Assignment 2: Differentially Expressed Genes in HER2-Amplified Breast Cancer

Caolán Maguire

1 Introduction

1.1 Background

HER2-positive breast cancer is a breast cancer that tests positive for a protein called human epidermal growth factor receptor 2 (HER2). This protein promotes the growth of cancer cells.

In about 1 of every 5 breast cancers, the cancer cells have extra copies of the gene that makes the HER2 protein. HER2-positive breast cancers tend to be more aggressive than other types of breast cancer.

ERBB2 gene amplification on chromosome 17 involves an abnormal increase in the number of copies of the ERBB2 (HER2) gene located on the long arm (q arm) of chromosome 17, specifically in the region 17q12. This genetic alteration is a critical driver event in approximately 15–20% of breast cancers and some other solid tumours.

HER2-positive (HER2+) breast cancer is aggressive, growing fast due to excess HER2 protein, often having a less favourable prognosis than Luminal (hormone-driven) types, especially without targeted therapy, but prognosis has dramatically improved with drugs like trastuzumab (Herceptin) and pertuzumab (Perjeta), making outcomes similar to HR+ disease. Prognosis varies within HER2+, depending on hormone receptor status (HR+/HER2+ vs. HR-/HER2+) and other factors, but generally, it's better than Triple Negative (TNBC) but worse than Luminal A, with HER2-enriched (ER-/PR-/HER2+) subtypes being particularly aggressive but highly responsive to targeted treatment.

1.2 Treatment Landscape

Despite the availability of multiple HER2-targeted therapies (Table 1), clinical efficacy remains limited. Trastuzumab monotherapy in the metastatic setting achieves response rates of only 11–26%, with over 50% of patients either failing to respond initially or developing resistance to trastuzumab treatment. Approximately 70% of patients with HER2-positive breast cancers demonstrate intrinsic or secondary resistance to trastuzumab, substantially limiting therapeutic efficacy. When trastuzumab is effective, the duration of response ranges from only 5 to 9 months, indicating that acquired resistance frequently develops. These challenges underscore the critical need for improved understanding of the molecular mechanisms driving HER2-positive cancers and their resistance pathways to develop more effective therapeutic strategies and overcome treatment failure.

TABLE 1
HER2-targeted therapies used to treat breast cancers

Drug name	**Brand name(s)**	**Used to treat early or metastatic breast cancer?**	**Pill, injection under the skin, or IV drug (given by vein through an IV)?**
Trastuzumab*	Herceptin (IV drug) and Herceptin Hylecta (injection)	Early and metastatic breast cancer	IV drug or injection
Pertuzumab	Perjeta (IV drug) and Phesgo (injection combined with trastuzumab)	Early and metastatic breast cancer	IV drug or injection
Margetuximab	Margenza	Metastatic breast cancer	IV drug
Ado-trastuzumab emtansine (T-DM1)	Kadcyla	Early and metastatic breast cancer	IV drug
Trastuzumab deruxtecan (fam-trastuzumab deruxtecan)	Enhertu	Metastatic breast cancer	IV drug
Tucatinib	Tukysa	Metastatic breast cancer	Pill
Neratinib	Nerlynx	Early and metastatic breast cancer	Pill
Lapatinib	Tykerb	Metastatic breast cancer	Pill

Summary

- Despite these therapies, response rates only 40%
- Resistance develops in many patients
- Need better understanding of molecular mechanisms driving HER2+ cancers

1.3 Why This Research Matters

Understanding the molecular differences between HER2-positive and non-HER2-positive breast cancers is essential for improving patient outcomes. By identifying differentially expressed genes, we can discover new therapeutic targets beyond HER2 itself. This research can also reveal prognostic biomarkers that predict which patients will respond to treatment and which may develop resistance. Since current therapies show variable response rates despite uniform HER2 amplification, comprehensive molecular profiling beyond HER2 status alone is critical for developing personalized treatment strategies and improving clinical outcomes.

1.4 Study Aims

This study had three main aims. First, to identify differentially expressed genes between HER2-amplified and non-amplified breast cancers using TCGA RNA-seq data. Second, to characterize the biological pathways that are altered in HER2-positive tumors through pathway enrichment analysis. Third, to develop a gene expression signature that can stratify patients by survival risk, potentially informing clinical decision-making and treatment intensity.

