Identification of biomarkers for breast invasive carcinoma using TCGA data.

Adam Elí Davíðsson, Margrét Dís Stefánsdóttir

Abstract

The accuracy of biomarkers used to predict cancer lethality is not consistent, but they can still provide statistically significant correlations and help distinguish between healthy and diseased tissues. Through gene expression data obtained from The Cancer Genome Atlas (TCGA), we can identify potential biomarkers for breast invasive carcinoma and make prognosis predictions. By using classifiers, clustering, and differential expression analysis, we can determine the most relevant genes for tissue identification and prognosis.

By correlating TCGA gene biomarkers with survival rates in breast-invasive carcinoma patients, we can identify potential biomarkers for future treatment plans. RNA expression profiles clearly differentiate between healthy and cancerous tissue samples, allowing for specific genes to be used as biomarkers for healthy and diseased tissue indications. However, analysis of these genes in survival Kaplan-Meier plots shows varying correlations between gene expression and increased mortality.

This analysis creates new opportunities for further research on genes in breast invasive carcinoma tissue and tumors with different grades that may affect prognosis and could be used as indicators for early aggressive treatment. Therefore, the identification of potential biomarkers for breast invasive carcinoma using TCGA gene expression data has significant implications for cancer diagnosis, prognosis, and treatment.

Introduction

Breast invasive carcinoma is the most common type of breast cancer, accounting for approximately 15% of all cancer deaths among women worldwide. It is a type of cancer characterized by the abnormal growth of cells in the breast tissue that can invade surrounding tissues and spread to other parts of the body [1]. According to the World Health Organization, breast cancer is the most common cancer among women, with an estimated 2.3 million new cases and 685,000 deaths reported in 2020 [7].

Currently, the main therapeutic approach for breast invasive carcinoma is surgery followed by chemotherapy [2]. However, advancements in genetics and gene profiling have led to the discovery and identification of biomarkers that can act as indicators of pharmacologic responses, normal biological processes, or pathogenic processes to a therapeutic intervention [3]. Biomarkers can be identified in any cells of interest and can be used as diagnostic tools to identify patients with abnormal conditions or as indicators of disease prognosis [4]. RNA sequencing-based differentiation gene expression analysis can be used to discover potential cancer biomarkers and therapeutic targets [5]. The hyperexpression of certain genes, such as the EGFR gene, has been correlated with a worse prognosis in breast invasive carcinoma [6]. Our aim in this report is to use RNA sequencing data to identify and cluster differences in gene expression between healthy and cancerous breast tissue and to identify highly expressed genes that may be potential biomarkers for breast invasive carcinoma prognosis.

Results

We used the TCGAbiolinks package to simplify data retrieval and facilitate analysis in R. The GDCquery function allowed us to retrieve data directly from the TCGA database and download it seamlessly, while the GDCprepare function helped us read the data in an R-accessible format by converting it into R data files. Our dataset consisted of 1224 samples, with 113 normal solid tissue samples serving as controls and 1111 breast invasive carcinoma primary tumor samples. The raw expression data contained 60,660 elements and was preprocessed and normalized for analysis.

We performed a principal component analysis (PCA) to assess the differences in gene expression patterns between normal breast tissue and primary breast cancer tissue samples. Our results showed that the two groups are well separated in the first two principal components, indicating distinct gene expression profiles between the normal and cancerous tissue samples (**Fig. 1**). This finding is consistent with previous studies that have reported significant differences in gene expression patterns between normal tissue and breast cancer tissue.

