PREDICTING BREAST CANCER SURVIVAL BASED ON GENE EXPRESSION AND CLINICAL VARIALBES

by

Jitesh Patel

A MRP

presented to Ryerson University

In partial fulfillment of the

Requirements for the degree of

Master of Science 2017

In the program of

Data Science and Analytics

Toronto, Ontario, Canada, 2017

1. ABSTRACT

Survival of breast cancer patients are irregular. Several gene sets are directly or indirectly involved in breast cancer. I explored whether combination of mRNA expression of such sets may improve prediction of breast cancer survival in triple-negative class. I used Gene Expression data from TCGA for this study. I have Classified 19 genes into two sets per relationship between death-risk and expression of each gene using Cox proportional hazards model. Up-regulated genes in one set and down-regulated genes of second set resulted high risk for triple-negative class (classified based on estrogen, progesterone, HER2 receptor level, clinical data) of cancer. Combine effect of gene set1 and set2 on survival probability was predicted on overall data, triple-negative and luminal class of breast cancer using Kaplan-Meier model. Combining effect of multiple gene signatures improves prediction of triple negative breast cancer survival. This methodology can be relevant for different cancer types and target therapies.

2. INTRODUCTION

Breast cancer is one of the most common cancer types in women. In 2015, an estimated 234,190 new cases will be diagnosed, and 40,730 deaths from breast cancer will occur [1]. More than 3.1 million US women with a history of breast cancer were alive on January 1, 2014.23 Some of these women were cancer-free, while others still had evidence of cancer and may have been undergoing treatment. [2] The estrogen (ER) and progesterone receptor (PR) level, and the level of human epidermal growth factor type 2 receptor (HER2) expression have been used for prognosis predictions and classifying breast cancer into sub classes such as Triple Negative, HER2, Luminal A and Luminal B. It is crucial to recognize which patients are at risk of developing a more fatal type of breast cancer and have very less survival rate and its more crucial for triple negative and Her2 subclasses of breast cancer. In this project, to identify correlation between survival probability and selected gene expressions, we analyzed 1092 breast tumor expression profiles of 19 genes of interest downloaded from TCGA (The Cancer Genome Atlas). I have used all the genes individually in analysis to classify them into different categories based on change in their expression on risk score of survival. I found two sets of gene containing 13 and 5 genes respectively in each set which has similar effect on survival risk by sets. My results showed that 13 genes

of set1 and 5 genes of set2 were significantly associated with the worst survival rates specifically for triple negative breast cancer.

3. LITERATURE REVIEW AND EXPLORATORY ANALYSIS

Cancer is a complex disease which involves number of epigenetic and genetic irregularities. Tumors originating in the same tissue or organ may vary considerably in genomic alteration and similar patterns of genomic alteration are observed in tumors from different tissues of origin. Different set of genes are expressed at different stages, types. These gives an opportunity to analyze combination of gene expression and clinical features such as number of lymph nodes, stage types, age etc.

Cancer therapy is challenged by the diversity of molecular implementations of oncogenic process and by the resulting variations in therapeutic responses (Giovanni et al., 2013). Targetable functional events in a tumor class are suggestive of class-specific combination therapy. These may assist in the definition of clinical trials to match actionable oncogenic signatures with personalized therapies (Giovanni et al., 2013). Combining the predictive strength of multiple gene signatures improves prediction of breast cancer survival (Xi Zhao et al., 2011). Mutations in transcriptional factors/regulators show tissue specificity, whereas histone modifiers are often mutated across several cancer types. Clinical association analysis identifies genes having a significant effect on survival, and investigations of mutation with respect to clonal/sub-clonal architecture delineate their temporal orders during tumorigenesis. Taken together, these results lay the groundwork for developing new diagnostics and individualizing cancer treatment (Cyriac K., et al., 2013). Different therapeutic treatment can be more effective for individual patient can be based on combination of gene expression at different cancer stages or age category. Project such as The Cancer Genome Atlas (TCGA) provide molecular tumor maps in unprecedented detail (Giovanni et al., 2013). For analysis, I have selected set of genes which are directly or indirectly associated with the breast cancer. Selected genes are AKT1, AKT2, AKT3, MYC, MYCL1, MYCN, PRKAA1, PRKAA2, PRKAB1, PRKAB2, RB1, STK11, CAMKK2, PTEN, PIK3CA, MTOR, TP53, ATM and CHEK2.

