

Relating Gene Expression Profiling and Copy Number Changes with Different Sub-types
of Breast Cancer

AS
36
2016
MATH
.P45

A thesis presented to the faculty of
San Francisco State University
In partial fulfillment of
The Requirements for
The Degree

Master of Arts
In
Mathematics

by

Rachael Phillips

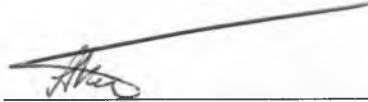
San Francisco, California

May 2016

Copyright by
Rachael Phillips
2016

CERTIFICATION OF APPROVAL

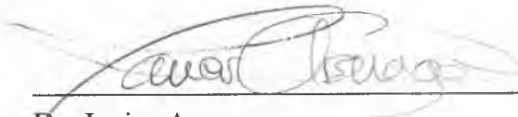
I certify that I have read *Relating Gene Expression Profiling and Copy Number Changes with Different Sub-types of Breast Cancer* by Rachael Phillips and that in my opinion this work meets the criteria for approving a thesis submitted in partial fulfillment of the requirements for the degree: Master of Arts in Mathematics at San Francisco State University.



Dr. Alexandra Piryatinska
Associate Professor of Mathematics



Dr. Joseph Gubeladze
Professor of Mathematics




Dr. Javier Arsuaga
Professor of Mathematics

Relating Gene Expression Profiling and Copy Number Changes with Different Sub-types
of Breast Cancer

Rachael Phillips
San Francisco State University
2016

Throughout the lifespan of the human body, trillions of cell divisions occur. Every time a cell divides, a regulatory process can fail causing alterations to the DNA. These alterations can cause diseases that include different types of cancer. Our work focuses on different sub-types of breast cancer such as Luminal A, Luminal B, Basal (Triple Negative), and HER2; molecular classifications that have been found through biological discovery. In our study, we focus on two different data sets: copy number changes and gene expression profiling. This data helps detect important changes in the breast cancer tissue. We apply methods such as persistent homology and correlation to confirm with Chin et al. the regions where copy number aberrations are affecting the expression of certain genes.

I certify that the Abstract is a correct representation of the content of this thesis.

 / ALEXANDRA PRYATINSKA, 05/24/2016

Chair, Thesis Committee

Date

ACKNOWLEDGMENTS

I would like to personally thank my family and friends for supporting me through this endeavor. My sister, Ginger Gaskill for always supporting me and helping me through some tough times. My mother for pushing me to be better and showing me how proud she is. My father for being so kind and supportive by telling me that I can do anything that I set my mind to! My brother Nick Voorhees and my sister-in-law Courtney Voorhees for letting me stay at their home while I finished my thesis. I would also like to thank my friends Patrick, and Holly for being two of my best friends and being there when I needed them the most! Madhu, Stephanie, Jeff and John for always supporting me and helping me stay sane! Wiseley, Judy, Michelle, Robert and Maxime for helping me with biology, programming or mathematics! I would mostly like to thank my advisor Dr. Javier Arsuaga for all his help during these last 3 years. He has been supportive, kind and very generous! Thank you Dr. Arsuaga for helping me get through this and for being such a great influence on my life, I couldn't have done it without you!!! Dr. Piryatinska at SFSU for being helpful and kind during my undergraduate and graduate programs, you have been an inspiration to me! Thank you all for your love and support for the past 3 years, I am eternally grateful, I couldn't have done it without you!

TABLE OF CONTENTS

1	Introduction	1
2	Biology Background	4
2.1	Biology of Breast Cancer	4
2.1.1	Breast Anatomy and Types	4
2.1.2	Clinical Characteristics of Breast Cancer	5
2.1.3	Molecular Subtypes of Breast Cancer	6
2.2	Epigenetic and Genetic Changes	13
2.2.1	Gene Expression	14
2.2.2	Copy Number Changes (aCGH)	15
2.2.3	DNA Microarray	17
2.2.4	Relationship between aCGH and Gene Expression	19
2.2.5	Tumor Suppressor Genes and Oncogenes	19
3	Data and Methods	21
3.0.1	Data Processing	22
3.0.2	Assigning genes to clones	25
4	Methods	35
4.1	Method from Chin et al.	35
4.1.1	Frequencies	36
4.1.2	Clustering	40

4.1.3	Regression	40
4.1.4	Association of Copy Number and Gene Expression	41
4.2	Our methods to verify Chin et al.	41
4.2.1	Frequencies	41
4.2.2	Clustering	42
4.2.3	Association of Copy Number and Gene Expression	42
4.3	Finding Significant Regions in the Genome	45
4.3.1	Segments, Sliding Window, and Point Cloud	46
4.3.2	Connected Components and Filtration	49
4.3.3	Choosing Filtration Parameter	50
4.3.4	Betti Curves	52
4.3.5	Permutation Testing	53
4.3.6	Correlation with Significant Regions	55
4.3.7	Pearson's Product/-Moment Correlation	59
4.4	Adjusting P-values	62
4.4.1	Holm Adjusted P-value	62
4.4.2	Bonferroni	63
5	Results and Discussion	64
5.1	Results for Chin et al.	64
5.1.1	Frequencies	65

5.1.2	A Correlation Study: Copy Number and Gene Expression in Four Chromosomes	74
5.1.3	Clustering with Intrinsic Genes	92
5.1.4	Conclusion	96
5.2	Results using persistent homology	97
5.2.1	Significant regions	97
5.2.2	Betti Curves	101
5.2.3	Correlation Between Copy Number and Averaged Gene Expression	106
5.2.4	Conclusion	124
6	Appendix A	126
	Bibliography	134

LIST OF TABLES

Table	Page
2.1 Molecular Classifications of Breast Cancer	12
2.2 Differences in two DNA microarray experiments	18
3.1 Phenotypes Studied	23
3.2 Copy Number Data	24
3.3 Gene Expression Data	24
4.1 Chin et al. Increases and Decreases for Molecular Subtypes of Breast Cancer	39
4.2 Number of Genes	44
5.1 Correlation of Genes and Copy Number from our study	87
5.2 Correlation of Genes and Copy Number from our study continued	88
5.3 Significant regions for each subtype	99
5.4 Number of Copy Number Clones and Averaged Genes found correlated .	112
6.1 First 180 intrinsic genes from Stanford	127
6.2 Second 180 intrinsic genes from Stanford	128
6.3 Third 144 intrinsic genes from Stanford	129
6.4 Last 48 intrinsic genes from Stanford	130
6.5 66 Genes found Correlated with Copy Number from Chin et al.	132
6.6 Genes Found Correlated with Copy Number	133

LIST OF FIGURES

Figure		Page
2.1	Average hierarchical clustering of patients and published markers of breast epithelial subtypes found in Chin et al. [7]. Green indicates lower expression levels, while red indicates higher expression levels. Genes are seen on the right and the bottom contains tumor samples. Subtypes of the 95 patients; erBb2, Basal, Luminal A, Luminal B, and Normal-Like are represented in colors green, red, light blue, orange and pink, respectively; these can be found at the top of the figure. ER and PR status are represented in rows at the top of the figure as well. Dark blue indicates negative ER/PR status and light blue indicates positive status.	8
2.2	Average hierarchical clustering for copy number in the entire genome. Green indicates high copy number value while red represents low copy number. To the left, chromosomes are colored in black and grey; black indicates odd chromosomes, grey is for even chromosomes. Chromosomes are in numerical order from 1 (top) to 23 (bottom). At the top, we see the same ER,PR and Subtypes rows as described for Figure 2.1.	11
2.3	Copy number variations [14]	16
2.4	DNA microarray process [24]	18
3.1	Algorithm for assignment of clones to genes based on location in the genome.	27

3.2	Scatter plots for chromosome 17q in bins for a single HER2 tumor. A, B and C represent bin 1, bin 2 and bin 3 for 17q, respectively. The x -axis represents copy number and the y -axis represents gene expression.	28
3.3	aCGH profile (1), gene expression profile (2) and averaged gene expression profile (3) for a Her2 patient in 17q. x -values represent the bp position in the chromosome arm 17q, while y - values are the copy number, gene expression and averaged gene expression, respectively. Gene expression probes were averaged for all genes assigned to the same clone.	30
3.4	Profiles HER2 Patient Point Cloud before Averaging Genes. x -values in this figure are copy number values for (1) in Figure 3.3; y -values are gene expression values for profile (2).	32
3.5	Profiles HER2 Patient Point Cloud after Averaging Genes. x and y values, similar to Figure 3.4 are from profiles (1) and (3) above.	33
4.1	Subdividing one chromosome arm into segments	47
4.2	Sliding window method for patient b0243 patient in 13q	48
4.3	Boxplot of copy number over all patients.	51
4.4	Betti curve for basal vs. others for 4q segment 7	53
4.5	Correlation between clones and averaged genes for 4q segment 7	56
4.6	Sections studied for high correlation	57
4.7	Correlation between clones and genes from the diagonal of Figure 4.5	58
4.8	Strong Positive Correlation	60

4.9	Strong Negative Correlation	60
5.1	Frequency of Copy Number Values for basal subtype. The x -axis displays the BP position along the entire genome, while the y -axis is the frequency for each copy number clone. Frequency for each clone is calculated by the number of tumor samples above a 0.2 threshold for amplifications and below -0.2 for deletions, divided by the total number of tumor samples. Black lines indicate where the chromosomes begin and end; black dotted lines imply the position of the centromeres of each chromosome. Chromosomes are located from left to right in order from 1 to 23.	66
5.2	Frequency of Copy Number Values for HER2 patients. Figure 5.2 shows increases and decreases in copy number based on location in the genome. Similar to Figure 5.1, solid black lines indicate where the chromosomes begin and end, while dotted black lines are representations of the centromeres for each chromosome.	68
5.3	Frequency of Copy Number Values for Luminal A. Black straight and dotted lines are the same as Figures 5.1 and 5.2 above. Here we have the gains and losses from the luminal A patients.	70
5.4	Frequency of Copy Number Values for Luminal B.	71
5.5	Frequency of Copy Number Values over all patients. Similar to Figures 5.1 - 5.4, this is a frequency graph for all clones, however, this graph represents all tumor samples.	73

5.6	Heat map of correlation for bin 1, chromosome 8p11 - 12. Dark red indicates a higher, positive correlation (near 1); darker blue implies a stronger, negative correlation (near -1). Columns are clones and rows are genes. Genes on the left that are labeled numbers are expression probes that did not have the gene name for that probe.	77
5.7	Heat map of correlation for bin 1, chromosome 11q13 - 14.	79
5.8	Heat map of correlation for bin 1, chromosome 17q11 - 12	80
5.9	Heat map of correlation for bin 2, chromosome 20q13	81
5.10	Heatmap sharing Correlation between genes and clones	83
5.11	Heatmap sharing Correlation between genes and clones. Unlike Figure 5.10, Figure 5.11 has some of the genes that are represented by the gene expression probes from Figure 5.10. Since duplicate names are not repeated, we only see the first row for that gene.	85
5.12	Gene TACC1 and copy number clone	89
5.13	Gene ADAM9 and copy number clone	89
5.14	Gene IKBKB and copy number clone	90
5.15	Gene POLB and copy number clone	90
5.16	Gene CCND1 and copy number clone	90
5.17	Gene GRB7 and copy number clone	90

5.18 Hierarchical Clustering with Intrinsic Genes. Subtypes are color coordinated as red (basal), green (ERBB2), light blue (luminal A), orange (luminal B) and purple (normal-like). The subtypes are shown in the row above the heatmap. Above the row of subtypes, are the ER/PR status for each tumor; dark blue and light blue represent negative and positive ER/PR status, respectively. Green in the heat map indicates low copy number and red, high.	93
5.19 Hierarchical Clustering with Intrinsic Genes. Similar to Figure 5.18, 5.19 in hierarchical clustering of “intrinsic” genes. However, in this case, we delete the 33 genes that were found significantly correlated by Chin et al.	95
5.20 Betti curves for Luminal A versus others in 10q segment 2	102
5.21 Betti curves for Luminal B versus others in 8q segment 3	103
5.22 Betti curves for HER2 versus others in 17q segment 1	104
5.23 Betti curves for Basal versus all other subtypes in 11q segment 6. In these figures the blue is the test curve, and the red curve is the control curve. . .	105
5.24 Correlation between copy number (columns) and averaged gene expression (rows) for 22 basal tumors in 2p 25.3 - 11.2. Darker red indicates strong, positive correlation and dark blue indicates strong, negative correlation, white indicates no correlation.	107

5.25	Correlation of Copy Number and Averaged Gene Expression for 17q 11.2 - q21.31 for 15 HER2 tumor samples. Red and blue are the same as Figure 5.24 and rows are averaged genes, while columns are copy number clones.	109
5.26	Correlation of Copy Number and Averaged Gene Expression for 8q 22.2 - 24.3 for 13 luminal B tumor samples. Again, we have red and blue coloring, same as Figures 5.24 and 5.25.	111
5.27	Correlation of Copy Number and Genes for 13 Luminal B patients in 8q 22.2 - 24.3. Dark red is a strong positive correlation, while dark blue is a strong negative correlation. Genes are the rows and copy number clones are the columns.	114
5.28	Correlation of Copy Number and Genes for 15 HER2 patients in 8q 22.2 - 24.3.	116
5.29	Correlation of Copy Number and Genes for 39 HER2 patients in 10q 21.1 - 22.2.	118
5.30	Correlation of Copy Number and Genes for 22 basal patients in 4q 31.21 - 32.3.	120
5.31	Point cloud for basal patients in 8p 21.2 - 11.2	123
5.32	Point cloud for non-basal patients in 8p 21.2 - 11.2	123
5.33	Point cloud for basal patients in 4q 31.21 - 32.3	123
5.34	Point cloud for non-basal patients in 4q 31.21 - 32.3	123

Chapter 1

Introduction

Every person begins as a single cell; the cell that is created during reproduction. This cell undergoes the cell division process and creates trillions of identical cells. When a cell is mature, it goes through a cell cycle to produce its daughter cell. During this cycle a regulatory process may fail, which can cause unregulated cell division; this is the source of many diseases, including cancer.

After lung, breast cancer is the most common type of cancer among women [1]. Therefore, it is important to study the aspects of this disease that effect progression. A way to accomplish this is by studying the instabilities in genetic material that have been altered. Copy number aberrations are one of the causes of these instabilities; changing the DNA during replication by repeating or deleting segments in the genome. Previous studies have shown that copy number changes have led to the expression of genes to be altered [20],[9],[7].

In recent years, array Comparative Genomic Hybridization (aCGH), copy number data, has been used to study these genetic changes. The experiment for aCGH is a technique that counts the number of copies of DNA throughout the entire genome. Normal and tumor samples are extracted and colored to hybridize in an array; providing a method for approximating relative copy number for thousands of DNA sequences.

Gene expression profiling is an experiment similar to aCGH; measuring the amount of cDNA between two samples. For this experiment, two samples of mRNA are extracted from tumor samples and reverse transcribed. The complementary DNA is then placed in an array to be hybridized; producing an amount for the expression of genes in those cDNA sequences. Our goal is to study these two experiments together to locate genes that are regulated by copy number.

Two studies were done for this project; the first includes verifying the results from Chin et al. [7] and the second was to extend the method using persistence homology introduced in Arsuaga et al., [4], with both copy number and gene expression to find genes that are regulated by copy number.

First, we use frequencies of copy number aberrations to verify the same common regions of amplifications in Chin et al.; correlation was then found in these regions between gene expression and copy number. chin et al. found 66 genes in regions 8p11-12, 11q13-14,

17q11-12 and 20q13 that were correlated with copy number. Due to issues with mapping; 43 of the 66 genes found in Chin were the same as our genes and out of those 43 genes, only 29 were found correlated with copy number. This only verified that roughly 64 percent of our results were the same as theirs. Then, using “intrinsic” genes, we used hierarchical clustering analysis without the 29 correlated genes found by our study. This resulted in similar classification to clustering analysis done before the correlated genes were removed.

Secondly, we use β_0 numbers to understand the connectivity of our data of one subtype versus all other subtypes for each iteration of a filtration parameter, ϵ . At each iteration of ϵ , betti curves are calculated and graphed; permutation testing is then done to see if the two curves for one subtype versus all others are statistically the same. If they are not the same, this implies that the genes in this region are either regulated or deregulated by copy number. Once we found the regions that were significant, we then found correlation between genes and copy number in that region.

Significant regions found for all subtypes included 2p, 2q, 3q, 4q, 5q, 11q, and 16q for basal, 10q for luminal A, 8q for luminal B and 17q for HER2. Out of the 66 genes found correlated with copy number, only 2 were found using persistent homology. Five of our genes found using persistent homology were located in the intrinsic gene list from Stanford.

Chapter 2

Biology Background

2.1 Biology of Breast Cancer

Breast cancer is the uncontrollable division of damaged cells in the breast tissue. It is the most common cancer among American women; about 1 in 8 women will be diagnosed with invasive breast cancer in their lifetime [1]. The American Cancer Society estimates that in 2016, 246,660 women will be diagnosed with an invasive breast cancer and about 40,450 women will die from breast cancer [1]. To better understand breast cancer, we will explore biological characteristics that underly this disease.

2.1.1 Breast Anatomy and Types

In healthy female breasts there are 12 - 20 sections called lobes, each made up of several smaller lobules. These lobules are connected by milk ducts and produce milk in nursing mothers. Since breast cancer often begins in these specific locations, they are often

categorized according to the tissue from which breast cancer cells originate. Lobular Carcinoma in Situ (LCIS) is the type of breast cancer that originates in the lobules. The other type, Ductal Carcinoma in Situ (DCIS), originates in the milk ducts. LCIS and DCIS are further categorized by “invasive” and “non-invasive”: non-invasive implies that the breast cancer cells have the potential to spread to other areas of the body, but have yet to do so and invasive cancer cells have spread outside of their origin [23].

2.1.2 Clinical Characteristics of Breast Cancer

In the clinic, breast cancers are classified according to four main characteristics: ER, PR, HER2, and Ki-67 [39]. The receptors, ER and PR, are proteins found in the plasma membrane that respond to the hormones estrogen and progesterone, respectively [39]. An excessive amount of these receptors in breast tissue induces cell proliferation, enabling progression of cell division [39]. For patients with an abnormal amount of ER, a drug called tamoxifen is commonly recommended to target these receptors [8]. This drug acts as a replacement for the estrogen hormone, therefore, preventing the tumor from growing further and in some cases, reducing tumor size [39].

HER2, also called *erbB2* or *Neu*, is an oncogene (a mutated growth-controlling gene) which promotes the health of breast cells. When this gene is in excess, there is uncontrollable growth and division of these cells [5]. To mitigate this abnormality, Herceptin targets all cells, but affects cells with HER2 more [25]. Patients who were prescribed this drug

while undergoing standard chemotherapy treatment had slower disease progression and longer overall survival [39].

The protein Ki-67 is a biomarker in breast cancer for cell proliferation [12]. During cell division, Ki-67 increases, therefore, the higher the concentration of Ki-67 the quicker the cells divide [6]. An excessive amount of this protein is frequently seen in different malignant tissues and associated with worse survival in multiple types of cancers, including breast cancer [12]. A staining process is used to measure the percentage of tumor cells that have positive Ki-67 [6].

These characteristics are used to categorize tumors into distinct molecular classifications. By analyzing these distinctive properties, physicians have a better understanding of what treatments to perform on patients.

2.1.3 Molecular Subtypes of Breast Cancer

Patterns of certain characteristics are measured by DNA microarrays (a method that calculates the activity of genes in tissue samples, see Section 2.2.3) to better classify tumors into significant subgroups [35]. The molecular subtypes that have been commonly categorized by physicians and scientists are related to the clinical characteristics as described in the previous section. The method for classification of these subtypes; Basal, Her2, Luminal A, Luminal B and Normal-like, is described briefly here: [35]:

Data: Double-stranded complementary DNA (cDNA) microarrays were analyzed with 122 tissue samples and 500 “intrinsic” genes. These samples and genes were used to find molecular classes for the data set from Perou et al. [32].

Intrinsic Genes: Genes are considered “intrinsic” if their expression for tissue samples varied minimally in the same patient, but showed maximal variation between patients [35].

Tumor Classes: After genes were median-centered, both samples and genes were classified by the average hierarchical clustering technique. This technique begins with every element in the data set as it’s own cluster; a distance matrix is then calculated to find the minimal average distance between all clusters. The average distance is the average over all distances from one cluster to another. The two clusters with the smallest average distance are combined to one cluster. In the next iteration, a new distance matrix is then calculated for every cluster and the two clusters with the minimal average distance are combined. This process continues until all the elements are in the same cluster. An example of hierarchical clustering is shown in Figure 2.1 below.

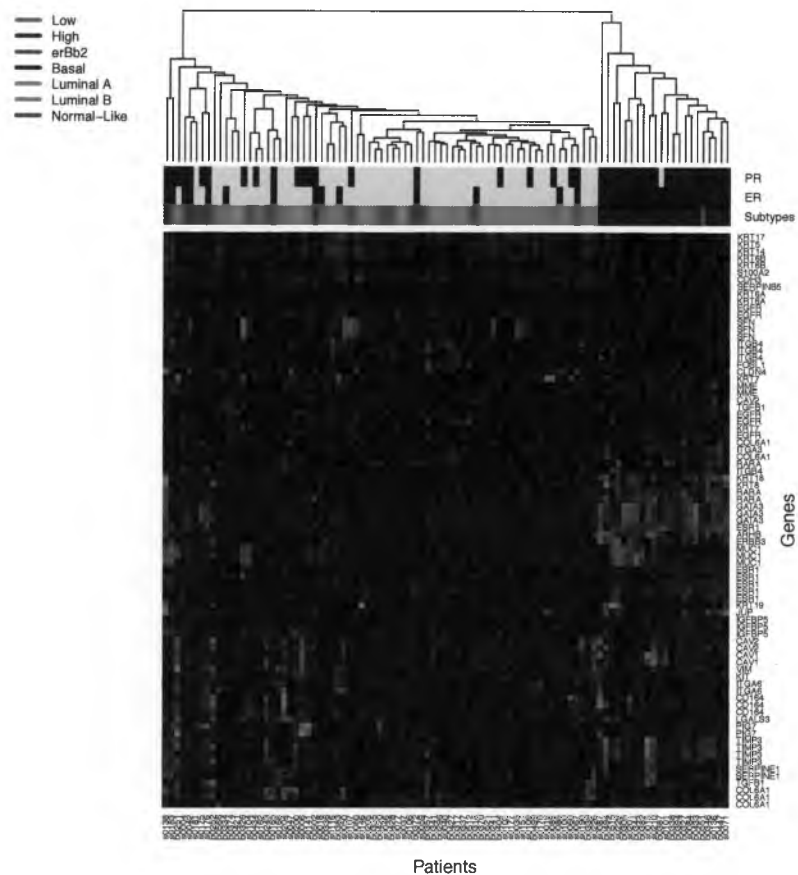


Figure 2.1: Average hierarchical clustering of patients and published markers of breast epithelial subtypes found in Chin et al. [7]. Green indicates lower expression levels, while red indicates higher expression levels. Genes are seen on the right and the bottom contains tumor samples. Subtypes of the 95 patients; erBb2, Basal, Luminal A, Luminal B, and Normal-Like are represented in colors green, red, light blue, orange and pink, respectively; these can be found at the top of the figure. ER and PR status are represented in rows at the top of the figure as well. Dark blue indicates negative ER/PR status and light blue indicates positive status.

The genes in Figure 2.1 are a subtype of the “intrinsic” genes that were not found correlated with copy number. Basal subtypes clustered well at the top right corner with the ER/PR negative tumors, while subtypes such as luminal and erBb2 are scattered, but with some clustering in the middle and to the left.

Classifying: Correlation (the strength of a linear relationship between two variables) was calculated pairwise between all tumors in the same class. Tumors that showed high correlation within the same class were used to calculate centroids (averages within each cluster) of those five classes.

Verification: Verification of the samples from Perou et al. ([36]) and West et al. ([40]) was done by assigning a tumor sample to the class with the highest correlation to the centroid of that class [35].

Prediction: Prediction analysis of microarrays (PAM) is used to predict classes for the data set from Sorlie et al. This analysis uses a gene selection step, Δ , that is integrated into the algorithm to balance prediction accuracy with a minimal set of genes. The value of Δ that was chosen for its prediction accuracy with a minimal set of genes was used for training on the data set in Sorlie et al. and the samples from van't Veer et al. and West et al. were predicted. For more details about the method PAM, please see [19].

Correlation was calculated for each tumor sample from Chin et al. with each of the five centroids for each cluster found in Sorlie et al.; the molecular classification for each tumor was chosen by the highest correlation between centroid and tumor sample. Figure 2.1 shows the hierarchical clustering of intrinsic genes that were not correlated with copy number values according to Chin et al. Clustering for copy number was also done to observe the subgroups of different phenotypes for all tumors. This is illustrated in Figure 2.2 below [7].

Figure 2.2 is similar to Figure 2.1, however, does not cluster as well. The basal subtypes, in red, still group well with the ER/PR negative types. However, the molecular subtypes do not cluster together as well. This is due to clustering the entire copy number data set, explaining why Figure 2.1 clusters better.