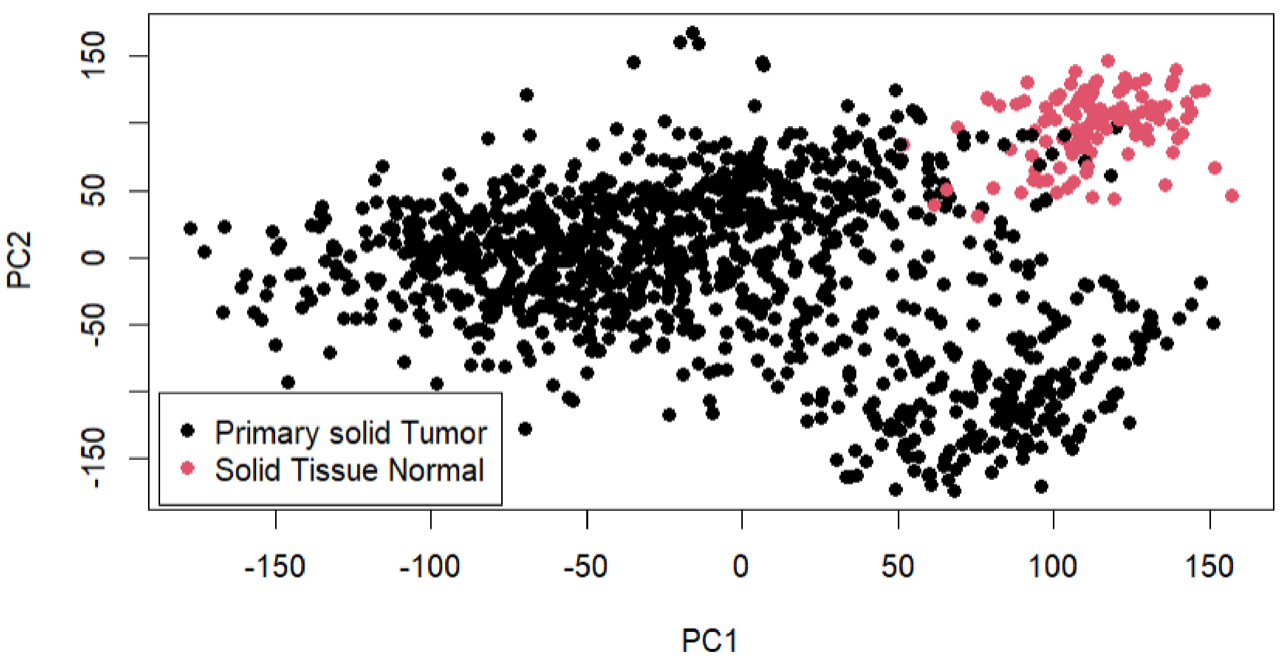

Figure 1 | Separation between healthy normal tissue and unhealthy tumor tissue. (**Black dots**) represent unhealthy tumor tissue while the (**Red dots**) represent healthy normal solid tissue.

By using PCA, we were able to reduce the dimensionality of our data while retaining most of the variation in the dataset. This allowed us to identify the key genes that contribute to the separation between normal and cancerous tissue samples.

The data was classified using differential expression analysis through the limma_pipeline function and an Elastic Net classifier model. The latter model was chosen due to the higher dimensionality of the data compared to the number of samples used. The model selected genes that were most relevant for predicting whether the tissue samples were from diseased or normal healthy tissue. The genes selected were used to determine any overlap between the Elastic Net model and the differential expression analysis. The relevant genes from the Elastic Net model were hierarchically clustered based on their predictive qualities and highlighted in green on the left axis. The top bar indicated unhealthy primary tumor tissue with black and healthy normal tissue with red. The row Z-score was used to visually enhance the gene clusters that had similar trends in expression data. The heatmap showed that the majority of the genes resulted in very low expression in the normal control group, except for a few exceptions, which indicated very high expression in the control and were also selected by both differential expression and Elastic Net. There were no clear patterns of expression seen from the selected genes.

