UCD School of Medicine
Scoil an Leighis UCD

Assignment 2
Gene Expression Analysis and Interpretation

Biological Principles and Cellular Organization
ANAT40040

Anastasiia Deviataieva

Student Number: 24100519

December 2023

1 Introduction

A major challenge in the most common cancer type among women, breast cancer, arises from the disease's heterogeneous nature. The analysis of gene expression profiles has revealed the presence of at least five distinct types of breast cancer, each exhibiting unique biological properties: Luminal A, Normal-like, Luminal B, HER2-enriched, Triple Negative. (Morra et al. 2021; Arribas et al. 2011; Wu et al. 2022). The standardised diagnostic evaluation of hormone receptors (ER and PR) and HER2, in accordance with international guidelines, is crucial for determining these subtypes.

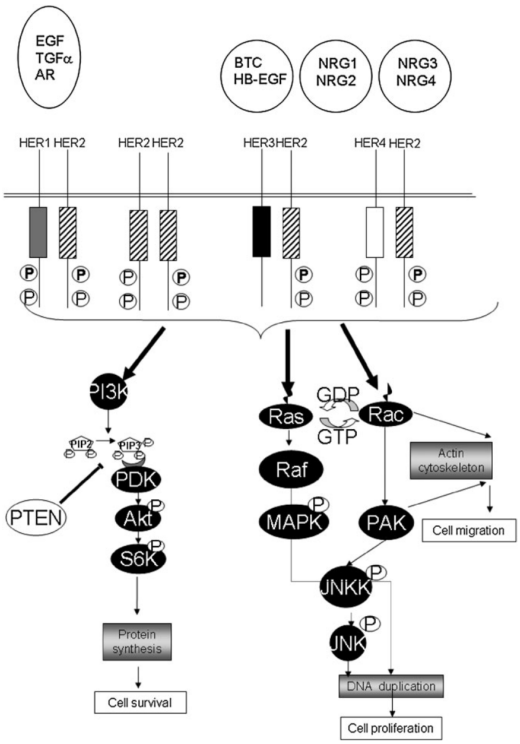

Figure 1. A condensed picture of the main pathways engaged in downstream signalling following HER2 activation. Ligands (top) initiate receptor dimerization, leading to HER2 heterodimers (e.g., HER2-HER3). Receptors are phosphorylated at intracellular tyrosine residues, initiating downstream signaling pathways. PI3K/AKT Pathway: HER2 activates PI3K, which catalyzes the conversion of PIP2 to PIP3. This process recruits AKT to the membrane, where it is activated by PDK1. Once activated, AKT phosphorylates multiple downstream targets, regulating cellular processes such as survival, growth, proliferation, and metabolism. PTEN counteracts PI3K signaling by inhibiting the conversion of PIP2 to PIP3. RAS/MAPK/ERK: HER2 activates RAS, which initiates the MAPK cascade that regulates transcription factors associated with cell proliferation and survival. Rac/PAK/JNK: Rac activates PAK and JNK, which regulate DNA duplication, cell proliferation, and migration through changes in the actin cytoskeleton (important for metastasis).(Valabrega et al. 2007)

ERBB2 (HER2), a proto-oncogene on chromosome 17q21, which belongs to the EGFR, encodes a transmembrane receptor tyrosine kinase involved in cell growth and differentiation.(Raghav and Moasser 2023) HER2-positive cancer represent 11-30% of all breast cancer cases and are defined by the overexpression of the receptor resulting from gene amplification (Schettini et al. 2020), significantly contributes to increased cancer aggressiveness, and poor prognosis.

