Assignment 2 - Gene expression analysis and interpretation

Written Report

Davids Jalisevs

ID - 23205391

Top of Form

1. Background/Introduction (Draft)

Breast cancer is the most prevalent cancer in women, exhibits diverse subtypes characterized by distinct genetic aberrations. Within this spectrum the human epidermal growth factor receptor 2 amplified (ERBB2+) breast cancer subtype stands out as particularly aggressive. Despite the availability of targeted therapies, the response rate to these interventions remains at approximately 40%. One of the questions: What molecular characteristics distinguish ERBB2+ breast cancer from other subtypes?

The problem stems from the need to solve the molecular intricacies contributing to the aggressiveness of ERBB2+ breast cancer. The specific objective is to identify differentially expressed genes that play a pivotal role in distinguishing ERBB2+ tumors from other breast cancer subtypes. Leveraging TCGA RNASeq data and employing computational tools, this analysis seeks comprehensive insights into the genetic landscape, aiming to uncover potential biomarkers and shed light on the biological pathways that drive the aggressiveness of ERBB2+ breast cancer.

The rationale for undertaking this analysis is deeply rooted in the clinical significance of ERBB2+ breast cancer. Despite advancements in targeted therapeutic interventions, the limited response rate emphasizes the need for a more profound exploration of the underlying molecular landscape. By pinpointing differentially expressed genes associated with ERBB2+ tumors, scientists strive to identify potential biomarkers that may serve as targets for more effective treatments. This project has the potential to provide important new understandings of the complex molecular processes that underlie the aggressiveness of ERBB2+ breast cancer.

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 steps, 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 will be read into the R environment. To ensure better integration and as stated in instructions patient IDs from the RNASeq data will be matched with corresponding IDs in the Copy Number Aberrations (CNA) data and Patient Data.

For biomarker identification, a metadata column reflecting ERBB2+ status will be 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.

Normalization of the RNASeq data will be performed using the DESeq2 package, ensuring accurate representation of gene expression levels. Subsequently, a differential expression analysis will be used to identify genes exhibiting significant expression differences between HER2-amplified and non-amplified breast cancer tumors.

Pathway enrichment analysis will then be carried out to elucidate the biological pathways associated with HER2 amplification, employing gene sets derived from the differential expression analysis.

variance-stabilized transformation (VST) will be applied to the expression values using the DESeq2 package, providing a more stable representation of gene expression. The resulting VST values will undergo Principal Component Analysis (PCA), assisting in the creation of a picture of the data structure and possible clusters.

Then optional analyses will be pursued for additional marks. Cluster analysis, utilizing the VST values, will reveal potential subgroups within the breast cancer dataset. Additionally, a Cox regression model with differentially expressed genes will be generated, and lasso cross-validation will be employed to identify a set of genes predicting survival most effectively.

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 (draft)

The results section will provide an overview of the outcomes derived from the analyses performed on the breast cancer dataset. Using the outlined methods, each (sub)analysis contributes to unraveling the molecular intricacies of ERBB2+ breast cancer.

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

To elucidate the distinct gene expression patterns associated with ERBB2+ breast cancer, a differential expression analysis will be conducted between HER2-amplified and non-amplified tumors. This analysis aims to identify genes exhibiting significant expression differences, potentially serving as key markers for the aggressive ERBB2+ subtype.

A screenshot of a computer code

Description automatically generated

P value – 0.05

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

As a pivotal subset of the differential expression analysis, the top 10 differentially expressed genes will be ranked based on their fold change values. This focused approach allows for the identification of genes with the most substantial alterations in expression, shedding light on potential drivers of ERBB2+ breast cancer aggressiveness.

A screenshot of a computer code

Description automatically generated

**Pathway Enrichment**

Pathway enrichment analysis will be performed to unveil the biological pathways associated with the differentially expressed genes. By exploring the enriched pathways, we aim to gain insights into the functional implications of the identified gene alterations, providing a broader context for the observed molecular changes.

**PCA Plot**

A Principal Component Analysis (PCA) plot will be generated using the variance-stabilized transformed (VST) expression values. This visualization technique aids in understanding the overall structure of the dataset, revealing potential clusters or patterns that may correlate with ERBB2 amplification. Each data point on the PCA plot represents a sample, allowing for an intuitive interpretation of the underlying gene expression variation.

A screen shot of a computer screen

Description automatically generated

**Key Figures and Tables**

A screenshot of a computer code

Description automatically generatedDifferentially Expressed GenesTop of Form

Top 10 Differentially Expressed Genes Ranked by Fold Change

A screenshot of a computer

Description automatically generated