MQP-ID-DSA-0901
MQP-ID-CR-0901

Analysis of Microarray Data to Confirm Novel Subtype of Breast Cancer

An Interdisciplinary Major Qualifying Project Report
Submitted to the Faculty of the
WORCESTER POLYTECHNIC INSTITUTE
in partial fulfillment of the requirements for the
Degrees of Bachelor of Science in
Biology/Biotechnology and Computer Science

By:

Philip Davidson

&

Justin Liu

4/29/2010

APPROVED:

Professor Carolina Ruiz Computer Science WPI Project Advisor	Professor David S. Adams Biology and Biotechnology WPI Project Advisor	Professor Karl Simin Department of Cancer Biology University of Massachusetts Medical School

Table of Contents

Abstract

Microarray data of 1,056 breast tumors from NCBI's GEO database were collected and analyzed. The goal was to discover microarrays most similar to a potentially novel subtype of breast cancer characterized by low claudin expression. Claudins form tight junctions, which regulate cell-cell movement of solutes and ions; faulty tight junctions have been correlated with cancer metastasis. Results showed little support for a Claudin-Low cluster. Nine clusters were found, when 5-6 were expected. These additional clusters may reveal novel subtypes.

Acknowledgements

We would like to thank Dr. Karl Simin of the University of Massachusetts Medical School's (UMass med) Cancer Biology department for mentoring this project and Professors Carolina Ruiz and David Adams for advising it. We would also like to thank Konstantin Zeldovich of UMass med's Bioinformatics Department for his aid in setting up the normalization procedures.

1. Introduction

Microarray technology allows the analysis of global gene expression in cell samples of different cell types or exposure to varying treatment conditions. One practical application of this involves the field of cancer research. Herschkowitz et al. (2007) profiled the expression patterns of breast cancers from women and compared them to tumors derived from a variety of mouse breast cancer models. During this process, a new, rare, and potentially dangerous subtype was discovered. This new subtype was characterized by low expression of genes involved in tight junctions and cell-cell adhesion (Herschkowitz, et al., 2007). The subtype was referred to as the "Claudin-Low" subtype since several *Claudin* family members segregated among this group of low expressing genes. Claudin proteins form intercellular tight junctions between cells, which control the paracellular flow of molecules. Intracellularly, tight junctions are the docking site for signaling molecules, such as growth factor receptors, as well as actin fibers that are critical for the cytoskeleton. Defective tight junctions can cause unnatural cell proliferation and may increase the metastatic potential of tumor cells, which are hallmarks of malignant forms of cancer. Reduced intercellular adhesion may permit carcinoma cells to migrate in other areas of the body, which makes them especially dangerous, since tumor metastasis is associated with the high morbidity of cancer.

However, the Claudin-Low subtype was identified in only about 6% of microarrays in the data used in the Herschkowitz (2007) paper.

In order to confirm that tumors exhibiting the Claudin-Low subtype exist outside of the original data set considered, and to potentially discover more tumors that demonstrate the Claudin-Low subtype, a larger data set was analyzed using a similar approach. Microarray data of 1,056 breast tumors from NCBI's Gene Expression Omnibus database were downloaded and analyzed using various data mining techniques. The discovery of multiple microarrays exhibiting similar gene expression profiles to Claudin-Low tumors would confirm the existence of this subtype, and a gene classifier could be developed to identify tumors of this type. The primary tool used for this analysis was hierarchical clustering supplemented with various filters and information enhancement algorithms. Additionally, once a Claudin-Low-like expression pattern was discovered, different distance metrics were used to compare the microarrays, to this expression pattern, in order to find the most similar tumor samples. Hierarchical clustering using a simple threshold of gene variation as a filter showed little support for a distinctly Claudin-Low cluster in the larger cohort of microarrays. This is not conclusive evidence that Claudin-Low tumors are nonexistent, and several factors may explain why a distinct cluster of Claudin-Low tumors was not identified in our current analysis. One reason may be an ascertainment bias. Since Claudin-low tumors can show a distinct metaplastic morphology when compared to other tumors of the breast (Hennessy, et al., 2009), Claudin-Low tumors may have been inadvertently excluded from studies focused on analyzing breast adenocarcinomas. Second, because the microarrays were performed on bulk tumor cells, it is possible that Claudin-Low cells make up a fraction of cells in a heterogeneous sample. If they contribute to only a minor proportion of cells in the tumor, then their expression signature would be swamped by the expression of the majority of tumor cells. Claudin-Low cells show similar expression profiles to cells with enhanced "tumorigenicity" (Al-Hajj et al., 2003), which may be indicative of cancer stem cell phenotype (Hennessy, et al., 2009) and chemotherapy resistance (Li, et al.,

2008). So, although the results do not support a distinct subtype, understanding the biology of cells with the Claudin-Low phenotype may still have broad implications for more effective treatment of aggressive breast cancers. While a definitive Claudin-Low subtype was not extracted from the data, nine distinct clusters were found, when five were expected. These results may help to further stratify previously identified subtypes into subcategories with distinct biology and clinical outcomes.

2. Background

2.1.Gene Expression

The central dogma of molecular biology illustrates the flow of biological information between the three major bio-molecules: DNA, RNA, and protein. Generally, DNA is used to create RNA, which is then used to create proteins; this is shown in Figure 1.

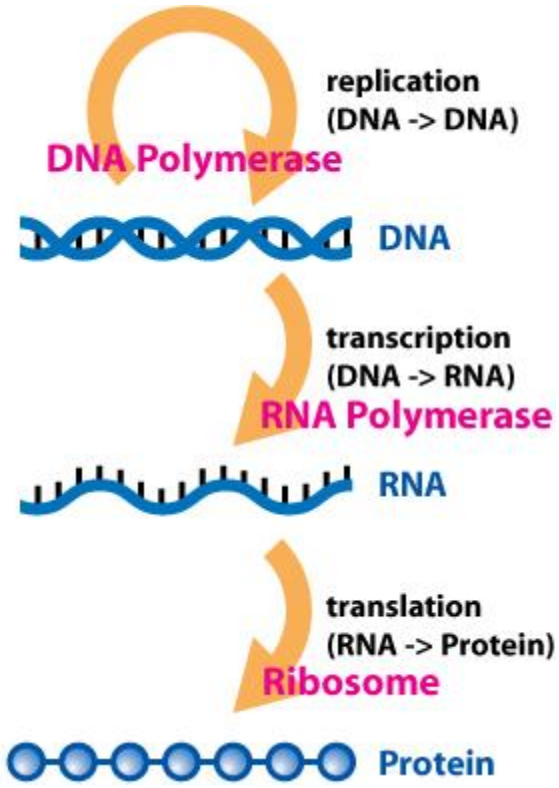

Figure 1 – The Central Dogma of Molecular Biology. Diagram shows the normal flow of genetic information from DNA to RNA to protein. Profiles of RNA types expressed in tumor cells were analyzed in this project. (Horspool)

In order to convert from one bio-molecule to another, a complex process is required. The production of RNA from DNA is known as transcription, and the production of protein from DNA is known as translation. The necessity of these conversions is due to the different roles that each bio-molecule plays as a result of its structure.

2.1.1. Bio-molecules Overview

DNA is made up of two anti-parallel strands (same information, opposite directions), arranged in the shape of a double helix and connected through hydrogen bonds. Each strand consists of two ends: the 5’ end and the 3’ end; the number of each end corresponds to the number of the terminal carbons on the nucleotide. Each strand is also comprised of a sequence of four nucleotides: adenine (A), guanine (G), cytosine (C), and thymine (T). These nucleotides have specific pairing behaviors amongst each other:

adenine will bond with thymine, and guanine will bond with cytosine; this is referred to as complementary base pairing.

These properties of DNA allow it to be a suitable medium for long-term storage of genetic information. Being double stranded provides many benefits: the molecule is more stable against physical forces and chemical degradation, replication can occur easily and accurately with either strand as a template (due to complementary base pairing), and transcription can be carried out faster since either strand of the DNA can be used as a template (Freeman, 2005).

RNA is a nucleic acid like DNA, but there are three main differences. First, RNA is usually single stranded and shorter than DNA. This means that RNA is able to adopt structures other than the double helix found in DNA, allowing it to act as a functional molecule. Secondly, RNA contains a ribose sugar (instead of deoxyribose) as the backbone. This causes RNA to be much more reactive than DNA, because the attached group is a hydroxyl (-OH) instead of the more stable hydrogen (Freeman, 2005). Also, RNA contains no thymine. This is controlled during transcription, discussed in section 2.1.2. With respect to this project, not all the information residing in DNA is transcribed into RNA, so the profile of RNA molecules varies from cell to cell, and between cells at different stages of tumor formation.

Proteins are the main functional molecules in cells. They perform diverse functions, including defense, movement, catalysis, signaling, structure, and transport (Freeman, 2005). All these various functions are made possible by the numerous conformations that a protein can form. Protein structure can be described in four levels of complexity: primary structure, secondary structure, tertiary structure, and quaternary structure. The primary structure is the linear sequence of the amino acids forming the proteins (instead of nucleotides, amino acids are the building blocks of proteins). The secondary structure is the arrangement or folding of the peptide's structure due to local interactions of amino acids; the three possible conformations are helices, sheets, and turns. The tertiary structure is the three-dimensional structure of the polypeptide (including interactions amongst multiple peptides). Finally, quaternary structure describes the arrangement and interaction of multiple polypeptides; it is possible for proteins to be made up of multiple polypeptide chains (Voet et al., 2008).

2.1.2. Transcription

In the process of transcription a DNA sequence is used as a template to output a functional mRNA. The process occurs from the 3' to 5' direction on the template DNA strand so that the RNA chain can grow in the 5' to 3' direction; this occurs in the nucleus of the cell. In eukaryotes, RNA Polymerase II binds to General Transcription Factors (GTF), which bind to the promoter region (usually downstream of a TATA chain of nucleotides). After this binding, RNA Polymerase II then continues down the 5' direction of the template strand, building the RNA on the 3' end. RNA Polymerase II continues to travel down the DNA until it reaches the end. It is important to note that the template strand of DNA being used to create the transcript RNA will have sequences that are complementary to the transcript RNA formed (Griffiths et al., 2008).

The RNA transcript is significantly different from the DNA sequence it is built from. When an A is encountered in the template strand, rather than pairing it with a T in the new mRNA, a different amino

acid, uracil (U), is used. Aside from replacing all T's with U's, transcription also goes through numerous processing steps that modify the transcript RNA so that it is not just a DNA clone. These processing events occur as RNA Polymerase II travels along the template DNA. One modification is the addition of a cap at the 5' end of the transcript. The cap prevents RNA degradation and is used to initiate translation. Another modification is the terminating RNA transcription. This occurs when a sequence that contains the nucleotides AAUAAA or AUUAAA shows up on the 3' end of the transcript RNA. An enzyme recognizes this site and proceeds to cut it. Finally, the last two processing events go hand-in-hand: splicing and removal of introns. In eukaryotic DNA, there are sequences known as introns and exons. Exons are regions containing nucleotides that code for amino acids while introns are regions containing nucleotides that are not translated. Therefore, during transcription, all the introns are usually removed, and exons are aligned next to each other. Additionally, it is possible for the arrangement of exons to differ. For example, if a gene has three exons, it is possible that one transcript RNA will contain exons 1 and 2 while another transcript may contain only exons 1 and 3. This is known as alternative splicing and is what allows a single gene to code for multiple proteins (Griffiths et al., 2008).

2.1.3. Translation

Unlike DNA replication and transcription, translation travels down its template strand (an mRNA) in the 5' to 3' direction and occurs in the cytoplasm (except for the translation of secretory proteins, which occurs within the endoplasmic reticulum (ER)). The key players of translation are the ribosome and tRNA. Ribosomes attach to the cap of the mRNA and proceed down the mRNA until it finds a nucleotide start sequence (AUG). When it finds this sequence, a tRNA containing the complementary sequence will attach to the ribosome and drop off a methionine molecule. This will proceed with subsequent triplets of nucleotides until a stop sequence is reached on the mRNA (UAA, UAG, UGA). Once this stop sequence is reached, the ribosome is released from the mRNA in order to find another mRNA to translate. Figure 2 shows which amino acid is coded by each nucleotide triplet (Griffiths et al., 2008).

