

Assignment 2 – Gene Expression Analysis and Interpretation

Amritansh Tiwari 25205761

Introduction

Breast cancer is one of the most prevalent cancers in women. This type of cancer has different drivers which are molecular or genomic that drive patient outcomes. One such outcome which is genetic is the amplification of the gene ERBB 2 which codes for a human growth factor known as HER 2. Cancers where HER2 is amplified leads to aggressive tumor growth and subsequent poor clinical outcomes. Since it is a growth factor which is amplified, it leads to enhanced growth of tumor and metastatic properties.

There have been targeted therapies for HER2 associated tumors, such as trastuzumab, clinical studies also show that the outcomes with such therapies were heterogeneous and many patients developed a resistance to the same. Hence it is felt that a molecular understanding of the HER2 amplification is needed to properly assess the therapeutic status.

Gene expression analysis is a useful tool to understand the disease in more detail, taking help from the expression data, such as copy number, differential expression and others. We can combine the copy number data that shows amplification with the clinical patient outcome data, to gain further insights into the phenomenon.

Here we will use The Cancer Genome Atlas known as TCGA, which has a vast and comprehensive catalogue of data relating to cancers patients. We have utilised three types of data in this study : First is the RNA expression data , second is the Copy Number or CNA , third is the Clinical data , which has the patient outcomes. Comparing the EBR2 genes versus the non amplified genes we have applied studies like survival modeling, pathway enrichment, PCA and others to explore the impact of this amplification on real patient outcome.

Methods

For this study we have used the TCGA Breast cancer dataset which was available on the cBioPortal. This is a hub for storing and displaying the genetic, and clinical sample data of various cancer types and patients. Three main files were used : the RNA sequencing profile

was used to assess the expression profile of the amplified gene using transcriptomic data. We have used the copy number CNA file to assess which of the genetic subtypes contribute to the cancer causing outcome and then finally we have used the Clinical patient data to check whether there is any relation between the gene amplification and poor prognosis and subsequent outcomes. We have used the statistical analysis programming language R combined with the suite, R studio for the study, utilising the various libraries made for gene analysis.

The study was conducted using these three datasets matched with the patients' unique identifiers, ie. the TCGA patient identifiers. For this, where each clinical data entry had a sample patient identity, it was converted to the patient level ID. Now that patient identity was solidified, we retained those patients that had data entries in all three files. Now that the data was sorted properly, we could check for which patients showed ERBB2 amplification using the copy number data available. The criteria for amplification was set to be the CNA number greater than 1.

With the data obtained thus, we calculated the gene expression levels which was normalised using the DESeq2. Next step was to check the comparison between tumors that had amplified levels of ERBB2 with the non amplified tumors. For this we used the p values to check for significance using FDR and then ranked the genes based on a log2 change. We used gene expression to identify the patterns in the biology of these tumors, particularly to visualize the various sub types of tumors, at an overall level.

Pathway enrichment analysis was used to identify which processes within the gene were responsible for ERBB2 amplification using differential expression. We then used dimensionality reduction tools like PCA to detect the sample level variations. Finally a heatmap was constructed to assess the expression levels of amplified and non amplified tumors. Using this differential expression gene data with variance stabilised values we built a LASSO regularised cox model to estimate the survival of the patient. This model helps us to select the relevant features on the basis of which we classified the patients as high and low risk. We finally made a Kaplan Meier curve to visualize the two classes of patients.

Results

1. Description of Cohort and status of Amplification :

We used three types of data : RNA, CNA and patient clinical data to combine with TCGA patient identifier and curate a data set of patients having all three types of data, cleaned the data set, found 1068 patients in this process and were able to see amplified and non amplified with the criteria as $CNA > 1$ as amplified.

2. Analysis of differential gene expression :

Differential gene expression analysis was performed between ERBB2 amplified vs non amplified tumors using DESeq2. We set the FDR false discovery rate at 0.05, and

identified 9695 genes were identified as significantly differentially expressed. This includes both upregulated and downregulated genes.

