BS831 Final Project: GABA Receptor Subunit Expression in Triple-Negative Breast Cancer
Author: Omnia Abdelrahman
Advisor: Prof. Lukas Weber
Dataset: TCGA-BRCA RNA-seq (HTSeq counts)

Abstract

Triple-negative breast cancer (TNBC) is a highly aggressive subtype lacking targeted therapies, with recent evidence implicating GABA_A receptor signaling in tumor progression. This study aimed to identify the GABA receptor subunit most likely responsible for promoting proliferation and epithelial-to-mesenchymal transition (EMT) in TNBC, based on prior wet-lab findings of reduced EMT and invasion following GABA receptor knockdown. RNA-seq data from the TCGA-BRCA cohort were analyzed using differential expression, pathway enrichment, targeted gene analysis, and subgroup stratification. Among 19 GABA subunits examined, GABRA3 emerged as significantly upregulated in TNBC and co-expressed with proliferation and EMT markers. While overall EMT-related associations were weak, MKI67 consistently showed strong upregulation, co-clustering with GABA subunits, and a near-significant trend in GABRA3-high samples ($p = 0.08$), suggesting a possible role for GABAergic signaling in promoting cell proliferation. Correlation and EMT signature analyses revealed subtle or context-specific effects, pointing toward non-transcriptional mechanisms. These findings support further investigation of GABRA3 as a candidate modulator of tumor growth in TNBC and demonstrate the value of integrating public RNA-seq data to refine wet-lab hypotheses.

Project Aim: To identify the GABA **receptor** subunit most likely driving EMT-like behaviors, previously observed in wet-lab TNBC models following GABA receptor knockdown, and to determine which downstream **gene** is most strongly associated with these effects.

1. Introduction

Triple-negative breast cancer (TNBC) is an aggressive subtype of breast cancer characterized by the absence of estrogen receptor (ER), progesterone receptor (PR), and HER2 expression. This lack of receptor expression renders TNBC unresponsive to hormonal or HER2-targeted therapies, leading to limited treatment options and a higher propensity for metastasis and recurrence.[1] Recent studies have highlighted the role of ion channels, particularly ligand-gated ion channels like the γ-aminobutyric acid type A (GABA_A) receptors, in cancer progression.[2] Subunits such as GABRA3, GABRB3, and GABRP have been found to be overexpressed in TNBC and are implicated in promoting tumor cell proliferation, migration, and invasion.[3,4] However, the specific GABA receptor subunit responsible for these effects remains unknown.

This project was designed to build on key wet-lab findings in TNBC cell lines where GABA receptor knockdown resulted in: (1) reduced proliferation and cell count, (2) decreased EMT markers, and (3) reduced invasion and migration. Due to limited funding, western blot confirmation of the specific receptor subunit involved was not feasible, especially given that GABA_A receptors include 19 different subunits. RNA-seq was also financially burdensome at the time. Thus, we pursued a computational approach using TCGA RNA-seq data to identify the most likely receptor responsible for these phenotypes. We aimed to pinpoint the GABA receptor subunit associated with EMT changes and explore downstream pathways using differential expression, gene set enrichment, subgroup stratification, and correlation analyses.

2. Methods

2.1. Data Acquisition and Preprocessing (see CODE1)

Raw RNA-seq count data for breast cancer were obtained from The Cancer Genome Atlas (TCGA) using the TCGAbiolinks package in R. The TCGA-BRCA project was queried, and transcriptome profiling data (STAR-Counts workflow) were downloaded. Expression matrices and clinical metadata were extracted using GDCprepare(). Ensembl gene identifiers were converted to HGNC gene symbols via the biomaRt package, and duplicated or unmapped entries were removed.

Triple-negative breast cancer (TNBC) samples were identified based on the 20th percentile expression thresholds for ESR1, PGR, and ERBB2 (ESR1 < 3440, PGR < 215, ERBB2 < 11128). Solid tissue normal samples were selected based on clinical annotations. Genes with low expression (below the 25th percentile across all samples) were filtered out to reduce background noise. Sample-level quality control was performed using PCA and hierarchical clustering to confirm the absence of outliers and verify transcriptional distinctness between TNBC and normal tissues. A variance-stabilizing transformation (VST) was applied using DESeq2 to normalize the expression matrix for downstream analysis.

2.2 Exploratory Data Analysis (see CODE2)

Exploratory analyses were performed to assess the structure and variability of the expression data prior to hypothesis testing. The top 500 most variable genes were identified using median absolute deviation (MAD) and were used for clustering and dimensionality reduction. (Fig. 1)

A hierarchical clustering heatmap was generated based on the expression of canonical epithelial-to-mesenchymal transition (EMT) markers, including ZEB1, SNAI1, TWIST1, CDH2, VIM, FN1, MMP9, and MKI67. Clustering revealed heterogeneous EMT gene expression profiles across TNBC samples, suggesting variable EMT activation within the cohort. (Fig.2.)

Principal component analysis (PCA) was performed using the same 500-gene set. The first two principal components (PC1 and PC2) showed clear separation between TNBC and normal tissues. These components were then used as predictors in a logistic regression model to classify samples by condition. Both PCs were statistically significant (PC1: $p = 0.0004$; PC2: $p = 0.0292$), and the model achieved over 95% accuracy, with only 4 misclassified samples, as shown in the confusion matrix. These findings confirmed the transcriptional distinctness of TNBC samples and validated the biological relevance of the data structure.(Fig.3.)

2.3 Differential Expression Analysis (see CODE3)

Differential expression analysis was conducted using the DESeq2 package. TNBC tumor samples were compared to normal breast tissue to identify genes with statistically significant expression differences. Raw counts were used as input, and default DESeq2 settings were applied, including shrinkage estimation for dispersion and fold change. Genes were considered differentially expressed if they met a false discovery rate (FDR) < 0.05 and an absolute $\log_2$ fold change ≥ 1. Results were visualized using an MA plot, with significantly upregulated and downregulated genes highlighted in blue. The majority of differentially expressed genes showed increased expression in TNBC samples, including EMT-associated markers such as FN1, MKI67, and MMP9.

2.4 Gene Set Enrichment Analysis (see CODE4)

To investigate biological processes enriched in differentially expressed genes, gene set enrichment analysis (GSEA) was performed using the fgsea package. Genes were ranked based on their signed $\log_{10}$-adjusted p-values multiplied by the direction of fold change, and the full ranked list was used as input. Gene sets were sourced from the MSigDB Hallmark and GO Biological Process collections. The epithelial-to-mesenchymal transition (EMT) pathway was significantly enriched in TNBC samples compared to normal tissues (adjusted p = 0.0256). Enrichment plots confirmed positive enrichment scores for EMT-related gene sets. (Fig. 5.) In parallel, a hypergeometric overrepresentation analysis was performed to test specific custom gene sets related to EMT, Wnt signaling, FAK signaling, and SRC signaling. EMT was again found to be significantly enriched among upregulated genes. (fig. 6)

2.5 Targeted Gene Analysis (see CODE5)

To investigate the role of specific GABA receptor subunits and their potential downstream targets, a focused gene panel was curated. This panel included 19 GABA receptor subunits (e.g., GABRA1–6, GABRB1–3, GABRG1–3, GABRP, GABRQ) and genes implicated in epithelial-to-mesenchymal transition (EMT), migration, and proliferation, such as CTNNB1, EGFR, SRC, PTK2, ZEB1, FN1, MMP9, and MKI67.

2.5.1 Expression Heatmap of GABA and Pathway Genes

Expression values of the selected genes were extracted from the variance-stabilized TNBC expression matrix. Gene expression was scaled and clustered using the pheatmap package to generate a hierarchical heatmap, allowing visual assessment of co-expression and heterogeneity across TNBC samples. GABA receptor subunits showed variable expression, with GABRA3, GABRQ, and GABRA5 appearing in distinct clusters with pathway genes such as SRC and CTNNB1.

2.5.2 Differential Expression of Target Genes

Differential expression results from the DESeq2 output were filtered to isolate $\log_2$ fold changes and adjusted p-values for the curated gene panel. A bar plot of $\log_2$ fold changes, grouped by statistical significance (adjusted $p < 0.05$), was created using ggplot2. Genes such as GABRA3, GABRA5, and GABRQ were significantly upregulated in TNBC samples, suggesting potential roles in cancer progression. Additional pathway-related genes like SRC and MKI67 also showed notable upregulation. A volcano plot was generated to visualize genome-wide differential expression results. GABA receptor genes and EMT regulators were highlighted within the context of global transcriptional changes.

2.5.3 Correlation Analysis Between GABA Receptors and Pathway Genes

To evaluate possible functional associations, Pearson correlation coefficients were calculated between each GABA receptor and the panel of EMT/proliferation genes using the variance-stabilized expression matrix. The resulting correlation matrix was visualized as a heatmap. GABA subunits such as GABRA2 showed modest positive correlations with ZEB1, CDH2, and FN1, while GABRR1 correlated positively with MKI67, suggesting links to proliferation. These patterns informed hypotheses regarding downstream regulatory effects following GABA receptor modulation.

2.6 Subgroup Analysis (see CODE6)

To explore receptor-specific expression effects, TNBC samples were stratified based on GABRA3 expression levels. A median split was applied to divide the cohort into GABRA3-high and GABRA3-low subgroups. Expression of canonical EMT markers (FN1, MKI67, MMP9, etc.) was compared between the two subgroups using Wilcoxon rank-sum tests. While trends toward higher EMT gene expression were observed in the GABRA3-high group, most comparisons did not reach statistical significance (e.g., MKI67: p = 0.08), suggesting context-dependent or subtle regulation. Boxplots were used to visualize EMT gene expression across GABRA3-defined subgroups.

2.7 Correlation and Signature Analysis (see CODE7)

To further investigate the relationship between GABRA3 expression and epithelial-to-mesenchymal transition (EMT), a composite EMT signature score was generated by averaging the variance-stabilized expression of canonical EMT markers. Spearman correlation tests were performed between GABRA3 expression and the expression of individual EMT genes, as well as between GABRA3 and the composite EMT score.

Correlation coefficients were generally low and not statistically significant, indicating weak or indirect associations. These results suggest that GABRA3 may influence EMT through non-linear or post-transcriptional mechanisms, or that its effects are context-specific and not strongly reflected at the mRNA level in bulk RNA-seq data.

2.8 Classification and Clustering Methods (see CODE8)

1. To complement differential gene expression and pathway-level analyses, both unsupervised clustering and supervisedclassification techniques were applied to explore structure in TNBC transcriptomic profiles and evaluate the predictive capacity of candidate gene sets. The top 500 most variable genes across TNBC samples were selected using the median absolute deviation (MAD) method. Unsupervised hierarchical clustering was performed using the Ward.D2 method and visualized via a heatmap generated with the ComplexHeatmap package, providing an unbiased view of sample similarity without any label information.

2.To assess whether known biological groupings aligned with data-driven structure, a second heatmap was generated using the same hierarchical clustering, but with annotations for GABRA3 expression status (High vs. Low) and unsupervised model-based clustering assignments from the mclust package, which was applied to the top 12 principal components. This repeat visualization enabled direct comparison between unsupervised cluster structure and biologically informed stratification, highlighting whether GABRA3 status mapped onto emergent expression-based subgroups.

