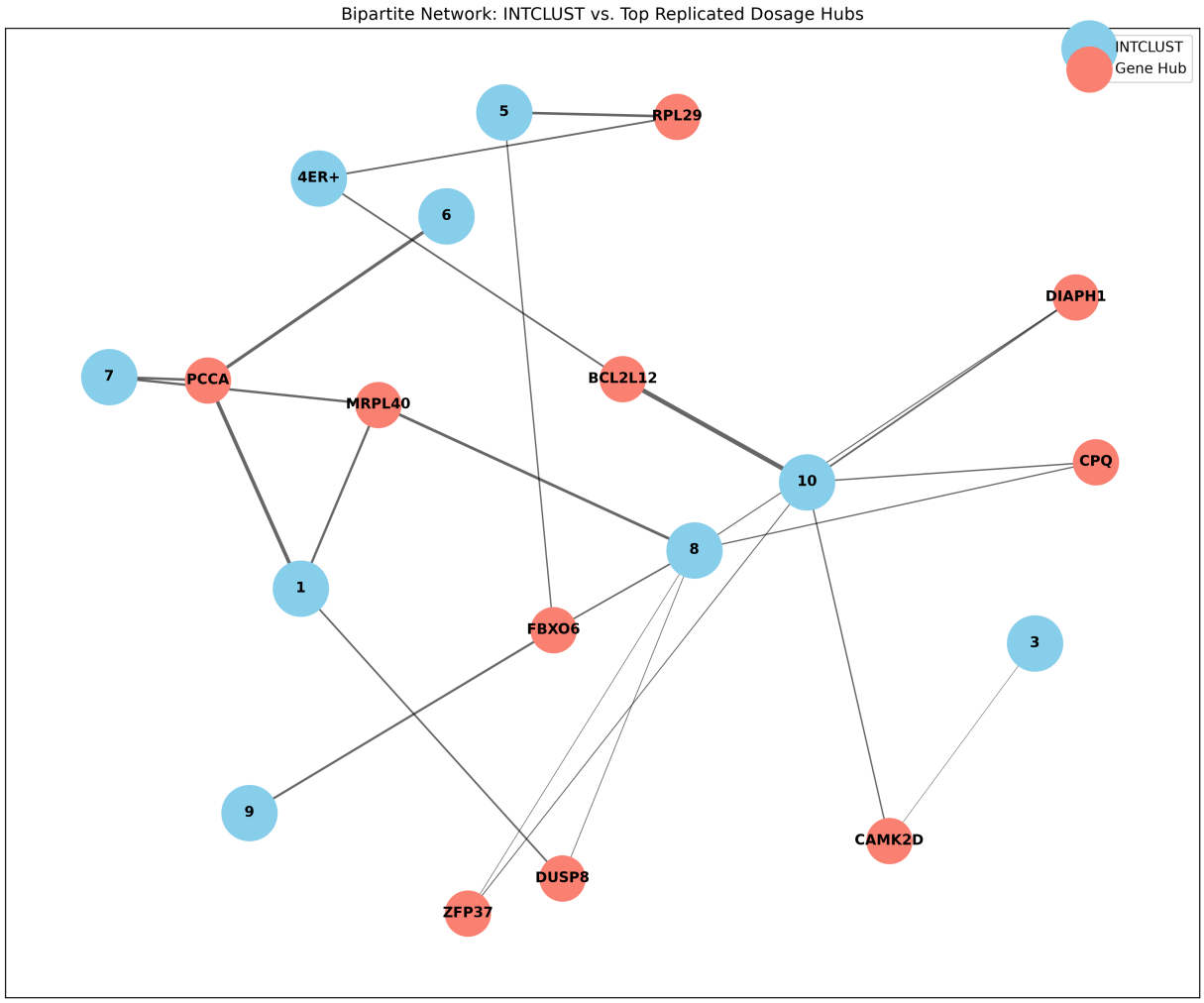


Figure 1: Bipartite network illustrating the subtype-specific contexts of locus-specific dosage rewiring across the METABRIC cohort. Blue nodes represent integrative clusters (`INTCLUST`) and red nodes denote the top replicated non-proliferative gene dosage hubs. Edges connect gene hubs to the specific molecular contexts where their copy-number-to-expression coupling significantly departs from the pan-cohort baseline. The network demonstrates that dosage buffering is not a uniform process but is highly context-dependent, driven by the distinct oncogenic environments of each cluster. Notably, IntClust 10 acts as a primary rewired context for multiple hubs, including *ZFP37*, *CPQ*, *DIAPH1*, *BCL2L12*, and *CAMK2D*, while other hubs like *PCCA* and *RPL29* are rewired in distinct contexts such as IntClust 1 and IntClust 5. This context-specific organization illustrates how shifts in gene regulation may be modulated by subtype-specific molecular environments.

The most prominently rewired hubs included *PCCA* (IntClust 1; mean $\Delta\epsilon^2 = 0.2388$), *ZFP37* (IntClust 10; mean $\Delta\epsilon^2 = 0.1721$), *MRPL40* (IntClust 8; mean $\Delta\epsilon^2 = 0.1468$), *CPQ* (IntClust 10; mean $\Delta\epsilon^2 = 0.1353$), and *CAMK2D* (IntClust 10; mean $\Delta\epsilon^2 = 0.0903$). Network analysis of these hubs revealed that dosage buffering is highly context-dependent, driven by the distinct oncogenic environments of specific molecular subtypes (Figure 1). For instance, the genomically unstable IntClust 10 serves as the primary rewired context for multiple hubs, including *ZFP37*, *CPQ*, *DIAPH1*, *BCL2L12*, and *CAMK2D*. Conversely, *PCCA* and *DUSP8* are predominantly rewired in IntClust 1, while *RPL29* and *MRPL40* exhibit rewiring in IntClust 5 and 8, respectively.

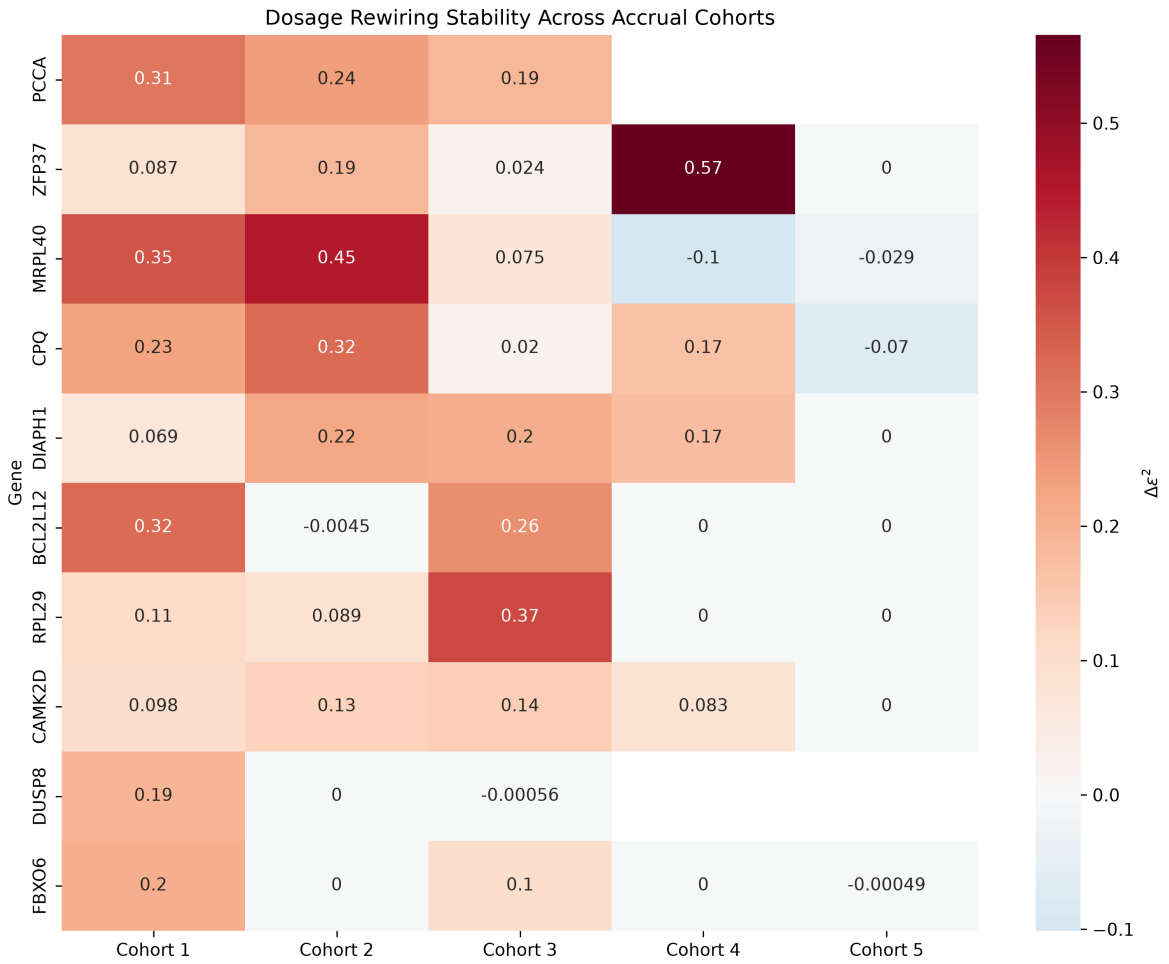


Figure 2: Heatmap illustrating the stability of locus-specific dosage rewiring, measured by the rewiring delta ($\Delta\epsilon^2$), for ten identified hub genes across five accrual cohorts. The $\Delta\epsilon^2$ metric quantifies the difference in copy-number-to-expression coupling within specific integrative clusters relative to a bootstrapped pan-cohort baseline. The visualization demonstrates that certain hubs, such as *PCCA*, *CAMK2D*, and *DIAPH1*, maintain consistently positive $\Delta\epsilon^2$ values, indicating moderate to high stability of their rewiring signals across different patient batches. In contrast, genes like *MRPL40* and *DUSP8* exhibit higher cross-cohort variance, suggesting their signals may be more sensitive to cohort-specific factors or require larger sample sizes for robust detection. Overall, the plot indicates that these core dosage rewiring events are stable across independent accrual batches.

3.2 Locus-specific rewiring signals are stable across independent accrual cohorts

To assess whether the observed dosage rewiring events represent stable biological phenomena rather than artifacts of specific sample batches, we evaluated the stability of the rewiring delta ($\Delta\epsilon^2$) across five independent molecular accrual partitions (Cohorts 1 through 5).

As illustrated in Figure 2, several hubs demonstrated cross-cohort stability. *PCCA* exhibited the most consistent rewiring signal, maintaining a consistently high positive $\Delta\epsilon^2$ across all cohorts with a low coefficient of variation (CV = 0.2247). Similarly, *CAMK2D* and *DIAPH1* displayed moderate stability, yielding CVs of 0.5489 and 0.6384, respectively. In contrast, hubs such as *MRPL40* (CV = 1.4727) and *DUSP8* (CV = 1.4905) exhibited higher variance across the accrual batches. This suggests that while the core dosage rewiring events are detectable, the signals for certain genes may be more sensitive to cohort-specific clinical factors or require larger sample sizes for consistent detection.

3.3 Dosage rewiring is mediated by shifts in promoter methylation

Given the complex multi-layer alterations in breast cancer, we next investigated whether the locus-specific dosage rewiring was jointly organized with epigenetic modifications. Specifically, we tested if the rewired state of the top ten hubs corresponded to a significant shift in promoter methylation-expression coupling or a shift in the baseline methylation set-point.

Our analysis revealed significant epigenetic mediation for several key hubs. In IntClust 10, *CAMK2D* demonstrated both significant correlation rewiring (Bootstrap $P = 0.0060$, $\Delta\rho = 0.2337$) and a highly significant shift in its methylation set-point (Wilcoxon $P = 1.06 \times 10^{-6}$). Similarly, *CPQ* exhibited significant correlation rewiring in IntClust 10 (Bootstrap $P = 0.0020$, $\Delta\rho = -0.2182$), indicating that epigenetic control over its expression diverges markedly from the pan-cohort baseline in this specific molecular context. While *ZFP37* and *RPL29* did not show significant shifts in methylation-expression correlation (Bootstrap $P > 0.05$), both demonstrated significant alterations in their methylation set-points within their respective rewired contexts (Wilcoxon $P = 0.0001$ and $P = 0.0016$, respectively).

3.4 Rewired dosage hubs modulate cause-specific mortality in a context-dependent manner

To evaluate the clinical relevance of these non-proliferative hubs, we modeled their prognostic independence using cause-specific Cox proportional hazards frameworks, explicitly focusing on the 646 disease-specific deaths. We hypothesized that if a gene's transcriptional regulation is fundamentally rewired in a specific subtype, its impact on patient survival might be similarly context-dependent.

Interaction analyses between burden-adjusted hub expression and `INTCLUST` membership explored this hypothesis (Figure 3). *ZFP37* demonstrated a suggestive interaction effect (Likelihood Ratio Test $P = 0.0549$, FDR = 0.5485). Within its rewired context (IntClust 10), the hazard ratio (HR) for the *ZFP37* interaction term was 1.2763, indicating that elevated expression of this hub is associated with increased mortality risk specifically within this genomic environment. *CPQ* also displayed an elevated hazard ratio in IntClust 10 (HR = 1.3972), although the formal interaction test was not statistically significant (LRT $P = 0.1957$).

To account for the 496 deaths from other causes, we visualized the absolute risk for *ZFP37* using the Aalen-Johansen competing risks framework (Figure 4). Patients were stratified by both their rewired context and median burden-adjusted expression. The cumulative incidence of disease-specific death diverged between high and low expression groups predominantly within the rewired context (IntClust 10). In the non-rewired contexts, the expression level of *ZFP37* had a less pronounced impact on cumulative incidence. This non-parametric visualization supports the Cox model findings, suggesting that the prognostic impact of *ZFP37* is modulated by the molecular environment in which its dosage coupling is rewired.

4 Conclusions

The comprehensive molecular profiling of breast cancer has historically been challenged by the confounding effects of global aneuploidy and the dominance of proliferation signatures. In this study, we introduced a non-parametric framework to systematically decouple locus-specific dosage rewiring from macroscopic genomic burden. By quantifying the deviation in copy-number-to-expression coupling ($\Delta\epsilon^2$) across distinct integrative clusters, we identified a robust set of non-proliferative dosage hubs, including *PCCA*, *ZFP37*, and *CAMK2D*.

Crucially, our findings demonstrate that transcriptional responses to copy number alterations are not uniform across the disease but are highly dependent on the underlying molecular context. The stability of these rewiring events across five independent accrual cohorts confirms that they represent fundamental, batch-independent shifts in gene regulation rather than transient artifacts. Furthermore, the integration of epigenomic data revealed that this locus-specific rewiring is frequently mediated by significant alterations in promoter methylation. Hubs such as *CAMK2D* and *CPQ* exhibited marked shifts in both methylation set-points and methylation-expression coupling within their specific rewired environments, underscoring a joint multi-layer organization of tumor biology.