2 Methods

TCGA breast cancer RNA-seq data (n=1,068 samples: 328 HER2+, 740 non-HER2+) were obtained from cBioPortal. HER2+ status was defined by ERBB2 CNA > 0. Genes were filtered to retain those with $\geq$10 counts in $\geq$70% of samples.

Differential expression analysis was performed using DESeq2 (version 1.40.2) with significance thresholds of adjusted p-value < 0.05 and $|\log 2 \text{ fold change}| > 1$. Multiple testing correction was performed using the Benjamini-Hochberg method.

Variance stabilizing transformation (VST) was applied for principal component analysis and heatmap visualization. The top 10 genes ranked by adjusted p-value were used for heatmap clustering. All analyses were conducted in R version 4.3.2.

3 Results

3.1 Differential Expression Analysis

Analysis of 14,030 genes identified 83 differentially expressed genes between HER2-amplified and non-amplified breast cancers. 36 genes were upregulated and 47 genes were downregulated in HER2+ samples. ERBB2 was the most highly upregulated gene (log2FC=2.85, padj¡0.001), validating the HER2+ molecular classification. Volcano plot visualization (Figure 1) shows clear separation between significant and non-significant genes. Red points indicate genes meeting significance thresholds (padj¡0.05, —log2FC—¿1). The distribution shows both upregulated and downregulated genes, indicating balanced transcriptional changes in HER2+ breast cancers.

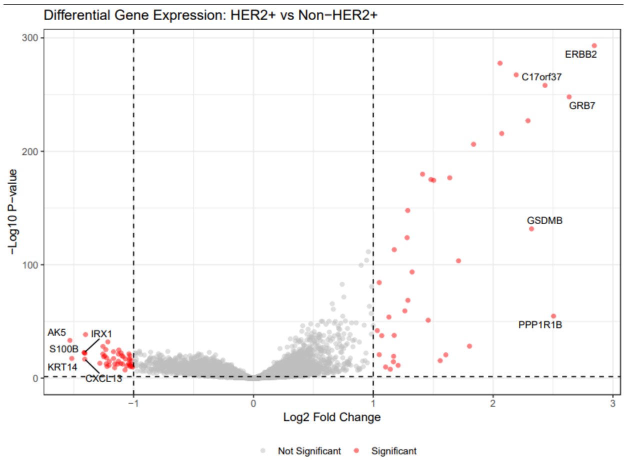

Fig. 1. Volcano plot of differential gene expression between HER2+ and non-HER2+ breast cancer. Red points indicate significantly differentially expressed genes (padj¡0.05, —log2FC—¿1).

3.2 Top 10 Genes

The top 10 differentially expressed genes are shown in Table 2. Among upregulated genes, ERBB2 encodes the HER2 receptor tyrosine kinase, which drives cell proliferation and survival signaling through MAPK and PI3K pathways. GRB7 is an adaptor protein located on chromosome 17q12, frequently co-amplified with ERBB2, and involved in receptor tyrosine kinase signaling. PPP1R1B (protein phosphatase 1 regulatory subunit 1B), also on chromosome 17, regulates dopamine signaling and is part of the ERBB2 amplicon. C17orf37 is an uncharacterized gene on chromosome 17 within the ERBB2 amplification region. GSDMB (gasdermin B) regulates apoptosis and pyroptotic cell death, located in the 17q12 amplicon.

Among downregulated genes, AK5 (adenylate kinase 5) is involved in nucleotide metabolism and cellular energy homeostasis. KRT14 (keratin 14) is a basal epithelial marker; its downregulation indicates HER2+ tumors are not basal-like subtype. S100B is a calcium-binding protein involved in cell cycle regulation and differentiation. IRX1 (Iroquois homeobox 1) is a transcription factor involved in developmental patterning. CXCL13 is a chemokine that recruits B cells and follicular T helper cells; its downregulation may indicate immune evasion.

Notably, the top upregulated genes cluster on chromosome 17q12, reflecting the ERBB2 amplicon. Downregulated genes include basal markers (KRT14) and immune-related factors (CXCL13), suggesting HER2+ tumors have distinct molecular features from basal-like breast cancers and may employ immune evasion mechanisms.

