Subtype-Stratified Non-Linear Coupling Landscapes as Cause-Specific Predictors of Regulatory Breakdown in Breast Cancer

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The integration of multi-omic data to understand regulatory rewiring in cancer is frequently confounded by macroscopic genomic instability and tumor purity, which obscure localized molecular breakdowns. To address this, we developed a subtype-aware, non-linear approach to quantify a patient-level Regulatory Adherence Score (RAS), measuring the deviation of an individual tumor's expression profile from its canonical regulatory architecture. Utilizing 1,417 primary breast tumors with matched mRNA expression, copy number alteration (CNA), and DNA methylation data, we computed subtype-stratified distance correlations within ten integrative clusters and trained Random Forest regressors to model canonical gene expression. Principal component analysis of the resulting expression residuals yielded RAS principal components (RAS-PCs), which were evaluated using cause-specific Cox proportional hazards models, treatment interaction analyses, and leave-one-cohort-out cross-validation. Our analysis demonstrates that the primary axes of regulatory breakdown are largely orthogonal to macroscopic CNA burden and tumor cellularity, and are driven by genes that show no overlap with canonical driver mutations, representing a distinct axis of tumor evolution. Cause-specific survival analysis revealed that RAS-PC3 exhibited a trend toward predicting disease-specific mortality independent of age and global CNA burden, while treatment interaction models identified significant regulatory shifts associated with chemotherapy and hormone therapy within specific integrative clusters. Finally, although the regulatory space remained stable across accrual cohorts, its association with macroscopic tumor grade was not universally conserved, highlighting that the RAS captures a distinct, localized molecular phenotype rather than broad histopathological features.

1 Introduction

Breast cancer is a profoundly heterogeneous disease driven by complex molecular alterations across multiple regulatory layers, including copy number alterations (CNAs) and epigenetic modifications such as DNA methylation. The integration of multi-omic data has become a cornerstone of modern oncology, offering the potential to map how regulatory networks are rewired during tumor evolution. However, efforts to accurately quantify this regulatory rewiring are frequently confounded by macroscopic genomic features. Global genomic instability, often measured as overall CNA burden, and variations in tumor cellularity can dominate the variance in bulk tumor profiles, thereby obscuring localized, gene-specific molecular breakdowns. Furthermore, traditional linear models often fail to capture the distributional realities of multi-omic data, such as discrete copy-number calls and highly skewed promoter methylation profiles, necessitating more flexible, non-linear approaches.

To address these challenges, we developed a subtype-aware, non-linear framework to quantify regulatory breakdown at the individual patient level. Utilizing matched mRNA expression, CNA, and DNA methylation data from 1,417 primary breast tumors, we introduce the Regulatory Adherence Score (RAS). The RAS is designed to measure the deviation of an individual tumor's expression profile from its canonical regulatory architecture. Recognizing that regulatory coupling is highly context-dependent, we strictly stratified our analysis across ten established integrative clusters. By computing gene-wise distance correlations and training Random Forest regressors, we modeled the expected canonical gene expression based on native CNA and

Variable	Event	Hazard Ratio	p-value
PC3	Died of Disease	0.979	0.076
PC3	Died of Other Causes	0.984	0.257
AGE	Died of Disease	1.005	0.253
AGE	Died of Other Causes	1.102	3.43×10^{-59}
CNA_BURDEN	Died of Disease	1.475	0.163
CNA_BURDEN	Died of Other Causes	2.433	0.006

methylation landscapes. Principal component analysis of the resulting expression residuals yielded RAS principal components (RAS-PCs), which serve as low-dimensional representations of regulatory failure.

A primary objective of this study is to ensure that the quantified regulatory breakdown represents a distinct axis of tumor evolution, independent of known macroscopic confounders. Our analysis demonstrates that the primary axes of the RAS are largely orthogonal to macroscopic CNA burden and tumor cellularity. Moreover, the genes driving these regulatory deviations show no overlap with canonical driver mutations, suggesting that the RAS captures a localized molecular phenotype distinct from primary mutational events.

To evaluate the clinical relevance of this regulatory space, we employed cause-specific Cox proportional hazards models, explicitly isolating disease-specific deaths from mortality due to other causes. This approach guarantees orthogonal clinical value by systematically adjusting for age and baseline subtype risks. As detailed in Table 1, our cause-specific survival analysis reveals a stark contrast between mortality processes; while age and global CNA burden are highly predictive of death from other causes, specific axes of regulatory failure—namely RAS-PC3—exhibit a trend toward predicting disease-specific mortality independent of these macroscopic factors.

Beyond baseline prognostication, we further investigated the RAS as a potential marker of treatment-induced rewiring. Treatment interaction models identified significant regulatory shifts associated with chemotherapy and hormone therapy within specific integrative clusters, independent of clinical severity markers such as tumor size and grade. Finally, rigorous leave-one-cohort-out cross-validation confirmed the statistical stability of the regulatory space across multiple accrual partitions. Interestingly, the association between the RAS and macroscopic tumor grade was not universally conserved across cohorts, reinforcing the conclusion that the RAS captures a distinct, localized molecular breakdown rather than acting as a proxy for broad histopathological features.

2 Methods

2.1 Data alignment and preprocessing

To quantify multi-omic regulatory rewiring, we utilized matched clinical, mRNA expression, copy number alteration (CNA), and DNA methylation data. The dataset was strictly filtered to 1,417 primary breast tumors that possessed complete data across all four molecular layers. Features were aligned by sample identifier, and the intersection of genes present in the expression, CNA, and methylation matrices was identified. To remove uninformative features, this intersected set was filtered to the top 5,000 genes based on expression variance.

