Capstone Project Interim Write-Up - Genomics Project

Tomas Meade

Problem Description

Gynecologic and breast cancers have a strong prevalence throughout the world and in the United States. In 2017, there were 350,000 cases of Pan-Gyn cancers in the United States and many more outside of the U.S. Specifically, breast cancer alone has the highest rate of diagnosis of any other cancer in the world according the World Health Organisation. Breast cancer accounted for 12% of new cancer cases in the world in 2021 and resulted in 650,000 deaths in 2020. However, survival rates have shown improvement in the few decades. The American Cancer Society reported in 2017 that over the last 25 years, the number of deaths has dropped by 40%.

This is a promising improvement that can be partially accredited to clinical advancements like targeted therapies. These targeted therapies focus on the molecular make up of tumors unlike chemotherapy which targets the cancer cells. In order to continue to develop and improve targeted therapies, recent research has attempted to identify molecular features and therapeutic targets based on genomic data.

The aim of this project is to replicate some of the recent molecular research on breast cancer, specifically parts of the research done in the paper titled "A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers" by Berger et al. We will attempt to find shared and unique molecular features of breast cancer tumors as well as identify significant clinical characteristics and their relation to the genetic make up of patients. Of particular interest to the project is the HER2 gene which is known to have a strong relationship to breast cancer based on previous research. To examine this relationship, we account for general HER2 mutation and whether the gene exhibited amplification or deletion in each patient. We then analyze clinical outcomes, specifically survival rates and how they relate to these characteristics. We will prove that HER2 positive cancers tend to be more severe when it comes to survival outcomes.

Data Description

The data source for our project is the Genomic Data Commons or GDC. The GDC is part of the National Cancer Institute (NCI) and is a cancer research database with a specific focus on genomic data. The data sets are generated by NCI designated cancer centers around the U.S like university medical centers and hospitals. It includes data sets from two of the largest programs devoted to examining genomic data related to cancer research and treatment. The two programs being The Cancer Genome Atlas (TCGA) and Therapeutically Applicable Research to Generate Effective Therapies (TARGET) Mission. The mission is to provide data to assist with genomic cancer research projects.

For our project we choose to narrow down the focus and concentrate on the TCGA and specifically the TCGA-BRCA project which contains data on patients with breast cancer. The TCGA-BRCA project data is organized by case so each observation is a patient with breast cancer. There are a total of 1,098 cases in the TCGA-BRCA project. Of the 1098 cases 1,085 are females and 12 are males, 945 are alive and 152 are dead, and there are 21,058 different genes present. Of special interest to us is the HER2 gene (see figure 1). In the data set there are 233 cases with HER2. As for the general structure of the data, each observation is a case and the features are different types of genomic data, clinic data and bio specimen data which contains information about the actual physical sample that was taken from the patient. The other features contain information about mutations and genes present in the sample and also images of the samples. To conduct our analysis we will utilize the only clinical data and parts of the genomic data. Since there are many different types of genomic data, we decided to focus on a few key types that were identified as significant by the paper mentioned earlier. For the first part of our analysis we use copy number variation (CNV) data. CNV data can be converted into focal scores which are a centralized way of determining if the variation resulted in amplification or deletion, or if no significant variation occurred.

Clinical Data	**N=1098**
Sex	
Male	12 (1.1%)
Female	1085 (98.9%)
Vital Status	
Alive	945 (86.1%)
Dead	152 (13.9%)
AJCC Stage	
Stage 1	183 (16.8%)
Stage 2	621 (57.2%)
Stage 3	249 (22.9%)
Stage 4	20 (1.9%)
Stage X	13 (1.2%)

Figure 1: Summary Statistics of the TCGA-BRCA dataset.

For the second part of our analysis we will use other features of genomic data that were identified as significant in previous literature. The key variables are protein expression of ER and PR, ERBB2 amplification status, hyper mutator phenotype, AR protein expression and mutation status of PTEN, TP53, H-RAS/K-RAS/N-RAS, ERBB2, PIK3CA, and POLE. We will then look at similarities and differences between cases based on these genomic features and then connect them to clinic data to examine how they relate things like race,

cancer stage and survival. To conduct the survival analysis we will have to account for the built-in time-based bias since when patients can be in different stages of cancer when they are diagnosed. The plots below show how the survival rates differ by AJCC pathological stage which is A system to describe the amount and spread of cancer in a patient's body. In stage IV, we can see that patients with HER2 amplification and deletion have survival curves that drop much faster than those who have no HER2 mutation.