3.3 Pathway Enrichment

Gene ontology enrichment analysis of upregulated genes revealed 12 significantly enriched pathways, predominantly related to transcriptional regulation and cell signaling. These included positive regulation of transcription elongation by RNA polymerase II, regulation of protein kinase activity, regulation of kinase activity, peptidyl-tyrosine phosphorylation, and peptidyl-tyrosine modification. The

TABLE 2
Top 10 Differentially Expressed Genes

Gene	log2FC	adj. p	Function
ERBB2	2.85	¡0.001	RTK
GRB7	2.64	¡0.001	Adaptor
PPP1R1B	2.50	¡0.001	Regulator
C17orf37	2.43	¡0.001	Chr17
GSDMB	2.32	¡0.001	Apoptosis
AK5	-1.53	¡0.001	Kinase
KRT14	-1.52	¡0.001	Basal
S100B	-1.41	¡0.001	Binding
IRX1	-1.41	¡0.001	Homeobox
CXCL13	-1.41	¡0.001	Chemokine

enrichment of kinase activity regulation pathways is consistent with ERBB2's role as a receptor tyrosine kinase, while transcription elongation pathways reflect the presence of transcriptional regulators MED1, MED24, and CDK12 within the chromosome 17 amplicon. These findings indicate that HER2+ tumors exhibit enhanced proliferative signaling and transcriptional activity, consistent with their aggressive clinical phenotype.

In contrast, 142 pathways were enriched among downregulated genes, primarily involving epithelial differentiation and tissue development. Major categories included intermediate filament organization, keratinocyte differentiation, epidermal cell differentiation, epidermis development, keratinization, skin development, urogenital system development, ossification, nephron development, and nervous system myelination. The strong enrichment of keratinocyte and epidermal differentiation pathways, along with multiple keratin genes (KRT5, KRT14, KRT16, KRT17, KRT6B), confirms that HER2+ breast cancers lack basal epithelial characteristics. Additional enriched pathways included skeletal system morphogenesis, proximal/distal pattern formation, and multiple developmental patterning processes involving HOX transcription factors.

The suppression of epithelial differentiation pathways suggests HER2+ tumors maintain a less differentiated, more proliferative state. Downregulation of intermediate filament and cytoskeletal organization pathways may facilitate cellular plasticity and migration. The loss of tissue-specific differentiation programs, including urogenital development and neural myelination pathways, indicates dedifferentiation characteristic of malignant transformation. Notably, the enrichment of HOX gene-related developmental pathways among downregulated genes suggests disruption of normal developmental programs that typically maintain epithelial tissue architecture.

Collectively, pathway analysis reveals HER2+ breast cancers are characterized by increased transcriptional activity and kinase signaling coupled with loss of epithelial differentiation and structural organization. This molecular profile explains both the aggressive proliferative tumor phenotype and the loss of normal tissue architecture. The coordinate upregulation of transcription elongation machinery alongside ERBB2 amplification may represent a mechanism for sustained oncogenic signaling, while suppression of differentiation pathways enables continued proliferation characteristic of HER2+ disease.

3.4 PCA

Principal component analysis revealed PC1 explains 21.5% of variance and PC2 explains 8.5% of variance in gene expression (Figure 2). While HER2+ and non-HER2+ samples showed general separation along PC1, there was notable overlap between groups, particularly in the central region of the plot. This overlap indicates molecular heterogeneity within HER2+ and non-HER2+ subtypes. HER2 amplification status is not the sole determinant of gene expression patterns. Some HER2+ samples cluster near non-HER2+ samples, suggesting additional molecular features beyond HER2 status drive transcriptional variation. This heterogeneity may explain variable treatment responses observed among HER2+ patients.