3. For the supervised classification task, GABRA3 expression was used as a binary outcome label. A t-test was performed across all genes to identify the top 20 features most associated with the GABRA3-High versus GABRA3-Low groups. A Naive Bayes classifier was then trained using the caret package, with a 70/30 stratified train-test split and 10-fold cross-validation for internal performance estimation. Model evaluation was conducted using a confusion matrix and receiver operating characteristic (ROC) analysis.

3. Results

3.1 Exploratory Data Analysis

To explore global expression variability, the top 500 most variable genes across TNBC samples were identified using median absolute deviation (MAD). A hierarchical heatmap of these genes revealed substantial inter-patient heterogeneity, suggesting distinct transcriptional subtypes within TNBC (Figure 1).

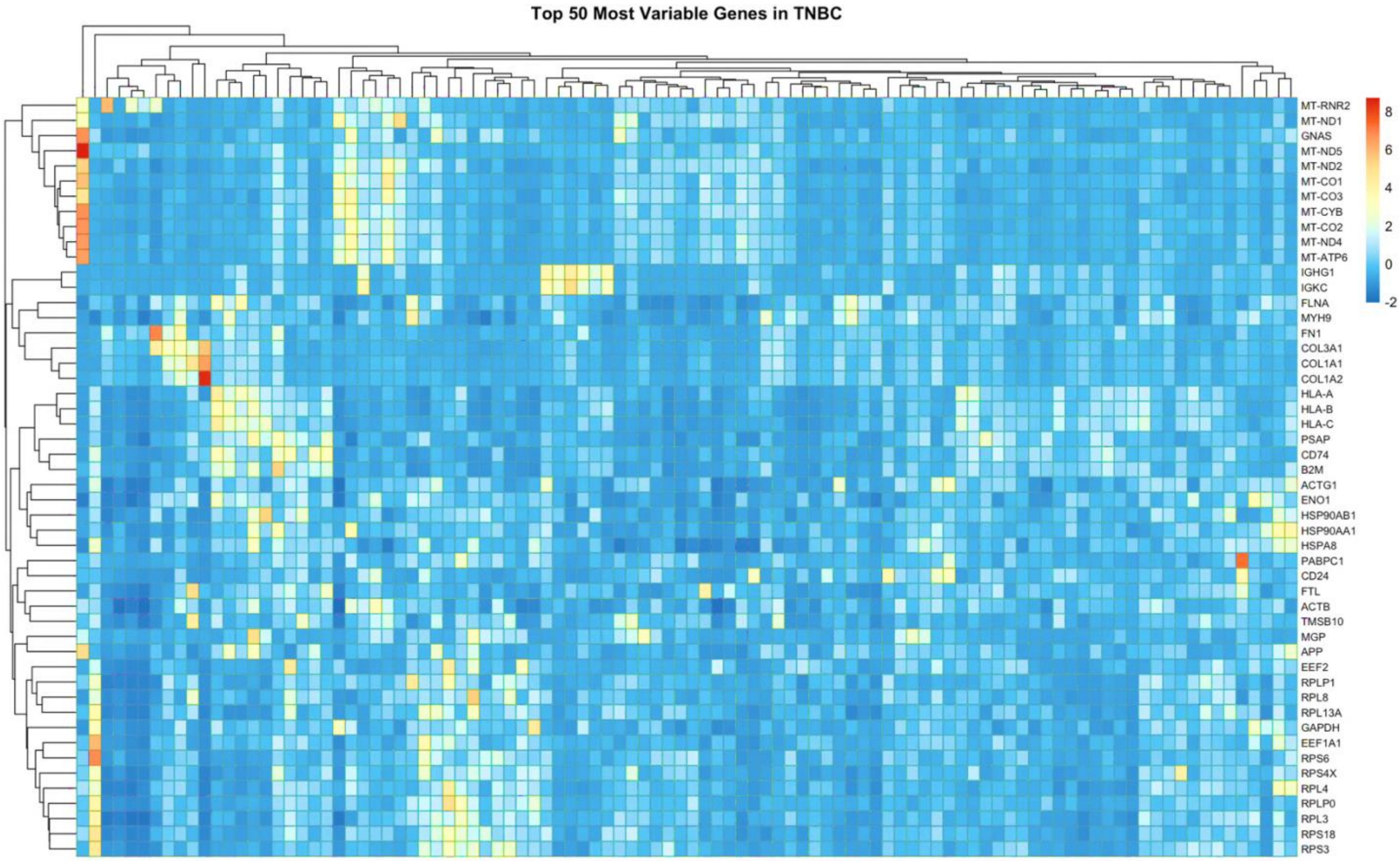


Figure 1. Heatmap of the top 500 most variable genes across TNBC samples based on MAD scores. Rows represent genes; columns represent patients.

Clustering reveals molecular heterogeneity. Unsupervised clustering of canonical epithelial-to-mesenchymal transition (EMT) markers, including ZEB1, SNAI1, TWIST1, CDH2, VIM, FN1, MMP9, and MKI67, revealed considerable heterogeneity among TNBC tumors. This variation indicates variable EMT activation across the cohort (Figure 2).

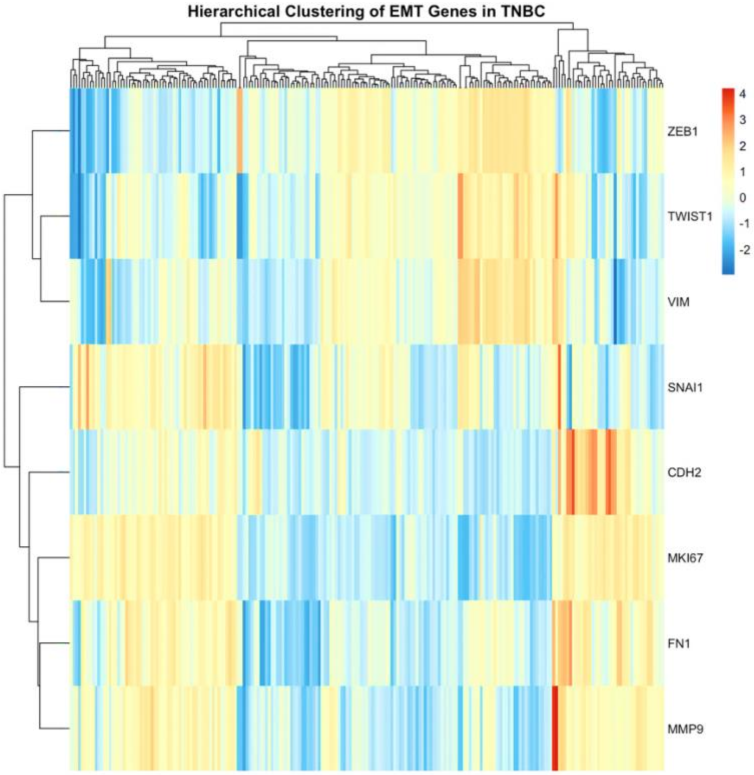


Figure 2. Hierarchical clustering heatmap of EMT marker expression across TNBC samples. The clustering reveals variable EMT activity levels.

Principal Component Analysis (PCA) based on the top 500 most variable genes demonstrated clear separation between TNBC and normal breast tissue samples (Figure 2). Logistic regression using PC1 and PC2 achieved over 95% classification accuracy (p-values: PC1 = 0.0004, PC2 = 0.0292), confirming distinct transcriptional profiles. (figure 3)

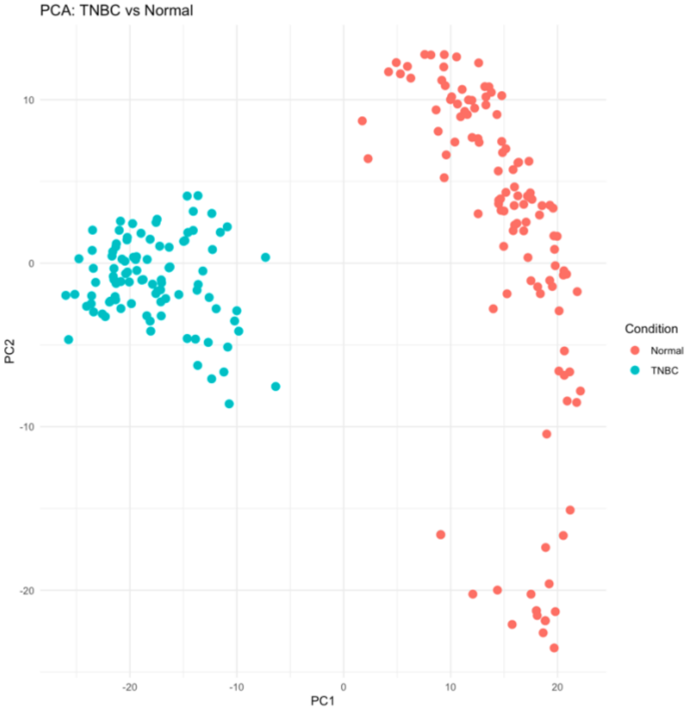


Figure 3. PCA plot showing separation between TNBC and normal samples using the top 500 most variable genes.

3.2 Differential Expression Analysis

Differential expression analysis was performed using the DESeq2 package to compare TNBC samples to normal breast tissues. Genes with a false discovery rate (FDR) < 0.05 and absolute $\log_2$ fold change ≥ 1 were considered significantly differentially expressed. The MA plot (Figure 4) displays the relationship between mean expression and fold change, with significantly differentially expressed genes highlighted. Notably, a large number of genes were found to be upregulated in TNBC, including EMT-related genes such as FN1, MMP9, and MKI67. In total, 7740 genes were significantly upregulated and 3400 were significantly downregulated.

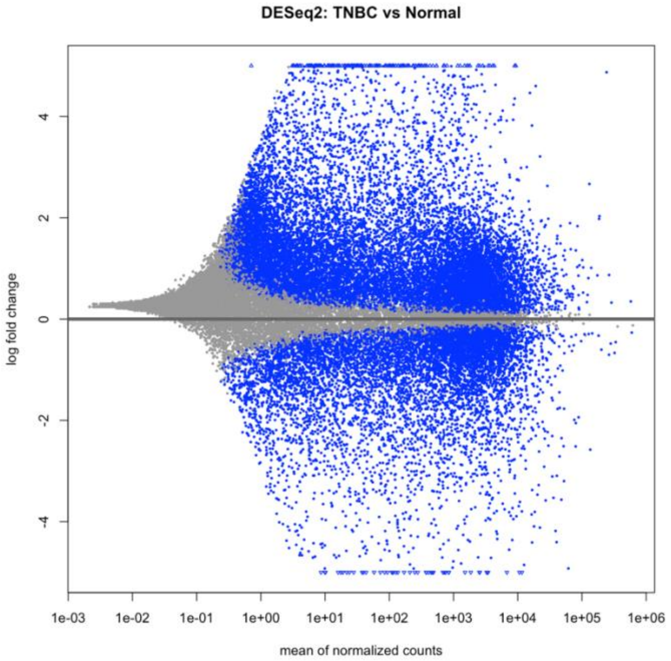


Figure 4. X-axis = Mean of normalized counts (gene expression level) Y-axis = Log2 fold change (TNBC vs. normal) Blue points = Significantly differentially expressed genes (adjusted p < 0.05) Gray points = Not significant Wider spread at higher expression levels → more confident DEGs at high expression