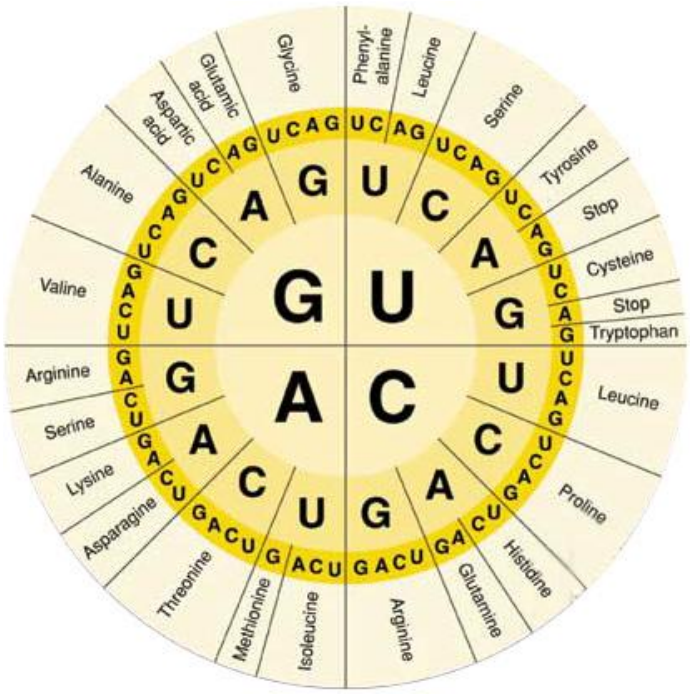

Figure 2 – Graphic Illustrating the Universal Genetic Code. The table shows a list of which amino acid is encoded by each nucleotide triplet. (Cooper & Hausman, 2007)

2.1.4. Expression Regulation

In eukaryotes, there are many methods to regulate gene expression. With respect to this project, the main category of gene regulation that we are concerned with is transcriptional regulation. Normally, eukaryotic DNA does not undergo transcription. This is because most genes are packaged in a tightly wound structure known as chromatin. The DNA is wound tightly around histone octamers in structures known as nucleosomes. In order to unwind the chromatin, complexes add or remove molecules at the histone tails protruding from the nucleosomes. Another transcription regulation tool is the enhancer sequences. These regions allow multiple transcription factors to bind, forming enhanceosomes, which are large protein complexes that activate transcription mainly through reshaping the DNA so that RNA Polymerase II can bind and travel along it. Analyzing the levels of mRNA in a cell allows one to estimate what genes are actually being expressed (translated into protein), and infer the biology of the cell.

2.2.Epithelial Cell Development

An example of epithelial cell development can be found in skin cells. In skin, pluripotent epidermal stem cells are situated in the bulb of hair follicles. Above this layer are more specialized stem cells; these cells are committed to becoming epidermal cells and are found in the basal layer of the skin. The next layer is the spinosum layer, and this houses the transit-amplifying cells. Above the spinosum is the granular layer, which contains numerous cytokeratin-filled granules. Accumulation of cytokeratin by skin cells causes the skin cells to lose their structure, forming the outer layer of the corneum (top-most layer of skin) (Majumder, 2009, p. 4).

Early studies by DeOme (1959) showed that a subset of mammary cells can act as stem cells to repopulate the mammary gland. Although the hierarchy of cells in the mammary gland, or the epithelial structures that comprise the other organs of the body, is less clearly defined than skin, the development of these diverse tissues likely involve analogous mechanisms.

2.3.Cancer Biology

Cancer is characterized by uncontrolled cell growth. This uncontrolled cell growth is generally the result of DNA mutations that affect genes regulating the cell cycle. Some mutations that lead to this phenotype are mutations that cause cells to: ignore apoptosis signals, self-produce growth factors, ignore growth inhibition signals, and fail to differentiate. Although some viruses are able to cause tumors by mutating human DNA, the majority of tumors are caused by radiation and chemical carcinogens (Cooper & Hausman, 2007).

The specific mechanics behind tumor behaviors are still being elucidated today. At the moment, there are two major competing models for the development of cancer: the stem cell theory of cancer and the de-differentiation theory of cancer (Majumder, 2009, pp. 7-8).

2.3.1. Stem Cell Theory of Cancer

Cancer is a problem of developmental biology (Majumder, 2009, pp. 2-3). In normal development, the two main phases are expansion and determination (Majumder, 2009, p. 3). After fertilization of the egg, the initial stem cell will divide symmetrically for the first five to six divisions (Majumder, 2009, p. 3). The daughter cells produced during these divisions are completely totipotent, thus during this expansion phase, the number of embryonic stem cells expands (Majumder, 2009, p. 3). Once this expansion is completed, the cells move into the determination phase; during this phase, divisions occur asymmetrically. As a result of the asymmetrical divisions, one of the daughter cells remains a stem cell, while the other daughter cell begins to travel down the path of differentiation, slowly losing its stem cell characteristics (Majumder, 2009, p. 4). The stem cell theory of cancer states that the cause of cancer lies in a failure to properly complete such a progression (Majumder, 2009, p. 4), leaving immortalized cells with stem-like properties as a seed for the tumor.

Although the cells in cancer tissue are mostly undifferentiated and do not mature properly, the same types of cells are observed in both normal and cancer tissues: stem cells, transit-amplifying cells, and terminally differentiated cells (Majumder, 2009, p. 4). The difference between the two tissues is the *quantity* of the different cell types; more specifically, the quantity of transit-amplifying cells (Majumder, 2009, p. 4). In normal cells, the transit-amplifying cells divide a finite number of times before becoming terminally differentiated cells, which are unable to divide further. However, in cancer cells, the transit-amplifying cells divide without approaching a terminally differentiated state, thus resulting in unnatural accumulation of this cell type (Majumder, 2009, p. 4). The physical product of this can actually be seen (a tumor), but it does not necessarily explain how metastasis occurs. Cancer stem cells are this theory's solution to the metastasis issue (Majumder, 2009, pp. 13-15).

Although recent findings in cancers of various tissues show some support for the stem cell theory of cancer, the strongest support for this theory come from teratocarcinomas (cancers in germ

tissue) (Majumder, 2009, p. 9). Since germ cells can differentiate into both the embryo, and yolk sac and placenta, they are true stem cells (fully totipotent). In 1964, Leoy Stevens transplanted normal testicular germ cells into unnatural locations in a mouse model; teratocarcinomas appeared at these sites, suggesting that normal stem cells can transform into cancer cells if removed from their native environment (Majumder, 2009, pp. 9-10). Another experiment that showed strong support for this theory was conducted in skin cells. A reliable way to induce cancer in the skin involves a 2-step process of initiation and promotion (Majumder, 2009, pp. 12-13). In initiation, the skin is exposed to Benz(o)pyrene, which binds to DNA in skin cells, permanently altering the genes. Cancer will not rise until these cells are treated with phorbol ester, which promotes the damaged cells to proliferate. In this specific experiment, the time elapsed between initiation and promotion was varied, and the results were surprising. Regardless of the time elapsed, cancer will always result after the promotion procedures (assuming initiation occurred). Since skin cells shed every 2 – 3 weeks in mice, the initiation must have affected the stem cells responsible for replacing the skin cells (Majumder, 2009, pp. 12-13).

2.3.2. De-differentiation Theory of Cancer

The de-differentiation theory of cancer was the standard model for cancer formation until the stem cell theory recently re-immerged. This theory states that cancers occur when mature cells return to a less differentiated form as a result of external stimuli like chemicals, parasites, and viruses (Majumder, 2009, p. 8). Current clinical prognosis of cancers is determined by assigning a tumor grade by specially trained surgical pathologists. Tumor grade is largely based on the level of differentiation of a tumor, indicated by the morphology of the cancer cells. Nevertheless, the molecular basis for tumor differentiation is largely undefined. Intriguingly, the Weinberg lab (Mani, et al., 2008) described a series of experiments that may help to unify the stem cell and de-differentiation models. Through an epigenetic process called epithelial to mesenchymal transition (EMT), a normal cellular transformation that occurs during development and wound healing where epithelial cells acquire stem call characteristics. This process is governed by a series of transcriptional repressor proteins, including Snail, Slug, and Twist, which were first described in *Drosophila* developmental pathways. Genes encoding these proteins show elevated transcript levels among Claudin-Low tumors (Herschkowitz, et al., 2007). More importantly, the targets of these transcriptional repressors include the low expressing genes that co-segregate with the *Claudin* genes in Claudin-Low tumors, leading us and others (Hennessy, et al., 2009) to speculate that Claudin-Low tumor cells represent mammary epithelial precursor cells or cells that have undergone an EMT.

2.3.3. Breast Cancer

Breast cancer starts in the tissues of the breast and can be classified in numerous ways. Current clinical practice relies heavily on the expression of several growth factor receptors, including the estrogen and progesterone receptors, and HER2/ERBB2. The expression of these proteins (or their absence), along with pathological evaluation, and whether a tumor has metastasized locally or distantly, will all be considered by a physician when deciding a therapeutic regimen to recommend to a patient. An alternate way to classify breast cancers that is emerging is to categorize them into subtypes based on their gene expression profiles. The type of breast cancer this project is focused on, Claudin-Low, would be described as a “Triple Negative” (none of the three receptor proteins is expressed) by standard

clinical criteria (Herschkowitz, et al., 2007). Triple Negative tumors generally have a poor clinical outcome, and no specific targeted therapies exist, so these patients are typically treated with harsh chemotherapies. The current subtypes of breast cancer that have been confirmed by gene expression signatures are Luminal A, Luminal B, Normal Breast-Like, ERB2-Overexpressing, and Basal Epithelial-Like (Majumder, 2009, p. 93); these subtypes vary in terms of prognoses for metastasis-free or overall survival (Majumder, 2009, p. 93).

2.3.4. Tight Junctions

Tight junctions (Figure 3) regulate the movement of solutes and ions between adjacent cells through passive, electro-osmotic gradient-driven transport (Gonzalez-Mariscal, 2006, p. 33). The gradients are typically a side-effect of active transports and external exposures like food consumption (Gonzalez-Mariscal, 2006). Three attributes that are variable in tight junctions are electrical conductance, size, and charge selectivity (Gonzalez-Mariscal, 2006, p. 34). Conductance is the most variable and is classified in five orders of magnitude from leaky to tight; leaky epithelium is characterized by having at least half of their total conductance occurring through tight junctions (Gonzalez-Mariscal, 2006, p. 34). The size of tight junctions vary, junctions of epithelial cells generally range from 7 – 15 Å (Gonzalez-Mariscal, 2006, p. 34). Lastly, selectivity is defined as the permeability ratio of sodium to chloride; most epithelia are cation-selective, meaning that the sodium permeability is higher than the chloride's (Gonzalez-Mariscal, 2006, p. 34).

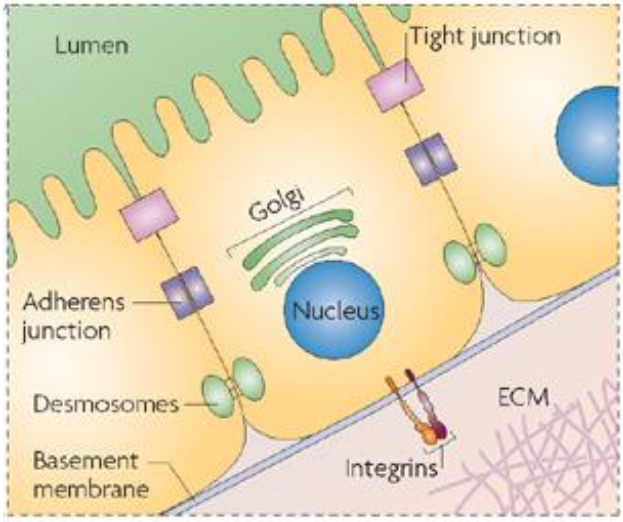

Figure 3 – Conceptual Illustration of Tight Junctions. (Bryant & Mostov, 2008)

The two kinds of proteins that form or interact with tight junctions are cytoplasmic proteins and membrane proteins (Gonzalez-Mariscal, 2006, p. 118). The cytoplasmic proteins are further divided into two groups: PDZ domain-containing adaptor proteins and non-PDZ domain-containing adaptor and signaling proteins (Gonzalez-Mariscal, 2006, p. 118). Claudins, JAMs, occludins, and Crumbs constitute the membrane proteins; occludins and claudins are linked to the regulation of epithelia proliferation, while JAMs and Crumbs interact with signaling complexes controlling epithelial polarization (Gonzalez-

Mariscal, 2006, p. 110). Furthermore, the intertwining of some of these membrane proteins between adjacent cells form physical barriers that further regulate the movement of solutes and ions between cells (Gonzalez-Mariscal, 2006, p. 34). Since tight junctions are not very adhesive, they work in conjunction with the adherens junctions and desmosomes, forming junctional complexes between cells (Cooper & Hausman, 2007) (Figure 3).

2.3.5. Claudin

Contrary to the behavior of occludins and JAMs, claudin expression is required for the formation of tight junctions in cells (Gonzalez-Mariscal, 2006, p. 117). The claudin family of proteins consists of at least 24 members in humans and mice, with each member varying in frequency and localization throughout the body (Gonzalez-Mariscal, 2006, p. 37). These proteins cross the cell membrane four times (Figure 4) and range in size from 20 – 25kDa (Gonzalez-Mariscal, 2006, p. 34). The most distinguishing characteristics of claudins are the presence of a WGLWCC motif of conserved amino acids located in the first loops away from the cytoplasm, and termination of the proteins in PDZ binding motifs, which bind PDZ domains in cytoplasmic proteins like ZO-1 (Gonzalez-Mariscal, 2006, p. 34). Proteins that are most homologous to claudins are the membrane proteins of the eye lens (MP2)), epithelial membrane proteins (EMP), and peripheral myelin proteins (PMP) (Gonzalez-Mariscal, 2006, p. 36).

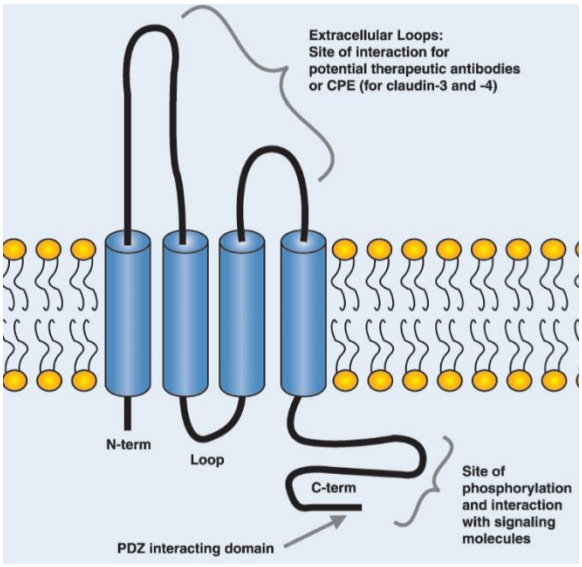

Figure 4 – Diagram of Claudin Structure. Figure shows Claudins as transmembrane proteins that pass the membrane four times. (Morin, 2005).

