A Panoptic Approach to Studying MicroRNA Expression in Breast Tumors

By

Priyanka Bhatt

A Dissertation Submitted to

Rutgers University

School of Health Related Professions

In partial fulfillment of the Requirements for the Degree of

Doctor of Philosophy

Department of Health Informatics

May 18, 2014

A Panoptic Approach to Studying MicroRNA Expression in Breast Tumors

By

Priyanka Bhatt

Approved by:

Professor Shankar Srinivasan, Chairman	**Date**
Professor Dinesh P. Mital, Co-Chairman	**Date**
Professor Masayuki Shibata, Co-Chairman	**Date**
Professor David Fenyo, Co-Chairman	**Date**

Abstract

Breast cancer is the second leading cause of cancer related death amongst women and has a major impact on the lives of those affected by it. Previous research has uncovered that miRNAs, a type non-coding RNA, play a key role in the onset and development of breast tumors. Genomics, transcriptomics and proteomics studies have shed light on the intricate involvement of miRNA in mediating breast cancer progression. These small molecules are involved in many biological processes and we have yet to understand all the levels of complexity. Studying miRNA can lead to more efficient methods of screening and treatment for breast tumors while providing us with other key information from the molecular level. As confirmed by previous studies, miRNAs are good candidates for diagnostic as well as prognostic markers.

This project takes a panoptic approach to studying miRNA in breast cancer genomics and proteomics and will strive to accomplish the following goals:

- **Visualize miRNA expression data through heatmaps and identify sub-types.**

  - We hypothesize that we will see a difference in expression of miRNAs across subtypes.

- **Use heatmaps to examine miRNA expression across race.**

- We hypothesize that miRNA expression will vary across race.

- **Perform PCA to deduce potential miRNA biomarkers for each subtype of breast cancer.**

  - We hypothesize that we will find at least one unique biomarker for each subtype of breast cancer.

- **Find the top 20 miRNA pairs with statistically significant correlation in expression.**

  - We hypothesize that these pairs will have miRNAs from the same family.

- **Study networks and pathways.**

  - We hypothesize that we will find networks within subtypes of breast cancer. We also feel that miRNAs will be involved in more than one disease pathway.

Data from The Cancer Genome Atlas (TCGA) data portal, which is an open source data portal open to the public, was utilized for the purposes of this project. The data is generated through miRNA-sequencing techniques with the aid of an Illumina Genome Analyzer. We used a combination of self-generated scripts in Perl and R and web based databases to filter, sort and analyze our data [1, 2].

On visual interpretation of our heatmaps we found different patterns of miRNA expression in the various sub-types of breast cancer. Patterns across race were not as significantly different upon visual interpretation but this may be attributed to data from a small cohort. Our PCA revealed potential biomarkers for each subtype of breast cancer. We suggest hsa-mir-127 and hsa-mir-379 as potential biomarkers of the basal subtype, hsa-mir-19a, hsa-mir126, hsa-mir-20a and hsa-mir-30a as potential biomarkers of the HER2 subtype, hsa-mir-222 as a potential biomarker of the Luminal A subtype and hsa-mir-152, hsa-mir-26b and hsa-mir-200c as potential biomarkers of the Luminal B subtype. By graphing these potential biomarkers across all subtypes we have visually confirmed our findings. We found 20 pairs of miRNAs with statistically significant correlation in expression and of those 6 were pairs that with miRNAs that are not related. We found miRNA-gene networks within two subtypes of breast cancer and generated a schematic to show the first layer of interaction. We used an online tool to generate another schematic that illuminates miRNA involvement in various pathways.

We feel that our study has led to some interesting findings. We encourage future studies to further validate our proposed biomarkers as well as improve our methods. Our study lacked data from normal tissue samples; the inclusion of which we feel would have yielded more thorough results. In conclusion, we believe that computational methods of data analysis in

studying miRNA expression data are truly powerful and the results of these methods will only get more accurate as more data is made available.

Acknowledgements

David Fenyo, Ph.D.

Shankar Srinivasan, Ph.D.

Table of Contents

List of Figures

List of Tables

Chapter I

INTRODUCTION

I. INTRODUCTION

1.1 Breast Cancer Statistics

An estimated 232,340 women and 2,240 men will be diagnosed with breast cancer in the year 2013 [3]. The Cancer Genome Atlas reports that there are 1.3 million cases and 450,000 breast cancer deaths worldwide [4]. It is one of the most common types of cancers diagnosed in women in the United States and it is the second leading cause of cancer related death in American women [5-8]. Although rare, breast cancer has also been diagnosed in males. While this disease is life altering for the patient it also has a major impact on the affected individual's family members. A study found that a significant number of family caregivers of patient's with breast cancer are clinically depressed [9]. To fully understand the gravity of a disease like breast cancer it is important to first understand the disease then examine its impact.

1.2 Understanding Breast Cancer

Cancer occurs when the cells in our body start to function in an abnormal manner. In cancer, the normal cell cycle is disrupted due to the malfunction of

signaling mechanisms that monitor and control cell growth, division and apoptosis. The result is a rapid uncontrolled growth and division of cells. For a cell to qualify as a malignant cancer cell it must have both uncontrolled growth and the ability to invade other cells [5]. Tumors that do not have the ability to invade other cells are known as benign and are in most cases not as harmful as malignant tumors. Genetic mutations inherited from parents may significantly increase the probability of developing cancer. Hereditary breast cancer results from mutations in the BRCA1 and BRCA2 genes that are passed on in an autosomal dominant manner [10]. Generally, most breast cancers are caused by somatic mutations that occur in breast cancer cells and are not inherited [11].

1.3 Breast Cancer Pathology

Breast cancer, as the name suggests, is a cancer that occurs in breast tissue. It usually originates in the nodules or ducts that aid in milk production and transport, respectively, but can also occur in stromal tissues [3, 12]. There are two major types of breast cancers: **non-invasive** (also known as in-situ) breast cancer does not spread beyond the epithelial cells of the nodules or ducts while **invasive** breast cancer is known to spread to other tissues and has a poorer prognosis [7, 13]. Non-invasive breast cancer has the potential to become invasive and therefore more harmful over a period of time [3, 7, 13]. Ductal Carcinoma In-situ (DCIS) and Lobular Carcinoma In-

situ (LICIS) are non-invasive breast cancers that occur in the ducts and lobules, respectively. DCIS is more common and is further classified into sub-types based on the presence or absence of certain hormone receptors [6, 14-16].

1.4 Molecular Sub-types

Table 1 lists the four main distinguishable subtypes of breast cancer and the hormone receptors that appear in each. Luminal A sub-type has the best prognosis of the four sub-types and is positive for the estrogen receptor (ER), positive or negative for the progesterone receptor (PR) and negative for Human Epidermal Growth Factor Receptor 2 (HER2). Luminal B is positive for ER, positive or negative for PR and positive also for HER2. Basal like or triple negative sub-type tumors are negative for all three receptors and are especially common in African American women [17]. HER2 type is relatively uncommon and is positive for HER2 and negative for both ER and PR [14, 16, 18, 19]. Hormone therapy can be used to block estrogen from binding to the ER positive cancer cells in order to restrict their growth and survival. The prognosis for tumors that are negative for hormone receptors is poorer when compared to tumors positive for the ER [15, 19].

Table 1 – Molecular Sub-Types of Breast Tumors

Subtype	ER	PR	HER2
Luminal A	+	+ or -	-
Luminal B	+	+ or -	+
Basal (Triple Negative)	-	-	-
HER2	-	-	+

Table 1: Molecular sub-types of tumors and their specific characteristics. Luminal tumors are positive for estrogen receptors and negative for HER2 – these tumors also have the best prognosis of all the sub-types. Triple negative tumors (basal like tumors) are negative for all receptors and generally have the worst prognosis.

1.5 Risk Factors

Many factors, environmental and genetic contribute to the development and growth of breast cancers. Research validated risk factors are age, geographical location, obesity, alcohol consumption, cigarette smoking, sedentary lifestyle and family history [20]. Younger age at menarche and

older age at the birth of the first child and/or menopause can also put women at a greater risk for developing breast cancer [21]. The probability of developing breast cancer is further increased in women who have a mutation present in one of the breast cancer related genes in addition to being at risk for non-hereditary reasons. A full list of the risk factors involved in breast cancer is listed in **table 2**.

Table 2 – Risk Factors of Breast Cancer

Table 2. A list of compiled major risk factors for breast cancer.

Risk Factors
Age
Environment
Family History
Genetics
Gender
Race and Ethnicity
Breast Tissue Density
Age at first menstrual period
Age at menopause
Age at first pregnancy
Diet
Alcohol Consumption
Body Weight
Socioeconomic Status

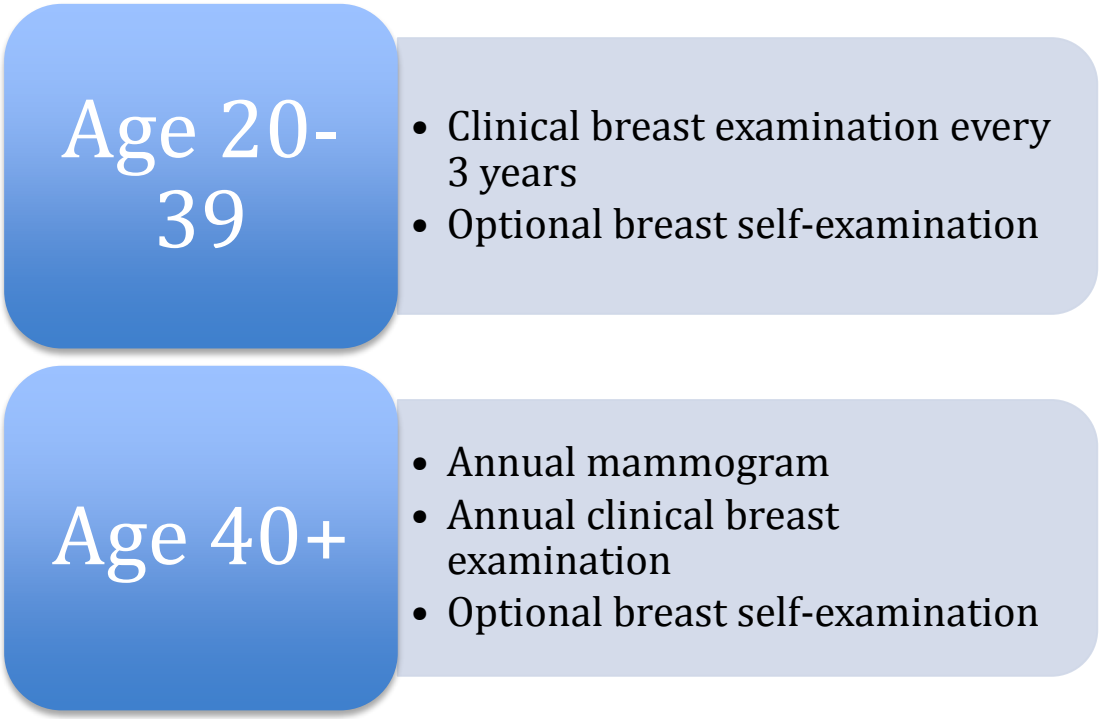
1.6 Diagnosis and Treatment

The diagnosis of breast cancer is generally made either by mammography screening or the discovery of signs and symptoms associated with the

disease. The detection of breast cancer through mammography screening has a more favorable prognosis because tumors found through screening are detected at an early stage, are smaller in size and are also treated early [15, 22, 23]. High mammographic density is correlated with an increased risk of breast cancer [24, 25]. Mammographic density has been found to decrease with age and this decrease contributes to a lower overall risk of breast cancer [26]. The American Cancer Society suggests certain guidelines to screen for breast cancer in the general population of women (**Table 3**). An annual mammogram and clinical breast examination is recommended for women over the age of 40.

Table 3 – Screening and early detection guidelines by the American Cancer Society

Table 3. The American Cancer Society suggests that women under the age of 40 years and over 20 years opt for a clinical breast examination every 3 years while women over 40 years of age should get an annual mammogram in addition to an annual clinical breast examination. A self-breast examination is optional for both age groups [8].



In addition to screening and symptoms, other diagnostic tests and biopsies are performed to confirm a diagnosis and evaluate the tumor [5].

Once a diagnosis is made the tumor is thoroughly assessed before proceeding with treatment options. Treatments for breast cancer include surgery, chemotherapy, hormonal therapy, biological therapy and radiation [20].

1.7 Statement of Problem

Breast cancer affects a significant number of lives each year, leaving a lasting psychological and sometimes physical impact. For many years scientists have studied the genomics of breast tumors to gain a better understanding of the pathology and to unravel a potential cure for the disease. Along the way, the discovery of key biological pathways has provided a probable solution for early detection and treatment [27]. The importance of early detection in a disease like breast cancer has been discussed for years but there is still a lack of an effective tool or biomarker that can replace older methods of screening. There is also an imminent need to uncover details of breast tumor progression at the molecular level via a new panoptic approach. A deeper understanding of pathways and networks can bring to light novel and more effectual treatment options.

1.8 Breast Cancer Genomics, Transcriptomics and Proteomics

Genomic data provides us with the sequence of our genetic information – allowing us to study the changes in this information that can

lead to disease states. Specific changes in the gene sequence can be used to identify different types and sub-types of breast cancer. A fairly recent high-throughput technique of sequencing data, called next-generation sequencing, has cut the time and cost of sequencing genetic information. Next generation sequencing's greatest advantage over the old method of using DNA microarrays is that it is fast. It also allows us to profile the nucleotide in DNA and RNA to further expound the regulation and expression of genes [28]. The steps to perform DNA and RNA-sequencing are almost identical. To perform next-gen RNA-sequencing (RNAseq), the RNA molecules must first be purified by 2-D Polyacrylamide Gel Electrophoresis (PAGE) and fragmented. After applying a primer to each end, a cDNA library is created using reverse transcriptase and PCR. A genome sequencer like the Iluumina Genome Analyzer is used to perform the rest of the experiment [28]. The results are short reads that may be aligned to a reference genome depending on the needs of the experiment. Sequencing of the genome and transcriptome (all RNA molecules) can provide us with crucial information like RNA expression levels in a cell at a certain time. This expression analysis of the transcriptome is a snapshot of information that among other things tells us which genes are active and which are inactive.

Although genomic and transcriptomic data is invaluable, it is incomplete without proteomics data. The proteome represents the full protein complement of a cell at any given time and accounts for posttranslational

modifications and other molecular changes [29]. Proteomics studies are important in profiling proteins and discovering protein-protein interactions [30]. The process of studying proteins begins with purifying the proteins. The purified proteins are then fragmented into peptides and in turn these peptides are identified via mass spectrometry. All mass spectrometers have an ionizer, mass analyzer and detector but different methods may be used [29, 31]. Mass spectrometers record the mass to charge ratio of peptides as well as the intensity – this data is represented on a graph [32]. The results from mass spectrometry experiments are then used to identify proteins. Tandem mass spectrometry is a type of mass spectrometry used for sequencing and identifying unknown proteins. A disadvantage to mass spectrometry is that it requires a sequence database for analysis but this is outweighed by its high-throughput capability [31].

Emerging methods in genomics and proteomics have changed the way we study genes and proteins. We owe our widespread understanding of cellular processes and pathways to potent tools like RNA-sequencing and mass spectrometry. Further advancements in these technologies will be beneficial to the medical community as they have the potential to structure and support applications of personalized medicine and drug discovery [28].

1.9 miRNA and Breast Cancer

microRNA (miRNA) are small non-coding RNA molecules that have the potential to be prospective biomarkers as well as targets for therapy in cancer [33-36]. Other non-coding RNAs have also been associated to cancer and **table 4** is a list of these RNAs including miRNAs and their functions.

Table 4 – Non-coding RNAs and their functions

Table 4 – This table has been adapted from Rossi et al and lists a few different types of non-coding miRNA and their functions [36]. The main function of miRNAs is either post-transcriptional repression and activation of genes.

Non-coding miRNA Name	Function
microRNA (miRNA)	**Post-transcriptional repression/activation**
Long Non-coding RNA (lncRNA)	**Regulate gene expression**
Small Interfering RNA (siRNA)	**Silence gene expression**
Small Modulatory RNA (smRNA)	**Transcriptional activation of neuronal differentiation**

miRNA are approximately 22 nucleotides in length and were first discovered in 1993. Primary-miRNA (pri-miRNA) are transcribed by RNA

polymerase II and exist as self-folding stem-loop structures. The Drosha-DGCR8 complex recognizes the loop structures and cleaves them out to form a pre-miRNA in the nucleus. The pre-miRNAs are then transported to the cytoplasm with help from the exportin-5 protein. In the cytoplasm, Dicer cuts the loop of the pre-miRNA to form 2 strands of miRNA of which one will bind to a RNA-induced Silencing Complex (RISC) and Argonaute protein to become the final mature miRNA (**Figure 1**) [27, 37, 38]. This cluster can then bind to mRNA sites for translational repression and degradation of the mRNA.

Figure 1 – Biogenesis of miRNA

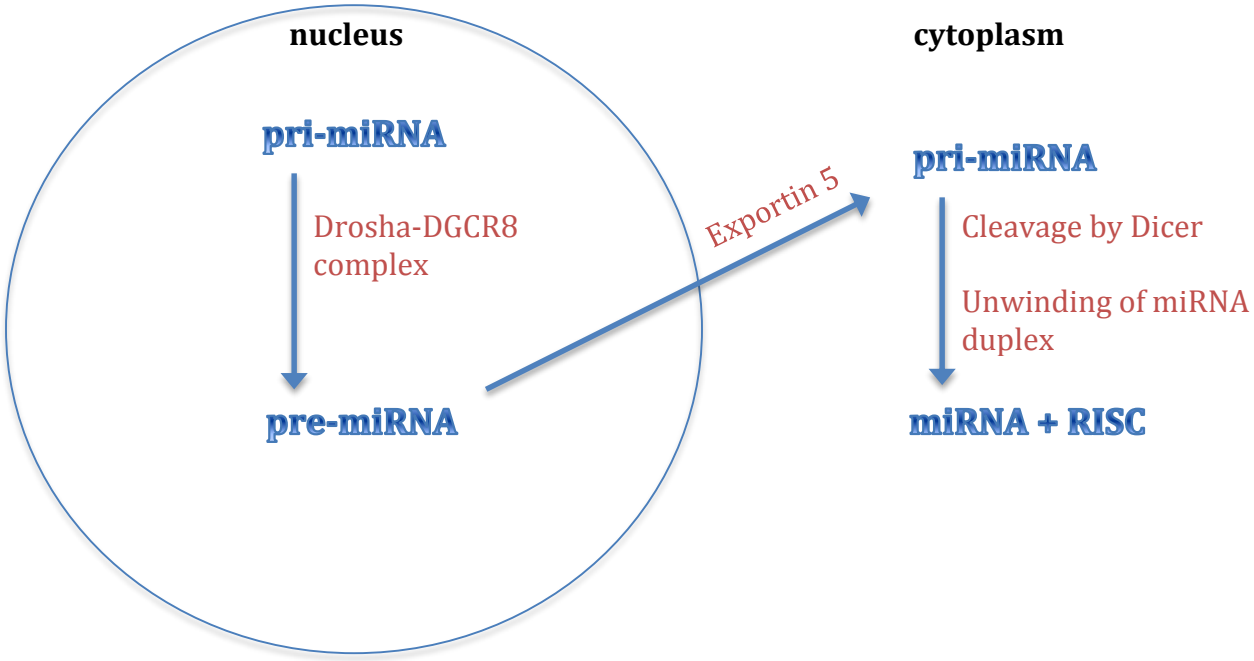


Figure 1. The biogenesis of miRNA begins in the nucleus where RNA-polymerase II transcribes primary miRNAs that are cleaved by the Drosha-DGCR8 complex to make pre-miRNA. These pre-miRNA are exported out to the cytoplasm by exportin 5 and are further processed by Dicer before becoming a part of the RNA-induced Silencing Complex (RISC) [39].

Deregulated expression or mutations of miRNA have been found to lead to the pathogenesis of diseases like breast cancer [35, 40]. They are also known to play an essential role in post-transcriptional modification and regulation of breast cancer related genes by binding to complementary sequences on the 3' un-translated region (UTR) of multiple target mRNA [38, 41, 42]. **Figure 2** is an illustration that highlights the interaction of miRNA with mature RNA molecules.

Figure 2 – Biological Interactions

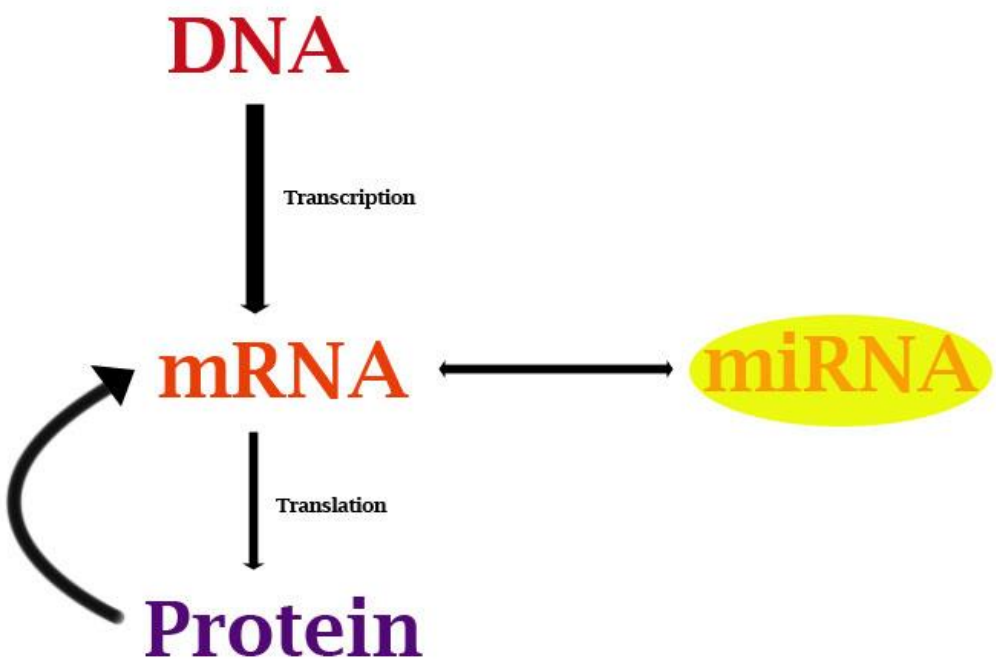


Figure 2. Protein and miRNA molecules bind to mRNA for post-translational modifications and regulation of genes. miRNA that binds to mRNA can cause it's degradation and indirectly stop translation and expression of target genes.