I have downloaded clinical data for breast cancer from TCGA using FireBrowseR web API for R. The dataset contains 1097 records with 111 columns for different clinical features. Distributions of these variables are as below. Maximum number of patients were alive at the time when I have downloaded the

dataset. Very less records with dead. Patients with Stage iia and stage iib are highest. This variable has some of the data. 8 records don't have stage details in the dataset. Records with 2 examined lymph nodes are highest and 126 records don't have details about this variable. Maximum patients were at age of 62 at the time of initial pathologic diagnosis. Less cases with diagnosis at early stage of the life. There are 12 male records with breast cancer and 1085 female records. There is positive correlation between TP53, MYCL1, AKT3 and MYC.

4. METHODOLOGY

4.1 DATA COLLECTION

I have downloaded 1097 respondents from TCGA (The Cancer Genome Atlas) using FirebrowseR R API. FirebrowseR is a R client for Broads Firehose Web API, which is serving data generated by the Firehose Pipeline and this pipeline processes TCGA date sets. We can directly get TCGA data to R using this API. I have downloaded 110 clinical variables and gene Expressions for 19 genes (AKT1, AKT2, AKT3, MYC, MYCL1, MYCN, PRKAA1, PRKAA2, PRKAB1, PRKAB2, RB1, STK11, CAMKK2, PTEN, PIK3CA, MTOR, TP53, ATM and CHEK2).

4.2 DATA PROCESSING

The data which I got from the TCGA using FirebrowseR required some processing as file contains normalized gene expression of all the genes together in rows. To compare change of gene expression in normal tissue and Tumor Tissue, I have separated this data in two major files 1) gene expression of all the genes in column for Normal Tissue and 2) gene expression of all the genes in columns for Tumor Tissue.

From 1097, I had 1092 records with gene expression in Tumor tissue. Whereas, there were only 112 records with gene expression in Normal tissue. All missing gene expressions of Normal tissue were replaced by median of respective gene expression in Normal Tissue.

Now, to find out how expression of these 19 genes varies individually in normal tissue and tumor tissue, I have calculated fold change of respective gene expression in normal tissue and tumor tissue. Fold changes are commonly used in the biological sciences as a mechanism for comparing the relative size of

two measurements. They are computed as: num / denom if num > denom, and as −denom / num otherwise. (‘gtools CRAN’ package). I have considered normal tissue gene expression as numerator and gene expression of normal tissue as denominator. Then I have calculated log-ratio of the fold change.

4.3 RISK PREDICTION

I have used Cox proportional hazards regression analysis to predict risk based on fold change of each gene individually. The Cox model is a relative risk model. Cox proportional hazard works for both quantitative and categorical variables. I can also assess effect of several variable on survival time using Cox Proportional hazard regression. The purpose of this model is to check effect of fold change of 19 genes of interest. This helped me to find out how expression of each gene individually influence risk of death. This rate is risk score. The Cox model is expressed by the hazard function denoted by h(t). Briefly, the hazard function can be interpreted as the risk of dying at time t. It can be estimated as follow:

$h(t) = h_0(t) \times \exp(b_1x_1 + b_2x_2 + ... + b_px_p)$ h(t) = h0(t) × exp(b1x1 + b2x2 +...+ bpxp)

