Gene Expression Analysis and Interpretation[1]: ERBB2+

[1] Github: https://github.com/evqizig/Gene-expression-analysis-and-interpretation-ERBB2

Introduction

Background

Breast cancer has become the most common cause of cancer deaths and the fifth leading cause of cancer deaths in women. In the United States, there were an estimated 268,670 cases of invasive cancer in 2018, 6% of which presented as stage IV disease[1]. This report analyses a publicly available breast cancer dataset focusing on the human epidermal growth factor receptor 2 amplified (ERBB2+) breast cancer, one of the most aggressive subtypes.

The report uses data from Breast Invasive Carcinoma (TCGA, PanCancer Atlas)[2]. The data analysis assignment is to preprocess the TCGA RNASeq data for breast cancer from cbioportal and identify the differentially expressed genes between ERBB2+ and other breast cancer tumours. Detailed steps include:

1. Differential Expression Analysis HER2 Amplified and Not Amplified
2. Top 10 Differentially Expressed Genes Ranked by Fold Change
3. Pathway Enrichment
4. PCA Plot

Based on the data analysis, this report is to obtain statistical results of differential expression between ERBB2+ and other breast cancer tumours, as well as statistically based characterisation and visualisation of the data, leading to a biological interpretation of the data.

Methods

Differential Expression Analysis

DESeq2 is a differential gene expression analysis method based on the negative binomial distribution[3]. It estimate variance-mean dependence in count data from high-throughput sequencing assays and test for differential expression based on a model using the negative binomial distribution.

Pathway Enrichment

The clusterProfiler is a universal enrichment tool for interpreting omics data[4]. This package supports functional characteristics of both coding and non-coding genomics data for thousands of species with up-to-date gene annotation. This report completes KEGG pathway enrichment based on this package.

Principal component analysis

ncipal component analysis (PCA) is a technique for reducing the dimensionality of such datasets, increasing interpretability but at the same time minimizing information loss[5]. It does so by creating new uncorrelated variablesthat successively maximize variance.

Results

1. Differential Expression Analysis HER2 Amplified and Not Amplified

```
out of 17767 with nonzero total read count
adjusted p-value < 0.1
LFC > 0 (up)       : 594, 3.3%
LFC < 0 (down)     : 167, 0.94%
outliers [1]       : 0, 0%
low counts [2]     : 0, 0%
```

The results of the differential expression analysis shows that of the 17,767 genes with non-zero total reads, 594 genes (3.3%) are up-regulated (LFC > 0) and 167 genes (0.94%) are down-regulated (LFC < 0). The adjusted p-value is less than 0.1, indicating that the genes in the dataset differs statistically significantly between HER2 amplified and unamplified samples. No outliers or low count genes were identified in this analysis, meaning that no extreme anomalies affected the results.

The 594 up-regulated genes in the sample may be positively associated with cancer progression, as HER2 amplification is commonly associated with aggressive tumour behaviour. These genes may be part of pathways that promote cell division, survival and metastasis. In contrast, the expression of 167 down-regulated genes may be suppressed in cancer cells. Based on this result, appropriate gene-targeted

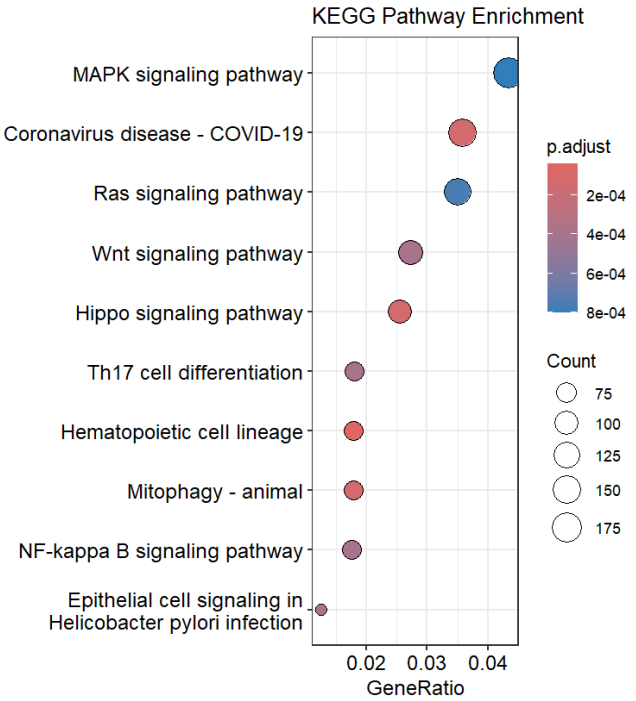
therapies or drugs can be designed to regulate the expression of the corresponding genes to achieve remission and treatment of breast cancer.

2. Top 10 Differentially Expressed Genes Ranked by Fold Change

```
log2 fold change (MLE): ERBB2Amp
Wald test p-value: ERBB2Amp
DataFrame with 10 rows and 6 columns
         baseMean log2FoldChange     lfcSE      stat      pvalue        padj
        <numeric>      <numeric> <numeric> <numeric>   <numeric>   <numeric>
SPANXA2   2.34144        4.35182  0.636214   6.84018 7.90927e-12 1.33874e-10
GAGE12D  10.26037        4.29671  0.796617   5.39369 6.90254e-08 5.43345e-07
SPANXC    2.42982        4.17240  0.524366   7.95704 1.76199e-15 5.74503e-14
GAGE2B    1.52257        4.05693  1.412766   2.87162 4.08371e-03 9.81452e-03
FAM9C     1.69448        3.37450  0.297925  11.32667 9.68212e-30 1.32310e-27
PNMT    164.98858        3.30954  0.179622  18.42506 8.26917e-76 5.91087e-73
GAGE4     4.45391        3.30592  0.670896   4.92762 8.32372e-07 5.09372e-06
KRT20     3.54436        3.04173  0.436004   6.97638 3.02876e-12 5.58982e-11
TBX10     8.67972        2.99093  0.398741   7.50094 6.33600e-14 1.59343e-12
GAGE2D    4.09089        2.93442  0.760270   3.85970 1.13525e-04 4.12083e-04
```

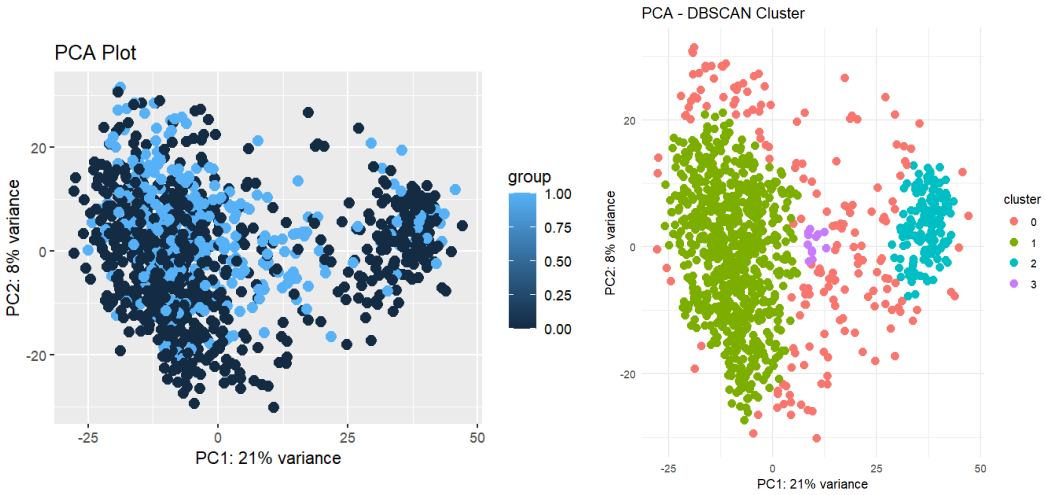
This image shows a table of the top 10 differentially expressed genes (DEGs) with high log2-fold change, indicating that the expression of these genes is most significantly increased in ERBB2-amplified samples, are critical in pathways affected by ERBB2, are strongly correlated with breast cancer progression, and can be a prime target for targeted therapy.

3. Pathway Enrichment



The size of the dots in the graph indicates the number of genes involved, with larger dots representing more genes, and the x-axis indicates the proportion of genes involved in each pathway in the dataset. It can be seen that the MAPK signalling pathway is highly represented and plays a key role in the development of breast cancer, and the enrichment of the COVID-19 pathway may indicate that breast cancer is associated with viral infection and immune response, while the Ras and Wnt signalling pathways are the classical pathways commonly associated with cancer.

4. PCA Plot and Gene Expression Cluster in PCA



Because of the difficulty in estimating the number of clusters, the DBSCAN cluster method was used for this CLUSTER analysis. In the PCA Plot, the data does not show a good separation, while in the PCA after performing the DBSCAN cluster process, two clear clusters can be seen: cluster 1 and cluster 2.

In the context of HER2 amplified and unamplified gene analyses, the two main clusters identified may be distributed to represent the amplified and unamplified portions of HER2-associated samples. Such segregating features may influence prognostic and therapeutic decisions.

Discussion

Differential expression analysis of HER2 amplified and unamplified and datapath enrichment analyses demonstrated the important role that HER2 and related genes may play in breast cancer development. Statistical results and analytical and visualisation tools such as PCA make this gene expression relationship more readable. Just as 594 genes (3.3%) are up-regulated (LFC > 0) and 167 genes (0.94%) are down-regulated, the relationship between breast cancer and the HER2 gene is extensive and involves numerous other related genes and pathways.

References

[1] Arciero, C.A., Guo, Y., Jiang, R., Behera, M., O'Regan, R., Peng, L. & Li, X. 2019, "ER+/HER2+ Breast Cancer Has Different Metastatic Patterns and Better Survival Than ER−/HER2+ Breast Cancer", Clinical breast cancer, vol. 19, no. 4, pp. 236-245.

[2] https://www.cbioportal.org/study/summary?id=brca_tcga_pan_can_atlas_2018

[3] Love, M.I., Huber, W. & Anders, S. 2014, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2", Genome Biology, vol. 15, no. 12, pp. 550-550.

[4] Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., Fu, X., Liu, S., Bo, X. & Yu, G. 2021, "clusterProfiler 4.0: A universal enrichment tool for interpreting omics data", Innovation (New York, NY), vol. 2, no. 3, pp. 100141-100141.

[5] Jolliffe, I.T. & Cadima, J. 2016, "Principal component analysis: a review and recent developments", Philosophical transactions of the Royal Society of London. Series A: Mathematical, physical, and engineering sciences, vol. 374, no. 2065, pp. 20150202-20150202.