3. What were the top differentially expressed genes :

baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	gene
18.5279328	-4.563540134	0.845434	-5.39787	6.74E-08	5.16E-07	CSN2
2.34148268	4.351821772	0.636642	6.83559	8.17E-12	1.34E-10	SPANXA2
10.2609423	4.296768714	0.79662	5.393752	6.90E-08	5.26E-07	GAGE12D
2.42992757	4.172453979	0.524708	7.95195	1.84E-15	5.75E-14	SPANXC
1.52260623	4.056977805	1.412768	2.871652	0.004083	0.009572	GAGE2B
47.1476103	-3.709707129	0.492304	-7.5354	4.87E-14	1.21E-12	CSN3
1.69449408	3.374502147	0.29813	11.31891	1.06E-29	1.40E-27	FAM9C
165.000596	3.309590338	0.179624	18.42513	8.26E-76	5.71E-73	PNMT
4.45378472	3.305703828	0.661619	4.996387	5.84E-07	3.59E-06	GAGE4
22.7838397	-3.290684712	0.729214	-4.51265	6.40E-06	3.09E-05	LALBA

Table 1 : Table shows the genes and log2fold change values of the differentially expressed genes.

Genes are ranked in order of magnitude of expression levels using log2 fold change. This includes both genes with enhanced and repressed expressions levels in the amplified tumors. Examples of these genes are CSN2,SPANXA2 etc.

4. Pathway enrichment analysis :

Several processes were found to be significantly enriched such as immune signalling, cell proliferation among others. this indicates that these changes affect different functions of tumor behavior rather than a single isolated effect.

5. Principle Components Analysis :

PCA was studied by using variance controlled gene expression values which would show us a global trend relating to the two classes, amplified and non amplified , with amplified being true. This figure shows very little separation.