Our study is based on four of these molecular subtypes of breast cancer: Basal, Her2, Luminal A, and Luminal B. These sub-types are categorized by the clinical characteristics described in Section 2.1.3 [17]. Other features have an impact on these subtypes, however, we focus on these four, shown in Table 2.1.

Table 2.1: Molecular Classifications of Breast Cancer

Molecular Subtype	ER status	PR status	HER2 status/Ki-67 status
Luminal A	+/-	+/-	-/low Ki-67
Luminal B	+/-	+/-	+/high Ki-67
HER2	-	-	+
Basal (triple-negative)	-	-	-

Luminal A: This subtype is found in forty percent of all breast cancers, therefore, it is the most common type. These subtypes generally have ER+ and/or PR+ with no HER2 enrichment and low Ki-67 expression. Therefore, making it a slow-growing, less aggressive tumor than the other subtypes of breast cancer [34].

Luminal B: 10 - 20 percent of breast cancers are luminal B. A luminal B tumor is similar to luminal A in that it can have ER+ and/or PR+, however, it can have HER2 expression with high Ki-67. This can potentially cause the tumor to have high proliferation rates (number of cancer cells actively dividing) [34].

Basal: This type is also found in 10 - 20 percent of all breast cancers and is generally referred to as “triple-negative” because there is no excess traces of ER, PR or HER2. This subtype is more difficult to treat due to the fact that no hormonal therapy will help. Therefore, this subtype has poorer short-term prognosis [34].

HER2: These tumors are found in 10 percent of breast cancers and grow and spread more aggressively. Like the basal subtype, they have poorer short-term prognosis, however, the discovery of targeted therapies have had a positive influence on these subtypes by reversing prognostic impacts on HER2 over expression [34].

In this study, we further our understanding of how the concentration of DNA (i.e. copy number) affects the expression of genes (active and non-active genes) in any given tumor.

2.2 Epigenetic and Genetic Changes

It is commonly known that every human cell contains 23 pairs of chromosomes. Chromosomes are thread-like structures that are made up of protein and DNA (deoxyribonucleic acid) [30]. Many environmental factors, internal and external, affect the expression of genes by causing epigenetic and genetic changes. DNA is known as the “blue-print” of a cell and damages that alter the underlying DNA sequence are defined as genetic changes [39]. On the other hand, epigenetic changes are changes to proteins, but do not necessarily affect the DNA directly [2]. When alterations occur in the breast tissue, this can cause breast

cancer. In this study, we will look at a type of genetic change known as copy number.

2.2.1 Gene Expression

For genes to be expressed, DNA is transcribed into mRNA and mRNA is translated into functional proteins [39]. DNA is made up of compounds known as nucleotides that are connected in a long strand by phosphates and sugars. These nucleotides are Adenine (A), Cytosine (C), Guanine (G) and Thymine (T). Each nucleotide binds to an explicit base pair (A,C,G,T), specifically A binds to T and G binds to C. The binding of one strand to its corresponding strand of base pairs gives us the double-helix. In this double-helix, a specific sequence is called a gene. The processes that use DNA to produce proteins is described as the following:

Transcription: A single strand of messenger RNA (mRNA) is produced through transcription by the enzyme RNA polymerase II. The RNA polymerase II will “unzip” the DNA, read one of the strands, and produce the corresponding strand known as mRNA, which contains the nucleotides, A, G, C and U.

Translation: mRNA leaves the nucleus and binds with ribosomes that are translated to amino acids [33]. One or more chains of amino acids in a specific order become proteins. [28].

Activity of proteins make up most functions of the cell. For example, receptors are proteins that bind to a specific substance in the cell. Other functions such as breaking down food for energy are also functions of proteins, therefore, proteins are very important for any organism. However, DNA microarrays were introduced in the 1980s as a surrogate of protein concentration to measure the expression of genes in the cell [18]. They are also used to count the number of copies of genes in a tissue sample (i.e. copy number) [39].

2.2.2 Copy Number Changes (aCGH)

As discussed in the previous section, every human has 23 pairs of chromosomes. Alterations to DNA in these chromosomes may result in damaged cells, which can be caused by mutations. Mutations are permanent changes of a nucleotide sequence in the genome, which may change the transcription product of the DNA [26]. Copy Number variations are a subset of mutations, changing the DNA sequence with either a deletion or amplification (addition) [39]. These variations are illustrated in Figure 2.2.

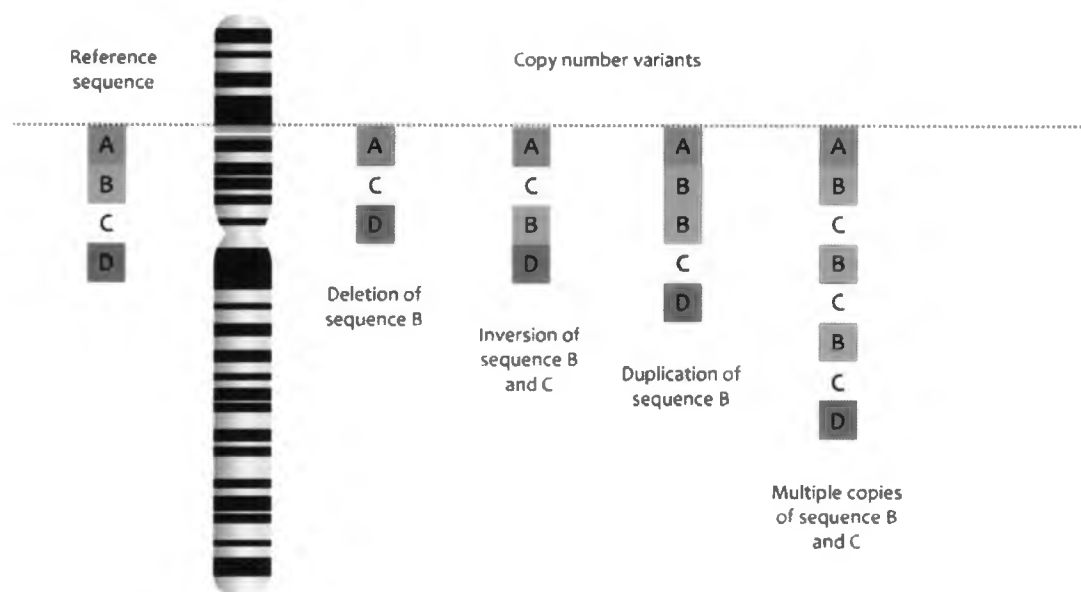


Figure 2.3: Copy number variations [14]

This figure shows the ideogram of a "healthy" chromosome with a reference sequence A,B,C and D, colored blue, green, yellow and red, respectively. This reference sequence can be changed in multiple ways: deletion, inversion, duplication and multiple copies. Deletion is caused by the removal of one or more nucleotides from a reference sequence. For example, from figure 2.1, there is a clear deletion on the right of the full chromosome. This will result in a smaller copy number during the DNA microarray experiment. Inversion is another type of copy number alteration (change), however, in the DNA microarray study (counting the number of copies of sequences in the DNA) there is no detection of copy

number change. When duplication occurs, there is an extra nucleotide in the sequence, as shown in the sequence second to the left. The last copy number change would be multiple copies where a sequence will be repeated multiple times. As the figure shows with the sequence on the right, B and C are repeated two extra times resulting in a copy number increase.

2.2.3 DNA Microarray

The experimental method for measuring copy number changes and gene expression profiling is DNA microarray and sequencing [24]. There are two main experiments for the DNA microarray: counting the number of copies of a DNA sequence (copy number) and counting the amount of gene activity (gene expression).

The first step in the DNA microarray experiment is to extract two tissue samples to compare. One sample is from the tumor and the other is from the general diseased area, these samples are considered our test and control samples, respectively. This is demonstrated in figure 2.3 with “Normal” and tumor tissue samples in the top left corner. mRNA is extracted from these samples and reverse transcribed. Both samples are labeled with fluorescent colors, Cyanine 3 (green) and Cyanine 5 (red). Hybridization, the binding of the DNA, is then performed on the microarray and scanned with a computer to get the intensity of those colors. The values that are assigned to each well represent the activity of the genes in that sequence. Figure 2.4 illustrates this process.

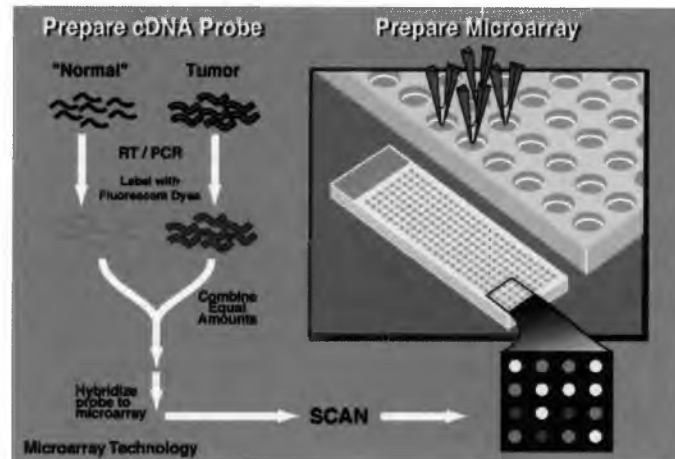


Figure 2.4: DNA microarray process [24]

Table 2.2: Differences in two DNA microarray experiments

	Copy Number	Gene Expression
Test sample color	green	red
Control sample color	red	green
Tissue extracted	DNA	mRNA transcribed into cDNA

Depending on the study, different approaches are used to calculate the genomic or epigenomic activity; Table 2.2 shows the distinctions between the experiments. The intensity of the color in each well estimates which DNA sample was more abundant. For gene expression, the color being brighter red implies that the tumor's sample was more abundant, while green indicates the control sample is more abundant [24]. This is reversed for copy number since the colors are opposite for each sample type. Yellow, for both experiments indicates that the samples bound about the same. After hybridization is complete, the

microarray is scanned, values are then given to the intensity of the color producing a data set of values that represents biological features and the data sets are both normalized using different methods, described in [7].

2.2.4 Relationship between aCGH and Gene Expression

DNA microarrays produce information about key biological features that help understand disease. aCGH data present information regarding genetic changes in tumor samples, while gene expression profiling gives genetic and possibly epigenetic changes. It is our goal to locate the regions of the genome that have high amplification of copy number (more copies of DNA) and also have over expressed genes (genes that are being expressed more than normal).

2.2.5 Tumor Suppressor Genes and Oncogenes

Damaged cells that uncontrollably divide is the main cause of any kind of cancer. There are two types of genes that inhibit the proliferation of damaged cells: tumor suppressor genes and oncogenes. Proto-oncogenes inform the cell when to continue dividing during the division process. Damage to these genes may then cause them to become oncogenes, making division faster to a damaged cell and in some cases, causing damages to the cell directly. Tumor suppressor genes function oppositely to proto-oncogenes and keep the cell from dividing too quickly. If these genes work incorrectly and a cell is damaged, they will

not initiate apoptosis (a cell committing suicide) [39].

Oncogenes and tumor suppressor genes can be found using aCGH and gene expression data. From these experiments, patients can have many amplifications/deletions in regions that are not relevant to the tumor, these are known as “passenger aberrations”. We perform statistical methods on a sample of the breast cancer tumor population to identify “driver aberrations” (regions that have relevance to the tumor) and discard “passenger aberrations”.

Chapter 3

Data and Methods

Copy number and gene expression profiling data were extracted from cells in the study done by Chin et al. [7]. Expression profiling for this study was done with the Affymetrix High Through-put Array (HTA) GeneChip system in a 96-well format. This system contains 96 microarrays positioned on a single plate in parallel, therefore, analyzing whole-genome expression for as many as 96 samples at a time. Transcription level is measured for each sequence represented by eleven pairs of oligonucleotide probes (a sequence of 13 to 25 nucleotides that are designed to hybridize to DNA or RNA sequences) in each well. Simplifying running multiple microarrays at one time, while increasing the standardization across samples [2]. The expression data from Chin et al. contained only probes from the HG-U133A Array [7].

Two types of arrays were used for genome copy number; scanning and oncoBAC arrays. Scanning arrays contained 2464 Bacterial Artificial Chromosomes (BACs) that were

selected approximately in megabase (1 million base pairs) intervals along the genome. OncoBAC arrays contained 960 P1 (form of chromosome derived through biological manipulation), PAC (P1-derived artificial chromosomes) or BAC clones. DNA samples were labeled for cancer and normal female genomes with CY3 (green) and CY5 (red), respectively. Once hybridization occurred, slides were washed and imaged with a 16-bit CCD camera with filters [7].

These data, analyzed by Chin et al., identified recurrent genomic and transcriptional abnormalities. Their findings included 66 genes in high-level amplification regions whose expression levels were correlated with copy number [7]. Our study includes reproducing the results that were found in Chin et al. which is discussed in chapter 4.

3.0.1 Data Processing

Both gene expression and copy number data were processed as described previously by Chin et al. [7]. Raw data was available for all three data sets: gene expression, copy number and phenotype, however, we used the already processed copy number data since it was also accessible. Multiple steps were taken to clean the raw data that was used for gene expression and phenotype.

75 patients were deleted from the raw data; 48 of those patients were deleted because they did not have a molecular subtype categorization and 32 were deleted due to one

of the three data sets not having that sample. Furthermore, 76 of the genes in the gene expression were removed from the data set because they were from chromosome Y. Since male breast cancer is a different disease entirely, we chose to only study female breast cancer. Information for the arms and cytobands was not provided in the data sets, therefore, was found using code in the programming language R and the cytoband information found in the UCSC genome browser [16]. Cytoband information consists of the base pair begin and end positions for each cytoband, giving us the ability to locate the clone or probe based on the chromosome and base pair position. Table 3.1 represents the phenotypes studied for these data.

Table 3.1: Phenotypes Studied

Data Set	Chin 2006	Raw Data Available	Yes
Normalized	Yes	Data Type 1	Gene Expression
Data Type 2	Copy Number	Total Patients	170
Patients we used	95	Subtypes (Y/N)	Yes
Basal	22	Luminal A	39
Luminal B	13	Normal-Like	6
HER2	15	ER positive	59
PR positive	53	Grade 1	12
Grade 2	37	Grade 3	46
Ki-67 Mutation Status	$\frac{79}{95}$ patients	Lymph Node Status	No
Number of Genes	21339	Number of CGH Clones	2149

Table 3.1 describes phenotype data over all tumors. This includes ER, PR, Ki-67, and molecular classifications; the raw phenotype data also included other phenotypes, but since

we do not focus on these, they will not be discussed.

Tables 3.2 and 3.3 below show samples of the data after processing; copy number (top) and gene expression (bottom). The average of copy number is 0, while the average of gene expression is 6. This is because the copy number data has been normalized with \log_2 ratios of the color intensity from the DNA microarrays and gene expression is normalized in another way.

Table 3.2: Copy Number Data

Clone	Chrom	Arm	bp	Cytoband	b0165	b0241
RMC01P070	1	p	5918606	p36.31	0.024341	-0.235773
RP11-51B4	1	p	6069000	p36.31	0.074727	-0.002416
RP11-60J11	1	p	6817000	p36.31	0.081012	0.066367
BAL01B2579	1	p	7827384	p36.23	-0.282227	-0.051416
RMC01P009	1	p	9421348	p36.22	0.0692	0.278853
RP11-199O1	1	p	10284000	p36.22	-0.084206	0.006058

Table 3.3: Gene Expression Data

Probe	Chrom	Arm	bp	Cytoband	b0165	b0241
211050_x_at	1	p	740099	p36.33	5.418485	5.294758
220399_at	1	p	801449	p36.33	5.437932	5.315859
219337_at	1	p	923262	p36.33	6.419116	5.946706
208023_at	1	p	1052764	p36.33	4.498784	4.371343
217855_x_at	1	p	1058369	p36.33	9.249178	9.554619
221972_s_at	1	p	1058369	p36.33	7.507093	8.06238

The first five columns represent the copy number clones and gene expression probes (seg-

ments of DNA or cDNA in the microarray experiment) and their locations in the genome. For example, clone RMC01P070 is located in chromosome 1, arm p, at base pair position 5918606 inside cytoband p36.31 (a small section in the arm of the chromosome). Columns preceding this information are the values produced from the microarray experiments for each sample, including b0165 and b0241 in the tables above. Copy number data is a 2149 by 100 matrix, while expression profiling data is a 21339 by 100 matrix. Since our study must contain the same number of clones and genes (probes), a method was needed to combine these two data sets properly.

Multiple probes can contain the same gene; this happens because the regions of cDNA can “signal” different aspects of that same gene. Therefore, multiple probes that are associated with the same gene will produce different gene expression values, even though they contain the same gene.

3.0.2 Assigning genes to clones

Multiple methods have been used to associate aCGH clone values and expression values, below is the process used by Chin et al. [7].

1. Copy number clones were deleted that did not have greater than or equal to 0.2 copy number values for five or more samples.
2. Bins of size 20Mbp (Mega base pairs) were created for each chromosome arm.

3. Clones and genes were placed inside these bins if their bp position was located in that bin.
4. Pairwise distances between genes and clones locations were calculated in each bin.
5. The clone and gene with the minimal distance were then assigned to each other.
6. This was repeated for all genes, clones and every bin.

Figure 3.1 below demonstrates this method.

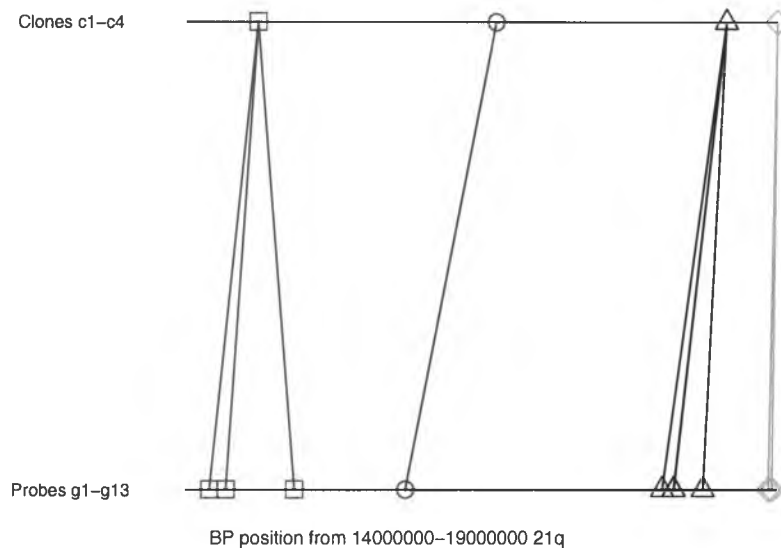


Figure 3.1: Algorithm for assignment of clones to genes based on location in the genome.

Illustrated in Figure 3.1, there are several more genes than there are clones, therefore, we must assign multiple genes to one clone. For example, the red lines indicate that genes g1 through g5 are assigned to clone c1, green lines mean that gene g6 is assigned to clone c2, blue lines indicate genes g7 through g11 are assigned to clone c3 and yellow lines indicate genes g12 and g13 are assigned to clone c4. Therefore, this process creates a lot of noise in our data (meaningless data) since we will be repeating the clone information multiple times while the gene information is used once. To clarify these issues, Figure 3.2 shows

the scatter plot of copy number, x , and gene expression, y , in chromosome 17, arm q for a HER2 patient.

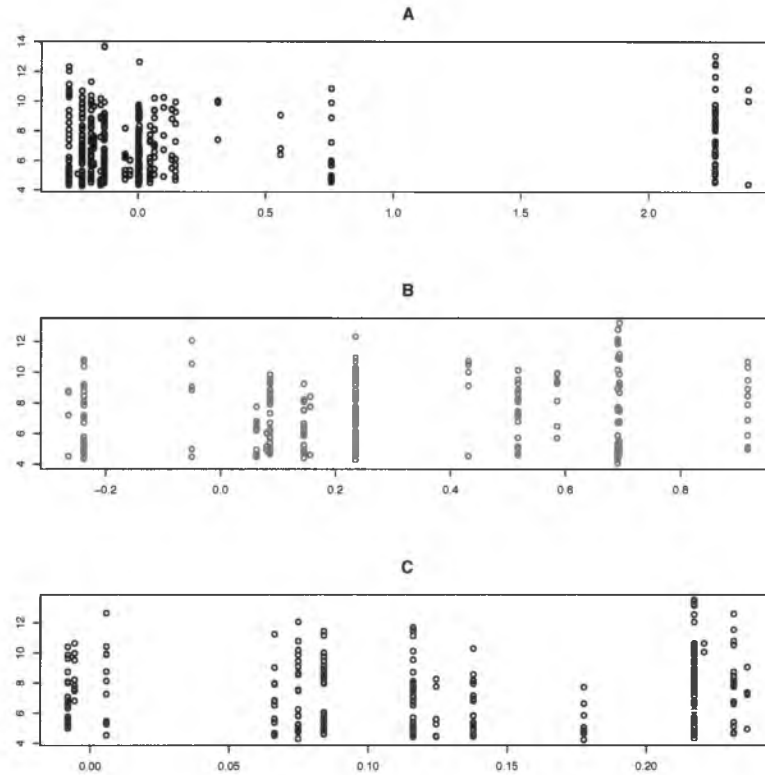


Figure 3.2: Scatter plots for chromosome 17q in bins for a single HER2 tumor. A, B and C represent bin 1, bin 2 and bin 3 for 17q, respectively. The x -axis represents copy number and the y -axis represents gene expression.

Figure 3.2 demonstrates variation of clones and expression of genes in a point cloud. In this case, we are specifically referring to the 2-dimensional coordinate system since there

are only two variables. The point cloud for each bin tells a different story; for example, it is clear in bin 1 there are a few highly amplified clones with assigned genes that are over expressed. This is shown on the right hand side of the point cloud for bin 1 where the copy number value is past 2. There are several y values (gene expression) in that line of multiple points that are higher than the average expression value, 6. This could indicate that there are genes that are regulated by copy number in this region of the genome for this HER2 patient.

The method used by Chin et al. to associate genes to clones is not informative enough due to poor point clouds created. Instead, we took the average over all the genes that were assigned to the same clone. This was not done for bins of size 20Mbp, but for the entire genome since we need to separate the data in a different fashion than the bins. However, by combining both data sets, the number of clones was decreased to 1827 from 2149, therefore deleting 322 clones. Therefore, biological information is lost for removing clones and averaging genes. Although this does lose information, we hypothesized that this still provides enlightenment about the expression of a common region, while creating a one-to-one correspondence between copy number and gene expression values over all patients. Figure 3.3 demonstrates the difference between the profiles before and after averaging genes.

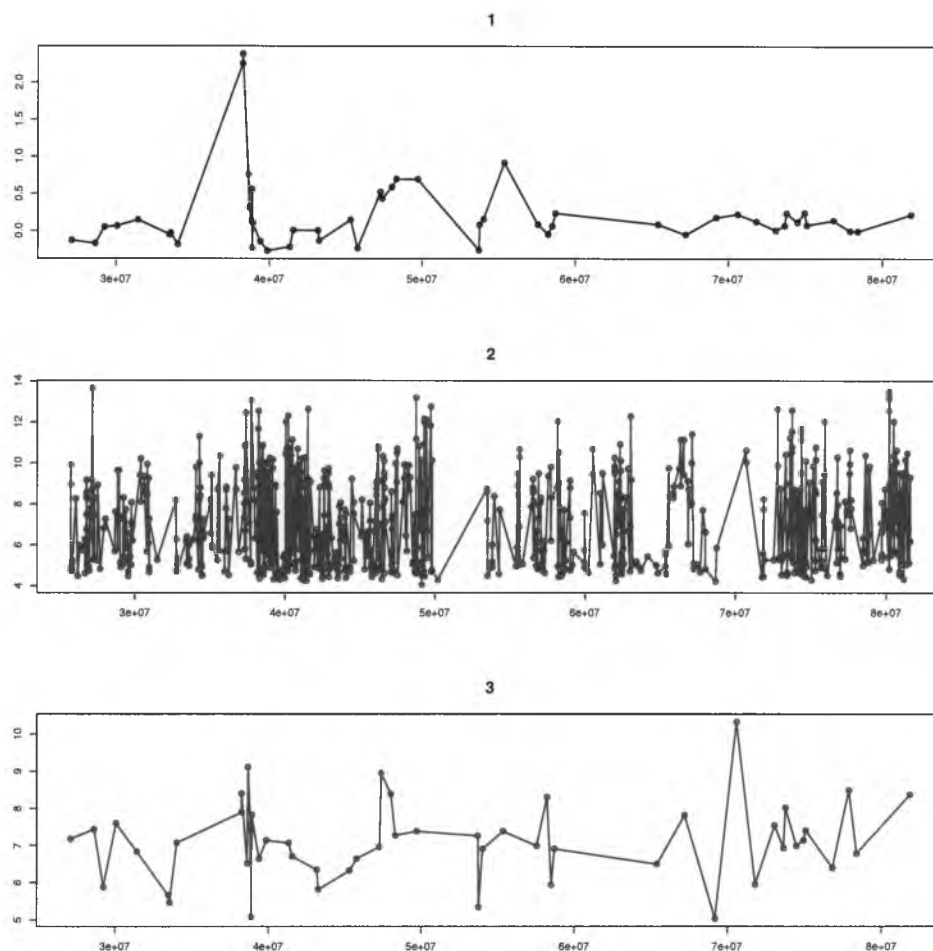


Figure 3.3: aCGH profile (1), gene expression profile (2) and averaged gene expression profile (3) for a Her2 patient in 17q. x -values represent the bp position in the chromosome arm 17q, while y - values are the copy number, gene expression and averaged gene expression, respectively. Gene expression probes were averaged for all genes assigned to the same clone.