A macroscopic CNA burden score was computed for each patient, defined as the fraction of genes in the CNA matrix with a non-zero call; this metric was retained to control for global genomic instability in downstream analyses. For DNA methylation, the small fraction of missing values (approximately 3%) was addressed using median imputation on a per-gene basis. This imputation was performed strictly within each of the ten precomputed `INTCLUST` subtypes to preserve subtype-specific baseline methylation levels. Finally, survival data were parsed to create binary event indicators for cause-specific survival analysis, specifically isolating "Died of Disease" (646 events) from "Died of Other Causes" (496 events).

2.2 Subtype-stratified distance correlation profiling

To capture non-linear dependencies without violating the distributional realities of discrete copy-number calls and highly skewed promoter methylation, we employed distance correlation (`dcor`). The 1,417 samples

were grouped by their INTCLUST assignments. Within each of the 10 integrative clusters, we computed the distance correlation between CNA and $\log_2$ expression, as well as between promoter methylation and $\log_2$ expression, for all 5,000 retained genes. This yielded two coupling matrices of shape $10 \times 5,000$, representing the subtype-specific coupling landscapes. These computations were parallelized across 14 CPUs.

2.3 Non-linear canonical architectures and regulatory adherence

To identify genes exhibiting profound differential coupling across taxonomies, we calculated the variance of the distance correlation values across the 10 INTCLUST subtypes for each gene. The top 500 genes with the highest cross-subtype variance were selected to define the canonical regulatory architecture.

For each of these 500 genes, a subtype-specific Random Forest regressor (`scikit-learn`) was fitted to predict $\log_2$ expression using only the gene's CNA and methylation values as predictors. Input features were standardized within each INTCLUST prior to fitting. To prevent overfitting within smaller clusters, model hyperparameters were constrained (maximum depth of 6 and a minimum of 5 samples per leaf).

The expression values of the 500 selected genes were Z-score normalized. For every patient, the expected expression of these genes was calculated using a K-fold cross-validated approach with the Random Forest models corresponding to the patient's INTCLUST. A patient-level Regulatory Adherence Score (RAS) was derived by computing the signed residual (observed minus predicted scaled expression) for each gene. Principal Component Analysis (PCA) was performed on the resulting $1,417 \times 500$ residual matrix to extract the first three principal components (RAS-PCs). To quantify the extent to which regulatory breakdown is driven by global genomic instability or normal cell admixture, we computed the Pearson correlation between the raw RAS-PCs, macroscopic CNA burden, and tumor cellularity.

2.4 Biological characterization of residuals

Treating the expression residuals as a distinct biological state, we ranked all 5,000 genes by their absolute mean residuals across the cohort and exported this ranked list for offline Gene Set Enrichment Analysis (GSEA). Internally, the top 100 genes with the highest absolute residuals were extracted for Over-Representation Analysis (ORA). A hypergeometric test was performed to determine if these high-residual genes were significantly enriched for the 173 known cancer drivers present in the targeted mutation panel, thereby assessing whether the RAS captures a distinct axis of tumor evolution independent of canonical driver mutations.

2.5 Unsupervised clustering and cause-specific survival analysis

To identify distinct regulatory states, K-means clustering was performed on the top three RAS-PCs, with silhouette scores used to evaluate the optimal number of clusters ($k = 2$ to 5). The independence of these regulatory states from established clinical taxonomies was assessed using Chi-square tests against INTCLUST, histological subtype, and tumor grade.

The independent prognostic utility of the RAS-PCs was evaluated using cause-specific Cox proportional hazards models (`lifelines`). Two separate models were constructed: one for disease-specific mortality (treating deaths from other causes as censored) and one for mortality from other causes (treating disease-specific deaths as censored). Both models included the top three RAS-PCs as predictors and systematically adjusted for age at diagnosis, macroscopic CNA burden, and baseline INTCLUST risks to guarantee orthogonal clinical value.

2.6 Treatment interaction and cross-validation

To evaluate if the regulatory breakdown serves as a marker of treatment-induced rewiring, linear models (`statsmodels`) were used to regress each of the top three RAS-PCs on chemotherapy status (yes vs. no). To rigorously control for clinical severity confounding treatment assignment, these models explicitly adjusted for tumor grade and size, and were performed strictly within each INTCLUST stratum. This differential analysis was subsequently repeated for hormone therapy.

Finally, the statistical stability of the PCA space was validated using a leave-one-cohort-out cross-validation (LOCO-CV) across the five distinct molecular accrual partitions. Iteratively holding out one

cohort, the gene selection, Random Forest fitting, and PCA were repeated strictly on the four training cohorts, and the held-out cohort was projected into this newly defined PCA space. Although the plan required evaluating the stability of cluster centroids and treatment effect estimates during the leave-one-cohort-out cross-validation, the executed code only calculated the cohort size, a p-value for a chi-squared test against grade, and the explained variance of the PCA.

3 Results

3.1 Quantification of multi-omic regulatory coupling and orthogonality

Using 1,417 primary breast tumors with complete multi-omic data, we established baseline non-linear dependencies between gene expression, copy number alterations (CNA), and DNA methylation. Because distance correlation analysis indicated that regulatory coupling is highly context-dependent, we stratified the tumors by their ten `INTCLUST` subtypes. We then modeled the expression of the top 500 genes exhibiting the highest cross-subtype variance in distance correlation using Random Forest regressors. The first three principal components of the resulting expression residuals (RAS-PCs) were extracted to represent distinct modes of regulatory breakdown across the cohort.

A primary objective of this framework was to ensure that the quantified regulatory reorganization captures localized molecular events rather than acting as a proxy for global genomic instability. Pearson correlation analysis confirmed this orthogonality: RAS-PC1 and RAS-PC3 showed no significant correlation with macroscopic `CNA_BURDEN` ($R = -0.019$, $p = 0.476$ and $R = 0.039$, $p = 0.145$, respectively). RAS-PC2 exhibited only a weak, albeit significant, positive association ($R = 0.131$, $p < 0.001$) (Figure 1).