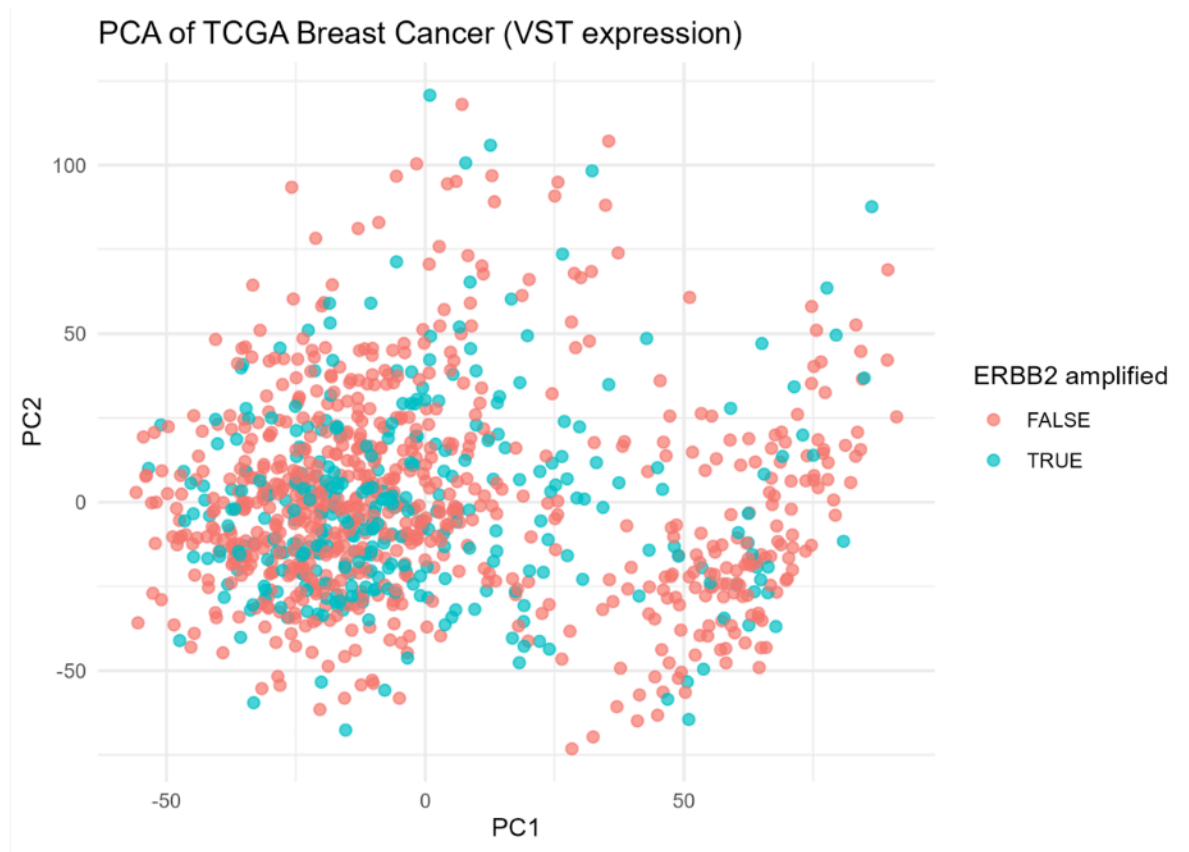


Figure 1 : Figure shows Principle Components Analysis of the two classes : Amplified vs Non amplified as True and False respectively.