Figure 3.5 shows the point cloud of the copy number (x) and expression (y) values before (1 and 2 from Figure 3.3) and after (1 and 3 from Figure 3.3) averaging genes.

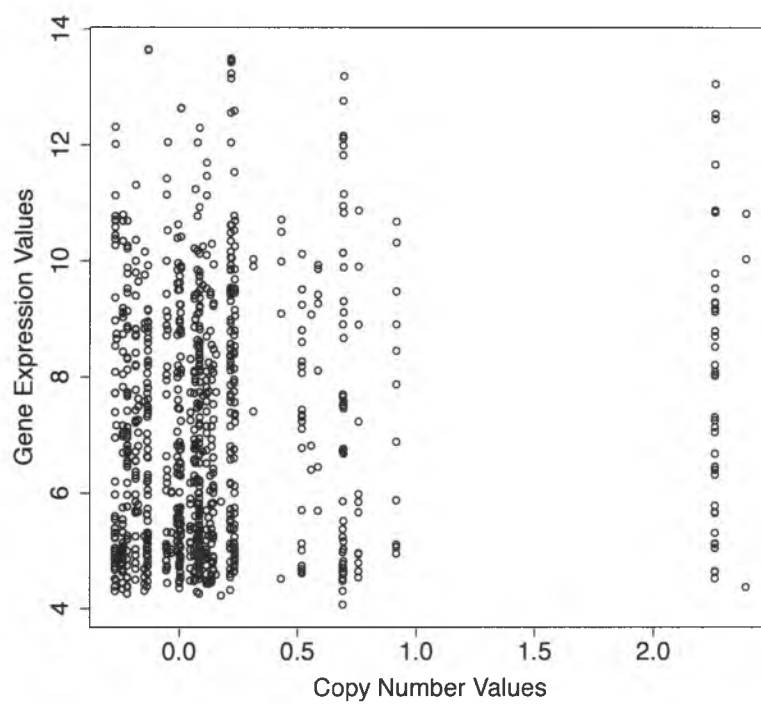


Figure 3.4: Profiles HER2 Patient Point Cloud before Averaging Genes. x -values in this figure are copy number values for (1) in Figure 3.3; y -values are gene expression values for profile (2).

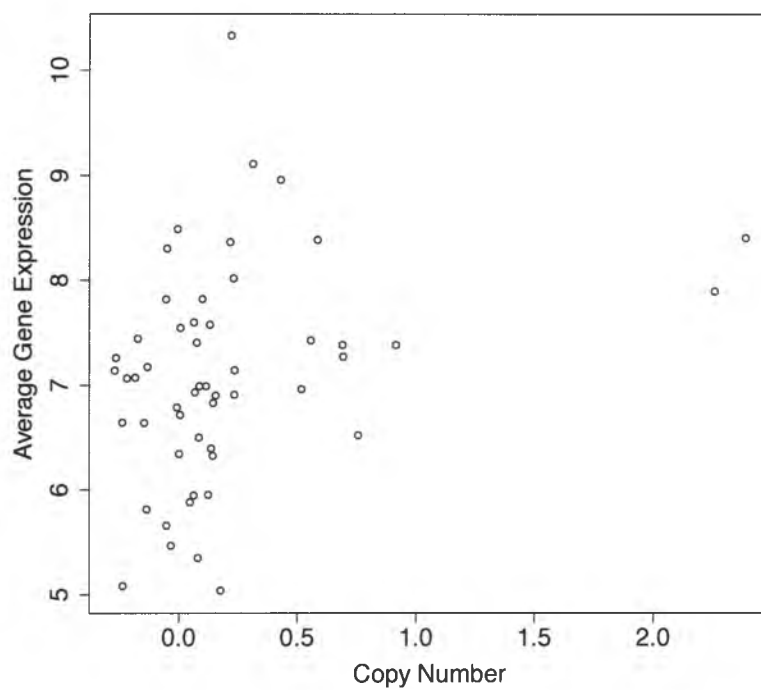


Figure 3.5: Profiles HER2 Patient Point Cloud after Averaging Genes. x and y values, similar to Figure 3.4 are from profiles (1) and (3) above.

Figure 3.4 contains more biological information, however, it is difficult to use in mathematical procedures because of the scattered points. As seen in Figure 3.3, there are clones that are highly amplified and genes assigned to those clones that are over expressed; this is also observed in Figure 3.4 and 3.5. Genes and clones of this nature are the main focus of our project; finding genes that are regulated by copy number clones. First, we reproduce the results from Chin et al. and second, we use persistent homology to locate significant genes and clones in the genome for different subtypes of breast cancer.

Chapter 4

Methods

Chin et al., 2006, analyzed gene expression and copy number data to identify genes that contribute to breast cancer pathophysiologies that accompany these diseases [7]. In regions of high-level amplifications: 8p11-12, 11q13-14, 17q11-12, and 20q13, there were 66 genes that were correlated with genome copy number variations. A subset of these genes have been found to be therapeutic targets for breast cancer subtypes. To verify these results, we execute some of the same methods as Chin et al. [7].

4.1 Method from Chin et al.

To find aberrations linked to breast cancer pathophysiologies, Chin et al. [7], used several methods to analyze 101 breast tumors. Their methods are described in Sections 4.1.2, 4.1.3, 4.1.4 and 4.1.5. From the supplemental data provided at the end of the paper, we were able to clarify which genes were not in our data set. Out of the 66 genes found in

high-level amplification regions that were correlated with copy number, 43 were found in our data. This gave us about 65 percent of the genes that they had found correlated with copy number. Out of those 43, we found that 29 of them were correlated with copy number clones from our analysis of a similar method to Chin.

4.1.1 Frequencies

Using circulatory binary segmentation (CBS), Chin et al. analyzed intensity measurements into regions of equal copy number. CBS is also a method to identify those discrete copy number aberrations that can be overlooked; a detailed description of this process can be found here [31, 37]. After analyzing the copy number data with CBS, missing values for clones were imputed. If a clone was within a segmented region of equal copy number, the value of the corresponding segment was imputed for that missing value. On the other hand, if a probe was located between two segmented regions, the maximum value for adjacent segments was imputed. Imputed values are referred to as “smoothed” values.

Tumor-specific experimental variation was estimated with scaled median absolute deviation (MAD) of the difference between the original and smoothed values. MAD is calculated by

$$constant * median(|x_i - median(x_i)|)$$

where $constant = 1.4826$ and where x_i is the difference of the original and smoothed value

for the i -th copy value. The constant 1.4826 is used for normally distributed data.

Gain and loss status for each copy number clone was assigned using the mergeLevel procedure that is described here [41]. All clones that corresponded to the copy number level with the minimal absolute median value were deemed unchanged, however, other clones were gains or losses depending on the sign of the segment mean.

Frequency of alterations was found for each copy number clone locus (location) by computing the proportion of samples showing an aberration at that locus. If the observed copy number value was more than four tumor-specific MAD away from the smoothed data, the clone was then assigned its original value; identifying the single outlier high level amplifications.

Amplification for each clone was determined by the width of the segment (0 if outlier) to which the clone belongs and the minimum difference between the smoothed value (observed if outlier) of the clone and the segment means of the neighboring segments. If the minimum difference is great than e^{-x^3} , where x is the final smoothed value of that clone and the clone belongs to the segment spanning less than 20Mb, the clone is amplified. This allows for the small valued clones to be declared amplified if they are surrounded by segments with the required difference becoming larger as the value of the clone gets smaller [7].

Several regions of the genome have been found to have high-level and low-level gains, as well as high-level amplifications, these regions found by Chin et al., are described in Table 4.1 below.

Table 4.1: Chin et al. Increases and Decreases for Molecular Subtypes of Breast Cancer

Subtype	Basal	Her2	Luminal A	Luminal B
Increased CNAs	3q,8p11-12, 8q,10p,11q13-14,17q11-12,17q21-24,20q13	1q,7p,8q,16p,17q,20q	1q,16p	1q, 8p11-12,8q,1113-14,17q,20q
Decreased CNAs	3p, 4p, 4q, 5q,12q,13q,14q,15q	1p,8p,13q,18q	16q	1p, 8p, 13q, 16q, 17p, 22q

After copy number aberrations were analyzed, associations between gene expression and copy number were tested for four common regions of amplification 8p11-12, 11q13-14, 17q11-12 and 20q13. In these areas, the 66 genes explained in Section 4.1.1 were found correlated with copy number.

4.1.2 Clustering

Unsupervised hierarchical clustering for all copy number clones was done to observe how the phenotypes of breast cancer group together. They concluded that ER/PR negative tumors grouped well with basal-like expression [7]. Following the method in which genes that were correlated with copy number; clustering was done with the “intrinsic” gene list from Stanford without the 66 genes found correlated with copy number.

4.1.3 Regression

Multivariate and univariate regressions were performed on different histopathological features such as size of tumor and nodal status (lymph node status) with survival duration and/or disease recurrence (univariate analysis). Both analyses are statistical methods; multivariate regression models 2 or more dependent variables and univariate regression contains one variable. These processes can be described in further detail in Hidalgo and Goodman [21].

4.1.4 Association of Copy Number and Gene Expression

High-level amplifications were associated with reduced survival duration and/or distant recurrence overall and within the luminal A expression subgroup. 66 genes in these regions were correlated with copy number and were deregulated by these high-level amplifications. Clones and genes were placed in bins of size 20Mb for each of the four regions, each gene was assigned to a clone and correlation was calculated for the 186 genes in those regions 8p11-12, 11q13-14, 17q11-12 and 20q13.

4.2 Our methods to verify Chin et al.

To verify the genes that were found correlated with copy number, similar methods to Chin et al. were used. With these methods we were able to confirm that 29 out of the 43 genes that we had from the 66 genes found in Chin et al. were correlated with copy number. This was done with the following procedures.

4.2.1 Frequencies

For each clone, the frequency of copy number value was found by calculating the number of times that clone was above 0.2 for each sample and dividing by the total number of samples. Therefore, a clone was a copy number gain if it was above the 0.2 threshold. A similar process is used for losses, except the number of times a clone is below -0.2

is counted. An entire arm was considered a gain or loss if the percentage of gains or losses of any clones in an arm is above forty percent. Figures 6.1, 6.2, 6.3, 6.4 and 6.5 in Chapter 6 show the frequencies of basal-like, her2, luminal A, luminal B and all 95 tumors, respectively.

4.2.2 Clustering

Copy number clones were used for unsupervised hierarchical clustering to understand what breast cancer pathophysiologies group together. Chin et al. indicates that ER/PR negative tumors cluster with basal-like subtypes. Figure 2.2 in Section 2.1.2 demonstrates the average hierarchical clustering of the copy number clones for the entire genome. Similar to Chin et al., [7], clustering was done for the intrinsic genes from Stanford without the genes that were found correlated with copy number. This can also be viewed in Section 5.1.3 as Figure 5.18. The method of average hierarchical clustering is described in Section 2.1.2.

4.2.3 Association of Copy Number and Gene Expression

For the regions 8p11-12, 11q13-14, 17q11-12 and 20q13, correlation was found for each copy number clone and gene. This was done by the description below.

Extraction: Genes and clones for these four regions were extracted from the original data.

Bins: Genes and clones were placed into bins of size 20Mb if their loci were in that bin.

Correlation: Correlation between all genes and clones was calculated for each bin.

Adjusted P-values: Holm adjusted p-values were adjusted for each correlation.

Interpretation: Point clouds were observed to better understand the connection between the gene expression and copy number relationship.

Correlation and Holm adjusted p-values are explained in the Methods chapter. The 43 genes found in our data and Chin et al. are shown in Table 4.2 below.

Table 4.2: Number of Genes

Gene Name	Number Copies	of	Gene Name	Number Copies	of
PROSC	3		BRF2	2	
ASH2L	1		LSM1	1	
BAG4	1		FGFR1	4	
TACC1	3		ADAM9	1	
AP3M2	1		POLB	1	
VDAC3	3		SLC20A2	1	
IKBKB	3		FNTA	2	
CCND1	3		FGF3	1	
FADD	1		PPFIA1	4	
FOLR3	1		NEU3	1	
LHX1	1		ACACA	2	
PSMB3	1		PIP5K2B	2	
FLJ20291	1		PPARBP	2	
TCAP	1		PNMT	1	
ERBB2	2		GRB7	1	
PSMD3	1		NR1D1	3	
BCAS1	1		CSTF1	2	
RAE1	3		PCK1	1	
TMEPAI	1		RAB22A	1	
VAPB	1		NPEPL1	2	
GNAS	6		TH1L	1	
WHSC1L1	2				

Most of these genes are repeated for different gene expression probes. The values to the right of each gene name represents the number of times this gene is present in our data.

Repeated for this study was the association of copy number and gene expression with the entire genome. This was again produced using the method to place genes and clones into bins, assign genes to clones and then find correlation within each bin between each clone and gene.

Outcome of these methods are found in Results and Discussion (Chapter 6).

4.3 Finding Significant Regions in the Genome

Our study extends the method proposed by Arsuaga et al.,[4], to locate significant regions of copy number changes in cancer. In that study, the following procedure was done to find significant regions in the genome based on one group versus another.

1. Segmenting the data such that 20 clones are in a segment with 10 overlapping.
2. Using a patient profile with the sliding window algorithm to produce a point cloud for each patient.
3. Check the minimum distance between all points in the cloud to choose the optimal increment values of the filtration parameter, ϵ .

4. Choose a filtration parameter, ϵ , to connect vertices in the point cloud at each iteration of the filtration.
5. β_0 numbers are calculated at each iteration of the filtration, ϵ .
6. The average is calculated at each ϵ increment for one group versus another and is then graphed; this is referred to as the Betti curves.
7. Permutation testing statistics is calculated to state whether the curves are statistically different; adjusted p-values for this are also calculated.

We extend this study by using both copy number and gene expression data to find significant regions of the genome. These regions will tell us if genes are being regulated or deregulated by copy number.

4.3.1 Segments, Sliding Window, and Point Cloud

The first step in this method includes sectioning the data into segments along the entire genome. Each segment consists of 20 clones with 10 overlapping clones from the previous and subsequent segment. Figure 4.1 demonstrates the process of segmenting the data.

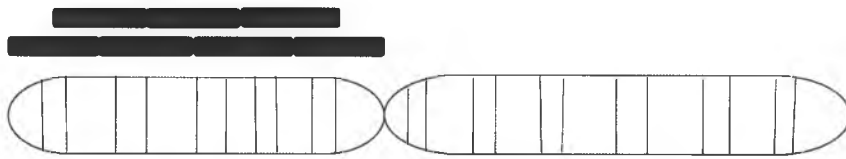


Figure 4.1: Subdividing one chromosome arm into segments

The sliding window in Arsuaga et al. uses only copy number profile for the sliding window [4]. For this study, we use the sliding window in a different way. Figures 4.2 represents our sliding window method for persistent homology.

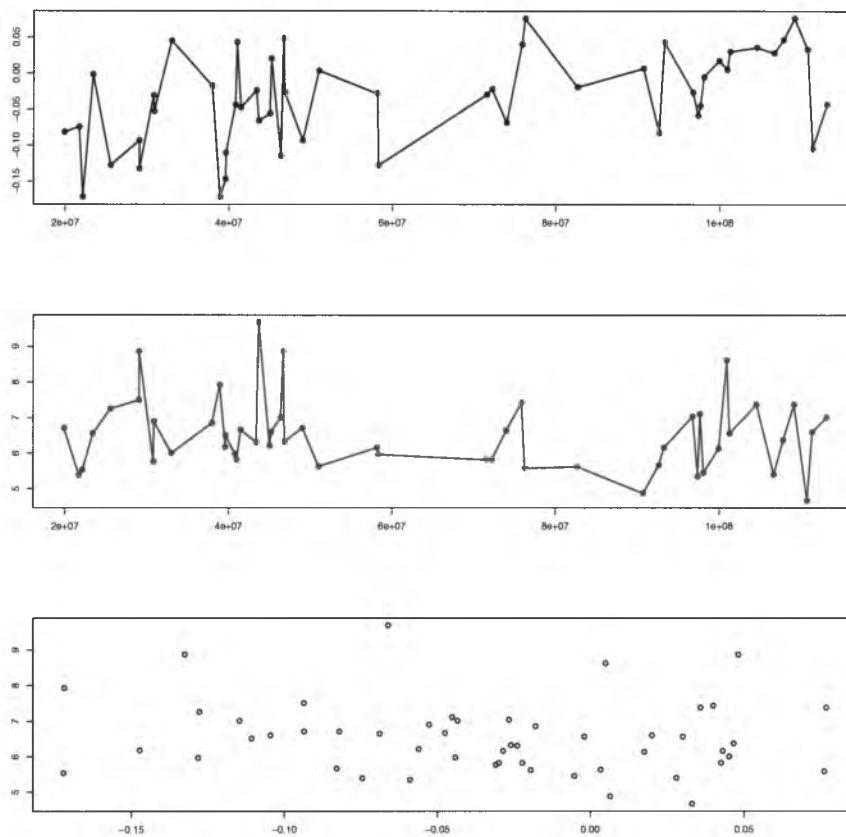


Figure 4.2: Sliding window method for patient b0243 patient in 13q

The graph at the top is the copy number profile for patient b0243 in 13q; there are 51 clones in this region. Each clone from left to right is used as the x -value for our 2-dimensional point cloud. The middle graph is the averaged gene expression profile for patient b0243 in 13q; there are 51 averaged genes in this region. Each averaged gene from left to right is used as the y -value for the point cloud. At the bottom is the point cloud that we create with these profiles. If we imagine that profile 1 (copy number) is labeled c_1, c_2, \dots, c_n and profile 2 (gene expression) is labeled g_1, g_2, \dots, g_n then our point cloud would have the labels $(c_1, g_1), (c_2, g_2), \dots, (c_n, g_n)$.

4.3.2 Connected Components and Filtration

For all of our given point clouds, a fixed number, ε , is used on our 2-dimensional point cloud to connect vertices; these vertices connect using the Vietoris-Rips complex [13],[38].

Definition 4.1. Given a set of vertices, $\{x_i\}$, the **Vietoris-Rips Complex** R_ε is the abstract complex such that the pairwise distance of each vertice is less than ε , $d(x_k, x_j)$ for all $i, j \in \{x_i\}$.

In other words, when two vertices are within ε away from each other, they will connect. Therefore, the 2-dimensional Vietoris-Rips simplicial complex is a triangle when three points connect pairwise.

For each value of ε , the number of connected components decreases. For example, when $\varepsilon = 0$, the number of connected components will be the number of vertices in the point

cloud; when two vertices connect, the number of connected components goes down by one and so on. Before connected components are counted at each filtration, ϵ must be chosen appropriately.

4.3.3 Choosing Filtration Parameter

To choose the optimal ϵ , we use R code to calculate the minimum, first quartile, median, third quartile, maximum and the average of all distances between every point in each segment. Quartiles divide the data into four equal parts where the third quartile constitutes the upper 25 percent of the data. The first quartile is the bottom 25 percent of the data; median is the middle. This was calculated over all patients for every segment. A boxplot for each copy number patient is displayed in Figure 4.3.

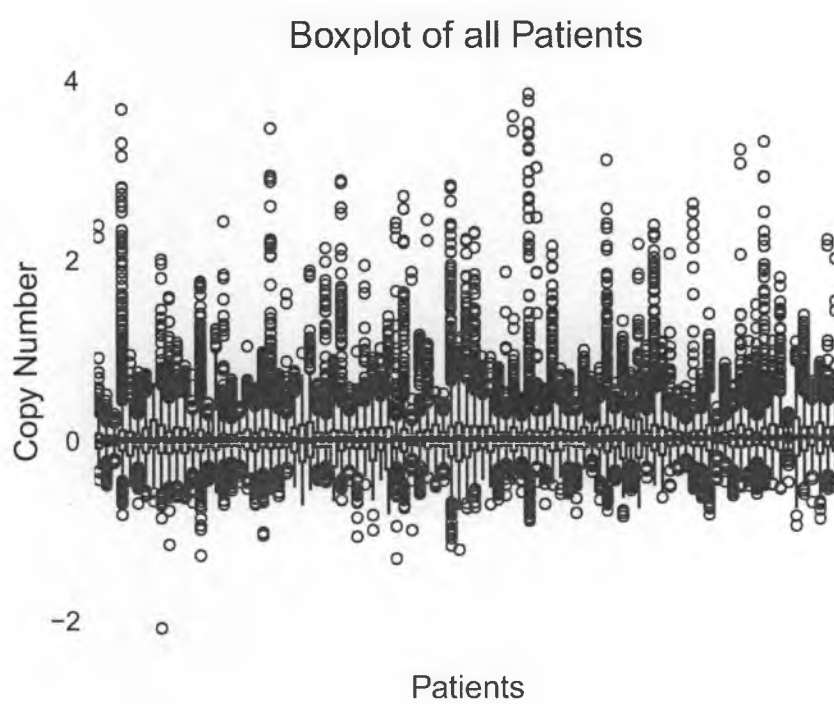


Figure 4.3: Boxplot of copy number over all patients.

In this figure we can see quartiles for each patient. Each box represents the spread of the copy number data for that patient. The points at the ends of the boxplot are outliers, the top of the box represents the first quartile, the middle line inside the box is the median and the bottom of the box is the third quartile. Therefore, the box contains 75 percent of the data. IQR (Interquartile range) = $(Q_3 - Q_1)$ where Q_1 is the first quartile and Q_3 is the third quartile. Outliers consist of values smaller than $Q_1 - 1.5 IQR$ and larger than $Q_3 + 1.5 IQR$. The highest point above the box is the maximum and the lowest point below the box is the minimum. Using this idea, we find the minimum distance of the minimum pairwise distances over all patients.

4.3.4 Betti Curves

At each iteration of ε , the number of connected components, β_0 , is calculated for each patient and recorded. This process ends when every vertex is connected to at least one other vertex. This creates a function $\beta_0(\varepsilon)$ that is dependent on ε . As the ε increases, the number of connected components will decrease. Once the algorithm terminates, the average β_0 value for each ε is calculated for the two groups that we are testing. For example, in Figure 4.4 we have basal vs. all other subtypes. Therefore, for all $\varepsilon_1, \dots, \varepsilon_n$, we calculate the average β_0 for the subtype basal and calculate the average for all other subtypes.

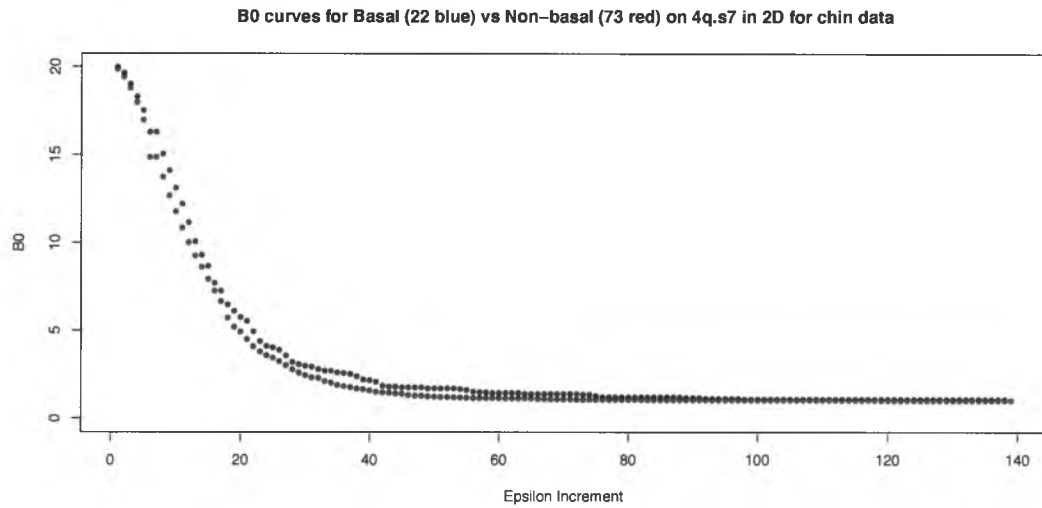


Figure 4.4: Betti curve for basal vs. others for 4q segment 7

Figure 4.4 shows the β_0 curves, the test curve (blue curve) and the control curve (red curve). Notice that the blue curve is above the red curve; we only consider this case in our study because it indicates that the vertices for the test are further away from the center than the vertices for the control; indicating that the test group has some amplification/deletion and over or under expression. Next, we determine if these two curves are statistically different by running a permutation test.

4.3.5 Permutation Testing

Betti curves, like the one in Figure 4.4 consist of the average between one group (test) and the average of all other groups (control). The test group is denoted by the vector

$\vec{t} = (\vec{t}_1, \vec{t}_2, \dots, \vec{t}_n)$ and the control group is denoted by $\vec{c} = (\vec{c}_1, \vec{c}_2, \dots, \vec{c}_n)$. Each \vec{t}_i is a vector of β_0 values for the i th ε iteration for the test group and \vec{c}_i is a vector of β_0 values for the i th ε iteration for the control group.