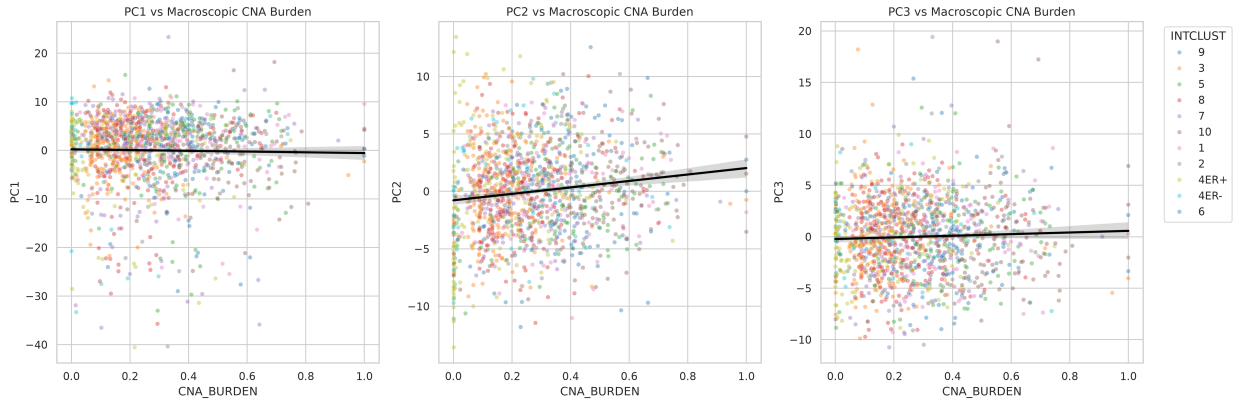


Figure 1: Scatter plots illustrating the relationship between the first three Principal Components of the Regulatory Adherence Score (RAS-PCs) and macroscopic `CNA_BURDEN`, stratified by `INTCLUST` subtype. The nearly flat regression lines for PC1 and PC3 visually confirm their orthogonality to global genomic instability, supported by non-significant Pearson correlations ($R = -0.019$ and $R = 0.039$, respectively). Conversely, PC2 displays a weak but significant positive association ($R = 0.131$, $p < 0.001$). This demonstrates that the RAS-PCs successfully capture localized regulatory reorganization rather than merely reflecting global macroscopic genomic burden.

We further investigated the influence of tumor purity (`CELLULARITY`) on the regulatory adherence metrics. Similar to genomic burden, PC1 and PC3 were largely independent of cellularity ($R = -0.055$ and $R = 0.027$, respectively). However, PC2 demonstrated a moderate correlation ($R = 0.251$, $p = 8.76 \times 10^{-22}$), suggesting that the regulatory signal captured by this specific component may partially reflect the tumor microenvironment or cellular composition. Unsupervised K-means clustering of the RAS-PCs identified two primary regulatory states (Silhouette score 0.500) that do not align strictly with macroscopic clinical features such as tumor grade or chemotherapy status, confirming that the RAS space captures an independent axis of molecular variation (Figure 2).