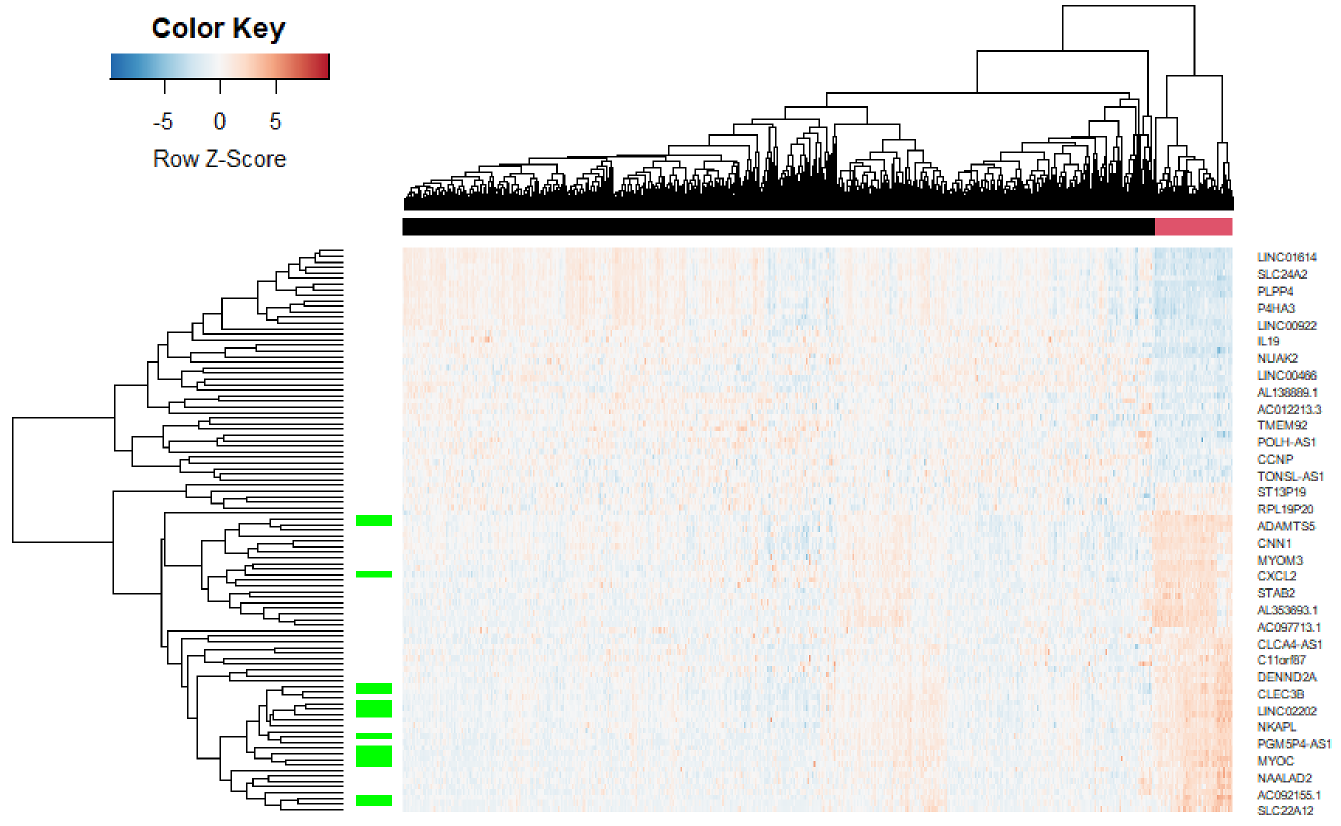

Figure 2 | Heatmap from the Elastic net model and top genes from DE analysis (Black/red bar on top), black indicates unhealthy tumor tissue and red indicates healthy normal solid tissue. (**Green boxes on the left**) indicate highlighted genes.

We used gene expression to divide the patient into groups, to see whether up or down regulation of genes affects survival. We extract the table we had already made, which has differentially expressed genes, ordered by significance. Then we made a for loop to look at the six most differentially expressed genes, one by one. We then visualized the gene expression distribution on the diseased samples versus the healthy samples of the selected gene. We take the expression values and the median of the gene to produce a Kaplan-Meier plot. The top six genes selected were: VEGFD, CA4, LYVE1, MYOC, PAMR1 and CD300LG. (**Fig. 3**). We used the Kaplan-Meier survival plot to see if some of them provide a significant effect on survival. Patients were divided into two groups, up and down regulated. If the patient expression was greater or equal to the median, then it is claimed to be “up-regulated”, meaning it has high gene expression, otherwise it is “down-regulated”, meaning it has low gene expression. From these figures, we were able to see that none of these expressions appeared to make any difference for prognosis.