6. Heatmap :

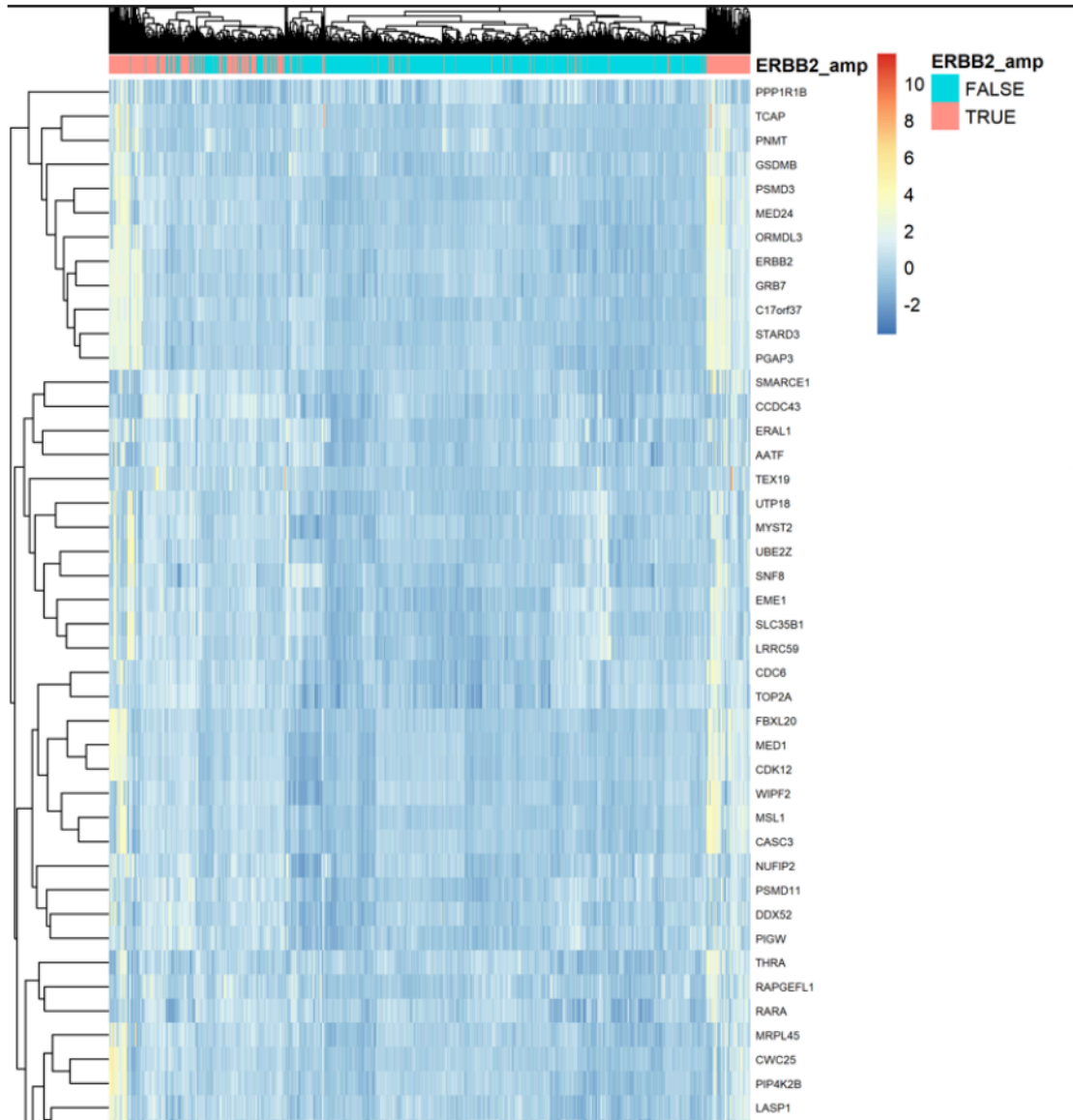


Figure 2 : Figure shows heatmap of about 50 genes that are differentially expressed.

Here we see that there is a difference in the expression pattern of amplified compared to non amplified genes which visually confirms that there are expression level changes involving HER2 amplification.

7. Cox LASSO regression survival analysis :

Survival analysis was performed using the Cox regularised regression based on the expression values obtained. Methods of feature selection like regularisation enabled us to reduce the number of dimensions required to solve the problem and potentially counter

any over fitting as well. We found that we could divide this outcome into two classes, high risk and low risk.

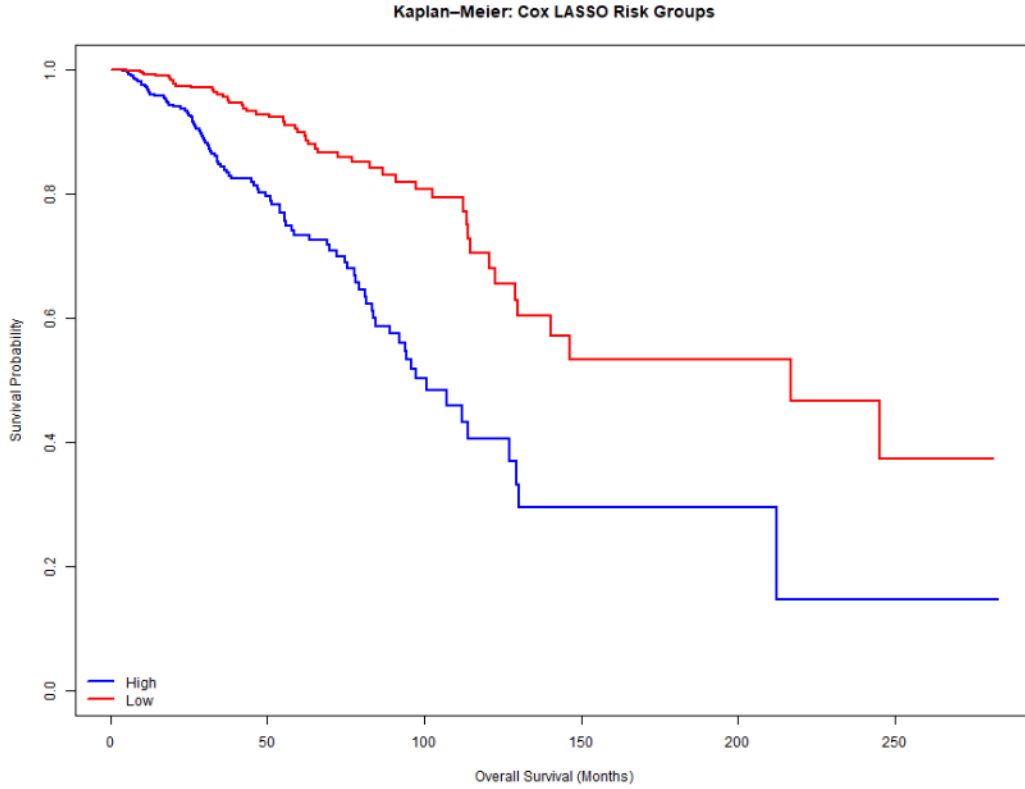


Figure 3 : Figure shows the Kaplan Meier plot for the two classes of patient outcomes.