It is likely that most miRNAs are tumor suppressors based on observations that show an overall decrease in miRNA expression in cancers [43]. The actual mechanism of regulation by miRNAs is more complicated than our current understanding. A single miRNA can have multiple target mRNAs – many of these targets are yet unknown. In addition miRNAs have

shown that they sometimes operate in networks [27]. miRNAs are named sequentially and sometimes have an alphabet suffix denoting different forms of the same miRNA. The first 3 letters of a precursor miRNA designate the species (ex. hsa-miR-100) and are left out when writing the name of a mature miRNA (ex. miR-100) [44]. Until now, only a few of these miRNA networks have been discovered in certain diseases. Next-generation sequencing techniques like miRNA-seq are used to study miRNA expression and detect novel miRNAs. Several hundred miRNAs have been identified in the past and more will be discovered through de-novo sequencing methods [45].

1.10 The Cancer Genome Atlas

The Cancer Genome Atlas (TCGA) is a data portal that was initially set up by investment from the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI). The National Institute of Health (NIH) supports the efforts of TCGA by providing major resources. The TCGA network generates open source data for researchers who can download and analyze the data sets. As of the date of our download, TCGA has data for 957 total cases of breast invasive carcinoma out of which miRNA data is available for 894 cases, clinical data for 905 cases and methylation data for 931 cases. The participants donate a tumor tissue sample as well as a normal tissue sample, both of which are collected at the Tissue Source Sites (TSS) and sent to Biospecimen Core Resources (BCRs) for further processing. Although

currently, data on normal tissue in breast cancer is unavailable. The mission of TCGA is to coordinate an effort to accelerate our understanding of cancer at the molecular level using genome analysis technologies [46].

1.11 Goals and Objectives

The goal of this project is to utilize the recently aggregated TCGA data on breast tumors to comprehensively study the expression and other unique aspects of miRNAs in breast cancer. Available TCGA data makes the following objectives of the study possible:

1. ***Using TCGA data, study miRNA expression in breast tumors to understand whether miRNA profiling can be used to identify subtypes.***
    - Using the clinical data available for 905 cases of patients with breast cancer this project aims to identify the miRNAs that have the most variable expression profiles in the various sub-types of breast cancer. Previous studies have shown that miRNA profiling can be used to predict breast cancer sub-types – we hope to confirm this theory using new data from TCGA. The goal is to evaluate whether the difference in miRNA expression levels between sub-types is significant enough to use as a diagnostic tool. We plan to create a variety of heatmaps using R to visually see the difference in miRNA expression between sub-types [1]. The hypothesis is that miRNA

expression levels will be a good diagnostic indicator of breast cancer sub-types.

2. ***In order to define potential biomarkers for the subtypes of breast cancer we hope to identify miRNAs that are most differentially expressed in each subtype using Principal Components Analysis (PCA).***
   - We will perform PCA on miRNA expression data to decipher which miRNA or miRNAs account for the greatest variability in each subtype. We hope to find a different miRNA biomarker for each subtype of breast cancer. In order to ensure that the potential biomarkers are unique we will do a PCA on all samples to eliminate any miRNAs that are up or down regulated in the disease and are not subtype specific.

3. ***Compare miRNA expression across different races***
   - Previous studies have shown that some miRNAs are expressed differently in individuals of varying races [47-49]. These differences in expression may be related to polymorphisms that are characteristic to a particular race [50]. The plan is to generate heatmaps with supervised clustering by race and to look for patterns that suggest variable expression across race. The

prediction is that miRNA expression will vary across race. This should be visible in the clusters/annotations of the heatmaps we generate.

4. ***Search for miRNA-gene networks in breast cancer subtypes***

   - To elucidate miRNA networks in breast cancer subtypes this study will use web tools like miRBase and miRWalk to first find gene targets of miRNAs that are potential biomarkers of each subtype according to our PCA analysis [44, 51]. Next we will elucidate which miRNAs from each subtype have common gene targets. Previous studies have worked on miRNA networks in breast cancer and uncovered some intriguing interactions but our approach is different and much simpler [52, 53]. The hypothesis is that more than one miRNA will have the same gene target.

5. ***Find the 20 most highly correlating miRNA pairs and verify whether the paired miRNAs are related.***

   - miRNAs that are from the same family or related will have similar expression profiles as they usually belong to a pre-cursor that is located nearby on the same chromosome. To verify whether this is true for the TCGA dataset we will use a function in R to identify 20 pairs of miRNAs that have a statistically significant correlation

between their expression profiles in breast tumors. We will use a miRNA database to confirm that the pairs of miRNAs are related.

The following 3 objectives will require new data to be generated:

6. ***Study the expression of miRNAs in breast tumors over a period of time***

   - miRNAseq data from samples of different stages over time will be required for this analysis. This data is currently unavailable via the TCGA website. This type of study would be essential to miRNA analysis as it can reveal a lot about the role of miRNAs in tumor growth and development. We hypothesize that the expression profiles of miRNAs will vary at the different stages of the breast tumor and will also give us an idea of which miRNAs are good prognostic markers for the disease.

7. ***Examine drug induced changes in the expression of miRNA***

   - Data required for this goal is likely to be generated in the near future. To the best of our knowledge, currently no data is available on miRNA expression with drug use. If made available we would like to use that data to evaluate if and how drugs influence the expression of miRNAs in breast cancer. Studying any drug induced changes in miRNA expression will aid our understanding of miRNA interactions with

exogenous molecules and help us evaluate possible targets of therapy. The hypothesis is that drug use will alter the expression of miRNAs in breast tumors.

8. ***Compare the miRNA expression profiles between normal and breast tissue samples from the same patient.***
    - Currently the TCGA data portal only has miRNA sequencing data for breast tumor tissue samples. It would be interesting to compare this data to the miRNA expression profiles in normal tissue from the same individuals. We would expect that certain miRNAs will be up or down regulated in the breast tissue versus the normal tissue.

Chapter II

LITERATURE REVIEW

II. LITERATURE REVIEW

2.1 Current Breast Cancer Research

In reviewing the literature on breast cancer it was found that research in this field is multi-faceted and multi-dimensional. Studies that try to understand all aspects of breast cancer can achieve a comprehensive understanding of the disease. A retrospective study that collected discharge data on women with breast cancer from 2002-2009 found that there were racial and ethnic disparities in stage, co-morbidities, surgical treatment allocations and outcomes among the cohort [54]. While investigating the mortality risk between white and black women diagnosed with invasive breast cancer, a group of researchers discovered that the difference showed mainly in older women diagnosed with luminal A/p53- breast cancer [55]. An April 2013 study conducted at Stanford University School of Medicine identified novel gene fusions in breast cancer (CLTC/VMP1) as well as other cancers [56]. Studies like the ones mentioned above are contributing greatly to the way we currently understand and treat cancer. The future of breast cancer research will greatly benefit from the foundation of data that these studies have created.

2.2 miRNA in Other Cancers

The notion that miRNAs are involved in the molecular development of cancer has attracted attention from many researchers, especially those in the proteomics and bioinformatics sector. Furthermore, there has been a surge in the studies discussing miRNAs and bioinformatics over the last two years **[36].**

The role of miRNAs in ovarian cancer is similar to their role in breast cancer. Studying the dysregulation of miRNAs in ovarian cancer has led scientists to the discovery of new pathways in ovarian tumorigenesis thus paving the way for potential tools to battle the disease [57]. A study found that tumorigenesis in bladder cancer was linked with low levels of miR-34a in bladder tissue [58]. Another study closely investigated the differential expression of miRNAs in lung cancer versus normal lung tissue and found that has-miR-339-5p was especially suitable as a biomarker [59]. In order to validate a non-invasive early detection screening method for pancreatic cancer, a group of researchers examined the expression of 3 miRNAs (miR-21, miR-155 and miR-200) in blood and pancreatic tissue samples and found that they were highly expressed in the presence of pancreatic cancer – making them good candidates for biomarkers [60]. In a similar study miR-193a-3p, miR-23a and miR-338-5p were named as potential blood based

biomarkers for early detection of colorectal cancer [61]. Another study found that some miRNAs had a statistically significant difference in expression in cervical cancer [62]. An additional study confirmed that the same was true for prostate tumors which means that miRNAs may be used as prognostic and diagnostic evaluators in prostate cancer as well [63]. All of these studies have one thing in common – they highlight the importance of studying miRNAs in cancers. It is also true that by studying miRNAs, researchers are racing with time to develop new and effective ways to treat individuals affected by cancer. In the next few years we may witness the clinical use of multi targeted miRNA based anti-cancer therapies [64].

2.3 Personalized medicine and miRNA

Numerous studies have been published in the recent past on the subject of miRNAs and their function in breast cancer. In our review of the literature, We found that studies canvassing the function of miRNAs in breast cancer are recent. A large-scale study that was published in nature profiles miRNA expressions of 1,302 breast tumors to establish a framework of information for future studies [65]. In another study, researchers examined the expression of miRNAs in a rare type of breast cancer known as Pure Mucinous Breast Carcinoma (PMBC) and found that miR-143 and miR-224-5p were down regulated [66]. A research team led by Jincai Wang found that increased expression of miR-9 and miR-200c were predictors for lymph node

metastasis thus paving the way for those miRNAs to be potential diagnostic markers [67]. Results from pioneering studies like these contribute directly to personalized clinical medicine, which is promptly going to replace the one-size fits all approach in practice today. The benefit to taking a personalized approach to medicine is that it accounts for genetic variations that are substantially responsible for disease characteristics and response to therapeutics in each individual patient. Advances in the proteomics field have made it easier for scientists to study these variations in our genetic makeup, thus making it possible for medical tools and treatments to be designed for individuals and not groups [36]. New sequencing techniques have given us an unbelievable depth in studying RNA and peptide molecules in a particular cell at a specific time. **Table 5** is a list of databases and software used in the study and analysis of miRNAs. There are other miRNA related software tools and databases that also aid in the research process [37].

Table 5 – Databases and softwares used in miRNA analysis

Table 5. A list of online databases and software tools available to researchers for the purpose of studying and analyzing miRNAs.

miRNA Database/Tools	URL	Description
miRBase	http://www.mirbase.org/	Provides detailed information and target search for miRNAs.
miRPath	http://diana.imis.athena-innovation.gr/DianaTools/index.php?r=mirpath/index	Pathway analysis tool for miRNAs that can also be used to study miRNA interactions that have been validated experimentally.
TargetScan	http://www.targetscan.org/	A tool to find predicted and validated targets of miRNAs.
miRCancer	http://mircancer.ecu.edu/	A database that provides information on miRNA cancer association.
miRTrail	http://mirtrail.bioinf.uni-sb.de/mirtrail.php	Analyze relationship between miRNAs and mRNAs.

2.4 Expression profiling miRNA in breast cancer

A study confirmed that the difference in expression of certain miRNAs in breast cancer tissue versus normal tissue is significant. The most deregulated miRNAs were miR-125b, miR-145, miR-21 and miR-155 [33]. Other studies suggest that miR-21 and miR-155 are up regulated in breast

cancer and in contrast miR-106, miR-125b and miR-145 are down regulated [45, 68]. The interesting thing is that the expression of some up-regulated or down-regulated miRNAs can be correlated with the different molecular sub-types of breast tumors [34, 38, 69, 70]. A group of researchers validated the role of miRNAs in disease classification by using Artificial Neural Networks (ANNs) to identify miRNA signatures associated with estrogen (miR-342, miR-299, miR-217, miR-190, miR-135b, miR-218), progesterone (miR-520g, miR-377, miR-527-518a, miR-520f-520c) and HER2 (miR-520d, miR-181c, miR-302c, miR-376b, miR-30e) receptor status [70]. Not surprisingly, miRNAs are also believed to be involved in the migration, invasion and metastasis of a tumor [71]. According to an expression analysis study conducted in 2011, migration and metastasis of breast tumors was attributed to increased expression of miR-423 [69]. MiRNAs can be either tumor suppressing (miR-30a, miR-31, miR-34a, miR-125s, miR-200s, miR-203, miR-205, miR-206, miR-342) or oncogenic (miR-10b, miR-21, miR-135a, miR-155, miR-221, miR-222, miR-224, miR-373, miR-520c). Research studies have verified that miRNAs have overlapping functions – they can be involved in migration, invasion and metastasis simultaneously [42].

2.5 miRNAs as biomarkers

A majority of the research on breast cancer has focused on establishing a biomarker for the disease to help with making an early

diagnosis [72]. The primary reason for this is that early detection has been a major contributor to the decline in breast cancer death rates in the last decade [45]. Current screening methods (mammography and breast examination) are ineffective for early detection and are many times painful to the patients [3]. To procure a solution, miRNAs are being investigated as future biomarkers for breast cancer. A study published in the Journal of Translational Medicine, investigating miRNAs circulating in blood, concluded that certain miRNAs could potentially serve as blood-based biomarkers [68]. Unlike circulating mRNA, circulating miRNA is stable at room temperature and can withstand freeze-thaw cycles[73]. A different study measured the serum concentrations of miR-10b, miR-17, miR-34a, miR-93, miR-155 and miR-373 in breast cancer patients and healthy women. The researchers concluded that the serum concentrations of miR-34a, miR-93, miR-373 were significantly different in healthy women versus women with breast cancer and these findings highlight the diagnostic value of circulating miRNA [74]. In one study, researchers found an insignificant difference in the serum expression levels of certain miRNAs, but this may be attributed to the small cohort of women in the study [75]. There are a few challenges in using circulating miRNAs as biomarkers – one being the extraction of these small molecules from blood [73]. As biomarkers, miRNAs would have to: 1. Differentiate non-disease from disease tissue, 2. Determine correct prognostic group for patients and 3. Monitor response to therapeutics [45]. There are many

benefits to having efficient biomarkers for breast cancer and even breast cancer subtypes, which is why scientists are meticulously examining miRNAs in order to learn everything about their interactions with breast cancer genes [39].

2.6 miRNAs as targets for therapy

Exogenous analogs of tumor suppressing miRNAs injected into mouse models have shown positive results, thus verifying that miRNAs can be targets of therapy in addition to being biomarkers of disease [42, 76]. A paper published in nature this year found that the down regulation of miR-140 in normal breast epithelium leads to cancer stem cell formation and eventually DCIS. The same study also found that two most significantly activated stem cell factors SOX9 and ALDH1 are direct targets of miR-140 and that replenishing miR-140 in-vivo led to reduced tumor growth [77]. In a similar study, miR-221 was found to promote tumorigenesis in triple negative breast cancer; the scientists noticed inhibition of tumor growth in-vitro and in-vivo when miR-221 was knocked down [78]. Angiogenesis promoting microRNA, miR-27a, is another potential target for cancer therapy. miR-27a was found to increase tumor growth, metastasis and angiogenesis in an animal model [79]. Angiogenesis or the formation of new blood vessels provides nutrition and oxygen for tumor growth and promotes tumor metastasis. Increased angiogenesis is a signature characteristic of cancer and it is therefore

necessary to develop anti-angiogenesis therapies [80]. Results from the above mentioned studies provide hope for the clinical use of mimic miRNAs as therapeutic agents in human breast tumors and remind us that the distance between bench and bedside is constantly decreasing.

2.7 miRNAs as Prognostic Markers

In China, researchers found that miR-206 would be a good candidate for a novel prognostic indicator in breast cancer patients as low levels of miR-206 were found unfavorable for overall survival [81]. A different study found that miR-125b would be a good prognostic response marker for cancer therapy [82]. A drug interaction study, found that miRNAs play a role in regulating chemoresistance which is a major hurdle in the treatment of cancer [83]. Advanced studies on miRNA and drug interactions will greatly increase the depth of our understanding of miRNA pathways. The use of miRNAs as prognostic markers will greatly aid clinicians in treating breast cancer patients.

2.8 Networks and Interactions of miRNAs

miRNA research specific to breast cancer has come a long way in a short span of time but there is still much scope to learn and explore. For example, more studies are needed to ascertain miRNA networks in breast

cancer because current research on this topic is lacking [27]. In one particular study, researchers have tried to use computational methods to predict interactions between miRNAs but this work is preliminary [41]. A recent study conducted in Taiwan investigated miRNA regulated Protein Interaction Networks (PINS) in breast cancer. The researchers paired miRNAs with target mRNAs using expression profiles and target prediction databases to elucidate interactions and functions [53]. Complex networks of miRNAs have already been studied in other types of cancer and further studies will be required to expose miRNA networks specifically found in breast cancers [52].

2.9 Small Nucleotide Polymorphisms (SNPs) and the risk of breast cancer

A recent study examining genetic variants or Single Nucleotide Polymorphisms (SNPs) of miRNAs concluded that some SNPs can be linked to increased or decreased survival [50]. While inspecting the effects of the rs6505162 SNP, A > C polymorphism of miR-423 a group of researchers found that this polymorphism was correlated with a lower risk of breast cancer [84]. In a similar study another polymorphism (rs1161494913) was evaluated and the researchers concluded that the CC polymorphism was linked to a reduced risk of breast cancer while occurrence of the T allele was significantly coupled to an increased risk of breast cancer [85]. A study from 2012 suggests that a SNP in the binding site of miRNA of gene IL23R may lead to

a greater risk of cancer [86]. Another study evaluated common genetic variations in the biogenesis pathway of miRNA and found that these gene variants did not significantly impact the risk of breast cancer in Asian women [87].

CHAPTER III

METHODS

III. METHODS

3.1 Tumor Samples and Data Source

The data used in this study was downloaded from the TCGA web data portal. While the TCGA project is funded by the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI), the data is available to the public for the purpose of scientific research. We downloaded miRNA sequencing data and gene expression data for all samples available at the time of download. Although the TCGA data does not outline a specific miRNA sequencing method used to generate their data – we have outlined a general procedure for miRNA sequencing. **Figure 3** is a schematic depicting a detailed method of miRNAseq using the Illumina sequencing method.

Figure 3 – miRNAseq Preperation

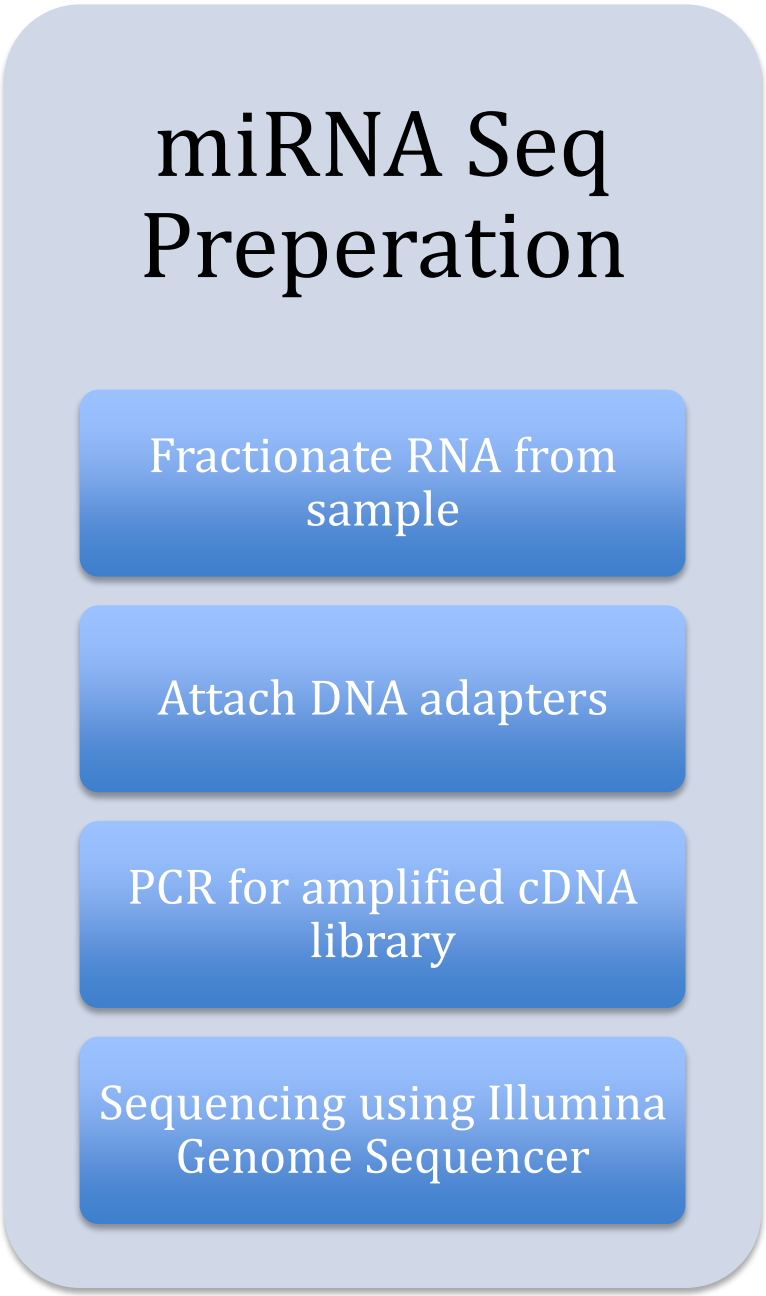


Figure 3. Step by step RNAseq procedure of the Illumina sequencing method. Isolated samples of RNA are fractionated and adaptors are applied before sequencing is performed in an Illumina Sequencer.