HER2 activation occurs through dimerization, forming homo- or heterodimers with other family members (e.g., HER1, HER3, HER4). This process phosphorylates tyrosine residues in the cytoplasmic domain, activating downstream signaling pathways, including PI3K/AKT, MAPK/ERK, and Rac/PAK (Fig. 1). HER2 amplification enhances these pathways, promoting cell proliferation and tumor progression. (Arribas et al. 2011)

Targeted therapies including trastuzumab, pertuzumab, and T-DM1 have significantly transformed the handling of HER2-positive breast cancer, leading to enhanced survival rates. Nonetheless, therapeutic resistance and incomplete responses continue to pose considerable challenges, highlighting the necessity of identifying further molecular targets and pathways implicated in HER2-driven tumorigenesis.(Valabrega et al. 2007; Loibl et al. 2021; Schlam and Swain 2021)

The aim of this analysis is to examine RNA-seq data from breast cancer patients, while comparing HER2-amplified and non-amplified tumours. It enables the identification of differentially expressed genes (DEGs), which clarify HER2-dependent mechanisms and facilitate the discovery of potential

biomarkers or therapeutic targets.

2 Methods

Data collection

The data for the analysis were obtained from the cBioPortal for Cancer Genomics platform, which is widely used to study multidimensional oncological genomic data. BRCA (Breast Invasive Carcinoma) TCGA PanCancer Atlas 2018 dataset were used: RNA-Seq data, copy number aberrations data, and clinical data. The analysis had a focus on the human epidermal growth factor receptor 2 (HER2) amplification, which is a distinctive marker of one of the most aggressive subtypes of breast cancer —HER2-positive breast cancer.

Software and packages used

Data analysis was performed in R (v4.4.2) using specialized packages:

- The *DESeq2* package was used to perform differential expression analysis (DEA) between groups (HER2-amplified and non-amplified tumors). This package allowed normalisation of RNA-Seq data, statistical testing, and identification of significant differentially expressed genes (DEGs).
- To visualize the DEGs, the *EnhancedVolcano* package was used to construct a Volcano Plot displaying log2FoldChange and p-value.
- *ReactomePA* and *clusterProfiler* packages were used to perform pathway enrichment analysis and functional annotation of significant genes. These tools allowed us to identify key biological processes and molecular pathways associated with HER2 amplification.
- The *pathview* package was used to integrate gene expression data with molecular pathways, providing a view of their activity in the context of biological processes. The *org.Hs.eg.db* package provided annotation for human genes. It was used to convert gene identifiers (e.g. SYMBOL) to other formats such as ENTREZID, which is necessary for enrichment analysis.
- To visualise the enrichment analysis results (e.g., Reactome), the *enrichplot* package was used to generate dot plots and tree plots.
- The *pheatmap* package was used to construct heatmaps.
- The *ggplot2* package was used to create complex and visually appealing plots, including PCA (Principal Component Analysis) to visualize differences between HER2 groups.
- The *survival* package was used to perform survival analysis, including the construction of Kaplan-Meier curves. The *glmnet* package was used to construct a Cox regression model with Lasso regularization assessing the relationship between gene expression levels and patient survival. This analysis allowed us to identify genes associated with clinical outcomes. To visualize the survival analysis, the *survminer* package was used, which allowed to plot Kaplan-Meier curves with division of patients into risk groups and add statistical significance (p-value) to improve the graphical presentation.