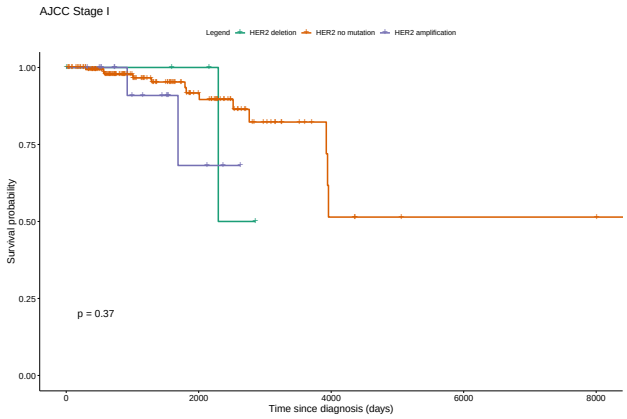

Figure 2: Survival Plot for patients in the AJCC stage I split by HER2 mutation.

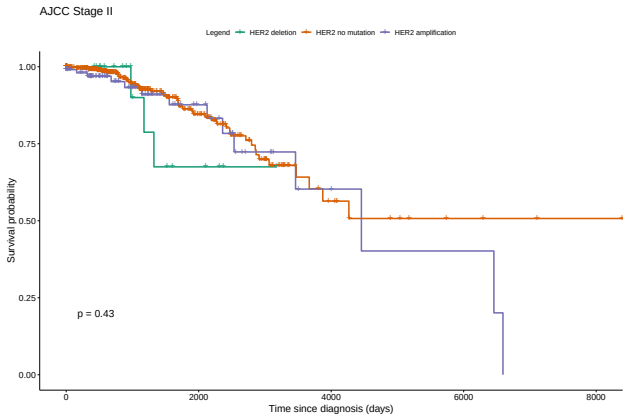

Figure 3: Survival Plot for patients in the AJCC stage II split by HER2 mutation.

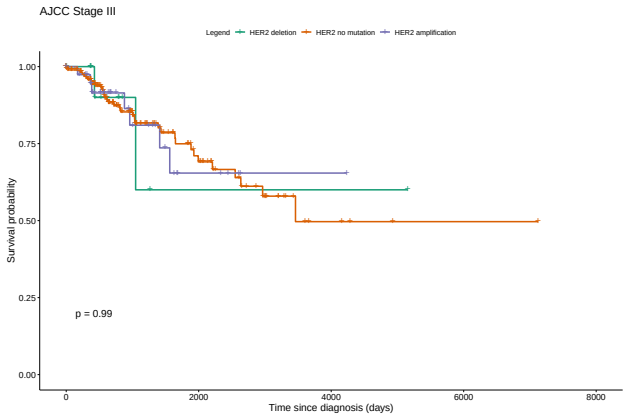

Figure 4: Survival Plot for patients in the AJCC stage III split by HER2 mutation.

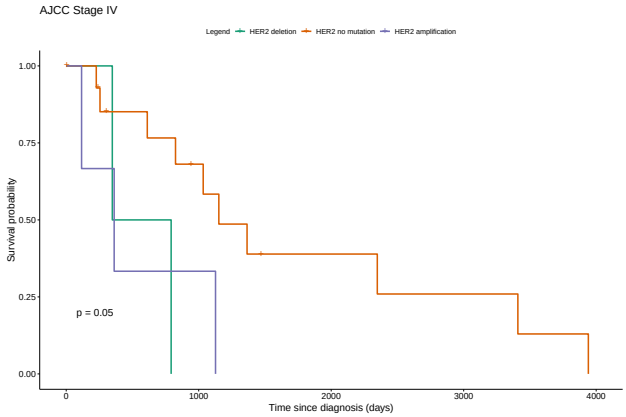

Figure 5: Survival Plot for patients in the AJCC stage IV split by HER2 mutation.