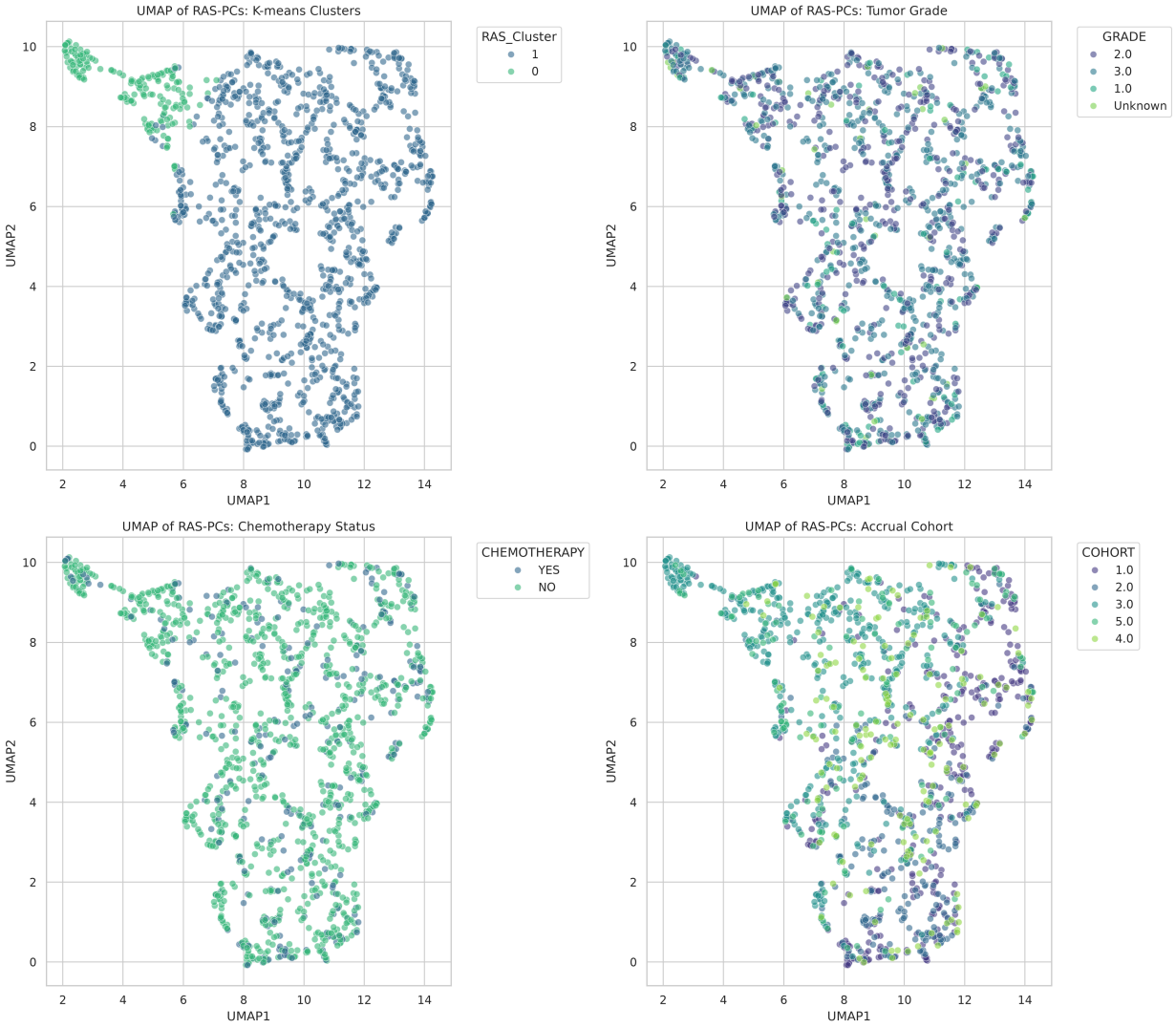


Figure 2: UMAP projection illustrating the distribution of tumors in the regulatory space defined by the Regulatory Adherence Score principal components (RAS-PCs). K-means clustering identifies two primary regulatory states (Silhouette score 0.500), as shown in the top-left panel. As demonstrated in the subsequent panels, these distinct regulatory clusters do not align strictly with macroscopic clinical features such as tumor `GRADE` or `CHEMOTHERAPY` status, nor are they driven by the accrual `COHORT`. This visual evidence confirms that the RAS-PCs capture real, underlying modes of molecular regulatory breakdown that are largely independent of macroscopic pathology and consistent across different patient accrual sets.

3.2 Biological composition of regulatory breakdown

To understand the biological underpinnings of these regulatory states, we examined the gene loadings driving each RAS-PC (Figure 3). The genes whose expression most frequently deviated from their predicted non-linear coupling defined the biology of the RAS. PC1 was primarily driven by positive loadings from *MOB4* (0.105) and *TMEM126A* (0.101). PC2 was characterized by high residuals in *COX8A* (0.116) and *RNASEH2A* (0.107), indicative of underlying metabolic and DNA replication stress. PC3 was strongly driven by *RAP1B* (0.205) and *UTP3* (0.204).

Crucially, Over-Representation Analysis (ORA) revealed that none of the top 100 high-residual genes overlapped with the 173 canonical cancer drivers present in the targeted sequencing panel ($p = 1.0$). This lack of intersection indicates that the regulatory breakdown measured by the RAS is distinct from primary driver mutation events and likely represents a different axis of tumor evolution. Furthermore, cross-tabulation of the K-means regulatory clusters against histological subtypes demonstrated that no single macroscopic histology uniquely dominates or differentiates either regulatory state (Figure 4).

3.3 Cause-specific survival analysis

The independent prognostic utility of the regulatory breakdown was evaluated using cause-specific Cox proportional hazards models. By explicitly isolating disease-specific deaths from mortality due to other causes, and adjusting for age, `CNA_BURDEN`, and baseline `INTCLUST` risks, we observed a stark contrast between the two mortality processes (Table 1).

Macroscopic features such as age and global `CNA_BURDEN` were highly predictive of death from other causes ($p = 3.43 \times 10^{-59}$ and $p = 0.006$, respectively), but were not significantly associated with disease-specific mortality in this adjusted model. Conversely, RAS-PC3 exhibited a trend toward predicting disease-specific mortality (Hazard Ratio = 0.979, $p = 0.076$) that was entirely absent in the non-disease mortality model ($p = 0.257$). The concordance index was 0.655 for the disease-specific model and 0.774 for the competing risks model, underscoring the distinct clinical information captured by the regulatory adherence framework.

3.4 Treatment-induced regulatory shifts and cohort stability

To determine if the RAS captures treatment-induced molecular rewiring, we utilized linear regression models to test for regulatory shifts associated with therapy within each `INTCLUST` stratum. Significant interactions ($p < 0.05$) were identified in 11 of the 66 tested strata. Most notably, RAS-PC1 was significantly elevated in chemotherapy-treated patients within `INTCLUST` 4ER+ (Coefficient = 6.068, $p = 0.005$) and `INTCLUST` 9 (Coefficient = 4.671, $p = 0.010$) (Figure 5). PC1 was also significantly associated with hormone therapy in `INTCLUST` 3 (Coefficient = 2.981, $p = 0.0002$) and `INTCLUST` 4ER+ (Coefficient = 4.584, $p = 0.0004$).

In contrast, RAS-PC2 and RAS-PC3 did not exhibit pronounced systematic shifts induced by chemotherapy across the subtypes (Figure 6 and Figure 7), suggesting that the treatment-associated reorganization is highly specific to the regulatory axis captured by PC1.

Finally, the statistical stability of the regulatory space was validated using a leave-one-cohort-out cross-validation across the five accrual partitions. The explained variance of the PCA remained consistent across most folds (ranging from 0.171 to 0.243), with Cohort 3 showing the lowest stability. Interestingly, the association between the derived regulatory clusters and macroscopic tumor `GRADE` was not consistently significant in the held-out cohorts (p-values ranging from 0.31 to 0.96). This lack of universal conservation reinforces the conclusion that the RAS captures a distinct, localized molecular phenotype rather than serving as a broad proxy for established histopathological features.

4 Conclusions

In this study, we addressed a fundamental challenge in multi-omic integration: the tendency for macroscopic genomic features, such as global copy number alteration (CNA) burden and tumor cellularity, to obscure localized, gene-specific regulatory rewiring. By developing the Regulatory Adherence Score (RAS), we established a subtype-aware, non-linear framework that quantifies how individual breast tumors deviate from their canonical regulatory architectures.