The first step of the procedure is to isolate tumor samples using a mirVana isolating kit. In the next step the isolated RNA from both samples is fractionated using denaturing PAGE before DNA adaptors are applied to both ends of the RNA. The purpose of the DNA adaptors is to act as primer binding sites during reverse transcription and PCR that are performed to create an amplified cDNA library of the samples [28]. The last step of the sequencing method is the actual sequencing which takes place in an IIlumina Sequencer [88]. The resulting sequence reads from the RNAseq can then be assembled by TopHat (a fast splice junction mapper for RNAseq reads) [89].

3.2 Data Processing

A Perl script was used to aggregate miRNA expression data from the TCGA files that were downloaded [2]. A similar Perl script was used for collecting the gene expression data. The first set of data generated has miRNAs with no missing values. The second set of data selected for all miRNAs with less than 80% missing values. We used another Perl script to read through clinical files and extract information on subtype, race, and receptor status. We merged the clinical data with the miRNA expression data sets in order to sort and divide the data by the subtype of breast cancer.

3.3 Studying Expression Data

R statistical programming was used to further process and study the expression data generated using the Perl scripts [1]. The data was normalized to correct for background error and account for any sample wise differences across our miRNA data set. We first took log2 of the data and then performed a miRNA wise (row-wise) z-score normalization. The z-score normalization sets the row mean to 0 and the standard deviation to 1. It is calculated by subtracting the mean of the row from each individual value then dividing the result by the standard deviation of the row. We did a sample-wise and miRNA-wise clustering of the data to group our data by similarity. The unsupervised clustering method we chose is Pearson correlation. We also scaled our data by rows (miRNAs). Next we generated a variety of heatmaps to analyze the expression profiles of miRNAs comprehensively. In order to clearly annotate the data with a color side bar we chose to use the R Pheatmaps package [90].

We generated heatmaps for 3 sets of data. Our first set of data included miRNAs with no missing data, the second set included miRNAs with less than 80% missing values and the third data set included only selected miRNAs that were verified via miRCancer (a miRNA cancer association database) to play a role in breast cancer [91]. For each data set we generated eight heatmaps. The first heatmap included all of the data, the next five were by subtype (basal, HER2, luminal A, Luminal B) and the final two were by

Coefficient of Variance (CV) cut-offs (CV>60, CV>100). We used the R package impute to deal with missing data – this package imputes missing expression data by nearest neighbor averaging [92]. It is necessary to impute the missing values in order to cluster all the heatmaps by pearson correlation. The heatmaps are annotated by subtype, race and receptor status where possible.

Figure 4 – Heatmaps Generated in R

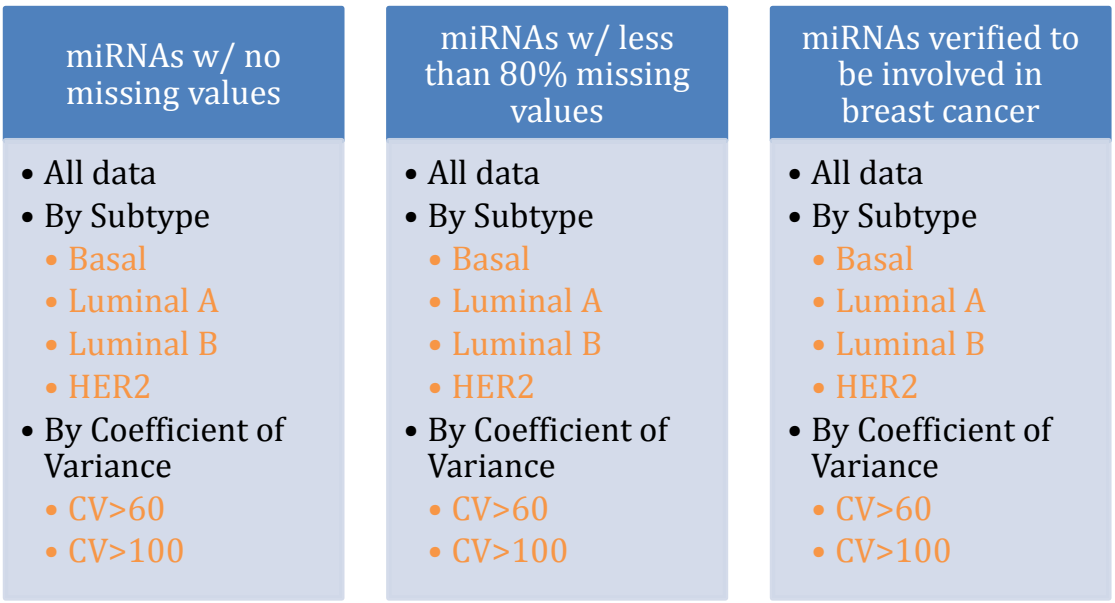


Figure 4. A list of the heatmaps that were created for each data set.

3.4 Principle Components Analysis

Before performing Principal Components Analysis (PCA) we filtered and prepared our data. The miRNA cohort for this analysis included only those

miRNAs that play a role in breast cancer. In addition, we chose only those samples for which we have gene expression data, as this would enable us to do an analysis of gene expression in the future. A list of miRNAs that are associated to breast cancer was downloaded from the miRCancer association database. An R script was used to select for miRNAs from our data that were on the downloaded list. We further filtered our miRNA cohort to exclude all miRNAs with missing values to reduce the amount of error and improve the accuracy of the results. The data was z-score normalized in R prior to performing PCA. To investigate potential biomarkers for each subtype of breast cancer we performed PCA on the selected cohort as well as the cohort divided by subtype [93]. Microsoft Excel was used to graph the results of the PCA. The miRNAs that only occurred in a particular subtype's PCA analysis uniquely were chosen as potential biomarkers of that subtype. In order to visually verify the potential biomarkers, we used excel to generate simple graphs that displayed the expression of potential biomarkers from each subtype, across all subtypes.

3.5 miRNA Correlations

To find the top 20 most highly correlated pairs of miRNAs we used an R function that would give us the pairs of highly correlated miRNAs along with the linear correlation coefficient [94]. To be statistically significant the correlation coefficient must be above 0.05. The data used for the correlation

analysis had no missing values and did not select for specific miRNAs. We then used an online miRNA database to verify whether the 20 paired miRNAs belong to the same family.

3.6 Networks, Pathways and Targets

To uncover interactions of the potential miRNA biomarkers for each subtype found via PCA analysis we studied the relationship between these specific miRNAs and their common verified targets. The online miRNA database miRWalk was used to find the verified targets of the selected miRNAs [51]. Only the common targets amongst the miRNAs were used. If a subtype did not have more than one potential biomarker we did not explore potential networks. For subtypes with miRNA networks we used web based databases to establish whether or not the miRNAs involved are related.

We used the same cohort of miRNAs that were used in the PCA analysis to generate a heatmap that indicates what other pathways our cohort of miRNAs are involved in. The web application we used for this analysis is part of the Diana Tools software and is called miR-Path [95]. We chose to include the 5’ and 3’ versions of our miRNA cohort for this analysis.

Chapter IV

RESULTS

IV. RESULTS

4.1 Data Summary and Statistics

Table 1 lists the major subtypes of breast cancer and describes the characteristics of each. The Luminal subtypes of breast cancer are positive for the ER and PR receptor while the Basal and HER2 subtypes are negative for both. **Figure 2** is a schematic showcasing the simplified interaction of miRNA, proteins and mRNA. miRNA and protein bind to mRNA causing post translational modification and regulation of genes.

Table 5 summarizes the TCGA data statistics. The unfiltered TCGA cohort comprises of expression data for 1046 miRNA from 931 tissue samples. When filtered for miRNA with <80% missing values the cohort has expression data for 537 miRNAs and 931 tissue samples. Further filtering the data to remove all missing values gives expression data for 183 miRNAs and 931 tissue samples. We filtered the data with <80% missing values and no missing values by selecting for miRNAs associated with breast cancer. There are 55 miRNAs associated with breast cancer that had <80% missing values for 931 tissue samples whereas there are 41 miRNAs associated with breast

cancer without any missing values for 931 tissue samples. For the data with 41 breast cancer associated miRNAs that exclude missing data we also filtered for only those samples with available gene expression data – this sample wise filtering gave us 512 tissue samples. We found that of the unfiltered 931 tissue samples 137 are basal, 45 are HER2, 529 are Luminal A, 132 are Luminal B and 88 are uncategorized.

Table 6 – TCGA Data Summary Statistics

Table 6. This is a table with summary statistics from the TCGA data that we downloaded and filtered for our purposes.

Data Type	# of miRNA	# of Samples
Raw	1046	931
<80% missing values	537	931
No missing values	183	931
Breast cancer associated miRNA (<80% missing values)	55	931
Breast cancer associated miRNA (no missing values)	41	931
Breast cancer associated miRNA and samples with gene expression data (no missing values)	41	512

4.2 Heatmaps – All Samples

Figure 5 is a heatmap generated by miRNA expression data that is filtered to exclude all missing values. One can visually see patterns in the expression of miRNAs. The upper right quadrant shows some miRNAs closer to the top as highly expressed (red) and the ones below as having low

expression (green). There is a large cluster on **figure 5** composed mostly by the Basal subtype where we note that the majority of the miRNAs are greatly expressed. In the large cluster of mostly Luminal A subtype samples we note that the miRNAs are not as highly expressed. The unsupervised clustering does not cleanly separate the samples by subtype or ER status – although some patterns can be seen.

Figure 5 – miRNA Expression (No Missing Values)

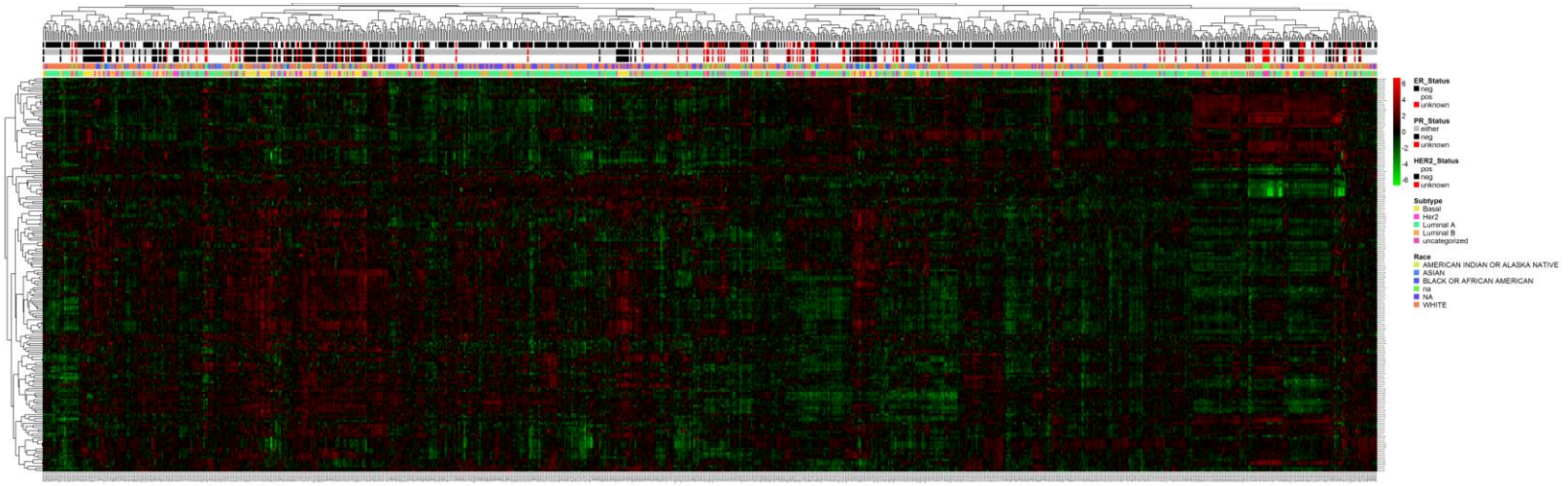


Figure 5. A heatmap of miRNA expression in breast cancer tissue samples that excludes all missing values. The heatmap is annotated by race, subtype, ER status, PR status and HER2 status.

Figure 6 is a heatmap of miRNA expression that includes miRNAs with <80% missing values. We noted a small cluster of the basal subtype near the center of the heatmap where miRNAs are greatly expressed. This is also a patch of miRNAs to the right of the heatmap that are highly expressed. In addition the lower left quadrant of the heatmap shows a group of highly expressed miRNAs. This pattern falls primarily under the Luminal A subtype of samples. The dark regions on this heatmap correspond to miRNAs that are neither greatly expressed nor lowly expressed. The clustering of the samples by Pearson correlation does not clearly separate the samples into known subtypes but we do however see multiple smaller clusters of the same subtype.

Figure 6 – miRNA Expression (<80% Missing Values)

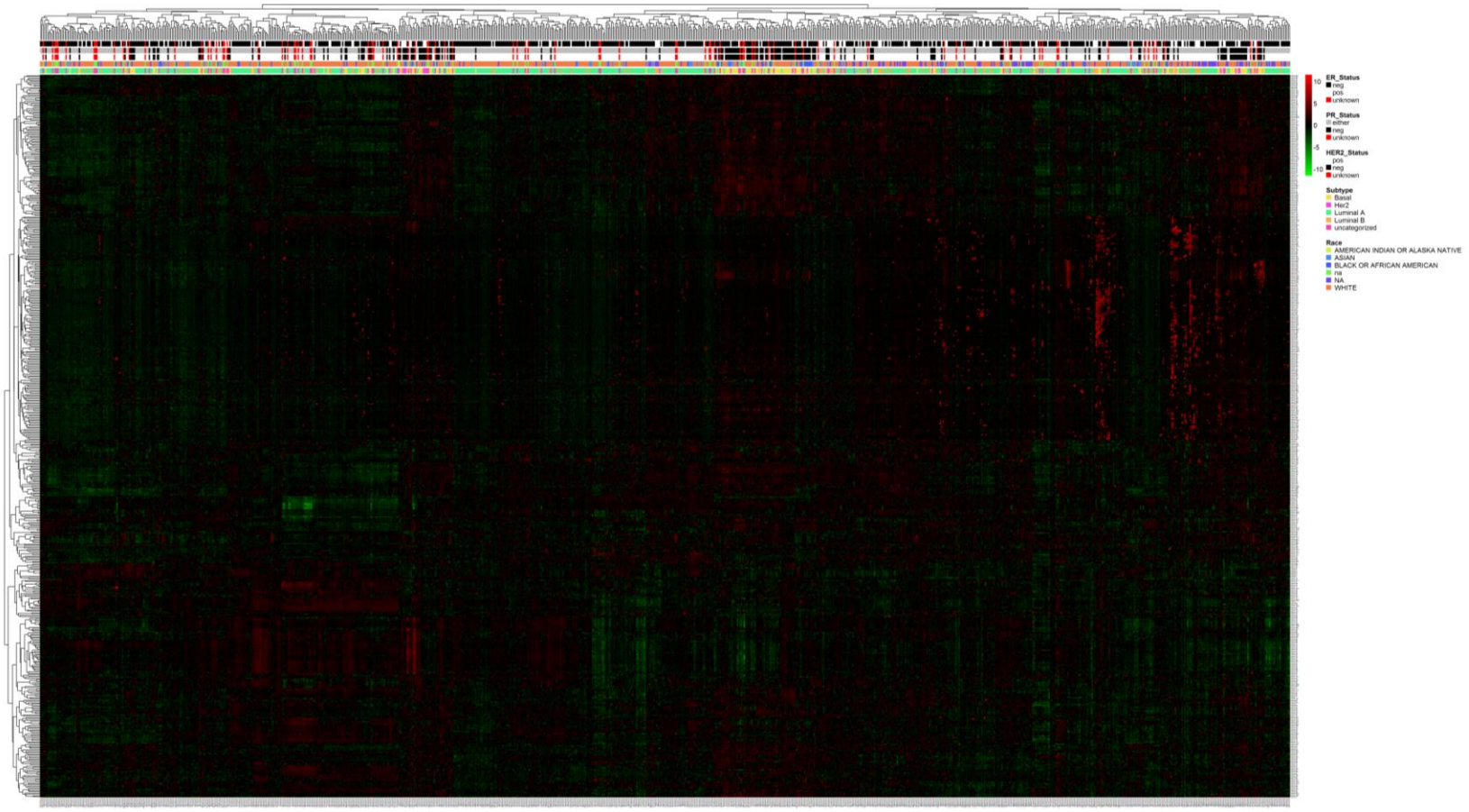


Figure 6. This figure is a heatmap of miRNA expression in breast cancer tissue samples. The data includes all miRNA with <80% missing values. Clustering is done by Pearson correlation and the data has been scaled by row.

Figure 7 is a heatmap of miRNA expression that filters for miRNAs that are associated with breast cancer and have <80% missing values. The notably expressed miRNAs form a group in the center of the heatmap. A remarkably low expression of miRNAs is seen towards the top right of the heatmap – this pattern falls primarily under samples that belong to the Luminal A subtype. There are 2 major and a few smaller clusters of the Basal or Triple Negative subtype that can be seen in the annotation.

Figure 7 – miRNA Expression (Select miRNAs and <80% Missing Values)

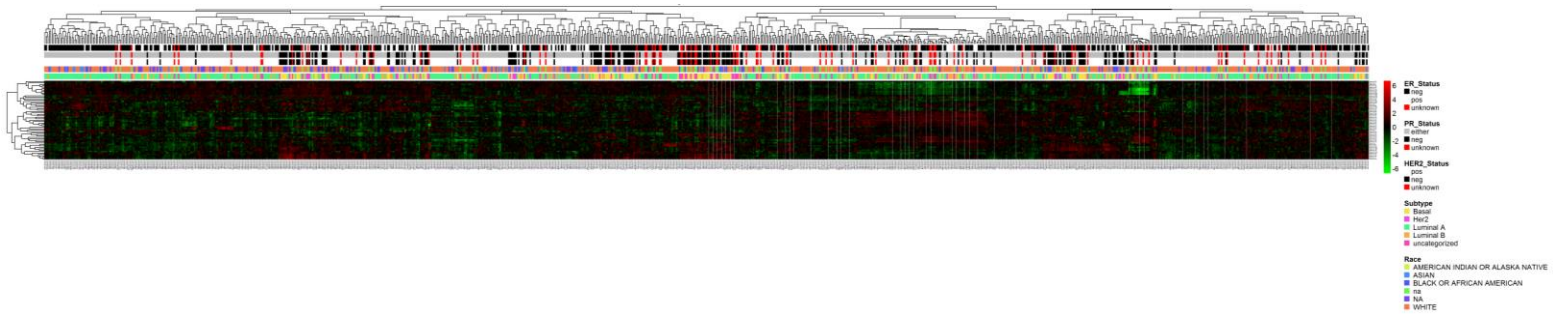


Figure 7. Heatmap of miRNA expression for miRNAs that are associated with breast cancer. The data allows miRNAs with <80% missing values.

4.3 Heatmaps of the Basal Subtype

The following 3 figures will show heatmaps with expression data only for the basal samples. **Figure 8**, for example excludes all miRNAs with missing values and selects for samples from the Basal subtype of breast cancer. In the basal subtype, the ER, PR and HER2 receptor status are all negative. The clustering is done by Pearson correlation and the samples are annotated by race. The notable thing about **figure 8** is the large cluster of red to the left that corresponds to miRNAs that are highly expressed. In the center is a slightly smaller group of miRNAs that are under-expressed and represented by a patch of green.