Some of the most important variables are ER and PR positive status and ERBB2 amplification status. ER positive breast cancer are those that have estrogen receptors and PR positive cancers are those with progesterone receptors. This means that cancer cells with these receptors need estrogen and progesterone to grow, which makes them important factors since treatments can target these receptors to inhibit tumor growth. Another key variable is ERBB2 amplification status. ERBB2, also known as HER2, is a gene that produces HER2 proteins which regulate the growth of breast cells. Having amplified ERBB2 or HER2 means that a patient is HER2 positive and that breast cells grow too rapidly. This amplification can result in breast cancer cells spreading a lot faster, resulting in more severe cases. Therefore, it is a important factor in understanding and treating breast cancer.

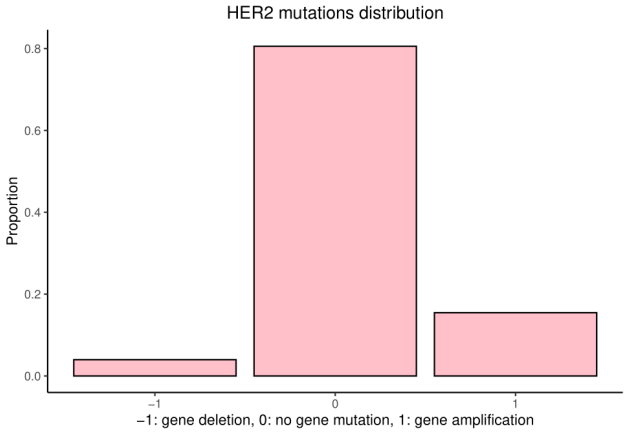

Figure 6: Proportion of HER2 amplifications, deletions and none significant mutation amoung the patients