3 Results

3.1 Differential expression analysis HER2 amplified and not amplified

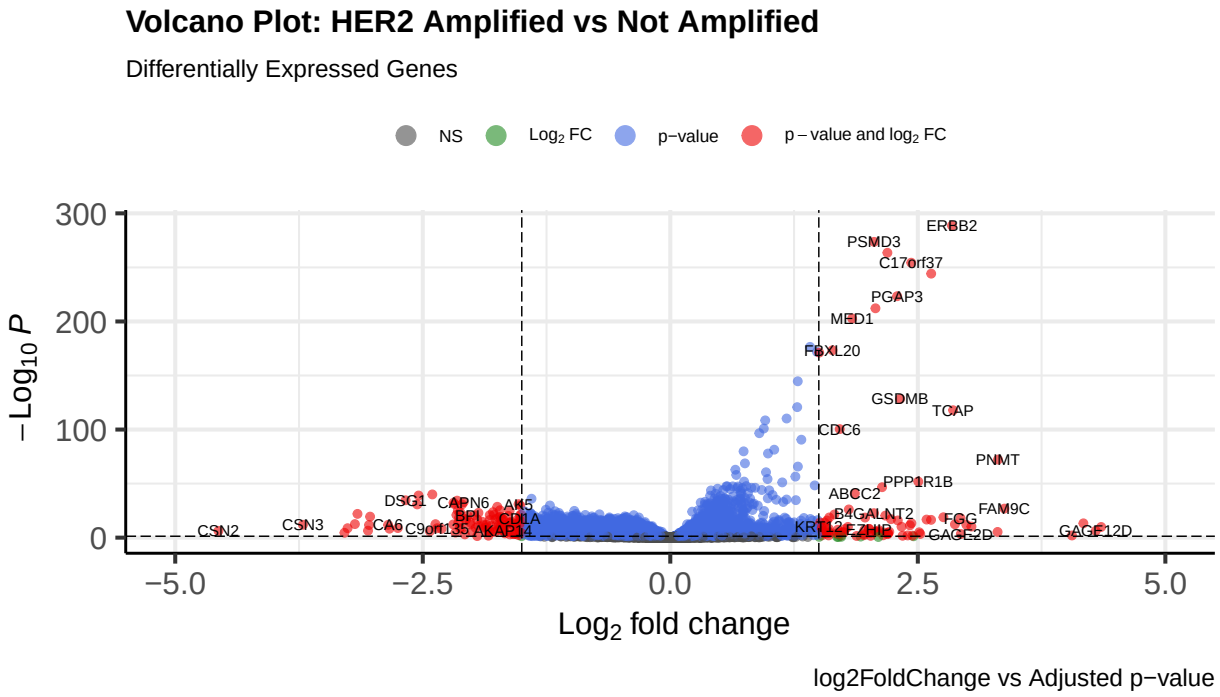

Figure 2. Differentially expressed genes between HER2-amplified and not-amplified groups. The Log2 Fold Change axis shows the degree of gene expression change. Genes to the right of zero (positive log2FoldChange) are more expressed in the HER2-amplified group; to the left, less expressed in the group. Y-axis (-log10 p-value) represents the inverse logarithmic value of the adjusted p-value. The higher the value, the greater the statistical significance of the expression change. The colours of the dots indicate significance by the criteria: red (along two axes), blue (along one of the axes), gray - insignificant changes. Captions correspond to the gene name.

To visualize the results Volcano Plot (Fig. 2) was constructed, which allows intuitive visualization of thousands of genes simultaneously, assessing their statistical significance (p-value) and expression change (fold change), making it extremely useful for interpreting large datasets such as RNA-seq. Volcano Plot uses two key thresholds to identify differentially expressed genes: padj and log2FoldChange. Since we test thousands of genes simultaneously in our analysis, this increases the risk of false positives, and the adjusted p-value takes into account multiple comparisons (Benjamini-Hochberg method), reducing the likelihood of false conclusions. Genes with padj < 0.05 are considered statistically significant, where 0.05 is the statistical standard (5% probability of chance). Log2FoldChange measures the relative change in gene expression levels between groups. Log2FoldChange > 1.5 means that gene expression increases by more than 2.8 times ($2^{1.5} \approx 2.8$), which is biologically significant, highlighting genes with significant expression changes that are most likely associated with biological processes.

18,599 genes were tested for differential expression between the HER2-amplified and not-amplified groups. ERBB2 is significantly overexpressed in the HER2-amplified group. This is an expected result, since HER2 amplification is directly associated with ERBB2 expression.

Also among the overexpressed genes, PSMD3 can be noted, which is involved in protein degradation via the proteasome pathway. Probably, the increase in PSMD3 expression is associated with the adaptation of tumour cells to stress. HER2 activates pathways associated with cell migration and adhesion, which probably leads to an increase in PGAP3. The interaction of HER2 and estrogen receptors enhances the activity of the transcriptional complex, of which MED1 is a component. Among the downregulated genes is CSN2 (Beta-Casein), which is involved in lactation and epithelial function, which may indicate tumor aggressiveness. A decrease in CSN3 (Kappa-Casein), which is involved in the structural stability of milk proteins, indicates suppression of epithelial-related processes, which is also characteristic of HER2-positive tumors.