The primary cause for the variances in tight junction permeability has been attributed to varying concentrations of the different claudins in the tight junctions (Gonzalez-Mariscal, 2006, p. 37). In one experiment, where claudin-1 expression was promoted in low resistance MDCK cells, the cells exhibited higher cell-cell adhesion, which suggests increased electrical resistance across the epithelium (Gonzalez-Mariscal, 2006, p. 37). In another experiment, claudin-2 expression was promoted in high resistance

MDCK cells, but the cells turned out leakier than before (Gonzalez-Mariscal, 2006, p. 37). When claudin-5 was knocked out from mice, the brain tissues between the wild type and mutant mice exhibited drastic differences. When molecular markers were passed through the tissues, the wild type tissues retained markers from 68kDa to 562Da while the mutant tissues became leaky up to 562Da but restricted the next highest marker (1862Da) (Gonzalez-Mariscal, 2006, p. 39). Despite their conflicting behaviors, claudins are still vital in the proper regulation of tight junctions.

2.3.6. Claudins, Tight Junctions, and Cancer

Studies have shown a correlation between lowered expression levels of tight junction proteins and increased rates of cancers; this suggests that tight junction proteins may be involved in controlling cell growth and differentiation (Gonzalez-Mariscal, 2006, p. 96). Additionally, the disruption of tight junctions is characteristic of malignant and metastatic cancers (Gonzalez-Mariscal, 2006, p. 126). Further evidence linking tight junctions and cancers can be seen in the microvessls, small blood vessels that branch out from the main arteries, which feed tumors; these vessels often have reduced occluding and claudin expression (Gonzalez-Mariscal, 2006, p. 120).

2.4.Microarray Technology

mRNA quantities in cells are a good indicator of gene expression, since mRNA is used as a template for the creation of proteins. Microarrays are a method to measure gene expression through mRNA quantities. A microarray consists of thousands of cDNA sequences hybridized to different wells on a glass slide or a biochip. Copy DNA (cDNA), are DNA sequences that have been derived from mRNA and therefore do not contain non-coding introns (although non-coding un-translated regions UTRs present in mRNAs may be represented on the chip). The construction of cDNA allows mRNA of opposite sense to attach to it. A biochip is a non-porous material containing miniaturized wells. Microarrays that use the biochip approach can be much smaller and can be created using several technologies not usable otherwise, including ink jetting and photolithography. Using these advanced fabrication techniques biochips are also able to have smaller wells, closer together. The number of wells per area is termed well density. Biochips can be built to have high enough well density such that the entire human genome can be represented in one chip (Freeman, 2005)

The following sample high-level procedure illustrates the theory behind the construction of microarrays and how they are used to quantify mRNA levels. First, a sample of the target cell's total cellular mRNA is procured. This mRNA sample is labeled with a fluorescent marker. Another mRNA sample, from a population that the first will be compared to, is also taken and labeled with a different fluorescent color. In practice, the target mRNA sample is labeled with a red color fluorescence and the second mRNA sample with a green. Both of the mRNA samples are then hybridized to the biochip at the same time. After the hybridization process is finished and the excess (non-hybridized) mRNA is cleared from the biochip it can be scanned and data collected. This is called a "two color" microarray analysis. Another approach is to only hybridize one each of the two samples to two different biochips and only measure the intensity of the fluorescence; this is termed a "one color" microarray analysis. It is common to have a reference in a single color analysis, usually a sample taken from a common cell line, which is run under the same conditions, as the other samples being observed and is used to normalize the data between different runs. For example, if under one condition genes of the reference are at one level and

under another condition the same genes of the reference are seen to have a lower intensity the data from the second run should be normalized to the first runs reference numbers.

After hybridization in a two color analysis, thereof ratio of the two labeled samples will reflect the abundance of transcript in a sample. A red color, or the color of the target mRNA, denotes an upregulation of the gene corresponding to the cDNA in that well under the conditions of the target sample as compared to the other sample. A green color, or the color of the second mRNA, denotes a downregulation of the gene corresponding to the cDNA in that well under the conditions of the target sample as compared to the other sample. A yellow color, or the combination of the two colors, denotes that both colored mRNAs hybridized to the same spot on the array, so the expression of the gene in both conditions is the same. A well with a black color after hybridization implies that neither sample's mRNA hybridized to it, meaning the gene is not being expressed at the time of sample collection in either sample. These shades are measured using sensor technologies that will convert the fluorescence under each light frequency to a number. These numbers are combined to produce the raw data from a microarray. The data from a microarray has indirectly been shown to have a correlation with the amount of protein that is produced. For example, DeRisi et al. (1997) showed that microarray data collected from yeast corroborates up or down regulation of gene expression expectations under different conditions (Voet et al., 2008).

All microarray data used in this project was gathered using Affymetrix microarrays. The Affymetrix microarray is a single color microarray on a biochip created using photolithography. The cDNAs used on these biochips are 25-mer oligonucleotides arranged in perfect sense/mismatch pairs (Freeman, 2005). The difference between the perfect sense and mismatch genes are that the mismatch are the same as the perfect sense except for the 13^{th} nucleotide. However, it has been shown that the information gained from the mismatch pairs is marginal (Bolstad et al., 2003). The Affymetrix microarray is measured using scanning confocal microscopes. All procedures involving Affymetrix microarrays have been automated, allowing for much more reproducible results that can be obtained more efficiently. Further, since the process has been automated, the there is less need for normalization, which may skew the data, especially for low signal levels.

2.5.Computational Biology

Due to the large amount of genetic and gene expression information inherent in many large studies conducted by biologists and others interested in life sciences it is becoming more common to analyze those datasets using techniques from computer science. Data mining techniques can be used on data from biological experiments to detect patterns, expose similarities, and make predictions. As researchers have realized the potential of this sort of analysis on data, tools to reduce time and effort of this analysis have been created. Further, as computer analysis of biological data has become more prevalent, tools have been developed especially for this domain.

In any area that data mining is applied, domain knowledge can promote better results through attribute selection and chosen form of analysis. Biology is no exception to this rule as several reviewed papers have shown machine learning techniques being applied to biological datasets. Hu et al. (2006) demonstrated that there are 306 intrinsic genes that affect breast cancer through attribute selection

techniques. Herschkowitz et al. (2007) showed that clustering, a form of computing similarities, can be used to analyze microarray data. Finally, Reyal et al. (2008) showed that classification techniques can be applied to gene expression information from microarrays.

2.5.1. Tools in Computational Biology

The main tools used for this project are RMAExpress, Cluster 3.0, Java TreeView, and dChip. Specific details about these tools can be found in the methods section.

2.5.2. Highlighted Techniques in Data Mining

Any task in data mining can be generalized to the following flow: data storage/retrieval, data pre-processing, pattern discovery, pattern evaluation, and pattern usage. Many techniques and approaches can be employed in conjunction with each of these steps. Data pre-processing refers to any technique that prepares the data to make pattern recognition more manageable and produce better results without losing any of the information present in the original data. Pattern discovery techniques attempt to find patterns, given sets of data.

Normalization is a technique that can be used to put numbers of an attribute onto a different scale. Examples of normalization include log normalization, where the logarithmic value, base two, is taken. Another, more advanced use of normalization would be to compare the reference gene values between two single color microarrays to obtain a normalization factor which can then be applied to the results, making them directly comparable. Although normalization does not change the intrinsic behavior of the data, it facilitates analysis of the data.

In data sets where data for each attribute is not necessarily available it may be required that techniques to determine a value for missing attribute is applied. In any case, the best solution to issues of this nature is to attempt to gather the actual value for that instance in the data set. If this is not possible, the mode or mean of values in that attribute from other instances may be the best to give to missing values. Other options include assigning values based on other attributes, which may require domain experience, or to assign a set value to the missing attribute. However, in some cases this may bias the data and another approach should be investigated. Missing values can be treated as a separate value in some forms of pattern discovery.

Attribute selection can be used to reduce the number of attributes that are present in a data set. The attributes that are selected should be the attributes that have the most effect on the outcome that is trying to be predicted. Generally, this means that attributes without much variation can be released from the data set as they do not help to classify an outcome and are not characteristic to any instance. Correlation based feature selection is one method that can be used to select only attributes which give the most information about the outcome. This method works by performing a heuristic search over subsets of attributes. The search gives the highest merit to a set of attributes which predict the target attribute the best but are also the most independent from each other (Witten 2005).

Classification algorithms are used to determine which sets of attribute values predict which outcomes. This is a common approach in any data set where some outcome is recorded for each instance. The construction of decisions trees is an approach that attempts to determine the outcome

based on splitting the data by attribute value. Another technique is to mine association rules which find within the data that some set of attribute values is associated with another set of values. If rules are selected such that the set of values is associated with an outcome they can become classification rules, or rules that predict outcomes. Classification rules can also come from domain knowledge and be applied to a dataset to verify the class of an instance.

Clustering algorithms can be used to measure similarity between instances. There are several methods used in clustering analysis. One of these methods is the k-means clustering algorithm which assigns some number of initial instances as initial cluster centers and then all of the most similar instances are assigned to that cluster. The mean values of each of those instances are calculated and that becomes the new cluster centroid, with the most similar instances being assigned to their new closest cluster centroid. This continues until the calculated mean value for each cluster is equal to the current point of the respective cluster centroid (Witten & Frank, 2005). Another method, termed incremental clustering, takes as input each instance and uses category utility to decide if it is more useful to make that input its own cluster, add it to another cluster or split another cluster into multiple clusters (Witten 2005).

3. Materials and Methods

3.1.Materials

3.1.1. Data

The data used for this project is the same data used in a recent meta-analysis of breast cancer expression data by Reyal et al. (2008). This data set consists of 1,056 microarrays of samples from the tumors of breast cancer patients. Each microarray consisted of fluorescence readings over 22,283 probe sets. A probe set is made up of various nucleotide sequences that are designed to attach to complementary sequences that correlate to known DNA sequences (GeneChip Mouse Expression Set 430). The data is freely obtainable through NCBI's Gene Expression Omnibus (GEO) data set links found through citations in Reyal et al. (2008).

- Links for Data Set Download
  - ftp://ftp.ncbi.nih.gov/pub/geo/DATA/supplementary/series/GSE7390/
  - ftp://ftp.ncbi.nih.gov/pub/geo/DATA/supplementary/series/GSE6532/
  - ftp://ftp.ncbi.nih.gov/pub/geo/DATA/supplementary/series/GSE3494/
  - ftp://ftp.ncbi.nih.gov/pub/geo/DATA/supplementary/series/GSE2603/
  - ftp://ftp.ncbi.nih.gov/pub/geo/DATA/supplementary/series/GSE1456/

The normalization software can only accept one chip type: HG-U133A. A program was written in Java by the authors to sort the *.CEL files into different folders based on chip type, as described by the DatHeader field in the .CEL file. Files with other types included HG-U133B and HG-U133AAofAv2, and were not used in this project.

3.1.2. Software

RMAExpress

RMAExpress (http://rmaexpress.bmbolstad.com/) is an open-source, freely distributed library written by Benjamin Boldstad who is an author on a paper outlining the Robust Multi-chip Average procedure (Bolstad et al., 2003). This software implements the RMA normalization procedure, a three step process including background adjustment, quantile normalization and summarization. Use of the program consists of three steps. First, the user loads a *.CDF file describing the chip type. Then the user loads all of the *.CEL files that need to be normalized (*.CEL). Finally, the user selects Compute RMA Measures and the program will perform the normalization. The output of this process is a *.TXT file containing the normalized expression values, which are also $\log_2$ transformed. It is organized so that the genes form the rows and the microarrays form the columns.

Cluster 3.0

Cluster 3.0 (http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/software.htm) builds off of the original Cluster application written by Michael Eisen of Stanford University. It is capable of running various clustering algorithms on 2-dimensional data sets; these algorithms include hierarchical, k-means, self-organizing maps, and principal component analysis (Hoon, 2002). However, only the hierarchical clustering functionality was used. Like RMAExpress, this program is open source.

Java TreeView

Java TreeView (http://jtreeview.sourceforge.net/) is an application developed by Alok Saldanha for the visualization of gene expression data. It chromatically displays the expression levels of all the genes for all the microarrays in a single block, and it allows the user to zoom in on a specific gene for more detailed observation. As with the previous programs, this is released under the GNU General Public License (Saldanha, 2003).

dChip

dChip (http://biosun1.harvard.edu/complab/dchip/) is a software package developed for handling Affymetrix microarrays. It is able to display and normalize *.CEL files, and performs more advanced functions like sample comparisons and hierarchical clustering (dChip: Introduction and Installation, 2007). In this project, dChip was used to extract lists of genes that are involved in specific functional groups for distance calculations.

Weka

Weka (http://www.cs.waikato.ac.nz/ml/weka/) is a tool that features various algorithms for data mining. These algorithms are involved in procedures like preprocessing, classification, regression, clustering, association rules, and visualization (Witten & Frank, 2005). For this project, Weka was used for calculating distance metrics.

3.2.Methods

The flow chart of the process followed in this project is depicted below in Figure 5. The following subsections describe each step in detail.