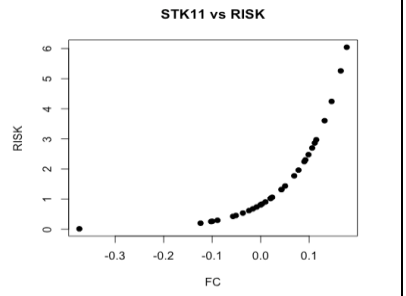
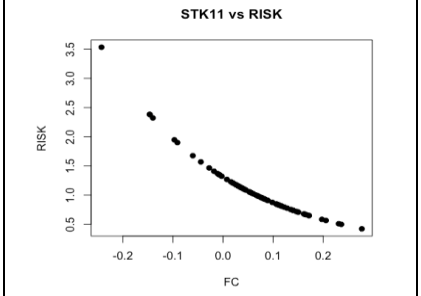
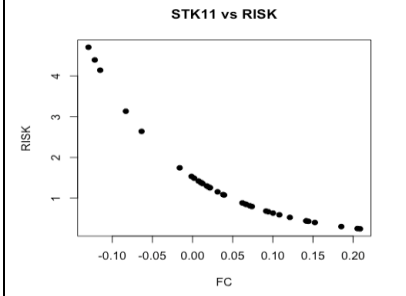
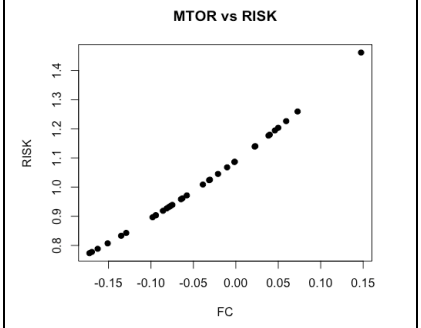
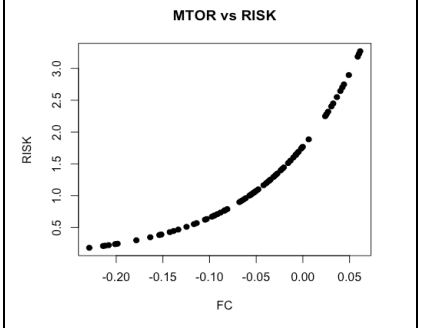
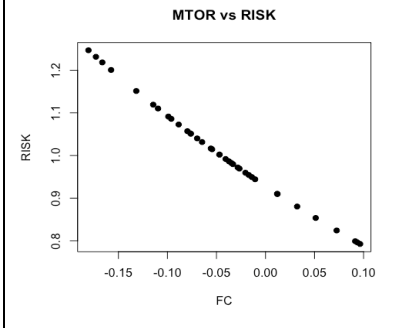
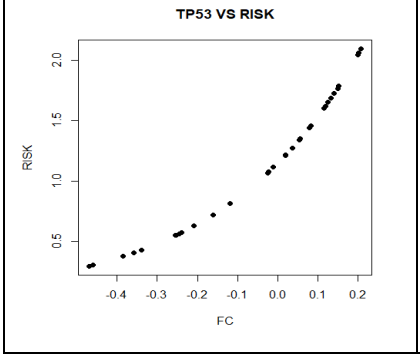
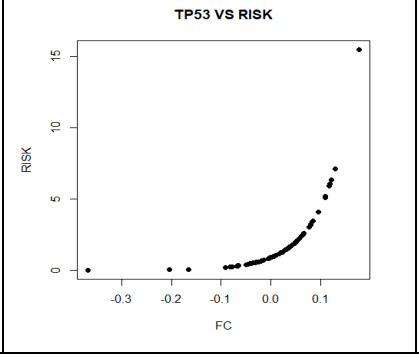
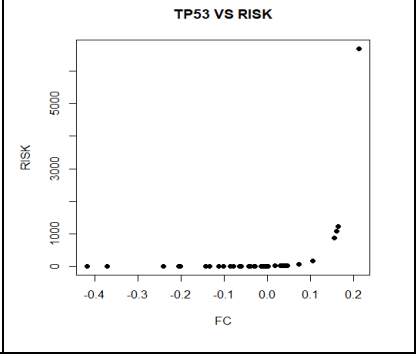
where,