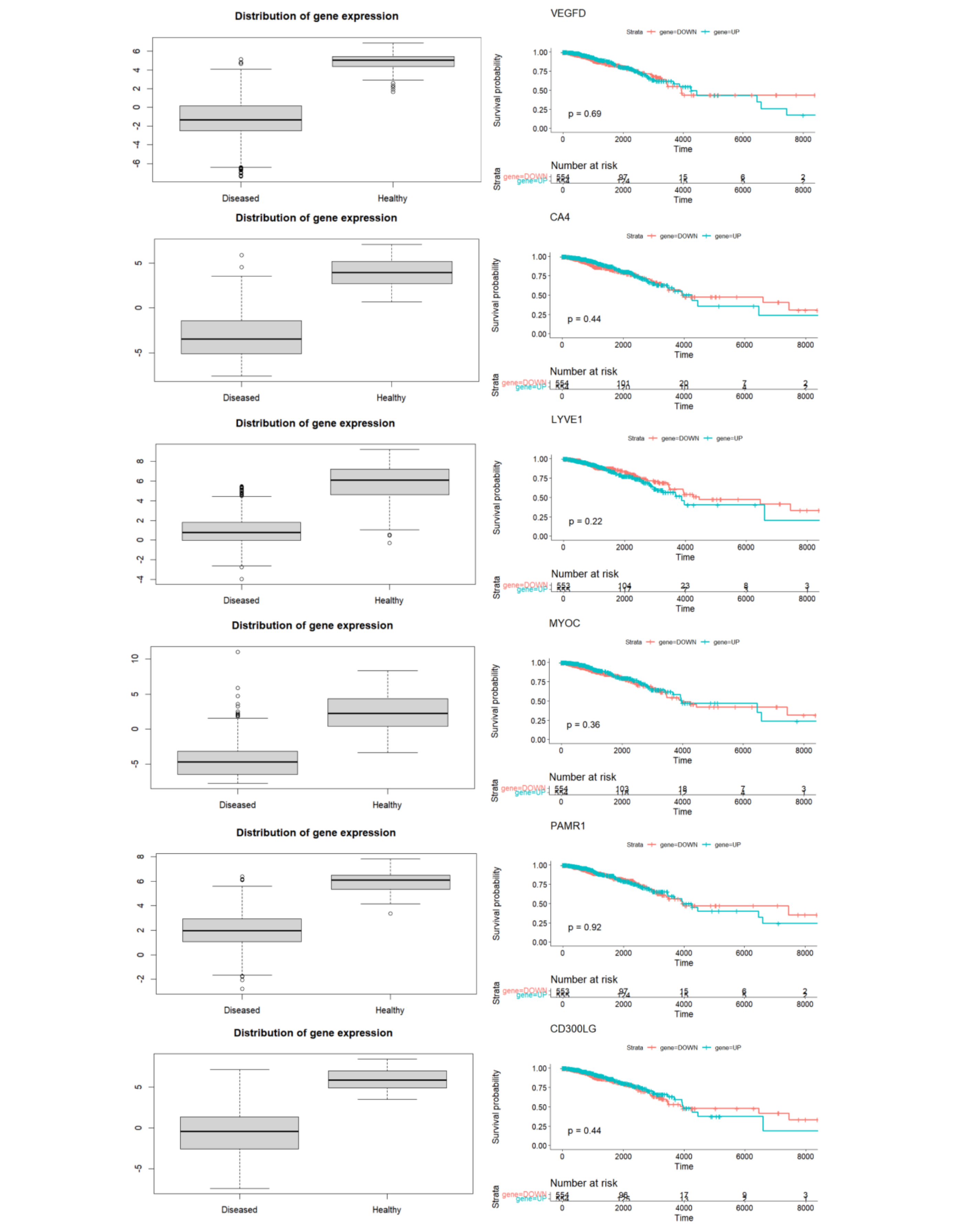

Figure 3 | Kaplan-Meier Survival plot on prognosis for clustered genes
This figure has in whole twelve images, two images per gene. Each gene has image A, on the left, and image B, on the right.
A. The difference in expression on average between diseased and healthy tissue of the chosen gene.
B. Pink stands for increased expression of the gene, and blue decreased expression of the gene.

Discussion

We initially used Python 3, which is a language we are familiar with, but we encountered difficulties acquiring data from the TCGA database. We searched for guides online and found that the most popular programming language for working with gene expression data was R and that it could quite simply be used to fetch and analyze TCGA data, so we switched to R. As programmers who are accustomed to working with Python, switching to R was challenging. The syntax and structure of R code are different from Python, making it difficult for us to write custom code efficiently. Moreover, navigating the many different libraries and packages available in R was complicated. However, after becoming familiar with the language and its unique features, we found R to be a powerful tool for data analysis and visualization. Its functions for statistical analysis and graphing were ultimately powerful, although they required more complexity compared to Python 3.