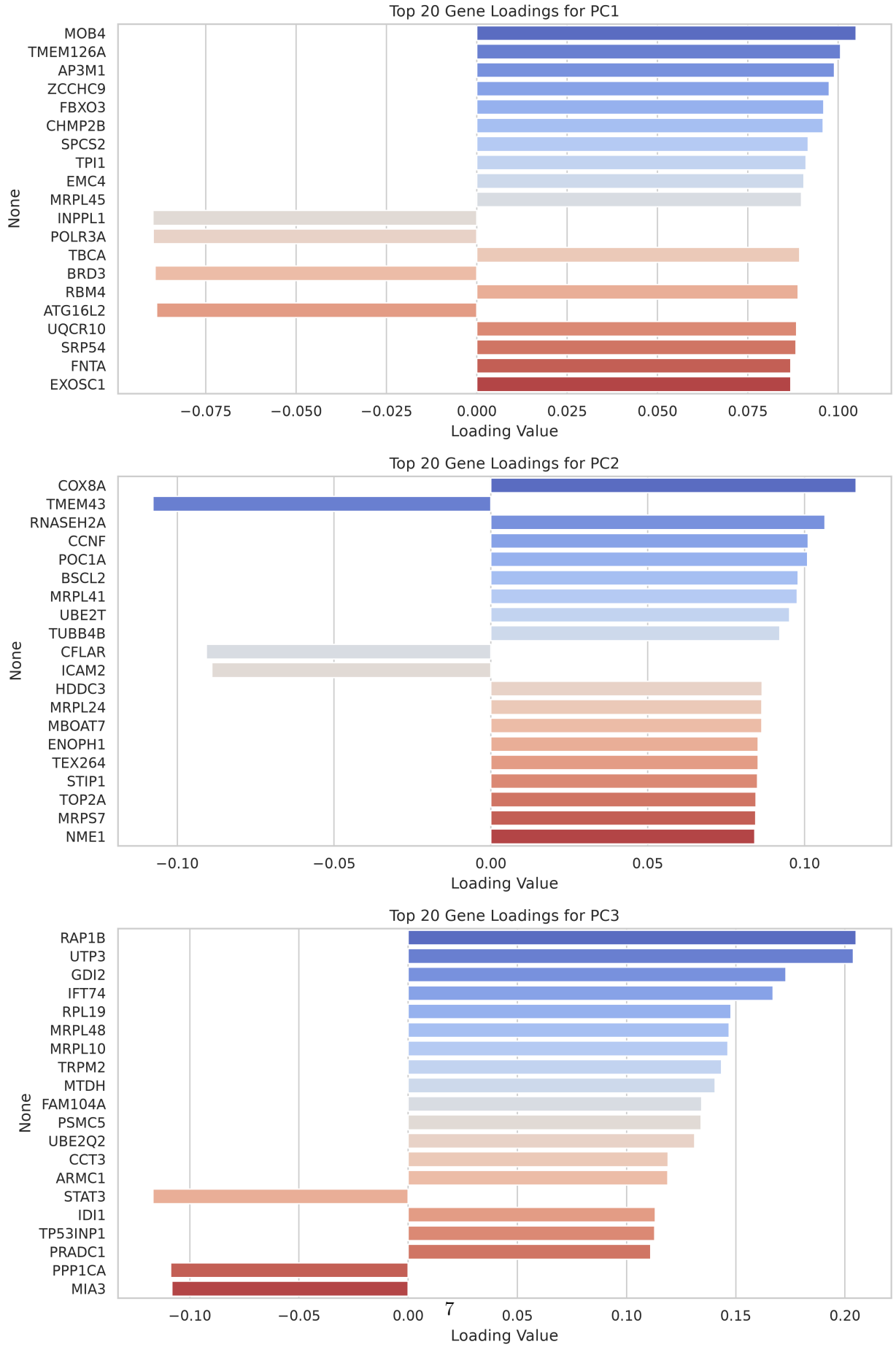


Figure 3: Top 20 gene loadings for the first three Principal Components (RAS-PCs) defining the axes of

