Determining the Efficacy of Utilizing CNA Patterns to Diagnose Breast Cancer

Gayatri Ratakonda

Yorktown High School

Acknowledgements

I'd like to thank my mentor for his constant support, revisions, and time spent for the betterment of the results of my project. I'd also like to recognize my teachers for their tireless effort to help make sure I communicated my results in the best possible way. Finally, I'd like to thank my parents and peers, whose words of encouragement motivated me throughout the process of experimentation.

Table of Contents

Table of Figures

Abstract

Breast cancer is a prevalent disease, affecting approximately 1 in 8 women during their lifetime (Komen, 2019). The advancement of modern technology has allowed for the 5 and 10-year survival rates for breast cancer patients after tumor removal to be 90 and 83% respectively. However, if the breast cancer is metastatic by the time the patient is diagnosed (approximately 30% of diagnoses), the 5-year survival rate drops to 27% (PHS, 2017). Seeing as almost a third of the patients are diagnosed post-metastases, a necessity for better diagnostic methods, which are less invasive and easier for the patients to undergo exists. In order to create a new computationally derived diagnostic method, a database created by Pereira et al. in 2016 which focused on Copy Number Aberration (CNA) data was used. CNAs are documented changes in the number of times a gene is expressed. In the database, the CNA values ranges from -2, meaning a double deletion of a gene from the DNA strand, up to 2, which denoted a double amplification of the gene. By utilizing *Matplotlib* in Python, the reported tumor oncogenes that were most highly correlated with metastatic activity were derived. Then, integrating the CNA data, and it was found that the WWTR CNA sequence was highly correlated with the presence of the known tumor oncogene, HERC2. With this finding, an algorithm was created by utilizing *SciKit Learn* with the input fields asking only for the CNA data of the WWTR sequence. By utilizing this algorithm, an accuracy of 63% was achieved. The algorithm's accuracy was stunted by the limitations of the database, and refining this algorithm by adding more CNA sequence data is necessary in order to get a higher level of accuracy. However, this experiment was substantial in that a definite correlation between cancer and CNA patterns was established, which promotes further studies to be undertaken in exploring CNA sequences within cancer diagnostic.

Introduction

Overview

Breast cancer is a deadly disease and growing cause of early death among women globally, affecting approximately 1 in every 8 women at some point during their lifetime (Komen, 2019). However, the advancement of modern technology in cancer treatment has allowed for the 5 and 10-year survival rates for breast cancer patients after tumor removal to be 90 and 83% respectively, which is extremely high compared to the survival rates 10 years ago, which were at 30 and 26% (CN, 2019). However, if the breast cancer is metastatic by the time the patient is diagnosed (approximately 30% of diagnoses), the 5-year survival rate drops to 27% (PHS, 2017). Metastatic breast cancer is characterized by recognizable lumps in or around the breast tissue, which only form after the cancer has been active for 2-5 years depending on its subtype (CN, 2019). As the mutated cells spread and divide, the tumor grows at an exponential rate. Catching the tumor after metastases has occurred leaves almost nothing possible in terms of successful treatments that the patient would overall benefit from. This brings about the general problem with treating breast cancer, being that the correct initial diagnoses must be made before metastatic activity in order for the highest level of recovery to be possible. Diagnostic procedures in place right now include biopsy and diagnostic mammography, both of which are highly invasive in differing ways. While biopsy requires tissue samples, mammography requires potential exposures to radiation from x-rays. Creating an accurate and noninvasive diagnostic procedure would be more efficient for patients due to the fact that they would not have to spend time or money getting surgery and/or scans. It would also have the capacity to predict the treatment method that would most benefit the patient based on their genomic data, which will ease the minds of future cancer patients by providing them with a realistic plan.

The Genomic Transcriptome of Breast Cancer Patients

The onset of breast cancer is caused by several malignant alterations in the DNA sequence known as mutations. Mutations of specific genes have been seen to be correlated with the onset/progression of breast cancer tumors (Workman, 2018). Some of the most commonly appearing and widely recognized tumor oncogenes are BRCA1, BRCA2, HERC2, and ERBB2, known to appear in patients with aggressive breast cancers. Mutations of these genes allow for them to become active in the sense that they have the ability to start creating proteins through the processes of transcription and translation (the central dogma). These proteins change cell function by altering the basic cellular process of mitosis through inhibiting portions of DNA from being replicated. Mitosis has several regulatory checkpoints looking for proper cell alignment and correctly replicated DNA that, if not passed, will revert the cell to a state known as the "G0 phase" or a non-dividing state. The proteins created from the mutated oncogenes also allow these checkpoints to be overridden through DNA silencing (Chung, 2010). Testing for the mutated genes requires biopsy and tissue samples because the only the affected tissue contains mutated DNA.

