Assignment (DRAFT)

Tilman Schaefer 23206466

Background

According to the International Agency for Resarch on Cancer, breast cancer is the most common cancer overall, accounting for about 12% of all cancer cases. In approx. 20% of invasive breast cancers the ERBB2 oncogene is overexpressed, which has also been linked to the promotion of breast cancer invasion and metastasis, resulting in poor patient survival (https://www.intechopen.com/chapters/53690).

ERBB2 belongs to the ERBB family of genes that encode a member of the epidermal growth factor (EGF) receptor family of receptor tyrosine kinases (RTKs) [https://www.ncbi.nlm.nih.gov/gene/2064#summary]. Although this protein has no ligand binding domain of its own and therefore cannot bind growth factors, it does bind with other EGF receptors and activates downstream signalling pathways such as MAPK and PI3K/Akt to the effect of promoting cell proliferation and suppressing apoptosis.

While the overexpression of ERBB2 has been established as a reliable biomarker for the diagnosis, treatment and prognosis of breast cancer, many of the underlying processes such as tumor progression and resistance to treatment are still not well understood. It is therefore important to understand the factors that contribute to therapy resistance of ErbB2-positive breast cancer tumors and to identify other genetic or transcriptomic factors in order to identify novel therapeutic strategies to overcome resistance.

In this study I am going to investigate whether and what other genes are overexpressed in patients with ERBB2+ breast cancer. This could provide insights into other contributing factors that promote/supress tumor progression or enhance/inhibit therapy resistance, ultimately leading to a better understanging and the development of alternative therapy targets.

Methods

For this investigation I performed a Differential Gene Expression analysis of patients with BRCA, using the DESeq2 R-package from BiocManager. The patient data was obtained from … and contains a dataset with about 1080 observations. For the analysis I divided the patient's data into two groups, one group with an amplified level of ERBB2 (CNA level > 0) and the

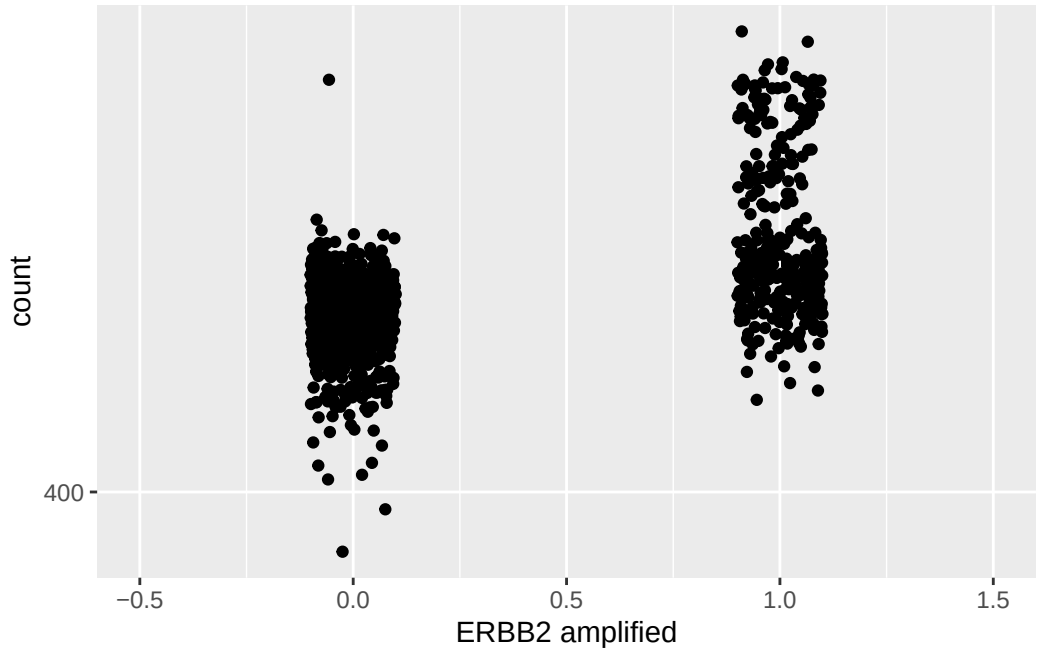
other one with an expression level <= 0. A KEGG Enrichment Analysis was subsequently performed using the clusterProfiler library from BiocManager. Lastly, I performed a Principal Component Analysis.

Results

Data preparation

For data preparation I filtered the RNASeq dataset to only contain read counts of >= 10.

In order to assess the overall quality of the filtered data I evaluated the normalised counts for the two patient groups. For this I plotted the distribution of counts of the ERBB2 gene. For non-amplified ERBB2 patients we can observe a generally lower count compared to ERBB2 amplified patients, together with a smaller distribution.