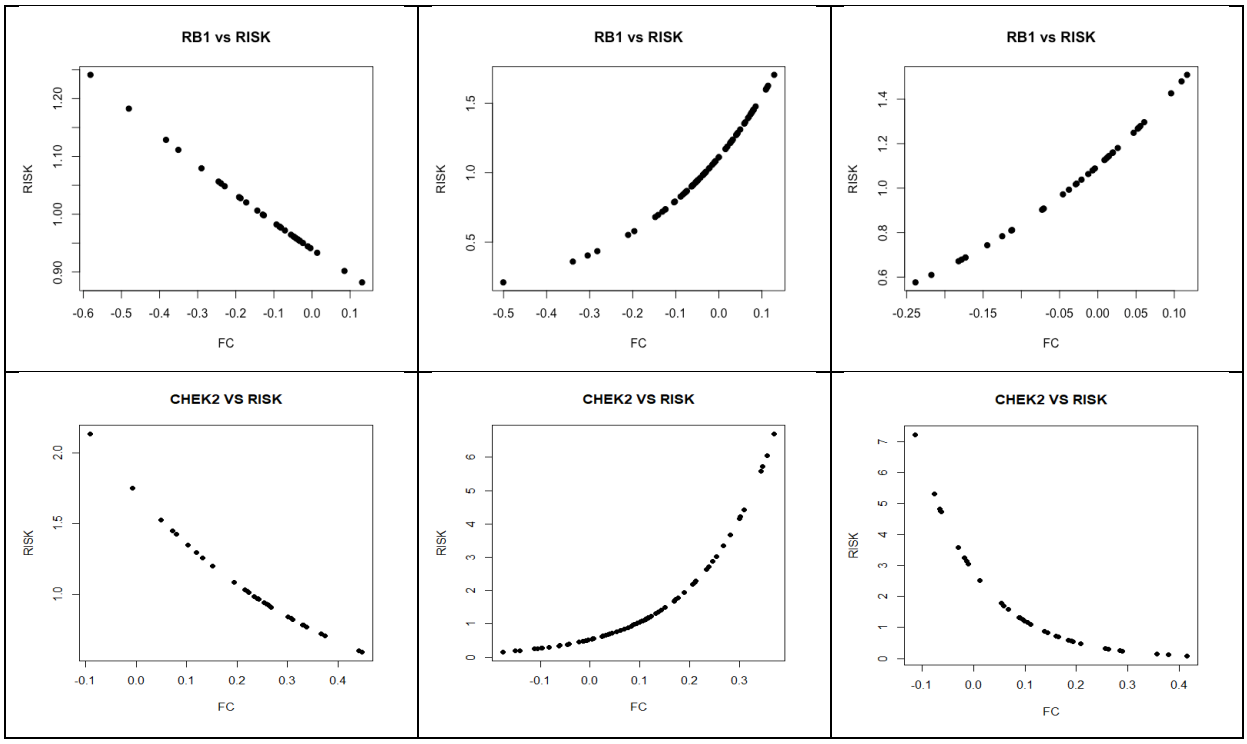
- t represents the survival time
- h(t)h(t) is the hazard function determined by a set of p covariates (x1, x2, ... ,xpx1,x2, ..., xp)
- the coefficients (b1, b2, ..., bpb1, b2, ..., bp) measure the impact (i.e., the effect size) of covariates.
- the term h0h0 is called the baseline hazard. It corresponds to the value of the hazard if all the xixi are equal to zero (the quantity exp (0) equals 1). The ‘t’ in h(t) reminds us that the hazard may vary over time.

I have plotted risk score of individual gene across fold change to find out any correlation between fold change of gene and risk score. To investigate genes related to triple negative breast cancer, I have classified records downloaded from the TCGA into three groups (Triple Negative, Luminal A and Luminal B) using clinical variables (Estrogen Receptor(ER)-er_level_cell_percentage_category, HER 2 Receptor (HER2)-her2_erbb_pos_finding_cell_percent_category, Progesterone Receptor (PG): progesterone_ receptor_level_cell_percent_category). Records with more than 10% level for any of these receptors were recoded as positive for respective receptor. I have predicted correlation between up-regulated gene set and down regulated gene set on risk.

