HER2 Biomarker and the Gene Expression Omnibus (GEO) Database

Madeline Barry, Lauren Driscoll, and Aria Moss

Breast cancer is one of the most diagnosed cancers in women in the United States, with 1 out of 8 women being diagnosed at some point in their life [1]. To effectively treat breast cancer, it is imperative that the diagnosis be made early, potentially by preemptively screening women who may be at high risk. To identify these high-risk patients, genetic biomarkers can be used to identify when there are genetic abnormalities relating to a cell's estrogen receptors (ER), progesterone receptors (PR), or human epidermal growth factor receptor 2 (HER2) [2]. This project focuses on HER2-positive (HER2+) patients, or those with an overexpression of the HER2 gene. HER2 makes receptor proteins in breast cells, and when it is overexpressed, it can drive cell growth and proliferation. This, in conjunction with the reduction in anti-apoptotic signals, can lead to cancerous tumor growth. Identifying this biomarker in a clinical setting, frequently by taking a direct tissue sample and undergoing gene amplification, has both predictive and prognostic value. It can help determine not only a patient's likelihood of developing breast cancer, but also impact potential treatment plans [3].

Gene Expression Omnibus (GEO) is a free database composed of high-throughput genomic data submitted by researchers. It has a wealth of data on various genes and their gene expression profiles with the aim of condensing research from a wide array of related experiments and putting the quantitative results into one database. Learning to use databases like GEO and analyzing the data can be a critical resource in modern medicine. The enormous amount of data available can be used to do more research on biomarkers such as HER2 which can be imperative to understanding, predicting, and treating illnesses such as breast cancer. It is a key tool that can be important for innovation in medicine and potentially alter the approach physicians take in a clinical setting. For HER2 specifically, it is a gene that can be used as a biomarker. A biomarker is a gene that can be used to give a snapshot in time of what is going on biologically within a system. These biomarkers can help in real-time to give more information on a patient's condition. When data is collected on them and uploaded into a database such as GEO, improvements can be made in the treatment and diagnosis of conditions associated with that biomarker.

To learn how to use the database, the GEO website has detailed information, though learning materials can also be easily acquired from other sources such as NIH, university websites, and manufacturers such as ThermoFisher Scientific. All the data found in GEO is originally submitted by researchers. Each sample submitted is from one experiment and includes a description of how it was processed and the results. When there are enough samples on the same subject, the experiments were completed using the same technology and conducted in a similar fashion, they are combined into data series. These series are then processed and revised by employees at GEO, summarizing and approving the information in a series, making it a dataset that can be used for gene profiles. Due to the high volume of data collected, the datasets in the Gene Expression Omnibus are more efficient than analyzing results one experiment at a time.

To choose a dataset for further analysis, important variables to consider include the source of the data, the sample size, and the technology involved in the original research. The dataset analyzed here is labeled as GDS5027, which is comprised of the series GSE50948, a set of 156 *Homo sapiens* formalin-fixed, paraffin-embedded core biopsy samples used in a study by Prat et al. [4]. 114 samples are from HER2+ patients and 42 samples are from HER2- patients. It was expected that the HER2 biomarker would be overexpressed in HER2+ patient samples, identifiable by the fold change in different HER2+ probes.

The reporter for this sample is the Affymetrix Human Genome U133 Plus 2.0 Array. This is a microarray chip that analyzes gene expression. After a biopsy is taken and RNA is extracted from the sample, it undergoes an amplification process called a polymerase chain reaction (PCR). During PCR, fluorescent tags are added to the amplified fragments of RNA, which are then washed over the chip. The chip has zones with "probes," short segments of DNA that the RNA attaches to if it is complementary. By knowing which probes correlate with which genes in the human genome, the fluorescence of each probe can be measured, which provides information on how much RNA was in the sample. Thus, the level to which genes were expressed could be detected and compared to the other genes on the chip. We chose three HER2 probes to use (210930_s_at, 216836_s_at, and 234354_x_at) for this analysis, picking the three that were related to our gene of interest and had the greatest fold change between the positive and negative groups, based on a p-value adjusted for multiple comparisons.