Copy Number Aberrations (CNAs)

Copy number aberrations (CNAs) are alterations to the DNA sequence that appear in all genomic data in the entire body (Dee, 2016). Specifically, they are changes in the number of times a gene is expressed. They can be the amplification or silencing of genes, changing the overall expression levels of the DNA. In a study conducted by Sun in 2017, it was seen that the observed CNA could be indicative of an expressed or silenced gene, which is important in determining if certain cancers are likely to appear in a patient during their lifetime.

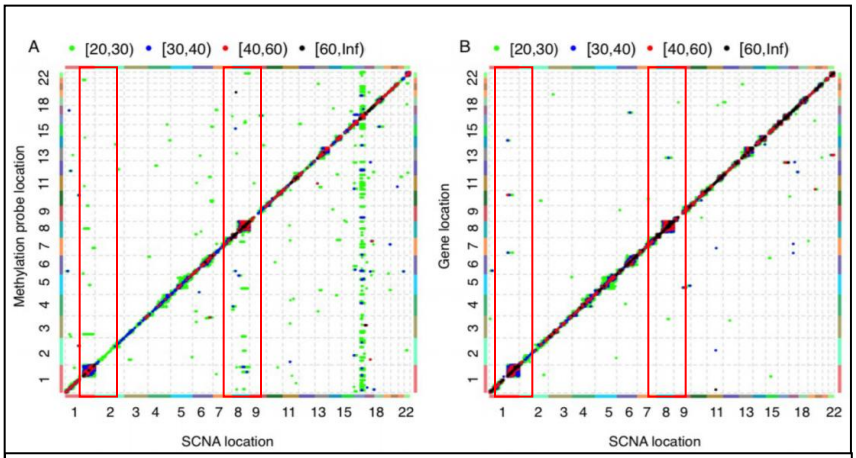

Fig. 1 – Demonstrates that the areas of high methylation (gene silencing) and gene expression are located in areas of extremely high CNA expression. *thorax.bmj.org*

DNA methylation is a marker for gene silencing due to an extra methyl groups binding to the end of the DNA sequence and making it unreadable to the transcription enzyme, which is responsible for translating DNA into RNA that will direct the creation of proteins. By observing that a high amount of CNA activity has been shown around the methylation probes used in the study (seen by the clusters in Figure 1), it was confirmed that CNAs are an accurate tool used to measure gene silencing (Sun, 2017). High amounts of CNA activity had also been observed around the location of the gene on the DNA strand itself, confirming that the expression of the gene is regulated by the number of times it is repeated due to the alteration. This can be seen in the highlighted areas of part B in Figure 1, in which small clusters of activity can be discerned as gene emulsification.

CNA Patterns in Gene Recognition

Usually, CNAs will not appear singularly, emulsifying or silencing only one gene. Instead, they emulsify large groups of DNA, meaning many genes, through a series of CNAs. When considering all identified tumor oncogenes and the genes that lie on

the DNA sequence, many aberrations take place in which genes are expressed an irregular amount of times. It was previously observed that the frequency of CNA activity was seen to be

correlated with the mutation of genes, as can be seen in section 8 of Figure 2, in which a high rate of CNA activity corresponds with a high mutation rate in that section of DNA. Utilizing this ability of clustered CNAs to determine the presence of genes has been virtually unexplored in cancer diagnostics due to the fact that previous research in CNAs have only used them for medical cases in which an established correlation between gene presence and sickness progression has been made.

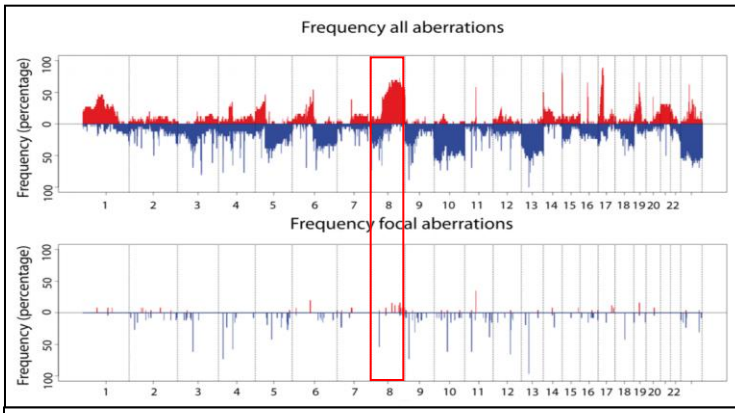

Fig. 2 – Demonstrates the correlation of high levels of copy number aberration (CNA) activity with the overall high rates of mutations in genes in an analyzed string of DNA induced to modify itself.

Problem Statements

P1: Can an established correlation between CNA patterns and tumor oncogenes expression allow for the elimination of invasive breast cancer diagnostic procedures such as biopsy and mammography?

P2: Can the CNA pattern also determine the most effective treatment for a patient after a positive diagnosis?

Objectives

OBJ1: Use given CNA data to determine the likelihood of a tumor oncogene being expressed or silenced in a patient's genomic data.

OBJ2: Using three other parameters (Age, Tumor Stage at Diagnosis, Survival After Removal) and using a machine learning algorithm, given by the dataset, determine the best treatment for the patient's specific case of breast cancer with the highest likelihood of success.

Hypotheses

H1: It is hypothesized that by analyzing the CNA data of 2,433 breast cancer patients, specific CNA patterns will emerge that are indicative of the presence of tumor oncogenes.

- It is believed that this will happen because in the past, it has been observed that the presence of CNA clusters have been indicative of certain genes (Sun, 2017).

H2: Once the positive tumor oncogenes have been discerned, using the three other parameters will allow for the treatment with the highest likelihood of success to be identified.

- It is believed that this will be possible because the machine learning algorithm can take into account all four parameters and produce a formula that will maximize the likelihood of success for the breast cancer treatment.

This space was intentionally left blank.

Methodology

Role of Mentor

The research plan was developed by the student and revised with the mentor prior to experimentation. The student independently programmed all aspects of the project pertaining to the dataset, including procedures utilizing the software Matplotlib, SciKit Learn, and Tensor Flow at the suggestion of the mentor. The creation of the prediction algorithm was done through the acquired results of the *Matplotlib* analysis by the student.

Formatting 2,433 Samples in Jupyter Notebook

The de-identified and anonymous samples were taken from a publicly viewable database created by Pereira et al. in 2016 as a part of a larger study in cancer related Copy Number Aberrations (CNAs) and refining breast cancer types. Of the 68 physical parameters provided about the patients with cancer, 7 of them were isolated and merged to create each patient's profile. These seven parameters were "Tumor Stage at Diagnosis", "Tumor Size", "Metastatic Activity", "Cancer Type", "Age at Diagnosis", "Breast Surgery" and "Vital Status". For the tumor stage at diagnosis, patients who were diagnosed with grade 3 tumors and above (grades range form 1-5) were taken into account in order to get the maximum number of CNA matches. For metastatic activity, I only took into account patients who were positive for metastases. For cancer type, patients with Invasive Ductal or Invasive Lobular Carcinomas were taken into account due to the fact that these two carcinomas comprised about 95% of the total dataset. All ages above 40 were accepted, along with all tumor sizes, so the algorithm had a wide range of data to "learn" from. Only patients who went through radiotherapy or chemotherapy along with

some type of breast surgery were accepted in order to be able to analyze which therapy was the most effective for each CNA pattern.

Choosing Mutagens with Matplotlib

Matplotlib allows for analysis of two parameters in relation with each other. This is done through the creation of a logistic regression line. Each point is plotted on a cartesional plane with an (x, y) coordinate. Each patient has a numerical or (yes/no) response for each parameter. If the parameter was yes/no, the numerical values of 0 and 1 were assigned to yes and no respectively. The goal was to choose genes whose mutations were highly correlated with deadly tumors, so all patient data (2433 profiles) were used. The "X" or "independent" variable was the presence of the chosen tumor oncogene (HERC2, PLY1, ERBB2 etc.), while the other was the "Tumor Grade" (intensity ranging from 1-5). The more correlated the two parameters are, the more they will congregate around the "line of best fit" that is drawn between the two values using a regression algorithm.

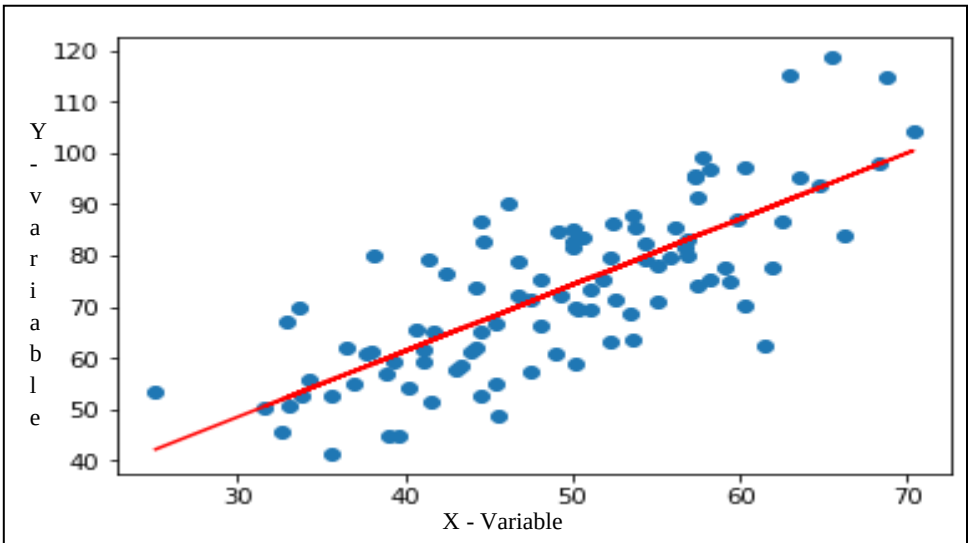

Fig. 3 – Demonstrates a regression line made on Matplotlib that tests the correlation between two parameters. The majority of the dots gathered around the red line suggest a high level of correlation between the two values due to the low residual rate. [res = predicted-observed]

Genes with high levels of correlation to tumors grade 3 or higher (grades range from levels 1-5) are chosen to undergo CNA analysis in the next protocol due to the fact that their presence will most likely lead to the growth and progression of a large and deadly tumor. By catching the presence of these genes through CNAs, it is possible to catch tumors earlier.

Preliminary CNA Analysis with SciKit Learn

SciKit Learn is a prediction model that utilizes previous data in order to state the likelihood of an event occurring in the future. The Copy Number Aberration (CNA) data comes in series, with an activity range of -2 to 2. -2 demonstrate a CNA double deletion of the gene, while -1 is a single deletion. 0 is an indicator that no disturbance had occurred with the patients DNA. 1 and 2 demonstrate emulsifications of 1x and 2x respectively of the specific gene. Each patient has 22,545 relevant pieces of CNA data. The relevancy of the CNA data was decided through the spot analysis of being in a 10,000 base proximity to the gene location on the DNA. The learning library of SciKit Learn will take into account the patient's positive genes and read each piece of CNA data across all series. While examining the 1,965 relevant samples that were relevant after the filtering process done through Jupyter Notebook, the prediction model of the CNA analysis was more and more accurate. Different prediction models are created based upon the positive genes found in the sample.

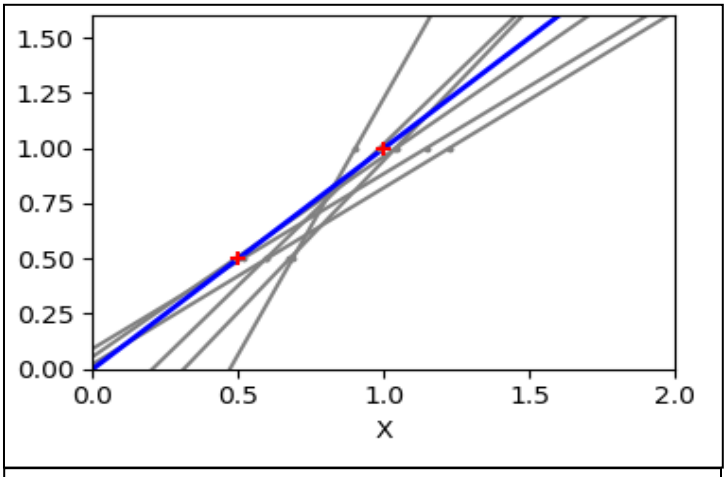

Fig. 4 – In grey, the load sets are plotted in order to show the algorithm learning the trend between x and y. The line in blue shows the most likely trend to be seen when comparing the two parameters.

Predicting Likelihood of Getting Breast Cancer with a Prediction Algorithm

The Copy Number Aberration (CNA) sequence determined from previous CNA analysis that yielded high levels of correlation to the presence of known tumor oncogenes was used to make a simple genetic algorithm. The known CNA levels (reported from -2 to 2) from 200 control samples without breast cancer are averaged for each CNA value of the sequence (5 in total) and used as the "baseline" score. If a patient has a higher level of amplification for one of the CNAs (indicated by a higher positive number), it increases the likelihood of their developing a type of breast carcinoma, which is reflected in their "likelihood" percentage given by the algorithm. Two different carcinomas were accounted for in this study, so the scope of the algorithm could only predict the patient's likelihood of getting either Ductal or Lobular Carcinoma. By measuring the patient's distance from the "baseline" their distances are squared and then added together and divided by 4 (due to there being 5 CNAs in the sequence) and multiplied by 100 in order to get the percentage of likelihood of developing a certain type of breast carcinoma.

This space was intentionally left blank.

Results

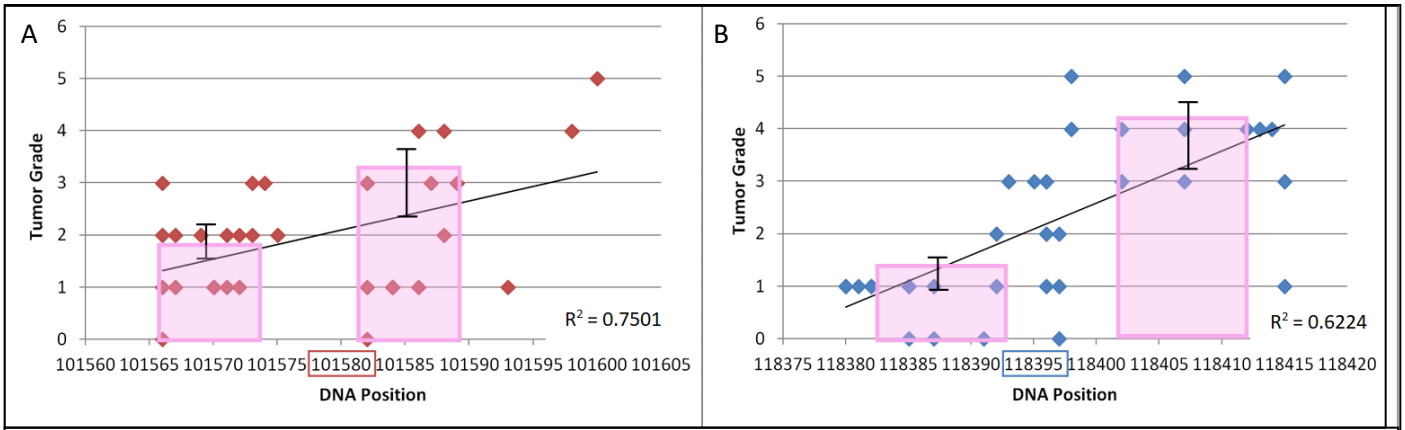

Fig. 5 – **(A)** The tumor grade increases from clustering around tumor grades 1-2 after reaching the 101580th position on the DNA, at which the expression of HERC2 begins, denoted by the red rectangle. The pink boxes take into account the clusters of data and the error bars mark the first and third quartiles. The black line is the linear regression line, relating the expression of HERC2 to tumor grade. **(B)** The tumor grade experiences increases from clustering around tumor grade of 1 after reaching the 118395th position on the DNA, at which the expression of ERBB2 begins, denoted by the blue rectangle. (Figure by G. Ratakonda)

HERC2 and ERBB2 are key regulators of metastatic breast cancer development.

It was observed through *Matplotlib* analysis that the presence of HERC2 is highly correlated with an elevated grade and rate of metastases. The Human Epidermal Radioactive Complex 2, also known as the HERC2 gene belongs to the Human Epidermal Radiation Complex (HERC) family. It is derived from the Human Epidermal Growth Factor Receptor (EGFR) family after undergoing radioactive mutations, and is first located at the 101580th position of DNA [Figure 5A]. The presence of the HERC2 gene indicates that the Her2 gene has entered its malignant form and has been proven in this study to be regulated with tumor grades level 3 and up. Tumor grades are ranked from levels 1-5, with level 3 being the benchmark for the beginning of metastases. Level one is cell abnormalities with odd cell build ups and tumors ranging from 3cm to 4cm at the tumor's widest diameter. Level two is mild cell buildups with tumors ranging from 5cm to 8cm. Level three describes multiple tumors at varying sizes that have begun to metastasize. Level four and five are the same as level three with increasing levels

of intensity in tumor size and expansion level. R^2 is a value determining how much of the variation in tumor grade is accounted for by the linear regression model shown in Figure 1A. Seeing as 75% of the variation is accounted for by the model, it means that the level of expression of HERC2 is strongly correlated with the increasing grade of tumor intensity. A majority of the data in Figure 1A can be sorted into 2 observed clusters, shown by the pink boxes. The first cluster was of relatively low tumor grade levels with a median of 1.83 out of 5, and the other cluster was of metastatic tumors, with a median of 3.23 out of 5. It was seen that that the increased presence of HERC2 expression was correlated to the growth of tumors grades 3 and higher ($p > 0.05$). ERBB2 is a gene that works in tandem with the HERC2 family (Dankort et al., 2017). It is known in full form as the Erythroblastic oncogene B and is part of another oncogene family, and is located at the 118395th position on the DNA strand [Figure 5B]. Utilizing *Matplotlib* for regression data, it was seen that ERBB2 has an R^2 value of 0.6224. Seeing as 62% of the variation is accounted for by the model, it can be inferred that the level of expression of HERC2 is has a moderately strong, positive, linear relation to tumor grade.

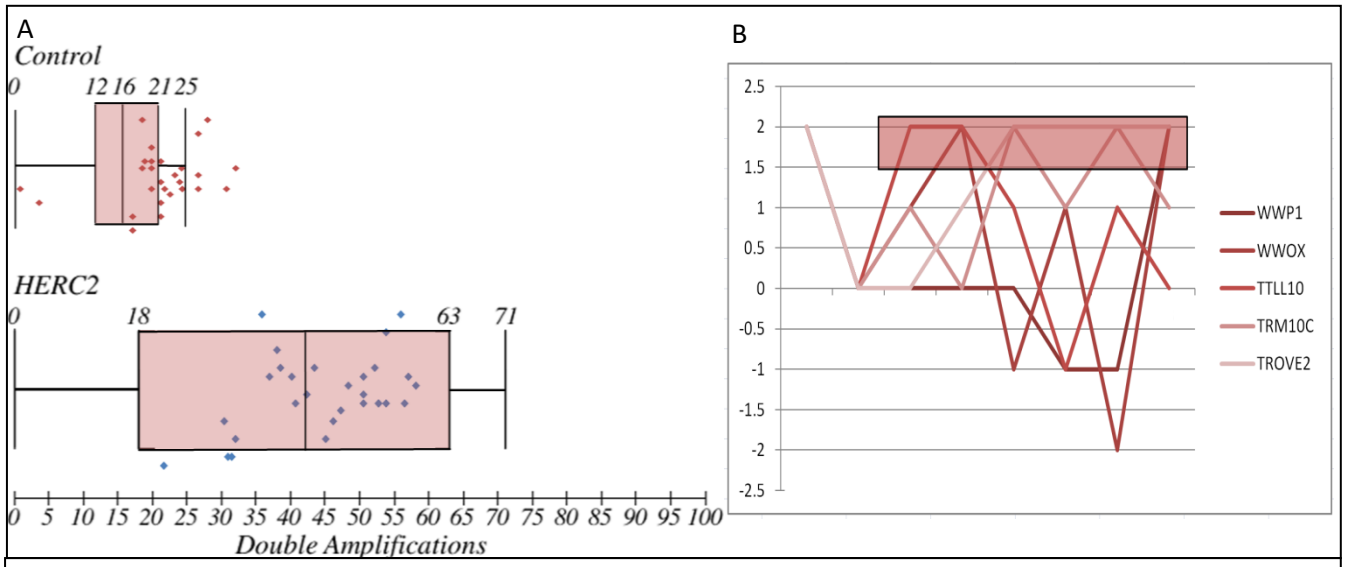

Fig. 6 – (**A**) This top of this figure depicts the control patients with no HERC2 active in the body's normal number of double amplifications resulting from the presence of CNAs. This is compared to the plot below in which the patients with HERC2 active have a lot more double amplifications occurring on average. (**B**) This depicts what the WWTR CNA sequence looks like in patients with HERC2 marked as positive. The shaded box highlights the double amplifications that were seen to be higher for HERC2 patients in part A. (Figure by G. Ratakonda)

The WWTR CNA sequence is correlated with the presence of HERC2 and other HER family genes.

The CNAs, or copy number aberrations, were marked within the dataset with a value of -2 to 2. -2 denoted a double deletion of a gene, -1 meant a single gene deletion, 0 meant no change in the gene expression, 1 meant a single amplification of the gene, and 2 meant a double amplification. -2 and 2 both show high levels of correlation with the presence of genes (Pereira, 2016). It was seen that in the presence of HERC2, the level of double amplifications were increased by 125% from the total control. The WWTR CNA sequence was hypothesized to be the sequence that HERC2 was correlated to because of the high number of double amplifications seen in the CNA data in Figure 2A. This sequence is comprised of the CNAs WWP1, WWOX, TTLL10, TRM10C and TROVE2 [Figure 6B]. It was seen on average, that the patients with HERC2 present in their oncogenic data had a significantly higher level of 2's (higher level of amplification) within their CNA data within this sequence, with a $(p < 0.05)$. This was observed first in the patient data that showed a high number of double amplifications for the WWTR sequence in the area of HERC2 activity in comparison to the control [Figure 6A], which is what tied the WWTR sequence to the HERC2 gene.

The prediction algorithm created based upon HERC2 detection is 63% accurate in detecting ductal carcinoma.

In order to detect breast cancer non-invasively, it was necessary to establish a significant Copy Number Aberration (CNA) sequence that was correlated with the presence of a certain tumor oncogene. It was found that the WWTR sequence could determine whether the known tumor oncogene, HERC2, would end up positive or negative within the patient [Figure 6B].

Through testing the other 1,966 profiles for the presence of HERC2 using their WWTR sequence data, it was found that the algorithm produced was 63% accurate in determining whether or not the patient would contract ductal carcinoma, which is the most common form of breast cancer that affects patients ages 30-65.

Discussion

Matplotlib is a regression tool that allows analysis of two vectors in comparison to each other. It was first utilized in order to compare different tumor oncogenes to tumor grade in order to determine which had the highest level of correlation to metastatic tumor grades (grade 3+ out of 5). It was seen that out of the 357 total chosen genes, 16 of them were significantly correlated to tumors of grades 3 and above [Figure 5A]. Of these genes, the focus of the study became the presence of the HERC2 gene. It was correlated to increased metastatic activity and decreased cell function (Zhang, 2018). HERC2 was seen to have a higher level of "double amplified genes" than the control, which was realized as a result of the *SciKit Learn* analysis. The presence of HERC2 was then tested to determine whether any Copy Number Aberration (CNA) sequences responded with a pattern of high double amplifications. It was seen that the WWTR CNA sequence was highly correlated to the presence of double amplifications, and a pattern was seen when looking at the actions of the WWTR sequence in the area of HERC2 activity [Figure 5B]. Once this correlation was established, the algorithm for breast cancer detection was made through the utilization of *SciKit Learn*. Through utilizing a patient's genomic data, it is possible to extract the data regarding the WWTR CNA sequence through DNA splicing. When testing patient's from the dataset, it was seen that the accuracy of the algorithm was 63%, which is comparatively higher than previous models made, which only had an accuracy of around 30-40% (Feizi, 2018).

Application

There were several recurrent sources of error in the duration of the experiment, many of which can be eliminated with further testing. The age group of the individuals was highly varied, from 30-65. The immune system of a 30 year old and that of a 65 year old have the potential to be incredibly different, depending on various health choices undertaken during the course of the patients' lives. These choices were not reflected in the database on account of potentially introducing bias and decreasing the possible sample size significantly. However, the preliminary findings in this study can be applied directly to larger databases with more parameters that account for the disparity in immune systems of different women. Having multiple tumors also brought the overall tumor grade up per patient, which is a traditional aspect of Invasive Ductal Carcinoma, even though the tumors themselves may not be malignant. The research ignoring number of tumors can be refined in the future by adding a parameter that reports how many tumors are present within each patient pre and post – diagnosis.