3.2 Top 10 differentially expressed genes ranked by fold change

Gene	baseMean	log2FoldChange	padj
SPANXA2	2.34148	4.35182	1.33804e-10
GAGE12D	10.26094	4.29677	5.43107e-07
SPANXC	2.42993	4.17245	5.73377e-14
GAGE2B	1.52261	4.05698	9.81590e-03
FAM9C	1.69449	3.37450	1.32362e-27
PNMT	165.00060	3.39059	5.90828e-73
GAGE4	4.45378	3.30569	5.10220e-06
KRT20	3.54389	3.04149	5.60668e-11
TBX10	8.67965	2.99093	1.59234e-12
GAGE2D	4.09091	2.93445	1.12149e-04

Figure 3. Top 10 differentially expressed genes with largest Fold Change

Gene	baseMean	log2FoldChange	padj
CSN2	18.52793	-4.56354	5.32431e-07
CSN3	47.14761	-3.70971	1.25238e-12
LALBA	22.78384	-3.29068	3.18561e-05
LACRT	30.64575	-3.25904	1.84574e-09
SMR3B	40.97939	-3.18522	2.87416e-13
NTS	43.68043	-3.15924	1.04029e-22
CARTPT	307.06416	-3.05279	3.49719e-07
RLBP1	3.44557	-3.04264	1.87511e-12
UCP1	6.86728	-3.03187	3.58603e-20
CA6	10.49175	-2.85011	1.59234e-12

Figure 4. Top 10 differentially expressed genes with smallest Fold Change

Tables (Fig. 4, 3) were constructed that present the 10 genes with the highest and lowest log2FoldChange values between groups. SPANXA2 and SPANXC genes are among the most logarithmic expression changes. Log2FoldChange: $4.35, 4.17$ ($\approx$ 20-fold increase in expression) and padj: 1.33e-10 (high statistical significance), 5.73e-14 (extremely significant). Members of the sperm homolog (SPANX) family play a role in reproduction and cell division. Therefore, they may be associated with mechanisms aimed at increasing cell growth and survival in HER2 Amplified tumours. GAGE genes (GAGE12D, GAGE2B, GAGE4, GAGE2D) (Log2FoldChange:$\approx 2.934.30$ ($8 - 20x$ increase), padj: $\approx$5e-07 to 1e-04 (very significant results)) are frequently expressed in many different types of cancer, and associated with tumour cell adaptation to immune evasion.(Gjerstorff and Ditzel 2008) The breast proteins with the smallest logarithmic change in expression, as in Fig. 2, are CSN2 and CSN3 (Log2FoldChange: -4.56 and -3.71 (21x and 13x reduction). padj: 5.32e-07 and 1.25e-12 (high significance)), the decrease in which is likely related to the loss of the epithelial phenotype characteristic of aggressive tumors with HER2 Amplified. The decrease in LALBA (Lactalbumin Alpha) is associated with the suppression of normal breast function, as it is involved in regulating the production of lactose in the milk.(Pradman K. Qasba and Brew 1997)

3.3 Pathway enrichment

For the enrichment visualisation method, I chose the Reactome Pathway Enrichment Analysis (Fig. 5) method because it focuses on complete biological pathways, providing a deeper understanding of the interactions between molecules. Additionally, it offers greater detail,