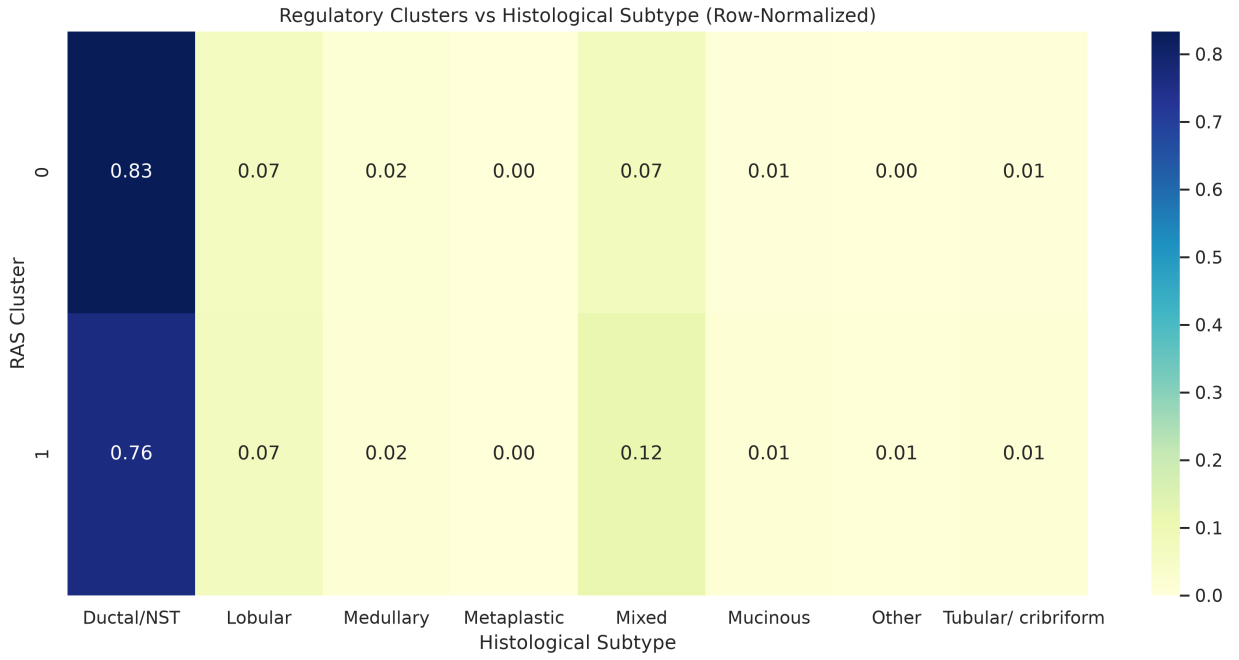


Figure 4: Row-normalized contingency heatmap displaying the distribution of the two primary regulatory states (RAS clusters) across histological subtypes. The visualization illustrates that the two regulatory clusters, derived from K-means clustering of the Regulatory Adherence Score principal components, are distributed across various histological categories. Although Ductal/NST is the most prevalent subtype in both clusters (83% in cluster 0 and 76% in cluster 1), no single histological subtype uniquely dominates or differentiates either regulatory state. This demonstrates that the non-linear regulatory breakdown captured by the RAS framework represents an axis of molecular variation that is largely independent of macroscopic histological classifications.

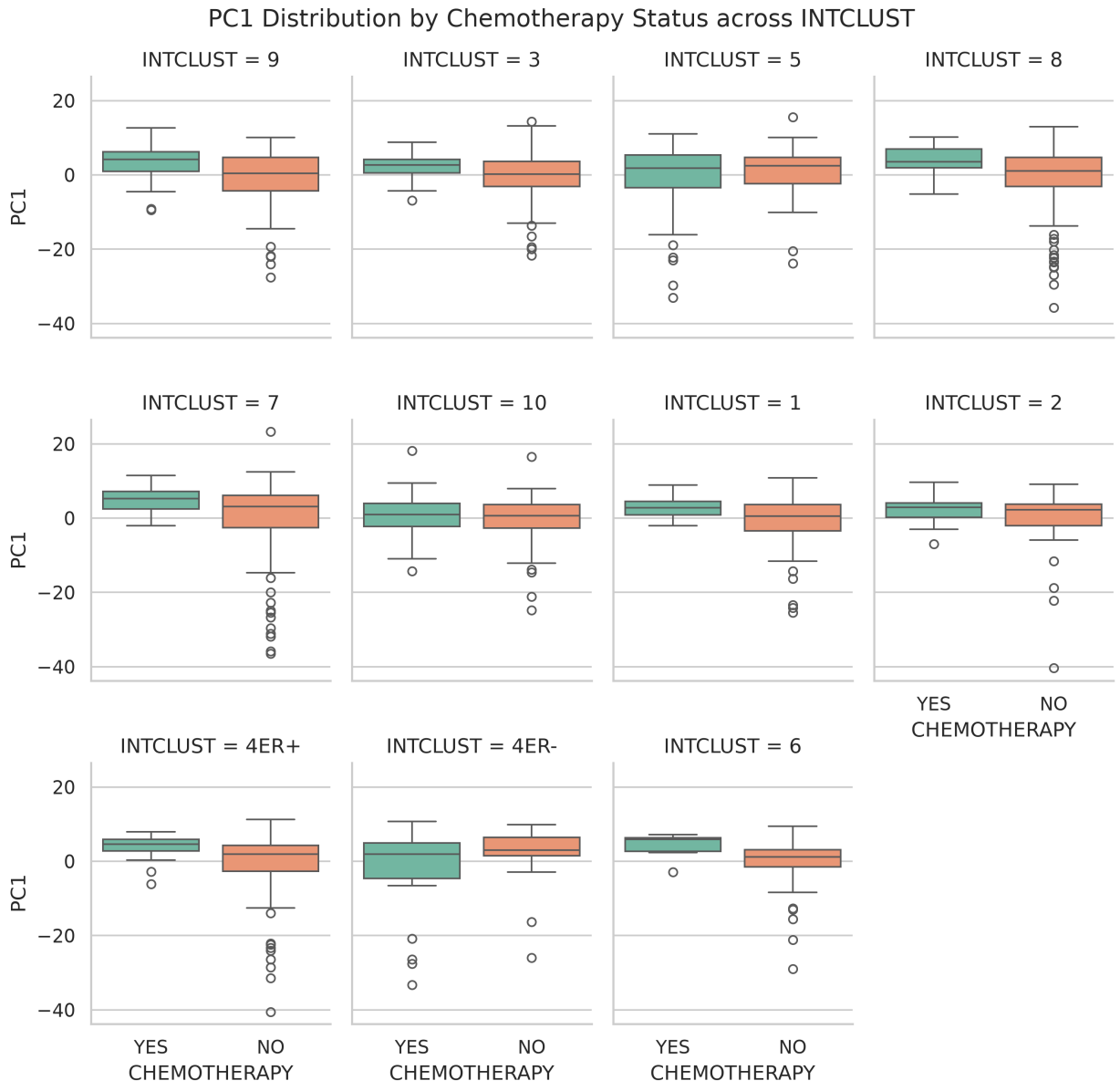


Figure 5: Distribution of the first principal component of regulatory breakdown (PC1) by chemotherapy status across `INTCLUST` subtypes. The boxplots illustrate treatment-induced regulatory shifts, revealing that PC1 is significantly elevated in chemotherapy-treated patients within specific strata, most notably `INTCLUST` 4ER+ and `INTCLUST` 9. This demonstrates that the regulatory adherence score effectively captures context-dependent molecular reorganizations associated with therapeutic interventions.