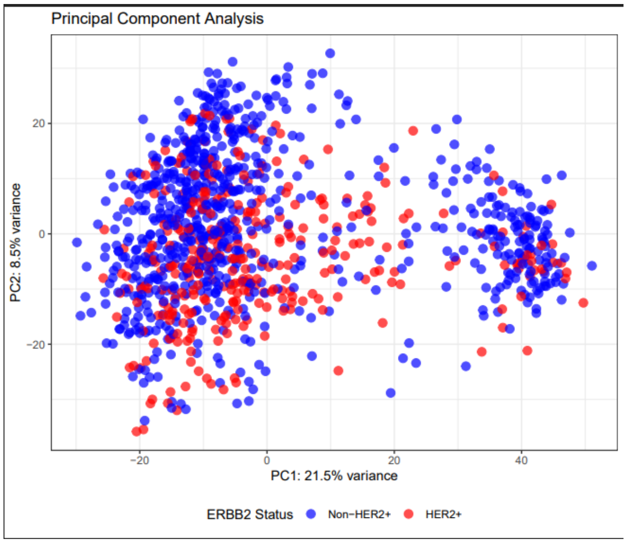

Fig. 2. Principal component analysis of gene expression profiles by HER2 status. PC1 and PC2 explain 21.5% and 8.5% of variance, respectively.

3.5 Heatmap

Hierarchical clustering of the top 50 most significant differentially expressed genes revealed distinct expression patterns between HER2+ and non-HER2+ samples (Figure 3). The heatmap displays z-score normalized expression values, with red indicating high expression and blue indicating low expression. Samples largely clustered according to HER2 status, with HER2+ samples (red annotation bar) forming a distinct group on the right side of the heatmap characterized by high expression of chromosome 17 amplicon genes. While some heterogeneity within groups was evident, the overall clustering pattern demonstrates clear molecular distinction between HER2+ and non-HER2+ breast cancers. Genes formed distinct co-expression modules, with upregulated genes (ERBB2, GRB7, PPP1R1B, C17orf37, GSDMB) showing coordinated high expression in HER2+ samples. This visualization confirms that differential expression analysis identified biologically meaningful gene expression patterns.

3.6 Survival Analysis

Lasso-penalized Cox proportional hazards regression was initially attempted to develop a prognostic gene signature.

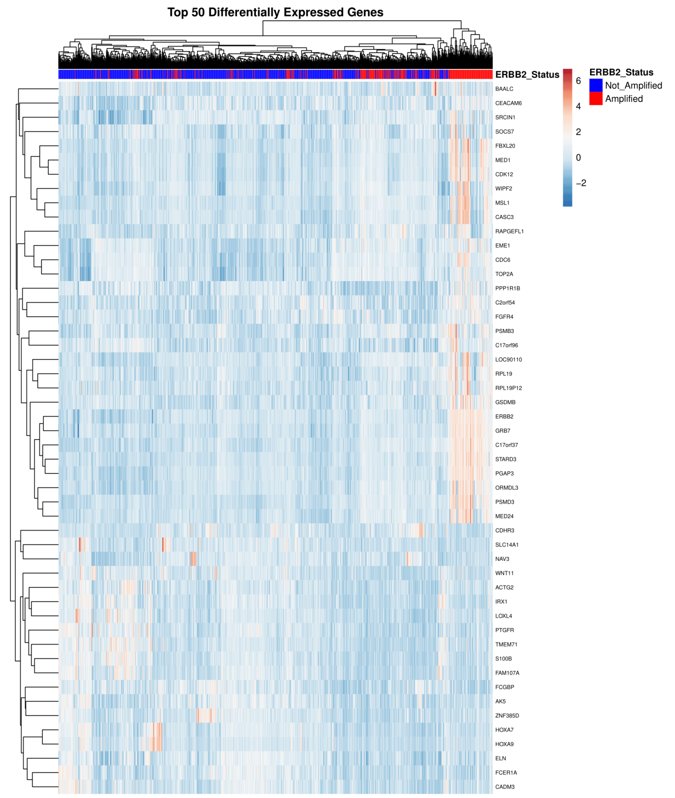

Fig. 3. Heatmap of top 10 differentially expressed genes showing hierarchical clustering of samples and genes.

Cross-validation selected lambda=0.041, however this resulted in complete regularization with 0 genes retained. As an alternative approach, the top 20 most variable differentially expressed genes were selected for Cox regression modeling. These genes included: LTF, DHRS2, CEACAM6, KRT5, KRT14, KRT17, CEACAM5, PPP1R1B, SFRP1, KRT6B, KRT16, SOX10, CXCL13, COL17A1, PI15, CXCL17, DSC3, GLYATL2, ACTG2, and C7.