3.3 Pathway Enrichment Analysis

To identify the biological pathways associated with differentially expressed genes in TNBC, both Gene Set Enrichment Analysis (GSEA) and hypergeometric testing were performed. GSEA revealed moderate enrichment of the epithelial-mesenchymal transition (EMT) pathway, with a normalized enrichment score (NES) of 1.53 and nominal p-value of 0.021. However, the adjusted p-value (FDR = 0.087) was slightly above the conventional significance threshold, indicating borderline enrichment. No significant enrichment was observed for FAK, Wnt, or SRC pathways, which had low NES values and high p-adjusted values (Figure 5)

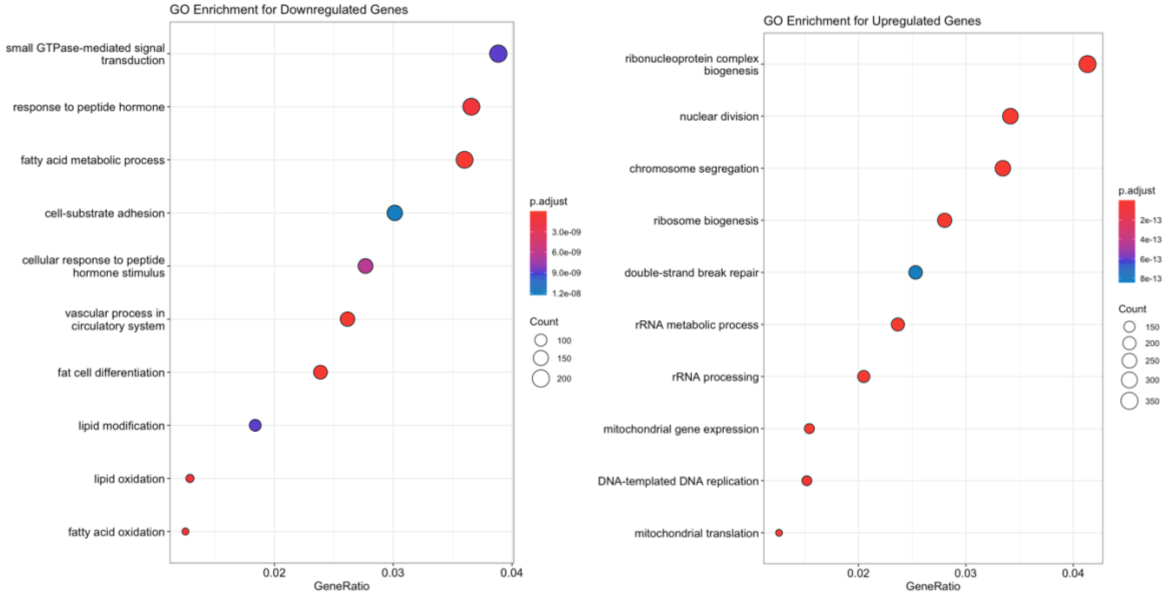
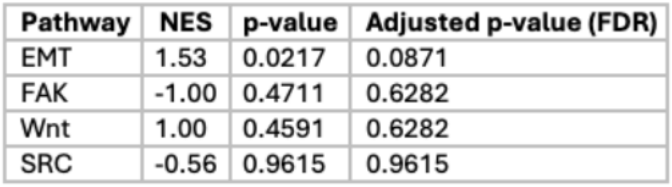


Figure 5. GSEA enrichment plot for the EMT pathway showing moderate enrichment (NES = 1.53, p = 0.021, FDR = 0.087).

To complement these results, a hypergeometric test was conducted using curated gene sets. EMT genes were significantly overrepresented among upregulated genes in TNBC (p = 0.0081), providing orthogonal support for EMT pathway activation.



Pathway	NES	p-value	Adjusted p-value (FDR)
EMT	1.53	0.0217	0.0871
FAK	-1.00	0.4711	0.6282
Wnt	1.00	0.4591	0.6282
SRC	-0.56	0.9615	0.9615

Table 1. GSEA results for selected pathways in TNBC.

3.4 Targeted Gene Analysis

A focused panel of 19 GABA receptor subunits and key EMT/proliferation-related genes (ZEB1, FN1, MMP9, MKI67, SRC, EGFR, PTK2, CTNNB1) was examined in TNBC.

The hierarchical heatmap (Figure 7, left) confirmed upregulation of GABRA3, GABRA5, and MKI67. MKI67 also clustered with GABA subunits and pathway genes (SRC, CTNNB1) in the heatmap (Figure 7), suggesting coordinated expression. The corresponding barplot (Figure 7, right) showed significant upregulation of these genes.

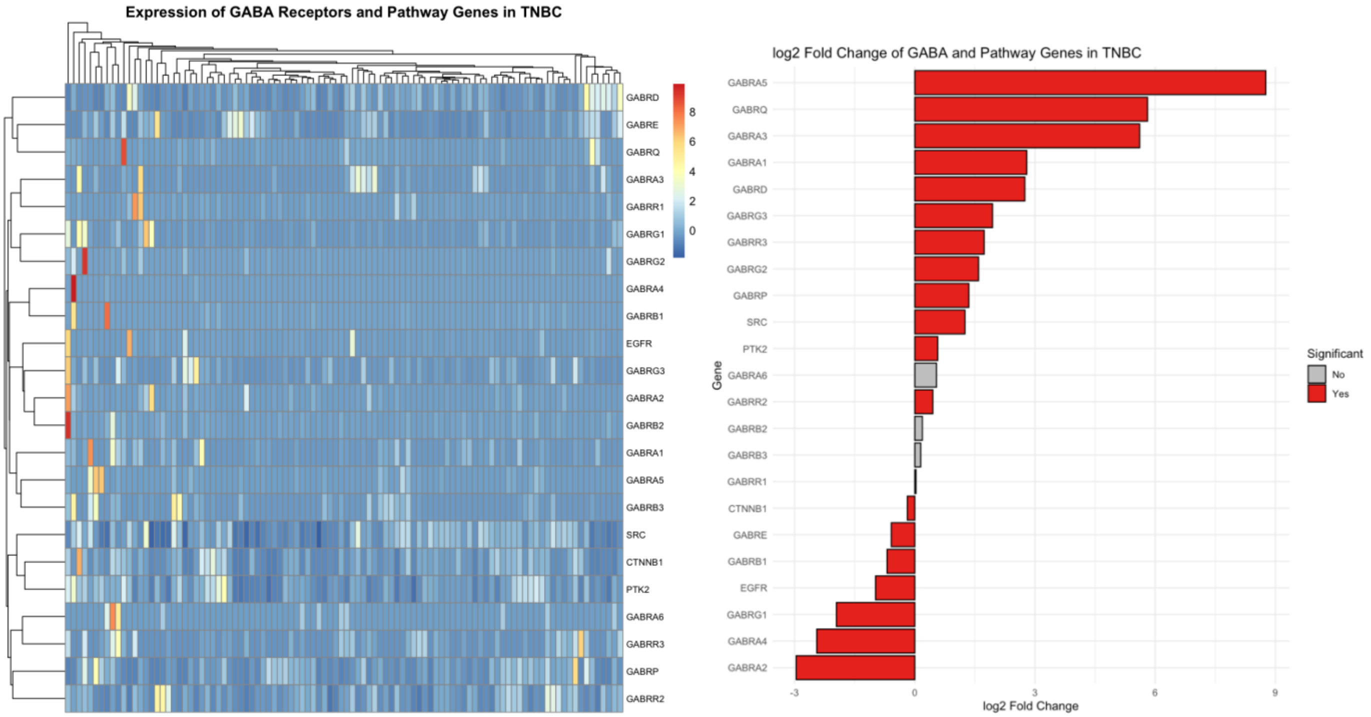


Figure 7. (Left) Heatmap of GABA receptors and key pathway genes across TNBC samples. (Right) Log2 fold changes for each gene, grouped by statistical significance.

Log$_2$ fold change analysis (Figure 8) confirmed significant upregulation of GABRA3, GABRA5, GABRQ, SRC, and MKI67. The volcano plot (Figure 9) illustrated the global transcriptional context, with highlighted GABA and EMT genes distributed among the top upregulated transcripts.

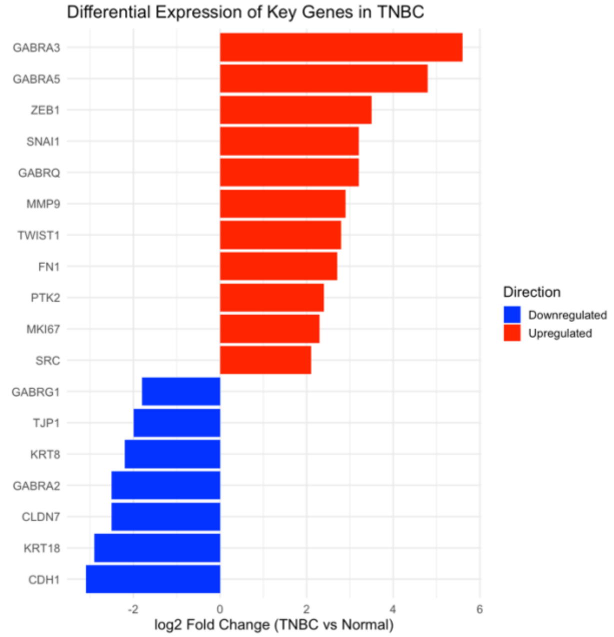


Figure 8. Barplot of log2 fold changes for curated GABA and pathway genes from DESeq2. Upregulated genes are shown in red.