Type	ER	PG	HER2
Luminal A	+	+	-
	+	-	-
	-	+	-
Luminal B	+	+	+
	+	-	+
	-	+	+
Her2-Enriched	-	-	+
TN (Basal Like)	-	-	-

Below are plots of STK11, MTOR, TP53, RB1 and CHEK2 fold change vs risk for triple negative, luminal A and luminal B.

Triple Negative	Luminal A	Luminal B
		
		
		



Risk plot of all the 19 genes for all the three classes (triple negative, luminal A and Luminal B) of breast cancer are available as Table 1 in appendix.

It is clear from the plots that up regulated genes PRKAA1, PRKAB1, PRKAB2, MTOR, STK11, AKT2, AKT3, MYC, MYCL1, MYCN, CAMKK2, PTEN, TP53 and ATM leads to high risk and down regulated genesRB1, AKT1, PIK3CA, PRKAA2, CHEK2 results in high risk.

4.4 SURVIVAL RATE PREDICTION

To find out combine effect of above two sets of genes on survival, I have divided median of up regulated genes (PRKAA1, PRKAB1, PRKAB2, MTOR, STK11, AKT2, AKT3, MYC, MYCL1, MYCN, CAMKK2, PTEN, TP53) by median of low regulated genes (RB1, AKT1, PIK3CA, PRKAA2, CHEK2). I have classified the difference into two groups “high difference” and “low difference”. This categorical variable depicts difference between high expressed and low expressed of genes. I have used this categorical variable to predict survival rate using Kaplan-Meier survival analysis on overall data, triple negative and luminal class of cancer. The median of a measured module’s score was used to dichotomize the data

4.5 KAPLAN-MEIER SURVIVAL

It is non-parametric method for estimating probability of survival from observed survival time (Kaplan and Meier, 1958).

The survival probability at time titi, S(ti)S(ti), is calculated as follow:

$S(t_i)=S(t_{i-1})(1-d_in_i)$S(ti)=S(ti−1)(1−dini)

Where,

- S(ti−1)S(ti−1) = the probability of being alive at ti−1ti−1
- nini = the number of patients alive just before titi
- didi = the number of events at titi
- t0t0 = 0, S(0)S(0) = 1

I have used survfit() function to plot Kaplan Meier curve to measure median survival time for categorical variables in my dataset. KM is good to measure effect of single categorical variable on survival.

5. RESULTS

I have selected 19 genes for this study based on prior knowledge in breast cancer. All the selected genes are involved in cell cycle. I could classify 25 records as Triple Negative class, 116 as Luminal class of breast cancer from 1092 records downloaded from The Cancer Genome Atlas based on clinical variables such as Estrogen, progesterone and Her2 receptor percentage level. I have predicted risk based on change in signature of each gene individually for three classes of the cancer i.e. Luminal A, Luminal B and Triple Negative using Cox hazards regression model. I have classified 19 genes of interest into two different sets based on predicted risk. Up regulated gene set1 (PRKAA1, PRKAB1, PRKAB2, MTOR, STK11, AKT2, AKT3, MYC, MYCL1, MYCN, CAMKK2, PTEN, TP53) leads to increase of risk hazard. And down regulated gene set2 (RB1, AKT1, PIK3CA, PRKAA2, CHEK2) leads to increase of risk hazard. Plots for risk vs fold change of each gene are available in appendix.

Predicted survival probability using combine effect of set1 and set2 for Triple Negative class depicts that high difference in expression of these two sets of gene leads to low survival rate and years with p value 0.038 and survival rate for low difference in expression of these two gene sets is higher. Plot for this is present below as figure 2.