Future Research

In the final algorithm, an accuracy of 63% was achieved. With the refining of more CNA sequences, there may be established correlations present with other tumor oncogenes. Gaining more insight on the issue of breast cancer diagnosis with noninvasive diagnostic methods will allow future patients a plan in that they have time to prepare for when they do contract cancer, and can take preventative measures like surgery while the survival rates are still high. Further applications of this research can also help define the most effective breast cancer treatment depending on which active CNA sequences are present within the patient's genomic data.

Conclusion

Breast cancer is a deadly disease and one of the leading causes of early death for women globally. While it affects approximately 1 in every 8 women, breast cancer today is still seen to be one of the “most treatable” cancers. However, the overall survival rates are highly optimistic and drop dramatically when the cancer that is diagnosed is already metastatic, which accounts for approximately one third of all breast cancer diagnoses. In order to find a way to combat this issue while also noninvasively diagnosing breast cancer, data was extracted from a public database. CNAs are changes made to the DNA that alter the number of times a gene can be expressed and they occur everywhere in the body, meaning that even samples of a patient’s saliva will have their CNA data. It was hypothesized that the presence of certain patterns of CNAs would allow for the detection of certain tumor oncogenes, which in turn would allow for the detection of breast cancer. It was found through the use of Matplotlib regression data that a correlation existed between the activities of the CNA sequence WWTR was correlated with the presence of well known tumor oncogene, HERC2. When building the algorithm, the only inputs needed were the five CNA amplification numbers ranging from -2 to 2. From entering these in, we were able to detect the presence of ductal carcinoma with an accuracy of 63%, which is 1.5 times higher than preexisting non-invasive methods of diagnosing breast cancer, which were only accurate 53% of the time (Chi, 2014). In the future, improvements can be made on with the algorithm now that initial correlations between the presence of tumor oncogenes and CNA sequences have been established. With these alterations, this free-to-run method of breast cancer detection can be universalized and help save the lives before cancer even becomes an issue within their lives.

Works Cited

1. Alonso, M. H., Aussó, S., Lopez-Doriga, A., Cordero, D., Guinó, E., Solé, X., . . . Moreno, V. (2017, July 06). Comprehensive analysis of copy number aberrations in microsatellite stable colon cancer in view of stromal component. Retrieved from https://www.nature.com/articles/bjc2017208
2. Chi, C., Murphy, L. C., & Hu, P. (2018, January 29). Recurrent copy number alterations in young women with breast cancer. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5837756/
3. Hegselmann, S., Gruelich, L., Varghese, J., & Dugas, M. (2018, November 29). Reproducible Survival Prediction with SEER Cancer Data. Retrieved from http://proceedings.mlr.press/v85/hegselmann18a.html
4. Hieronymus, H., Murali, R., Tin, A., Yadav, K., Abida, W., Moller, H., . . . Sawyers, C. L. (2018, September 04). Tumor copy number alteration burden is a pan-cancer prognostic factor associated with recurrence and death. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6145837/
5. Pereira, B., Chin, S., Rueda, O. M., Vollan, H. M., Provenzano, E., Bardwell, H. A., . . . Caldas, C. (2016, May 10). The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. Retrieved from https://www.nature.com/articles/ncomms11479
6. STAFF, C. (2019, February 28). Breast Cancer - Statistics. Retrieved from https://www.cancer.net/cancer-types/breast-cancer/statistics
7. Both, J., Krijgsman, O., Bras, J., Schaap, G. R., Baas, F., Ylstra, B., & Hulsebos, T. J. (2014, December 31). Focal Chromosomal Copy Number Aberrations Identify CMTM8 and GPR177

as New Candidate Driver Genes in Osteosarcoma. Retrieved from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0115835

8. Chung, J. H., Zhang, Y., & Bunz, F. (2010, June 01). Checkpoint bypass and cell viability. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/20505362
9. Osborne, C., Wilson, P., & Tripathy, D. (2004). Oncogenes and Tumor Suppressor Genes in Breast Cancer: Potential Diagnostic and Therapeutic Applications. Retrieved from http://theoncologist.alphamedpress.org/content/9/4/361.full.pdf html
10. PHS, S. (2017). Ask an Expert: Breast cancer growth rate. Retrieved from https://oregon.providence.org/forms-and-information/a/ask-an-expert-breast-cancer-growth-rate/
11. Song, B., Wang, L., Zhang, Y., Li, N., Dai, H., Xu, H., . . . Yan, J. (2019, March). Combined Detection of HER2, Ki67, and GSTP1 Genes on the Diagnosis and Prognosis of Breast Cancer. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/30585764