Figure 8 – miRNA Expression in Basal Tumors (No Missing Values)

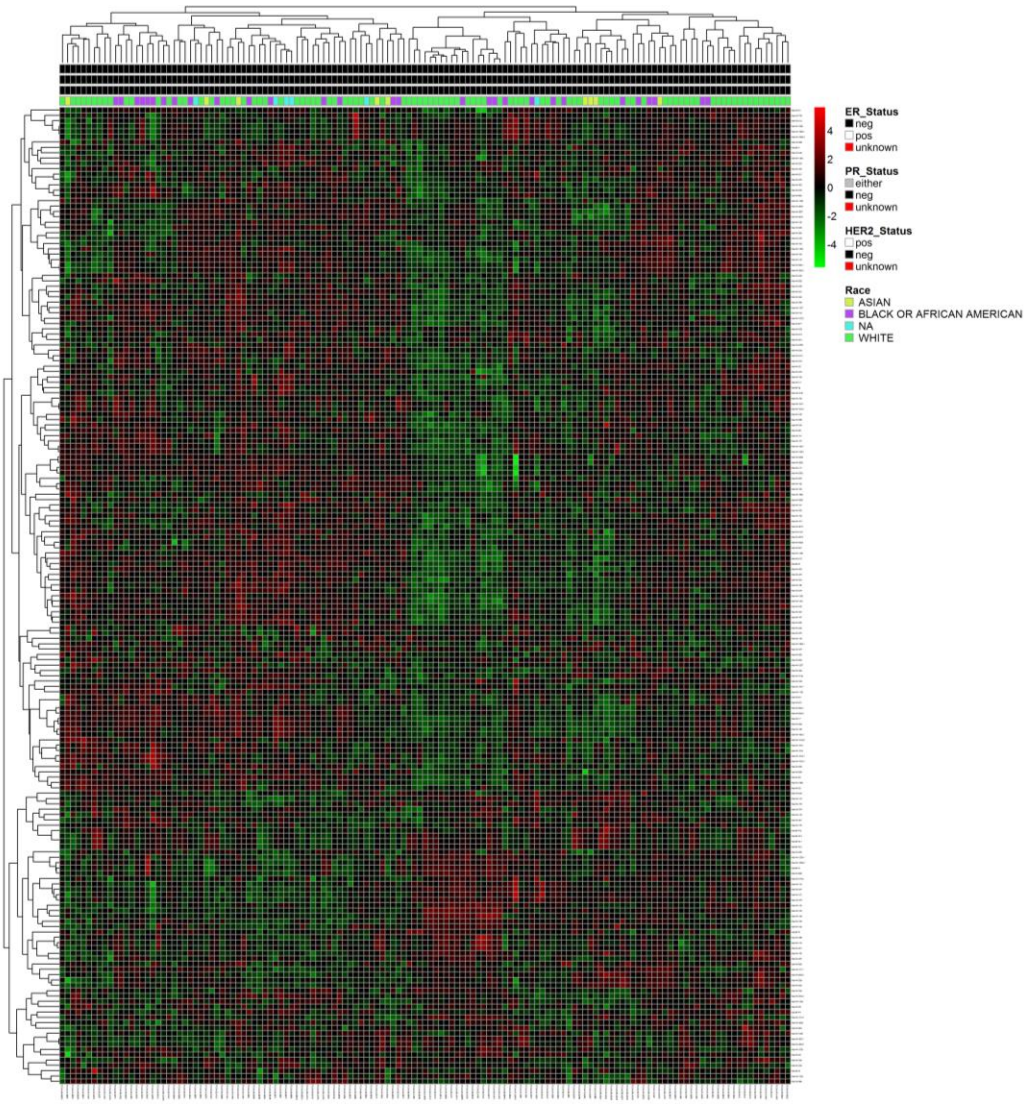


Figure 8. This heatmap shows miRNA expression in the basal subtype and excludes all missing data. There is a large cluster of red to the left of the heatmap that represents highly expressed miRNAs in that region.

In **figure 9** below the heatmap data includes miRNAs with <80% missing values but the samples all belong to the basal subtype. The patterns in this heatmap are similar to those found in **figure 8** but not as remarkable. The miRNAs with high expression are more towards the left of the heatmap.

Figure 9 – miRNA Expression in Basal Tumors (<80% Missing Values)

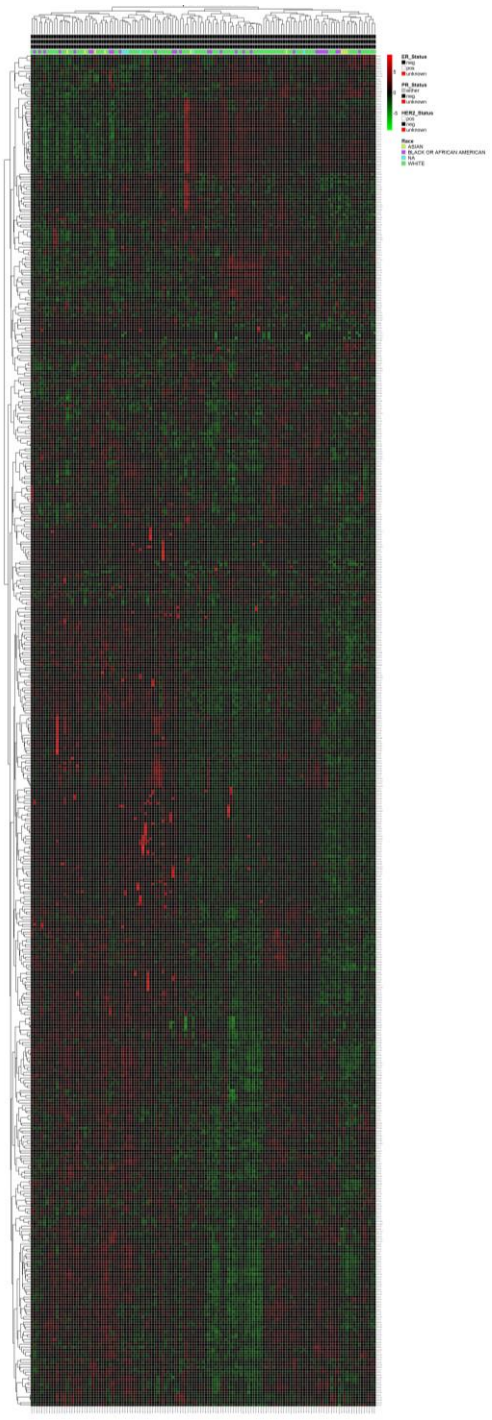


Figure 9. miRNA expression in basal tumors with miRNAs that have <80% missing values.

The third heatmap for the basal subtype (**Figure 10**) shows the expression of select miRNAs in basal tumors. There is an interesting pattern of expression just off center where the miRNAs on top are all greatly expressed while the miRNAs below them are under-expressed. There are four main sample wise clusters of the samples.

Figure 10 – miRNA Expression in Basal Tumors (select miRNAs)

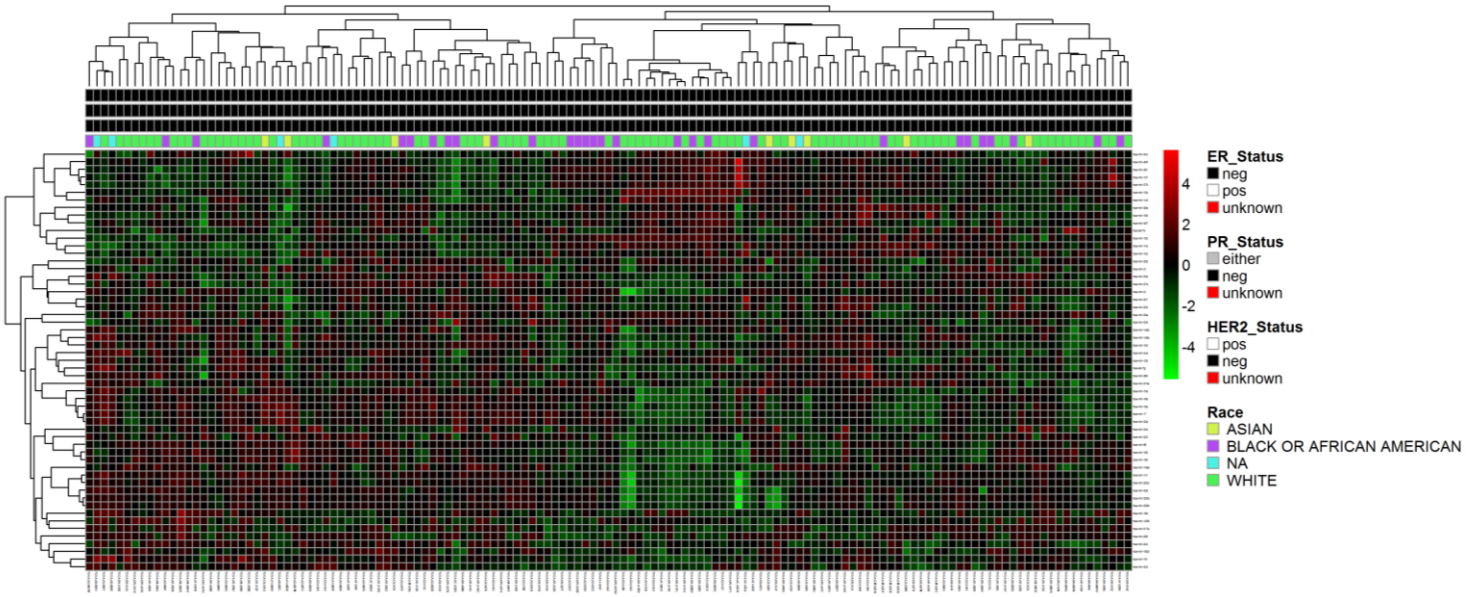


Figure 10. This heatmap of miRNA expression in basal tumors includes only miRNAs that are associated to breast cancer.

4.4 Heatmaps – HER2 Tumors

The next few heatmaps showcase the expression of miRNAs in the HER2 tumor samples. **Figure 11** is a heatmap of expression that excludes all missing data. The most notable patterns of expression on this heatmap are near the top left and bottom right. The miRNAs in these regions of the heatmap are highly expressed. There is a small patch in the right central region of the heatmap where the expression of miRNAs is very low.

Figure 11 – Expression of miRNA in Her2 Tumors (No Missing Values)

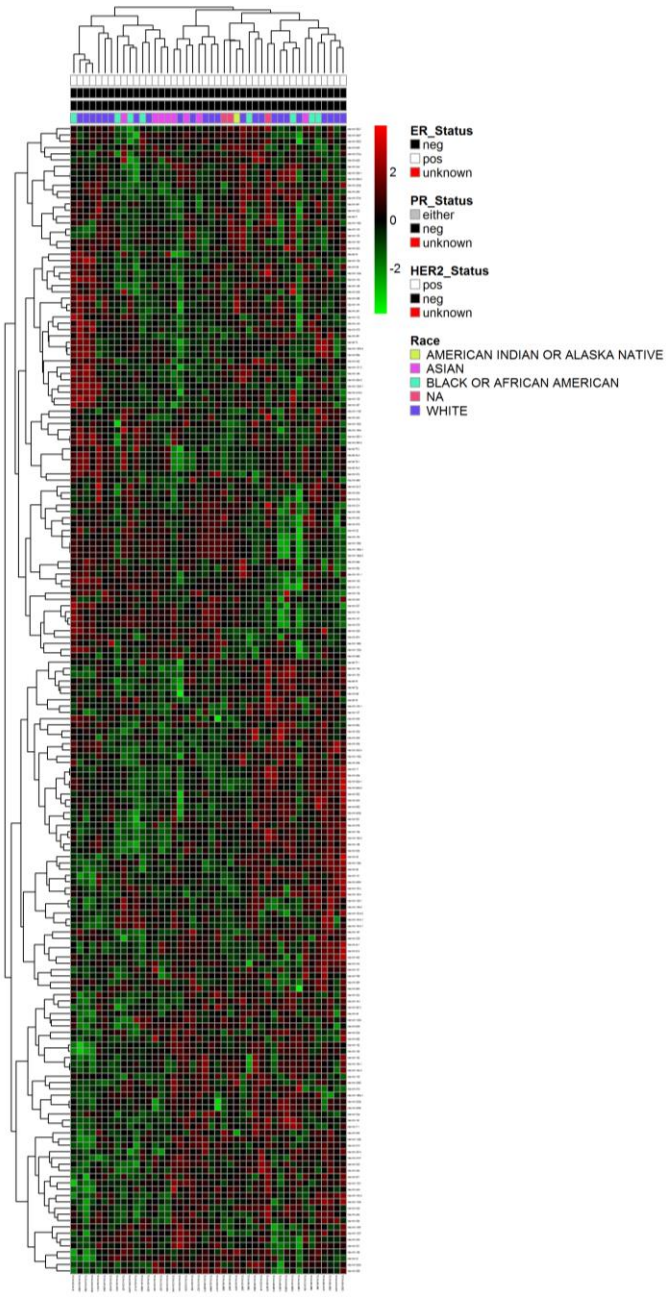


Figure 11. This heatmap of miRNA expression in the Her2 subtype excludes miRNAs with missing values.

Figure 12 shows the expression of miRNAs in HER2 tumors while including miRNA data with <80% of the values missing. The top center of this heatmap is dominated by red which means that the expression of miRNAs in that region of the heatmap is high.

Figure 12 – miRNA Expression in HER2 Tumors (<80% Missing Values)

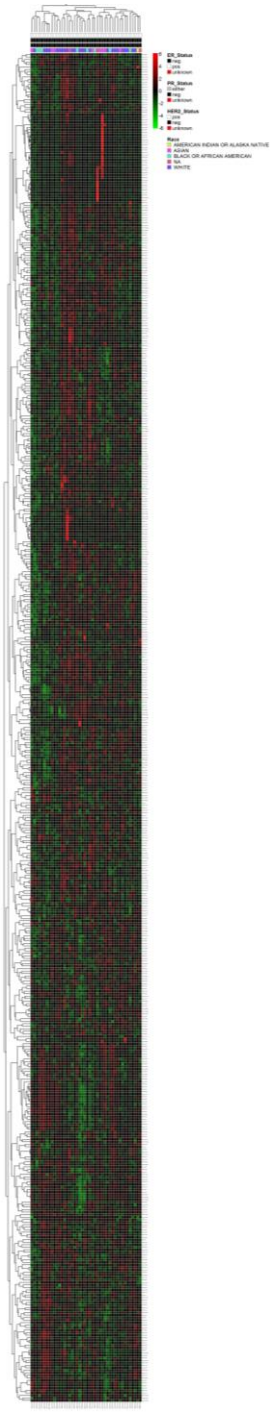


Figure 12. miRNA expression in HER2 tumors that includes miRNAs with <80% missing data.

In the next heatmap (**Figure 13**), the data has been filtered to include only select miRNAs that are associated with breast cancer and have <80% missing values. This heatmap shows 2 major clusters of miRNAs and 4 major clusters of tissue samples. The expression of miRNAs is varied throughout the heatmap but visibly it looks like the majority of miRNAs are highly expressed.

Figure 13 – miRNA Expression in HER2 Tumors (Select miRNAs)

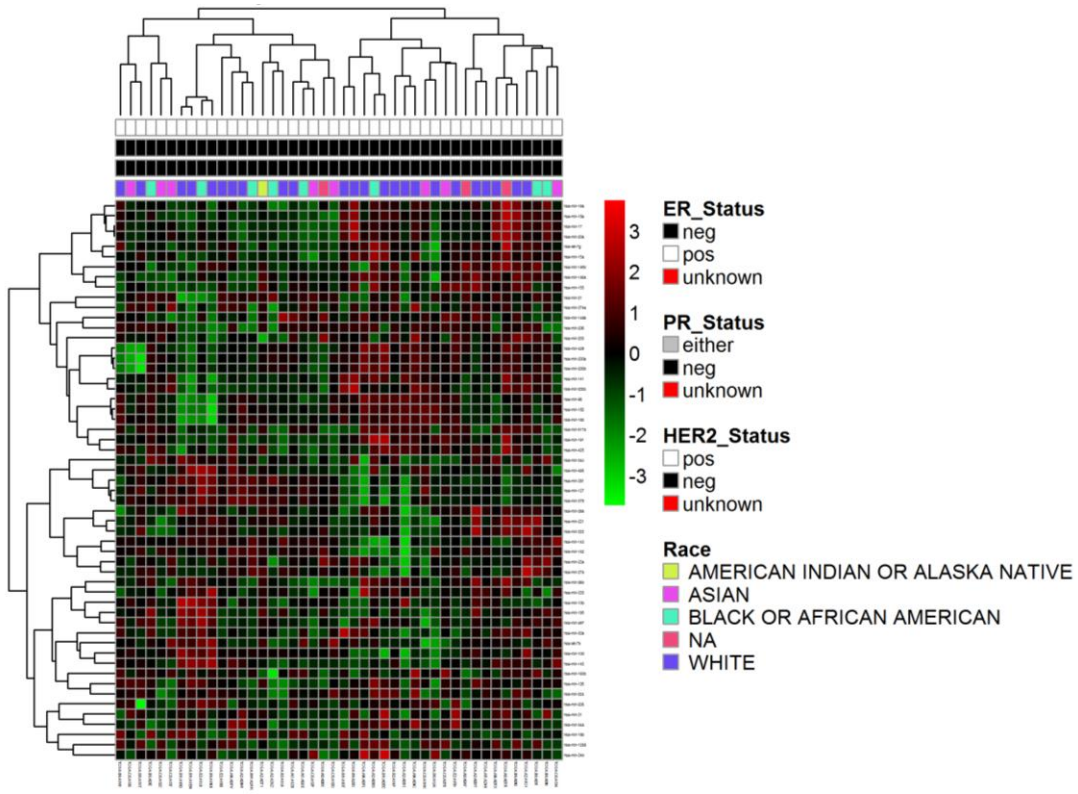


Figure 13. miRNA expression in HER2 tumors – the data excludes miRNAs that are not associated with breast cancer and includes miRNAs with <80% missing data.

4.5 Heatmaps – Luminal A Tumors

The majority of samples in our data (529 samples to be specific) belong to the Luminal A subtype. The following 3 heatmaps highlight the expression of miRNAs in Luminal A tumors. The first figure of these 3, **Figure 14,** is a heatmap of expression that excludes miRNAs with missing values. The more highly expressed miRNAs are towards the left part of the heatmap and the under-expressed miRNAs are towards the right side. In the annotation of race we note some clusters from the NA category.

Figure 14 – miRNA Expression in Luminal A Tumors (No Missing Values)

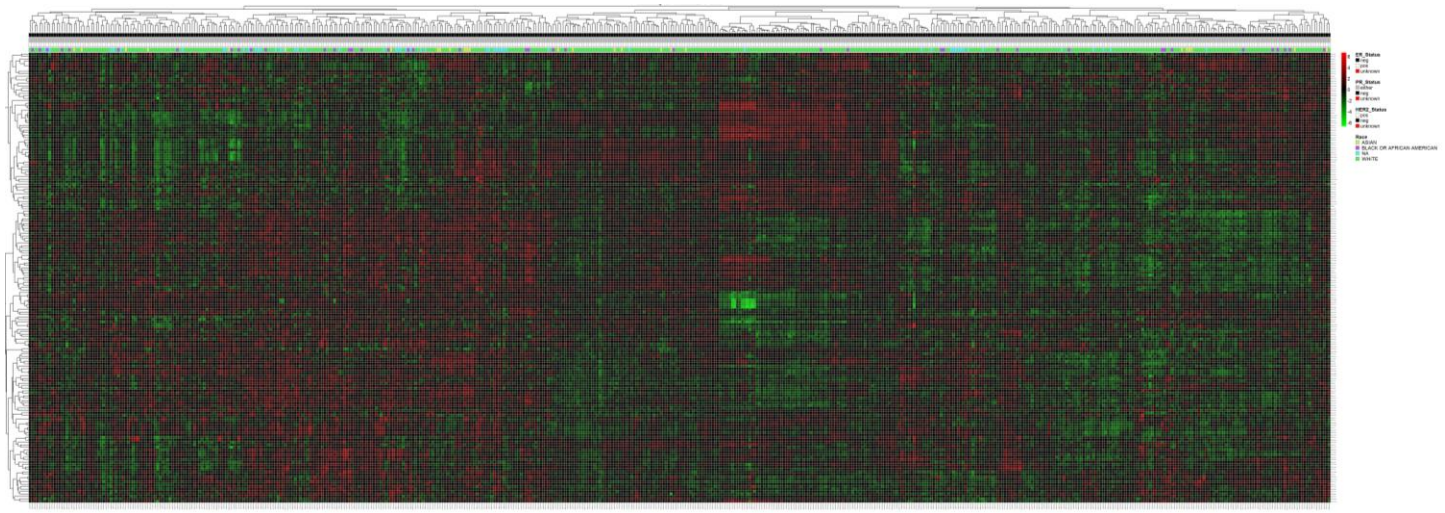


Figure 14. This is a heatmap of miRNA expression in Luminal A tumors. The data has no missing values. The clustering method is Pearson correlation.

The next heatmap in the luminal A category is **Figure 15**. It includes miRNAs with <80% missing values. The heatmap itself is not as remarkable as the one with no missing data. miRNAs that are highly expressed are interspersed throughout the heatmap but tend to be notable mostly near the bottom middle and bottom right side.

Figure 15 – miRNA Expression in Luminal A Tumors (<80% Missing Values)

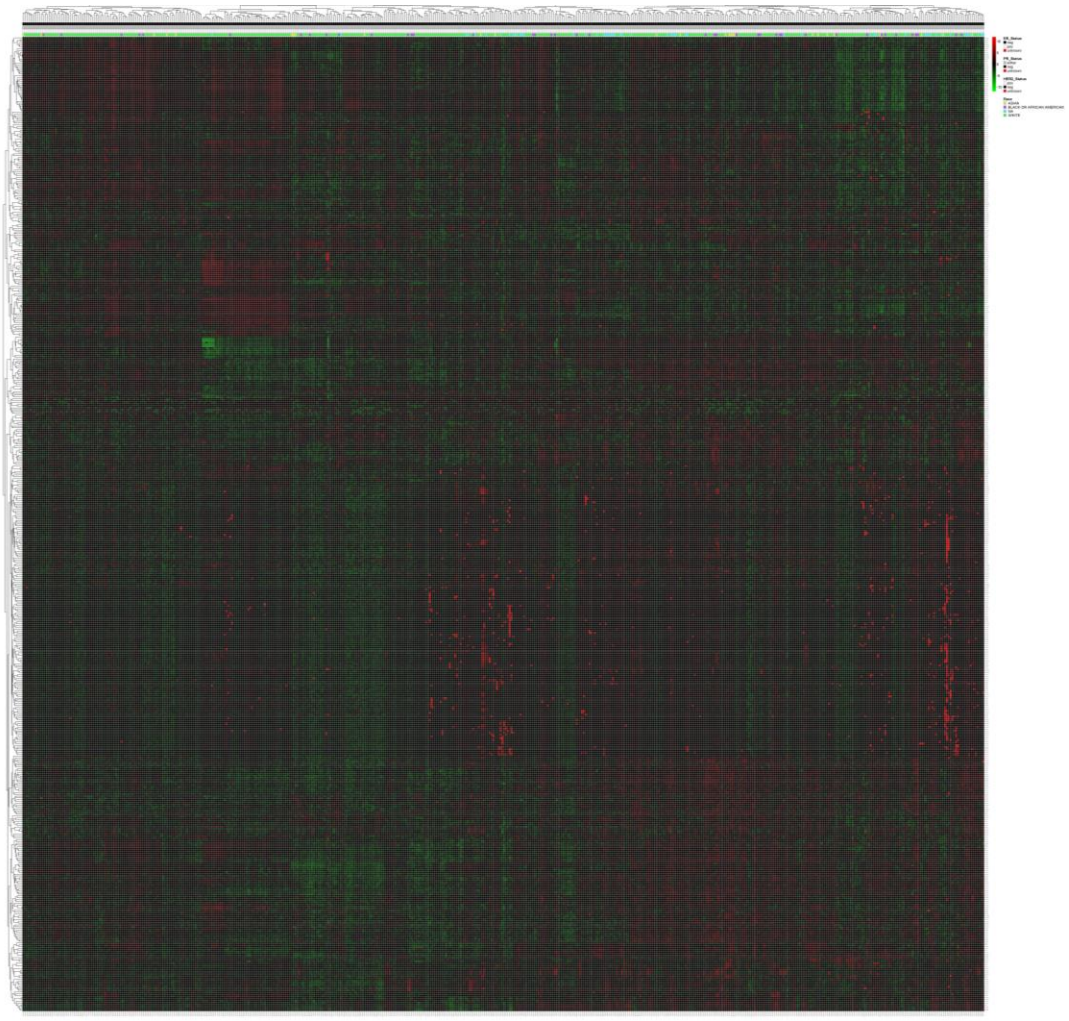


Figure 15. miRNA expression in Luminal A tumors. The data includes <80% missing values.

In **Figure 16** we see expression data for Luminal A samples that excludes miRNAs that are not associated with breast cancer and includes miRNAs with <80% missing values. This heatmap shows some interesting patterns of expression. Just off center to the left we see a bright patch of green above a dark patch of red. In that area the miRNAs are really under-expressed.

Figure 16 – miRNA Expression in Luminal A Tumors (Select miRNAs and <80% Missing Values)

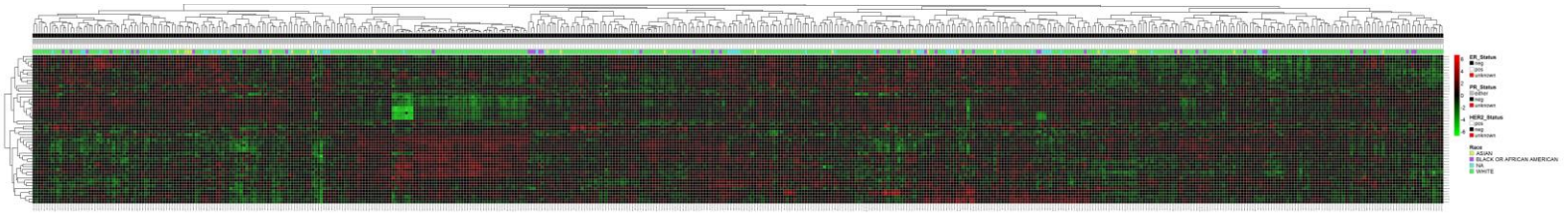


Figure 16. Heatmap of miRNA expression in Luminal A tumors with miRNAs associated with breast cancer and <80% missing values.

4.6 Heatmaps – Luminal B Tumors

The following 3 heatmaps all represent the expression of miRNAs in the luminal B subtype. **Figure 17** is the first heatmap and excludes miRNA with missing values. The top left corner of the heatmap is dominated by red whereas the bottom right corner is mostly green. In addition the bottom center of the heatmap has a notable red region where the miRNAs are highly expressed.

Figure 17 – miRNA Expression in Luminal B Tumors (No Missing Data)

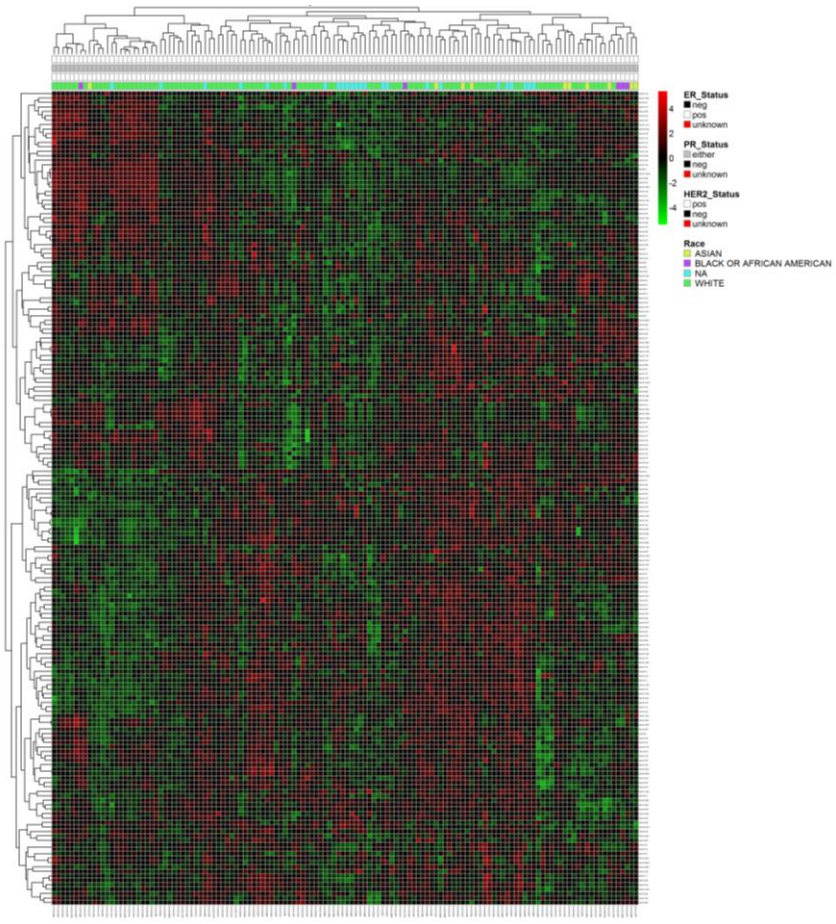


Figure 17. miRNA expression in Luminal B tumors with no missing data. The heatmap is annotated by race.

Figure 18 is a heatmap of miRNA expression in the Luminal B samples that includes miRNAs with <80% missing values. There are only a few visible patterns of expression. One region that is interesting is on the top right side of the heatmap where there is a bright region of red. miRNAs in that specific region are highly expressed.

Figure 18 – miRNA Expression in Luminal B Tumors (<80% Missing Values)

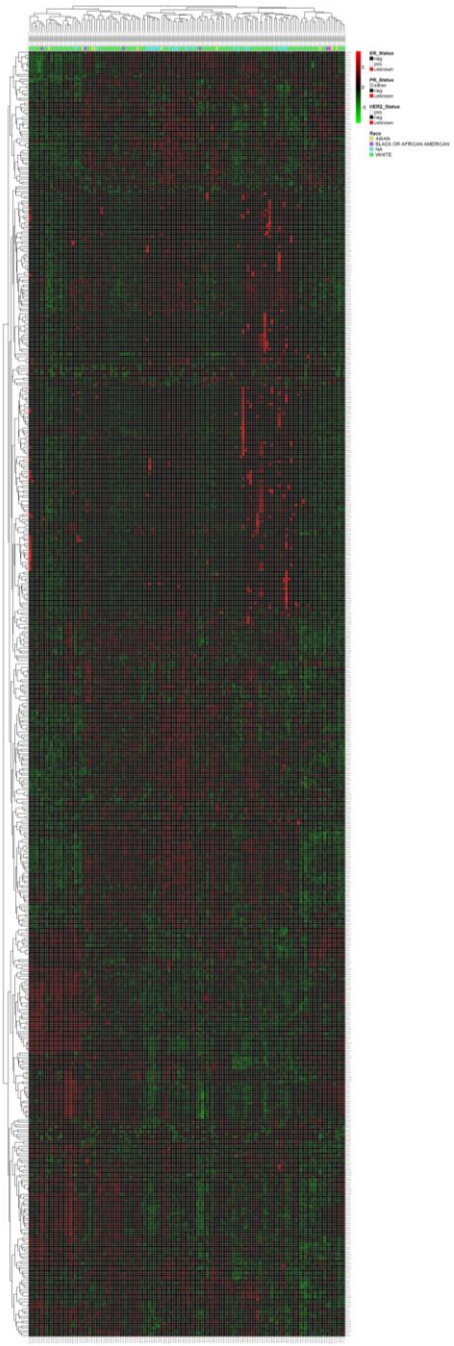


Figure 18. A heatmap of miRNA expression in Luminal B tumors. Includes miRNAs with <80% missing values.

The final heatmap of expression in Luminal B tumors is **Figure 19**. This heatmap filters to exclude all miRNAs that are not associated with breast cancer and includes <80% missing values. There is an interesting region on the left side of the heatmap where there is a bright green patch right above a bright red patch. The majority of the samples in this region belong to the White race. Finally there are 6 major clusters of samples and 2 major clusters of miRNAs.

Figure 19 – miRNA Expression in Luminal B Tumors (Select miRNAs and <80% Missing Values).

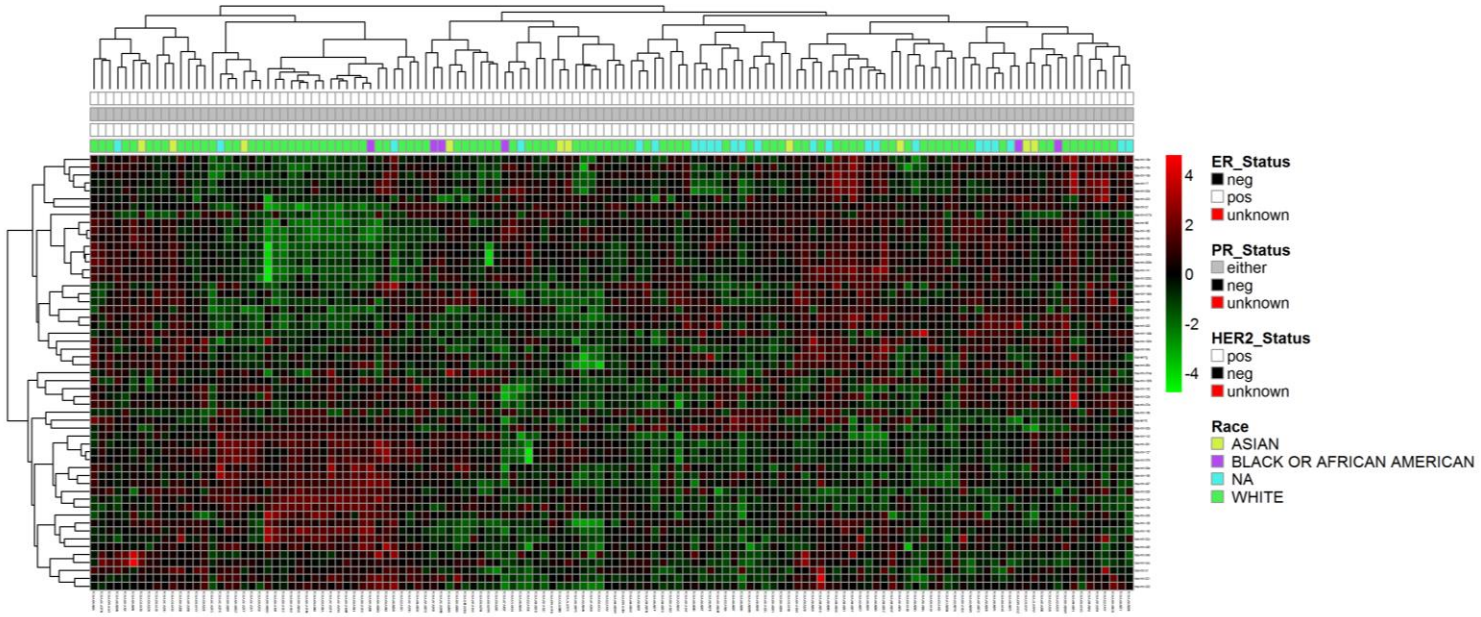


Figure 19. Expression of miRNAs in Luminal B tumors. The heatmap is annotated by race and is clustered by Pearson correlation.

4.7 Heatmaps – Coefficient of Variance > 60%

The next series of heatmaps select for miRNAs that have a coefficient of variance above 60%. **Figure 20** is the first in this series and does not include any miRNAs that have missing data. There are some noteworthy patterns in this heatmap – especially in the upper regions. The miRNAs are highly expressed in the upper left region whereas they are under-expressed in the upper right region. Similar patterns can be seen in the lower half of the heatmap.

Figure 20 – miRNA Expression for CV>60% (No Missing Data)

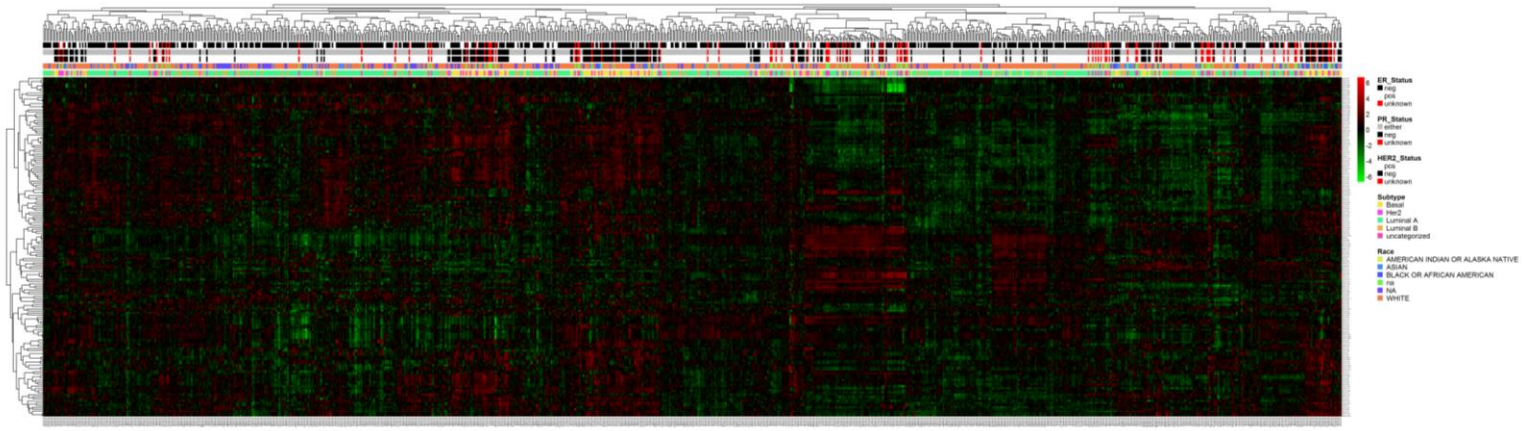


Figure 20. This heatmap selects for miRNAs that have a coefficient of variance greater than 60% and excludes all missing data.

In **Figure 21** the data for the heatmap includes miRNAs with <80% missing values. There is a cluster of the basal subtype close to the center of the heatmap. The miRNAs that fall under this cluster of basal samples are highly expressed.

Figure 21 – miRNA Expression for CV>60% (<80% Missing Values)

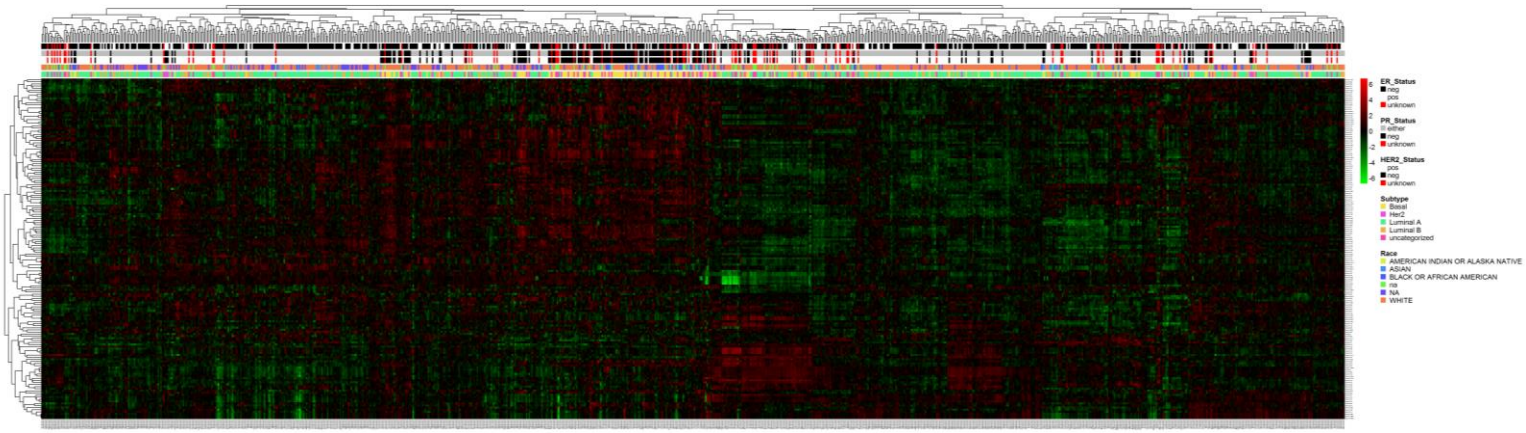


Figure 21. This heatmap selects for miRNAs that have a coefficient of variance greater than 60% and includes miRNAs with <80% missing values.

The last in the series of heatmaps with coefficient of variance >60% is **Figure 22**. This heatmap features only those miRNAs that are associated with breast cancer and also includes miRNAs with <80% missing values. As in **figure 21** there is a cluster of the basal subtype but it is located towards the right of the heatmap. Below the cluster of basal samples the miRNA expression is high. There is an eye-catching bright green area in the bottom left part of the heatmap that represents the under expression of miRNA in that region.

Figure 22 – miRNA Expression for CV>60% (select miRNA with <80% Missing Data)

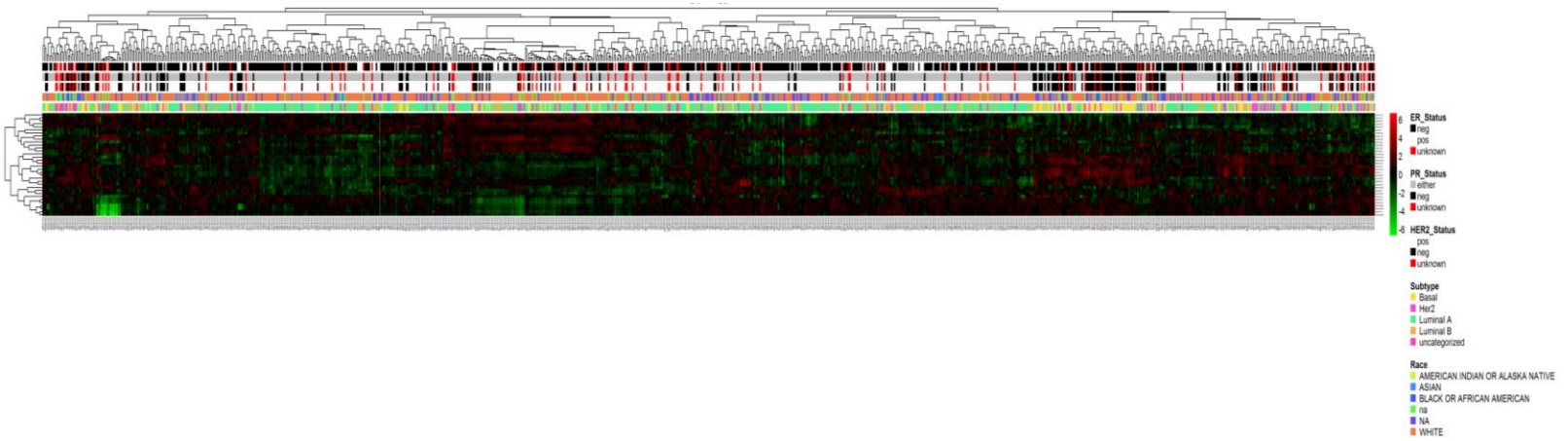


Figure 22. This heatmap selects for miRNAs that have a coefficient of variance greater than 60%, excludes all miRNAs not associated with breast cancer and includes miRNAs with <80% missing values.

4.8 Heatmaps – Coefficient of Variance >100%

In this last set of heatmaps we present expression of miRNAs that have a coefficient of variance that is >100%. **Figure 23** has miRNA with no missing values. There are 2 large clusters of the basal subtype that can be seen in the annotation region. The bottom left part of the heatmap falls under the basal samples and is predominantly red in color.