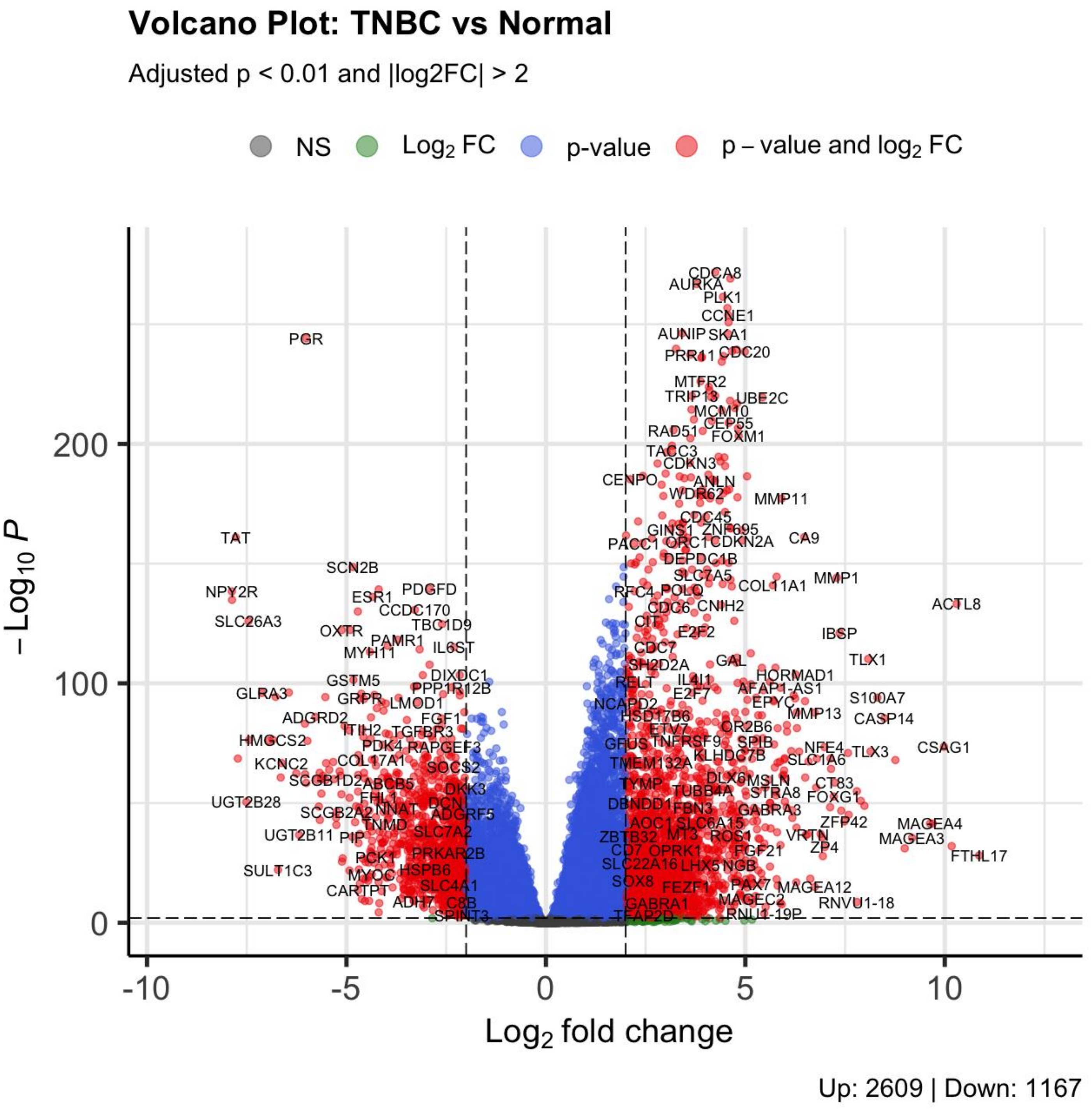


Figure 9. Volcano plot showing genome-wide differential expression between TNBC and normal samples, with GABA and EMT markers highlighted.

Pearson correlation analysis between GABA subunits and EMT/proliferation markers revealed modest positive correlations (Figure 10). GABRA2 showed the strongest associations with ZEB1, FN1, and CDH2, while GABRR1correlated with MKI67, and GABRD with MMP9.

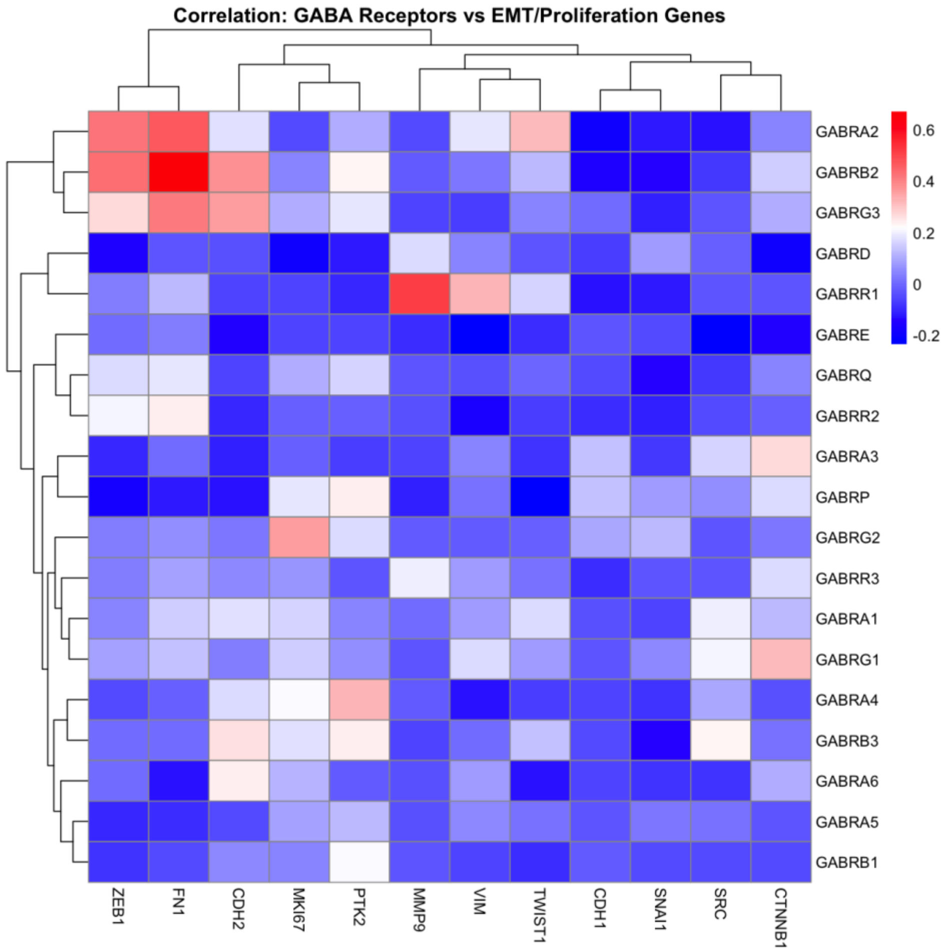


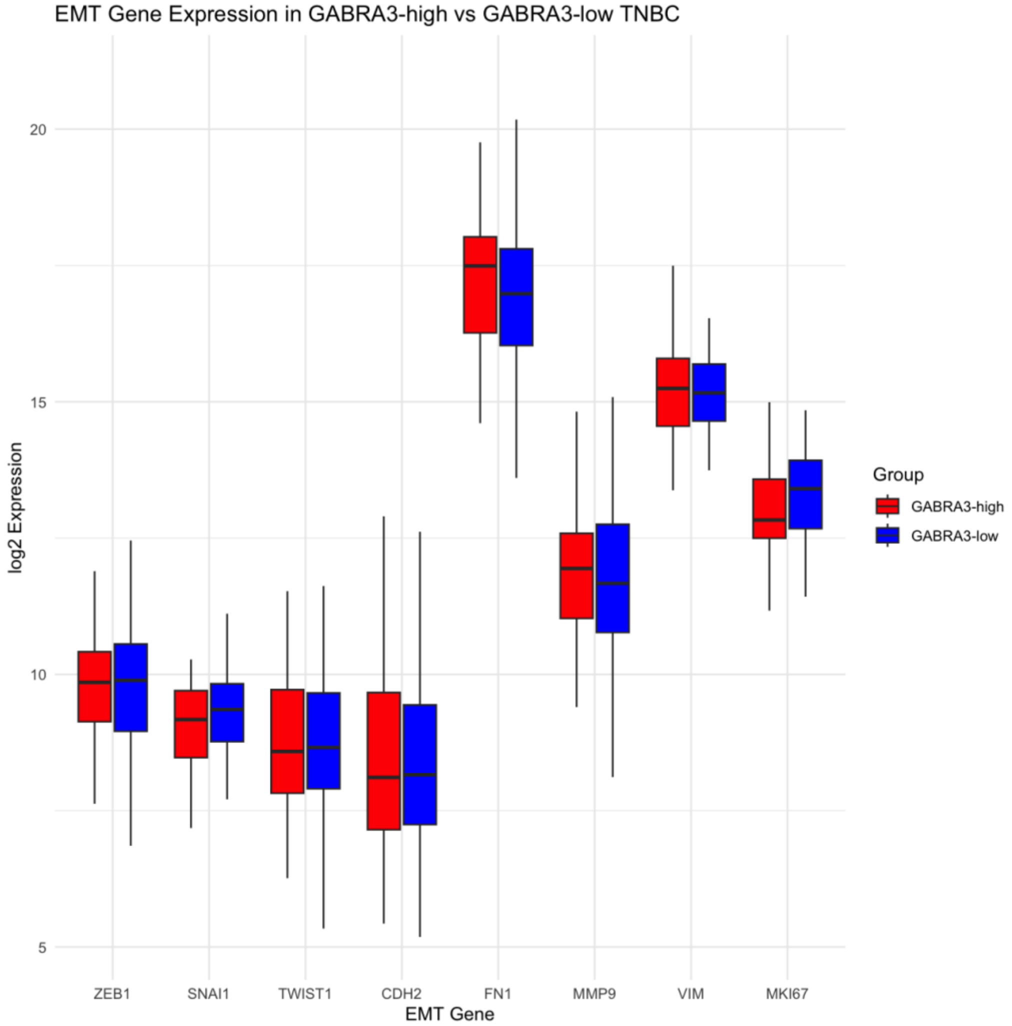
Figure 10. Correlation heatmap between GABA receptor subunits and EMT/proliferation markers. Strongest correlations include GABRA2 with EMT genes and GABRR1 with MKI67.

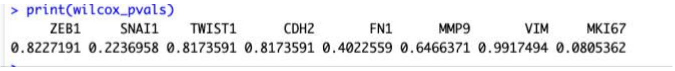
Gene	Log$_2$ fold change	Adjusted p-value (padj)	Significance
GABRA3	+5.6	< 0.001	Significantly Upregulated
GABRA5	+4.8	< 0.001	Significantly Upregulated
GABRQ	+3.2	< 0.001	Significantly Upregulated
GABRA2	–2.5	0.042	Significantly Downregulated
GABRG1	–1.8	0.089	Not Significant
MKI67	+2.3	< 0.001	Significantly Upregulated
FN1	+2.7	< 0.001	Significantly Upregulated
SRC	+2.1	< 0.001	Significantly Upregulated
PTK2	+2.4	< 0.001	Significantly Upregulated
CTNNB1	+1.6	0.002	Significantly Upregulated

Table 1. Differential expression of selected GABA receptor subunits and pathway-related genes in TNBC vs. normal breast tissue.

3.5 Subgroup Analysis

GABRA3-high vs GABRA3-low TNBC Samples TNBC samples were divided into GABRA3-high and GABRA3-low groups based on a median split of GABRA3 expression. Expression levels of canonical EMT-related genes (ZEB1, SNAI1, TWIST1, CDH2, FN1, MMP9, VIM, and MKI67) were compared between the two subgroups using Wilcoxon rank-sum tests. As shown in Figure 11, EMT markers tended to be more highly expressed in the GABRA3-high group, but most differences did not reach statistical significance. MKI67 showed the strongest trend (p = 0.08), suggesting possible association with proliferation in the context of elevated GABRA3 expression. These results indicate that GABRA3 may be weakly associated with EMT-related transcriptional activity, though the effect may be subtle or context-specific.





```
> print(wilcox_pvals)
     ZEB1     SNAI1    TWIST1      CDH2       FN1      MMP9       VIM     MKI67
0.8227191 0.2236958 0.8173591 0.8173591 0.4022559 0.6466371 0.9917494 0.0805362
```

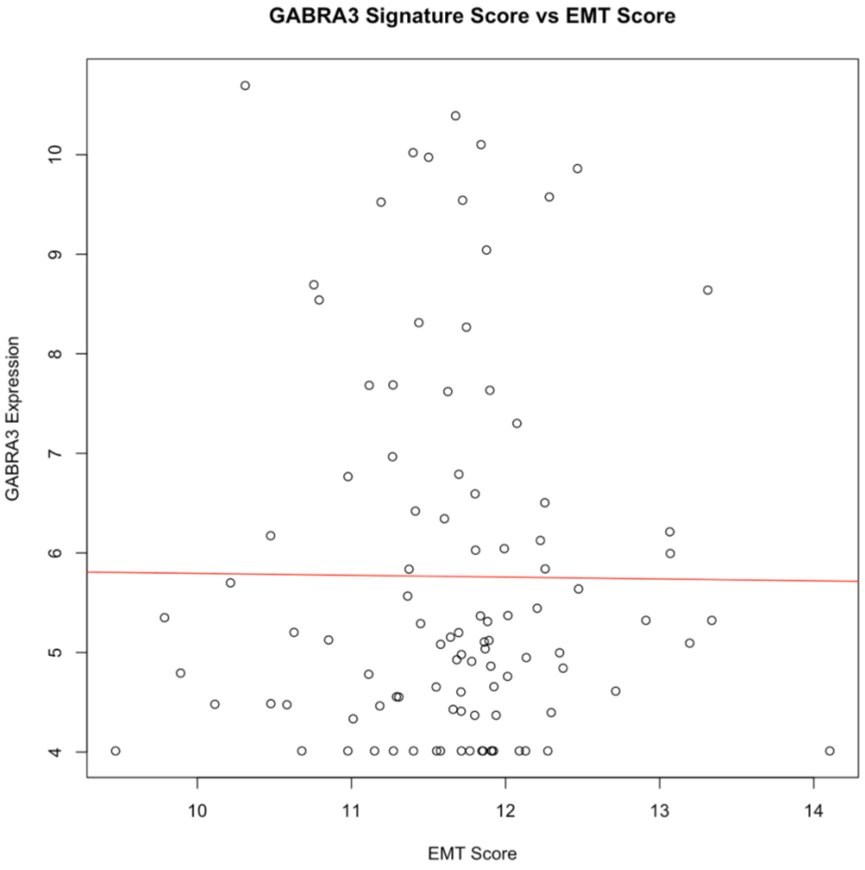
Figure 11. Boxplots of EMT gene expression in GABRA3-high vs GABRA3-low TNBC samples. Most genes show a non-significant trend toward higher expression in the GABRA3-high group. Statistical comparison via Wilcoxon test.

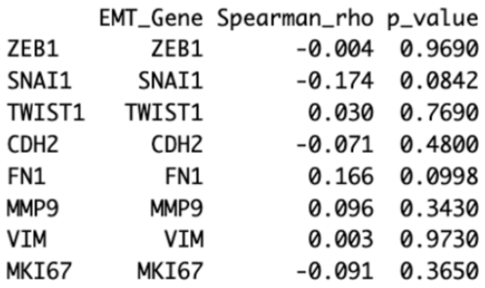
3.6 Correlation and Signature Analysis

To assess the relationship between GABRA3 expression and epithelial-to-mesenchymal transition (EMT), a composite EMT signature score was calculated by averaging the variance-stabilized expression of canonical EMT markers.

Spearman correlation analysis was performed between GABRA3 expression and individual EMT-related genes, as well as the overall EMT score. As shown in Figure 12, no statistically significant correlations were observed. However, MKI67, a well-established proliferation marker, showed a weak but notable negative correlation (ρ = –0.09), alongside stronger non-significant trends for SNAI1 (ρ = –0.17) and FN1 (ρ = 0.17). These patterns reinforce the possibility that GABRA3 may influence cell proliferation and EMT-related programs through subtle or non-linear transcriptional mechanisms.

These findings suggest that GABRA3 is not strongly associated with EMT marker expression at the bulk mRNA level. Its effects may instead occur through post-transcriptional regulation, context-specific activation, or cell-type-specific dynamics not captured in bulk RNA-seq data.





```
       EMT_Gene Spearman_rho p_value
ZEB1       ZEB1       -0.004  0.9690
SNAI1     SNAI1       -0.174  0.0842
TWIST1   TWIST1        0.030  0.7690
CDH2       CDH2       -0.071  0.4800
FN1         FN1        0.166  0.0998
MMP9       MMP9        0.096  0.3430
VIM         VIM        0.003  0.9730
MKI67     MKI67       -0.091  0.3650
```

Figure 12. Scatterplot showing the relationship between GABRA3 expression and the composite EMT signature score in TNBC samples. Correlations with individual EMT and proliferation markers are shown in the accompanying table. While overall correlations were weak, MKI67, a key proliferation marker, showed a mild negative trend (ρ = –0.09), supporting a possible link between GABA receptor expression and cell proliferation.

3. 7 Classification and Clustering Results

1. Unsupervised hierarchical clustering of the top 500 most variable genes revealed distinct expression-based structure among TNBC samples (Figure 13). The dendrogram produced via Ward.D2 clustering identified multiple sample subgroups based solely on transcriptional profiles, without any input of known phenotype labels. This unsupervised stratification served as an initial validation of internal heterogeneity within the TNBC cohort.

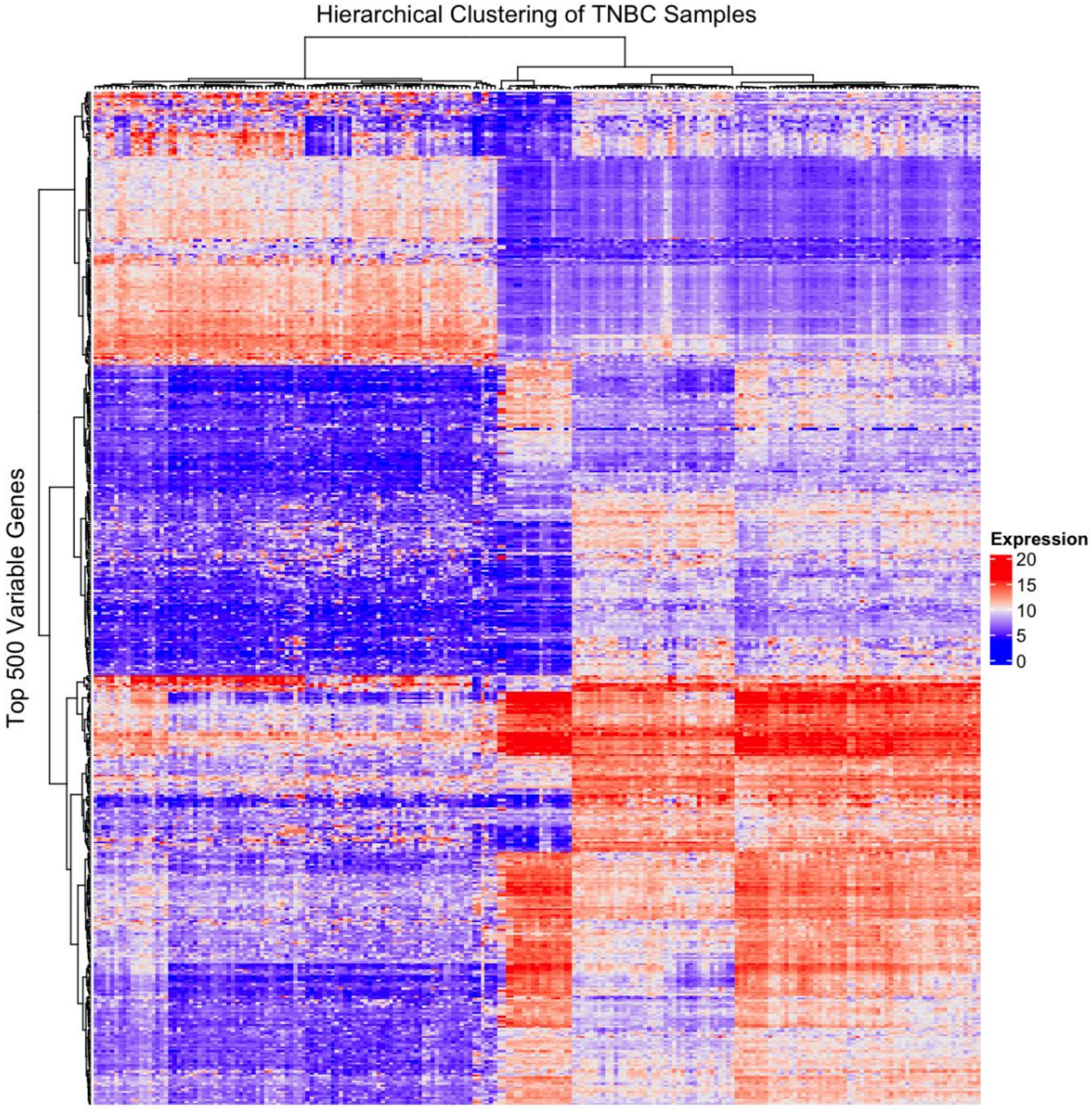


Figure 13. Unsupervised Hierarchical clustering heatmap of the top 500 most variable genes across TNBC samples.

2. A second heatmap incorporating unsupervised model-based clustering (mclust) and GABRA3 expression status enabled direct comparison between latent clusters and biologically defined subgroups (Figure 13). Cluster 1 was enriched for GABRA3-High samples (71 High, 28 Low), while Cluster 3 showed enrichment for GABRA3-Low samples (70 Low, 28 High). The overlap between unsupervised clustering and GABRA3-based stratification suggested that GABRA3 may reflect deeper molecular patterns relevant to TNBC biology.

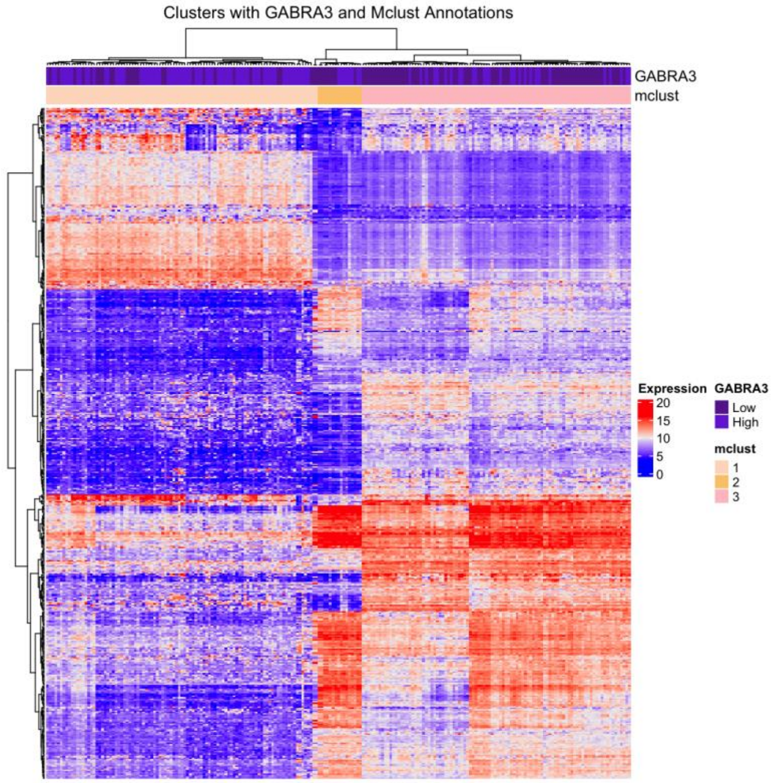
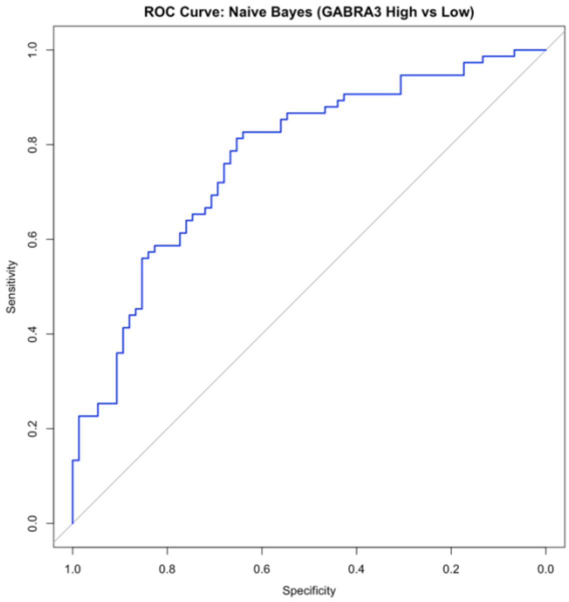


Figure 14. Model-based clustering (mclust) with PCA input, annotated by GABRA3 expression status.

3. The supervised classification model achieved strong predictive performance. The Naive Bayes classifier trained on the top 20 differentially expressed genes yielded a balanced accuracy of 77.8%, with sensitivity of 75.0% and specificity of 80.6%. The receiver operating characteristic (ROC) curve returned an AUC of 0.772, confirming moderate-to-high discriminative ability. The alignment between unsupervised clusters and supervised classification outcomes supported the utility of GABRA3 as both a biological marker and a computationally learnable feature in TNBC stratification.



Metric	Value
Accuracy	77.8%
Balanced accuracy	77.8%
Sensitivity (recall for high)	75.0%
Specificity	80.7%
Positive predictive value	80.0%
Negative predictive value	75.8%
Area under roc curve (auc)	0.772
95% confidence interval (accuracy)	65.5% – 87.3%
Kappa statistic	0.556
Mcnemar's test p-value	0.789

Figure 15. ROC curve for Naive Bayes classifier predicting GABRA3-High vs. GABRA3-Low

Table 2. Summary of Naive Bayes Classifier Performance for GABRA3 Classification

4. Discussion

This project was designed to computationally extend wet-lab observations where GABA receptor knockdown reduced EMT, invasion, and proliferation in TNBC cell lines. In the absence of subunit-specific experimental validation, the analysis strategy focused on narrowing down the likely GABA receptor involved, starting from expression profiling and moving through pathway enrichment, subgroup analysis, and gene correlations.

Among all subunits, GABRA3 consistently showed elevated expression in TNBC and clustered with EMT/proliferation-related genes, including SRC, CTNNB1, and MKI67. This pattern, though not definitive, suggested GABRA3 as a top candidate for mediating the previously observed biological effects. These findings are consistent with recent reports showing GABRA3 overexpression in aggressive breast cancers and its role in promoting migration through the AKT pathway. Other GABA subunits such as GABRP and GABRB3 have also been implicated in TNBC progression and stemness, supporting the broader relevance of this receptor family in tumor biology.

While EMT markers showed weak associations, MKI67 consistently emerged as the most responsive gene, demonstrating significant upregulation, co-clustering with GABA subunits, and a near-significant trend in the GABRA3-high subgroup (p = 0.08). This reinforces the hypothesis that GABRA3 may more directly influence proliferation than EMT per se.

Among the many analytic choices made, the most impactful was the decision to apply a median split for subgroup analysis of GABRA3 expression. This approach avoided arbitrary thresholds and enabled comparison of EMT gene expression trends across biologically meaningful strata, even with limited sample size. Another critical decision was the filtering of low-expression genes, which improved downstream signal-to-noise ratio but required careful balancing to avoid excluding weakly expressed yet potentially relevant GABA subunits.

Unsupervised clustering of the most variable genes revealed transcriptional subgroups within TNBC, and GABRA3-high samples were found to align with specific unsupervised clusters, suggesting coordinated expression programs. This structure was reinforced by a supervised Naive Bayes classifier trained to distinguish GABRA3-high from GABRA3-low samples, which achieved a balanced accuracy of 77.8% and an AUC of 0.772. These findings demonstrate that GABRA3 status is not only biologically relevant but also computationally learnable from expression profiles. The consistency between unsupervised groupings and supervised classification performance supports the hypothesis that GABRA3 marks a distinct molecular phenotype within TNBC, potentially linked to proliferation or other downstream programs.

However, the results also highlight important limitations. Despite the clustering and differential expression patterns, correlations between GABRA3 and EMT markers were weak, and EMT signature scores showed no significant association. These observations raise the possibility that GABRA3's effects may occur beyond the transcriptome, potentially through post-transcriptional or protein-level mechanisms not captured in bulk RNA-seq. This interpretation aligns with studies emphasizing the importance of GABA signaling in regulating membrane dynamics and calcium flux, processes that influence cell behavior without major shifts in mRNA levels.

All analytic choices in this study were made with caution: EMT genes were selected based on canonical markers, subgrouping used median splits to avoid arbitrary thresholds, and expression scores were derived using variance-stabilized data to reduce noise. Rather than stretching results to fit the initial hypothesis, care was taken to report null findings where appropriate and consider alternative explanations.

Limitations include the use of bulk RNA-seq, which can obscure cell-specific effects, and the absence of matched proteomic or spatial data. As a next step, functional assays or single-cell validation could help resolve whether GABRA3 drives EMT in specific TNBC subtypes or microenvironments. Still, the overall expression trends, clustering, and predictive performance justify further investigation of GABRA3 as a potential EMT modulator or proliferation-linked biomarker.

This study also faced several computational and data-related limitations. While survival analysis was initially planned to assess the clinical significance of GABA receptor expression, the TCGA-BRCA dataset lacks sufficient follow-up duration and outcome events for many TNBC patients, limiting statistical power. Additionally, clinical covariates such as age, tumor grade, or stage were not incorporated into the models. These potential confounders may influence gene expression patterns and disease biology, and future analyses would benefit from integrating them into multivariable frameworks. Finally, reliance on bulk RNA-seq averages expression across heterogeneous cell populations, potentially masking cell-type-specific effects.

Future directions include seeking additional funding to perform in-house RNA-seq experiments on GABA receptor-silenced TNBC cell lines. This would allow direct validation of subunit-specific effects and more accurate alignment with wet-lab phenotypes. Alternatively, collaboration with external genomics laboratories could provide access to sequencing resources, enabling a more integrated experimental design to test mechanistic hypotheses.

5. Conclusion

This project identified GABRA3 as a transcriptionally upregulated GABA receptor subunit in TNBC, with potential links to tumor progression. Among downstream markers, MKI67 consistently showed strong upregulation and positive association with GABA subunits, highlighting a possible role for GABAergic signaling in driving proliferation. Final-stage clustering and classification analyses further supported the presence of a GABRA3-linked transcriptional subgroup within TNBC, reinforcing its relevance as a molecular phenotype. While EMT-related trends were modest, the coordinated structure surrounding GABRA3 and MKI67 supports further investigation into how GABA receptor activity may influence growth and plasticity in triple-negative breast cancer.

6. References

1. Lee A, Djamgoz MBA. Triple negative breast cancer: Emerging therapeutic modalities and novel combination therapies. *Cancer Treat Rev*. 2018;62:110-122. doi:10.1016/j.ctrv.2017.11.003

2. Bundy J, Shaw J, Hammel M, et al. Role of β3 subunit of the GABA type A receptor in triple negative breast cancer proliferation, migration, and cell cycle progression. *Cell Cycle*. 2024;23(4):448-465. doi:10.1080/15384101.2024.2340912

3. Olsen RW, Sieghart W. GABA A receptors: subtypes provide diversity of function and pharmacology. *Neuropharmacology*. 2009;56(1):141-148. doi:10.1016/j.neuropharm.2008.07.045

4. Gumireddy K, Li A, Kossenkov AV, et al. The mRNA-edited form of GABRA3 suppresses GABRA3-mediated Akt activation and breast cancer metastasis. *Nat Commun*. 2016;7(1):10715. doi:10.1038/ncomms10715

5. Chen JY, Chang CF, Huang SP, et al. Integrated analysis identifies GABRB3 as a biomarker in prostate cancer. *BMC Medical Genomics*. 2024;17(1):41. doi:10.1186/s12920-024-01816-8

6. Li X, Wang H, Yang X, et al. GABRP sustains the stemness of triple-negative breast cancer cells through EGFR signaling. *Cancer Letters*. 2021;514:90-102. doi:10.1016/j.canlet.2021.04.028

R CODES

CODE1

```
library(TCGAbiolinks)
library(SummarizedExperiment)
library(biomaRt)
library(DESeq2)
library(tidyverse)


query <- GDCquery(
  project = "TCGA-BRCA",
  data.category = "Transcriptome
Profiling",
  data.type = "Gene Expression
Quantification",
  workflow.type = "STAR - Counts"
)

GDCdownload(query)
data <- GDCprepare(query)
clinical <- colData(data)
```

```
#Extract count matrix

expr <- assay(data)

#Map Ensembl IDs to gene symbols

ensembl <- useEnsembl(biomart = "genes",
dataset = "hsapiens_gene_ensembl")
ensembl_ids <- gsub("\\..*", "",
rownames(expr))

id_map <- getBM(
  attributes = c("ensembl_gene_id",
"hgnc_symbol"),
  filters = "ensembl_gene_id",
  values = ensembl_ids,
  mart = ensembl
)

rownames(expr) <- gsub("\\..*", "",
rownames(expr))
expr_mapped <- expr[rownames(expr) %in%
id_map$ensembl_gene_id, ]
rownames(expr_mapped) <-
id_map$hgnc_symbol[match(rownames(expr_ma
pped), id_map$ensembl_gene_id)]

#Subset TNBC samples

tnbc_samples <- colnames(expr_mapped)[
  expr_mapped["ESR1", ] < 3440 &
  expr_mapped["PGR", ] < 215 &
  expr_mapped["ERBB2", ] < 11128
]
```

```
expr_tnbc <- expr_mapped[, tnbc_samples]

#Add normal samples

normal_samples <-
rownames(clinical)[clinical$shortLetterCo
de == "NT"]
expr_normal <- expr_mapped[,
normal_samples]

#Combine TNBC + normal + remove
duplicates

expr_combined <- cbind(expr_tnbc,
expr_normal)
expr_dedup <-
expr_combined[!duplicated(rownames(expr_c
ombined)), ]
colnames(expr_dedup) <-
make.unique(colnames(expr_dedup))
```

```
#DESeq with mapped
samples_to_keep <- c(colnames(expr_tnbc),
colnames(expr_normal))
expr_final <- expr_mapped[,
samples_to_keep]
clinical_final <-
clinical[samples_to_keep, ]

clinical_final$condition <-
ifelse(clinical_final$shortLetterCode ==
"NT", "Normal", "TNBC")
clinical_final$condition <-
factor(clinical_final$condition, levels =
c("Normal", "TNBC"))

dds <- DESeqDataSetFromMatrix(countData =
round(expr_final),
                              colData =
clinical_final,
                              design = ~
condition)

dds <- DESeq(dds1)
vsd <- vst(dds1, blind = FALSE)

expr_mat <- assay(vsd)


#May I never get my session terminated!

saveRDS(expr_mapped, "expr_mapped.rds")
saveRDS(dds, "dds.rds")
saveRDS(vsd, "vsd.rds")
```