Another issue that we encountered during our study was related to the choice of cancer type. We initially sought to retrieve melanoma data from the TCGA database, and hence we downloaded TCGA-SKCM data which corresponded to Skin Cutaneous Melanoma. However, we were perplexed during the PCA analysis, as we found only one healthy tissue sample in the dataset, which could have led to misleading results. Therefore, after considering other options, we decided to shift our focus to TCGA-BRCA (Breast Invasive Carcinoma) data, where we found 113 normal tissue samples, which we considered to be more suitable for obtaining a more accurate final outcome.

Methods

Software and Packages

We used R programming language (version 4.2.0) in this analysis, along with several packages, including: TCGAbiolinks, limma, edgeR, glmnet, factoextra, FactoMineR, caret, SummarizedExperiment, gplots, survival, survminer, RColorBrewer, gProfileR, genefilter.

Data Retrieval

We retrieved the gene expression data from the TCGA-BRCA dataset using the GDCquery function of TCGAbiolinks package. We included only the samples that were categorized as “Primary Tumor” or “Solid Tissue Normal”. We then used the GDCprepare function to prepare the dataset for analysis.

Differential Gene Expression Analysis

We performed differential gene expression analysis using the limma package. The limma_pipeline function was used to perform the following steps: Filter genes based on their expression levels using the filter filterByExpr function of edgeR package, normalize the dataset using the trimmed mean of M-values (TMM) method followed by voom function of limma package, fit a linear model to the data using lmFit function, estimate the statistical significance of differential expression using empirical Bayes moderated t-statistics using eBayes function and finally to select the top 100 differentially expressed genes based on their p-values using the topTable function. We used the condition_variable argument of the limma_pipeline function to specify the variable that describes the sample type. We set the reference_group argument to “Solid Tissue Normal” to define the reference group for differential expression analysis. The resulting object of the limma_pipeline function was saved in the limma_res.RDS file.

Principle Component Analysis

We performed principal component analysis (PCA) using the plot_PCA function. The function takes as input the voom object resulting from the limma_pipeline function and the condition variable to be used for grouping the samples. The resulting PCA plot shows the samples in a two-dimensional space defined by the first two principal components. The function also saves the PCA object in the res_pca variable.

Public Github repository containing all the code used for this report:

https://github.com/adamelid/LVF601M-Project

References

1. American Cancer Society. (2022). Breast Cancer. https://www.cancer.org/cancer/breast-cancer.html

2. National Cancer Institute. (2021). Breast Cancer Treatment (PDQ®)–Patient Version. https://www.cancer.gov/types/breast/patient/breast-treatment-pdq

3. European Medicines Agency. (2016). Guideline on the use of pharmacogenetic methodologies in the pharmacokinetic evaluation of medicinal products. https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-use-pharmacogenetic-methodologies-pharmacokinetic-evaluation-medicinal-products_en.pdf

4. Zhang, J., & Shu, C. (2017). Biomarkers of breast cancer. Handbook of Experimental Pharmacology, 249, 93-108. https://doi.org/10.1007/164_2017_25

5. Shyr, Y., & Kim, K. (2014). Overview of RNA sequencing and applications. In RNA sequencing (pp. 1-23). Springer. https://doi.org/10.1007/978-1-4939-1392-4_1

6. Wang, S. Y., & Li, H. (2017). Correlation of EGFR gene mutations with prognosis in patients with breast invasive carcinoma. Oncology Letters, 13(3), 1573-1576. https://doi.org/10.3892/ol.2017.5622

7. World Health Organization. Breast Cancer: Prevention and Control. https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/

8. Isabelle K. T., Viktor Á., Kolbeinn S. H. (2022). TCGA Data Analysis on Cancerous Glioblastoma Survival Rate Using Gene Biomarkers. https://drive.google.com/drive/folders/1_-7U_nHVjIJLinUIbOsUpOnKgAaXbmMa