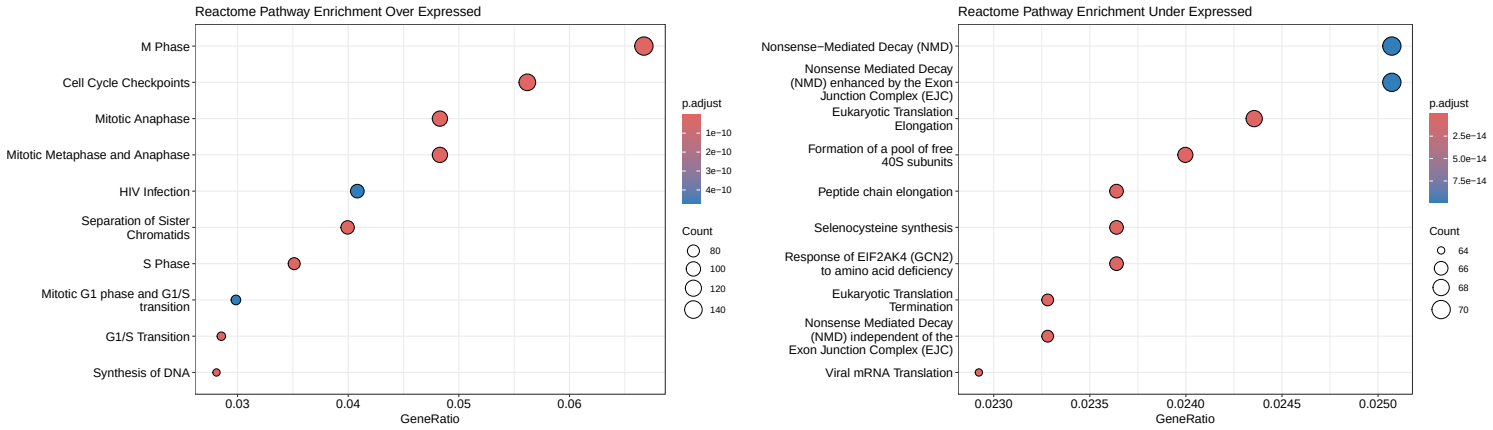

Figure 5. Reactome Pathway Enrichment, ranked by adjusted p-values (p.adjust) and gene ratio, where richer colors indicate more significant pathways. The size of the circles is proportional to the number of genes involved in each pathway. Left: over expressed, Right: under expressed

supports a larger number of species, and is updated more frequently. Unlike GO, which classifies genes based on broad categories (e.g., biological process, molecular function), Reactome focuses on well-defined pathways, allowing for more targeted analysis. While KEGG is widely used, it is sometimes less detailed (mainly focusing on metabolic pathways and some signalling processes); Reactome covers not only metabolism but also high-level processes such as the cell cycle, DNA replication, and signalling cascades.

For genes with increased expression (left panel), enriched pathways indicate increased cell proliferation. This is typical for cells that are actively dividing, such as cancer cells, or cells that are repairing damage. Activation of cell cycle and DNA replication pathways indicates increased cell division, which may be associated with physiological regeneration, pathological hyperplasia, or malignancy.

We can compare these results with Fig. 2, where high expression of genes such as ERBB2, CDC6, FBXL20, and PNMT indicates increased proliferation, which is typical for HER2-positive cancers. For genes with reduced expression, signaling pathways such as Peptide Chain Elongation, Eukaryotic Translation Termination, and Eukaryotic Translation Elongation, which are involved in protein translation, can be noted. Therefore, a decrease in the activity of translation pathways may indicate the suppression of protein synthesis. A decrease in NMD (Nonsense-Mediated Decay) (responsible for the elimination of defective mRNA) indicates a decrease in mRNA quality control.

Figure 6. Principal Component Analysis (PCA) displays differences in gene expression between two groups of samples: group 1: HER2 amplified, group 0: HER2 not amplified.

3.4 Principal Component Analysis (PCA)

In the PCA plot (Fig. 6), PC1 component explains 20% of the variation in the data. Values

along this axis reflect major differences between groups due to the expression of genes associated with HER2 amplification. The PC2 component explains 8% of the variation. It may reflect minor biological differences between samples (e.g., variations in metabolic pathways). Together, the components (28%) account for a meaningful portion of the variation in the dataset, although most of the variance is likely explained by higher-order components. This highlights the complexity of the data and the influence of additional biological factors. Little overlap or blurring between groups may be due to biological variability or confounding effects of other factors. However, there is little separation between groups, indicating molecular differences between HER2-amplified and non-amplified tumours, likely supporting its role in tumour phenotype.