The null hypothesis (H_0) is that $\sum_i (\text{avg}(\vec{t}_i) - \text{avg}(\vec{c}_i))^2 = 0$ for all i .

$$T_{obs} = \sum_{i=1}^n (\text{avg}(\vec{t}_i) - \text{avg}(\vec{c}_i))^2$$

is the observed statistic, while the permutation (test) statistic is calculated as

$$T_{perm} = \sum_{i=1}^n (\text{avg}(\vec{t}_i) - \text{avg}(\vec{c}_i))^2$$

T_{perm} represents the permutation test statistic.

Sum of the squared differences of the original betti curves is our observed statistic; the permutation (test) statistic consists of randomizing the patients first into two groups of the same size as the test and control groups and finding the sum of the squared differences of the average at each ε increment.

Segments that were found significant were further analyzed by computing the correlation between copy number and averaged gene expression. Figure 4.5 shows the correlation of one segment found significant from our study. They adjust the p-values based on the

methods Holm, False Discover Rate (FDR), Hochberg, Hommel, Bonferroni and BY.

For this study, we considered significant regions to be the ones where all 6 of the adjusted p-values were less than 0.05.

4.3.6 Correlation with Significant Regions

Rows of the correlation matrix below in Figure 4.5 are the averaged genes and the clones are the columns. The diagonal of these correlation matrices are the clones and averaged genes that were assigned to each other. Therefore, the clones and averaged genes along the diagonal are the closest in genomic location to each other. Dark red indicates a strong, positive correlation while dark blue is a strong, negative correlation.

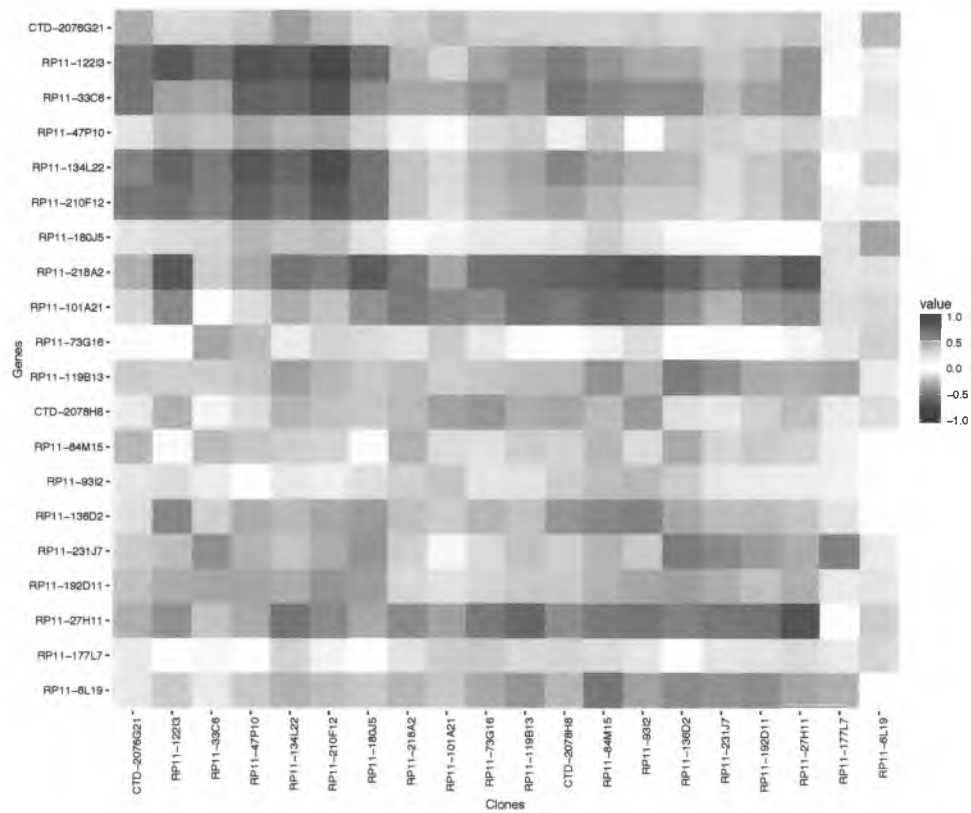


Figure 4.5: Correlation between clones and averaged genes for 4q segment 7

We can see from Figure 4.5 that there is strong positive correlation in the diagonal region of this figure; these are most important to us because they are the clones and assigned genes that were assigned to each other. Therefore, we look at all the diagonal elements in the matrix and the adjacent elements of the diagonal. These were chosen because the clones and averaged genes in the adjacent diagonal regions are also near each other in genomic location. Figure 4.6 illustrates the region that is of most importance.

Figure 4.6: Sections studied for high correlation

$$\begin{pmatrix} \begin{matrix} \blacksquare \\ \blacksquare \end{matrix} & \begin{matrix} \blacksquare \\ \blacksquare \end{matrix} & \begin{matrix} \blacksquare \\ \blacksquare \end{matrix} & \begin{matrix} \blacksquare \\ \blacksquare \end{matrix} & \begin{matrix} \blacksquare \\ \blacksquare \end{matrix} & \begin{matrix} \blacksquare \\ \blacksquare \end{matrix} \\ \begin{matrix} \vdots \\ \vdots \end{matrix} & \begin{matrix} \vdots \\ \vdots \end{matrix} & \begin{matrix} \vdots \\ \vdots \end{matrix} & \begin{matrix} \vdots \\ \vdots \end{matrix} & \begin{matrix} \vdots \\ \vdots \end{matrix} & \begin{matrix} \vdots \\ \vdots \end{matrix} \\ \begin{matrix} \vdots \\ \vdots \end{matrix} & \begin{matrix} \vdots \\ \vdots \end{matrix} & \begin{matrix} \vdots \\ \vdots \end{matrix} & \begin{matrix} \vdots \\ \vdots \end{matrix} & \begin{matrix} \vdots \\ \vdots \end{matrix} & \begin{matrix} \vdots \\ \vdots \end{matrix} \\ \begin{matrix} \vdots \\ \vdots \end{matrix} & \begin{matrix} \vdots \\ \vdots \end{matrix} & \begin{matrix} \vdots \\ \vdots \end{matrix} & \begin{matrix} \vdots \\ \vdots \end{matrix} & \begin{matrix} \vdots \\ \vdots \end{matrix} & \begin{matrix} \vdots \\ \vdots \end{matrix} \\ \begin{matrix} \vdots \\ \vdots \end{matrix} & \begin{matrix} \vdots \\ \vdots \end{matrix} & \begin{matrix} \vdots \\ \vdots \end{matrix} & \begin{matrix} \vdots \\ \vdots \end{matrix} & \begin{matrix} \vdots \\ \vdots \end{matrix} & \begin{matrix} \vdots \\ \vdots \end{matrix} \\ \begin{matrix} \vdots \\ \vdots \end{matrix} & \begin{matrix} \vdots \\ \vdots \end{matrix} & \begin{matrix} \vdots \\ \vdots \end{matrix} & \begin{matrix} \vdots \\ \vdots \end{matrix} & \begin{matrix} \vdots \\ \vdots \end{matrix} & \begin{matrix} \vdots \\ \vdots \end{matrix} \end{pmatrix}$$

Figure 4.6 shows the matrix of correlations with black boxes to demonstrate the areas of the correlation matrices we are looking for. This is to determine the specific genes that drive significance in this segment.

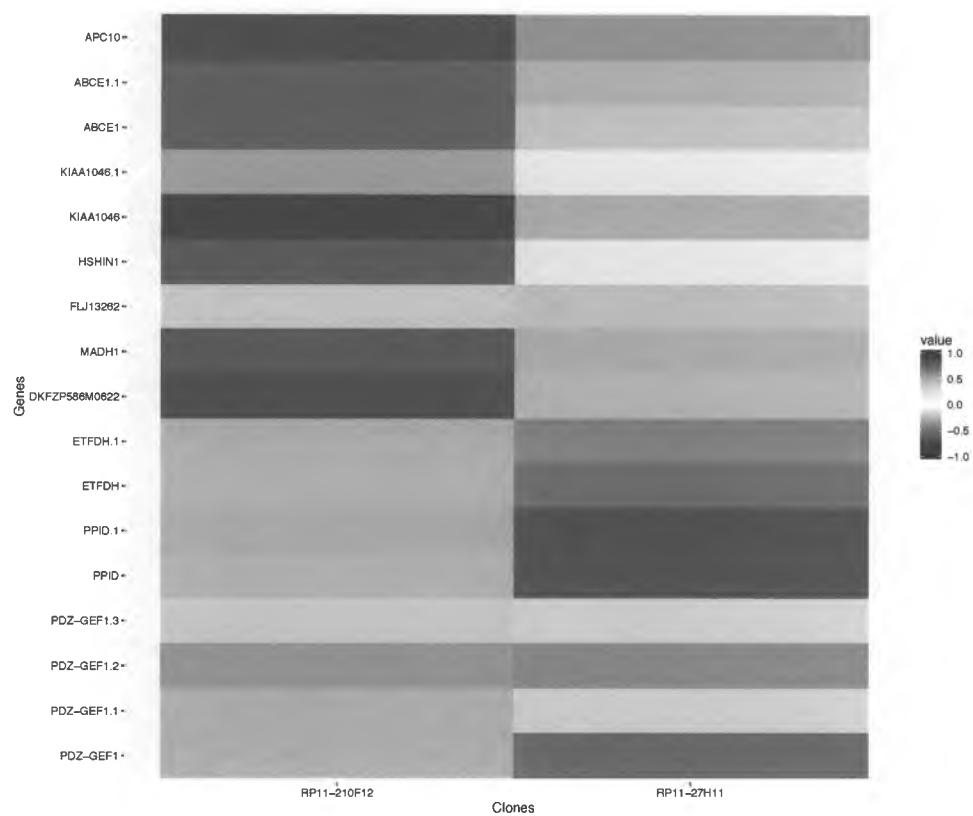


Figure 4.7: Correlation between clones and genes from the diagonal of Figure 4.5

Rows represent genes and columns represent clones and the entries show the correlation between gene expression value and copy number value in the sections previously identified in Figure 4.6. The genes that are correlated with copy number are candidates to be regulated by copy number.

4.3.7 Pearson's Product/-Moment Correlation

This description of correlation is based off of Marx and Devore [27],[11].

Definition 4.2. For two variables, x and y , the **Pearson's Product-Moment Correlation** is given by

$$\rho = \frac{\text{Cov}(x, y)}{s_x s_y}$$

The correlation coefficient, ρ , measures the strength of the linear relationship between x and y by calculating the covariance of the two variables and scaling it by the standard deviations of x and y .

Therefore, if we have a sample of $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ the r , the **Pearson's Product-Moment Correlation** can be expanded by the following equation.

$$\begin{aligned} r &= \frac{\frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y}}{\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}} \\ &= \frac{n \sum_{i=1}^n x_i y_i - (\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)}{\sqrt{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \sqrt{n \sum_{i=1}^n y_i^2 - (\sum_{i=1}^n y_i)^2}} \end{aligned}$$

In our case, we have c as our x variable and g for our y variable, which represent the copy number and gene expression, respectively.

$$= \frac{n \sum_{i=1}^n c_i g_i - (\sum_{i=1}^n c_i) (\sum_{i=1}^n g_i)}{\sqrt{n \sum_{i=1}^n c_i^2 - (\sum_{i=1}^n c_i)^2} \sqrt{n \sum_{i=1}^n g_i^2 - (\sum_{i=1}^n g_i)^2}}$$

The numerator of the equation above is equivalent to $\sum_i \sum_i (c_i - \bar{c})(g_i - \bar{g})$. The variance of c_i from it's mean is $(c_i - \bar{c})$ and $(g_i - \bar{g})$ is the variance of each g_i from it's mean. This means that if there is a strong positive linear correlation between the two, then when $(c_i - \bar{c}) > 0$, so will $(g_i - \bar{g})$ and when $(c_i - \bar{c}) < 0$, then $(g_i - \bar{g})$ will also be < 0 . The opposite holds for a strong negative correlation. In Figures 4.8 and 4.9, we see that both figures have strong linear correlation.

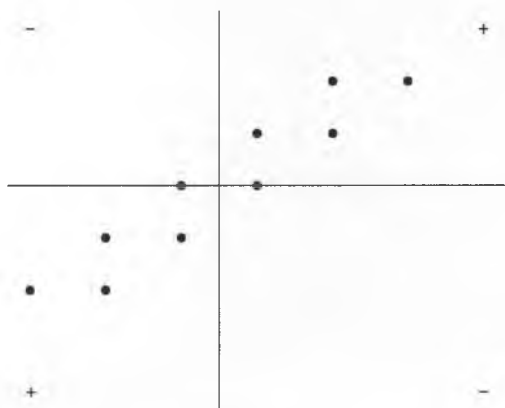


Figure 4.8: Strong Positive Correlation

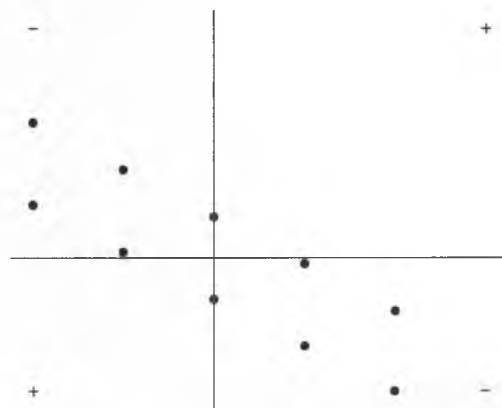


Figure 4.9: Strong Negative Correlation

Figure 4.8 has a positive correlation, while Figure 4.9 has a negative correlation. The vertical lines represent \bar{c} ; horizontal lines are \bar{g} . Therefore, the top portion of the ratio

for r gives us the strength of the relationship between c and g . s_c and s_g help give this relationship a dimensionless measure of dependency. Therefore, giving r values in between -1 and 1 . If there is perfect positive correlation, then $r = 1$ and if there is perfect negative correlation, then $r = -1$. For copy number and gene expression, we would like to find the genes that are regulated by copy number increases or decreases; indicating that we are looking for genes and clones with strong positive correlation. It is possible to find genes and clones that have strong negative correlation, however, we mainly focus on positive correlation between the two. The p-value for the correlation is also important because that tells us how accurate our correlation is; we will define a p-value.

Definition 4.3. A **p-value** is associated with an observed test statistic, measuring the probability of getting a value for that test statistic as extreme as or more extreme than what was observed (alternative hypothesis, H_a) given that the null hypothesis (H_0) is true.

For correlation, the null hypothesis is $r = 0$ and the alternative hypothesis means there is a linear relationship between the two variables. The calculation of this specific p-value is found by permutation testing. g_i values are permuted, denoted, $g_{i'}$ and the correlation coefficient for that permutation is then calculated.

$$r' = \frac{n \sum_{i'=1}^n c_i g_{i'} - (\sum_{i'=1}^n c_i) (\sum_{i=1}^n g_{i'})}{\sqrt{n \sum_{i=1}^n c_i^2 - (\sum_{i=1}^n c_i)^2} \sqrt{n \sum_{i'=1}^n g_{i'}^2 - (\sum_{i=1}^n g_{i'})^2}}$$

Multiple permutations of the g_i values occur and the p-value is the number of times the permutation test statistic is larger than the original correlation coefficient divided by the

number of permutations.

$$p - \text{value} = \frac{\text{number of times } r' > r}{\text{number of permutations}}$$

A small p-value of < 0.01 is significant for this study and implies that we reject the null hypothesis; indicating that the two variables are related.

4.4 Adjusting P-values

For permutation testing with the persistent homology method, we use Holm, Hommel, Bonferroni, BY, FDR, and Hochberg. In this section, we will discuss Holm and Bonferroni; Holm is used for both permutation testing and correlation, while Bonferroni is strictly for permutation testing.

4.4.1 Holm Adjusted P-value

Using holm adjustment p-values for correlation includes the following method. Let H_1, H_2, \dots, H_m be a family of hypotheses with corresponding p-values P_1, P_2, \dots, P_m . We arrange these in increasing order; $P_{(1)} \leq \dots \leq P_{(m)}$. Each P_i is compared with $\frac{\alpha}{(n-i+1)}$ for rejection. Therefore, H_j is rejected if $P_j \leq \frac{\alpha}{(n-j+1)}$ for all $j \leq i$ [3].

4.4.2 Bonferroni

With multiple testing, the null hypotheses are $H_i (i = 1, \dots, n)$ and corresponding P_i values for each. If we assume that t of the n hypotheses are true, a type I error can occur only if one of the events $P_i \leq \frac{\alpha}{n}$ occurs for one true hypotheses.

Chapter 5

Results and Discussion

Results from two studies will be discussed in this chapter. The first study, we verified some of the results found by Chin et al. Secondly, we focus on our own approach to find significant regions in the genome for different molecular subtypes. Within these regions, we determine genes that are being regulated by copy number by using pearson's correlation coefficient.

5.1 Results for Chin et al.

Frequencies were produced to clarify regions of amplifications and deletions for different subtypes over the entire genome. Table 4.1 indicates the areas in the genome that have increased and decreased copy number, according to Chin et al. [7]. The method used to find frequencies for each copy number clone is described in Sections 4.1.1 and 4.2.1.

5.1.1 Frequencies

To better understand regions with genomic aberrations linked to breast cancer, Chin et al. produced frequencies to locate regions of amplifications or deletions based on the four most common subtypes of breast cancer; basal-like, HER2, luminal A, and luminal B. Areas of common amplifications and deletions for all tumor samples were then studied using correlation. By creating similar frequency figures, we were able to find similar regions of amplifications and deletions; similar to Table 4.1. We observe this by looking at the frequency graph for the subtype basal in Figure 5.1.

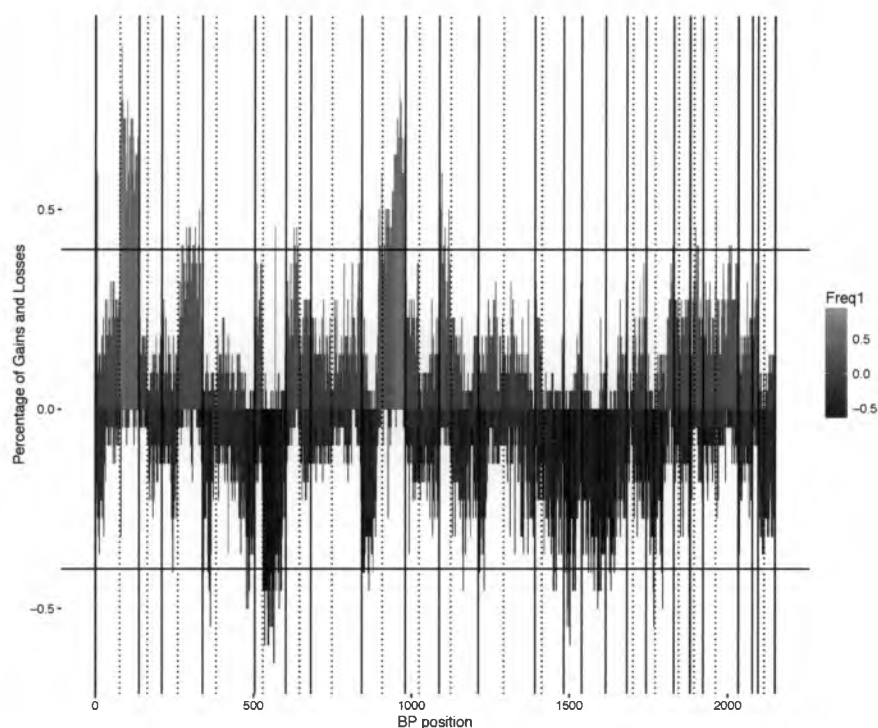


Figure 5.1: Frequency of Copy Number Values for basal subtype. The x -axis displays the BP position along the entire genome, while the y -axis is the frequency for each copy number clone. Frequency for each clone is calculated by the number of tumor samples above a 0.2 threshold for amplifications and below -0.2 for deletions, divided by the total number of tumor samples. Black lines indicate where the chromosomes begin and end; black dotted lines imply the position of the centromeres of each chromosome. Chromosomes are located from left to right in order from 1 to 23.

Figure 5.1 shows several regions of gains and losses for copy number clones from the 22 basal tumors. The regions with amplifications are 1q, 3q, 5q, 6p, and 10p, with slight amplifications in regions 8p, 8q, 17q, 19q, and 20q. Losses were found at 4q, 5q, 8p, 12q, 13q, and 14q; only a few deletions in clones were found in 3p, 4p, 10q, 11p, 15q, 17p and 17q . These results include all regions found with increased and decreased copy number from Table 4.1, except 11q, which was found to be an amplification by Chin et al. [7]. Regions such as 1q, 5q, 6p, 19q and 20q were not found amplified in Chin et al., however, were found in our study. Losses that were found in 10q, 11p, 17p and 17q were not claimed deletions by Chin et al. [7].

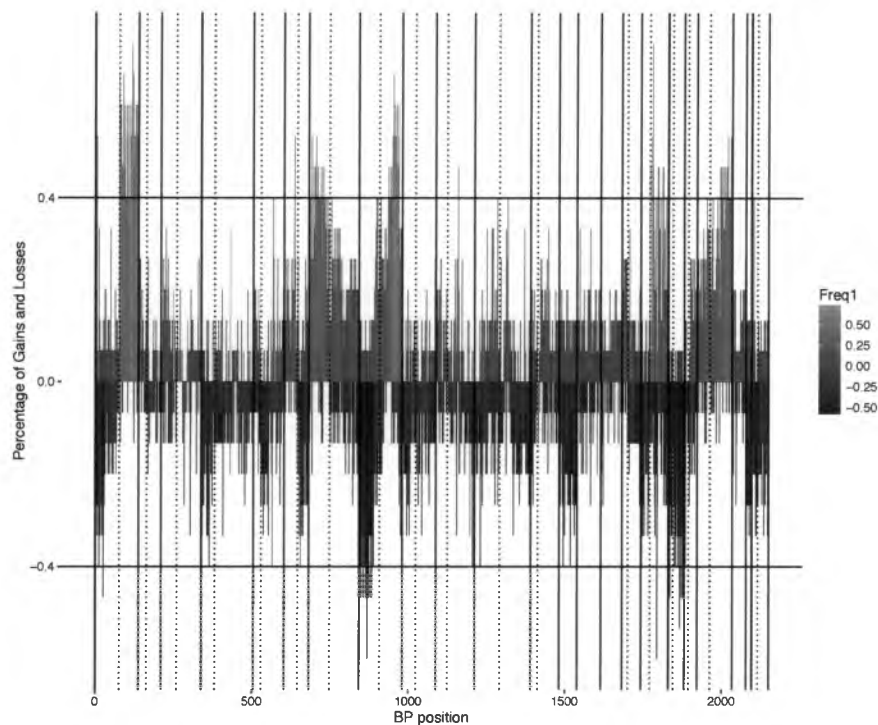


Figure 5.2: Frequency of Copy Number Values for HER2 patients. Figure 5.2 shows increases and decreases in copy number based on location in the genome. Similar to Figure 5.1, solid black lines indicate where the chromosomes begin and end, while dotted black lines are representations of the centromeres for each chromosome.

In Figure 5.2, we see that 15 HER2 patients have copy number increases in 1q, 7p, 8q, 17q, and 20q; small peaks of copy number amplifications were found in 10q, 11q, 20p and 23p. Decreases in 1p, 8p, and 18q were found for these Her2 patients and slight losses were indicated in regions 4p, 5q, 6q, 11p, 13q, 17q and 18p. The region 16p was not found to be an amplification for our study, however, regions such as 10q, 11q, 17q, 20p and 23p were claimed amplifications in our study that were not declared by Chin et al. Decreases such as 4p, 5q, 6q, 11p, 17q and 18p were found in Figure 5.2, but were not found in Chin et al.

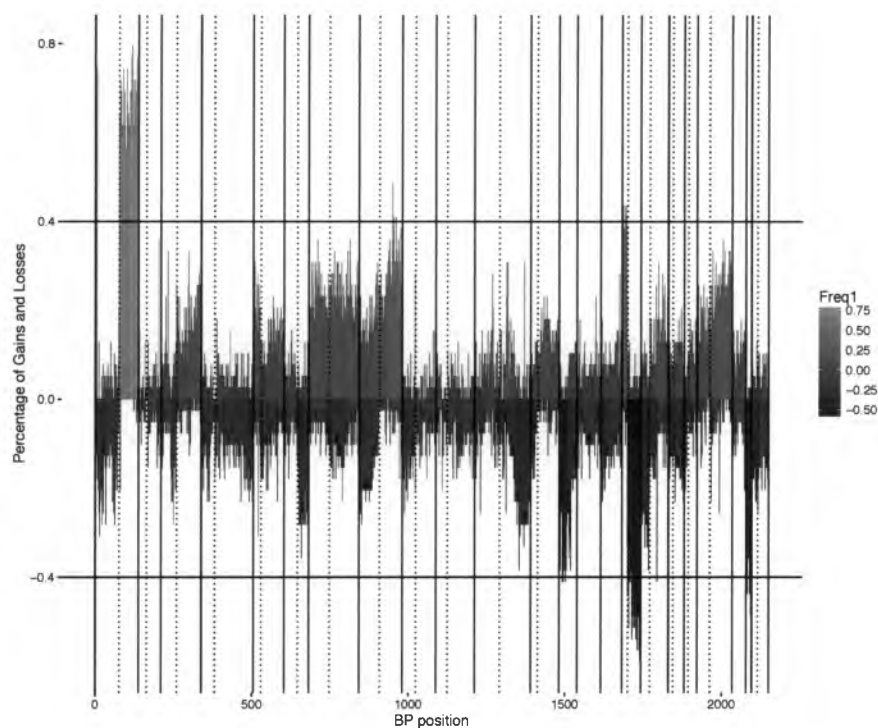


Figure 5.3: Frequency of Copy Number Values for Luminal A. Black straight and dotted lines are the same as Figures 5.1 and 5.2 above. Here we have the gains and losses from the luminal A patients.

39 luminal A tumors found gains in 1q, 8q and 16p; losses were located in 13q, 16q and 22q, with a slight loss in 17q. Region 8q was not considered an increase in Chin et al., however, was found to be an amplification in Figure 5.3. High-level amplifications were declared to be in regions 8p11-12, 11q13-14, 12q13-14, 17q11-12, 17q21-24 and 20q13 by Chin et al., but were not found with this method.

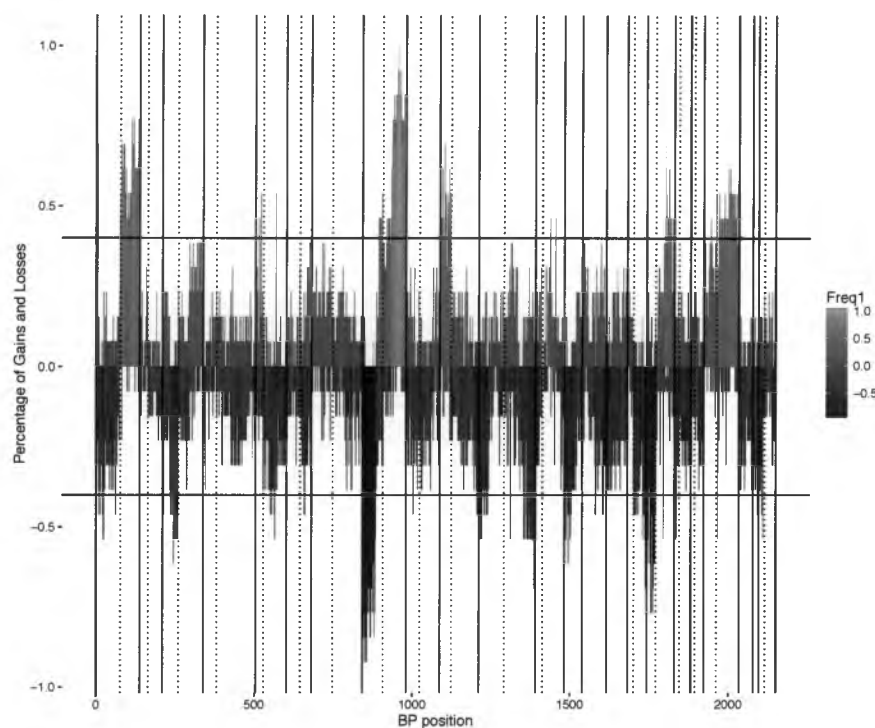


Figure 5.4: Frequency of Copy Number Values for Luminal B.

The last subtype we study, luminal B, has increases in 1q, 8q, 10p, 17q, and 20q; amplifications were slightly in sections 5p, 5q, 8p, 12q, and 20p. Losses are present in 1p, 3p, 8p, 11p, 13q, 16q, and 17p, with less deletions in regions 5q, 6q, 9p, 10q, 11q, 14q, 17q, and 23p. In regions 1q, 8q, 17q, and 20q, similarities with Chin et al. for amplifications were found. Losses in 1p, 8p, 13q, 16q, and 17p were also found in Chin et al.; 22q, however was a loss not found by Figure 5.4, but was found in Chin et al. The regions that were declared copy number increases or decreases for this study, but not Chin et al. were 10p,

5p, 5q, 8p, 12q, 3p, 11p, 6q, 9p, 11q, 14q, and 23p.

For all subtypes, several more amplifications and/or deletions were found in our study than in Chin et al. Reasons for this could potentially be due to using different methods with similar data and/or data provided to the public (our data) could include copy number clones that were not included in Chin et al. or vice versa. Since different methods were used for this portion of the study, this could result in some dissimilarities between the results. As for the data, we concluded that some of the clones that were in Chin et al. were not provided to us in our data and vice versa; leading to different outcomes for comparing the results for these two methods.

From Chin et al. and Figures 5.1 - 5.4, regions of high-level gains and losses have been found in chromosomes 1, 8, 11, 12, 13, 16, 17, and 20. Chin et al. locates regions of recurrent high-level amplifications in 8, 11, 12, 17 and 20, specifically 8p11-12, 11q13-14, 12q13-14, 17q11-12, 17q21-24 and 20q13. These regions of amplification can be viewed in Figure 5.5 below.

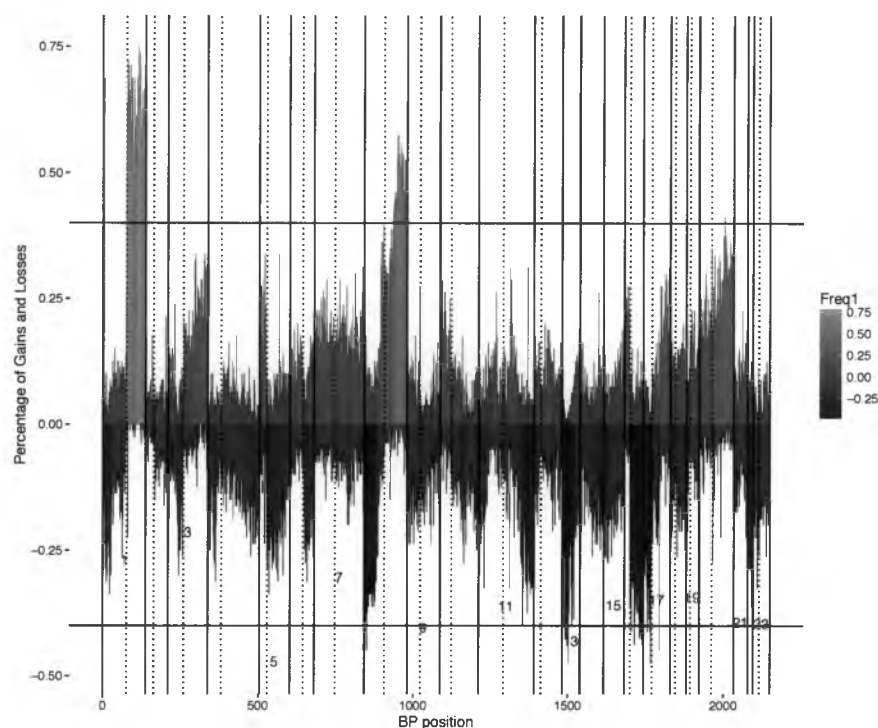


Figure 5.5: Frequency of Copy Number Values over all patients. Similar to Figures 5.1 - 5.4, this is a frequency graph for all clones, however, this graph represents all tumor samples.

For all tumor samples, illustrated in Figure 5.5, gains were found in 1q, 8q, 17q, and 20q; amplifications in 17q are not as high as other amplifications, but was found to have an amplification, detected by the methods described in Section 4.1.1., therefore, we consider this region to be amplified as well. Slight increased copy number in 5p, 5q, 8p, 12q, and 20p was found. Losses included 8p, 11q, 13q, 16q and 17q; a few clones showed deletions in 1p, 3p, 5q, 6q, 9p, 11p, 14q, 18p, 17p, 20q, and 23p. The study done by Chin et al.

included normally undetected amplifications in 8p11-12, 11q13-14, 17q11-12, and 20q13, therefore, they are not as apparent in our study in Figure 5.5.

5.1.2 A Correlation Study: Copy Number and Gene Expression in Four Chromosomes

Before running correlation, we checked the overlap of our data, to verify which genes were in our data set; 43 out of the 66 in Chin et al. were found in our data set. When running the correlation with bins of size 20Mb in the four regions (8p11-12, 11q13-14, 17q11-12 and 20q13), we found that only 37 of these genes were present. Therefore, we ran two different studies to find genes correlated with copy number. The first study included using bins of size 20Mb, placing the genes and clones in these bins based on genomic location, assigning genes to clones and then calculating correlation. Secondly, we found correlation directly with the 43 genes found in our data with copy number. We have a similar situation to the missing genes from the original data set from Chin et al., some of the clones were not found in the provided data set. 1948 of the 2149 clone IDs that are in the data set provided were also found in the supplemental data from Chin et al. Therefore, causing some of the genes to be assigned to different clones than what was done in Chin et al.

Chin et al. computed the pearson's correlation coefficients for 186 genes whose expression levels were significantly associated with copy number. This was done using the method of assigning genes to clones described in Section 3.0.2 and finding the genes that were

significantly correlated with copy number in all four chromosomes that were found to have amplifications in copy number (8p11-12, 11q13-14, 17q11-12, and 20q13). Out of these sections in the genome, they found that 66 genes were significantly correlated with copy number based on an adjusted p-value ($p < 0.05$).

In our study, pearson's correlation coefficient was computed for 914 genes in the four chromosomes that were found in sections of chromosomes 8p, 11q, 17q and 20q. There are several more genes studied in our study than Chin's study. We hypothesize this difference is because they found association between copy number and gene expression first, this resulted in 186 genes. These genes were then used to calculate correlation with copy number; resulting in 66 genes that were significantly correlated with copy number.

Our study included calculating correlation between 914 genes and their assigned clones. Of these 914 genes, 214 were found significantly correlated ($\text{Holm} < 0.05$) with copy number, making about 23 percent of the genes to be significantly correlated with copy number. 18 of the 214 significant genes matched those 66 found in Chin et al. For the 37 that were present, about 48 percent of them were found correlated with copy number.

When placing genes and clones into bins of size 20Mb, each of the four regions had either one or two bins. 8p11-12 had two bins, 11q13-14 also had two, 17q11-12 had one and 20q13 had two. Figures 5.6, 5.7, 5.8 and 5.9 below show heat maps that represent the

correlation value for all genes and clones in four of the seven bins. Rows represent genes, columns represent clones and the colors range from dark blue to dark red with white in between. The darker the red color, the stronger the positive correlation is; the darker the blue, the stronger the negative correlation is and white indicates that there is no correlation.

As discussed in the Biological Background, there are changes that directly affect the DNA (genetic changes); these changes are interesting since the DNA is transcribed into mRNA. Gene expression values measure the amount of mRNA in breast cancerous tissue versus a non-cancerous tissue sample. Therefore, significant, positive correlation between gene expression and copy number suggests that there are regulations of gene expression by copy number.

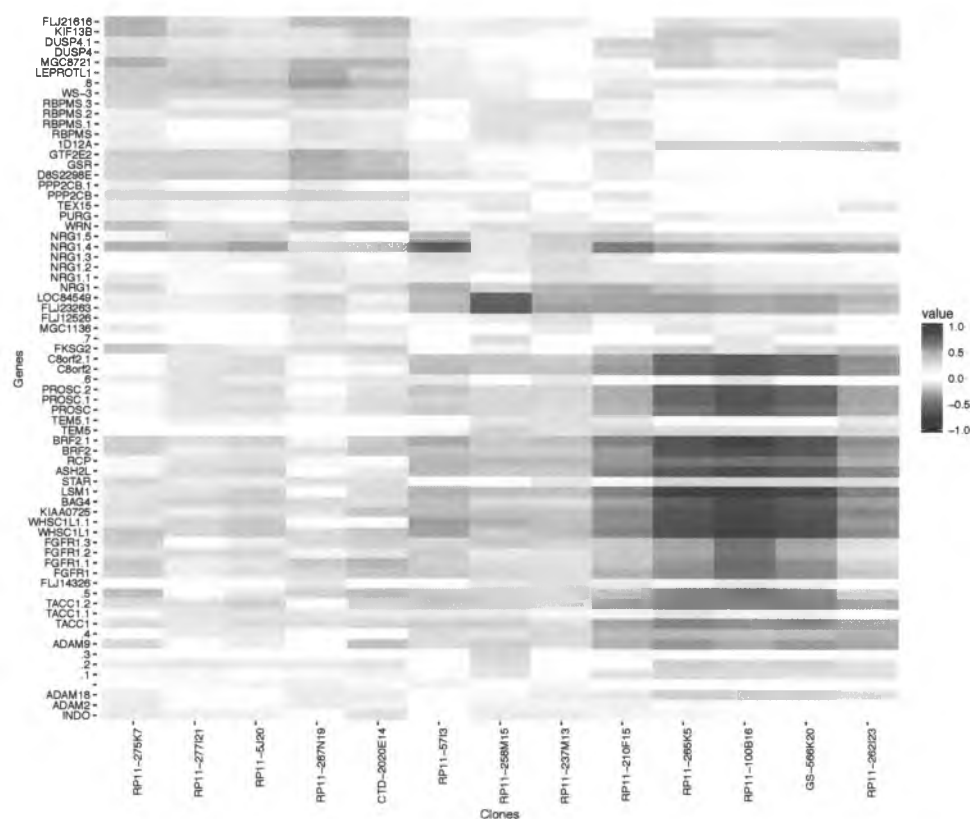


Figure 5.6: Heat map of correlation for bin 1, chromosome 8p11 - 12. Dark red indicates a higher, positive correlation (near 1); darker blue implies a stronger, negative correlation (near -1). Columns are clones and rows are genes. Genes on the left that are labeled numbers are expression probes that did not have the gene name for that probe.

Figure 5.6, shows correlation in 8p11-12 between gene expression and copy number. In the bottom right corner there is high correlation (dark red). When interpreting these correlation matrices, one needs to take into consideration the exact location of these genes and clones is not known, therefore correlation may extend over multiple genes/clones. Figure 5.6,

contains 9 of the significantly correlated genes that were also found in Chin et al. in the dark red region. These genes include PROSC, BRF2, ASH2L, LSM1, BAG4, WHSC1L1, FGFR1, TACC1 and ADAM9. ADAM9 is considered to be druggable due to presence of protein folds that like interactions with drug-like compounds [7]. Genes PROSC, BRF2, ASH2L, LSM1, BAG4 have the assigned clone RP11-265K5, WHSC1L1 is assigned to the clone RP11-100B16 and FGFR1, TACC1 and ADAM9 are assigned to GS-566K20.

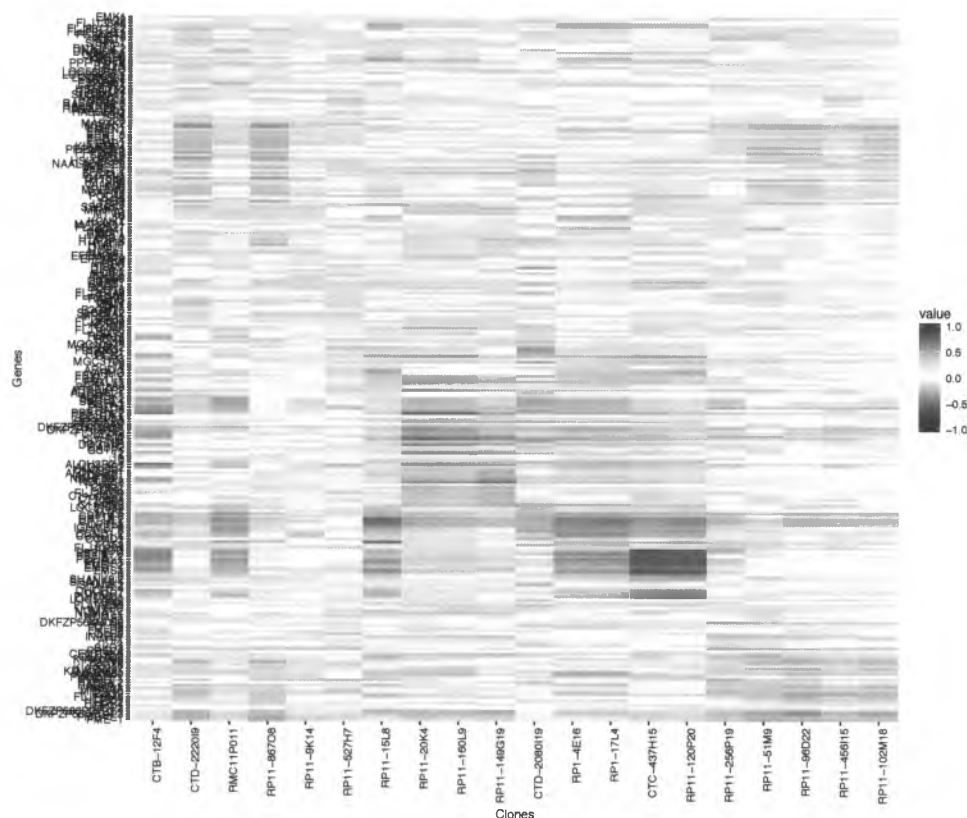


Figure 5.7: Heat map of correlation for bin 1, chromosome 11q13 - 14.

Figure 5.7 (11q13-14) shows the amount of strong correlation is minimal, however, there is a small area in the bottom right corner of the figure. This area of dark red contains 3 of the 66 genes that were found in Chin et al. [7]. Included here are the genes *CCND1*, *FADD* and *PPFIA1*. While none of these genes are druggable, they are an important part of molecular functions. *CCND1* has been known for being 10 Kbp within a site of recurrent tumorigenic viral integration in the mouse [7]. Clones associated with these

Figure 5.9: Heat map of correlation for bin 2, chromosome 20q13

Figure 5.9 clearly shows that there is more correlation in the upper middle region. In this area, 6 genes found in Chin et al. were found to be correlated. They include CSTF1, RAE1, PCK1, RAB22A, VAPB and NPEPLI. Their corresponding clones include GS-32I19, LLNLBAC-255K9, RMC20P073, RMC20P073, RMC20P073, and RMC20P073, respectively.

For our second study described at the beginning of this section, correlation was found for all of the 43 genes out of the 66 found in Chin et al., since there were mapping issues with our data set. For example, ERBB2 is a popular oncogene found in 17q12, however, in our data, the bp ERBB2 is mapped to is 17q21.1. This error is due to genomic remapping efforts done every few years and changing the bp position. In our data, the gene is located at 17q bp position 38231358, however, when found in a genome browser such as the UCSC genome browser, the bp position is 39700080. Therefore, when running this algorithm, certain genes may be left out of the final correlation. The heat map for only the 66 genes found by Chin et al. is shown in Figure 5.10.

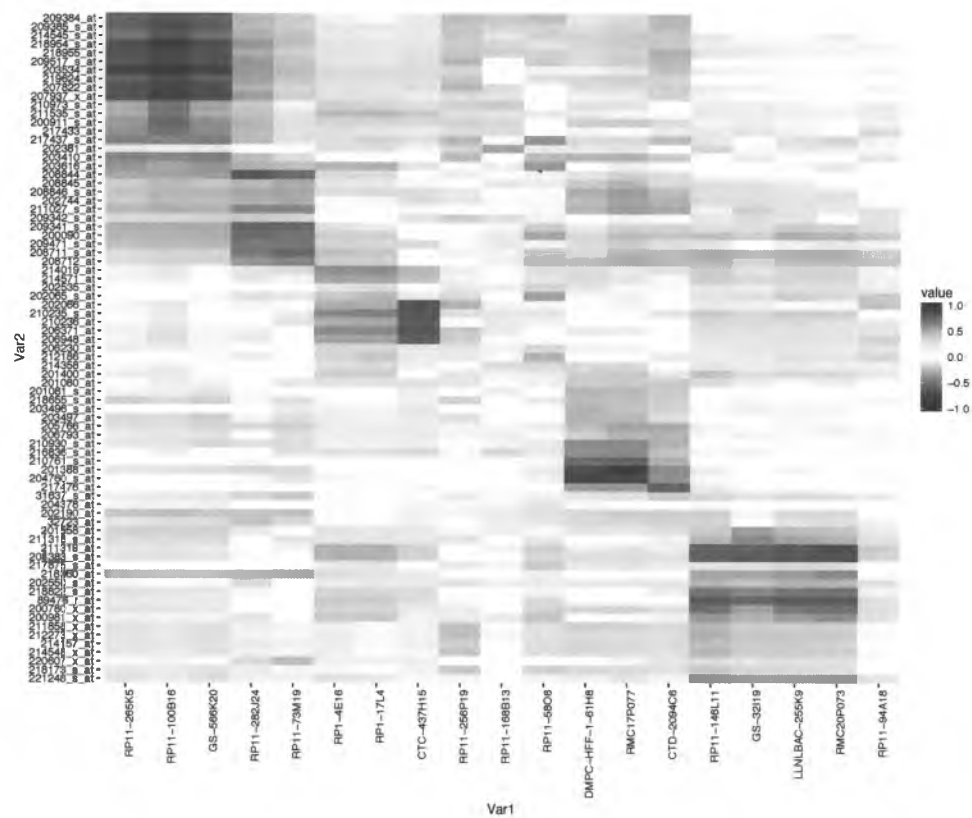


Figure 5.10: Heatmap sharing Correlation between genes and clones

Figure 5.10 shows the correlation matrix for 43 out of the 66 genes found significantly correlated with copy number in Chin et al. The figure shows a positive correlation along the diagonal; due to the fact that each clone (column) has one or more genes that are associated with it. Therefore, what can be observed here is the block-diagonal dark red areas in this figure.

These genes include FGFR1, FNTA, PNMT, ERBB2, NR1D1, GRB7 and TACC1; antibody inhibitors have been developed for genes such as FGFR1 and ERBB2. FNTA, PNMT, and NR1D1 are considered druggable based on proteins that favor interactions with drug-like compounds. PNMT and GRB7 are also therapeutic targets in these specific regions of amplification [7].

Figure 5.11 below shows the same correlation matrix with the names of the genes. The program R will not put duplicate genes and therefore, the first gene with that name is used.

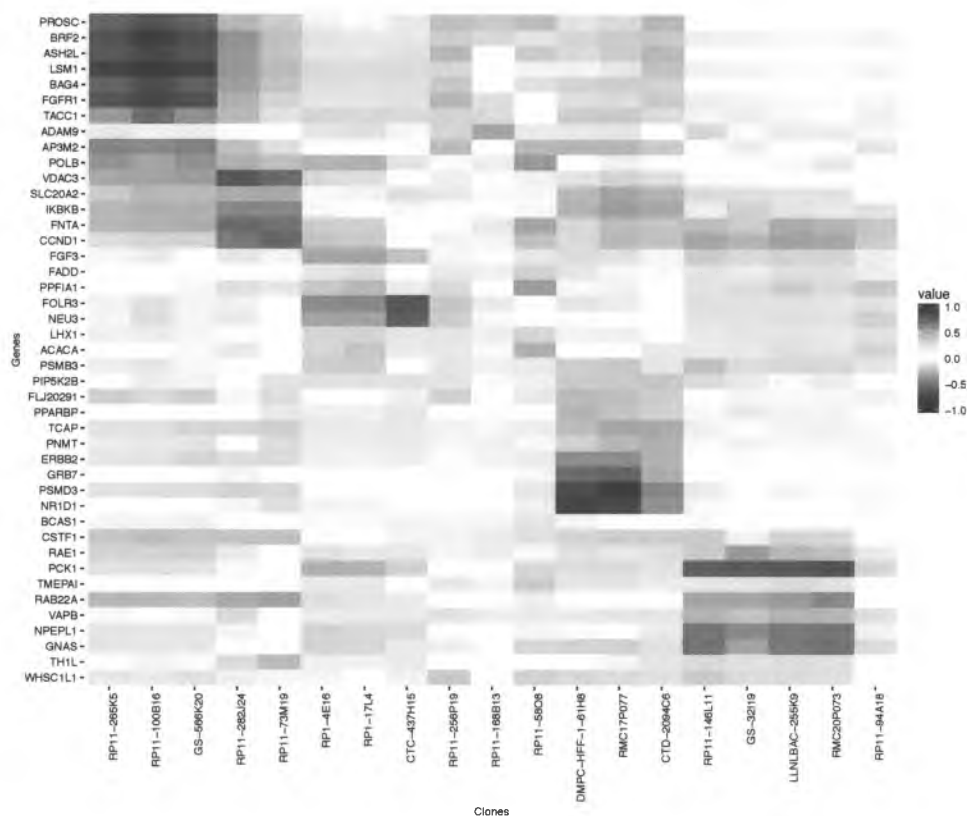


Figure 5.11: Heatmap sharing Correlation between genes and clones. Unlike Figure 5.10, Figure 5.11 has some of the genes that are represented by the gene expression probes from Figure 5.10. Since duplicate names are not repeated, we only see the first row for that gene.

Tables 5.1 and 5.2 show the genes that were significantly correlated with copy number. These tables present the numerical results corresponding to the heat maps in Figures 5.10 and 5.11. Gene expression probes, gene names, unigene ids, copy number clones, correlations between that gene and clone, and the adjusted p-values are listed below, from left to right.

Table 5.1: Correlation of Genes and Copy Number from our study

aCGH_label	exp_label	corr	p_values	affy
RP11-265K5	209384_at	0.695384071	7.51E-12	PROSC
RP11-265K5	209385_s_at	0.708756736	1.26E-12	PROSC
RP11-265K5	214545_s_at	0.629724218	1.13E-08	PROSC
RP11-265K5	218954_s_at	0.779430726	0	BRF2
RP11-265K5	218955_at	0.770916308	0	BRF2
RP11-265K5	209517_s_at	0.756221657	0	ASH2L
RP11-265K5	203534_at	0.856618368	0	LSM1
RP11-265K5	219624_at	0.7081149	1.26E-12	BAG4
RP11-100B16	218173_s_at	0.842756785	0	WHSC1L1
RP11-100B16	221248_s_at	0.858077987	0	WHSC1L1
GS-566K20	207822_at	0.485527883	0.000817533	FGFR1
GS-566K20	207937_x_at	0.455283761	0.004689487	FGFR1
GS-566K20	210973_s_at	0.550617776	1.01E-05	FGFR1
GS-566K20	211535_s_at	0.563213515	3.87E-06	FGFR1
GS-566K20	200911_s_at	0.608708052	8.31E-08	TACC1
GS-566K20	217437_s_at	0.604856338	1.18E-07	TACC1
GS-566K20	202381_at	0.530167941	4.46E-05	ADAM9
RP11-282J24	203410_at	0.771844746	0	AP3M2
RP11-282J24	209342_s_at	0.464769332	0.002762721	IKBKB
RP11-282J24	203616_at	0.567997616	2.66E-06	POLB
RP11-282J24	208845_at	0.704282589	2.51E-12	VDAC3
RP11-282J24	208846_s_at	0.67474429	9.17E-11	VDAC3
RP11-282J24	202744_at	0.625429878	1.72E-08	SLC20A2

Table 5.2: Correlation of Genes and Copy Number from our study continued

aCGH_label	exp_label	corr	p_values	affy
RP11-73M19	200090_at	0.705429871	1.88E-12	FNTA
RP11-73M19	209471_s_at	0.603904675	1.28E-07	FNTA
RP1-4E16	208711_s_at	0.553969429	7.86E-06	CCND1
RP1-4E16	208712_at	0.49474883	0.000463086	CCND1
CTC-437H15	202535_at	0.786296796	0	FADD
CTC-437H15	202065_s_at	0.769821312	0	PPFIA1
CTC-437H15	202066_at	0.748929532	0	PPFIA1
CTC-437H15	210235_s_at	0.782659289	0	PPFIA1
CTC-437H15	210236_at	0.747313151	0	PPFIA1
DMPC-HFF-1-61H8	205766_at	0.557888837	5.84E-06	TCAP
DMPC-HFF-1-61H8	206793_at	0.544103021	1.64E-05	PNMT
DMPC-HFF-1-61H8	210930_s_at	0.682891404	3.50E-11	ERBB2
DMPC-HFF-1-61H8	216836_s_at	0.825907867	0	ERBB2
RMC17P077	210761_s_at	0.839275473	0	GRB7
CTD-2094C6	201388_at	0.726891407	0	PSMD3
GS-32I19	202190_at	0.523638787	7.00E-05	CSTF1
GS-32I19	32723_at	0.564862024	3.40E-06	CSTF1
LLNLBAC-255K9	201558_at	0.774365788	0	RAE1
LLNLBAC-255K9	211318_s_at	0.772119197	0	RAE1
RMC20P073	208383_s_at	0.593799463	3.13E-07	PCK1
RMC20P073	218360_at	0.655701235	7.77E-10	RAB22A
RMC20P073	202550_s_at	0.75179977	0	VAPB
RMC20P073	218822_s_at	0.673983768	9.98E-11	NPEPL1
RMC20P073	89476_r_at	0.573748355	1.67E-06	NPEPL1

For Tables 5.1 and 5.2 a total of 29 genes and clones in these two tables; this is roughly 44 percent of the total genes that were found in our data and Chin et al. [7].

Point clouds (x -axis copy number and y -axis gene expression) from our data are displayed below to visualize the correlation between these genes and their copy number assignments. Genes that were within 100Kb of sites of recurrent tumorigenic viral integration in the mouse; TACC1, ADAM9, IKBKB, POLB, CCND1, and GRB7 point clouds are shown below in Figures 5.12 - 5.17.

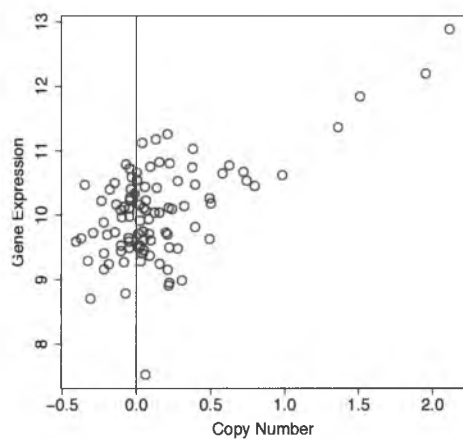


Figure 5.12: Gene TACC1 and copy number clone

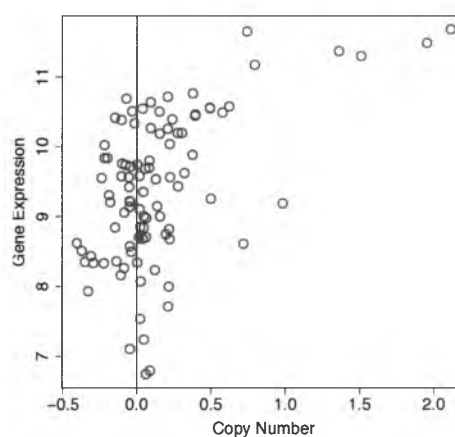


Figure 5.13: Gene ADAM9 and copy number clone

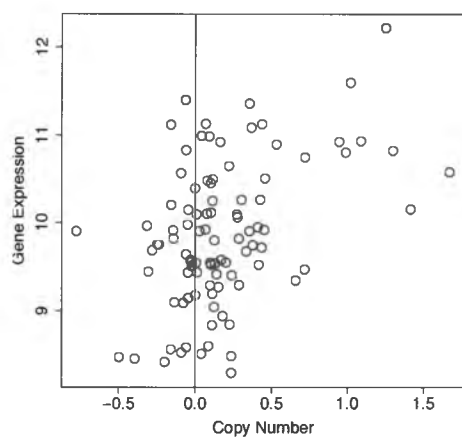


Figure 5.14: Gene IKBKB and copy number clone

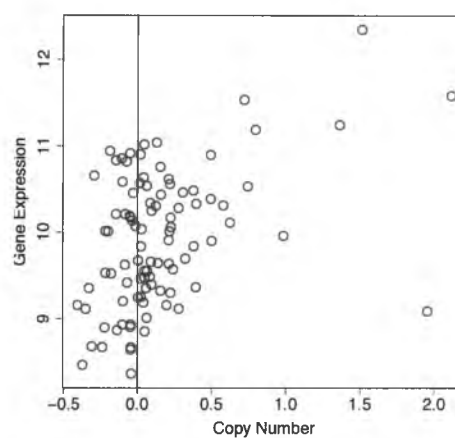


Figure 5.15: Gene POLB and copy number clone

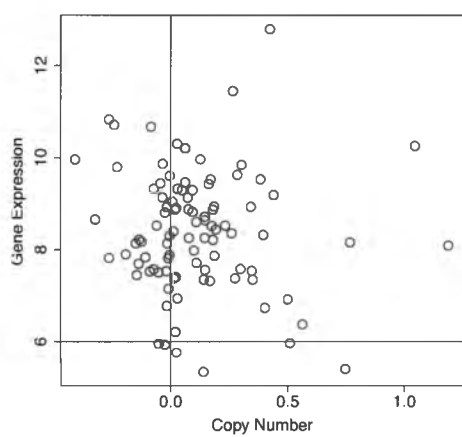


Figure 5.16: Gene CCND1 and copy number clone

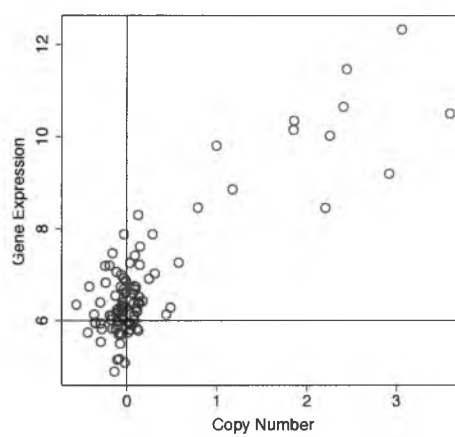


Figure 5.17: Gene GRB7 and copy number clone

These point clouds represent 6 genes found significantly correlated with copy number clones in Chin et al. [7]. Each point in the cloud is a different patient for that gene and it's copy number assignment. x -values represent copy number value and the y -values represent gene expression value for a probe containing that gene. Vertical lines represent the copy number average, 0 and horizontal lines represent the gene expression average, 6.

The figures that represent ADAM9, TACC1, IKBKB, and POLB have noisy data for copy number, since the points are mostly around the $x = 0$ line; however, all four also have high expression since they are higher than the average. There is a slight linear direction for the gene expression and copy number. This suggests that these genes are regulated by copy number. As the value of the copy number increases for all patients, the value of gene expression also increases. GRB7 demonstrates several points near the "origin", while having a linear relationship between the gene and copy number clone over all patients. ERBB2 also has a formation with a small cluster of noisy data that is near the "origin" and a linear relationship between the two variables. CCND1 is the least intuitive since it does not seem to have much of a pattern at all, with a correlation of roughly .55, but has a very small p-value; indicating that we cannot reject the hypothesis that the two variables are linearly related.

For both of our methods, we were able to find correlation between some of the 66 genes found significantly correlated with copy number in Chin et al. Our first study, with the

bins, produced 18 genes that were correlated with copy number. The second study, directly finding correlation between the 43 genes with the corresponding copy number clones, found that 29 genes were associated with copy number clones.

5.1.3 Clustering with Intrinsic Genes

In the study done by Chin et al., hierarchical clustering was done for “intrinsic” genes (discussed in Section 2.1.3) minus the genes that were statistically correlated with copy number for the expression data. Their results showed that tumors still resolve into the basal-like and luminal classes, however, ERBB2 (HER2) is lost. We reproduced this clustering in Figure 5.18.

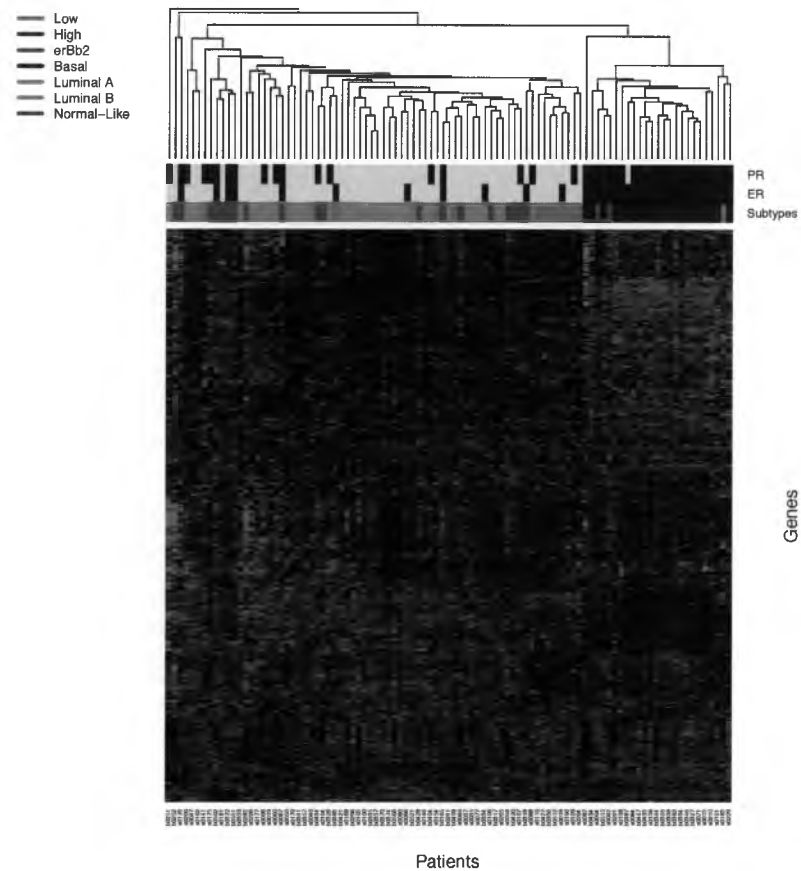


Figure 5.18: Hierarchical Clustering with Intrinsic Genes. Subtypes are color coordinated as red (basal), green (ERBB2), light blue (luminal A), orange (luminal B) and purple (normal-like). The subtypes are shown in the row above the heatmap. Above the row of subtypes, are the ER/PR status for each tumor; dark blue and light blue represent negative and positive ER/PR status, respectively. Green in the heat map indicates low copy number and red, high.

Similar to Figures 2.1 and 2.2, Figure 5.18 is the average hierarchical clustering of genes and patients. These genes are the 552 “intrinsic” genes from Stanford, of which 424 were found in our data. Figure 5.18, includes the 43 genes that were found to be correlated with copy number. It can be observed that basal clusters together the best with ER/PR negative tumors; the luminal patients cluster somewhat well, but the clusters for ERBB2 and normal-like are lost. These results were similar to the hierarchical clustering done by Chin et al.

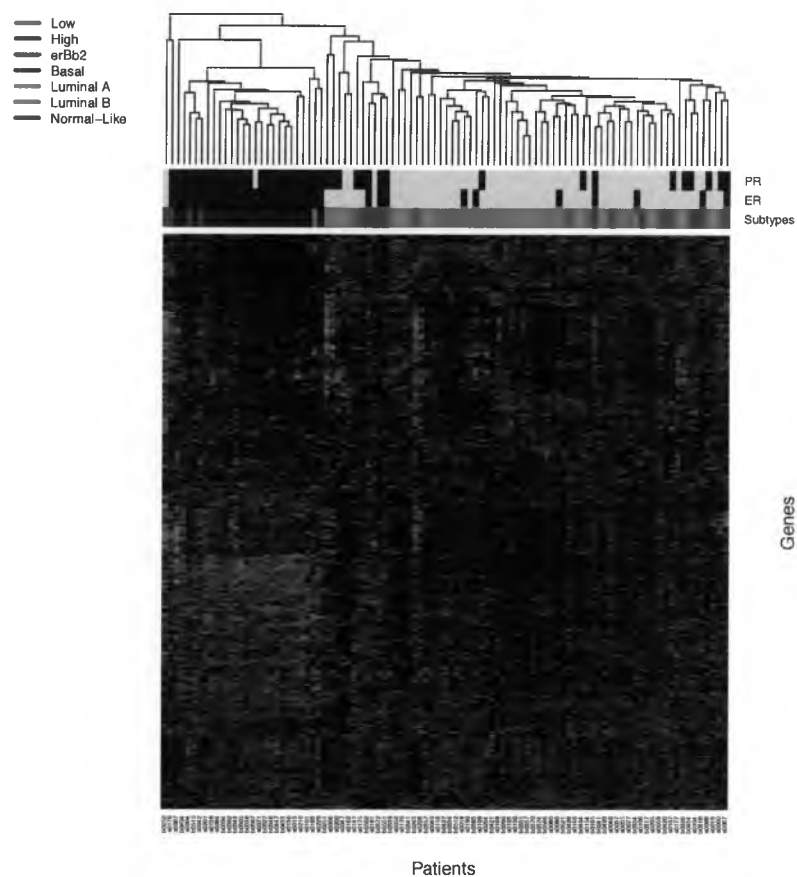


Figure 5.19: Hierarchical Clustering with Intrinsic Genes. Similar to Figure 5.18, 5.19 in hierarchical clustering of “intrinsic” genes. However, in this case, we delete the 33 genes that were found significantly correlated by Chin et al.

Figure 5.19 represents the average hierarchical clustering with the “intrinsic” genes from Stanford, without the 29 genes that were found significantly correlated in Section 5.1.1. This figure has roughly the same outcome as Figure 5.18, the basal-like and luminal subtypes cluster well; the ERBB2 tumors, although not entirely clustered together have clustered better than before.

5.1.4 Conclusion

Using CBS to find the frequencies of each subtype, Chin et al. located four regions of high-level amplifications between all tumor samples; 8p 11-12, 11q 13-14, 17q 11-12 and 20q 13. To verify this with the data provided to us, we created frequencies representing the percentage of tumor samples above 0.2 or below -0.2. We were able to confirm that we also had these regions as amplifications for all tumor samples across the entire genome.

Our first study with correlation included finding 214 significantly correlated genes with copy number ($Holm < 0.01$) in regions 8p11-12, 11q13-14, 17q11-12 and 20q13, out of 914 total genes in these regions. 18 of those 214 genes found correlated were in Chin’s 66 genes.

Secondly, we studied correlated with the 43 genes that were the same as the 66 genes found in Chin et al., 29 of these genes were significantly correlated with copy number. Patients were clustered using average hierarchical clustering before and after deleting

correlated genes. This study showed that genes regulated by copy number do not represent a significant proportion of the genes defining the molecular subtypes.

Our next study consists on finding significant regions in the genome specific to different subtypes of breast cancer using our persistent homology method. The results to this method is described in the next section.

5.2 Results using persistent homology

Using the method described in 4.1.3, regions of significance were found using copy number as our x -value and gene expression as our y -value. Significant regions were found for each of the subtypes vs. all other subtypes.

5.2.1 Significant regions

Before running the program to produce results for persistent homology, we chose the optimal increment of the filtration parameter. This process is described in Section 4.2.4; we used the average of the minimum over all patients for all segments, which was the value 0.027. 0.02 was used as our first ϵ , however, 0.01 was also chosen as an iteration because it produced more significant regions. With the 1827 clones and 1827 averaged genes, significant regions were found based on a point clouds for patients of one subtype vs. all other subtypes. These regions of significance do not fully inform the relationship

between copy number and gene expression. For example, a cluster in the first quadrant that is roughly at the same distance away from the “origin” then there will be no way of detecting which quadrant it is in. In this case, the “origin” is (0,6). Quadrant 1 represents cases where over expression of genes and amplifications of clones are; therefore, this is the situation where genes are regulated by amplifications. Quadrant 2 contains excess expression, with underlying deletions genes are deregulated by copy number. The third quadrant negatively detects that these genes are regulated by copy number. The fourth quadrant, shows amplifications with under-expression of genes. Table 5.3 shows significant regions found by the persistent homology method.

Table 5.3: Significant regions for each subtype

Subtypes	Chrom	Arm	Segment	Cyto.Begin	Cyto.End
Basal	2	p	1	p25.3	p11.2
Basal	2	q	1	q11.1	q31.1
Basal	2	q	2	q22.3	q34
Basal	3	q	2	q21.3	q25.1
Basal	4	q	7	q31.21	q32.3
Basal	5	q	3	q15	q31.1
Basal	11	q	6	q22.1	q23.2
Basal	16	q	1	q12.1	q22.1
Basal	16	q	2	q21	q24.3
Her2	17	q	1	q11.2	q21.31
Lum A	10	q	1	q11.21	q21.3
Lum A	10	q	2	q21.1	q22.2
Lum B	8	q	2	q12.1	q22.1
Lum B	8	q	3	q21.11	q24.13
Lum B	8	q	4	q22.2	q24.3

More regions were found significant for copy number from Arsuaga et al. using the Horlings et al. data set [4], [22]. In Arsuaga et al., [4], significant regions for copy number were identified between different subtypes. Table 4.1 in Chapter 4 shows the regions of copy number changes for each subtype. We do not see many similar regions for each of the subtypes. For basal, we did find sections of 3q, 4q, and 5q significant; HER2 subtype found a section of 17q significant and luminal B has significance in most of 8q. Arsuaga and colleagues found that 8q for luminal B was significant with copy number aberrations, HER2/ERBB2 significant regions included 17q11-11.2, 17q12-q21.31 and 17q21.31-q22. For basal, Arsuaga and colleagues found 3q and 5q significant.

In the study by Natrajan et al., [29], genes whose expression was correlated with copy number were found in 1q, 8p, 8q, 11q13, 17q12-21, 17q22-25 and 20q13. These regions were similar to those reported by Chin et al., however, most of these were not found significant with the persistent homology method. For basal-like subtypes, Natrajan et al. found that 1q22-24, 1q44, 8q24.1, 10p15 and 19q12-13 contained genes whose expression correlates with copy number. HER2 genes that were correlated with copy number were found in 8p12, 8q22-24, 11q13-14, 17p11, and 17q23. In our study, part of 17q was considered significant as well. The results produced by persistent homology do not match these results well, therefore, we must understand why that is. To dive deeper into these issues, we analyzed point clouds.

5.2.2 Betti Curves

As discussed in Section 5.1.4, the Betti curves for these significant regions must have the test above the control. This is because we need the test (subtype) to connect at a “slower” rate than the control (other subtypes). Figures 5.20, 5.21, 5.22 and 5.23 show the Betti curves for regions 10q s2 luminal A vs others, 8q s3 luminal B versus others, 17q s1 HER2 vs others, and 11q s6 for basal versus others.

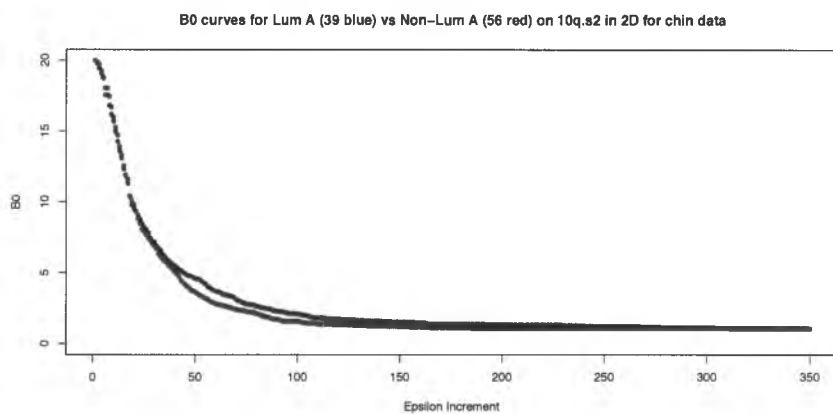


Figure 5.20: Betti curves for Luminal A versus others in 10q segment 2

Figure 5.20 shows that the blue curve (test) is above the red curve (control). The x -axis is the β_0 increment and the y -axis is the average β_0 values for all patients in either the test or control at each ε iteration.

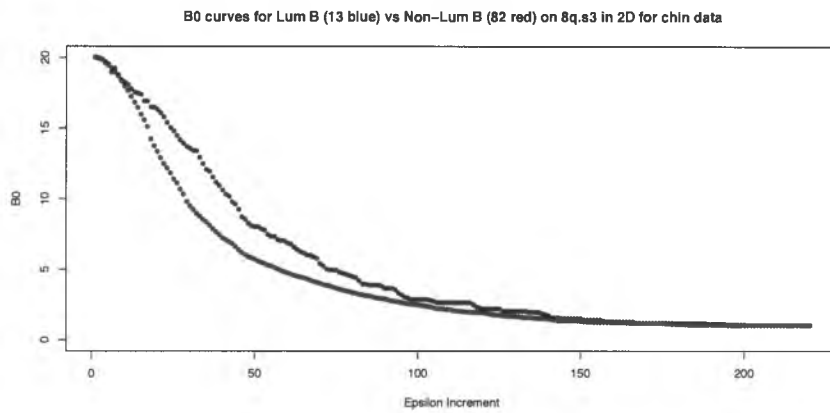


Figure 5.21: Betti curves for Luminal B versus others in 8q segment 3

Figure 5.21 illustrates the Betti curves for luminal B patients (test) and the other subtypes (control). This figure also has the test above the control and that is because the points connect at a “slower” rate or at a higher ϵ .

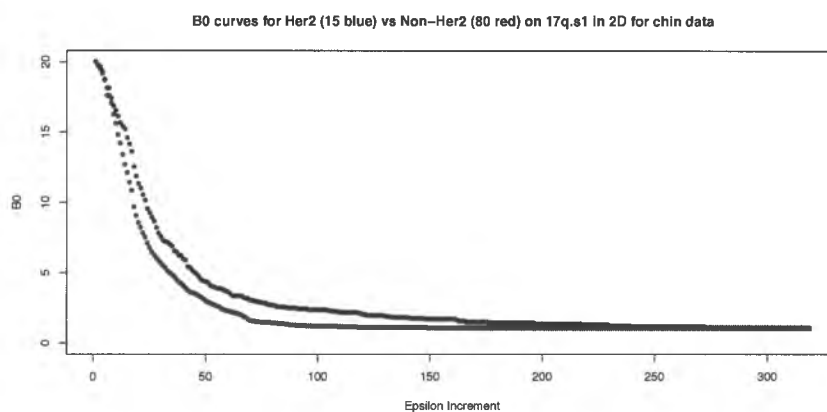


Figure 5.22: Betti curves for HER2 versus others in 17q segment 1

Here is the Her2 subtype in 17q segment 1. It shows that the test above the control for the average β_0 numbers (y-axis) and the ϵ increments (x-axis)

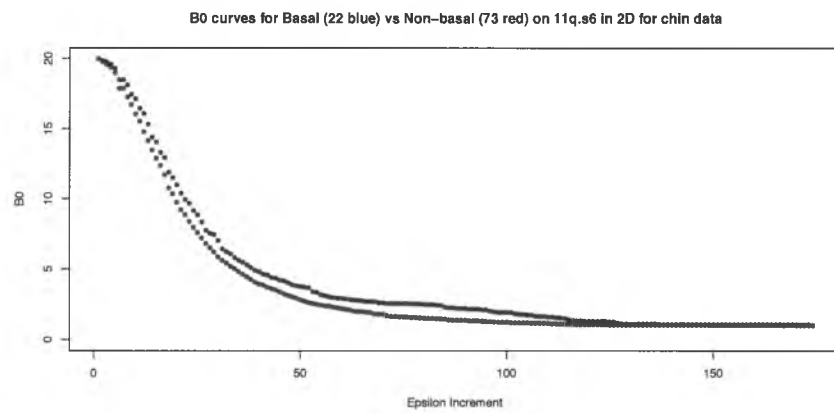


Figure 5.23: Betti curves for Basal versus all other subtypes in 11q segment 6. In these figures the blue is the test curve, and the red curve is the control curve.