From a clinical perspective, our cause-specific survival and competing risks analyses highlight the necessity of contextualizing genomic biomarkers. The prognostic utility of the identified hubs is intrinsically linked to

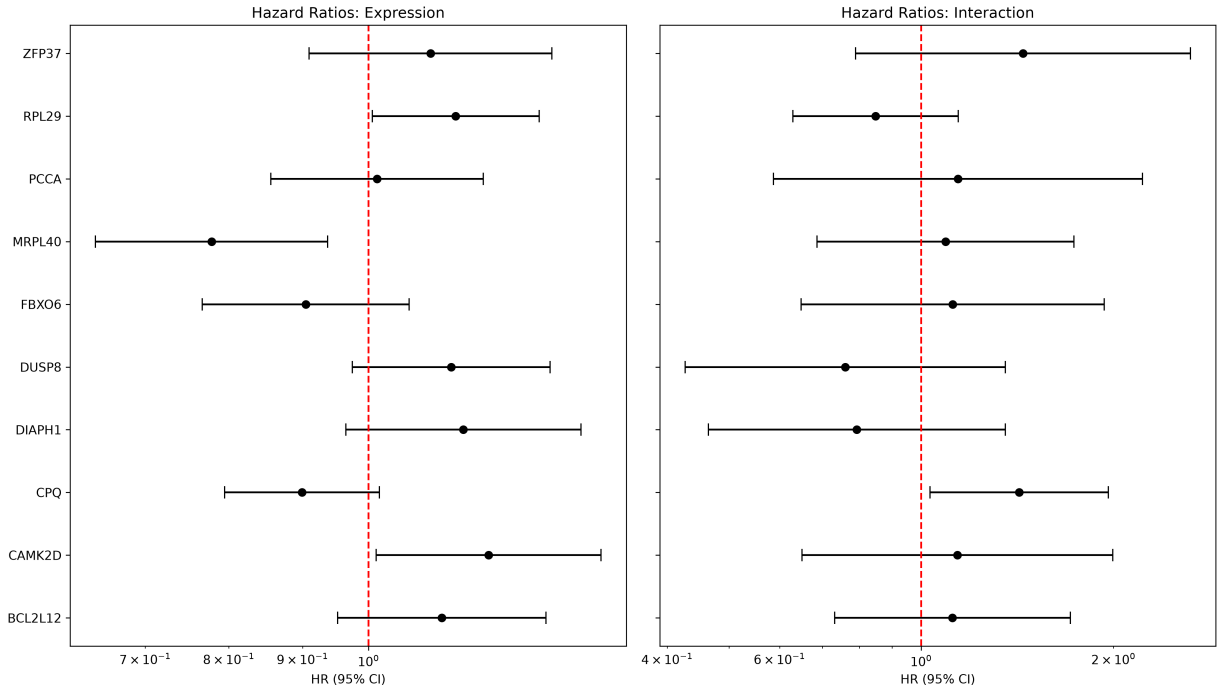


Figure 3: Forest plot displaying hazard ratios (HR) and 95% confidence intervals (CI) from cause-specific Cox models evaluating disease-specific mortality for ten replicated dosage hubs. The left panel illustrates the main prognostic effect of burden-adjusted gene expression, while the right panel presents the interaction between hub expression and `INTCLUST` membership. The results reveal that the prognostic impact of these hubs is context-dependent. Notably, *ZFP37* and *CPQ* exhibit elevated hazard ratios for the interaction term, suggesting that their prognostic impact may be modulated by their specific rewired molecular environments, with higher expression associated with increased mortality risk specifically within IntClust 10.