The Cox model generated risk scores for each patient based on weighted gene expression levels. Patients were stratified into high-risk and low-risk groups using median risk score as cutoff.

Kaplan-Meier analysis revealed significant survival differences between risk groups (log-rank p¡0.0001, Figure 4). Patients were stratified into high-risk (n=527) and low-risk (n=528) groups. Curve separation became evident at approximately 2-4 years. High-risk patients showed worse overall survival, with 5-year survival rates of approximately 90% for high-risk vs 87% for low-risk patients. These results demonstrate that the gene expression signature successfully stratifies HER2+ patients by survival outcome, suggesting potential clinical utility for prognostic assessment.

4 Discussion

This study identified 83 differentially expressed genes between HER2-amplified and non-amplified breast cancers using TCGA RNA-seq data from 1,068 patients (328 HER2+, 740 non-HER2+). Pathway enrichment revealed HER2+ tumors exhibit enhanced transcriptional regulation and kinase signaling alongside suppression of epithelial differentiation, tissue development, and structural organization pathways.

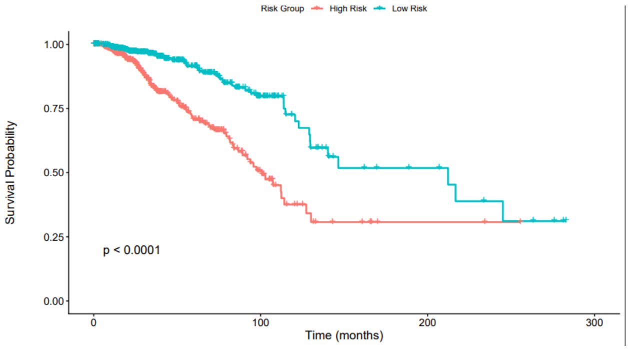

Fig. 4. Kaplan-Meier survival curves for high-risk and low-risk groups. Log-rank p¡0.0001 indicates significant survival difference between groups.

A gene expression signature was developed using the top 20 most variable differentially expressed genes after Lasso Cox regression failed to select genes (lambda=0.041), with significant prognostic value (p¡0.0001).

These findings provide molecular insights into HER2+ breast cancer biology and highlight genes for potential therapeutic targeting.

ERBB2 was the most highly upregulated gene (log2FC=2.85), validating HER2+ molecular classification and confirming the biological relevance of our analysis. Notably, multiple co-amplified genes from the chromosome 17q12 amplicon were also highly upregulated, including GRB7 (log2FC=2.64), PPP1R1B (log2FC=2.50), C17orf37 (log2FC=2.43), and GSDMB (log2FC=2.32). This co-amplification pattern indicates that HER2+ phenotype results from coordinated overexpression of multiple genes within the amplicon, not ERBB2 alone. GRB7, an adaptor protein involved in HER2 signaling, represents a potential additional therapeutic target, as inhibiting both HER2 and GRB7 may provide superior tumor control. The coordinated upregulation of 17q12 genes may contribute to treatment resistance by maintaining signaling pathway activation even when HER2 is blocked.

Downregulated genes included multiple basal keratins (KRT14 with log2FC=-1.52, and others including KRT5 and KRT17 from the survival signature genes), demonstrating that HER2+ breast cancers are molecularly distinct from basal-like (triple-negative) breast cancers. This distinction is clinically important as basal-like and HER2+ subtypes have different treatment strategies and prognoses. The downregulation of CXCL13 (log2FC=-1.41), a chemokine involved in B cell and T cell recruitment, suggests potential immune evasion mechanisms in HER2+ tumors. Suppression of epithelial differentiation pathways including keratinocyte differentiation, epidermal cell differentiation, and intermediate filament organization may contribute to immune evasion in HER2+ breast cancers. Loss of structural organization and developmental patterning pathways may facilitate metastatic dissemination by reducing structural constraints on tumor cells and disrupting normal tissue architecture.

These changes collectively explain the aggressive clinical behavior characteristic of HER2+ breast cancers.