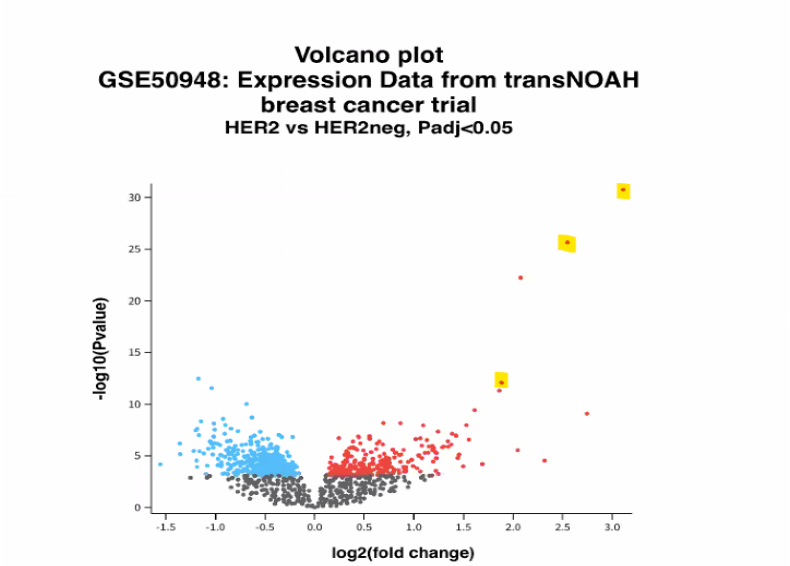

Figure 1: The volcano plot is the log transformation of the fold change and P-Value of the different probes used in this study. The highlighted data points represent the three probes of interest in our analysis of the HER2 gene. These probes were selected because of their P values and fold change values which indicated an over expression of the HER 2 gene.

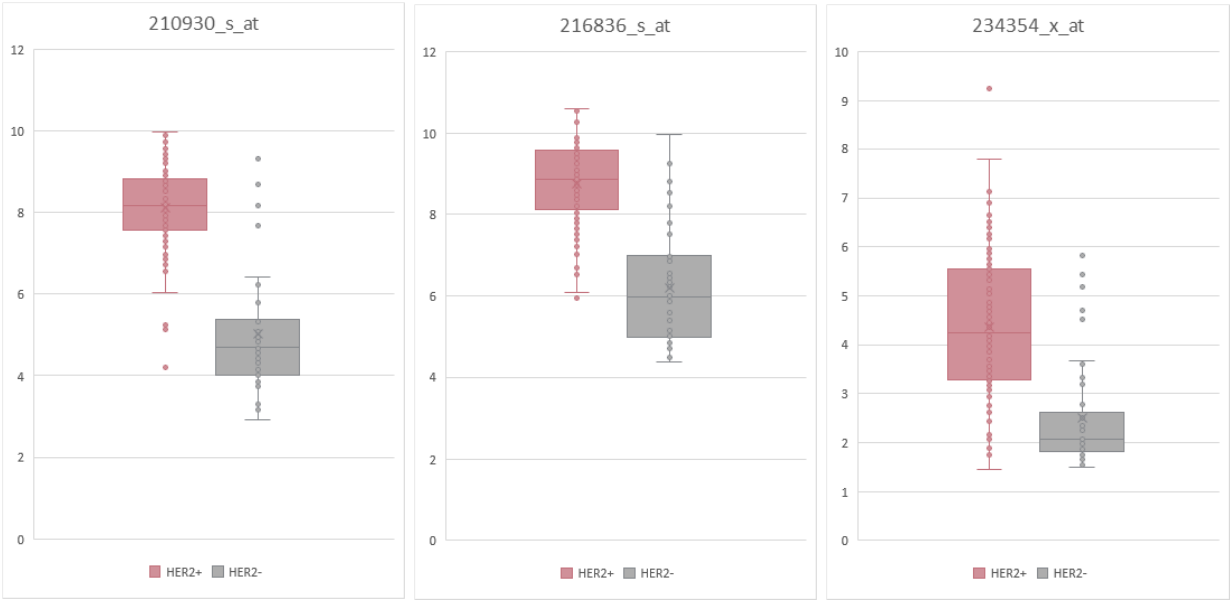

Figure 2: The data from the 210930_s_at, 216836_s_at, and 234354_x_at probes have been separated into two groups, HER2 positive (+) and HER2 negative (-). The HER2 + group is shown to have a higher expression value across the three different probes. The boxes represent the interquartile ranges, with each individual points representing one patient. Points outside the range of the plots are considered outliers, which have values that deviate from the first or third quartile by at least 1.5 times the interquartile range.