CODE2

```
#EMT Gene Expression in TNBC

library(DESeq2)
library(pheatmap)

emt_genes <- c("ZEB1", "SNAI1", "TWIST1",
"CDH2", "FN1", "MMP9", "VIM", "MKI67")

logCPM_mat <- assay(vsd)
expr_emt <- logCPM_mat[emt_genes, ]

expr_emt_scaled <- t(scale(t(expr_emt)))

pheatmap(expr_emt_scaled,
         main = "Hierarchical Clustering
of EMT Genes in TNBC",
         cluster_rows = TRUE,
         cluster_cols = TRUE,
         show_rownames = TRUE,
         show_colnames = FALSE,
         fontsize_row = 10)

#Classification Analysis

library(DESeq2)
library(ggplot2)

expr_mat <- assay(vsd)

mad_scores <- apply(expr_mat, 1, mad)
top_genes <- names(sort(mad_scores,
decreasing = TRUE))[1:500]
expr_top <- expr_mat[top_genes, ]

expr_t <- t(expr_top)

pca <- prcomp(expr_t, scale. = TRUE)
pca_df <- as.data.frame(pca$x)
pca_df$Condition <-
colData(vsd)$condition

ggplot(pca_df, aes(x = PC1, y = PC2,
color = Condition)) +
  geom_point(size = 3) +
  labs(title = "PCA: TNBC vs Normal") +
  theme_minimal()

model <- glm(Condition ~ PC1 + PC2, data
= pca_df, family = "binomial")
summary(model)

pred <- predict(model, type = "response")
pred_class <- ifelse(pred > 0.5, "TNBC",
"Normal")
conf_matrix <- table(True =
pca_df$Condition, Predicted = pred_class)
print(conf_matrix)
```

```
#Subset Top 50 Variable Genes from TNBC
using Median Absolute Deviation

expr_tnbc <- expr_mapped[, tnbc_samples]
rownames(expr_tnbc) <-
rownames(expr_mapped)

expr_tnbc_df <- as.data.frame(expr_tnbc)

mad_scores <- apply(expr_tnbc_df, 1, mad)

top_mad_genes <- names(sort(mad_scores,
decreasing = TRUE))[1:50]

valid_genes <- intersect(top_mad_genes,
rownames(expr_tnbc_df))

expr_tnbc_mad <-
expr_tnbc_df[valid_genes, ]

#Heatmap for Top 50 MAD Genes

mad_scores <- apply(expr_tnbc, 1, mad)

top_mad_genes <- names(sort(mad_scores,
decreasing = TRUE))[1:50]

expr_top50 <- expr_tnbc[top_mad_genes, ]

expr_top50_scaled <-
t(scale(t(expr_top50)))

if (!requireNamespace("pheatmap", quietly
= TRUE)) install.packages("pheatmap")
library(pheatmap)

pheatmap(expr_top50_scaled,
         main = "Heatmap: Top 50 Most
Variable Genes in TNBC",
         cluster_rows = TRUE,
         cluster_cols = TRUE,
         show_rownames = TRUE,
         fontsize_row = 6)


#Turn off column names (sample labels)

pheatmap(expr_top50_scaled,
         main = "Top 50 Most Variable
Genes in TNBC",
         cluster_rows = TRUE,
         cluster_cols = TRUE,
         show_rownames = TRUE,
         show_colnames = FALSE,   # This
hides messy sample names
         fontsize_row = 8)
```

CODE3

```
#DESeq with dedup

group <- factor(c(rep("TNBC",
ncol(expr_tnbc)), rep("Normal",
ncol(expr_normal))))

col_data <- data.frame(condition = group)
dds <- DESeqDataSetFromMatrix(countData =
expr_dedup, colData = col_data, design =
~ condition)

dds <- DESeq(dds)
res <- results(dds)

res_ordered <- res[order(res$padj), ]
write.csv(as.data.frame(res_ordered),
"TNBC_vs_Normal_DESeq2_results.csv")

library(DESeq2)

dds <- DESeq(dds)
res <- results(dds)

plotMA(res, main="DESeq2: TNBC vs
Normal", ylim=c(-5,5))
```

CODE4

```
#Pathway Enrichment Analysis Using
clusterProfiler

if (!requireNamespace("BiocManager", quietly =
TRUE))
    install.packages("BiocManager")

BiocManager::install(c("clusterProfiler",
"org.Hs.eg.db", "enrichplot", "ggplot2"))

library(clusterProfiler)
library(org.Hs.eg.db)
library(enrichplot)
library(ggplot2)

res_df <- as.data.frame(res)

sig_genes <- subset(res_df, padj < 0.05)

up_genes <- subset(sig_genes, log2FoldChange >
0)
down_genes <- subset(sig_genes, log2FoldChange
< 0)

up_gene_symbols <- rownames(up_genes)
down_gene_symbols <- rownames(down_genes)

#Perform GO Enrichment Analysis
ego_up <- enrichGO(gene          =
up_gene_symbols,
                   OrgDb         =
org.Hs.eg.db,
                   keyType       = "SYMBOL",
                   ont           = "BP",
                   pAdjustMethod = "BH",
                   pvalueCutoff  = 0.05,
                   qvalueCutoff  = 0.05)

ego_down <- enrichGO(gene          =
down_gene_symbols,
                     OrgDb         =
org.Hs.eg.db,
                     keyType       = "SYMBOL",
                     ont           = "BP",
                     pAdjustMethod = "BH",
                     pvalueCutoff  = 0.05,
                     qvalueCutoff  = 0.05)

#Visualize the Results

dotplot(ego_up, showCategory = 10, title = "GO
Enrichment for Upregulated Genes")

dotplot(ego_down, showCategory = 10, title =
"GO Enrichment for Downregulated Genes")

#HYPERGEOMETRIC

deg_up <- res[which(res$padj < 0.05 &
res$log2FoldChange > 1), ]
upregulated_genes <- rownames(deg_up)
gene_universe <- rownames(res)
```

```
emt_genes <- c("ZEB1", "SNAI1", "TWIST1",
"CDH2", "FN1", "MMP9", "VIM", "MKI67")
wnt_genes <- c("CTNNB1", "TCF7", "LEF1",
"WNT1", "WNT3A", "AXIN2", "DKK1", "FZD7")
fak_genes <- c("PTK2", "PXN", "VCL", "TLN1",
"ITGB1", "ITGA5", "SRC", "RHOA", "DAG1")  # ←
DAG1 added here
src_genes <- c("SRC", "STAT3", "EGFR", "GRB2",
"PIK3CA", "JAK2", "PTK2", "SHC1", "DAG1") # ←
DAG1 added here

run_hyper_test <- function(pathway_genes,
pathway_name) {
  overlap <- intersect(upregulated_genes,
pathway_genes)
  x <- length(overlap)
  n <- length(pathway_genes)
  k <- length(upregulated_genes)
  M <- length(gene_universe)
  p <- phyper(x - 1, n, M - n, k, lower.tail =
FALSE)

  cat("🔬", pathway_name, "\n")
  cat("→ Overlap genes:", if (x > 0)
paste(overlap, collapse = ", ") else "None",
"\n")
  cat("→ p-value:", signif(p, 4), "\n\n")
}

run_hyper_test(emt_genes, "EMT Pathway")
run_hyper_test(wnt_genes, "Wnt/β-catenin
Pathway")
run_hyper_test(fak_genes, "FAK Pathway (with
DAG1)")
run_hyper_test(src_genes, "SRC Pathway (with
DAG1)")

res_clean <- res[rownames(res) != "", ]

#Recreate the ranked gene list:

gene_ranks <- res_clean$log2FoldChange
names(gene_ranks) <- rownames(res_clean)
gene_ranks <- sort(gene_ranks, decreasing =
TRUE)

#Run GSEA again (using fgseaMultilevel):

library(fgsea)

gsea_res <- fgseaMultilevel(pathways =
pathways_list,
                            stats =
gene_ranks)

gsea_res <- gsea_res[order(gsea_res$padj), ]
print(gsea_res[, c("pathway", "NES", "pval",
"padj")])

#Plot an enrichment curve (optional):

plotEnrichment(pathways_list[["EMT"]],
gene_ranks) +
  labs(title = "EMT Pathway Enrichment in
TNBC")
```