Figure 23 – miRNA Expression for CV>100% (No Missing Data)

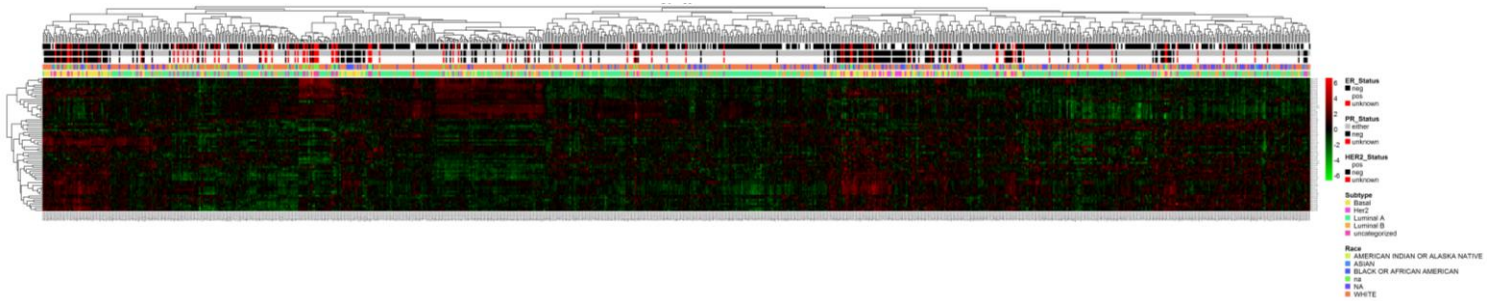


Figure 23. Expression of miRNAs with CV>100%. This heatmap excludes all missing data.

The next heatmap of expression for miRNAs with CV>100% (**Figure 24**) has included miRNAs with >80% missing values. The majority of the red regions lie to the left of the heatmap. miRNAs are expressed highly below a cluster of basal tissue samples.

Figure 24 – miRNA Expression for CV>100 (<80% Missing Data)

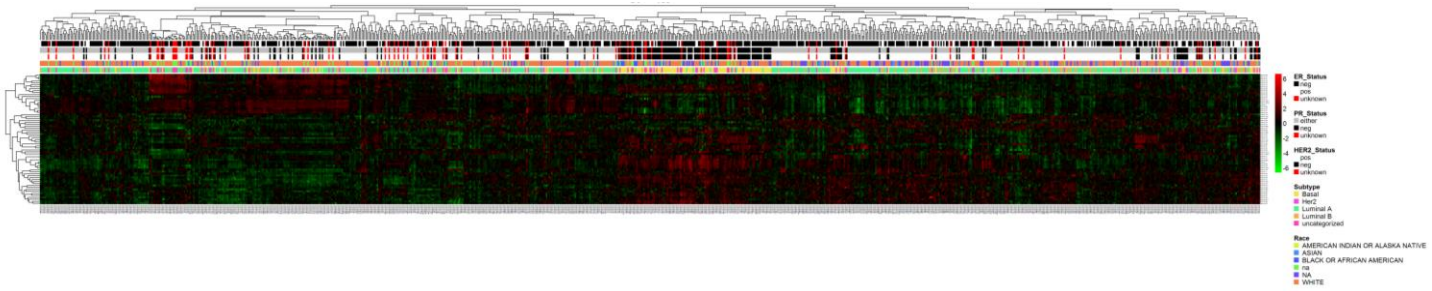


Figure 24. Expression of miRNAs with CV>100%. This heatmap includes miRNAs with <80% missing data.

Figure 25 is the last heatmap for CV>100% and filters for miRNAs associated with breast cancer and includes <80% missing values. The right half of the heatmap has red regions on the top half and green on the bottom while the left half of the heatmap has red regions predominantly in the bottom and green regions on top. There are 2 major clusters of miRNAs.

Figure 25 – miRNA Expression for CV>100 (Select miRNA and <80% Missing Values)

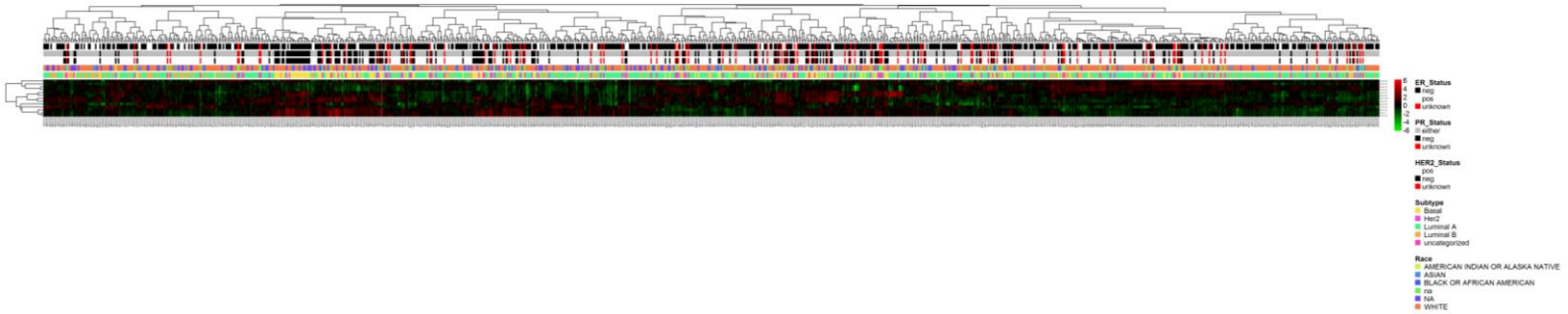


Figure 25. Expression of miRNAs with CV>100%. This heatmap excludes miRNAs that are not associated with breast cancer and includes miRNAs with <80% missing values.

4.9 Results of Principle Components Analysis

The Principal Components Analysis (PCA) was performed on miRNA expression data for all samples and then by subtype. The data included only those miRNAs that are associated to breast cancer and excluded all miRNAs with missing values. The samples were filtered to include only those for which we had gene expression data. The PCA for all the samples shows that the first component explains 99.98% of the variance in the data. **Figure 28** is a graph of the scores for the first principle component that results from a PCA on data with all samples. The 7 most variable miRNAs in order are hsa-mir-21, hsa-mir-374a, hsa-mir-23a, hsa-mir-203, hsa-mir-19a, hsa-mir-205 and hsa-let-7g.

Figure 26 – PCA for All Samples

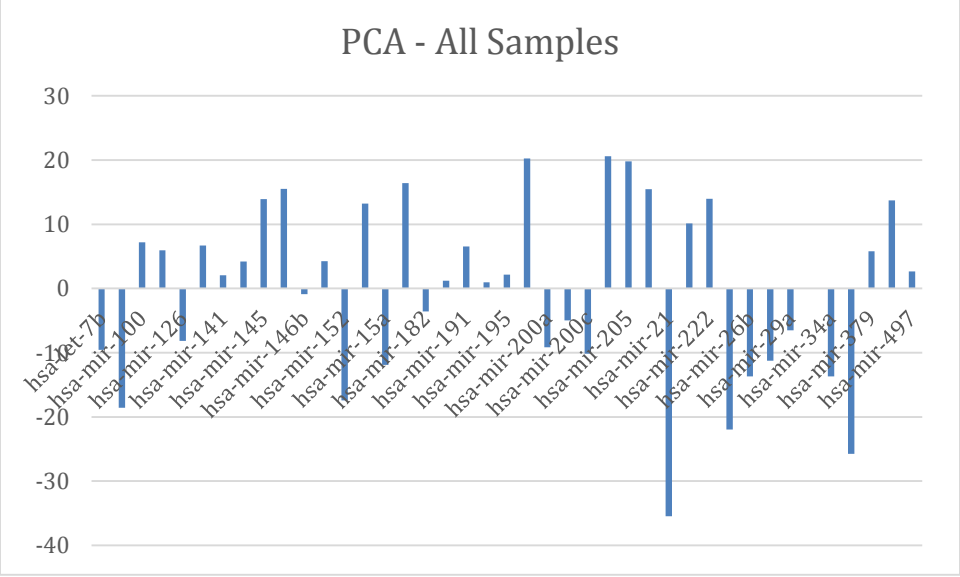


Figure 26. A graph of the scores for the first principle component in the PCA for all samples shows us the sign and magnitude of each miRNA.

Figure 27 is a graph of the scores of the first principal component in the PCA analysis for the Basal subtype. The 7 most variably expressed miRNAs in the PCA of the Basal subtype in order are hsa-mir-21, hsa-mir-374a, hsa-mir-203, hsa-mir-205, hsa-mir-127, hsa-mir-23a and hsa-mir-379. The first component explains 99.8% of the variance.

Figure 27 – PCA for Basal Samples

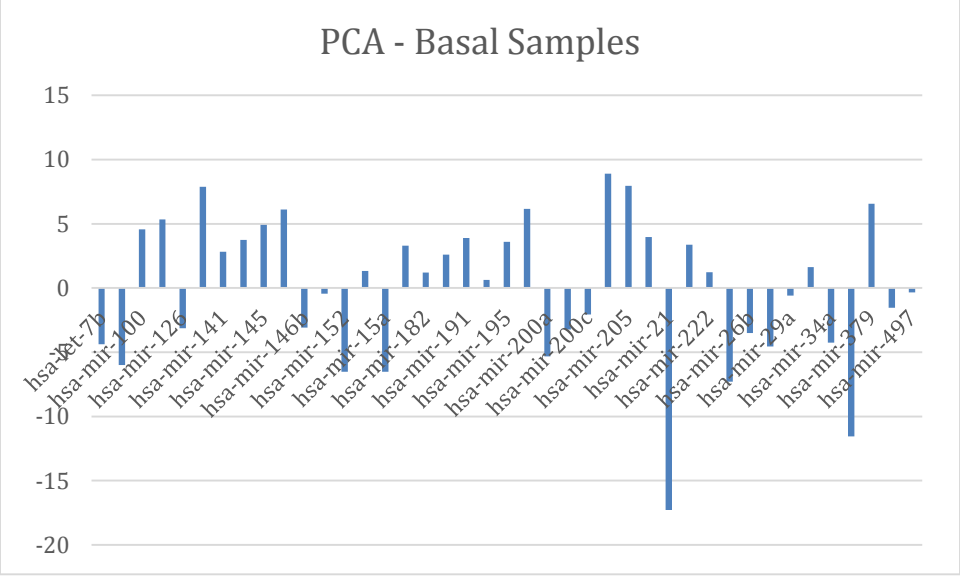


Figure 27. A graph of the scores for the first principle component in the PCA for Basal samples shows us the sign and magnitude of each miRNA.

The miRNAs hsa-mir-127 and hsa-mir-379 occurred uniquely in the PCA analysis of the Basal subtype which means they can further be explored

as potential Basal subtype biomarkers in breast cancer. To visually see how these specific miRNA express differently in the Basal subtype, we graphed the expression of both miRNAs across all subtypes (**Figure 28**).

Figure 28 – Expression of Potential Basal Subtype Biomarkers

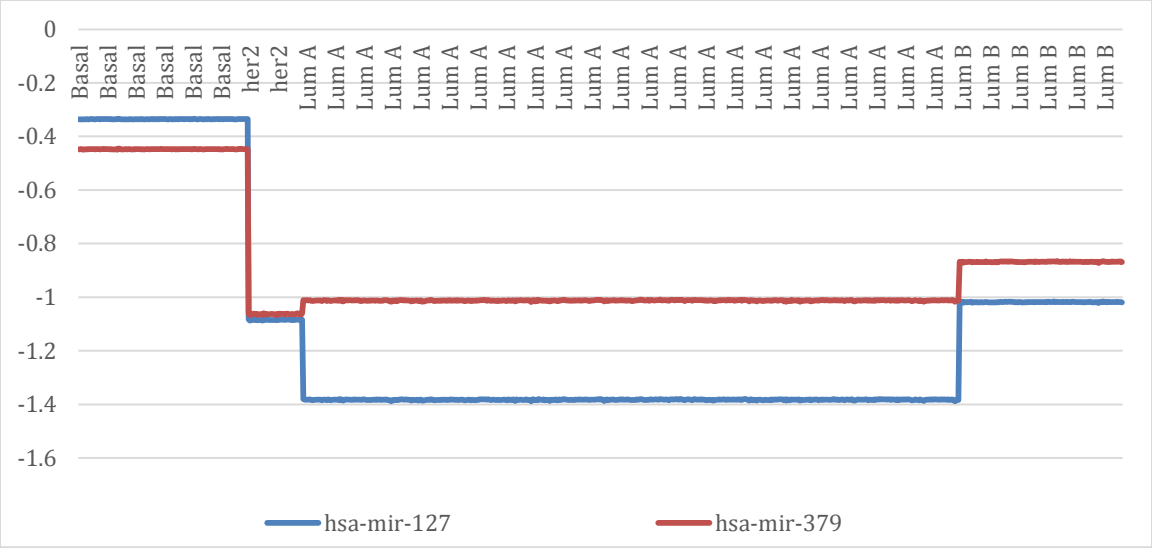


Figure 28. A graph of the expression of has-mir-127 and has-mir-379 (proposed Basal subtype biomarkers) across all four subtypes. According to the graph, both miRNAs are highly expressed in the Basal subtype of breast cancer.

The results of the PCA performed on the HER2 samples are displayed in **Figure 29**. The first component accounts for 99.99% of the variance in the data. The 7 most variably expressed miRNAs in the HER2 subtype are hsa-mir-374a, hsa-mir-21, hsa-mir-23a, hsa-mir19a, hsa-mir-126, hsa-mir-20a and hsa-mir-30a respectively.

Figure 29 – PCA for HER2 Samples

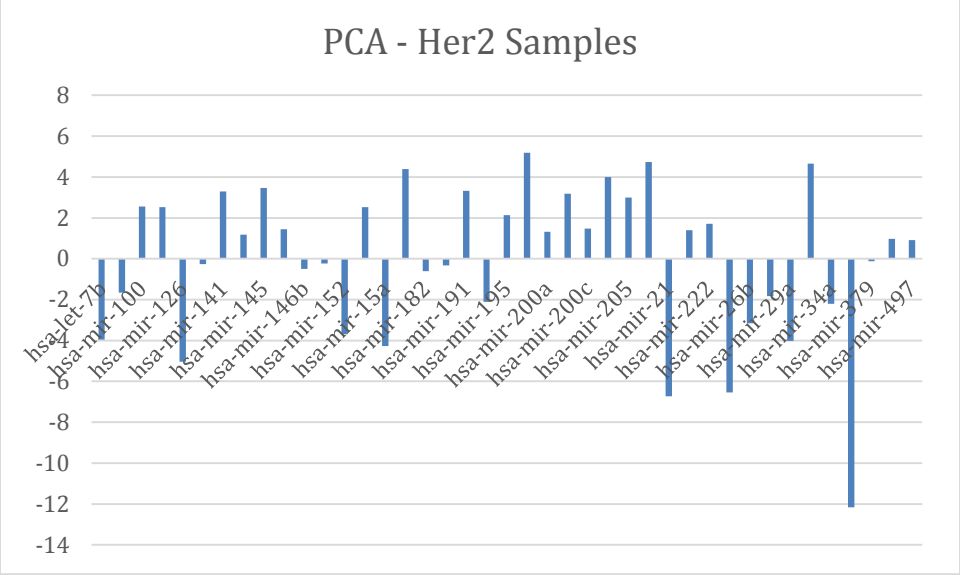


Figure 29. A graph of the scores for the first principle component in the PCA for HER2 samples shows us the sign and magnitude of each miRNA.

The PCA analysis of the HER2 subtype yielded four miRNAs that we did not find in the PCA analysis of any other subtype. Those four miRNAs are hsa-mir-19a, hsa-mir-126, hsa-mir-20a and hsa-mir-30a. We graphed the expression of these potential biomarkers for the HER2 subtype across all subtypes in **Figure 30**.

Figure 30 – Expression of Potential HER2 Subtype Biomarkers

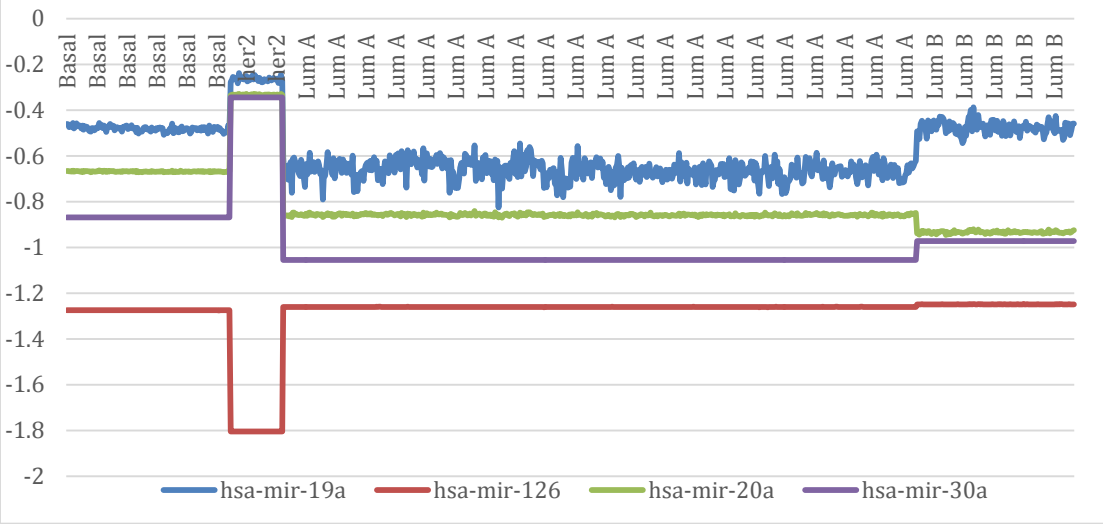


Figure 30. A graph of the expression of hsa-mir-19a, hsa-mir-126, hsa-mir-20a and hsa-mir-30a (proposed HER2 subtype biomarkers) across all four subtypes. According to the graph, hsa-mir-126 is down regulated while the other 3 miRNA are up-regulated in the HER2 subtype of breast cancer.

Figure 31 is a graph of the results from a PCA of Luminal A samples. The first principal component explains 99.96% of the variance in the data for the Luminal A. The top 7 most variable miRNAs from the PCA of luminal A tumors in order are hsa-mir-21, hsa-mir-23a, hsa-mir-374a, hsa-mir-205, hsa-mir-222, hsa-let-7g and hsa-mir-203.

Figure 31 - PCA for Luminal A Samples

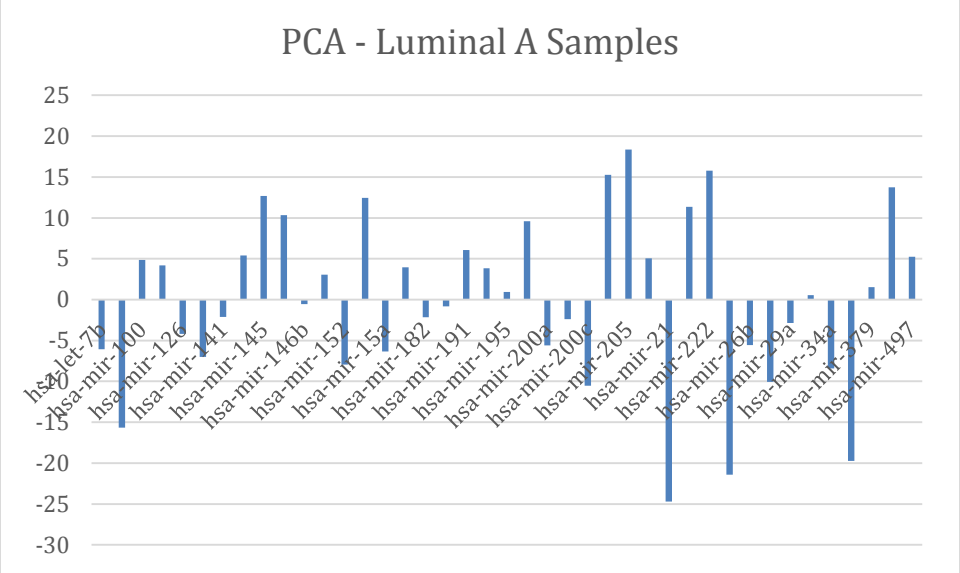


Figure 31. A graph of the scores for the first principle component in the PCA for Luminal A samples shows us the sign and magnitude of each miRNA.

Only one of the 7 miRNAs (hsa-mir-222) from the PCA analysis was unique to the Luminal A subtype. This expression of this potential biomarker for the Luminal A subtype was graphed across all subtypes in **Figure 32**. It can be noted from the graph that hsa-mir-222 is greatly up regulated in Luminal A subtypes.

Figure 32 – Expression of Potential Luminal A Subtype Biomarkers

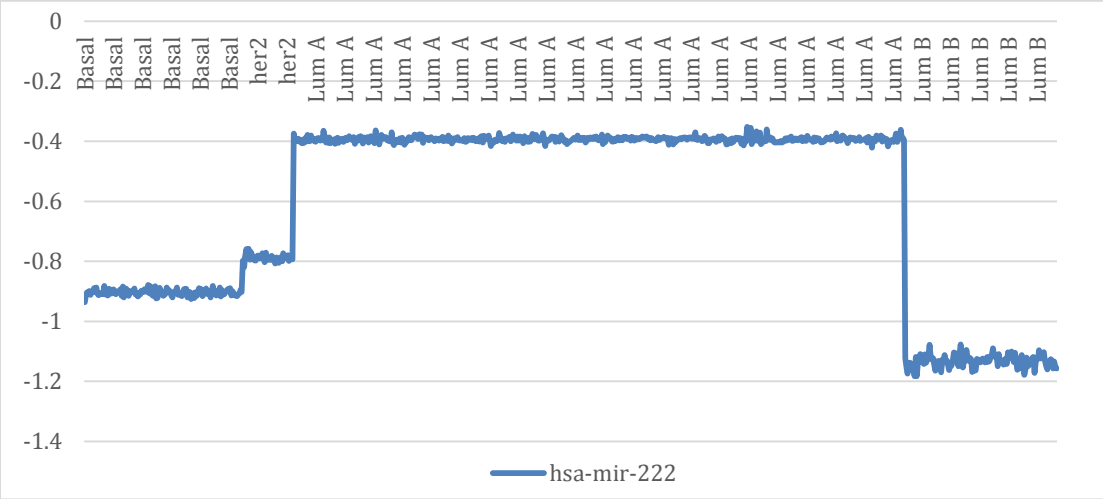


Figure 32. A graph of the expression of hsa-mir-222 (a proposed Luminal A subtype biomarker) across all four subtypes. According to the graph, hsa-mir-222 is up-regulated in the Luminal A subtype of breast cancer.