The expression of the HER2 gene is elevated in the HER2+ population when compared to the HER2- population in all analyzed probes. For each probe, we compared the two populations using a two-tailed t-test assuming unequal variance with an alpha value of 0.05. For the probes 210930_s_at, 216836_s_at, and 234354_x_at, the p-value produced by the t-tests was 3.91E-17, 4.05E-15, and 3.49E-13, respectively. These are all well below our alpha value of 0.05, indicating that our results are statistically significant. Furthermore, figure 2 uses box and whisker plots to illustrate similar trends for each of the three HER2 probes. These expression patterns are reasonable when compared to scientific literature; HER2 was first found to have a correlation with human breast cancer in 1987 by Slamon et al. [5], who have been cited throughout the years since [6, 7, 8]. An article by Gutierrez & Schiff states that overexpression of HER2 is correlated to “tumor development and progression for this subset of breast cancer,” and indicates higher mortality and lower time to relapse on average [9]. The literature suggests that HER2 overexpression is common in about 30% of cancer patients, so the conclusion that patients with HER2+ tumor growth show a higher expression of HER2 on average than those with HER2- tumors is reasonable.

A few limitations and considerations must be considered while analyzing the data obtained from the GEO database. Data obtained from different probes can not be combined and averaged due to the different specificity of each probe. Since each probe is designed to bind with unique segments of DNA or RNA, the rate and frequency of the binding can not be combined. However, multiple probes can show a significant trend and provide evidence to support the conclusions further. Additionally, it is important to recognize that these samples are paraffin-embedded, which can cause degradation of the RNA and may cause biases in the data. However, the significant differences between the HER2+ groups and HER2- groups across multiple probes still strongly suggest that there are marked differences in HER2

expression. Despite these considerations, this analysis demonstrates that measuring HER2 gene expression through RNA analysis is a good indication of when a patient has developed HER2+ tumors as opposed to HER2- tumors. With this technology, treatment plans can be better tailored to patients and overall create a higher level of quality healthcare for those at risk or diagnosed with breast cancer.

References

[1] *Breast Cancer Facts and Statistics*. Breastcancer.org. (2022, March 10). Retrieved April 8, 2022, from https://www.breastcancer.org/facts-statistics

[2] *Breast cancer biomarkers*. Breast Cancer Biomarkers | Choose the Right Test. (2022, March). Retrieved April 8, 2022, from https://arupconsult.com/content/breast-cancer

[3] Bertozzi, S., Londero, A. P., Seriau, L., Vora, R. D., Cedolini, C., & Mariuzzi, L. (2018, November 5) *Chapter: Biomarkers in Breast Cancer*. IntechOpen. Retrieved April 8, 2022, from https://www.intechopen.com/chapters/62000

[4] Prat, A., Bianchini, G., Thomas, M., Belousov, A., Cheang, M. C. U., Koehler, A., Gómez, P., Semiglazov, V., Eiermann, W., Tjulandin, S., Byakhow, M., Bermejo, B., Zambetti, M., Vazquez, F., Gianni, L., & Baselga, J. (2014, January 16). *Research-based PAM50 subtype predictor identifies higher responses and improved survival outcomes in HER2-positive breast cancer in the noah study*. American Association for Cancer Research. Retrieved April 8, 2022, from https://aacrjournals.org/clincancerres/article/20/2/511/78518/Research-Based-PAM50-Subtype-Predictor-Identifies

[5] Slamon, D. J.; Clark, G. M.; Wong, S. G.; Levin, W. J.; Ullrich, A.; McGuire, W. L. (1987, January 9). *Human breast cancer: Correlation of relapse and survival with amplification of the HER-2/neu oncogene*. Science (New York, N.Y.). Retrieved April 8, 2022, from https://pubmed.ncbi.nlm.nih.gov/3798106/

[6] Yarden, Y. (2001). *Biology of HER2 and its importance in breast cancer*. Oncology. Retrieved April 8, 2022, from https://pubmed.ncbi.nlm.nih.gov/11694782/

[7] Carney, W. P. (2005, November 2). *HER2 status is an important biomarker in guiding personalized HER2 therapy*. Personalized medicine. Retrieved April 8, 2022, from https://pubmed.ncbi.nlm.nih.gov/29788570/

[8] Moasser, M. M. (2007, October 4). *The oncogene HER2: Its signaling and transforming functions and its role in human cancer pathogenesis*. Oncogene. Retrieved April 8, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3021475/

[9] Gutierrez, C., & Schiff, R. (2011, January). HER2: Biology, detection, and clinical implications. Archives of pathology & laboratory medicine. Retrieved April 8, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3242418/

[10] *Structure & function of GeneChip Microarrays*. (n.d.). Retrieved April 9, 2022, from https://www.csus.edu/indiv/r/rogersa/bio181/genechipho.pdf