Discussion

This study was conducted to identify the key differences in the transcriptomic profiles of two types of tumors, ERBB2 amplified and non amplified, using a public dataset available on the cBioPortal. We used three types of data mainly the transcriptomic, copy number and clinical patient sample data to identify the molecular reasons associated with HER2 amplification as well as the relevance of amplification on the outcome of the patient or patient survival. Overall we found that this ERBB2 amplification is linked to changes in the gene expression levels that differentiates the two tumor groups.

We observed an impact of HER2 amplification using differential gene expression analysis. We identified differences in the level of alteration in many such genes in amplified and non amplified tumors. We also found that there was upregulation and downregulation of the genes which means that this amplification has a larger cell wide impact rather than an isolated event. e

used pathway enrichment which also supported this idea, finding that it led to changes in cell signalling, cell proliferation and other processes which match the previously known background of these aggressive tumors in HER2 positive breast cancer studies.

PCA was performed based on stabilised values of expression data which revealed a very low level of separation between amplified and non amplified tumors. This may show that even though we have genetic variation in these two types of tumors it is not the only contributor to the molecular heterogeneity of these tumors. There might be other genetic and molecular factors that influence the behavior of breast cancer tumors.

Survival analysis was performed using the Cox regularised regression based on the expression values obtained. Here we found a set of genes that were responsible for the survival of the patients. Methods of feature selection like regularisation enabled us to reduce the number of dimensions required to solve the problem and potentially counter any over fitting as well. We found that we could divide this outcome into two classes, high risk and low risk. This may show that there is some contribution of ERBB2 amplification in the outcome of the tumor growth, notwithstanding other genetic and molecular factors involved.

There were some challenges which could not be addressed in this study regarding the type of data obtained. This study was done on a single cohort as well as no individual validating dataset was available. Additionally, the survival modeling and differential gene studies were done both on the same data which may cause bias in the model. We can confirm the outcome of this study with external data from different cohorts.

In conclusion this study aims to understand the role of transcriptomic data analysis combined with clinical patient data for ERBB2 associated breast cancer. The results show that HER2 amplification is linked to several gene expression changes as well as its link to patient outcomes. This also shows the value of analysing public cancer data for understanding the role of such individual gene level changes in formulating our hypothesis about the disease.

References

- Ménard, S., Pupa, S.M., Campiglio, M. and Tagliabue, E. (2003). Biologic and therapeutic role of HER2 in cancer. *Oncogene*, [online] 22(42), pp.6570–6578. doi:<https://doi.org/10.1038/sj.onc.1206779>.
- Loibl, S. and Gianni, L. (2017). HER2-positive breast cancer. *The Lancet*, [online] 389(10087), pp.2415–2429. doi:[https://doi.org/10.1016/s0140-6736\(16\)32417-5](https://doi.org/10.1016/s0140-6736(16)32417-5).
- Deng, M., Brägelmann, J., Schultze, J.L. and Perner, S. (2016). Web-TCGA: an online platform for integrated analysis of molecular cancer data sets. *BMC Bioinformatics*, 17(1). doi:<https://doi.org/10.1186/s12859-016-0917-9>.

- Love, M., Anders, S. and Huber, W., 2014. Beginner's guide to using the DESeq2 package. *Genome Biol*, 15(55010.1186).
- Insuk Sohn, Jinseog Kim, Sin-Ho Jung, Changyi Park, Gradient lasso for Cox proportional hazards model, *Bioinformatics*, Volume 25, Issue 14, July 2009, Pages 1775–1781, <https://doi.org/10.1093/bioinformatics/btp322>
- Chen, Y., Ma, Y., Li, Y., Yu, Y., Lu, B., Liao, L., Li, F., Wen, Z., Jiang, W., Guo, P., Fang, D. and Lu, G. (2023). Bioinformatics combined with clinical data to analyze clinical characteristics and prognosis in patients with HER2 low expression breast cancer. *Gland Surgery*, [online] 12(2), pp.197–207. doi:<https://doi.org/10.21037/gs-22-747>.

GenAI Use Declaration

GenAI tools like ChatGPT was used to understand some of the concepts used in the study in more detail as well as to form a better structure of the report. All data analysis and writing work has been done entirely by me, Amritansh Tiwari .

Github Link : https://github.com/Amt11cdr/TCGA_BRCA_ERBB2_Expression_Analysis