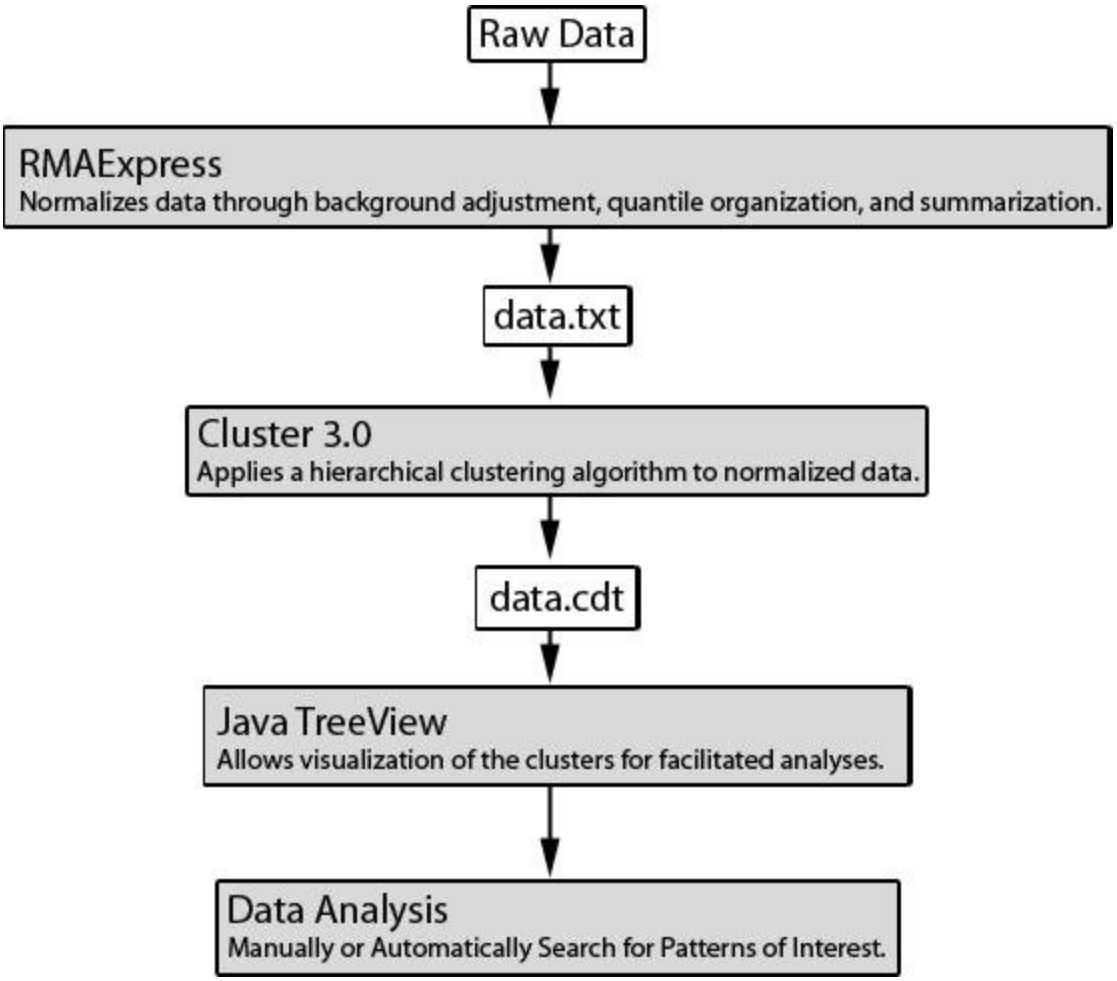

Figure 5 – Flow Chart of Tools Used in this Project.

3.2.1. Data Preprocessing

Normalization is necessary when comparing the results of different microarray chips, because it removes variation caused by the manufacture, preparation or experimental handling of the chip. That variation can obscure the interesting variations, which may exhibit important differences in the same genes between different microarrays. If the obscuring variations are not removed, improper results may be collected (Irizarry, et al., 2002). The procedure used to normalize Affymetrix chips is Robust Multichip Average (RMA) as implemented by RMAExpress. When the RMA'd data was written to a file the values were also converted to a $\log_2$ scale. This was done to accommodate algorithms that produce better results when applied to data that has been converted to a $\log_2$ scale.

RMAExpress was used with the following settings:

- Input: Raw Data (*.CDF, *.CEL files)
- Settings:
  - Min Fold-Change < 20%
  - Percent Missing > 50%
- Output: data.txt – normalized data

3.2.2. Clustering & Viewing the Clusters

Next, the normalized data was filtered and clustered. The filtering process used was implemented by Cluster 3.0. In Cluster 3.0, the only field modified for this project was the SW (Gene Vector) field. This filters genes based on how variable they are over the entire given data set, accepting genes with variation higher than the value set by the filter. An SW value of 0.5 will filter out fewer genes than with a value of 1.5. In this project, we initially used an SW value of 1.0. For latter experiments, we varied this value to observe changes in clustering results. Eventually, we decided to use 0.5 in order to work with a larger set of genes.

The clustering method used was an agglomerative hierarchical clustering. Clustering is a technique used in data mining to find groups of similar elements and measure their similarity. Hierarchical clustering begins with an arbitrary point and finds the next closest point. It then puts them into a cluster. The next closest point to this cluster's centroid is then taken and if the cluster's centroid is the closest point, a branch is created between the cluster and the point. If another point is closer than the original cluster's centroid a new cluster is created. This continues for each point. See Figure 6 for a diagram of this process.

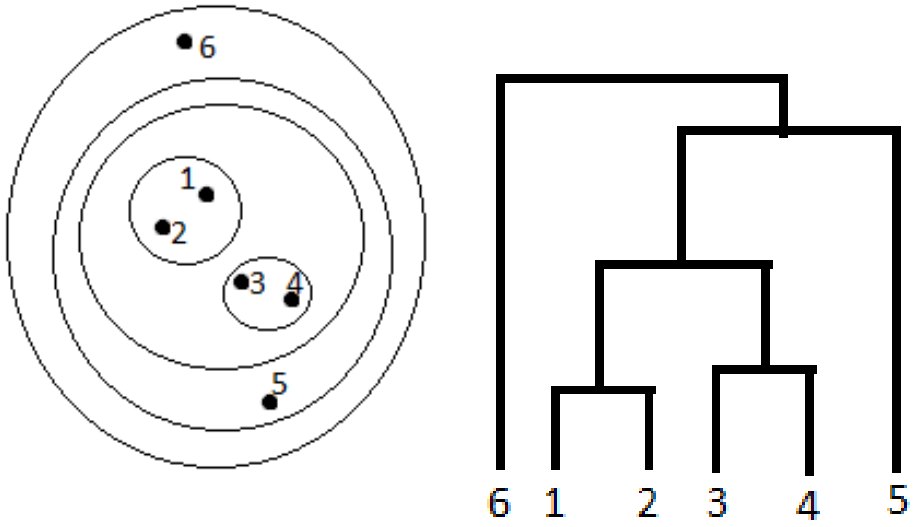

Figure 6 – Diagram of Agglomerative Hierarchical Clustering.

The clustering was done using Cluster 3.0, which outputs a *.CDT file containing the genes (rows) and microarrays (columns) in order of similarity across genes and microarrays. Cluster 3.0 also outputs a *.ATR (contains information on microarray tree branches in the visualization) and *.GTR (contains information on gene tree branches) file. These files are then loaded into Java TreeView, which displays the dendrogram (a hierarchical tree) and relative expression values of the microarrays across the filtered genes. Each gene for each microarray is equivalent to one pixel in the grand representation. That pixel will be green if the expression for that gene is below median expression and red if it is above median expression; median expression in this case refers to the median expression value of the gene across the entire data set. The gene dendrogram is displayed with its branches extending horizontally

towards the data and the microarray dendrogram is displayed with its branches extending vertically towards the data. Figure 7 shows a screenshot of the visualization that Java TreeView provides.

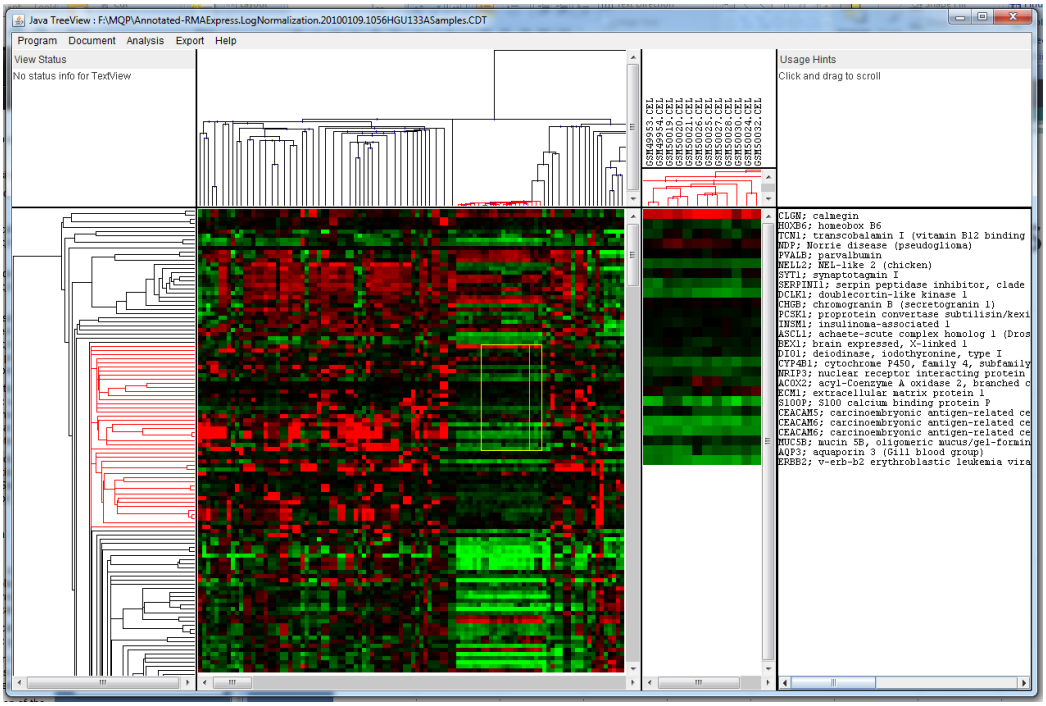

Figure 7 – Screenshot of Java TreeView GUI

The following settings were used for Cluster 3.0 and Java TreeView:

- Cluster 3.0
  - Input: data.txt
  - Settings:
    - Filter Data Tab: SW (Gene Vector) at 0.5 → Apply Filter
    - Adjust Data Tab: Median Center Genes, then Median Center Arrays
    - Hierarchical Tab: Cluster Genes & Array → Centroid Linkage
  - Output: data.cdt, data.gtr, data.atr
- Java TreeView
  - Input: data.cdt, data.gtr, data.atr

3.2.3. Cluster Summarization

Cluster summarization refers to the grouping of sets of microarrays based on the similarity that is shown through Java TreeView. Basically, a branch node is selected and the children (microarrays) of that branch are grouped together. During the selection of branch nodes, correlation values were also taken into consideration. Correlation described how similar the microarrays in a given branch were, and we selected branches that had a correlation of 0.4 – 0.5. This was implemented through a java program that took input in the form of an *.ATR file and returned a list of microarray names corresponding to the children of a given node within the *.ATR file. More specifically, this was done by rebuilding the tree below that point and then recording each of the microarrays that made up the leaves of the tree. The

average and median centroids were obtained by taking a list of microarrays and a *.CDT file, and computing the average of the values of each gene for each of the given microarrays obtained from the *.CDT file. These were then written to a file, that mimicked a *.CDT file, so it could be displayed by Java TreeView.

3.2.4. Unsupervised Ranking by Distance

The distance between two points can be calculated in multiple ways. Two common distance functions are Euclidean and Manhattan. Expressions for the Euclidean and Manhattan distances are shown below:

$$Euclidean(\boldsymbol{a}, \boldsymbol{b}) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}$$

$$Manhattan(\boldsymbol{a}, \boldsymbol{b}) = \sum_{i=1}^{n} |a_i - b_i|$$

In both distance metrics, **a** and **b** represent the points, and **n** represents the number of dimensions. When given three points: if the distance between points **a** and **b** is lower than the distance between points **a** and **c**, then point **b** is closer, or more similar, to point **a** than point **c** according to that distance metric.

Both methods of measuring distance were employed to rank similarity to a point of interest. A microarray was considered more similar to that point if the distance between the microarray and the point was relatively low. For these experiments the distance functions were calculated for each of the 1056 samples to the point of interest. The Euclidean distance function used was implemented by Weka. The Manhattan distance function used was implemented by summing the absolute value of the difference for each gene between the microarray and a point of interest. This was implemented in a java program written by the authors.

Once the distances between each microarray and the point of interest were calculated, the list of microarrays was ordered based on distance from smallest to largest. This was given as input to another java program which, using the *.CDT file of the initial clustering, ordered the microarrays based on the given ordering. By passing this *.CDT file into Java TreeView, we were able to observe the similarities and differences in gene expression levels with a new ordering of the microarrays.

3.2.5. Supervised Ranking by Distance (Functional Gene Groups)

In the previous ranking, the distances were calculated over every gene that passed through the filter. In order to generate better results, we decided to focus on groups of genes that are involved in specific functions. Using dChip, we extracted lists of genes involved in functions that may be relevant to Claudin-Low tumors (dChip: Introduction and Installation, 2007). The functional groups that we extracted were: apoptosis, basement membrane, cell-cell adhesion, cell communication, cell cycle, cell

proliferation, DNA repair, epidermis development, and tight junction. We then performed distance calculations within each of these functional groups between the microarrays and a point of interest.