It is necessary for the blue curve to be above the red curve for this to be significant because it implies that it took longer for the points in the point cloud to connect. Originally, our study had 21 significant segments in the genome (adjusted p-value < 0.05), however, when the betti curves were observed, we had to reject some of these due to the control curve above the test.

5.2.3 Correlation Between Copy Number and Averaged Gene Expression

To understand which genes and clones were producing the results from Table 5.3, correlation was found between copy number clones and averaged genes in regions. Averaged genes for each clone were named the same as their corresponding clone for simplicity. This produces an $m \times m$ matrix of correlations and adjusted p-values. Figures 5.24, 5.25 and 5.26 show the correlation between copy number clones and averaged genes for regions 2p segment 1 (basal tumors), 17q segment 1 (HER2 tumors), and 8q segment 4 (luminal B tumors), respectively.

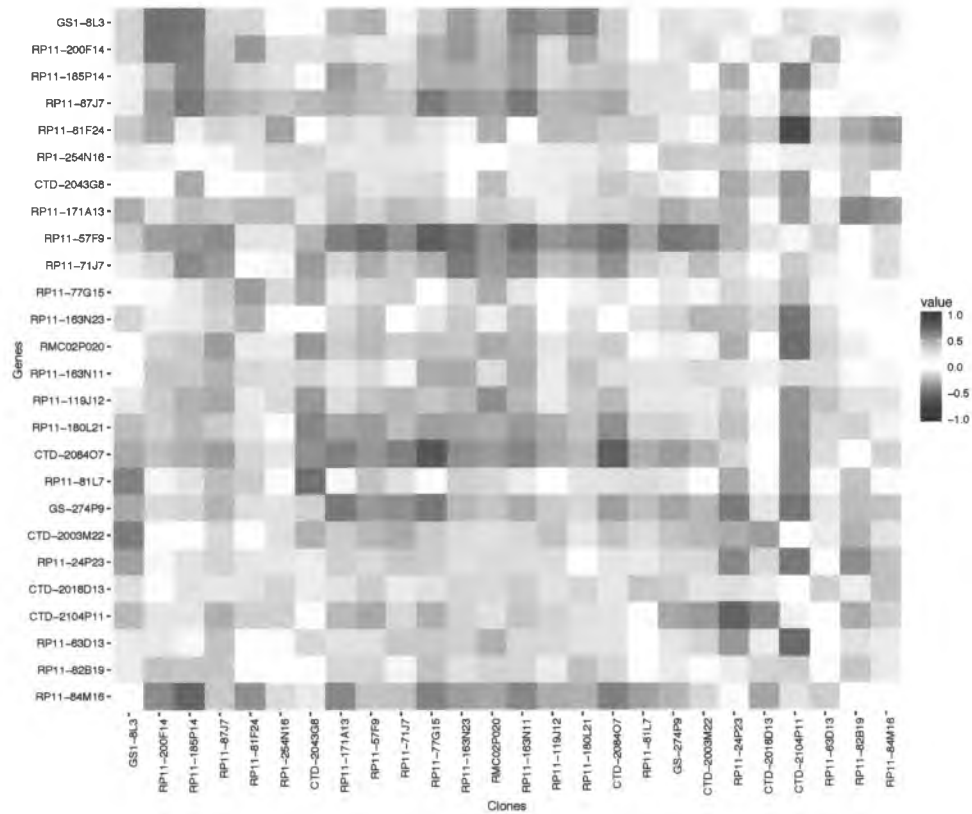


Figure 5.24: Correlation between copy number (columns) and averaged gene expression (rows) for 22 basal tumors in 2p 25.3 - 11.2. Darker red indicates strong, positive correlation and dark blue indicates strong, negative correlation, white indicates no correlation.

Figure 5.24 indicates a poor correlation between copy number and averaged genes in 2p25.3-11.2; along the diagonal region there is little correlation, indicating many of the clones and averaged genes are not significantly correlated. For the correlation of these clones and averaged genes, only one clone and averaged gene was found significant, however, it was not in the diagonal region, therefore, we did not study this region further. This is an example of a region that lacks regulation of averaged genes by copy number. However, this could be because we are looking at the average of genes, as opposed to the genes themselves. Figure 5.25 illustrates correlation between copy number and averaged gene expression for the significant region for all HER2 subtypes.

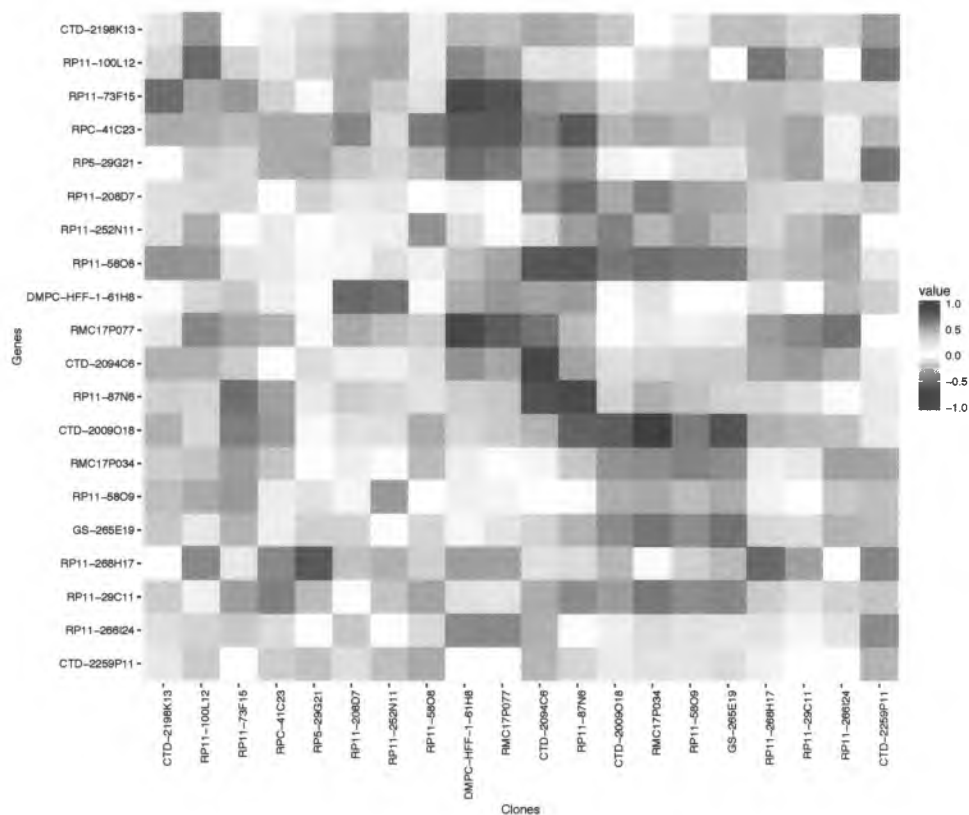


Figure 5.25: Correlation of Copy Number and Averaged Gene Expression for 17q 11.2 - q21.31 for 15 HER2 tumor samples. Red and blue are the same as Figure 5.24 and rows are averaged genes, while columns are copy number clones.

By observing Figure 5.25, we can see that there are regions along the diagonal from clones and averaged genes RMC17P077 - CTD-2009O18. Three sections out of this region were found significantly correlated; averaged genes associated with RMC17P077 and clone DMPC-HFF-1-6IH8, CTD-2094C6 for both averaged genes and copy number clone and CTD-2009O18 averaged genes and RMC17P034 were all found correlated with an adjusted p-value less than 0.05.

The last heatmap of correlation is luminal B tumors, which shows more correlation along the diagonal than Figures 5.24 and Figure 5.25.

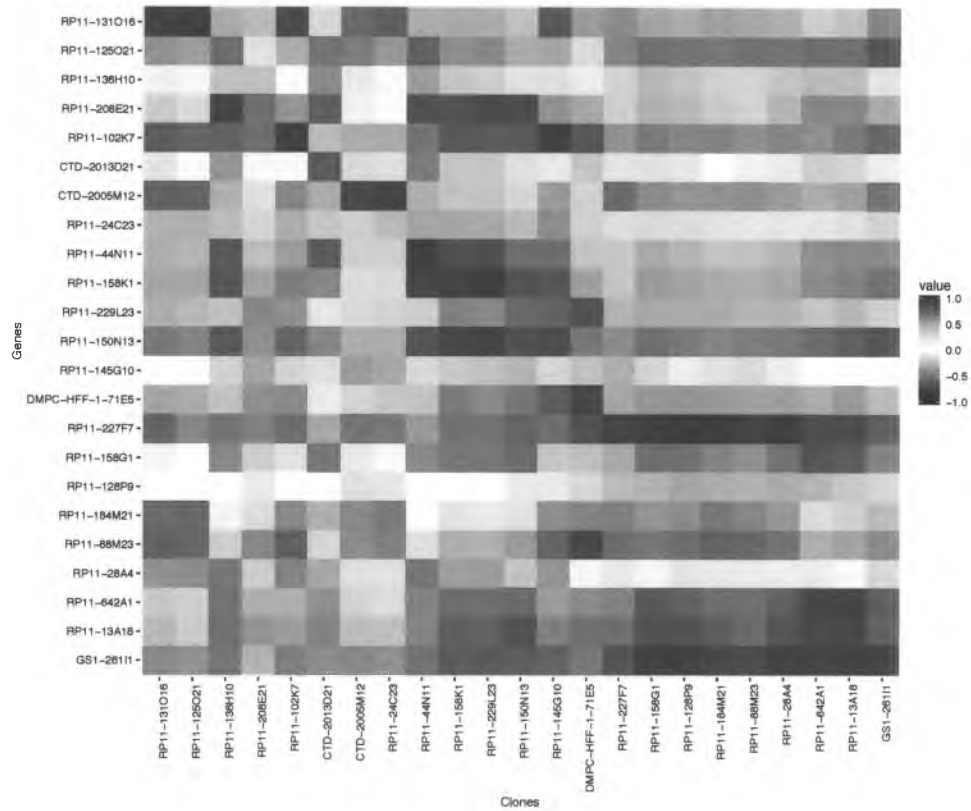


Figure 5.26: Correlation of Copy Number and Averaged Gene Expression for 8q 22.2 - 24.3 for 13 luminal B tumor samples. Again, we have red and blue coloring, same as Figures 5.24 and 5.25.

Figure 5.26 indicates several areas along the diagonal with correlation. Clones and averaged genes with significant correlation include averaged genes representing id RP11-131O16 and clone RP11-125O21, clones RP11-102K7, RP11-44N11, and GS1-261I1, with averaged genes representing the same clone id, averaged genes representing RP11-158K1 and clone RP11-229L23, clone RP11-158G1 and averaged genes representing clone id RP11-227F7, and clone RP11-13A18 with averaged genes representing GS1-261I1. Table 5.4 shows the number copy number clones and averaged genes found correlated for the four different subtypes and regions in Table 4.4.

Table 5.4: Number of Copy Number Clones and Averaged Genes found correlated

Number of Correlated Clones and Genes	Subtype	Chrom	Arm	Cytobands
2	Basal	3	q	22.3, 23, 24
5	Basal	4	q	31.21, 31.22, 31.3, 32.3
2	Basal	11	q	23.2
9	Basal	16	q	12.2, 22.1, 23.3, 24.1
3	HER2	17	q	12, 21.1, 21.2
4	Luminal A	10	q	22.1, 22.2
8	Luminal B	8	q	22.2, 24.13, 24.21, 24.3

With clones and average genes from Table 5.4, we calculated the correlation for the clones and genes for these diagonal regions. Figures 5.27, 5.28, 5.29 and 5.30 show the correlation for these diagonal regions for each subtype.

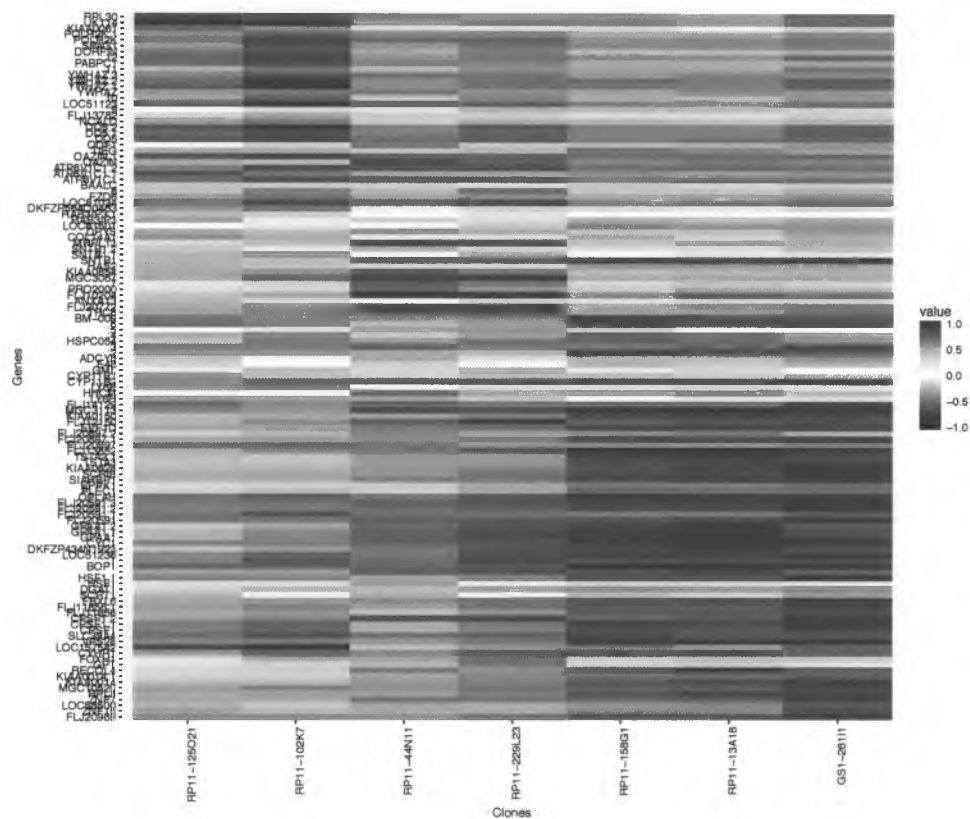


Figure 5.27: Correlation of Copy Number and Genes for 13 Luminal B patients in 8q 22.2 - 24.3. Dark red is a strong positive correlation, while dark blue is a strong negative correlation. Genes are the rows and copy number clones are the columns.

Figure 5.27 indicates positive correlation along the general diagonal section. With this heat map, there are 18 genes that are regulated by copy number in this region of the genome for luminal B subtypes. Two of these genes YWHAZ and SCRIB (CRIB1, SCRB1, SCRIB1) have been considered cancer causing genes by mouse insertional mutagenesis experiments [10]. SCRIB has been identified for encoding a protein involved in tumor suppressor pathways [15]. Figure 5.28 shows the correlation between genes and copy number for the HER2 tumor samples in 17q. None of these genes were found in the 66 genes in Chin et al. or the intrinsic genes.

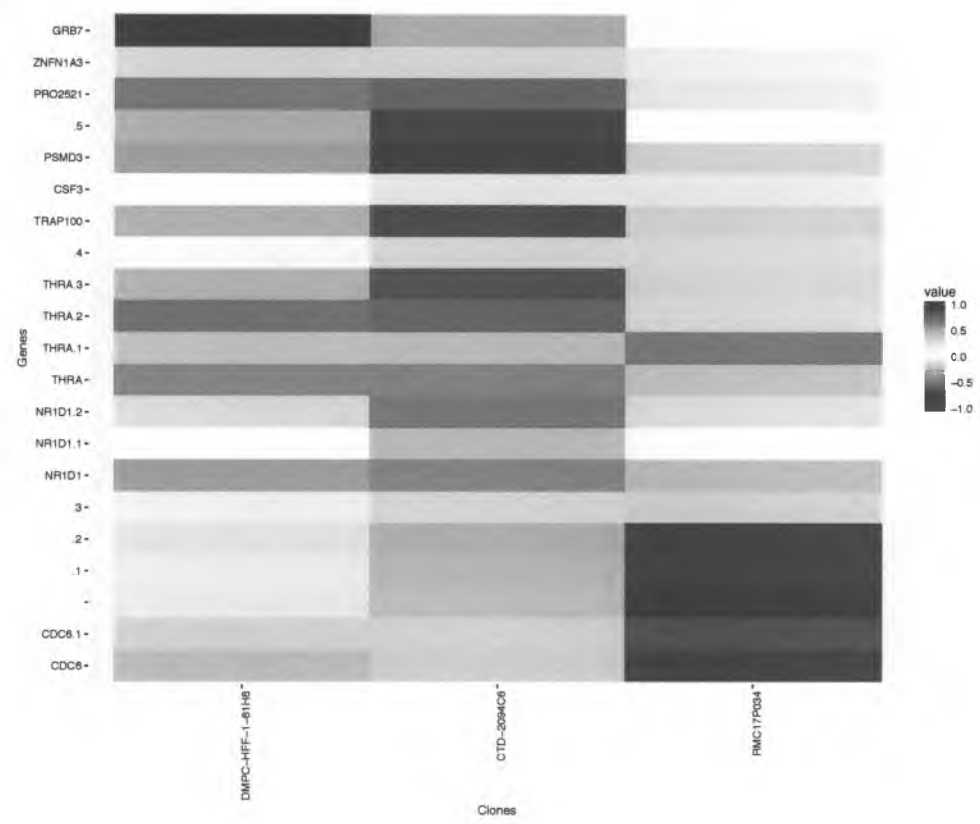


Figure 5.28: Correlation of Copy Number and Genes for 15 HER2 patients in 8q 22.2 - 24.3.

For Figure 5.28, we found two genes that were correlated, as they were in Chin et al.; GRB7 and PSMD3. These genes are not therapeutic targets, however, GRB7 is within 10Kbp of sites that are recurrent tumorigenic viral integration in the mouse, as discussed before [7]. GRB7 was also found to be regulated by copy number in Natrajan et al. [29]. GRB7 and TRAP100, found in our results are in the intrinsic gene list. The correlation of these genes can be observed in the heatmap in Figure 5.28.

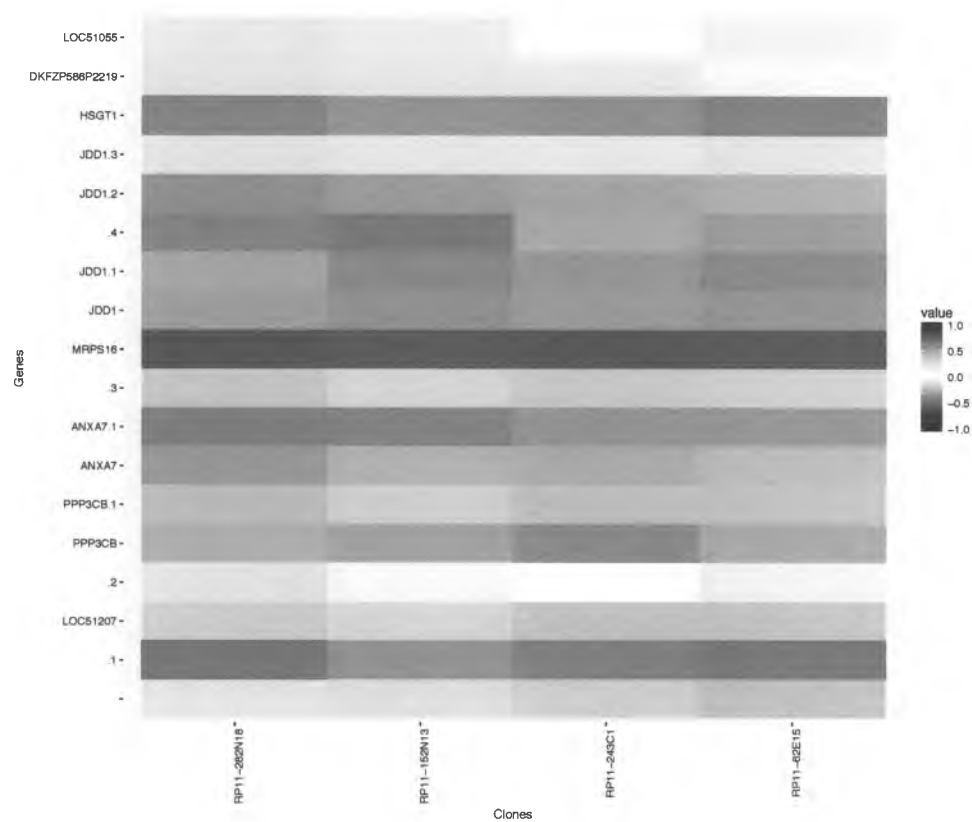


Figure 5.29: Correlation of Copy Number and Genes for 39 HER2 patients in 10q 21.1 - 22.2.

Five genes,(HSGT1, JDD1, MRPS16, ANXA7, and PPP3CB), were found significantly correlated with copy number in Figure 5.29. One of these genes, ANXA7 was found in the intrinsic gene list from Stanford. These genes were not directly related to breast cancer, however, PPP3CB is a protein coding gene with a related pathway that includes the immune system. The last heat map for this study that will be displayed contains a section of 4q for the basal tumor samples.

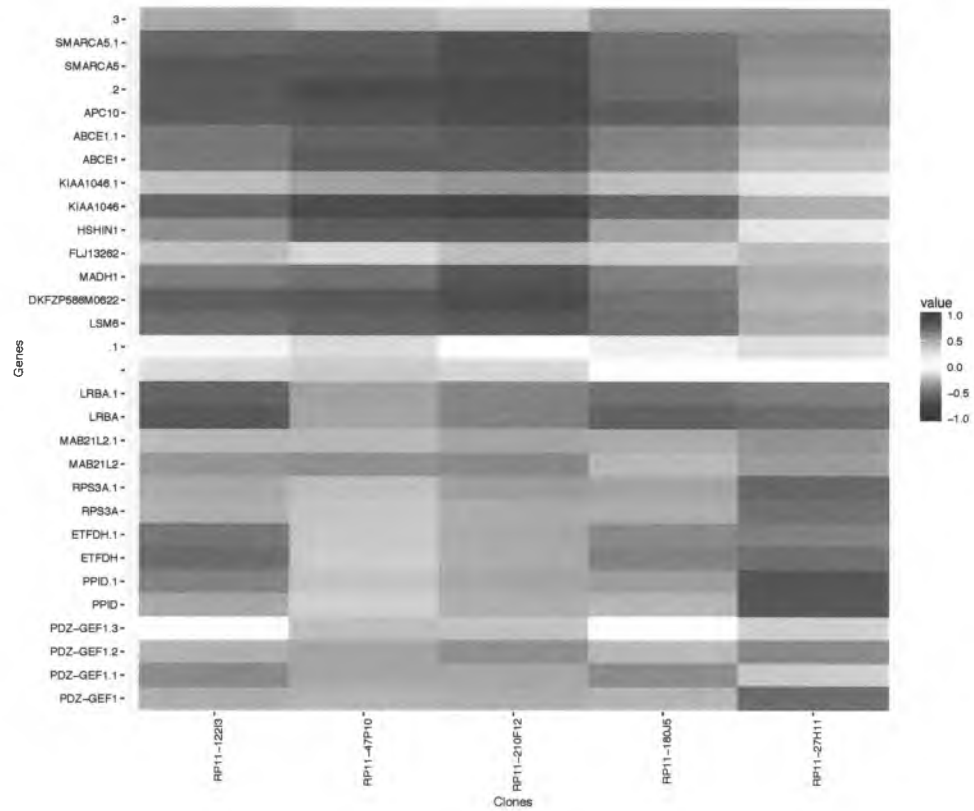


Figure 5.30: Correlation of Copy Number and Genes for 22 basal patients in 4q 31.21 - 32.3.

For this section of the genome, 11 genes were found correlated with copy number, was found in the intrinsic gene list; LRBA. Genes PDZ-GEF1 (RAPGEF2) and ANAPC10 have related pathways to the immune system as well and the gene SMARCA5 is thought to be a transcription regulator of certain genes; related pathways of this gene include DNA damage and cell cycle.