The results of the PCA for the Luminal B samples are in **Figure 33**. The first component of this analysis explains 99.96% of the variance in the data. The most variable miRNAs from this analysis are hsa-mir-21, hsa-mir-203, hsa-mir-205, hsa-mir-152, hsa-mir-26b, hsa-let-7g, and hsa-mir-200c.

Figure 33 – PCA for Luminal B Samples

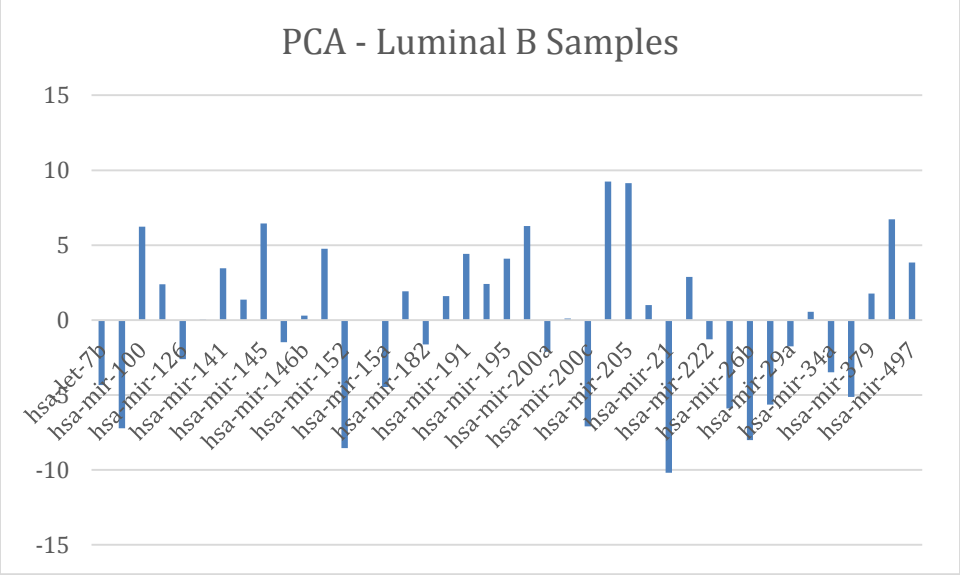


Figure 33. A graph of the scores for the first principle component in the PCA for Luminal B samples shows us the sign and magnitude of each miRNA.

There are three miRNAs from the result of the PCA analysis of the Luminal B subtype that do not occur in the PCA analysis of any other subtype. These potential biomarkers of the Luminal B subtype are hsa-mir-152, hsa-mir-26b and hsa-mir-200c. The expression of all three potential biomarkers of the Luminal B subtype was graphed across all the subtypes of breast cancer. **Figure 34** shows the resulting graph. It is interesting to note that although the expression of all 3 potential biomarkers is down regulated in the Luminal B subtype, the expression of hsa-mir-200c is up regulated in the HER2 subtype.

Figure 34 – Expression of Potential Luminal B Subtype Biomarkers

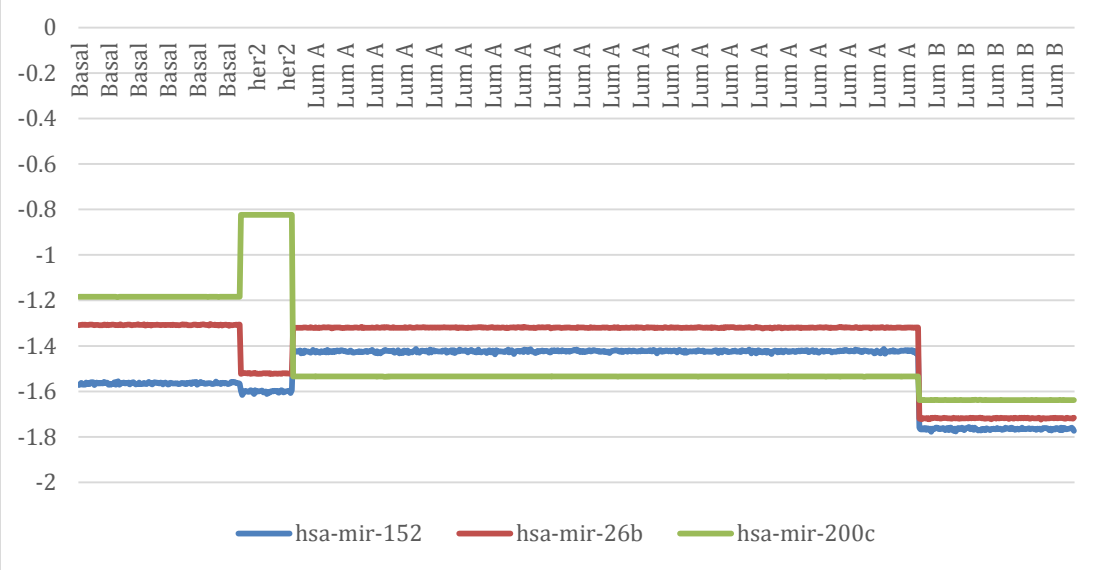


Figure 34. A graph of the expression of hsa-mir-152, hsa-mir-26b and hsa-mir-200c (proposed Luminal B subtype biomarkers) across all four subtypes. The miRNAs are down regulated in the Luminal B subtype of breast cancer. It is also important to note that according to this graph, hsa-mir-200c is up regulated in the HER2 subtype.

4.10 Results of Correlation Analysis

Through correlation analysis we found the 20 pairs of miRNAs that are most highly correlated in their expression. **Table 7** lists the 20 pairs of miRNAs along with the correlation coefficient for each. The pair of miRNA with the highest correlation coefficient is hsa-let-7a.1 and hsa-let-7a-2.

Fourteen of the 20 pairs that have correlating expression belong to the same family of miRNAs.

Table 7 – 20 Pairs of Most Highly Correlated miRNAs

Table 7. This is a table listing the 20 most highly correlated pairs of miRNAs and the correlation coefficients for those pairs.

Pair	First Variable	Second Variable	Correlation	Related
1	hsa-let-7a-1	hsa-let-7a-2	0.9997511	Yes let-7 family
2	hsa-let-7a-2	hsa-let-7a-3	0.9996304	Yes let-7 family
3	hsa.let.7a.1	hsa-let-7a-3	0.9996081	Yes let-7 family
4	hsa-mir-9-1	hsa-mir-9-2	0.9988614	Yes mir-9 family
5	hsa-mir-199a-1	hsa-mir-199a-2	0.9944959	Yes mir-199 family
6	hsa-mir-199a-2	hsa-mir-199b	0.9918414	Yes mir-199 family
7	hsa-mir-29b-1	hsa-mir-29b-2	0.9837018	Yes mir-29 family
8	hsa-mir-365-1	hsa-mir-365-2	0.9783261	Yes mir-365 family
9	hsa.mir.194.1	hsa-mir-194-2	0.9765156	Yes mir-194 family
10	hsa-mir-199a-1	hsa-mir-199b	0.9752805	Yes mir-199 family
11	hsa-mir-128-1	hsa-mir-128-2	0.9511545	Yes mir-128 family
12	hsa-mir-200a	hsa-mir-200b	0.9443755	Yes mir-8 family
13	hsa-mir-17	hsa-mir-20a	0.934769	Yes mir-17 family
14	hsa-mir-144	hsa-mir-451	0.9334398	No
15	hsa-let-7c	hsa-mir-99a	0.9178015	No
16	hsa-mir-125b-2	hsa-mir-99a	0.9159054	No
17	hsa-let-7c	hsa-mir-125b-2	0.9111898	No
18	hsa-mir-92a-1	hsa-mir-92a-2	0.9056862	Yes mir-25 family
19	hsa-mir-182	hsa-mir-183	0.9048191	No
20	hsa-mir-127	hsa-mir-379	0.8975715	No

4.11 Results of Networks, Pathways and Targets Analysis

In order to elucidate the involvement of miRNAs in multiple pathways we queried an online database with our cohort of miRNAs that are associated with breast cancer. **Figure 35** is a heatmap displaying the results of this search. We note that hsa-miR-15a-5p, hsa-miR-193b-3p, hsa-miR-34a-5p, has-let-7g-5p, has-miR-17-5p, hsa-miR-20a-5p are involved in the most number of pathways according to the heatmap.

Figure 35 – Heatmap of miRNA Expression in Various Pathways

Figure 35. This heatmap was generated using a we tool called miRPath which is a part of Diana Tools [95].

We found miRNA-gene networks in the 2 of the 4 subtypes we analyzed. **Figure 36** displays the results of our analysis for the Basal subtype. There are 5 genes that are common targets of has-mir-127 and has-mir379.

Figure 36 – miRNA-Gene Network in Basal Subtype

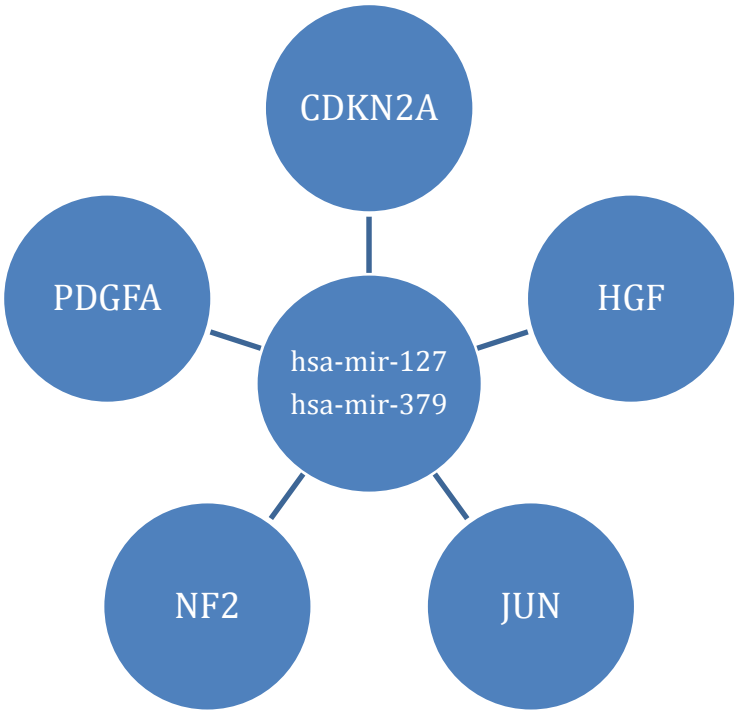


Figure 36. hsa-mir-127 and hsa-mir-279 both have 5 common gene targets as shown above.

The HER2 subtype has 4 potential biomarkers for which we investigated common gene targets. We found that there are 8 genes that are common targets of hsa-mir-19a, hsa-mir-126, hsa-mir-20a and hsa-mir-30a

which are miRNAs that we propose as potential biomarkers of the Her2 subtype in breast cancer. We did not find any gene-miRNA networks in the Luminal A and Luminal B subtypes. **Figure 37** displays the results of this analysis.

Figure 37 – miRNA-Gene Network in Her2 Subtype

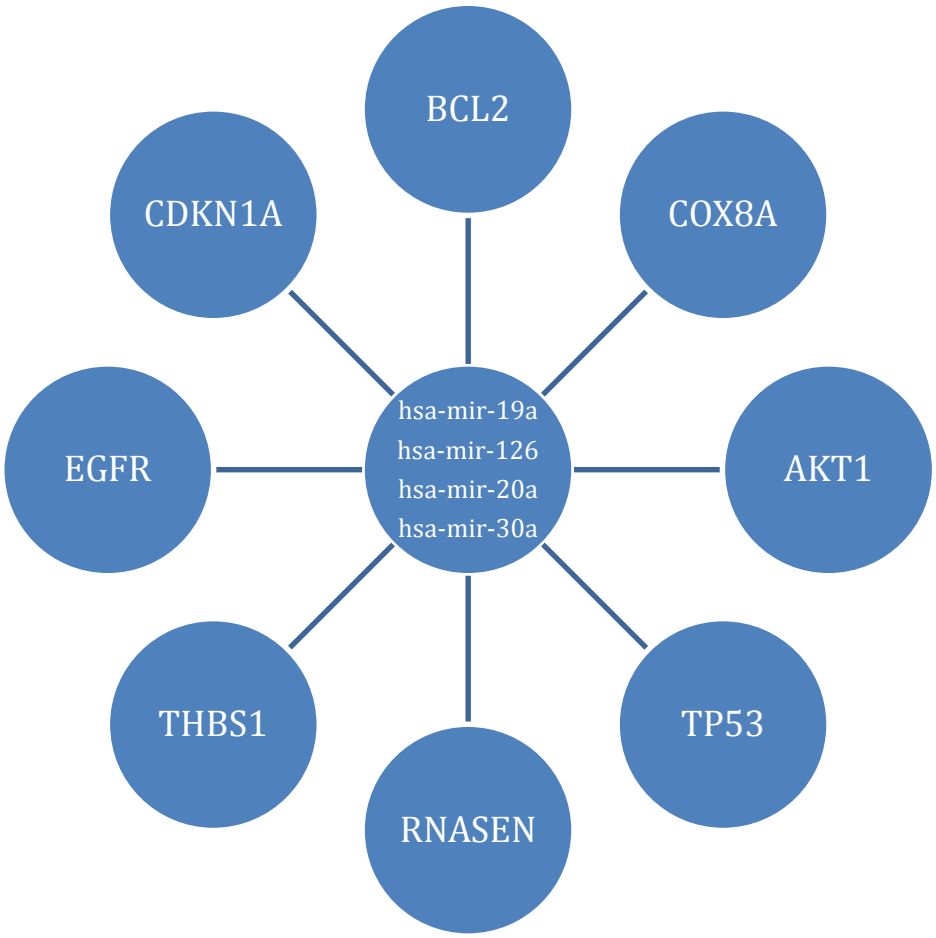


Figure 37. hsa-mir-19a, hsa-mir-126, hsa-mir-20a and hsa-mir-30a have 8 gene targets in common as shows by this schematic.

Chapter V

DISCUSSION

V. DISCUSSION

Investigating the small and impactful miRNAs has provided us with a wealth of information on the biological interactions of these molecules and how those interactions affect the outcome of a disease like cancer. The deregulation or expression of miRNAs in cancer tissue can reveal a lot about the characteristics of the specific tumor. In our study we wanted to take a panoptic approach to visualizing the miRNA expression data generated by TCGA. Although miRNA expression has been studied before – we believe that we have been broad in our approach and method.

We generated various heatmaps in order to comprehensively visualize how miRNA expression can vary by subtype, race and statistical manipulation of data. The results are intriguing. The heatmaps that are sorted by subtype show visually how different the expression patterns can be between subtypes of breast cancer. The clustering of our heatmaps with all samples included did not completely group together the known subtypes. We also did not see any significant grouping by race in heatmaps by subtypes. This may be attributed

to the method of clustering or normalization we used. We however chose unsupervised clustering over supervised clustering of known groups to find new patterns of expression. One such recurring pattern we noted on our heatmaps is a small grouping of the basal samples that corresponded to miRNAs that were highly expressed.

The reason for filtering the miRNAs by a cut off for coefficient of variance was to highlight the expression of miRNAs with the most variable expression in the cohort that we had. The result was a heatmap with patches of bright red and green that show how variable the expression of miRNAs can be across samples. The inclusion of miRNAs with <80% missing values changed the way the heatmap clustered and also displayed new patterns of expression.

The difference in expression of miRNAs between tumor subtypes is advantageous to the field of science as this non-conformity in expression allows us to use miRNAs as biomarkers of disease. In order to propose potential biomarkers we used Matlab to conduct Principal Components Analysis on our cohort of miRNAs for each subtype. We then graphed the potential biomarkers to visually confirm our results. To the best of our knowledge our study is the first to use PCA on miRNA expression data from TCGA to suggest potential biomarkers.

The results of our PCA are quite promising. Our results suggest that hsa-mir-127 and hsa-mir-379 are potential biomarkers for the Basal subtype of breast cancer. A previous study found that hsa-mir-127 is down regulated in breast cancer and regulates cell proliferation by acting on BCL6 [96]. Past work has also shown that hsa-mir-379 is down regulated in breast cancer [97]. The graph in **figure 28** shows that both hsa-mir-127 and hsa-mir-379 are up regulated in the basal subtype of breast cancer. For the HER2 subtype our analysis reveals that hsa-mir-19a, hsa-mir-126, hsa-mir-20a and hsa-mir-30a would be good potential biomarkers. Although not much information is available, hsa-mir-19a and hsa-mir-126 are known to be down regulated in breast cancer from previous studies [98, 99]. In contrast, according to one study, hsa-mir-20a is up regulated in breast cancer [100]. The miRNA, hsa-mir-30a, has been more thoroughly researched in other studies that have highlighted its potential as a prognostic marker and a target for therapeutic intervention as it is a tumor suppressing miRNA that is deregulated in breast cancer [101, 102]. Our findings in **figure 30** show that apart from hsa-mir-126 which is down regulated, the other 3 suggested biomarkers are all up regulated in the HER2 subtype of breast cancer. PCA of Luminal A samples provided us with one miRNA (has-mir-222) that our data supports as a good potential biomarker for the Luminal A subtype. A study from 2012, has also concluded that has-mir-222 is significantly up regulated in breast cancer and

would be a great biomarker for the disease [103]. This result in **figure 32** supports the previous finding – hsa-mir-222 is up regulated in the Luminal A subtype according to our analysis. The Luminal B samples in our study yielded 3 potential biomarkers after PCA and they are hsa-mir-152, hsa-mir-26b and hsa-mir-200c. Our findings are in line with studies that were published fairly recently - that claim that has-mir-152 and has-mir-26b are down regulated in breast cancer and should be explored further as a therapeutic targets [104-106]. On the other hand hsa-mir-200c is up regulated in breast cancer [107]. When we graphed the expression of these 3 proposed biomarkers across all subtypes we found that all 3 were down regulated in the Luminal B subtype (**figure 34**). It is important to note however that hsa-mir-200c is the least down regulated of the 3 and it is also looks like it is up regulated mostly in the HER2 subtype. The potential biomarkers for each subtype were chosen because they occurred only once when comparing a list of the top 7 most variably expressed miRNAs in each subtype. The miRNAs that we found as most variably expressed across the subtypes would be good biomarkers for breast cancer. Those miRNAs include hsa-mir-21, hsa-mir-374a and has-mir-23a, hsa-mir-203 and hsa-mir-205. Results from other studies have also put forth these miRNAs as suggestions for potential biomarkers and or targets of therapeutic intervention after extensive research [108-111].

We had originally hypothesized that our analysis of the highly correlated pairs of miRNAs would belong to the same family of miRNAs. This was supported by knowledge that miRNAs that belong to the same family may be transcribed from the same primary-miRNA or may have common mRNA targets. Although, our analysis mostly supports the hypothesis we made, we did find that 6 of the 20 pairs of miRNAs with highly correlating expression were unrelated. This is definitely an intriguing finding.

The fact that miRNAs play a role in multiple pathways is not new but we felt it would be good to visualize the cross involvement of select miRNAs in other disease pathways. The results of the pathway analysis are in **figure 35**. In our mission to uncover miRNA-gene networks in breast cancer subtypes we found 2 networks where the most variable miRNAs in the subtype have common gene targets. We created schematics to help visualize these interactions in **figure 36** and **figure 37**. We believe that the up or down regulation of these common targets plays a major role in breast cancer. We chose only to explore the first level of any potential network to reveal immediate interactions in networks that are in all probability highly intricate.

Chapter VI

SUMMARY AND CONCLUSIONS

VI. SUMMARY AND CONCLUSIONS

Genomics and proteomics applications have helped us solve some of the many mysteries of molecular interactions that were unknown to us. In addition these applications have been competent in the way they generate, manage and analyze large amounts of data – thus alleviating most of the burden faced by scientists. The ability to sequence genomes efficiently and in a short span of time has also given us the ability to examine and treat patients as individuals and not groups. In the long run, the practice of personalized medicine will optimize treatment and save lives [112]. Through this study we aimed to contribute to the growing depth of knowledge that can redefine our approach to treating cancer and potentially save lives.

Although comprehensive we feel our approach to studying miRNA expression could have been more in-depth if applications for the comparison of heatmaps were available as web based tools. A computer program would be better at recognizing and comparing patterns across heatmaps. Visual comparison relies too heavily on an individual doing the comparisons and therefore is prone to error. Regardless, it is interesting to note that we could

deduce via visual analysis only that miRNA expression varies quite a bit when compared between subtypes. We did not note any significant differences in the expression patterns of miRNA across race but there could be several reasons for this. One such reason could be a limited cohort with only a small number of samples representing the racial minorities. We also lacked data from normal tissue samples and therefore could not compare the expression of miRNA in breast tumors versus normal tissue samples. The availability of such data would allow us to be even more confident in our analysis and conclusions.