3.2.6. Other Computer Programs

Several other computer programs were written to facilitate the clustering and visualization process. Most of these were written by the author to parse or manipulate files for use with the tools mentioned above. These Java programs had various functions, including:

- Filter and order genes in a *.CDT file based on a given gene list
- Rename/move files; this was used to rename experiment files based on the study they came from
- Move files based on the header of the raw *.CEL file
- Add gene names and descriptions to a *.CDT file, replacing probe set names using a table that mapped probe sets to genes found in BRB ArrayTools.

4. Results

The result of the initial hierarchical clustering was the discovery of 9 sub-clusters exhibiting different expression patterns. Furthermore, a highly correlated cluster of microarrays was discovered. These microarrays were actually derived from cell line (MDA-MB-231) that was originally derived from the pleural effusion of a breast cancer patient, and widely used in breast cancer research as an *in vitro* cell culture model for highly aggressive and invasive breast cancer. Because these cell lines are known to show an expression pattern similar to Claudin-Low tumors, we used their profiles as a proxy for Claudin-Low tumor expression. The microarrays for these cell lines were grouped as cluster 1 and the other samples were compared to that cluster's centroid without supervision and with supervision (incorporating functional gene groups). The final results were the discovery of two microarrays that were very similar to the cell lines, and 12 microarrays that were found to be similar to the cell lines in some aspects, but not in some aspects characteristic of Claudin-Low tumors.

4.1. Initial Clustering

After the data was run through RMA, log normalized, median centered by gene and by microarray, and filtered, the resulting genes were used to perform hierarchical clustering. The goal of clustering was to find groupings of microarrays that exhibited similar genetic expression. This would allow the classification of different groups and the discovery of microarrays exhibiting similar expression profiles like the recently discovered Claudin-Low subtype (Herschkowitz, et al., 2007).

In the filtering step, 6,754 genes out of 22,283 passed the variability filter (SW > 0.5). This amount of filtering allowed analysis over a significant number of genes compared to those used in the Herschkowitz et al. (2007) profiling paper. These roughly 7,000 genes were used in the clustering shown in Figure 8. The horizontal dendrogram (rows) is a hierarchical clustering of the genes based on similar expression pattern throughout the samples. It was expected that the gene clustering would group genes that are regulated similarly and therefore likely have similar function. For example, some subtypes express a group of genes that induce a higher rate of proliferation while others do not. The genes in this set are regulated similarly; they are upregulated in the subtypes that have higher rates of proliferation but downregulated in those that do not. Since this group of genes shows a similar expression pattern they will show high correlation and are grouped together in the hierarchy. The vertical dendrogram shows a hierarchical clustering of the microarrays. Microarrays showing similar expression patterns will have a higher correlation and will be grouped together.

The height of each branch of the dendrogram is based on the level of correlation between the two nodes it connects. A branch with little or no correlation between the nodes on the left and right will be very far in height from the colored pixels, while a very highly correlated branch will be very close in height to the pixels. The actual microarray expression values are seen in Figure 8 as the red and green pixels. Since all of the data was median centered by gene and by microarray, the color of the pixels illustrates whether it is above the median expression level (red) or below (green). Clusterings were also performed on a set of genes of greater variance (SW > 1.0, yielding around 700 genes) and on all of the genes (around 22,000). These clusterings were made for referential purposes and the microarrays and genes clustered similarly. They are not shown here.

One property of dendrograms generated from hierarchical clustering is that their branches can be flipped over a node. In other words, the left-most branch at a node can switch positions with the right-most branch without compromising the data. This property allowed us to manipulate the microarray orderings so that the arrays were ordered by cancer based on the genes that were expressed above median level. This ordering shows the progression of cancer from stem cells, to basal cells, and finally to luminal cells. The microarrays are labeled on the bottom of the image based on the developmental cell type for which they exhibit the expected gene expression pattern. SC stands for stem-like cells. In basal cells, keratins 5, 14, 6, and 17 exhibit above-median levels of expression while in luminal cells, keratins 8 and 18 exhibit above-median levels of expression. These keratins are marked in Figure 8 and are expected to be seen as red pixels for the given subtype where they are expected to exhibit above-median levels of expression.

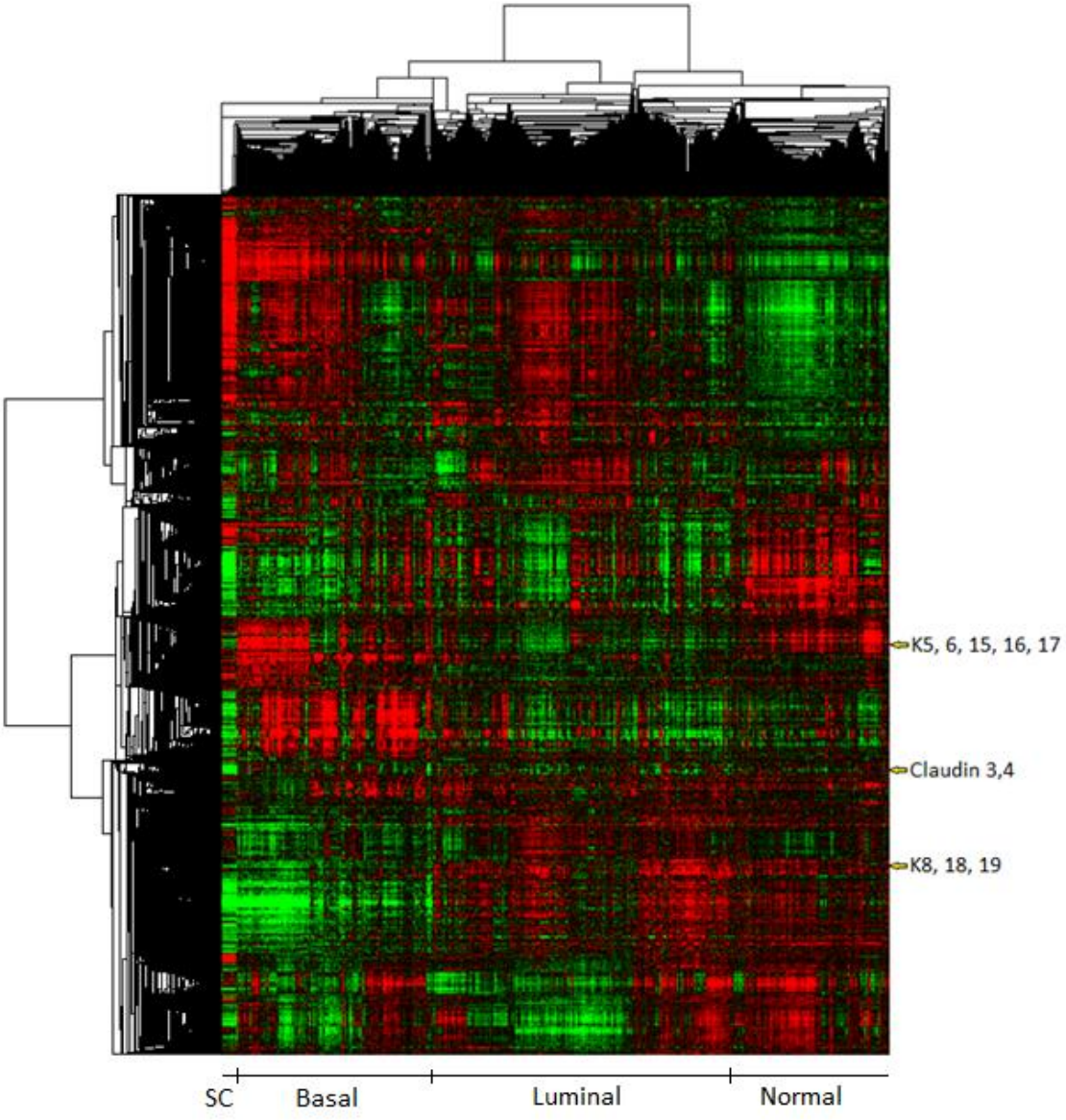

Figure 8 – Microarrays from Hierarchical Clustering Ordered by Development

4.2. Cluster Summarization

Small clusters of microarrays can be merged into larger clusters based on their expression patterns; these larger clusters were chosen by hand based on homogeneity of the groups of microarrays as well as having a correlation value between 0.4 and 0.5. Figure 9 shows the average and median centroids of each of the nine resulting clusters.

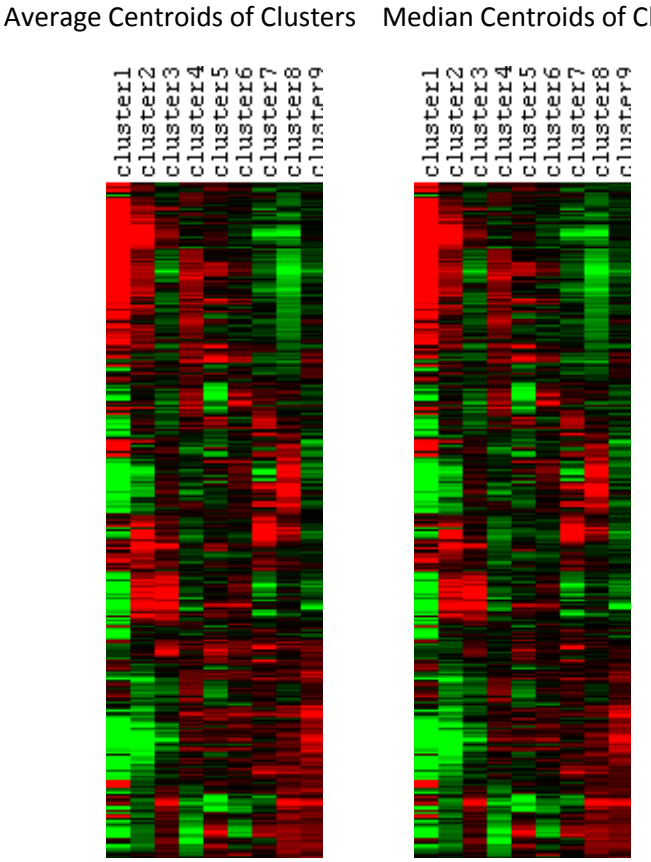

Figure 9 – Average and Median Centroids of Cluster Summarizations

The microarrays on the left region of Figure 8, which display very high intercluster correlation and fall under the SC development label, has a gene expression pattern similar to the Claudin-Low subtype. However, these microarrays are all from the same study. Additionally, they also have much higher and lower expression values than the median as shown by the brighter reds and greens in these samples. It was discovered through inspection of the study that these microarrays came from cell lines derived from a tumor sample that is the current best model for the Claudin-Low subtype. Specifically, the study was looking for tumor cells that metastasize to the lung (Minn, et al., 2005). All of these microarrays were added to the same cluster, cluster1, during cluster summarization, along with two other clinical samples (GSM178055 and GSM177944). These two clinical samples are considered Claudin-Low since they have such a high correlation with Claudin-Low cell line samples. The GEO

database identification values for these microarrays can be found in Table 1. Samples GSM49953 through GSM50034 listed in Table 1 are the cell line microarrays.

Table 1 - Cluster 1 Members (Claudin-Low cell lines and two clinical samples)

GSM49953	GSM49954	GSM49955	GSM49956
GSM49957	GSM50018	GSM50017	GSM50019
GSM50020	GSM50021	GSM50024	GSM50034
GSM50032	GSM50024	GSM50022	GSM50025
GSM50027	GSM50026	GSM50028	GSM50030
GSM50031	GSM50023	GSM50029	GSM50033
GSM178055	GSM177944		

4.3. Ordering by Distance Metrics

Once the information about the cell lines was available it became clear that while these are the best candidates for being Claudin-Low subtype breast cancer they are also already known to be of this subtype and their discovery is not novel. Instead, the microarrays of these cell lines were used as a comparison to find tumor samples that were most like the Claudin-Low subtype. This allowed us to use distance metrics to find more subtle relationships in the data set. Figure 10 shows the same data as Figure 8 (genes in the same order for both figures) but the microarrays are ordered by average Euclidean distance to all of the microarrays from cluster 1. Figure 11 and Figure 12 are the same as Figure 10, but the microarrays were ordered by Euclidean distance to the cluster 1 centroid and by Manhattan distance to the cluster 1 centroid respectively.