In the region 16q 21 - 24.3, another gene that we found correlated with copy number was found in the intrinsic gene list; HRIHFB2206 is the name of that gene. This gene was found correlated with copy number in both regions 16q 21 - 24.3 and 16q 12.1 - 22.1 due to overlapping of clones and averaged genes for the persistent homology method.

6 genes were identified as correlated with copy number for the luminal B tumor samples in 8q 21.11 - q24.13, however, none were found in Chin et al. or in the intrinsic gene list. In section 3q 21.3 - q25.1 only 2 genes that were correlated with copy number.

Our study showed that out of the 58 genes correlated with copy number, only 5 of them were from the intrinsic gene list. Two out of these 58 were found to be correlated in Chin et al. from the 66 genes in regions 8p11-12, 11q13-14, 17q11-12 and 20q. All regions reported by Chin et al., except 17q11-12, were not found significant using persistent homology. We observe the point clouds for regions that were expected to be significant. Figures 5.31 and 5.32 show point clouds for basal subtype patients vs. all other subtypes

for a region that was found significant (4q 31.21 - 32.3) and a region that was not found significant (8p 21.2 - 11.2). Figures 5.31 and 5.33 are all basal subtypes and Figures 5.32 and 5.34 are all non-basal subtypes for the two regions. We also found that genes such as PPP2R3A, APC10, DKFZP586M0622, ZW10, FTS (AKTIP), TERF2, CDC6, and SCRIB found correlated with copy number are genes that encode for proteins involved in the cell cycle.

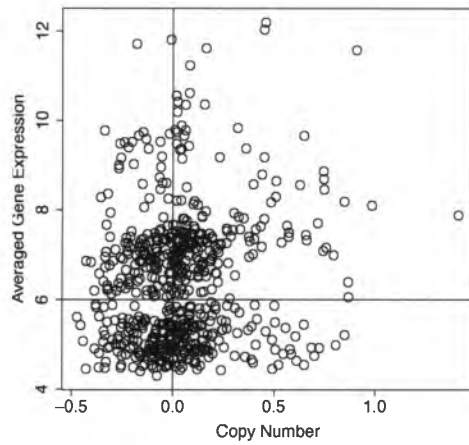


Figure 5.31: Point cloud for basal patients in 8p 21.2 - 11.2

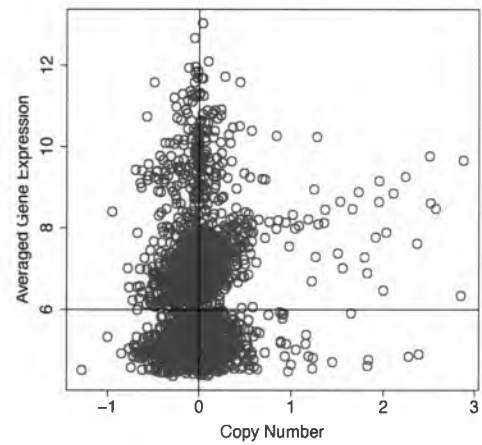


Figure 5.32: Point cloud for non-basal patients in 8p 21.2 - 11.2

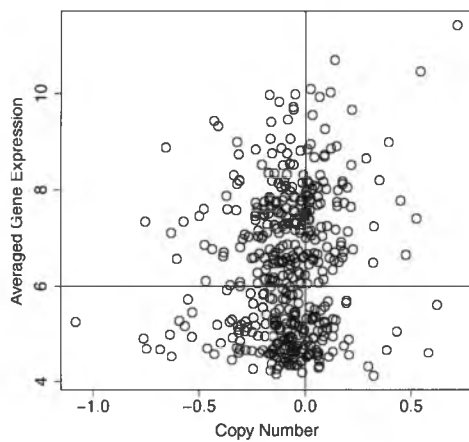


Figure 5.33: Point cloud for basal patients in 4q 31.21 - 32.3

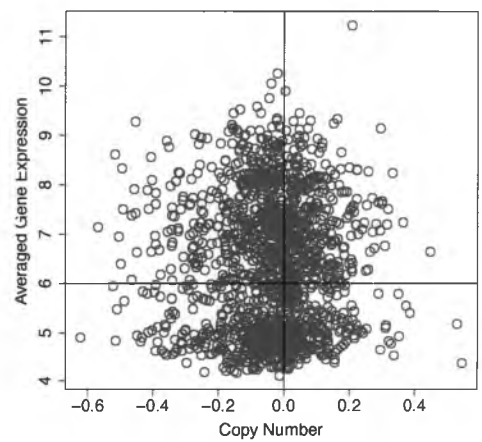


Figure 5.34: Point cloud for non-basal patients in 4q 31.21 - 32.3

These four figures show the relationship between copy number and gene expression for different regions in the genome for all patients. It is clear that in 8p, more points in the cloud are further away from the center for the non basal subtypes than the basal subtypes; therefore, confirming the lack of significance for this region with the persistent homology method. On the other hand, Figures 5.33 and 5.34, representing 4q, produces the opposite situation. The black (basal) point cloud is more spread out than the red (non-basal) point cloud, therefore confirming why this region would be significant for this subtype. For all regions that were considered significant with persistent homology, this is the case.

5.2.4 Conclusion

Using persistent homology, we located 15 segments in the genome that were significant. Three of these segments were found in luminal B subtypes vs. all other subtypes, one segment was found significant for HER2 tumor samples, two were found in luminal A and nine were found in basal. From these 15 segments, we were able to confirm that a total of nine diagonal elements in the correlation matrices had significantly correlated clones and averaged genes. It was then confirmed that from these correlations, 58 total genes were correlated with copy number. Five of these genes were found in the intrinsic and two genes were found in the 66 genes Chin et al. found correlated with copy number.

Some significant regions that were found using persistent homology were also found in

Chin et al. and Natrajan et al. [7], [29]. These areas that agreed with our findings include 3q, 4q, 5q, and 16q, for basal subtypes, 17q for HER2, and 8q for luminal B. Several regions that have previously been studied for amplifications or deletions of copy number were not produced in our results. However, in this case, we extended our study to copy number and gene expression, which provides different information. These are not the significant regions of copy number amplifications and deletions only, but areas of regulated and deregulated genes by copy number, therefore, it is not necessarily the case that these regions should be significant for the persistent homology study.

Other differences between the studies done by Arsuaga et al. and Natrajan et al. can confirm the inconsistent results. Persistent homology studies the topology of one group versus another group. In our case, we studied one subtype versus all other subtypes. By doing this, our results may not be the same since the study done by Chin et al. was for all tumor samples. Another issue that arises when executing the persistent homology method is the assignment of genes to clones. Averaging genes loses a lot of biological information that could be key to our study; new methods of assigning genes to clones could provide a more accurate result.

Chapter 6

Appendix A

Here are the list of intrinsic genes from Stanford.

Table 6.1: First 180 intrinsic genes from Stanford

FLJ10700	KIAA0205	EPHX2	FZD6	NQO1
FLOT2	MGC12685	JAM1	PPP6C	LOC51189
SMARCE1	PRO0149	HSPC157	C9orf12	ATP6V1G1
TLK1	GLUD1	FLJ10956	DKFZP564L2423	RAB9A
TRAP100	H1F0	PAPSS2	FLJ10509	ETFA
PPARB	ARPC5	SLC9A7	AA019332	MPHOSPH6
ERBB2	BART1	TCN1	LAMA5	LOC55829
GRB7	B4GALT1	MADH4	ARFRP1	P115
ERBB2	COPA	FZD8	ELF3	KIAA0233
DKFZp761B0319	FECH	ALDH3A2	C20orf55	CTSH
TBPL1	TCF12	IGF1R	ENTPD6	HSPBP1
DKFZp564O2364	SLC20A1	LXN	MGC20576	PIR
CEACAM6	F2R	ZIM2	CITED2	MGC2628
S100P	GNS	FADD	RAI17	LOC129642
EPB72	UNG	PPP2CB	HK1	AA598508
FLJ10607	MGC2752	LGALS9	ANK3	KIAA0429
HNF3G	DEGS	BST2	NME4	EFNB2
FLJ39293	FLJ20481	MX1	RANBP6	SLC4A3
GSTT1	ANXA7	G1P3	LIG3	PPIF
DKFZp566O084	FGFR4	BTN3A3	SH3YL1	GPI
PPP1R15A	RAB31	TAP1	PBP	CD24
PTMS	RAB31	MB	LOC91689	MMP15
FLJ34425	FLJ14059	PSMB10	TM9SF2	SLC20A2
N33	SEMA3C	PLSCR1	CLN5	CTL2
COLEC12	PTK9	FKBP11	GYG	PIP5K1A
FJX1	PXMP3	ITM3	POLR2C	JAM1
TGFBR3	STX7	ID4	ATOX1	SLC3A2
RPL27A	INPP4B	IGL@	AP2S1	STX5A
RB1CC1	FLJ25604	IGLL1	HSPC023	FLJ20452
TNC	KIAA1463	CCL18	ATP6V0B	KDELR3
PK428	FZD1	PSPHL	MGC2477	CNN3
KIAA1340	ESTs,	DRIM	ADRM1	STK38
KIAA0193	WNT5A	LIPH	FXD3	DKFZP434L0718
MTMR6	CSDA	H2BFQ	CDH1	DATF1
HBXAP	KIT	H1F2	W86859	SIAT4C
PHCA	FLJ21069	CANX	NQO1	CPA3

Table 6.2: Second 180 intrinsic genes from Stanford

CRABP1	KDELR2	TRIM29	CCK	EPAC
GLDC	KIAA1691	KRT17	AF1Q	ACAS2
SDC2	NRBF-2	MFGE8	MT1G	KIAA0857
GSTP1	CCNE1	ZDHHC5	MT1L	MEN1
MCM3	CALU	CX3CL1	AGXT	PIK3R1
PLOD2	LANP-L	FZD7	FABP7	AKR1C1
PLOD	POLR2F	FLJ11796	ESTs	FACL2
SERPINH2	SQLE	CHI3L2	PEMT	PAM
PTK7	S100A10	ESTs	FLJ31373	FLJ20811
NRG1	T54544	FLJ14525	RAD52	GNB2L1
PREP	LAD1	B3GNT5	MGC3207	RPL32
RFC3	STK24	FLJ11796	FLJ11196	T49282
NEK4	MAFG	SLC5A6	MGST1	Ells1
RPL10	CTPS	FVT1	CDKN2D	FLJ33034
HRIHFB2206	TMSNB	FLJ22678	ACAA2	PTPRM
SDHA	CUL1	ACTG2	MGC29654	FLJ37284
LOC157378	TP53BP2	VCL	BRP44I	LTF
PNAS-131	CBR1	SGCE	CDC42EP4	AQP3
MDS029	ADAM9	MMP14	LOC55862	ABLIM1
HEAB	CDK2AP1	LAMC2	MYH9	KL
ABCD3	PTPRK	CD59	PDGFRA	GSTA4
ADSL	S100A11	CP	LOX	NPAS2
BTG3	KIP2	RCL	COL6A3	KRT13
ATP5G1	KIAA1971	CABC1	OSF-2	KRT13
D123	SR-BP1	FLJ12517	LMO7	TFAP2C
PRNPIP	EXT2	PRAME	PEA15	CA2
EBNA1BP2	FLJ10697	DGUOK	IGFBP5	SSFA2
NSEP1	FLJ31360	HSF2	AA054451	FZD4
GGH	FLJ14761	PDK3	IGFBP5	BC008967
LC27	CXCL1	UBE1	H11	GPX4
LC27	CDH3	PITPN	H11	HRASLS3
PRDX4	SLPI	ICAP-1A	IFRD1	HOXB5
HSPC163	TONDU	BRP17	SENP3	HOXB6
GARS	GABRP	KCNK1	ATP6V1B1	PCSK1
FLJ12442	ANXA8	GALNT3	MAPK3	CHGB
TMSB10	KRT5	SAT	MGC10500	PON3

Table 6.3: Third 144 intrinsic genes from Stanford

GRIA1	DKFZp586J2118	RARRES3	CYP2A7
PRKACB	PRO1489	POLYDOM	AGTR1
IGFBP2	PRO1489	FLJ40165	PLAT
MKNK2	DKFZP586B2420	FLJ11280	NPEPPS
DUSP4	DKFZp761F2014	DKFZp313L231	HIS1
KIAA0222	ENPP5	VAV3	HIS1
DKFZp434C184	ShrmL	LRBA	KIAA0239
FLJ20920	MAP2K4	KIAA1243	FLJ39082
SLI	PTPRN2	LOC51313	RNASE4
EIF4G3	RGS5	DKFZp434E033	FMOD
GRIA2	KIAA1157	DKFZp686F18109	ADRA2A
PRO1331	KIAA0544	KIAA0876	FLJ23160
CYFIP2	CCND1	FLJ10980	TCEA3
ZNF236	FLJ11730	TLE3	DDB1
DKFZP564D0372	GRLF1	NAT1	KIAA0182
CA11	LU	NAT1	SLC7A6
IRS1	SIAH2	LIV-1	KIAA0310
NPY1R	BF	MGC:22588	DUSP16
AKR7A2	FMO5	HNF3A	FLJ38045
QDPR	SCNN1A	XBP1	ZNF75
DXS1283E	STAT6	FLT1	ODF2
MAIL	SELENBP1	GATA3	RAB14
MSX2	MUC1	GATA3	RAB2L
MSX2	MUC1	ESR1	SFRS6
SNK	MUC1	RAB5EP	MGC9042
NRBP	NUMA1	PTP4A2	ACADVL
FLJ20568	FLJ13322	RERG	FLJ12592
BECN1	ABCD3	RERG	CMAR
COX6C	BIRC1	CEGP1	FLJ10948
IGBP1	FLJ40901	ACADSB	FLJ22566
GSTM2	FLJ40901	FBP1	SPTAN1
ZFP36L2	SMA3	MGC27171	DOM3Z
GSTM4	MCCC2	HSD17B4	CXYorf1
GSTM1	ASAH1	FLJ11796	PLXNB1
GSTM3	BLVRA	KIAA1025	76P
DKFZp586J2118	TCEAL1	KIAA0303	MST1

Table 6.4: Last 48 intrinsic genes from Stanford

FLNB	DKFZp586H0623	SLC16A2
ZFX	ESTs	TFAP2B
RALGPS1A	PDCD4	HMGCS2
ECE1	SLC11A2	FLJ39952
DIP13B	CRAT	EPOR
CIT	FER1L3	OS-9
MGST2	FLJ35016	NGFRAP1
NDP	ZNF220	DJ79P11.1
D5S346	LRRFIP1	TRPS1
C4B	FLJ22269	SERPINA1
CYB5	BMP4	S100A1
ALCAM	KIAA1253	NALP2
GTF2H2	ACOX2	TFPI2
ESTs	ESTs,	MRPS14
AMFR	WWP1	APOD
SLC11A3	APOD	SLC11A3

Here are the 66 genes that were found correlated with copy number by Chin et al.

Here are the genes that we found correlated with copy number for one subtype versus the rest in our method.

Table 6.5: 66 Genes found Correlated with Copy Number from Chin et al.

SPFH2	LHX1
PROSC	ACACA
BRF2	DDX52
RAB11FIP1	TBC1D3
ASH2L	SOCS7
LSM1	PGGF2
BAG4	PSMB3
DDHD2	PIP5K2B
WHSC1L1	FLJ20291
FGFR1	PPARBP
TACC1	STARD3
ADAM9	TCAP
GOLGA7	PNMT
SLD5	PERLD1
MYST3	ERBB2
AP3M2	GRB7
IKBKB	GSDML
POLB	PSMD3
VDAC3	NR1D1
SLC20A2	2NF217
THAP1	BCAS1
FNTA	CSTF1
LOC441347	RAE1
CCND1	RNPC1
FGF3	PCK1
FADD	TMEPAI
PPFIA1	RAB22A
CTTN	VAPB
NADSYN1	STX16
KRTAP5-9	NPEPL1
FOLR3	GNAS
NEU3	TH1L
N-PAC	C20orf145

Table 6.6: Genes Found Correlated with Copy Number

Basal	Her2	Luminal A	Luminal B
PPP2R3A	GRB7	HSGT1	UK114
FLJ10618	CDC6	JDD1	YWHAZ
SMARCA5	PSMD3	PPP3CB	LOC51123
APC10	TRAP100	MRPS16	DD5
ABCE1	THRA	ANXA7	MRPL13
KIAA1046			FLJ10204
HSHIN1			BM-009
DKFZP586M0622			MGC3113
MADH1			FLJ13852
LSM6			GPAA1
LRBA			BOP1
PPID			MGC3113
PDZ-GEF1			SCRIB
TMPRSS5			FLJ20591
ZW10			CYC1
FTS			LOC51236
KIAA1005			HSF1
GOT2			VPS28
KIAA1464			FLJ20989
HRIHFB2206			
TERF2			
HSBP1			
MBTPS1			
KIAA1609			
NOC4			

Bibliography

- [1] ACS. What are the key statistics about breast cancer? <http://www.cancer.org/cancer/breastcancer/detailedguide/breast-cancer-key-statistics>. Last Revised: 02/22/2016.
- [2] Affymetrix. Genechip ht human genome u133 array plate set. http://www.affymetrix.com/catalog/131441/AFFY/HT+Human+Genome+U133+Array+Plate+Set#1_1.
- [3] Mikel Aickin and Helen Gensler. Adjusting for multiple testing when reporting research results: the bonferroni vs holm methods. *American journal of public health*, 86(5):726–728, 1996.
- [4] Javier Arsuaga, Tyler Borrmann, Raymond Cavalcante, Georgina Gonzalez, and Catherine Park. Identification of copy number aberrations in breast cancer subtypes using persistence topology. *Microarrays*, 4(3):339–369, 2015.
- [5] BreastCancer.Org. Breast cancer facts and figures 2013-2014. <http://www.cancer.org/acs/groups/content/@research/documents/document/acspc-042725.pdf>. 2013-2014.
- [6] BreastCancer.org. Rate of cell growth. http://www.breastcancer.org/symptoms/diagnosis/rate_grade. Last modified on October 23, 2015 at 10:05 AM.
- [7] Koei Chin, Sandy DeVries, Jane Fridlyand, Paul T Spellman, Ritu Roydasgupta, Wen-Lin Kuo, Anna Lapuk, Richard M Neve, Zuwei Qian, Tom Ryder, et al. Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer cell*, 10(6):529–541, 2006.

- [8] Ashley MD Cimino-Mathews and Gang MD PhD Zheng. Breast cancer and breast pathology. <http://pathology.jhu.edu/breast/biomarker-testing.php>.
- [9] J Climent, JL Garcia, JH Mao, J Arsuaga, and J Perez-Losada. Characterization of breast cancer by array comparative genomic hybridization this paper is one of a selection of papers published in this special issue, entitled 28th international west coast chromatin and chromosome conference, and has undergone the journal's usual peer review process. *Biochemistry and Cell Biology*, 85(4):497–508, 2007.
- [10] cosmic. Catalogue of somatic mutations in cancer. <http://cancer.sanger.ac.uk/census/>.
- [11] Jay Devore. *Probability and Statistics for Engineering and the Sciences*. Cengage Learning, 2015.
- [12] Quest Diagnostics. Ki-67, ihc with interpretation. http://www.questdiagnostics.com/testcenter/testguide.action%3Fdc%3DTS_Ki-67. Content reviewed 12/2012.
- [13] Maria A Doyle, Jason Li, Ken Doig, Andrew Fellowes, and Stephen Q Wong. Studying cancer genomics through next-generation dna sequencing and bioinformatics. *Clinical Bioinformatics*, pages 83–98, 2014.
- [14] eMedMD.com. Genomics an introduction. <http://www.emedmd.com/>.
- [15] GeneCards. Human gene database. <http://www.genecards.org/>.
- [16] UCSC genome browser. Ucsd genome bioinformatics. <https://genome.ucsc.edu/>.
- [17] Reina Haque et. al. Impact of breast cancer subtypes and treatment on survival: An analysis spanning two decades. *Cancer, Epidemiology, Biomarkers and Prevention*, 21:1848–1855, 2012.
- [18] Lissa Harris. The dna microarray. <http://www.the-scientist.com/?articles.view/articleNo/16657/title/The-DNA-Microarray/>. August 29, 2005.
- [19] Trevor J. Hastie. Pam prediction analysis of microarrays users guide and manual. <http://statweb.stanford.edu/~tibs/PAM/pam.pdf>.

- [20] Peter M Haverty, Jane Fridlyand, Li Li, Gad Getz, Rameen Beroukhi, Scott Lohr, Thomas D Wu, Guy Cavet, Zemin Zhang, and John Chant. High-resolution genomic and expression analyses of copy number alterations in breast tumors. *Genes, Chromosomes and Cancer*, 47(6):530–542, 2008.
- [21] Bertha Hidalgo and Melody Goodman. Multivariate or multivariable regression? *American journal of public health*, 103(1):39–40, 2013.
- [22] Hugo M Horlings, Carmen Lai, Dimitry SA Nuyten, Hans Halfwerk, Petra Kristel, Erik van Beers, Simon A Joosse, Christiaan Klijn, Petra M Nederlof, Marcel JT Reinders, et al. Integration of dna copy number alterations and prognostic gene expression signatures in breast cancer patients. *Clinical Cancer Research*, 16(2):651–663, 2010.
- [23] National Breast Cancer Foundation Inc. Breast anatomy. <http://www.nationalbreastcancer.org/breast-anatomy>.
- [24] National Human Genome Research Institute. Dna microarray technology. <https://www.genome.gov/10000533/dna-microarray-technology/>.
- [25] Susan G. Komen. Tamoxifen. <http://ww5.komen.org/BreastCancer/Tamoxifen.html>. Updated 07/31/15.
- [26] Harvey Lodish. *Molecular cell biology*. Macmillan, 2008.
- [27] Morris L Marx and Richard J Larsen. *Introduction to mathematical statistics and its applications*, volume 31. Pearson/Prentice Hall Upper Saddle River, NJ, USA, 2006.
- [28] MedicineNet.com. Definition of proteins. <http://www.medicinenet.com/>. Last Editorial Review: June 14, 2012.
- [29] Rachael Natrajan, Britta Weigelt, Alan Mackay, Felipe C Geyer, Anita Grigoriadis, David SP Tan, Chris Jones, Christopher J Lord, Radost Vatcheva, Socorro M Rodriguez-Pinilla, et al. An integrative genomic and transcriptomic analysis reveals molecular pathways and networks regulated by copy number aberrations in basal-like, her2 and luminal cancers. *Breast cancer research and treatment*, 121(3):575–589, 2010.

- [30] NHGRI. What is a chromosome? <https://www.genome.gov/26524120/chromosomes-fact-sheet/>. Last Updated: June 16, 2015.
- [31] Adam B Olshen, ES Venkatraman, Robert Lucito, and Michael Wigler. Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics*, 5(4):557–572, 2004.
- [32] Charles M Perou, Therese Sørlie, Michael B Eisen, Matt van de Rijn, Stefanie S Jeffrey, Christian A Rees, Jonathan R Pollack, Douglas T Ross, Hilde Johnsen, Lars A Akslen, et al. Molecular portraits of human breast tumours. *Nature*, 406(6797):747–752, 2000.
- [33] Martin Raff, Bruce Alberts, Julian Lewis, Alexander Johnson, and Keith Roberts. Molecular biology of the cell 4th edition. 2002.
- [34] American Cancer Society. Her2 status. <http://www.breastcancer.org/symptoms/diagnosis/her2>. Last modified on October 23, 2015 at 9:53 AM.
- [35] Therese Sørlie, Robert Tibshirani, Joel Parker, Trevor Hastie, JS Marron, Andrew Nobel, Shibing Deng, Hilde Johnsen, Robert Pesich, Stephanie Geisler, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. *PNAS*, 100(14), 2003.
- [36] Laura J Van't Veer, Hongyue Dai, Marc J Van De Vijver, Yudong D He, Augustinus AM Hart, Mao Mao, Hans L Peterse, Karin van der Kooy, Matthew J Marton, Anke T Witteveen, et al. Gene expression profiling predicts clinical outcome of breast cancer. *nature*, 415(6871):530–536, 2002.
- [37] ES Venkatraman and Adam B Olshen. A faster circular binary segmentation algorithm for the analysis of array cgh data. *Bioinformatics*, 23(6):657–663, 2007.
- [38] Anne Vincent-Salomon, Nadège Gruel, Carlo Lucchesi, Gaëtan MacGrogan, Remi Dendale, Brigitte Sigal-Zafrani, Michel Longy, Virginie Raynal, Gaëlle Pierron, Isabelle de Mascarel, et al. Identification of typical medullary breast carcinoma as a genomic sub-group of basal-like carcinomas, a heterogeneous new molecular entity. *Breast Cancer Research*, 9(2):R24, 2007.
- [39] Robert A. Weinberg. *the biology of CANCER*. Garland Science, Taylor and Francis Group, LLC, an informa business, 2007.

- [40] Mike West, Carrie Blanchette, Holly Dressman, Erich Huang, Seiichi Ishida, Rainer Spang, Harry Zuzan, John A Olson, Jeffrey R Marks, and Joseph R Nevins. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Sciences*, 98(20):11462–11467, 2001.
- [41] Hanni Willenbrock and Jane Fridlyand. A comparison study: applying segmentation to array cgh data for downstream analyses. *Bioinformatics*, 21(22):4084–4091, 2005.