3.5 Heatmap

In the heatmap (Fig. 7), the row (gene) and column (sample) clusters show how genes and samples are grouped based on the similarity of their expression profiles. Genes ERBB2, GRB7, and others in this region (e.g., C17orf37, STARD3) are highly expressed in the HER2-amplified group. This is expected since these genes share the amplified region of the chromosome with HER2. (Kauraniemi and Kallioniemi 2006; Tan et al. 2010) Genes associated with HER2 signaling and cell proliferation (e.g., CDK12, CDC6) show similar expression levels and are clustered together. (Lin et al. 2023)

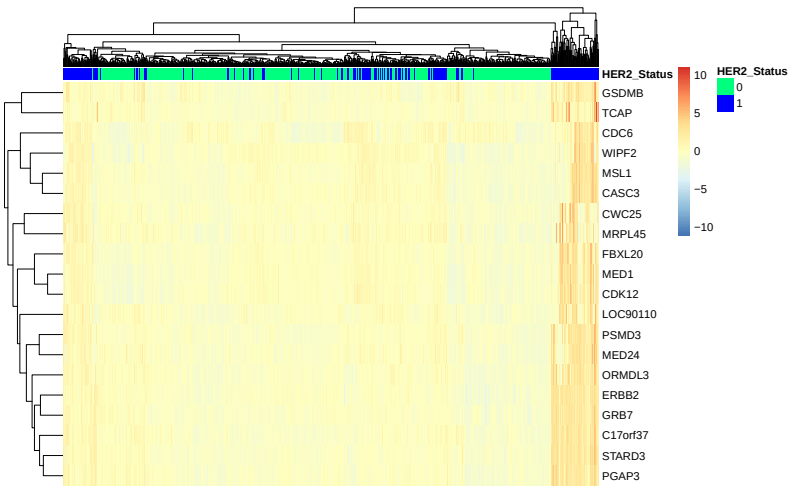

Figure 7. Heatmap of expression levels of the top 20 genes with the most significant changes between samples with different HER2 statuses (amplified: 1, non-amplified: 0) The colour scale shows the level of gene expression, where red: high expression, blue: low.

3.6 Lasso Regularized Cox Regression with DE Genes

The graph (Fig. 8) shows the survival curves for two groups of patients, "High Risk" and "Low Risk", based on the variance stabilised transformed values of the DE genes and the median risk predicted by the model. A p-value < 0.0001 indicates a statistically significant difference between the two groups. The "High Risk" group demonstrates a significantly lower probability of survival over time. These patients are likely to have unfavourable biological characteristics, such as high expression of genes (associated with aggressive tumour growth) and active signalling pathways associated with metastasis or treatment resistance. Patients in this group may require more aggressive forms of therapy, such as targeted therapies (e.g., HER2 inhibitors) or combination therapy. Patients in the low-risk group are likely to have less aggressive biological characteristics. These patients can receive standard treatment with lower risks of toxicity and better long-term outcomes. These results may also show that the expression of genes linked to HER2 status or other clinical factors does affect survival.

4 Discussion

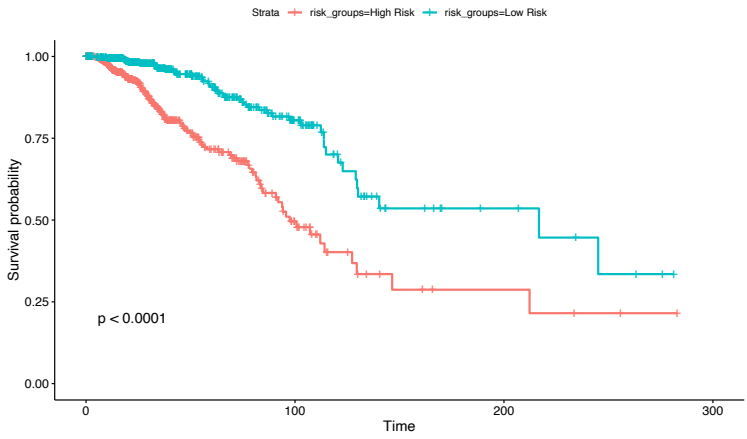

Figure 8. Kaplan-Meier survival curves illustrating the survival probabilities over time for two patient groups: 'High Risk' and 'Low Risk', categorized based on median risk scores predicted by the Lasso Regularized Cox Regression model. Vertical tick marks along the survival curves represent censored data points, where patients were either lost to follow-up or their survival time exceeded the observation period. The solid lines depict the estimated survival probabilities for each group.