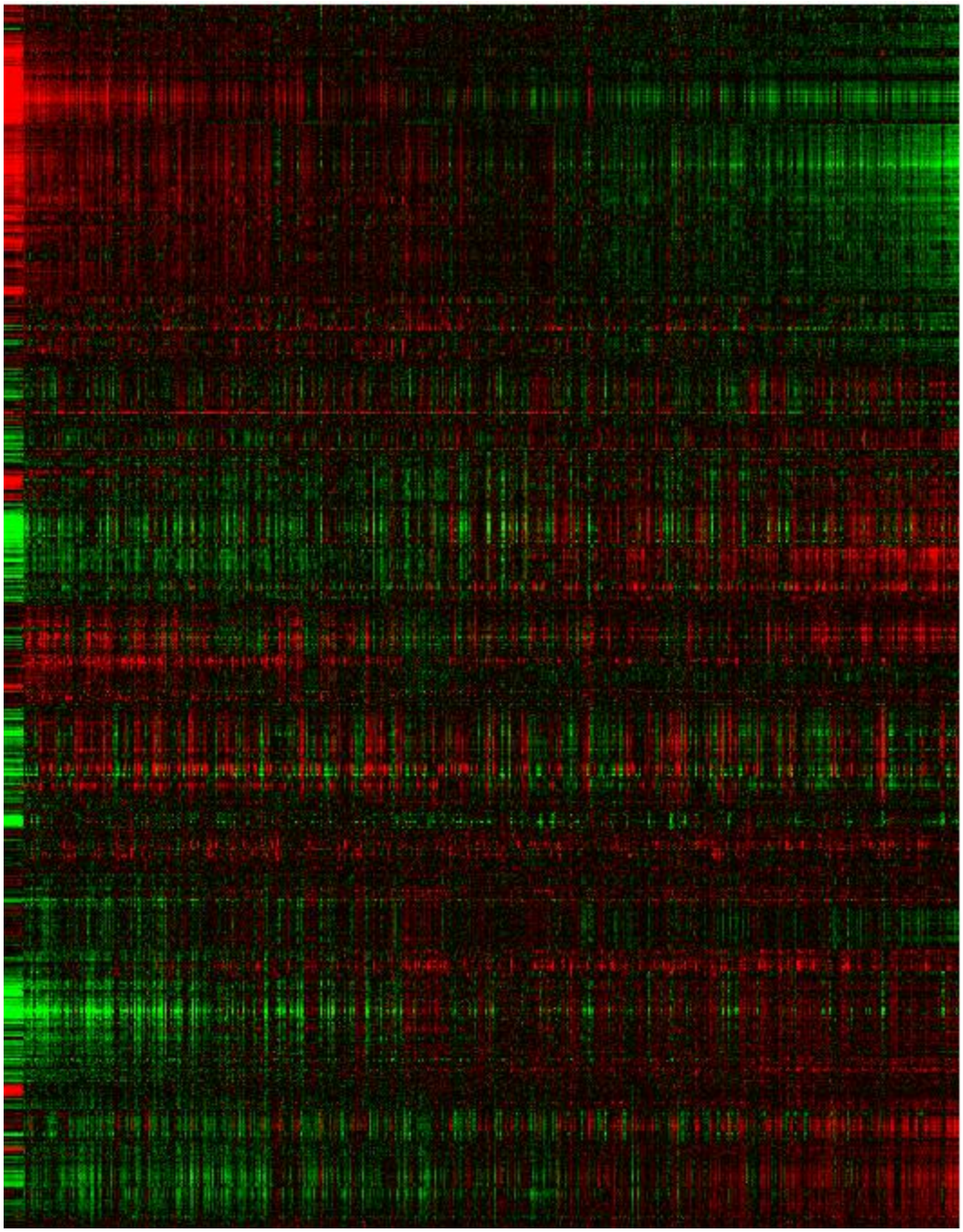

Figure 10 – Microarrays ordered based on Average Euclidean Distance to Cluster 1 Microarrays

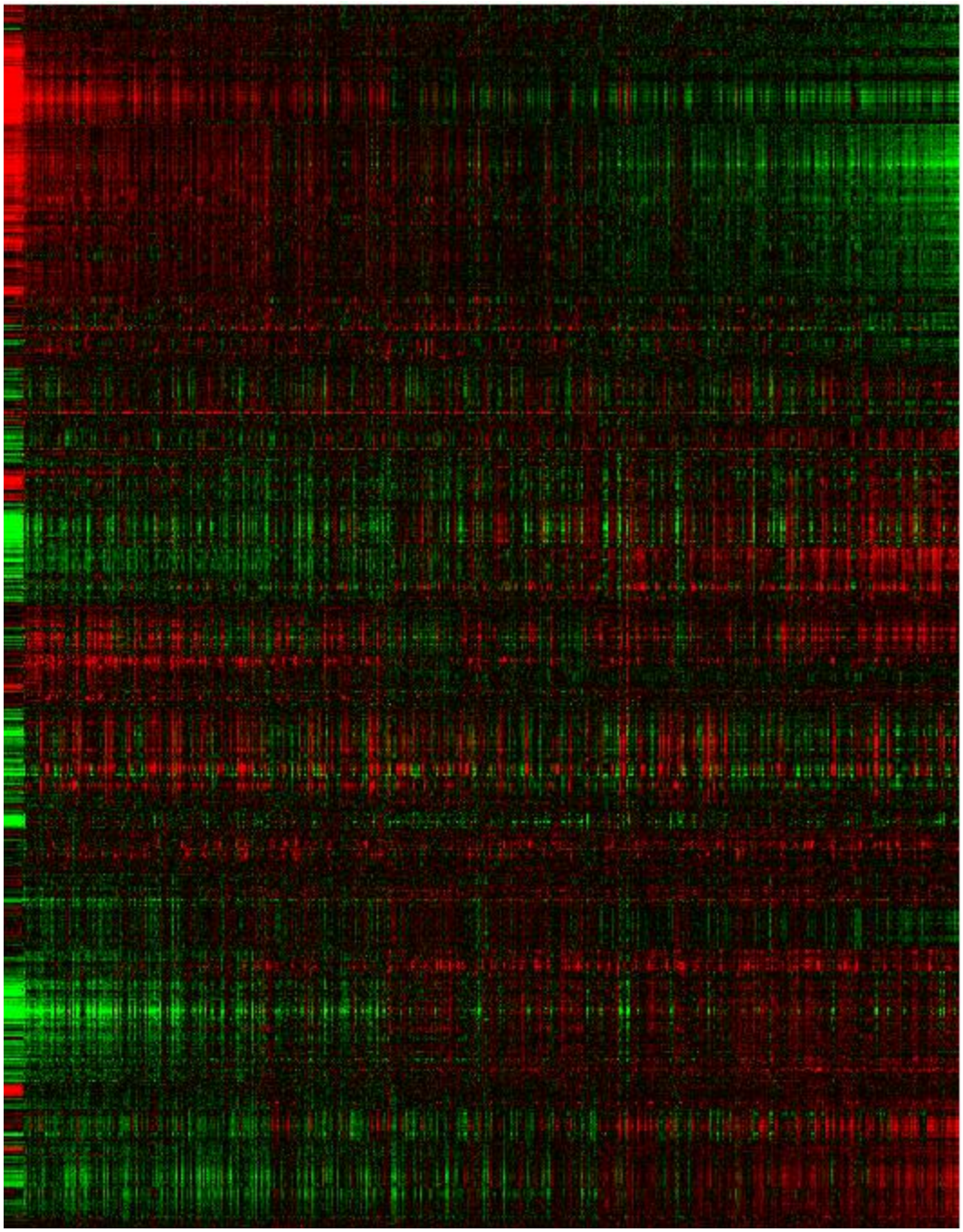

Figure 11 – Microarrays Ordered by Euclidean Distance to Cluster 1 Centroid

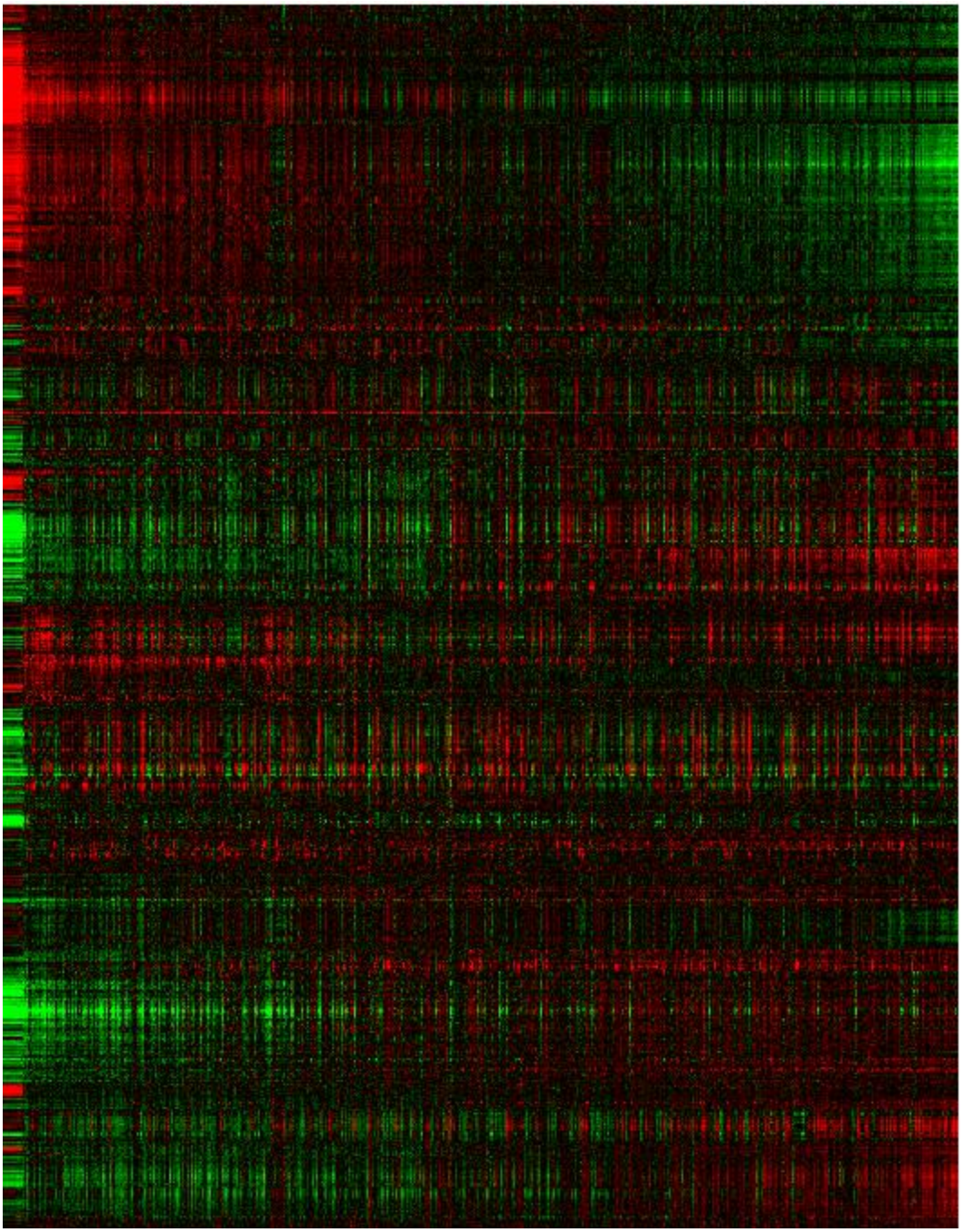

Figure 12 – Microarrays Ordered by Manhattan Distance to Cluster 1 Centroid

Figure 13 plots the Euclidean distances from each microarray to the cluster 1 centroid ordered smallest to largest. From this figure, it is clear that the cell lines listed in Table 1 made up the first 22 points, marked with A in the diagram. However, an interesting result is that several other microarrays ranked closer to the cluster 1 centroid than the two non-cell-line microarrays that were part of cluster 1. A list of these microarrays can be seen in Table 2. The area where these samples can be seen in the graph is marked B in the diagram. These microarrays correspond with the microarrays before the first inflection point, about 0.9 in Figure 13, and the tumors they represent could be considered as more

Claudin-Low-like. However, after these microarrays were investigated closely, they were found to have too many basal markers to be considered purely Claudin-Low. Given that the Claudin-Low subtype is very similar to the basal subtype, this is not an entirely unexpected result. It is also possible that these tumors from these microarrays have sub-populations of Claudin-Low cells which skew the microarrays towards the Claudin-Low cell line's centroid.