In addition to visually studying miRNA expression in breast cancer we used Principal Components Analysis (PCA) to support with evidence any potential biomarkers for breast cancer subtypes. A literature search confirms that this type of PCA analysis has not previously been performed on TCGA breast cancer data. We hope that future studies validate our computational results by experimental study of the miRNAs we suggest as biomarkers or targets of therapeutic intervention. Once again, having data on normal breast tissue samples would aid in further validating our results.

In **table 7**, we found 6 pairs of miRNAs that share highly correlated expression but are unrelated. Further studying this interesting finding could lead to the discovery of new families or new molecular interactions of miRNAs.

Our analysis of the pathways and networks in breast cancer was straightforward and simple. We feel that future studies can build on our findings and do a more in-depth analysis that explores important protein and molecular interactions apart from the gene targets that we focused on.

In conclusion we believe that our results and the results of any follow up studies will help greatly in accomplishing the purpose of TCGA – which is to aid in the comprehensive and multi-layered analysis of cancers to discover novel theories rapidly and efficiently. As more and more data is generated the results of any computational analyses will have fewer errors and more accuracy. The inclusion of more clinical data and data on co-morbidities and environmental factors would be invaluable for future studies that wish to study miRNAs in cancer. Environmental factors for example may play a major role in the expression of miRNAs and may even explain why some miRNAs are up or down regulated in individuals. Thus far, studies have concentrated more on the internal interactions of miRNAs but we feel that external factors should also be explored by scientists in future studies.

An all-inclusive approach is vital in the study of miRNAs and personalized medicine – which will completely change the way we treat patients in the near future. We also support the use of computational methods in studying big data and validating experimental research. Computational

research offers great precision and multiple-approaches to data analysis. We hope that more researchers utilize these computational methods to interpret an ever growing amount of scientific data that is continually generated.

References

1. Team, R.C., *R: A Language and Environment for Statistical Computing*, 2014, R Foundation for Statistical Computing: Vienna, Austria.

2. Wall, L., *Perl*, 2014.

3. NCI. 2013 [cited 2013; Estimated New Cases and Deaths]. Available from: http://www.cancer.gov/.

4. *Comprehensive molecular portraits of human breast tumours.* Nature, 2012. **490**(7418): p. 61-70.

5. 2013 [cited 2013; Available from: http://www.cancer.org/.

6. Bombonati, A. and D.C. Sgroi, *The molecular pathology of breast cancer progression.* The Journal of pathology, 2011. **223**(2): p. 307-17.

7. Richie, R.C. and J.O. Swanson, *Breast cancer: a review of the literature.* Journal of insurance medicine, 2003. **35**(2): p. 85-101.

8. DeSantis, C., et al., *Breast cancer statistics, 2011.* CA: a cancer journal for clinicians, 2011. **61**(6): p. 409-18.

9. Nik Ruzyanei Nik Jaafar, S.H.S.D., Suriati Mohamed Saini, Siti Nor Aizah Ahmad, Marhani Midin, Hatta Sidi, Umi Adzlin Silim, Azlin Baharudin, *Clinical Depression While Caring for Loved Ones with Breast Cancer.* Comprehensive Psychiatry, 2013.

10. Hedenfalk, I., et al., *Gene-expression profiles in hereditary breast cancer.* The New England journal of medicine, 2001. **344**(8): p. 539-48.

11. NIH. 2007 [cited 2013; Available from: http://ghr.nlm.nih.gov/condition/breast-cancer.

12. Breastcancer.org. 2013 [cited 2013; Available from: http://www.breastcancer.org.

13. Sainsbury, J.R., T.J. Anderson, and D.A. Morgan, *ABC of breast diseases: breast cancer.* BMJ, 2000. **321**(7263): p. 745-50.

14. Malhotra, G.K., et al., *Histological, molecular and functional subtypes of breast cancers.* Cancer biology & therapy, 2010. **10**(10): p. 955-60.

15. Sorlie, T., et al., *Repeated observation of breast tumor subtypes in independent gene expression data sets.* Proceedings of the National Academy of Sciences of the United States of America, 2003. **100**(14): p. 8418-23.

16. Carey, L.A., et al., *Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study.* JAMA : the journal of the American Medical Association, 2006. **295**(21): p. 2492-502.

17. Schnitt, S.J., *Classification and prognosis of invasive breast cancer: from morphology to molecular taxonomy.* Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc, 2010. **23 Suppl 2**: p. S60-4.

18. Bernstein, L. and J.V. Lacey, Jr., *Receptors, associations, and risk factor differences by breast cancer subtypes: positive or negative?* Journal of the National Cancer Institute, 2011. **103**(6): p. 451-3.

19. Perou, C.M. and A.L. Borresen-Dale, *Systems biology and genomics of breast cancer.* Cold Spring Harbor perspectives in biology, 2011. **3**(2).

20. CDC. *Risk Factors*. 2013 [cited 2013; Available from: http://www.cdc.gov.

21. McPherson, K., C.M. Steel, and J.M. Dixon, *ABC of breast diseases. Breast cancer-epidemiology, risk factors, and genetics.* BMJ, 2000. **321**(7261): p. 624-8.

22. Senie, R.T., et al., *Method of tumor detection influences disease-free survival of women with breast carcinoma.* Cancer, 1994. **73**(6): p. 1666-72.

23. Sihto, H., et al., *Molecular subtypes of breast cancers detected in mammography screening and outside of screening.* Clinical cancer research : an official journal of the American Association for Cancer Research, 2008. **14**(13): p. 4103-10.

24. McCormack, V.A. and I. dos Santos Silva, *Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis.* Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology, 2006. **15**(6): p. 1159-69.

25. Vachon, C.M., et al., *Mammographic density, breast cancer risk and risk prediction.* Breast cancer research : BCR, 2007. **9**(6): p. 217.

26. Lokate, M., et al., *Age-related Changes in Mammographic Density and Breast Cancer Risk.* American journal of epidemiology, 2013.

27. Elizabeth O'Day, A.L., *MicroRNAs and Their Target Gene Networks in Breast Cancer.* Breast cancer research : BCR, 2010. **12**(201).

28. Ansorge, W.J., *Next-generation DNA sequencing techniques.* New biotechnology, 2009. **25**(4): p. 195-203.

29. Pandey, A. and M. Mann, *Proteomics to study genes and genomes.* Nature, 2000. **405**(6788): p. 837-46.

30. Walther, T.C. and M. Mann, *Mass spectrometry-based proteomics in cell biology.* The Journal of cell biology, 2010. **190**(4): p. 491-500.

31. Yarmush, M.L. and A. Jayaraman, *Advances in proteomic technologies.* Annual review of biomedical engineering, 2002. **4**: p. 349-73.

32. Monti, M., et al., *Functional proteomics.* Clinica chimica acta; international journal of clinical chemistry, 2005. **357**(2): p. 140-50.

33. Iorio, M.V., et al., *MicroRNA gene expression deregulation in human breast cancer.* Cancer research, 2005. **65**(16): p. 7065-70.

34. Mattie, M.D., et al., *Optimized high-throughput microRNA expression profiling provides novel biomarker assessment of clinical prostate and breast cancer biopsies.* Molecular cancer, 2006. **5**: p. 24.

35. Hafner, M., et al., *Barcoded cDNA library preparation for small RNA profiling by next-generation sequencing.* Methods, 2012. **58**(2): p. 164-70.

36. Rossi, S. and G.A. Calin, *Bioinformatics, Non-coding RNAs and Its Possible Application in Personalized Medicine.* Advances in experimental medicine and biology, 2013. **774**: p. 21-37.

37. Kala, R., et al., *MicroRNAs: an emerging science in cancer epigenetics.* Journal of clinical bioinformatics, 2013. **3**(1): p. 6.

38. Blenkiron, C., et al., *MicroRNA expression profiling of human breast cancer identifies new markers of tumor subtype.* Genome biology, 2007. **8**(10): p. R214.

39. Bartels, C.L. and G.J. Tsongalis, *MicroRNAs: novel biomarkers for human cancer.* Clinical chemistry, 2009. **55**(4): p. 623-31.

40. Jansson, M.D. and A.H. Lund, *MicroRNA and cancer.* Molecular oncology, 2012. **6**(6): p. 590-610.

41. Yoon, S. and G. De Micheli, *Prediction of regulatory modules comprising microRNAs and target genes.* Bioinformatics, 2005. **21 Suppl 2**: p. ii93-100.

42. Tang, J., A. Ahmad, and F.H. Sarkar, *The Role of MicroRNAs in Breast Cancer Migration, Invasion and Metastasis.* International journal of molecular sciences, 2012. **13**(10): p. 13414-37.

43. Lu, J., et al., *MicroRNA expression profiles classify human cancers.* Nature, 2005. **435**(7043): p. 834-8.

44. miRbase. *miRNA Database*. [cited 2013; Available from: http://www.mirbase.org/.

45. Zhu, W., et al., *Circulating microRNAs in breast cancer and healthy subjects.* BMC research notes, 2009. **2**: p. 89.

46. TCGA. 2013; Available from: https://tcga-data.nci.nih.gov/tcga/.

47. Bensen, J.T., et al., *Association of germline microRNA SNPs in pre-miRNA flanking region and breast cancer risk and survival: the Carolina Breast Cancer Study.* Cancer causes & control : CCC, 2013. **24**(6): p. 1099-109.

48. Wei, J.J. and P. Soteropoulos, *MicroRNA: a new tool for biomedical risk assessment and target identification in human uterine leiomyomas.* Seminars in reproductive medicine, 2008. **26**(6): p. 515-21.

49. Lian, H., L. Wang, and J. Zhang, *Increased risk of breast cancer associated with CC genotype of Has-miR-146a Rs2910164 polymorphism in Europeans.* PloS one, 2012. **7**(2): p. e31615.

50. Sung, H., et al., *Common genetic polymorphisms of microRNA biogenesis pathway genes and breast cancer survival.* BMC cancer, 2012. **12**: p. 195.

51. Dweep, H., Sticht, C., Pandey, P., Gretz, N., *miRWalk - database: prediction of possible miRNA binding sites by "walking" the genes of 3 genomes*, in *Journal of Biomedical Informatics*2011. p. 839-7.

52. Yang, D., et al., *Integrated analyses identify a master microRNA regulatory network for the mesenchymal subtype in serous ovarian cancer.* Cancer cell, 2013. **23**(2): p. 186-99.

53. Chia-Hsien Lee, W.-H.K., Chen-Ching Lin, Yen-Jen Oyang, Hsuan-Cheng Huang, Hsueh-Fen Juan, *MicroRNA-Regulated Protein-Protein Interaction Networks and Their Functions in Breast Cancer.* International journal of molecular sciences, 2013. **14**: p. 11560-11606.

54. Dehal, A., A. Abbas, and S. Johna, *Racial disparities in clinical presentation, surgical treatment and in-hospital outcomes of women with breast cancer: analysis of nationwide inpatient sample database.* Breast cancer research and treatment, 2013. **139**(2): p. 561-9.

55. Ma, H., et al., *Mortality risk of black women and white women with invasive breast cancer by hormone receptors, HER2, and p53 status.* BMC cancer, 2013. **13**: p. 225.

56. Giacomini, C.P., et al., *Breakpoint analysis of transcriptional and genomic profiles uncovers novel gene fusions spanning multiple human cancer types.* PLoS genetics, 2013. **9**(4): p. e1003464.

57. Di Leva, G. and C.M. Croce, *The Role of microRNAs in the Tumorigenesis of Ovarian Cancer.* Frontiers in oncology, 2013. **3**: p. 153.

58. Wang, W., et al., *Expression and role of miR-34a in bladder cancer.* Indian journal of biochemistry & biophysics, 2013. **50**(2): p. 87-92.

59. Guo, W.G., et al., *Bioinformatics analyses combined microarray identify the desregulated MicroRNAs in lung cancer.* European review for medical and pharmacological sciences, 2013. **17**(11): p. 1509-16.

60. Tang, S., et al., *Sweating the Small Stuff: MicroRNAs and Genetic Changes Define Pancreatic Cancer.* Pancreas, 2013. **42**(5): p. 740-59.

61. Yong, F.L., C.W. Law, and C.W. Wang, *Potentiality of a triple microRNA classifier: miR-193a-3p, miR-23a and miR-338-5p for early detection of colorectal cancer.* BMC cancer, 2013. **13**(1): p. 280.

62. Gocze, K., et al., *Unique MicroRNA Expression Profiles in Cervical Cancer.* Anticancer research, 2013. **33**(6): p. 2561-7.

63. Walter, B.A., et al., *Comprehensive microRNA Profiling of Prostate Cancer.* Journal of Cancer, 2013. **4**(5): p. 350-7.

64. Ju, J., J. Jiang, and A. Fesler, *miRNA: the new frontier in cancer medicine.* Future medicinal chemistry, 2013. **5**(9): p. 983-5.

65. Dvinge, H., et al., *The shaping and functional consequences of the microRNA landscape in breast cancer.* Nature, 2013. **497**(7449): p. 378-82.

66. Zhou, F., et al., *MicroRNA and histopathological characterization of pure mucinous breast carcinoma.* Cancer biology & medicine, 2013. **10**(1): p. 22-7.

67. Wang, J., et al., *Overexpressions of MicroRNA-9 and MicroRNA-200c in Human Breast Cancers Are Associated with Lymph Node Metastasis.* Cancer biotherapy & radiopharmaceuticals, 2013.

68. Wu, X., et al., *De novo sequencing of circulating miRNAs identifies novel markers predicting clinical outcome of locally advanced breast cancer.* Journal of translational medicine, 2012. **10**: p. 42.

69. Farazi, T.A., et al., *MicroRNA sequence and expression analysis in breast tumors by deep sequencing.* Cancer research, 2011. **71**(13): p. 4443-53.

70. Lowery, A.J., et al., *MicroRNA signatures predict oestrogen receptor, progesterone receptor and HER2/neu receptor status in breast cancer.* Breast cancer research : BCR, 2009. **11**(3): p. R27.

71. Luo, D., et al., *A systematic evaluation of miRNA:mRNA interactions involved in the migration and invasion of breast cancer cells.* Journal of translational medicine, 2013. **11**: p. 57.

72. Manuel Mayr, A.Z., Peter Willeit, Johann Willeit and Stefan Kiechl, *MicroRNAs Within the Continuum of Postgenomics Biomarker Discovery.* Arteriosclerosis, Thrombosis, and Vascular Biology, 2013. **33**: p. 206-214.

73. Asaga, S. and D.S. Hoon, *Direct Serum Assay for MicroRNA in Cancer Patients.* Methods in molecular biology, 2013. **1024**: p. 147-55.

74. Eichelser, C., et al., *Deregulated Serum Concentrations of Circulating Cell-Free MicroRNAs miR-17, miR-34a, miR-155, and miR-373 in Human Breast Cancer Development and Progression.* Clinical chemistry, 2013.

75. Godfrey, A.C., et al., *Serum microRNA expression as an early marker for breast cancer risk in prospectively collected samples from the Sister Study cohort.* Breast cancer research : BCR, 2013. **15**(3): p. R42.

76. Scott, G.K., et al., *Coordinate suppression of ERBB2 and ERBB3 by enforced expression of micro-RNA miR-125a or miR-125b.* The Journal of biological chemistry, 2007. **282**(2): p. 1479-86.

77. Li, Q., et al., *Downregulation of miR-140 promotes cancer stem cell formation in basal-like early stage breast cancer.* Oncogene, 2013.

78. Nassirpour, R., et al., *miR-221 Promotes Tumorigenesis in Human Triple Negative Breast Cancer Cells.* PloS one, 2013. **8**(4): p. e62170.

79. Tang, W., et al., *miR-27a regulates endothelial differentiation of breast cancer stem like cells.* Oncogene, 2013.

80. Hanahan, D. and R.A. Weinberg, *Hallmarks of cancer: the next generation.* Cell, 2011. **144**(5): p. 646-74.

81. Li, Y., F. Hong, and Z. Yu, *Decreased expression of microRNA-206 in breast cancer and its association with disease characteristics and patient survival.* The Journal of international medical research, 2013. **41**(3): p. 596-602.

82. Wang, H.J., et al., *MiR-125b regulates side population in breast cancer and confers a chemoresistant phenotype.* Journal of cellular biochemistry, 2013.

83. Kutanzi, K.R., et al., *MicroRNA-mediated drug resistance in breast cancer.* Clinical epigenetics, 2011. **2**(2): p. 171-185.

84. Smith, R.A., et al., *A genetic variant located in miR-423 is associated with reduced breast cancer risk.* Cancer genomics & proteomics, 2012. **9**(3): p. 115-8.

85. Linhares, J.J., et al., *Evaluation of single nucleotide polymorphisms in microRNAs (hsa-miR-196a2 rs11614913 C/T) from Brazilian women with breast cancer.* BMC medical genetics, 2012. **13**: p. 119.

86. Wang, L., et al., *A miRNA binding site single-nucleotide polymorphism in the 3'-UTR region of the IL23R gene is associated with breast cancer.* PloS one, 2012. **7**(12): p. e49823.

87. Sung, H., et al., *Common genetic variants in the microRNA biogenesis pathway are not associated with breast cancer risk in Asian women.* Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology, 2012. **21**(8): p. 1385-7.

88. *Illumina Sequencer*. [cited 2013; Available from: http://www.illumina.com/.

89. *TopHat*. [cited 2013; Available from: http://tophat.cbcb.umd.edu/.

90. Kolde, R., *Pretty Heatmaps*, 2013.

91. Xie, B., et al., *miRCancer: a microRNA–cancer association database constructed by text mining on literature.* Bioinformatics, 2013. **29**(5): p. 638-644.

92. Trevor Hastie, R.T., Balasubramanian Narasimhan and Gilbert Chu, *Impute: Imputation for microarray data*, 2014.

93. MathWorks, *Matlab Student Version*, 2012.

94. Avril Coghlan, R.R. *Using R for Multivariate Analysis*. 2010 [cited 2014; Available from: http://little-book-of-r-for-multivariate-analysis.readthedocs.org/en/latest/src/multivariateanalysis.html.

95. I. S. Vlachos, N.K., T. Vergoulis, G. Georgakilas, M. Reczko, M. Maragkakis, M. D. Paraskevopoulou, K. Prionidis, T. Dalamagas, A. G. Hatzigeorgiou, *DIANA miRPath v.2.0: investigating the combinatorial effect of microRNAs in pathways Nucleic Acids Research 2012*, 2012.

96. Chen, J., et al., *miR-127 regulates cell proliferation and senescence by targeting BCL6.* PloS one, 2013. **8**(11): p. e80266.

97. Khan, S., et al., *miR-379 regulates cyclin B1 expression and is decreased in breast cancer.* PloS one, 2013. **8**(7): p. e68753.

98. Yang, J., et al., *MicroRNA-19a-3p inhibits breast cancer progression and metastasis by inducing macrophage polarization through downregulated expression of Fra-1 proto-oncogene.* Oncogene, 2013.

99. Zhu, N., et al., *Endothelial-specific intron-derived miR-126 is down-regulated in human breast cancer and targets both VEGFA and PIK3R2.* Mol Cell Biochem, 2011. **351**(1-2): p. 157-64.

100. Schwarzenbach, H., et al., *Diagnostic potential of PTEN-targeting miR-214 in the blood of breast cancer patients.* Breast cancer research and treatment, 2012. **134**(3): p. 933-41.

101. Cheng, C.W., et al., *MicroRNA-30a inhibits cell migration and invasion by downregulating vimentin expression and is a potential prognostic marker in breast cancer.* Breast cancer research and treatment, 2012. **134**(3): p. 1081-93.

102. Zeng, R.C., et al., *Down-regulation of miRNA-30a in human plasma is a novel marker for breast cancer.* Med Oncol, 2013. **30**(1): p. 013-0477.

103. Wu, Q., et al., *Analysis of serum genome-wide microRNAs for breast cancer detection.* Clinica chimica acta; international journal of clinical chemistry, 2012. **413**(13-14): p. 1058-65.

104. Xu, Q., et al., *A regulatory circuit of miR-148a/152 and DNMT1 in modulating cell transformation and tumor angiogenesis through IGF-IR and IRS1.* J Mol Cell Biol, 2013. **5**(1): p. 3-13.

105. Liu, X.X., et al., *MicroRNA-26b is underexpressed in human breast cancer and induces cell apoptosis by targeting SLC7A11.* FEBS Lett, 2011. **585**(9): p. 1363-7.

106. Li, J., et al., *MiRNA-26b inhibits proliferation by targeting PTGS2 in breast cancer.* Cancer Cell Int, 2013. **13**(1): p. 7.

107. Korpal, M., et al., *Direct targeting of Sec23a by miR-200s influences cancer cell secretome and promotes metastatic colonization.* Nat Med, 2011. **17**(9): p. 1101-8.

108. Li, X., et al., *c-MYC-regulated miR-23a/24-2/27a cluster promotes mammary carcinoma cell invasion and hepatic metastasis by targeting Sprouty2.* The Journal of biological chemistry, 2013. **288**(25): p. 18121-33.

109. Yan, L.X., et al., *MicroRNA miR-21 overexpression in human breast cancer is associated with advanced clinical stage, lymph node metastasis and patient poor prognosis.* Rna, 2008. **14**(11): p. 2348-60.

110. Ru, P., et al., *Anti-miR-203 Upregulates SOCS3 Expression in Breast Cancer Cells and Enhances Cisplatin Chemosensitivity.* Genes Cancer, 2011. **2**(7): p. 720-7.

111. Cai, J., et al., *MicroRNA-374a activates Wnt/beta-catenin signaling to promote breast cancer metastasis.* J Clin Invest, 2013. **123**(2): p. 566-79.

112. Rafii, A., et al., *Where cancer genomics should go next: a clinician's perspective.* Hum Mol Genet, 2014.