The results of this analysis provide information about the molecular differences between HER2-amplified and non-amplified breast cancer tumours. A total of 18,599 genes were analysed to identify genes with differential expression between the groups. HER2 amplification is associated with ERBB2 overexpression, confirming its role in the activation of signaling pathways. Expression of genes involved in stress adaptation (PSMD3) and cell migration (PGAP3), as well as proliferation factors (MED1) is increased. Reduced expression of CSN2 and CSN3 indicates the loss of epithelial functions and aggressiveness of HER2-positive tumours. Also, the mechanisms underlying HER2-associated cancer are highlighted by the results in Tables 4, 3. The genes SPANXA2 and GAGE12D (more than 20-fold increase) are associated with cell growth and immune evasion. CSN2 and CSN3 genes (21-fold decrease) indicate loss of epithelial phenotype and suppression of normal mammary gland functions.

Pathway enrichment analysis showed that overexpressed genes in HER2-amplified tumours are associated with the cell cycle, DNA replication and cell division. This is typical for actively proliferating cancer cells. Reduced gene expression is associated with protein translation and mRNA control pathways, which may indicate suppression of protein synthesis and mRNA quality control mechanisms.

PCA identified two key components explaining 28% of the variation in the data. The components explain a significant part of the variation in the data, reflecting molecular differences between HER2-amplified and non-amplified tumours, which confirms the role of HER2 in the formation of the tumour phenotype. However, it also highlights the complexity of the data and the influence of additional biological factors.

The heat map visualized the high expression of key genes such as ERBB2, GRB7, and STARD3 in the HER2-amplified group, confirming their localization in the amplified chromosomal region. Clusters of genes associated with cell proliferation and HER2 signalling highlight their role in tumour progression.

Survival analysis demonstrated a correlation between gene expression profiles and clinical outcomes of patients. Cox regression model with Lasso regularization identified gene sets associated with high- and low-risk groups, highlighting the prognostic value of molecular markers and the need for targeted therapy and personalized treatment strategies.

Project code link: GitHub

References

Arribas, Joaquín, José Baselga, Kim Pedersen, and Josep Lluís Parra-Palau (Mar. 2011). "p95HER2 and Breast Cancer". In: *Cancer Research* 71.5, pp. 1515–1519. ISSN: 0008-5472. DOI: 10.1158/0008-5472.CAN-10-3795. eprint: https://aacrjournals.org/cancerres/article-pdf/71/5/1515/2660200/1515.pdf. URL: https://doi.org/10.1158/0008-5472.CAN-10-3795.

Gjerstorff, M. F. and H. J. Ditzel (2008). "An overview of the GAGE cancer/testis antigen family with the inclusion of newly identified members". In: *Tissue Antigens* 71.3, pp. 187–192. DOI: https://doi.org/10.1111/j.1399-0039.2007.00997.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1399-0039.2007.00997.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1399-0039.2007.00997.x.

Kauraniemi, P and A Kallioniemi (2006). "Activation of multiple cancer-associated genes at the ERBB2 amplicon in breast cancer". In: *Endocrine-Related Cancer Endocr Relat Cancer* 13.1, pp. 39–49. DOI: 10.1677/erc.1.01147. URL: https://erc.bioscientifica.com/view/journals/erc/13/1/0130039.xml.

Lin, Shanshan, Qingna Jiang, Xiuwang Huang, Jianhua Xu, Lixian Wu, and Yang Liu (2023). "Synthesis of Novel Dual Target Inhibitors of CDK12 and PARP1 and Their Antitumor Activities in HER2-Positive Breast Cancers". In: *ACS Omega* 8.28, pp. 25574–25581. DOI: 10.1021/acsomega.3c02912. eprint: https://doi.org/10.1021/acsomega.3c02912. URL: https://doi.org/10.1021/acsomega.3c02912.