The gene expression signature successfully stratified patients into high-risk and low-risk groups with significantly different survival outcomes (p¡0.0001). This signature could guide treatment intensity decisions, with high-risk patients potentially benefiting from more aggressive or combination therapies. The modest difference in 5-year survival rates (90% vs 87%) suggests that while the signature has statistical significance, its clinical utility may be enhanced by integration with traditional clinical variables such as tumor stage, grade, and hormone receptor status. Nevertheless, the highly significant p-value demonstrates that gene expression patterns beyond HER2 amplification carry prognostic information and warrant further investigation in prospective clinical trials.

The PCA analysis revealed substantial molecular heterogeneity within HER2+ samples (PC1: 21.5% variance, PC2: 8.5% variance), with notable overlap between HER2+ and non-HER2+ groups. This heterogeneity explains the variable treatment responses observed clinically where only 40% of patients respond to trastuzumab. These findings emphasize the importance of comprehensive molecular profiling beyond HER2 status alone for treatment selection. Future biomarker development should incorporate multiple molecular features to capture the full spectrum of HER2+ tumor biology.

This study has several limitations. First, as an observational analysis of existing data, causation cannot be inferred from gene expression associations. Second, the Lasso Cox regression failed to select genes (lambda=0.041 resulted in 0 genes), requiring use of an alternative variance-based selection approach selecting the top 20 most variable genes, which may not optimally weight genes for prognostic performance. Third, no independent validation cohort was analyzed to confirm signature performance. Fourth, important clinical variables including age, tumor stage, treatment received, and ER/PR status were not incorporated into survival models. Fifth, RNA-seq data reflects bulk tumor tissue and does not capture cellular heterogeneity or tumor microenvironment composition.

Future work should validate the prognostic signature in independent cohorts such as METABRIC or external institutional datasets. Integration of multi-omic data including DNA methylation, somatic mutations, copy number alterations, and proteomics would provide more comprehensive molecular characterization. Functional studies using cell line and animal models are needed to determine which identified genes causally contribute to HER2+ phenotypes and represent viable therapeutic targets. Drug repositioning analyses could identify existing approved drugs targeting downregulated pathways that might be repurposed for HER2+ breast cancer. Single-cell RNA-seq would reveal cellular heterogeneity and identify rare cell populations contributing to treatment resistance.

References

Baselga, J. and Swain, S.M. (2009) 'Novel anticancer targets: revisiting ERBB2 and discovering ERBB3', *Nature Reviews Cancer*, 9(7), pp. 463–475.

Benjamini, Y. and Hochberg, Y. (1995) 'Controlling the false discovery rate: a practical and powerful approach to multiple testing', *Journal of the Royal Statistical Society: Series B*, 57(1), pp. 289–300.

Cancer Genome Atlas Network (2012) 'Comprehensive molecular portraits of human breast tumours', *Nature*, 490(7418), pp. 61–70.

Jegg, A.M. et al. (2012) 'HER2-amplified breast cancer: mechanisms of trastuzumab resistance and novel targeted therapies', *Future Oncology*, 8(9), pp. 1205–1217.

Love, M.I., Huber, W. and Anders, S. (2014) 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome Biology*, 15(12), p.550.

Nahta, R. and Esteva, F.J. (2006) 'HER2 therapy: molecular mechanisms of trastuzumab resistance', *Breast Cancer Research*, 8(6), p.215.

Slamon, D.J. et al. (2001) 'Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2', *New England Journal of Medicine*, 344(11), pp. 783–792.

Wong, A.L. and Lee, S.C. (2012) 'Mechanisms of resistance to trastuzumab and novel therapeutic strategies in HER2-positive breast cancer', *International Journal of Breast Cancer*, 2012, article 415170.

Xu, M. et al. (2022) 'Resistance mechanisms and prospects of trastuzumab', *Frontiers in Oncology*, 12, article 1006429.

Yu, G. et al. (2012) 'clusterProfiler: an R package for comparing biological themes among gene clusters', *OMICS: A Journal of Integrative Biology*, 16(5), pp. 284–287.

GitHub Repository

The GitHub repository contains all analysis scripts, generated figures, intermediate results files, and a README with instructions for reproducing the analysis.

Github repository