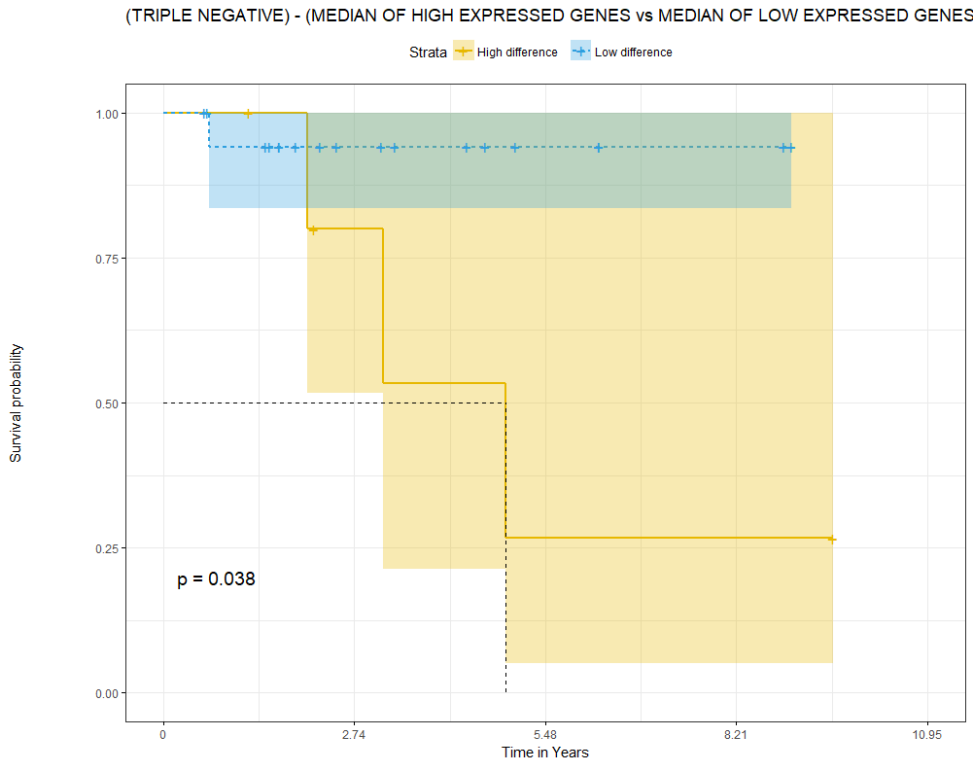


Figure 2

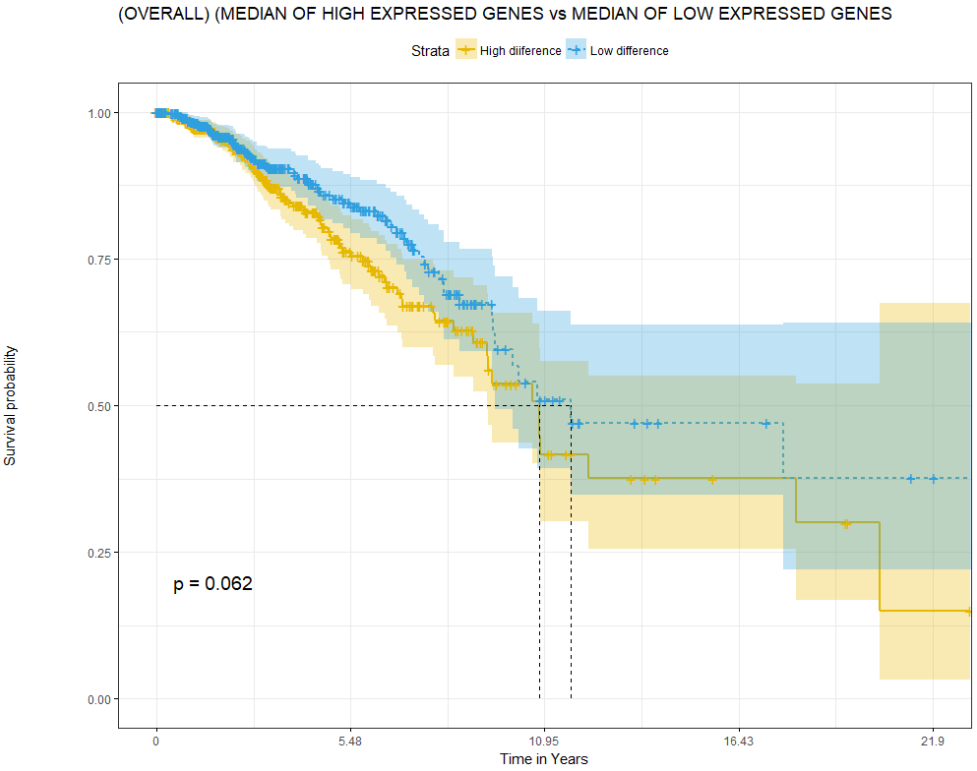


Figure 3

Predicted survival plot for overall data and luminal class of breast cancer are available in appendix. In both, overall data (Figure 3) and luminal class of cancer (Figure 4), combine effect of these gene expression are not able make distinctive prediction. For overall data, median survival for high difference and low difference between these genes are almost closer to 10 years with p value 0.77 whereas for luminal class of cancer survival rate is not reaching to median.

I have also tried to predict survival using different sets of gene expression based on prior knowledge of triple negative breast cancer and risk score predicted using Cox hazards regression model. I have used MYC, MYCL1, MYCN as set1 and RB1 as set2, PRKAA1, PRKAB1, PRKAB2, MTOR as set1 and RB1 as set2 and lastly PRKAA1, PRKAB1, PRKAB2, MTOR, STK11 as set1 and RB1 as set2. Survival rate is not reaching to median for any of these combinations in triple negative. Survival Plots for each experimented set of genes are available in appendix Figure 5 to 12.