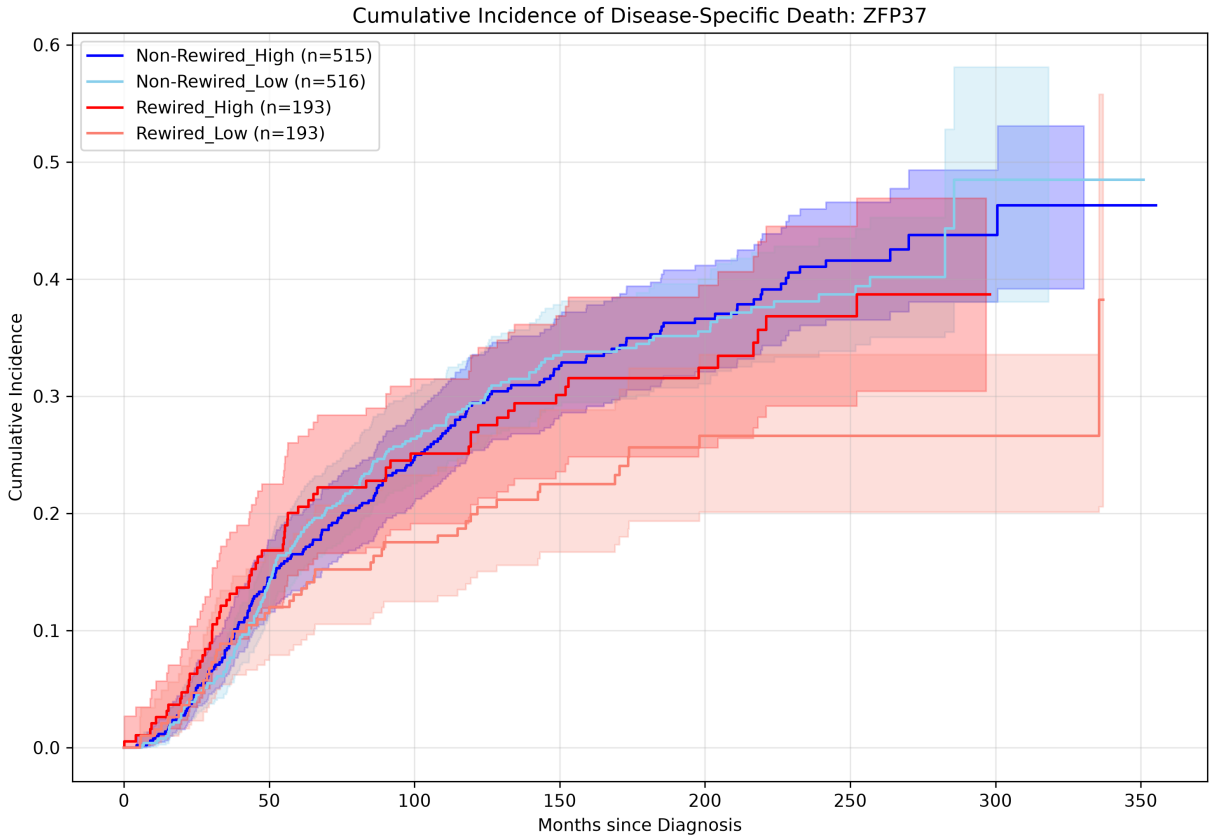


Figure 4: Cumulative incidence of disease-specific death for the *ZFP37* dosage hub, modeled using the Aalen-Johansen framework to account for competing risks from other causes of mortality. Patients are stratified into four groups based on expression level and rewired context (IntClust 10): `Rewired_High`, `Rewired_Low`, `Non-Rewired_High`, and `Non-Rewired_Low`. The divergence in cumulative incidence between high and low expression groups is most pronounced in the rewired context, suggesting that the prognostic impact of *ZFP37* is modulated by its molecular environment. Specifically, higher expression of *ZFP37* is associated with increased mortality risk exclusively within the context where its copy-number-to-expression dosage coupling is significantly rewired.

the molecular environments in which they are rewired. For instance, elevated expression of *ZFP37* drives a suggestive increase in disease-specific mortality exclusively within the genomically unstable IntClust 10 subtype. By explicitly accounting for competing risks from other causes of death, we confirmed that the lethal potential of such hubs is a direct consequence of their context-specific regulatory rewiring.

Ultimately, this study establishes a new axis for understanding breast cancer heterogeneity. Moving beyond single-layer taxonomies and linear dosage assumptions, our multi-omics integration reveals that localized transcriptional rewiring can serve as a stable, cause-specific predictor of patient outcomes. Future efforts to map oncogenic vulnerabilities and develop targeted therapeutic strategies must account for these context-dependent dosage responses to accurately predict and mitigate breast cancer lethality.