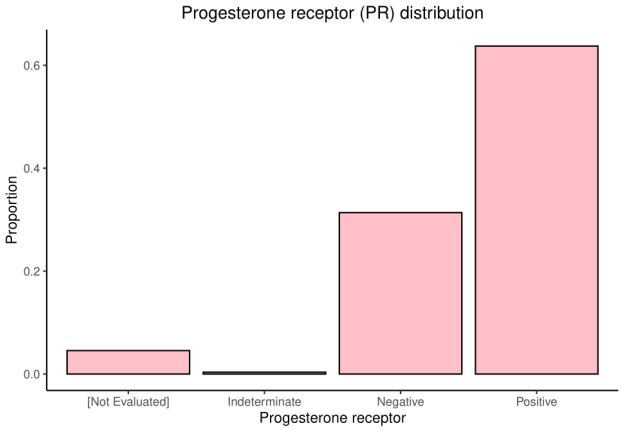

Figure 7: Proportion of PR positivity amoung the patients

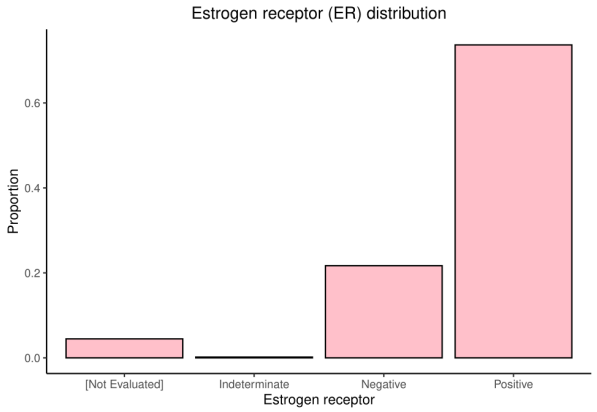

Figure 8: Proportion of ER positivity amoung the patients

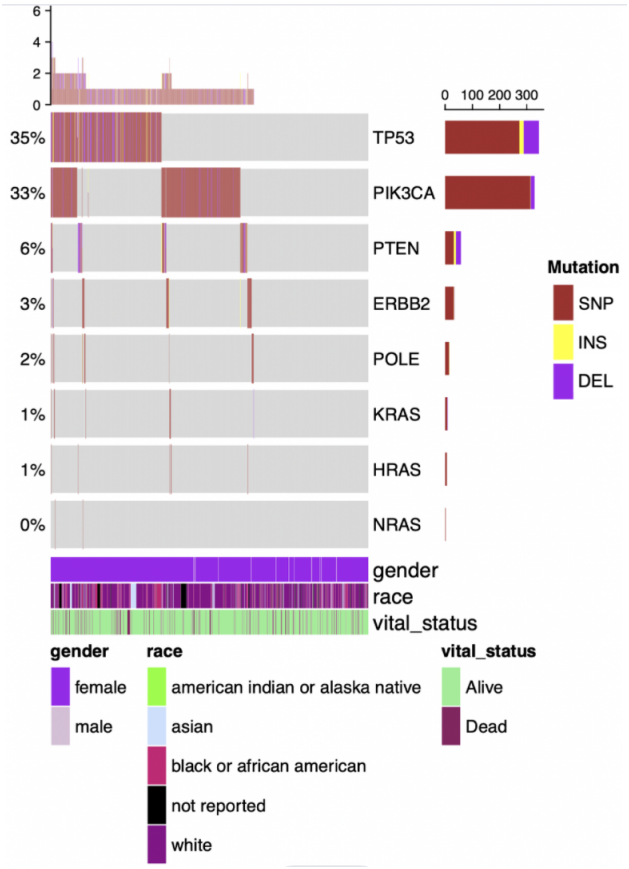

Figure 9: Diagram showing the types of mutations for the key genes present in the dataset. Each of thin colored lines represent a patient and color specifies the type of mutation that occured for the genes listed on the right. The gender, race and vital status for each of the patients is also listed. The percentages are the the proportion of times the gene exhibited mutation out of the total samples the gene was present in.

In order to access the data, we queried the GDC platform from an R environment. The GDC processes data using bioinformatics pipelines which then can be accessed by researchers. One way to access this is using the Bioconductor software package which is based primarily in R, but can be used with other software as well. In order to access the TCGA data we used an R package called TCGAbiolinks which utilizes the Bioconductor package and allows you to write queries directly in R to to extract TCGA project specific data. For the first part of our analysis, we pulled the clinical and CNV focal score data from the TCGA-BRCA project into R using TCGAbiolinks.

Methods

To prepare the data for analysis, all the variables were converted to factors (categorical) or integers/numeric (continuous). To analyze shared and unique molecular features of breast cancer tumors, we used unsupervised learning to group similar patients together. Specifically, we used hierarchical agglomertive clustering to group the patients into multiple clusters.We first cluster the patients using only the CNV focal scores and then we also cluster the patients using both the focal scores and the clinical data. In the future we preform the clustering on all the key features mentioned in the data section.

To preform the clustering on the CNV data, we used euclidean distance to quantify the distances between observations which is given by

$$d_ij = d(i,j) = sqrt((i_1 - j_1)^2 + (i_2 + j_2)^2 + \ldots + (i_n + j_n)^2).$$

To preform the clustering on both the clinical and CNV data, pairwise dissimilarities between observations, in this case patients, are calculated. These dissimilarities or distances are computed using the 'daisy()'. function in R. Since we have both categorical and continuous variables, we use a generalization of Gower's formula which is the gowers metric in hclust(). This means that the variables will be standardized and then the distance between observations is the sum of all the distances for each variable.

Gower's method starts with standardization which is done by dividing each individual value by the column's range after subtracting the minimum value. This means that each variable has a range of between 0 and 1. Then the dissimilarity is calculated as the weighted mean of the contributions of each variable or

$$d_ij = d(i,j) = sum(k = 1:p; w_k\delta(ij;k)d(ij,k))/sum(k = 1:p; w_k\delta(ij;k)).$$

The equation above shows that d_ij is a weighted mean of $d(ij,k)$ with weights w_k and $\delta(ij;k)$, w_k are the weights for k, $delta(ij;k)$ is 0 or 1, and $d(ij,k)$ is the distance between $x[i,k]$ and $x[j,k]$. The weight $delta(ij;k)$ is zero when $x[,k]$ is missing in one of the rows or when the variable is asymmetric binary and both values are zero, otherwise it is always one. The contribution of binary variables given by $d(ij,k)$ is zero if both values are equal and one otherwise. The contribution to the distance for the other variables is the absolute difference of both values, divided by the range of that variable.