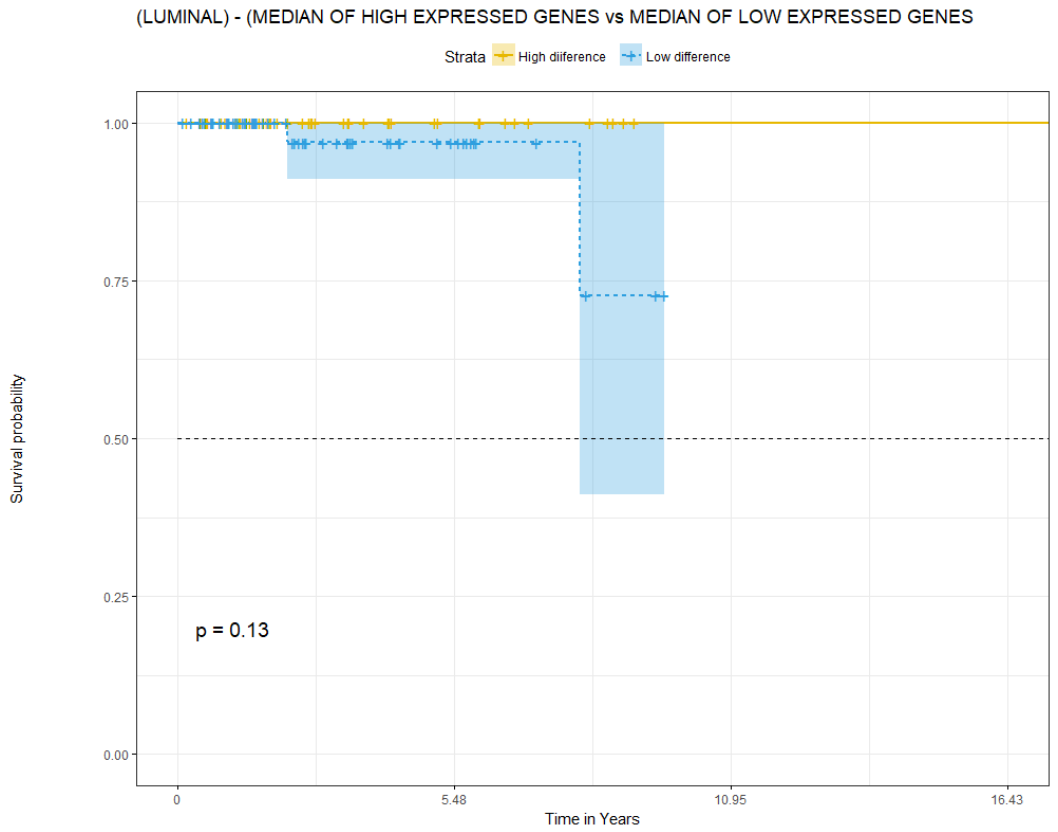


Figure 4

6. DISCUSSION

The discovery of prognostic factors is crucial work in breast cancer research and even more crucial for triple negative subtype of cancer. In this study, using 19 genes of interest based on prior knowledge of breast cancer, we found that two types of prognostic gene sets are strongly correlated with poor patient survival having triple negative subtype of breast cancer. When comparing risk score and gene expression, tumors with high risk scores were more likely to have upregulated PRKAA1, PRKAB1, PRKAB2, MTOR, STK11, AKT2, AKT3, MYC, MYCL1, MYCN, CAMKK2, PTEN, TP53 genes in triple-negative subtype, and worse survival rates. The 13 up regulated genes and 5 down regulated genes used as a prognostic gene in this research. All the selected 19 genes involved in the cell cycle process. Expression levels of these genes were strongly associated with poor survival. Combined effect of highly expressed PRKAA1, PRKAB1, PRKAB2, MTOR, STK11, AKT2, AKT3, MYC, MYCL1, MYCN, CAMKK2, PTEN, TP53 and low-expressed RB1, AKT1, PIK3CA, PRKAA2, CHEK2 revealed poor survival in triple negative subtypes of breast cancer and found to be prognostic factor.

7. CONCLUSION

My conclusion presents the score of prognosis prediction model which is sturdily correlated with shortened survival times in breast cancer, and the score of the model is consistently high in aggressive breast cancer types Triple Negative. Therefore, we recommend the consideration of combine effect of these up regulated and down regulated genes as new prognostic markers and we expect that these findings can be adapted to research on target therapies for ill-defined breast cancer types. As future work, I am working on dataset from ICGA to prove similar hypothesis.