Top 10 Differentially Expressed Genes Ranked by Fold Change

The aim of differential expression analysis is to discover significant changes in expression levels of genes. The difference in expression levels from a control group can then function as an indicator of a common underlying factor in the development of a disease. It also serves as a pre-requisite for performing a Pathway Enrichment Analysis. In this section I obtained the top 10 genes that were differentially expressed based on the observed log2 fold change. Since

the log2 fold change can be positive (increased expression) or negative (decreased expression), I retrieved the genes with the top absolute value of the log2 fold change and ordered them accordingly. The result is summarised in @tab-top10-results.

	baseMean	log2FoldChange	pvalue	padj
CSN2	18.331467	-4.540222	0.0000000	0.0000003
SMR3B	60.736503	-3.768110	0.0000000	0.0000000
CSN3	47.977244	-3.728425	0.0000000	0.0000000
LALBA	26.227637	-3.495534	0.0000012	0.0000074
GAGE4	4.394711	3.332880	0.0000007	0.0000046
FAM9C	1.673085	3.393388	0.0000000	0.0000000
GAGE2B	1.503136	4.067200	0.0041139	0.0099585
SPANXC	2.410721	4.138504	0.0000000	0.0000000
GAGE12D	10.123610	4.322133	0.0000001	0.0000004
SPANXA2	2.312353	4.359869	0.0000000	0.0000000

Pathway Enrichment Analysis

The DGE analysis performed before allows to subsequently identify more specific groups or categories of genes that together play a role in signalling pathways and intra-cellular processes. @tab-top5-paths shows the result of the pathway enrichment analysis ordered by ascending p-value. From this table we can see that the first 2 pathways, P13K-Akt and JAK-STAT belong to the signal transduction subcategory. The P13K pathway is one of the most important intracellular pathways, which regulates cell growth, survival, metabolism, and angiogenesis. It is also believed to be deregulated in a wide spectrum of human cancers. (https://molecular-cancer.biomedcentral.com/articles/10.1186/s12943-019-0954-x) The JAK-STAT pathway on the other hand plays an important role in the regulation of the immune system (https://biosignaling.biomedcentral.com/articles/10.1186/s12964-017-0177-y).

Table 2: Top 5 paths

	category	subcategory	Description	p.adjust	Count
hsa04151	Environmental Information Processing	Signal transduction	PI3K-Akt signaling pathway	0.0015329	333
hsa04630	Environmental Information Processing	Signal transduction	JAK-STAT signaling pathway	0.0036038	158
hsa04142	Cellular Processes	Transport and catabolism	Lysosome	0.0036038	127
hsa04360	Organismal Systems	Development and regeneration	Axon guidance	0.0042483	172

	category	subcategory	Description	p.adjust	Count
hsa04144	Cellular Processes	Transport and catabolism	Endocytosis	0.007280723	2

PCA

The purpose of conducting a PCA is to reduce the number of dimensions in the data, which facilitates the recognition of patterns and the visualisation of the data. For this I first transformed the data using the Variance stabilizing transformation and then obtained the PCA via the prcomp() function.

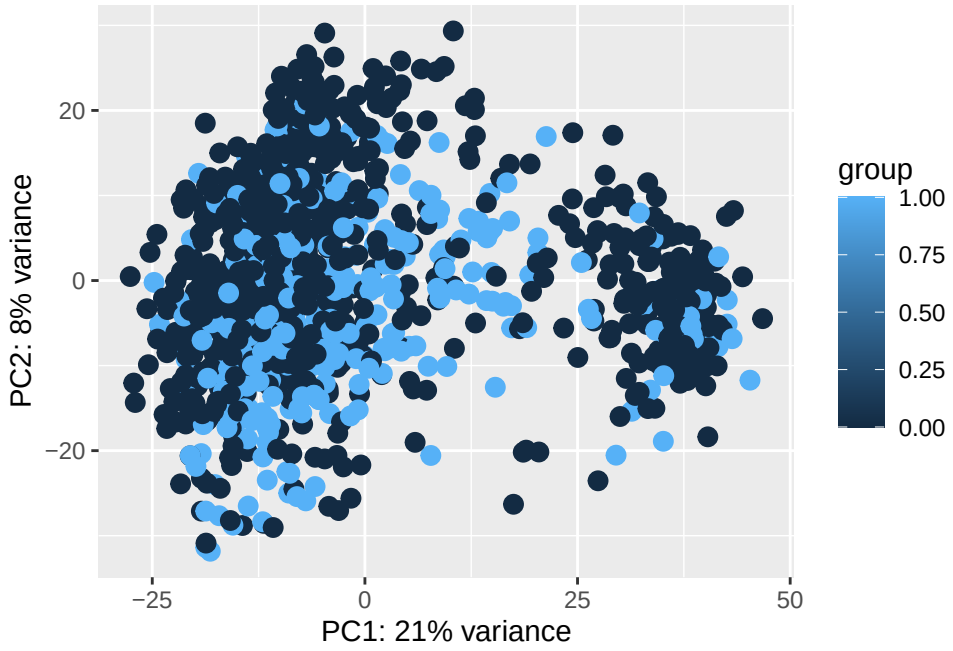


Figure 1: **?(caption)**

The PCA plot is shown in fig-pca. In our case the PCA has not resulted in a clear clustering of data.

Discussion

This study investigated … in patients with BRCA that exhibited an amplified expression of the ERBB2 oncogene. For this I performed a Differential Gene Expression analysis followed by a Pathway Enrichment analysis.

The result obtained from the DGE analysis are difficult to interpret. What I found confusing is the fact that some of the genes (SPANXC and SPANXA2) are overexpressed but belong to a family of genes that are only present in the male reproductive organs (testis). Since the data set comprised of mainly female patients I would not have expected a male-specific gene to be significantly overexpressed. In the literature I could also not find any links between the other genes in that result set and breast cancer.

The PEA on the other hand exhibited the more interesting result, indicating a significant deregulation of 2 major pathways, P13K-Akt and JAK-STAT, which both play a role in cell growth, survival, and the immune system. Given the aggressive nature of BRCA in patients with amplified ERBB2 gene expression, it is therefore plausible that these affected pathways contribute to the overall promotion of breast cancer invasion and metastasis.

References

TO BE DONE