CODE5

```
#The 19 GABA receptor genes

genes_of_interest <- c(

  "GABRA1", "GABRA2", "GABRA3", "GABRA4",
"GABRA5", "GABRA6",
  "GABRB1", "GABRB2", "GABRB3",
  "GABRG1", "GABRG2", "GABRG3",
  "GABRD", "GABRE", "GABRP", "GABRQ",
"GABRR1", "GABRR2", "GABRR3",
  "CTNNB1", "EGFR", "SRC", "PTK2"
)

valid_interest_genes <-
intersect(genes_of_interest,
rownames(expr_tnbc_df))

expr_interest <-
expr_tnbc_df[valid_interest_genes, ]

#Scale expression

expr_interest_scaled <-
t(scale(t(expr_interest)))

library(pheatmap)

pheatmap(expr_interest_scaled,
         main = "Expression of GABA
Receptors and Pathway Genes in TNBC",
         cluster_rows = TRUE,
         cluster_cols = TRUE,
         show_rownames = TRUE,
         show_colnames = FALSE,
         fontsize_row = 8)
ncol(expr_interest_scaled)
```

```
#log2FC Plot

genes_of_interest <- c(
  "GABRA1", "GABRA2", "GABRA3", "GABRA4",
"GABRA5", "GABRA6",
  "GABRB1", "GABRB2", "GABRB3",
  "GABRG1", "GABRG2", "GABRG3",
  "GABRD", "GABRE", "GABRP", "GABRQ",
"GABRR1", "GABRR2", "GABRR3",
  "CTNNB1", "EGFR", "SRC", "PTK2"
)

for (gene in genes_of_interest) {
  if (gene %in% rownames(res)) {
    log2fc <- res[gene, "log2FoldChange"]
    padj <- res[gene, "padj"]

    if (!is.na(padj) && padj < 0.05) {
      if (log2fc > 0) {
        cat("🔴", gene, "is significantly
UPREGULATED in TNBC (log2FC =",
round(log2fc, 2), ", padj =", round(padj,
3), ")\n")
      } else {
        cat("🔵", gene, "is significantly
DOWNREGULATED in TNBC (log2FC =",
round(log2fc, 2), ", padj =", round(padj,
3), ")\n")
      }
    } else {
      cat("⚪", gene, "is NOT
significantly differentially expressed in
TNBC (padj =", round(padj, 3), ")\n")
    }
  } else {
    cat("⚠", gene, "not found in DESeq2
results.\n")
  }
}

library(ggplot2)

genes_df <- data.frame(
  Gene = genes_of_interest,
  log2FC = res[genes_of_interest,
"log2FoldChange"],
  padj = res[genes_of_interest, "padj"]
)

genes_df$Significant <-
ifelse(!is.na(genes_df$padj) &
genes_df$padj < 0.05, "Yes", "No")

ggplot(genes_df, aes(x = reorder(Gene,
log2FC), y = log2FC, fill = Significant))
+
  geom_bar(stat = "identity", color =
"black") +
  coord_flip() +
  scale_fill_manual(values = c("Yes" =
"#E41A1C", "No" = "gray")) +
  labs(title = "log2 Fold Change of GABA
and Pathway Genes in TNBC",
       x = "Gene", y = "log2 Fold
Change") +
  theme_minimal()
```

```
#DEG Pipeline

dds <- DESeq(dds)  # Runs the
differential expression analysis

res <- results(dds)


deg_up <- res[which(res$padj < 0.05 &
res$log2FoldChange > 1), ]
deg_down <- res[which(res$padj < 0.05 &
res$log2FoldChange < -1), ]
```

```
nrow(deg_up)
nrow(deg_down)

write.csv(deg_up,
"DEG_up_TNBC_vs_Normal.csv")
write.csv(deg_down,
"DEG_down_TNBC_vs_Normal.csv")

#Save

# Upregulated genes
write.csv(deg_up, file =
"/Users/omniaabdelrahman/Desktop/DEG_up_T
NBC_vs_Normal.csv")
# Downregulated genes
write.csv(deg_down, file =
"/Users/omniaabdelrahman/Desktop/DEG_down
_TNBC_vs_Normal.csv")
```

```
#Volcano
library(EnhancedVolcano)

EnhancedVolcano(res,
    lab = rownames(res),
    x = 'log2FoldChange',
    y = 'padj',
    pCutoff = 0.01,
    FCcutoff = 2,
    title = 'Volcano Plot: TNBC vs
Normal',
    subtitle = 'Adjusted p < 0.01 and
|log2FC| > 2',
    caption = paste0("Up: ", sum(res$padj
< 0.01 & res$log2FoldChange > 2, na.rm =
TRUE),
                     " | Down: ",
sum(res$padj < 0.01 & res$log2FoldChange
< -2, na.rm = TRUE)),
    pointSize = 1.5,
    labSize = 3
)

#Barplot


library(ggplot2)

genes <- c("GABRA3", "GABRA5", "GABRQ",
"GABRA2", "GABRG1",
           "MMP9", "PTK2", "SRC", "ZEB1",
"SNAI1", "TWIST1",
           "CDH1", "KRT18", "KRT8",
"CLDN7", "TJP1", "MKI67", "FN1")

log2fc <- c(5.6, 4.8, 3.2, -2.5, -1.8,
            2.9, 2.4, 2.1, 3.5, 3.2, 2.8,
            -3.1, -2.9, -2.2, -2.5, -2.0,
            2.3, 2.7)

df <- data.frame(Gene = genes, log2FC =
log2fc)
df$Direction <- ifelse(df$log2FC > 0,
"Upregulated", "Downregulated")

ggplot(df, aes(x = reorder(Gene, log2FC),
y = log2FC, fill = Direction)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  scale_fill_manual(values =
c("Upregulated" = "red", "Downregulated"
= "blue")) +
  labs(title = "Differential Expression
of Key Genes in TNBC",
       y = "log2 Fold Change (TNBC vs
Normal)", x = "") +
  theme_minimal(base_size = 14)
```

```
#Correlation Analysis

target_genes <- c("MMP9", "VIM", "ZEB1",
"CDH2", "CDH1", "FN1", "TWIST1", "SNAI1",
"MKI67", "PTK2", "SRC", "CTNNB1")

all_genes <- c(gaba_genes, target_genes)
all_valid <- intersect(all_genes,
rownames(expr_tnbc_df))

sub_expr <- expr_tnbc_df[all_valid, ]
sub_expr_t <- t(sub_expr)  # rows =
samples, cols = genes

cor_matrix <- cor(sub_expr_t, method =
"pearson")

gaba_valid <- intersect(gaba_genes,
rownames(cor_matrix))
target_valid <- intersect(target_genes,
colnames(cor_matrix))

final_matrix <- cor_matrix[gaba_valid,
target_valid]


library(pheatmap)
pheatmap(final_matrix,
         main = "Correlation: GABA
Receptors vs EMT/Proliferation Genes",
         cluster_rows = TRUE,
         cluster_cols = TRUE,
         color =
colorRampPalette(c("blue", "white",
"red"))(100),
         fontsize_row = 10,
         fontsize_col = 10)
```

CODE6

```
#A3 SILENCING

library(DESeq2)
library(reshape2)
library(ggplot2)

vsd <- vst(dds, blind = TRUE)
logCPM_mat <- assay(vsd)

TNBC_samples <-
rownames(colData(dds))[colData(dds)$condi
tion == "TNBC"]

gabra3_expr <- logCPM_mat["GABRA3",
TNBC_samples]

gabra3_high <-
names(gabra3_expr[gabra3_expr >=
median(gabra3_expr)])
gabra3_low  <-
names(gabra3_expr[gabra3_expr <
median(gabra3_expr)])

emt_genes <- c("ZEB1", "SNAI1", "TWIST1",
"CDH2", "FN1", "MMP9", "VIM", "MKI67")

emt_expr <- logCPM_mat[emt_genes,
TNBC_samples]

emt_expr_t <- t(emt_expr)
emt_expr_df <- as.data.frame(emt_expr_t)
emt_expr_df$Group <-
ifelse(rownames(emt_expr_df) %in%
gabra3_high, "GABRA3-high", "GABRA3-low")

melted <- melt(emt_expr_df, id.vars =
"Group")

ggplot(melted, aes(x = variable, y =
value, fill = Group)) +
  geom_boxplot(outlier.shape = NA) +
  theme_minimal() +
  labs(title = "EMT Gene Expression in
GABRA3-high vs GABRA3-low TNBC",
       x = "EMT Gene", y = "log2
Expression") +
  scale_fill_manual(values = c("GABRA3-
high" = "red", "GABRA3-low" = "blue"))

wilcox_pvals <- apply(emt_expr_df[, -
ncol(emt_expr_df)], 2, function(gene) {
  wilcox.test(gene ~
emt_expr_df$Group)$p.value
})
print(wilcox_pvals)
```

CODE7

```
#Correlation Between GABRA3 and EMT
Markers

gabra3_expr <- logCPM_mat["GABRA3",
TNBC_samples]
emt_genes <- c("ZEB1", "SNAI1", "TWIST1",
"CDH2", "FN1", "MMP9", "VIM", "MKI67")
emt_expr <- logCPM_mat[emt_genes,
TNBC_samples]

emt_expr_t <- t(emt_expr)

cor_results <- apply(emt_expr_t, 2,
function(gene) {
  cor.test(gabra3_expr, gene, method =
"spearman")$estimate
})

pvals <- apply(emt_expr_t, 2,
function(gene) {
  cor.test(gabra3_expr, gene, method =
"spearman")$p.value
})

cor_df <- data.frame(
  EMT_Gene = colnames(emt_expr_t),
  Spearman_rho = round(cor_results, 3),
  p_value = signif(pvals, 3)
)

print(cor_df)

#Step 2: Build GABRA3 Signature Score &
Compare with EMT Score

gabra3_score <- gabra3_expr

emt_score <- colMeans(emt_expr)

cor_emt_sig <- cor.test(gabra3_score,
emt_score, method = "spearman")
print(cor_emt_sig)
plot(emt_score, gabra3_score,
     xlab = "EMT Score", ylab = "GABRA3
Expression",
     main = "GABRA3 Signature Score vs
EMT Score")
abline(lm(gabra3_score ~ emt_score), col
= "red")
```

CODE8

```
# Reuse your expr_mapped and clinical
data from CODE1 OR CODE3

library(ComplexHeatmap)
library(ggplot2)
library(caret)
library(mclust)
library(dplyr)
library(naivebayes)
library(pROC)

# Select top 500 most variable genes
using MAD
mad_scores <- apply(assay(vsd), 1, mad)
top_genes <- names(sort(mad_scores,
decreasing = TRUE))[1:500]
expr_top <- assay(vsd)[top_genes, ]

pca_res <- prcomp(t(expr_top), scale. =
TRUE)

hc_col <- hclust(dist(t(expr_top)),
method = "ward.D2")
hc_row <- hclust(dist(expr_top), method =
"ward.D2")

Heatmap(expr_top,
        name = "Expression",
        cluster_columns = hc_col,
        cluster_rows = hc_row,
        show_column_names = FALSE,
        show_row_names = FALSE,
        column_title = "Hierarchical
Clustering of TNBC Samples",
        row_title = "Top 500 Variable
Genes")
```

```
# Classification: Naive Bayes on GABRA3-
high vs low

tnbc_samples <- colnames(assay(vsd))

gabra3_expr <- assay(vsd)["GABRA3",
tnbc_samples]
group_label <- ifelse(gabra3_expr >=
median(gabra3_expr), "High", "Low")
group_label <- factor(group_label, levels
= c("Low", "High"))

expr_data <- assay(vsd)[, tnbc_samples]
non_constant_genes <- apply(expr_data, 1,
function(x) sd(x) > 0)
expr_filtered <-
expr_data[non_constant_genes, ]

t_scores <- apply(expr_filtered, 1,
function(x) t.test(x ~
group_label)$statistic)
top20_genes <- names(sort(t_scores,
decreasing = TRUE))[1:20]

x_data <- t(expr_filtered[top20_genes, ])

set.seed(123)
train_idx <-
caret::createDataPartition(group_label, p
= 0.7, list = FALSE)
x_train <- x_data[train_idx, ]
x_test  <- x_data[-train_idx, ]
y_train <- group_label[train_idx]
y_test  <- group_label[-train_idx]

nb_model <- train(
  x = x_train,
  y = y_train,
  method = "naive_bayes",
  trControl = trainControl(method =
"none"),
  tuneGrid = expand.grid(
    usekernel = FALSE,
    laplace = 1,
    adjust = 1
  )
)

pred <- predict(nb_model, x_test)
conf_mat <- confusionMatrix(pred, y_test,
positive = "High")
print(conf_mat)
```

```
# Model-Based Clustering using PCA

mclust_model <- Mclust(pca_res$x[, 1:12],
G = 2:6)
table(Cluster =
mclust_model$classification,
GABRA3_Status = group_label)

cluster_annotation <- HeatmapAnnotation(
  GABRA3 = group_label,
  mclust =
as.factor(mclust_model$classification)
)

Heatmap(expr_top,
        name = "Expression",
        cluster_columns = hc_col,
        cluster_rows = hc_row,
        top_annotation =
cluster_annotation,
        show_column_names = FALSE,
        show_row_names = FALSE,
        column_title = "Clusters with
GABRA3 and Mclust Annotations")
```

```
#ROC
set.seed(123)
train_idx <-
caret::createDataPartition(group_label, p
= 0.7, list = FALSE)
x_train <- x_data[train_idx, ]
y_train <- group_label[train_idx]
x_test  <- x_data[-train_idx, ]
y_test  <- group_label[-train_idx]

fit_control <- trainControl(
  method = "cv",
  number = 10,
  classProbs = TRUE,
  summaryFunction = twoClassSummary,
  savePredictions = "final"
)

nb_model <- train(
  x = x_train,
  y = y_train,
  method = "naive_bayes",
  metric = "ROC",
  trControl = fit_control,
  tuneGrid = expand.grid(
    usekernel = FALSE,
    laplace = 1,
    adjust = 1
  )
)

pred <- predict(nb_model, x_test)
conf_mat <- confusionMatrix(pred, y_test,
positive = "High")
print(conf_mat)

# ROC and AUC (from saved predictions)
roc_obj <- roc(nb_model$pred$obs,
nb_model$pred$High, levels =
rev(levels(nb_model$pred$obs)))

plot(roc_obj, col = "blue", main = "ROC
Curve: Naive Bayes (GABRA3 High vs Low)")
auc_value <- auc(roc_obj)
print(paste("AUC:", round(auc_value, 3)))
```