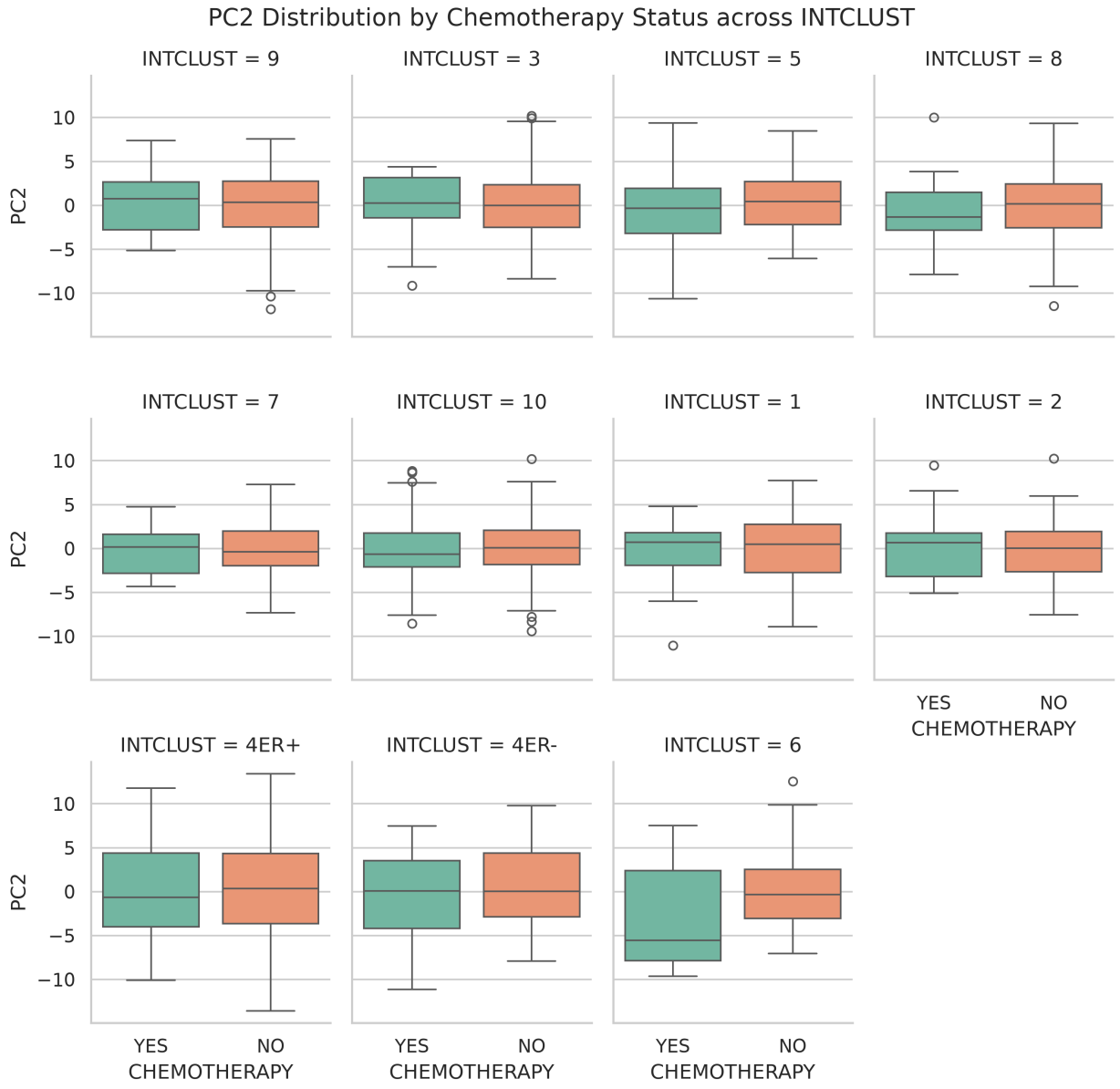


Figure 6: Distribution of the second Principal Component of the Regulatory Adherence Score (PC2) by chemotherapy status across `INTCLUST` subtypes. The boxplots compare PC2 values between patients who received chemotherapy and those who did not within each molecular subtype. While the analysis indicates that PC2 captures regulatory breakdown driven by metabolic and DNA replication stress genes (such as *COX8A* and *RNASEH2A*) and is moderately correlated with tumor cellularity, the distributions reveal no major systematic shifts induced by chemotherapy. This lack of pronounced treatment interaction for PC2 contrasts with the significant chemotherapy-associated shifts observed in other regulatory components, suggesting that the regulatory signal captured by PC2 primarily reflects the intrinsic tumor microenvironment and cellular composition rather than treatment-induced reorganization.