Loibl, Sibylle, Philip Poortmans, Monica Morrow, Carsten Denkert, and Giuseppe Curigliano (2021). "Breast cancer". In: *Lancet (London, England)* 397.10286, pp. 1750–1769. DOI: 10.1016/S0140-6736(20)32381-3. URL: https://doi.org/10.1016/S0140-6736(20)32381-3.

Morra, Anna et al. (Apr. 2021). "Breast Cancer Risk Factors and Survival by Tumor Subtype: Pooled Analyses from the Breast Cancer Association Consortium". In: *Cancer Epidemiology, Biomarkers & Prevention* 30.4, pp. 623–642. ISSN: 1055-9965. DOI: 10.1158/1055-9965.EPI-20-0924. eprint: https://aacrjournals.org/cebp/article-pdf/30/4/623/3100604/623.pdf. URL: https://doi.org/10.1158/1055-9965.EPI-20-0924.

Pradman K. Qasba, Soma Kumar and K. Brew (1997). "Molecular Divergence of Lysozymes and -Lactalbumin". In: *Critical Reviews in Biochemistry and Molecular Biology* 32.4, pp. 255–306. DOI: 10.3109/10409239709082574. eprint: https://doi.org/10.3109/10409239709082574. URL: https://doi.org/10.3109/10409239709082574.

Raghav, Kanwal P.S. and Mark M. Moasser (July 2023). "Molecular Pathways and Mechanisms of HER2 in Cancer Therapy". In: *Clinical Cancer Research* 29.13, pp. 2351–2361. ISSN: 1078-0432. DOI: 10.1158/1078-0432.CCR-22-0283. eprint: https://aacrjournals.org/clincancerres/article-pdf/29/13/2351/3341613/2351.pdf. URL: https://doi.org/10.1158/1078-0432.CCR-22-0283.

Schettini, F., T. Pascual, B. Conte, N. Chic, F. Brasó-Maristany, P. Galván, O. Martínez, B. Adamo, M. Vidal, M. Muñoz, A. Fernández-Martinez, C. Rognoni, G. Griguolo, V. Guarneri, P. F. Conte, M. Locci, J. C. Brase, B. Gonzalez-Farre, P. Villagrasa, S. De Placido, and A. Prat (2020). "HER2-enriched subtype and pathological complete response in HER2-positive breast cancer: A systematic review and meta-analysis". In: *Cancer Treatment Reviews* 84, p. 101965. DOI: 10.1016/j.ctrv.2020.101965. URL: https://doi.org/10.1016/j.ctrv.2020.101965.

Schlam, Ilana and Sandra M. Swain (2021). "HER2-positive breast cancer and tyrosine kinase inhibitors: the time is now". In: *npj Breast Cancer* 7, p. 56. DOI: 10.1038/s41523-021-00265-1. URL: https://doi.org/10.1038/s41523-021-00265-1.

Tan, M., P. Li, K. S. Klos, Y. Ran, G. Li, Z. Wang, W. Chen, J. Lu, X. Zhou, Z. Liu, K. Liang, Y. Sun, C. Kang, R. G. Pestell, and D. Yu (2010). "ErbB2 promotes Src synthesis and stability: novel mechanisms of Src activation that confer breast cancer metastasis". In: *Journal of Biological Chemistry* 285.30, pp. 22686–22695. DOI: 10.1074/jbc.C110.114124.

Valabrega, G., F. Montemurro, and M. Aglietta (2007). "Trastuzumab: mechanism of action, resistance and future perspectives in HER2-overexpressing breast cancer". In: *Annals of Oncology: Official Journal of the European Society for Medical Oncology* 18.6, pp. 977–984. DOI: 10.1093/annonc/mdl475. URL: https://doi.org/10.1093/annonc/mdl475.

Wu, Diana, Lilian U. Thompson, and Elena M. Comelli (2022). "MicroRNAs: A Link between Mammary Gland Development and Breast Cancer". In: *International Journal of Molecular Sciences* 23.24. ISSN: 1422-0067. DOI: 10.3390/ijms232415978. URL: https://www.mdpi.com/1422-0067/23/24/15978.