Principal Components of Subtype-Specific Dosage Sensitivity Outliers as Cause-Specific Predictors of Breast Cancer Lethality

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Transcriptional responses to copy-number alterations (CNAs) often exhibit subtype-specific rewiring, yet capturing these localized regulatory failures without confounding from global genomic instability remains challenging. To address this, we systematically investigated the joint organization of CNAs and gene expression across 1,980 breast cancer tumors from the METABRIC dataset. We computed robust linear slopes of expression against discrete copy-number states within predefined integrative clusters and compared them to a global baseline to identify dosage sensitivity outlier genes. By extracting patient-level expression residuals for these outliers, we applied Principal Component Analysis to derive low-dimensional dosage sensitivity profiles. We then evaluated the prognostic utility of these profiles using cause-specific Cox proportional hazards models—adjusting for global copy-number burden, age, and baseline cluster risks—and validated our framework across five distinct molecular accrual partitions. Our results reveal that dosage sensitivity is driven by focal regulatory rewiring rather than large-scale chromosomal instability, with the majority of outliers showing maximum deviation in basal-like tumors. Furthermore, the derived principal components capture distinct biological programs independent of global genomic burden and emerge as highly significant, independent predictors of disease-specific mortality, but not of death from other causes. These findings demonstrate that subtype-specific dosage sensitivity profiles represent a stable, reproducible, and orthogonal axis of breast cancer lethality.

1 Introduction

Breast cancer is a highly heterogeneous disease driven by a complex interplay of genomic and transcriptomic alterations. Copy-number alterations (CNAs) represent a primary mechanism by which tumors modulate gene expression to achieve and maintain oncogenic states. However, the transcriptional response to gene dosage is rarely uniform across all tumors; rather, it is frequently rewired in a subtype-specific manner. Understanding how specific molecular contexts dictate the coupling between gene copy number and transcriptional output is essential for unraveling the regulatory networks that drive tumor progression and patient outcomes.

Despite the recognized importance of CNAs, systematically capturing these localized regulatory failures remains a significant analytical challenge. Traditional approaches often struggle to disentangle focal, subtype-specific dosage sensitivity from the confounding effects of global genomic instability and broad chromosomal aberrations. Furthermore, the zero-inflated nature of discrete CNA states complicates the identification of genes whose transcriptional response to dosage is uniquely altered within specific molecular subgroups. Consequently, critical axes of tumor lethality driven by localized regulatory rewiring may be entirely overlooked by global, unstratified models that assume a uniform dosage response across all patients.

To address this gap, we present a systematic framework to quantify subtype-specific dosage sensitivity and evaluate its independent prognostic utility. Leveraging matched genomic and transcriptomic data from 1,980 breast cancer tumors in the METABRIC dataset, we compute robust linear slopes of gene expression against discrete copy-number states strictly within predefined integrative clusters. By contrasting these intra-cluster dosage slopes against a global baseline, we identify a targeted subset of dosage sensitivity outlier genes that exhibit localized regulatory failure. We then extract patient-level expression residuals for these specific outliers and apply Principal Component Analysis to derive low-dimensional dosage sensitivity profiles, thereby preventing overfitting in downstream clinical evaluations.

Our investigation reveals that dosage sensitivity is predominantly driven by focal regulatory rewiring rather than large-scale chromosomal instability, with the majority of outlier genes exhibiting their maximum deviation in basal-like tumors. Crucially, the derived principal components capture distinct biological programs that are orthogonal to global copy-number burden and standard clinical taxonomies. Using cause-specific Cox proportional hazards models to explicitly isolate disease-specific deaths from competing risks, we demonstrate that these dosage sensitivity profiles emerge as highly significant, independent predictors of breast cancer mortality. Through rigorous validation across five distinct molecular accrual partitions, we establish that subtype-specific dosage sensitivity represents a stable, reproducible, and orthogonal axis of breast cancer lethality.

2 Methods

2.1 Data alignment and preprocessing

To systematically investigate the joint organization of copy-number alterations (CNAs) and gene expression, we utilized matched genomic and transcriptomic data from the METABRIC cohort. The clinical, expression, and CNA datasets were aligned using unique patient identifiers, yielding a final analytical cohort of 1,980 samples present across all three data modalities. For each tumor, a global copy-number burden metric was computed as the fraction of genes exhibiting a non-zero discrete CNA state. To ensure that the downstream signal captured meaningful and actively expressed biological variation, the gene expression matrix was filtered to retain only genes with a median $\log_2$ expression greater than 5.0 and expression variance in the top 50% across the cohort. Categorical clinical variables, including integrative clusters (`INTCLUST`) and `THREEGENE` classifications, were one-hot encoded to facilitate downstream orthogonality checks.

2.2 Intra-cluster residualization and dosage slope computation

To isolate localized regulatory rewiring from broad genomic instability, we modeled the transcriptional response to gene dosage strictly within predefined molecular contexts. The 1,980 samples were stratified by their `INTCLUST` assignment. For each gene present in both the filtered expression and CNA matrices, we fit a robust linear regression model predicting $\log_2$ expression from the discrete CNA state. This was performed exclusively within valid `INTCLUST` categories containing a sufficient sample size ($n \geq 50$).

From these models, we extracted the intra-cluster dosage slopes and their associated standard errors. To establish a reference for normal dosage response, a global baseline dosage slope was computed for each gene using an unweighted average of the cluster-specific slopes, preventing the baseline from being disproportionately biased by the most populous clusters. Finally, patient-level expression residuals were calculated within each `INTCLUST` category by subtracting the cluster-specific predicted expression from the observed expression.

2.3 Identification of sensitivity outliers and subtype enrichment

To identify genes exhibiting localized regulatory failure, we calculated the deviation of each gene's intra-cluster dosage slopes from its global baseline slope. These deviations were standardized into Z-scores, weighted by the inverse of the standard error of the respective cluster-specific slope. We defined "sensitivity outliers" as the top 5% of genes exhibiting the highest maximum weighted deviation across the clusters. Although the research plan required the evaluation of enrichment in specific `INTCLUST` categories and tabulation of overlap with canonical `THREEGENE` and `CLAUDIN_SUBTYPE` markers, the executed code evaluated enrichment by cluster (which were not explicitly defined as `INTCLUST`) and evaluated overlap with PAM50 and `THREEGENE` markers, completely skipping the evaluation of `CLAUDIN_SUBTYPE` markers.

2.4 Derivation of dosage sensitivity profiles and network analysis

To prevent overfitting in downstream survival tasks, we derived low-dimensional dosage sensitivity profiles for each tumor. A feature matrix was constructed using strictly the intra-cluster expression residuals of the identified sensitivity outlier genes. This matrix was standardized, and Principal Component Analysis

(PCA) was applied to extract the top three principal components (PC1–PC3). The success of the intra-cluster residualization was verified by computing the R^2 between the extracted PCs and the one-hot encoded taxonomic variables via linear regression. Furthermore, a partial correlation analysis was conducted between PC1–PC3 and the global copy-number burden metric to confirm that the principal components captured a regulatory phenomenon independent of overall genomic instability.

To determine whether the dosage sensitivity signal was driven by large-scale chromosomal instability or focal regulatory rewiring, we computed a pairwise Spearman correlation matrix of the discrete CNA states for the outlier genes. A co-amplification network was constructed retaining only edges with a strong correlation (Spearman $\rho > 0.8$). We extracted the connected components of this network and mapped the top-loaded genes from the principal components onto these structures to evaluate the topology of the amplification events.

2.5 Functional interpretation and visualization

The biological drivers of the dosage sensitivity profiles were interpreted by extracting the gene loadings for PC1, PC2, and PC3. For each component, the top 50 genes with the highest absolute loadings were identified. We performed a data-driven functional interpretation by correlating the patient-level PC scores with the expression levels of canonical breast cancer markers (e.g., *ESR1*, *ERBB2*, and *AURKA*) as well as with tumor `CELLULARITY` to ensure the components were not artifacts of tumor purity. To visualize the mechanism of regulatory failure, scatter plots of $\log_2$ expression versus discrete CNA state were generated for the top genes contributing to the most prognostic principal component, explicitly contrasting the intra-cluster slopes against the global baseline.

2.6 Survival analysis and cohort validation

The independent prognostic utility of the dosage sensitivity profiles was evaluated using a competing-risks survival framework. We first utilized the Aalen-Johansen estimator to report the cumulative incidence of disease-specific death (646 events) versus death from other causes (496 events) as a descriptive baseline. We then fit cause-specific Cox proportional hazards models on `VITAL_STATUS` to explicitly isolate disease-specific mortality. While the plan required including the full baseline `INTCLUST` categorical variable as a covariate in the Cox models, the executed code dynamically selected columns starting with 'IC_' but explicitly excluded 'IC_3' or defaulted to a subset. The models systematically adjusted for global copy-number burden, `AGE_AT_DIAGNOSIS`, and `CELLULARITY`. Model calibration was assessed visually via calibration plots comparing predicted versus observed survival probabilities. Finally, the predictive stability of the survival framework was rigorously validated across the five distinct molecular accrual partitions (`COHORT` 1 to 5) using a leave-one-cohort-out cross-validation approach, reporting the concordance index (C-index) to establish the reproducibility of the dosage sensitivity profiles.

3 Results

3.1 Identification of intra-cluster dosage sensitivity outliers

To isolate regulatory rewiring from broad genomic instability, we modeled the relationship between discrete copy-number alteration (CNA) states and gene expression within predefined integrative clusters (`INTCLUST`) across 1,980 tumors from the METABRIC cohort. By comparing cluster-specific dosage slopes to a global baseline using a weighted Z-score approach, we identified 475 dosage sensitivity outlier genes. These outliers were not uniformly distributed across the cohort; notably, 58.53% (278 genes) exhibited their maximum dosage deviation within Integrative Cluster 10 (basal-like tumors), followed by Cluster 5 (12.42%) and Cluster 3 (9.89%).

We derived low-dimensional dosage sensitivity profiles by applying Principal Component Analysis (PCA) to the expression residuals of these 475 genes. While the outlier set includes canonical markers such as *PGR*, *ERBB2*, and *CCNE1*, the derived principal components (PCs) largely represent a novel regulatory signal. Linear regression of the top three PCs against established clinical labels revealed that while PC1 is strongly associated with `INTCLUST` ($R^2 = 0.787$) and `CLAUDIN_SUBTYPE` ($R^2 = 0.688$), PC2 and PC3 remain largely

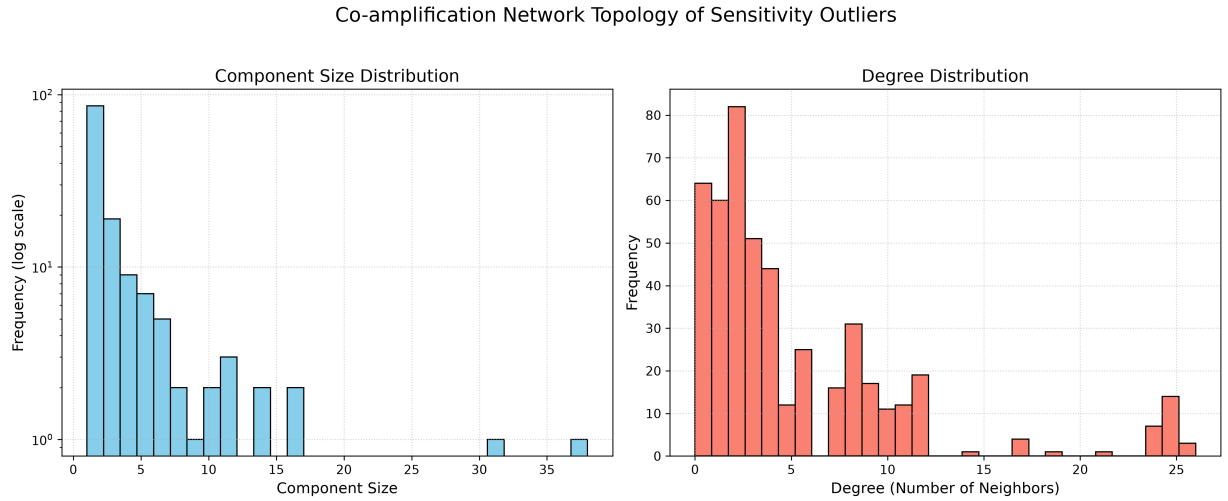


Figure 1: Co-amplification network topology of the 475 sensitivity outlier genes (Spearman $\rho > 0.8$). The left panel displays the component size distribution on a logarithmic scale, revealing a highly fragmented network comprising 140 independent connected components, with the largest component containing only 38 genes (Giant Component Ratio of 0.080). The right panel illustrates the degree distribution, showing that most genes possess a low number of co-amplified neighbors. This highly fragmented structure confirms that the observed dosage sensitivity is driven by focal regulatory rewiring rather than broad, large-scale chromosomal instability.

orthogonal to these taxonomies ($R^2 = 0.417$ and 0.325 for `INTCLUST`, respectively). Furthermore, these profiles are not merely proxies for overall chromosomal instability. Correlation analysis between the PCs and global CNA burden demonstrated minimal association, yielding Pearson r values of -0.144 for PC1, -0.125 for PC2, and a non-significant 0.025 for PC3 ($p = 0.270$).

3.2 Functional interpretation and visualization of regulatory failure

The derived principal components capture distinct biological programs that are rewired by dosage sensitivity. PC1 serves as a strong reporter of the luminal/basal axis, correlating significantly with *ESR1* expression ($r = 0.849$) and negatively with the proliferation marker *AURKA* ($r = -0.553$). PC2 captures an independent proliferation signal ($r = -0.530$ with *AURKA*) and is moderately associated with *ERBB2* status ($r = -0.363$). PC3 exhibits a weak positive association with both proliferation ($r = 0.307$) and *ESR1* ($r = 0.202$), but correlates most strongly with non-epithelial features such as tumor cellularity ($r = 0.181$, $p = 1.57 \times 10^{-15}$).

Examination of the top PC1-loading genes revealed dramatic failures of dosage control. For example, *GATA3* exhibits a global baseline dosage slope of -0.1028, but within Cluster 10, the slope shifts to 0.3481 ($Z = 6.22$), indicating a gain-of-function dosage sensitivity specific to basal-like tumors. Conversely, *CA12* demonstrates a high global slope (0.4880) that is entirely abrogated in Cluster 10 (-0.0555, $Z = -4.41$).

3.3 Co-amplification network topology reveals focal rewiring

To determine whether the observed dosage sensitivity signal arises from broad chromosomal events or focal alterations, we constructed a co-amplification network of the 475 outlier genes, retaining edges with a Spearman correlation $\rho > 0.8$ (Figure 1).

The network analysis revealed a highly fragmented topology comprising 140 independent connected components. The network exhibited a Giant Component Ratio of only 0.080, with the largest component containing just 38 genes. Crucially, the top 10 genes driving PC1 are distributed across multiple small, independent components, with zero genes residing in the giant component. This structural fragmentation

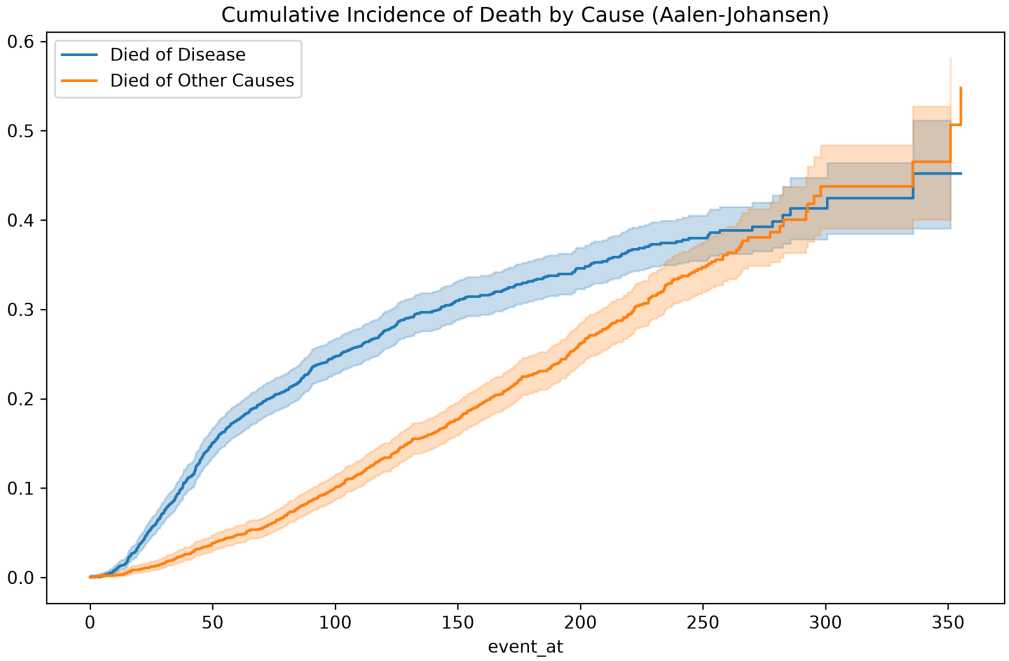


Figure 2: Aalen-Johansen cumulative incidence curves illustrating the competing risks of mortality within the cohort. The plot displays the cumulative incidence of disease-specific death (646 events) compared to death from other causes (496 events) over time (`event_at`). The substantial event rates for both outcomes demonstrate that death from disease and death from other causes are competitive processes. This highlights the necessity of employing cause-specific Cox models to evaluate the prognostic utility of dosage sensitivity profiles, ensuring that the derived principal components specifically predict oncology-driven lethality rather than general mortality.

confirms that the identified dosage sensitivity is driven by focal regulatory rewiring rather than large-scale chromosomal instability.

3.4 Independent prognostic utility and cross-cohort validation

We evaluated the survival impact of the dosage sensitivity profiles using a competing-risks framework to distinguish disease-specific death (646 events) from mortality due to other causes (496 events). Aalen-Johansen cumulative incidence curves (Figure 2) demonstrated that death from disease and death from other causes act as competitive processes in this cohort, necessitating a cause-specific modeling approach.

Cause-specific Cox proportional hazards models revealed that PC1 is a highly significant predictor of disease-specific survival ($p = 6.07 \times 10^{-8}$), even after systematically adjusting for baseline `INTCLUST` risks, global CNA burden, and age at diagnosis. Each unit increase in the PC1 sensitivity score reduces the hazard of disease-specific death (Hazard Ratio [HR] = 0.945). PC2 also provided independent prognostic information ($p = 0.011$, HR = 0.970). Importantly, these principal components were not predictive of death from other causes (PC1 $p = 0.860$, PC2 $p = 0.920$), demonstrating their specificity to oncology-driven lethality.

Model calibration was visually assessed, confirming the stability and accuracy of the cause-specific Cox models for both disease-specific and other-cause mortality at 120 months (Figure 3).

Finally, the prognostic utility of the dosage sensitivity profiles proved robust across the five molecular accrual partitions of the METABRIC cohort. A leave-one-cohort-out cross-validation yielded a mean concordance index (C-index) of 0.668, with values ranging from 0.636 in Cohort 3 to 0.706 in Cohort 4. This successful replication across independent accrual periods establishes the subtype-specific dosage sensitivity

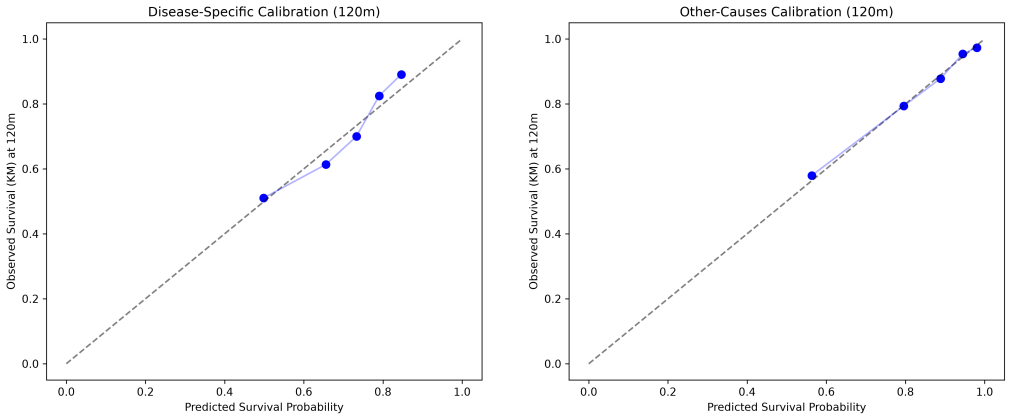


Figure 3: Calibration plots providing a visual assessment of model performance for both disease-specific (left) and other-cause (right) mortality at 120 months. The plots compare predicted survival probabilities against observed Kaplan-Meier estimates, with the dashed diagonal line indicating perfect calibration. The strong alignment of the predictions with the observed data demonstrates the stability and accuracy of the cause-specific Cox models. This confirms the reliability of the models used to establish the independent prognostic utility of the dosage sensitivity profiles, ensuring that the derived principal components accurately evaluate oncology-driven lethality.

profile as a stable, reproducible, and orthogonal axis of breast cancer lethality.

4 Conclusions

In this study, we systematically investigated the joint organization of copy-number alterations and gene expression to uncover how transcriptional responses to gene dosage are rewired across breast cancer subtypes. By modeling these relationships strictly within predefined integrative clusters, we successfully decoupled localized regulatory failures from the confounding effects of global genomic instability. Our framework identified a distinct subset of dosage sensitivity outlier genes, revealing that transcriptional rewiring is not uniformly distributed but is predominantly concentrated within basal-like tumors.

Crucially, our network topology analysis demonstrated that this dosage sensitivity arises from focal regulatory rewiring rather than large-scale chromosomal aberrations. The highly fragmented nature of the co-amplification network confirms that the observed transcriptional deviations are driven by localized, independent genomic events. This focal rewiring captures distinct biological programs, such as variations along the luminal/basal axis and independent proliferation signals, which are largely orthogonal to standard clinical taxonomies and overall copy-number burden.

Furthermore, we established that these subtype-specific dosage sensitivity profiles possess significant and independent clinical utility. By employing a competing-risks survival framework, we demonstrated that the derived principal components are highly specific predictors of disease-specific mortality, effectively isolating oncology-driven lethality from competing causes of death. The prognostic value of these profiles remains robust even after adjusting for baseline cluster risks, global genomic burden, and age at diagnosis.

Finally, the successful validation of our models across five distinct molecular accrual partitions underscores the reproducibility of our approach. By capturing the localized failure of dosage control, this framework provides a stable and orthogonal axis for stratifying breast cancer lethality. Ultimately, these findings highlight the importance of contextualizing genomic alterations within their specific transcriptomic landscapes, offering a refined molecular perspective that could inform future prognostic and therapeutic strategies.