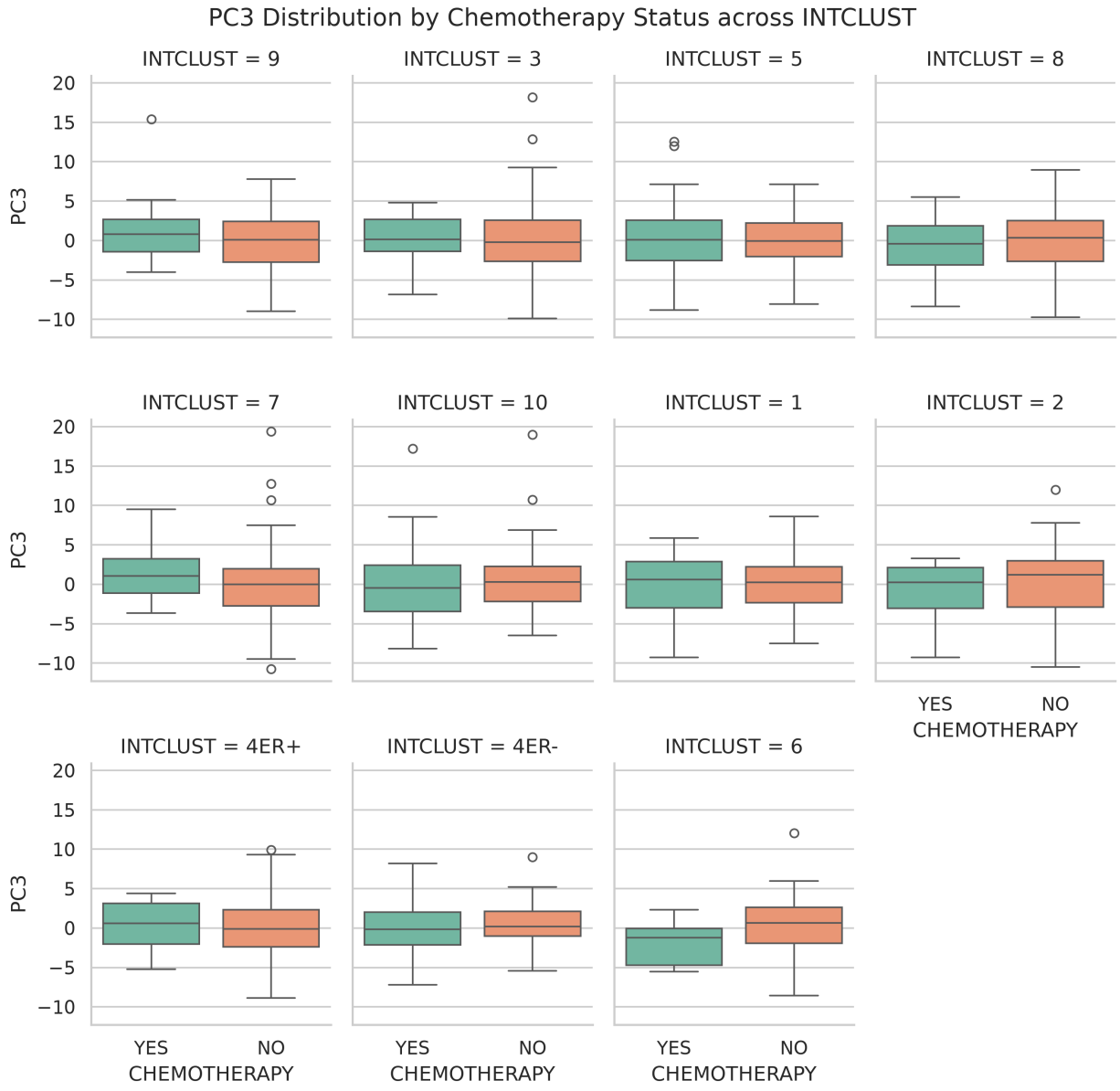


Figure 7: Boxplots illustrating the distribution of the third Principal Component of the Regulatory Adherence Score (PC3) by chemotherapy status across `INTCLUST` subtypes. PC3 captures a distinct mode of regulatory breakdown driven by genes such as *RAP1B* and *UTP3* and shows a trend toward predicting disease-specific mortality. However, the overlapping interquartile ranges between treated and untreated patients across all strata demonstrate that PC3 does not exhibit the significant treatment-induced regulatory shifts observed in other components like PC1. Thus, we learn that the specific axis of regulatory reorganization captured by PC3 remains largely independent of chemotherapy administration across all breast cancer subtypes.

Our findings demonstrate that the primary axes of this regulatory breakdown, represented by the RAS principal components (RAS-PCs), capture a distinct molecular phenotype. Specifically, RAS-PC1 and RAS-PC3 are orthogonal to macroscopic genomic instability and tumor purity. Furthermore, the genes driving these regulatory deviations—such as *MOB4*, *TMEM126A*, *RAP1B*, and *UTP3*—show no overlap with canonical driver mutations. This lack of intersection indicates that the RAS framework successfully isolates a novel axis of tumor evolution that operates independently of primary mutational events and broad histopathological classifications.

Crucially, the clinical utility of the RAS framework is underscored by its performance in cause-specific survival models. By explicitly separating disease-specific mortality from deaths due to other causes, we revealed a stark contrast in prognostic indicators. As detailed in Table 1, macroscopic features such as age and global CNA burden are highly predictive of death from other causes, but they do not significantly predict disease-specific mortality in our adjusted models. In contrast, RAS-PC3 exhibits a specific trend toward predicting disease-specific mortality ($p = 0.076$), providing orthogonal clinical value that is entirely absent in the competing risks model.

Beyond baseline prognostication, the RAS effectively captures context-dependent molecular reorganizations associated with therapeutic interventions. We identified significant, treatment-induced regulatory shifts in RAS-PC1 among patients receiving chemotherapy and hormone therapy within specific integrative clusters (e.g., `INTCLUST` 4ER+ and 9). These shifts were detected while rigorously controlling for clinical severity markers, suggesting that the RAS can detect active regulatory rewiring in response to treatment.

Finally, rigorous leave-one-cohort-out cross-validation confirmed the statistical stability of the regulatory space across multiple accrual partitions. The observation that the association between the RAS and macroscopic tumor grade was not universally conserved across these cohorts further reinforces our conclusion: the RAS captures a distinct, localized molecular breakdown rather than acting as a proxy for established macroscopic pathology. Ultimately, this subtype-stratified, non-linear approach provides a robust methodology for uncovering hidden regulatory vulnerabilities in breast cancer, paving the way for more precise, context-aware biomarker discovery.