After the distances are calculated then the clustering is performed using the distances between observations. For bottom up or agglomertive clustering, each patient starts out as its own cluster and similar patients are

grouped together. The 'hclust()' function preforms the clustering in R. The algorithm joins the two closest clusters and then repeats until there is only one cluster containing all the patients. The cluster grouping is done using the complete linkage method calculates the distance between two clusters by the largest distance between a point in the first cluster and a point in the other cluster. For each step of the algorithm the distances between clusters are computed using the Lance-Williams dissimilarity update formula. Given two clusters, K and L the dissimilarity between them and another cluster M is given by

$$D(K \cup L, M) = \alpha_1 * D(A, C) + \alpha_2 * D(B, C) + \beta * D(A, B) + \phi * |D(A, C) - D(B, C)|$$

where the four coefficients $(\alpha_1, \alpha_2, \beta, \phi) are (.5, .5, 0, +.5)$.

After the clustering is complete, a dendrogram can be constructed using the groups at each iteration. Then the number of clusters can be decided by visualizing the dendrogram and choosing the number where there is a large separation between steps of the dendrogram. We split our dendrogram into 5 clusters using 'cutree()' which cuts the dendogram and labels each point into its respective cluster. In the future we will be more rigorous about choosing k. We chose 5 initially because that is what the paper we are replicated choose as the optimal number.

Results

The two plots below show the results of clustering on the CNV data and then both the CNV data and the clinical data. The heatmap shows the genes and varies instances of amplification and deletion. The dendrogram shows the hierarchical clustering of the patients and the red line shows where we cut and separated the patients into clusters. These results are only preliminary and we want to cluster on more genomic data before looking at characteristics of the clustering and conducting survival analysis.

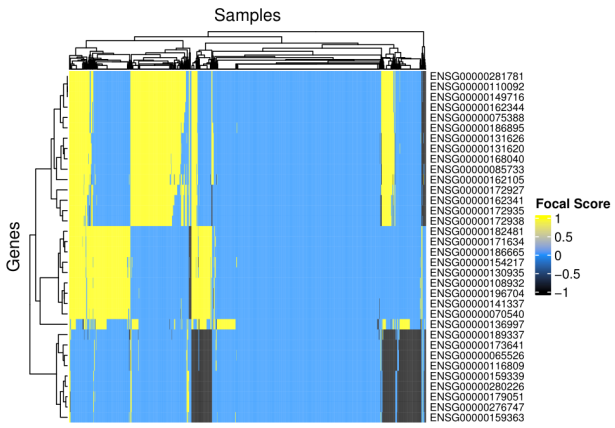

Figure 10: Heatmap and cluster analysis of patients based on CNV focal scores

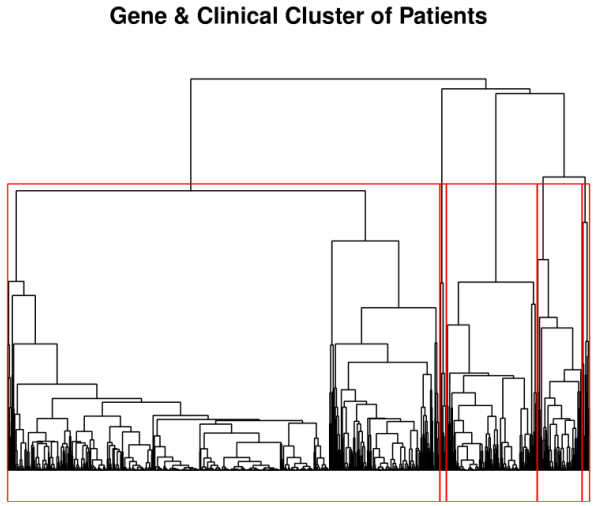

Figure 11: Dendrogram of patients based on clinical data and CNV focal scores

Conclusions

In the future we want to incorporate the other key features, mutations of selected genes, hyper-mutator status, ERBB2 amplifications, CNV load, immune score, ER, PR, and AR protein expressions. Following the paper, we will also convert the data into present/absent (for discrete features like mutations) and high/low (for continuous features) in each sample. To do this we will attempt to obtain the thresholds for low/high categorization by modelling the variables with a bimodal Gaussian distribution and setting the threshold

as the value between the two modes. We will also attempt to revise the existing thresholds for AR, ER and PR and potentially identify new thresholds by maximizing the AUC (using the Youden index) for the continuous-valued expressions available for AR/ER/PR status. We will then use these data to preform cluster analysis following the same steps listed in the methods section. Lastly, we will construct a descion tree to group the patients into clusters and preform survival analysis.

Reference

Berger et al. A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers