Assignment 2 - Gene expression analysis and interpretation

Written Report

Davids Jalisevs

ID – 23205391

Github Link - https://github.com/DavidJalisevs/BioPrincipleAssignment2

Top of Form

1. Background/Introduction

Breast cancer is the most diagnosed cancer in women, exhibits distinct subtypes with varying genetic characteristics. Among these subtypes, the human epidermal growth factor receptor 2 amplified (ERBB2+) breast cancer stands out due to its aggressive nature. Despite the availability of targeted therapies, the response rate remains around 40%. The fundamental question driving this analysis is: What molecular characteristics differentiate ERBB2+ breast cancer from other subtypes?

The rationale for addressing this question lies in the clinical significance of ERBB2+ breast cancer. Although targeted therapeutic interventions exist, the limited response rate emphasizes the need for a deeper exploration of the molecular landscape. Identifying specific genetic aberrations associated with ERBB2+ tumors can lead to the discovery of potential biomarkers, enabling more effective and personalized treatment strategies. By deciphering the molecular intricacies underlying ERBB2+ breast cancer aggressiveness, researchers aim to contribute valuable insights that may ultimately improve patient outcomes.

The analysis leverages publicly available TCGA RNASeq data, obtained from the cbioportal platform, to conduct a comprehensive investigation. The focus is on identifying differentially expressed genes between ERBB2+ and other breast cancer subtypes. This approach aims to uncover key molecular players and pathways contributing to the aggressiveness of ERBB2+ breast cancer, providing a foundation for future research and therapeutic advancements.

The problem statement revolves around understanding the unique molecular features of ERBB2+ breast cancer and the rationale lies in the pursuit of improving treatment strategies for this aggressive subtype. The analysis utilizes advanced computational methods to unravel the complexities of gene expression patterns associated with ERBB2+ tumors.

Methods (draft)

In conducting this comprehensive analysis of breast cancer data, In order to extract useful information from the given dataset, a set of organized methods is included in the approach. To initiate this process the dataset will be retrieved from the cbioportal platform, using following link <https://www.cbioportal.org/study/summary?id=brca_tcga_pan_can_atlas_2018>.

The obtained dataset, comprising RNASeq, Patient Data, and Copy Number Aberrations Data files, will then be extracted after untarring the folder.

In the next step, the R language will be used as the primary tool for data processing and analysis. The RNASeq, Patient Data, and Copy Number Aberrations Data files are read into the R environment. To ensure better integration and as stated in instructions patient IDs from the RNASeq data are matched with corresponding IDs in the Copy Number Aberrations (CNA) data and Patient Data.

For biomarker identification, a metadata column reflecting ERBB2+ status is created, using the CNA levels (specifically, values greater than 0 indicating amplification). This metadata will serve as a crucial basis for subsequent analyses focused on this breast cancer subtype.

**Differential Expression Analysis**

I have used DESeq2 package to normalize RNASeq data and identify differentially expressed genes between ERBB2+ and non-amplified tumors. The metadata column reflecting ERBB2+ status was created based on CNA levels. The analysis resulted in a list of differentially expressed genes, with significant p-values and adjusted p-values.

**Top Differentially Expressed Genes**

From the list, the top 10 differentially expressed genes were selected based on their fold change values. This subset provides insight into genes with the most substantial alterations in expression between ERBB2+ and non-amplified tumors.

**Pathway Enrichment Analysis**

I have performed pathway enrichment analysis using the **clusterProfiler package**, mapping gene symbols to Entrez IDs. This revealed 176 enriched terms, including categories such as "Signal transduction" and "Infectious disease: viral," shedding light on biological pathways associated with ERBB2+ breast cancer.

**VST and Principal Component Analysis (PCA)**

VST Calculation

Variance-stabilized transformation (VST) is computed using the **vst** function from DESeq2. This step minimizes heteroscedasticity in gene expression data, providing a stabilized representation.

PCA Plot

A Principal Component Analysis (PCA) plot is generated using the VST values. The **plotPCA** function from DESeq2, coupled with **ggplot2** for visualization, enables a comprehensive exploration of breast cancer dataset structure.

Throughout this methodology, the utilization of R language and specialized packages such as DESeq2 will ensure a rigorous and reproducible analysis, allowing for a detailed exploration of the molecular characteristics of ERBB2-positive breast cancer.

Results

Breast cancer, a heterogeneous disease, exhibits diverse subtypes with unique genetic features. Among them, the aggressive ERBB2-amplified (ERBB2+) subtype poses clinical challenges despite available targeted therapies. This analysis aims to unravel molecular intricacies, identifying key genes and pathways distinguishing ERBB2+ tumors from others.

**Differential Expression Analysis HER2 Amplified and Not Amplified**

The differential expression analysis revealed 9743 genes with significant expression differences between ERBB2+ and non-amplified tumors (p < 0.05). Among these, EZHIP exhibited the highest log2 fold change (2.01), suggesting its potential role in ERBB2+ aggressiveness. The significant alterations in genes like GTPBP6 and A2LD1 shed light on distinct molecular mechanisms driving ERBB2+ breast cancer.

A screenshot of a computer code

Description automatically generated

P value – 0.05

**Top 10 Differentially Expressed Genes Ranked by Fold Change**

The top 10 differentially expressed genes, ranked by fold change, provide a focused view. CSN2, with a remarkable log2 fold change of -4.56, stands out as a potential biomarker. Meanwhile, genes like SPANXA2, GAGE12D, and SPANXC exhibit substantial upregulation in ERBB2+ tumors, emphasizing their relevance.

A screenshot of a computer code

Description automatically generated

**Pathway Enrichment**

The pathway enrichment analysis explores the functional implications of gene alterations, shedding light on biological processes associated with ERBB2+ breast cancer. Pathway enrichment analysis uncovered 176 enriched terms, providing functional context to gene alterations. Key pathways include "PI3K-Akt signaling" and "MAPK signaling," aligning with known pathways implicated in cancer progression.

A graph with dots and numbers

Description automatically generated

**Plot Description**

The pathway enrichment plot provides a comprehensive overview of 176 terms along the Y-axis. Each dot's position on the X-axis signifies GeneRatio, ranging from 0 to 0.05. The color gradient, from blue to red, corresponds to p.adjust values (ranging from 3e-06 to 1e-06), while dot size represents the count (incrementing by 50).

**Insights**

The results showcase a landscape of enriched pathways, spanning various categories such as "Environmental Information Processing" and "Human Diseases." Notably, the "PI3K-Akt signaling pathway" emerges as a significant finding, with a highly significant p-value and a low false discovery rate (q-value). This pathway plays a pivotal role in regulating cell proliferation, survival, and growth, implicating its potential involvement in the aggressiveness of ERBB2+ breast cancer.

Specific pathways, such as "PI3K-Akt signaling" and "MAPK signaling," exhibit significant **X-axis** shifts, indicating substantial enrichment. The count increase in certain pathways, like "Human Papillomavirus infection," suggests a larger number of involved genes.

The involvement of pathways like "MAPK signaling pathway" and "Calcium signaling pathway" further suggests intricate signaling networks driving ERBB2+ breast cancer progression. These pathways are associated with fundamental cellular functions, and their dysregulation can contribute to the invasive and proliferative characteristics of cancer cells.

Enriched pathways provide potential therapeutic targets for intervention. Targeting the dysregulated pathways identified in this analysis could disrupt the molecular mechanisms driving ERBB2+ breast cancer, potentially leading to more effective and personalized treatment strategies for patients with this aggressive subtype.

**A screen shot of a computer screen

Description automatically generatedPCA PLOT**

The Principal Component Analysis (PCA) plot serves as a crucial visualization tool to decipher the molecular landscape of ERBB2+ breast cancer. Two main clusters emerge, suggesting potential molecular subgroups within breast cancer. Notably, the smaller cluster in the right quadrant, enriched with black dots, hints at a distinct gene expression pattern associated with ERBB2+ status.

**Larger Cluster (Left):**

* **Location:** Spans from -25 to 12 on the X-axis and -20 to 20 on the Y-axis.
* **Composition:** Exhibits a mix of black and blue dots.
* **Interpretation:** The coexistence of black and blue dots suggests a shared gene expression pattern between ERBB2+ and non-amplified tumors. This cluster implies commonalities in molecular profiles, potentially capturing genes critical for both ERBB2+ and non-amplified breast cancers.

**Smaller Cluster (Right):**

* **Location:** From X 25 to around 40 and Y from 10 to around 15.
* **Composition:** Predominantly comprises black dots.
* **Interpretation:** The dominance of black dots in this smaller cluster signifies a distinct gene expression pattern primarily associated with ERBB2+ status. While smaller in size, this cluster suggests a more homogenous molecular profile unique to ERBB2+ tumors.

The spatial distribution of clusters provides important insights into the heterogeneity of ERBB2+ breast cancer. The larger cluster indicates shared molecular features, potentially reflecting underlying biological processes common to both ERBB2+ and non-amplified subtypes. In comparison the smaller cluster highlights specific gene expression patterns exclusive to ERBB2+, underscoring the uniqueness of this aggressive breast cancer subtype. This interpretation deepens comprehension of the molecular diversity within ERBB2+ breast cancer, providing valuable insights for the development of targeted therapeutic interventions.