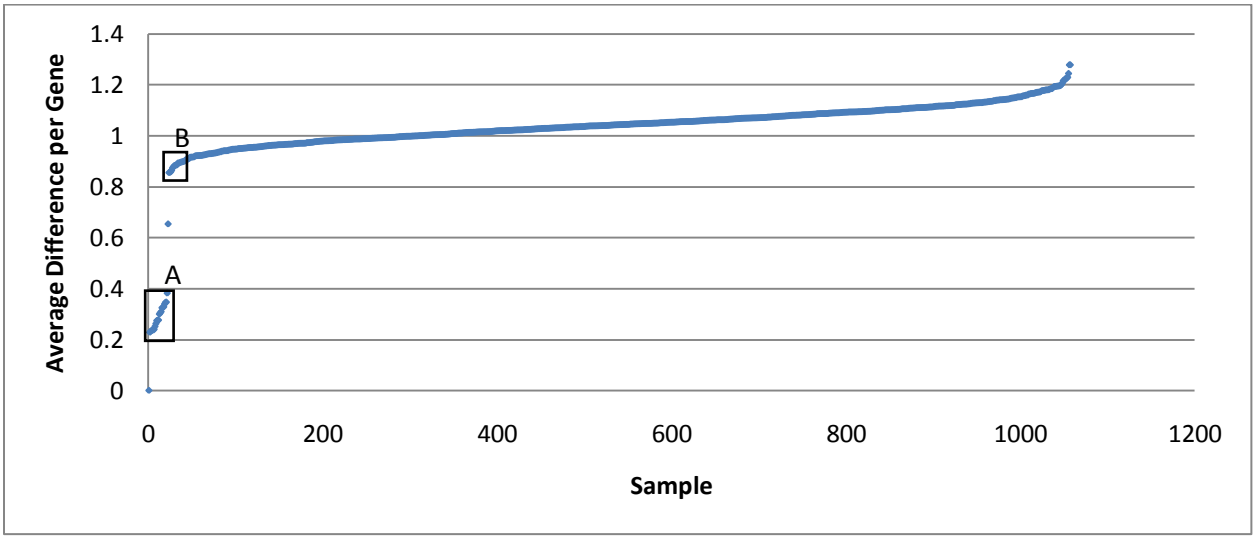

Figure 13 – Values for Ordering by Manhattan Distance from Smallest to Largest

Table 2 - Microarrays more similar than Cluster1 Clinical Microarrays by Manhattan Distance

GSM107204
GSM50067
GSM178025
GSM178079

4.4. Ordering of Samples Considering Functional Gene Groups

The genes can also be categorized into functional groups. This functional grouping of genes can be used to rank similarity in a more supervised manner. The average difference per gene in a given functional gene group between a microarray and the cluster 1 centroid was calculated. The average difference per gene can be conceptualized as the Manhattan distance for a given functional group divided by the number of genes in that functional group. The average value for each of the average difference per gene per gene group was calculated and used to order the microarrays from smallest to largest. Figure 14 shows this ordering. While it is not significantly different than the previous orderings, the values obtained seem to exaggerate the differences. The equation below shows the formula that was used to calculate this. In this formula 1,…,m each denote one functional gene group containing n_j genes. Each of 1,…,n_j refers to a gene within the functional group j. **a** and **b** each refer to a microarray and the subscripts refer to a gene value.

$$avg.diff.per\ gene\ per\ func.gene\ group(\boldsymbol{a},\boldsymbol{b}) = \frac{\sum_{j=1}^{m}\frac{\sum_{i=1}^{n_j}|a_{ji}-b_{ji}|}{n_j}}{m}$$

Mathematically speaking, this analysis simply places equal weight on the differences between the genes of each functional gene group. Standard Manhattan distance can be normalized by the number of genes used to calculate it and would then be equal to the average difference per gene, placing equal weight on each gene. For the average difference per gene per gene group within a given gene group all of the genes hold equal weight and are condensed to one average. Thus no matter how many genes are in a given gene group the values can be compared since they have the same dimensionality. When the averages for each of the gene groups are averaged each gene groups have the same weight regardless of how many genes were used to calculate it.

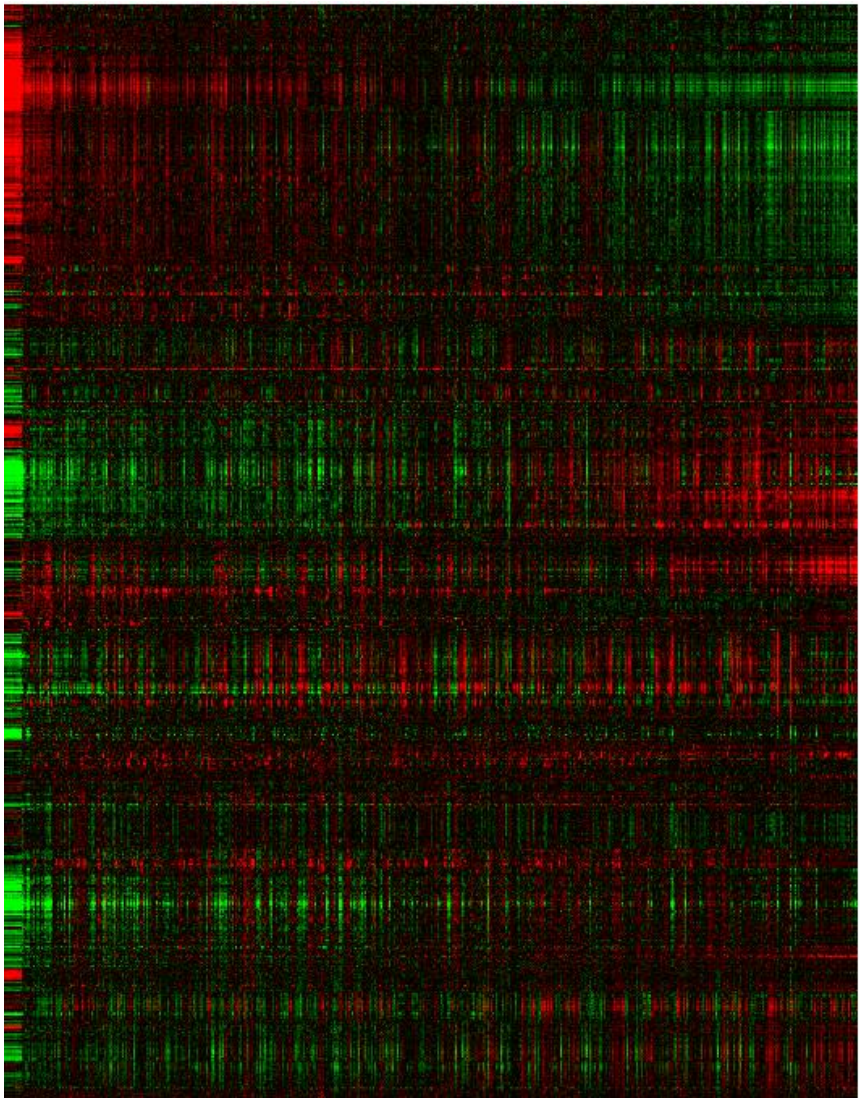

Figure 14 – Microarrays Ordered by Average Difference per Gene per Functional Gene Group

Figure 15 is a graph of the values used to order by the average difference per gene per functional gene group (Figure 14). The inflection point is about 1.0 and again, corresponds roughly with one of the clinical samples from cluster 1. Again, several microarrays ranked closer than one of the clinical microarray samples that were grouped in cluster 1. These samples are listed in Table 3.

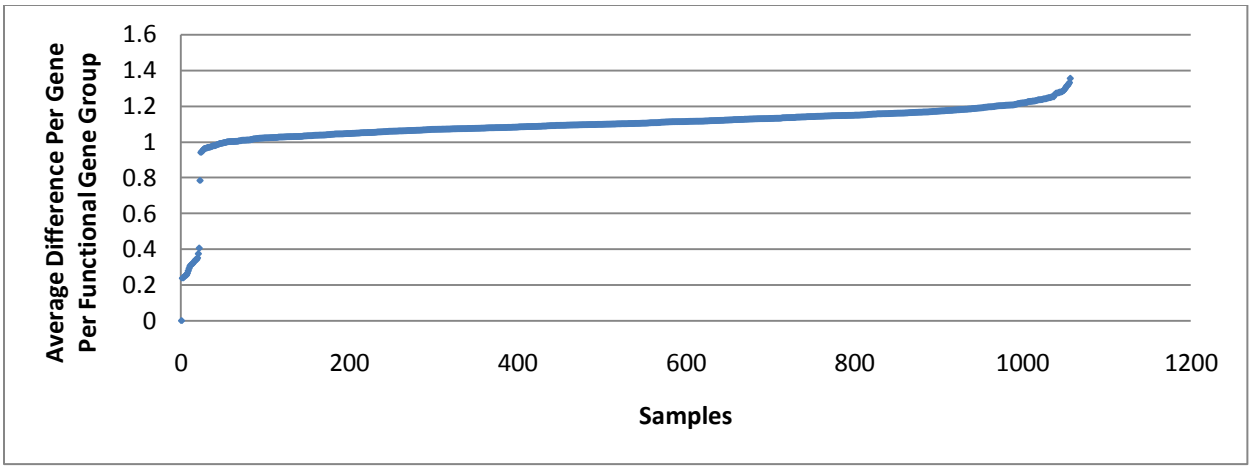

Figure 15 – Values for Ordering by Average Difference per Gene per Functional Gene Group

Table 3 - Microarrays more similar than Cluster1 Clinical Samples by averaged average difference per gene per gene group

GSM50067
GSM107218
GSM79189
GSM177980
GSM65845

Another method of reordering the microarrays using the functional groups is to perform an order summarization. To do this, the microarrays are sorted in increasing order according to their Euclidean distances to the cluster 1 centroid on each of the functional groups. Each microarray gets a score equal to the sum of its placement within each ordering. For example, given three functional gene groups, a microarray might have the fifth, seventh, and tenth lowest distance from the cluster 1 centroid, giving it an order summarization score of twenty-two. This was done using the functional gene groups as in Figure 14. Figure 16 shows a graph of the placement sums. The group with the smallest placement values and the next point after that are Claudin-Low cell lines. The point after that are two microarray samples; one of these is a clinical microarray that was grouped with cluster 1 and the other was a different microarray sample. The second clinical microarray that was added to cluster 1 is found ranked number 36 out of 1056. This sample corresponds with the inflection point seen in the graph below. This implies that the samples ranked from 24 to 36 all show very similar expression patterns to the Claudin-Low expression pattern when considered on a per gene group level. However, after investigation these 12 samples, listed in

Table 4 is a set of the several microarrays found to rank more closely to cluster 1's centroid than the two clinical tumor microarrays in the manhattan distance orderings and have already been investigated and found to be more basal-like than Claudin-Low like.

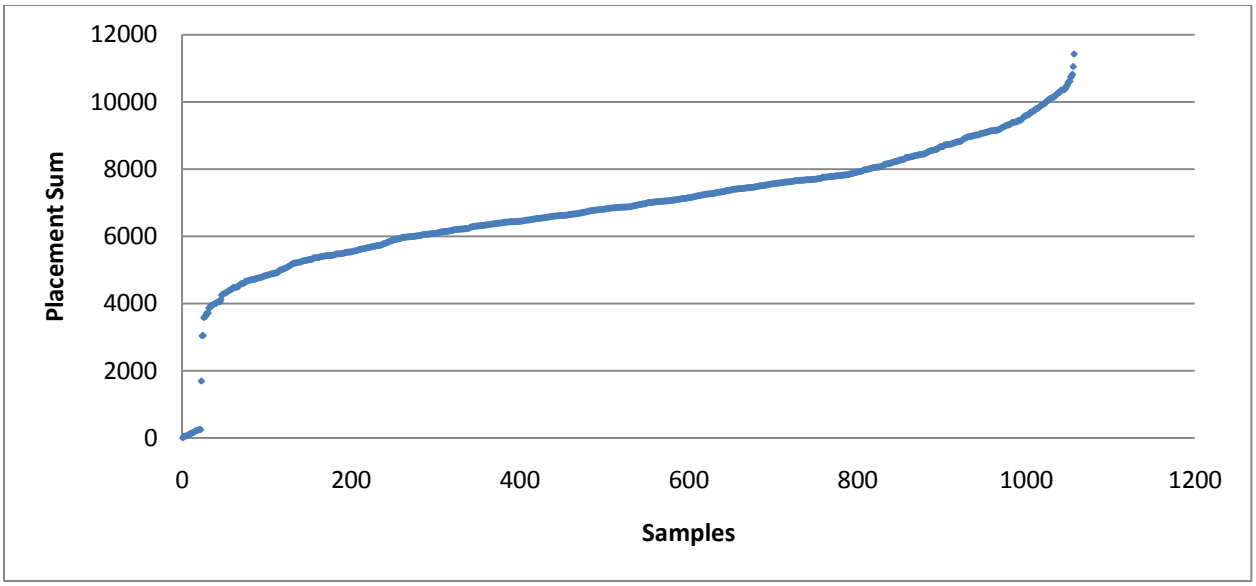

Figure 16 – Functional Gene Group Distance Order Summarization

Table 4 - Microarrays more similar to Cluster 1 Samples by Order Summarization

GSM79189
GSM50037
GSM65792
GSM79270
GSM150964
GSM65857
GSM50043
GSM107107
GSM79344
GSM79340
GSM65838

A comparison of the per functional gene distances between the microarrays and cluster 1 can be seen by plotting the distance values for each group in a parallel coordinates graph (Figure 17). This method allows us to compare how similar each microarray is to the cluster 1 centroid by looking at one functional gene group at a time; the similarity between two microarrays over different functional groups may vary and we want to explore the possibility that certain functional groups are more useful in clustering Claudin-Low tumors than others. The collection of microarrays with the lowest values over each functional group are the cell line samples, listed in Table 1, which exhibited low overall distance in Figure 15. This collection is marked with an arrow in Figure 17. It is notable that there are no other microarrays that have a lower distance value in any of the functional groups than the cluster 1 microarrays.