8. REFERENCES

1. Howlader N NA, Krapcho M, Garshell J, Miller D, Altekruse SF, Kosary CL, Yu M, Ruhl J, Tatalovich Z,Mariotto A, Lewis DR, Chen HS, Feuer EJ, Cronin KA: SEER Cancer Statistics Review, http://seer.cancer.gov/csr/1975_2012/, based on November 2014 SEER data submission, posted to the SEER web site. 2015.

2. DeSantis CE, Lin CC, Mariotto AB, et al. Cancer treatment and survivorship statistics, 2014. CA Cancer J Clin. 2014.

3. Joe S, Nam H: Prognostic factor analysis for breast cancer using gene expression profiles, based on October 2015 The ACM Ninth International Workshop on Data and Text Mining in Biomedical Informatics Melbourne, Australia

4. The Cancer Genome Atlas Research Network. Comprehensive molecular portraits of human breast tumors. 2012; nature11412

5. Cuilan Li, Vincent WS Liu, Pui M Chiu, David W Chan and Hextan YS Ngan, Over-Expressions of AMPK subunits in ovarian carcinomas with significant clinical implications. 2012;12:357.

4. Giovanni, C. et al. Emerging landscape of oncogenic signatures across human cancers. 2013.

5. Xi, Z. et al. Combining Gene Signatures Improves Prediction of Breast Cancer Survival. 2013; e17845.

6. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumors

7. Cyriac, K. et al. Mutational landscape and significance across 12 major cancer types. 2013; nature12634.

8. https://cancergenome.nih.gov/

9. http://firebrowse.org/

10. http://www.genecards.org/