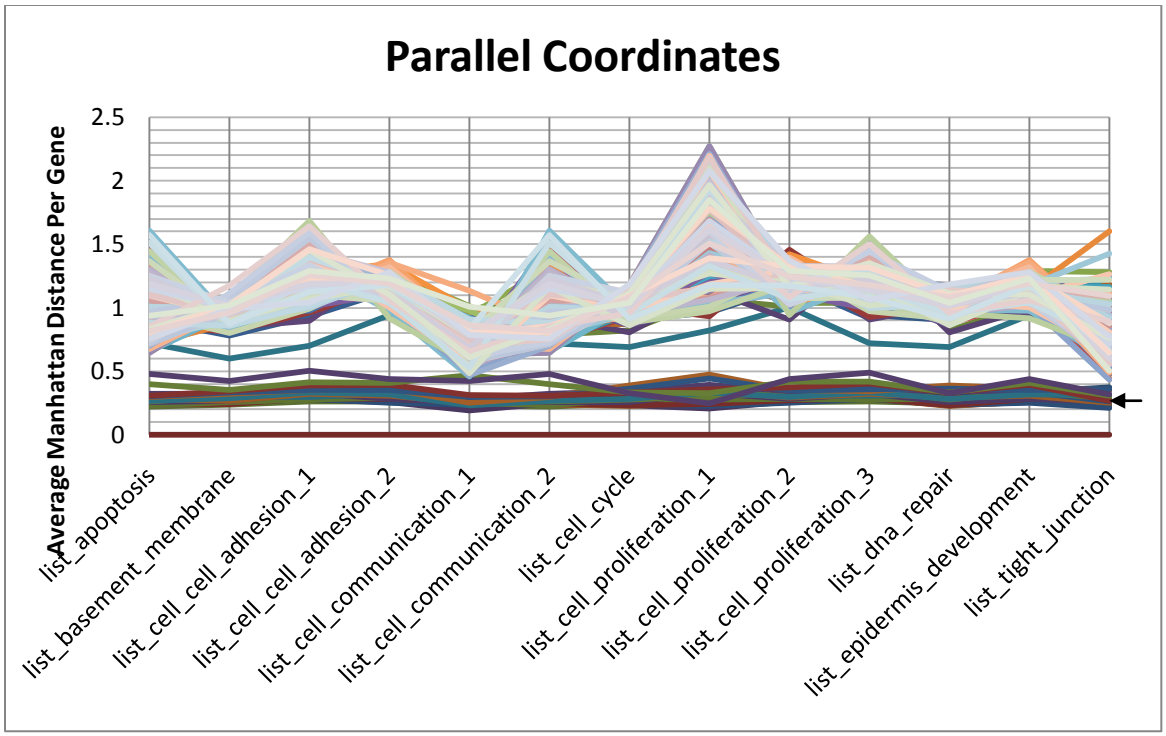

Figure 17 – Parallel Coordinates Graphing the Distance for each Functional Gene Group to Cluster 1's

5. Discussion

5.1.Existence of Claudin-Low Subtype

In the comparative profiling of murine and human breast cancer tumors (Herschkowitz, et al., 2007), roughly 2 to 3% of microarrays were found to exhibit genetic expression profiles similar to the new Claudin-Low subtype. In this project, it was expected that by analyzing a larger number of microarrays, this proportion of Claudin-Low like microarrays would be maintained in a meta-analysis of human breast cancer. Since this project's data set contained 1056 microarrays of human breast cancer samples, 20 to 40 microarrays were expected to be identified as Claudin-Low. However, only two to three microarrays were found. There are several possible explanations for this result.

The simplest reason is that there could be less Claudin-Low tumors found in Humans than in mice. Since the Minn breast cancer cell lines (Minn, et al., 2005) exhibit all of the characteristics of Claudin-Low tumors, and were derived by sorting human breast cancer tumors, it is known that Claudin-Low cells do exist in humans.

Second, Claudin-Low cells exhibit a very different morphology than other breast cancer cells; they are spindloid and metaplastic (Hennessy, et al., 2009). It is possible that when tumors were collected and classified, Claudin-Low tumors were clinically classified as non-breast tumors. This excludes them from the data sets used in this project. This is a problem often associated with survey studies, and is known as an ascertainment bias.

Third, it is possible that the samples given in the dataset are not made up of one cell type. The image below, Figure 18, depicts tumor samples stained with hemotoxylin and eosin dyes that show cell morphology. These samples are examples of tumors and show that tumors do not contain only one type of cell. There is evidence within the genetic expression data that, while some of the tumors exhibit above-median expression of genes that are expected to be upregulated in Claudin-Low tumors, they also exhibit above-median expression of other genes that are characteristic of another subtype. This could be an artifact of tumor samples which contain some Claudin-Low cells and some of another subtype. A model explaining this follows: Claudin-Low tumors begin as stem cells and develop into cells of other stages without going through the intermediate stages. For example, a Claudin-Low tumor arises, originating from a stem-like cell, and when there are very few cells making up the tumor, one of the cells differentiates into a Luminal type cell and continues to divide. Some of the cells will be progenitors of the stem-like cells while the rest of the cells will be progenitors of the differentiated cell (luminal). The result of this would be a tumor where some of the cells are identified as Luminal while the others would be identified as Claudin-Low. Once the tumor cells achieve these morphologies (such as luminal or basal), they will exhibit characteristics similar to cancers of their respective cell types, despite their Claudin-Low origins. Another possible model of multi-type tumors is that the tumor arises in a developed cell (Luminal, for example) and then goes through an Epithelial-Mesenchymal Transition due to mutations or environmental factors. Both models are currently being considered by breast cancer research experts. The result here shows evidence for both of these models, but does not give stronger evidence for one over the other.

Even if Claudin-Low tumors were not a major subtype or if their numbers were much lower than those of other breast tumors, they continue to merit further research. With many of the luminal and basal tumors, the cancerous cells grow locally and are thus clinically maintainable. However, Claudin-Low tumors are known to travel from their original environment, and since claudins are proteins vital for proper cell-cell communication, the spread of these cells can induce cancer formation in cells at other locations of the body, leading to metastasis.

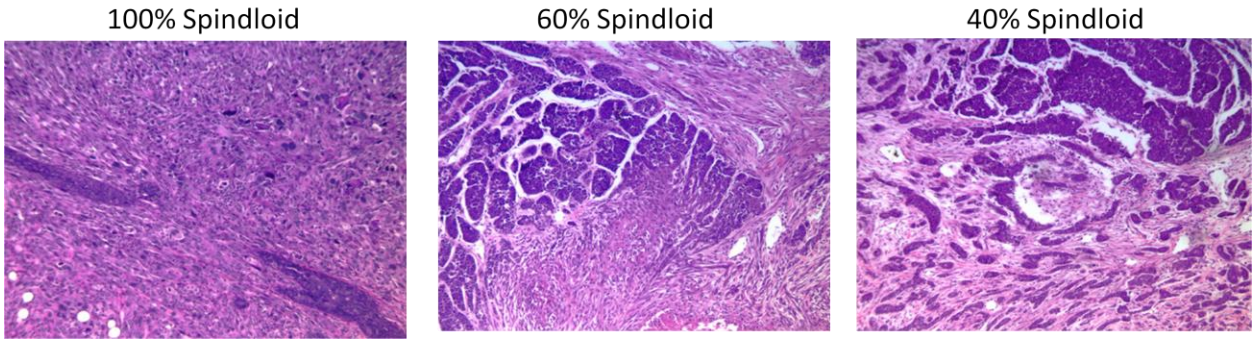

Figure 18 - HE Stains of Mucine Breast Tumors exhibiting varying levels of splindloid morphology

5.2.Clustering Summary Lead to Nine Subtypes

Sorlie et al. (2001) used roughly 80 tumor microarrays and stratified them into subclasses based on genetic expression and survival data. This led to five subtypes: basal-like, normal-breast-like, ERBB2-overexpression, and two types of luminal-like. They note that the luminal type tumors can be stratified into at least two subtypes. Hu et al. (2006) tried to find a set of genes that can classify tumors into subtypes and found six subtypes; the sixth subtype was a novel type called IFN-like.

The cluster summarization of the initial hierarchical clustering found sets of microarrays that displayed similar expression profiles and showed a branch correlation of 0.4 to 0.5. This information was used to separate the microarrays into sub-clusters. Figure 9 shows the mean and median centroids of each of the nine sub-clusters. The resulting nine clusters had low enough correlation to each other that they can be considered different. This differs from the result of other subtype classifications, which produced five or six subtypes. Since these clusters are different enough, they may warrant further investigation to determine if more subtypes exist; a new subtype could be useful in determining prognosis and treatment.

The goal of this project was not to find new classifications of breast cancer tumors, but some preliminary classification of the nine clusters can be offered. Cluster 1 was created using microarray data of the Minn cell lines, as well as two other microarrays (GSM178055 and GSM177944), which demonstrate similar genetic expression to the Minn cell lines. Cluster 1's centroid was used throughout this project as the point of comparison in attempting to find the tumor microarrays that exhibit similar expression to Claudin-Low. Clusters 2 and 3 demonstrate above-median expression of keratins, which are expected to be upregulated in basal-like tumors despite demonstrating different levels of expression in other gene groups. The fact that the two basal-like clusters could be separated through correlation suggests that there may be a major sub-class within basal-like tumors; this is a previously undiscovered result. Cluster 4 has above-median expression of HER2. Clusters 5 and 6 have above-median expression

of luminally expressed keratins. Cluster 7 did not have any identifying genes over or under expressed. Clusters 8 and 9 exhibit normal breast tissue expression; these samples could have arisen due to tumor samples containing half basal-like cells and half luminal-like cells. Alternatively, the sample could have been taken improperly, and may actually contain normal breast cells as opposed to tumor cells.

6. Works Cited

Al-Hajj, M., Wicha, M. S., Benito-Hernandez, A., Morrison, S. J., & Clarke, M. F. (2003). Prospective identification of tumorigenic breast cancer cells . *Proceedings of the National Academy of Sciences of the USA* , 3983-3988.

Bolstad, B. M., Irizarray, R. A., Astrand, M., & Speed, T. P. (2003). A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Variance and Bias. *Bioinformatics* , 185-193.

Bryant, D. M., & Mostov, K. E. (2008). From cells to organs: building polarized tissue. *Nature Reviews Molecular Cell Biology 9* , 887-901.

Cooper, G. M., & Hausman, R. E. (2007). *The Cell A Molecular Approach Fourth Edition.* Sunderland, MA: Sinauer Associates, Inc.

dChip: Introduction and Installation. (2007, 12 12). Retrieved 4 22, 2010, from Cheng Li Lab: http://biosun1.harvard.edu/complab/dchip/install.htm

DeOme, K. B., Faulkin, L. J., Bern, H. A., & Blair, P. B. (1959). Development of Mammary Tumors from Hyperplastic Alveolar Nodules Transplanted into Gland-free Mammary Fat Pads of Female C3H Mice. *Cancer Research* , 515-526.

DeRisi, J. L., Iyer, V. R., & Brown, P. O. (1997). Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale. *Science* , 680-686.

Freeman, S. (2005). *Biological Science Second Edition.* Upper Saddle River, NJ: Pearson Prentice Hall.

GeneChip Mouse Expression Set 430. (n.d.). Retrieved 4 27, 2010, from Affymetrix: http://www.affymetrix.com/support/help/faqs/mouse_430/faq_8.jsp

Gonzalez-Mariscal, L. (2006). *Tight Junctions.* New York: Springer Science + Business Media.

Griffiths, A. J., Wessler, S. R., Lewontin, R. C., & Carroll, S. B. (2008). *Introduction to Genetic Analysis 9th Edition.* New York, NY: Freeman and Company.

Hennessy, B. T., Gonzalez-Angulo, A.-M., Stemke-Hale, K., Gilcrease, M., Krishnamurthy, S., Lee, J.-S., et al. (2009). Characterization of a Naturally Occurring Breast Cancer Subset Enriched in Epithelial-to-Mesenchymal Transition and Stem Cell Characteristics. *Cancer Research* , 4116-4124.

Herschkowitz, J. I., Simin, K., Weigman, V. J., Mikaelian, I., Usary, J., Hu, Z., et al. (2007). Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. *Genome Biology* , R76.1-R73.17.

Hoon, M. d. (2002). *Cluster 3.0 for Windows, Mac OS X, Linux, Unix*. Retrieved 1 20, 2010, from University of Tokyo, Human Genome Center: http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/manual/

Horspool, D. *Central Dogma of Molecular Biochemistry with Enzymes.* Wikimedia.

Hu, Z., Fan, C., Oh, D. S., Marron, J. S., He, X., Qaqish, B. F., et al. (2006). The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics* , 1-12.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., et al. (2002). Exploration, Normalization, and Summaries of High Density Oligonucleotide Array probe Level Data. *Biostatistics* .

Li, X., Lewis, M. T., Huang, J., Gutierrez, C., Osborne, C. K., Wu, M. F., et al. (2008). Intrinsic resistance of tumorigenic breast cancer cells to chemotherapy. *Journal of the National Cancer Institute* , 672-679.

Majumder, S. (2009). *Stem Cells and Cancer.* New York: Springer Science + Business Media.

Mani, S. A., Guo, W., Liao, M.-J., Eaton, E. N., Ayyanan, A., Zhou, A. Y., et al. (2008). The Epithelial-Mesenchymal Transition Generates Cells with Properties of Stem Cells. *Cell* , 705-715.

Minn, A. J., Gupta, G. P., Siegal, P. M., Bos, D. P., Shu, W., Giri, D. D., et al. (2005). Genes that mediate breast cancer metastasis to lung. *Nature* , 518-524.

Morin, P. J. (2005). Claudin Proteins in Human Cancer: Promising New Targets for Diagnosis and Therapy. *Cancer Research* , 9603-9606.

Reyal, F., Vliet, M. H., Armstrong, N. J., Horlings, H. M., Visser, K. E., Kok, M., et al. (2008). A comprehensive analysis of prognostic signatures reveals the high predictive capacity of the Proliferation, Immune response and RNA splicing modules in breast cancer. *Breast Cancer Research* , 1-15.

Saldanha, A. (2003). *Welcome to Java TreeView*. Retrieved 1 20, 2010, from http://jtreeview.sourceforge.net/index.html

Schena, M. (1999). *DNA Microarrays: A Practical Approach.* Oxford University Press.

Sorlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences of the United States of America* , 10869-74.

Voet, D., Voet, J. G., & Pratt, C. W. (2008). *Fundamentals of Biochemistry Third Edition.* Hoboken, NJ: Wiley & Sons, Inc.

Witten, I. H., & Frank, E. (2005). *Data Mining (Second Edition).* Morgan Kaufmann Publishers.